Predicting Loan Default with Lending Club

THOMAS FRISS

THINKFUL DATA SCIENCE BOOTCAMP

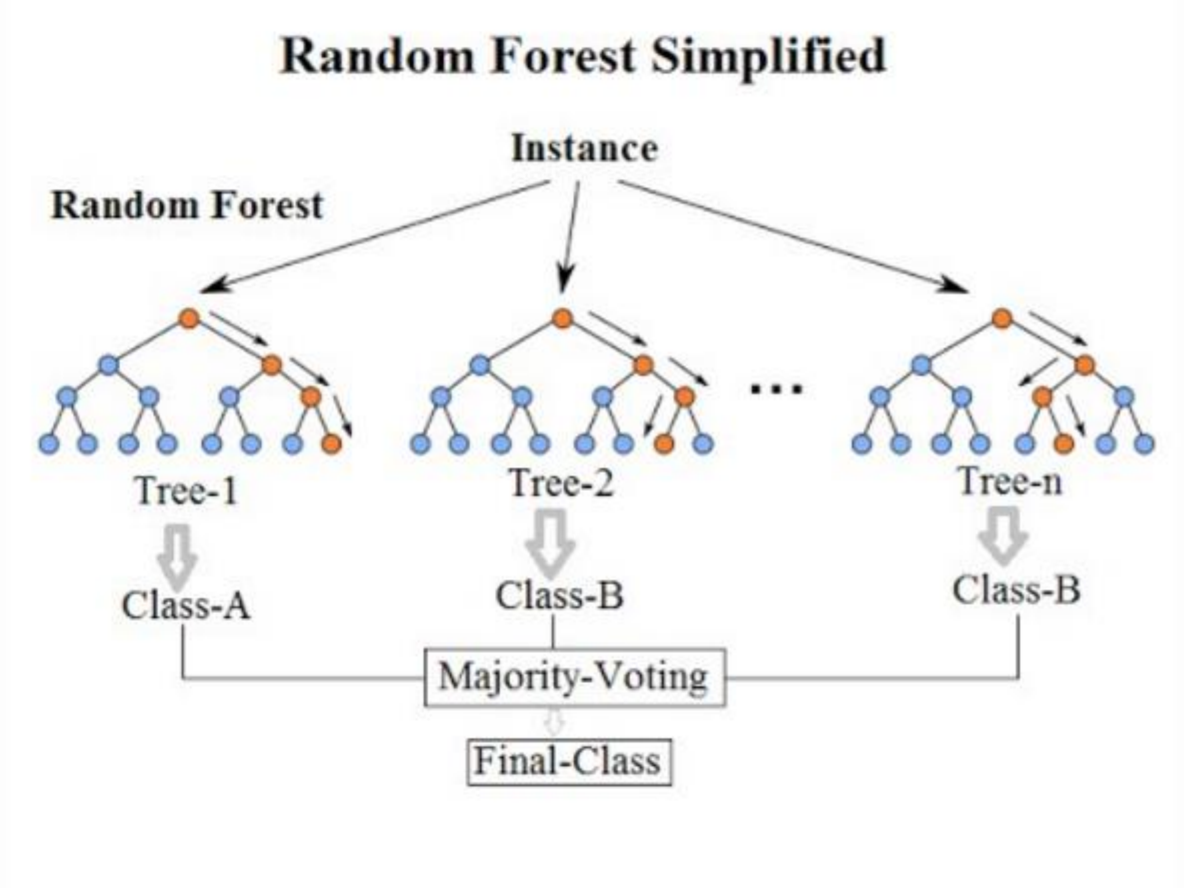
LendingClub

What is Lending Club?

- Lending Club is a financial institution that allows for investors to lend money to borrowers seeking loans for medical, business or personal reasons.
- Investors browse loan requests by prospective borrowers and choose whether or not to contribute to the loan request.
- Individual investors can invest as little as $25 per loan allowing investors with small amounts of capital to still create a portfolio of invested loans.
- Investors are able to examine information about the borrower and the loan itself to help in deciding whether or not to contribute to the loan.

