Loan Approval Prediction:

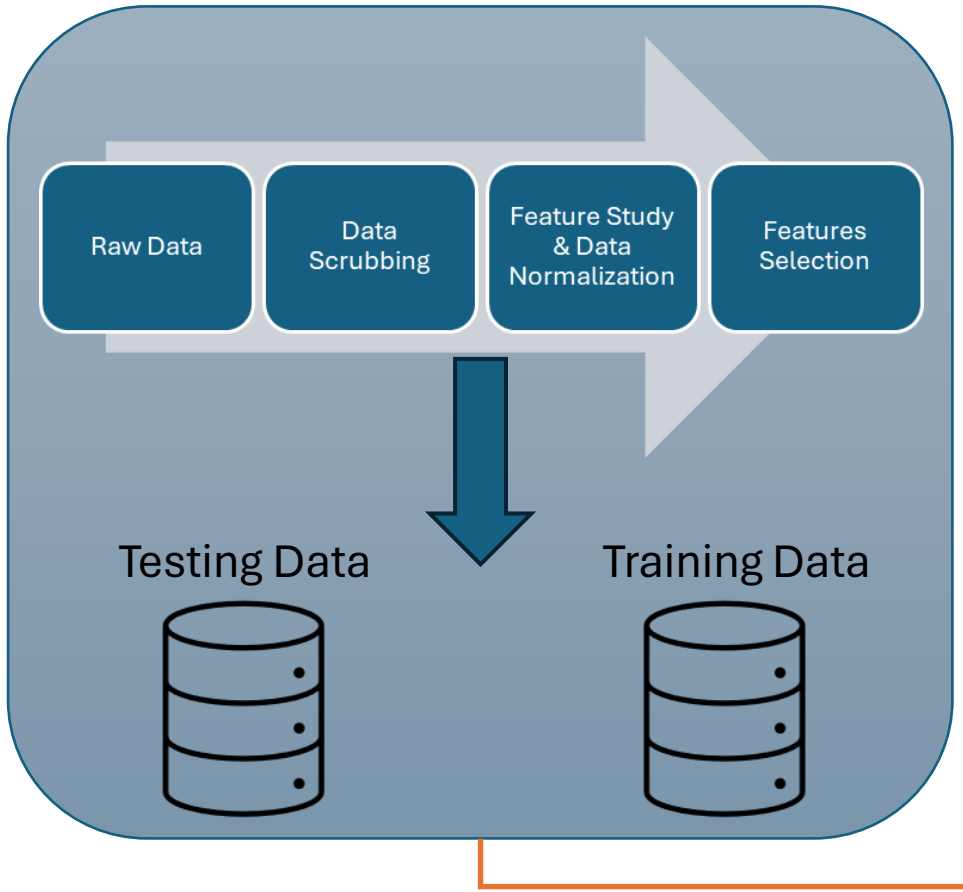
Best ML Model Comparison

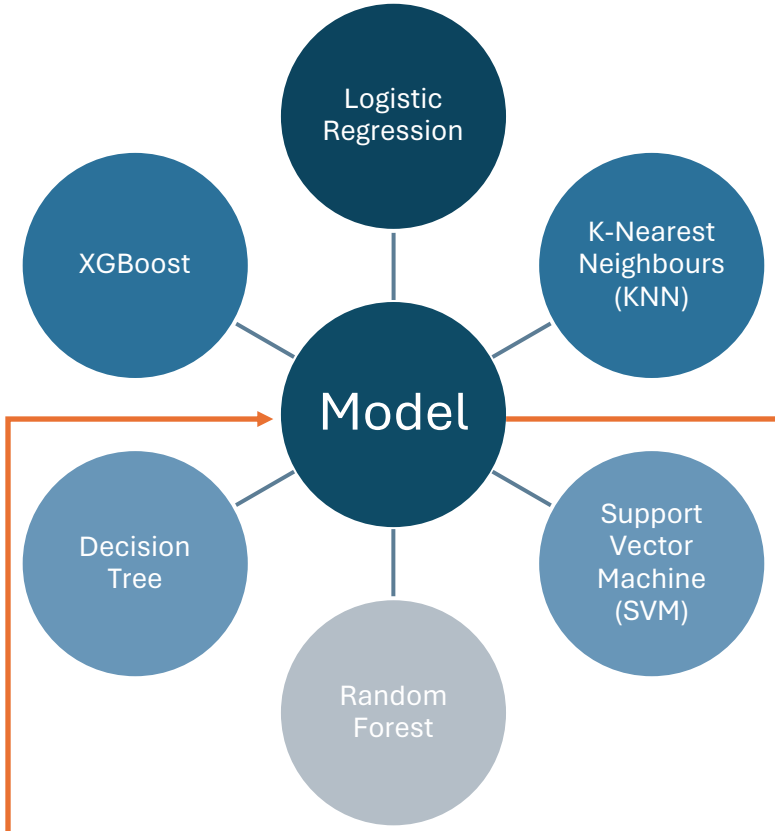
PROJECT OBJECTIVE

- To build the ML model which will make accurate predictions and will help Mumbai Trust Bank to streamline loan approval

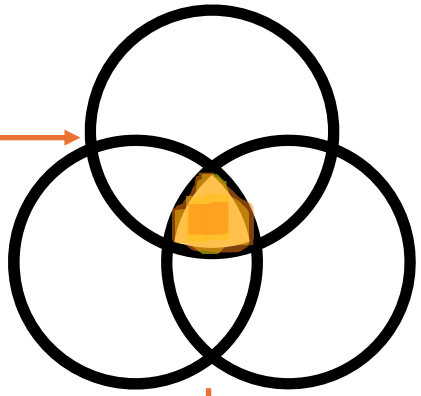
Mumbai Trust Bank

METHODOLOGY





Model Comparison



Model Selection

DATA SCRUBBING

4.5k records (individual loan decisions)

