Data Storm
V 1.0

Team Appstat / ds1061

- Rahul Kalubowila
- Senuri Guneratne
- Anuradha Dissanayake

https://github.com/r3babba/datastorm-repo

F1: 0.82366

EXECUTIVE SUMMARY

The financial crisis of 2007-2009 highlighted the importance of risk management within financial institutions. Particular attention has been given to the risk management practices and policies at the mega-sized banks at the center of the crisis in the popular press and the academic literature. Secondly, unlike many traditional loans, credit card borrowing does not require consumers to post collateral which may place a greater risk on the lender. Hence it is important to identify the lead indicators of credit card default. From the data given in the case study the following insights were extracted.

Next month default risk decreased as the balance limit increased. Usually financial institutes allocate balance limits based on a resolver's financial capacity. According to Abdul-Muhmin and Umar (2007), the tendency to revolve is significantly higher among males. We can observe it here as well.

In Marital factor we have a drawback in the dataset. As married resolvers aren't represented.

According to Wickramasinghe and Gurugamage (2009), age has been found to be one of the significant demographic and socio-economic characteristic in describing consumer credit card practices. From the dataset we can see that the lowest risk of default is for the ages between 31-45. And the highest for 65 and above.

Focusing on to the engineered features we can see that risk is minimum for resolvers who had never defaulted during the six months. Looking at the consecutive default payments we see that the risk is significantly higher for resolvers with 6 consecutive defaults.

Finally, when we standardize and categorize the times of resolvers paying habits, we were able to see that resolvers who are more likely to pay late are 48.6% likely to default.

Hence using analytics, we can Identify the lead indicators of credit card default and based on these indicators, generate a list of priority clients with high probability of credit card defaults. Thus, by identifying these priority clients we can develop personized marketing interventions to reduce credit card default.

BUSINESS INSIGHTS

High Risk Resolver

- Targeted Counselling for high risk resolvers.
- Following up on default payments made in the current month. i.e. Communicating via email, telephone etc.
- Consumer education on healthy spending behavior.
- Have high screening process for senior citizens.
- Better financial assessment procedures to identify resolving patterns.
- Focus on consecutive defaulters.

Low Risk Resolver

- Monitoring unusual behavior
- Screening out credit risky customers
- Issue more cards to high income consumers.
- Target graduates.
- General financial awareness campaign

Model Analysis

Our baseline for the analysis was a random forest model, we used sampling techniques to balance the response variable and with the combinations used achieved a 0.93 F1 score but it performed poorly in the test dataset at Kaggle. Then on day 2 we worked with various classifiers such as Nearest Neighbors, Neural Net, AdaBoost, Naive Bayes, QDA and the AdaBoost model gave a good score both on site and on the Kaggle Set.

Finally, on day 3 we optimized a XGBoost model which gave us a marginally better score at the Kaggle Test Set which we used as the final model.

Interpretability wise we would recommend the random forest model since decision nodes can be analyzed further to deeply understand hidden patterns. Overall all models performed similarly except for the first model that we suspect was overfitted for the training set.

Methodology

Python Programming language & SPSS Software was used to carry out the Model Fitting for Predictive Analysis & Descriptive Analysis respectively.

Feature Engineering

- ZPay_Time – Nominal

This is the summation of standardized values of PAY_JULY, PAY_AUG, PAY_SEP, PAY_OCT, PAY_NOV, PAY_DEC categorized into 3. Higher positive values suggested typically late payments by the resolver. And negative values suggested early payments. Thus, the variable was categorized into 3 for positive, 0, negative values. We were able to see a high probability of defaulting next month for typically late resolvers.

- No_Def_Paym

This is the total number of payments that the resolver has defaulted. Here we were able to Identify that the probability for a resolver to default is significantly lower if he or she has not made any default payments in the past 6 months.

- Consec_Def

This is the total number of consecutive default payments made by a resolver. We were able to see a significance increase in the probability for defaulting for a person with 6 consecutive defaults.