



FRAUD DETECTION

CREDIT CARD FRAUD DETECTION – PREDICTIVE MODEL

Thanks to Springboard mentor – DIPANJAN SARKAR, Data Science Lead – Google, Author

PROBLEM STATEMENT

- Fraudulent transactions increased drastically in recent years
- As payment modes and platform increased, so did fraud landscape
- Thwarting fraud became a difficult task against sophisticated hacking tools and new techniques
- Customer satisfaction is compromised with manual anti-fraud controls



SOLUTION USING MACHINE LEARNING

SUPERVISED LEARNING ALGORITHMS CAN PREDICT CLASSIFICATION PROBLEM EFFICIENTLY

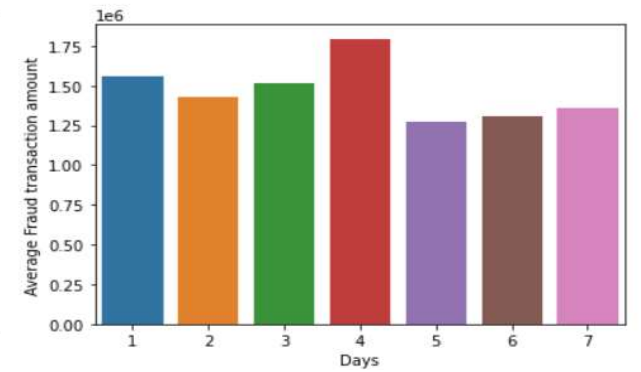
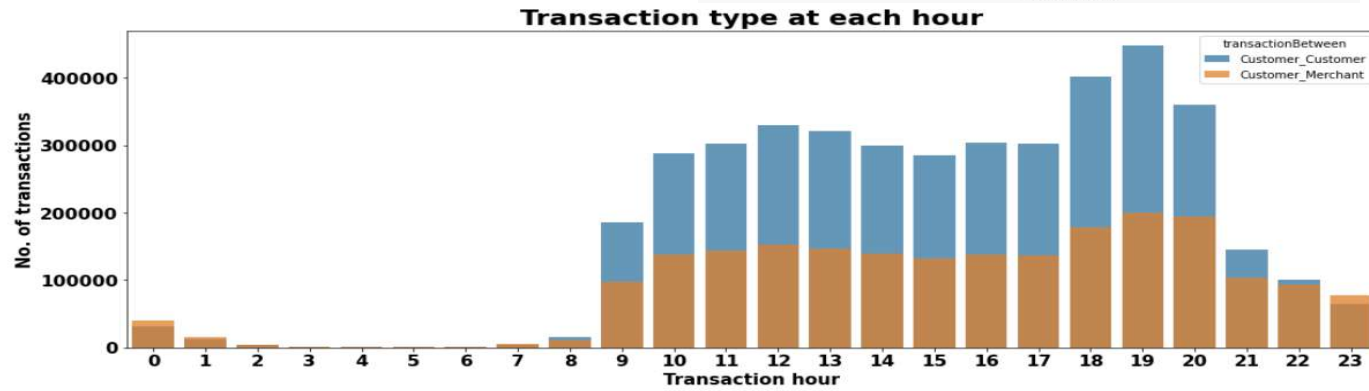
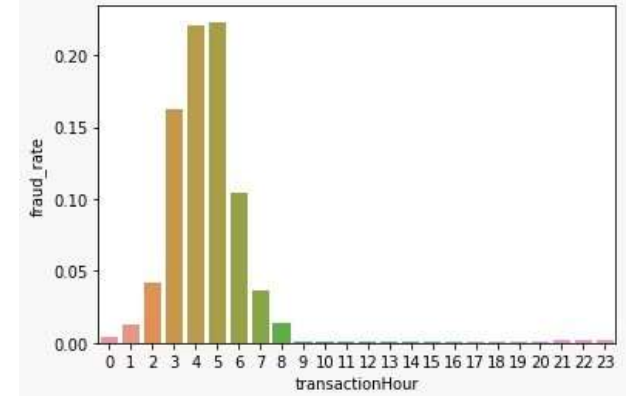
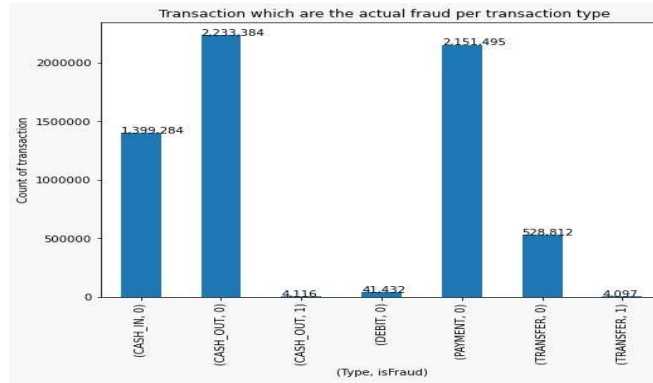
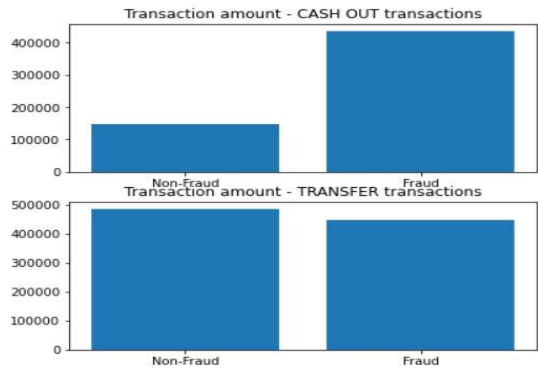


