Data Storm 1.0

Submission by AlphaZero (ds1048)

Introduction

Quoted from the Problem Description

The task was to work on the dataset given and predict the customers who might default credit card payments. A default can occur when a borrower is unable to make timely payments, misses payments, or avoids/stops making payments.

As a result of credit card default, bank accounts close, non-performing loans and bad debts get written-off. Therefore, the bank needs help in predicting and preventing credit card default to improve their Bottom line.

Methodology

Tools

The tools that we used can be summarized as follows.

1. R and R Studio and related technologies
2. Support Vector Machines
3. Generalized Linear Model
4. Random Forest
5. XGBoost

Feature Engineering

The following features were generated from the dataset

1. Balance = Due Payment - Paid Amount
2. Due Payment as a percentage of Credit Limit
3. Balance as a percentage of Credit Limit

While the following have been removed for they had no effect on the response variable. 1. Gender 2. Marital Status

Modelling Approaches

Workflow

The workflow was the three days went with two days of Exploration and one day of Modelling. The training dataset and testing dataset were first preprocessed and saved. After that the training dataset was split into training and testing again to fit the model and test the results on.

The first approach was to use a Support Vector Machine(SVM) as it was capable of generalizing well on a given problem. The SVM was tuned on a validation set and the best gamma and cost variables were found from the range of $[2^{-2} : 2^{2}]$. The values were computed seperately and the resulting values were used straight away to reduce reproducing time. The F1 score on the testing split was approximately 0.8588

The second approach was to use a Generalized Linear Model(GLM) as it was a simple tool that was able to provide intepretations on the variables. However there was a caveat with GLMs as they tended to have low performance when a higher number of predictors were used. Nevertheless the F1 score on the testing set was approximately 0.8234. The variable significance testing showed that there were only a few variables that had low significance such as the amount paid in july, the amount due in august etc. When rechecking with the visualizations, those variables seemed to be significant towards predicting the response variable.

The third approach was to use a Random Forest Classifier(RF) as it was versatile to handle high dimensional data while avoiding overfitting. The RF was fitted with 500 trees and the variable importances was also taken into consideration. (Howver as the computation increased with the number of variables this was only used for an intiial inspection). The F1 score on the testing set was on par with the SVM at 0.8684

Final Approach

The final approach was to use a stacked model. The predictions on the training and testing set(from the original csv files) were recorded for the three model types specificed above. Subsequently a eXtreme Gradient Boosting (XGB) classifier was fitted on the predictions to get the final output. The XGB model scored a F1 score of 0.9524. From the feature importances of the XGB model it was evident that the predictions from RF model was the most influential on the outcome.

Insights and recommendations

Business insights

Marketing Interventions

- It was seen that when the credit limit increases then there is a higher chance that the credit card will become a default credit card. Therefore, when increasing the credit limit it will be better to check the client's current financial status, their current bank transactions. And it will be better if the customers with a lower credit limit are given a special focus as they tend to default even more. Hence attention should be given to the new customers who have a lower credit limit. Therefore, when handling them it has to be completed in a sensible manner.
- The customers who tend to pay the amount less than the due amount should also be considered. Normally customers pay a certain amount but less than the due in order to avoid the card being cancelled. Therefore it will be better to analyze the ratio of the paid and the due amount frequently of the customers in order to identify the customers behavior.
- These factors should not be taken into consideration strictly when judging whether a person will have a credit card default.
    - Gender
    - Marital Status
    - The amount paid with respect to their credit limit
    - Education Status
- It would be a good measure to check if the due amount of a client exceeds 25% of his credit limit as a measure of ensuring that he doesnt default.
- It would also be wise to check if the balance of a client exceeds 20% of his credit limit as a measure of ensuring that he doesnt default.

Dataset Recommendations

- It would be even more helpful to consider the time since getting the credit card as well when predicting defaulting behaviour. In addition details such as whether the client is employed and if so what is his salary? what might be the assets that he owns etc. can be added to investigate further.

- And also we can block the credit cards which exceed the credit limits continuously at least for two months continuously or frequently with just paying the due. Because there is a high chance that the customer is not going to pay the full amount but just the due in order to take the advantage of the credit since the card won't be default if a due is paid.
- In addition it is expected that clients who get credit card default have missed their payments for a long time. But in reality it is the clients who have missed their payments recently are the ones who are getting credit card default. This could be for the reason that it is hard for them to keep up with the debt and hence they are letting the card be defaulted. It is also evident in the instance where clients who have paid in advance are also defaulting their credit cards as it is hard for them to keep up with the debt.