CREDIT CARD FRAUD DETECTION CAPSTONE PROJECT UPGRAD – DATA SCIENCE BOOTCAMP

Sharjeel Ahmed

PROJECT OUTLINE

- A credit card is one of the most used financial products to make online purchases and payments. Though the Credit cards can be a convenient way to manage your finances, they can also be risky. Credit card fraud is the unauthorized use of someone else's credit card or credit card information to make purchases or withdraw cash.
- It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.

PROBLEM STATEMENT

- The dataset contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.
- We have to build a classification model to predict whether a transaction is fraudulent or not.

PROBLEM STEPS

Exploratory Data Analysis: Analyze and understand the data to identify patterns, relationships, and trends in the data by using Descriptive Statistics and Visualizations.

Data Cleaning: This might include standardization, handling the missing values and outliers in the data.

Dealing with Imbalanced data: This data set is highly imbalanced. The data should be balanced using the appropriate methods before moving onto model building.

Feature Engineering: Create new features or transform the existing features for better performance of the ML Models.

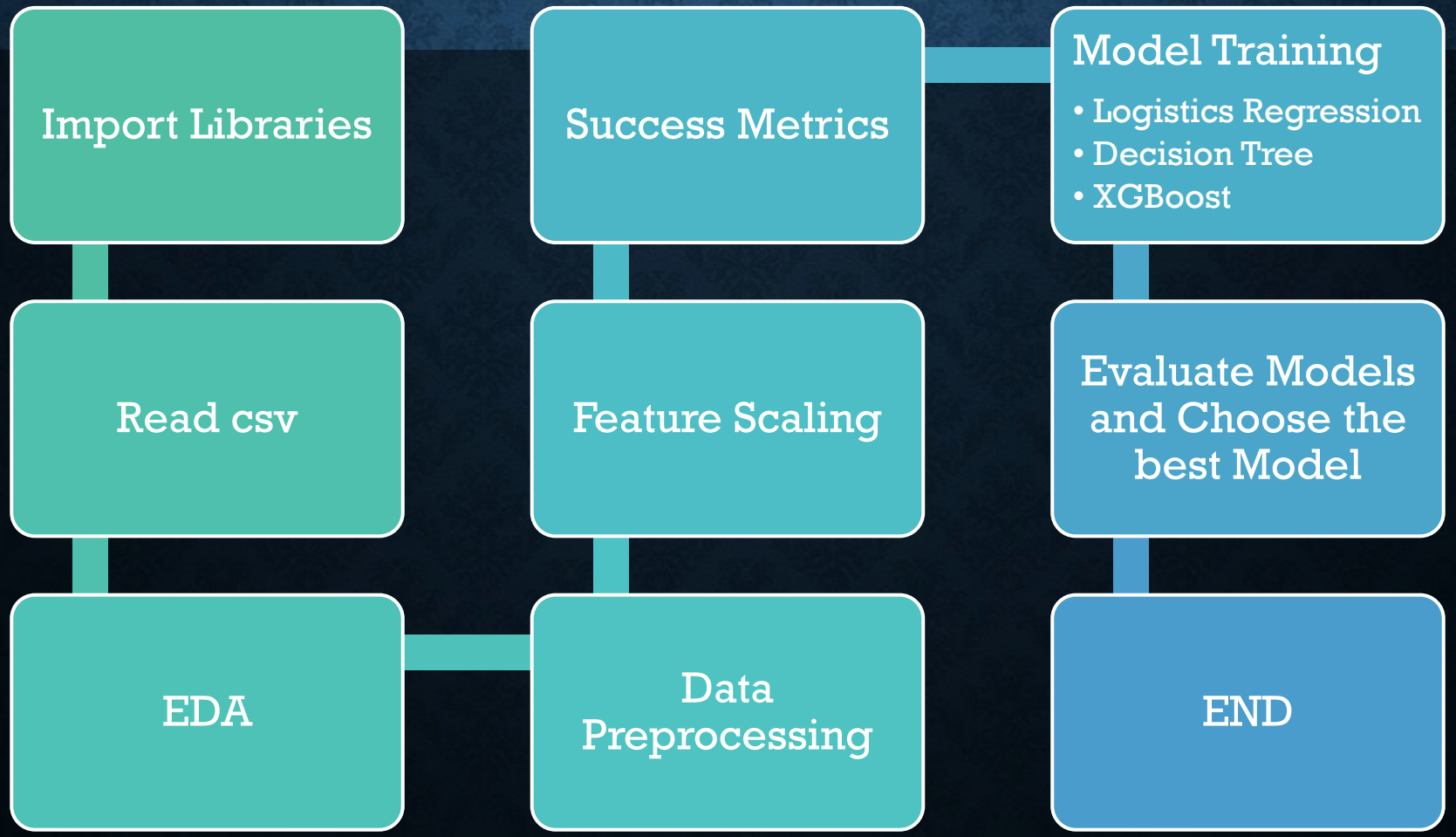
Model Selection: Choose the most appropriate model that can be used for this project.

Model Training: Split the data into train & test sets and use the train set to estimate the best model parameters.

Model Validation: Evaluate the performance of the model on data that was not used during the training process. The goal is to estimate the model's ability to generalize to new, unseen data and to identify any issues with the model, such as overfitting.

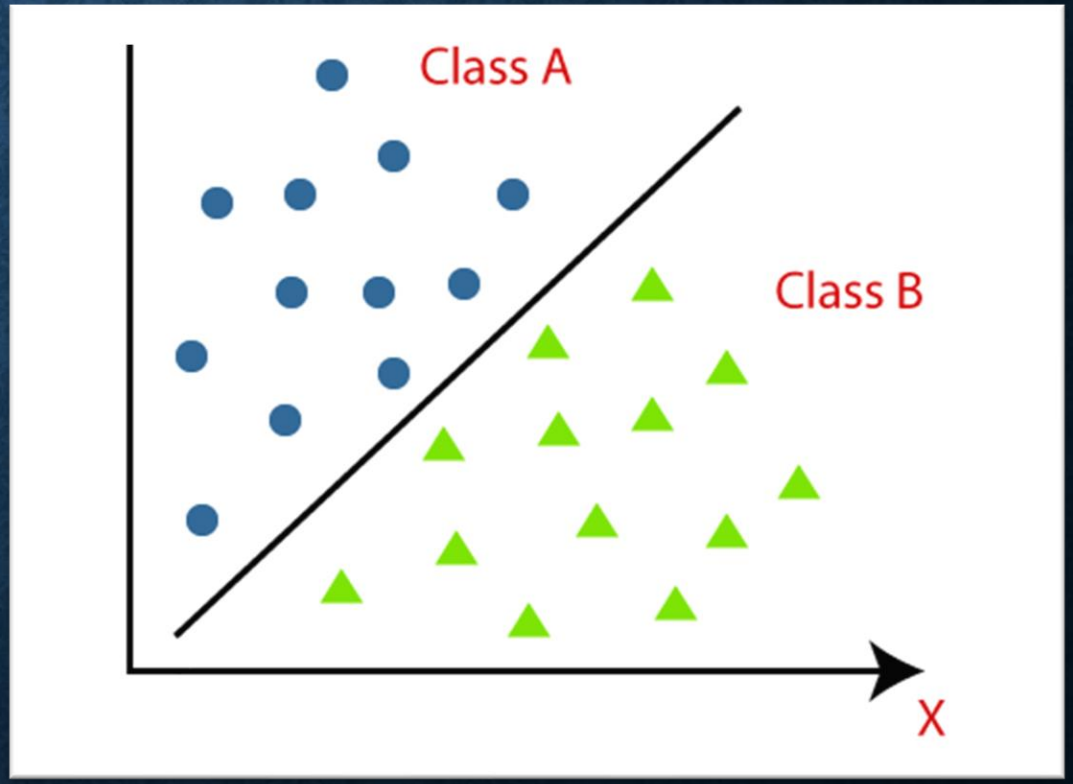
Model Deployment: Model deployment is the process of making a trained machine learning model available for use in a production environment.

PROJECT STEPS



CLASSIFICATION

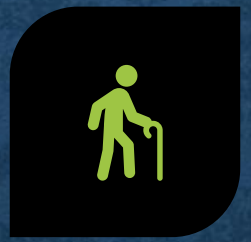
- The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data.
- Binary Classification
- If the classification problem has only two possible outcomes, then it is called as Binary Classifier.
- 0-1, Yes-No, positive-negative, True-False, Pass-Fail, Alive-Dead.
- Fraud/Non Fraud.



DATASET OVERVIEW

The dataset consists of 31 columns, namely time, V1, V2 ……. amount, class.

The time columns refers to the elapsed time in seconds since the first transaction



There are 28 columns from V1 – V28 consisting of features relating to data due to confidentiality issues, relating to age, income and other confidential details

The amount column has transaction value in dollars for each transaction.

