Examining Predictions for Credit Card Defaults

Dongyan Sun

Data Science Institute

December 6, 2023

https://github.com/EmmaSun19902023/Examining_Predictions_for_Credit_Card_Defaults_Project/tree/main

1. Introduction

Motivation

Predicting Overdue

Lending Decision

Risk Management

Financial Inclusion

Dataset Description

American Express

Kaggle

45,528 Rows

19 Columns

IID data

4.43% Missing

Target Variable

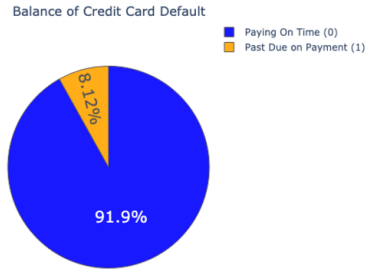
credit _card_default

1 for Past Due

0 for On Time

Imbalance

Classification



Previous Work

Credit Scoring Approaches Guidelines

World Bank Group Suggests:

1. Regression

2. Decision Tree

3. Random Forest

4. Gradient Boosting

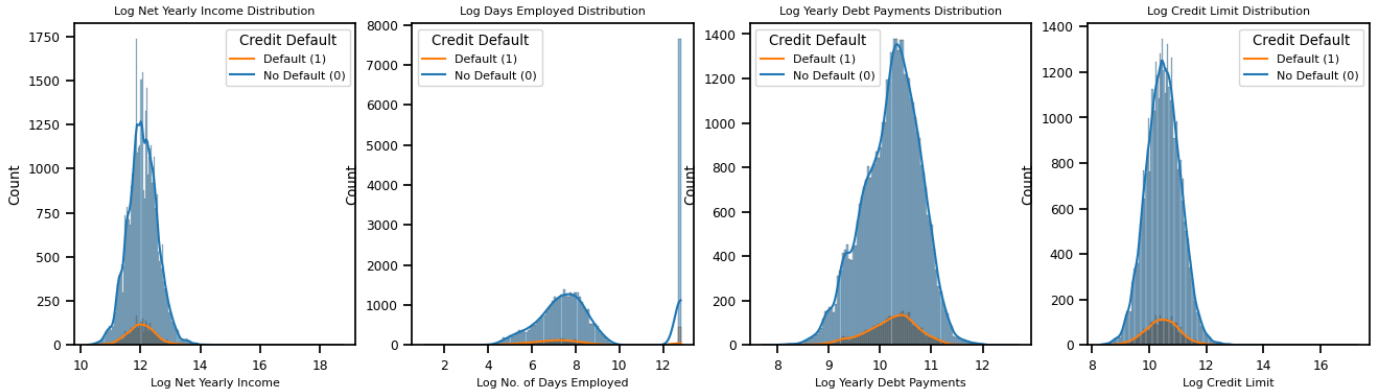
5. Deep Neural Networks

6. SVM

2. EDA

Histograms

- Skewed Distributions
- Log Transformation
- Consistent Patterns

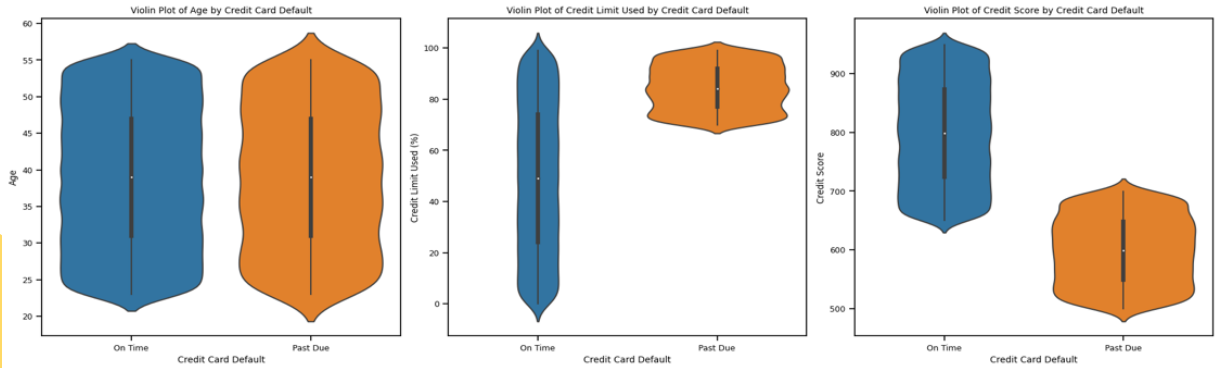


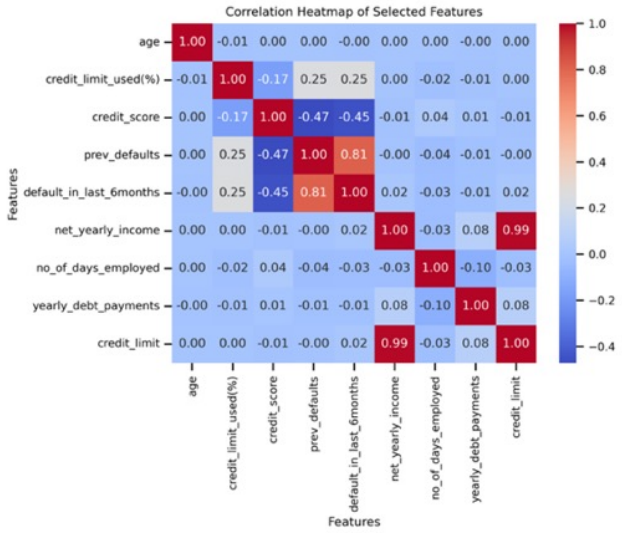
Heatmap

- Net Yearly Income VS. Credit Limit
- Default in last 6 month VS. Previous Default
- Strong Correlation

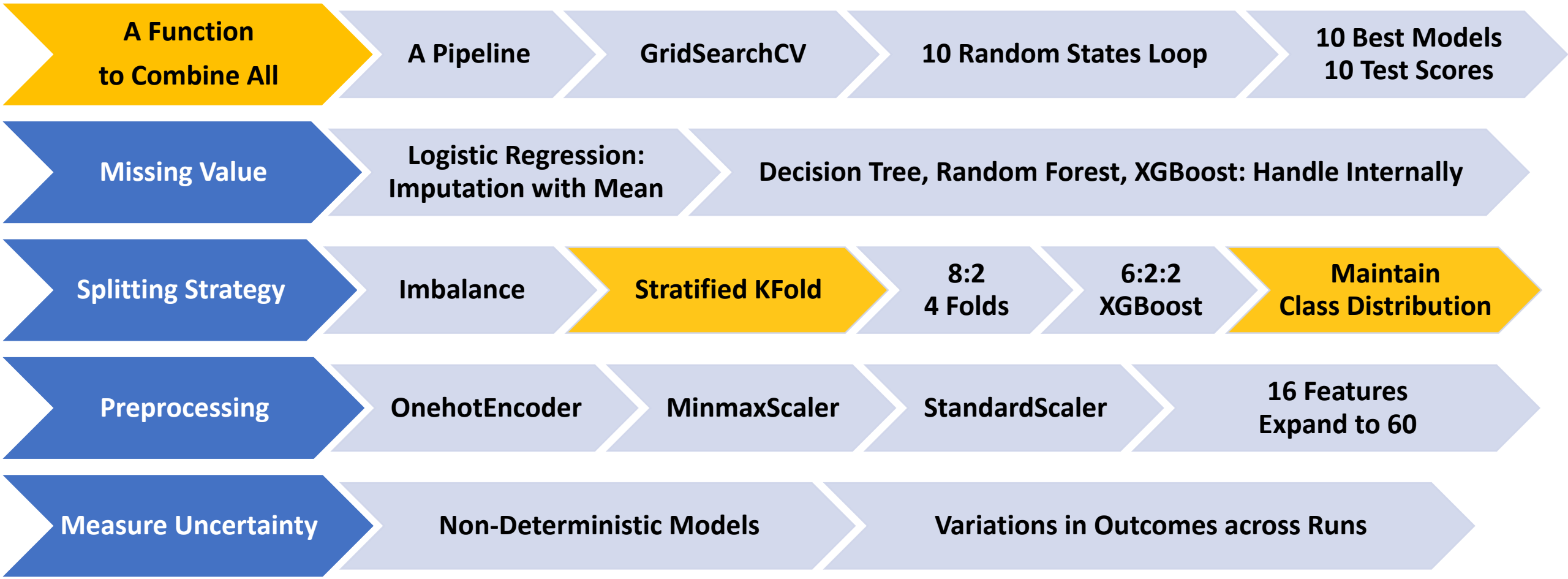
Violin plots

- Age: Consistent
- Credit Limit Used (%): Wider Spread for Class1
- Credit Score: Distinct Patterns

