Credit Scoring in Microfinance Using Non-Traditional Data

by

Saulo Neftali Carpio Ruiz

For the degree of Master of Science, Data Analytics

Supervised by

João Gama
Pedro Gomes

Faculdade de Economia

Universidade do Porto

2017

Biographic Note

Saulo Neftali Carpio Ruiz was born on November 09th, 1991, in Higüey, Dominican Republic.

After finishing high school with outstanding academic performance, he looked for another challenge, enrolling in the mathematics with concentration in statistics and actuarial science program at the Instituto Tecnologico de Santo Domingo in 2009. Carpio is part of the first even promotion of actuarial science professionals at the Instituto Tecnologico de Santo Domingo and also in Dominican Republic. Finish the math program in 2013.

In the last year of the collage, we started labor activities in Banco Leon, at the time 4th bank in Dominican Republic in terms of wallet share. Carpio was Business Intelligence Analyst, which allowed him use the statistical knowledge in a financial environment.

Later in 2014, Carpio moved to the Banco de Reservas de la Republica Dominicana, which is the first bank founded in Dominican Republic and manage the found of the state. Here Carpio took care of similar tasks related to Business Intelligence.

In 2015, Carpio was awarded with a KITE-Erasmus scholarship to undertake the Data Analytic Msc at the Faculty of Economics of the University of Porto.

Acknowledgments

I would like to thanks the KITE-Erasmus Scholarship program and Faculty of Economics of University of Porto, which allowed me to pursue a higher level of education through the scholarship grant. To my supervisors João Gama and Pedro Gomes for their guidance through this project. To my family, starting with my mother Maria, that is always taking care of me regardless the distance, with messages of support every morning. My sister Rosemary and brother Emmanuel, they always give me support to keep going forward. My aunts and uncles that always try to keep me close to the family even on the other side of the Atlantic Ocean. And to Seedstars which gave me the opportunity to freely try the different data mining features.

Resumo

A grande maioria da população mundial vive em mercardos emergentes. Não obstante deste facto e do grande número de habitantes, continua a faltar uma infraestrutura financeira apropriada. O acesso a empréstimos é uma das maiores dificuldades sentidas pelos clientes nestes mercados. Esta limitação deriva do facto de grande parte destes clientes não terem um histórico bancário que seja facilmente verificável o que implica que os bancos tradicionais não consigam fornecer empréstimos. Esta tese tem como objectivo principal propor a modelação da pontuação de crédito baseada em dados não tradicionais obtidos através de smartphones para o processo de classificação do empréstimo. Nós usamos os modelos Logistic Regression (LR) e Support Vector Machine (SVM) que são os principais modelos usados na banca tradicional. Comparamos também a transformação do conjunto de dados de treino criando indicadores boleanos relativamente ao "recoding" usando Weight of Evidence (WoE). Os nossos models melhoraram o desempenho do processo de seleção de empréstimos manual, melhorando o rácio de aprovação e diminuindo o rácio de empréstimos em atraso. Comprando com a nossa linha de base, o nosso modelo SVM melhorou o rácio de aprovação em 251.5% sendo capaz também de reduzir ao mesmo tempo em -196.80% o rácio de empréstimos em atraso. Esta tese mostra que a pontuação de crédito pode ser útil em mercados emergentes e que os dados não tradicionais podem ser usados para construir algoritmos robustos que são capazes de identificar bons clientes na banca tradicional.

Abstract

Emerging markets contain the vast majority of the world's population. Despite the huge number of inhabitants, these markets still lack a proper finance infrastructure. One of the main difficulties felt by customers is the access to loans. This limitation arises from the fact that most customers usually lack a verifiable credit history. As such, traditional banks are unable to provide loans. This thesis proposes credit scoring modeling based on non-traditional-data, acquired from smartphones, for loan classification processes. We use Logistic Regression (LR) and Support Vector Machine (SVM) models which are the top performers in traditional banking. Then we compared the transformation of the training datasets creating boolean indicators against recoding using Weight of Evidence (WoE). Our models surpassed the performance of the manual loan application selection process, improving the approval rate and decreasing the overdue rate. Compared to the baseline, the loans approved by meeting the criteria of the SVM model, presented -196.80% overdue rate. At the same time, the approval criteria of the SVM model generated 251.53% more loans. This thesis shows that credit scoring can be useful in emerging markets. The non-traditional data can be used to build strong algorithms that are able to identify good borrowers as in traditional banking.

Index

List of Tables

List of Figures

Acronyms

MFI Microfinance Institutions

SMS Short Message Service

MNO Mobile Network Operator

CRISP-DM Cross Industry Standard Process for Data Mining

NPL Non-Performing Loan

ANN Artificial Neural Networks

DT Decision Tree

RF Random Forest

BN Bayesian Networks

LR Logistic Regression

LR-A LR trained with dataset A

LR-B LR trained with dataset B

SVM Support Vector Machine

SVM-A SVM trained with dataset A

SVM-B SVM trained with dataset B

WoE Weight of Evidence

IV Information Value

GDP Gross Domestic Product

SMS Short Message Service

MMS Multimedia Messaging Service

P2P Peer-to-Peer

AUROC Area under Receiver Operating Characteristic

GDP Gross Domestic Product

USA United States of America

NBFC Non-Banking Finance Company

FINCA Foundation for International Community Assistance

Chapter 1

Introduction

Finance in emerging markets is an exciting and growing market that is completely distinct from what one can find in developed countries. Even though studies show that 85% of the world population is in emerging markets Barnes (2016), they still lack a proper finance infrastructure. According to the World Bank, it is estimated that there are 2.5 billion unbanked adults who lack of access to financial services Bank (2013). From these financial services, loans are the most relevant and more requested services. Yet in these markets customers cannot rely on banks to have access to a loan as they usually lack a verifiable credit history. Microfinance Institutions (MFI) target these customers, by providing local access to basic financial services. However, due to the risks involved with this kind of service, MFI's loan process tends to be slow and cumbersome. Customer requests frequently include identification card, employment letter, utility bills, loan application letter, or guarantors. Although it is a common practice to request this type of information in developed economies, most customers in emerging markets do not have them or it is hard to get them. Furthermore, MFI apply high interest rates which can directly affect the utility of this service. These factors reduce significantly the number of customers that can apply for a loan.

Digital technologies bring a new dynamic to the finance market in emerging markets. Smartphone adoption in these markets is approaching the numbers of developed economies Poushter (2016) and new fintech solutions for unbanked people are surfacing. As the trend of using mobile phone to make financial operations was growing(Katakam et al., 2016), several companies proposed loan products across emerging markets where one can use a simple mobile app to apply for a loan Fifer Mandell et al. (2015). By being more flexible than MFI, they can target different customers. However, challenges in customer classification and eligibility for a loan arise. Credit scoring has been the way to go in traditional credit institutions, and normally rely on reliable user data such as his credit history. These new loan products lack access to traditional data. They only have access to input from customer and data collected from their smartphones, such as call logs, Short Message

Service (SMS) logs, apps installed and social network relationships.

1.1 Motivation

This work is based in a real business problem from a microlender based in a emerging market. This company believe that technology is the best way to deal with daily problems usually present in emerging markets. Embracing this philosophy, the MFI has built a novel approach to micro-credit. This approach is alongside the issues the company need to face in order to succeed, create the perfect scenario for the use of data mining technique and Big Data.

The MFI is totally digital, interacting with the customers by a mobile application. Through the application, customers do all the process from the loan request to the payment of it, generating useful information on each step. All starts when the customer register as a customer of the MFI, on this step the customer needs to fill a form on which the demographics and personal information is captured. Added to the information the customer provides, the MFI also obtain access to the information of the Mobile Network Operator (MNO). The data from the MNO contains information about services use by the customer (calls messages). This data is also rich for the use of data mining, since it allows to create and study social network of the customers of the MFI. After the customer provides the MFI with the information requested and apply for a microcredit, the MFI faces another challenge, which is to estimate the risk of the applicant in order to grant or deny the loan request. This challenge will be the main problem to tackle during this study.

1.2 Problem definition

As presented before, the case of study is based on a branchless microlender operating in an emerging market that is built upon an android application. As seen on figure 1.1, the highlighted node it's the one on which the approval or denial of the loan based on customer data is made. Currently this process is done by humans operators taking up to 10 minutes per application. At the moment, the quantity of loan applications surpasses the capacity of evaluation of the operations team. This situation creates a queue for the application that can translate into weeks from the time of the application to the evaluation of the application. For the time that the loans is disbursed, the customer already borrow the money from other lender. Other issue on this process high default rate. These defaults are the result customers that didn't pay on time (overdue) and didn't pay the new agreement made on the recovery process.

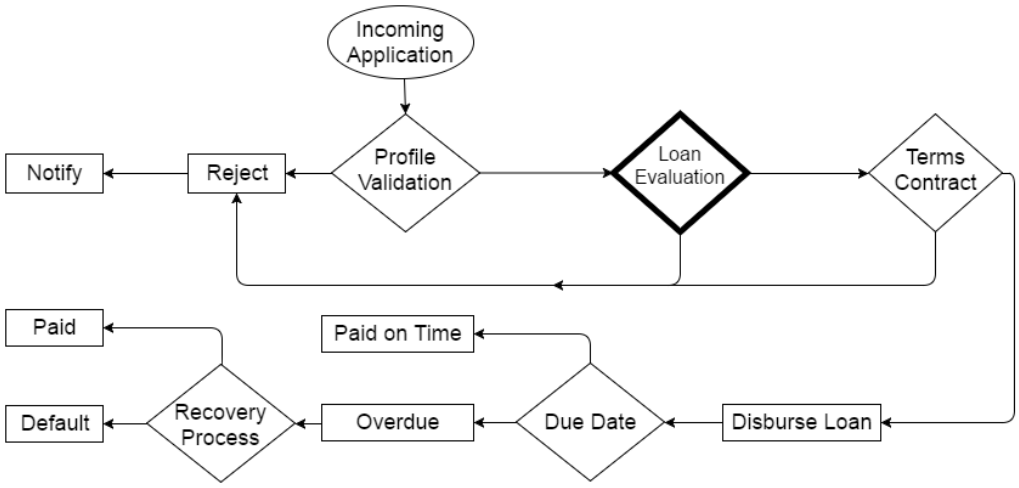

Figure 1.1: Loan Application Process

1.3 Contributions

In this thesis we try to solve these problems by implementing a decision support system *"score"* to evaluate the loans on less than 1 minute and able to process multiple applications at the same time. With the credit scoring implementation, not only the time of evaluation will decrease substantially but also the company can focus on reaching more customers increasing the volume of the business. Also, the decision support system is expected to perform better than human evaluation in terms of default rate, a successful model should reduce the current default rate by at least 50%.

1.4 Organization

The remaining chapters of this thesis will be following the Cross Industry Standard Process for Data Mining (CRISP-DM).

So far, the business understanding has been covered. The next steps of the CRISP-DM include:

- **Data Understanding**: A detailed description of the data set will be done followed by an exploratory analysis.
- **Data Preparation**: Alternate variables will be created from existing data. After this step, the data set will be ready to train different algorithms.
- **Modeling**: Different classification algorithms will be trained with the processed data set with cross-validation as model validation technique.