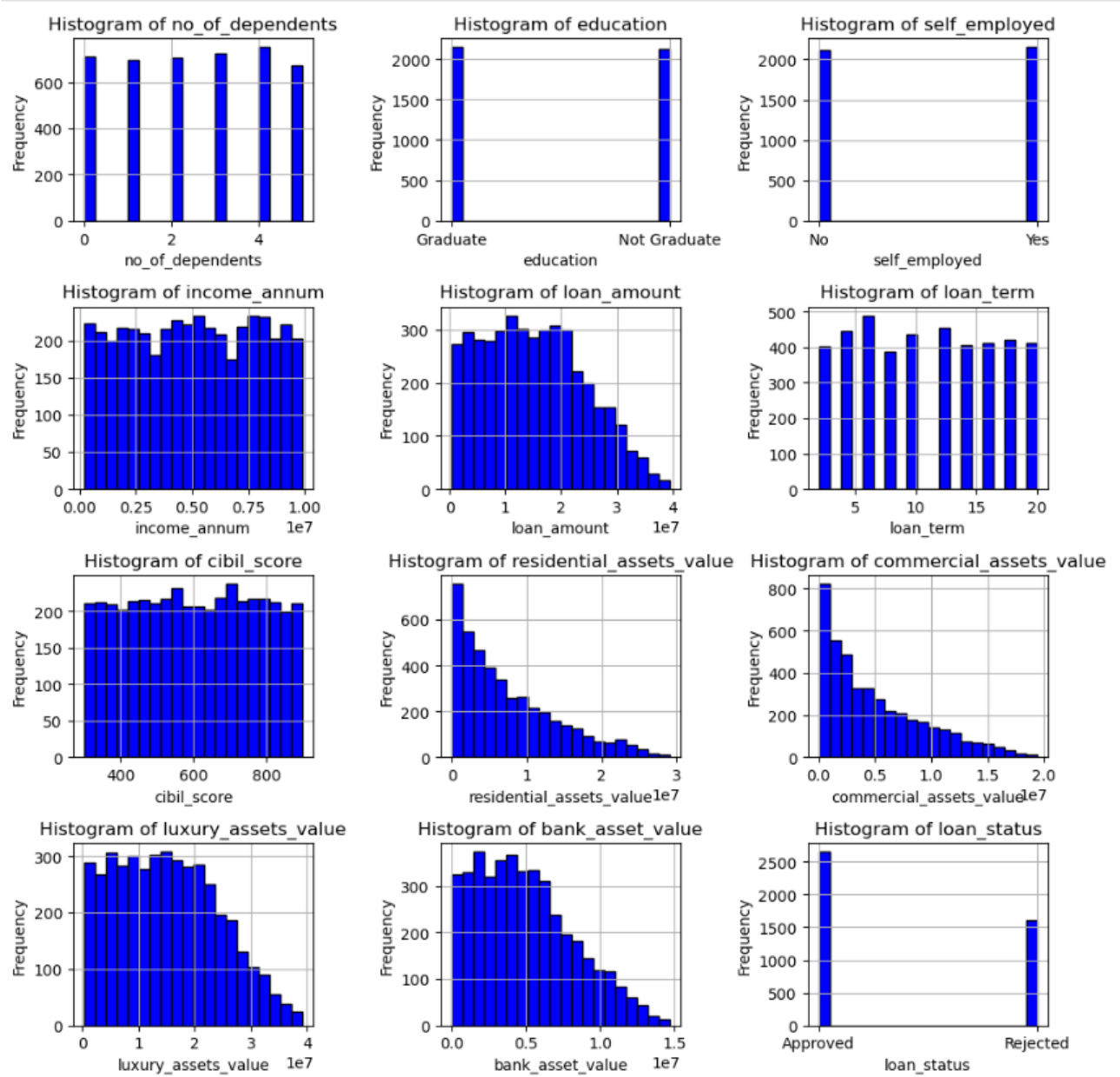
12 features, e.g. annual income, assets value, cibil score (credit score), loan amount, loan terms, etc.

