

Loan Default Prediction

Capstone Project Report | Author: Shouvik Nandy

Contents

	Slide Nbr#
Executive Summary	3
Problem Statement	4
Solution Summary	5
Recommendations	6
Appendix	10

Executive Summary

This report provides analysis and proposed solution of leveraging Machine Learning and predictive methods to optimize the current loan approval process in the banks consumer credit department for Home Equity lines of credit.



Important factors that determine whether a customer will default a loan include –



Debt to Income Ratio



Delinquent Credit Lines



Major delinquency or late payment

The data gathered by the bank under the existing loan underwriting process will be used to develop the predictive models. Data analysis technique include exploration, data preparation and predictive model development.



Key success factors for the solution is to reduce number of wrongly approved loans that will increase NPA (Non Performing Asset).

The proposed solution is to automatically predict whether a customer will default a loan based on the available data and share the predictions with credit department to optimize the loan approval process. The predictions will be supported by important factors that determined the result of prediction.



Key information that determine the likelihood of a customer defaulting the loan are not captured for many existing loans that were approved or rejected. Improve processes to capture these information. Continuously refine the model to improve performance and prediction accuracy.

Problem Statement

Current State

Resource intensive and time consuming loan under writing process. This is further impacted by human biases for approval of loans.

Gap/Opportunity

Lack of predictive methods to aide decision making in the loan approval processes.

Future State

Faster loan underwriting by aiding decision making with predictive power of Machine Learning.

2

Weeks

Average time taken by the current loan underwriting processes to approve or reject the loan.

