Credit Card Fraud Detection Project Report

Concepts or topics that I learnt:

I learnt about Python basics, numpy, pandas. Then I learnt started learning ML, I learnt basics of ML, how to preview data using ML, doing exploratory data analysis on a dataset. Learnt about the different ML models like logistic regression, SVM, Naive Bayes, Decision Trees, etc. I learnt about how these models can be improved. I learnt the ways of dealing with imbalanced data. I also learnt how to use Neural Networks Testing and also made a Deep Neural Network model.

Explanation of Code and Insights:

1) We first mounted the Google Drive. Then imported the numpy, pandas, matplotlib libraries. Also, then imported our dataset.
2) Then we checked whether there is any null value in our dataset so that it can be filled. There was not any null value.
3) Then we checked the datatype of all the columns.
4) Then using a countplot we compared the number of fraudulent versus non-fraudulent transactions. Number of fraudulent transactions were very less as compared to the number of non-fraudulent transactions. Thus, the model can be biased because of imbalanced dataset.
5) Using a scatter plot we observed the correlation of Class with time and amount column. There was not any dependence of Class on Time , so we dropped Time column.
6) Then we plotted frequency distribution of Amount column of our dataset using Histogram. Then we splitted the whole dataset into two separate data frames i.e. Dataframe with Normal Transactions and Dataframe with Fraudulent Transactions and also plotted their frequency distribution of Amount column using Histogram.
7) Then we plotted distribution plots of V1 to V28 columns of original dataframe and also of Dataframe with Normal Transactions and Dataframe with Fraudulent Transactions.
8) Then we separated target and feature variables. V1 to V28 columns were already scaled. So, we scaled Amount Column. Then we eliminated skewness from data.
9) Then we splitted the data into training and testing data, and train the models on the training data.
10) Models we used are Logistic Regression Model, Decision Tree Classifier, Random Forest Classifier, Support Vector Machine, XgBoost Model, DNN and also calculated Accuracy, precision , recall and F1 score.

11) As the data was imbalanced we used the SMOTE sampling technique to deal with it and trained the model on the resample dataset and again calculated Accuracy, precision , recall and F1 score.

Table of models and Metrics

Without SMOTE

Models	Accuracy(%)	Precision(%)	Recall(%)	F1(%)
Logistic Regression Model	99.91924440855307	87	62	73
Decision Tree Classifier	100	100	100	100
Random Forest Classifier	100	100	100	100
Support Vector Machine	99.87711105649381	80	38	51
XgBoost Model	100	100	100	100
DNN	99.96	100	100	100

With SMOTE

Models	Accuracy(%)	Precision(%)	Recall(%)	F1(%)
Logistic Regression Model	99.95786664794073	80	100	89
Decision Tree Classifier	100	100	100	100

Random Forest Classifier	100	100	100	100
Support Vector Machine	100	100	100	100
XgBoost Model	100	100	100	100
DNN	99.99	100	100	100