**CREDIT CARD FRAUD ANALYSIS USING AZURE**

**Team members:**

Devi Sumanth Pasunoori

Tinku Naga Sai Pavan Adapa

**Video link :** https://drive.google.com/file/d/19cxViUaTWrOD8Zxm\_f1dd3OHTTdoU3wR/view?usp=share\_link

**Github link:**<https://github.com/sumanth1904/sdai_project>

**Goals and Objectives:**

* **Motivation:**

Financial fraud cases, like credit card fraud, have increased because of recent increase e-commerce and e-payment systems. As a result, it is critical to put in place mechanisms that can detect credit card fraud. Due to the increase in fraud rates, researchers started exploring different techniques to identify and analyse online fraudulent transactions. One such technique which seems promising is using machine learning models. When using machine learning for credit card fraud detection, features of credit card frauds play an important role, and they must be chosen carefully.

* **Significance:**

As there is lot of fraud occurring on the internet with e-payment systems. There is scope for systems which can detect such activities. In order to detect them, one must learn which transaction is fraudulent based on previous data. Only machine learning can be useful when there is a chance to learn from data. By using machine learning, we can train the system to identify which transaction is most probably fraudulent and isolate them without leading anyone to be a cyber-crime victim. Our aim is to identify the best machine learning model which can classify a credit card transaction as valid or fraudulent.

* **Objectives:**

1. To implement machine learning techniques for fraudulent credit card analysis.
2. To test various algorithms against the same dataset
3. To find out the best model among them based on testing accuracy
4. To tune the model so that it is neither overfitted nor underfitted.

* **Features:**

Features of given project are:

1. It is supervised learning model.
2. Dataset type is .csv
3. It is a classification model
4. This project uses numpy, pandas, scikit learn modules.

**Increment 1:**

**DATASET:**

for this project we chose a dataset from kaggle which contains transactions carried out by credit cards in Europe in September 2013. This dataset contains transactions that are done during two days. Out of 284,807 transactions, 492 transactions turned out to be fraud. Even though one class is less dense than the other, we try to comprehend the patterns in the fraudulent transactions. SMOTE (Synthetic minority oversampling method) is used to increase the number of fraudulent transactions in the dataset. In order to maintain privacy and security the features are renamed v1, v2, v3……v8 from original features. These variables are result of PCA transformation. Apart from these 28 features there are also two more features ‘Amount’ and ‘Time’. These features are not transformed using PCA.

**IMPLEMENTATION:**

In the first phase of our project we need to implement credit card fraud detection using machine learning. Since we are classifying a transaction as fraudulent or not. It is a classification algorithm. We intend to implement any one of the many widely used classification algorithms. After we achieve a desire result we would like to go to the next phase that is implementing different ML algorithms for the same problem.

By solving the problem using different machine earning algorithms, we get an idea of which model is best suited for the given problem. We can interpret using plots.

In the last phase we can implement software development techniques on the given machine learning code by using cloud computing software like azure.

**PROJECT MANAGEMENT:**

**Planning**:

1. Defining the tasks to be done.
2. To check the feasibility of project
3. Dividing the tasks among the team.

**Data** **collection**:

1. Collecting the data
2. Understanding data
3. Cleaning the data
4. Building data ingestion pipelines

**Model selection:**

1. Selecting the best suited model.
2. Refining Model.

**Testing**:

1. Testing models using pytest.

**Deployment**

**Model maintenance**

**IMPLEMENTATION STATUS REPORT:**

**Work completed:**

* **Description:** we collected the dataset. Preprocessing and cleaning the dataset is done. We tried a couple of algorithms which are best suited for our problem.
* **Responsibility:**

Data collection: Tinku Naga Sai Pavan Adapa

Increasing dataset using SMOTE: Devi Sumanth Pasunoori

Training KNN and Decision trees: Tinku Naga Sai Pavan Adapa

Training logistic regression and random forests: Devi Sumanth Pasunoori

* **Contributions:**

Tinku Naga Sai Pavan Adapa 50%

Devi Sumanth Pasunoori 50%

**Work to be completed:**

* **Description:** making an analysis. Testing the software. Building models using azure, testing.
* **Responsibility:**

