Loan Approval Prediction

Project IV: Model Deployment

Vivien Ho

Overview

Goal:

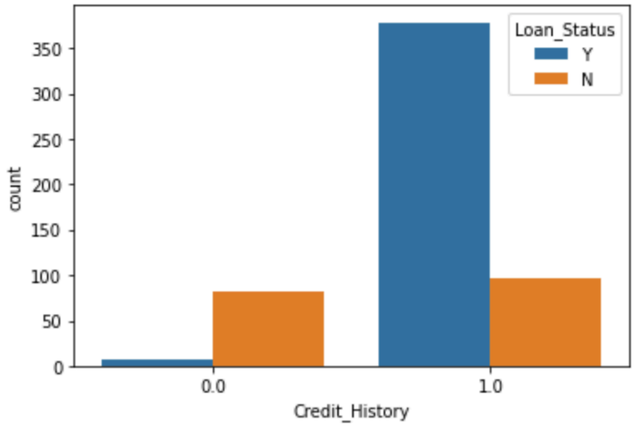
Created a model to predict which of the applicants will have their loan approved.

Process:

- Generate Hypothesis
- Data Exploration and Cleaning
- Feature Engineering
- Build Pipelines
- Serialization of models
- Flask app
- AWS Deployment

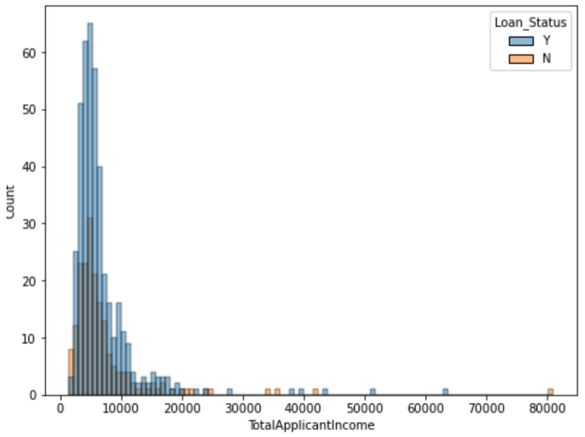
Exploring Factors Affecting Loan Approval

Credit History



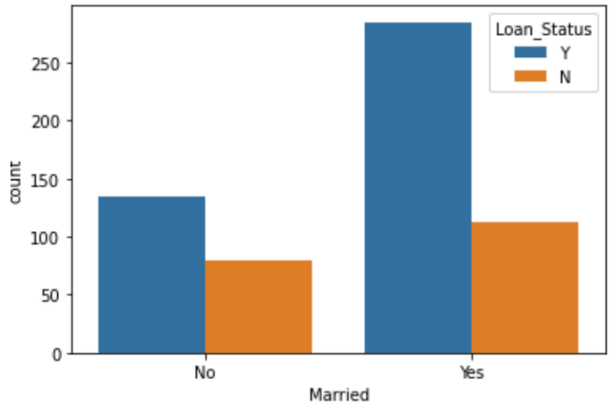
- Majority of the applicants who have a loan approved, have a credit history.

Total Income



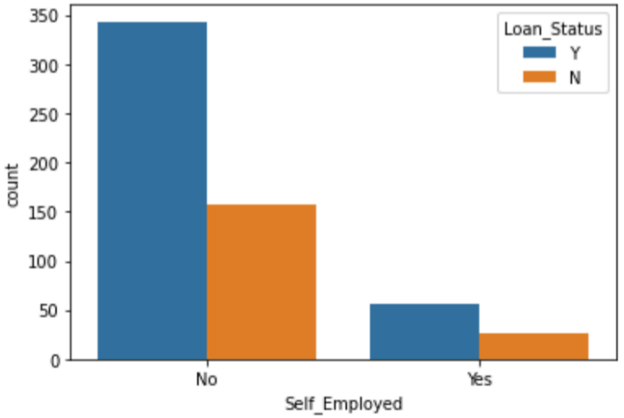
- Applicants with high total income equates to more loan count

Married



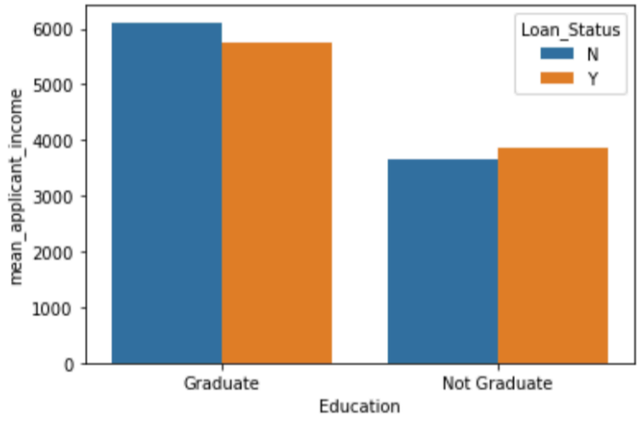
- More married people have a loan

Self-Employment



- Less people who are self-employed have a loan

Education Level



- Applicant's education does not seem to affect their loan status.

