Title	Design and evaluation of a credit decision system using mobile data in agricultural micro-finance
Sub Title	
Author	Simumba, Naomi(Kotake, Naohiko) 神武, 直彦
Publisher	慶應義塾大学大学院システムデザイン・マネジメント研究科
Publication year	2018
Jtitle	
JaLC DOI	
Abstract	
Notes	修士学位論文. 2018年度システムエンジニアリング学 第272号
Genre	Thesis or Dissertation
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=K040002001-00002018-0001

Master's Dissertation 2018

Design and Evaluation of a Credit Decision System Using Mobile Data in Agricultural Micro-Finance

Naomi Simumba

(Student ID Number: 81634517)

Supervisor: Prof. Naohiko Kohtake

September 2018

Graduate School of System Design and Management

Keio University

Major in System Design and Management

ABSTRACT

Student Identification Number	81634517	Name	Simumba, Naomi

Title

Design and Evaluation of a Credit Decision System Using Mobile Data in Agricultural Micro-Finance

Abstract

Financial exclusion has a major socio-economic impact on poor and unbanked smallholder farmers. They face challenges accessing credit facilities to fund their farming activities because they lack the financial history data required by financial institutions to create credit scores for credit risk evaluations. Financial institutions may also face challenges with regards to collecting data from farmers who may live in areas where physical access is a difficult.

To overcome these issues, non-financial data, i.e. data not related to a person's financial activities, has been proposed to develop credit scores for financially excluded persons including smallholder farmers. Some types of non-financial data provide additional advantages. Mobile data, data collected through a mobile phone, is an example of non-financial data that can be used to simply data collection from those in far flung areas thus eliminating the need for physical access during data collection. Other suitable tools for the collection of non-financial data include satellites which can gather useful environmental information remotely.

Although existing research has applied non-financial data for credit score development, this often relies on financial history data for the initial development of credit scoring models. A method must be developed which does not use financial data for score development. Further, to use mobile and other non-financial data, the exact types of data that should be collected must be clearly defined. Finally, appropriate data analysis methods to be used in the context of score development for smallholder farmers must be clearly identified.

This research proposes a Credit Decision System using mobile data to assist financial institutions in making credit decisions for financially excluded farmers, thus tackling these issues in the field of agricultural micro-finance. The system collects mobile data and other types of non-financial data such as satellite data. Alternative scoring factors which do not require financial data are developed by identifying stakeholders' concerns with regards to lending and evaluated using the non-financial data collected. Resulting scores are then

provided to financial institutions, allowing them to make credit decisions regarding financially excluded farmers.

The design is evaluated through the creation of a prototype using data collected by means of mobile phones and surveys from farmers in rural Cambodia. Stakeholders' concerns are identified through a workshop and interviews. To address these concerns, three alternative scoring factors are selected: reliability, revenue, and interaction scores. It is found that the types of data that should be collected to evaluate these alternative scores are personal information and farm activity information. The former may include factors such as age, gender, and length of residence, while the latter may include types of crops being cultivated, use of irrigation, and harvest methods. The alternative scores are developed by building classification models, with 80% of the data collected, using multiple logistic regression and support vector machine algorithms. These algorithms are compared by Area Under the Receiver Operating Characteristics Curve values and found to be similar with support vector machines outperforming multiple logistic regression.

Prototype Verification is done chiefly through model testing. This involves using the remaining 20% of the original data to determine accuracy of the scoring models. Additional considerations such as complexity and cost are made to determine the most suitable model in this context. Prototype validation is conducted through interviews with farmers and financial institutions, as well as a workshop with personnel of a mobile data collection company. These validation exercises show that mobile data can be a convenient source of information for financial institutions making lending decisions provided the right types of data are collected. Further, stakeholders' concerns can be used to develop appropriate alternative scoring factors independent of financial data.

The system proposed in this research can be used to improve access to credit services for smallholder farmers, thus increasing their livelihood and standard of living. It can also be used for other financially excluded groups to reduce the impact of financial exclusion.

Keywords

Mobile Data, Micro-finance, Agriculture, Credit scoring, Machine Learning

TABLE OF CONTENTS

LIST OF IMAGES

TABLE OF FIGURES

LIST OF TABLES

DEFINITION OF KEY TERMS

The following are some of the key terms used and their meaning in the context of this thesis paper:

Financial exclusion – lack of access to useful and affordable financial products and services that meet needs of the user and are delivered in a responsible and sustainable way [1]

Smallholder farmers - farmers who manage areas of varying size which tend to be smaller, sell part of their produce, and may be semi-subsistent

Non-financial data – data not related to a person's financial activities

Credit scoring factors – evaluation criteria used to make a credit decision

Credits scoring – a set of decision models and their underlying techniques that aid lenders in granting consumer credit

Financial institution – an institution that provides financial services to individuals including credit services

Data collection company – a company whose business includes collection of data from individuals

Satellite data collection company - a company which collects and offers satellite data as part of its business

Mobile data – data that is collected through a mobile phone

Survey data – data collected through a paper-based survey

Buddies – people who use the mobile application of the data collection company used for prototyping (Agribuddy LTD) directly to collect data from farmers on behalf of the company

ACKNOWLEDGEMENTS

I would like to show my appreciation to Professor Kohtake, my primary supervisor, whose support and encouragement has been invaluable throughout my master's program. Kohtake-sensei went above and beyond in not only providing research guidance, but also the opportunity to participating in the research project that made possible the collection of data used in this research.

I would also like to express gratitude to the members of Kohtake laboratory, both past and present, who have repeatedly given great insights and feedback that have helped improve my research. My special gratitude goes to Mr Okami, who gave guidance and support with regards to the data analysis processes used here. I would further like to thank Mr Kodaka and Mr Nishino who also provided additional reviews.

My thanks goes to Professor Ogi, my secondary supervisor, for his helpful critique and advise which allowed me to further clarify the key research points of this thesis. I would like to thank the entire team at the G-Spase Human Resource Development Program for their many timely comments.

Acknowledgement is also made of Mr. Kitaura, Chief Executive Officer, and the rest of the team at Agribuddy Ltd (https://www.agribuddy.com/) for their kind assistance in providing the data used in this research. In addition, their insights into the farming activities in the prototyping areas have been highly useful in constructing the basis of this research.

I. INTRODUCTION

In 2016, 815 million men, women and children (11% of the world's population) went hungry [2]. Food, a basic human need, is something many in the developed world take for granted. Yet as the world becomes a more uncertain place to live in, it becomes increasingly difficult for many to obtain. The fact remains that man cannot survive without food, and food would not be produced without agriculture in its various forms. However, agriculture is more than simply a source of food. It is also a key economic activity, providing 3.8% of global Gross Domestic Product (GDP) [3] and employment for 30% of the world's population [4]. There are 2.5 billion people globally involved in smallholder agriculture on 500 million small farms [5].

Smallholder farmers form a significant portion of the agricultural sector which supports entire economies in parts of the developing world. They also supply 80% of food consumed [6] and are responsible for much of the food security in these parts of the world [7]. Beyond this, smallholder agriculture has been identified as a key tool for poverty reduction [8]. However, smallholder farmers are finding it increasingly difficult to remain competitive. Global changes such as climate change have contributed to make smallholder farmers one of the most vulnerable groups on the planet [5]. Growing competition from the global market [6] and climate change [9] threaten their livelihoods.

Given the importance of the smallholder farmer to the world's economy and food security, it is vital that they receive the support needed to maintain their status as valuable contributors to the global society. Financing of smallholder farmers, or agricultural micro-finance for smallholder farmers, is one of the major challenges in this regard. Financial resources have a huge impact on a farmer's productivity [10], financing the hiring of labour and the purchase of equipment and agriculture inputs which may be used to maximize production. As many smallholder farmers come from poor households [6], the lack of credit facilities leads to limited productivity which poses a serious problem for farmers and the economies in which they participate. According to the United States Agency for International Development (USAID), there is an estimated US$430 billion shortfall in financing the needs of smallholder farmers [11]. Part of this shortfall may be attributed to financial exclusion which limits access to credit services. Financial exclusion is the lack access to "…useful and affordable financial products and services that meet their needs – transactions, payments, savings, credit and insurance – delivered in a responsible and sustainable way…" [1]. It is highly prevalent, with the World Bank approximating that 44% percent of adults globally do not have an account with formal

financial institutions or mobile money lenders [12], one of the indicators of financial exclusion. Some of the factors linked to financial exclusion are low financial literacy, poverty, and long distances to financial institutions [13]. It should be noted that many smallholder farmers tend to live in rural areas where these issues are predominant.

Moving forward, the question of how to support financially excluded smallholder farmers by giving them access to credit services is critical. The activities conducted for this research will attempt to answer this question by designing a Credit Decision System that can be utilized by lending institutions providing credit services to smallholder farmers. Several questions will be asked. **What data should be collected and how? What scoring factors should be used to evaluate borrowers? Which tools should be used to carry out these evaluations?** Data collection methods and appropriate data types will be explored. A method of developing scoring factors will be proposed to evaluate borrowers. Various data analysis tools will be considered.

As with all social issues, financial exclusion is multifaceted in nature; involving technical, environmental, and social aspects. All of this adds to the complexity of the required solution. A systematic method of identifying stakeholder concerns then collecting and analysing data based on these concerns is needed to design a suitable result. Systems engineering, and systems thinking as part of it, incorporates the entire context of the problem in the design process making it the ideal approach for creating solutions to solve social problems using technology. Various techniques will be used to understand the complex interaction of issues; Customer Value Chain Analysis (CVCA), a method used to analyse the flow of money and other value between stakeholders, and Causal Loop Diagrams, which analyse the interactions between related factors, among others. This broader perspective with multiple viewpoints is one of the core ideas of systems thinking and will prove invaluable in the design and evaluation of the system.

I.1 BASIC CONCEPTS

I.1.1 AGRICULTURE

Agriculture is major economic activity in many parts of the world. According to the International Labour Organisation (ILO), 1.3 billion people rely on agricultural activities for their livelihood [14]. It is also a vital economic contributor. Its economic importance varies by country. Image 1 shows the contribution agriculture makes to global Gross Domestic Product

(GDP). In developing countries especially, agriculture is a chief employer. Its global significance makes the field of agriculture highly relevant as a target of research.

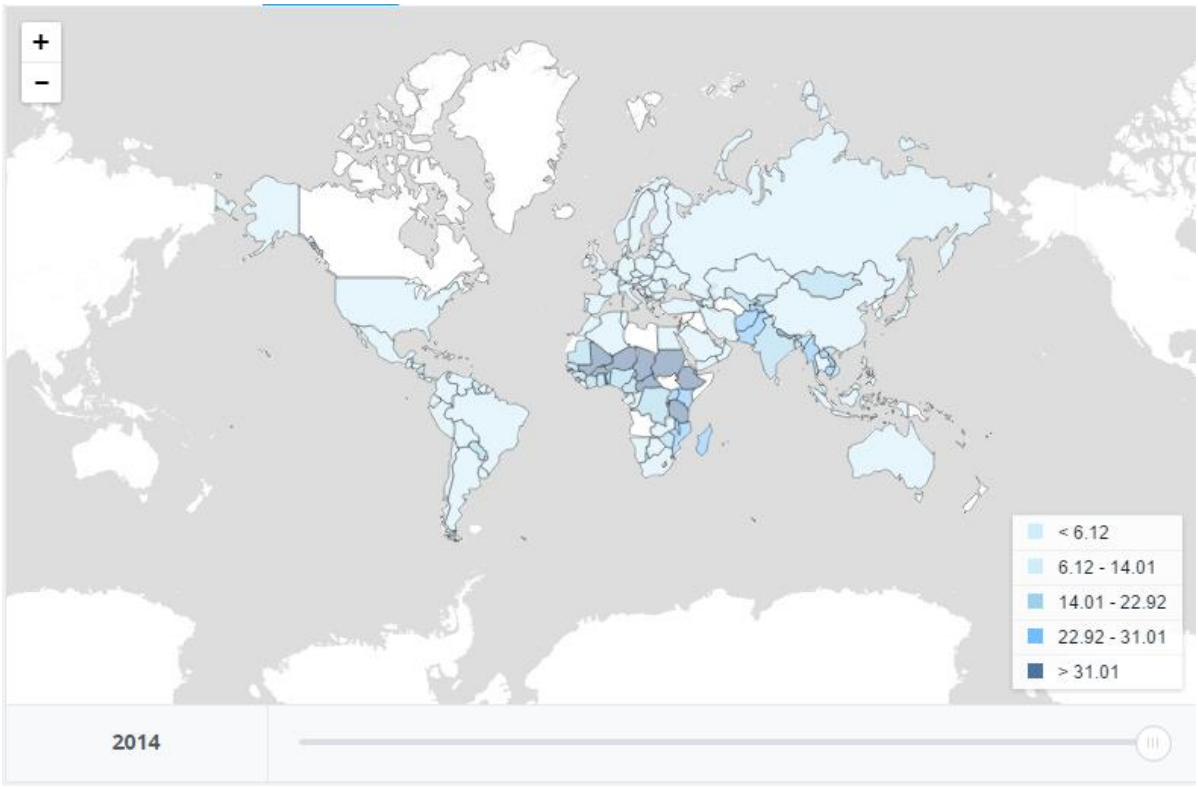


Image 1 world map showing agriculture value added to GDP, 2014 [15]

I.1.2 SMALLHOLDER FARMERS

Categorizing farmers is a difficult task since there are many variations in terms of their land size, productivity, resources available, labour and other factors. Location also factors into this categorization. For instance, 500 hectares is considered a small holding in Australia, while this may not be the case in parts of Europe. The Food and Agriculture Organisation classifies farms by several metrics. These include farm management, economics analysis, financial value, family labour effort, bio-mechanical energy, and water consumption. By this classification, it is possible to obtain six basic farm types [16]. The targets of this research are type 2 and type 3 farmers who may who may or may not be semi-subsistent, but who sell part of theirs produced for income.

i. **Type 1. Small subsistence-oriented family farms**: these are farmers who raise one or several crops and animals for consumption by the family
ii. **Type 2. Small semi-subsistence or part-commercial family farms:** these possess land ranging from small one or two-hectare parcels to much larger parcels (20- to 30-hectares). They grow mixed produce which may include several crop types and/or

livestock. Their produce is intended first for consumption by the family, and second to generate additional income for goods that cannot be grown on the farm such as clothes and farm inputs.

iii. **Type 3. Small independent specialized family farms:** these are farmers who are specialized in growing particular crops or livestock. They may be semi-subsistence, using part of their produce for family consumption, or completely commercial.

iv. **Type 4. Small dependent specialized family farms:** these farmers are similar to type 3. However, because they are dependent on other organisations for their income generation, they have less power in choosing the types of crops they grow and how the grow them. This may be due to debt to agro-industrial corporations, tenancy on farms owned by other parties, or integration into a larger farming/processing system, government directives, or lack of alternative markets.

v. **Type 5. Large commercial family farms:** they are similar to commercial farming estates except the main beneficiaries are families. Many produce single crop type or multiple crops. Often specializing.

vi. **Type 6. Commercial estates:** these range from 200 to 2000 hectares although they can be bigger or smaller. They are usually mono-crop and with hired management and absentee ownership. Having close relationships with their buyers, they have strong marketing operations and are treated entirely as resource generating enterprises.

I.1.3 PRODUCTIVITY

The productivity of a farmers is influenced by a wide range of factors. Among these is access to financing which can allow a farmer to fund their farming activities, buy irrigation equipment, hire more workers, buy farming implements such as seed, or farming equipment which can all allow a farm to cultivate larger tracts of land.

I.2 PROBLEM ANALYSIS

The smallholder farmer's lack of access to credit services is a multidimensional problem whose root lies in financial exclusion. Financial exclusion results in limited or missing financial records, including credit history. A credit history is required by credit bureaus and financial institutions to evaluate risk by creating credit scores before making lending decisions. Since, the financially excluded tend to be low income members of society living in less developed parts of the world [17], these institutions are often unable to provide them with credit services for two reasons. Firstly, they cannot easily collect data from them and, secondly, there are no

established methods of analysing the data that can be collected. As a result, credit bureaus tend to develop credit scores for the more educated members of society with higher income [18] which limits the market for credit services available to lending institutions. An analysis of the problem facing financial institutions is given in figure 1. Financial institutions looking for ways increase their profits may do so by extending their services to a new market, namely financially excluded smallholder farmers. However, lending to this market segment poses a risk for the lender because it is not possible to evaluate credit risk using traditional methods which require financial records from the borrower. These may not be not always available for a wide range of reasons. As a result, financial institutions miss out on the financial gains from the large unserved market made up of low income persons with limited or inexistent credit histories. Despite this, a study by LexisNexis, an institution dealing in research and risk management, estimated 2 in 3 customers with little or no credit history are in fact low risk borrowers [19], meaning that financial institutions could extend credit to this segment with less risk than assumed.

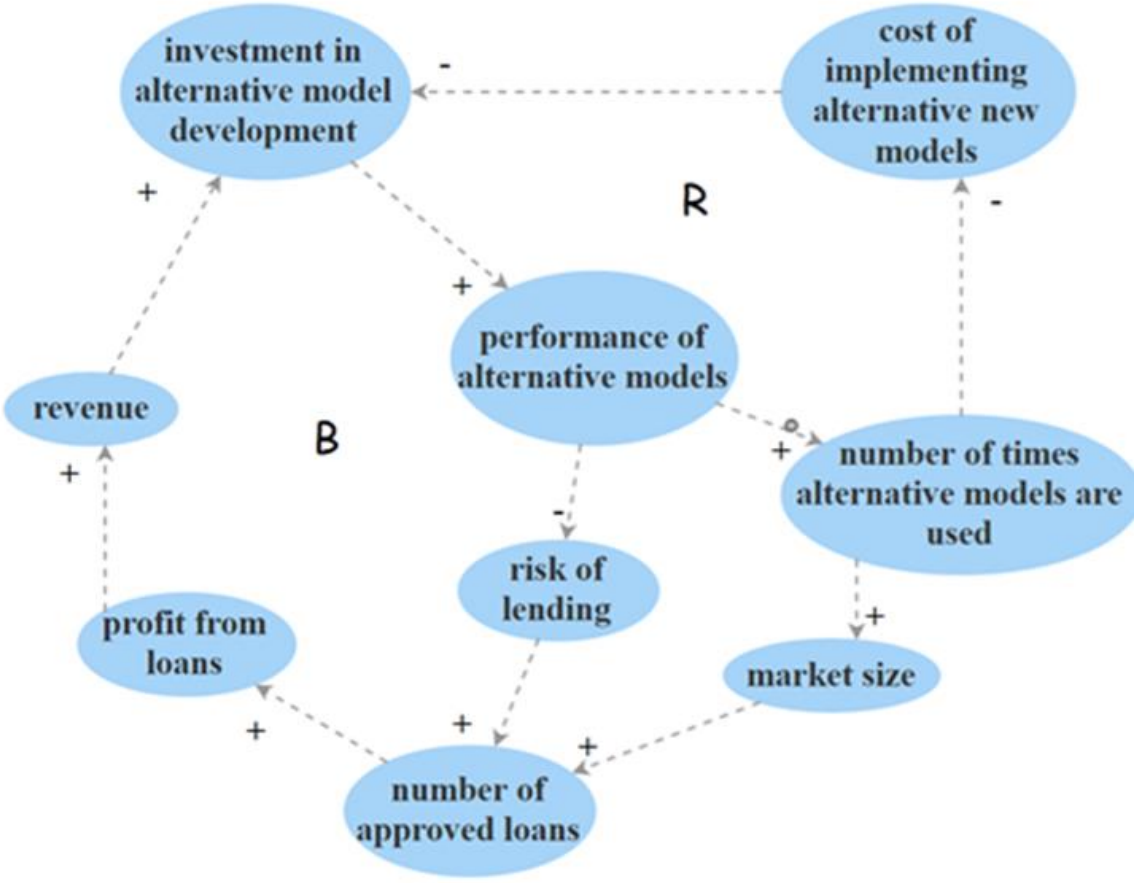


Figure 1causal loop analysis for financial institutions

Various factors affect the smallholder farmer's ability to generate revenue from their farming activities. Some of these are shown in figure 2. Financial exclusion itself has many causes. These may include long distances reducing physical access to financial institutions, illiteracy, or lack of funds among others [17]. This makes collection of data for credit risk evaluations difficult. For individuals without credit histories due to the reasons stated above, it may be a challenge to obtain access to credit facilities at financial institutions. This in turn leads to insufficient financial resources for agricultural inputs, ending in low income. A vicious cycle, shown in figure 3, is formed wherein the farmer is unable to obtain the resources needed to increase productivity and is therefore unable to increase productivity to obtain more resources.

