Loan Defaulter Classification

Yingqing Qiu
Metis Data Science Bootcamp

The problem

Goal and Impact

- Predict if a person will be a loan defaulter or not.

- Prevent bank losses and lower the potential of impacting country's economic growth.

Data

- Kaggle Dataset: Bank Loan Defaulter Prediction

- 67,463 rows and 35 columns

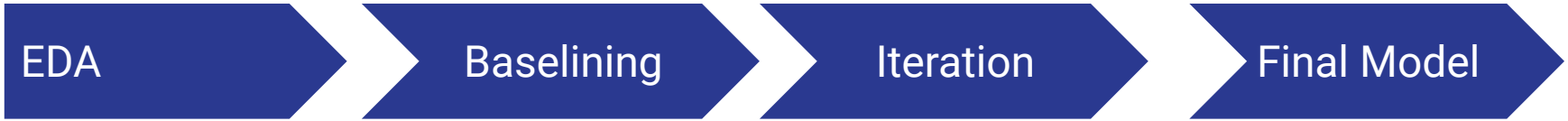
- Target: Loan Status (1 or 0)

Method

- KNN, Logistic Regression, Decision Tree, Random Forest, and XGBoost

- Oversampling, SMOTE, Cross-validation, GridsearchCV

Classification Work-Flow



1. Data Cleaning

2. Pair Plot

3. Prelim. Feature Engineering

1. Starting with features generated in EDA.

2. Using Logistic Regression as the baseline model.

