Lending Club Data Analysis

Predicting the Likelihood of Loan Default in Peer-to-Peer Lending

J. Slaga

- World’s largest peer-to-peer lending company
- Links investors and borrowers
  - Unsecured personal loans between $1000 and $40,000
  - Interest rates ranging from 5.6% - 35.8%
  - Default rates between 1.5% and 10%

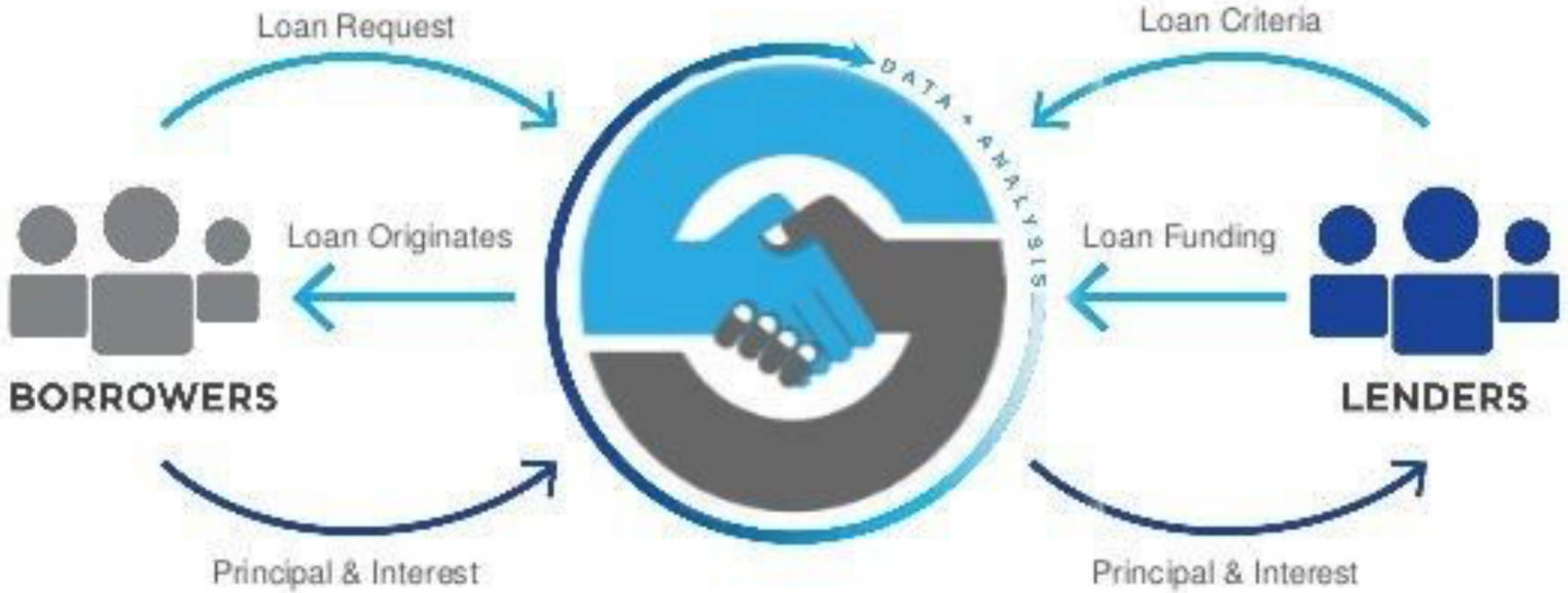
Peer-To-Peer Lending

- Lending Club is a peer-to-peer lending site matching lenders and borrowers through their online platform.
- Lower overhead costs allow for mutually a beneficial relationship:
  - Borrowers obtain loans at lower interest rates
  - Lenders earn higher returns
- Lending Club charges borrowers an origination fee and investors a service fee
-

Peer-To-Peer Lending

- Interest rates are determined by risk.
  - Borrowers deemed risky will have higher interest rates.
  - Borrowers deemed less risky will have lower interest rates.
- For investors, return of investment is also tied to risk
  - Higher risk / higher interest provide higher return
  - Lower risk / lower interest provide lower return

Value Exchange between Borrowers and Lenders



Which loans to invest in?