Rectified some data formatting issues

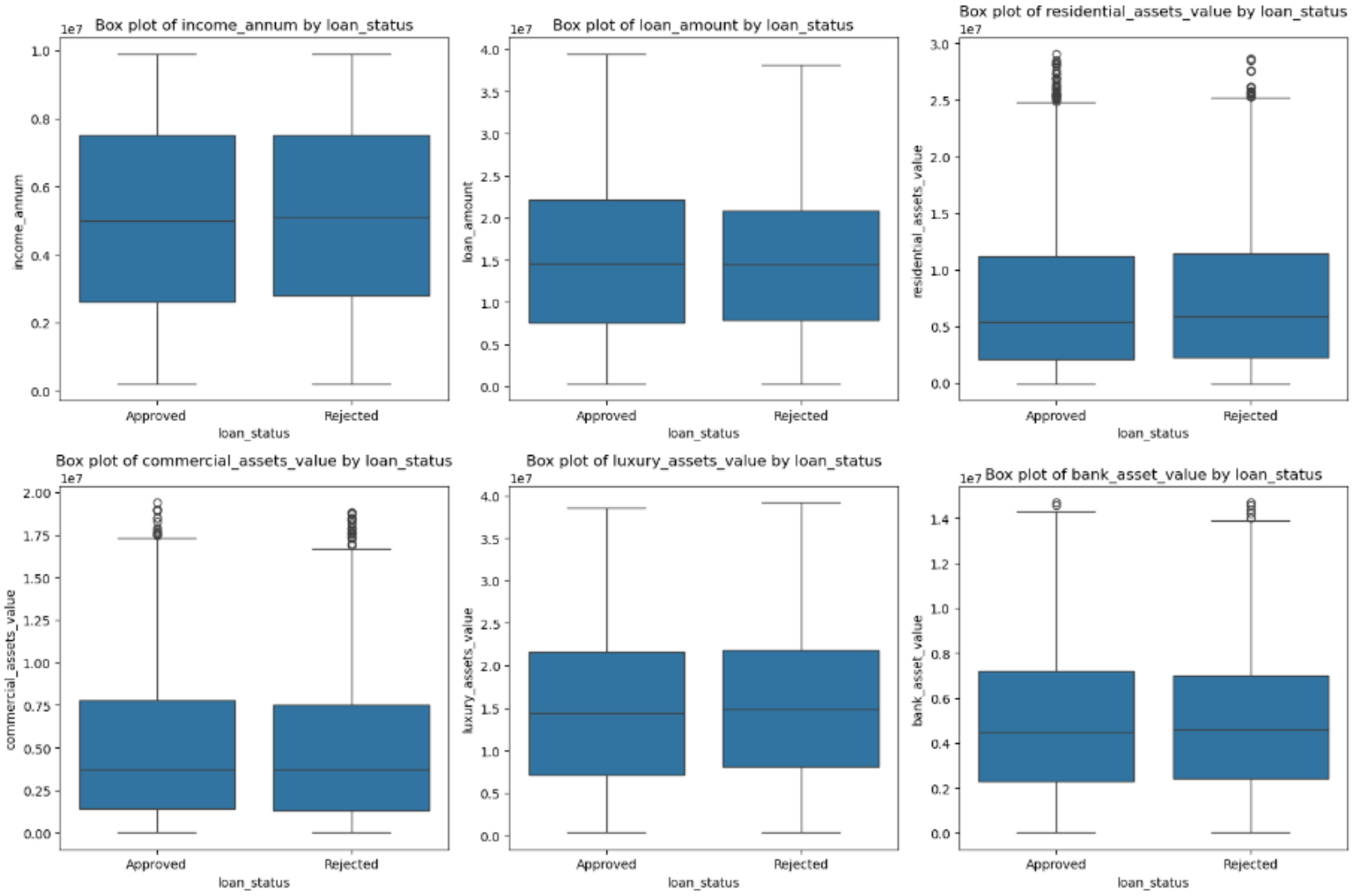
Prepared data frame for further analysis

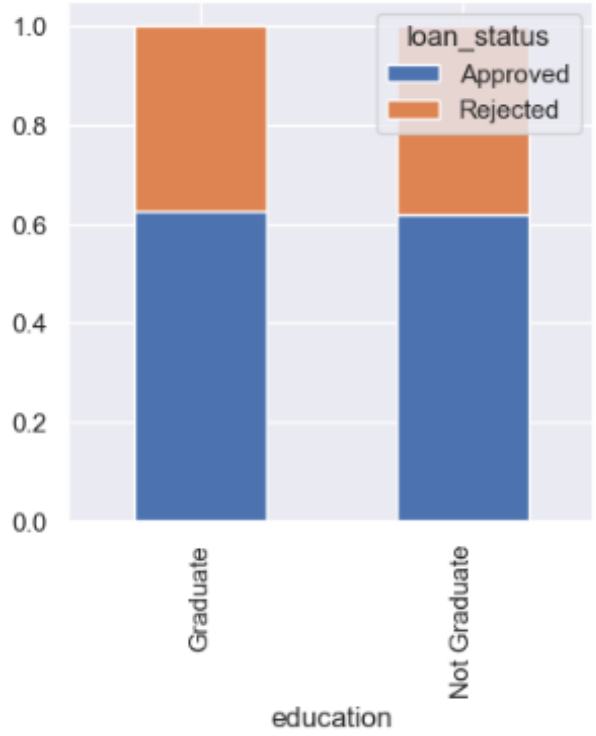
FEATURE STUDY & DATA NORMALIZATION

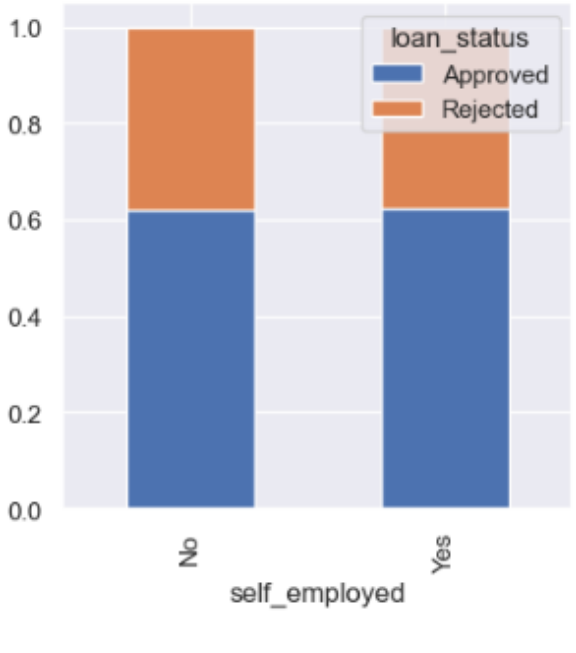
- Numerical and Categorical Features
- Category Split (education, employment, loan status)
- Outliers