DATA ANALYSIS & WRANGLING

- DATA <https://www.kaggle.com/ealaxi/paysim1>
 - Simulated mobile money transactions
 - 6362620 of records
 - 0.129 % transactions are fraud
 - oldbalanceDest and newbalanceDest info is not available for the transaction to Merchants
 - Highly Imbalanced dataset

	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrg	nameDest	oldbalanceDest	newbalanceDest	isFraud	transactionHour
2	1	TRANSFER	181.0	C1305486145	181.0	0.0	C553264065	0.0	0.0	1	1
3	1	CASH_OUT	181.0	C840083671	181.0	0.0	C38997010	21182.0	0.0	1	1
251	1	TRANSFER	2806.0	C1420196421	2806.0	0.0	C972765878	0.0	0.0	1	1
252	1	CASH_OUT	2806.0	C2101527076	2806.0	0.0	C1007251739	26202.0	0.0	1	1
680	1	TRANSFER	20128.0	C137533655	20128.0	0.0	C1848415041	0.0	0.0	1	1

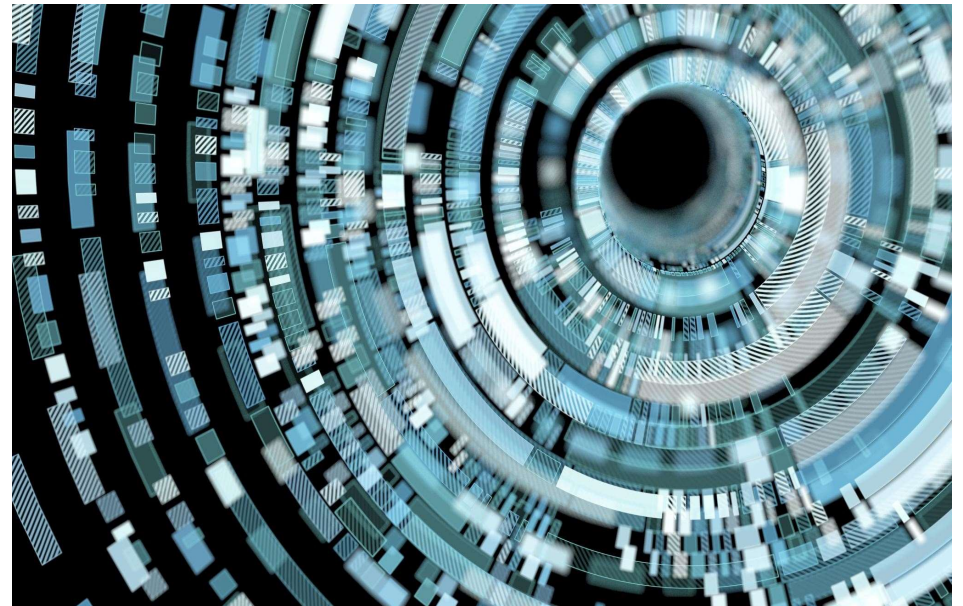




EXPLORATORY DATA ANALYSIS

EXPLORATORY DATA ANALYSIS SUMMARY

- 8213 fraud transactions:
 - 4116 in CASH_OUT
 - 4097 in TRANSFER
- The highest Fraud transaction is 10,000,000
- Most fraud transactions happened between 12:00am and 8:00am
- Transactions are stable during banking hours
- 16 System flagged fraud transaction in payment and cash-out
- Customer to Customer transactions are greater than Customer to Merchant transactions during banking hours
- Customer to Merchant transactions are more during off-peak hours