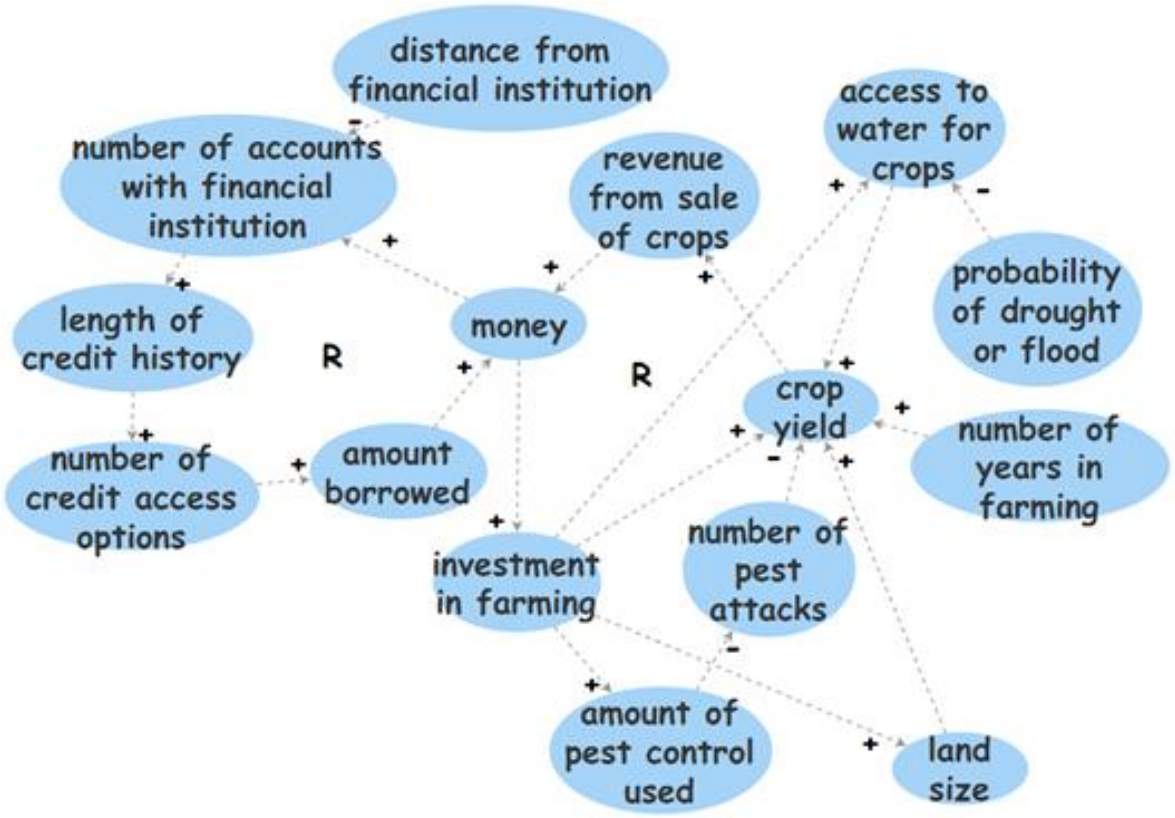


Figure 2 causal loop diagram for financially excluded farmers

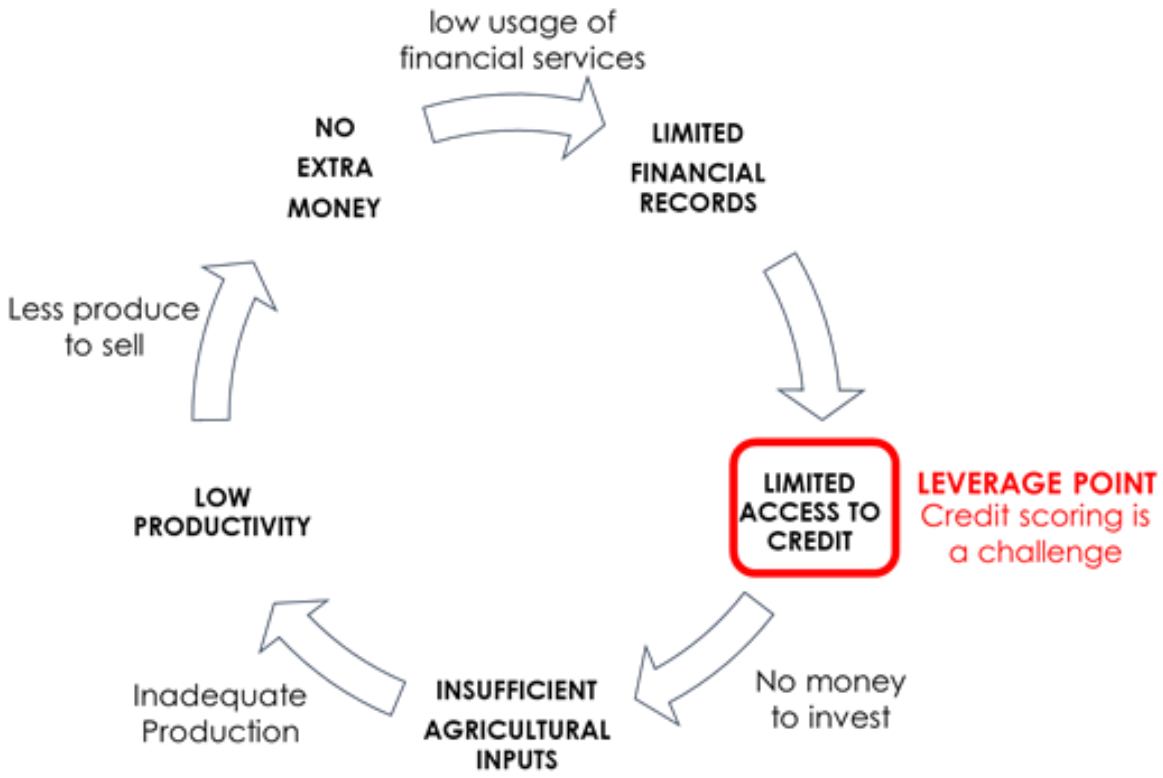


Figure 3 cycle of loop of farmer's funding problem

To break this cycle, we must ask: **How might we increase the number of credit service options available to smallholder farmers while managing the risk incurred by financial institutions?** A probable solution to this is a new system of assessing the risk of lending to smallholder farmers by financial institutions. The problems that must be solved by this system can be summarised as shown in table 1.

Table 1summary of problems

PROBLEMS
Data collection is challenging due to factors such as long distance. In addition, appropriate non-financial data is not known
Suitable data analysis methods for evaluation of scoring factors in this context are not known
Scoring factors typically require financial data

I.3 RESEARCH GOAL

The goal of this research is to design a Credit Decision System which can be used by financial institutions to provide credit services to financially excluded smallholder farmers.

I.4 RESEARCH SCOPE

Although there are many avenues to support smallholder farmers' agricultural activities, the focus of this research is on improving access to credit service, since this is a leverage point in the problem analysis. Therefore, the focus here will be on the factors that affect the design of the Credit Decision System.

I.5 RESEARCH OBJECTIVES

The issue of credit access for financially excluded farmers is a critical social issue which can be solved by designing a new credit decision system for use by financial institutions in assessing risk of lending to smallholder farmers. To do this, the following three interrelated challenges must be tackled. Because these issues are interrelated, a holistic approach must be taken to solve them. This solution is multidisciplinary, merging the fields of data analysis, mobile applications, agriculture, as well as credit scoring.

I.5.1 NON-FINANCIAL DATA (E.G. MOBILE DATA)

As stated earlier, long distances to financial institutions is one of the reasons for financial exclusion. Because financially excluded persons often reside in areas where physical access to financial institutions is a challenge, data collection tools that do not require such access must be used to collect data from potential borrowers (farmers). Further, non-financial data (alternative data) must be used since financial data is not available for this demographic. Therefore, this research will identify suitable methods of collecting non-financial data that do not require physical access. In addition, the exact types of data that must be collected to generate scores will be identified.

I.5.2 DATA ANALYSIS

Various data analysis tools have been used to analyse data for credit decision making. Suitable data analysis tools must be selected for this application; tools which balance the various requirements of the system. While many different machine learning algorithms are in use, it is critical that the tools used for this system are able to operate well in the context of financing smallholder farmers where issues such cost of computation and complexity will affect their expediency. This is yet another objective this research will attempt to fulfil.

I.5.3 ALTERNATIVE SCORING FACTORS

The credit scoring factors typically used for credit scoring cannot apply here since they depend on credit histories which this group of borrowers do not possess. As a result, this research will

propose a method of designing relevant scoring factors that describe the risks of lending to this group of borrowers.

This information is summarized in table 2 below:

Table 2 aligning research objectives with identified problems

PROBLEM	OBJECTIVE
Data collection is challenging due to factors such as long distance. In addition, appropriate non-financial data is not known	Identify **simplified data collection methods** and which **types of data** are needed
Suitable data analysis methods for evaluation of scoring factors in this context are not known	Determine most suitable data **analysis method**
Scoring factors require financial data	Design **alternative scoring factors** that do not require financial data

I.6 ORIGINALITY AND FEATURES OF RESEARCH

Although other researchers have tackled the issue of credit access for financially excluded farmers, this past research required the use of financial data to create scoring models. This means that the scoring factors developed required financial data. However, there is a lack of research showing what should be done when financial data cannot be obtained. In this research thesis, this question is answered by proposing a method of generating alternative scoring factors. Further, because the development of scoring factors is heavily dependent on data collection and analysis methods, these are also considered as part of this research. The details of how each of these aspects will be tackled are given below.

I.6.1 NON-FINANCIAL DATA (E.G. MOBILE DATA):

Collection and analysis of non-financial data is needed. This type of information is also known as alternative data. The basic concept of using non-financial data for credit risk analysis lies in the fact that many of the unbanked use other non-financial services from which a wealth of information can be collected to form the basis of credit decisions. Examples include social media information and usage of utilities among others [20]. This research proposes the use of mobile phones and satellites for data collection. Additionally, a data collection method that does not require physical access must be used to over overcome the problems of long distances which currently prevent farmers from accessing financial services.

Mobile data (i.e. data collected through a mobile phone) is a good fit in this case since data can be collected quickly. Further, many different types of data can be collected from the farmer that may be relevant for credit decision evaluations. Examples include personal information such as age, and number of family members. Modern mobile phones, with their wide range of features, provide a rich source of information. Coupled with this is the fact that mobile phone usage has increased in the past few decades. There were an estimated 7.1 billion mobile phone subscriptions as of 2017 [21]. This trend provides insight into demographics, such as smallholder farmers, from whom data is notoriously difficult to collect. The application of **mobile data** for credit decisions has seen a corresponding growth as several firms have incorporated it into their systems [22]. Some example of the mobile data used are information on mobile bank account transactions, duration of phone calls, and so on. One of the major advantages of using mobile application data is that it provides an insight into people's behaviour and personality; characteristics which are key for credit risk assessment.

Satellite Imagery can be used to provide data on physical aspects, such as vegetation cover and hydrology of the farm area, which affect agricultural productivity. This includes soil moisture data and the vegetation index. The former provides information of the soil moisture content in various locations. Naturally, soil moisture a key component of crop growth. In addition, such data could also be used to assess drought and flood risk which could impact productivity. The vegetation index, or Normalized Difference Vegetation Index (NDVI), is a metric that is commonly used to determine vegetation cover. Varying from -1 to 1, the lowest values tend to represent water, then rocks and other barren areas. Meanwhile higher values indicate the presence of healthy vegetation and as such can be used for crop yield prediction. Other methods of collecting this imagery data may also be used. One examples would be to use drones. Additional data can be collected from a variety of sources.

Open source data can be especially valuable since it can be collected without additional cost while containing a wealth of information. Some sources would include surveys carried out by governmental and non-profit organisation such as population, transport and health surveys. Social Media and News also provide good source of open data. This data can be analysed using machine learning tools such as Logistic regression and Support Vector Machines.

I.6.2 ANALYSIS METHODS

Machine learning is an approach to analysis that depends on a commuter's ability to learn from data and apply these lessons to a new data set. There are many different machine learning

methods that can be applied. Classification analysis is methods are the method of choice in credit scoring and so can be applied here. The major advantage given by machine learning is that it massive amount of data can be analysed at a single time. Further, machine learning algorithms can recognize trends in data that may not be readily apparent to humans without great effort. These factors make machine learning a suitable. There are many machine learning classification methods which apply different mathematical approaches and have different accuracies. However, the appropriate method for analysing data in this context is not known and must be identified through this research.

Geospatial data gives some relevant kind of information while also specifying its location or spatial properties. It may be analysed using different Geographical Information System (GIS) software including ArcGIS and QGIS. Such data may be collected through both satellites and mobile phones, among other methods. The analysis of the collected data can be analysed with various machine algorithms. The most suitable can be determined by comparing the performance of these models.

I.6.3 ALTERNATIVE SCORING FACTORS

Credit factors are used as evaluation criteria for credit analysis. Examples include loan default and bankruptcy risk among others. One of the desired objectives of this research is to design a process of generating credit scoring factors that can be used to evaluate the risk of lending to a farmer in this context and can be analysed using available alternative data. From the definition of credit scoring, the core value of a credit score is that it provides the lender with an idea of how much risk he/she will incur by granting credit. The selection of credit factors is context driven, relying on its intended use and the socio-economic characteristics of its target demographic. Using systems thinking would create a system that fits better into the larger context. Further, the credit factors for analysis must be based on requirements and improved upon through repeated validation and redesign. Here, the system development is determined by specific requirements in the context of Agricultural Micro-finance. Therefore, the interdisciplinary, iterative, and requirements-based nature of System Engineering [23] makes it suitable for this research. It proposes a methodical design process to fulfill requirements. This information is summarized in table 3 below.

Table 3 aligning research approaches with objectives and problems

PROBLEM	OBJECTIVE	APPROACH
Data collection is a challenge due to distances, etc. and appropriate data is not known	Identify **simplified data collection methods** and which **types of data** are needed	Use data collection methods that do not require physical access such as **mobile data**
Scoring factors require financial data	Design **alternative scoring factors** that do not have this requirement	Use **stakeholder requirements** as basis for alternative scoring factors
Suitable data analysis methods for evaluation of scoring factors in this context are not known	Determine **most suitable data analysis** method	Compare different analysis methods to choose method **most suitable for this context**

I.7 STRUCTURE OF THESIS

Section I (INTRODUCTION) gives an introduction of this research thesis with detailed subsection allocated as stated here. Basic concepts are discussed in section I.1 (BASIC CONCEPTS). In section I.2 (PROBLEM ANALYSIS), the current challenges facing smallholder farmers are explored. Causal Loop diagrams are used to show the interactions of different issues and their impact on the agricultural productivity of smallholder farmers. An analysis of these issues leads to a clear statement of the problem. The goal of this research is specified in section I.3 (RESEARCH GOAL). The scope of this research is outlined in section I.4 (RESEARCH SCOPE). Objectives are delineated in section I.5 (RESEARCH OBJECTIVES). Research features, as well originality, are stated in section I.6 (FEATURES AND ORIGINALITY OF RESEARCH). It will be shown that the research outlined in this thesis constitutes a new approach to solve a critical global problem. Research related to the objectives outlined in section I.5 will be explored in section II (RELATED WORK). Included is research concerning various data collection tools, analysis tools and credit/alternative scoring factors.

The System Engineering approach will be used to design a system appropriate to the context and stakeholder requirements in section III (PROPOSED SYSTEM DESIGN). The system design is evaluated in section IV (SYSTEM EVALUATION). Prototyping will be conducted in section IV.1 (PROTOTYPING) to show how the Credit Decision System design could be prototyped. Verification and Validation efforts are described in sections IV.2 (VERIFICATION) and IV.3 (VALIDATION). These sections outline the exercises conducted

to ensure that the prototype meets system requirements (verification) and stakeholder requirements (validation). Section V (DISCUSSION) reviews the findings with regards to each of the stated research objectives. In addition, the consideration is made of the issues that would affect a practical implementation of the Credit Decision System. A conclusion is provided in section VI (CONCLUSION), along with ideas for future work in section VII (FUTURE WORK).

II. RELATED WORK

II.1 CREDIT SCORES AND ALTERNATIVE SCORES

Credit scoring is defined as "…a set of decision models and their underlying techniques that aid lenders in granting consumer credit" [24]. This is done by assessing the risk of lending to prospective borrower. Historically, credit decisions were based the lender's knowledge of the borrower. With time, the essence of these credit decisions was distilled into a set of criteria. Known as the 5C's of credit analysis, the following were considered when making consumer lending decisions:

Character – does the borrower have good character (as judged by his/her repayment history)?

Capacity - is the borrower able to repay the loan based on their current disposable income?

Capital – what financial reserves does the borrower have?

Collateral – can the borrower commit any of their resources towards the loan?

Condition – are the market conditions sound?

In modern times, statistical and machine learning approaches have taken precedence. The goal of these approaches to credit scoring is to distinguish between "good" and "bad" borrowers. Because they are developed based on previous borrowers about whom the repayment patterns are known [24], the development of a credit scoring model typically relies on the availability of some type of financial information. For instance, the classification of "bad" may be taken to mean borrowers who have defaulted on the loan, whereas as those without default would be classed as "good". In this case, the credit score would denote the risk of default which is a commonly used evaluation criterion for credit scoring applied by several researchers [25]. Other applications of this approach have been found in mortgage scoring, fraud prevention, bankruptcy prediction [26] and even selection customers for marketing.

II.2 DATA ANALYSIS

Credit scoring is treated as a classification task with several methods of varying complexity from traditional statistical to artificial intelligence methods being used for analysis. Statistical methods such as Logistic Regression also has been outperformed by other, more sophisticated methods like random forests and gradient boosting trees in terms of accuracy [27]. More complex methods including Artificial Neural Networks, Markov Models, Support Vector

Machines and Hybrid models have also been used [28] [29]. On the other hand, comparison with Support Vector Machines proved ambiguous as logistic regression was outperformed by non-linear kernel but performed better than linear kernels [30] Other than classification accuracy, complexity cost and cost of misclassification also play a role in the selection of the final model.

II.3 NON-FINANCIAL DATA

Several studies using machine learning algorithms have developed credit scoring systems with alternative data, such as mobile data and social media data, to increase financial inclusion for those with limited or lacking credit histories [18]. **Mobile and social media data** has been shown to enhance credit scoring for consumers [31]. In their research, Weke and Ntwiga [32] established credit scores with data from M-Shwari, a mobile phone application for financial transactions with a focus on the poor and unbanked. Matsuyin showed the value of employing social media data to predict default and fraudulency [33]. Moreover, accuracy in predicting loan repayment using mobile phone use data has been proven [34]. Despite this wealth of research, there has been little work done utilizing this type of data for credit decisions to finance smallholder farmers. In classification analysis, the dependent variable is taken to be the "good" or "bad" classification and various borrowers'' characteristics such age, income level, and so on are taken as independent variables. The use of borrowers' characteristics to distinguish between "good" and "bad" loans was originally proposed by Durand in 1941. His study found relationships between loan repayments and characteristics such as length of employment and employment [35].

When dealing with smallholder farmers, additional characteristics must be factored in. Since the main source of income of such farmers is their farming activities, farming productivity is a major factor when assessing repayment capacity. Research suggests that aspects other than financial history may be used to assess the risk associated with lending to a farmer. This entails considering the various factors which impact the productivity of the farmer. Beyond financial resources, productivity is affected by several issues. The most obvious of these would be the availability of technology and natural resources [10]; for instance, land size and water availability. Such data can be obtained from high resolution satellite images and analysed using a Geographical Information System (GIS), a system designed to acquire, analyse, store, and manage data, using geographical or spatial indexing to relate the data. Modern GIS software offers powerful tools to present spatial information and conduct predictive modelling. This has applications in early warning to reduce the spread of epidemics [36]. By determining the

geographical distribution and variation of diseases, it allows policy makers to easily understand and visualise the problems in relation to the resources, and effectively target resources to those communities in need. Other applications include land suitability assessment based on multiple factors [37]. In assessing the viability of off-grid renewable energy sources in rural areas [38]. GIS has also been widely used in transport planning, analysis and modelling [39]. Yet another application is in Early Warning Systems for disaster management. Besides this, personal characteristics of farmers have been linked to productivity. For instance, research has established that there is a correlation between factors such as education, age, farming experience, and household size, and the technical efficiency of the farmer [40]. Other qualitative aspects are social influences such as a farmer's attitude [41].

A standard of collection and analysis of non-financial data for credit decision applications is yet to be set and it remains an area of great debate and research. Predominantly, previous research has focused on applications in consumer credit. Applying mobile application and satellite data to credit decisions for smallholder farmers poses a unique set of considerations since the factors affecting repayment ability are closely related to crops, land, and farm activity, among others.

III. PROPOSED SYSTEM DESIGN

III.1 CONCEPT OF OPERATION

A system is proposed where data is collected from farmers. The collection of information will be done through mobile devices by a data collection company. The information collected will include personal information about the farmers and farm location, in addition to information about farm activities such as the type of crops being grown. Additional data is collected through a satellite data collection company. The combined data is input into the system where it is used to assess each farmer. The processing of information will be carried out by the system. This information with then be provided to the financial institution to justify the provision of credit services to a borrower, i.e. make a credit decision. Beyond the provision of credit decision information to the financial institution, the system will also provide information to farmers regarding reasons for their assessment, allowing them to make improvements for the future. Through the data collection company's system, the financial institutions can benefit from an existing network to collect data. The data collection company may also benefit from the added revenue generated by providing this data for analysis. This extended operational concept showing the operation of the system in the wider context is described in figure 4 below:

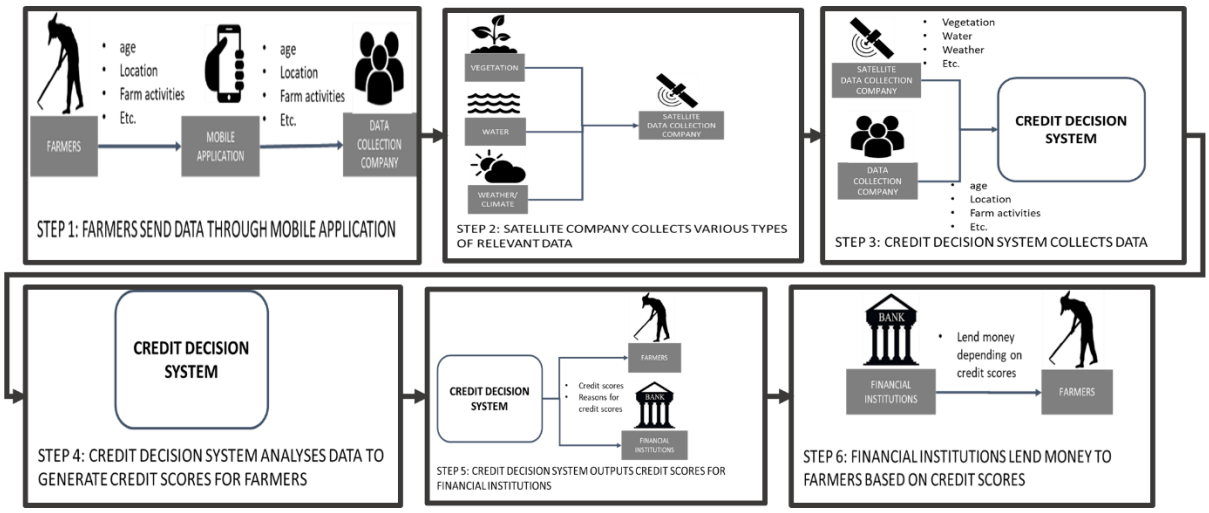


Figure 4 process of system's operation

III.2 STAKEHOLDERS

As with any system, various stakeholders interact with or have an influence on the Credit Decision System. To design a system in line with their requirements, system stakeholders and their various interests in the system are identified. While an exhaustive list of stakeholders could not be generated, it was vital that no key stakeholders were excluded from this process.

Based on the concept of operation, the key stakeholders that interact directly with and have the greatest impact on the system are financial institutions, smallholder farmers, and data collection companies (of both mobile and satellite data). Other stakeholders interacting with the system indirectly are markets and providers of logistics. Markets interact with the farmers, purchasing produce and supplying the money that will eventually be used to repay the lender. Logistics services provide transportation that allows the farmers to access markets.

i. **Financial institutions** are institutions that provide financial services to the general public. Here we use the term to refer to those that provide credit services to individuals, i.e. lending money.
ii. **Financially excluded smallholder farmers** are smallholder farmers who do not have safe, convenient access to financial services.
iii. **Data collection company** is a company in the business of collecting data from farmers through a mobile application. This may be carried out as part of their own business operations. For instance, a company that gives advice to farmers based on the data farmers input to their mobile platform.
iv. **Satellite data collection company** is a company in the business of collecting satellite imagery.

The key requirements of these stakeholders are as shown in table 4. Although the stakeholders have a wide range of needs, not all of these needs will be tackled by the system. for instance, the system itself does not provide loans to farmers.

Table 4 stakeholder requirements

STAKEHOLDER	NEEDS
Financial institution[1]	1. For each potential borrower, information on existing level of risks that can be used to make a lending decision 2. Simple output with clear reasons for each decision
Financially excluded smallholder farmers[2]	3. Simple output with clear reasons for each decision

[1] Based on interview with Mr. Kawai of Aeon Specialize Bank Cambodia PLC, see section IX.4 in appendix

[2] The farmers' needs from the system are limited to information since the system is not responsible for actual loan provision

III.3 LIFECYCLES

The farming lifecycle begins with the pre-planting stage, during which the farmer prepares for planting. Preparation activities may include clearing land, obtaining funds for agricultural input, etc. During the planting stage, the farmer plants crops. Next comes the growing stage when the farmer tends to his/her crops by watering, weeding, replanting, and so on. The harvesting stage occurs during the harvest season when the farmer reaps his/her crop yield. Finally, the post-harvest stage is when the farmer is able to sell his/her crops and return any borrowed funds to the lender. This process in shown in figure 5.

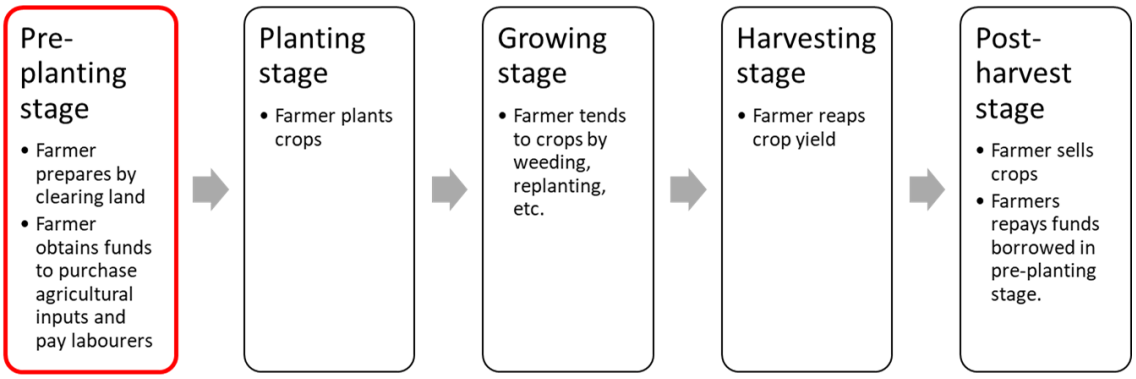


Figure 5 farming lifecycle stages

The design proposed here focuses on the operation of the system during the pre-planting stage which is when the farmer attempts to obtain funds to purchase agricultural inputs and pay labourers.

