Aalto University
School of Science

Predicting and Preventing Credit Card Default

Project Plan

MS-E2177: Seminar on Case Studies in Operations Research

Client: McKinsey Finland

Ari Viitala
Max Merikoski (Project Manager)
Nourhan Shafik

21.2.2018

1. Background

1.1. Credit Card Background

Credit cards are one of the major consumer lending products in the U.S., representing roughly 30% of total consumer lending (USD 3.6 tn in 2016). Credit cards issued by banks hold the majority of the market share with approximately 70% of the total outstanding balance. [1] [2] Bank's credit card charge offs have stabilized after the financial crisis to around 3% of the outstanding total balance. [3] However, there are still differences in the credit card charge off levels between different competitors. Being able to predict accurately which customers are most probable to default represents significant business opportunity for all banks. Bank cards are the most common credit card type in the U.S., which emphasizes the impact of risk prediction to both the consumers and banks.

Accepting a credit card means that you agree to certain terms. For instance, you have to pay your bills by the due date listed on your credit card statement. If you are severely lacking in payment ability, the credit card will be defaulted, which will affect your general credit status. A charge-off occurs when the bank decides it is not able to collect the payment. At this point it is usually handed off to debt collection agencies. This results in financial losses to the bank on top of the damaged credit rating of the customer and thus it is an important problem to be tackled.

Due to the significance of credit card lending, it is a widely researched subject. Many statistical methods have been applied to developing credit risk prediction, such as discriminant analysis, logistic regression, and probabilistic classifiers such as Bayes classifiers. Advanced machine learning methods including decision trees and artificial neural networks have also been applied [4]. The large extent of studies in this field will aid the project team in determining an appropriate methodology to achieve good results.

1.2. Project Background

Our client Kuutti Bank has approached us to help them to predict and prevent credit card defaulters to improve their bottom line. While the client has a proper screening process in place, they don't have active credit card default mitigation strategies leading to substantially higher default rates compared to their peers. The client has collected a rich data set on their customer base, but unable to leverage it properly due to lack of analytics capabilities.

In short, our goal is to implement a proactive default prevention program for the client by identifying customers with high default probability to improve their bottom line.

The client has collected a data set of 30,000 customers. It contains some demographic and payment history related features for each customer with a total of 26 variables. The features in the data can be divided into two categories, demographic data, and payment data. The payment data will not be available for new customers, since it contains a history of bills and payments. Demographic data includes features such as age, sex, education and marital status. Finally, the data set includes a binary indicator of default in the next month.

The data set is originally from a Taiwanese bank, collected from October 2005. However, two additional data features (location, employer) were added by McKinsey to add further possibilities and depth into the analysis. At least one published study by I-Cheng Yeh and Che-hui Lien uses the original data to compare the predictive accuracy of probability of default of six different data

mining methods [1]. However, the data used in this project is not a one-to-one match. In addition to academic research, the data set has been analysed in community-based platforms, such as Kaggle.

2. Objectives

The fundamental objective of the project is implementing a proactive default prevention program and identifying customers with high probability of defaulting to improve the client's bottom line. The challenge is to help the bank to improve its credit card services for the mutual benefit of customers and the business itself. An emphasis on creating a human-interpretable solution must be put into consideration in each stage of the project.

Even though plenty of solutions to the default prediction using the full data set have been previously done, even in published papers, the scope of our project extends beyond that, as our ultimate goal is to provide an easy-to-interpret default mitigation program to the client bank.

In addition to default prevention, the case study includes a set of learning goals. The team must understand key considerations in selecting analytics methods and how these analytics methods can be used efficiently to create direct business value. McKinsey also sets the objective of learning how to communicate complex topics to people with different backgrounds.

3. Tasks

3.1. Default Prediction Algorithm

The default prediction algorithm is, like its name says, a model used to predict credit card defaults based on our dataset. We think our best bet is to implement a machine learning algorithm since this has been previously done with a similar dataset [1]. There are many approaches to this problem and we also have to take into account the situation of the bank. For the bank, it would be the most beneficial to prevent defaults by filtering out risky customers by not giving them credit cards at all. However this means that we cannot use the credit card payment history for the prediction since that would not be available at time of issuing the credit card. This would naturally make predicting much harder, as we would be using just a fraction of the data but it would be more beneficial for the bank financially. The financial benefit of the bank must be kept close to the algorithm development as predictive accuracy is not the only metric with which the algorithm performance is evaluated.

3.2. Financial Model

The financial model aims to simulate the bank's credit card functions and cash flows connected to them. The goal of the financial model is to provide a connection from the default prediction algorithm to the actual financial performance of the bank. The primary utility of the financial model is to provide a validation tool for the default prediction algorithm. With it is easy to see how actions taken based on our suggestions would affect the bank, its customer base and most importantly the bottom line. The financial model also helps us to identify the key aspects of the bank's cash flows and provides input on how the prediction algorithm should be improved in order to gain the most financial benefit. The financial model could be thought of as the objective function that we want to optimize and the optimization is done by filtering away people with our default prediction algorithm but as stated above it also yields other benefits.

