**Module 4 Final Project — Milestone 2: Draft Report**

**Credit Card Approval Prediction using Data Mining Techniques**

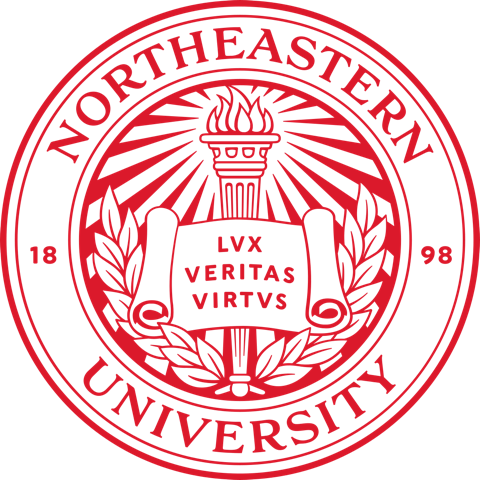
Akshay Kirit Kalavadia

Sunil Raj Thota

Jyothi Chandrakanth

Li-Ching Hwang

Nissy Asish Gantepalli



College of Professional Studies, Northeastern University ALY 6040 – Data Mining Application

Prof. Justin Grosz

October 24, 2021

**Understanding Business Problem**

The credit card system has been extensively employed as a device to boost the global economy toward remarkable development during the last two decades. However, offering credit cards to high-risk users might result in financial crises, which can lead to events similar to 2008. As a result, in this project, we will utilize data mining applications to examine a credit card application in order to assist credit lending organizations in deciding whether or not to approve or reject it. We'd utilize the information provided by applicants to forecast the possibility of credit card payment default. The applicant's credit score is one of the most crucial elements we evaluate while making this choice. The degree of risk may be precisely measured using a credit score. We can build a model based on the applicants' previous data that will offer regulators with a forecast of whether a client is a "good" or "risky" consumer.

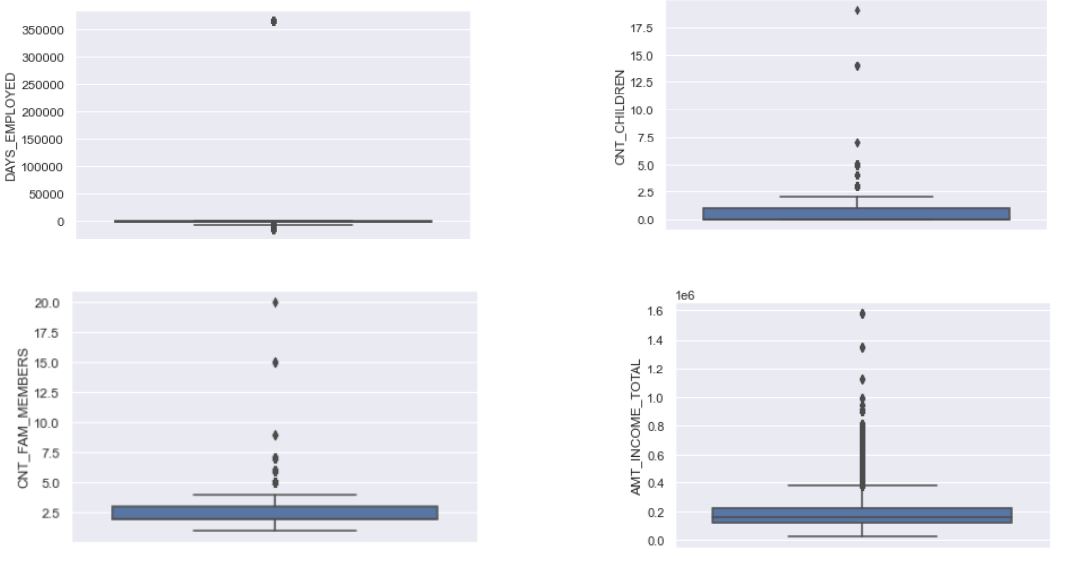
**Data Quality:**

The application dataset contains 438,557 records and 18 columns while credit record dataset contains 1,048,575 records and 3 columns. During the exploration of the dataset, we found that there were 134,203 null values in the *Occupation\_type* column which accounts for 31% of the total rows in that column. To check statistics of the dataset, i.e., mean, median, mode and quartile values, we check the describe function.

