Which Loans Might Default? Lending Club Data Analysis and Prediction

Meskerem Goshime

September 22, 2022

Data Wrangling and Data Cleaning

➔ I started with loan data from 2007-2015 with 74 columns and 759,339 rows.

➔ Columns with 50k or more missing values were dropped.

➔ Null values were imputed with 0 or 'Other' where it made sense.

➔ After that, remaining rows with missing values were dropped.

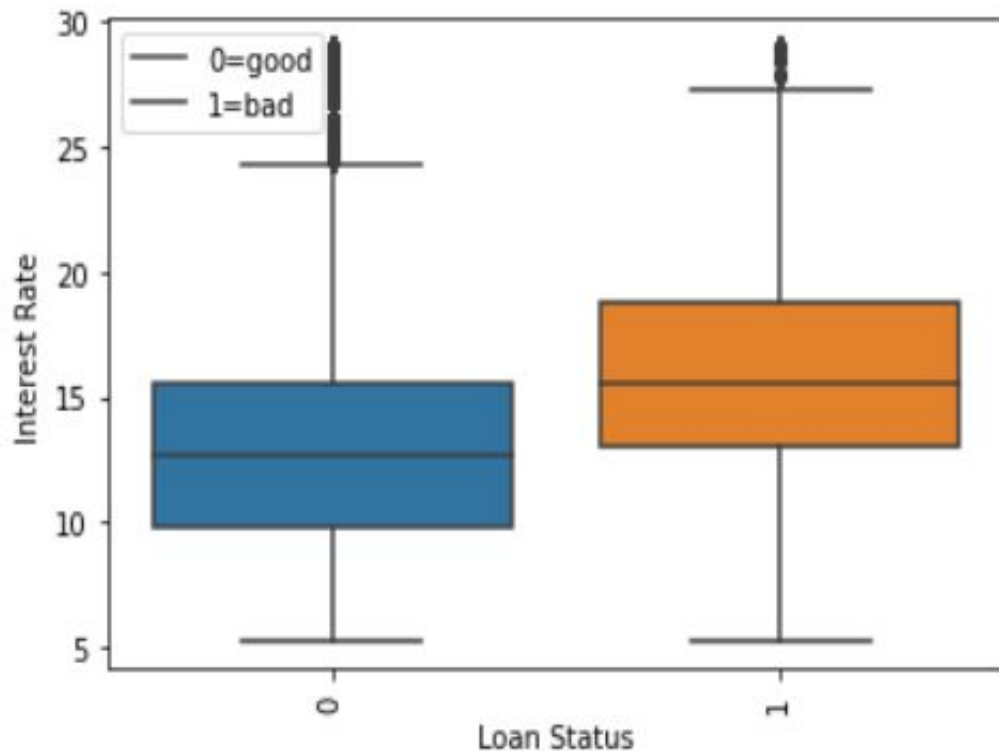➔ Took care of inconsistency in data entry by combining similar values in some columns.

Outliers Handling

➜ The most significant outliers were in annual_inc and dti columns.