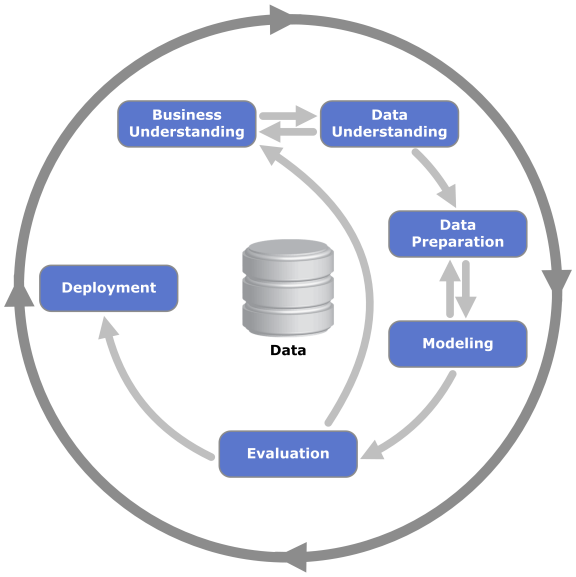

Figure 1.2: Crisp Data Mining Process

- **Evaluation**: The selected evaluation measures will be compared among the models in order to select the best two algorithms that have met the business expectations.

- **Deployment**: The performance of the selected algorithms will be evaluated with an A|B Testing in real applications.

This process can translate into chapters, in Chapter 2 we describe the previous works done relevant to this problem. Going from traditional credit scoring which is based in score cards and collection of data by the loan officer. To the latest uses of mobile data used to predict the default in credit cards. Chapter 3 review the construction of the different datasets used for training and testing the scoring models. This chapter also contains the variable selection methods, including also the reasoning behind non-common features.

In Chapter 3.4 we test the different classification techniques. We select the algorithms with the best performance on the selected metrics. The selected algorithms are use in further experimentation with real data. The experiments are described in Chapter 4. We explain the scenarios and settings of each experiments, with a detailed view of the transformation processes tested.

Chapter 3.4 present an analysis of results of the experiments shown previously. It relate the performance of the scoring models against the previous evaluation process carried by microlender. Finally, the Chapter 5 contains an overview of the whole project, evaluating the reasons of success or failure and setting future works in for this problem.

Chapter 2

State-of-the-Art

"Yes, they will. They do. Unlike the rich, the poor cannot risk not repaying. This is the only chance they have".

-Muhammad Yunus

In the 1960s, economist Muhammad Yunus was a student and professor of economics in the United States of America (USA), when he decided to return to Bangladesh in 1972 (Büthe et al., 2000). Bangladesh was a poor country where the economics theories learned on the United States were useless. With most of the population in poverty and with banks refusing to lend tiny loans at reasonable interest rates due to high risk of default, Yunus decided to pursuit a new business model to help people with financial needs. One day Yunus lent $27 to 42 poor villagers on the Jobra area not expecting repayment or generated interest. Yunus did not set a deadline for collection as he stated that they could pay whenever they can. Later on, they paid him back hence fulfilling the informal agreement. An unexpected discovery on this experiment was made over upcoming the next months and years. Not only the poor pay back their loans even without any collateral, but also that this small amount of money was inaccessible for them on formal banks, and informal lenders had extremely high interest rates (Büthe et al., 2000)

After having this experience, Yunus started building an organization in order to make this type of loans on a larger scale. He founded the Grameen Bank (Village Bank) (Büthe et al., 2000), known as the first organization focused on providing micro-credit to poor people. While others banks focused on lending to individuals, Yunus required borrowers at Grameen Bank to create peer support groups and use the money only for small businesses. This model helped not only to create relations among the entrepreneurs, but also to reduce the default risk by splitting the risk into the group. Yunus found a way to help the poor, and at the same time created a profitable business model on the path. As consequence of the success of this model, MFI's started proliferating in poor countries.

With the goal of improving Grameen model in mind, new MFI started focusing on being more profitable each time. Therefore, new models emerged all over the world, being the most notorious:

- MC2 promoted by Dr. Paul K. Fokam.
- The village banking model of Foundation for International Community Assistance (FINCA) developed by John Hatch.
- The SKS and Non-Banking Finance Company (NBFC) model in India.

For a detailed review and comparison of these models alongside with Grameen model of Yunus see Fotabong (2011).

In the race for a more profitable model, MFI started to rise interest rates or requesting physical collateral. Was not until almost the end of the century that MFI started using a tool commonly used in traditional banking, credit scoring. This tool allowed banks to keep a reasonable interest rate by selecting applicants with low risk of default. A new race began with the use of scoring models. That opens a new question, what is credit scoring?

2.1 What is credit scoring?

To define credit scoring Anderson (2007) proposes to break it into two components: *credit* and *scoring*. First, following the simple definition of *credit* which means, *buy now, pay later*. This word comes from the old Latin word *credo*, which means *trust in*, or *rely on*. Second, defining scoring as the use of a numerical tool to rank order cases according to some real or perceived quality in order to discriminate between them, and ensure objective and consistent decisions. Therefore credit scoring can be defined as the use of statistical models to transform relevant data into numerical measures that guide credit decisions.

Thomas et al. (2002) shows a similar approach defining credit scoring as a set of decision models and their underlying techniques that aid lenders in the granting of consumer credit. Finally, Schreiner (2000a), whose focus is microfinance, describes the credit scoring model as a formula that puts weights on different characteristics of a borrower, a lender, and a loan. This formula produces an estimate of a probability or risk that an outcome will occur. Based on these definitions is evident that a credit scoring is a flow that receives several characteristics and transforms the given variables in order to produce an output that can be interpreted as a score or probability. The flow can be described as shown on figure 2.1. For a more detailed flow see Chen et al. (2016).

Anderson (2007) present a detailed process when building a credit scoring model. In this study, we will provide a brief explanation of the core steps of building process,

Figure 2.1: Pipeline of credit risk assessment (Chen et al., 2016).

following the procedures needed to achieve our particular goal. This document will focus in the data input and data sources, the methods used to deal with data gathered, including the filtering of the cases and the selection of features. We also present the model training and selection alongside the results of the selected models.

2.1.1 Data Input

Usually the data comes from financial variables. The main source used for creating a credit score originates from credit bureaus. This guarantee a complete view of the customer on the market on which the financial service is required. Therefore helping to achieve a more accurate estimation of the reality. Internal data of the company normally us the second main source. It includes past deals with the customer and financial behavior based on his transactions. The internal data is very useful for customers with previous relations with the institution but fails when having a new customer. The third source is the application form, this source deals with new customers, usually on the applications the customer has to provide financial statements, prove the source of income, and other significant variables (Anderson, 2007).

2.1.2 Preprocessing

After the variables have been summarized by customer, the data is preprocessed. This include filtering customers with special conditions as bank employees, shareholders or any kind of special financial agreement. This part of the process should deal with missing values and outliers. Transformation of categorical variables should be done. Dummy variables transformation or Weight of Evidence (WoE) are ways to deal with the transformation (Van Gool et al., 2012). All variables become numerical after this transformation. Finalizing the preprocessing flow, normalization ensures that high scale variables do not make an unbalanced database.

2.1.3 Dimension reduction

If available, every source of information is used, but this usually creates noise due to correlated variables and variables with no impact on the prediction. Not only the noise is a problem, but the computer processing time can increase significantly. Hence, feature selection methods are then used for dimension reduction through linear or nonlinear transformation when needed.

2.1.4 Training

Several classifiers have been implemented successfully since Fair & Issac used statistical methods in order to classify on the financial sector in the 1950s. Ideally these methods use data of the performance of previous loans in order to differentiate future applications or customers. This part of the process assign weights to the different variables of the input so a new customer with similar input values of a previous customer will adopt the performance of the previous customer.

2.1.5 Validating

The prediction results are evaluated in terms of performance metrics used for comparison among different classifiers. However, sometimes a matrix cost is required in credit scoring when validating a single classifier, since cost of default overweight the cost of a miss opportunity (Sousa et al., 2016). This fact is a consequence of finance institutions aim to maximize revenue and not in all cases high accuracy translate into high revenue.

2.2 Can credit scoring work in microfinance?

In the early 2000s Mark Schreiner, who can be considered as one of the main contributors of credit scoring for microfinance, started using structured databases for building models that can be replicated. This work on structured databases was implemented mostly for MFIs located in Latin America. Credit scoring models of Schreiner were based on scorecards that include details from customer, loan and loan officer. The scorecard system showed positive results, however it was difficult to implement since loan officers had to do the process manually in order to fill the scorecard. In some cases the loan officer even had to visit the customer in order to validate some information such as household goods. These models can be seen on Schreiner (2001), Schreiner (1999) and Schreiner (2000b). Schreiner proved to MFIs that credit scoring can work for their institutions. His approach worked both with non-traditional variables and also with financial system information from the bureaus. Despite the promising results of his models, Schreiner concluded that due to the high difficulty of implementation and low added value, these models were

not able to replace the loan officers with an automated process. Nevertheless, this proved that the credit scoring (as statistical method) worked for MFIs business case. The main issue was to have it integrated with a scaling business model.

2.3 The arrival of Big Data

Two decades ago, the collection of data was an inefficient process for rural areas and/or poor countries. From private surveys to national census, there is always a group that cannot be reached due to difficult access conditions, because they live far away or are in small frequencies, hence considered insignificant. Insignificant because of the distance but also because they usually lack of goods and services. Usually, this group is mainly consisted of poor people that works on agriculture, livestock or other some kind of personal manufacture (Büthe et al., 2000). This group of people is not only excluded for counting or for their idea on the next product on the market, they are also excluded from different services including financial services.

Banks will never lend to someone who they cannot verify on their systems. This verification requires data that most of this people do not have. The problem is that banks always look in the same place: declaration of income, saving accounts, bureaus, etc. (Anderson, 2007). But with the arrival of the digital era, social media, and smartphones, customers now have both a financial and digital footprint. Every day billions of people use their mobile phones generating data on every interaction (Katakam et al., 2016), such as: SMS and call data patterns, social media activity, navigating on the web, or installing apps. Furthermore, even when they are not directly using their phones they generate geo-location points useful for movement patterns or frequent places.

This new type of non traditional data captured the attention of MFI. The struggle for obtaining and verifying data has always been a problem for MFI. It was usually an expensive and time consuming process. But with the arrival of digital footprint this process has now become an easy, fast and cheap procedure that provides huge amounts of data useful for credit scoring. This benefits bothMFI and users. Specially users, by granting them a digital identity that can be reached by scoring models. The implementation of an intelligent automatic process helps to solves both the time for recollecting data and the price of the data gathering, since surveys with human operators were no longer needed.

2.4 Non-traditional data in Micro finance

Over the past few years some companies have built successful businesses using classification algorithms to evaluate the risk of customers. These algorithms are usually trained with non-traditional data. The data used to build the models comes from

different sources, such as: social, mobile, data of payment of bills, and location. Some assumptions that can be made based on non-traditional data are:

- **Social**: A highly connected individual is assumed to be established in a location that creates a social commitment. Can also be a business person which use calls or SMS for the business relations.
- **Mobile**: If the individual tends to recharge the mobile on a regular basis, it can indicate a formal mobile service contract. It also shows there could be a source of income that allow this recharge to be made on the same period of time. Other variables such as manufacturer can indicate the financial level of the person.
- **Payments**: Recurrent payment on time of services is also an indicator of steady source of income as well as a payment behaviour.
- **Location**: Sparse location or dense location can both be used to have an understanding of the behaviour of an individual. Dense location on weekdays can indicate that customer has a job or some recurrent activity on a given location. On the other hand, sparse data can indicate two things. If it is on week days and has a route pattern can be a job indicating transportation (bus or truck driver, security service, etc.). Sparse data on holiday periods or weekends can indicate than the person is able to have vacations outside his living area.