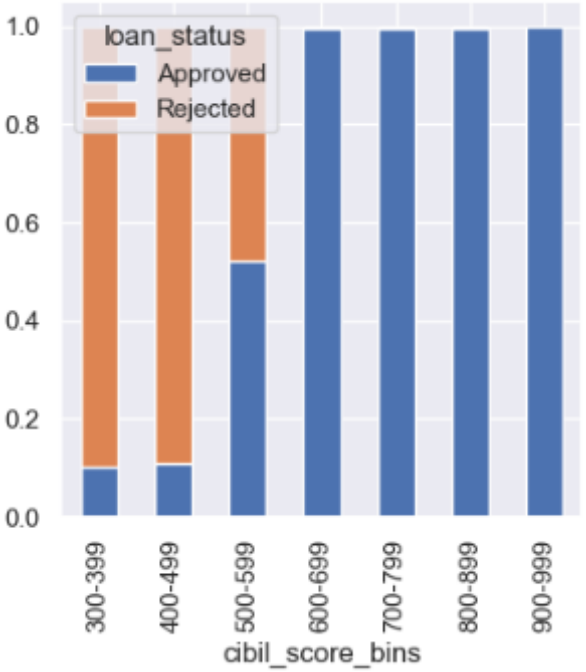


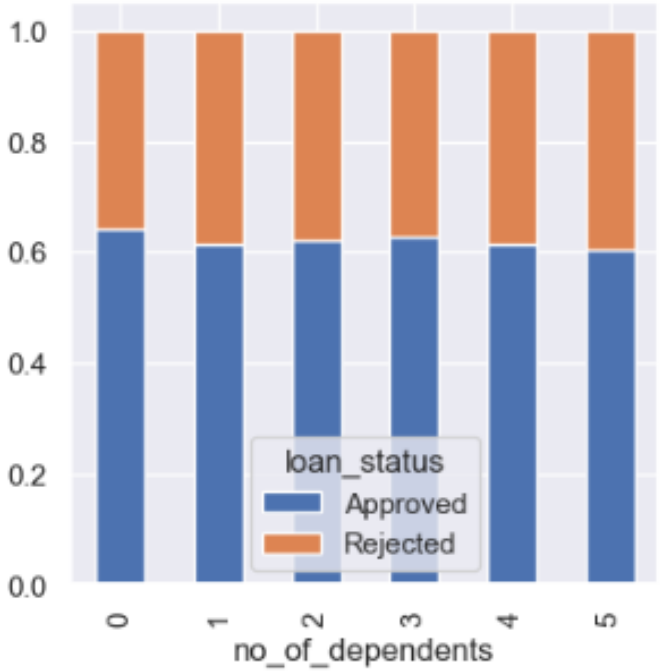
FEATURE STUDY & DATA NORMALIZATION

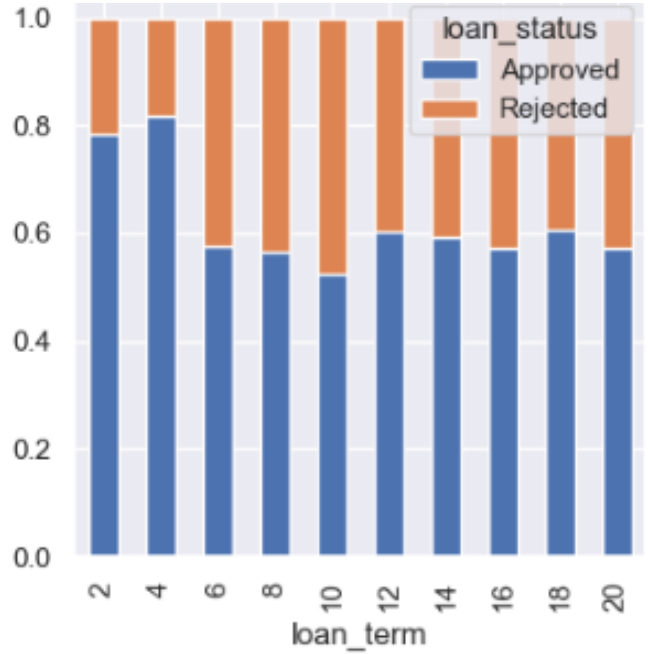






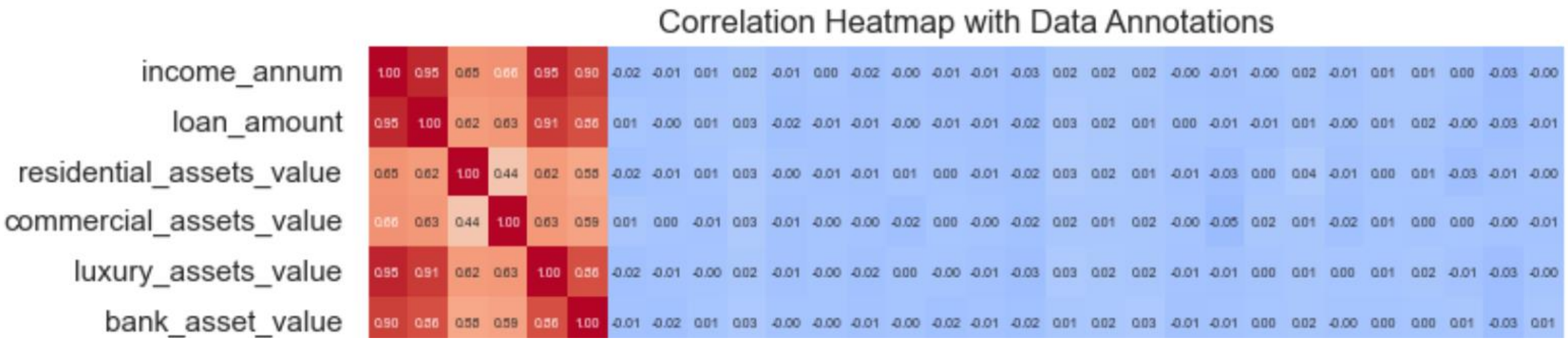






