id/x partners

CREDIT RISK PREDICTION WITH MACHINE LEARNING

04 DECEMBER 2023

JULIAN SAPUTRA

ABOUT ME

Hi, my name is Julian. I am an undergraduate student majoring in computer science with over 2 years of hands-on experience in data science and intelligent systems. Currently serving as a Freelance Data Science Research Assistant to a Ph.D. Candidate, I actively contribute to cutting-edge research. I am actively seeking opportunities in roles such as data scientist, data analyst, machine learning engineer, artificial intelligence engineer, business intelligence, and business analyst.

BACKGROUND

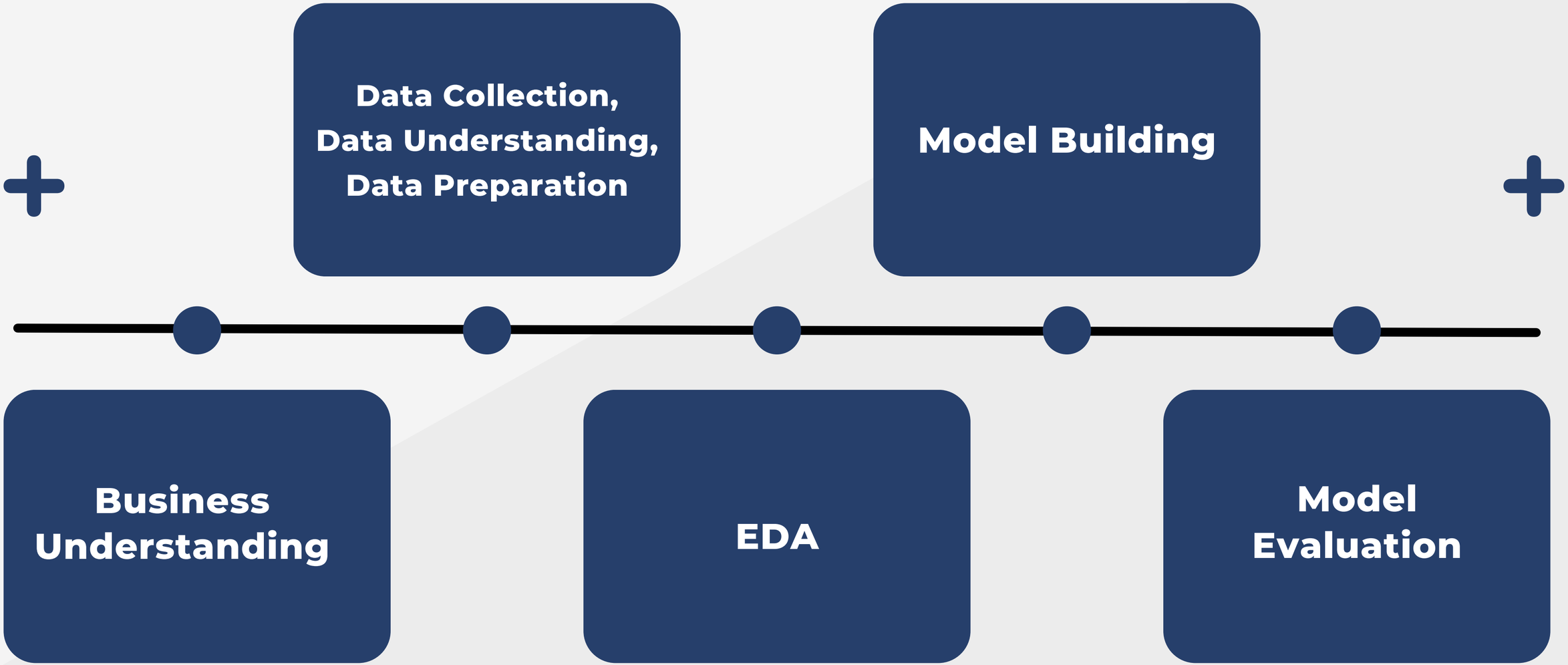
A lending company is facing significant challenges in its credit decision-making process due to the continued reliance on manual analysis to identify potential credit risks. This manual method not only consumes time but is also susceptible to errors and uncertainties, leading to inaccurate credit decisions and a higher risk of payment defaults.

As an intern Data Scientist at ID/X Partners, we are engaged in this project to provide a solution that enhances efficiency and accuracy in the credit decision-making process. We are employing various analytical approaches, including descriptive, exploratory, diagnostic, and predictive analyses using machine learning models.

PROBLEM STATEMENT

- The challenge at hand involves the inefficiency of manual assessment, which demands a considerable investment of time. This is particularly problematic as, among a total of 460 thousand customers, 11% are currently encountering difficulties with payment defaults. The need for a more streamlined and timely assessment process is evident to address and mitigate these issues effectively.

END-TO-END SOLUTION



BUSINESS UNDERSTANDING

- Credit risk is the potential for financial loss: It occurs when borrowers fail to meet payment obligations, posing a threat to a business's revenue and financial stability.
- Assessing credit risk involves evaluating creditworthiness: Lenders analyze factors like credit history and financial stability to gauge the likelihood of a borrower defaulting, aiding in informed credit decisions.
- Credit risk affects profitability and liquidity: Striking a balance between extending credit for sales and mitigating the risk of non-payment is crucial, as it directly impacts a business's profits and cash flow.

DATA COLLECTION

- The data provides by company, comprising information on both accepted and rejected loan applications.
- The dataset is composed of a single data file in CSV format, accompanied by a data dictionary in XLSX format.
- The data dictionary provides details about each column present in the data file, offering insights into the information contained within the dataset.

Features	Description
int_rate	Indicates if income was verified by LC, not verified, or if the income source was verified
issue_d	The month which the loan was funded
loan_amnt	Last month payment was received
last_pymnt_amnt	Last total payment amount received
loan_status	Current status of the loan
next_pymnt_d	Next scheduled payment date
recoveries	Indicates if a payment plan has been put in place for the loan
etc.	Other features

DATA UNDERSTANDING & PREPARATION

- Num of features = 74
- Num of rows = 466285
- There are 17 features that have all missing values in the rows of the data
- There are 52 numerical features and 22 object features
- Considered to choose the 'loan_status' feature as the target for credit risk classification
- This step is about the first preprocessing data to ensure that the data is ready for EDA

Univariate Analysis

Examining one variable at a time to understand its distribution and characteristics.

Bivariate Analysis

Analyzing the relationship between two variables to uncover patterns or correlations.

