Aalto University
School of Science

Predicting and Preventing Credit Card Default

Interim report

MS-E2177: Seminar on Case Studies in Operations Research

Client: McKinsey Finland

Max Merikoski (Project Manager)
Ari Viitala
Nourhan Shafik

11.4.2018

1. Objectives

The general objective of improving the bank's credit card operations by preventing credit card client default is clear and does not change. We aim to implement a proactive default prevention guideline to help the bank identify and take action on customers with high probability of defaulting to improve their bottom line.

The project plan divides the work into four tasks which are the default prediction algorithm, the financial model, customer segmentation, and finally implementing the program by applying these tasks into a suitable method of default prevention.

The end product should include a recommended set of actions to mitigate the default and a clear explanation of the business implications. The interpretability and adaptability of our solution needs to be emphasized when constructing the solution. The bank needs to be represented a solution that can be understood and applied by people with varying expertise, so that no further outside consultation is required in understanding the business implications of the decisions.

Several solutions to closely related credit risk prediction problems have been presented in literature, however the goal of our project extends beyond that, as we must provide an easy-to-interpret default mitigation program to the client bank.

2. Tasks

2.1. Default prediction algorithm

In implementing an accurate algorithm for default prediction, we have tried several machine learning algorithms based on the literature review, client suggestion and experience. Factors such as model interpretability and accuracy must be weighed when considering which is the best. I-Cheng and Che-hui [1] found that artificial neural networks produced the most accurate results on predicting real probability of default in the original dataset. Classification trees were the second best choice in this regard, with other methods such as discriminant analysis, logistic regression and K-nearest neighbours classifiers producing inferior results. The project team found discriminant analysis to produce relatively low error rates, but high share of false negatives. Decision trees avoid this problem, but do not improve accuracy much.

By suggestion of McKinsey, a gradient boosting method was also implemented. Considering the positive results of the decision tree classifiers and with gradient boost also being a tree-based method, it is expected to have good results. What we are ultimately aiming for is a compromise between models that offer interpretability for the bank to develop their business and for us to communicate the results, but also the predictive power of a black box e.g. a combination of a decision tree and gradient boosting or a random forest.

The next steps in this task involve improving validation methods to optimize our results and testing new algorithms. This will be partly done in cooperation with McKinsey. In validation, things such as considering the financial impact of false negatives or positives versus general prediction accuracy can lead to better results.

2.2. Financial model

At this stage of the project, the financial model of the bank has been formed according to the goals in the project plan and it fills the requirements set for it. The financial model describes the current financial situation of the bank and reveals customer base's relationship to the bottom line of the bank. It can also be used to calculate the financial effect an individual customer has on the bank. This means that we are also able to use it to investigate the effect our actions have on the bank's financial situation. For example, finding out the monetary value of the filtering of high risk customers.

Certain assumptions had to be made in the making of the model since our dataset is not a perfect sample of the customer base. For example, we had to estimate a certain level of bias towards defaults in the dataset since it contained roughly three times more defaulters than other comparable banks. This also means that we are not quite able to directly translate results estimated from the dataset to the entire customer base. For example, the bank has a large number of customers that do not actively use their credit card hence, there is no risk of losing money in default. However, these customers still provide passive income in the form of yearly fees. Based on the data, the winning strategy for minimizing money lost on default, would be to get rid of all the customers who actively use their credit card, which clearly is not the optimal solution when considering the bigger picture.

The work is still in progress on finding what would be the "objective function" of our optimization and this is also closely related to the debt profiles of people our default prediction algorithm will be able to single out effectively.

2.3. Customer segmentation

The main goal of the customer segmentation task is to provide a foundation for the interpretability of our end product. Segmentation allows us to focus on our default prediction, and measures the financial impact of decisions at a customer segment basis instead of a single customer level. This will help generalize our results and create a more easily adoptable solution for the client bank.

Currently, customer segmentation has been done using a decision tree classifier tested with different parameters to obtain appropriately sized groups. This approach is suitable for our purposes, since new trees can be trained very quickly and effortlessly, and the results can be input into our workflow easily. Decision trees also have the great advantage of handling different types of variables well and creating an intuitive logical representation of the segments that can also be visualized. Also, the choice of using a decision tree to classify groups into low or high default risks is supported by their accuracy in the analysis of I-Cheng et al. [1] A uni- and bivariate analysis of the data was also conducted in the early stages of the project, which can be referred to when interpreting and describing the segments.

The actual segmentation is done by selecting end leaf nodes of the tree that have a high ratio of defaulting clients and targeting those as our risk groups. Depending on the distribution of default ratios in different segments, a low-medium-high risk grouping could also be done instead of a low-high risk grouping. The choice of parameters in the decision tree classifier can affect this, but decision trees tend to be more sensitive to change in data than parameters. The downside of decision trees can be their tendency to overfit, however that is partly mitigated by our large dataset and the limitations in the parameters of the tree, since this classifier is not used for prediction purposes.