FEATURE STUDY & DATA NORMALIZATION

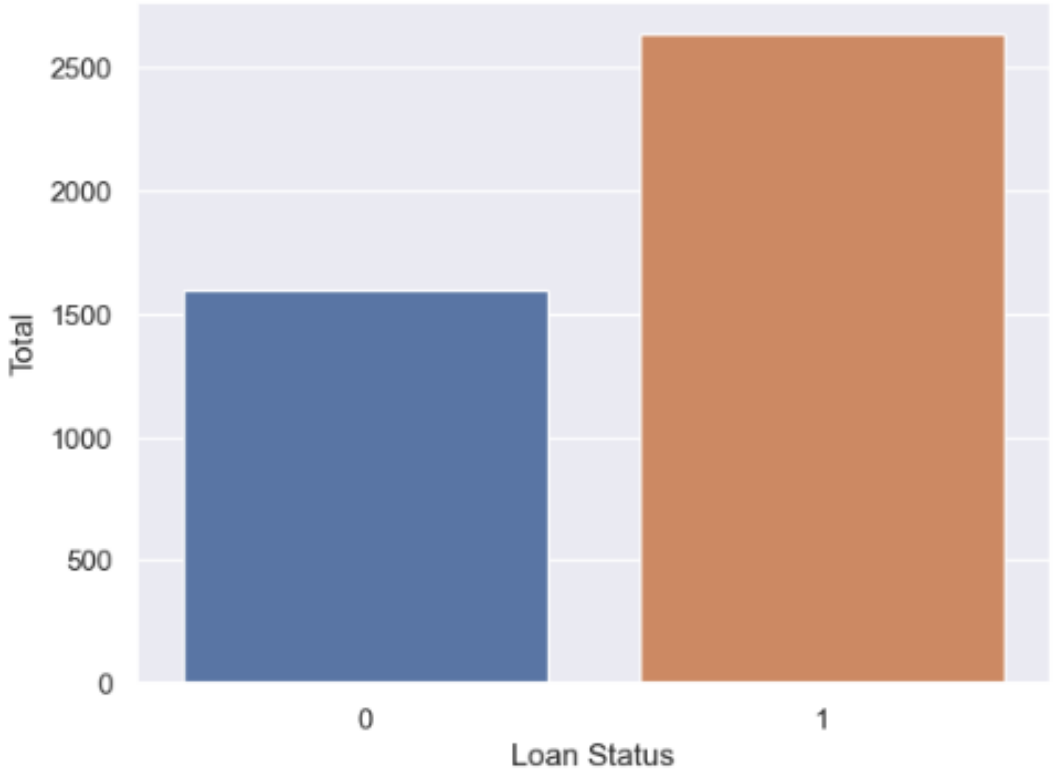
FEATURE STUDY & DATA NORMALIZATION



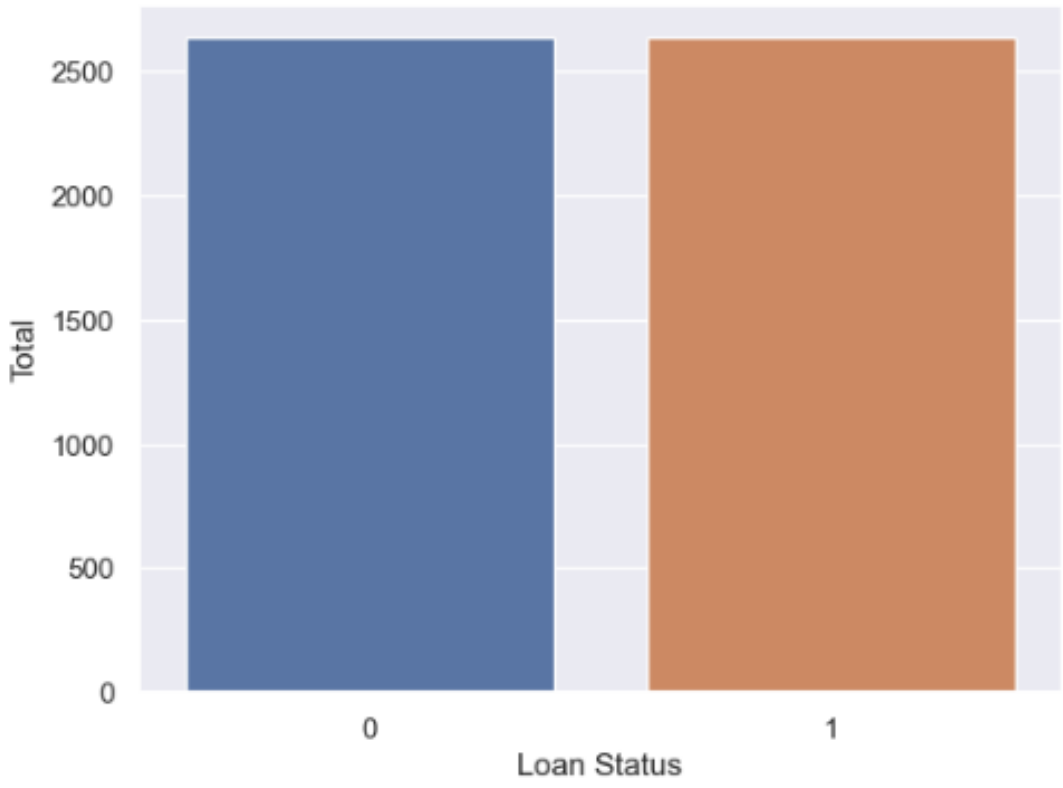
FEATURE STUDY & DATA NORMALIZATION