MODELING



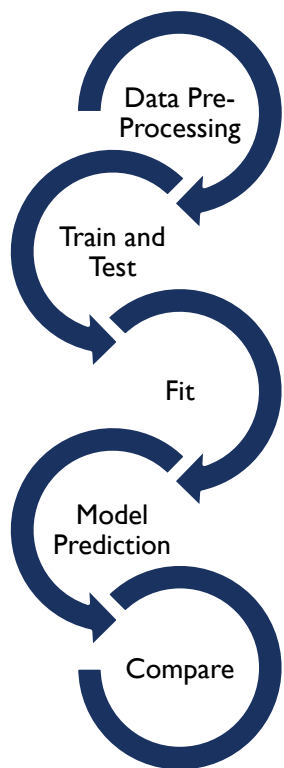
Supervised Learning

Binary Classification

Highly Imbalanced Data

Tools : Python, Sklearn, numpy, Matplotlib, Jupyter Notebook, Seaborn, Pandas

MODELING...

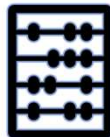


Feature Engineering

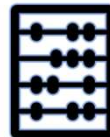
Derived Features

Split Data - 80/20

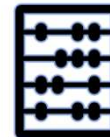
Hyper Parameter Tuning



Random Forest
Classifier



XgBoost



Logistic
Regression



Decision Tree

Accuracy

Precision

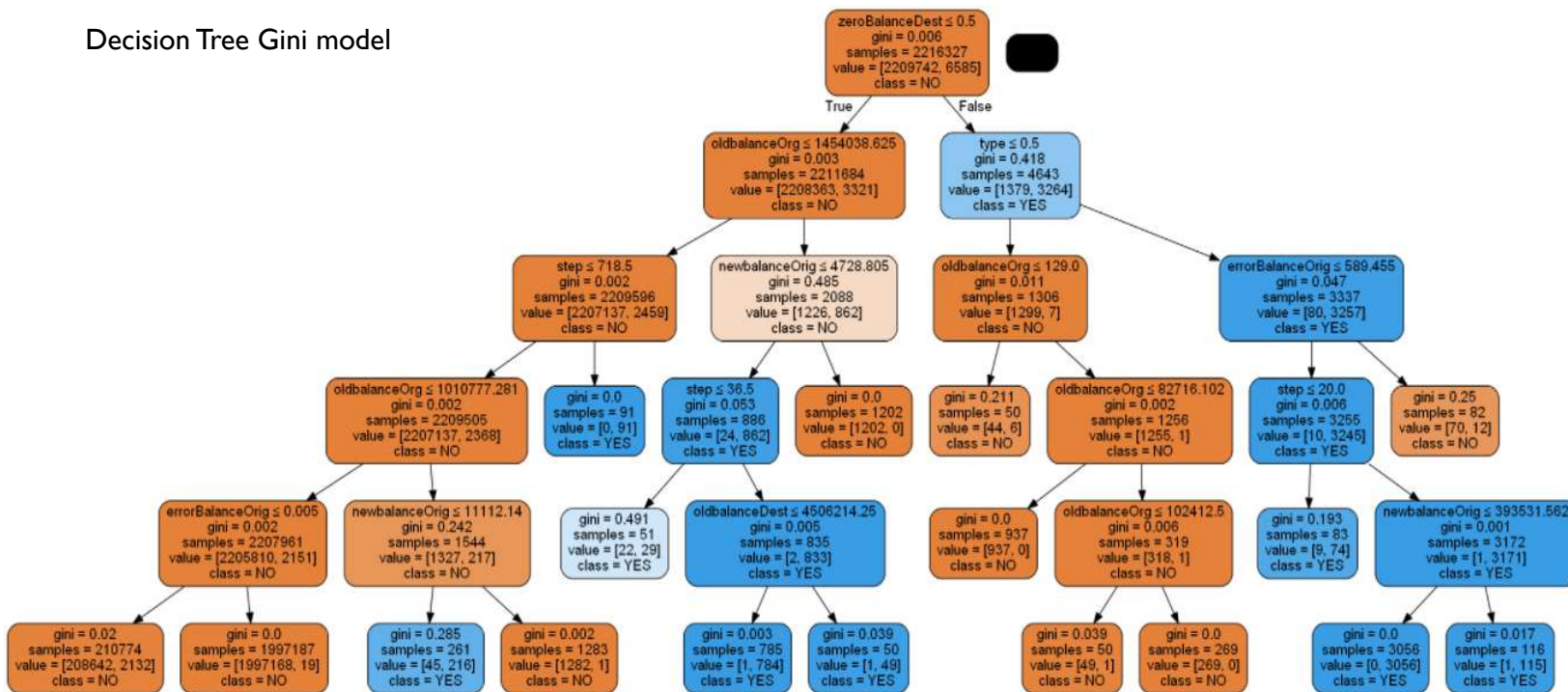
F1-Score

Recall

AUC/ROC Bend

DECISION TREE MODELS

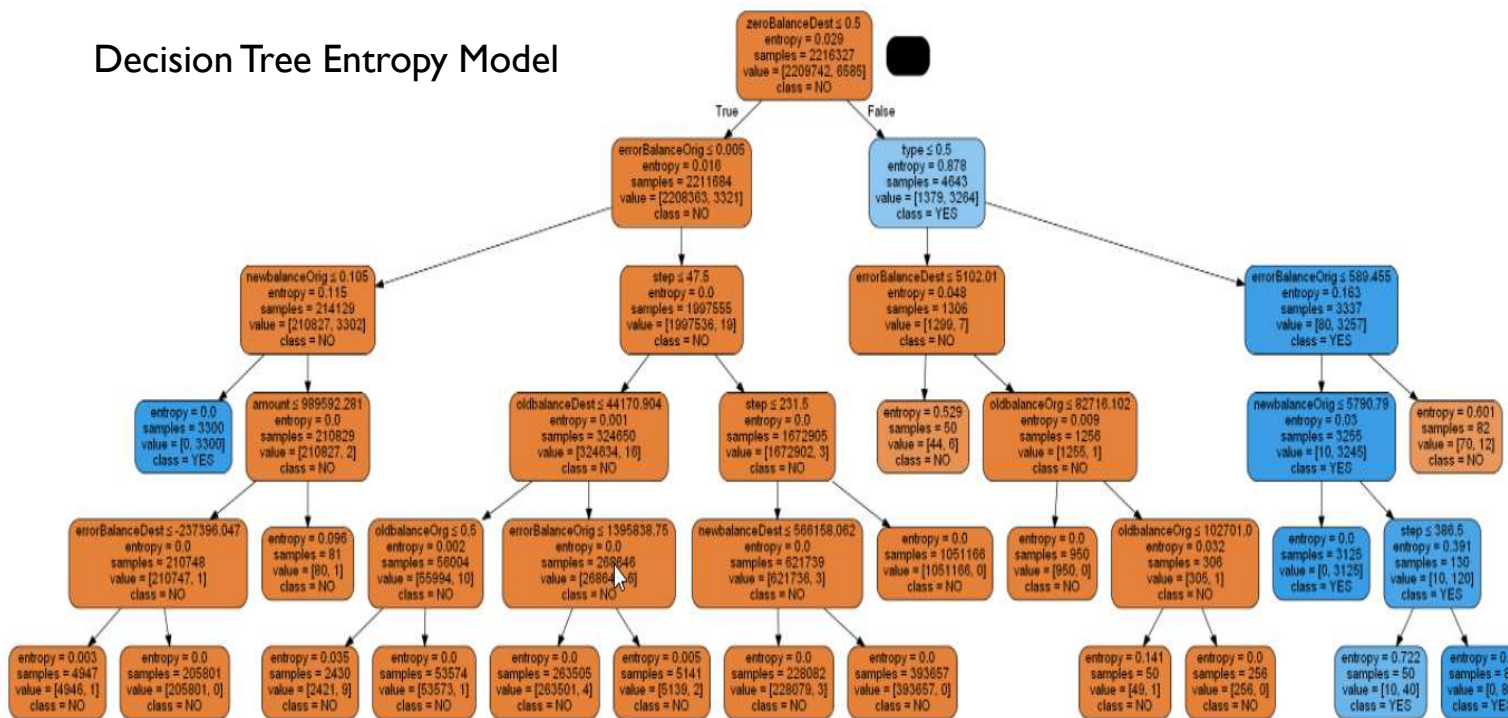
Decision Tree Gini model



- Root node split in this model started with zero balance destination
- Max-depth – 5, MIN sample leaf - 50, MIN sample split - 50
- Class – Yes=>Fraud, No=> Non-Fraud
- FI score – 99.89%
- Precision score – 98.72%