The class column is the target variable required to classify fraud and non fraud denoted by 1 and 0 in dataset respectively

DATASET OVERVIEW

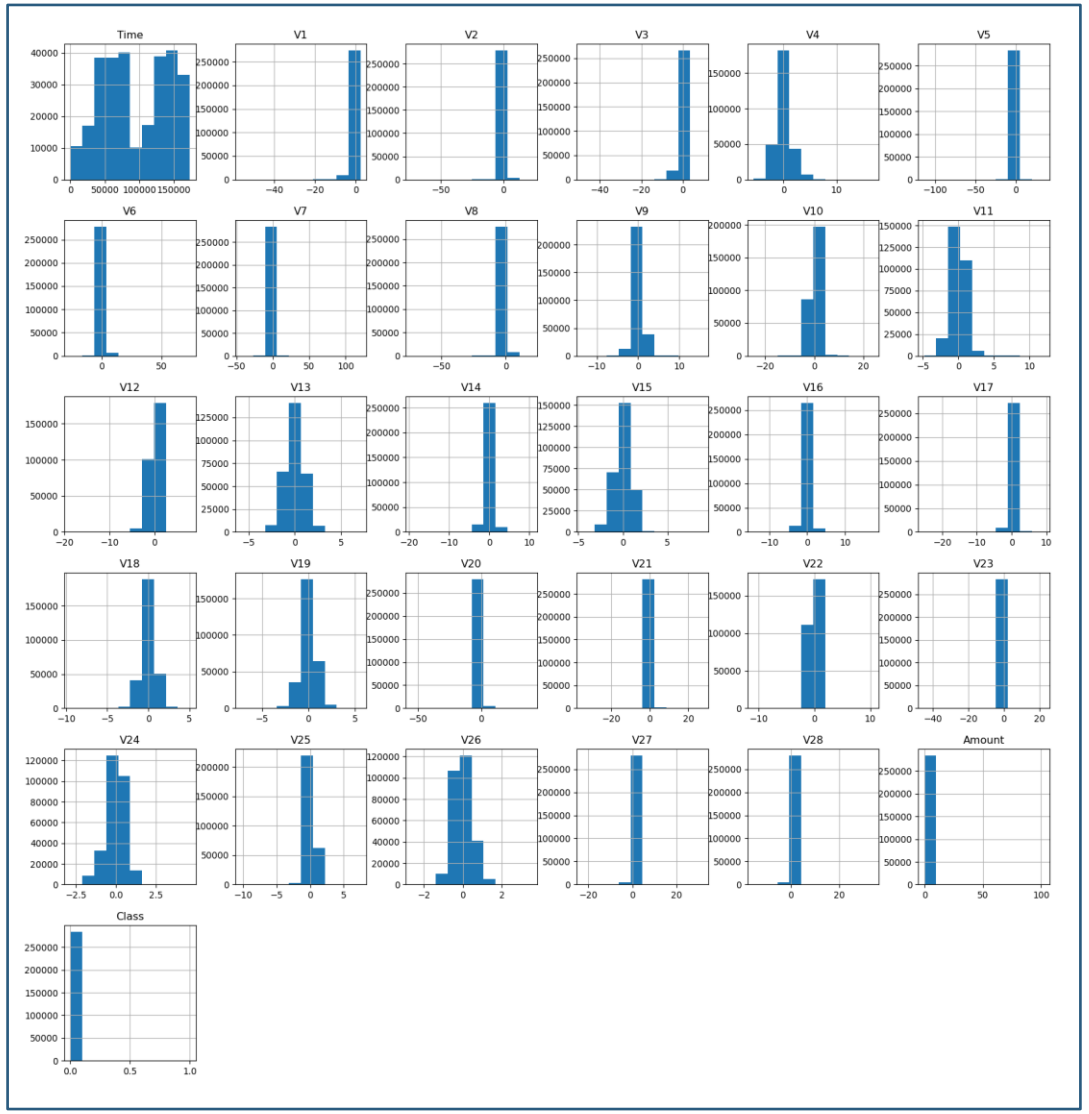
- The Columns V1 – V28 all the features went through a PCA transformation (Dimensionality Reduction technique) except for 'time' and 'Amount'.
- That means the dataset has been transformed to maintain the confidentiality of that data
- For PCA features need to be previously scaled so we can assume that they are (except for 'time' and 'Amount').

Time	V1	V2	V3	...	V26	V27	V28	Amount	Class
0	-1.35981	-0.07278	2.536347	...	-0.18912	0.133558	-0.02105	149.62	0
0	1.191857	0.266151	0.16648	...	0.125895	-0.00898	0.014724	2.69	0
1	-1.35835	-1.34016	1.773209	...	-0.1391	-0.05535	-0.05975	378.66	0
1	-0.96627	-0.18523	1.792993	...	-0.22193	0.062723	0.061458	123.5	0
2	-1.15823	0.877737	1.548718	...	0.502292	0.219422	0.215153	69.99	0

DESCRIPTIVE STATISTICS

- The Columns don't appear to have null values
- There is a possibility of outliers for feature V1 – V28
- The Amount column has a big gap from Third Quartile (Q3, 75%) to Max Value indicating Outliers for the column

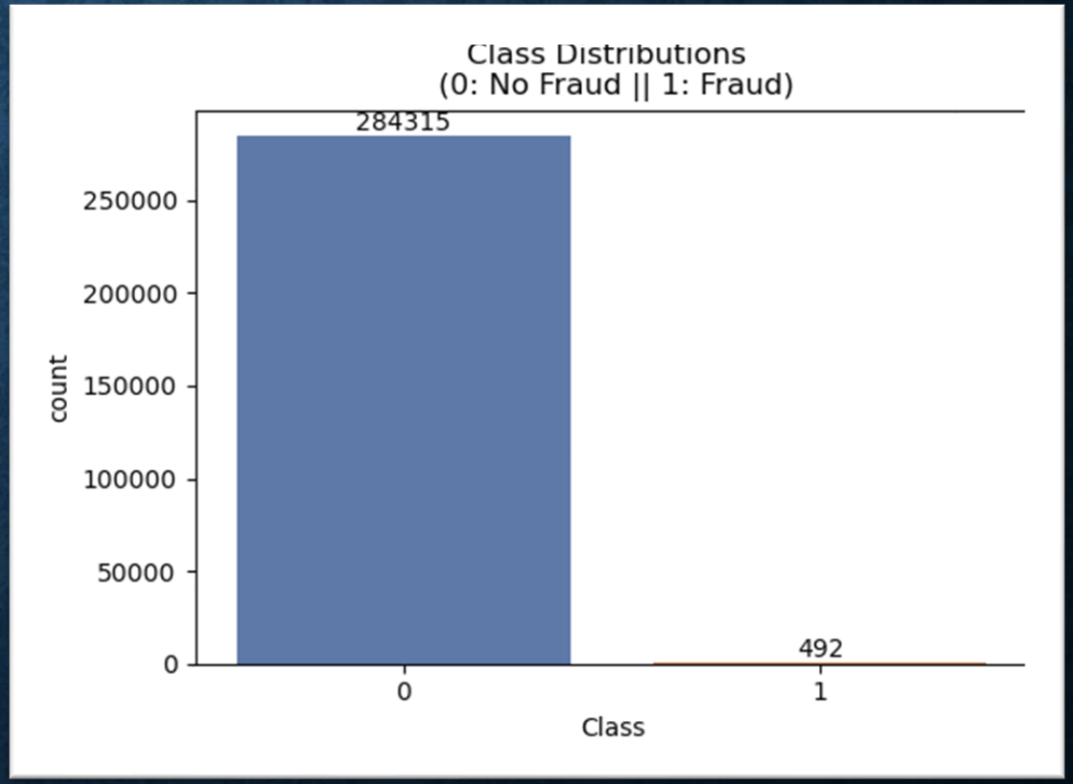
	count	mean	std	min	25%	50%	75%	max
Time	284807	9.48E+04	47488.15	0	54201.5	84692	139320.5	172792
V1	284807	1.76E-12	1.958696	-56.4075	-0.92037	0.018109	1.315642	2.45493
V2	284807	-8.25E-13	1.651309	-72.7157	-0.59855	0.065486	0.803724	22.05773
V3	284807	-9.65E-13	1.516255	-48.3256	-0.89037	0.179846	1.027196	9.382558
V4	284807	8.32E-13	1.415869	-5.68317	-0.84864	-0.01985	0.743341	16.87534
V5	284807	1.65E-13	1.380247	-113.743	-0.6916	-0.05434	0.611926	34.80167
V6	284807	4.25E-13	1.332271	-26.1605	-0.7683	-0.27419	0.398565	73.30163
V7	284807	-3.05E-13	1.237094	-43.5572	-0.55408	0.040103	0.570436	120.5895
V8	284807	8.78E-14	1.194353	-73.2167	-0.20863	0.022358	0.327346	20.00721
V9	284807	-1.18E-12	1.098632	-13.4341	-0.6431	-0.05143	0.597139	15.595
V10	284807	7.09E-13	1.08885	-24.5883	-0.53543	-0.09292	0.453923	23.74514
V11	284807	1.87E-12	1.020713	-4.79747	-0.76249	-0.03276	0.739593	12.01891
V12	284807	1.05E-12	0.999201	-18.6837	-0.40557	0.140033	0.618238	7.848392
V13	284807	7.13E-13	0.995274	-5.79188	-0.64854	-0.01357	0.662505	7.126883
V14	284807	-1.47E-13	0.958596	-19.2143	-0.42557	0.050601	0.49315	10.52677
V15	284807	-5.23E-13	0.915316	-4.49895	-0.58288	0.048072	0.648821	8.877742
V16	284807	-2.28E-13	0.876253	-14.1299	-0.46804	0.066413	0.523296	17.31511
V17	284807	-6.43E-13	0.849337	-25.1628	-0.48375	-0.06568	0.399675	9.253526
V18	284807	4.95E-13	0.838176	-9.49875	-0.49885	-0.00364	0.500807	5.041069
V19	284807	7.06E-13	0.814041	-7.21353	-0.4563	0.003735	0.458949	5.591971
V20	284807	1.77E-12	0.770925	-54.4977	-0.21172	-0.06248	0.133041	39.4209
V21	284807	-3.41E-13	0.734524	-34.8304	-0.2284	-0.02945	0.186377	27.20284
V22	284807	-5.72E-13	0.725702	-10.9331	-0.54235	0.006782	0.528554	10.50309
V23	284807	-9.73E-13	0.62446	-44.8077	-0.16185	-0.01119	0.147642	22.52841
V24	284807	1.46E-12	0.605647	-2.83663	-0.35459	0.040976	0.439527	4.584549
V25	284807	-6.99E-13	0.521278	-10.2954	-0.31715	0.016594	0.350716	7.519589
V26	284807	-5.62E-13	0.482227	-2.60455	-0.32698	-0.05214	0.240952	3.517346
V27	284807	3.33E-12	0.403632	-22.5657	-0.07084	0.001342	0.091045	31.6122
V28	284807	-3.52E-12	0.330083	-15.4301	-0.05296	0.011244	0.07828	33.84781
Amount	284807	8.83E+01	250.1201	0	5.6	22	77.165	25691.16
Class	284807	1.73E-03	0.041527	0	0	0	0	1