My Project

- The goal of the project was to examine the data provided from Lending Club and see if a useful predictive model could be constructed from the information provided.
- Lending Club's data is publicly available on both Kaggle and the web page for Lending Club
- Several different models were used to predict loan performance with the most successful being a random forest model.
- Photo Source: https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d



Model Results

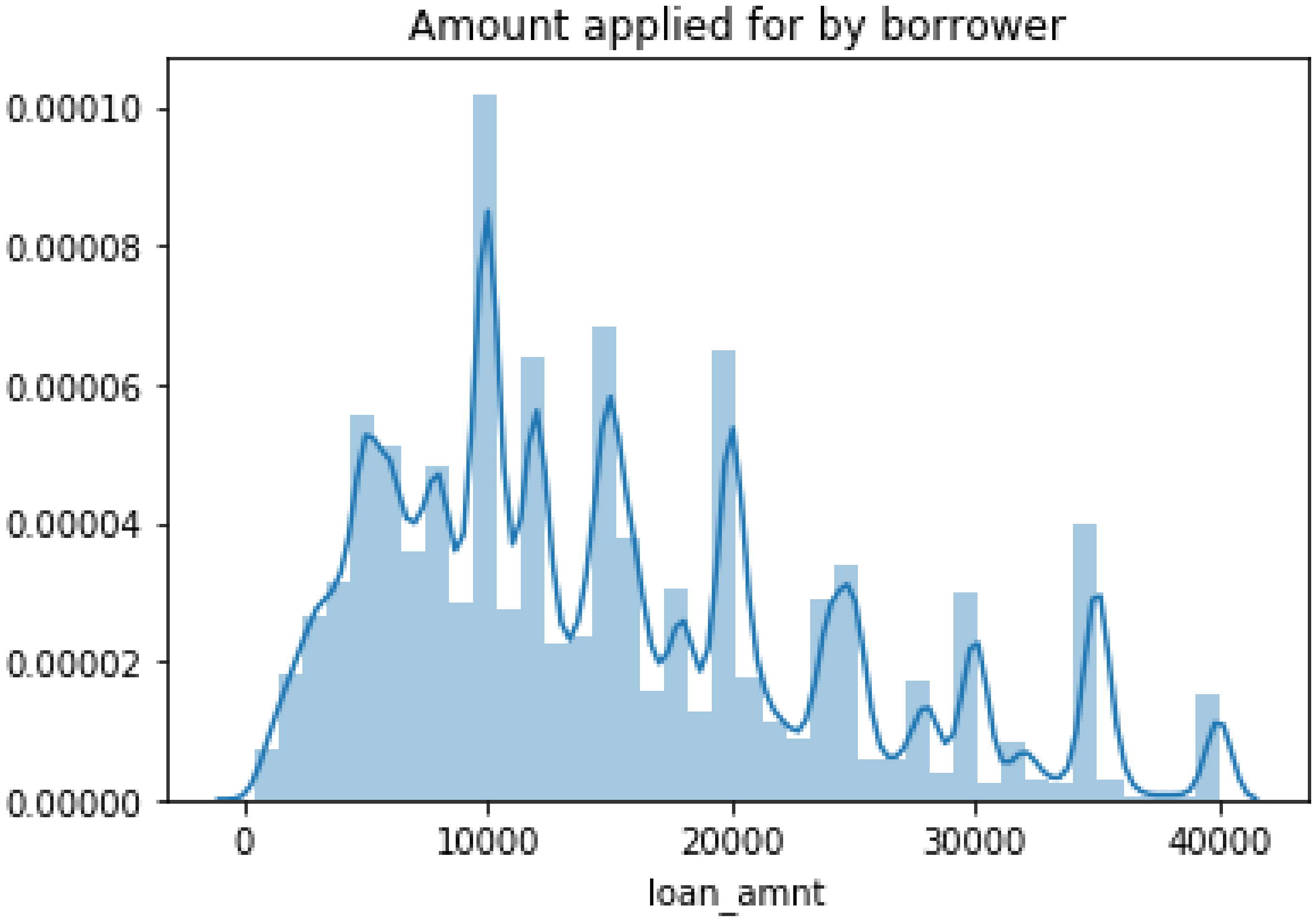
Model	Accuracy	Precision (True Positive)	Specificity(True Negative)
Random Forest 1	98%	99%	93%
Random Forest 2	98%	99%	93%
Gradient Boost	98%	99%	89%
XGB Classifier	98%	98%	87%
Gaussian Naïve Bayes	95%	95%	60%