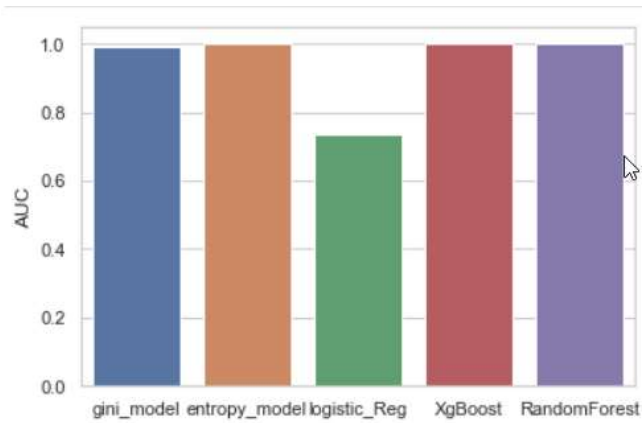
DECISION TREE MODELS...

Decision Tree Entropy Model

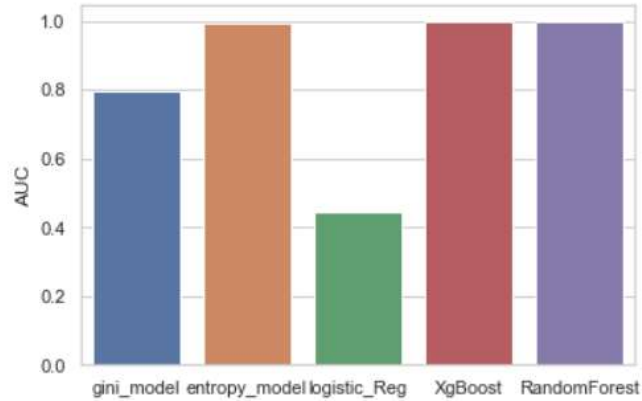


- Root node split in this model started with zero balance destination
- Class – Yes=>Fraud, No=> Non-Fraud
- Max-depth – 5, MIN sample leaf - 50, MIN sample split - 50
- FI score – 99.99%
- Precision score – 99.75%