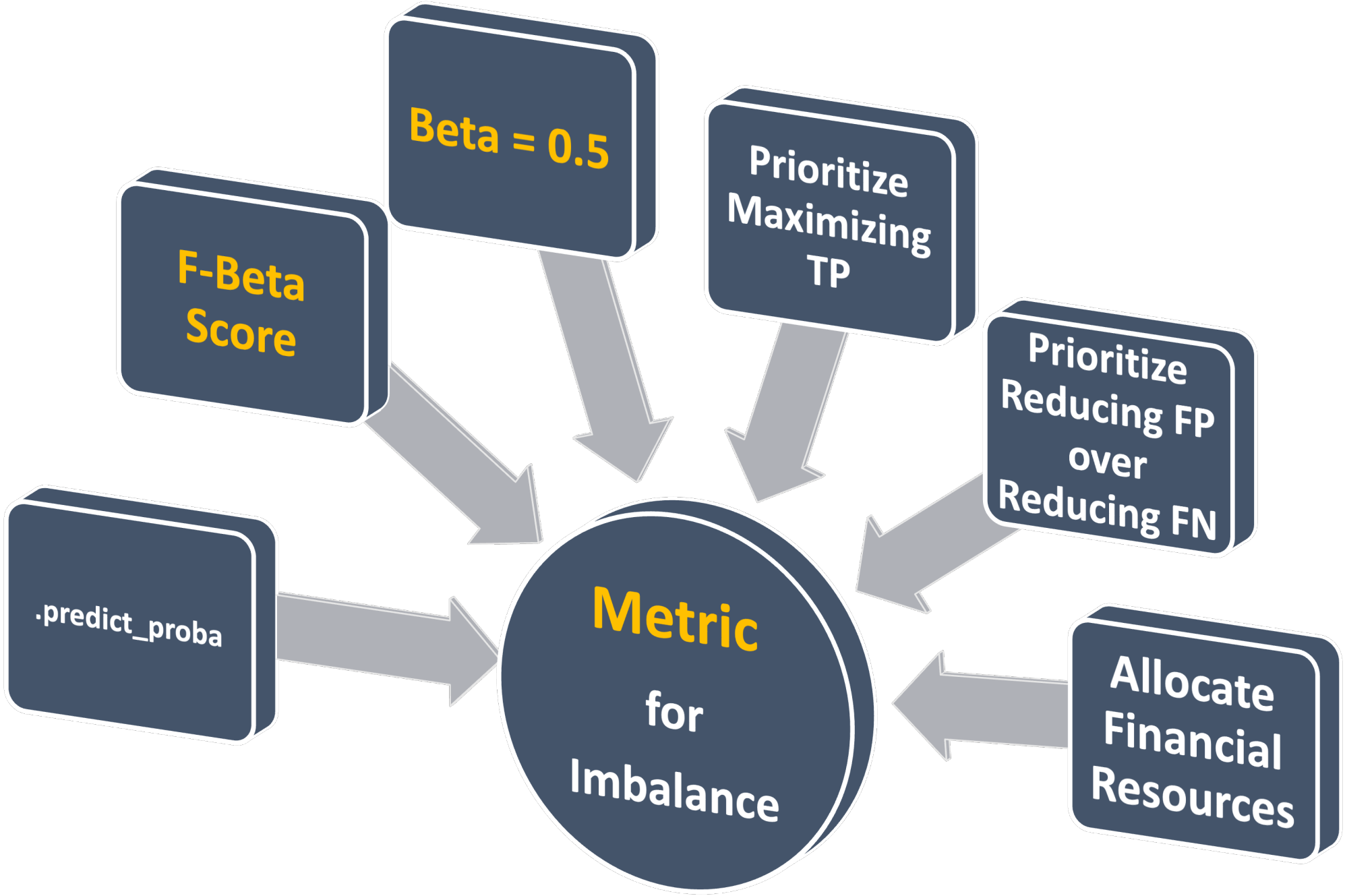




3. Cross Validation



Algorithms	Rationale	Parameters	Rationale	Best Parameters
Logistic Regression	famous for binary classification	C [0.01, 0.1, 1, 10]	allow fine-tuning of the regularization strength	C =0.1
Decision Tree	reduce overfitting and capture non-linear patterns	max_depth [5, 10, 15, None]	prevent the decision tree from becoming overly complex	max_depth =5