Making an analysis: Tinku Naga Sai Pavan Adapa

Building model using Azure: Devi Sumanth Pasunoori

Testing the software: Tinku Naga Sai Pavan Adapa

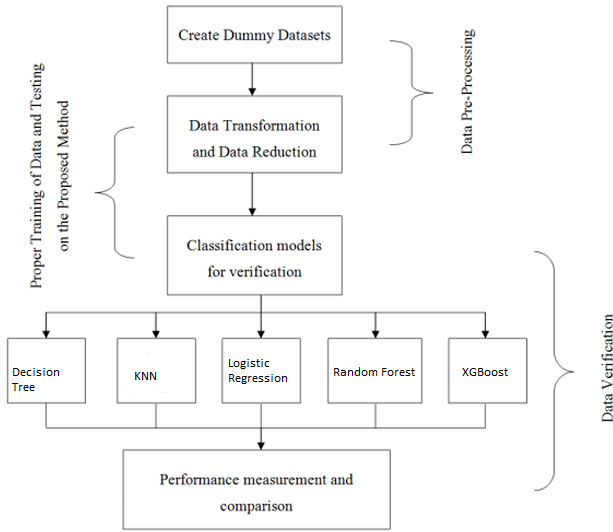
**Increment 2**

**Background:**

Credit card industry is big. Data associated with it is also huge. So, the data can be handled differently than other machine learning models. Data mining and big data techniques looks promising to handle such high volume and constantly incoming data. Based on self-organizing map and neural network, ROC curve is over 95% for fraud cases without any false alarms rate. It is evident from the given dataset that instances of genuine transactions are significantly higher than fraudulent transactions. There are only few fraud transactions in the whole dataset. Few researchers worked with considering this real world scenario taking the data set as it is. Whereas, the rest duplicated the fraud data in order make it 50% of the original dataset, In order to give the model more chance to learn the patterns of fraudulent transactions.

**Model**

**Workflow:**

****

We already discussed about how the dataset is unbalanced. In this project we opted to create dummy data set. We did this to make the quantity of fraudulent data which is less in number to be as much as genuine data. To achieve this this we used a statistical technique called SMOTE using which we could create dummy data without affecting the remaining original data set.

After creating dummy data we need to work on duplicate and redundant data. There may be rows with NA values which must be dropped. Since we have the data which is almost ready to use, there is no need of transforming data.

In this project, we use 5 different machine learning models, to make an analysis on which algorithm is best suited for the given dataset. Each model runs the same dataset and trains the model.

After training, we perform evaluation metrics to know which algorithm is best and make a plot on the results.

**Implementation:**

**Decision tree:**

Decision tree is a classification algorithm, it uses tree like model. Each node consists of a condition upon checking condition it is traversed to branches. Each branch’s probability of occurrence can be calculated easily. Here we used decision tree as it seemed promising with our dataset.

**KNN**:

K-nearest neighbours algorithm saves all the available train data into it’s model and learns from. Whenever there is an incoming data it assigns that data-point to one of its classification. Here, we have two groups, therefore the number of classifications will be 2, which gives us k=2. It uses Euclidean distance to assign incoming data-points to one of the classification.

**Logistic regression:**

Logistic regression is the basic kind of classification algorithm. As the name suggests it uses regression line. This regression line is used and drawn in such a way that it divides the dataset into two classes.