DATA DISTRIBUTION

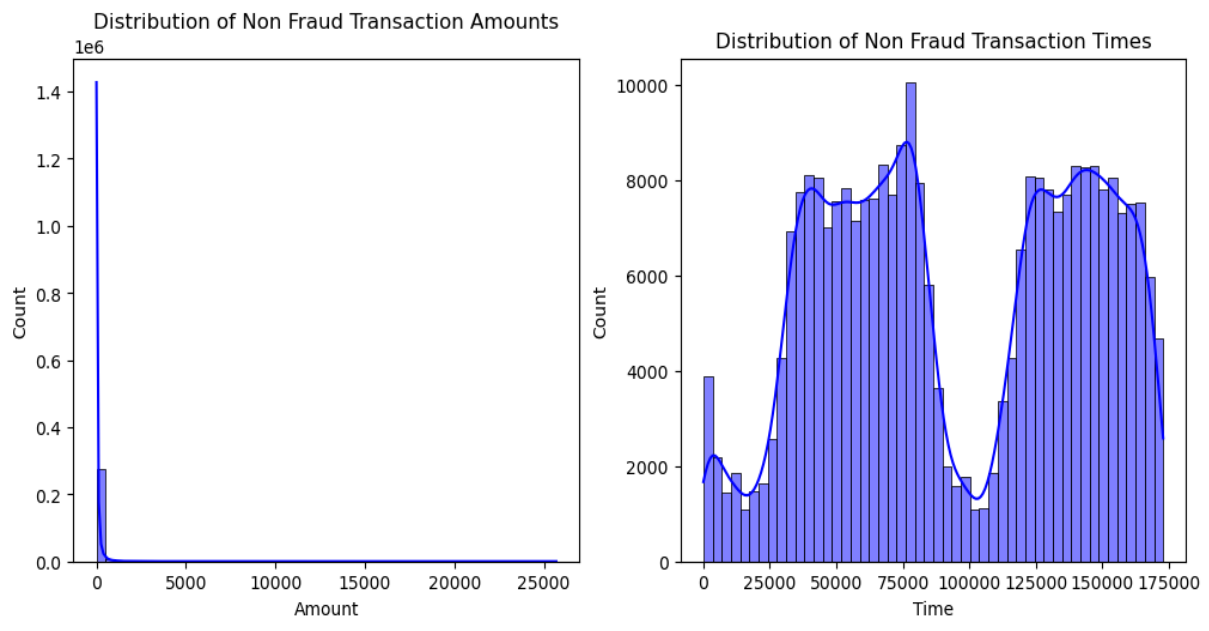
CLASS IMBALANCE OF TARGET VARIABLE

- The Target Variable column 'Class' has Huge Imbalance with respect to Non Fraud and Fraud Values,
- No Frauds 99.83 % of the dataset
- Frauds 0.17 % of the dataset.
- The huge imbalance leads to a biggest problem of for evaluating the performance of model.
- To evaluate performance we cant use Accuracy because 99% values will be correct on just overview