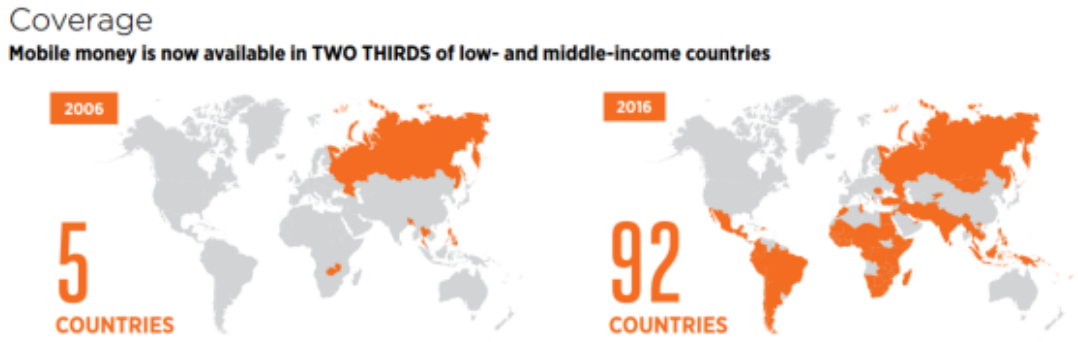

Figure 2.2: Comparison of mobile money coverage 2006 against 2016. *Data and figure from Katakam et al. (2016).*

Some emerging market countries use mobile money. This form of *money* refers to payment services operated in financial regulation and performed from or via a mobile device. This system, provided by a MNO, works as a bank account for the user. It allows several types of financial transactions, for example bills payment

and money transfer. These transactions generate the same type of data that a savings account in a traditional bank would generate. Since the use and coverage of mobile money is a growing trend, it has not been fully explored yet. As in 2005 only 5 countries supported this technology which has spread already to two thirds of emerging markets as seen in Fig. 2.2. This figure shows only the availability of the service in a given country. However, not only more MNO are providing these services in different countries, but also the individuals in those countries are using the service more often for their different financial needs. In Figure 2.3 we can observe an evolution from one million active accounts in 2008 to 174 millions at the end of 2016. People in emerging markets are embracing this service to meet their financial needs.

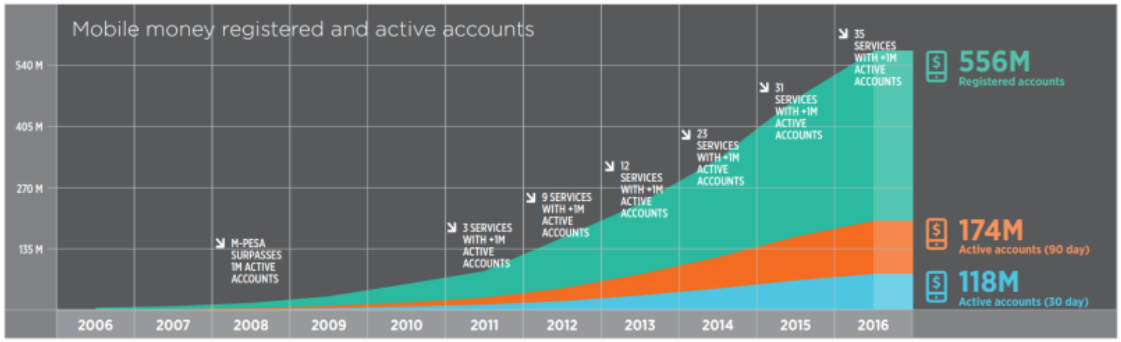

Figure 2.3: Evolution of mobile money accounts. *Data and figure from Katakam et al. (2016).*

During the past year, MNO were processing around 30,000 transactions per minute (Katakam et al., 2016). The amount of transactions per year grew from 1.2 billion transactions to 268 billion over the last decade. From these transactions, 68.7% are Peer-to-Peer (P2P) transactions and 12.5% are bill payment transactions. The huge proportion of P2P transactions is due to the way some small businesses operate. Instead of using cash or credit card, a person can buy fish to a local fish shop and pay by transferring from the mobile money account of the customer to the mobile money account associated with the fish shop. The bills payment transaction works the same as in developed and more mature markets. For example, a customer subscribes to a service, such as electric or telecommunication service. Bill payments can provide better insights about the acquisition level of the customers. Even if the P2P transactions can be used to acquire goods, it is not exclusive to such use. The data from these kind of transactions can be shared between the MNO and a financial services company.

Companies like M-Shwari (Cook and McKay, 2015) and InVenture (today known as Tala) (Fifer Mandell et al., 2015) have built services on top of mobile money to

offer mobile savings and mobile credit. Mobile money data is a good variable to discriminate among defaulters and non-defaulters. Nonetheless, the use of MNO data limits the ability of a business to grow since some of the emerging markets have a relatively low penetration of mobile money service (Katakam et al., 2015). On the other hand, companies like Lenddo (Stewart, 2014) and Wonga (Harkness, 2016) focus more on the social media footprint. This approach is useful to evaluate customers with medium to high presence on the social media since is based on *likes*, *friends* and *shares* (De Cnudde et al., 2015). On the flip side, it struggles to discriminate when low or no information on social media is available, which can be the case in rural areas.

Other companies in the same sector claim to use machine learning on their evaluation pipeline. Based on their operational description, Kreditech (Kreditech, 2017), Branch.co (Branch.co, 2017) and Cignifi (Cignifi, 2017) state that the use of machine learning is core on their loan evaluation process.

A more similar case to ours is the MobiScore approach (San Pedro et al., 2015). This approach is based on customers demographics and device logs as we intent to do. The main difference is that MobiScore studies the default for credit cards while we focus on short-term micro loans. The MobiScore proved that a scoring model trained with mobile network usage data can be useful to estimate the financial risk of a person. Macroeconomic variables can also be a factor to consider when lacking formal sources of financial data. The case presented in Blanco et al. (2013) proposes the integration of macroeconomic variables to build a reliable scoring model. The models include variables like rate of annual change of Gross Domestic Product (GDP) during loan term and rate of annual change in cost of electricity during loan term. These models performed better in terms of misclassification than the classic approach.

Regardless the main source of data (social/online, mobile or MNO), MFIs have been benefited from these new technologies, they now can create a branchless structure. With this new structure MFIs solved two issues. First, MFIs have now better criteria to evaluate their customers. And second, the huge cost of a brick and mortar schema was eliminated. For more about branchless banking see Pickens et al. (2009).

After reviewing the cases of success shown above, we started experimentation with our own data. First, we did exploration of variables and exploration of the target to predict. As much of the details on this field remain unpublished due to business reasons, we could not rely on the plain variables as some of the cases relate since they do not show the total process of transformation.

We also needed a clear definition of the target variable, for business purposes there is no difference between a customer that pays on time and a customer that pays five hours late. In the next chapter we define and test the moment on which a customer will be labeled as "Overdue" and how the other features might affect this label.

Chapter 3

Datasets and features selection

Few of the works published in this area present a detailed process of variable selection and transformation. Some of them present the features on a general view as we intent to. The MobiScore (San Pedro et al., 2015), refers to the use of personal information alongside the use of features extracted from the use of the services of the MNO. Blanco et al. (2013), go further and propose the use of macroeconomic features mixed with customer data in order to estimate risk in a microeconomic environment. For this study we focus only in the features collected by the MFI.

We also analyze the effects of the definition of the target variable. In Van Gool et al. (2012), a straight definition is set in base on installments. However, our case is based on loans that are paid at the end of the period agreed between MFI and the customer. In order to define the target variable, we tested the options that were aligned with the business and choose the optimal one.

3.1 Structure of datasets

This section describes the datasets used to understand the behavior of our customers. Several datasets are used for experimentation. Even though all datasets contain the same features, the transformation process, selection process and time horizon considered are different.

In our first iteration with the data, we created the process to extract, transform and load the data. In this part of the process we removed variables that could not add value to the modeling process due their structure. Variables like name or mobile phone number of the customer were useless since they do not represent any characteristic of the customer neither are dependant of the customer itself. Therefore, this type of variables were removed. To be able to apply for a loan, the customer most completely fill their profile. Following this condition, we know that a missing value in a given variable means that the customer does not meet the criteria of the variable.

The missing values are there replaced by zeros. This case can be easily explained with the variable *Date of Employment*. This variable is only available to customers who have chosen *"Employed"* in the *Employment status* variable in the previous step. In Fig. 3.1 we present the process used to create the base for each dataset.

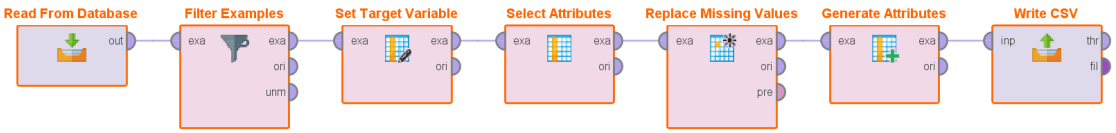

Figure 3.1: Process of extraction and cleaning of datasets prior to transformations.

Bellow we present and detail the task performed in each node:

- **Read from Database**: This node contains the SQL connection and query to extract the raw data from the database. Information is retrieved by loan, this means that a customer appears as many times as loan applications made.
- **Filter Examples**: Filters the dataset (by rows) to retain the loans that have been paid on time and the overdue loans. Can also filter the applications received in a given period of time, e.g. Applications of the last three months.
- **Set Target Variable**: Sets the target variable as a target for prediction.
- **Select Attributes**: Filters the dataset (by columns) in order to remove meaningless variables, e.g. Customer_ID.
- **Replace Missing Values**: Replaces missing values with zeros 0. This can only happen in some categorical variables. A missing value means that the customer did not comply with the given condition.
- **Generate Attributes**: Transforms date type variables into numeric variables, e.g. Date_of_Birth $\rightarrow$ Age
- **Write CSV**: Stores the processed dataset.

With this process flow we created the first dataset, denominated ETL dataset. This dataset is used for the model selection that will be presented in Chapter **??**.

The second dataset, hereinafter called dataset A, covers all the first loans granted to the customers. Only the completed loans where considered, meaning that only

paid loans and loans which have passed their respective due date were used. The third dataset, hereinafter called dataset B, consists of all loans granted and not only the first loan of each customer. As in dataset A, only completed loans were considered when generating this dataset.

Following the structure of dataset B we created two more datasets, dataset C and dataset D. As seen in Fig. 3.2, dateset C and datesets D consider different time windows. Dataset C considers the application of the previous three months while dataset D expand this time horizon up to five months. This will help us to test the hypothesis that models will require more fresh data in order to predict more accurately in this fast growing market. Notice that only dataset A considers only the first application of each customer. This approach was initially used because the scoring models will only be used in the first application of each customer. However, the conditions of the loans are the same regardless the number of previous granted loans. This means that a new customer have the same interest rate, same max amount to request and same max leght to request as a customer who had paid 5 loans already. With this into consideration, we decided to consider all applications in order to increase the sample size for the next datasets.