➜ Max Annual income = 9 million, 99.7 percentile 379 k

➜ Max dti = 380, 99.7 percentile 39

➜ Rows with values beyond the 99.7 percentile in the respective columns were removed.
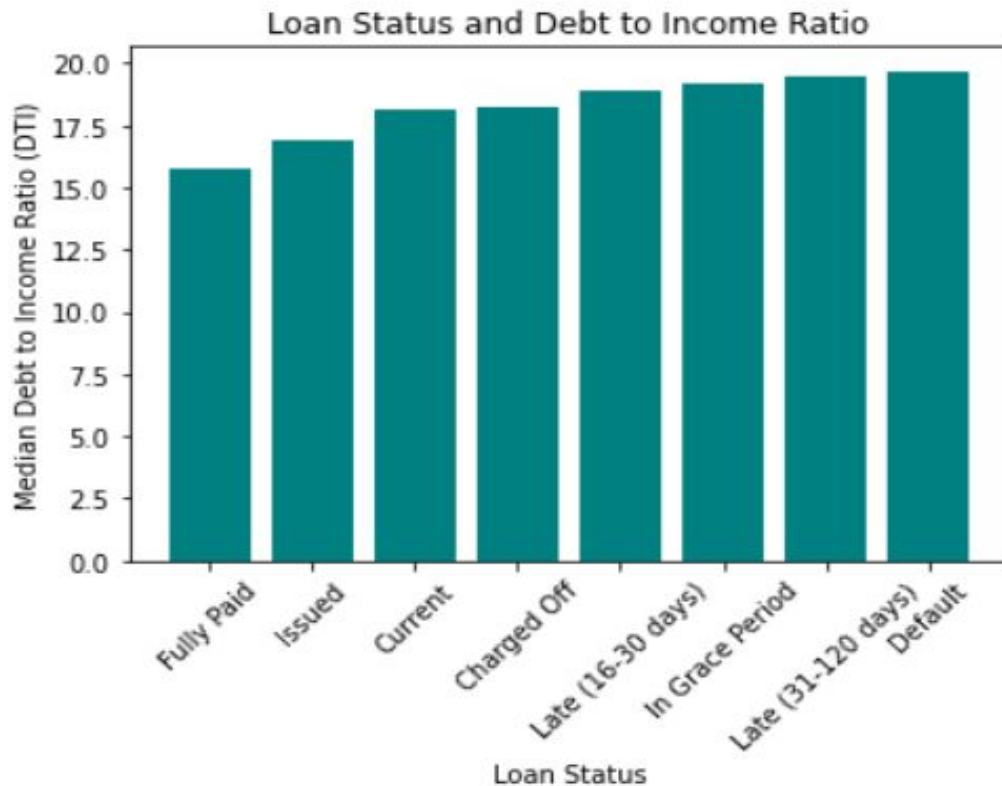
Feature Selection

➔ Columns with redundant data were dropped.
➔ Dropped columns that seem unnecessary based on my intuition from the data exploration
➔ I ended up with 33 columns from 74.
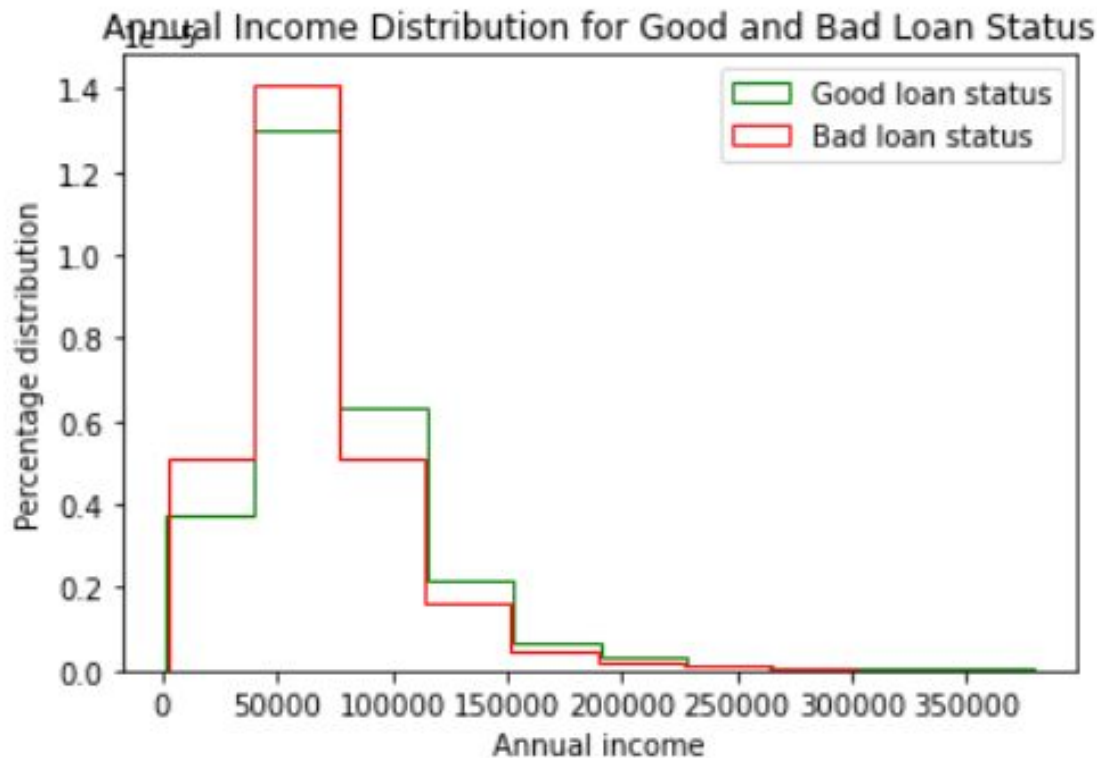➔ In the end I selected 10 columns out of the 33 using Feature Importance.