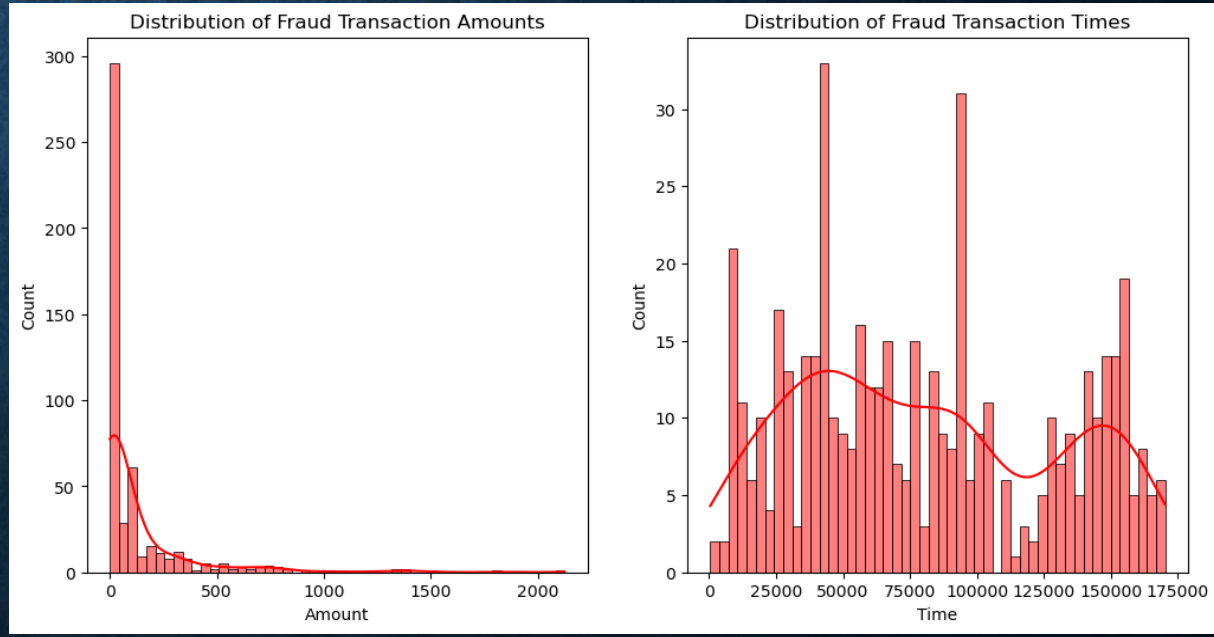


DISTRIBUTION OF TARGET VARIABLES CLASS

Distribution of 'Amount' and 'Time' For Non-Fraud

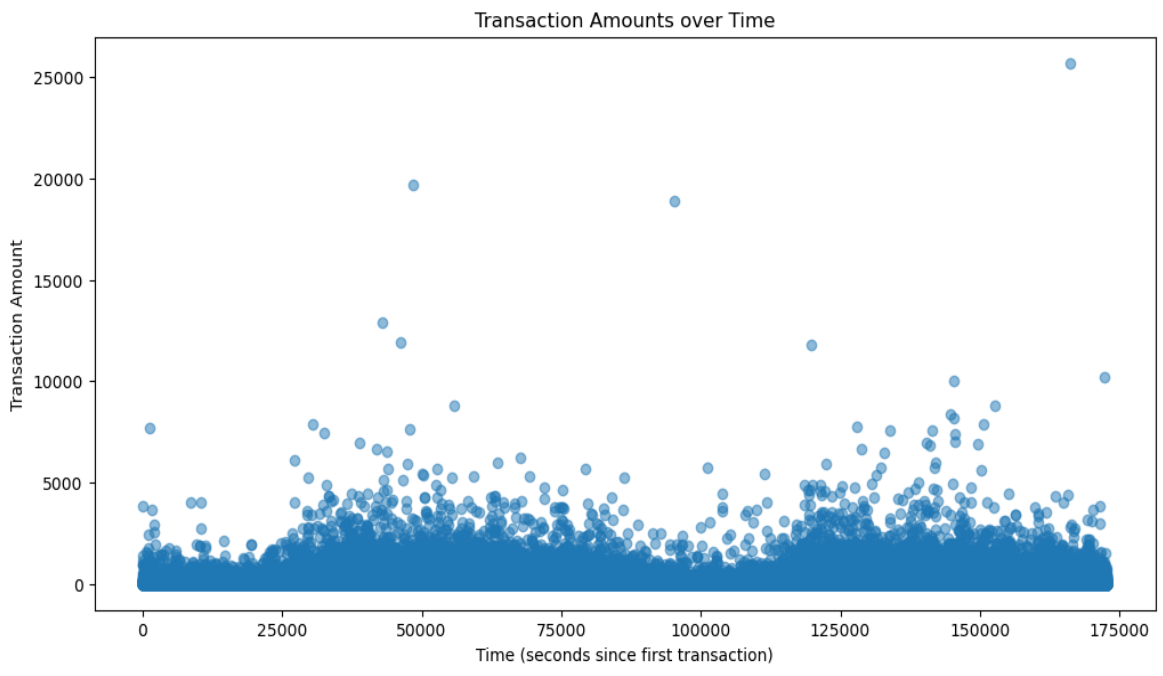


Distribution of 'Amount' and 'Time' For Fraud

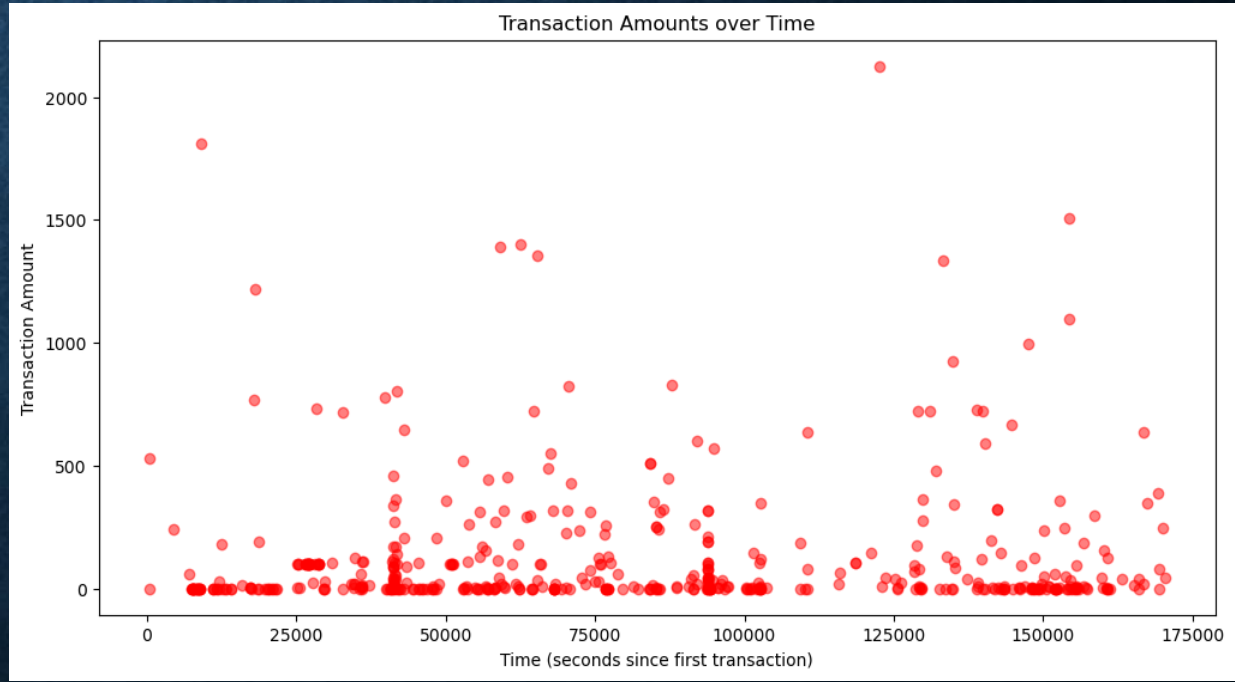


TIME VS. AMOUNT ANALYSIS

Time vs Amount for Non-Fraud Transactions

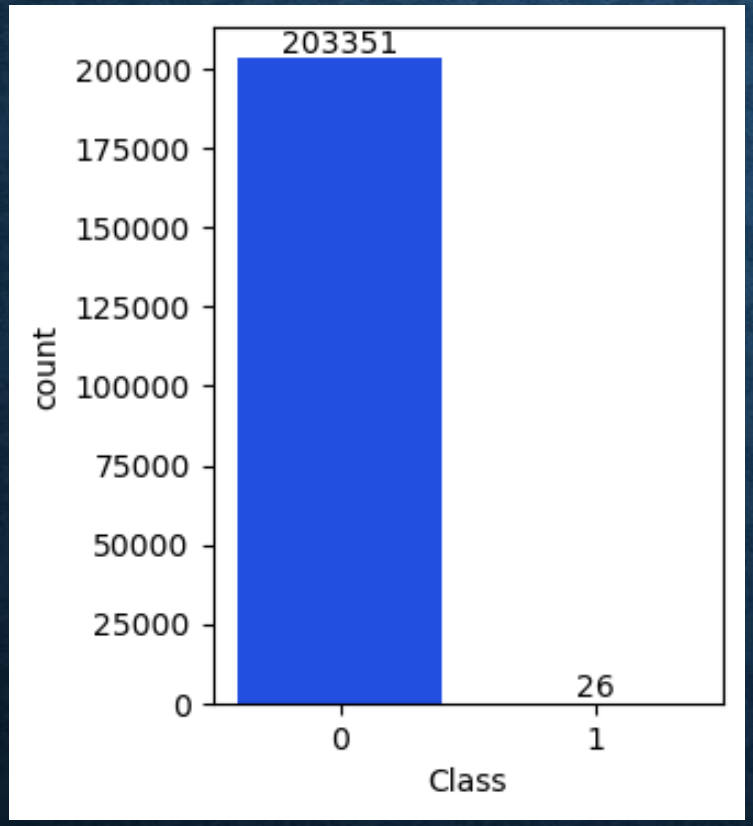


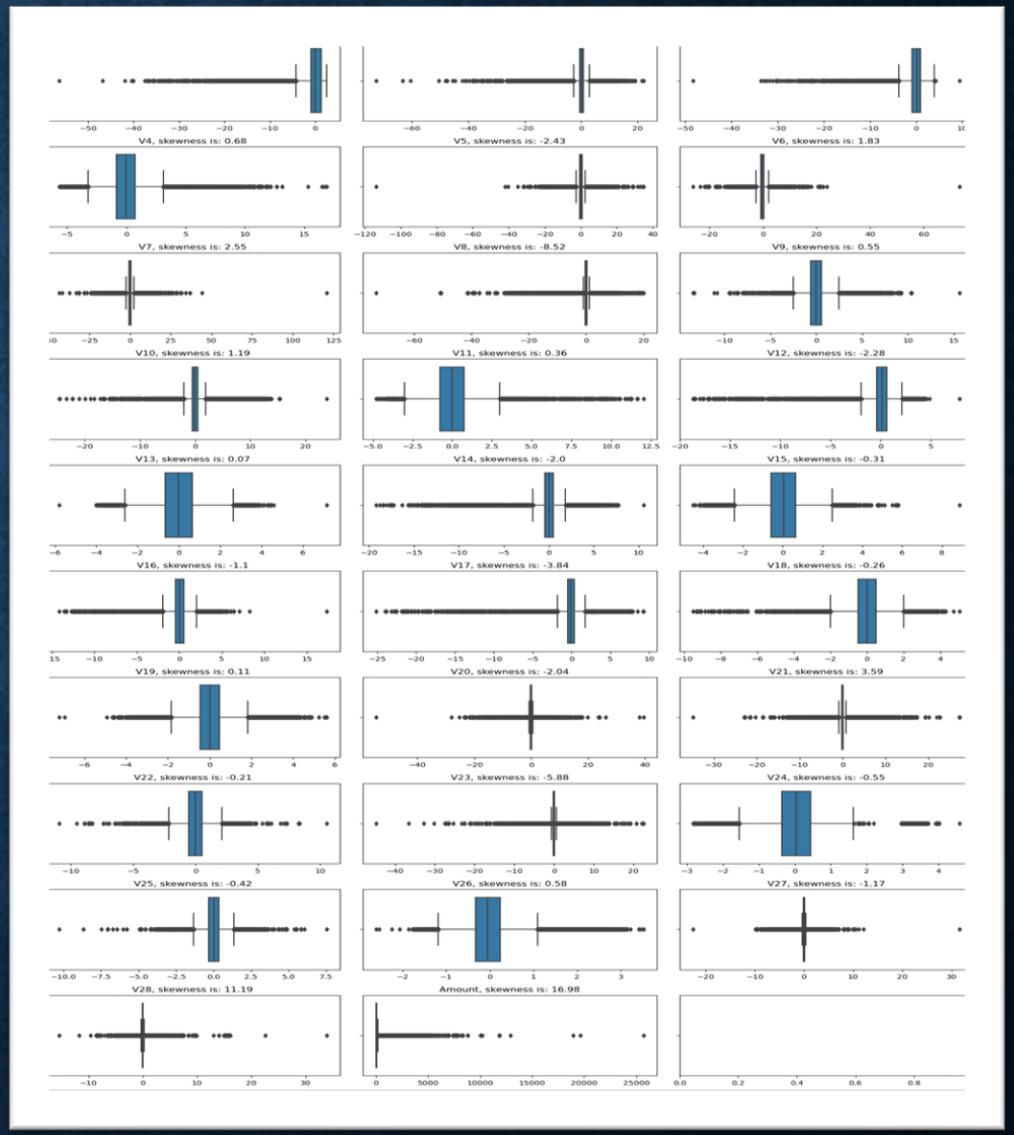
Time vs Amount for Fraud Transactions

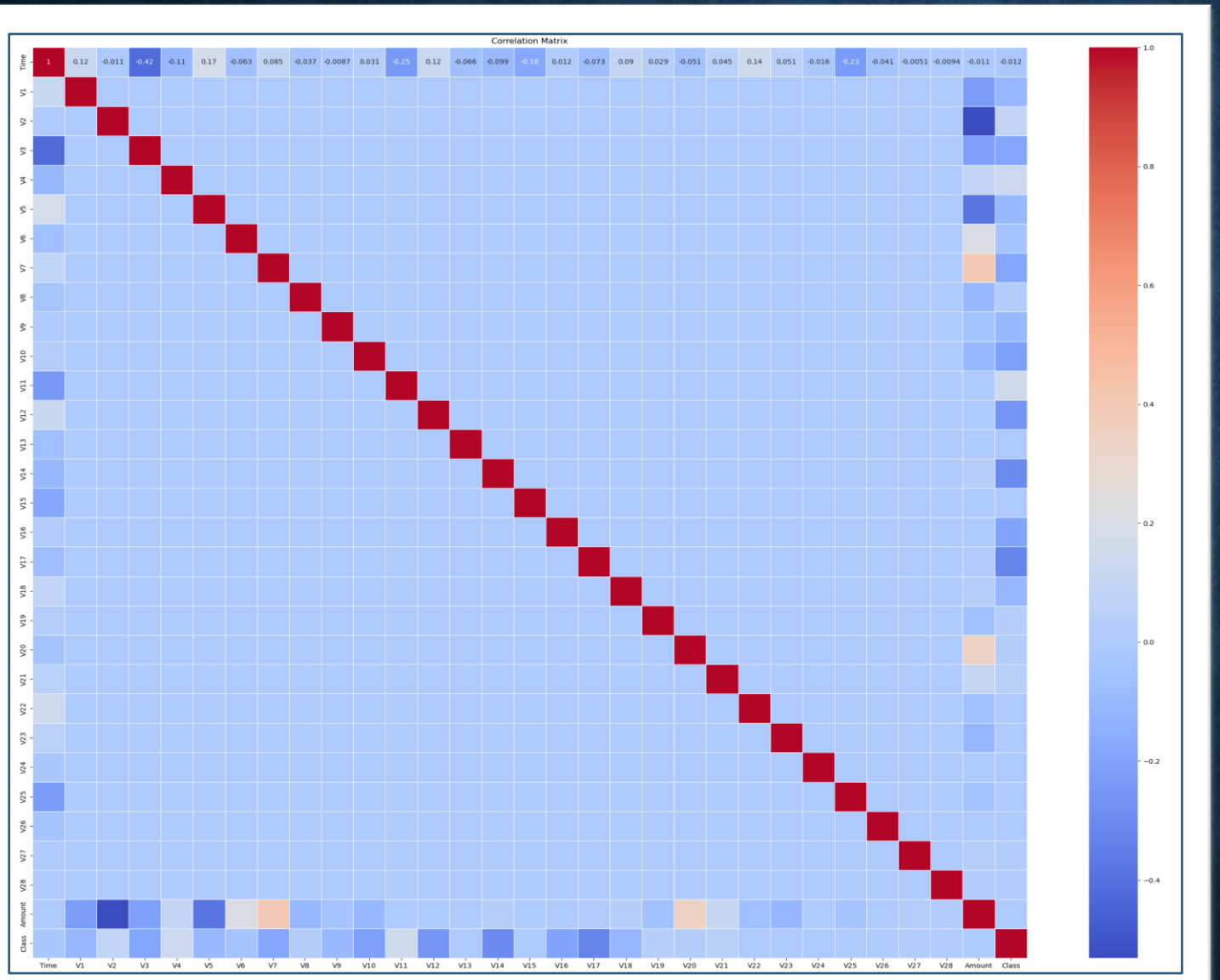


OUTLIER DETECTION USING BOX PLOTS

- There are only 26 Values left for Minority Class
- For a Highly Imbalanced Dataset such as this one where there are less number of Target Variable to begin with it would be detrimental