All datasets combine demographics of customers, personal information, loan details, mobile network usage and mobile features. We also included the performance of the loan (overdue or not) which is the target of prediction. In our scenario, a Non-Performing Loan (NPL) is a loan that has passed more than 3 days after due date and has not been totally repaid. Therefore this loan is considered as an overdue in the dataset and marked with 1 in the target variable.

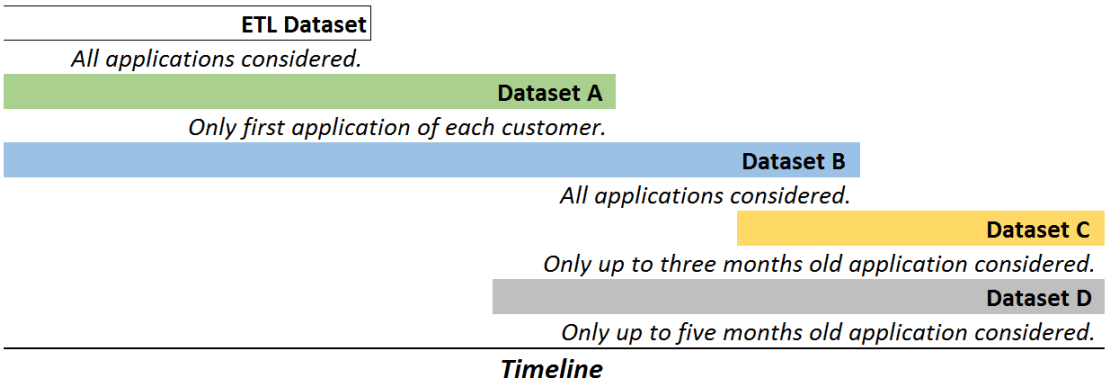

Figure 3.2: Timeline of creation of datasets and the restrictions considered for each dataset.

The dataset A consists of 38.5% overdue loans and 61.5% paid on time loans. The reminder datasets content 43.5% overdue loans and 56.5% paid on time loans. The summarized variables on each dataset use a 30 days window, using the previous 30 days from the moment of the loan application.

3.2 Description of features

The variables can be grouped by source. First, we focus on personal information of the customer. This personal information, including demographics, is collected when the customer opens the mobile app for the first time. In this first interaction with our service, the user also needs to fill a profile. Some of the variables in the profile can be changed later on, e.g. employment status. This information can provide better insights about the acquisition level of the customers. Some variables refer to the goods the customers have (e.g., house, car) and goes to their employment status. It also collect information about the dependants of the customer thought marital status and number of children. This type of data has been considered since the building of scorecards (Schreiner, 2001) to the use of advanced classification techniques as shown in Van Gool et al. (2012), San Pedro et al. (2015), Blanco et al. (2013), among others.

We also collected mobile phone network usage variables. These variables capture whenever the customer uses one of the services provided by the MNO, e.g. incoming or outgoing calls, SMS, Multimedia Messaging Service (MMS), etc. Biçer et al. (2010) show that the having a service of a given MNO can have impact when building a scoring model. The hypothesis is that high tier customers will relate with high tier MNO. From mobile phone we also collect system information, e.g. mobile applications installed. With the categories of the applications installed in the device, we can create a more complete profile of the customers. The applications installed work as the *"likes"* and *"shares"* presented in De Cnudde et al. (2015). Providing an idea of the real interests of the customers.

Finally, we add loan characteristics and conditions: length, amount and purpose of the loan.

Table 3.1 shows in detail the list of the features used. This list presents only the core variables, we combine some of this features to create more meaningful variables, e.g. ratio of amount requested over monthly income became the *"debt ratio"*.

Some previous works also have a similar selection of variables. Such is the case of the MobiScore (San Pedro et al., 2015), which considers SMS and calls as relevant features. Also, Bjorkegren and Grissen (2015) conclude that the use of mobile data usage patterns can be a valuable input to build a scoring model even when lacking formal financial history. Moreover, these variables were analyzed with independence tests in order to identify any relation between variables. Chi-Square independence test was conducted for categorical variables which reduced significantly the amount of variables to use for modeling. Variables like level of education, did not have enough impact in the target variable, therefore was not taken into account.

Personal Information
Age
Gender
Marital status
Education level
Number of children
Employment status
Ownership of house
Ownership of car
Monthly Income
MNO and Device Features
Airtime
Airtime top ups
Number of calls (Incoming, outgoing)
Number of SMS (Incoming, outgoing)
Device Manufacturer
Last mobile update
Mobile Applications installed
Loan characteristics
Loan amount
Loan length
Loan reason

Table 3.1: Table of features considered on datasets.

3.3 Exploration of variables

In this section we present the description of some variables alongside some of the analysis done previous to modelling. Note that due to business matters, we do not present absolute values, we present relative values of each variable.

Our first approach included a partial description of the population based on demographics. The population met some of the expectation from the business experts based in the emerging markets. The customers are mostly males between 26 to 35 years (Table A.7) that are settled in the evaluated region since only 10.3% live at their current address for less than 2 years. The customers also present low level of dependants, since 53.46% are single and 56.98% do not have children as presented in Table A.4 and Table A.6.

The population reflected some of the realities of the emerging markets. One of these realities is that most of the customers own a car but do not own a house (Table A.2, Table A.3). Cars in emerging markets are priority over residence, mostly because people need to cover long distances to get the goods and services they need.

Furthermore, 38.2% of the customers are self employed (Table A.8) and they might use their vehicle for their respective business. However, in terms of employment customers do not seem very steady, we observe a high proportion of customers have less than one year at their current employment (Table A.10).

We fund high levels of education in our population with 87.5% of the customers at university level. However, this does not seem to match in terms of salaries, since we focus in microcredtis and this microcredit represent between 10% - 30% of the monthly salary for 69.19% of the customers as seen in Table A.11 .

For business reasons we present the tests only for categorical variables. After doing a general overview of the population we went into further exploration, we analyze relations between the categorical variables and the target variable. We transformed some numerical variables into categorical (age_brackets and debt_ratio_brackets) in order to analyze any kind of existing relation.

For the evaluation of the categorical variables we used a balanced sample and evaluated the sample by Pearson Chi-square statistic and Cramer's V. These test are based contingency tables an will help us identify if there is any relation between variables and the strength of that relationship. We did the tests considering alpha = .05 .

For the Chi-Square tests we found significant relation between the target variables and the list of variables below:

- Age brackets
- Own car
- Number of children
- Duration at current address
- Employment status
- Years of employment
- Principal bank of customer
- Marital status
- Type of residence
- State of the customers

Failing to find significant relationship in the variables:

- Debt ratio bracket
- Education level

- Gender

Nonetheless, the relationship strength obtained by Crammer's V is relatively low. We believe that we did not find significant relation between the debt ratio and the target variable due to the constrains in the amount requested by customers. The customers cannot request pass a certain amount of monetary unit and most of them request the max amount allowed. Since most of the customers request the same amount then the variable turns into a generic variable such as gender and education level. Notice that to these categorical variables we added other categorical variables. We show the analysis of the personal information only, same type of test were used for the MNO related features and the loans related features. With a defined set of variables, selected through statistical tests, we formed a dataset that was use for the exploration of the algorithms.

3.4 Selection of algorithm and target definition

As in formal banking, we wanted to predict the likelihood that a customer will repay or not the loan requested. We had a mix of numerical and categorical variables as seen in the Chapter 3. However, we did not had a defined algorithm to deal with this problem. Over the last few years, different classification techniques have shown success in a diversity of scenarios related to credit scoring.

3.5 Algorithm selection

This section reviews the different implementations used to deal scenarios related to risk prediction. First, the MobiScore (San Pedro et al., 2015) shows that Decision Tree (DT), Support Vector Machine (SVM) and Logistic Regression (LR) algorithms can create reliable scores when lacking of financial history. The main goal of the MobiScore is to reduce the default rate in credit cards for customers in emerging markets. For this task, MobiScore train classification algorithms mostly with mobile phone data gathered from the customers. The MobiScore concludes that mobile phone data can indicate the financial risk of an individual without the needing bureau information which is the core component in formal banking and developed economies.

In other case, Blanco et al. (2013) shows that Artificial Neural Networks (ANN) outperformed the traditional techniques for scoring. The ANN model was more accurate when predicting defaulters. The ANN was trained with both financial and non-financial from a Peruvian MFI. As in our case, the main goal was to reduce the type II error which in this context is to classify a customer with bad credit as a customer with good credit, therefore granting financial products. The borrowers

considered in Blanco et al. (2013) are micro-entrepreneurs. Financial data from their micro-enterprise is used as input to train the scoring models.

Bjorkegren and Grissen (2015) highlights how useful can mobile usage data be when used for credit scoring. For this case, the use of Random Forest (RF) proved to be accurate when differentiating between good customers and bad customers. The dataset used for train the RF contains borrowers from a emerging market in the Caribbean area. This dataset also lack of traditional financial data used to train scoring models in developed countries.

Bayesian Networks (BN) have also been used for credit scoring and outperformed traditional techniques (Biçer et al., 2010). Although the implementation of the BN algorithm was used in formal banking in Turkey, the variables associated with the building of the net are not entirely financial. Variables that relate the customer with a given MNO were used to identify good borrowers, presenting once again that non-traditional data makes impact while creating a reliable credit score.

With this evidence, there is not a clear algorithm that outperform the others. Each scoring technique has been able to succeed when dealing with non-traditional data in order to predict financial risk in emerging markets. This is an indicator that the financial risk of an individual can be estimated using pure behavioral data when the financial data is poor or existent. Then, in order to select a model, we decided to train several algorithms with the same dataset and focus their performance in the test set.

First, we needed a clear definition of the target variable. For this purpose we analyzed the behavior of the overdue loans. The main idea of this analysis was to determine what should be called an *overdue* for the algorithms.

As presented in Fig. 3.3, most of the customers that do not pay on time, end up paying within the next three days after the due date. These effect is related to some of payment methods available to the customers. Even if the customer pays on the day their loan is due, the payment transaction can take up to 48 hours to be registered. Therefore, seems like the customer did not pay in time.

Furthermore, in terms of business, a customer with some hours or a day of late payment was not considered as a bad customer. In this order, the target variable for an application x was set as following:

$$Target(x) = \begin{cases} 1 & \text{If number of days in delay } > 3 \\ 0 & \text{If number of days in delay } \leq 3 \end{cases} \tag{3.1}$$

With a clear definition of the target variable and a dataset structured as dataset A, we trained several models in order to select those with good performance. To compare the performance of the models, we focused on two metrics: the Area under Receiver Operating Characteristic (AUROC) curve as a measure for the accuracy and the true positive rate, where overdue is the positive class. We used 10-fold cross validation to test the performance and grid-search for setting the hyper-parameters.