Some of the future work in this task includes optimizing the process by attempting to find a suitable amount of customer segments and improving the results with feature engineering. Decision trees handle continuous variables, but for the purposes of segmenting, variables such as age and balance limit were categorized into equal-length bins. The most complicated variable in respect to finding a proper way of representing is geographical location. Another challenge is that although the interpretation of decision trees is very easy, it does not mean that the results are instantly translated into real-world context. To satisfy the goals of the project and to justify the use of this segmentation approach, meaningful and descriptive segments with between-group variation must be obtained. Both a statistical and contextual approach to solving this should be taken. For example, Ho Ha and Krishan [2] used ANOVA to test for between- and within-segments differences among their segmentation of credit card customers to validate the optimal amount of segments, in addition to a descriptive analysis of the variables to identify and characterize the segments.

2.4. Implementing the program

The implementation of the default prevention program involves using the customer segmentation, default prediction and financial model to create a recommended set of actions to improve the bank's bottom line. This can be done by preventing defaults and describing the consequences and impacts of these actions in business terms. Currently, we are implementing a method of separating the data to train a prediction algorithm and performing customer segmentation, and then testing the model and measuring the financial impact of these segments using an unseen validation set of customers.

The benefit of this approach is that we can present risky customer segments and their financial effect in terms that the client can understand, while justifying our decision making with the default prediction algorithm.

Ho Ha and Krishan [2] used a similar approach to predict the duration of delinquency (neglecting debt payment) of credit card customers. First, customers were segmented into different groups based on their payment and transaction history, after which their predictive model was trained to predict delinquency in these groups. This way, they could make conclusions on how repayment of debt could be improved with certain types of customers. The approach is comparable to ours, since we are attempting to provide a solution by targeting a select group of customers.

3. Risk management plan

Table 1. The updated risk management plan of our project.

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 or stagnating progress	Moderate	Scheduling regular meetings and distributing the workload evenly on group members
Customer segmentation results being sub-optimal	Low	Bad segmentation can result in groups that the bank has difficulties targeting	Moderate to High	Validating the customer segmentation results properly
Problems with data	High	Not accurate nor desirable results	Low to moderate	Finding an algorithm that is robust with respect to false negatives

4. Schedule

Meetings and discussion with our client will shape the schedule and workload of the final weeks of the project. The last weeks will be more work intense than so far. More frequent communication both within the group and with our client McKinsey will be needed to achieve the goals that were set in the project plan.

The financial model is ready and currently satisfies its requirements. As for the other tasks, customer segmentation still has room for improvement, and the optimization of the prediction algorithm will be weighed more towards the end of the project when a minimum-requirement satisfying end product is improved on.

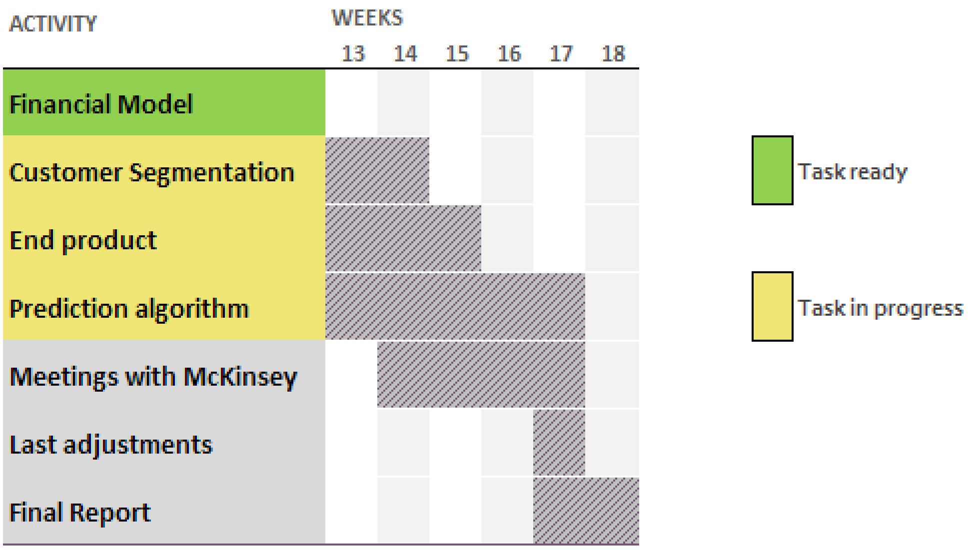

Figure 1: Schedule of the final weeks.

References

[1] 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.

[2] Sung Ho Ha and Ramayya Krishnan. (2012) “Predicting repayment of the credit card debt”, *Computers & Operations Research*, 39, pp. 765-773.