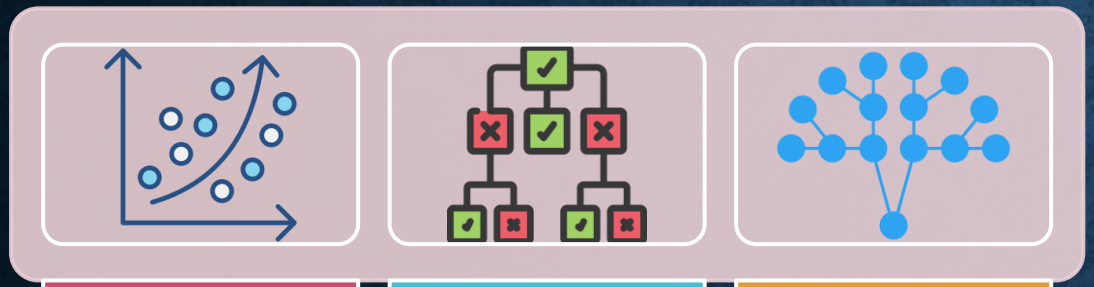




CORRELATION MATRIX

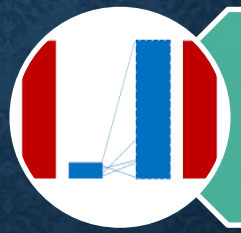
MODELING TECNIQUES USED AND HANDALING CALSS IMBALANCE

Models Used to Classify



- Logistics Regression
- Decision Tree
- eXtreme Gradient Boosting (XGB)

Sampling Technique Used to Treat Class Imbalance

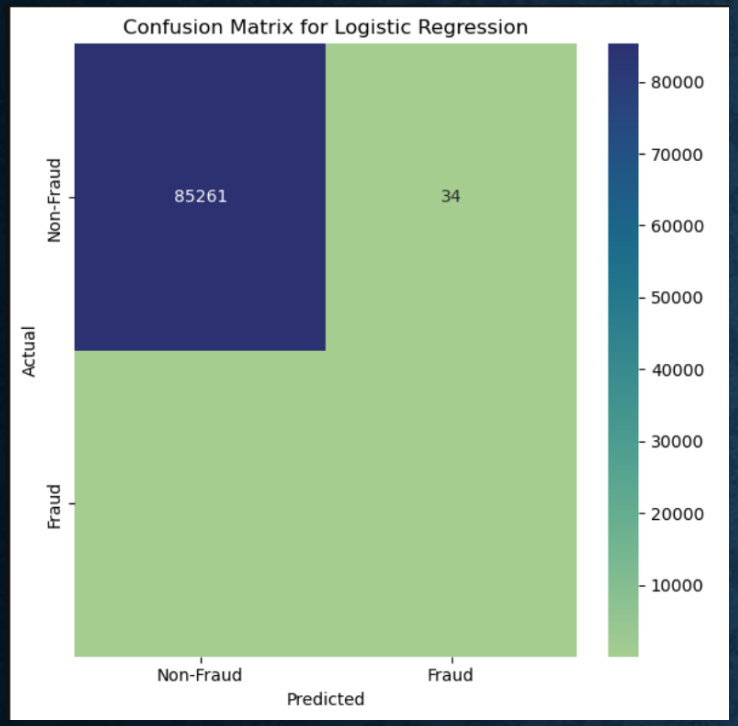


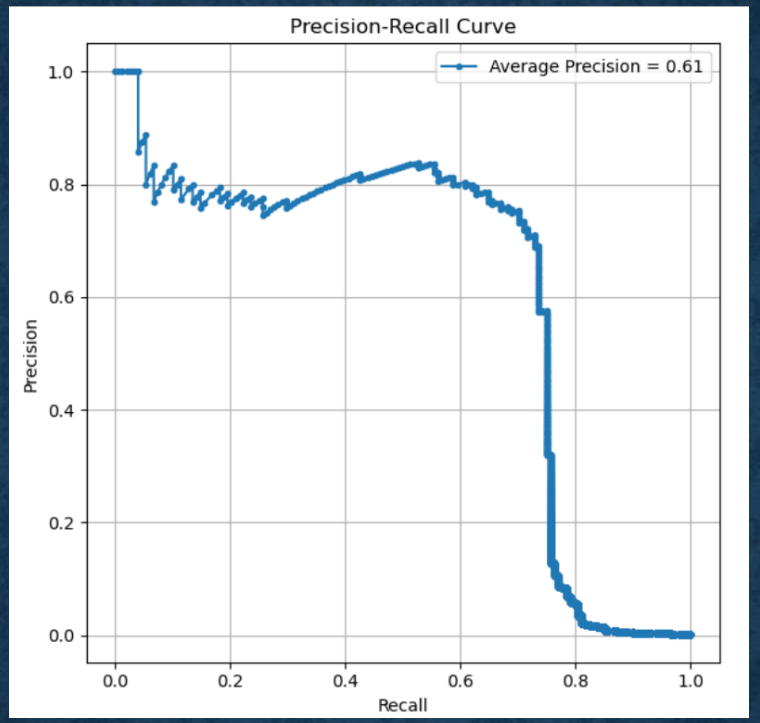
- Synthetic Minority Over-sampling TEchnique (SMOTE)

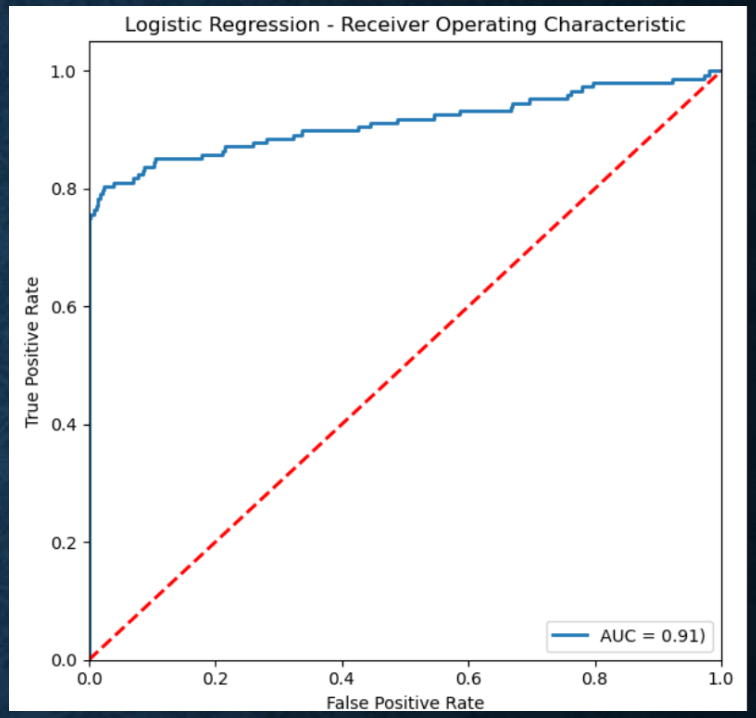
- SMOTE-Tomek

SUCCESS METRICS - OPTIMISING FOR RECAL

- Recall: The ability of a model to find all the relevant cases within a data set. The number of true positives divided by the number of true positives plus the number of false negatives.
- In most high-risk detection cases (like cancer), recall is a more important evaluation metric than precision.
- In the case of credit card fraud detection, we want to avoid false negatives as much as possible. Fraud transactions cost us a lot and thus we want to take appropriate measures to prevent them. A false negative case means that a fraud-positive transaction is assessed to genuine transaction, which is detrimental. In this use case, false positives (a genuine transaction as fraud-positive) are not as important as preventing a fraud.