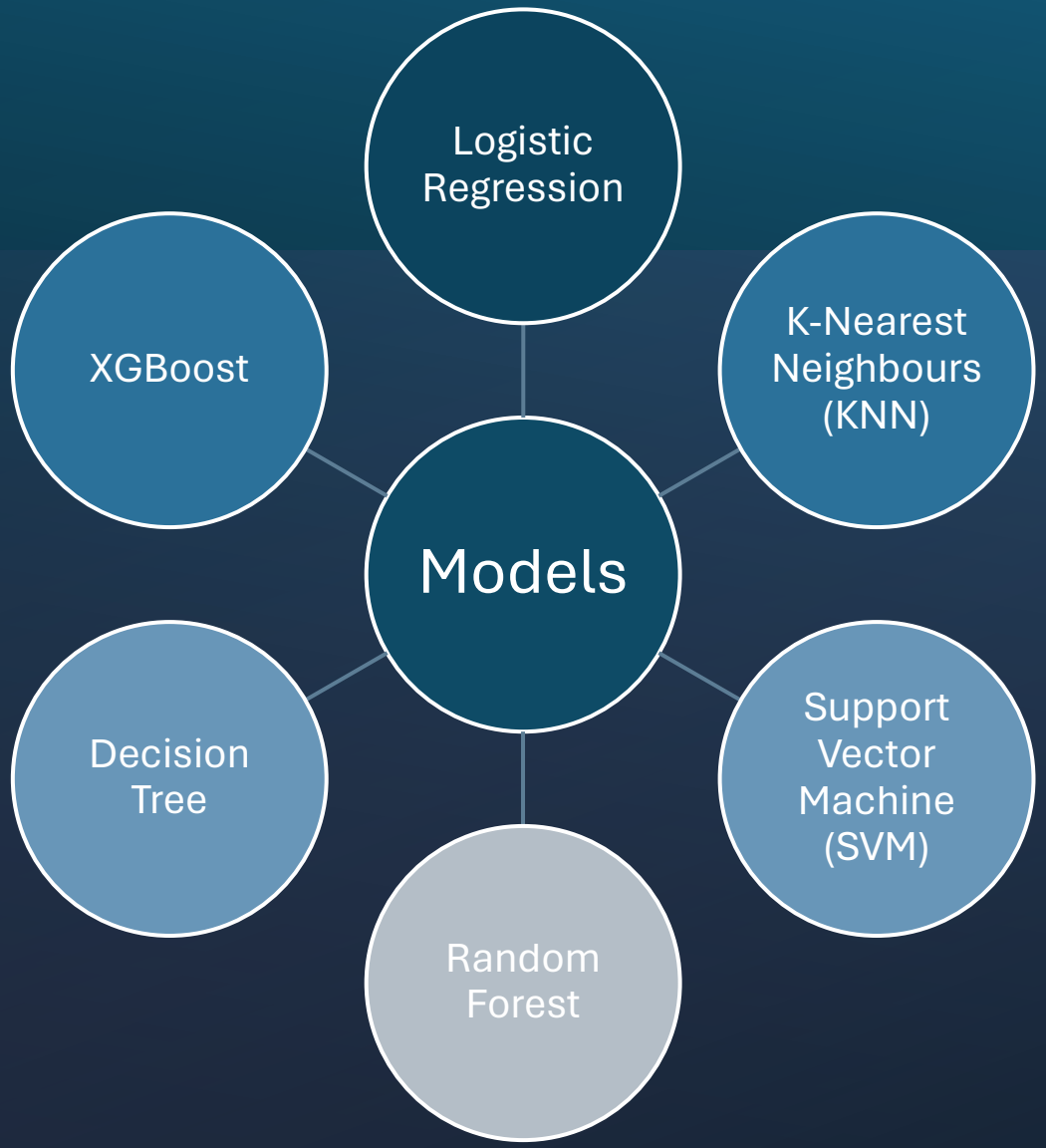
Synthetic Minority Over-sampling

Before SMOT



After SMOT




Logistic Regression
XGBoost
K-Nearest Neighbours (KNN)
Models
Decision Tree
Support Vector Machine (SVM)
Random Forest