1. Practicing with multiple Classifiers to generate new Feature set.

2. Consider imbalanced dataset.

3. Tuning parameters to optimize the models.

1. Select the most suitable model for the data set.

2. Interpret the results.

Summary of Features

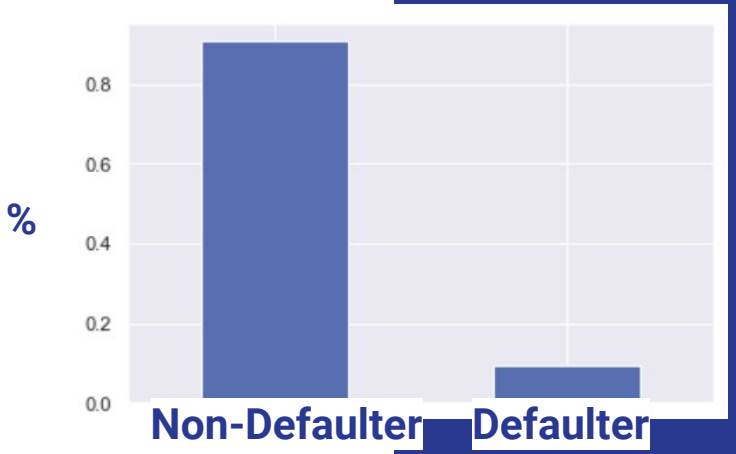
Loan Information

Loan Amount
Funded Amount
Grade
Verification Status
…

Personal Information

Home Ownership
Open Account
Public Record
Employment Duration
…

Target



Summary of Model Performances

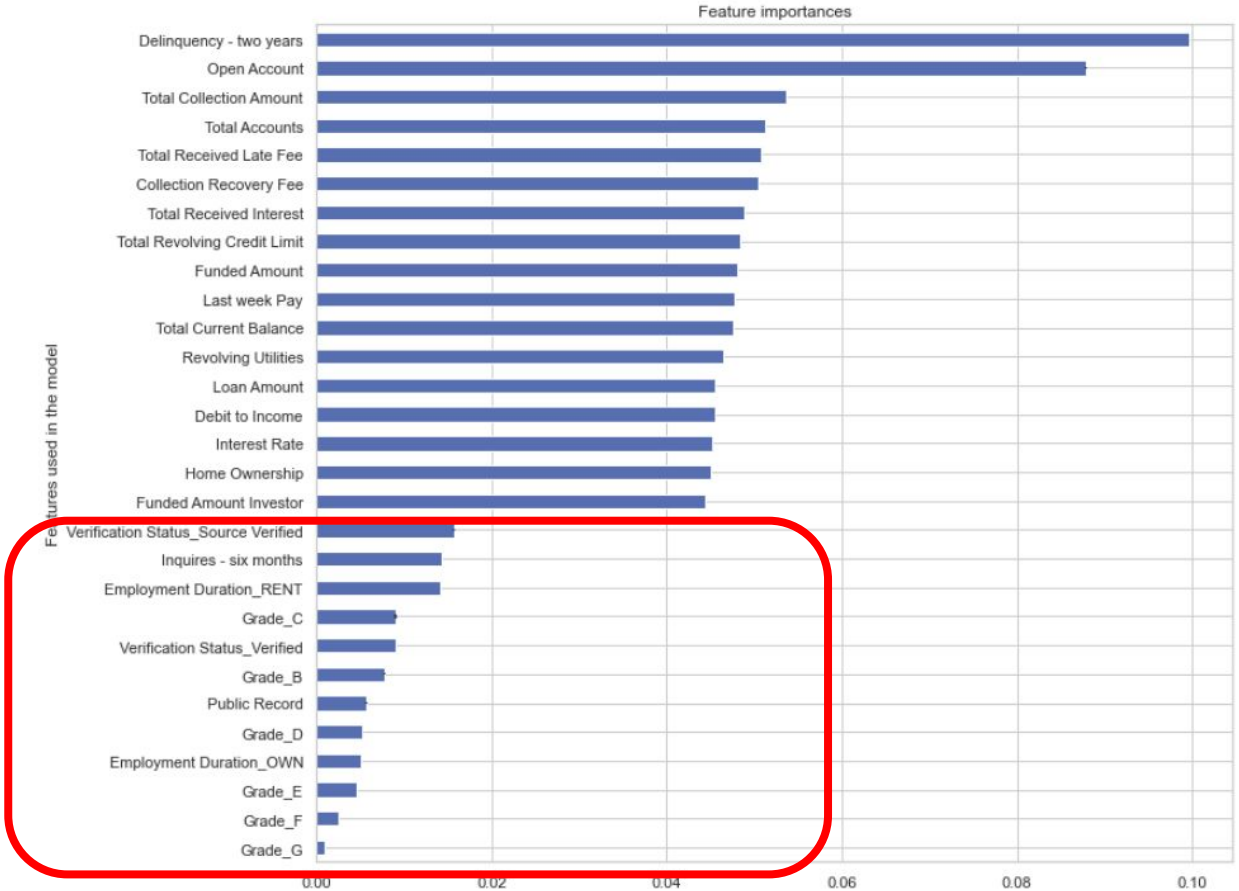
Oversampling + Cross-validation

	KNN	Logistic Regression	Decision Tree	**Random Forest**	XGBoost
Recall	0.179	0.134	0.118	0.095	0.119
Precision	0.092	0.100	0.102	0.089	0.098
F1	0.121	0.115	0.110	0.092	0.108
Accuracy	0.709	0.662	0.927	0.999	0.871