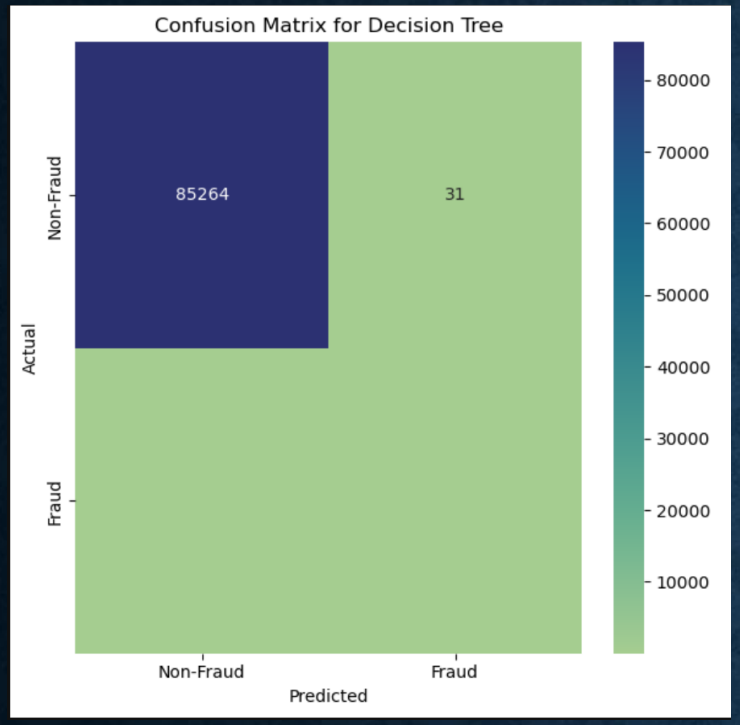


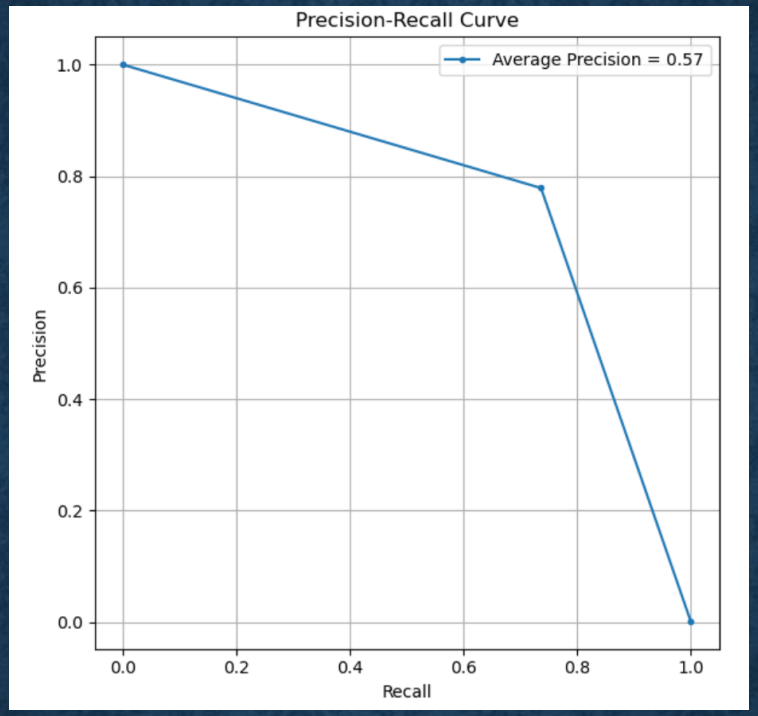


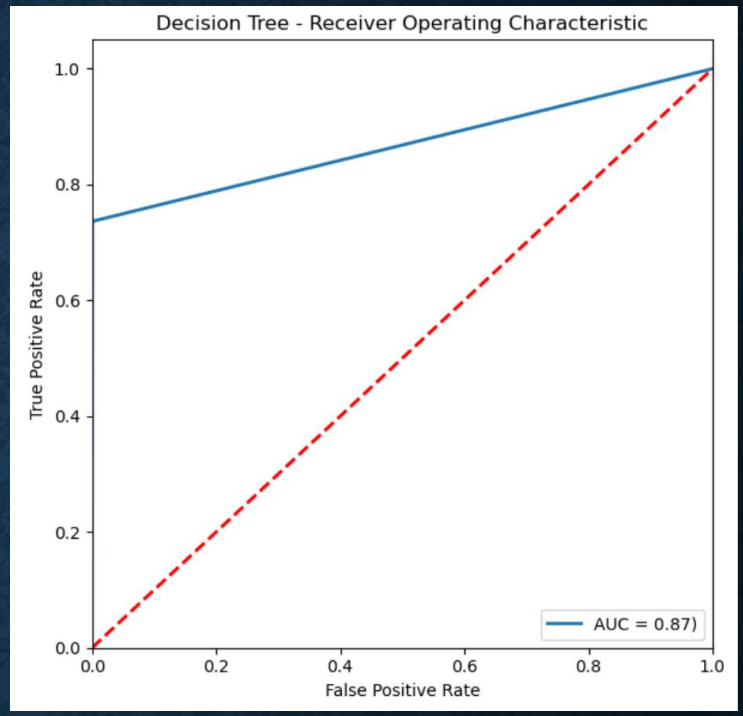
LOGISTICS REGRESSION

- The Model is Underfitting based of F1, Recall and Precision for the Most Parts with increase in Training Size
- The Classification Model has a good recall score it can be expected to improve after providing additional data to the model

Recall:	• 70.27%
Precision:	• 75.36%
F1 Score:	• 72.72%
Accuracy:	• 99.90%
ROC-AUC Score:	• 91.21%



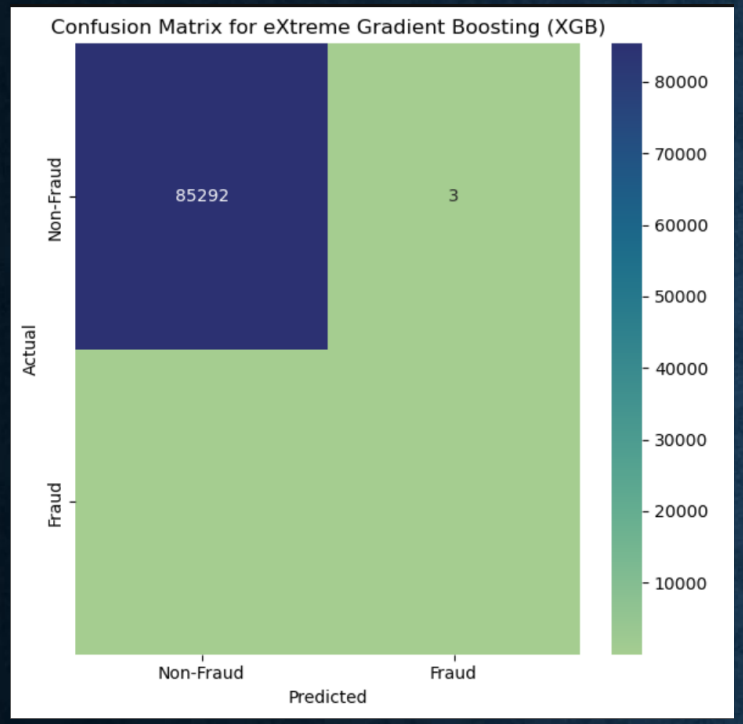


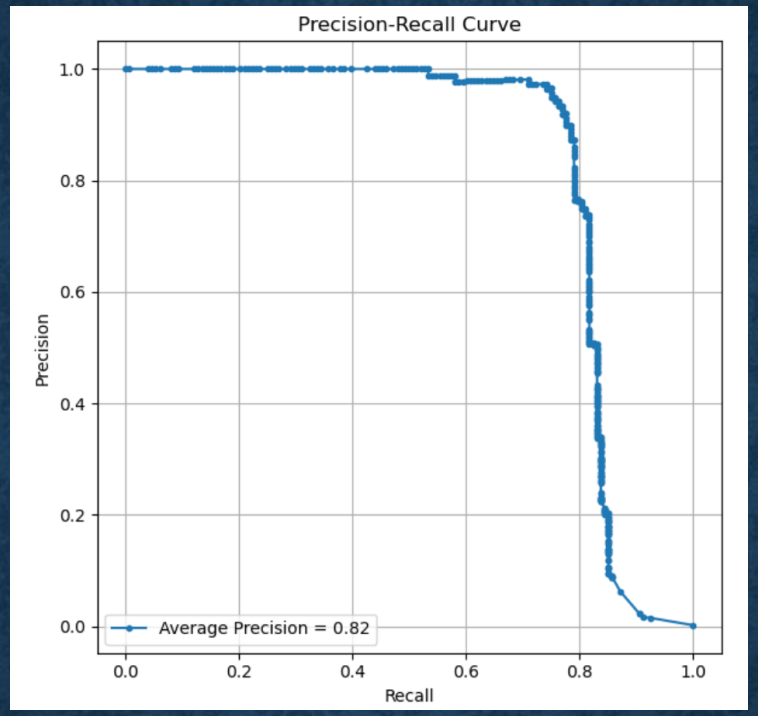


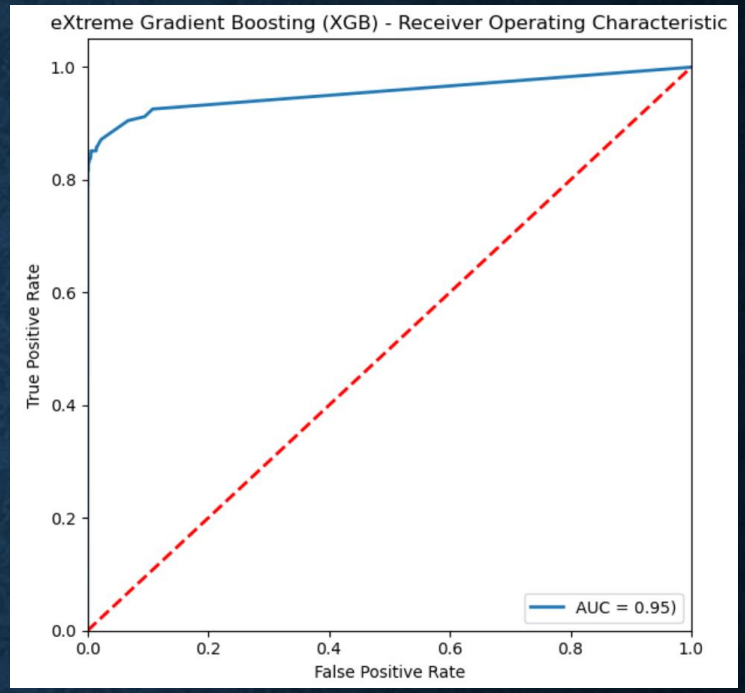
DECISION TREE CLASSIFIER

- The Model is Severely Underfitting based of F1, Recall and Precision as the Trianing Score Increases
- The Classification Model has a good recall score it can be expected to improve after providing additional data to the model
- Furthermore the Average Precision-Recall is less than 0.57 which is very poor performance for a classifier

Recall:	• 73.64%
Precision:	• 77.85%
F1 Score:	• 75..69%
Accuracy:	• 99.91%
ROC-AUC Score:	• 86.80%







EXTREME GRADIENT BOOSTING (XGB)