**Outliers and Suspicious data:**

In this project, after using describe function and boxplot, we were able to identify few outliers in the dataset as shown in the graph below:

**Figure 1: Boxplot of outliers in the dataset**

****

The anomalies that were discovered in the dataset and were deemed suspicious are provided by means of the charts above. It was one of the candidates' 20 children that caught our interest. Upon additional investigation, we learned that the candidate is "Single/Not married" and worked as a waitress/bartender, raising the issue of whether the data supplied is valid or inaccurate on the application. The most prominent outliers revealed in the applicants' earnings are the applicants with the highest income earned at 6.75 million dollars who worked as Labor, which would be the highest income earned.

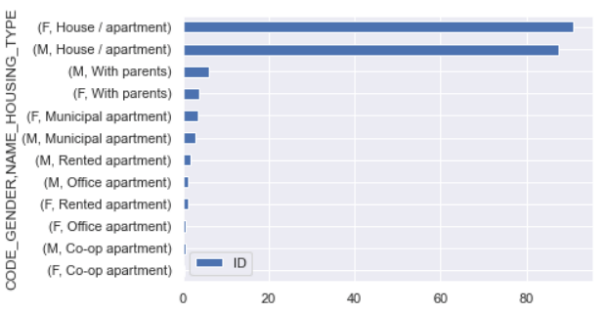
**Data Cleaning:**

In this project, accuracy is of utmost importance to assess an applicant’s profile without bias. Hence, we have removed all the records which are not complete or have missing values, that is, we have removed the rows where *occupation\_type* is null.

**Exploratory Data Analysis:**

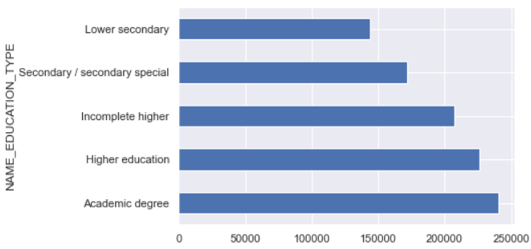
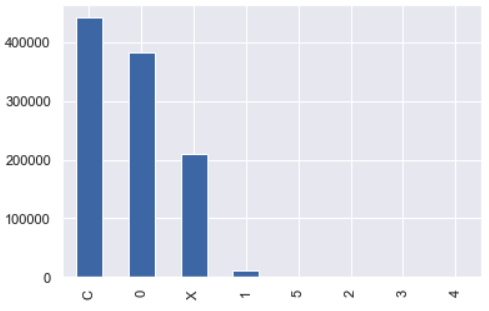
During the Exploratory Data Analysis process, we learned that the vast majority of applications were female, accounting for 67% of all applications, with just 33% of applicants being male. The kind of *housing* was one of the criteria that we were particularly interested in investigating. If a person lives in a rented apartment or does not own his or her own apartment, he or she will incur extra living expenses, which will influence their ability to make the credit card payment. We found that 90% of the applicants had their own home/apartment, which was surprising to us.

**Figure 2: Housing type based on gender**



Another intriguing finding, was that there is a significant correlation between *educational attainment* and the *annual income*. In terms of *annual income*, applicants with an *academic degree* seemed to have the highest annual income, while applicants with a lower secondary education level earned the lowest amount.

**Figure 3: Annual Income based on Level of Education and credit payment status**

** **

The most notable factor we discovered from the credit history data was, that the applicants' prior defaulting patterns were revealed via their credit history. As we can see from the graph, 42% of them made their payments before the due day, and 36% of them made their payments within a month past the due date.

**Analysis**

In order to determine the influence of the variables that lead to whether the bank approves the credit card application, the statistical method of predicting the binary class is used, with *target* as the dependent variable. We have converted all categorical variables, binary variables into dummy variables with a total of 29 independent variables. Here, the data set is divided into two parts at a ratio of 70:30. This means that 70% of the data will be used for model training and 30% will be used for model testing. The results would use an accuracy and confusion matrix to show whether this model is suitable. Next, we used logistic regression, decision trees, and random forest to test our model.

Since the dependent variable is discrete, logistic regression is an appropriate model for this situation. To determine if the independent variable has an influence on the target variable, we utilize regression to make predictions. According to the results of the logistic regression model's confusion matrix (Figure 5 in the appendix), the accuracy score is 0.61215, and the outcome we obtain in this model is not satisfactory. As a solution, we used decision trees and random forests to forecast our model.

In the decision tree model, we employ the same dependent variable and independent variable as in the previous model. The accuracy of the results obtained is increased to 0.82897. For example, in the confusion matrix (Figure 6 in the appendix), a true positive score of 0.7 and a true negative score of 0.96 were observed. As a result, the accuracy of a true negative forecast is very good.