EDA - Interest Rate of Borrowers in Bad Loan Status is Significantly Higher

EDA - Bad Loan Statuses Correspond with Higher Median DTI Value

Loan Status and Debt to Income Ratio

A bar chart titled "Loan Status and Debt to Income Ratio" showing Median Debt to Income Ratio (DTI) on the y-axis (0.0 to 20.0) versus Loan Status on the x-axis. The categories from left to right are: Fully Paid (~15.8), Issued (~17.0), Current (~18.2), Charged Off (~18.3), Late (16-30 days) (~18.9), In Grace Period (~19.2), Late (31-120 days) (~19.4), Default (~19.6).

EDA - Lower DTI Corresponds with Higher Annual Income

Annual Income Distribution for Good and Bad Loan Status

EDA - Zip Codes with Highest Default Ratio

Zip Code	Loans in Bad Status	Loan Count	% in Bad Status
415xx	12	75	0.160
736xx	12	83	0.144
237xx	26	191	0.136
126xx	28	209	0.133
638xx	20	154	0.129
668xx	13	105	0.123

Preprocessing & Training Data Development

➔ Grade and sub-grade columns columns were encoded as numeric columns.

➔ The Categorical columns were encoded using One Hot Encoding.

➔ The numeric columns were standardized using StandardScaler.

➔ Target and Predictor Variables.

◆ Loan status was chosen as the target variable (y).

◆ The rest of the columns became the predictor variables (X).

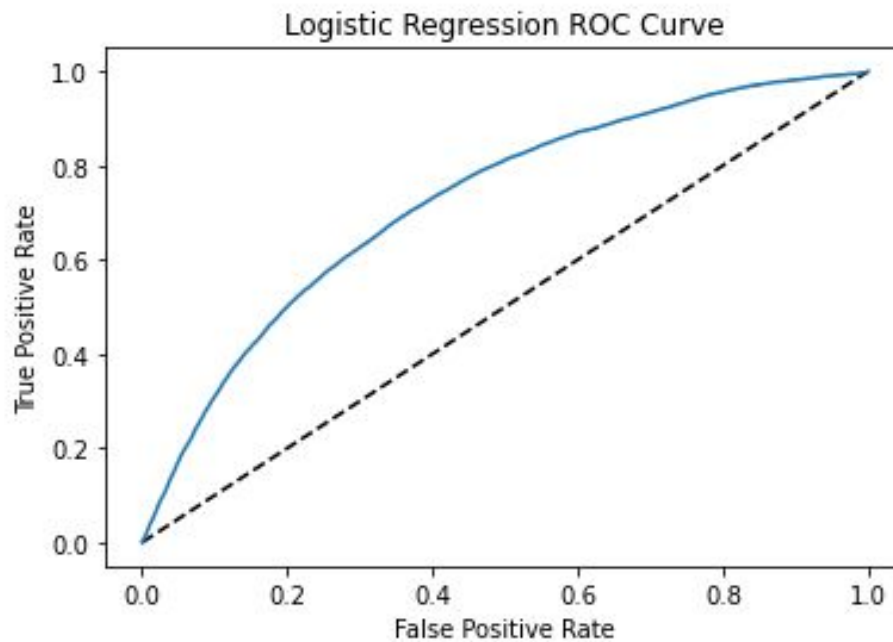➔ The data was split into Training Set, X_train, y_train (80%) and Test Set, X_test, y_test (20%).