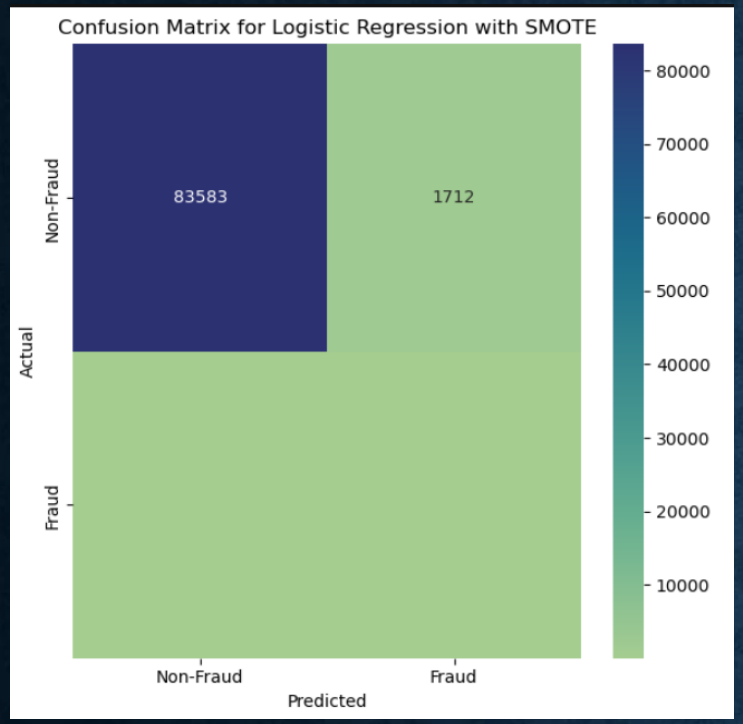
- The Model is Severely Underfitting based of F1, Recall and Precission, as the Trianing Score Increases model keeps converging indicating a possible improvement with addition of data
- The Classification Model has a good recall score it can be expected to improve after providing additional data to the model the Average Precision-Recallmoreless tha820.57 which is vgoodpoor performance for a classi with respect to anamoly detection Additionally the AUC Score is o.95 indicating a good performance with respect to Fitting of Data with Test Sets

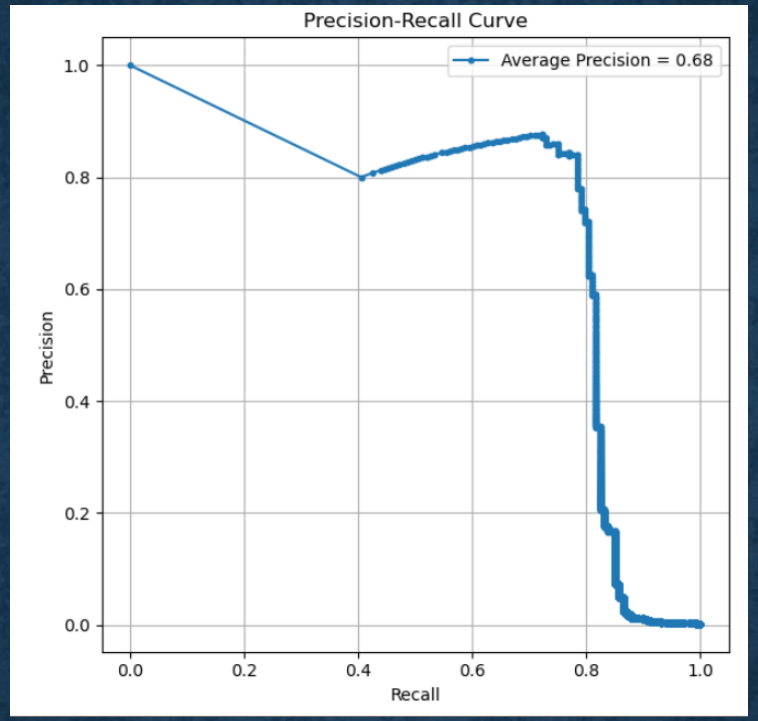
• 74.32%	Recall:
• 97.34%	Precision:
• 84.29%	F1 Score:
• 99.95%	Accuracy:
• 95.48%	ROC-AUC Score:

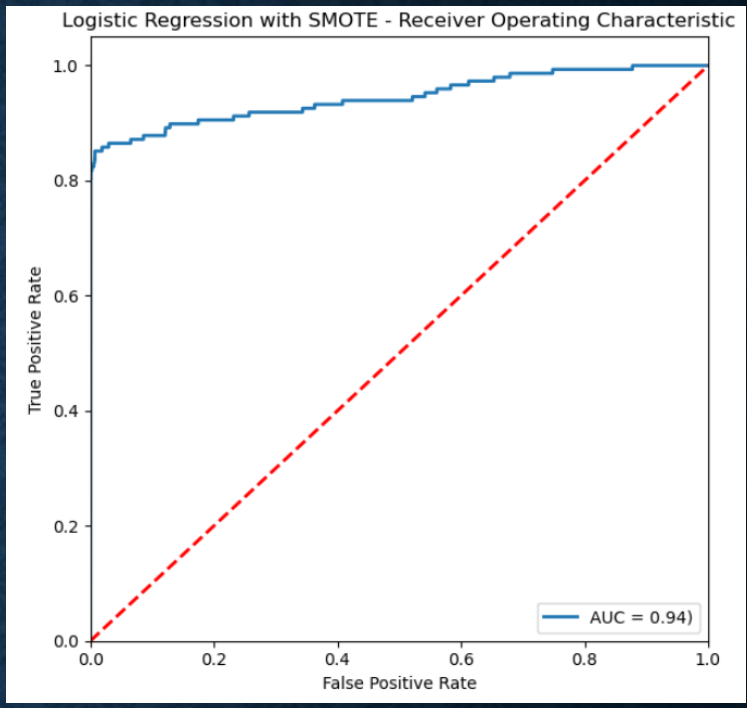
SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE (SMOTE)

SMOTE (Synthetic Minority Oversampling Technique) synthesize elements for the minority class.

SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line.



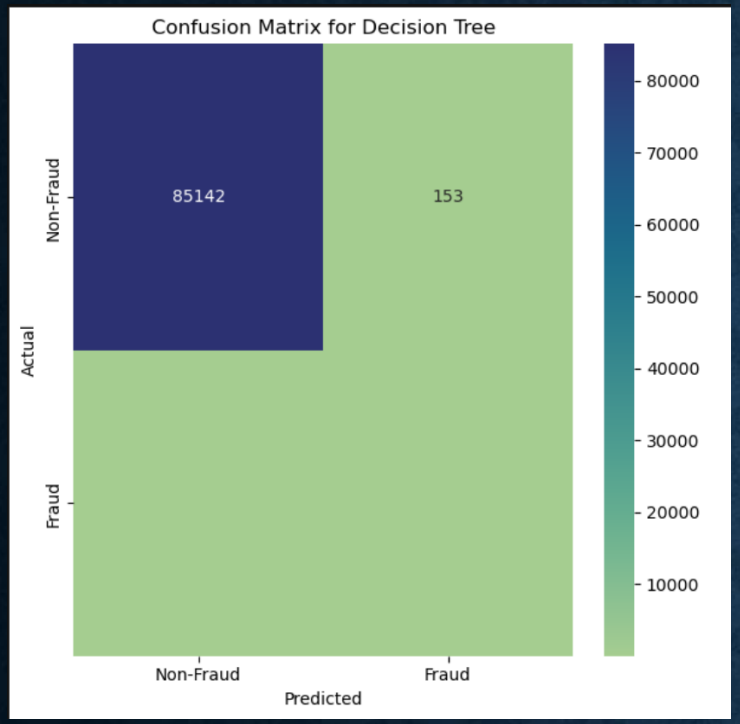


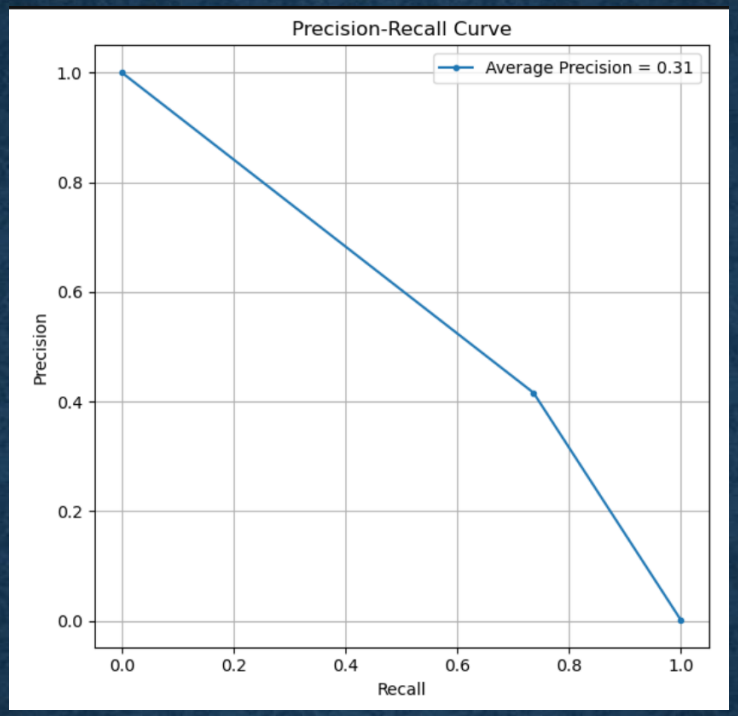


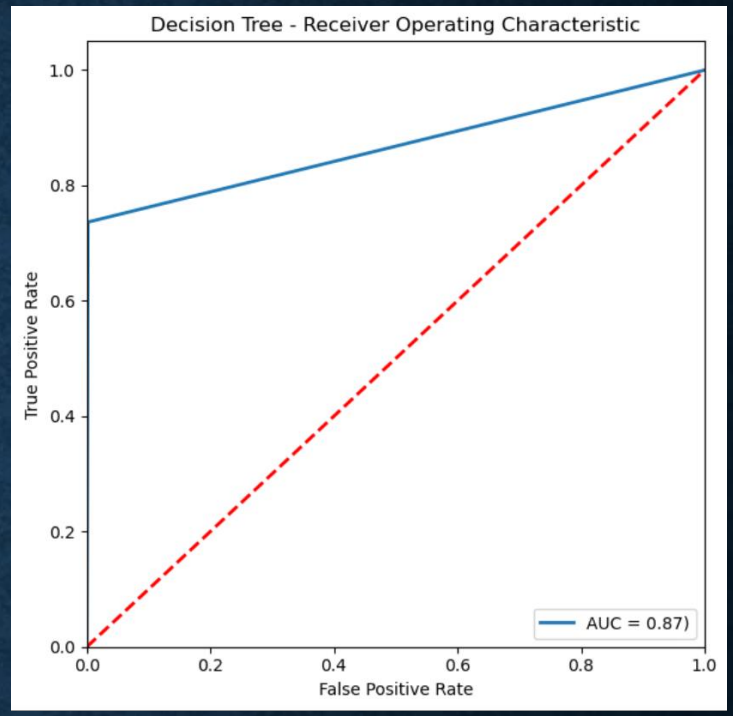
LOGISTICS REGRESSION WITH SMOTE

- After OverSampling using SMOTE Technique we Observe a Good Fit with respect to F1 score and Recall
- However With respect to Precision there is Overfitting with increase in Training Size
- Furthermore Average-Precision and Recall Score is observed to be under 0.68 which shows improvement but with respect to anamoly detection its is still underperforming