- What is the balance between risk and return?
- Investors want the highest interest loans that will be paid in full
- Goal: Create a machine learning model to predict likelihood of borrowers defaulting

Data

- Datasets obtained from Kaggle: https://www.kaggle.com/wendykan/lending-club-loan-data
    - Data dictionary of all feature columns
    - Complete loan data for all loans issued 2007-2015
- Includes current loan status and features such as loan amount, credit scores, and borrower information
- Contains 60 numerical features and 18 categorical features for assessing default risk

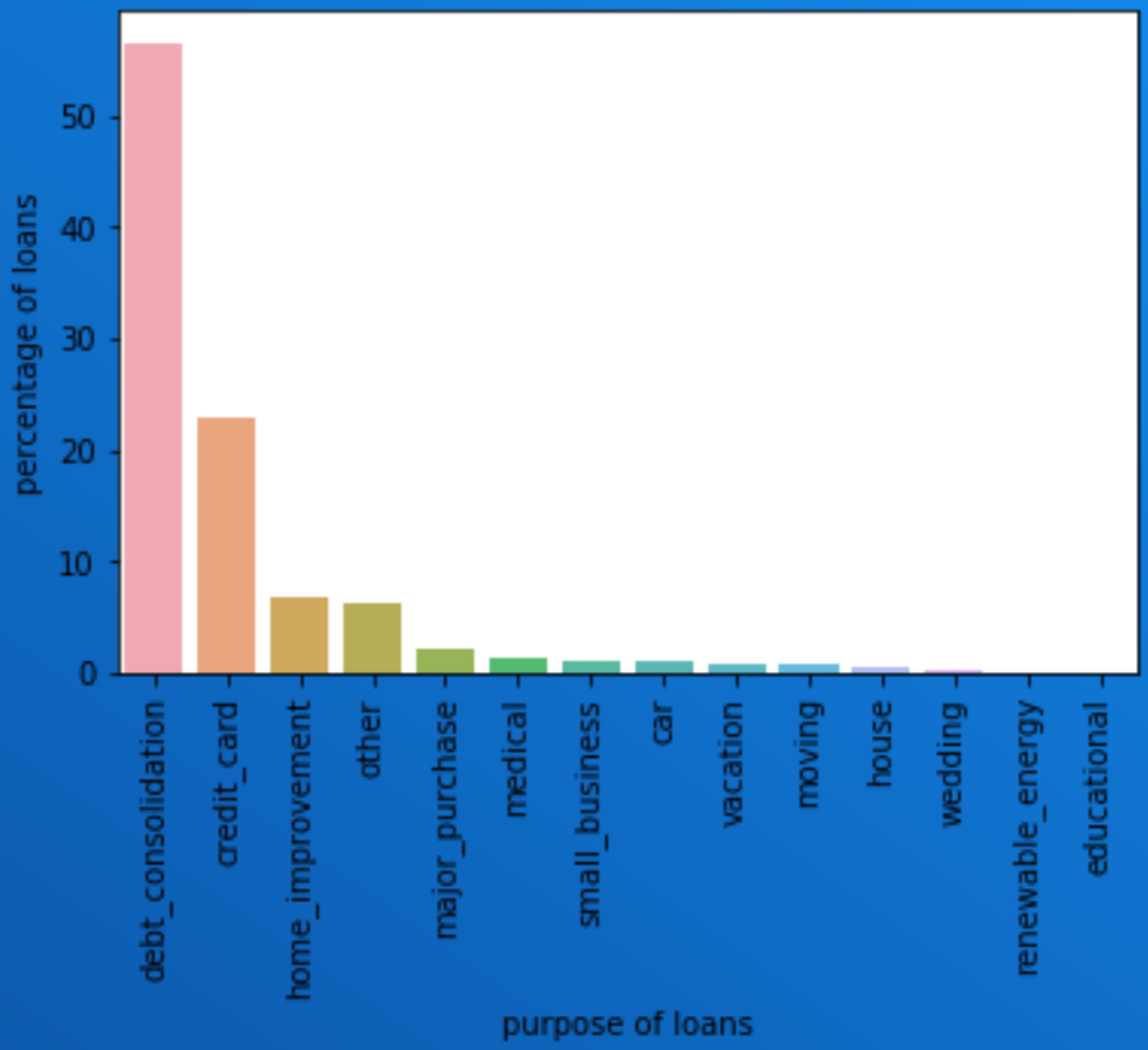
Data - Feature Information

- Profile variables: describe borrowers' basic information
  - Address, marital status, employment duration
- Financial history variables: describe borrowers' credit history
  - Annual income, open credit accounts and balances, bankruptcies
- Loan variables: describe loan
  - Loan amount, term, interest rate, assigned loan grade