50
%

Reduction in time required for loans processing

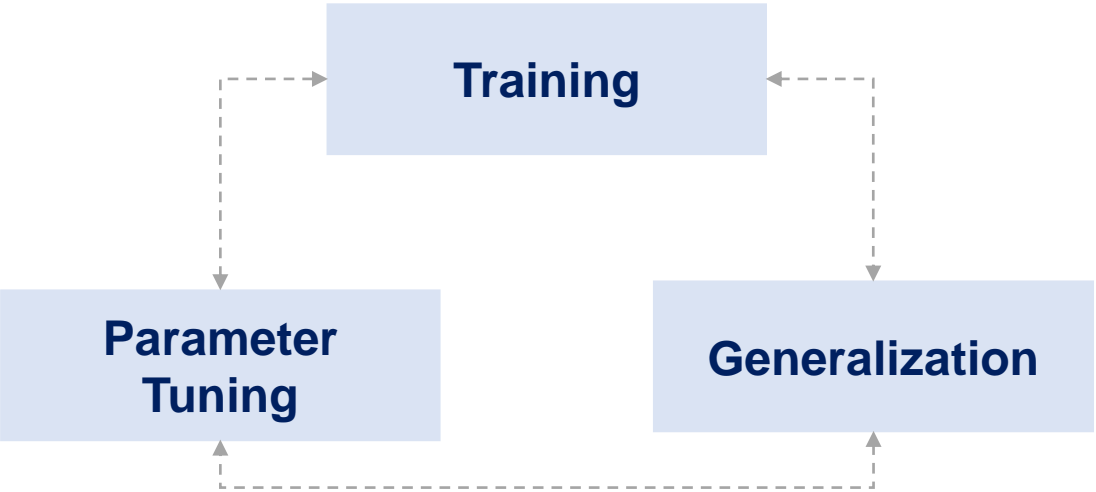
1

Week

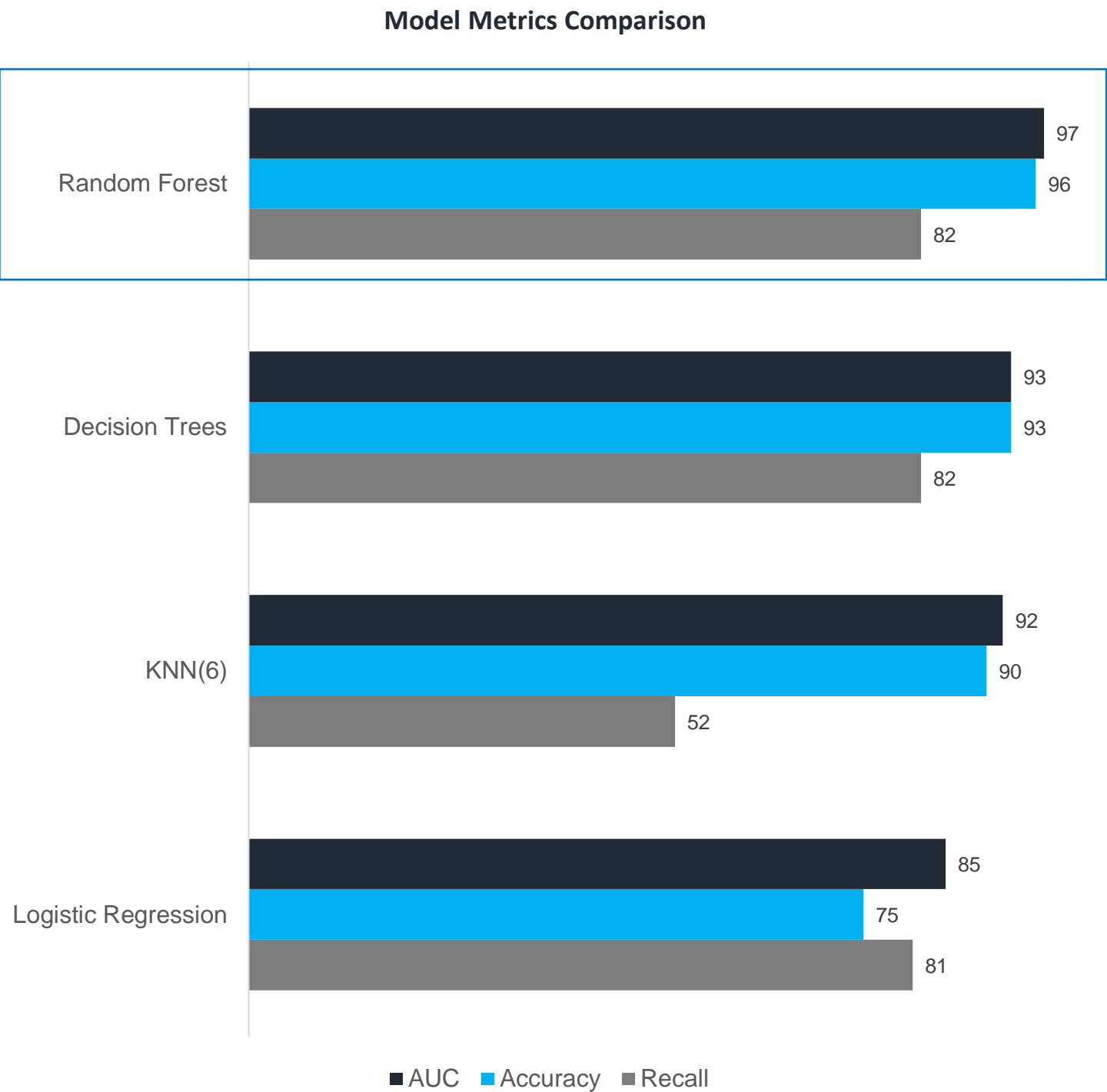
Loan underwriting completed faster as models predict the likelihood of customer defaulting a loan repayment.

Solution Design

- Identify patterns in data that influences the likelihood default loan
 - Identify data pre-processing requirements
- Handle missing values
 - Encode Categorical variables
 - Scaling of numerical variables
 - Split data in train , validate and test sets
 - Use SMOTE for oversampling of imbalance class
- Models: Logistic, Decision Tree & Random Forest
 - Tune hyperparameters using Random and Grid search methods
 - Generalization & Regularization for overfitting using cross validation methods
 - Utilized model performance metrics (Recall, Accuracy & AUC) for model optimization and selection
- Predict target class using the best model
- Measure Model performance
 - Identify and report important features of the model