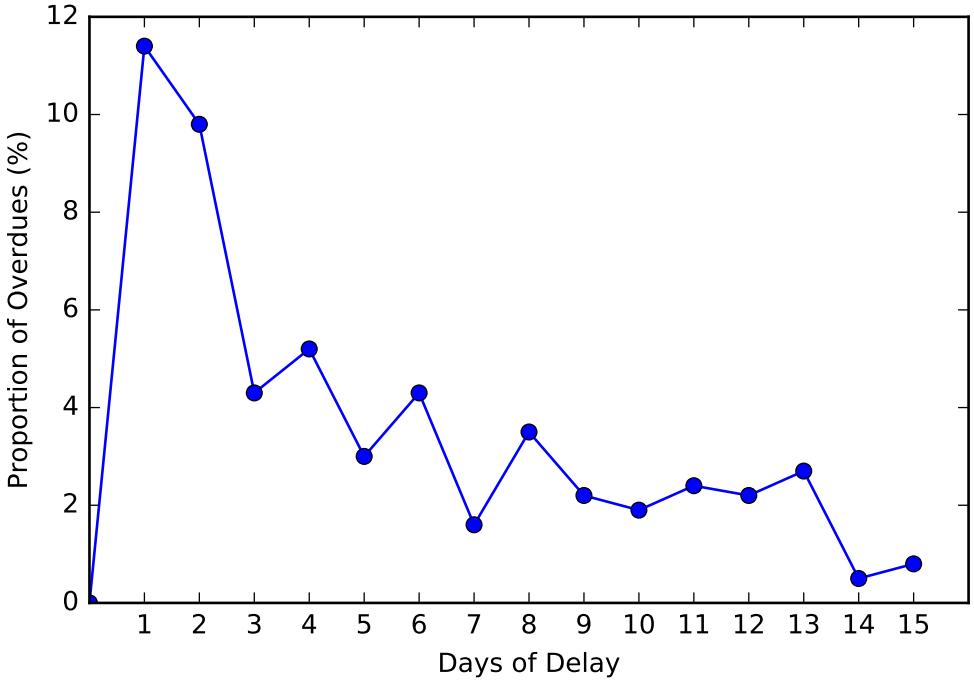

Figure 3.3: Distribution of overdue by days late.

Performance of Algorithms		
Algorithm	AUROC	True Positive Rate
Logistic Regression	90.41%	78.79%
SVM	85.06%	75.72%
Neural Network	89.29%	75.44%
Random Forest	80.65%	54.58%
Decision Rules	79.13%	67.36%
Decision Tree	71.84%	65.53%
Naive Bayes	78.41%	63.09%
Bayes Net	76.33%	56.20%

Table 3.2: Comparison of algorithms by AUROC curve and True Positive Rate.

As seen in Table 3.2, the top performers were LR, SVM and ANN. From these algorithms, we selected LR, SVM. This selection is the result of optimizing the cutoff of the algorithms, SVM prediction improved more than ANN with a sightly movement of the cutoff. As stated before, we would prefer to have a model with better prediction of the bad customers even if this means to lose some prediction

power in the good customers.

3.6 Set of target variable

We used a second test in order to validate the hypothesis used on the target variable definition. This test compared the AUROC changing the target definition to:

$$Target(x) = \begin{cases} 1 & \text{If number of days in delay } > 2 \\ 0 & \text{If number of days in delay } \leq 2 \end{cases} \tag{3.2}$$

Performance of Algorithms		
Algorithm	AUROC using 3 days	AUROC using 2 days
Logistic Regression	90.41%	84.97%
SVM	85.06%	74.60%
Neural Network	89.29%	81.50%
Random Forest	80.65%	79.31%
Decision Rules	79.13%	76.70%
Decision Tree	71.84%	70.88%
Naive Bayes	78.41%	75.65%
Bayes Net	76.33%	74.31%

Table 3.3: Comparison of algorithms for 3 days definition against 2 days definition.

As presented in Table 3.3, the performance of each algorithm is worst when using the 2 days definition for the target variable. The performance also decays when we set the definition to one day. It is evident that the smaller the *grace period* conceded to be labeled as good borrower the harder to set statistical difference.

For each algorithm, we selected the one which settings provide the best performance. We used the grid-search in order to test each combination of parameters, changing the definition of the target variable as well. The algorithms selected in this phase, LR and SVM, are the algorithms that we focus on for further experimentation. We also tried with different methods of variable selection. The features presented in Table 3.1 are the variables that optimize the key metrics set for evaluation. As a result of this test, we will use the target definition set in Equation 3.1 for further experimentation.

Chapter 4

Understanding credit scoring in emerging markets

In this chapter we present the structure of the different experiments undertaken. As presented in the Chapter 3, the target variable is a binary variable where 1 means overdue and 0 means paid on time. The classification task was modeled using supervised learning algorithms. Each entry will be processed in order to generate an output O which will be the estimated probability of going overdue.

We performed an experiment using each dataset described in Chapter 3. The main differences between the first two experiments were the variable selection procedure and the transformation method used. We test both transformation methods as presented in Van Gool et al. (2012), since it showed reliable scoring models when transforming the training dataset both with binary (dummy) variable transformation and with WoE transformation. We test both in order to see which one suits better to our case.

San Pedro et al. (2015) show how the different time windows can affect the performance of the credit scoring model. Following this reasoning, our third experiment focus on the time window considered for the training data. We test how the performance of our models will react when considering 3 months of data against 5 month of data.

4.1 First Experiment: Boolean Indicators

For the first experiment we used dataset A. Our first step was to extend the dataset A by transforming the categorical variables into several binary (dummy) variables. For variables with N categories we generated *N-1* new variables, e.g. *"Gender":{'Male','Female'}* was turned into *"Male":{'1','0'}*. Date variables were transformed into time unit elapsed since the date referenced, e.g. date of birth turned into age.

In the second step, we applied an information gain ranker in order to select the variables that contribute to the classification task and drop those that create noise. The final step was to build the model itself, for this task we applied supervised learning algorithms to the transformed dataset to build the models.

Following the results presented in Table 3.2, the two main classification methods that we focus on our experiments are LR and SVM. Not only LR and SVM performed better in terms of AUROC but also these models make no assumption on distribution of data, meaning they are quite tolerant to the input received (Sousa et al., 2016). We used 10-fold cross validation to test the performance of the dataset (Sousa et al., 2016). This validation technique can provide us better insights regarding the performance of the models in a generalized level. Through cross validation we can also avoid issues like overfitting. Our main metrics of evaluation were accuracy and approval rate. The recall is linked to the performance of the overdue class in the target variable. We choose this metrics so we can compare baseline set in the previous evaluation process. This comparison is key in order to know if we solved the main issue stated in Chapter 1 In our use case, we preferred to reduce the accuracy of the prediction of good borrowers in order to have better predictions in the overdue class, this approach is also presented in Blanco et al. (2013). For each algorithm we used grid-search for setting the hyper-parameters in order to optimize the evaluation measures. The output for each loan O_i was then transformed into a probability Pi following the expression:

$$P_i = \frac{e^{O_i}}{e^{O_i} + 1} \tag{4.1}$$

The models were applied to the new incoming loan applications on which the probability P for each new application was computed by the model and compared with a given threshold. Only applications from customers who applied for the first time and customers with only one paid loan before application where evaluated by the models.

4.2 Second Experiment: Implementing Weight of Evidence

The second experiment was based on dataset B. In order to obtain comparable results, we tested LR and SVM algorithms again. For this dataset we used the WoE coding of variables. The WoE focuses on the odds ratio. The WoE of a category C for an X variables is computed as:

$$WoE(X_C) = \left[ln \left(\frac{TotalGoods(X_C)}{TotalBads(X_C)} \right) \right] * 100 \tag{4.2}$$

In Equation 4.2, TotalGoods(X_C) refers to the number of borrowers that paid on time for category C in variable X. The same concept applies to TotalBads(X_C) but considering the borrowers that went overdue. Note that the value of the WoE is 0 when both distributions good and bad are equal. It indicates that the category evaluated does not allow to differentiate between classes. Using the recoded dataset, we calculated the Information Value (IV) of each variable. The IV for a category C of a variable X with n number of categories is as follow:

$$IV(X_C) = \frac{WoE_{Ci}}{\sum_{i=1}^{n} WoE_{C_i}} \tag{4.3}$$

The IV for variable X is the sum of the IV of each category C of X. It can be calculated as:

$$IV(X) = \sum_{i=1}^{C} \left[(TotalGoods_i - TotalBads_i) * ln \left(\frac{TotalGoods(X_C)}{TotalBads(X_C)} \right) \right] \tag{4.4}$$

After obtaining the IV for each variable, we follow the criteria presented in Siddiqi (2005) to select the relevant variables. Siddiqi (2005) propose brackets to relate the IV to the string of the relationship with the target variable. These brackets can relate the variable under analysis in order to determine a weak, medium, strong or existent relationship between the analyzed variable and the target variable. The relation can be set as follows:

When IV:

- Less than 0.02, the variable does not differentiate the Goods/Bads odds ratio.
- Between 0.02 to 0.1, the variable has only a weak relationship to the Goods/Bads odds ratio.
- Between 0.1 to 0.3, the predictor has a medium strength relationship to the Goods/Bads odds ratio.
- Equal 0.3 or higher, the predictor has a strong relationship to the Goods/Bads odds ratio.

In Table 4.1 we present an example of the calculation of IV and the corresponding WoE for each category of X. In the example, customers with value A in variable X will be replaced by -31,37. The IV for variable X is 0,156461 meaning it has a medium strength relationship to the Goods/Bads odds ratio.

For the experiment, we selected the variables with weak, medium and strong relationship to the Goods/Bads odds ratio and trained the models. In this experiment we trained both algorithms based on dataset B with the WoE transformation. As

Table of IV and WoE of variable X							
Categories of X	Goods	Bads	Total	Dist. of Goods	Dist. of Bads	Cumulative IV	WoE
A	163	144	307	0.1451	0.1986	0.016772	-31.37
B	78	15	93	0.0695	0.0207	0.059060	121.11
C	316	123	439	0.2814	0.1697	0.056534	50.60
D	23	12	35	0.0205	0.0166	0.000836	21.30
E	198	152	350	0.1763	0.2097	0.005775	-17.32
F	345	279	624	0.3072	0.3848	0.017483	-22.53
Total	1123	725	1848	1.0000	1.0000	0.156461	

Table 4.1: Example of IV calculation and WoE for variable X.

before, we used grid-search for setting the hyper-parameters in order to optimize the evaluation measures and 10-fold cross validation to measure the performance. The output for each loan O_i was transformed using equation 4.1 and then probability Pi of each application was compare to a given threshold. The models were applied to applications under the same criteria as the first experiment.

4.3 Third Experiment: Considering different time windows

The purpose of this experiment is to have an idea of how long should a model be valid. We intent to analyze two events that we believe could affect the performance of a credit scoring model in a long run. As Blanco et al. (2013) succeeded using macroeconomic variables in a microfinance problem. We decided to have a brief analysis and try to differentiate emerging markets from developed markets if in terms of financial indicators. First, we observe that emerging markets present a more volatile economy as presented in Fig. 4.1. This figure compares the GDP growth rate, the GDP refers to all the goods and services produced in a given country/regions in a period of time. By this definition, and based in Fig. 4.1, we can conclude that developed economies (US, UK, Switzerland) are more constant thought out the years in their productions of goods and services. This is different when we analyze emerging markets (Nigeria, Kenya, Tanzania). These constant movements into the economy can completely change the conditions of target population in a shorter period than expected. We also analyzed the yearly inflation in these markets, presented in Fig. 4.2. Inflation can be defined as the variation of the prices of goods and services over a period of time. It is evident that emerging markets suffer highers rates of inflation. This can easily translate into a lower acquisition level or payment capacity of an individual from an emerging market. Since, if the individual do not increase the corresponding income will be unable to keep up with financial

obligations.