III.4 SYSTEM OPERATOR

There are three options for the system operator, each choice will impact the context diagram of the system and the value flows between stakeholders.

i. **Financial institution:** the financial institution may choose to make the system a part of their internal processes. This may an advantage for them in that it will allow them to have more control over the development, ensure quality, and demand specific formats for output. On the other hand, the financial institution may have to employ additional people with the necessary expertise to develop and manage the system which would raise their expenses. The financial institution may find it less taxing to simply use the system output without developing the system themselves.

ii. **Data collection company:** for the data collection company, it may be advantageous to develop and operate the system as it would give them more control over the data of their users, in addition to providing a new revenue stream. However, they may shy away

from doing so for the same reasons a financial institution may do so: to avoid the taxing effort of system operation and development, as well as to negate the need of hiring additional staff.

iii. **Independent operator**: the system may be developed and operated by an independent third party who would collect the data from the data collection companies, develop context-appropriate Alternative scoring factors, evaluate them, and provide the results of this evaluation to the financial institution. The system context described here is of a system operated in such a manner. It removes the burden of system development from the financial institution and data collection companies, allowing them to focus on their core businesses, while still giving both the benefits they desire.

III.5 REQUIREMENTS ANALYSIS

The **context diagram** (figure 6) is developed to show the interactions between the system and external elements. A system boundary is established with the understanding that the system itself is not responsible for data collection or for the actual provision of credit services; its purpose is to facilitate the exchange to information and financial resources between stakeholders. Considering the boundary between the system of interest and its external systems, a requirement for data collection from the two data sources (satellite and mobile data) becomes apparent.

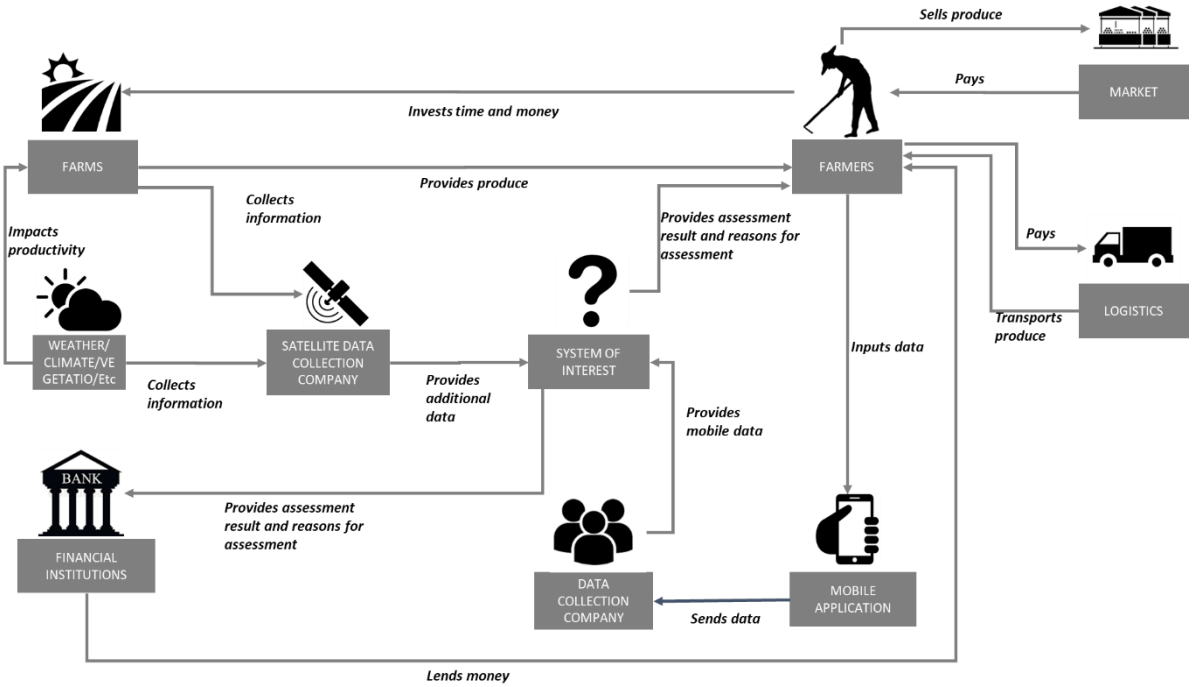


Figure 6 context diagram

The **Use Case diagram** (figure 7) of this system would be as shown below. This diagram introduces the requirement for the system to receive the request for a credit evaluation and to output the evaluation information to the financial institution and farmer.

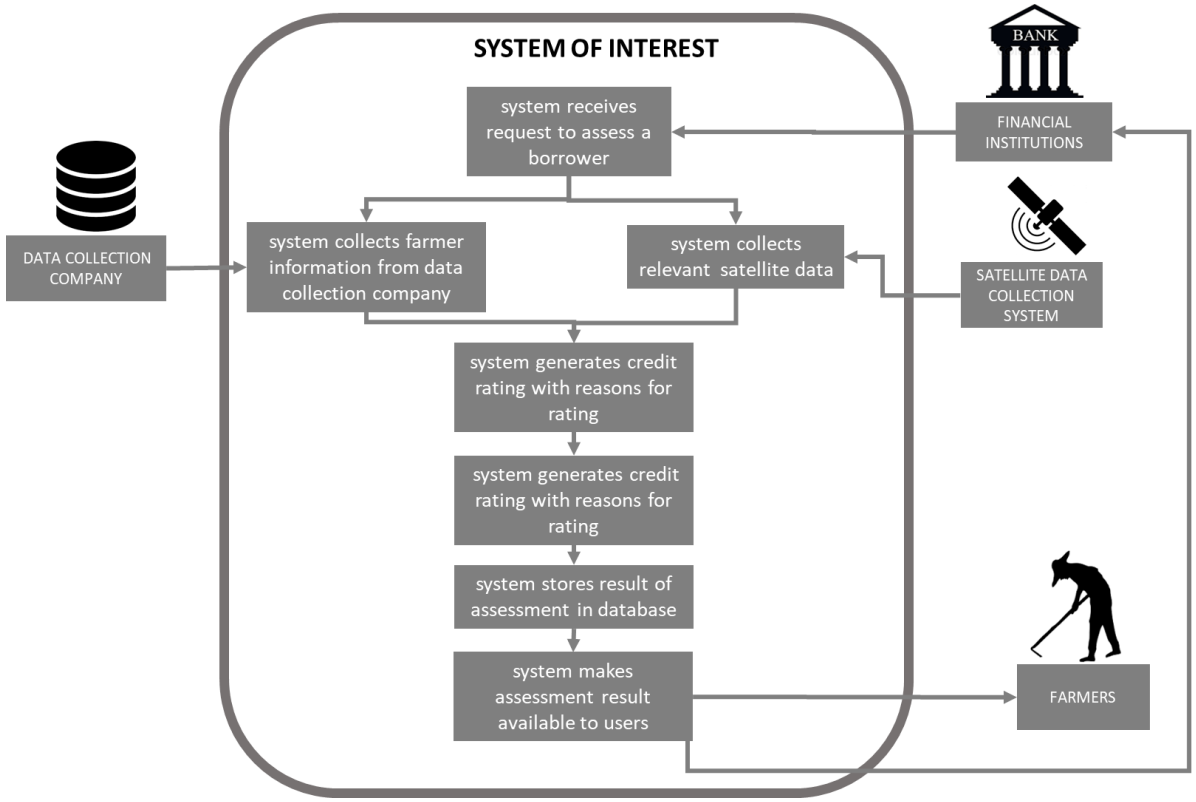


Figure 7 use case

This operational context and use case lead to the following **use case description** (table 5):

Table 5 use case description

No.	USE CASE DESCRIPTION
1	The system receives a request for credit rating for a farmer
2	The system collects mobile data from the data collection company
3	The system collects data from the satellite system
4	The system analyses the collected data in preparation for credit rating generation
5	The system generates a credit rating based on the risks of lending to the farmer
6	The system generates reasons for the given credit rating
7	The system stores results of evaluations
8	The system outputs credit rating and reasons for the credit rating to the financial institution and farmer

III.6 SYSTEM REQUIREMENTS

The system requirements drawn from the context diagram and use case are given in table 6. It is important to note that not all stakeholder needs are addressed by this system. In particular, the needs of the data collection company and satellite data collection company (need for revenue) are not factored into the design. This is done to focus the design on the technical requirements.

Table 6 system requirements

No.	**SYSTEM REQUIREMENT**	
1	The system shall receive a request for credit rating for a farmer	
2	The system shall collect data for evaluations	
	2.1	The system shall collect mobile data from the data collection company
	2.2	The system shall collect data from the satellite system
3	The system shall analyse the collected data in preparation for credit rating generation	
	3.1	The system shall analyse the collected non-geospatial data in preparation for credit rating generation
	3.2	The system shall analyse the collected geospatial data in preparation for credit rating generation
4	The system shall generate a credit rating based on the risks of lending to the farmer	
5	The system shall generate reasons for the given credit rating	
6	The system stores results of evaluations	
7	The system shall output credit ratings and reasons for ratings	
	7.1	The system shall output a credit rating and reasons for credits rating to the farmer
	7.2	The system shall output a credit rating and reasons for the credit rating to the financial institution

III.7 FUNCTIONAL DESIGN

The following is the **functional flow of the system** (figure 8)**:**

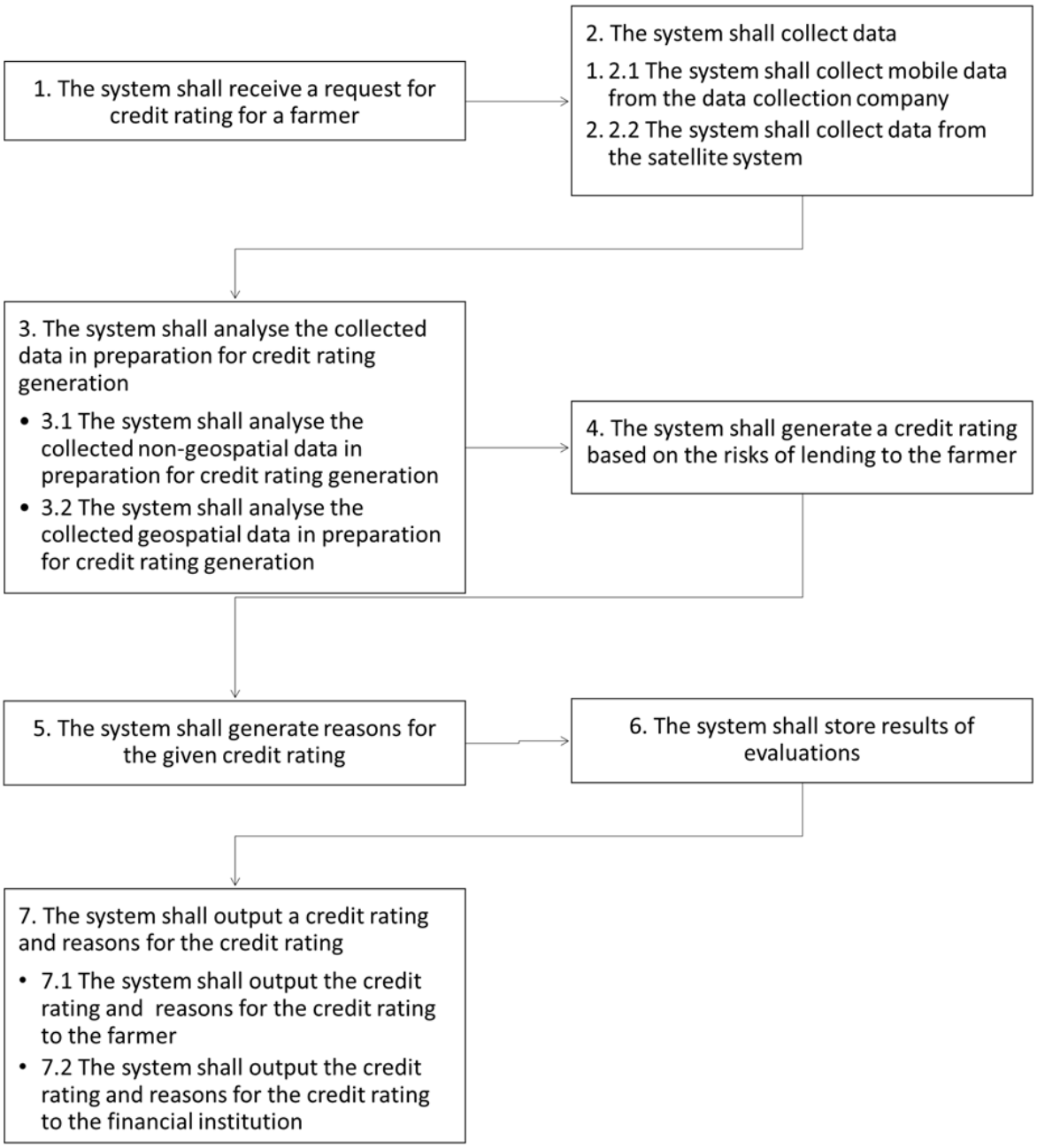


Figure 8 functional flow of system

III.8 PHYSICAL DESIGN

The **physical design** of the system is presented in figure 9 below.

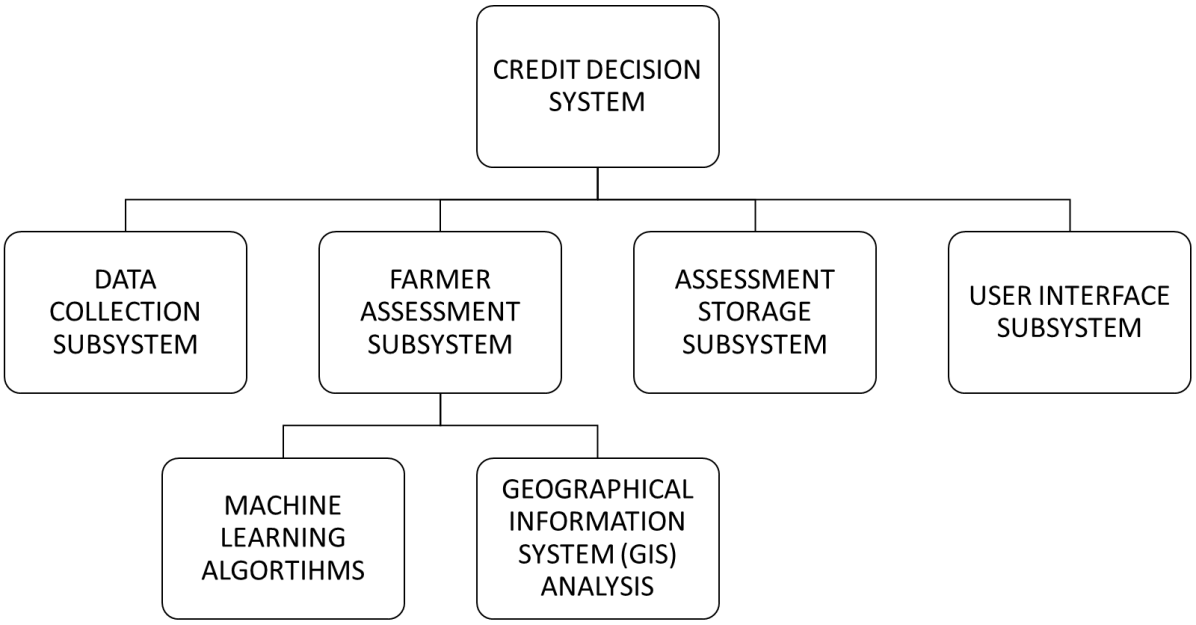


Figure 9 physical design

i. **Data Collection Subsystem:** The data collection subsystem may be implemented through software or by an individual with its function being to collect data from relevant sources. Access to the data will be impacted by factors such as cost, licensing, and data privacy laws, particularly for personal information. Mobile phones may be used to collect the majority of the data, thereby taking advantage of the convenience of this approach to data collection. Supplementary data may be collected through other means including surveys, and open source platforms. Types of data that should be collected will depend on the risks to be evaluated in a specific context.

ii. **Farmer Assessment Subsystem:** This subsystem is responsible for analysis of collected data and can be decomposed into the following:

   a. **Machine Learning Algorithms:** these are responsible for the analysis of non-geospatial data. machine learning algorithms, in particular classification algorithms are the standard analysis method for development of credit scores. Using these, a model can be trained based on a set of data and then tested with another set to check for model performance. Afterwards, the trained models can be used to generate scores for other farmers. The exact algorithms are not known and must be established through analysis and comparison in a specific context.

b. **Geographic Information System Analysis (GIS):** GIS analysis allows for the analysis of geospatial data such as location. This data can be combined using GIS analysis tools to determine the interactions between one type of data (e.g. location of farmers) and another type (e.g. temperature). This type of analysis will allow for the analysis of water distribution, temperature, and other factors that may impact farming activities.

iii. **Assessment Storage Subsystem:** storage of the system's evaluations is to be decided on based on the specific context. The latter may be implemented through a cloud service or on a dedicated server. Based on the cost, maintenance, and technical feasibility, various options may be explored and chosen.

iv. **User Interface Subsystem:** It is important that the use case be taken into full consideration during the design of user interface. In this case, financial institutions and farmers would receive the system. To optimise the experience of each user, two separate user interfaces must be designed; one for each user. Since farmers would already be using a mobile application to input their data for analysis, it may be beneficial to incorporate the system output either as part of the existing mobile application or as an additional mobile application service operated independently. In the former case, the system operator would have to be built in conjunction with the data collection company to deliver the output. Meanwhile, in the latter case the system operator would have more independence in delivering the output. The user interface to be used by the financial institution may also be integrated into existing bank procedures. For instance, if the bank makes evaluations using a web-based service then the service may be extended to include the output from the new system. The advantage of integrating the system output into existing infrastructure is that it reduces design time and takes advantage of an existing user-base. This architecture can be summarized as shown in figure 10.

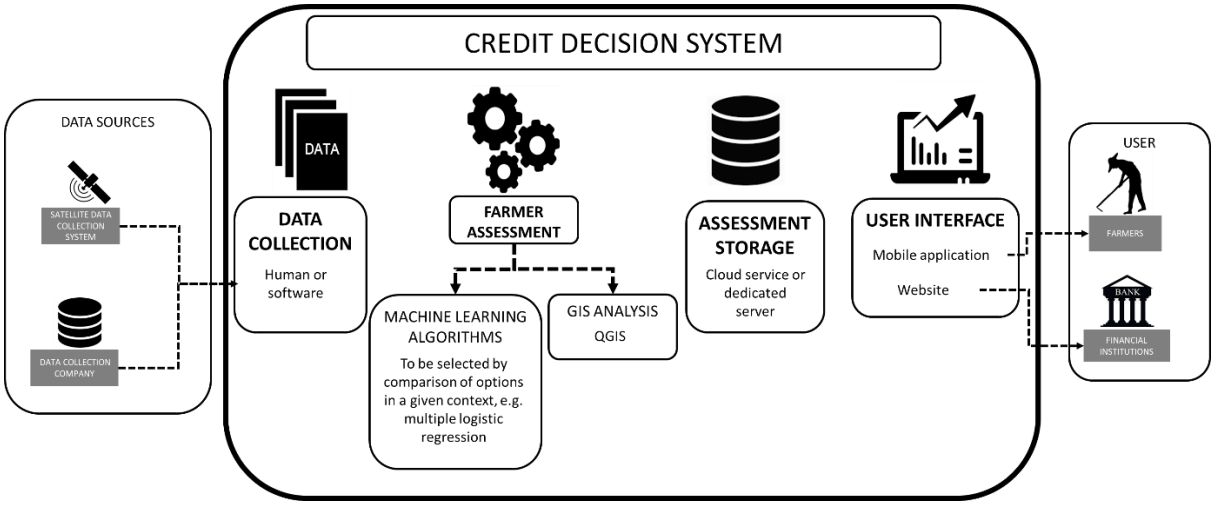


Figure 10 detailed physical design

III.9 SYSTEM ARCHITECTURE

The system was design based on the system engineering process, resulting in the architecture of figure 11:

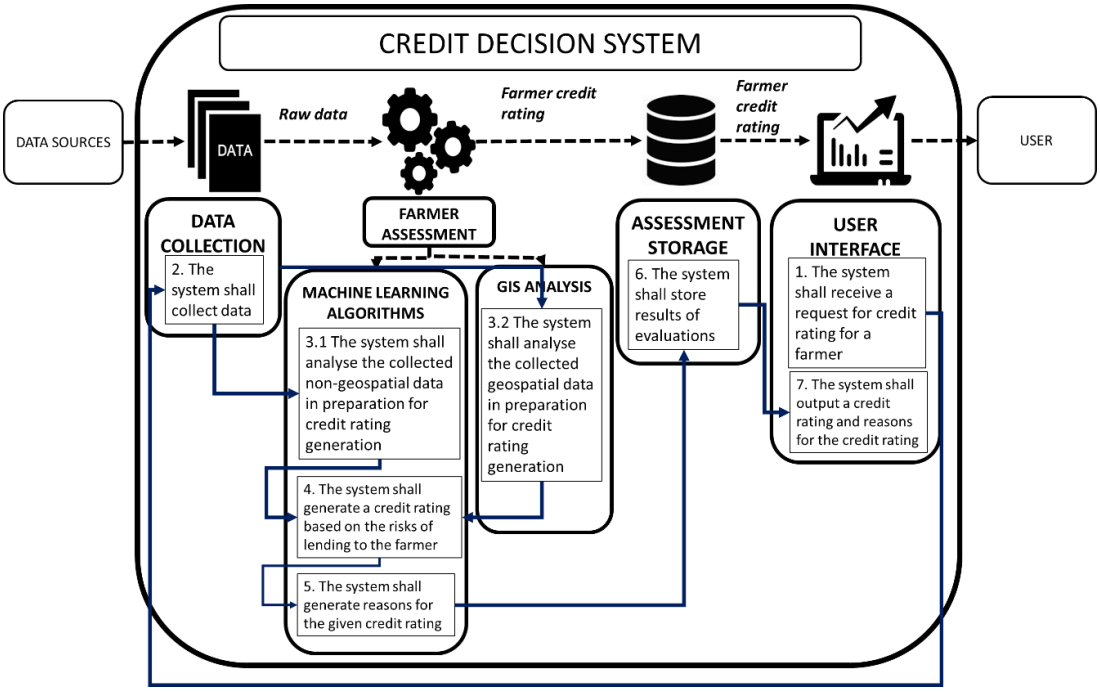


Figure 11 system architecture

III.9.1 SYSTEM INTERFACES

i. **Internal interfaces:** there are numerous interfaces between subsystems (highlighted in blue in system architecture diagram). The implementation of these interfaces is context specific, relying on the implementation methods for each subsystem.

a. **Data collection subsystem to Farmer assessment subsystem interface:** if the data collection subsystem is software, then this interface can be a software program used to transfer information from one subsystem to the other. However, if the data collection subsystem is a human, then the interface will require communication between the human and the software used for the farmer assessment subsystem.

b. **Farmer assessment subsystem to Assessment storage subsystem interface:** The interface from the farmer assessment subsystem to the assessment storage system can be implemented using a software program which transfers the results of an assessment to the Assessment storage server or cloud service.

c. **Assessment subsystem to User Interface subsystem interface:** If the user interfaces are electronic as proposed, software programs must be written to transfer the data from the Assessment storage subsystem to the User interface subsystem.

d. **User interface subsystem to data collection subsystem interface:** this interface will again depend on the nature of the Data collection subsystem and may be a human-to-software interface or a software-to-software interface.

ii. **External interfaces:** there are two main external interfaces between the system of interest and the systems external to it.

a. **Data sources to System of interest:** an interface (or more than one depending on how many data sources are used) exists between the system of interest and the data sources (both from the data collection company and the satellite data collection company). This nature of this interface may be software provided the data collection subsystem is also software and that the data sources can output the data in an electronic format. Otherwise, the interface will involve human-to-human communication or human-to-software communication

b. **System of interest to Users:** this interface provides a means for users to access the system output, where users here refers to farmers and financial institutions. The nature of this interface may be visual or audio, based on the user interface subsystem's physical implementation in a specific context.