Recommendation – Model Selection

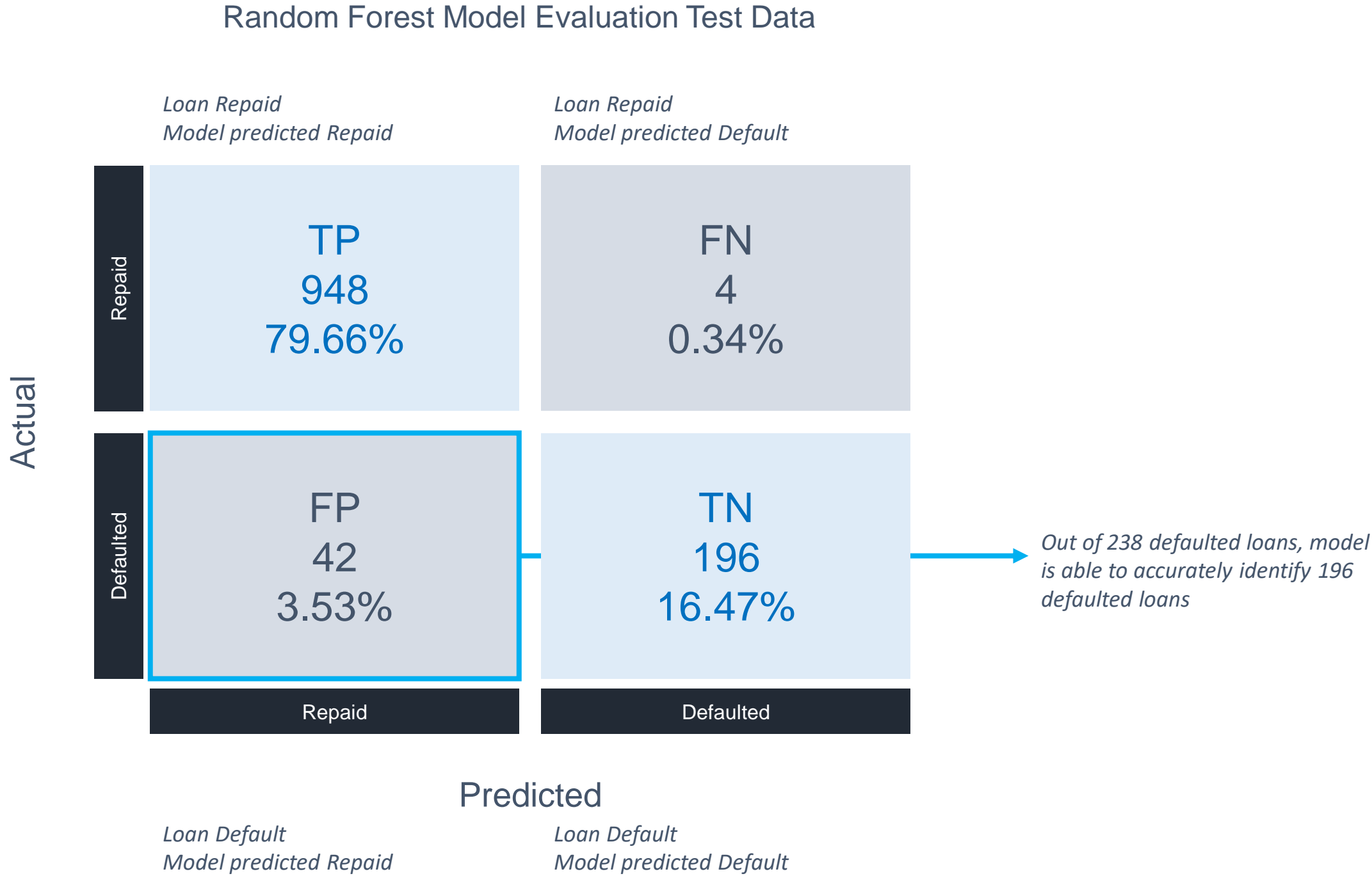


Based on model evaluation metrics, recommend Random Forest Model to be the final model.