Preliminary Explorations

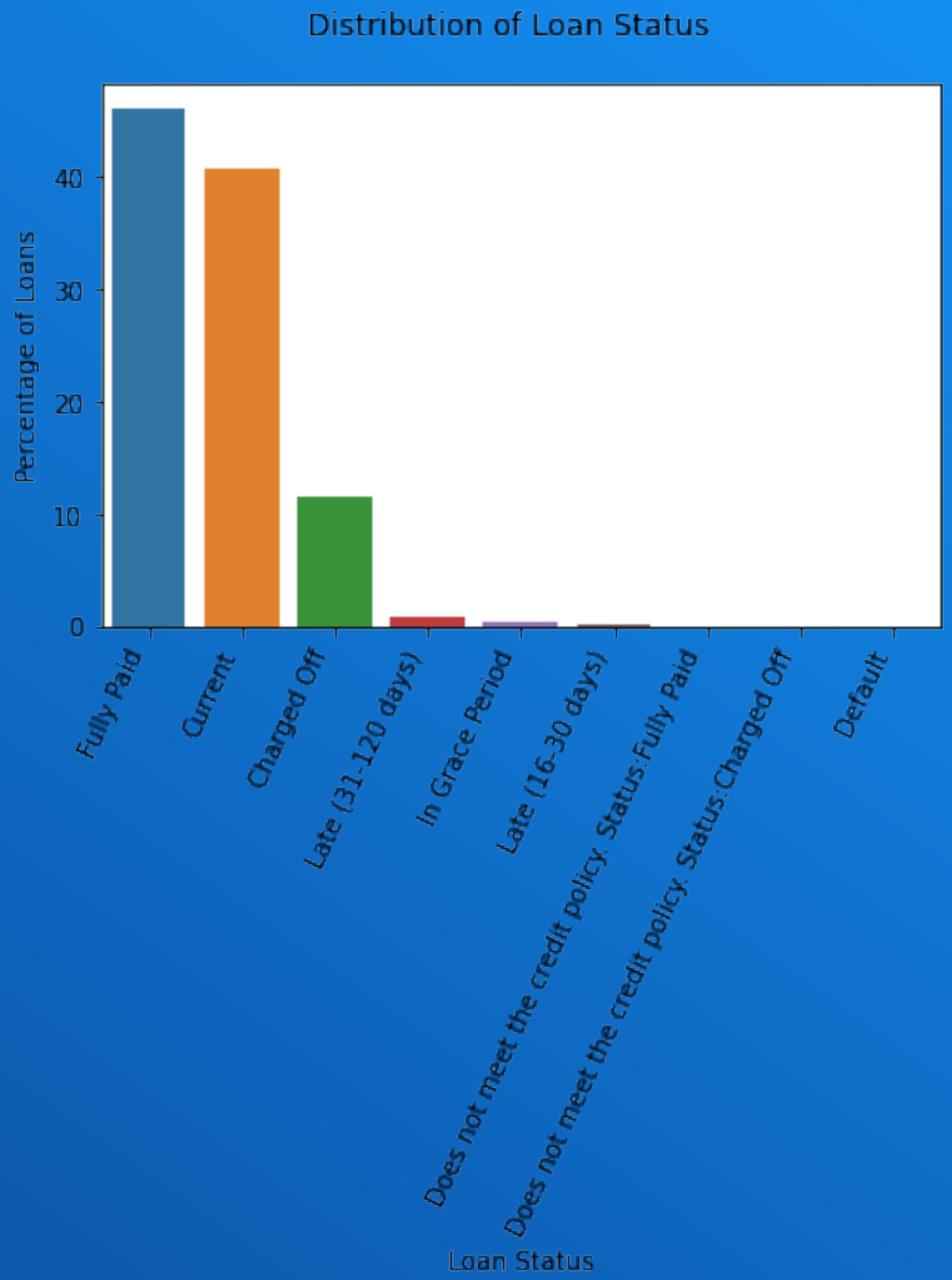
Top Purposes of Loans

- Debt consolidation (~60%)
- Paying off higher interest credit cards (~25%)
- Home improvement (<10%)

