# DSC630-T301 Predictive Analytic

# Milestone 5: Final Paper

Credit Card Approval Prediction



**DSC 630**

**Abhjit Mandal**

**Saurabh Shrestha**

**Ramizuddin Mohammed Shabuddin**

* **Executive Summary**

The purpose of this project is to predict credit card approval by analyzing the impact of different fields on the approval for a credit card. Being able to predict what type of customer are likely to default on their payment comes in handy when dealing with approval of applicants of the credit card to avoid future risk. According to a data from the Federal Reserve Bank of New York, in the first quarter of 2021, the 90-day failure to make outstanding payment was 9.98% compare with 9.09% at the same time in 2020.

The dataset for credit card approval for prediction was taken from Kaggle. Approval of application for the credit card depends on number of factors such as their credit payment status, their total income earned, age, year of employment and so on. Using all the required features and through the use of predictive analytics and machine learning, patterns and relationship among the variables were analyzed. To create a model, we used logistic regression, decision tree, random forest, support vector classifier, KNN and XGBoost.

* **Abstract**

The growth of the internet has led to a significant rise in credit card usage. It is one of the most used payment methods these days. As the world economy increases, credit card usage also increasing at an alarming rate. It is also evident that credit card defaulters have also increased significantly. Using machine learning algorithms like logistic regression, XGBoost and other algorithms, we are predicting if the customer to be approved with credit card or not.

* **Introduction Background of problem**

Credit card issuing institutions are becoming meticulous in approving credit cards for customers. In addition, the downturn of financial institutions during the US subprime mortgage and the European sovereign crisis has raised concerns about risk management properly. Hence, these challenges have attracted significant attention from researchers and practitioners. A wide range of statistical and machine learning techniques have been developed to solve credit card related problems. It is found that machine learning techniques are superior to other traditional statistical techniques in dealing with credit scoring.

The decision of approving a credit card or loan is majorly dependent on the personal and financial background of the applicant. There are two basic risks: one is a business loss that results from not approving the good candidate, and the other is the financial loss that results from approving the candidate who is at bad risk. It is very important to manage credit risk and handle challenges efficiently for credit decisions as it can have adverse effects on credit management. Therefore, evaluation of credit approval is significant before jumping to any granting decision.

* **Methods**

The dataset we retrieved from Kaggle is of credit card applications. There are two CSV files; one is application record, and another is a credit record. Application record dataset contains applicant’s personal information whereas credit record contains months’ balance and status of their payment. Factors like age, gender, income, education, Years employed, credit history, and other attributes all carry weight in the approval decision. Credit analysis involves the measure to investigate the probability of a third party to pay back the loan to the bank on time and predict its default characteristic.

With the support of the data, the plan is to analyze and apply necessary data preparation techniques which could be useful for creating an effective model. We followed the process as follows:

1. Load the dataset.
2. Converted column names for readability.
3. Dropped unnecessary columns.
4. Handled missing and null values.
5. Converted non-numeric to numeric value. Such as Education column had Lower secondary, Secondary / secondary special, Incomplete higher, Higher education and Academic degree as unique value which were converted to numerical value 0, 1, 2, 3, 4, 5 respectively.
6. Exploratory Data analysis – We created multiple graphs and analyzed the dataset. Such as for marital status, we had higher applicants who were married vs lowest applicants from widowed. For housing type; highest applicants came from housing/apartment vs lowest with co-op apartment housing type, for gender: we had higher female applicant vs Male
7. Train and test the machine learning model – We used sklearn's model selection for splitting dataset into train and test set.

We are training a supervised learning model where status is our target variable. Before going through the modeling process, we dropped columns that were irrelevant for our models such as numerical columns (has: cell phone, phone, work phone, months balance) and categorical columns such as Income category, Marital status, Housing type, Occupation. These columns do not have any correlation and are deemed unnecessary.

* **Results**

**EDA Analysis**: Here are examples of few graphs we created for EDA. The Status shows the binary values of either 1 or 0. 0 indicates that the applicant has paid their credit due on time or has no loan remaining. Whereas 1 indicates that they are behind on their payments.