Random Forest	robust to noise and overfitting	n_estimators [50, 100, 150, 200]	trade-off between low training error and low testing error	n_estimators =100
XGBoost	mitigate overfitting	learning_rate [0.01, 0.1, 0.3, 0.5]	the contribution of each tree control over the step size	learning_rate =0.01

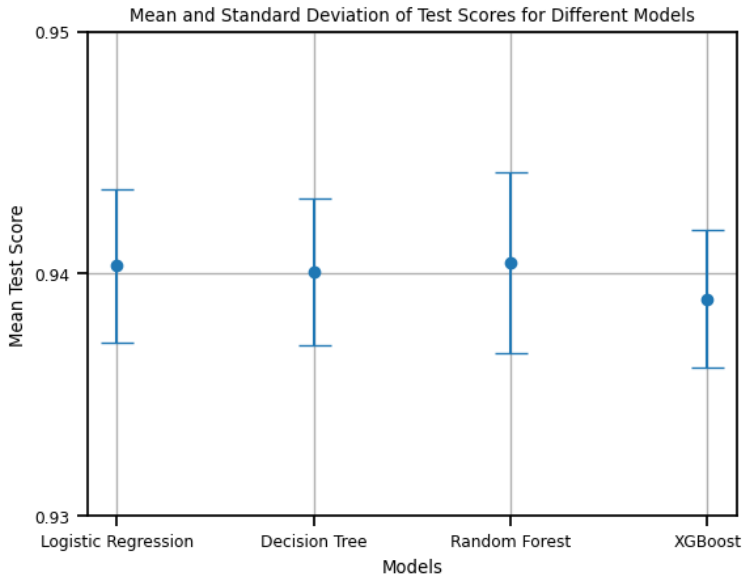

Beta = 0.5
F-Beta Score
Prioritize Maximizing TP
Prioritize Reducing FP over Reducing FN
.predict_proba
Metric for Imbalance
Allocate Financial Resources

