**Course Name:** ALY 6040 Data Mining Applications

Logo

Description automatically generated

**Instructor’s Name:** Prof. Dr. Justin Grosz

**Submitted by:**

Nalini Macharla

Damilola Odunuga

Pranay Joshi

**Business Problem**

In the last two decades, to rise the global economy towards incredible development, credit card system has widely been in usage. Nevertheless, issuing of credit cards to bad customers/non-payers can be considered as a high risk as it can cause financial crisis. To avoid this situation in our project, with the help of data mining techniques we will closely observe the credit card applications and help the financial institutions in deciding whether to approve the application or reject the application. With the help of the information that was provided by the applicants we can analyze and interpret if there are any defaulters amongst the credit card applicants. The most important factor that needs to be considered is the credit score of the applicant while making these interpretations. The risk factor can be calculated accurately with this credit score factor. Based on the credit card applicant’s previous data we can build a model, which will offer us regulators which will predict if the applicant is a good or a possible defaulter.

**Data Quality:**

For the project, we will combine two datasets. The first one is the application record dataset, there are 18 columns and 438557 observations. Another one is the credit record dataset having 3 columns and 1,048,575 observations. While exploring the dataset it is to be found that the dataset has

134,203 null values and they are present in the Occupation\_type column. These null values account a total of 31% rows in the entire column. We used the describe function to check the statistical values which are mean, median, mode and the quartile values.

**Are there any Outliers and Suspicious data?**

With the help of describe function and by using the boxplot we could spot that there are outliers in the taken dataset.

Table

Description automatically generated Chart

Description automatically generated

With the help of the boxplots above we discovered abnormalities in the dataset which are further considered to be doubtful. What caught our interest is the behavior of one applicant who supposedly has twenty children, however, after further analysis their marital status is recorded as single or unmarried in the application and their occupation was a bartender. Looking closely at this information, we can raise a concern whether the credit card applicant has provided correct information or not. The most noticeable outliers that were exposed in the earning of the applicants are those who have higher income values which was 6.75 million dollars and these applicants worked as Labors. This income earned is the highest income.

**Data Cleaning:**

In our analysis, assessing the credit card applicant application with utmost accuracy is important and this must be done without any unfairness. To reach this expectation we have cleaned the 134,203 null values which has 31% rows occupancy in the occupation\_type columns, because these null values represent missing information, therefore, we replace the null values in the occupation type with unemployed which means that 31% of the applicant are unemployed.

**Exploratory Data Analysis:**

The initial EDA we did was to evaluate the gender proportions in the dataset, and it is to be found that there are 67% Female applicants and 32% Male applicants. Other important factor that we were interested in exploring was the Housing Type. The factors that we considered were, if the applicant lives in a rented apartment, does not own an apartment. These types of applicants will be spending additional living expenses which can affect their credit card bill payments. What is surprising to us is that almost a 90% of the credit card applicants have their own living place.

Below is the graph supporting the above analysis:

Chart

Description automatically generated

Below is the graph supporting above analysis with respect to Gender proportions:

A picture containing table

Description automatically generated

Another interesting finding from our analysis was the correlation between the education type and the annual income. Applicants with higher education has higher annual income whereas, applicants with just lower secondary level education have lower annual income.

Below bar graph represents the annual income of the applicants with respect to their educational status:

Chart, bar chart

Description automatically generated

Below is the pie chart to show the credit card payment status of the above applicants:

Chart, pie chart

Description automatically generated

Based on the gathered information, the defaulters were detected by their previous payment patterns with the help of their credit history.

**ANALYSIS**

To decide the impact of the variables which will help the bank to decide whether to approve the credit card application or not, we have used the statistical method which will predict the binary class and for this the dependent variable we considered is ‘Target’. First, we started with converting all the categorical variables, binary variables into dummy ones with a total of 29 independent variables. Further we split the data into two parts, 70% as training data and 30% as testing data. We will display the results and see if the model is suitable or not with the help of accuracy and confusion matrix. To do this analysis and to acquire the desired results, we have used logistic regression to predict the likelihood of variables which could impact the decision of banks when approving credit application, decision tree for testing our model.

Also, eventually generated the new result of our model which were provided above, we recognize this as the new insight provided from the previous model to the new model. Additionally, we define oversampling and normalize the dependent variables. We also transform our dependent variable target to avoid oversampling of data, which was part of the causes of overfitting.

**Model result and Interpretation**

First, we run the logistic regression using the train set and predict the model using the test set which generates model intercept, coefficient of each independent variables, model accuracy, precision, and recall. Our logistic model generated approximately 67% accuracy score in predicting the impacts of the variables on credit card application. Our model precision score is generated which indicates that the ratio of correctly predicted positive observations to the total positive observations, that is the ratio of correctly predicted positive credit application is 67% of the total positive observation of the application. Since our model accuracy seems high, we will only look at the accuracy and precision score (classification report for logistic regression found in table 1 of Appendix).

Additionally, we generated coefficients for our model to determine which variables will help the banks make right decision to approve or decline an application. According to our model, the positive model coefficient of “employment years” of 35% for an applicant increases as the number of years employed increases, that is an older applicant has a higher chance of getting approved than a younger applicant based on work experience. This also indicates that 1-point increase in age correspond to 35% increase in status of applicant being approved. Furthermore, months of membership and years of employment of an applicant also contributes to the probability of been approved. Therefore, we assume that the bank should look out for applicants’ years of membership, and years of employment and others for approving of credit card application.

Next, the train set model was used to predict the test set of target variable after which we generate the confusion matrix of our logistic regression model to the count of observations that are predicted correctly. Our model generated true positives of 7979 and true negative of 0 observations, with false negative of 0 and false negative of 4022 (confusion matrix for logistic regression found in fig 2 of Appendix). This gives an overall view of how well our classification model is performing by showing the kinds of error its making. Logically, months of membership and years of employment alone is not enough for the bank to approve an application, therefore we run decision tree classifier to compare the result with the logistic regression model.