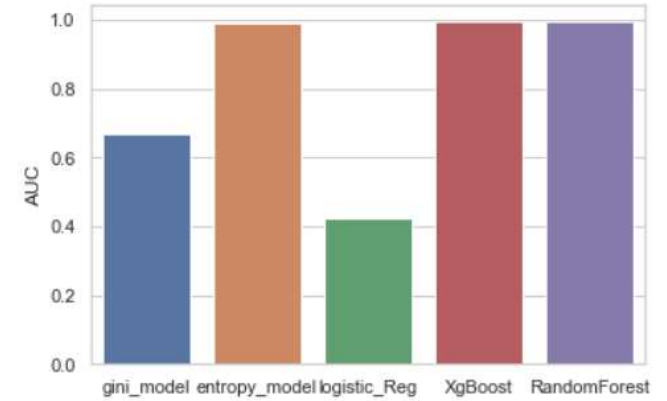
SCORE COMPARISON



ROC-AUC Score



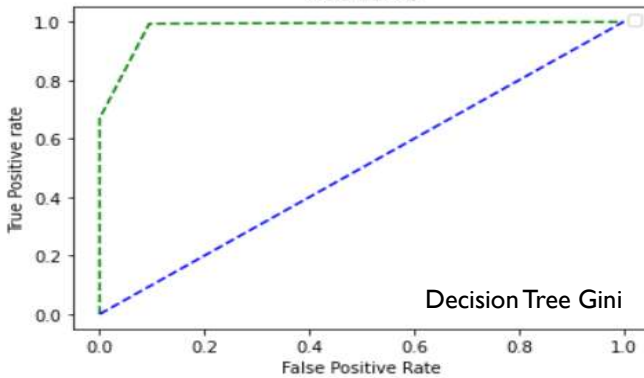
F-I Score



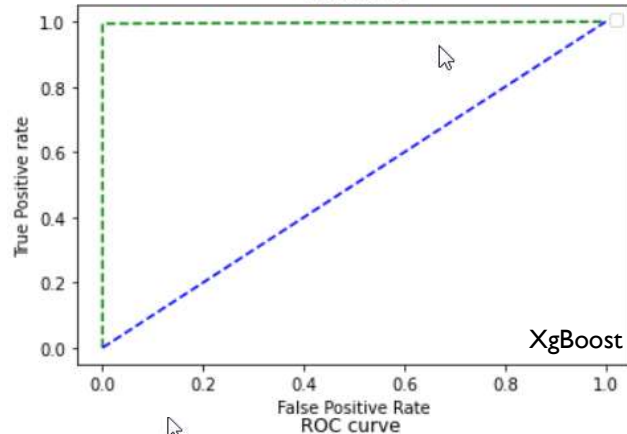
Precision Score

ROC-AUC CURVE COMARISON

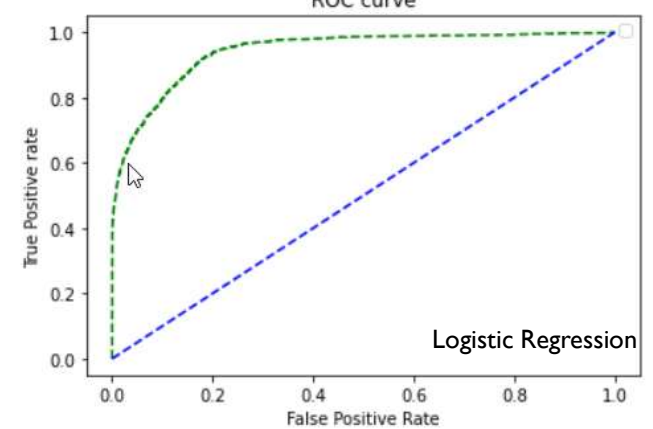
ROC curve



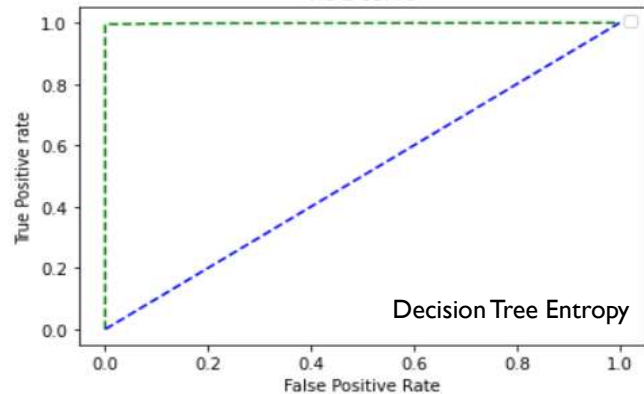
ROC curve



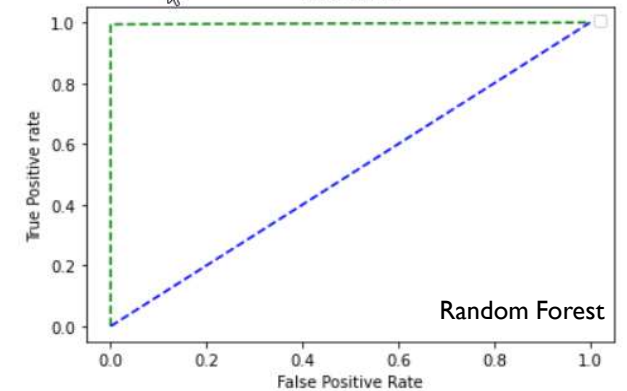
ROC curve



ROC curve

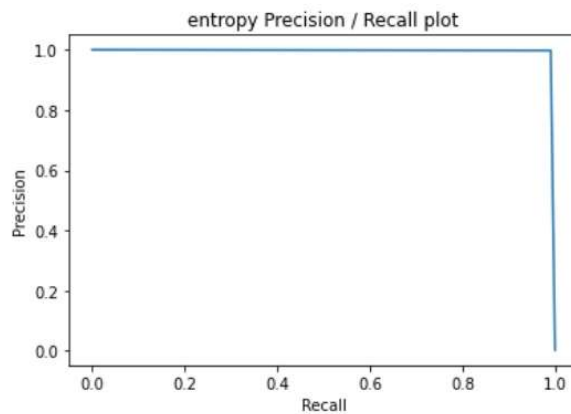
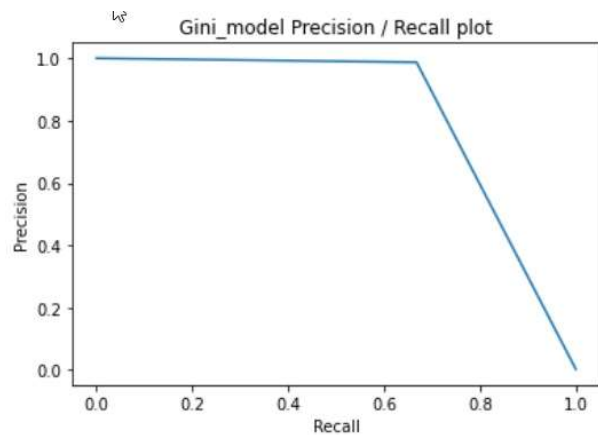
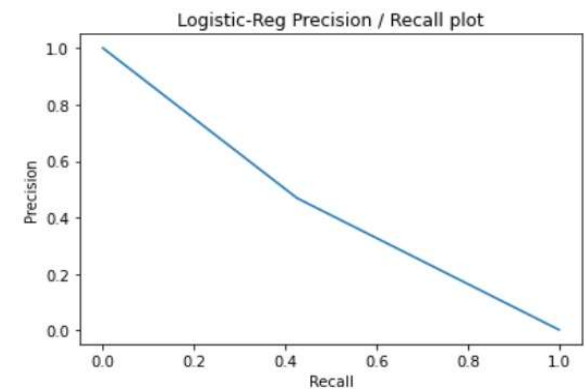
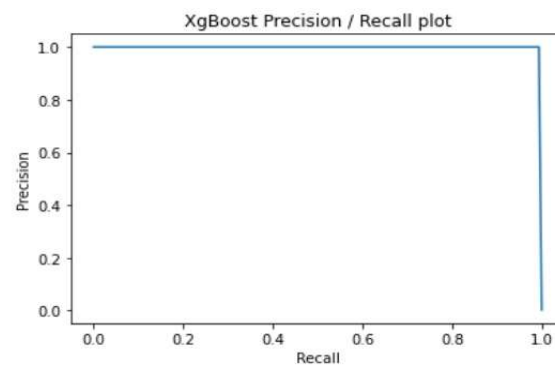
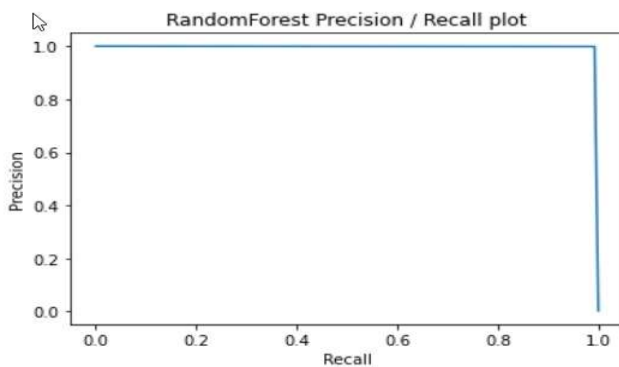


ROC curve



- Decision Tree Gini – 98.07%
- Decision Tree Entropy – 99.88%
- XgBoost – 99.83%
- Logistic Regression – 94.28%
- Random Forest – 99.69%

PRECISION/RECALL COMPARISON



- Decision Tree Gini – 98.72%
- Decision Tree Entropy – 99.75%
- XgBoost – 100%
- Logistic Regression – 46.88%
- Random Forest – 99.87%

CONCLUSIONS

- XgBoot Classifier Model provided the best result
- Reduced type-I error using Precision and F1- score
- Additional 5 features derived from the 11 features present in the initial data set
- Hyperparameter tuning techniques are performed to handle quantitative variables
- Logistic Regression used regularization tuning parameter 'C'
- Both Gini and Entropy model is performed in decision tree classification

Model Comparisons				
Model	Precision Score	F1-Score	Recall	ROC_AUC
Random Forest Classifier	99.87%	99.99	92.26%	99.69
XgBoost Classifier	100%	99.99%	99.38	99.83%
Logistic Regression	46.88%	99.68%	42.56%	94.28%
Decision Tree Gini Model	98.72%	99.89%	66.76%	98.07%
Decision Tree Entropy Model	99.75%	99.99%	99.07%	99.88%



THANK YOU

ARCHANA ROBIN