Exploring the Data

- The dataset has approximately 2.26 million rows and 145 columns with 3 columns that only contain null values and an additional 36 non-numeric columns.
- The data required extensive and aggressive cleaning due to high percentages of null values
- After cleaning was conducted we were left with 2,013,799 rows and 66 columns

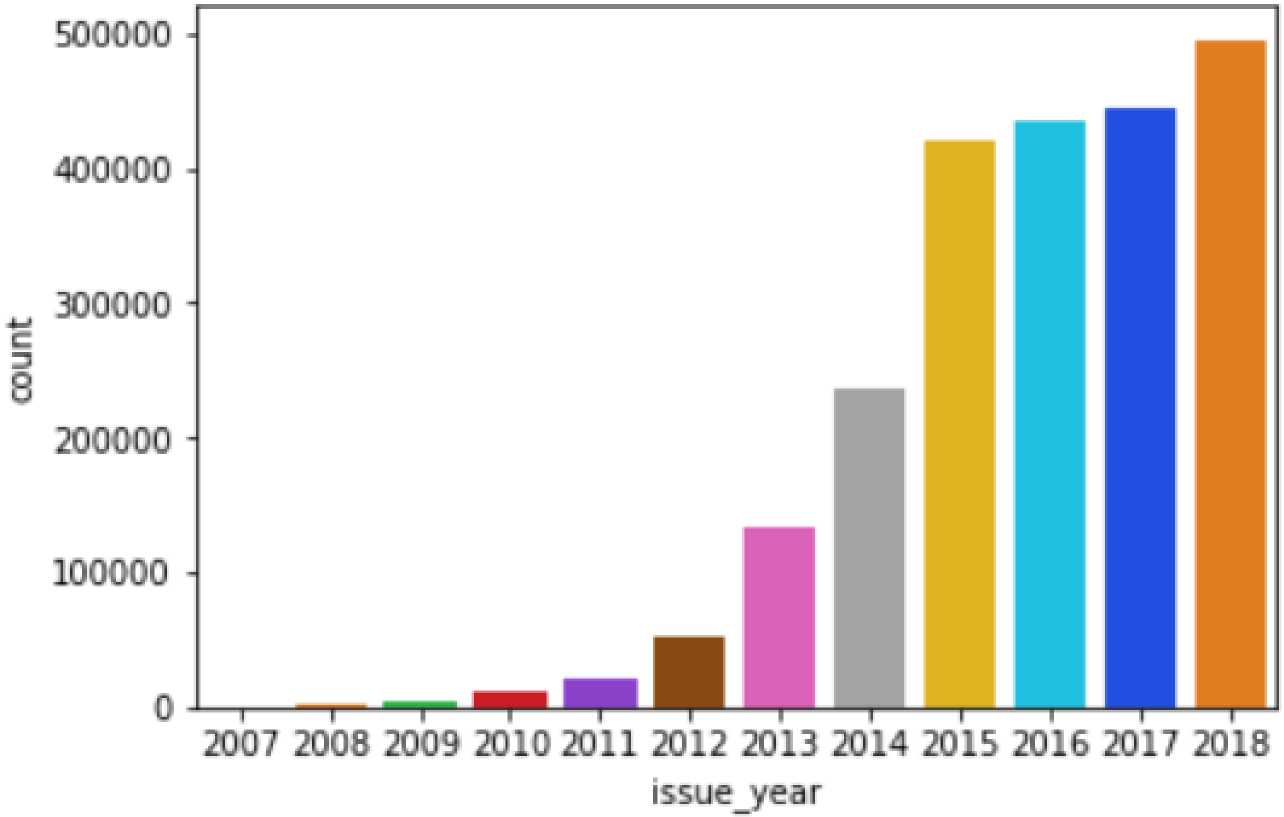
Loan Amount Distribution

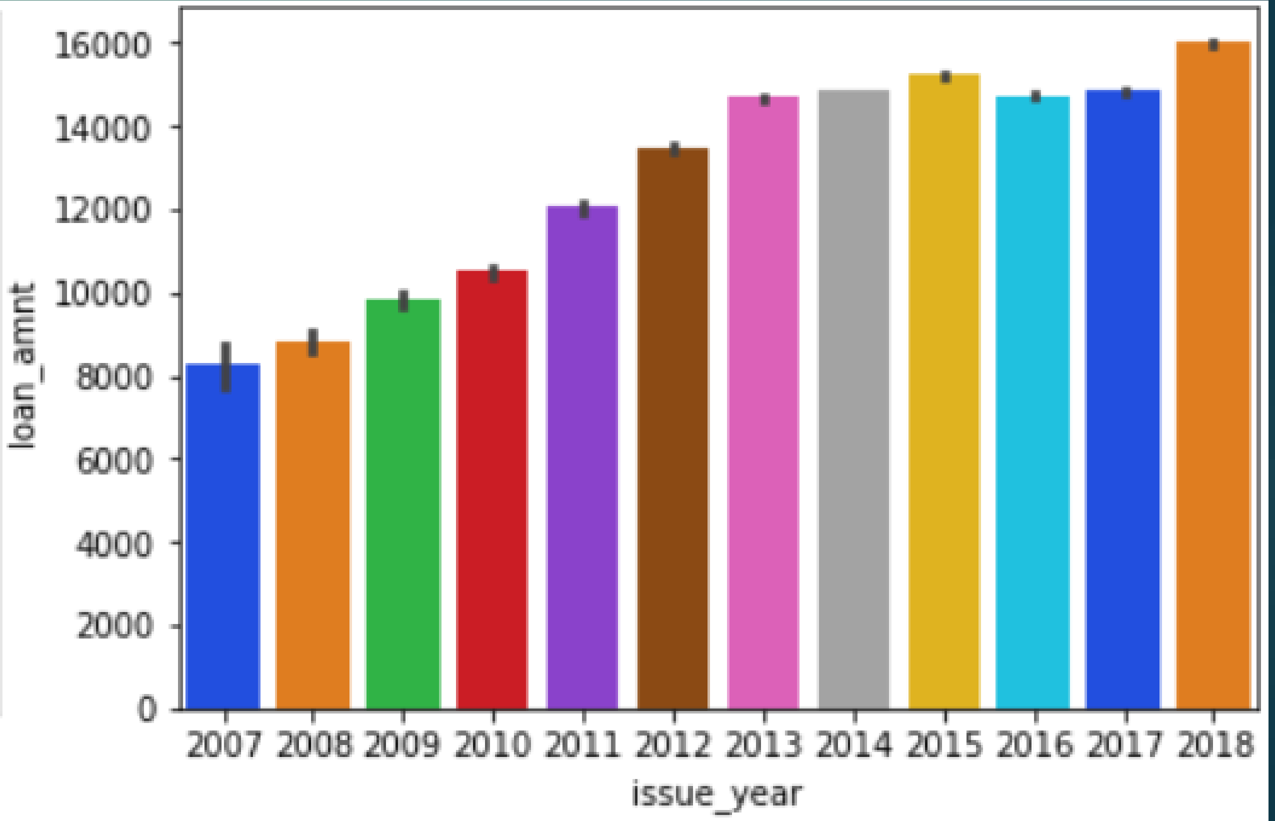
- We can see the distribution of loan amounts borrowers have applied for.
- There are upward spikes near increments of $10,000 with a general downward trend after $5,000