UNIVARIATE ANALYSIS

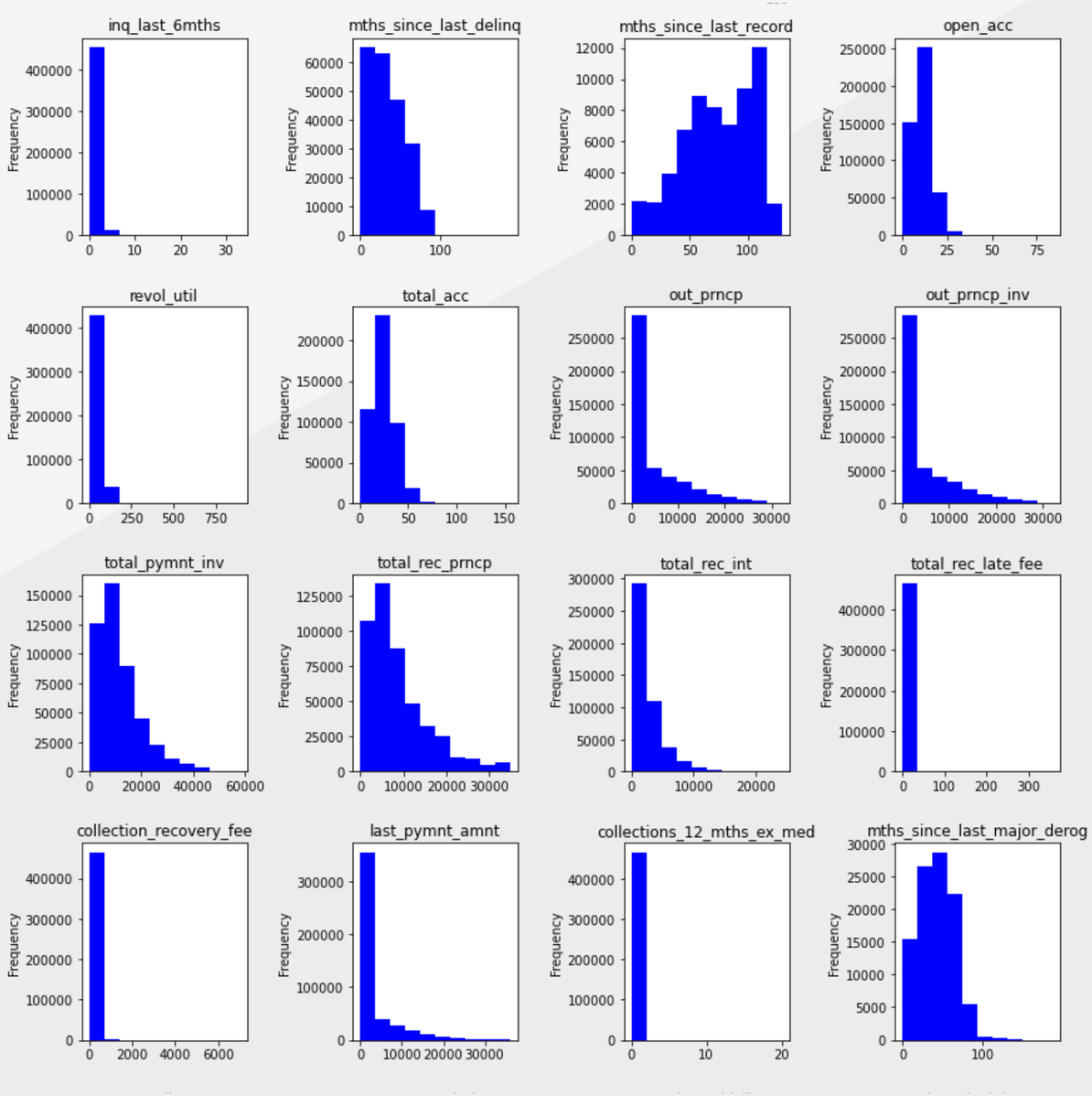
Categorical variables

INSIGHTS

- Feature: 36 months for loan term: Indicates a preference for shorter payment periods, possibly reflecting a desire to pay off the loan quickly and reduce interest payments.
- Grade Feature: Dominated by B: Suggests a significant portion of borrowers have a good credit score, lowering the risk for lenders.
- Emp_length Feature: Dominated by 10+ years: Implies that borrowers with longer employment histories are prevalent, potentially signaling increased financial stability.
- Home_ownership Feature: Dominated by MORTGAGE: Indicates that most borrowers with registered home ownership have a mortgage, offering insights into their financial stability.
- Verification_status Feature: Dominated by Verified: Indicates that a significant number of borrowers have their co-borrowers' joint income verified, adding credibility to income information.
- Loan_status (target feature): Dominated by good_loan: Implies that the majority of loans in the dataset are successfully meeting payment obligations, showcasing overall portfolio health.
- Purpose Feature: Dominated by debt_consolidation: Suggests that many borrowers seek loans for debt consolidation, potentially for better rates or more manageable terms.
- Initial_list_status: Dominated by f (fractional listing): Indicates that most loans are initially listed as fractional, allowing multiple investors to fund a single loan and spreading risk across lenders.

UNIVARIATE ANALYSIS

Numerical variables



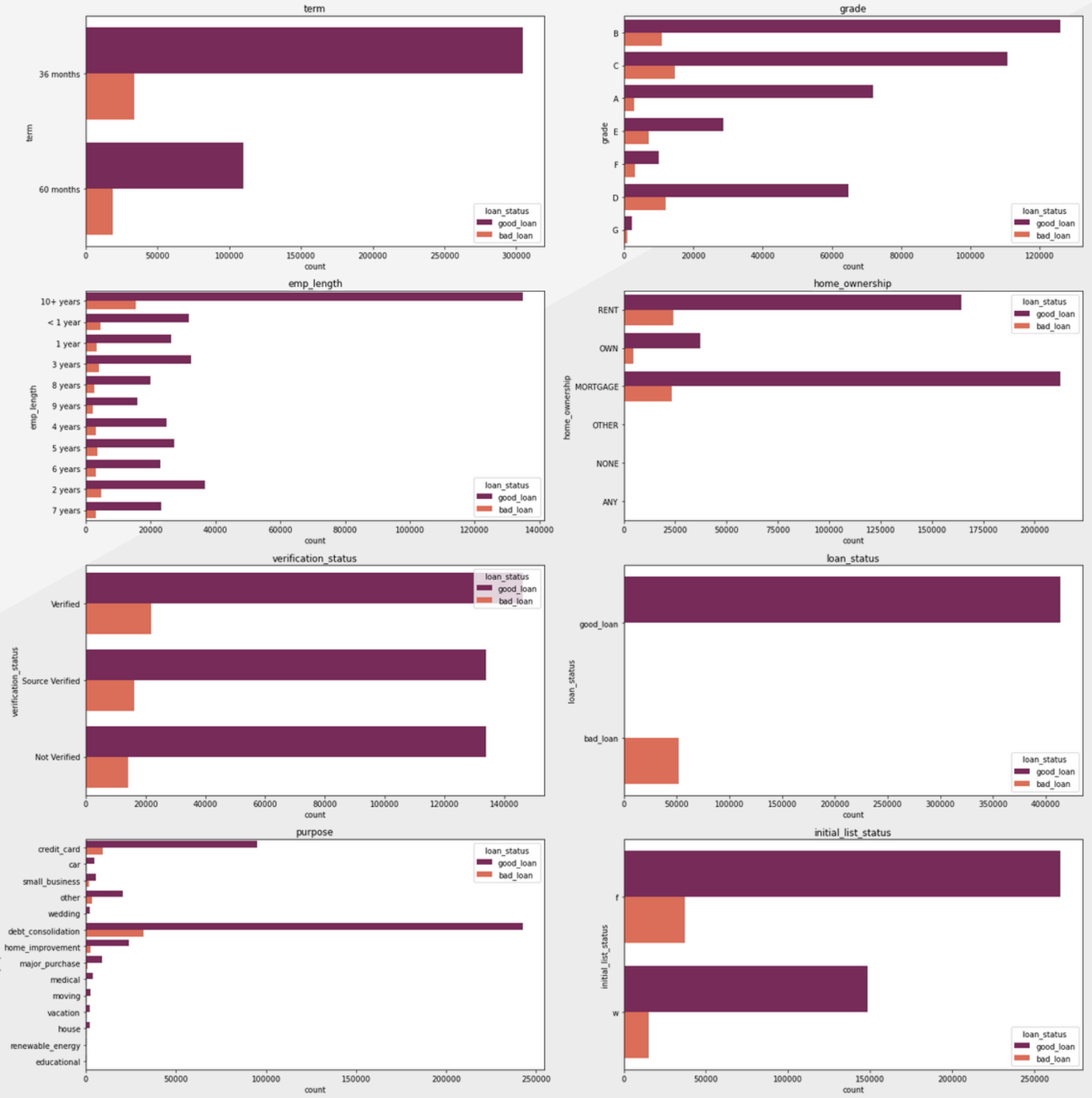
INSIGHTS

- The histogram analysis reveals that many features exhibit a normal distribution, indicating a balanced spread of values. However, there are several features showcasing a right-skewed distribution. Understanding these distribution patterns provides valuable insights into determining appropriate strategies for handling missing values.

- In cases where the data follows a normal distribution, filling missing values with the mean is a reasonable consideration. This approach leverages the central tendency of the data. On the other hand, for features exhibiting a right-skewed or left-skewed distribution, filling missing values with the median is a prudent alternative. The median, being less sensitive to extreme values, aligns well with the skewed nature of the data, offering a robust solution for handling missing values in such scenarios.

BIVARIATE ANALYSIS

Analysis is conducted on categorical features, and the coloration or distinction is based on the different values within the 'loan_status' variable.