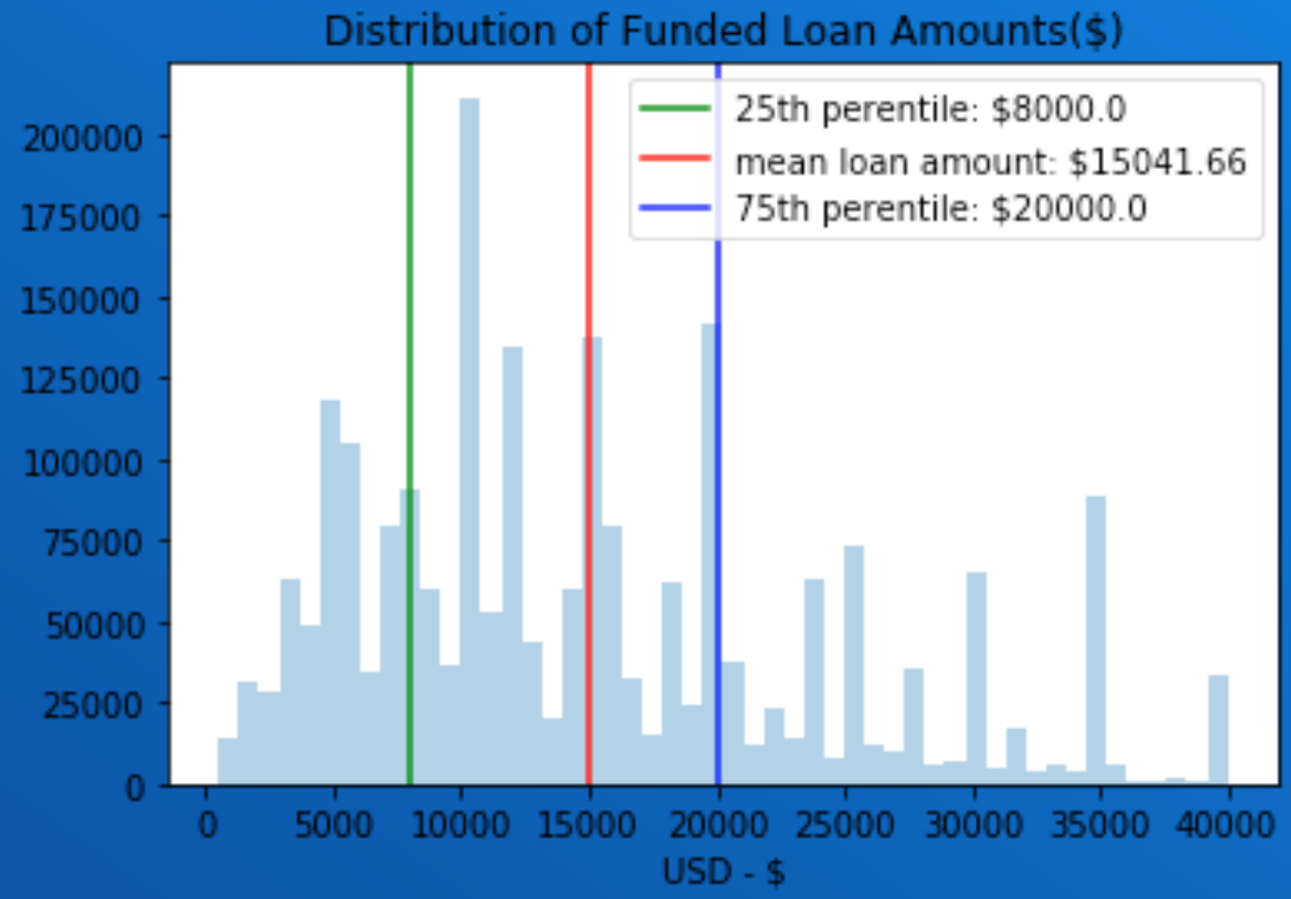


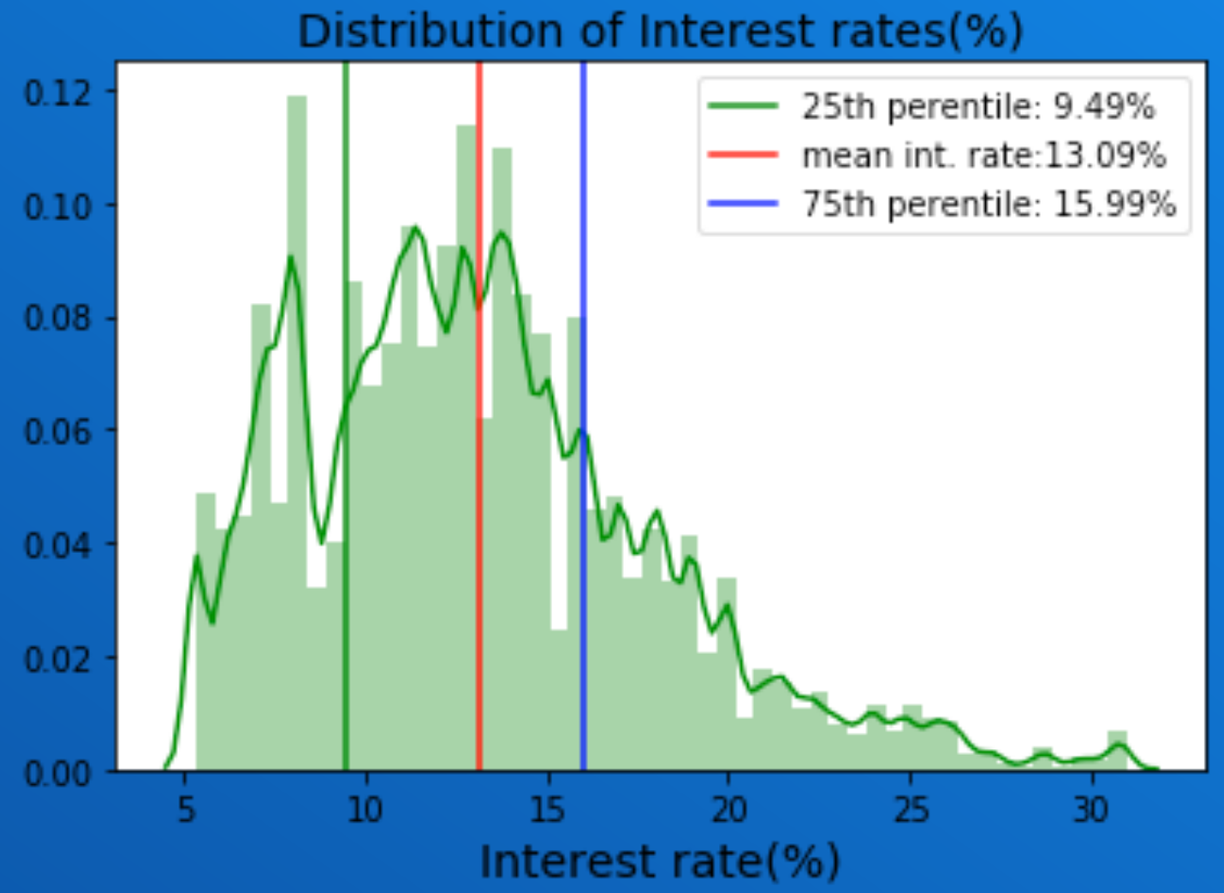
Preliminary Explorations

A distribution of the loan statuses shows that most loans (~88%) are fully paid off or current, while ~10% are charged off.