On the other hand, we have the constant growing of number of customers with an average of 18.7% per month. This means that in less than 6 month we have double the amount of users. This have direct impact in the description of the population. The demographics, proportion of goods and bad borrowers, etc., all change in a fast pace. This growing is not limited only to the number of customers but also to different regions in the targeted country. These regions mix with the already exiting customers creating a whole new description of the population.

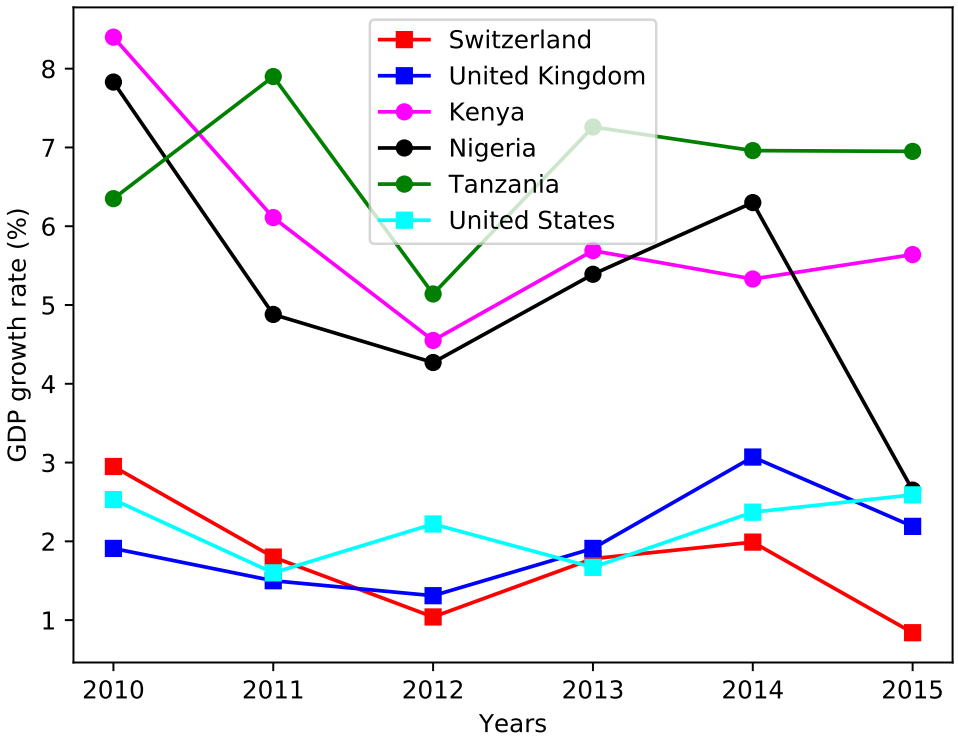

Figure 4.1: Timeline comparing the GDP growth rate of emerging markets against developed markets. *Data from World Bank (WorldBank, 2017)*

Therefore, we decide to train credit scoring models using different time windows. Based on the erratic movement of the GDP and the contently high inflation rate, our hypothesis states that the models should perform better when trained with a relatively short time window. The use of different time windows can be a determining factor as seen in San Pedro et al. (2015).

The third experiment is based in dataset C and dataset D. As seen in Fig 3.2, we created both datasets at the same time. However, we considered different time

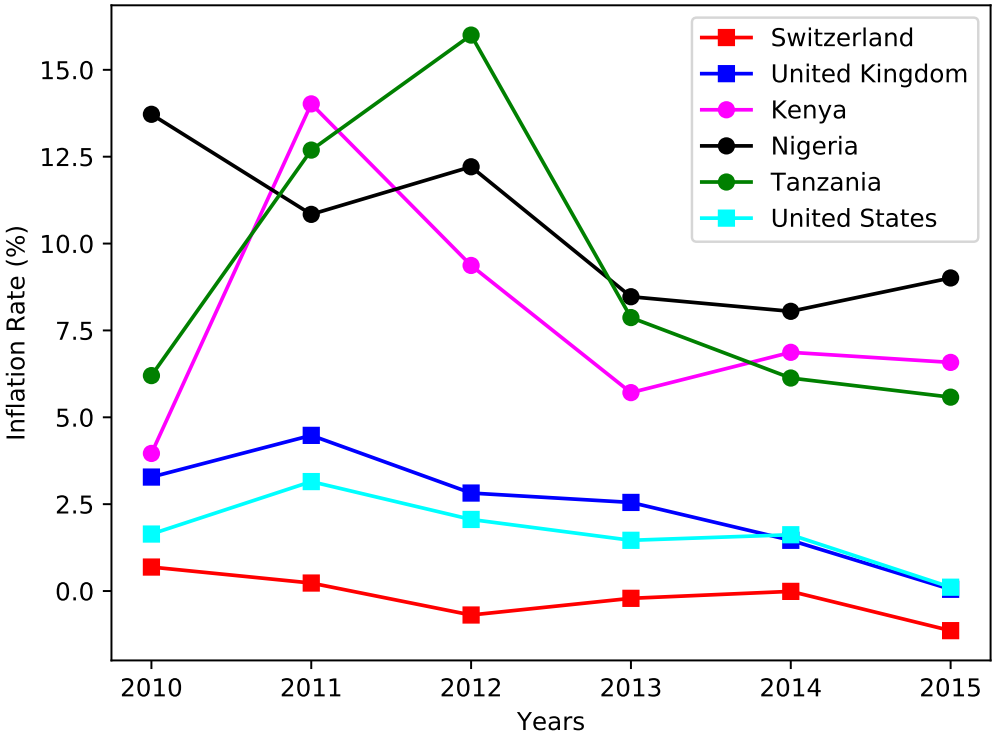

Figure 4.2: Timeline comparing the yearly inflation rate of emerging markets against developed markets. *Data from World Bank (WorldBank, 2017)*

windows. Using three months for dataset C and five months for dataset D. This means that if we generate both datasets today, an application made four months ago will not be considered for dataset C but will be considered for dataset D. We used transformed each dataset recoding with WoE. Afterwards, we selected the meaningful variables based in their respective IV. With the transformed variables we then trained a LR model and a SVM model for each dataset. For the training procedure we used grid-search for setting the hyper-parameters in order to optimize the evaluation measures and 10-fold cross validation to measure the performance as in previous experiments.

In terms of results and deployment, the first and second experiments were tested with real applications received after model building and deployment. As for the third experiment, the results will be obtained thought cross validation, comparing the AUROC and overall accuracy.

4.4 Analysis of results

In this section we analyze the result of the experiments described previously on this chapter. We group the results of the first and second experiment. As these experiments were applied to real cases, then we can compare with the performance of the previous evaluation workflow. Also, the first and second experiment both focus in the transformation method of the training dataset. The transformation method used for the third experiment was based on the result of the previous experiments.

The result of the experiments were compared to the baseline results. This baseline comprises the results of a relevant sample of loans that followed expert rules. In order to review the performance of the models, we need to compare the overdue rate. This rate represents the proportion of NPL over the total loans granted which met the approval criteria of the models. Only loans that surpassed their respective due date were considered. The length of both experiments is 40 days but were ran in different time periods. The second experiment kicked off 5 days after the end of the first experiment, with the deployment of the LR trained with dataset B (LR-B) model and the SVM trained with dataset B (SVM-B) model. At the end of the second experiment we proceeded to evaluate the time windows for the third experiment. The third experiment, as presented in Sec. 4.3, is evaluated by the AUROC.

4.5 Results of transformation experiments

In the first experiment we implemented the two selected algorithms referred in Sec. 3.5 and used A|B testing to select which model will evaluate the corresponding loan application. The number of the application was used to divide the population into two parts, this numbers is a sequence of natural numbers. Therefore, assigning the applications with odd application number to LR algorithm and the remaining (even application number) to the SVM algorithm. Even though we used A|B testing, we still calculated and stored probability of default for each application using both algorithms. This means that if the loan application in evaluation met the threshold for both models, then the result will be counted for both models as well. If the loan application only met the threshold for the scoring model, then the result will only be stored for that model in particular.

Therefore, the overdue rate represented in the first and second experiment shows the performance of each application that met the approval criteria of the corresponding algorithm.

The results shown in Table 4.2 compare the scoring models with the manual process.

The manual process refers to the all applications received before the beginning of the first experiment. This process was based on expert criteria pipeline. From the manual process, we considered only applications from customers who applied

Models Performance		
Model	Overdue Rate	Approval Rate
Experiment 1 LR	-26.45%	-4.33%
Experiment 1 SVM	442.19%	-91.91%
Experiment 2 LR	-115.50%	56.46%
Experiment 2 SVM	**-196.80%**	**251.53%**

Table 4.2: Improvements of models in relation to the baseline.

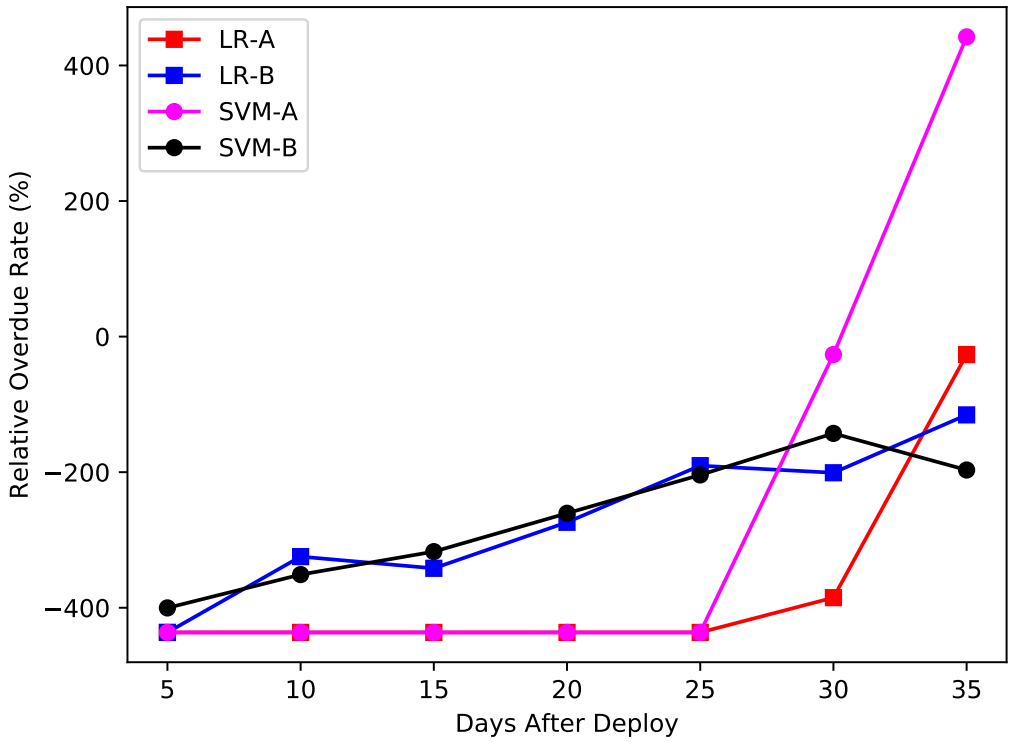

Figure 4.3: Comparison of overdue rate relative to baseline by days after deployment.

for the first time and customers with only one paid loan before. We selected these applications in order to match the same population targeted by the models.