DATA PREPROCESSING

Handling Missing Values

Feature Selection

Feature Encoding

Oversampling Data

Train & Test Split

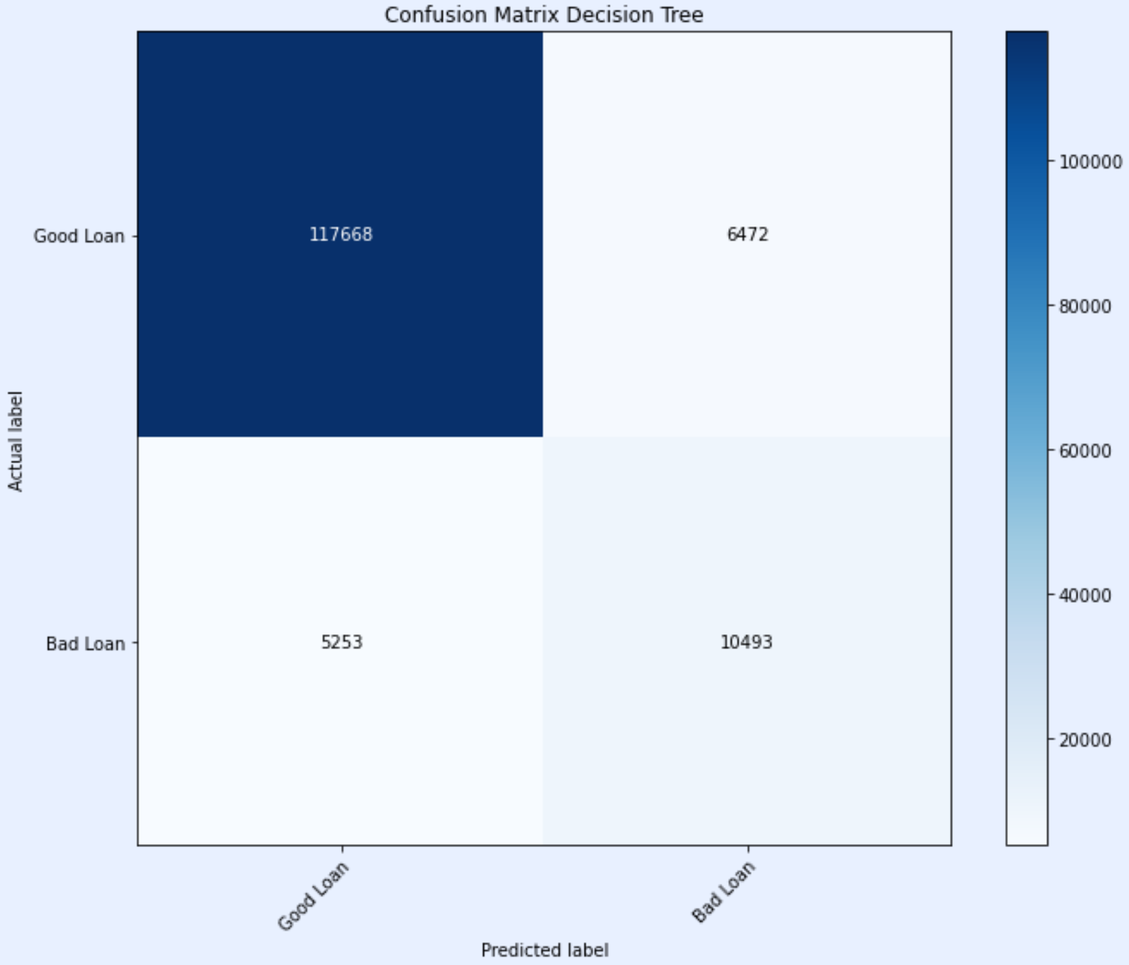
PREDICTIVE MODELING

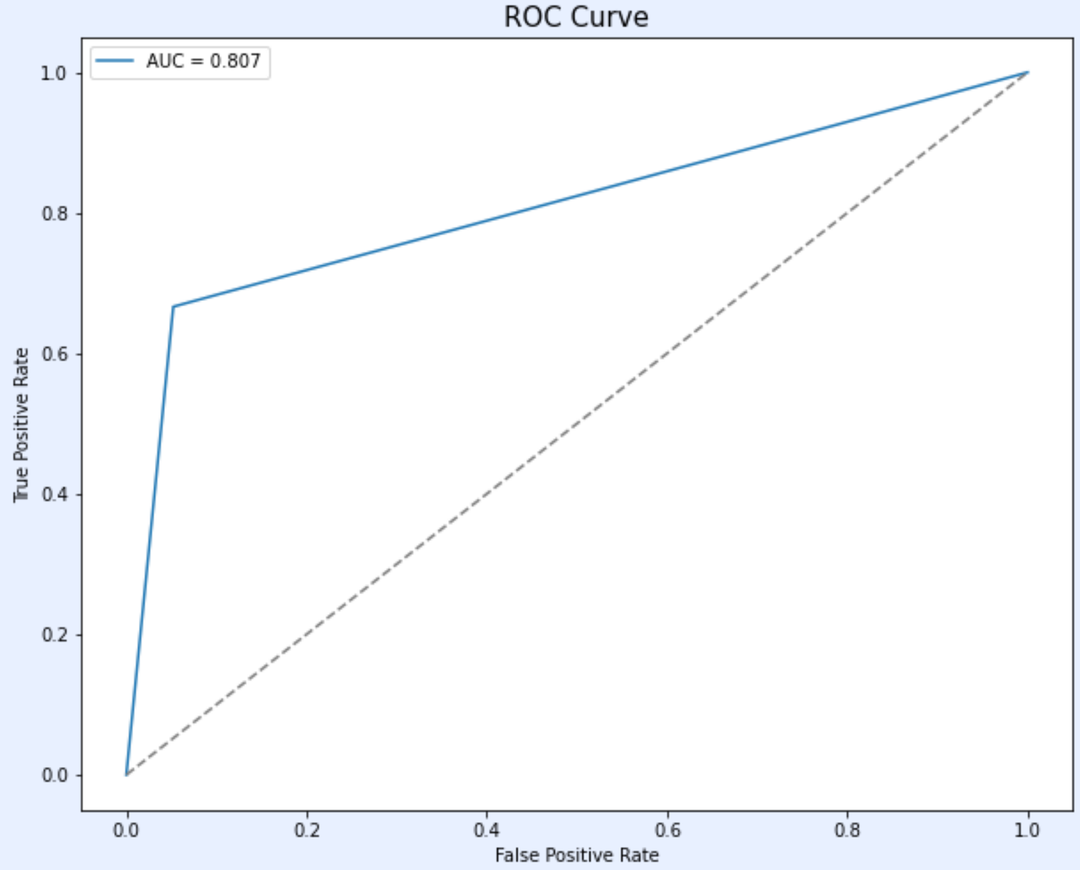
- Decision Tree Classifier
- Random Forest Classifier
- Logistic Regression
- XGBoost Classifier
- AdaBoost Classifier
- Bagging Classifier
- Stacking Classifier
- Decision Tree Classifier (Hyperparameter Tuning)

MODEL EVALUATION

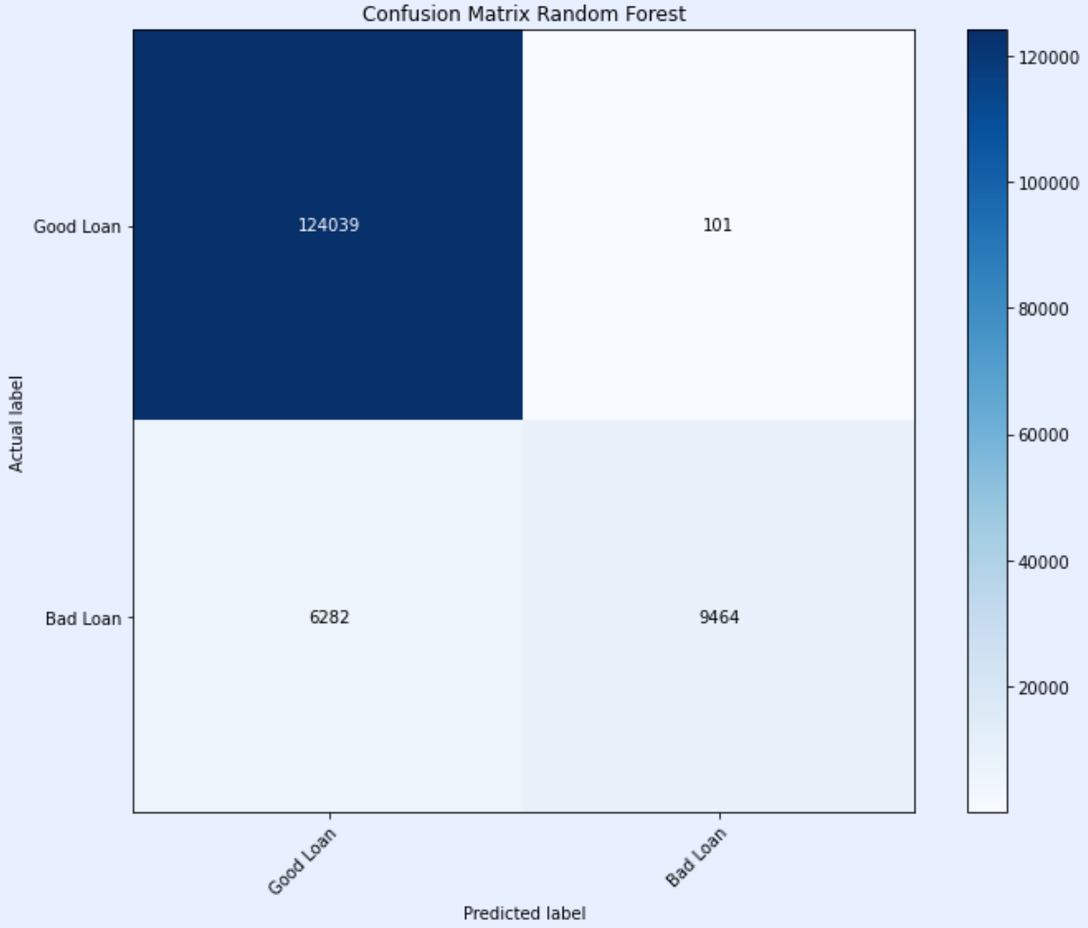
	Model	Akurasi	Precision	Recall	F1-Score	AUC-Proba Train	AUC-Proba Test
	Decision Tree Classifier	92%	79%	81%	80%	100%	80%
	Random Forest Classifer	95%	97%	80%	86%	100%	92%
	Logistic Regression	93%	90%	74%	79%	85%	79%
	XGBoost Classifer	95%	96%	80%	86%	99%	93%
	Adaboost Classifer	95%	94%	81%	86%	98%	92%
	Bagging Classifer	95%	95%	80%	86%	99%	89%
	Stacking Classifer	93%	84%	78%	81%	98%	77%
	Decision Tree Classifer (Hyperparameter Tuning)	95%	94%	79%	85%	98%	89%
	Decision Tree Classifier	92%	79%	81%	80%	100%	80%