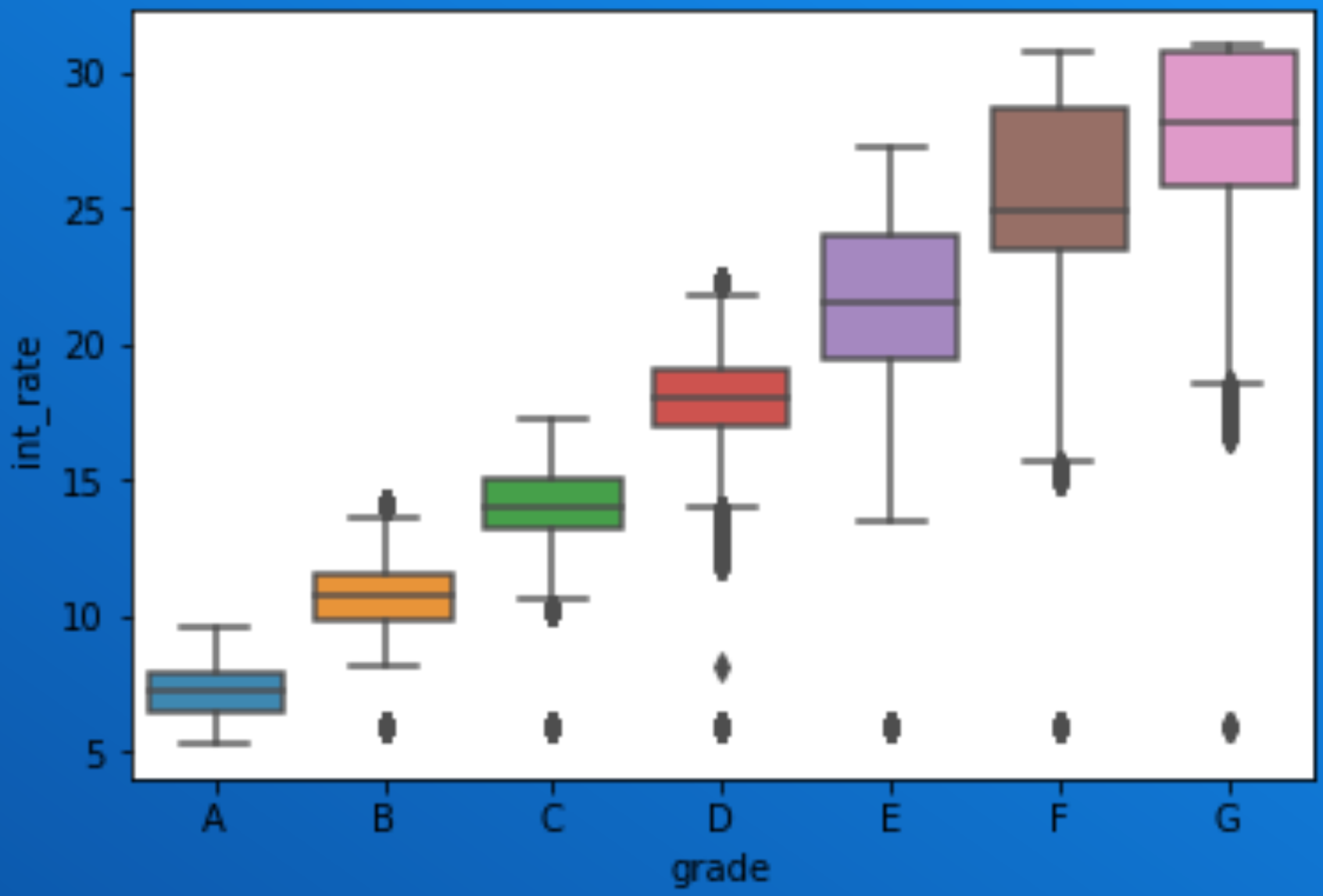


Preliminary Explorations





Preliminary Explorations

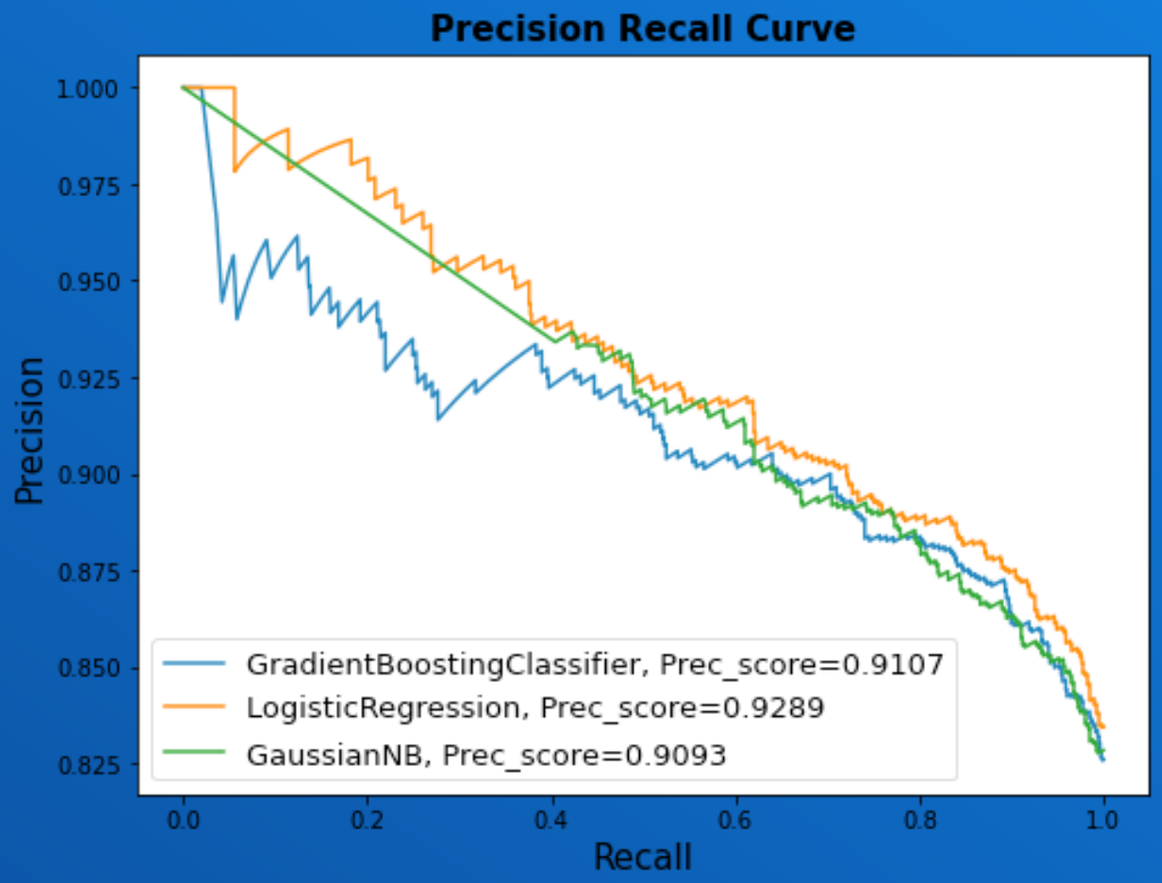


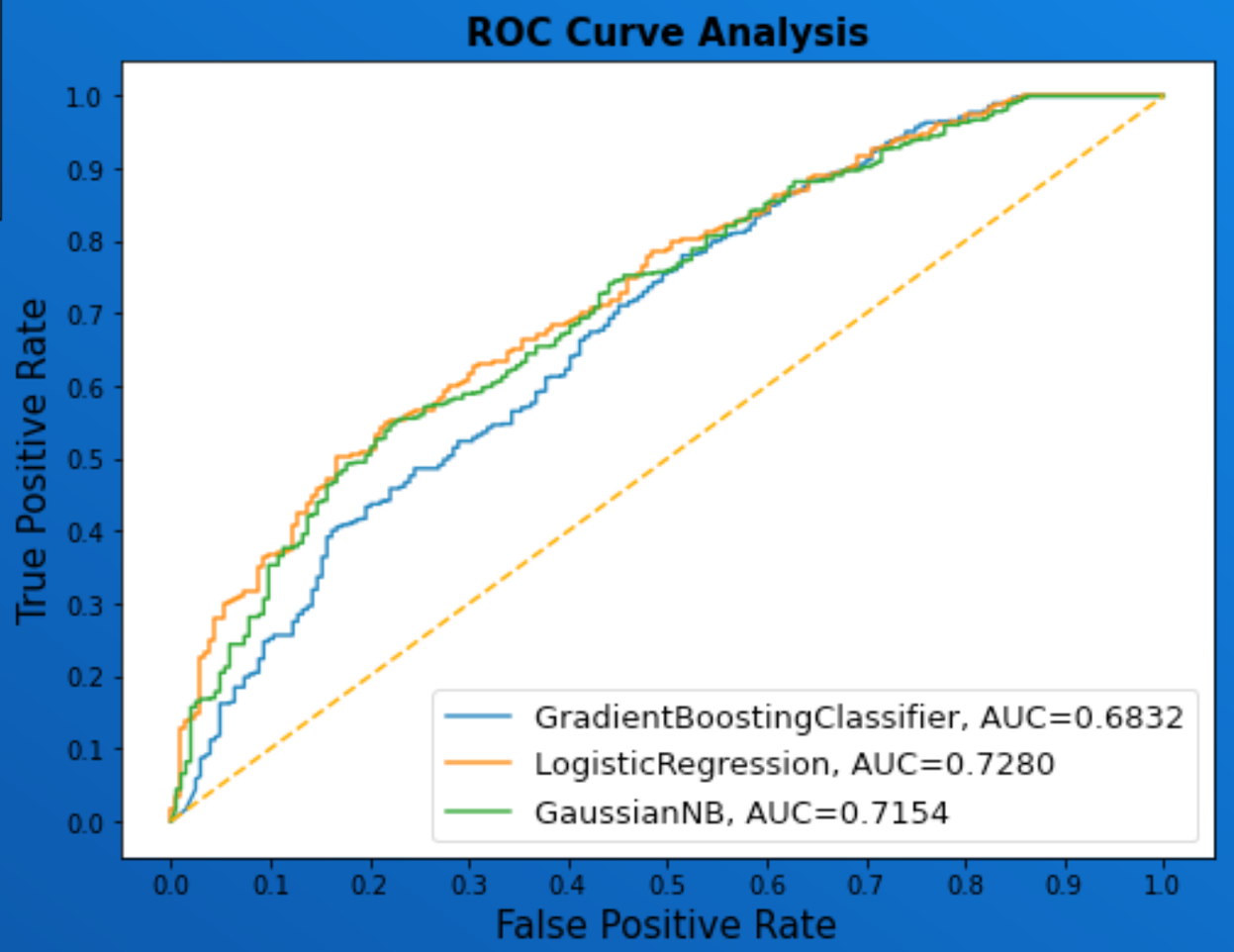
Lending Club assigns it a grade (A-G) to each loan according to its perceived credit risk.

Modeling

- Three classification models
  - Gradient Boosting Classifier (GBC)
  - Logistic Regression (LR)
  - Guassian NB (GNB)
- False positive rate (fpr), true positive rate (tpr), area under curve (auc), precision, recall, and precision score were calculated

Modeling


Precision Recall Curve
Precision
Recall
1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825
0.0
0.2
0.4
0.6
0.8
1.0
GradientBoostingClassifier, Prec_score=0.9107
LogisticRegression, Prec_score=0.9289
GaussianNB, Prec_score=0.9093


ROC Curve Analysis
True Positive Rate
False Positive Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
GradientBoostingClassifier, AUC=0.6832
LogisticRegression, AUC=0.7280
GaussianNB, AUC=0.7154

Conclusions:

- All three models have high precision scores and fall in an acceptable range for the ROC curve.
- Logistic Regression model performed the best.
- Next steps for improving the model should focus on refining the large number of features through feature importance and feature interdependence.

Thank you