Thereafter, we utilized the random forest (Figure 7 in the appendix) to build a prediction model. According to the results, the accuracy of the score is 0.89459.  The real positive sore is 0.88, which means that it performed best in the decision tree than the false positive score. Despite the fact that the true negative is 0.90 which is lower than the previous model, the accuracy obtained is high.

Next, we will continue to explore which of the 29 independent variables have the most influence on our dependent variables. Therefore, in the future, what factors should the bank pay more attention to when reviewing the process of customer application for credit cards, so as to reduce the workload of bank counter staff and reduce the risk of the customer being unable to repay the amount to the bank in the future.

**Finding, Exploration and Interest:**

We have trained multiple machine learning models with same data to find out which is the best model to predict the target variable based on accuracy and confusion matrix.

**Table 1: Accuracy of each model.**

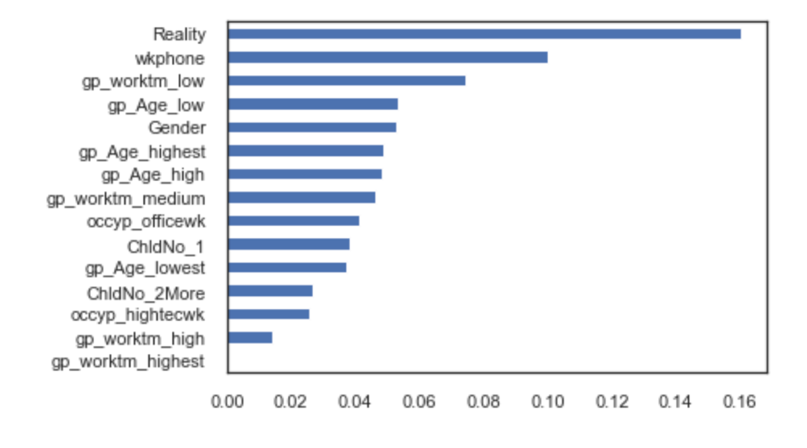
|  |  |
| --- | --- |
| **Model Name** | **Accuracy** |
| Logistic Regression | 64.5%. |
| Decision Tree | 80%. |
| Random Forest | 83.8%. |

From the above table, we can observe that the accuracy of the Logistic regression model is 64.5%. Based on the Logistic Regression model, the features that are having an impact on the model are *Gender, Car, Reality, days\_birth, days\_employed, flag\_mobil, wkphone,* and *phone.* By looking at the feature importance you can decide which features to possibly drop because they don’t contribute enough to the prediction process.

The accuracy of the Decision tree model is 80%. Based on the results of this model, the features that are having an impact on the model are *Reality, wkphone, Gender,* and *gp\_Age\_low*.

The quality of the random forest algorithm is that it is very easy to measure the relative importance of each feature on the prediction. The accuracy of the Random Forest model is 83.8%.

**Figure 4: Feature Selection using Random Forest Model**



From these three models, we can find that random forest has the highest accuracy. So, we find the most influential variables in the random forest model. Among these 29 variables, we use a bar graph to show which are top 15 variables that have the greatest impact on the target variable. The first one that has the greatest impact is whether the customer has reality (Own a House/Apartment), have they registered work phone, Duration of work (Between 10 to 20 Years) and Age (Between 30 to 40 Years). These four factors have the most impact on the entire model. Therefore, in the future, banks can review these conditions to determine whether bank can approve the credit card applications of customers.

**Conclusion**

From the model testing done above we can conclude that Random Forest model has the highest accuracy. While evaluating a credit card applicant, accuracy is very important for the judgement. We can also conclude that the most important features for consideration by banks are *house\_type*, *flag\_work\_phone, days\_employed,* and *days\_birth* column from the dataset.

**Wrapping up and Next Solution**

The Random Forest model that we developed can assist banks in the resolution of bad debts. According to our model, financial institutions should consider an applicant's type of housing when determining his or her capacity to repay the outstanding amount.  If a person is renting an apartment or a home, his or her finances would be impacted, and his or her ability to make credit card payments would be reduced. Registration of a business phone is yet another vital consideration to take into consideration. If the applicant's work phone number is recorded with the banks, the banks will be able to immediately reach the card user in the case of a default. If the customer's phone number is not in the system, it will be difficult for them to find them.

This model may also be used by applicants to examine their profile prior to submitting an application for a credit card. If their profile is not strong enough, their application may be disapproved. This will have a negative impact on their total credit score. In order to prevent this, they might leverage this model to improve their profile before submitting their application.