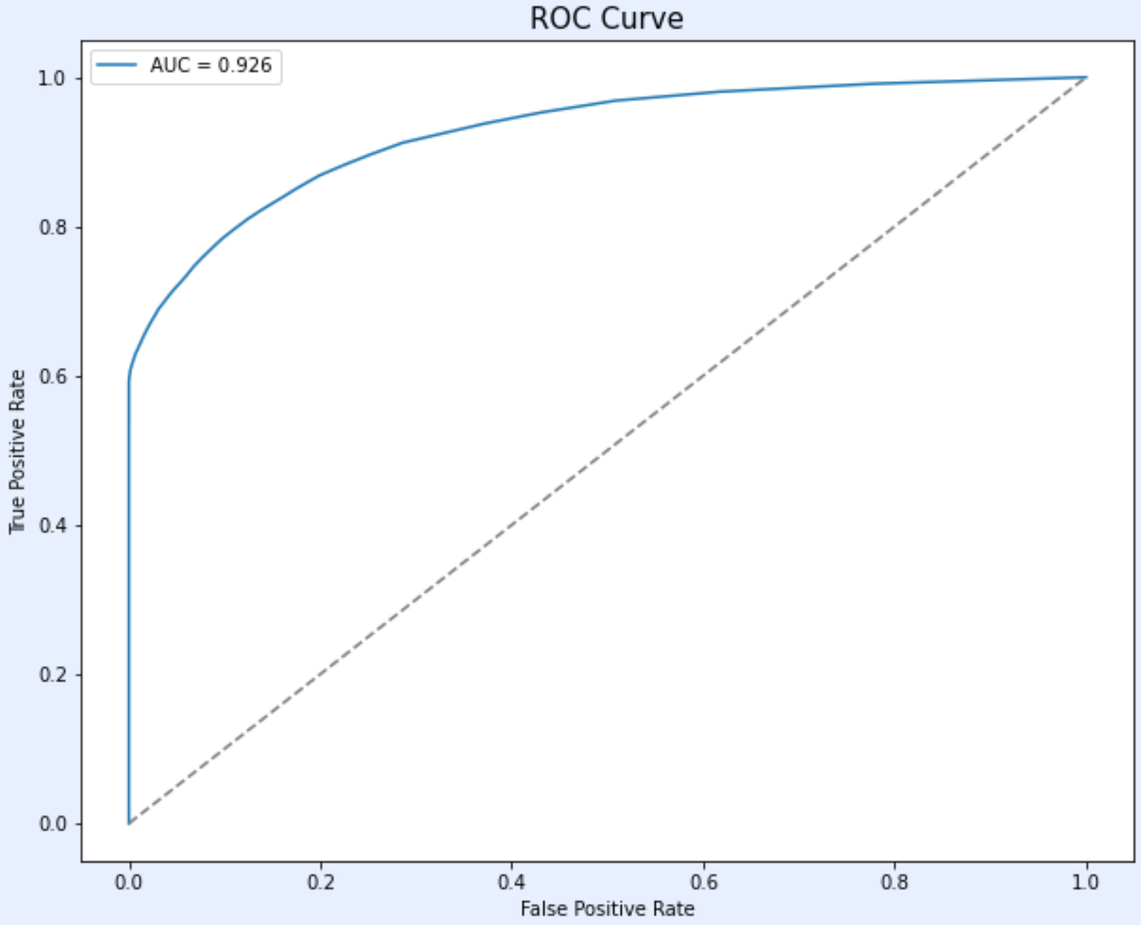
DECISION TREE CLASSIFIER



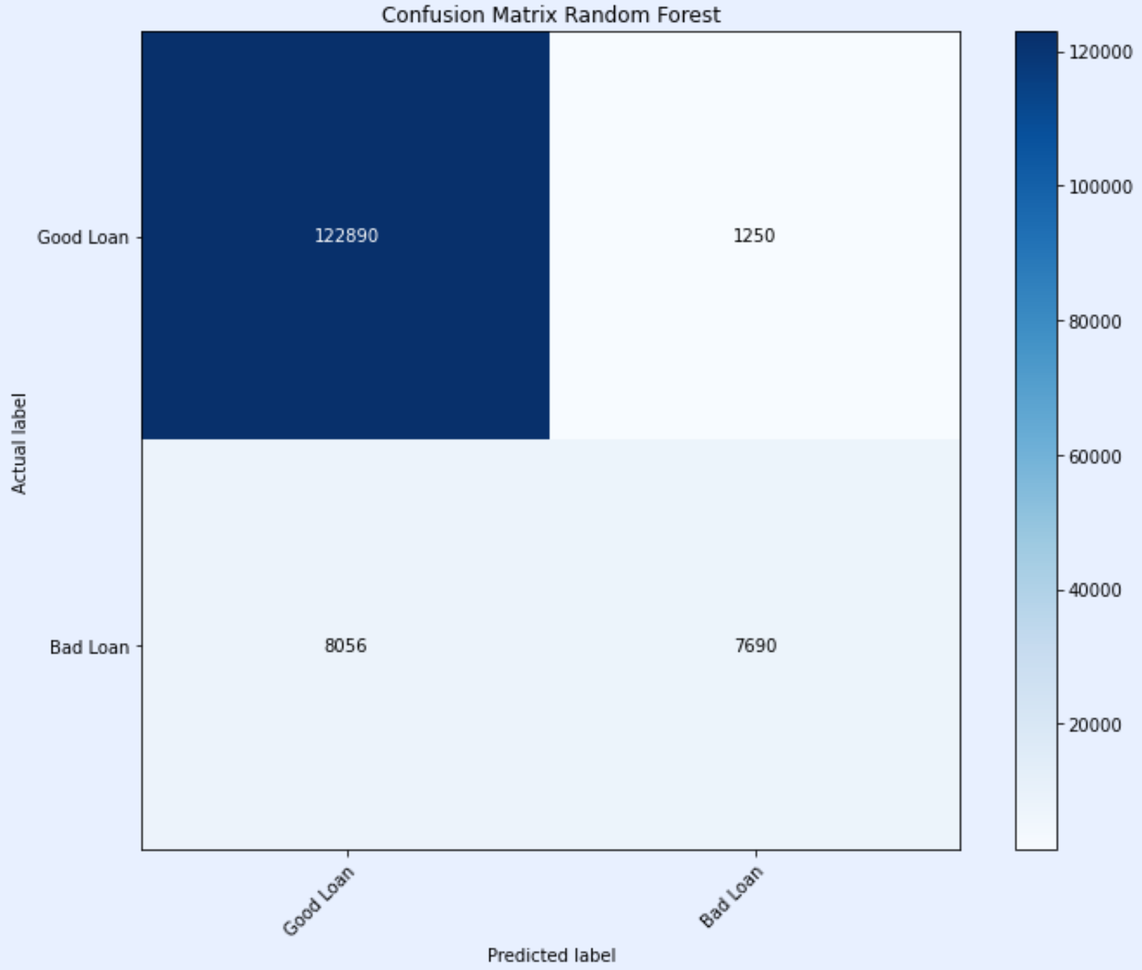


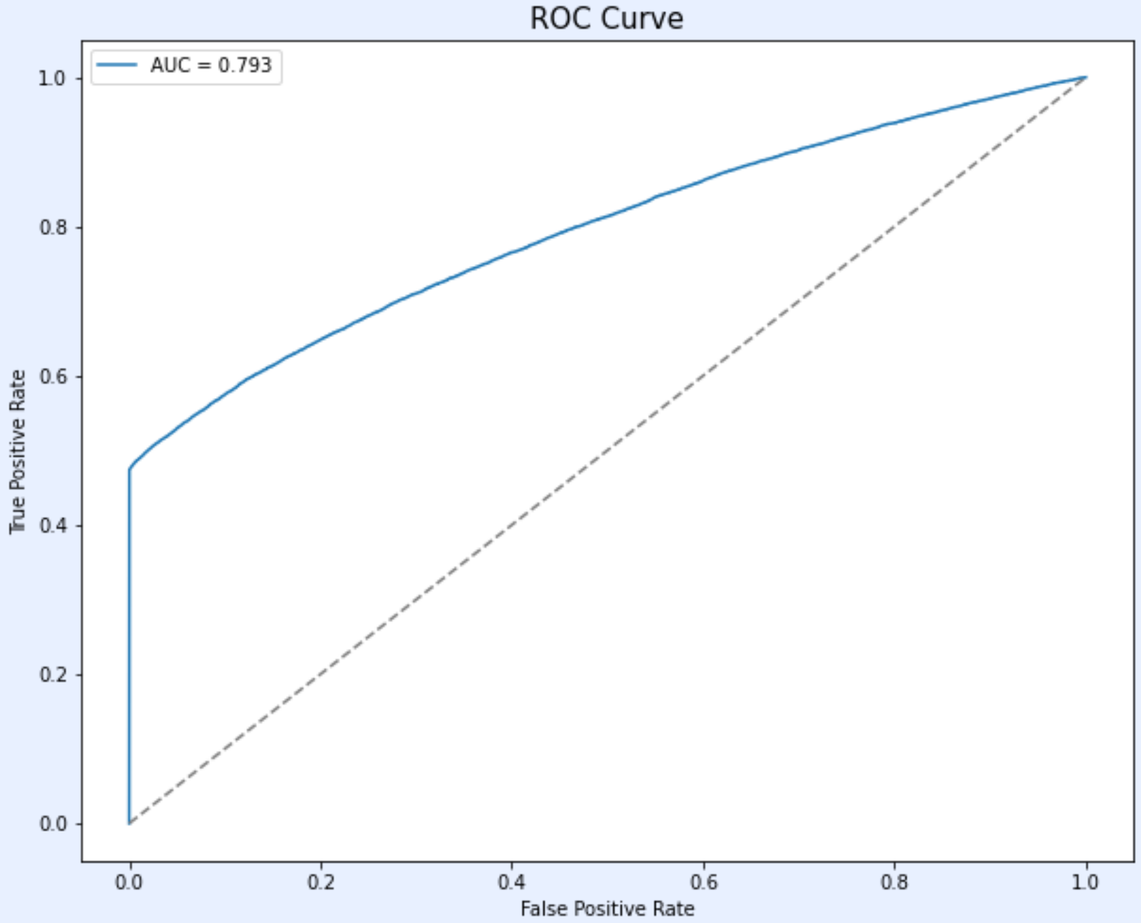
RANDOM FOREST CLASSIFIER