Company growth over time

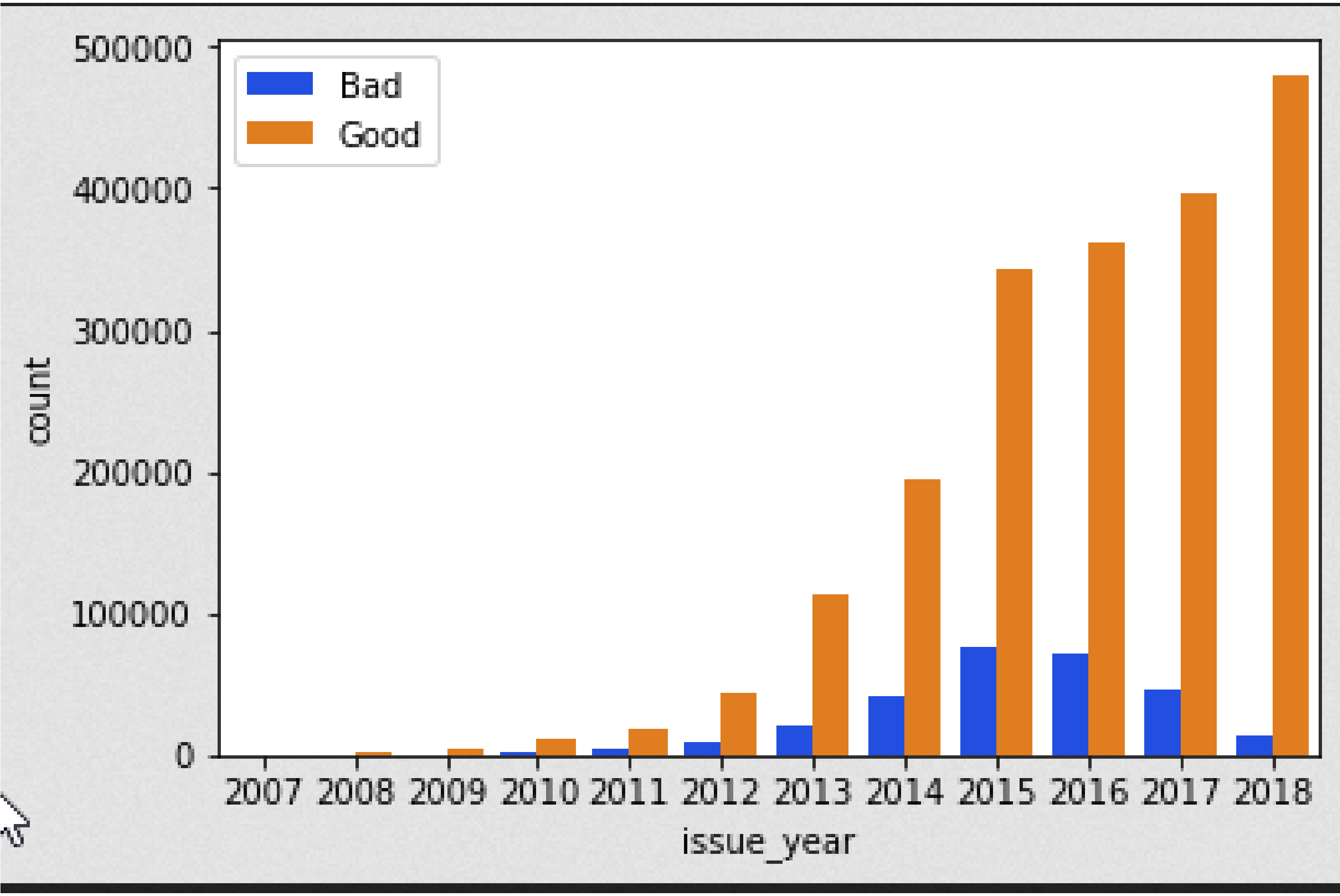
- The volume and average dollar value for each loan has increased since 2007 with the number of loans issues per year rising sharply between 2012 and 2015.