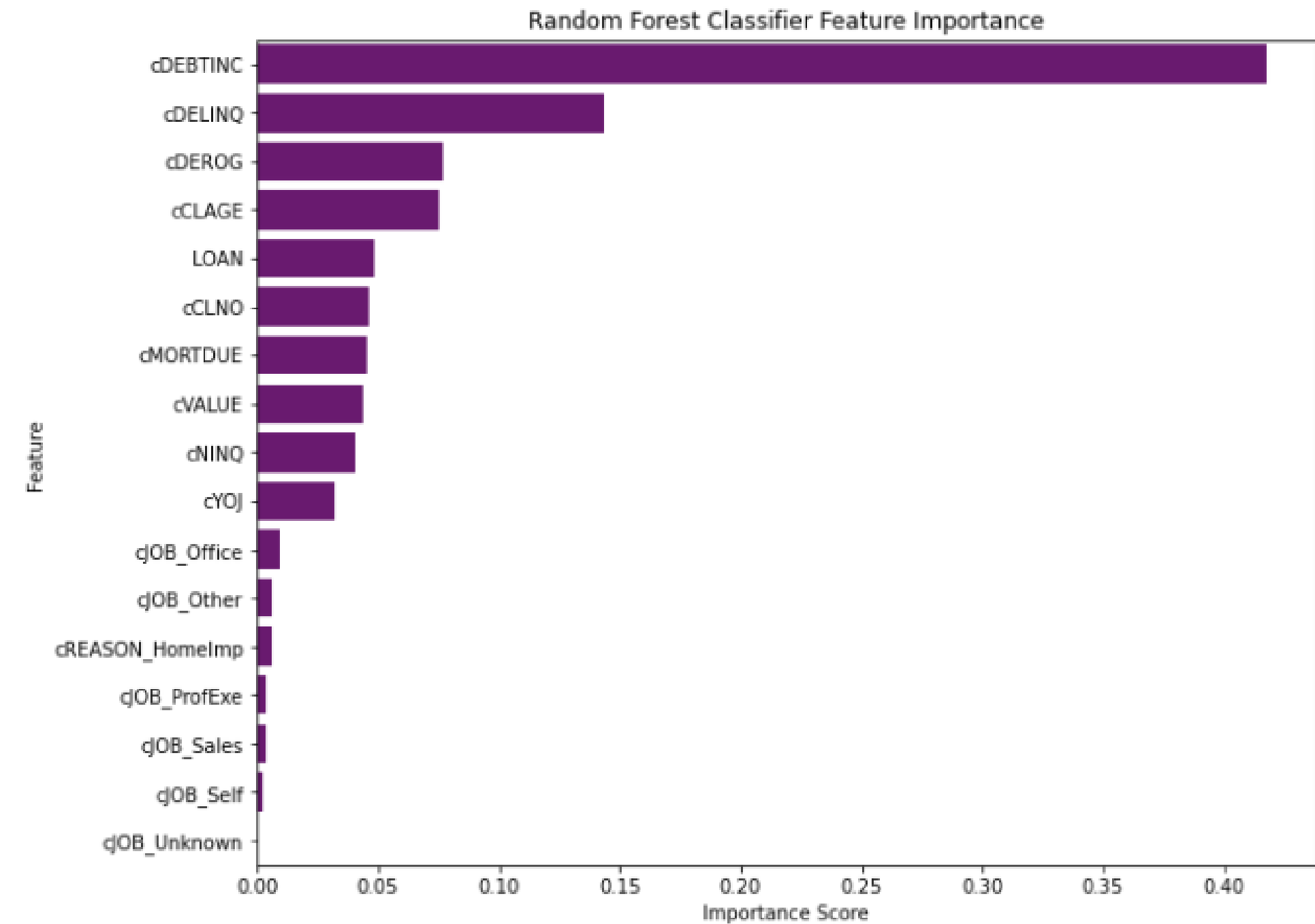
Score on Validation Data in %						
Model	Scaled Data	SMOTE	Hyper Parameter Tuned	Recall Defaulted Loans	Overall Accuracy	AUC
Logistic Regression	No	No	No	14	81	-
Logistic Regression	Yes	No	No	47	86	-
Logistic Regression	Yes	Yes	No	59	86	-
Logistic Regression	Yes	Yes	Yes	81	75	85
KNN(10)	Yes	Yes	No	49	88	-
KNN(6)	Yes	Yes	Yes	52	90	92
Decision Trees	Yes	Yes	No	84	95	-
Decision Trees	Yes	Yes	Yes	82	93	93
Random Forest	Yes	Yes	No	84	96	-
Random Forest	Yes	Yes	Yes	82	96	97

Recommendation – Model Evaluation

1190
#of loans evaluated



Recommendation – Important Features



The following top 3 feature are important for identifying a potential loan being defaulted by the customer

Feature	Interpretation
Debt to Income Ratio	Customer have no funds for repaying debt
Number of Delinquent Credit Lines	Customer have history of defaulting loans
Number of Reported Delinquencies and/or Late Payments	Customer have history of defaulting loans with major issues and late payments

Recommendations - Improvements

Threshold Analysis

We can evaluate updating the threshold to evaluate improvement in Recall score, as its is important that we don't approve a loan which has high probability of being defaulted.

Risks & Challenges

The model is not evaluated on Out of time data. The training data contain many imputed information, this might introduce bias in the model.

Implementation Considerations

There are lot of missing data in this dataset. The model performance can be increased if the missing data are being captured for the new loans.

Pilot the model on production data in parallel with manual operation to evaluate model performance on actual data.



Setup Data Engineering team to develop data pipelines required for Model implementation.

Periodically review Model performance and optimize continuously.

Explore new features which can improve model performance.

Verify the recommended features with Credit Analysts to evaluate and gaps in model interpretation.

Appendix

Notebook	Description	Format
 D:\Learning\Data Science\MIT-DS\Capstone Project\Notebooks\capstone.ipynb	Notebook entire capstone project	IPYNB
 Microsoft Edge HTML Document	Notebook entire capstone project	HTML