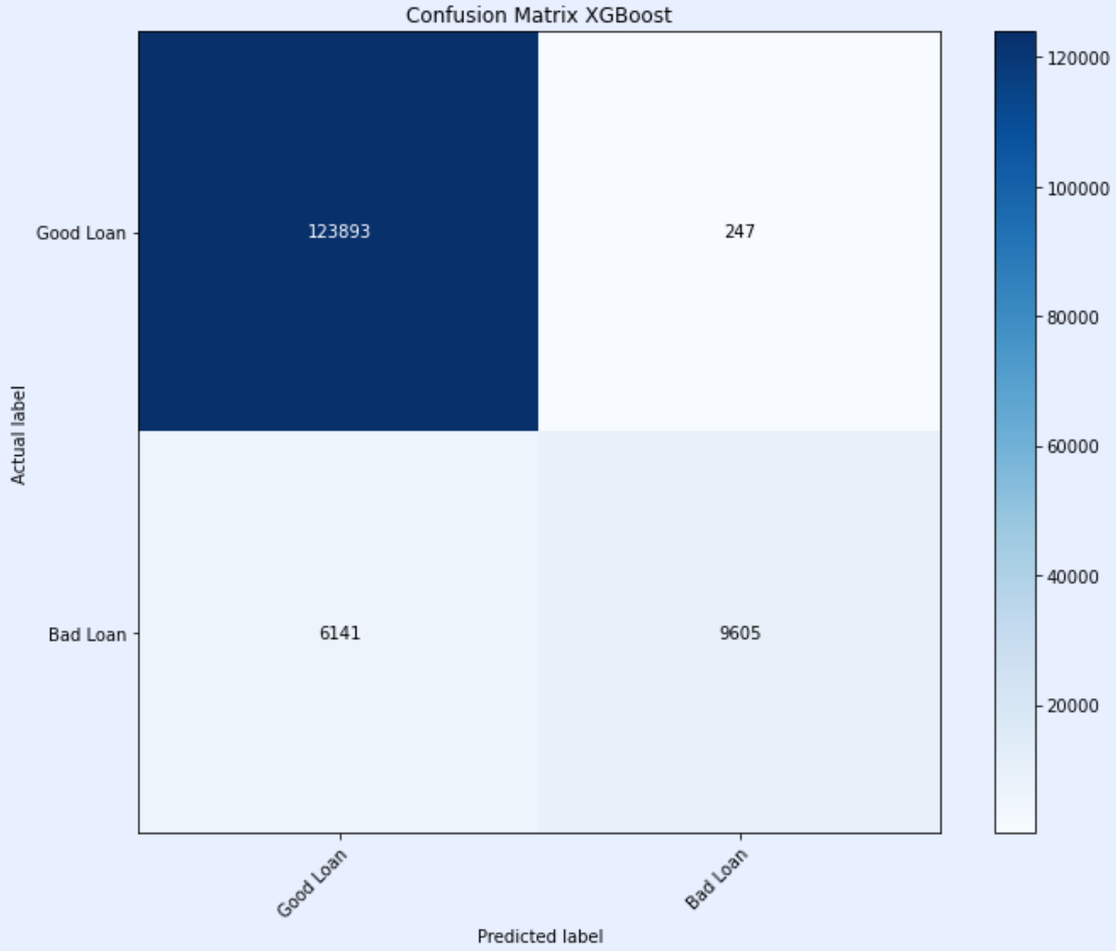


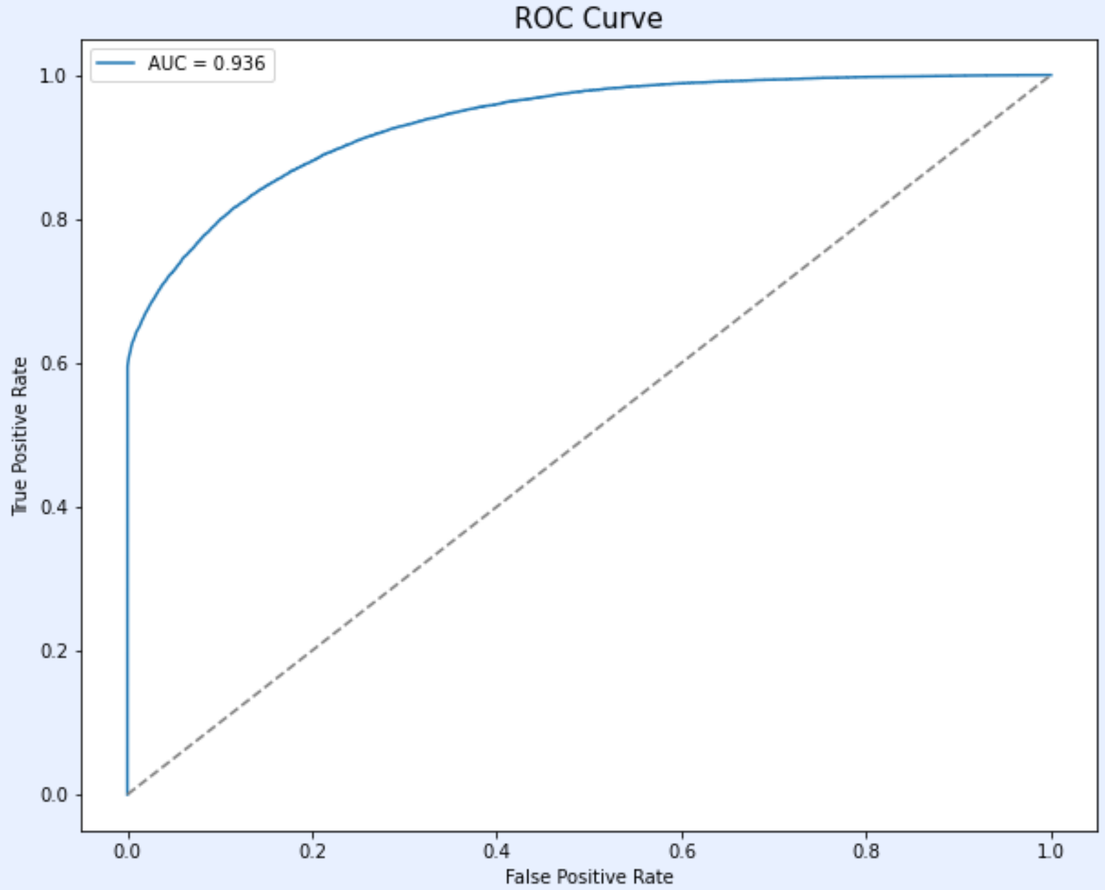
LOGISTIC REGRESSION



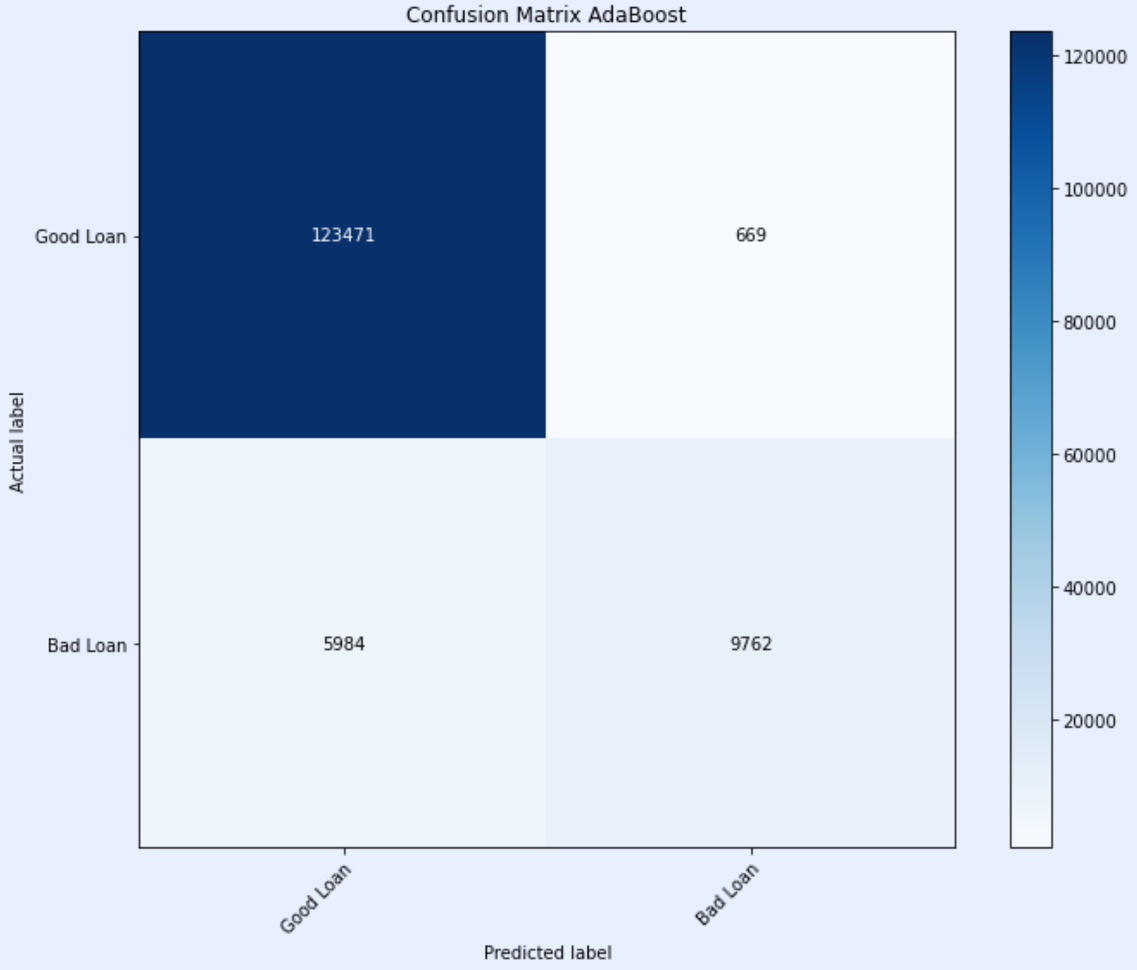


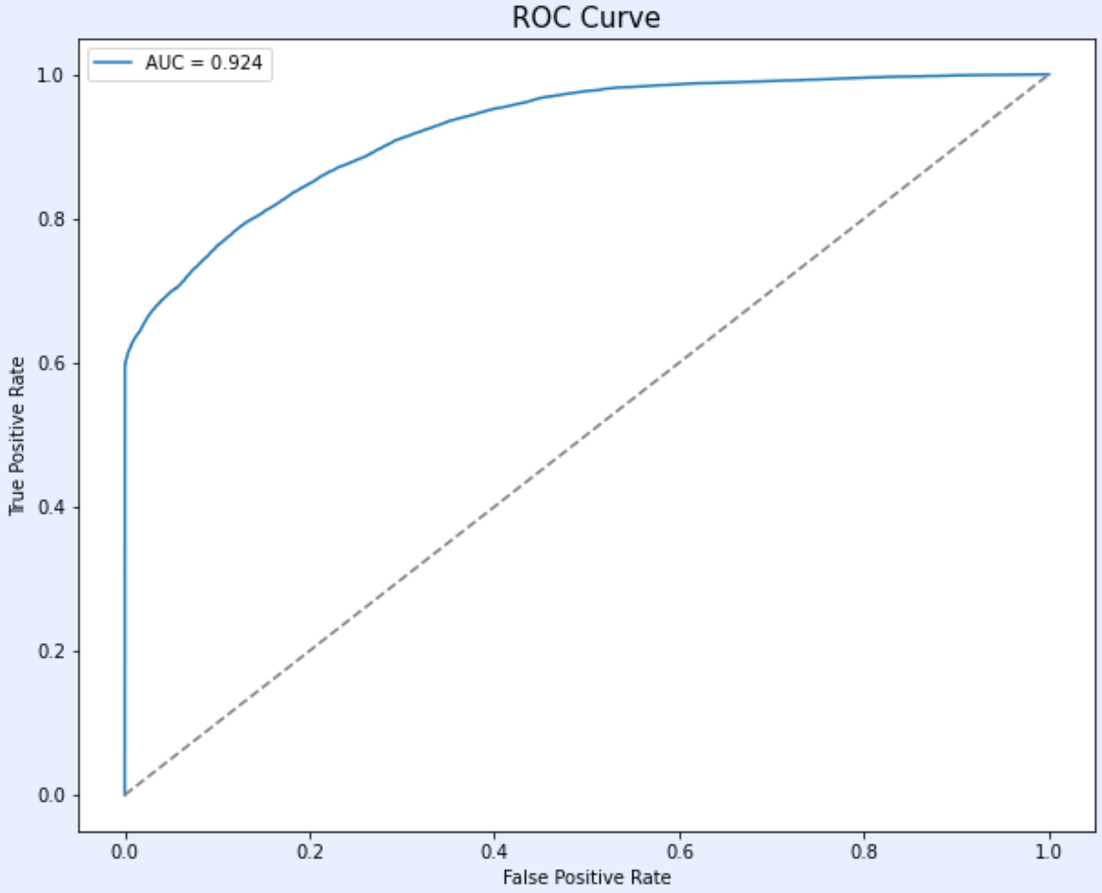