4. Results

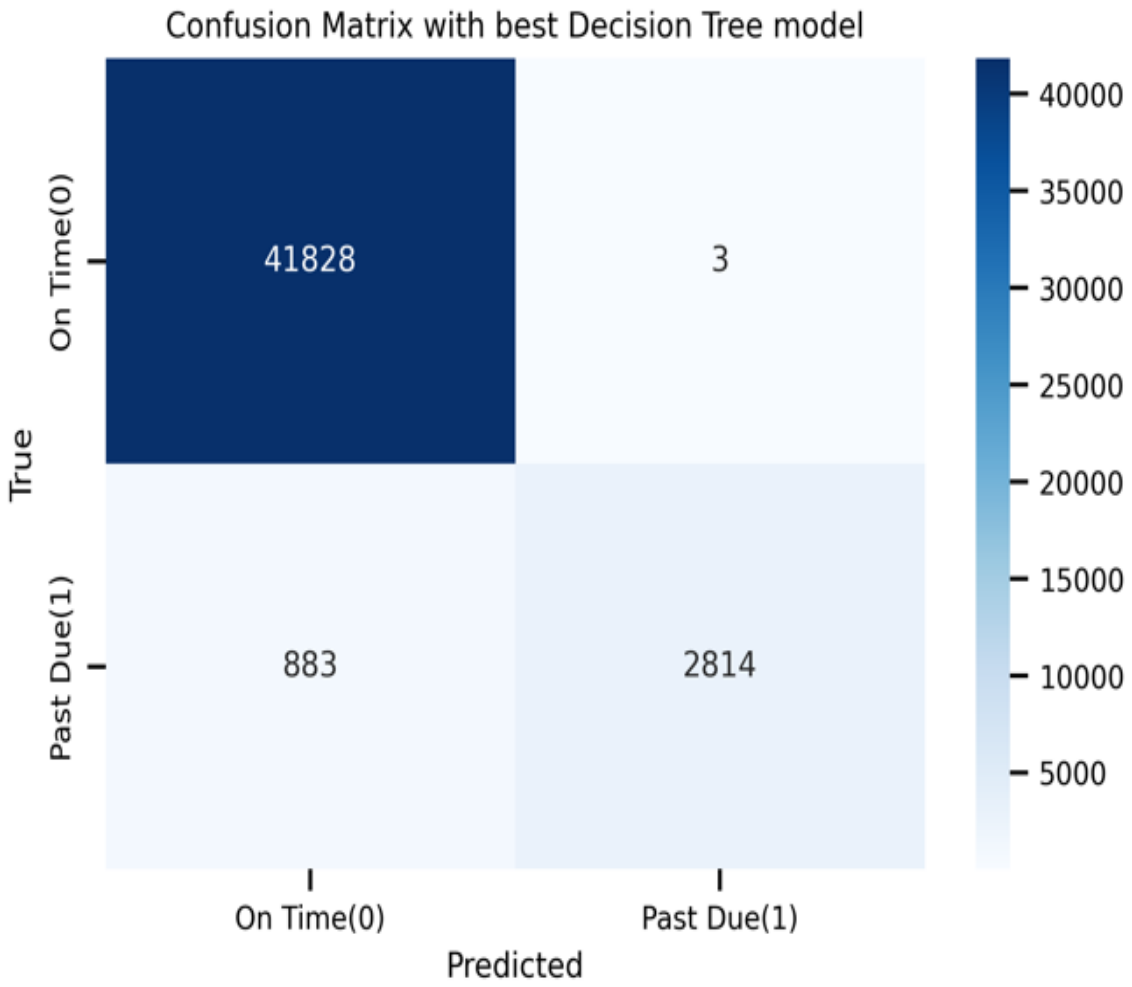
Baseline = 9.754%

Assume Baseline Model Predicts all Points as Class 1

Formula = $\frac{(1.25 * \text{Precision} * \text{Recall})}{(0.25 * \text{Precision} + \text{Recall})}$



Model	Mean of Test Score	Std of Test Score	Coefficient of Variation	Number Of Std Above Baseline
Logistic Regression	0.9403	0.003162	0.003363	266.48
Decision Tree	0.9400	0.003012	0.003205	279.63
Random Forest	0.9404	0.003711	0.0039462	227.12
XGBoost	0.9389	0.002861	0.0030470	294.09