III.9.2 IMPLEMENTATION PROCESS

The system design proposed here is general in that there is no specific context described: data collection company, financial institution, etc. These are elements that can only be identified

during implementation. Implementation of the system must be focused on fuflfilling described requirements using the chosen physical elements. In some cases, the best physical element must be chosen by first comparing the available options and selecting the most apt for implementation in that context.

III.9.3 DEVELOPMENT OF SCORING FACTORS

The main requirement of this system can be said to be requirment number 4 since it is the core reason for the system's existence. To fulfill it, it is necessary to establish the required criteria that must be considered when determining the risk of lending to smallholder farmers. In other words, the risks that will affect loan repayment by farmers in the implementation context must be known. **Alternative scoring factors** can then be designed to address each of the key risks identified.

IV. SYSTEM EVALUATION

As part of the evaluation of the system design, a prototype was developed of certain portions of the system to demonstrate that their proposed design and functionality was feasible.

IV.1 PROTOTYPING

IV.1.1 PROTOTYPING PURPOSE

The functions to be prototyped here are highlighted in the Figure 12 below. These functions were selected for prototyping because they represent the core function the system must be capable of performing in order to fulfil stakeholder requirements. Although the remaining functions are necessary for system operation, they are not as critical as those that are to be prototyped and so can be the subject of future prototyping efforts.

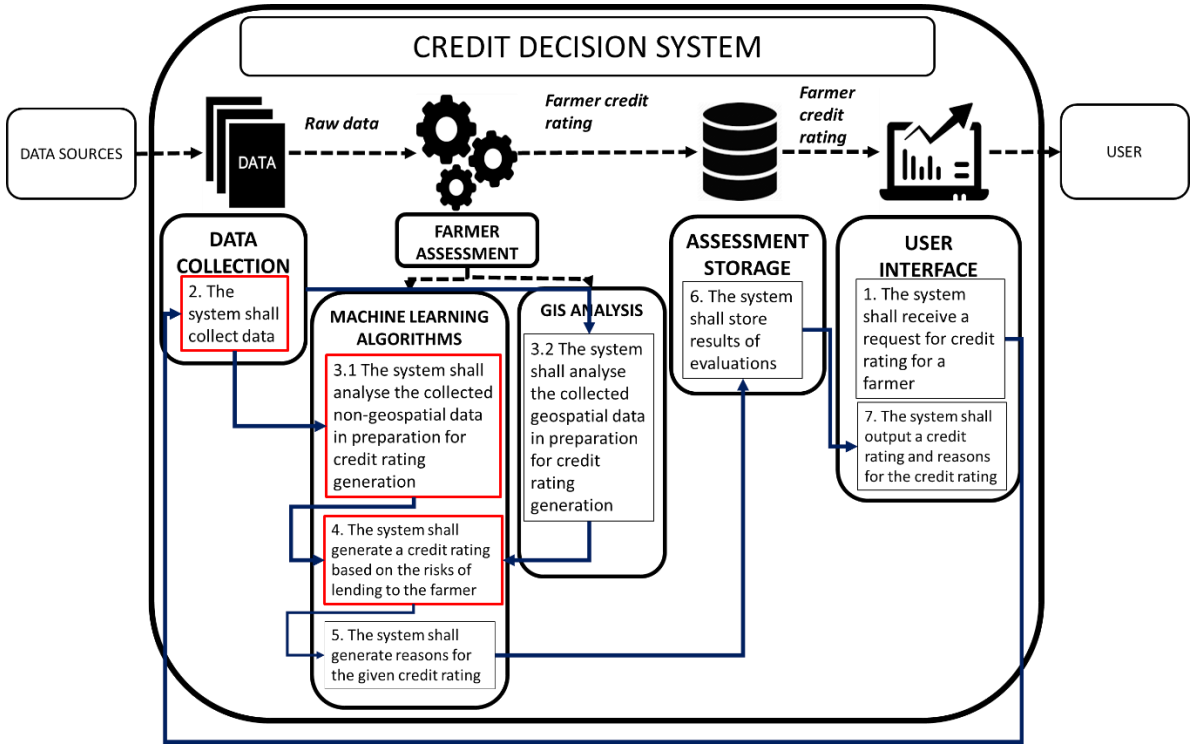


Figure 12 target functions for prototyping

IV.1.2 PROTOTYPING AREA

The area of data collection was Cambodia, in image 2, one of the less developed countries in Asia, with 14% of the population living below the national poverty line as of 2014 according to the Asian Development Bank [42]. Nearly 80% of the country's population of 15.4 million people resides in rural areas where the major economic activity is agriculture [43]. In fact, half the population in employed in the agricultural industry [44] the main crops being Rice, Cassava, and Corn. With little irrigation being implemented, most farmers depend on seasonal

rainfall for crop production. As there is only one rainfall season per year, farmers who rely entirely on rainfall are only able to grow one crop per year. Access to financial services remains a challenge with less than 30% of adults using formal lending services [45]. Rural farmers face significant challenges in obtaining sufficient funds for their agricultural activities. By instituting a credit decision system, more financial institutions can be encouraged to enter the agricultural microfinance space. This will make it more competitive and increase the number of options for rural farmers when it comes to obtaining funding.

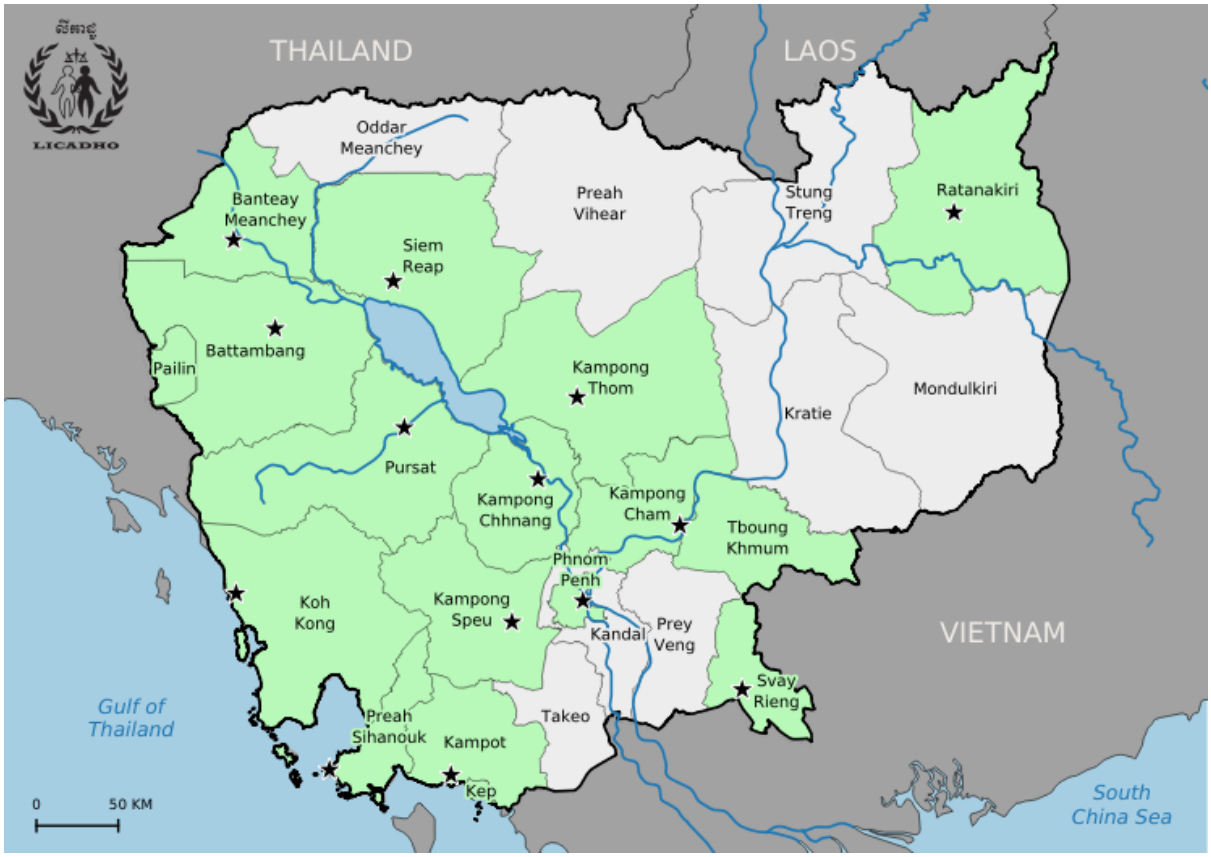


Image 2 Map of Cambodia [source:www.licadho-cambodia.org]

IV.1.3 RESEARCH PARTNERSHIP

Data was provided through a collaboration with Agribuddy Ltd[3]., a mobile platform company which offers various services to farmers and other players in the agricultural sector in Cambodia. Users of the mobile platform upload reports from farmers in their areas. The information available from the user records includes: personal information of users and farmers, location, and regular reports on farm activities such as harvest. Reports may be of different types including seeding (planting), harvest, tractor (machinery report), trouble (issues

[3] A more detailed introduction to the company structure and activities is given in the appendix (section IX.2).

with pests), and other farm activities. The areas of data collection are the Battambang, Siem Reap, Kampong Cham, Kampong Chhang, Kampong Speu, Kampong Thom, Kratie, and Prey Veng provinces of Cambodia.

IV.1.4 DETAILED PROTOTYPE DESIGN

A system design was created for the prototype system, beginning with identification of stakeholders and their requirements. The stakeholder described in Table 7 have an interest in the development of the system prototype.

Table 7 prototype system stakeholders' needs

STAKEHOLDER	ROLE	NEEDS
Aeon Specialized bank Cambodia PLC	Financial institutions	1. Credit decision information on each potential borrower (farmers) that accurately estimates the risks of lending to smallholder farmers 2. Simple output with clear reasons for each decision
Small holder farmers working with Agribuddy Cambodia LTD	Smallholder farmers	3. Simple output with clear reasons for each decision
Agribuddy Cambodia	Data collection companies	4. Revenue

Based on these needs, the stakeholder requirements to be fulfilled by the prototype are shown in table 8[4].

Table 8 prototype system stakeholder requirements

STAKEHOLDER	ROLE	REQUIREMENTS
Aeon Specialized bank Cambodia PLC	Financial institutions	1. Credit decision information on each potential borrower (farmers) that accurately estimates the risks of lending to smallholder farmers

The prototype system context is presented in Figure 13. It includes the collection of data from farmers by the data collection company. This prototype has no output as its function are to collect and analyse data. The physical design of the prototype is shown in Figure 14. It has a combination of the Data Collection subsystem and the selected aspects of the Farmer

[4] Not all stakeholder needs will be fulfilled by the prototype. The aim of the prototype is to focus on the feasibility of specific system functions

Assessment subsystem. The Data collection subsystem will be implemented by a human while the Farmer Assessment subsystem will be implemented using R programming software for development of machine learning algorithms. The specific algorithm to be used will be determined by comparing the results of actual analysis of available data.

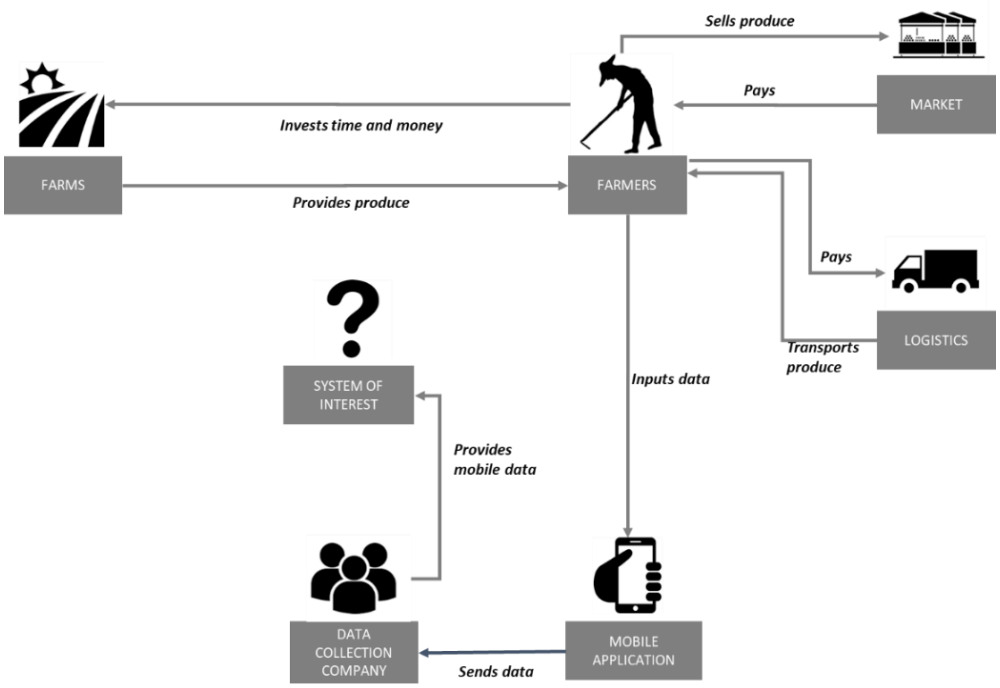


Figure 13 prototype context

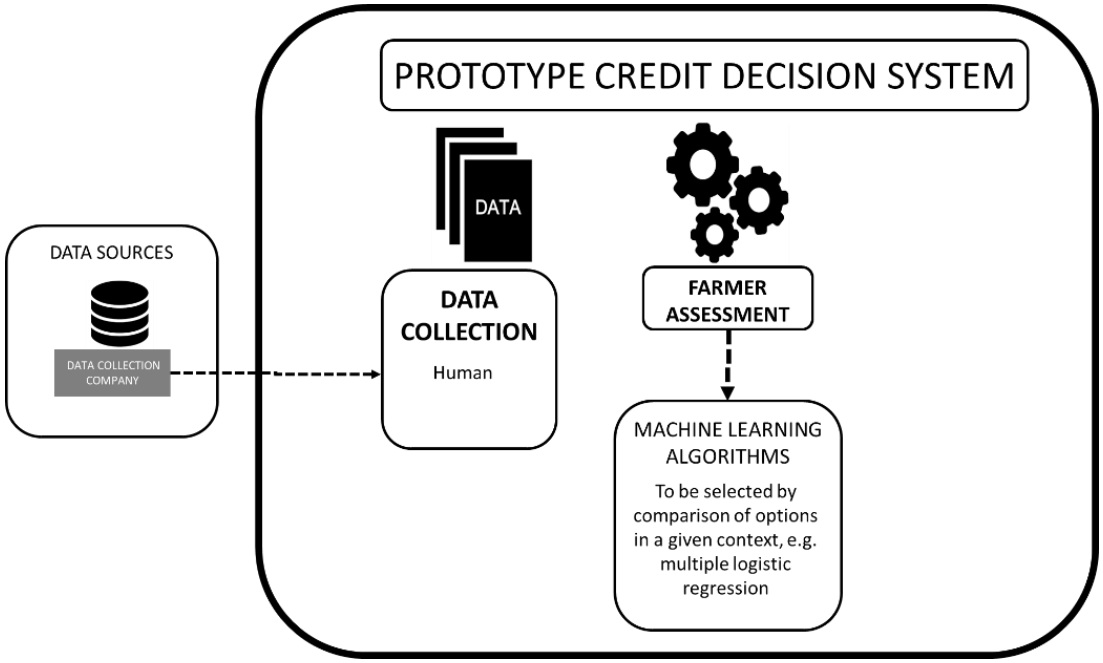


Figure 14 prototype physical design

The final prototype system architecture is shown in Figure 15 with data collected only from the data collection company.

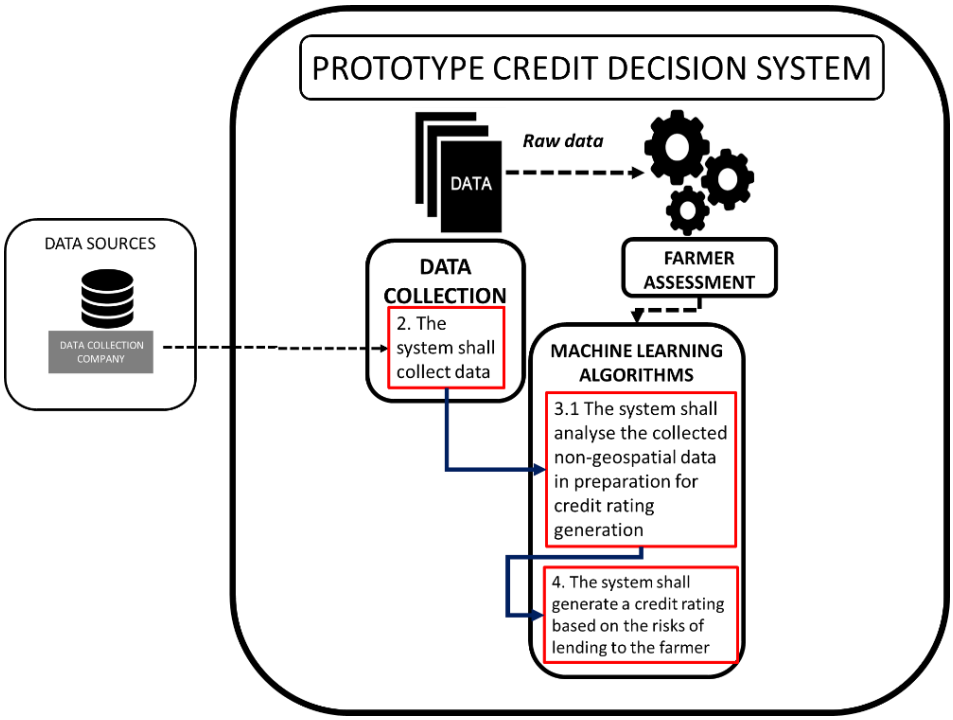


Figure 15 prototype system architecture

The prototype system has one internal interface and one external interface. The latter will be implemented by human-to-software communication where the human performing data collection will access the data from the data collection company on behalf of the system. The internal interface from the Data Collection subsystem to the Farmer Assessment subsystem will involve the human data collector making the information available to the software of the Farmer assessment subsystem.

IV.1.5 PROTOTYPING PROCESS OUTLINE

The process of prototyping will be conducted as shown in Figure 16.

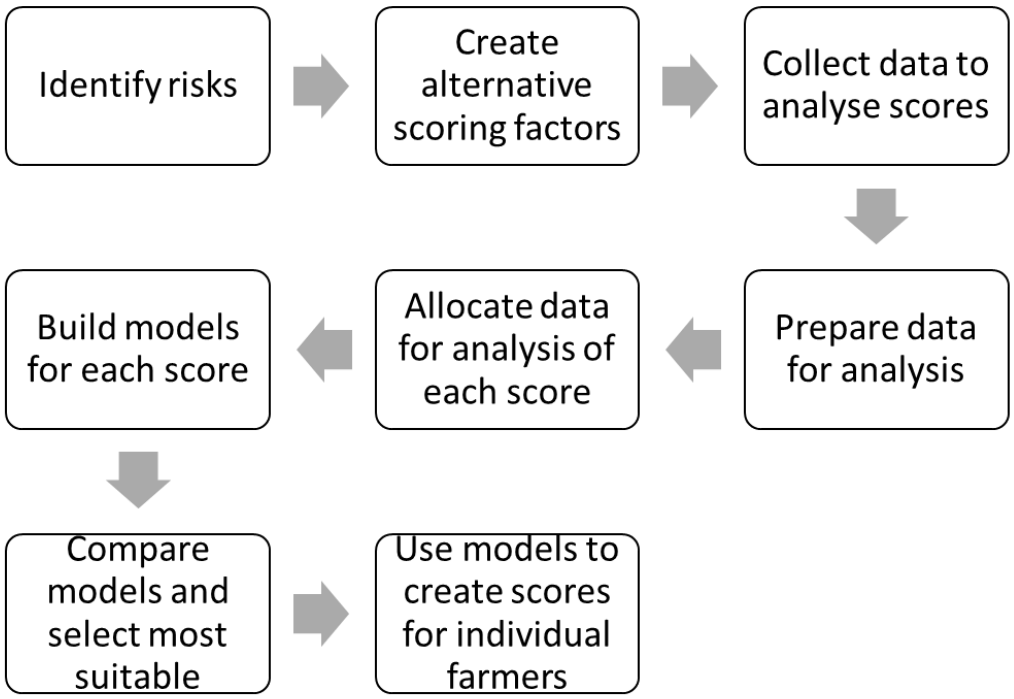


Figure 16 prototyping process outline

IV.1.6 FUNCTION 2: DATA COLLECTION BY DATA COLLECTION SUBSYSTEM

The data collection subsystem is responsible for fulfilling **System Function 2** by collecting data from the various available sources to be analysed by the system. This subsystem was prototyped by collecting data from the sources available in this context as mentioned below:

Mobile Data: The first is the use of a mobile application platform that provides various services to farmer and other players in the agricultural sector. This data was collected from users over a 5 year period (2012 to 2016) using the mobile application platform. Users of the application install it directly on their smartphones. These users then collect data on a regular basis from farmers in their areas who are unable to use the application directly due to factors such as inability to afford a smartphone or illiteracy. The data collected includes the personal and farm activity information from each farmer. The former consists of age, family size, length of residence, etc. while the latter consists of details of planting, harvesting, and other farm activities. These users may also be referred to as “Buddies”.

Survey Data: The second data collection method is surveying, which was used to supplement the data available through the mobile application. The data collected in this case includes crops

grown, farm size, irrigation and pest control methods used, crop yields and sale prices, among others. The survey was conducted by users of the mobile application who collected the data from farmers under their purview and provided the survey results to the data collection company, Agribuddy Ltd. The details of the types of data collected through Agribuddy Ltd are given in the Appendix.

IV.1.7 FUNCTION 4: ALTERNATIVE SCORE DEVELOPMENT FOR FARMER ASSESSMENT SUBSYSTEM

System Function 4 is fulfilled by assessing farmers based on the risks of lending to said farmers. Doing so requires that the risks are first identified. These risks may vary by application context and so risks must be identified for each context in which the system operates.

i. RISK IDENTIFICATION

In order to determine the risks of lending to farmers and thereby fulfil the requirement of evaluation borrowers based on these risks, interviews and workshops were conducted with various stakeholders. These risks would then form the basis of the credit factors to be evaluated. Field work was conducted in July 2017 to identify the risks that would impact loan repayment by farmers.

Image 3 farmer attempting to use pump for irrigation

Image 4 workshop conducted with employees of data collection company to identify risks

The first step in developing alternative scoring factors is identification of risks that may lead to loan default by smallholder farmers in the study area. Interviews were conducted with smallholder farmers. Image 3 shows a farmer who was interviewed attempting to use an irrigation system. The details of these interviews are given in the appendix. The first farmer inetrviewed had lived in the village for a very long time and so had a great amount of influence in the community. His reputation in the community was important to him which reduces the risk that he would default on the loan without good cause, lest he damage his reputation. The second farmer interviewed had a good personal profile and was hardworking. However, from a business perspective she was a poor risk because she was growing a crop (cassava) which has a very low market price and was unwilling to change this even though she was aware of the risks involved.