XGBOOST CLASSIFIER



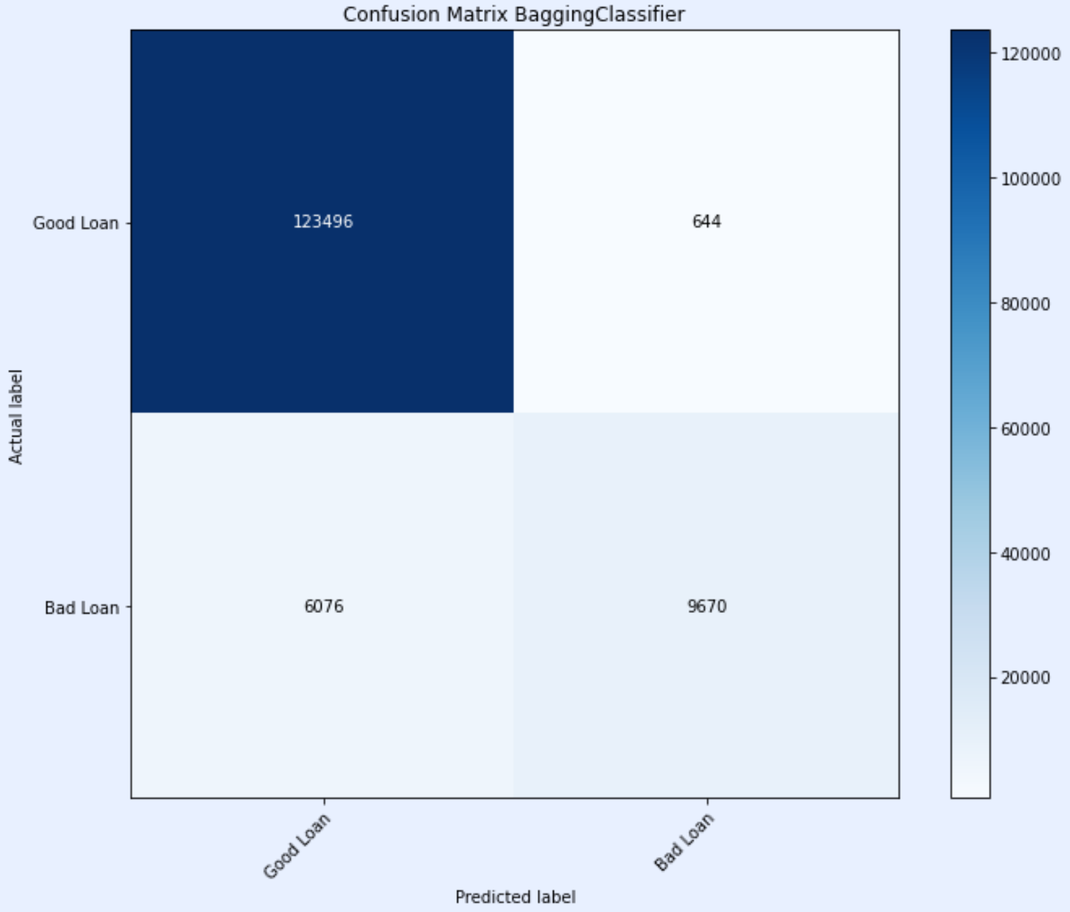


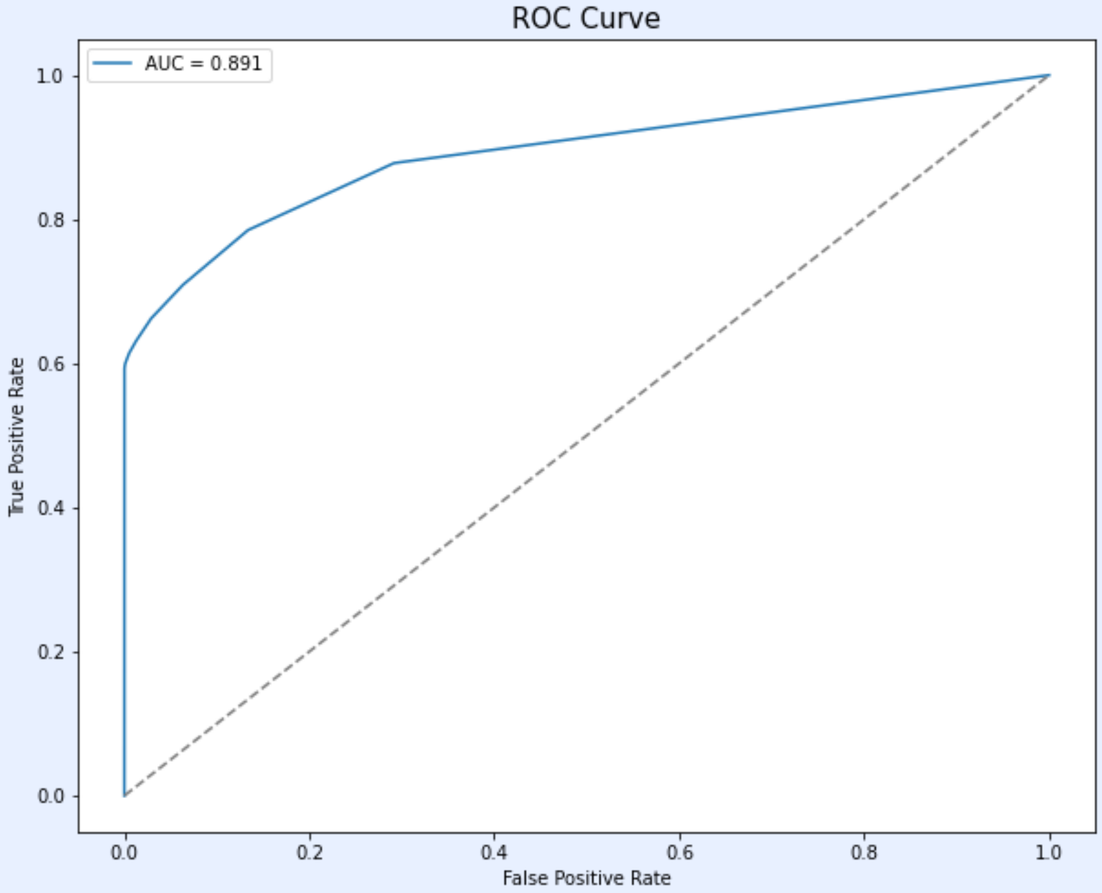
ADABOOST CLASSIFIER



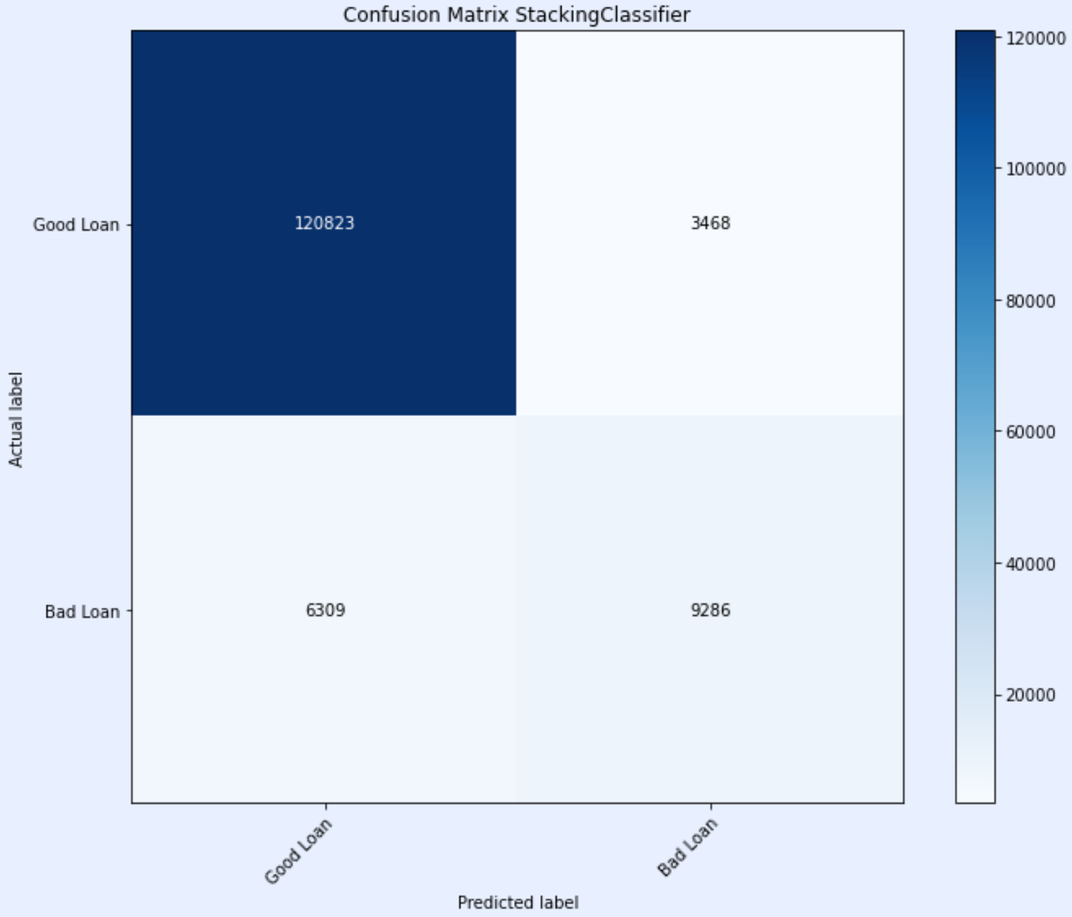


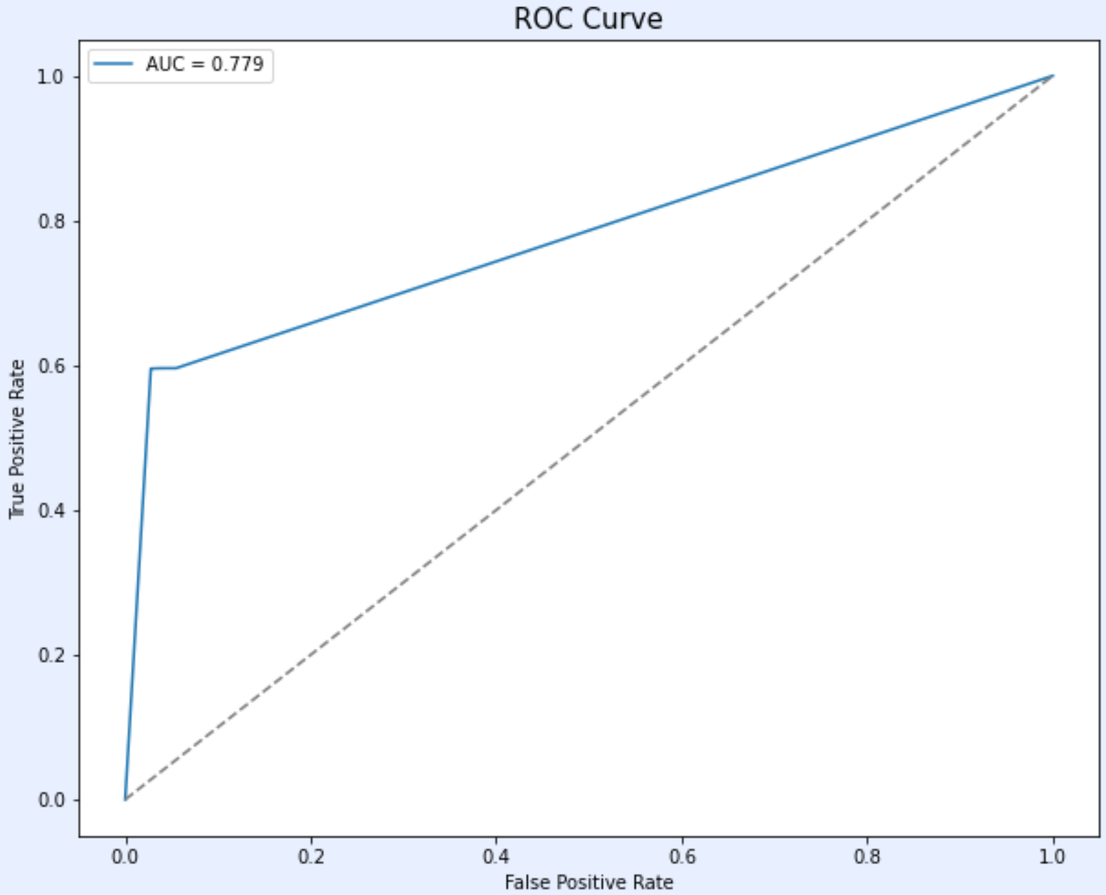