Modeling - Random Forest Classifier

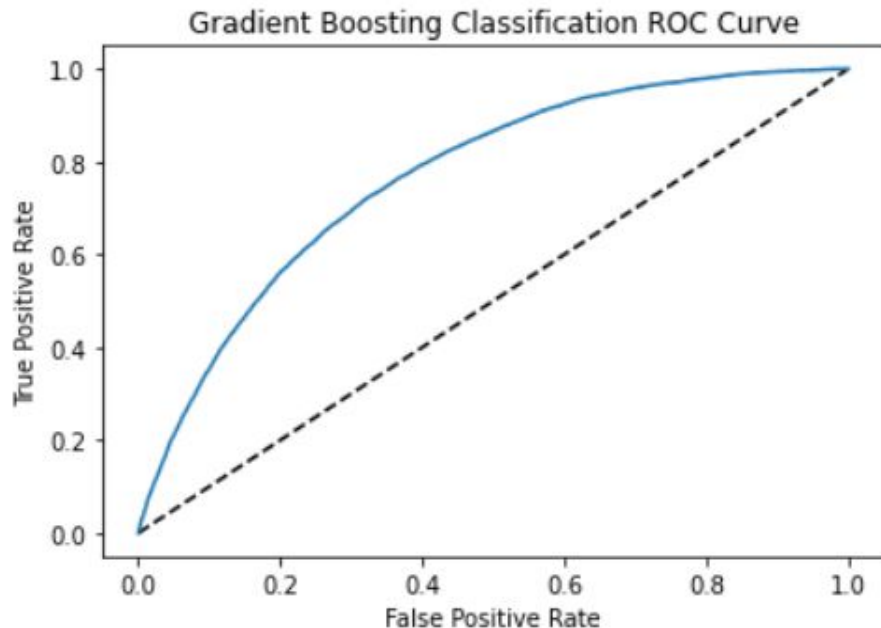➔ Chosen n-estimator value: 500

➔ ROC/AUC Score: 0.74

ROC Curve - Random Forest Classifier
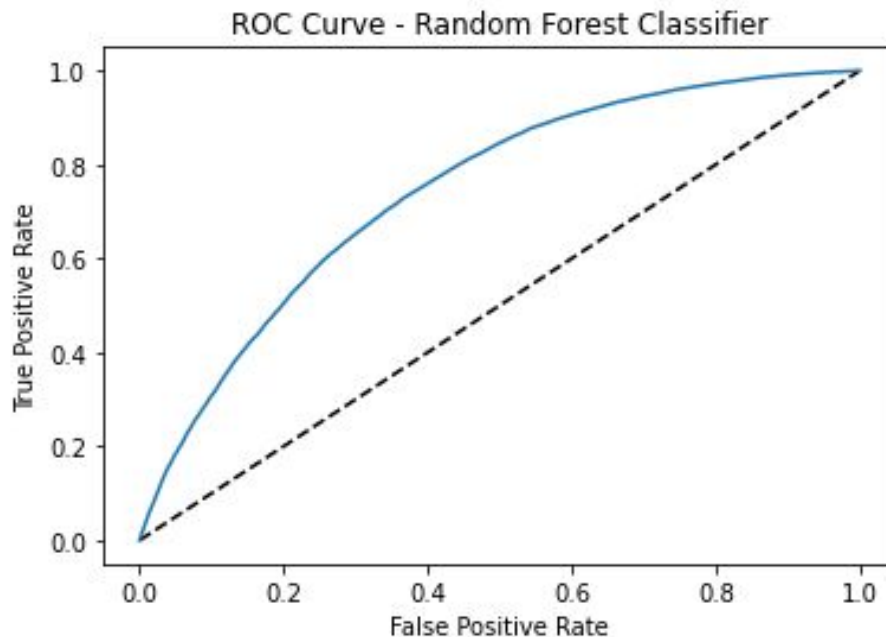
Modeling - Logistic Regression

➜ ROC/AUC Score: 0.72

Logistic Regression ROC Curve

Modeling - Gradient Boosting

➔ n_estimator value: 600

➔ Max_depth: 3

➔ ROC AUC Score: 0.77

Gradient Boosting Classification ROC Curve

Best Performing Model - Gradient Boosting

Best performing model - Gradient Boosting

➔ Best ROC/AUC score 0.77
➔ Minimizes the false positives while also keeping the false negatives low.

Planned Improvements

➔   Try PCA dimensionality reduction to see if the performance of the model will improve.
➔   Try resampling method to handle the imbalance in the data.

"All models are wrong, but some models are useful."

George E. P. Box

Project Files

[Project Notebooks](#)
[Project Report](#)

Special Thanks

➔ For Husain Battiwala for making [the data available on Kaggle](#)!
➔ For Tony Paek for his amazing mentorship!
➔ For my husband and boys for their encouragement and support!