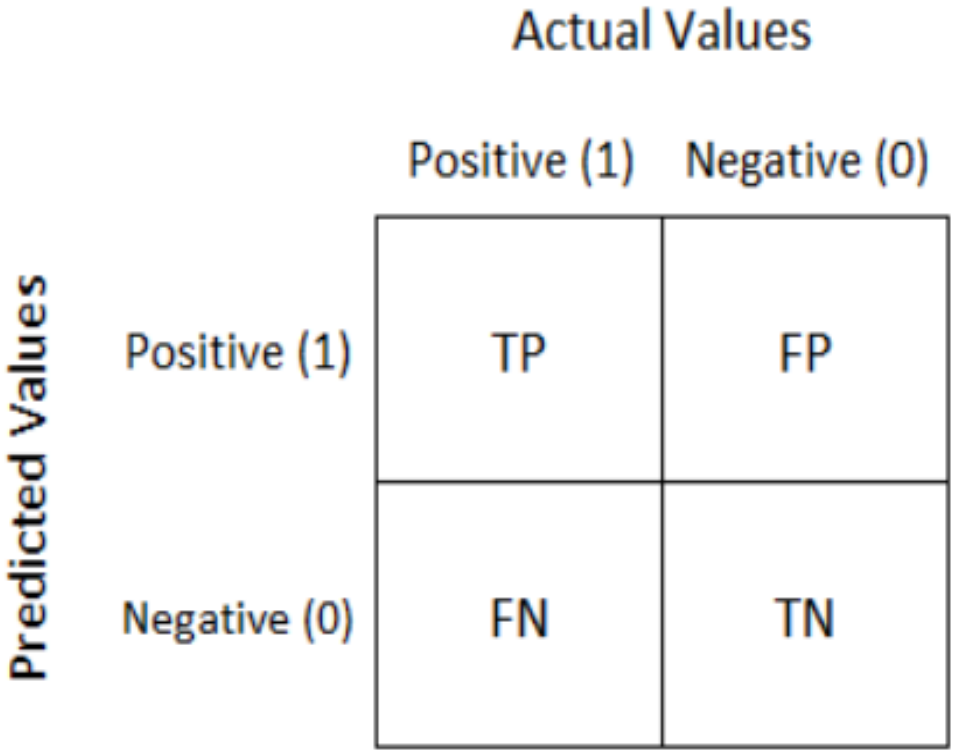
CONFUSION MATRIX

TP = True Positive: the model correctly predicts that a loan will be approved

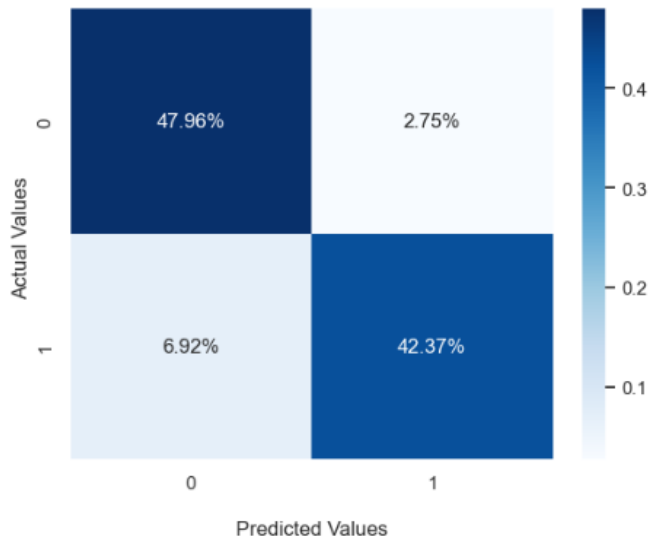
FP = False Positive: known as a Type I error. The model incorrectly predicts that a loan will be approved, but actually, the loan is rejected

TN = True Negative: the model correctly predicts that a loan will be rejected, and indeed the loan is rejected.

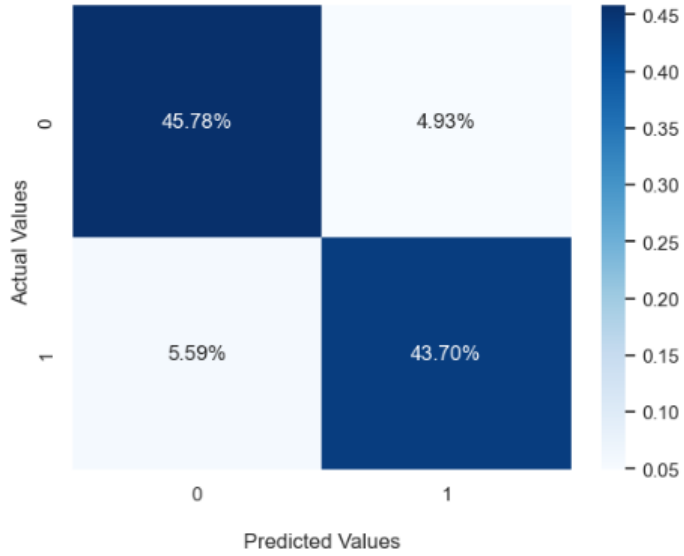
FN = False Negative: known as a Type II error. The model incorrectly predicts that a loan will be rejected, but actually, the loan should have been approved