Chart, scatter chart

Description automatically generated

The above graph shows that the applicants are not good candidates if Total income & years of Employment is less. **Chart, scatter chart

Description automatically generated**

The above graph shows that, majority of applicants who have higher income are more likely to pay their due on time. There is no correlation with age with their payments. We also analyzed the applicant’s distribution data, here are some results that we found:

**Chart, bar chart

Description automatically generated**A picture containing shape

Description automatically generated

Majority of applicant’s are married Majority of applicant's lives in House / Apartment

**Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated**

Majority of applicant's are 25- 65 years old Majority of applicant's are Employed for 0 -7 years

**Correlation**: Below we have the seaborn correlation heatmap which shows that the features are not highly correlated to each other. In addition to that, the features are also evenly split between positive and negative correlation between two variables. This graph also shows that there is no column (Feature) which is highly co-related with 'Status'

**A picture containing Teams

Description automatically generated**

**Data Modelling**:

To prepare and apply a model to this dataset, we split the dataset into two subsets. The first is the training set on which we developed the model. The second is the test dataset which we used to test the accuracy of our model. We allocated 80% of the items to Training and 20% items to the Test set.

For modeling we used Logistic regression, Decision tree, Random Forest, SVM, KNN and XGBoost models. Below is table which shows the accuracy based on used machine learning algorithm used.

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| Logistic Model Accuracy | 50.60 % |
| Decision Tree Model Accuracy | 69.55 % |
| Random Forest Model Accuracy | 76.00 % |
| Support Vector Classifier Accuracy | 49.79 % |
| KNN Model Accuracy | 45.98 % |
| XGBoost Model Accuracy | 84.14 % |

* **Discussion/conclusion**

1. As the dataset is highly imbalanced, we have used SMOTE (Synthetic Minority Oversampling Technique) technique to understand which model performs better.
2. We took 2 passes at the Machine learning models, one with initial data and other with balanced data after performing SMOTE technique and the two results differed significantly.

After applying all the Machine learning models on the balanced dataset, we got that XGBoost Model is giving the highest accuracy of 84.14 %. SMOTE Sampling methods provided much better results compared to raw data.

We will be refining our models and algorithms further in the coming weeks and if results remain the same then we will be choosing XGBoost as the preferred algorithm for any future credit card approval prediction.

For future work, the efficiency of the models can be improved if the dataset is larger, and balanced so, that the sampling method is not needed. If the original values of the dataset are known, then we can know how the data is correlated and which features are important and train accordingly. In the future different methods can be used to improve the results, more parameter tuning can be done.

* **Acknowledgments**

We would like to thank Kaggle from where we were able to retrieve the data. Additionally, we would like to thank our Professor, Fadi Alsaleem from Bellevue University who is teaching us the Predictive Analytics course as well as our classmates from DSC 630 for their contribution to the discussion on the subject matter. We would like to thank the authors; David Abbott and Eric Siegel for their book Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst and Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die respectively. With the help of the book, we were able to get a road map on how to work our project milestones. Finally, other gratitude goes to websites from where we could find help such as geekforgeeks, stackoverflow, python.org and countless others.

* **References**
* Abbott, D. (2014). *Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst* (1st ed.). Wiley.
* Siegel, E. (2016). *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die* (Revised and Updated ed.). Wiley.
* Li, S. (2019, February 27). *Building A Logistic Regression in Python, Step by Step*. Medium. <https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8>
* GeeksforGeeks. (n.d.). *Python Programming Language*. Retrieved October 2, 2021, from <https://www.geeksforgeeks.org/python-programming-language/>
* Credit Card Approval Prediction. (2020, March 24). Kaggle. <https://www.kaggle.com/rikdifos/credit-card-approval-prediction>
* *Change column type in pandas*. (2013, April 8). Stack Overflow. <https://stackoverflow.com/questions/15891038/change-column-type-in-pandas>
* *Hongmei, ChenaYaoxin, Xianga,”The Study of Credit Scoring Model Based on Group Lasso”, Elseiver, Volume 122, 2017, Pages 677-684.*
* Ashlesha Vaidya, “Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval”, ICCNT(2017), 10.1109/ICCCNT.2017.8203946