**References**

Feature selection in python sklearn. DataCamp Community. (n.d.). Retrieved October 25, 2021, from <https://www.datacamp.com/community/tutorials/feature-selection-python>.

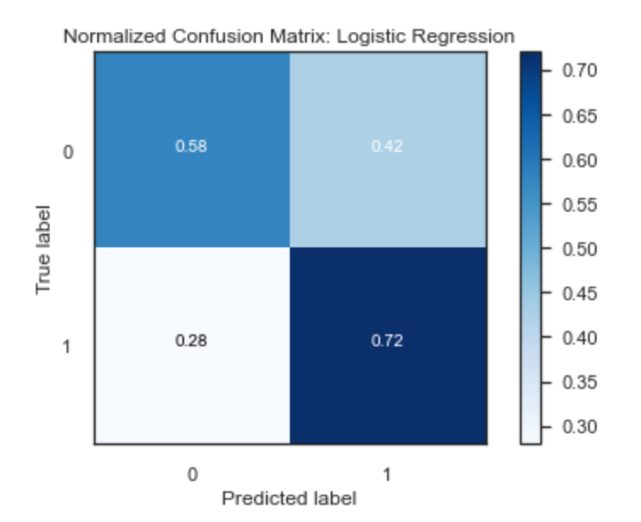
Luthi,B. Karp, G. (2021, March 22). *How to apply for a credit card so you'll get approved*. NerdWallet. Retrieved October 25, 2021, from <https://www.nerdwallet.com/article/credit-cards/apply-for-a-credit-card>.

Sklearn Random Forest classifiers in Python. DataCamp Community. (n.d.). Retrieved October 25, 2021, from <https://www.datacamp.com/community/tutorials/random-forests-classifier-python>.

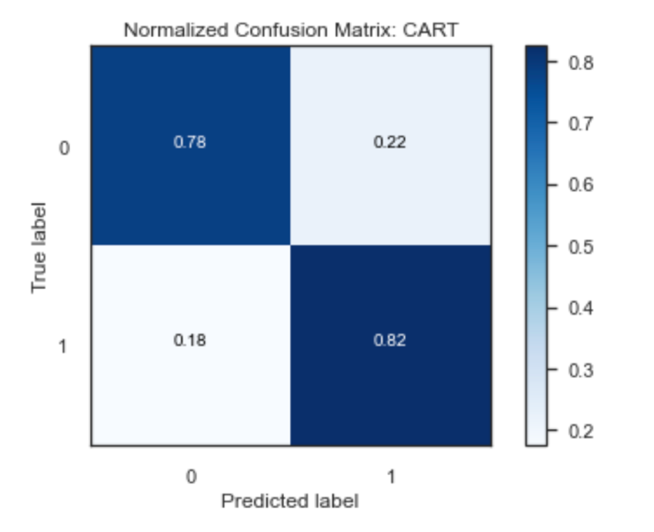
*The credit approval process & what to expect | credit.com*. (n.d.). Retrieved October 25, 2021, from <https://www.credit.com/personal-finance/credit-approval-process-what-to-expect/>.

**Appendix**

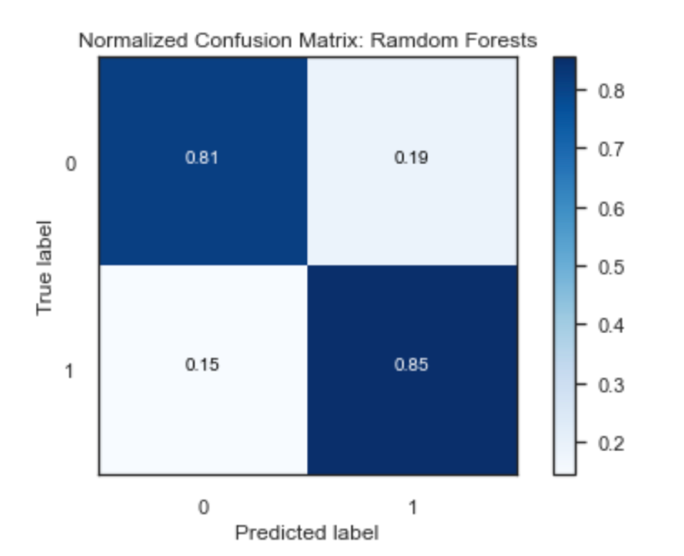
**Figure 5: Logistic regression confusion matrix**

****

**Figure 6: Deccision Tree confusion matrix**

****

**Figure 7: Random forest confusion matrix**

****