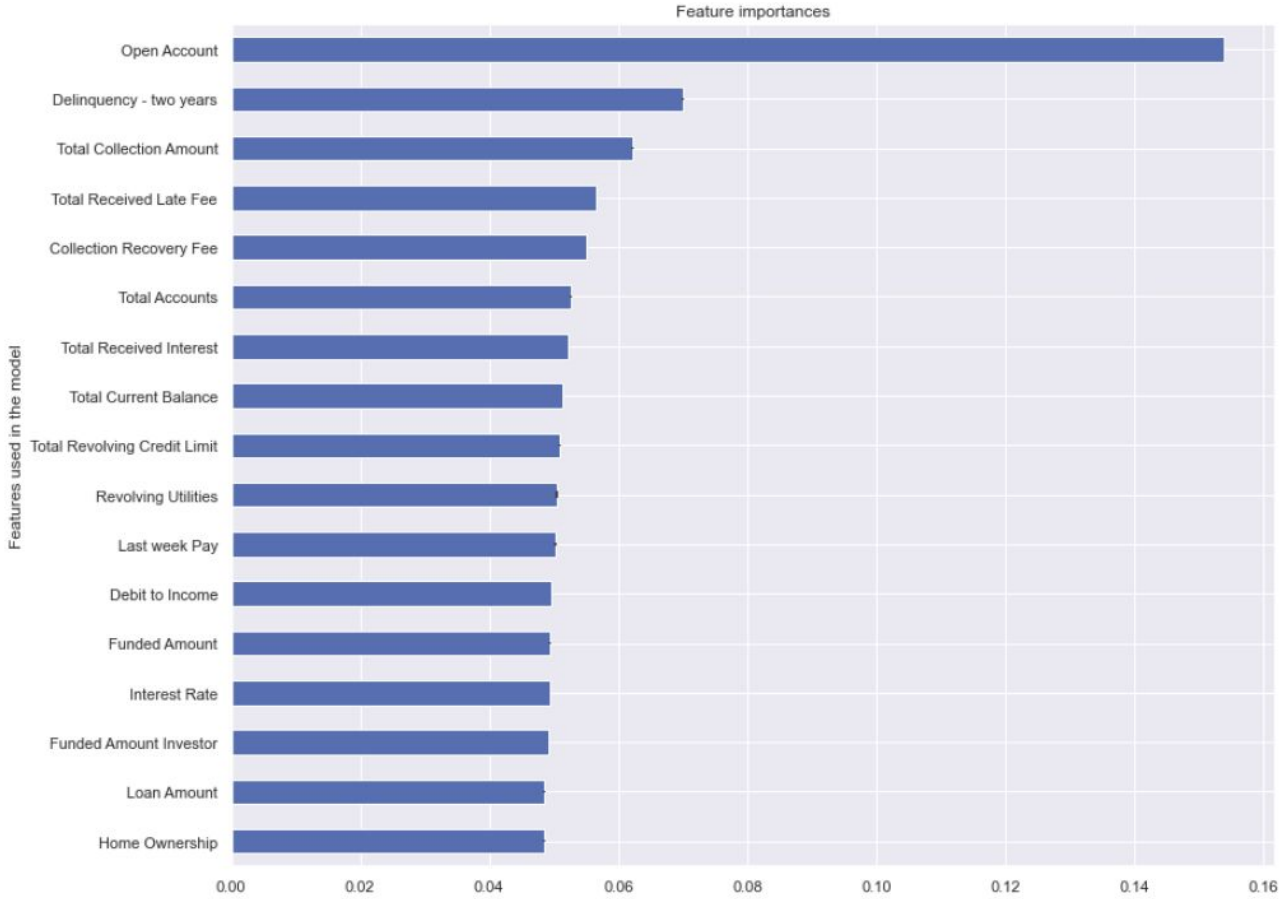
SMOTE + Cross-validation

	KNN	Logistic Regression	Decision Tree	**Random Forest**	XGBoost
Recall	0.261	0.155	0.992	0.139	0.062
Precision	0.092	0.088	0.090	0.089	0.075
F1	0.136	0.102	0.165	0.109	0.067
Accuracy	0.780	0.663	0.801	0.913	0.897

Optimizing Feature Selection



Optimizing Feature Selection



Final Model Interpretation

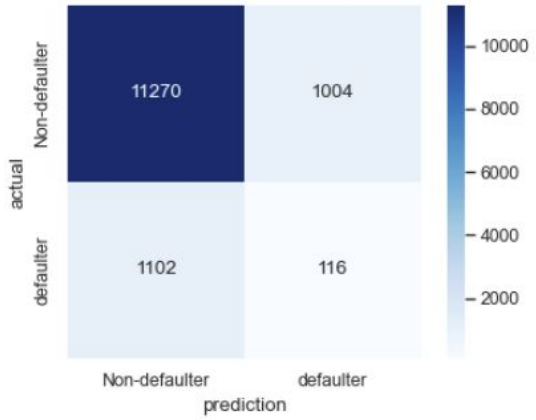
Model Parameters

'max_depth': 50,
'max_features': 'auto',
'min_samples_leaf': 1,
'min_samples_split': 2,
'n_estimators': 150

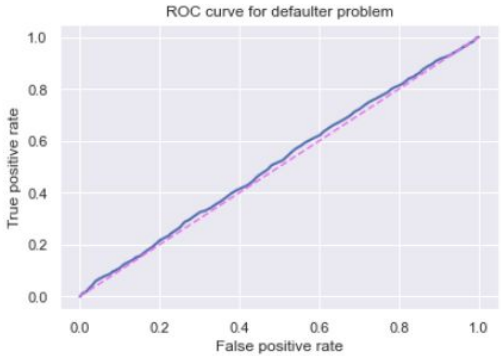
Results

'F1': 0.099,
'AUC': 0.515,
'Accuracy': 0.914,
'Precision': 0.104,
'Recall': 0.095

Confusion Matrix



ROC Curve



Conclusion & Future work

1. Personal Informations are more important features comparing to Loan Informations

2. The final model have 51% chance to distinguish between Loan defaulter and non-defaulter.

3. To adjust the model with collecting more valuable personal informations.

Thank you for your time!

Any Questions?