3.3. Customer Segmentation

This part includes a descriptive and basic quantitative analysis of the data set. The idea is to understand the distributions of each variable, how they correlate with default, and do they have useful overlaps within each other such that differing segments could be identified.

The analysis of data will help us understand the bank's business model and customer base. Customer segmentation should help by creating an easily interpretable default prevention solution due to the lack of inputs when dealing with new customers. A complicated "black box" model, the results of which cannot be put into real life application considering the end user, will not satisfy our goals. Customer segmentation is the process of providing a human interpretable interface for our model, so that the bank can draw meaning from our results. We also aim to increase the understanding of aspects of the data that are essential for improving our model and the business situation of the bank.

The inputs of the data can be divided into two clear categories, which are the demographic data available from new customers, and the historical payment data which will be mainly used for training the model.

Demographic, or categorical variables in the data are age, sex, marital status, education, location, employer, and a balance limit set by the bank in its original screening process. Historical data consists of payment amounts, bill amounts and a categorical indicator of payment status for the previous payment. This data is available from 6 months.

3.4. Implementing the Program

Implementation of the default prevention program as a task consists of applying the prediction algorithm (model), financial model and customer segmentation tasks into an end product that satisfies the project goal. The results of our analysis should be put together such that they can be easily incorporated into the bank's business.

The end product must satisfy a set of criteria and answer key questions that will help the bank improve its bottom line by giving instructions on how to handle clients and when to issue credit cards. The prevention program must communicate the results and methodology of the model in an efficient way. This should result in answering questions such as:

- What are the business implications of the program?
- What are the recommended steps for the bank?
- What mitigating actions should be taken to improve operations?

The team has been given a lot of freedom on how to approach the problem and which methods to use. The main focus of the project is the final goal of implementing a proactive default prevention program. No strict requirements are made on which methods the team should use or how the final solution should be presented. To develop the form and structure of the final application we will run through iterations of the development cycle. The end product of an iteration of the cycle will go through feedback which will redefine and clarify the goals for the next iteration so that the product can be improved in the development phase.

4. Schedule

As described in the implementation task, the schedule will be implemented in iterations of the development cycle. The timeline and due dates of the project are based around the deliverables of the course. The development cycle is described in Figure 1.

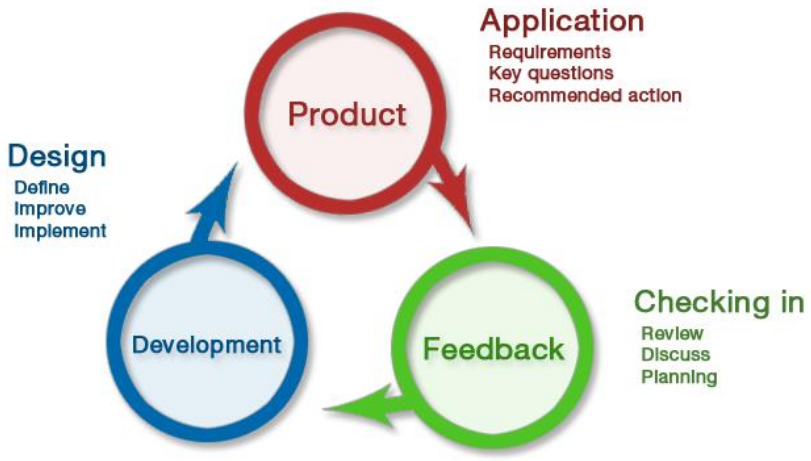

Figure 1: Development cycle.

The timeline of the project is shown in Figure 2. The development cycle phase will begin after the completion of the first draft of the product. After the product is complete, each development cycle continues with a round of feedback, which is applied in development and then put together into a new product. Each cycle is planned to last 2-3 weeks, which is why they are shown overlapping in the timeline figure. After the planned completion of the first draft at week 7, the team has enough time for about 3 development cycles depending on the time needed to complete them. Meetings with McKinsey and course management will be planned according to the progress of these cycles.