An interview was also conducted with Mr Kawaii of Aeon Specialized bank Cambodia (PLC). He stated that one of the key concerns the bank would have in lending to smallholder farmers would be determining the revenue that each farmer could generate. Personal characteristics of

the farmer were also a point consideration. Further risks that he wished to include were risks of drought, flood, etc[5].

In addition to interviews, a workshop was conducted with employees of Agribuddy to identify risks. The employees were all local to the area and worked with farmers on a daily basis, collecting data on their farm activities. The 11 workshop participants were asked to identfiy what risks may cause smallholder farmers in the area to default on a loan. This led to the list of risks given in the appendix. 79 risks were identified through brainstorming and distilled into the general categories listed in table 9.

Table 9 summary of risks identified during prototyping

RISKS	NUMBER OF TIMES MENTIONED
Other factors	10
Factors due to farming experience	11
Factors due to capital	4
Factors due to weather	3
Factors affecting market conditions	8
Unreliable data	13
Factors affecting crop yield	11
Trust between farmers. Data collectors and lenders	19

ii. ALTERNATIVE SCORING FACTORS

Considering the available data, three alternative scoring factors are proposed which can evaluated based on the available data and which represent the highlighted risks:

1. **Reliability Scoring Factor:** this scoring factor is proposed to denote the reliability of the data collector (i.e. user of the Agribuddy mobile application platform who inputs data on behalf of farmers). Since data is collected by platform users, any fraudulency or bad behaviour on the part of the platform user will result in broken trust among the farmer, user, and data collection company, **which was identified as one of the key risks.** A good/bad classification developed by the data collection company was used. The company has developed a classification of users based on good behaviour. In general, good behaviour is

[5] Details in appendix section IX.4

defined as trustworthiness, reliability, absence of fraudulency in the use of the mobile platform. The "user classification" is used as the dependent variable to be predicted by the Reliability Scoring Model.

2. **Revenue Scoring Factor:** this credit scoring factors is proposed to denote the ability of the farmer to generate income from their farming activities. **C**apacity to generate income was identified as **another of the key risks since is also affects a person's ability to repay loans**. Therefore, it must be factored into the credit decision system. For smallholder farmers, farming is often a major source of income. The revenue generated from farming is a function of the crop yield and market prices. Therefore, the farm revenue generation capacity would have to be determined in order to establish farmer's capacity to repay loans and would encompass these two factors. This can be evaluated by analysing the probability of the farmer generating a certain amount of revenue.
3. **Interaction Scoring Factor**: this scoring factor is proposed to denote the relationship between the farmer and the mobile platform user. We treat the level of interaction between the two as a proxy for trust. Taking each day a report is sent by the user as a record of interaction between the two. The "**interaction ratio**" is applied as the dependent variable to be predicted by the Interaction Scoring Model. This is the ratio of "number of days a report is sent" to the "period in days that the farmer has been registered on the platform".

Table 10 shows the alternative scores developed and the risks they are meant to address while tables 11, 12, and 13 give details of the data features used to predict each score.

Table 10 alternative scores and risks addressed

ALTERNATIVE SCORING FACTOR	MEANING	RISK ADDRESSED
Reliability Scoring Factor	Evaluates risk of unreliable data by data collector	• Evaluates risk of unreliable data by data collector
Revenue Scoring Factor	Measures probability of farmers generating above average level of revenue	• Low crop yield • Poor market conditions
Interaction Scoring Factor	Measures level of interaction between data collector and farmer as a proxy for trust between them	• Trust between farmers. Data collectors and lenders

Table 11 features used to predict reliability score

PREDICTORS FOR RELIABILITY SCORE
• Classification • Total Number of reports sent by user on behalf of all his/her farmers • Total number of farmers user reports on • Number of farmers user has registered but not reported on • Days since user joined the mobile application • User age • Number of years of user has resided in current location • Registered farmers living outside 4km radius of user • Registered farmers living within 4km radius of user • Total number of farm areas registered by user • Number of locations user has reported on • Number of crop variations grown by farmers registered to user • Number of farms user registered and reports on • Number of days user has sent reports • Average number of reports sent by a user per registered farm • Number of Farm activity types reported • Number of general farm activity reports • Number of Harvest activity reports • Number of seeding activity reports • Number of tractor activity reports • Number of trouble activity reports • Number of GPS reports • Number of photo reports • Number of farm areas user registered correctly • Number of actions of mobile platform • Number of calls from mobile application company • Average number of days farmers have used the service • Number male farmers • Number of female farmers • Average number of years farmers have resided in current location • Average age of registered farmers • Average number of family members of farmers • Number of registered farms • Average number of days person who invited user to mobile platform has used the service • Number of persons invited to the platform by the user • Number of family members of user • Availability of family photo • Availability of land document photo • Availability of id card photo • Availability of photo of user with his/her house • Determines if use was invited by another user • Determines if user communicates with mobile platform company • Earliest report time (of day) • Latest report time (of day) • Mean report time (of day) • Fraction of important data not sent by user

Table 12 features used to predict reliability score

PREDICTORS FOR INTERACTION SCORE
• Number of family members of farmers • Farmer age • Farmer gender

• Farmer Commune of Residence • Farmer District of Residence • Farmer Province of Residence • Farmer Village of Residence • Number of years of farmer has resided in current location • Number of crop types grown by farmer • Number of crop cycles per year • Number of Harvest activity reports • Number of Farm activity types reported • Number of days with reports over days registered on the platform • Number of crop variations grown (of the same crop type) • Number of reports sent on behalf of the farmer • Earliest report time (of day) • Total Number of reports sent by user on behalf of all his/her farmers • Total number of farmers user reports on • Average number of reports sent by a user per registered farm • Number of years of farmer has resided in current location

Table 13 features used to predict interactions score

PREDICTORS FOR REVENUE SCORE
• Farm size in hectares • Crop type • Number of crops cycles per year • Number of crop types • Number of labourers working on the farm • Chemical pest control method • Organic fertilizer type • Non-Organic fertilizer • Types of pest attack in the last year • Organic pest control method • Use of irrigation • Type of water source • Ploughing method • District • Village • Province • Crop variety • Harvest method

iii. DATA ANALYSIS METHODS

The following data analysis methods were used for analysis with their outputs compared to determine which would be the most suitable in this application:

1. **Logistic Regression**: Logistic regression is a predominant method in credit analysis and has become the benchmark method against which all other methods are compared for credit analysis [24]. This is in part due to its simplicity in comparison to other methods. For making credit assessments, logistic regression can be used to classify a potential borrower. The

classification is either "good" or "bad" with the associated characteristics used as predictors. The process finds the probability (Y) of obtaining a given class depending on the values of the predictors (X). Probability and predictors are related by the equation below. The values of the intercept (b0) and the coefficients (b1, b2…) are found using the maximum likelihood estimation method. This relationship is defined in equation (1) below. To classify a potential borrower as "good" or "bad", the probability (Y) is determined and a threshold is selected. A probability above the threshold is taken as one class (for instance: "good"), while the reverse situation implies the borrower is in the other class (in this case: "bad").

$$\ln\left(\frac{Y}{1-Y}\right) = bo + b_1X_1 + b_2X_2 + \cdots \quad (1)$$

2. **Support Vector Machines (SVM):** A support vector machine (SVM) can also be enlisted in credit analysis. It is a more sophisticated method of analysis which can be used for classification and regression tasks. In classification tasks, it defines a hyperplane which best separates the data points into classes while maximizing the margins around the hyperplane (i.e. the distance between the two classes). The data points lying on the margins are referred to as support vectors. Soft-margins are used to determine the impact of any points that fall into the wrong classification on the model. A cost parameter, C, is used to control the soft-margin so that larger values correspond to a model that is more forgiving of errors. A benefit of SVM classification is that it can easily fit linear and non-linear data. This is done through the kernel trick of transforming the feature space to higher dimensions. The transformation used depends on the type of kernel chosen.

   Linear Kernel: this method does not map the feature space to a higher dimension. The function used for mapping in this case is given in equation (2) where x and z are two feature vectors. The performance of this method depends on the value of the cost parameter, C. Higher values generally mean the model is more forgiving of errors.

$$K(x, z) = x \cdot z \quad (2)$$

   Gaussian Kernel (Radial Base Function, RBF): this method is better able to handle non-linear relationships in the data by transforming the features to higher dimensions. It is very flexible, able to map the input space into features spaces of infinite dimensions, and so can handle many different interactions in the data [30]. There is, however, a risk of overfitting the model. The kernel function is given in equation (3) where x and z are

two feature vectors and γ is a parameter to be optimized by tuning. The performance of this method depends on the value of the cost parameter of the soft-margin, C, and the γ parameter of the kernel. Higher values generally mean the model is more forgiving of errors [46]. This C value will be tuned during cross-validation as the model is developed to find the value which best fits the data.

$$K(x, z) = \exp\{-\gamma(||x - z||^2)\} \quad (3)$$

Evaluation Metrics: The evaluation metrics chosen for comparison of models are Receiver Operating Characteristic (ROC) curve, area under the ROC (AUC), and Accuracy. These are commonly used metrics for comparing different classification methods. This is mainly because they consider only the actual output of classification without regard to the underlying method used for classification. Evaluation parameters such as Root Mean Square Error (RMSE) which are normally used to evaluate logistic regression results could not be used here since the value would have no meaning in the context of a Support Vector Machine. The ROC curve plots the true positive rate (sensitivity) against false positive rate (1-specificity).

$$sensitivity = \frac{true\ positives}{true\ positives + false\ negatives} \quad (4)$$

$$specificity = \frac{true\ negatives}{true\ negatives + false\ positives} \quad (5)$$

The area under the ROC curve (AUC) has meaning as the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance [47]. Accuracy will also be used as an evaluation metric as it is an important consideration in assessing credit decisions and would the easily explainable to relevant stakeholders using the credit decision tool.

$$accuracy = \frac{true\ positives + true\ negatives}{total} \quad (6)$$

iv. ALTERNATIVE SCORING MODEL DEVELOPMENT AND EVALUATION

a. Reliability Score

From the collection via the mobile application [shown in the appendix], 46 variables were initially extracted and analyzed to identify users exhibiting trustworthiness and reliability, which are both important factors to consider in the credit decision process. The Reliability score was

developed with “user classification” as the dependent variable based on a total of 213 users, 90 of these were considered negative with the remaining 123 considered positive. By analyzing the personal and activity information generated by each user of the mobile platform, it is possible to distinguish the factors impacting classification, and hence the credit score. In preparation for model building, missing values in the data were imputed. The data was re-scaled to the [0,1] range using max-min rescaling as presented in equation (7). Here x is the input variable while x’ is the rescaled value.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (7)$$

Categorical variables were converted to binary form so that a variable with r categories was replaced with r-1 binary variables. Graphical tools were used to explore the data, as well as to identify and remove outliers. Tests for distribution found that only some of the variables, such as age, followed a normal distribution.

Feature selection was used to reduce the number of variables used for classification. The first selection of variables for model building was based on the correlations between variables as established through the correlation matrix and independence tests. Highly correlated variables were removed to prevent redundancy. Sparseness of the data for each variable was an additional consideration so that variable with fewer missing values were given preference. This resulted in the first features set used for analysis. Recursive feature elimination (RFE) was also carried out using the Random Forest method. RFE is a backwards selection method which uses resampling and external validation to select the best performing set of features [46]. The feature set with the highest accuracy was chosen. This further reduced the number of features to create the second feature set. For model building, three sets of data were used: one with selected features from recursive feature elimination, one with features from removal of redundant variables, and finally one with all the original features. This was done to determine the impact of the different feature selection methods on the model building process and thus obtain a superior model.

The data was randomly partitioned into a 20:80 ratio. 80% of the data was used for training of the models while the remainder was used for testing of the final model output. Each of the models were developed using the train function of the caret package in R programming language [46] . For each fold, the data was split into training and testing sets, the model was trained on the training set and tested on the test portion. For each 10-fold cross validation, the above procedure was repeated 10 times in such a way that each data point was part of the testing set at

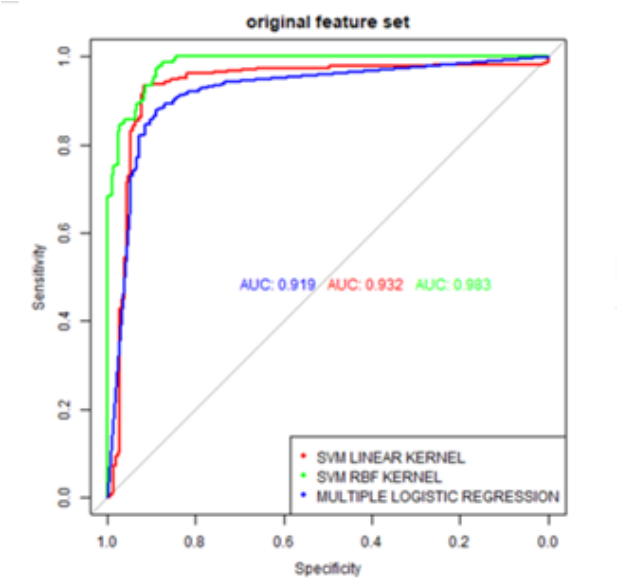
least once. 10-fold cross validation was iterated 3 times. In other words, the model was trained and tested 30 times to ensure that the results were a true representation of the data. The best model from each type was selected based on performance of the ROC curve.

Logistic Regression: To develop the logistic regression model, 10-fold cross validation was carried out 3 times as described above.

Linear SVM: To develop the linear SVM model, the cost parameter of the linear kernel function was tuned using 10-fold cross validation carried out 3 times as described above.

Gaussian (RBF) SVM: To develop the RBF SVM model, the cost and gamma parameters of the radial basis function were tuned using 10-fold cross validation carried out 3 times as described above.

For each model, the average of the ROC curves generated from each cross-validation iteration was plotted in figure 17. This resulted in the three ROC curves for each feature selection which will be used for evaluation. Further, the models are tested on the remaining 20% of the original data. The models were then evaluated based on the accuracy of the out-of-sample data tests.



(a)

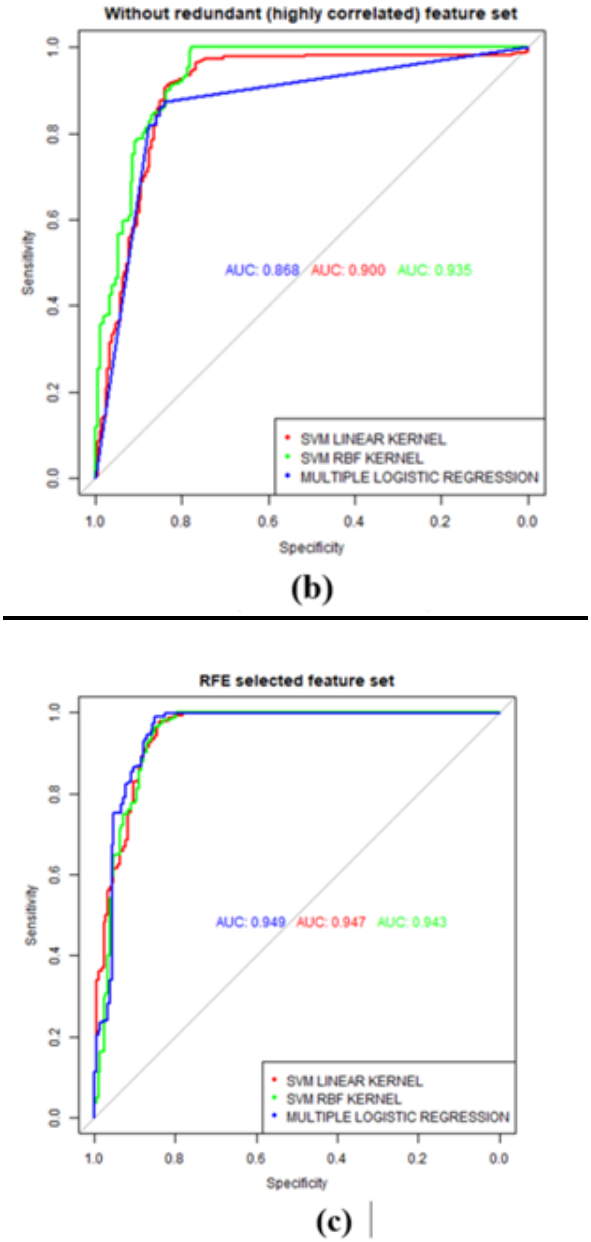
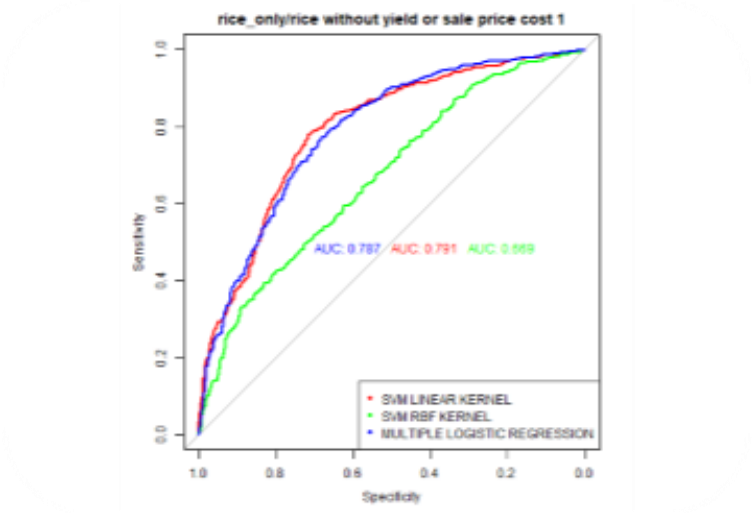


Figure 17: comparison of feature sets and algorithms from reliability score

b. Revenue Score

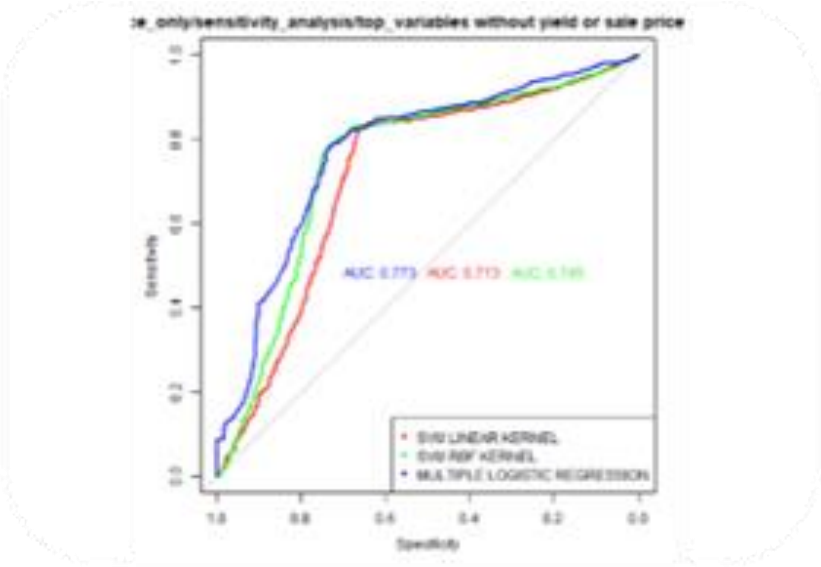
The Revenue score was developed with revenue as the dependent variable. Of the 1306 farmers evaluated, 653 farmers with revenue above the mean were given a positive classification while the remaining farmers were given a negative classification. The score building process used to develop the reliability score was repeated to generate scores that would denote the likelihood of a farmer creating an above average level of revenue. Various feature sets were created using sensitivity analysis as was done for the reliability scoring process. These included: original features, top 5 features selected through recursive feature elimination, and features selected

through sensitivity analysis of different models. Evaluating the results of this process led to the conclusion that the best option was to use the Multiple Logistic Regression method with a set of 12 features. The resulting ROC curves are shown in figure 18.



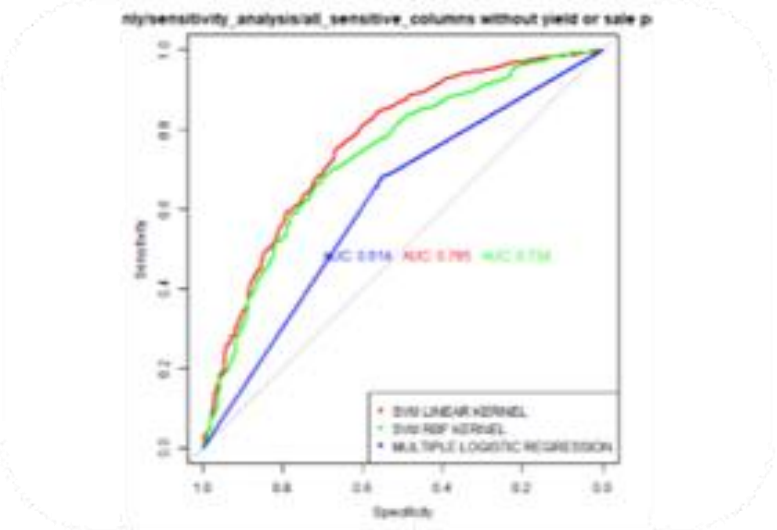
A) ORIGINAL FEATURES =

23 FEATURES



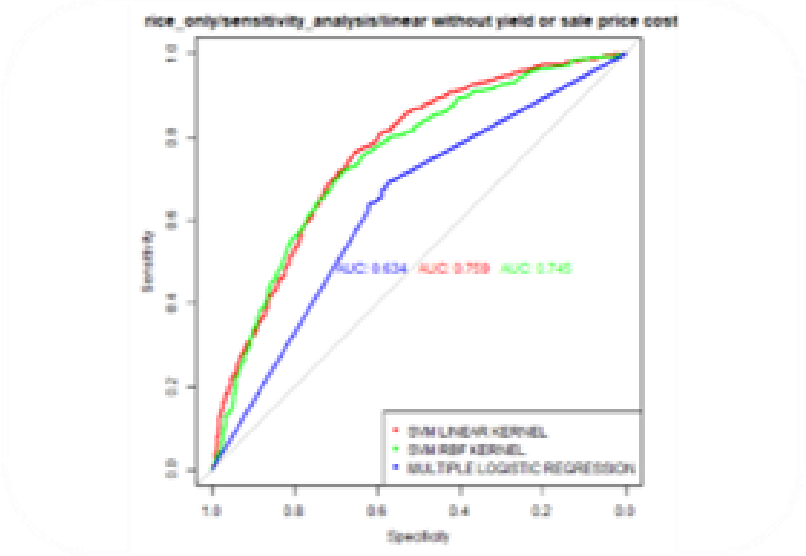
B) TOP RFE SELECTED =

4 FEATURES



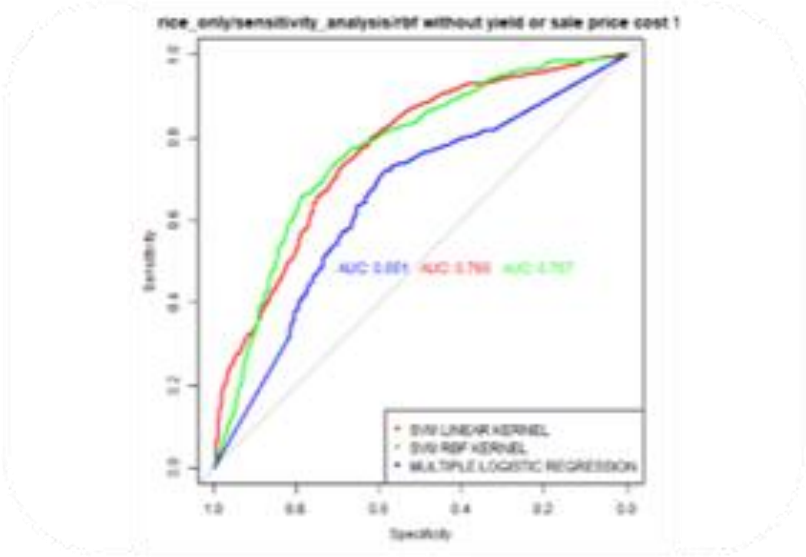
C) COMBINED =

19 FEATURES



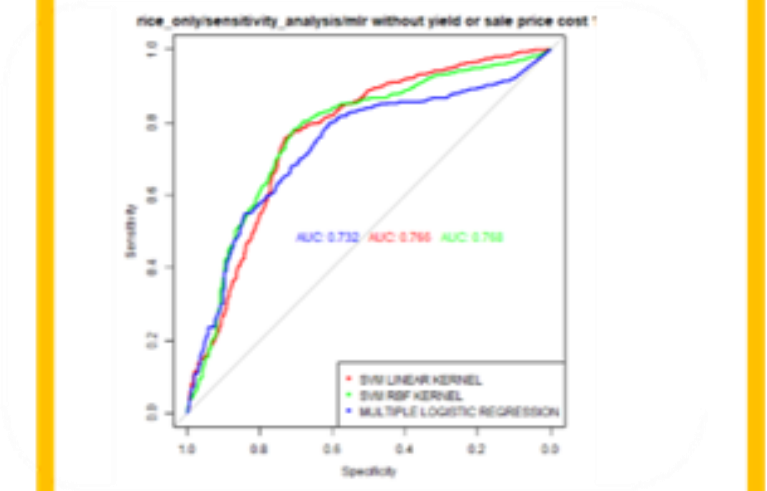
D) SVM LINEAR =

12 FEATURES



E) SVM RBF =

8 FEATURES



F) MULTIPLE LOGISTIC REGRESSION =

12 FEATURES

Figure 18 comparison of feature sets and algorithms for revenue score

c. Interaction Score

Data collected through the mobile application was used to evaluate this score. The Interaction score was developed with the interaction ratio as the dependent variable based on data of 494 farmers. 263 farmers with an interaction ratio above 0.003 were considered to be positive while the remaining values were considered negative. This value (0.003) was chosen to reflect the distribution of the variable which was bimodal with a clear cut off at the 0.003 mark. The list of variables used is given in table 12. The score building process used to develop the reliability score was repeated with the exception of the Recursive Features Elimination (RFE) process[6]. The score was developed by training models to predict the above average or below average levels of interaction. The resulting ROC curves are shown in figure 19.