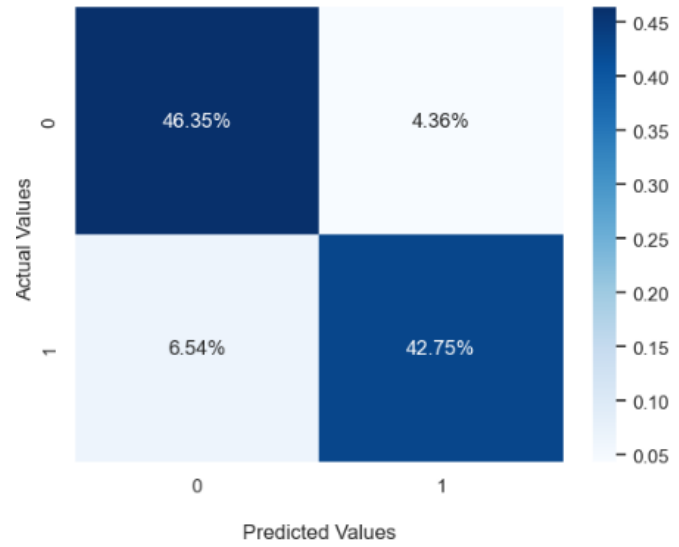
Support Vector Machine Confusion Matrix



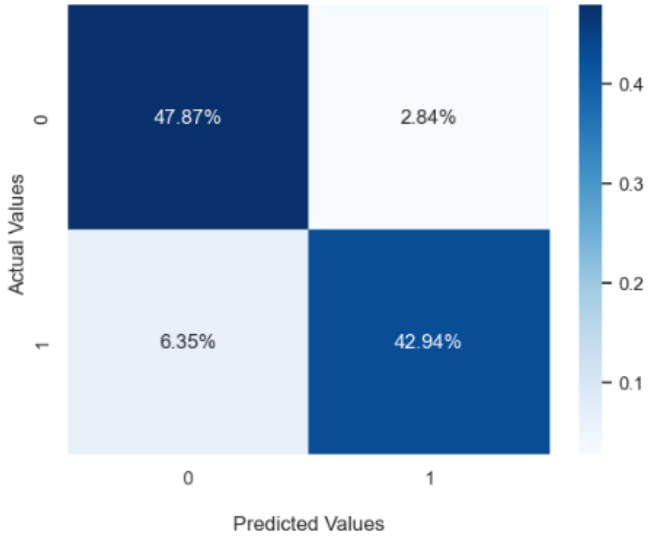
Decision Tree Confusion Matrix



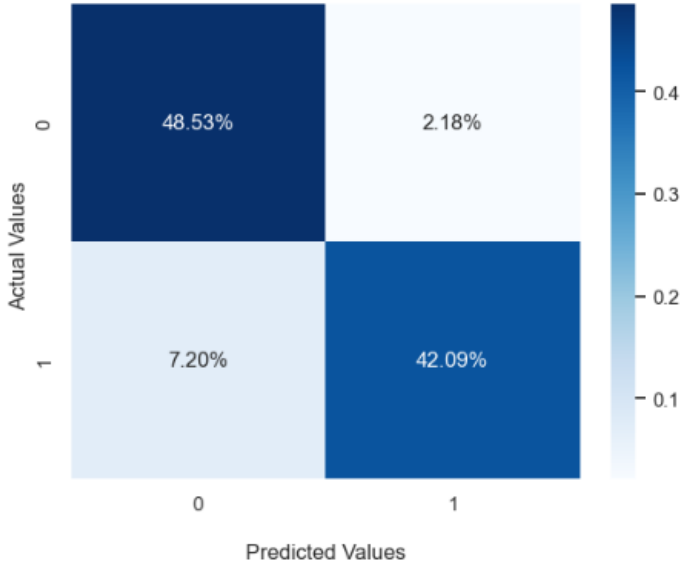
K-Nearest Neighbors Confusion Matrix



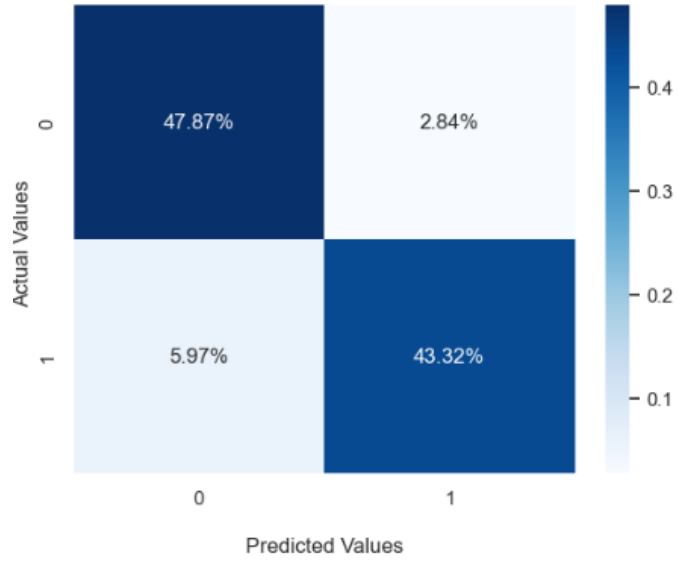
Random Forest Confusion Matrix



Logistic Regression Confusion Matrix



XGBoost Confusion Matrix



USEFUL METRIX

ACCURACY - Out of all the predictions we made, how many were true?

$$Accuracy = \frac{Number_of Correct Predictions}{Total Number_of Predictions} = \frac{TP + FN}{TP + FP + TN + FN}$$

PRECISION - Out of all the positive predictions we made, how many were true?

$$Precision/Positivepredictedvalue(P) = \frac{TP}{TP + FP}$$

RECALL - Out of all the data points that should be predicted as true, how many did we correctly predict as true?

$$Recall(R) = \frac{TP}{TP + FN}$$

F1 SCORE - F1 can therefore be used to measure how effectively our models make that trade-off

$$F - Score = \frac{2PR}{(P + R)}$$

MODEL COMPARISON

	Model	Accuracy	Precision Score	Recall Score	F1 Score
7	XGBoost	0.915	0.930	0.894	0.912
8	XGBoost with Tuning	0.913	0.942	0.877	0.908
4	Decision Tree with Tuning	0.907	0.912	0.898	0.905
0	Logistic Regression	0.906	0.949	0.856	0.900
6	Random Forest with Tuning	0.906	0.949	0.856	0.900
2	SVM	0.903	0.945	0.854	0.897
5	Random Forest	0.903	0.941	0.858	0.897
1	K Neighbors	0.894	0.905	0.877	0.891
3	Decision Tree	0.883	0.875	0.890	0.883

```
Max Accuracy: 0.915 (Model: XGBoost)
Max Precision Score: 0.949 (Model: Logistic Regression)
Max Recall Score: 0.898 (Model: Decision Tree with Tuning)
Max F1 Score: 0.912 (Model: XGBoost)
```

THANK YOU!

Email: grigorenko.jane@gmail.com

GitHub: https://github.com/JaneGrig

LinkedIn: linkedin.com/in/evgeniya-jane-grigorenko-47692211b