The two main indicators to measure are the overdue rate and the approval rate. Our experiments were focused on these two indicators since they are the core of a stable and sustainable business model. A high overdue rate will make the business model unprofitable, while a low approval rate will not grant enough loans to even cover the operational costs.

First, we notice that both models of the first experiment failed to classify cor-

rectly the overdues, this translate into direct costs for the MFI. The LR from first experiment improved the approval rate, however the quality of the loans was really poor.

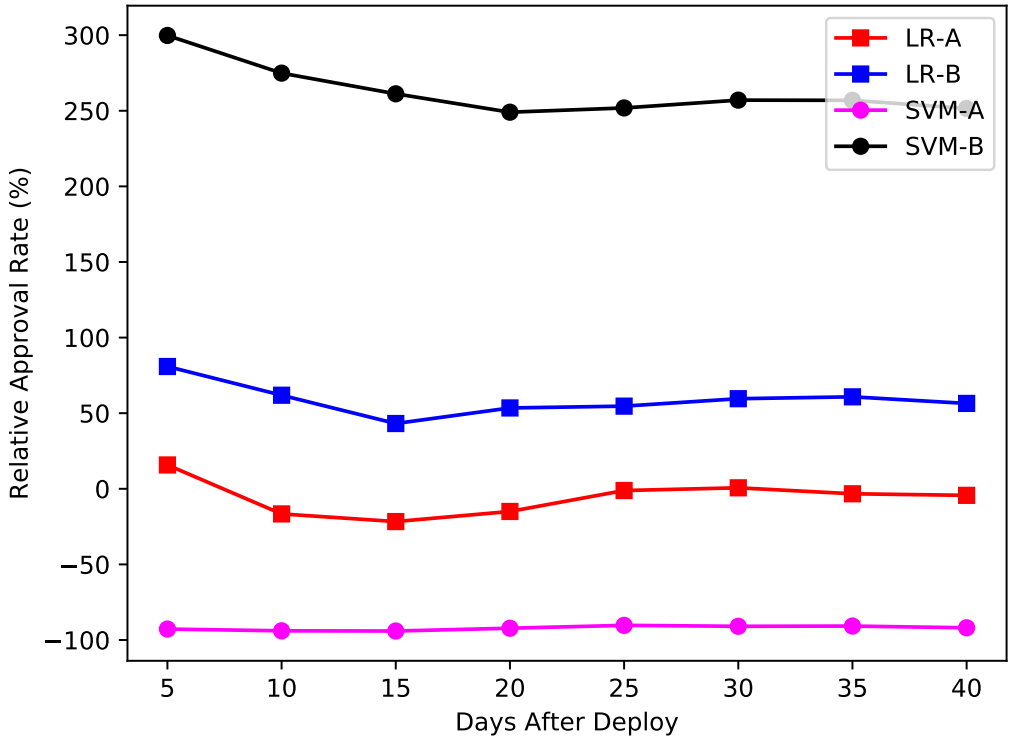

Figure 4.4: Comparison of approval rate relative to baseline by days after deployment.

In Figure 4.3 we compare the performance of the models in terms of classification error with real cases. The horizontal axis is the number of days after the deployment relative to the model in evaluation. Before the 30th day of deployment, both models trained with dataset A do not show overdue loans. This is due to the fact that the approval rate of LR trained with dataset A (LR-A) and SVM trained with dataset A (SVM-A) was relatively low and just a few loans with less than 30 days of length were granted. However, after the loans granted by LR-A and SVM-A reached their respective due date, an increase on the overdue rate can be observed. Even if LR-A and SVM-A achieved a good overdue rate, these models would not be suitable for the business due to lack of approval. The models approved less than the baseline with -4.33% and -91.91% decrease of the approval rate respectively. Models trained with dataset B show a better behavior overall. Loans evaluated by LR-B

and SVM-B presented less overdues and more evaluations were approved thought the evaluation by these models. As seen on Fig. 4.3, LR-B and SVM-B models are more stable on the overdue rate after the 30th day of deployment. These models are more constant and stable when compared to models created with dataset A. The LR-B and SVM-B models not only performed better in terms of overdue but also achieved higher approval rate than the manual process. As seen in Fig. 4.4, the SVM-B model was the best model in terms of approval rate. SVM-B was able to double the approval rate achieved by LR-B obtaining at the same time a sightly lower overdue rate. One of the main differences of the models, apart from datasets used for training, is the allocation of weights of the variables. Models trained with dataset A, gave more weights to variables related to the customer profile and loans conditions. The profile data of the customer is filled by the himself and cannot be verified, thus making this data not really reliable. Furthermore, loans amount conditions are capped to a specific value, meaning they can only make variations within a small interval.

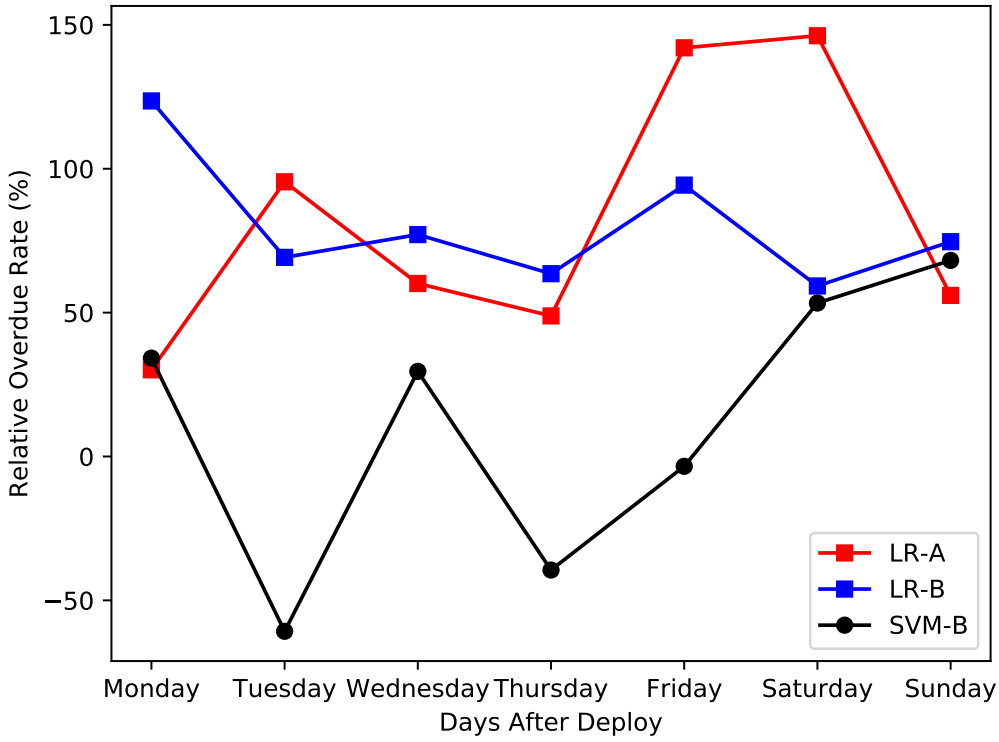

Figure 4.5: Overdue rate relative to baseline by day of the week.

On the other hand, models trained with dataset B allocated more weights in

variables related to historical data of the mobile phone of the customer. As the customer is not aware of its collection and/or cannot easily forged, these variables present a better profile of the customer and are more reliable, Both models trained with dataset B already surpassed the performance of the manual loan application selection process. More loans were granted when evaluated by the **LR-R!** (**LR-R!**) and SVM-B while the loans granted using the credit scoring evaluation presented less overdues.

While analyzing the performance of the models, we noticed some trends related to the nonpayment of the loans. As seen on Fig. 4.5, in general, the overdue rate tends to rise as the week progresses. This trend is also present in SVM-B which was the best model so far. LR-A also shows higher overdue rate in the last days of the week but concentrated on Friday and Saturday. LR-B shows less variation among the week days.

We did not consider SVM-A in this analysis due to the low number of approved loans with this model. The low number is directly connected to the extremely low approval rate.

Note that an overdue is not a total loss for the company since the overdue loans will enter into the recovery pipeline.

4.6 Results of third experiment

As for the third experiment, we compare the the AUROC. In Table 4.3, we present the comparison of the AUROC by time window considered. It is evident that the models trained with three months data perform better than the 5 months data. This verify the hypothesis presented in Sec. 4.3. The hypothesis suggested that the closest data (time related) can built more accurate models.

Models Performance for Third Experiment		
Model	3 months AUROC	5 months AUROC
Experiment 3 LR	95.42%	90.85%
Experiment 3 SVM	95.24%	90.77%

Table 4.3: Third Experiment: Comparison of AUROC using different time windows.

This kind of result is expected when looking at the complete overview of the business. For the current trimester the total number of customers increased by more than 50%. Moreover, the amount of granted loans for the same period of time resulted in more than 100% increase when compared to total granted loans prior to this period. These increases translate into a whole new population. The distributions by the different categorical variables is totally shifted. More important,

as the WoE transformation is based in the distribution of goods customers and bad customers per category. The weights assigned through the WoE analysis are outdated, thus the models are likely to decrease their performance.

This means that the models will be quickly outdated considering only the growing of the business itself. This fast business growing, combined with the instability of the market shown in Sec 4.3 can create a scenario when the scoring models are quickly outdated.

Chapter 5

Conclusions and Future Work

As we have shown in previous chapter, the use of credit scoring can be an useful tool to grant loans in emerging markets. Moreover, this problem seems to be one of the main upcoming challenges for MFIs in emerging markets. Everyday, these markets have access to more technology that can help them in a way we have shown in this study. This is the start of a new era for MFIs which have overcome obstacle since first implementations by Yunus and the Grameen Bank.

One of this obstacles was the jump from group lending to individual lending. However, at beginnings of 2000, Mark Schreiner proved that credit scoring could be an useful tool for individual lending at a microfinance level, but it creates another issue, the collection of data was too difficult. The collection of data is already history of the past, since mobile devices allow to collect thousands of data points in an instant, this data points are useful as we showed in the exploratory analysis. The exploration of the data collected indicates that non-traditional features can be significant when identifying the financial risk of the customer. This data, combined with classification algorithms, can deal with the problem we presented at the start of this study which was to estimate the risk of the applicant in order to grant or deny the loan request.

Even though the first experiment did not perform as expected, both models trained with dataset B and the WoE recoding, proved to be better than baseline in both metrics used for evaluation. SVM-B and LR-B not only improved the overdue rate and the approval rate, but also optimized the time of the loan approval pipeline. We believe that a two month gap between training phase and deployment phase affected the result for the first experiment. As we presented in this study, the fluctuations of the market combined with the accelerated pace of business growth can offset a model in a short period of time. Therefore, the gap between training and implementation could be a key factor for the success of the models.

5.1 Future Work

For future work, we will focus on using trends in the macro-economic environment. We believe this can be a determining factor of the performance of the credit scoring in emerging markets. Furthermore, we will study the predictive power of models with different time windows and models with streaming data as learning base. The deployment of the third experiment into the business will allow us to evaluate the performance with new cases. Since emerging markets present such a dynamic behaviour, we believe that a dynamic model would present better performance than the one achieved with static data. Both LR-B and SVM-B need to be studied for a longer period of time. This longer evaluation will help us identify for how long can a scoring model keep up with the initial performance.