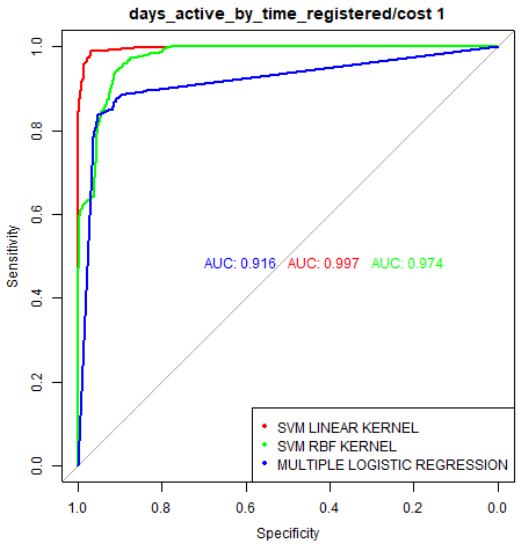


Figure 19 ROC curve for interaction score

[6] Excluded due to its long computation time

IV.2 VERIFICATION

Verification of the system prototype is an important part of ensuring that the prototyped system functions as intended. This is done by checking that the prototyped system satisfies all the system requirements it is meant to perform. It will mainly be carried out through demonstration and analysis (by testing the final scoring models using a portion of the original data). this verification plan is shown in table 14 and the result in table 15.

Table 14 verification plan

#	SYSTEM REQUIREMENT	METHOD OF VERIFICATION	VERIFICATION DETAILS	TIME	PASSING CRITERIA
2.1	The system shall collect mobile data from the data collection company for evaluation	Demonstration	• Use the system to collect mobile data for the specified farmers	• June 2017	• The system is able to collect the mobile data for the requested borrower credit rating
3.1	The system shall analyse the collected non-geospatial data in preparation for credit rating generation	• Analysis	• Make the system analyse the non-geospatial data for credit rating	• June 2017	• The system is able to analyse the non-geospatial data
4	The system shall generate a credit rating based on the risks of lending to the farmer	• Analysis	• Make the system generate a credit rating based on the risks of lending to the group of farmers. This is done by using the developed models to predict each type of score	• July 2017 •	• The system can generate each of the alternative scores which form the credit rating • System (scores) is able to predict scores with at least 70% accuracy

Table 15 verification result

#	SYSTEM REQUIREMENT	VERIFICATION DETAILS	PASSING CRITERIA	RESULT
2.1	The system shall collect mobile data	• Use the system to collect mobile data	• The system is able to collect the mobile data	• System was successfully able to collect the mobile data

	from the data collection company for evaluation	for the specified farmers	for the requested borrower credit rating	for the requested borrower credit rating
3.1	The system shall analyse the collected non-geospatial data in preparation for credit rating generation	• Make the system analyse the non-geospatial data for credit rating	• The system is able to analyse the non-geospatial data	• The system was able to analyse the non-geospatial data • This analysis is given in section IV.1.7
4	The system shall generate a credit rating based on the risks of lending to the farmer	• Make the system generate a credit rating based on the risks of lending to the group of farmers. This is done by using the developed models to predict each type of score	• The system can generate each of the alternative scores which form the credit rating • System (scores) is able to predict scores with at least 70% accuracy	• The system was able to generate each of the alternative scores which form the credit rating • System (scores) is able to predict scores with at least 70% accuracy for selected models

To verify the models used to predict each alternative score, 20% of the original data was used (the remaining data was used for model training, please see training results in System Prototyping section). The results of model testing for each type of score are given in the table 16 below:

Table 16 results of verification testing for alternative scoring models

MODEL	ACCURACY OF RELIABILITY SCORING MODEL	ACCURACY OF INTERACTION SCORING MODEL	ACCURACY OF REVENUE SCORING MODEL
Linear Support Vector Machine	**97.06%**	**100%**	**73.56%**
RBF Support Vector Machine	**100%**	**95.91%**	**66.66%**
Logistic Regression	**100%**	**95.91%**	**63.60%**

IV.3 VALIDATION

The core requirement of the system is to create scores that assess the risks of lending to smallholder farmers. In order to do this, risks were identified through consultation with various stakeholders as identified in the System prototyping section of this thesis. To analyse these risks, variables where selected through data analysis as part of the system operation. Validation exercises attempted to determine whether the selected variables were relevant for estimation of the risks evaluated. Interviews and a workshop were carried with stakeholders to ensure that the key stakeholder requirement had been met. Details of the validation plan are given in the table 16.

Table 17validation plans

ROLE	REQUIREMENTS	METHOD OF VALIDATION	VALIDATION DETAILS	TIME	PASSING CRITERIA
Financial institutions	1. Credit decision information on each potential borrower (farmers) that accurately estimates the risks of lending to smallholder farmers	Interviews Workshop	• Interview financial institutions to determine whether the credit rating given for the group of farmers agrees with what would be their own credit decision given the farmers' information • Interview farmers to determine whether the alternative scoring factors and the variables identified as important would affect the crop yield and revenue • Interview data collection to determine whether the alternative scoring factors and the variables identified as important would affect the crop yield and revenue	July 2018	• Financial institutions agree that the credit ratings are an accurate representation of the credit decision they would make given the farmers' information • Farmers agree that the factors evaluated by the alternative score would affect their crop yield and revenue • Data collection company personnel agree that the factors evaluated by the alternative score would affect their crop yield and revenue

i. DATA COLLECTION COMPANY

A workshop was held on 6th July 2018 with employees of the data collection company used for prototyping (Agribuddy LTD). The purpose of this workshop was to validate the outputs of the system so far.

Date: 6th July 2018

Workshop with: 18 employees of Agribuddy Cambodia

Location: Agribuddy Headquarters, Siem Reap, Cambodia

Objectives:

- Collect feedback concerning the validity of features identified through data analysis as being important for assessing the alternative scoring factors. In particular the reliability factor which was used to evaluate data collectors (buddies).

Details: Participants were asked to identify characteristics of a reliable user. This was done to counter-check whether features identified by participants would line up with those identified through analysis as being important. Factors identified included education, honesty, and providing accurate information. Secondly, participants were also asked to identify which activities or factors would increase revenue from farming activities. Some of the factors identified were use of technical farming methods, fertilizer, and size of labour force. Some of the work done during this workshop is shown in images 5 and 6[7].

[7] Details of the workshop are given in appendix (section IX.7)

Image 5 participant of validation workshop

Image 6 sharing results of 2 by 2 matrix

ii. **FARMERS**

The main beneficiary of the proposed system is the smallholder farmer who generates revenue through their farming activities. The system prototype assessed the revenue generation capacity of the smallholder farmer in rural Cambodia through analysis of collected data. As part of this analysis, the various factors affecting the crop yield and, thereby, revenue generation capacity of farmers were used. To validate that these factors did indeed have an impact on the crop yield (and by extension the revenue) interviews were conducted with farmers in and around Siem Reap, Cambodia.

Time: 7th and 8th July 2018

Location: Farm areas in and around Siem Reap

Objective: Validate that variables identified as being important for evaluation of the alternative scores were relevant

Details: The first day of field work was conducted in To Teong Thangai Village. Farmers in this village generally did not use irrigation. The second day of interviews was conducted in a village that was closer to the city (Siem Reap). A marked difference could be seen between this village and the one visited on the previous day. Farmers in the second village generally grew more than one crop, had built canals for irrigation, sold vegetables on a regular basis to middlemen to generate a more constant revenue. This was attributed to the proximity of the village to city, making it easier for NGO's to reach the village with training sessions for farmers and for middlemen to visit more regularly. The details of the interviews conducted are given in table 18.

Table 18 interviews with farmers for validation

Interviewee[8]	**Location**	**Date**	**Details**
A	To Teong Thangai Village	7th July 2018	• Has a 200m*21m plot of land which he uses for rice farming • Reuses his yield as seed for the next year's crop • At the time of the visit had a 6 month old crop of rice • Stated that his biggest problem was a lack of irrigation meaning he was reliant on rainfall

[8] Names are not used to protect the identity of interviewees

			• **The main factors that contributed to his yield are: Fertiliser, Weeding, and Water access** • He had pests last year and used chemical herbicide which was not effective • Sells his crop to middlemen • Harvest his crop by hand • Hires 15 labourers at 15000 local currency/labourer for harvesting, with the same done for transplanting • **He never felt the need to go to the bank directly, in part because he was too scared of what the requirements would be**
B	To Teong Thangai Village	7th July 2018	• Needs technical knowledge to improve his yield • **Identifies organic fertilizer and water as the most important factors for a high crop yield** • **Used to connect to the bank directly to obtain loans** • Currently has a loan • Stated that his biggest challenge was a lack of irrigation (he depends on rain) • Grows other crops to earn an income in the dry season
C	To Teong Thangai Village	7th July 2018	• Lives 30 minutes walk away from his farm ▪ **identifies trustworthy buddies in his community by the following factors: attitude, living situation, work ethic, social impact, respectability in community**
D	To Teong Thangai Village	7th July 2018	• has a 2 ha piece of land • At the time of the visit had a crop of rice that was 8 months old • Reuses part of her harvest as seed for the next year's crop • Expects to harvest 6 tons • **Most important factors to increase crop yield are irrigation, as well as fertiliser and herbicide** • Does not transplant • Never felt the need to ask the bank for a loan because she reuses her seed • Is eager to learn new techniques • Stated that she does not sell her crop, but then stated that she sell part of her crop to middlemen **this farmer did not seem to fully understand the concept of loan repayment with Agribuddy

E	Tatrav Village	8th July 2018	• Was in the process of transplanting rice at the time of the field visit with 10 hired labourers. The labour costs were roughly $4/day per labourer • Stated that his yield in the last year was 2 ton/ha • Felt that water was important for good yield
F	Tatrav Village	8th July 2018	• Was in the process of harvesting 200m^2 of spring onion and preparing it for market. This is something he is able to do about once a month with a different vegetable (spring onion, Chinese cabbage, etc.) • **Stated that the most important factor when it came to increasing his crop yield was irrigation.** • Has a canal which he uses for irrigation of his vegetables • His rice yield in the last year was 2tons/ha • Was taught newer farming methods by a non-governmental organization (NGO) including how to transplant rice and grow other crops. • Due to his proximity to the city was able to receive training and be closer to a market
G	Tatrav Village	8th July 2018	• In residence since 1984 • Grows spring onion and rice • Has 1.5ha of lad • Never calculates her yield • Sells crop to middlemen who visit the area • Uses irrigation (build 5 ponds for her own farm) • Has only 4 family members. She is currently working on her farm alone since her daughter is married with her own family and her husband is ill • **Feels that the most important factors to increase crop yield for rice are fertilizer (to make the rice grow) and herbicide (to kill weeds)** • Does not transplant her rice crop which leads to more weeds if herbicide is not used • **She has never desperately needed funds from a bank as she usually reuses part of**

			her previous yield as seed for the next year's farming

Image 7 in second village, farmers have better yields and are closer to the market

Image 8 farmers in second village transplanting rice

Image 9 farmers in second village grow other crops to supplement dry season income

Image 10works on her farm alone as her husband is ill

iii. FINANCIAL INSTITUTIONS

The intended user of the system output is the financial institution looking to lend to smallholder farmers. Therefore, the value of the system output to such a user had to be validated to determine whether the system output could be used by the financial institution to make lending decisions. This was done by conducting interviews with members of Aeon Specialized Bank Cambodia PLC, an institution that lends to locals in Cambodia and is looking to extend their services to smallholder farmers. Validation was conducted at the end of the prototyping process.

Date: 6th July 2018

Interview with: Mr Endo (Managing Director), Mr Sato (Head of E-business)

Location: Aeon Specialized Bank Cambodia PLC Offices, Phnom Pehn, Cambodia

Purpose: to receive validation from the perspective of potential lenders

Details: The two interviewees were shown the concept of operations and a mock-up of the system output. They explained that they currently have programs loaning to money for tractors to farmers. However, these are limited to those farmers who also some form of formal employment such as factory work. They expressed interest in the system output and stated that such a system would be useful in terms of allowing them to determine who had good personal characteristics as this was an important consideration for them when it came to lending. They stated that their biggest challenge in lending to farmers was verifying the farmer's identity. This is a challenge for the company given that many farmers will often have multiple identification documents. Therefore, data validity would be an important consideration for them before they could use the system. A second challenge they face is establishing the level of income/revenue has. The system output would be able to fulfil the need for this information.

V. DISCUSSION

V.1 NON-FINANCIAL DATA

The first objective of this research was to identify suitable methods of collecting non-financial data that do not require physical access and appropriate data for collection.

i. LIMITATIONS OF RELATED WORK

A review of related work showed that several data collection methods have been proposed for the evaluation of borrower's creditworthiness. Among these were social media services such as Facebook and Twitter. However, given that financially excluded smallholder farmers may not have access to such services or use them frequently, these methods are not viable for this application. Other methods applied included the use of data from mobile phone service accounts. However, such data collection methods may not be able to collect information that is specific to farming activities; information which is necessary for the evaluation a farmer's revenue. In addition, several researchers proposed the use of certain variables which may affect farming activities and farmer behaviour. However, there was no clear process given for the selection of such variables.

ii. PROPOSED SYSTEM DESIGN

The system design process proposed in this research suggests the use of a mobile application to collect data from farmers to make credit decisions. Data must be collected that can be used to evaluate the established alternative scoring factors. Collection of this data through a specially developed mobile phone service or application designed for smallholder farmers is proposed in addition to other data sources that may provide useful information such as Satellite imagery.

iii. PROTOTYPING

For prototyping purposes, data was collected by the data collection company, Agribuddy Ltd, through various means, including a mobile application and paper-based surveys. As expected, the data collected through the mobile application was easier to collect. Survey data took longer to collect as the data had to be written down on paper. The data was then passed on, sometimes over long distances, through company intermediaries until it reached the company. There it was inputted into the system by hand, a time-consuming process that required additional man power. On the other hand, mobile application data could be sent more easily at the click of a button. This process was less costly and time consuming, highlighting the major advantage of mobile data collection.

The data collected included farm activity data and personal information from farmers. Variables were allocated based on their relationship to the desired scoring factors[9].

In general, data collected from the mobile phone application and surveys provided interesting insights into the behaviour of both its users and the smallholder farmers they work with. From this data, it has been possible to identify the combinations of predictors which best predict the proposed alternative scores.

iv. SUGGESTIONS

a. Data Collection

A dedicated mobile application should be used for data collection. The information obtained can then be used to score other farmers quickly and efficiently based on their data. To improve the credit decision system, other factors which impact repayment of loans must be considered. This would require the collection of additional data. Since a mobile phone application is being used for data collection, the collection of new data simply involves adding a new data input field into the application. This brings the focus to a major advantage of using mobile phone application to collect data for making credit decisions. Large amounts of data are collected from users daily, meaning that the data is current and presents an image of the user's current situation. By adding such data fields, a more complete picture of the farmers can be created. Further, data can be collected by other means such as drones and Satellites to supplement data collected through the mobile application. The strategy used for data collection must ensure that the data required for evaluation of each alternative scoring factor is collected successfully. While constraints may make it impossible to collect all relevant data, the key to this process is to establish a balance that allows for the key scoring factors to be analysed.

b. Data Accuracy

The risk of data fraud is especially high given that the data is essentially self-reported by borrowers. This risk can be categorized into two types: **intentional** and **unintentional**. Intentionally inaccurate data would result from farmers maliciously sending in information that they know to be false to improve their credit decision

[9] The importance of each factor with regards to predicting each score was determined by performing logistic regression with the variable denoting the scoring factor as the dependent variable and each potential data type used as an input in turn.

outcome. Although it is not possible to entirely eliminate this risk, it can be mitigated by ensuring that the system performs data validation. One possibility is to counter-check information on farm activities provided by farmers with satellite data to verify the former's accuracy. For instance, Normalized Vegetation Indices can be used to observe farm areas and identify the timing of farm activities such as ploughing, seeding, and harvest. This information can then be checked against the information given by farmers to determine whether the farmer was truthful. Unintentionally inaccurate data may be reduced by communicating to farmers the importance of data accuracy to obtaining their desired outcome from the system.

c. **Data privacy**

Issues of data privacy have come to the fore in recent years. Companies such as Facebook have been embroiled in scandals because of their failure to prevent malicious use of their user's data. Such occurrences only serve to emphasise the need for data privacy when dealing with personal information. Not only is data privacy a legal requirement in many countries, it also helps to build trust between users and system operators which is an important part of building a sustainable system. User privacy should be ensured by masking, i.e. information that can be used to identify particular users should not be shared frivolously. Privacy considerations should made during **data storage** as well. This implies using storage platforms that are completely secure and are only accessible by authorized personnel.

V.2 DATA ANALYSIS

The second objective of this research was to identify appropriate data analysis methods for evaluation of alternative scoring factors with mobile data.

i. LIMITATIONS OF RELATED WORK

Classification is the typical data analysis method used for credit scoring. A look at the existing research in this field shows that methods of varying complexity have been applied for this purpose, including Logistic Regression, Support Vector Machines, and Neural Networks among others. However, there is no clear choice for the analysis of data in this context: where non-financial data is being used to make credit decisions for financially excluded smallholder farmers. In particular, the fact that smallholder farmers are part of the intended audience of the system output means that simpler methods must be considered. This is in addition to other factors such as balancing accuracy, cost of errors, and computational cost concerns.

ii. PROPOSED SYSTEM DESIGN

The proposed design suggests the use of analysis methods with a balance of accuracy, complexity, and computational cost. The appropriate classification method should be selected based on specific context by first comparing a series of applicable methods and comparing based on the suggested evaluation criteria.

iii. PROTOTYPING

To identify a suitable classification method for such analysis, Multiple Logistic Regression, Support Vector Machines with a Linear Kernel and Support Vector Machines with a Radial Basis Function Kernel were compared to determine the most apt method. The results were evaluated by Area Under the Receiver operating curve (AUC), where higher values of both indicate a well performing model. In terms of ROC curves, the Support Vector Machine with a radial basis function kernel was found to be the better performing method for the reliability credit scoring factor. However, it was outperformed by the Support Vector Machine with a linear kernel when applied to the other two scores.

Other considerations must be made when it comes to selecting the final model for implementation. In choosing the most appropriate analysis method to used when developing credit decision systems in Agricultural Microfinance, we must consider not only accuracy but also training time and ease of explanation. Other researchers comparing analysis methods of varying complexity for credit scoring have reached the conclusion that accuracy and model performance should not be the only reasons given for selection of analysis methods. Factors such as computational expense and complexity must also be considered. When the results are considered in this light it becomes apparent that although the Support Vector Machines had better performance, the long training time that would be necessary if the credit decision system was to be scaled up to larger datasets makes Support Vector Machines the less attractive options. The radial basis function kernel could be the most computationally expensive as it must map the features to higher dimensions. On the other hand, it may become beneficial to apply the more complex radial basis function (RBF) kernel model if new features are added which have complex relationships. The ease with which the results could be explained to potential borrowers is a further consideration, especially to those with low financial literacy, as may be the case among the financially excluded. Multiple Logistic Regression perhaps has the advantage here since the coefficients of the model make for an intuitive understanding of the model and the parameters impacting it. Support Vector Machines do not possess this

advantage. This could be overcome, however, by ranking the features of the model by their impact on classification.

Table 16 shows accuracies of the models as calculated based on out-of-sample tests on 20% of the original data. The test results gave near perfect accuracy for Reliability and Interaction scores. Although this result appears to be ideal, it should be noted that the sample size used for these two scores was small and so could have been strongly affected by any biases in contained. Therefore, testing should be repeated on larger sets of out-of-sample data if possible. Further data analysis is necessary for validation as well as to establish the most suitable algorithm.

iv. SUGGESTIONS

a. Accuracy

Accuracy is a commonly used metric to evaluate classification models. This is because a model with low accuracy could lead to a high rate of misclassification. For credit scoring, misclassification results in high risk borrowers being given loans while low risk borrowers are denied. Since this would prove to be expensive for the financial institution, high accuracy is a desirable feature in data analysis.

b. Complexity

Complexity of the selected data analysis methods plays a role in the performance of the system. Although more complex methods may give better accuracy, they may not be as easy to explain to stakeholders as simpler models. As such, consideration must be made of the complexity of the model used for the system.