Recall:	• 85.81%
Precision:	• 6.90%
F1 Score:	• 12.78%
Accuracy:	• 97.97%
ROC-AUC Score:	• 94.45%



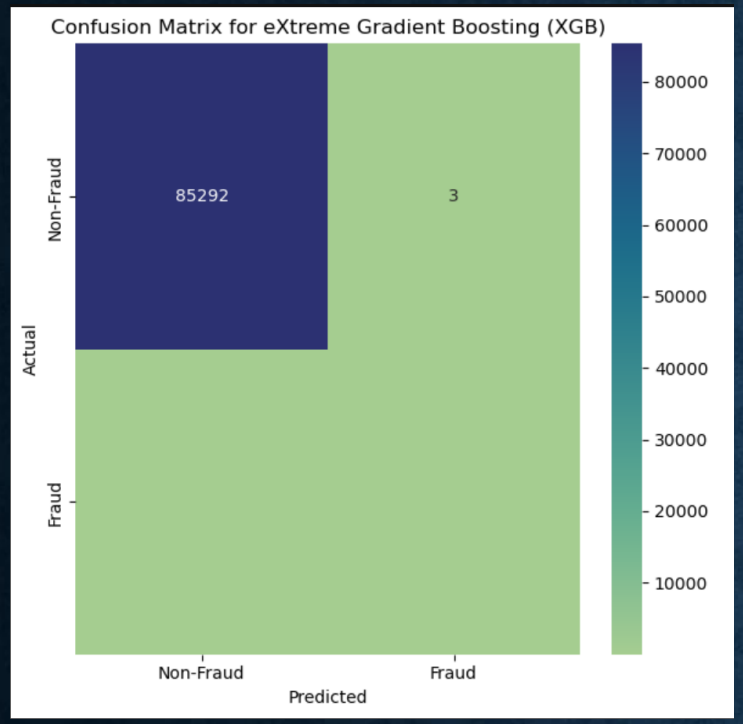


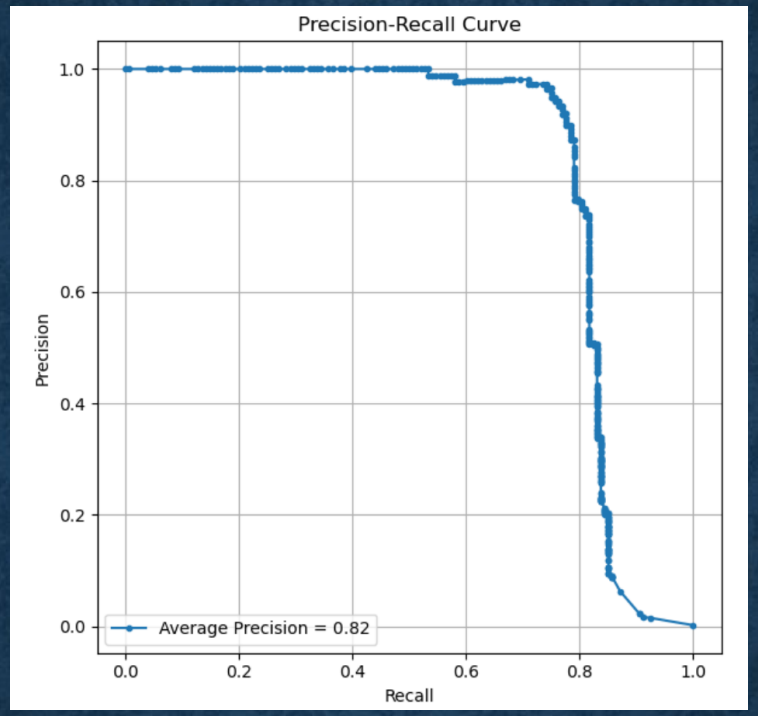


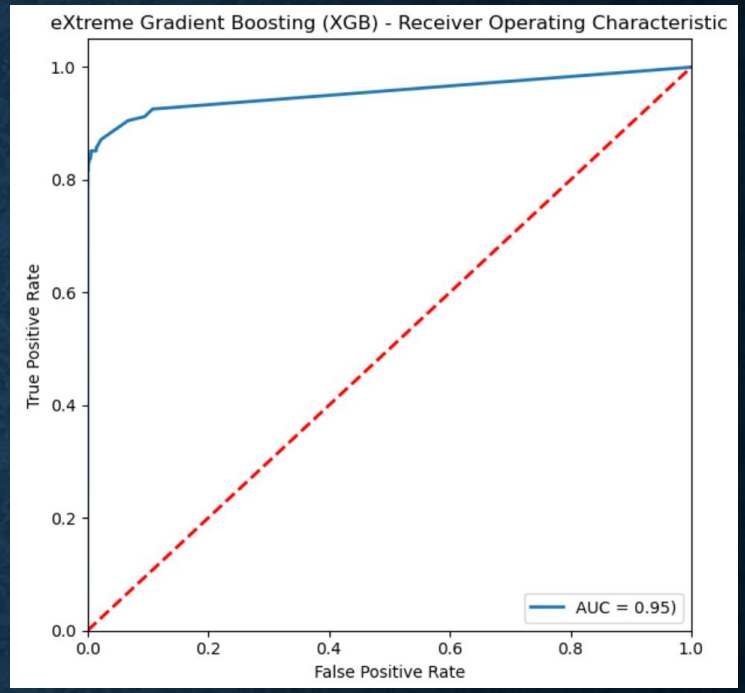
DECISION TREE CLASSIFIER WITH SMOTE

- The Model is Severely Underfitting after Oversampling with SMOTE based of F1, Recall and Precision as the Training Score Increases
- The Classification Model still has has a good recall score but it is still showing underfitting for small training sizes
- Furthermore the Average Precision-Recall is less than 0.31 which is very poor performance for a classifier after Oversampling the Datasets

Recall:	• 73.64%
Precision:	• 41.60%
F1 Score:	• 53.17%
Accuracy:	• 99.75%
ROC-AUC Score:	• 86.80%







EXTREME GRADIENT BOOSTING (XGB) WITH SMOTE

- After Oversampling using the SMOTE Techniques Model is showing severe underfitting and Low Precision and Recall Scores
- Furthermore the Average Precision-Recall Score is decresed indicating not a good performance with Oversampled data

Recall:	• 74.32%
Precision:	• 97.34%
F1 Score:	• 84.29%
Accuracy:	• 99.95%
ROC-AUC Score:	• 95.48%

BALANCING DATA USING SMOTE-TOMEK LINKS

- A combination of over-sampling the minority (abnormal) class and under-sampling the majority (normal) class can achieve better classifier performance than only under-sampling the majority class.
- After Sampling using SMOTE-Tomek Technique we Observe a Good Fit with respect to F1 score and Recall
- However With respect to Precision there is Overfitting with increase in Training Size
- Furthermore Average-Precision and Recall Score is observed to be under 0.68 which shows no improvement with Oversampling with SMOTE it is not performing well anamoly detection its is still underperforming

Logistics Regression Using SMOTE-Tomek	Decision Tree Using SMOTE-Tomek	eXtreme Gradient Boosting (XGB) Using SMOTE-Tomek
Recall : 86.49%	Recall : 73.65%	Recall : 85.14%
Precision : 8.20%	Precision : 41.92%	Precision : 22.46%
F1 Score : 14.98%	F1 Score : 53.43%	F1 Score : 35.54%
Accuracy : 98.30%	Accuracy : 99.78%	Accuracy : 99.47%
ROC_AUC : 95.53%	ROC_AUC : 86.74%	ROC_AUC : 95.87%

CHOSING BEST MODELS BASED ON RECALL

- After Comparing the Recall Scores
- The best performing model is eXtreme Gradient Boosting (XGB) on Imbalanced Dataset without sampling
- The Precision-Recall Score is more than 0.82 indicating good Performance and AOC Score of 0.95 indicating a good fit

Model	recall	precision	f1_score	Accuracy	roc_auc	confusion_matrix
Logistic Regression	70.27%	75.36%	72.73%	99.91%	91.21%	[[85261, 34], [44, 104]]
Decision Tree	73.65%	77.86%	75.69%	99.92%	86.81%	[[85264, 31], [39, 109]]
eXtreme Gradient Boosting (XGB)	74.32%	97.35%	84.29%	99.95%	95.49%	[[85292, 3], [38, 110]]
Logistic Regression with SMOTE	85.81%	6.91%	12.78%	97.97%	94.48%	[[83583, 1712], [21, 127]]
Decision Tree with SMOTE	73.65%	41.60%	53.17%	99.78%	86.73%	[[85142, 153], [39, 109]]
eXtreme Gradient Boosting (XGB) with SMOTE	84.46%	20.39%	32.85%	99.40%	95.20%	[[84807, 488], [23, 125]]
Logistic Regression with SMOTETomek	86.49%	8.20%	14.98%	98.30%	95.53%	[[83862, 1433], [20, 128]]
Decision Tree with SMOTETomek	73.65%	41.92%	53.43%	99.78%	86.74%	[[85144, 151], [39, 109]]
eXtreme Gradient Boosting (XGB) with SMOTETomek	85.14%	22.46%	35.54%	99.47%	95.87%	[[84860, 435], [22, 126]]

CLICK TO VIEW FULL PROJECT

Full Project : Fraud Detection in Minority Class of Credit Card Transactions

THE END