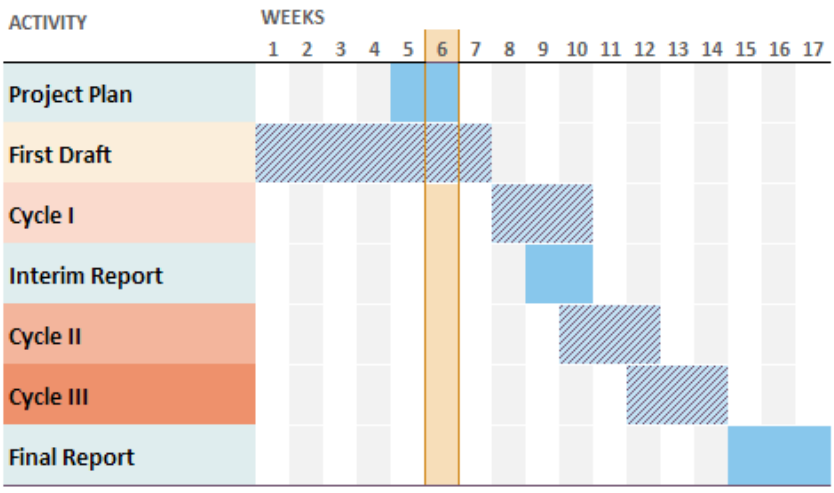

Figure 2: Timeline of the project, divided into 17 weeks.

5. Resources

Our project team includes three students in Mathematics and Operations Research. In our team, we are ensuring to distribute the work among the project members taking into consideration that each person has strengths in certain part in the project. Max has been assigned the role of the project manager.

During the project, we have a great support and assistance from both Professor Ahti Salo (Aalto University) and course assistant MSc Ellie Dillon (Aalto University). The course webpage includes a lot of helpful information. Our contact member inside Mckinsey is Arto who plays an essential role in guiding us and evaluating our progress and ensuring that our work meets the client's requirements.

In addition to the group's internal resources and the utilization of the project client McKinsey and course staff, resources include published papers and community-based research on the data set. The main resource of literature will be in academic research, but community resources offer starting points and tips as well. The academic literature offers an insight especially into the development of our predictive model and the methods used there.

6. Risks

Table 1. illustrates the risks affecting our project. Each risk is defined by its likelihood and impact and how to reduce its effect. The risks are also well described below the table.

Risks	Likelihood	Effects	Impact	Mitigation measures
Bad performance model	Low	Having no functional end product.	High	High qualified research
Not achieving the true implementation for the bank's current situation	Low	Final product is not satisfying the client's requirements.	Low to moderate	Working on the main objective together with the bank
Member absence	Low to moderate	Increasing the workload done by other group members	Low	Scheduling regular meetings and distributing the workload evenly on group members
Problems with data	High	Not accurate nor desirable results	Low to moderate	Finding an algorithm that is robust with respect to false negatives

Table 1: Risks that can be faced during the project.

A major risk in our project is that our model will not perform on the level that is required for a functional end product. One possibility could be that we are unable to find an algorithm that can predict defaults with good enough precision to be useful in filtering out customers with risk of defaulting. Also, since our dataset is heavily disproportionate there are roughly 4 times as many not defaults as defaults, we need an algorithm that is robust with regards to false negatives. In other words, an algorithm that labels everyone as not defaulting yields a high accuracy score but has no value in differentiating the defaulters. This risk is quite high compared to our other risks since our data doesn't seem to have any clear correlations between defaulting and other features. In addition, we are yet to come up with any well enough performing algorithm after a month of playing with the data. However, this risk can be mitigated by careful research since this phenomenon and even the same dataset has been previously investigated.

Another risk closely related to the previous one is that our model is not actually implementable to the bank's current situation. For example, we might be able to predict default reliably with six months of payment information but at that point, there would no options for the bank to act on it and the end result for the bank would be the same. This risk may not be as high as the previous one but it needs to be recognized because our goal cannot be pure predictive power but rather the applicability of our prediction to the current situation of the bank. This risk can be mitigated by keeping constantly in mind our objective and the whole process of credit card loans and thinking which approach is the most beneficial for the bank. Another aspect of this, is also the usability of the product by the bank since our product is not only a predictive model but also the interface by which it can used by the bank. We must ensure that we are speaking of the same things and getting feedback on the aspects that need improving as well.

A bit of a different risk which is not related directly to our product, is our functionality as a team. In order to draw the best possible result out of our team, we need clear communication and task management. This helps with even distribution of the workload, and making sure that everyone has something to work on and our project is constantly going towards its goals. As with every group project, communication can always be improved upon and it is a significant risk that task management is optimal. However, as long as the project is going forward, the impact of this is not necessarily that large. This risk can be mitigated with regular meetups and observing the work done by other group members. In addition, clear definitions of different tasks can help with evening out the workload but so far a more relaxed division of tasks has been working out fine.

References

[1] Federal Reserve. (2018) "Consumer Credit Historical Data", Federal Reserve G-19. [Online]. Available at: https://www.federalreserve.gov/releases/g19/HIST/default.htm

[2] Federal Reserve. (2017) "Report to the Congress on the Profitability of Credit Card Operations of Depository Institutions".

[3] Federal Reserve. (2018) "Charge-Off and Delinquency Rates on Loans and Leases at Commercial Banks". [Online]. Available at: https://www.federalreserve.gov/releases/chargeoff/default.htm

[4] I-Cheng Yeh and Che-hui Lien. (2009) "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients", *Expert Systems with Applications*, 36, pp. 2473-2480.