After running the decision tree classifier model using the split train and test set, our model generated an accuracy score of approximately 70% which seems similar to the accuracy result of our logistic regression. However, the decision tree model precision and recall score of our model is slightly different as it generates a precision score for 1 as 0.65 and for 0 is 0.70 (classification report of decision tree found in table 3 of Appendix). This indicates that our model is about 70% accurate in predicting the variables needed for the approval or rejection of credit card application by the bank.

We generated a confusion matrix for our decision tree model, which generated true positive of 7645 and true negative of 383 with false positive of 3257 and false negative of 716, since this shows the count of correctly predicted observation, and we can see that the true positives is higher that the false negatives, we will assume that the decision tree model accurately predicted the variables of importance for the bank (confusion matrix for decision tree model found in fig 3 in Appendix). The decision tree classifies all features of a dataset and predict features by importance, therefore, we proceeded to plot the feature importance generated by our model (feature importance plot found in fig 4 of Appendix).

According to our decision tree model, the feature importance plot shows age has the most important, followed by employment years, months of membership with the bank, and total income amount. The feature importance plot shows how much each features contribute; therefore, we assume that the bank should consider these features when approving or declining an application as these could contribute to the probability of a customer being non-payers.

After implementing and seeing the results from decision tree, we wanted to check whether the model can be more efficient as the accuracy achieved was 70%. Although we could have stopped here too, but just to make it more efficient, we went on testing with three more different models which boosted the accuracy like the gradient boosting which increased the accuracy and made it 69% and XGBoost it increased more with 68% accuracy so decided to stop here and develop a conclusion. And from the interpretation from these models, we also had roc\_auc\_score and f1 score which was 0.53 and 0.17.

When AUC=0.5, then the classifier is not able to distinguish between Positive and Negative class points. Meaning either the classifier is predicting random class or constant class for all the data points. So, one perspective of the interpretation from the AUC score which is 0.53 along with classification accuracy and low f1 score it may be interpreted that the classification training dataset has done a good job in ranking the test data. Our model originally had very high accuracy for both logistic regression and decision tree which is because of overfitting the model. The overfitting was caused by multicollinearity of three independent variables which are age, count of children and count of family members as shown by correlation matrix (Fig 1 of Appendix). So, the higher the AUC value for a classifier, the better its ability to distinguish between positive and negative classes.

Therefore, to eradicate the overfitting of model from our first report, we run our model again and we dropped some of the variables which caused overfitting for the model.

**CONCLUSION AND RECOMMENDATION**

Lending is the main way that banks meet the credit needs of customers. An inherent risk in a loan is that certain debts will not be repaid, these debts are eventually written off. The majority of debt that banks write off and sell to debt buyers is credit card debt. The goal of this analysis is to help banks determine which applications to approve and reject by analyzing applicants’ historical records. Through our model, we were able to determine the variables are not only significant but will help banks determine who to approve. As shown by our model, customer for months, and years of employment of an applicant should be thoroughly considered by the bank, as steady flow of income before applying for credit card.

Although been unemployed does not necessarily mean a customer will be delinquent in payment, but steady income through employment is a granted method for banks to make decision on approval. Furthermore, months of been a customer shows a serve as a strong indication of a loyal customer, as shown by our model, this means that applicant with numerous months of membership has stronger chances of being approved for another credit card by banks.

Concluding for model prediction which were mentioned in interpretation about the imbalanced dataset and overfitting of the model which we learn to handle by investigating the cause and running the model to generate a new model result, the problem of classification when there is an unequal distribution of classes in the training dataset. The imbalance in the class distribution may vary, but a severe imbalance is more challenging to model and may require specialized techniques.

The first thing which comes to my mind is the model is training the data of minority class based on the majority class, so for that we need to use resample technique and generate synthetic samples and if needed change the algorithm which is been used, if given more time we could have come up with the best optimized result for that. Moreover, if the model is overfitting, then probably, we can cross validate the dataset by training with more balanced data and after using algorithms and regularization methods and then coming up with ensembling modeling techniques.

We recommend that the banks should first consider the repayment history of older applicants who also have the most credible number of years of employment, since this will determine the income level, hence the number of delinquencies. Therefore, the banks can reduce the number of debts been written off. However, age and employment history are not enough.

For further analysis, we recommend that banks get more information on an applicant such as number of delinquencies, credit card utilization rate and hard inquiries of each applicant’s credit history before approving any applications. Also, banks could offer variety of payment options which could encourage repayment of credit card debts.

REFERENCES

OCC. (2014, August 4). *Consumer debt sales: Risk management guidance | OCC*. Office of the Comptroller of the Currency (OCC). <https://www.occ.treas.gov/news-issuances/bulletins/2014/bulletin-2014-37.html>

APPENDIX

Table 1: *Classification report of Logistic regression*

Table

Description automatically generated

Table 3: *Classification report of Decision Tree*

Table

Description automatically generated

Fig 1: Correlation Matrix

Chart

Description automatically generated

Fig 2: *Confusion matrix of Logistic regression*

Chart, treemap chart

Description automatically generated

Fig 3: *Feature Importance of Decision Tree*

Chart, bar chart

Description automatically generated

Fig 4: *Decision Tree*

Chart, line chart

Description automatically generated

Fig 4.1: *Confusion Matrix: Decision Tree*

Chart

Description automatically generated

Fig 5: *Confusion Matrix: Gradient Boost*

Chart, treemap chart

Description automatically generated

Fig 7: *Confusion Matrix: XGBoost*

Chart

Description automatically generated