V.3 ALTERNATIVE SCORING FACTORS

The design of methods of alternative scoring factors independent of financial data was a key objective of this research.

i. LIMITATIONS OF RELATED WORK

Existing research has focused on the development of credit scores using financial data. For instance, risk of default (which is a commonly used scoring factor) requires financial history data for model training. This requirement is difficult to fulfil when no financial data is available.

ii. PROPOSED SYSTEM DESIGN

A method of creating Credit Decision System was proposed along with a process of designing alternative scoring factors that are independent of the borrower's credit history. It was proposed that this be done by asking relevant stakeholders to identify the risks that farmer face. These

risks can then be evaluated using available data to create alternative scores. The proposed system has the benefit of requiring no financial history data as is typically required for credit scoring. As lending decisions are made and repayment patterns are observed, the alternative scoring factors can be validated by comparisons to loan repayment rates. Validation exercises must be conducted to validate the alternative scores by comparing them to actual loan repayment rates. If a higher repayment rate is observed among those with higher alternative scores, it can be said that the alternative scores are a valid alternative to existing credit scores in this context.

iii. PROTOTYPING

The Alternative scoring factors were created by considering the core requirement of a Credit Decision System which is to indicate the risk associated with lending to a particular borrower. To this end, risks associated with lending to smallholder farmers were identifed through discussions with stakeholders. A workshop was conducted with personnel at Agribuddy LTD who work with farmers on a daily basis. Risks that may affect the farmer were identified by asking workshop participants what risks farmers and buddies[10]. Interviews conducted with financial institution and farmers corroborated the risks identified. Risks identified were then condensed into categories. Selected categories were turned into Alternative scoring factors and evaluated.

In this context, three general risks were selected for analysis. These were risks concerned with trust among stakeholders, risks concerning reliability of data, and risks concerned with the borrower's ability to generate sufficient revenue by producing and selling crops. This resulted in three credit factors: an interaction scoring factor, a reliability credit scoring factor, and revenue credit scoring factor, respectively. Using these types of factors provides the lender with a better understanding of the borrower than can be gained from either one. It is interesting to note that these types of factors actually represent two of the five traditional credit factors. The Reliability and Interaction factors can be likened to the "Character" aspect of the 5 C's used in traditional credit analysis while the Revenue factor determines the "Capacity".

iv. SUGGESTIONS

a. Context Specific risks

[10] These are people who collect data from farmers in the prototyping system on behalf of the data collection company.

One of the basic tenets of Systems Engineering is the development of a system based on its context. In line with this, the specific risks identified may differ depending on the target group of borrowers and the context in which the Credit Decision System is being used. Therefore, the risk identification process must be repeated in each new context to determine what the risks are in that situation. Although the risks and evaluation factors chosen here are specific to agriculture and smallholder farmers in particular, the interdisciplinary nature of Systems Engineering means that a similar approach can be used to identify risks and design evaluation criteria suitable to any sector.

b. **Data availability**

Data availability played a role in selecting the exact scoring factors that were evaluated. This highlighted the impact of data collection on the types of evaluations that can be collected when alternative data is used. Where the data is available, as many of the identified risks as possible should be evaluated while each additional credit risk evaluation factor should be selected to represent a new perspective on the borrower and address the key concerns of stakeholders in the given scenario.

V.4 PROTOTYPING

A prototype was developed to check the feasibility of specific system functions. The system functions prototyped included the data collection, pre-processing and score development functions. This process involved the following: data collection, identification of relevant, design of alternative scoring factors, and analysis to determine the most suitable analysis methods. Prototyping is made possible by a collaboration with an agribusiness firm in Cambodia, Agribuddy Ltd., which collects data from local farmers through a mobile application service.

i. PROTOTYPE VERIFICATION

A portion of the data collected was used to test the final models developed with resulting accuracies being used to determine whether the final models would be able to properly categorize potential borrowers. the results of this analysis are given in the verification section of this thesis. It is clearly seen that the final model performance is dependent on the type of score being assessed and more specifically to the data used to develop the model.

The scoring models developed solely with the used of data collected via the mobile application showed performance that was significantly better than that shown by the scoring models

developed with data collected via the field survey. This raises the questions of which data collection method provides more accurate data.

ii. PROTOTYPE VALIDATION

Validation with various stakeholders was carried out. Financial institutions showed great interest in the system and in the concept of evaluating different aspects of the farmer. However, they stated that other types of data should also be collected before they could use the system to make lending decisions. They also stated that data reliability would have to be ensured.

Data collection company personnel agreed with some of the factors that had been used to evaluate the revenue scoring factor, showing that this type of data was appropriate for the desired analysis. Other factors were also introduced, including "networking to identify better market opportunities". These may be the subject of future prototyping efforts to improve system output. When it came to variables that could be used to identify reliable users (reliability score), variables such as "providing accurate information" were mentioned. This is a factor that is evaluated in the reliability score (since reliability score is based on user classification which is decided by such factors). However, some of the answers given were more ambiguous variables that would be difficult to quantity. Examples include answers such as "a person who is mature" Therefore, further consideration must be made of how these variables can be quantified for analysis purposes.

Upon interviewing farmers, it was found that many of the variables used for evaluation of the revenue score were important for their own crop yield (and thus revenue). These included use of irrigation, fertilizer, and transplanting. A distinct difference was observed between the perception farmers had of what factors had the greatest impact on crop yield based on the village they resided in. Farmers living in the same village had the same or similar perception. This introduces the idea that group behaviour may impact the behaviour of a single farmer. Therefore, farmers living in areas where good farming practices were used may be more likely to use good farming practices also.

It was also noted that the needs of smallholder famers differ. Not all farmers need credit services. These farmers are willing to get by using only available resources to reduce costs. For instance, some farmers are willing to use a portion of their harvest as seed for the next cycle of planting. Further, some farmers already had access to credit services. In this case, what they require is knowledge of farming techniques that can improve their crop yield. Thus, it should be note that, the designed system would not fulfil all the smallholder farmer's needs.

As only the reliability and revenue scores were validated here, the interaction score must also be validated separately before the system can be implemented.

iii. ADDITIONAL PROTOTYPING

The prototype system gave promising results. However, a larger prototype must be developed to determine the feasibility of other system functions such as the collection and analysis of satellite data. In addition, prototyping with a larger data set would allow for further insights to be obtained before actual implementation. Developing a prototype in other countries may also allow for insights on risk factors that would be at play in other contexts.

V.5 BENEFITS OF SYSTEM IMPLEMENTATION

a. ADVANTAGES FOR STAKEHOLDERS

Based on the problem analysis, the problems experienced by the key stakeholders that can be solved by this system are as shown in table 19.

Table 19 advantages for system stakeholders

STAKEHOLDER	CHALLENGE	BENEFIT FROM SYSTEM
Financial institution	Misses out on opportunity to increase market size (and revenue) by extending credit services to financially excluded farmers	Able to increase market size (revenue) for credit services while managing risk of lending
Financially excluded smallholder farmers	Unable to access credit services which can increase investment (and revenue) from farming	Able to obtain credit services and thereby increase investment in farming activities and revenue
Data collection company	Misses opportunity for additional revenue stream	Additional revenue stream from providing collected farmer data to system operator for use in evaluations

The data collection company was not included in the initial problem analysis since the issue of financial exclusion of farmers does not directly impact it. However, the system presents an opportunity for the data collection company to increase its revenue by introducing a new revenue stream. The company would have to invest in the collection of specific types of data

needed for analysis. Although, this may go beyond the scope of their current operations, they will be able to recoup their expenses by charging the system operator for use of the data.

ii. CUSTOMER VALUE CHAIN ANALYSIS

A key consideration when designing the operation of the system is the flow of value between stakeholders. It is essential that each stakeholder receives a benefit for their role in the system, thereby ensuring that they feel invested in doing their part to sustain the system. This analysis was carried out using the Customer Value Chain Analysis (CVCA) method. **Smallholder farmers who cannot access credit services from financial institutions** but are able to use mobile phone services, may experience challenges receiving value int the form of loans from these institutions. On the other hand, the financial institutions themselves face challenges when it comes to receiving information and profit from farmers. Therefore, before the implementation of the Credit Decision System, we see the scenario shown in figure 6 in which there is no value flowing between the financial institutions and the remaining stakeholders in this context. Rather the remaining stakeholders, interact amongst themselves, exchanging information, money and farm produce.

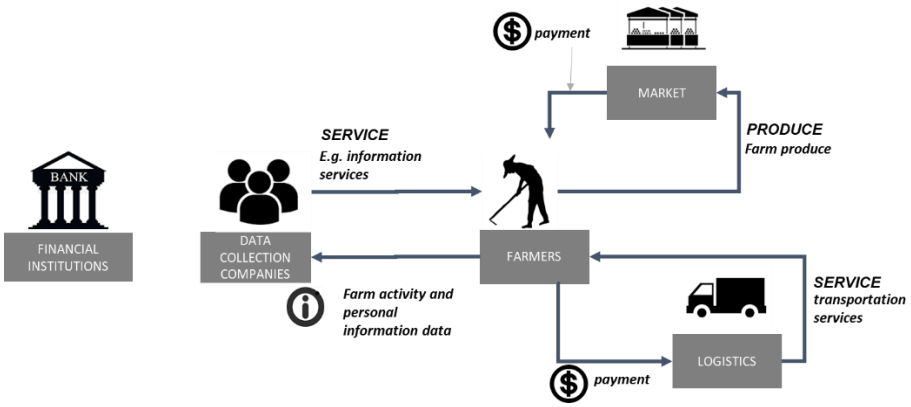


Figure 20customer value chain before implementation of new system

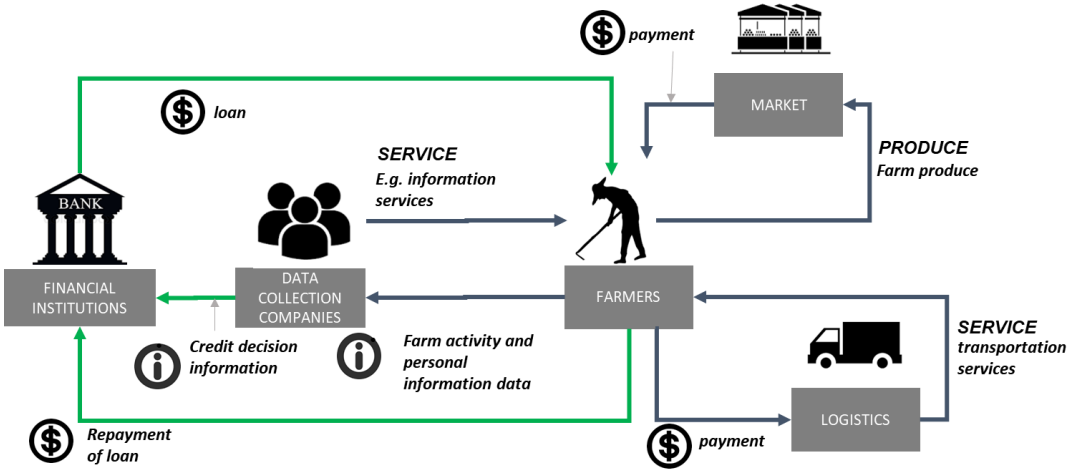


Figure 21 customer value chain after implementation of new system

V.6 GENERAL SUGGESTIONS FOR IMPLEMENTATION

Implementing the system in real world context may bring with it concerns beyond those factored into this design. Some of these are discussed here.

i. CULTURAL FACTORS:

Cultural factors may play a role in the behaviour of farmers in a given area. For instance, several farmers interviewed stated that they had learned their current farming methods from their parents. Farming methods and traditions passed down through the generations may be outdated or may benefit from new information gained through modern techniques. In affecting the farming methods, these factors also impact the revenue that can be generated by each farmer and so their ability to repay any loans. As such, cultural factors with a high impact on the behaviour of system users should be considered as part of the design process.

ii. SUSTAINABILITY

Farmers should not be lent money at exorbitant interest rates or in amounts that they would be unable to repay with income from their farming activities. Doing this would lead to an unsustainable system where the farmers use money from other sources to repay loans. Eventually, farmers without external sources of income would be unable to repay the loans, leading to default.

iii. GOVERNMENT POLICY

Government policies regarding agriculture vary by country. Countries with governments that provide more support for their farmers by providing subsidies or providing farming inputs may

have farmers with a higher, more stable income. In this case, the needs of the farmer will be different, and the system may have to provide different information to stakeholders.

iv. LEGAL RESTRICTIONS

Legal restrictions will affect the system operation. An example would be restrictions on the sharing of personal data with third parties without the permission of the farmer. This would affect the ability of the data collection company to provide information for system development.

v. EDUCATION:

The level and quality of education in the implementation context will affect the amount of access the farmer will have to information and financial resources. Better access may translate into a reduced need for the system. Further, language barriers between the financially excluded farmers and financial institutions may also be lowered if the farmer has better access to education, thus removing one of the key reasons for financial exclusion.

vi. **ECONOMIC STATUS:**

The economic status of farmers benefitting from this system will also impact the manner in which it may be deployed. Generally, in countries with a higher level of economic development, farmers like all citizens are likely to have higher income and be financially included. This may mean there is less demand for such a system in this context. They may also have access to more technology to increase their farming output.

VI. CONCLUSION

The goal of this research was to design a Credit Decision system which could be used by financial institutions to lend to financially excluded farmers. In order to solve the issue of credit access for financially excluded farmers, a Credit Decision System was designed. This system tackled the three objectives which were focused on non-financial data, data analysis, and design of alternative scoring factors.

The design process considered the data collection, analysis and evaluation methods existing in related research. This involved the technical processes of Requirements Analysis and Architectural Design. Since one of the main issues for the financially excluded is data collection due to long distance, data collection methods not requiring physical access were needed for the collection of data which was **non-financial** in nature. The design had various data sources for data collection including mobile data and satellite data. These methods provide **convenient and speedy data collection**. The key requirement of the system was to create scores based on the risks of lending and so for implementation, it was stated that risks must first be identified through discussions with stakeholders. Machine learning classification techniques were proposed for data analysis along with GIS tools for geospatial analysis.

A prototype was developed using data from farmers in rural Cambodia. The prototype focused not only a portion of the system functions and was designed to show the feasibility of the system's core functions. As part of this process, risks were identified by stakeholders and used to create the following **Alternative scoring factors: reliability score, revenue score, and interaction score**. These were found to be similar to existing factors used for credit analyis. Therefore, it can be said that the factors identified through the designed system aligned with current methods. Data was collected through a mobile phone application (**mobile data**) and a survey. Types of data collected included personal information such as ages, number of family members, and farming activity information such as farm size, seeding and harvesting times.

This data was applied to train scoring models for the alternative scores using three machine learning classification methods: **Multiple Logistic Regression**, **Support Vector Machines with a Linear Kernel**, and **Support Vector Machines with a Radial Basis Function Kernel**. Feasibility of applying each method in agricultural microfinance was considered. When these were compared to find the most suitable, it was decided that the selected option should be based on accuracy, complexity, and cost of computation in a given context.

Verification was successfully carried out through demonstration and analysis, while Validation was done through interviews and workshops. The latter showed that the main system user, financial institutions, would be interested in using the system output provided data reliability could be ensured. Further, some of the variables that had been used for evaluation of alternative score were found to be important as expected. However, new variables were also introduced which can be used to improve the system's performance in the future. Since only the reliability and revenue scores were validated, the interaction score must be validated separately in the future. Lastly, some additional considerations that must be made when using this system design were discussed. Some of these were the issues of data fraud and privacy. Other context specific concerns such as legal restrictions and government policy must also be factored into the design. This research could form the foundation of a credit decision tool using mobile application data. It has the potential to reduce the shortfall in financing for smallholder farmers by providing financial institutions with a viable means of assessing credit risk and thus encouraging them to enter the market of lending to smallholder farmers. This will in turn help to increase the productivity of smallholder farmers and allow them to continue to contribute in vital ways despite the changing global landscape. Additionally, it will allow for more lending institutions to enter the business of lending to smallholder farmers at a reduced risk to themselves.

VII. FUTURE WORK

Building on the results of the research conducted by this researcher at Master's Degree level, future research will refine the credit scoring factors identified previously. This will be done by considering the validation results of the previous research. In addition, the Interaction score must also be validated through interviews with relevant stakeholders. The relationship between credit scoring factors and loan repayments rates will be explored to determine their correlation. Based on this analysis, the credit scoring factors will be refined to improve their ability to predict loan repayment. Additionally, other credit scoring factors that may impact loan repayment will identified by communicating with relevant stakeholders, including lending institutions, farmers, and data collection companies. Potential considerations include current financial obligations and market prices. Flood and drought prediction will also be conducted since these present significant risks to crop yield and, as a result, loan repayment. Research has shown that two of the major reasons for default on agricultural loans are adverse weather affecting the crop cycle and market fluctuations in crop prices. The evaluation of these new credit scoring factors will require the collection of additional data. Satellite data may also be collected to determine the Normalized Difference Vegetation Index (NDVI) and Normalized

Difference Water Index (NDWI) which can be used to monitor crop growth and water content in each farmer's area. Mobile phones can be used to collect location data which can be analysed to determine the farmer's distant from market which could impact market prices, among other factors. To analyse this data, several analysis methods can be used. Geospatial data will be analysed using QGIS, an open source geospatial analysis tool. Additional analysis tools may also be applied. For instance, Markov Chains may be used to predict future behaviour from currently available data.

VIII. REFERENCES

[1] World Bank, “Financial Inclusion,” World Bank, 20 04 2018. [Online]. Available: http://www.worldbank.org/en/topic/financialinclusion/overview. [Accessed 18 05 2018].

[2] FAO, IFAD, UNICEF, WFP and WHO, “The State of Food Security and Nutrition in the World 2017. Building resilience for peace and food security,” Food and Agriculture Organization, Rome, Italy, 2017.

[3] World Bank, “Agriculture, value added (% of GDP),” 2015. [Online]. Available: http://data.worldbank.org/indicator/NV.AGR.TOTL.ZS?end=2016&start=1960&view=chart. [Accessed 30 06 2017].

[4] Interantional Labour Organisation (ILO), “Employment in Agriculture (% of total employment),” Wolrd Bank, 2010. [Online]. Available: http://data.worldbank.org/indicator/SL.AGR.EMPL.ZS?end=2010&start=2010&view=bar&year=2010. [Accessed 18 12 2016].

[5] International Fund for Agricultural Development (IFAD), “Smallholders, food security, and the environment,” International Fund for Agricultural Development (IFAD), 2013.

[6] Food and Agriculture Organization of the United Nations, “Small holders and Family Farmers,” Food and Agriculture Organization of the United Nations, 2012.

[7] Food and Agriculture Organization of the United Nations, “The State of Food Insecurity in the World,” Food and Agriculture Organization of the United Nations, 2015.

[8] World Bank, “Ending poverty and Hunger by 2030,” World Bank, 2013.

[9] K. S. R. Singh, “Climate change,agriculture and ICT: an exploratory analysis,” in *ICT for Agricultural Development under changing climate*, New Delhi, Narendra Publishing House, pp. 17-28.

[10] "Agricultural Productivity for Sustainale Food Security in Asia and the Pacific: the role of investment," in *Agricultural Investment nad Productivity in Developing Countires*, Economic and Social Development Department (FAO).

[11] United States Agency for International Development, "Guide to the Use of Digital Financial Services in Agriculture," United States Agency for International Development, 2016.

[12] World Bank, "Global Financial Index," World Bank, 2014.

[13] G. G. M., "Five challenges prevent financial access for people in developing countries," World Bank, 15 10 2015. [Online]. Available: https://blogs.worldbank.org/voices/five-challenges-prevent-financial-access-people-developing-countries. [Accessed 15 08 2017].

[14] Interntional Labour Organisation, "Agriculture; plantations; other rural sectors," 2015. [Online]. Available: http://ilo.org/global/industries-and-sectors/agriculture-plantations-other-rural-sectors/lang--en/index.htm. [Accessed 18 12 2016].

[15] World Bank, "Indicators," World Bank, 2014. [Online]. Available: data.worldbank.org/indicators. [Accessed 07 12 2017].

[16] Agriculture and Consumer Protection Department, "Farm Management for asia: a systems approach," Food and Agriculture Association, 2010. [Online]. Available: http://www.fao.org/docrep/w7365e/w7365e05.htm#TopOfPage. [Accessed 18 12 2016].

[17] K. L. S. D. A. S. H. J. Demirgüç-Kunt A., "The Global Findex Database," World Bank Group, Washington DC, USA, 2017.

[18] M. McEvoy, "Enabling financial inclusion," [Online]. [Accessed 15 08 2017].

[19] F. J., “Alternative Data and Fair Lending,” LexisNexis, 2013.

[20] M. N. Gordon M., “CFPB Insights on Alternative Data Use in Credit Scoring,” Law360, New York, USA, 2017.

[21] International Telecommunications Union, “Global and Regional ICT Data,” International Telecommunications Union, 2017.

[22] D. A. K. M. Costa A., “Big Data, Small Credit: The Digital Revolution and Its Impact on Emerging Market Consumers,” *Innovations: Technology, Governance, Globalization,* vol. 10, no. 3-4, pp. 49-80, 2015.

[23] SE Handbook Working Group, Systems Engineering Handbook, San Diego, California, USA: International Council on Systems Engineering, 2011.

[24] E. B. C. N. Thomas L.C., Credit Scoring and Its Applications, Philadelphia: Society for Applied and Industrial Mathematics, 2002.

[25] B. B. S. H. e. a. Lessmann S., “Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research,” *European Journal of Operational Research,* vol. 247, no. 1, pp. 124 - 136, 2015.

[26] R. V. Kumar P.R., “Bankruptcy prediction in banks and firms via statistical,” *European Journal of Operational Research,* vol. 180, pp. 1 - 28, 2007.

[27] E. H. F. T. M. Charpignon, “Prediction of Consumer Credit Risk,” Standford University.

[28] Y. Z. X. Li, “An Overview of Personal Credit Scoring: Techniques and Future Work,” *International Journal of Intelligence Science,* no. 2, pp. 181-189, 2012.

[29] C. M. L. Tsai C. F., “Credit rating by hybrid machine learning techniques,” *Applied Soft Computing,* vol. 10, pp. 374-380, 2010.

[30] M. Haltuf, “Support Vector Machines for Credit Scoring,” University of Economics in Prague, Faculty of Finance, Prague, 2014.

[31] P. W. D. B. Ntwiga, “Consumer lending using social media data,” *International Journal of Scientific Research and Innovative Technology,* vol. 3, no. 2, 2016.

[32] P. W. B. D. Ntwiga, “Credit Scoring for M-Shwari using Hidden Markov Model,” *European Scientific Journal,* vol. 12, no. 15, 2013.

[33] A. Masyutin, “ Credit scoring based on social network data,” *Business Informatics,* vol. 33, no. 3, pp. 15-23, 2015.

[34] D. a. D. Grissen, “Behavior Revealed in Mobile Phone Usage Predicts Loan Repayment,” Brown University, Department of Economics, 2016.

[35] D. D., Risk Elements in Consumer Instalment Financing, New York, USA: National Bureau of Economic Research, 1941.

[36] J. C. Johnson J., “GIS : A Tool for Monitoring and Management of Epidemics,” *Epidemiology*, no. February 2001, pp. 1-6, 2001.

[37]]. S. D.-M. E. E. L. Bojorquez-Tapia, “GIS-based approach for participatory decision making and land suitability assessment,” *International Journal of Geographical Information Science,* vol. 15, no. 2, pp. 129-151, 2001.

[38]]. Z. B. S. K. H. J. Bryne, “Evaluating the potential of small-scale renewable energy options to meet rural livelihoods needs,” *Energy Policy,* vol. 35, no. 8, pp. 4391-4401, 2007.

[39] B. S., “GIS in Transport Modelling Svante Berglund,” Swedeish Royal Institute of Technology, DEPARTMENT OF INFRASTRUCTURE AND PLANNING, 2001.

[40] R. O. J.C. Nwaru, “Credit Use and Technical Change in Smallholder Food Crop Production in Imo State of Nigeria,” *New york Science Journal,* vol. 3, no. 11, 2010.

[41] S. P. AMWA Herath, “Social factors influencing agricultural productivity in the non plantation agriculture in Sri Lanka,” in *27th Annual Conference of the Organization of Professional Associations of Sri Lanka*, Colombo, 2014.

[42] Economic Research and Regional Cooperation Department, ADB, “Basic Statistics 2017,” Asian Development Bank (ADB), 2017.

[43] National Institute of Statistics, Ministry of Planning, “Census of Agriculture in Cambodia 2013,” Ministry of agriculture, Forestry and Fisheries, 2015.