Strong Performance

Precision =0.9989

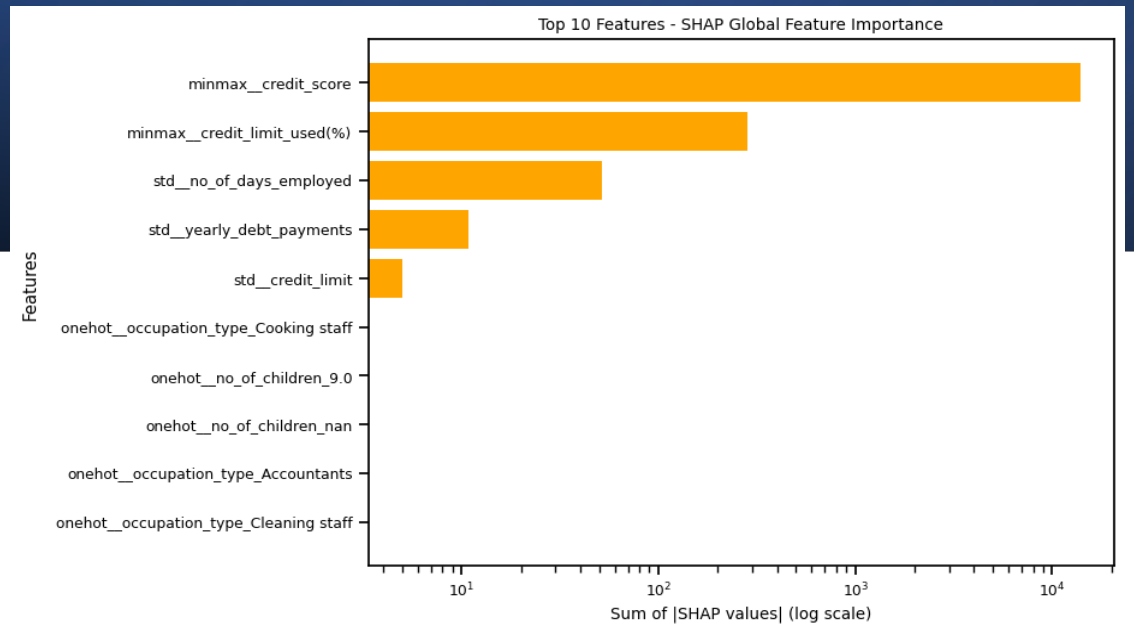
- accuracy of positive predictions

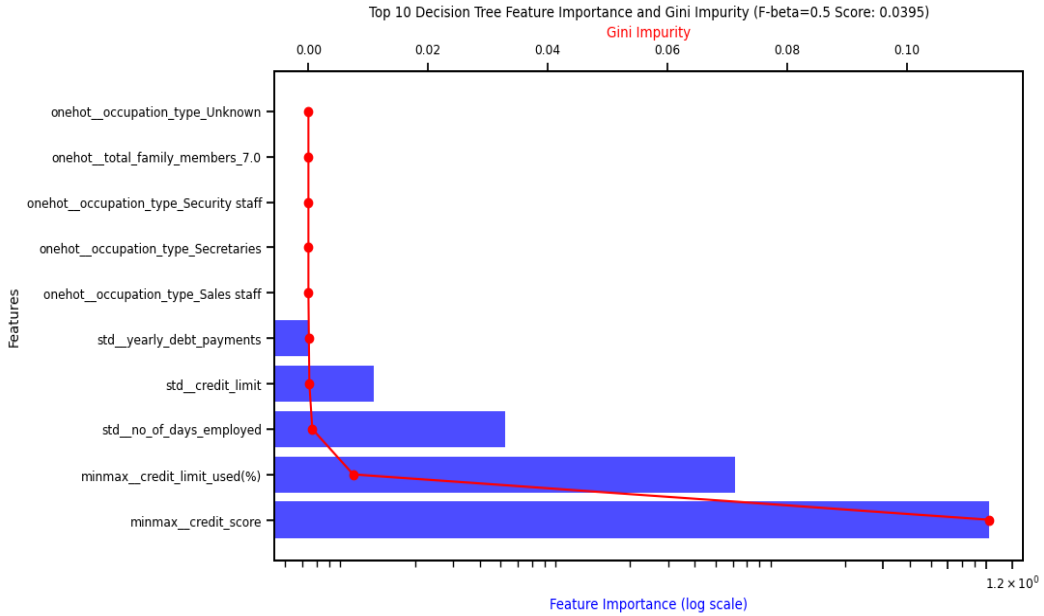
Recall =0.7612

- ability to capture all positive instances

F-beta Score =0.9402

- efficacy in classification tasks





Global: TOP 3

- credit_score
- credit_limit_used(%)
- no_of_days_employed

Metric

- Shap value
- .feature_importances_
- Gini Impurity

NO.9000 data point in Force Plot 3

credit_score

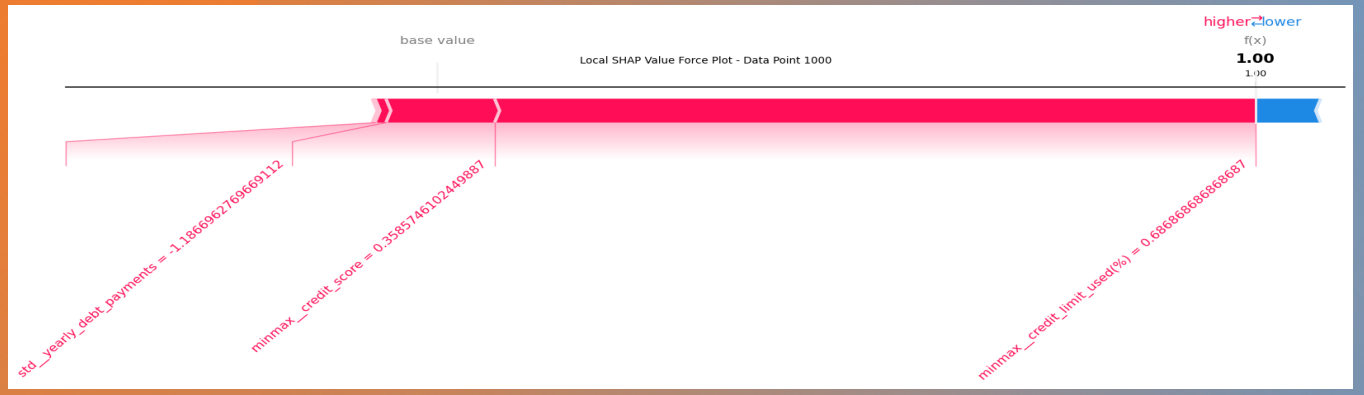
On Blue Arrow:
Lower-Than-Average & Negative Impact on the Prediction

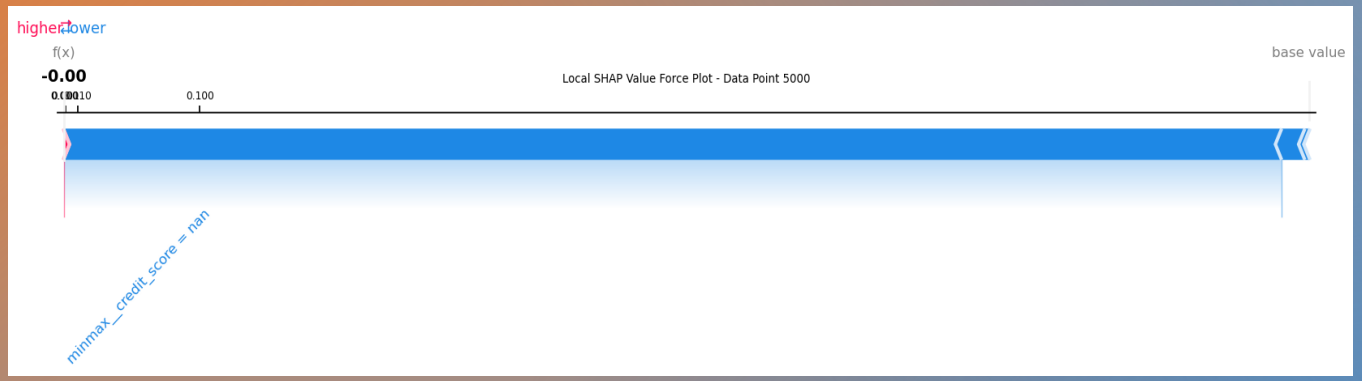
SHAP value for specific data point

credit_limit_used(%)

On Red Arrow:
Higher-Than-Average & Positive Impact on the Prediction

SHAP value for specific data point





5. Outlook

Possibility of being Overfitting

- Explore more Parameters and Values
- find sweet spot to Minimize Overfitting and Underfitting

Control Weight of credit_score

- Extremely Influential Variable in feature importance
- try Feature Scaling or Weighted Models

Drop certain Features

- Heatmap

Alternative Approaches to Handle Missing Values

Try various Beta Values

6. Reference

[1] Pradip Basak. AmExpert CodeLab 2021. Retrieved September, 2023. From https://www.kaggle.com/datasets/pradip11/amexpert-codelab-2021/code

[2] Knutson, M. L. Credit Scoring Approaches Guidelines - Final Web. World Bank Group. April 2, 2020. Page 15, from https://thedocs.worldbank.org/en/doc/935891585869845 1-0130022020/CREDIT-SCORING-APPROACHES-GUIDELINES-FINAL-WEB