BAGGING CLASSIFIER



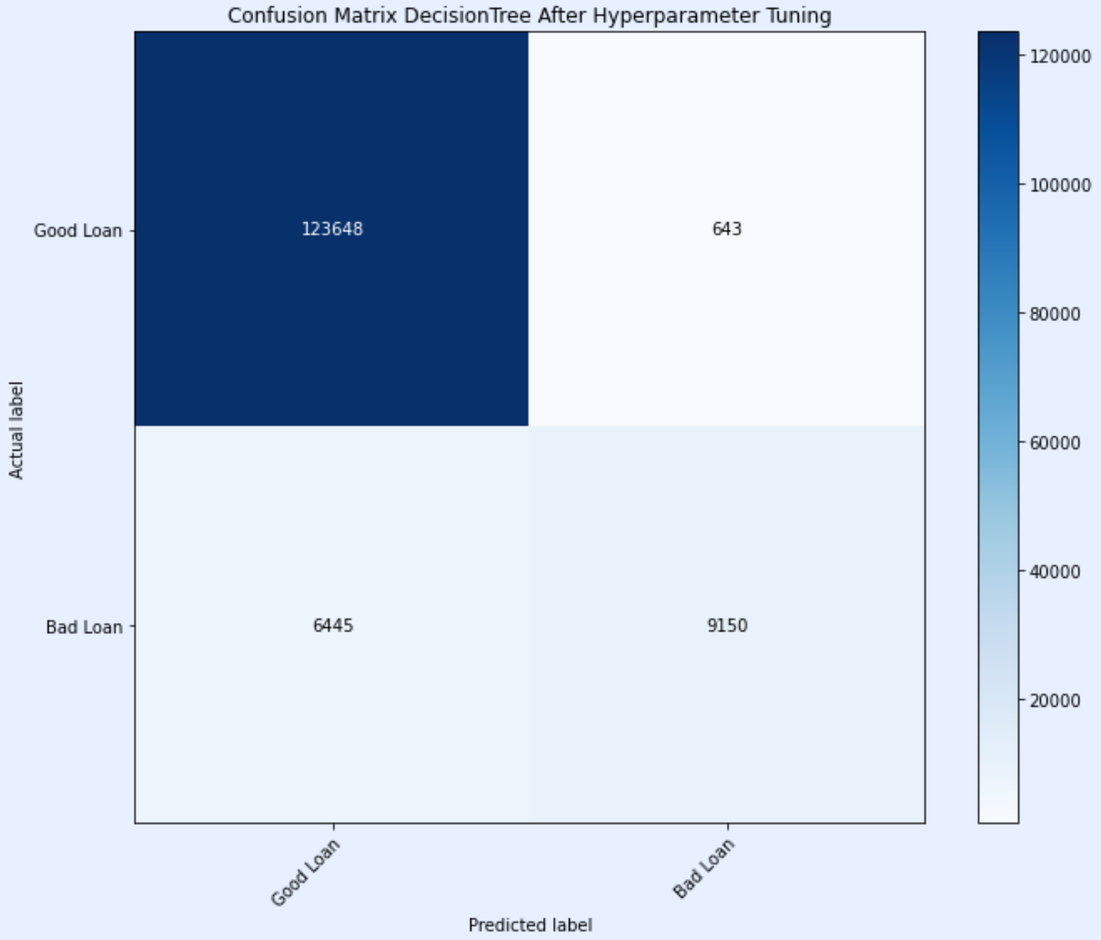


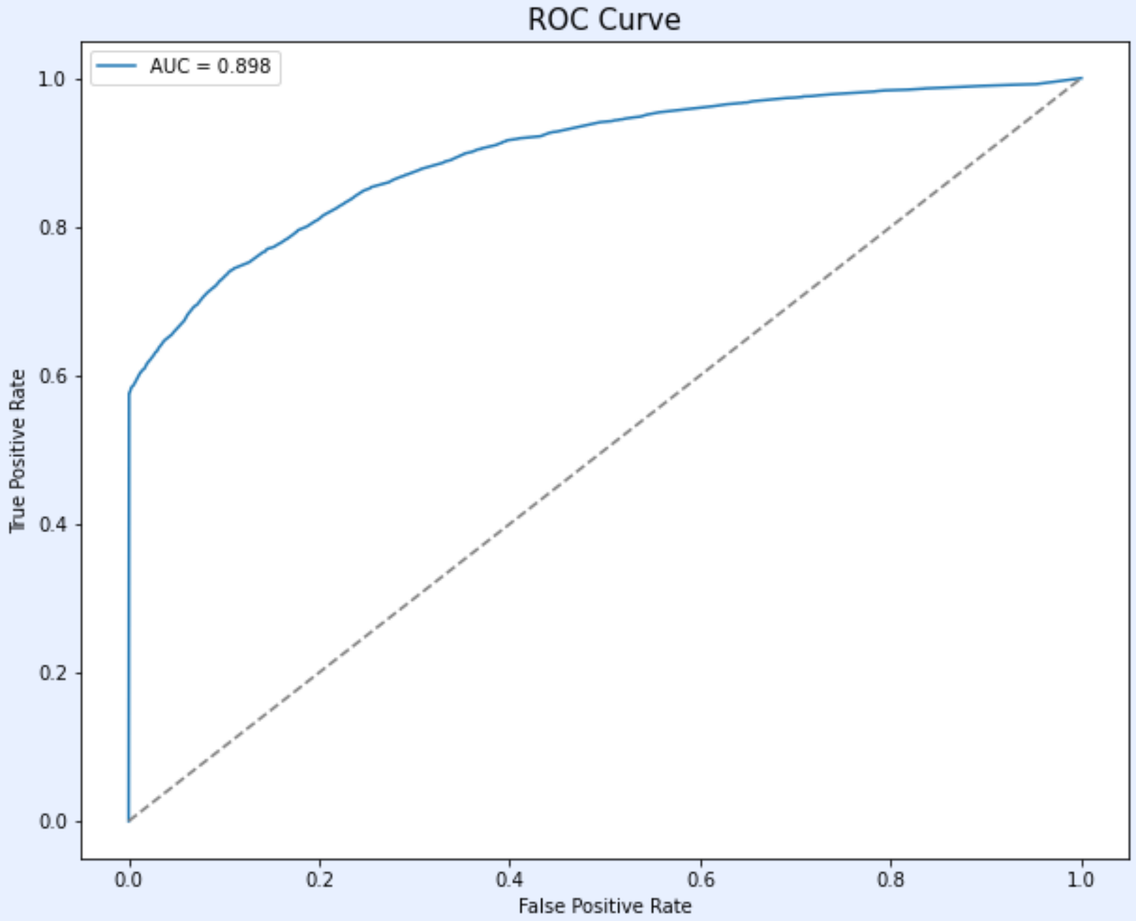
STACKING CLASSIFIER





DECISION TREE CLASSIFIER (HYPERTUNING MODEL)





