Capstone Project Submission

Name:- Ajit Kumar Toppo
Email:- ajitkumartoppo96@gmail.com

Github Link:- credit_card_default_prediction

Drive link:- (https://drive.google.com/drive/folders/1Lz-rxf1pebVUXu4PyBmc213Q5ngAwobk?usp=sharing)

Credit Card Default Prediction

Context:

With the change in time, more and more people are using credit cards. Not only people's living standards have improved but also people's consumption concept and consumption mode has drastically changed. With a large number of banks selling credit cards, the phenomenon of credit card default emerges one after another. It is very important for banks to effectively identify high-risk credit card default users.

Problem Statement:

The project aims to predict the customers who are more likely to default on their credit card payment. This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

Approach:

First I started with the understanding of the data then imported the required libraries. Checked the size of data and looked for missing values present in the dataset. Performed exploratory data analysis to understand and visualized the distribution of credit card users between male and females, understood their educational qualification and found out which age group are mostly using credit cards for their purchase. Did some feature engineering, found that the dataset is imbalanced and applied SMOTE technique to deal with it. Also there were many categorical features present for which I used One-hot encoding, also used StandardScaler to standardize the dataset. Then used train-test split to split the dataset into 80% and 20% for training and testing purposes respectively. Used various classification algorithms like Logistic Regression, k-nearest neighbors, Gaussian Naive Bayes, Support Vectors Classifier and Random Forest Classifier on the data. Evaluation metrics like classification report and confusion metrics are used to find out the best performing model

Conclusion:

Out of all the models Random forest gives the highest accuracy and seems to be a suitable model for this dataset among all with an accuracy score of 0.84(84%).