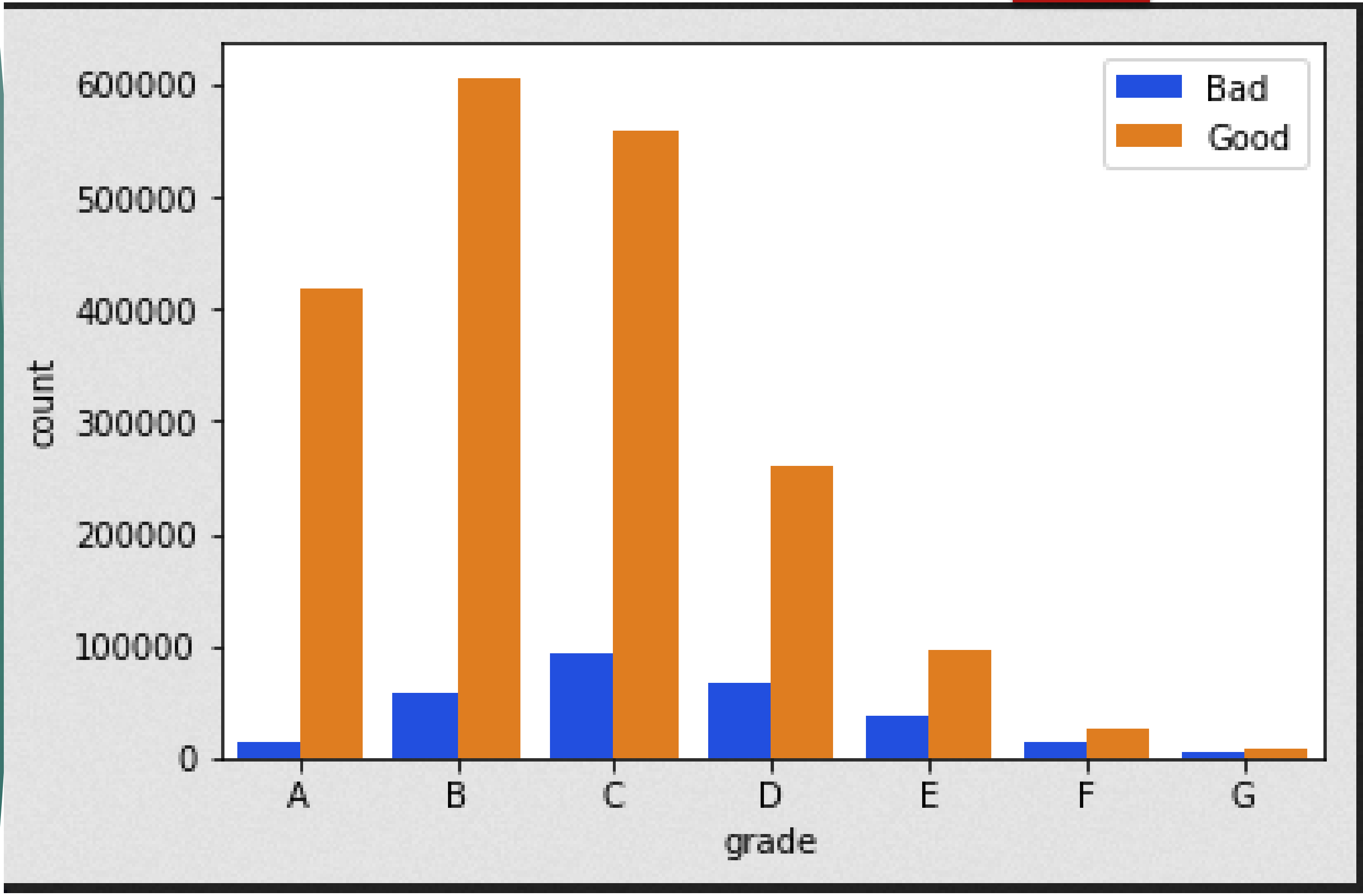
Good and bad loans by year

- On this graphic loans marked as “1”



