Team Name - DECODERS

Kaggle Username- DS1009

Display Name- Data Storm 1.0

Team members- Dimithri De Meraal

Pulasthi Ekanayeka

Chamath Ekanayeka

Github repository link - https://github.com/dimi-3/DECODERS-Datastorm

Highest F1 score achieved- 0.82516

1. Introduction

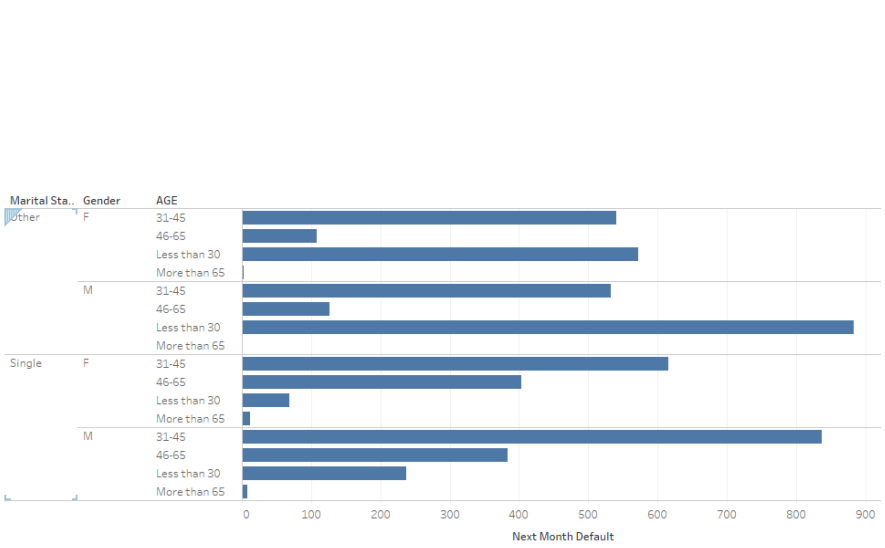
Credit Card facility is one of the major attraction to attract new customers and to upkeep the bank profitability with CC interest rates and charges. In order to up keep the profitability of the bank, it should not contain any bad debt or defaulted settlements.

Main objective of this challenge is to build a model which can identify priority customers who will default the instalments in the next month by using demographic and past transactional data of the priority customers. With the use of analytical tools, we will solve the problem in hand and provide recommendations from a business perspective.

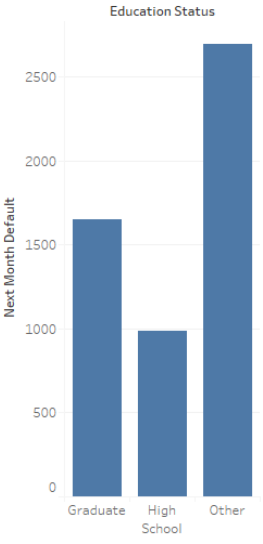
2. The approach in brief and the tools that were utilized to crack the case

We initially used Ms. Excel to start off our analysis by finding relationships within variables and used tableau to provide us with a graphical representation of the relationships and data. Then we decided to model it with python, but we used the following pre-processing steps before modelling it with python. We encoded the categorical and ordinal variables to numerical variables. Then we standardized the data to reduce over fitting and fit the model.

Task given to us was to identify priority clients who will default the next month. Approach used was, we identified how the paying patterns of clients are, the total due amounts and then we tried to identify the relationship between them through a regression and visuals.



(a)



(b)

As feature selection the columns we selected for the prediction were, Balance limit, gender, age, education status, marital status, sum of due amounts, client ID and next month default. Based on these columns we did feature engineering to increase the accuracy of the prediction model.

3. Feature Engineering ideas worked and reasons for designing them

Feature Engineering is when any useful attribute is being processed using domain knowledge to extract the true features in the raw data, using data mining.

Feature engineering process:

- **Brainstorming or testing features -** We initially used MS Excel and Tableau to create a simple regression and have a graphical representation of the data set to identify and decide which variables to use, what features to be selected and how to engineer those features to an advanced model.
- **Deciding and creating features -** We first cleaned the data through an encoder to convert the categorical/ordinal variables to continuous/numerical variables for machine reading purposes, converting columns to float type and standardized data and the model to reduce overfitting before creating the model to have a more accurate result. We created and summed the total due amount to July.
- **Check how the features work with the Model**-We were able to check the accuracy of our model through the accuracy checker provided kaggle.
- B**rainstorming, Creating and Improving the features -** We initially thought and decided to use random forest and linear regression to model but we were able to determine that it was not up to our expectations. Therefore, we decided to use Cat-Boost algorithm to create a model with more accuracy.

4. Final Model and how it was reached

The final model was reached using Cat-Boost algorithm. It gave us the highest accuracy score above linear regression, random forest and decision tree. Cat-Boost is known as an open source gradient boosting algorithm used to solve regression, classification and ranking problems outperforming most of the algorithms.

In our model, feature selection is done to allow the machine learning algorithm to train faster, improve accuracy of the model and to reduce overfitting. Through this we identify and select only the most accurate variables for the model to reduce complexity. Through feature engineering we developed the variables further to increase its usefulness and applicability.

Before running the algorithm, we took the sum of months to July and then ran the algorithm to create an accurate result after feature selection and engineering. Cat-Boosting allows the use of categorical variables without any pre-processing.

5. Business Insights (Specifications and Recommendations)

One of the main objectives of finding the priority clients who have a higher risk of defaulting credit cards is to gain a competitive advantage in terms of profitability. We were able to get a deeper insight on which variables were affecting the credit card defaulting patterns. We observed that married women and men aged less than 30 years defaulted higher while unmarried men and women aged between 31-45 default the most (Visual (a) above). We can observe that women and men with other qualifications default the most (Visual (b) above).

Recommendations

1. Reducing the credit limit by 25% for each credit card defaulted month.

Objective: Reducing the defaulted credit amount being a huge sum of money and maximum months a priority customer can default payments will be reduced to 4 months since credit limit is reduced by 25% each defaulted month.

2. Adjusting the credit card limits of potential credit defaults

Objective: Reducing the bad debt accumulation by reducing the credit given

Using the model, we developed for the competition, the banks will be able to determine potential credit card risks and risky priority customers and implementing a mechanism to automatically adjust the credit card limits.

3. Implementing an automated system to adjust interest rates according to the risk level of customers

4. Credit limit is normally determined at the initial stage considering their repaying abilities, but we have seen that some of the priority clients who had been considered to have a higher repaying capacity and was given a higher credit limit at initial stage have defaulted after some months. We recommend and urge the bank to monitor the repaying ability of the clients quarterly and adjust the limits and interest rates accordingly.