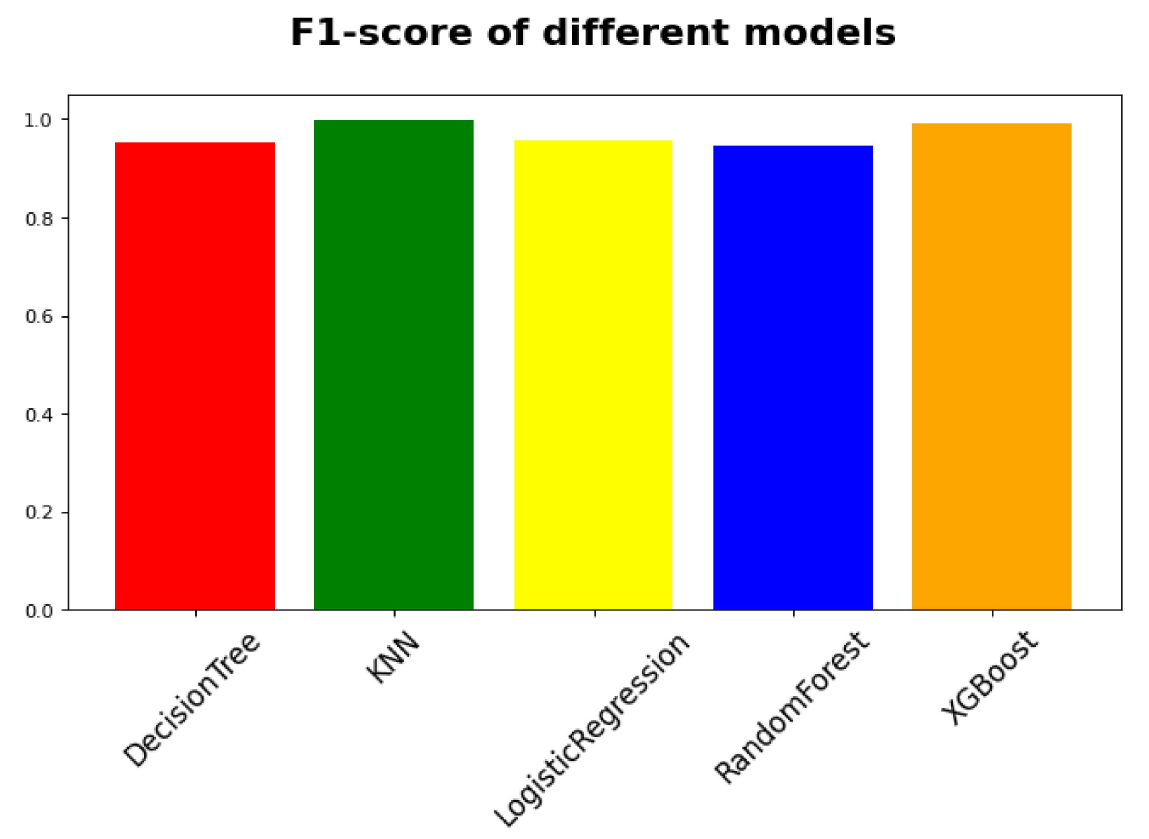
**Random Forest:**

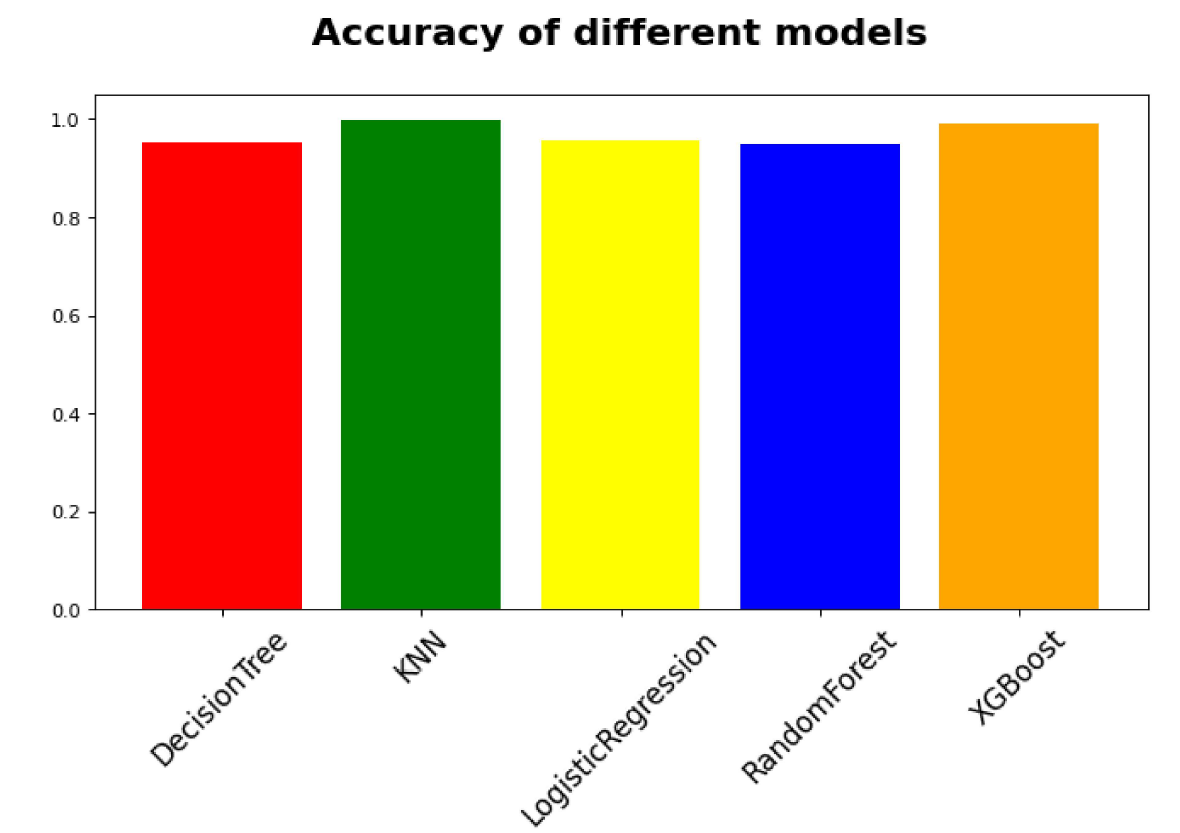
Random forest is a collection of decision trees. It generates decision trees and all of these are connected to one common tree and the average from all these sub-trees are collected at parent tree giving the best result.

**XGBoost:**

XGBoost is a ML library for regression and classification problems. It is decision tree which provides parallel tree boosting. In many ways XGBoost is considered as better model than random forest.

**Results:**





**IMPLEMENTATION STATUS REPORT:**

**Work completed:**

* **Description:** we collected the dataset. Preprocessing and cleaning the dataset is done. We tried a couple of algorithms which are best suited for our problem.
* **Responsibility:**

Data collection: Tinku Naga Sai Pavan Adapa

Increasing dataset using SMOTE: Devi Sumanth Pasunoori

Training KNN and Decision trees: Tinku Naga Sai Pavan Adapa

Training logistic regression and random forests: Devi Sumanth Pasunoori

Making an analysis: Tinku Naga Sai Pavan Adapa

Testing the software: Tinku Naga Sai Pavan Adapa

* **Contributions:**

Tinku Naga Sai Pavan Adapa 50%

Devi Sumanth Pasunoori 50%

**References:**

1. <https://www.kaggle.com/mlg-ulb/creditcardfraud>
2. <https://core.ac.uk/download/pdf/229268289.pdf>
3. <https://xgboost.readthedocs.io/en/stable/tutorials/model.html>