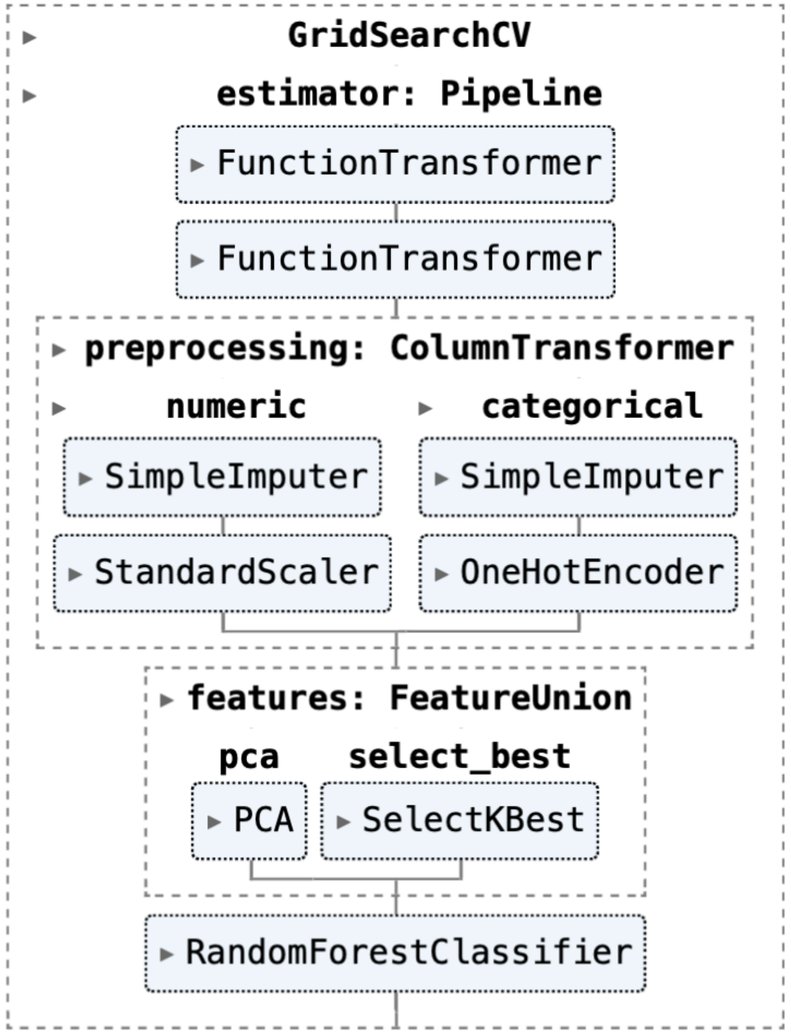
Pipeline Development

Preprocessing:

- Create total income - function transform
- Log 'Loan Amount' and 'Total Income' - function transform
- Numeric (mean) and Categorical features (mode) - column transform
- PCA (n=3) and Select K Best (k=6) - feature selection and union

Model:

- Random Forest



	Credit_History	Loan_Amount_Term	LogLoanAmount	Log_Total_Income	Gender_Male	Married_Yes	Dependents_1	Dependents_2	Dependents_3+	Education_Not Graduate	Self_Employed_Yes	Property_Area_Semiurban	Property_Area_Urban	Loan_Status_Y
0	1.0	0.279851	0.250640	0.008468	1	0	0	0	0	0	0	0	1	1
1	1.0	0.279851	-0.020226	0.082903	1	1	1	0	0	0	0	0	0	0
2	1.0	0.279851	-1.355202	-1.217363	1	1	0	0	0	0	1	0	1	1
3	1.0	0.279851	-0.150299	-0.301273	1	1	0	0	0	1	0	0	1	1
4	1.0	0.279851	0.174727	0.055266	1	0	0	0	0	0	0	0	1	1

Models

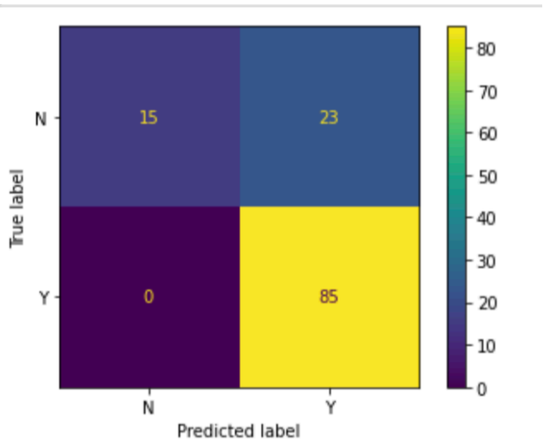
Models used in pipeline:

- Logistic Regression
- Random Forest
- KNN
- SVM

Logistic Regression	Accuracy	0.813
Logistic Regression	Recall	0.813
Logistic Regression	Precision	0.853
Logistic Regression	F1 Score	0.784
Random Forest	Accuracy	0.813
Random Forest	Recall	0.813
Random Forest	Precision	0.853
Random Forest	F1 Score	0.784
K-Nearest Neighbors	Accuracy	0.748
K-Nearest Neighbors	Recall	0.748
K-Nearest Neighbors	Precision	0.734
K-Nearest Neighbors	F1 Score	0.734
Support Vector Machines	Accuracy	0.813
Support Vector Machines	Recall	0.813
Support Vector Machines	Precision	0.853
Support Vector Machines	F1 Score	0.784

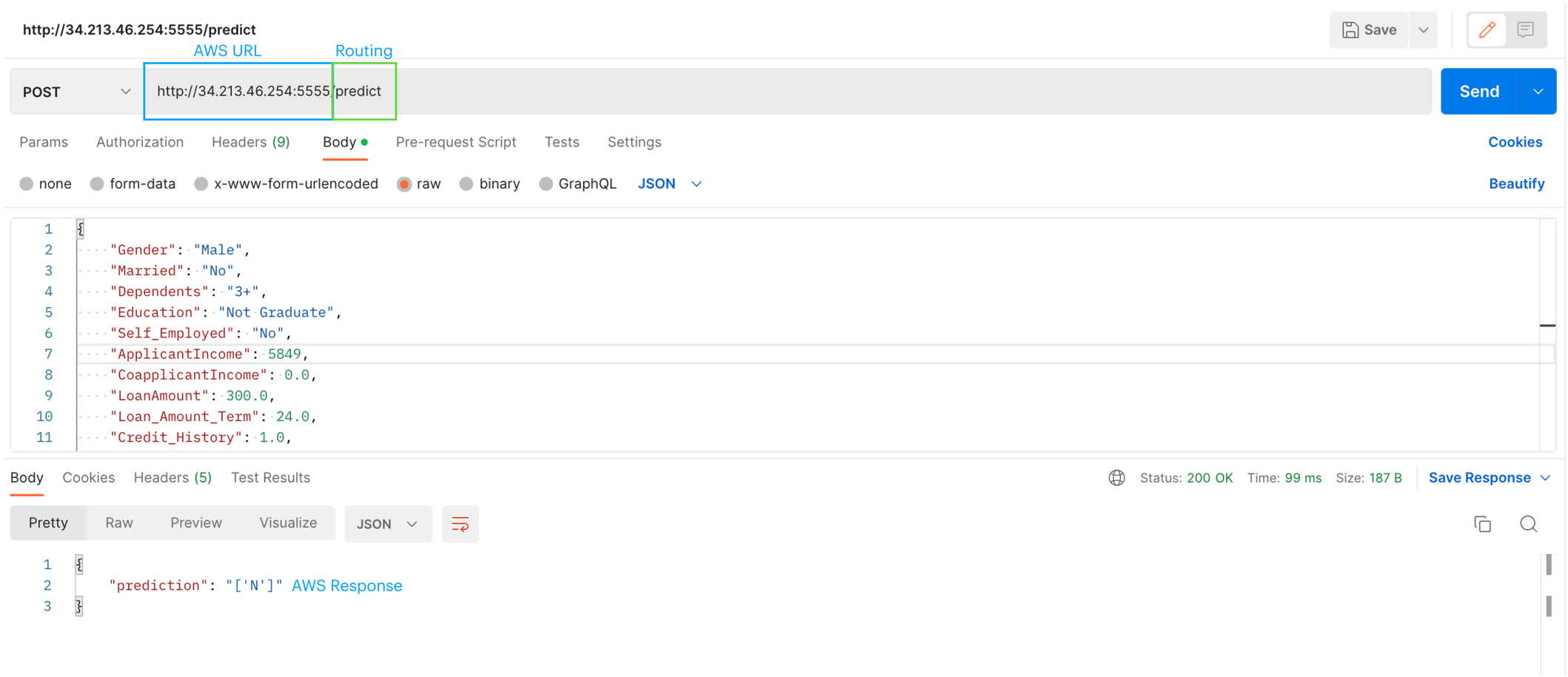
Summary:

- Models converge in accuracy and precision even if hyper-parameters are varied
- Selected Random Forest as model to deploy



Confusion Matrix for RF Model

Deployment to AWS



Challenges & Improvements

- Difficult and time-consuming to deploy to AWS

- Try different hyperparameters
- Try using different models to see their performance metrics

Thank you!