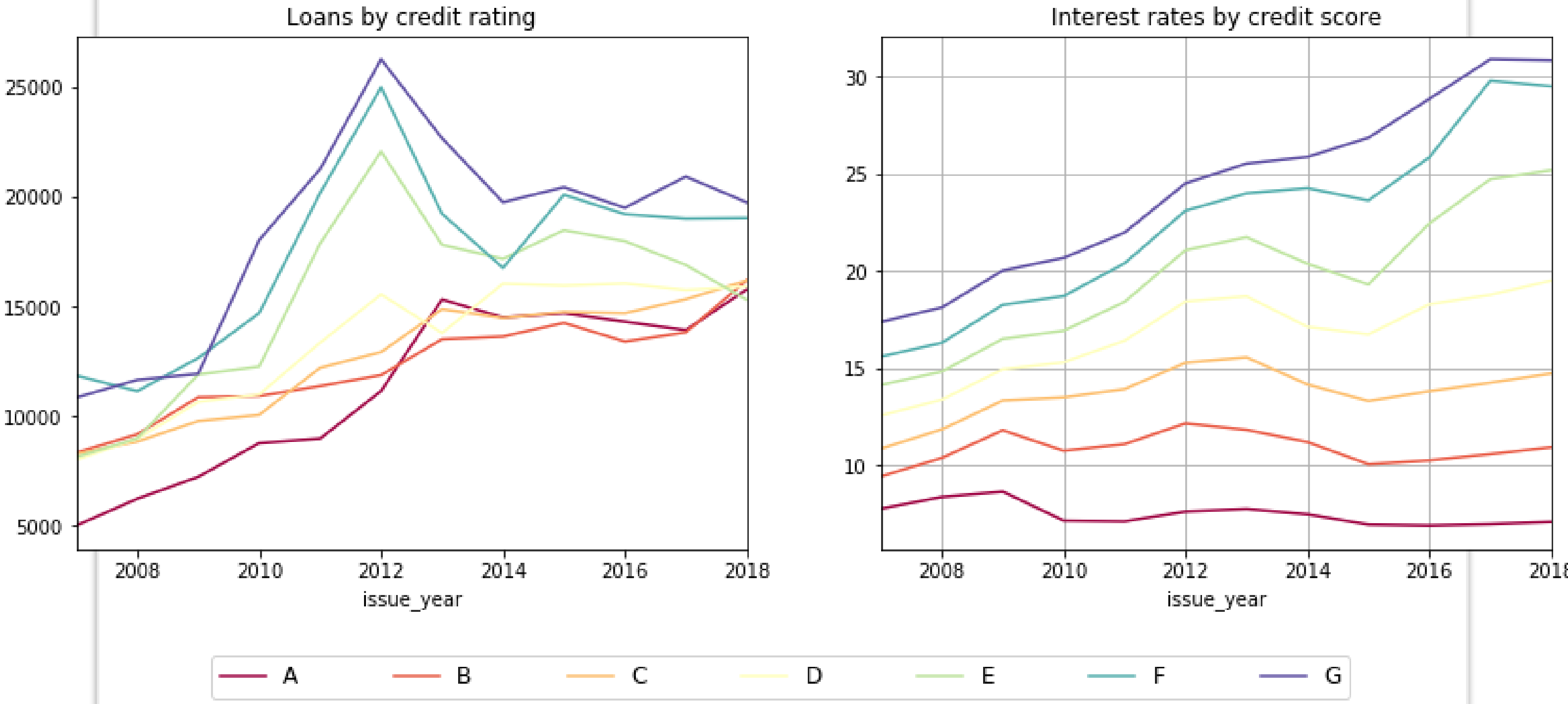
Loan Grades

- The notes are graded from A to E based on risk with A being the highest and E being the lowest.
- Riskier notes have higher interest rates but higher chance of default.