[44] Food and Agriculture Organisation of the United Nations, “FAO Statistical Pocketbook World Food and Agriculture,” Food and Agriculture Organisation of the United Nations, 2015.

[45] World Bank, “Global Financial Inclusion Statistics,” World Bank Group, 2014.

[46] K. M., “A Short Introduction to the caret Package,” 28 10 2016. [Online]. Available: https://cran.r-project.org/web/packages/caret/vignettes/caret.pdf.. [Accessed 05 07 2017].

[47] “An introduction to ROC analysis,” *Pattern Recognition Letters,* vol. 27, pp. 861 - 874, 2006.

IX. APPENDICES

IX.1 SYSTEM MODEL DIAGRAMS

SysML was used to create models of the Credit Decision System. This was done to assist in the design process. Some of the images from this modelling process are shown below:

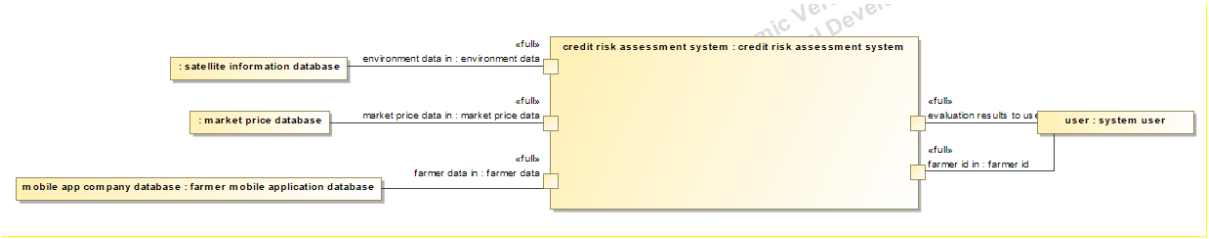


Figure 22 system context diagram

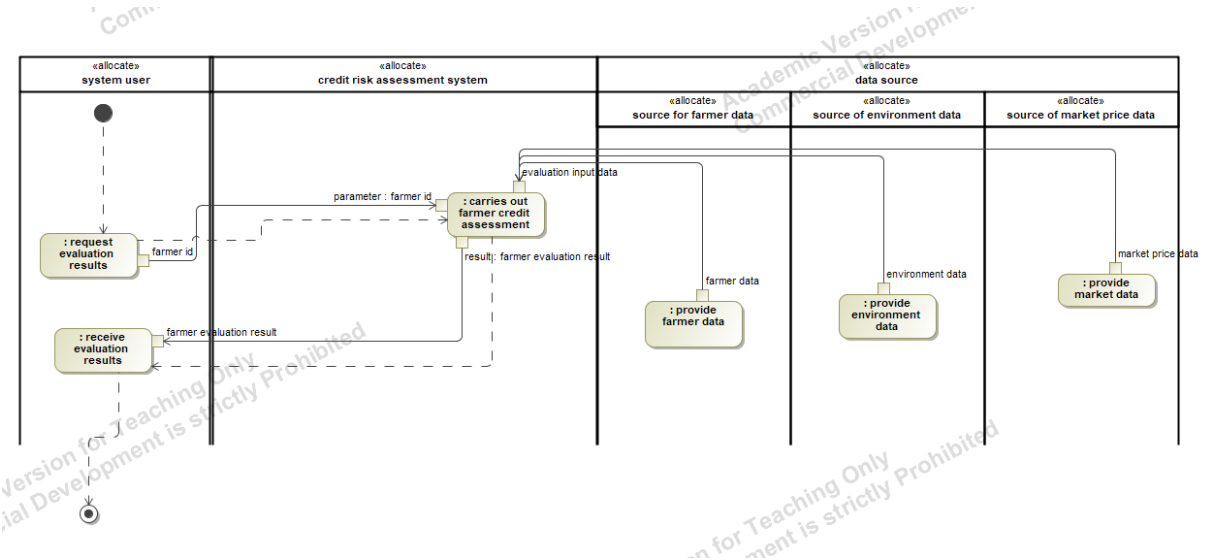


Figure 23 activity diagram of system context

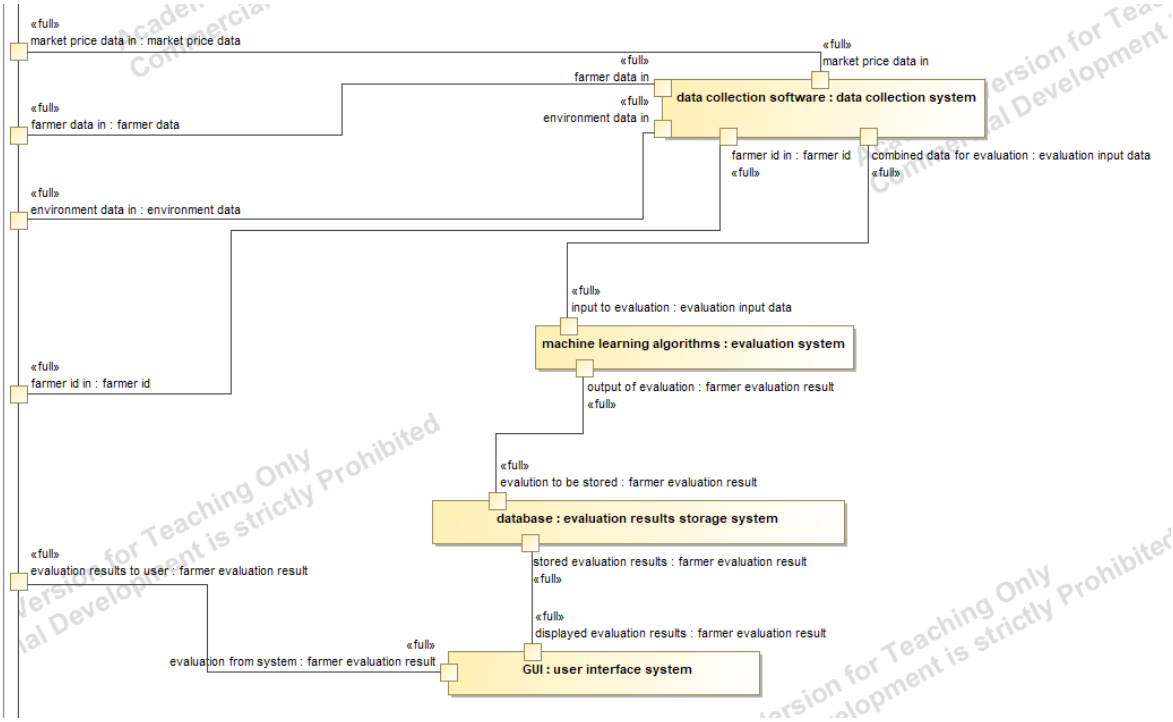


Figure 24 internal system block diagram

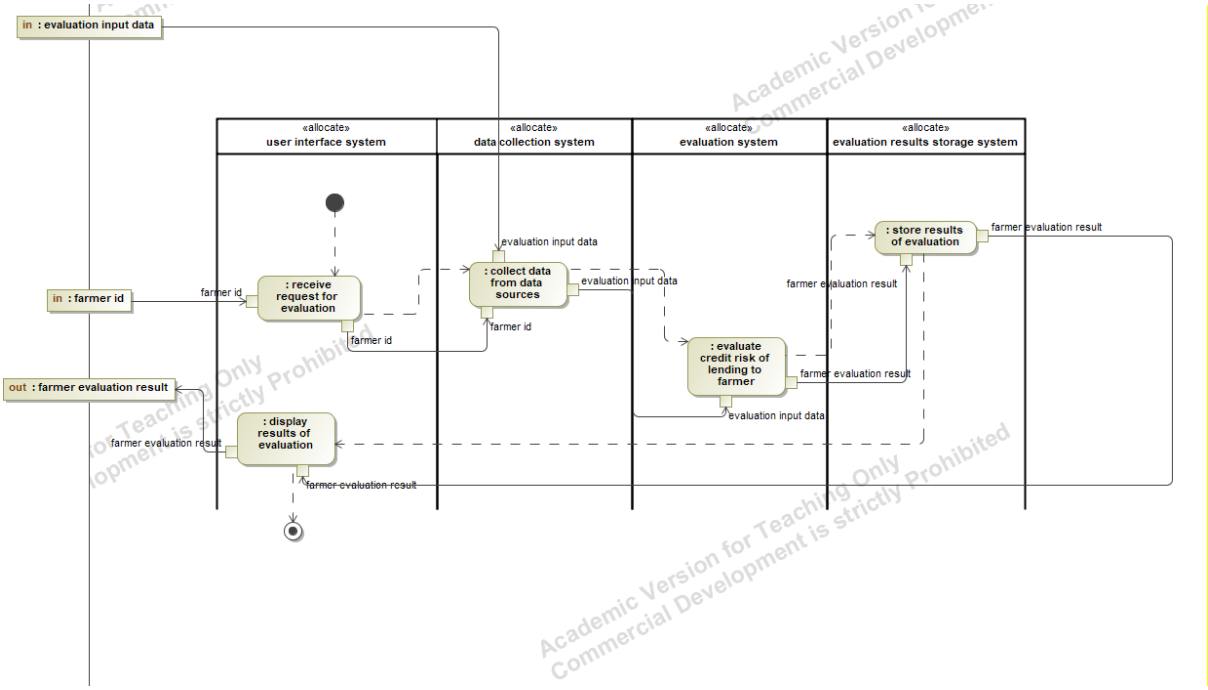


Figure 25 internal system activity diagram

IX.2 OVERVIEW OF AGRIBUDDY Ltd OPERATIONS

Agribuddy is a company operating in the field of agriculture in Cambodia, Mozambique and India. The company operates a mobile platform that offers various services to farmers and other players in the agricultural sector. A class of users referred to as "buddies" use the mobile platform directly to deliver reports on farming activity in their areas. A user of the Agribuddy mobile platform who registers and collects data from local farmers. A buddy is generally a member of the community who has a smartphone. He/she may or may not be living in a rural area. Each buddy can register with a maximum of 50 farmers. Each buddy is responsible for collecting and reporting information from the farmers under their purview. Through these activities, the company has been able to collect massive amounts of data on the farmers. Buddies garner points for each piece of data they upload onto the platform. When the minimum number of points has been reached, the buddy can redeem these points for cash. This is done by making an application on the platform. Buddies are divided into 3 classes according to their characteristics and ability to fulfil certain criteria.

- **Gold Buddies**: this class of buddies is chosen out the pool of buddies based on given criteria[11]. They are trained in the best data collection and reporting practices. Gold buddies who are caught falsifying data are removed. Gold buddies will enter a legal relation relationship with the Agribuddy wherein they will be trusted with loans and the responsibility of acting as agents of the company. Hence, Agribuddy will be incurring a legal and financial risk by working with these farmers and a selection method which identifies low risk buddies is needed.
- **Silver Buddies**: This is the initial classification each buddy is given upon joining the Agribuddy platform. It is only assigned once a buddy has uploaded both personal information and farm area scaling data correctly and the company has confirmed that the user is updating the correct information.
- **Failed Buddies**: these are buddies who have uploaded incomplete or untrue personal information and farm area scaling data. Further, they may have refused to make corrections when asked to do so by the company.

Mentors are employees who assist in training and advising buddies. Many of these are buddies who were identified as being exceptionally cooperative. They are responsible for training gold buddies and helping them connect with the bank. They are also responsible for marketing agriculture inputs to buddies and other farmers.

Call agents are employees of Agribuddy who are responsible for maintaining contact with buddies and farmers via phone. They call buddies to confirm details uploaded to the mobile platform when necessary. Additionally, they may also call a random selection of farmers registered with a buddy to check the validity of data provided by the buddy. Lastly, call agents check the satellite images of farm areas scaled by buddies to ensure that it was done correctly and without fraud.

IX.3 DATA COLLECTION

i. MOBILE APPLICATION DATA COLLECTION

FEATURE NAME	MEANING	FEATURE NAME	MEANING
classification	good or bad classification	activated_areas	number of farms registered correctly
num_reports	total number of reports sent by user	num_of_point_actions	number of of point actions
num_farmers	number of farmers user reports data from	num_of_calls	number of calls from the mobile application company
num_farms_not_reported	farms registered but not reported on	mean_period_active_owner_days	average days farmers have used the service
days_since_sign_up	days since joining the service	num_male_owners	number of male farmers
user_age	age of user	num_female_owners	number of female farmers
user_length_of_residence	years lived in current location	mean_length_of_residence_owner_years	years average farmer has lived in current location
Outside.of.4km.	registered farmers living outside 4 km radius of user	mean_owner_age	average age of farmers
Within.4km.	registered farmers living with 4 km of user	mean_owner_num_of_family	average number of family members of farmers
Total.number.of.areas.	number of registered farm areas	num_farms_on_record	number of registered farms
num_locations	number of locations user has reported from	mean_period_active_invitee_days	average days person invited by user has used the service
crop_variations	number of crop variations grown by user's farmers	num_of_invitees	number of persons invited by user
num_farms_reported	number of farms user has registered and reports on	Number.Of.Family	number of family members of farmers
num_days_active	number of days user has sent reports	Family.Photo	availability of family photos
mean_reports_per_farm	average number of reports per farm	Land.Document.Photos	availability of land document
activity_types	number of activity types reported on	ID.Card.Photos	availability of official identification
activity.Farm.activity	number of farm activity reports	Buddy.With.House.Photo	availability of geotagged photo of user's house
activity.Harvest	number of harvest activity reports	has_inviter	determines if the user was invited by another user
activity.Seeding	number of seeding activity reports	has_comm_history	whether or not user communicates with the company
activity.Tractor	number of tractor activity reports	earliest_report_time	earliest time of day reports are sent
activity.Trouble	number of trouble activity reports	latest_report_time	latest time of day reports are sent
num_GPS_reports	number of gps reports	mean_report_time	average time of day reports are sent
num_Photo_reports	number of photo reports	fraction_missing_fields	fraction of important data not sent by user

Figure 26 data collected from mobile application

ii. SURVEY DATA COLLECTION

VARIABLE NAME	MEANING	VARIABLE NAME	MEANING
Buddy_ID	Buddy ID	Pest_Attack_last_3_Year_1	Types of pest attack in the last year
Farm_ID	Farm ID	Pest_Attack_last_3_Year_2	Types of pest attack in the last year
Farm_Ha	Farm size in hectares	Pest_Attack_last_3_Year_3	Types of pest attack in the last year
Crop	Crop type	Pest_Attack_last_3_Year_4	Types of pest attack in the last year
num_crop_cycles_per_year	Number of crops cycles per year	Pest_Attack_last_3_Year_5	Types of pest attack in the last year
num_crops	Number of crop types	Organic_Pest_Control_1	Organic pest control method
num_Laborer_work_on_farm	Number of labourers working on the farm	Organic_Pest_Control_2	Organic pest control method
Chemical_Pest_Control_1	Chemical pest control method	Organic_Pest_Control_3	Organic pest control method
Chemical_Pest_Control_2	Chemical pest control method	Organic_Pest_Control_4	Organic pest control method
Chemical_Pest_Control_3	Chemical pest control method	Organic_Pest_Control_5	Organic pest control method
Chemical_Pest_Control_4	Chemical pest control method	Irrigation	Use of irrigation during farming
Chemical_Pest_Control_5	Chemical pest control method	Water_Farm_type	Type of water source used for watering
Organic_Fertilizer_Fert1	Organic fertilizer type	Plow_Farm_1	Ploughing method
Organic_Fertilizer_Fert2	Organic fertilizer type	Plow_Farm_2	Ploughing method
Organic_Fertilizer_Fert3	Organic fertilizer type	District	District of residence
Organic_Fertilizer_Fert4	Organic fertilizer type	Province	Province of residence
Organic_Fertilizer_Fert5	Organic fertilizer type	Village	Village of residence
Non_organic_Fertilizer_1	Non-Organic fertilizer	Last_Year_Yield_Kg	Yield obtained in the last year
Non_organic_Fertilizer_2	Non-Organic fertilizer	Sale_Price_Kg	Sale price of crops (per year)
Non_organic_Fertilizer_3	Non-Organic fertilizer	Variety	Crop variety
Non_organic_Fertilizer_4	Non-Organic fertilizer	Harvest_by	Harvest method
Non_organic_Fertilizer_5	Non-Organic fertilizer	Revenue	Revenue (Last_Year_Yield_Kg *Sale_Price_Kg)

Figure 27 data collected from surveys

IX.4 INITIAL INTERVIEW WITH FINANCIAL INSTITUTION

Date: 14th February 2018

Interview with: Mr Kawai (Aeon Specialized Bank Cambodia PLC)

Location: Aeon Specialized Bank Cambodia PLC Offices, Siem Reap, Cambodia

Objective: Collect requirements from financial institution

Details: An interview were conducted on 14th February 2018 with Mr Kawai of Aeon Specialized Bank Cambodia PLC, an institution focusing on microfinance for individuals in Cambodia. Mr Kawai verified that the bank does face challenges when it comes to lending to smallholder farmers. One of the major roadblocks experienced by the bank was **their inability to verify the farmer's income**. This is mainly because many smallholder farmers are not formally employed and do not have a consistent income level. He stated that the bank is currently only able to lend to people who are in formal employment. This is because the bank relies on the word of the employer who can verify that a given individual and a certain income which will be sufficient to repay a loan from the bank. Yet, smallholder farmers (who are in informal employment) do not have this benefit, making it **difficult to the bank to decided how much, or if at all, the farmer should be given a loan**. Therefore, he stated, **the bank would be interested in such a credit decision system which can predict revenue of the farmer as well as judge a farmer's behaviour**. However, he felt that the development and maintenance of such a system would be too involving for the bank to do directly. As such, there would need to be a third party willing to collect the data and carry out the analysis on behalf of the bank. He also mentioned that he would want the system to assess other risks such as the risks of drought and flooding which may greatly impact the crop yield and repayment ability of the farmer.

His expressed concerns regarding the output format **or user interface** through which the bank would be able to see the farmer evaluations. He suggested that the output format should be as **simple** as possible, allowing for an easy interpretation of evaluations.

IX.5 INITIAL WORKSHOP DETAILS

Date: 30th June 2017

Workshop with: 11 employees of Agribuddy Cambodia

Location: Agribuddy Headquarters, Siem Reap, Cambodia

Objective: Identify risks that may affect a repayment

Details: Risks for buddies (potential borrowers) were identified during a workshop session with Mentors, call agents and other personnel of Agribuddy. The potential risks were identified and then ranked by three groups: groups A, B, and C. Unranked, the main risks identified during the workshop were:

1) low productivity or crop yield
2) Lack of trust between the data collector (buddy) and farmer
    a. Some buddies and farmers have had bad experiences in the past with fraudulent agricultural initiatives. This has left them wary of trusting any new agricultural programs
    b. Data fraud by data collectors (buddies)
3) Government issues
    a. These included lack of infrastructure
    b. No government initiatives to support farmers
4) Transportation/travel
    a. Lack of transportation for buddies to travel to farm areas
5) Family situation
    a. situations such as alcoholism, illness or divorce in the family can drastically affect ability of the farmer to repay the loan
    b. having many children was connected to low repayment rates
6) Education
    a. Low financial literacy resulting in poor financial decisions on the parts of the farmers
    b. Lack of knowledge about the best farming methods to maximise productivity
    c. Lack of knowledge on how to use agricultural inputs correctly
7) Market
    a. Fluctuating market prices
    b. Lack of buyers for produce
    c. Low prices for produce
8) Technology:
    a. Difficulty accessing and using mobile technology which is a necessity for using the Agribuddy mobile application. For instance, poor mobile network coverage

b. Poor access to agricultural technologies which would simply farming processes such as tractors and harvesters

9) Water system
    a. Heavy dependence on rainfall for crop production
    b. This resulted in severe losses during droughts
10) Bad weather
    a. This includes droughts and floods which both have a detrimental effect of productivity
    b. Damage to roads due to heavy rainfall may affect transportation to an area
11) Other
    a. Economic crises
    b. Political turmoil

IX.6 INITIAL INTERVIEWS WITH FARMERS

Date: 1st July 2017

Interviewees: 2 farmers who also work with Agribuddy as data collectors

Location: farm areas in Siem Reap, Cambodia

Objective: identify the challenges faced by farmers and how the system could be used to evaluate them

Details: Field work was carried out by visiting several areas in and outside Siem Reap, Cambodia in July 2017. Two interviews were conducted.

1) FARMER 1

This farmer was 36 years old and had lived in the area with his family for his entire life. He had worked with other community initiatives in the past and so was considered a community leader. Some of the problems he faces include mobile network coverage in his area is poor so uploading data onto the Agribuddy platform can sometimes take a long time. He obtained income from his own farm, a pig farm, as well as buying and selling land in the village and receiving a commission from this. He would like to have good quality, affordable inputs.

Interviewers assessment: He has lived in the village for a very long time and so has a great amount of influence in the community. His reputation in the community is important to him which reduces the risk that he would default on the loan without good cause.

2) FARMER 2

This farmer was at the time growing cassava and rice on her land. Most of the necessary farm activity was done by herself and her husband, with additional labour being hired for harvest periods. She also hired harvest machines occasionally. Her main income source was from her farm.

Interviewers assessment: The second farmer interviewed had a good personal profile. She visited her farmers regularly and was hardworking. However, from a business perspective she was a poor risk because she was growing a crop (cassava) which has a very low market price and was unwilling to change this even though she was aware of the risks involved.

IX.7 VALIDATION WORKSHOP DETAILS

Date: 6th July 2018

Workshop with: 18 employees of Agribuddy Cambodia

Location: Agribuddy Headquarters, Siem Reap, Cambodia

Objective: Collect feedback concerning the validity of features identified through data analysis as being important for assessing the alternative scoring factors. In particular, the reliability factor which was used to evaluate data collectors (buddies).

QUESTION 1: participants were split into groups asked to identify factors affecting the reliability of a buddy (data collector).

a. Group A

1. - Good communication and leadership
2. - Potential and influence buddy in the village
3. - Educated person
4. - Responsible person
5. - Characteristic, Work commitment, Willing to learn/do new thing
6. - Matured person
7. - Sharing knowledge
8. - No alcohol
9. - No gambling
10. - Ability to work with farmers
11. - Have transportation to visit the farmers
12. - Honest person (can be trusted with work and problem solving)

b. Group B

1. - Accurate present address
2. - Honest person
3. - Educated person at least can read and write
4. - Influential person in the village
5. - Able to use technology
6. - Experience on agriculture such as CDOR/USAID and other organization
7. - Trusted person in the village
8. - Experience on volunteer work
9. - Good in problem solving, A person has good relationship with villager, Have experience working in community
10. - A willing person who want to help his/her community
11. - Willing and have goal to help farmers

QUESTION 2 participants were also asked to identify which activities or factors would increase revenue from farming activities.

1. Networking to get information on market price
2. Make farming in technical way and responsibility.
3. Change farmer's mind set
4. Model farming (i.e. using bets farming methods)
5. Planting with technical standard that lowers the cost and increase the yield
6. Using more technical farming methods and monitoring prices
7. Learn methods from farmers with higher revenue and use fertilizer properly
8. Learn new farming technique and identify good market price for selling their produce
9. Change farmer's mindsets
10. Identify good market for produce
11. Use more technical farming methods
12. Use more technical farming methods and have available market
13. Increase labour force
14. Borrow with low interest rate and find good markets
15. Use more technical farming methods

IX.8 VALIDATION INTERVIEW QUESTIONS FOR FARMERS

Interviews were conducted with farmers and buddies to validate the project output. Questions include:

1. What do you consider to be the most critical factor when you ae trying to increase your crop yield?
2. Have you ever approached a bank to borrow money? If not, why?
3. How/where you do you sell your produce
4. What are the biggest issues you are facing when it comes to farming