DISCUSSION AND RESULT

In terms of accuracy, precision, recall, F1-Score, AUC-Proba Train, and AUC-Proba Test, it is evident that two models stand out with the best performance—Random Forest Classifier and XGBoost Classifier. Random Forest achieves accuracy of 95%, precision 97%, recall 80%, F1-Score 86%, AUC-Proba Train 100%, and AUC-Proba Test 92%. On the other hand, XGBoost Classifier attains accuracy of 95%, precision 96%, recall 80%, F1-Score 86%, AUC-Proba Train 99%, and AUC-Proba Test 93%. This analysis indicates that both models, Random Forest Classifier and XGBoost Classifier, exhibit excellent performance in predicting credit risk, showcasing a high ability to classify data accurately.

The importance of assessing precision and recall is evident in this evaluation. Random Forest Classifier achieves a precision of 97%, highlighting its efficiency in identifying true positives and minimizing false positives. Simultaneously, a recall of 80% demonstrates the model's ability to detect as many true positives as possible. XGBoost Classifier, despite having a slightly lower precision (96%), maintains a good balance with an 80% recall. With an F1-Score of 86% for both models, it can be concluded that they successfully achieve a balanced trade-off between precision and recall.

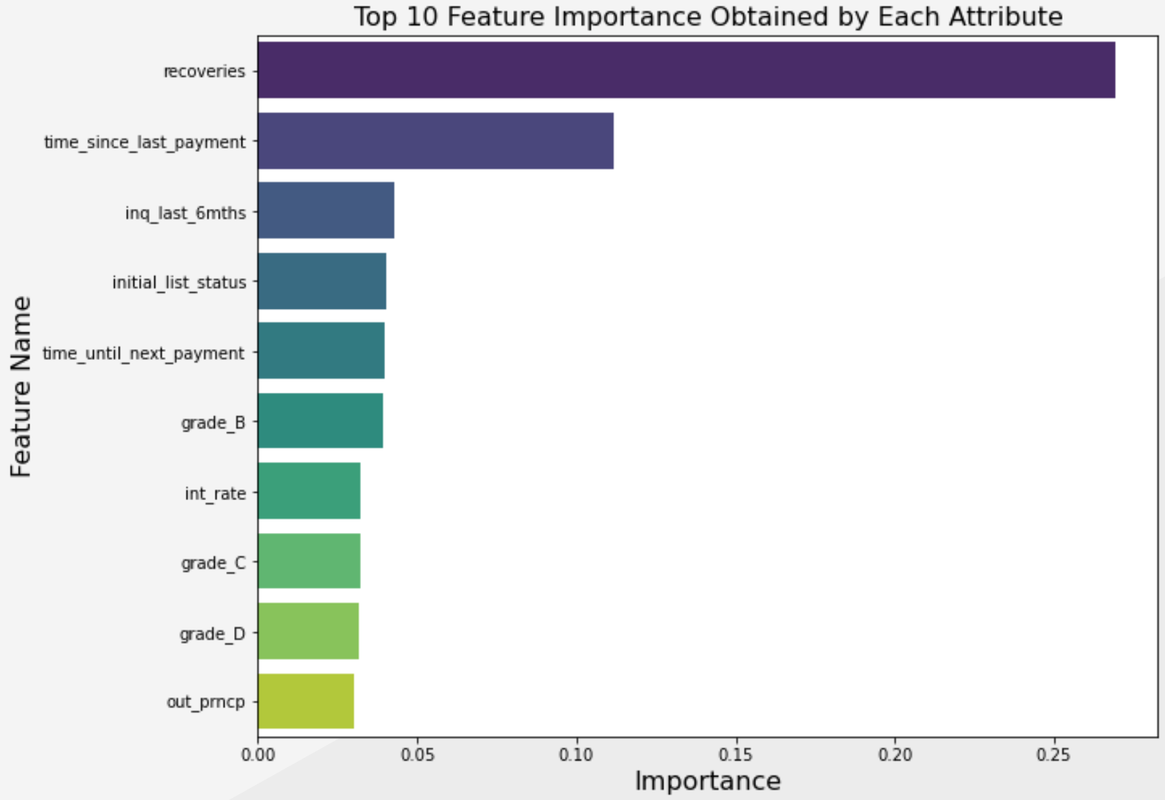
Additionally, the high AUC-Proba Train and AUC-Proba Test for both models indicate their ability to differentiate effectively between positive and negative categories. Random Forest Classifier achieves an AUC-Proba Test of 92%, while XGBoost Classifier reaches 93%. Overall, these results instill confidence that both Random Forest Classifier and XGBoost Classifier have strong potential for application in credit risk management, providing accurate and consistent predictions on both training and test data.

In detailing the confusion matrix evaluation results, RandomForestClassifier achieves 124,039 True Positives (TP), 101 False Positives (FP), 6,282 False Negatives (FN), and 9,464 True Negatives (TN). This indicates the model's capability to predict 124,039 good loans and 9,464 bad loans. Although the model still has the potential for prediction errors, as indicated by lower FP and FN values compared to TP and TN, these results affirm the RandomForestClassifier's proficiency in predicting both good_loan and bad_loan.

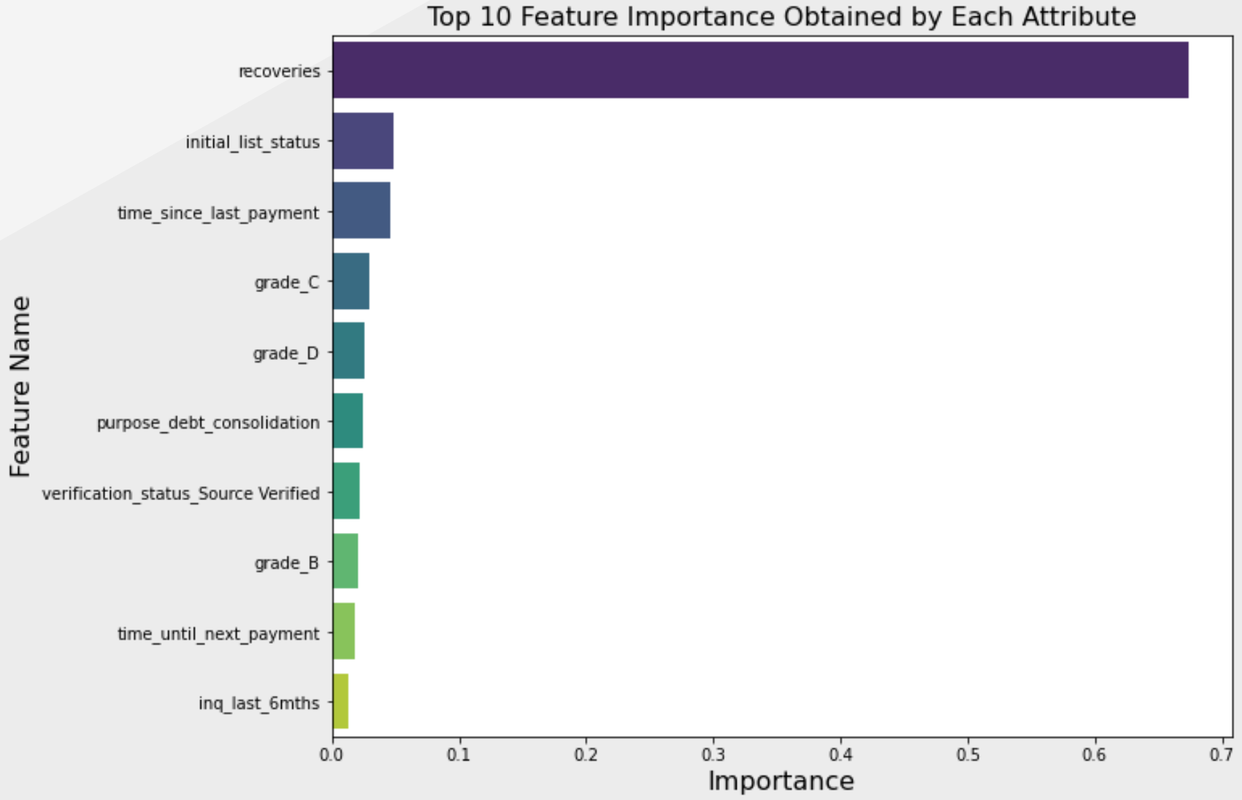
XGBoostClassifier, with 123,893 TP, 247 FP, 6,141 FN, and 9,605 TN, exhibits a similar pattern to RandomForestClassifier. The percentage of correct predictions is more dominant than prediction errors (FP and FN), confirming the model's reliability in predicting the categories of good_loan and bad_loan.

FEATURE IMPORTANCE

The results of the feature importances plot from RandomForestClassifier and XGBoostClassifier highlight the significance of features such as recoveries, time_since_last_payment, initial_list_status, time_until_next_payment, inq_last_6mths, and grade. Analysis of these features can serve as a foundation for credit risk assessment. Focusing on these features is expected to provide more accurate recommendations regarding the most influential factors in determining credit risk.



RandomForestClassifier



XGBoostClassifier

id/x partners

THANK YOU

JULIAN SAPUTRA