Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

NAMES	E-MAIL	CONTRIBUTION
Charan	charanbzact9865@gmail.com	All project

Please paste the GitHub Repo link.

GitHub link:- https://github.com/Charanpoojary/CREDIT-CARD-DEFAULT-PREDICTION-CLASSIFICATION-

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

This Project summary is entitled as "**Credit Card Default Prediction".** The objective of this project is to train various supervised learning algorithms to predict the client's behavior in paying off the credit card balance. In classification problems, an imbalanced dataset is also crucial to enhance the performance of the model, so different resampling techniques were also used to balance the dataset. We first investigated the datasets by using exploratory data analysis techniques, including data normalization. We started with the logistic regression model, then compared the results with traditional machine learning-based models. Then K-means SMOTE resampling method on Taiwan client's credit dataset.

In the end, the proposed method has also been deployed on the web to assist the different stakeholders. Therefore, when the financial institution considers issuing the client a credit card, the institution needs to check the payment history of that person because the decision on whether to pay on duly or owe the bill on a specific month usually relates to the previous payment history. For instance, if a person owes numerous bills already, he or she is likely to delay the payment of the current month unless this person gets a windfall so that the total arrears can be paid off. Besides the payment history, it is also imperative to look at the applicants' credit limit of their current credit cards. This is a result of a virtuous circle: people who pay on duly tend to have better credit scores, so the banks prefer to increase these people's credit lines by taking less risk. As a result, if a potential client already has a credit card with a high credit limit line, this person is unlikely to fail to pay the full amount owed in the future. Although the financial institution often collects clients' personal information such as age, educational level, and marital status when people apply for credit cards, this information also affects the default behavior. In other words, the financial institution should equally consider their potential clients who are men or women, obtain bachelor degrees or master degrees, single or married when deciding whether to approve their credit card/loan applications. We tried our best to make a thorough analysis, and there are still a few possible improvements that may require longer-term action. For the boosting models, only the GBDT method was trained, but various variants of boosting techniques may also be utilized in the future. The financial market changes rapidly every day, and people's economic status and performance are affected by the market all the time. So, if more economic indicators are added to the dataset, this will lead to a more generic model.

After exploring, manipulating and experimenting with different models on the credit card default data set we have obtained a maximum accuracy of 82% to determine whether a person defaults on their credit card or not. Ideally, we would have been able to increase this accuracy by trying out various ways of preprocessing the data, utilizing dimensionality reduction, fine-tuning the models' hyperparameters, and applying ensemble learning.

First, we derived new features from the data set. Since this resulted in about 100 features, we explored dimensionality reduction. Using the second form of reduction, LLE, we ran our initial classification models, Logistic Regression and Support Vector Classifier with the original training set and the reduced set. Seeing no improvement with the set produced in LLE we continued using our original training set.

Next, we fine-tuned the better performing model, Logistic Regression, with GridSearchCV as another attempt to improve the model. After using GridSearchCV no significant improvements were seen.

This led us to trying ensemble learning to see if our overall accuracy could be improved by combining various models' predictions. Once again there did not seem to be a significant improvement in both the Bagging Classifier and Voting Classifier when compared to our Logistic Regression model.

In the end, looking at the accuracy scores of each model was not enough information to choose which model performed best when trying to predict whether or not a person would

default on their credit card. Above one can clearly see how close all these accuracy scores fall. In fact, by just seeing the accuracy scores the Voting Classifier seems to perform the best. But if looked at performance via ROC curves the Logistic Regression model seems to perform the best and the Voting Classifier does significantly worse.

For future work, we think it would be interesting to develop more complex models, such as implementing a neural network and seeing if there could be a better performance of an 83% accuracy score since all our previous techniques did not seem to affect the accuracy.