Bibliography

Anderson, R. (2007). *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford University Press.

Bank, W. (2013). Financial inclusion: Helping countries meet the needs of the under-banked and under-served. Last visited 30-Apr-2017.

Barnes, P. (2016). E-commerce in emerging markets: The biggest growth opportunity. Last visited 28-Apr-2017.

Biçer, I., Sevis, D., and Bilgic, T. (2010). Bayesian credit scoring model with integration of expert knowledge and customer data. In *International Conference 24th Mini EURO Conference "Continuous Optimization and Information-Based Technologies in the Financial Sector" (MEC EurOPT 2010)*, pages 324–329. Vilnius Gediminas Technical University Publishing House "Technika".

Bjorkegren, D. and Grissen, D. (2015). Behavior revealed in mobile phone usage predicts loan repayment.

Blanco, A., Pino-Mejías, R., Lara, J., and Rayo, S. (2013). Credit scoring models for the microfinance industry using neural networks: Evidence from peru. *Expert Systems with applications*, 40(1):356–364.

Branch.co (2017). Branch. Last visited 15-Jan-2017.

Büthe, T., Yunus, M., and Jolis, A. (2000). Banker to the poor: Micro-lending and the battle against world poverty.

Chen, N., Ribeiro, B., and Chen, A. (2016). Financial credit risk assessment: a recent review. *Artificial Intelligence Review*, 45(1):1–23.

Cignifi (2017). Cignifi inc. Last visited 29-May-2017.

Cook, T. and McKay, C. (2015). How m-shwari works: The story so far. *Consultative Group to Assist the Poor (CGAP) and Financial Sector Deepening (FSD)*.

De Cnudde, S., Moeyersoms, J., Stankova, M., Tobback, E., Javaly, V., Martens, D., et al. (2015). Who cares about your facebook friends? credit scoring for microfinance. Technical report.

Fifer Mandell, A., Strawther, M., and Zhu, J. (2015). Inventure: Building credit scoring tools for the base of the pyramid.

Fotabong, L. A. (2011). Comparing microfinance models.

Harkness, T. (2016). *Big Data: Does Size Matter?* Bloomsbury USA.

Katakam, A., Frydrych, J., Murphy, A., and Naghavi, N. (2015). State of the industry 2015: Mobile money.

Katakam, A., Frydrych, J., Murphy, A., and Naghavi, N. (2016). State of the industry on mobile money: Decade edition: 2006 - 2016.

Kreditech (2017). Kreditech - providing access to credit for the underbanked. Last visited 29-May-2017.

Pickens, D., Porteous, D., and Rotman, S. (2009). Scenarios for branchless banking in 2020. *Washignton, DC: CGAP*.

Poushter, J. (2016). Smartphone ownership and internet usage continues to climb in emerging economies. Last visited 30-Apr-2017.

San Pedro, J., Proserpio, D., and Oliver, N. (2015). Mobiscore: towards universal credit scoring from mobile phone data. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 195–207. Springer.

Schreiner, M. (1999). A scoring model of the risk of costly arrears at a microfinance lender in bolivia. *Center for Social Development, Washington University in St. Louis.*

Schreiner, M. (2000a). Credit scoring for microfinance: Can it work? *Journal of Microfinance/ESR Review*, 2(2):105–118.

Schreiner, M. (2000b). A scoring model of the risk of costly arrears for loans from affiliates of women's world banking in colombia. *Women's World Banking.*

Schreiner, M. (2001). The risk of exit by borrowers from a microlender in bolivia. *Center for Social Development, Washington University in St. Louis, gwbweb. wustl. edu/users/schreiner.*

Siddiqi, N. (2005). *Credit risk scorecards: developing and implementing intelligent credit scoring.* John Wiley & Sons.

Sousa, M. R., Gama, J., and Brandão, E. (2016). A new dynamic modeling framework for credit risk assessment. *Expert Systems with Applications*, 45:341–351.

Stewart, J. (2014). Systems and methods for using online social footprint for affecting lending performance and credit scoring. US Patent 8,694,401.

Thomas, L. C., Edelman, D. B., and Crook, J. N. (2002). *Credit Scoring and Its Applications*. Siam.

Van Gool, J., Verbeke, W., Sercu, P., and Baesens, B. (2012). Credit scoring for microfinance: is it worth it? *International Journal of Finance & Economics*, 17(2):103–123.

WorldBank (2017). World bank. Last updated 27-Apr-2017.

Appendices

Appendix A

Distribution by gender	
Gender	Dist. by gender
Male	77,51%
Female	22,49%

Table A.1: Distribution by gender.

Distribution of customers with car	
Category	Dist. of customers
Own car	72,31%
Do not own car	27,69%

Table A.2: Distribution of customers with car.

Distribution of customers by type of home	
Type of home	Dist. of customers
Rented	63,27%
Family	26,78%
Bought	4,91%
Employer	5,04%

Table A.3: Distribution of customers by type of home.

Distribution of customers by marital status	
Marital status	Dist. of customers
Single	53.60%
Married	45.18%
Widowed	0.51%
Separated	0.71%

Table A.4: Distribution of customers by marital status.

Distribution of customers by duration at current address	
Years at current address	Dist. of customers
Less than 1 year	2,86%
1 years	7,44%
2 years	19,33%
3 years	15,38%
4 years	11,62%
5 years	7,57%
6 or more years	35,80%

Table A.5: Distribution of customers by duration on current address.

Distribution of customers by number of children	
Number of children	Dist. of customers
No children	56,98%
1	13,43%
2	15,58%
3	9,07%
4	3,41%
5 or more children	1,53%

Table A.6: Distribution of customers by number of children.

Distribution of customers by age bracket	
Age bracket	Dist. of customers
18 - 20	0,95%
21 - 25	14,67%
26 - 30	27,10%
31 - 35	29,24%
36 - 40	17,69%
41 - 45	6,54%
46 - 50	2,16%
51 - 55	1,08%
56 - 60	0,40%
More than 60	0,16%

Table A.7: Distribution of customers by age bracket.

Distribution of customers by employment status	
Employment status	Dist. of customers
Employed	61,60%
Self employed	38,16%
Retired	0,06%
Unemployed	0,05%
Student	0,14%

Table A.8: Distribution of customers by employment status.

Distribution of customers by level of education	
Level of education	Dist. of customers
Primary	0,12%
Secondary	12,40%
University	87,48%

Table A.9: Distribution of customers by level of education.

Distribution of customers by years of employment	
Years employed	Dist. of customers
Less than 1 year	37,74%
1 year	9,89%
2 years	10,33%
3 years	9,77%
4 years	8,00%
5 years	5,84%
More than 5 years	18,43%

Table A.10: Distribution of customers by years of employment.

Distribution of customers by debt ratio bracket	
Debt ratio bracket	Dist. of customers
[0% - 10%]	19,89%
(10% - 20%]	34,57%
(20% - 30%]	34,62%
(30% - 40%]	10,83%
(40% - 50%]	0,09%

Table A.11: Distribution of customers by debt ratio bracket.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	216,702[a]	9	,000
Likelihood Ratio	219,705	9	,000
N of Valid Cases	9356		

Figure A.1: Chi-Square test for age brackets

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	85,988[a]	1	,000		
Continuity Correction[b]	85,559	1	,000		
Likelihood Ratio	86,115	1	,000		
Fisher's Exact Test				,000	,000
N of Valid Cases	9356				

Figure A.2: Chi-Square test for own car variable

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	65,509[a]	5	,000
Likelihood Ratio	65,580	5	,000
N of Valid Cases	9356		

Figure A.3: Chi-Square test for number of children

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	5,754[a]	4	,218
Likelihood Ratio	5,756	4	,218
N of Valid Cases	9356		

Figure A.4: Chi-Square test for debt brackets

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	95,518[a]	6	,000
Likelihood Ratio	96,371	6	,000
Linear-by-Linear Association	34,372	1	,000
N of Valid Cases	9356		

Figure A.5: Chi-Square test for duration at current address

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1,588[a]	2	,452
Likelihood Ratio	1,589	2	,452
Linear-by-Linear Association	1,128	1	,288
N of Valid Cases	9356		

Figure A.6: Chi-Square test for education level

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	125,489[a]	4	,000
Likelihood Ratio	125,993	4	,000
Linear-by-Linear Association	103,748	1	,000
N of Valid Cases	9356		

Figure A.7: Chi-Square test for employment status

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	132,998[a]	7	,000
Likelihood Ratio	133,503	7	,000
N of Valid Cases	9356		

Figure A.8: Chi-Square test for years of employment

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	108,295[a]	17	,000
Likelihood Ratio	109,044	17	,000
N of Valid Cases	9356		

Figure A.9: Chi-Square test for principal bank of the customer

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	2,121[a]	1	,145		
Continuity Correction[b]	2,049	1	,152		
Likelihood Ratio	2,121	1	,145		
Fisher's Exact Test				,147	,076
N of Valid Cases	9356				

Figure A.10: Chi-Square test for gender

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	109,834[a]	3	,000
Likelihood Ratio	110,034	3	,000
Linear-by-Linear Association	92,792	1	,000
N of Valid Cases	9356		

Figure A.11: Chi-Square test for marital status

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	10,434[a]	3	,015
Likelihood Ratio	10,439	3	,015
Linear-by-Linear Association	,366	1	,545
N of Valid Cases	9356		

Figure A.12: Chi-Square test for type of residence

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	209,643[a]	36	,000
Likelihood Ratio	220,581	36	,000
N of Valid Cases	9356		

Figure A.13: Chi-Square test for state of the customer

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,152	,000
	Cramer's V	,152	,000
N of Valid Cases		9356	

Figure A.14: Symmetric measures for age bracket

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	-,096	,000
	Cramer's V	,096	,000
N of Valid Cases		9356	

Figure A.15: Symmetric measures for own car variable

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,084	,000
	Cramer's V	,084	,000
N of Valid Cases		9356	

Figure A.16: Chi-Square test for number of children

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,025	,218
	Cramer's V	,025	,218
N of Valid Cases		9356	

Figure A.17: Symmetric measures for debt brackets

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,101	,000
	Cramer's V	,101	,000
N of Valid Cases		9356	

Figure A.18: Symmetric measures for duration at current address

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,013	,452
	Cramer's V	,013	,452
N of Valid Cases		9356	

Figure A.19: Symmetric measures for education level

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,116	,000
	Cramer's V	,116	,000
N of Valid Cases		9356	

Figure A.20: Symmetric measures for employment status

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,119	,000
	Cramer's V	,119	,000
N of Valid Cases		9356	

Figure A.21: Symmetric measures for years of employment

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,108	,000
	Cramer's V	,108	,000
N of Valid Cases		9356	

Figure A.22: Symmetric measures for principal bank of the customer

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,015	,145
	Cramer's V	,015	,145
N of Valid Cases		9356	

Figure A.23: Symmetric measures for gender

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,108	,000
	Cramer's V	,108	,000
N of Valid Cases		9356	

Figure A.24: Symmetric measures for marital status

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,033	,015
	Cramer's V	,033	,015
N of Valid Cases		9356	

Figure A.25: Symmetric measures for type of residence

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,150	,000
	Cramer's V	,150	,000
N of Valid Cases		9356	

Figure A.26: Symmetric measures for customer state