Loans by Credit Score

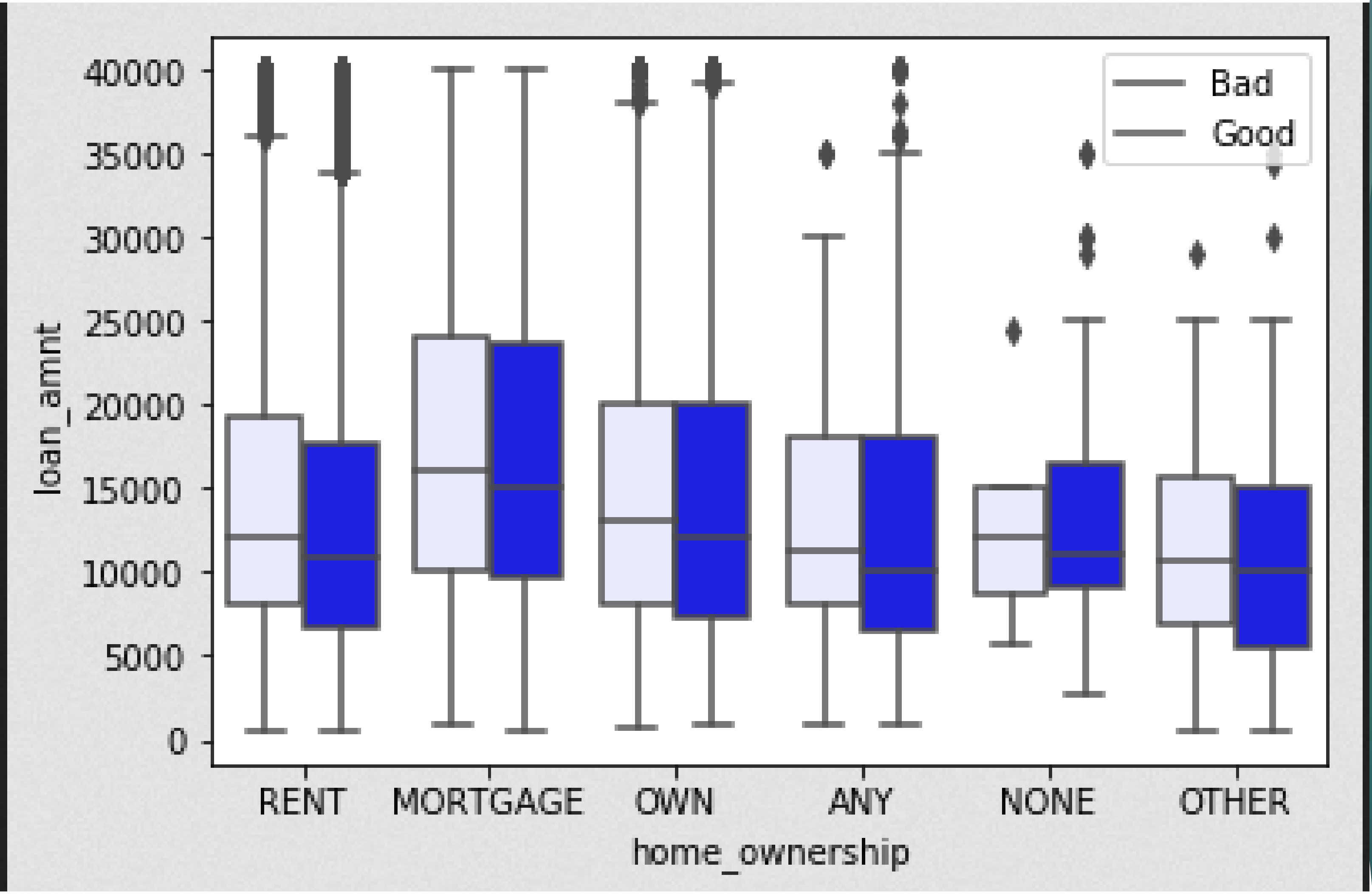
- An examination of the loans by credit rating. Loans are listed in the legend in order of safety with A being the safest and G being the most risky.



Loans by credit rating

Interest rates by credit score

Loan Statistics by Home Ownership Status



Analyzing the Data

- Approximately 14% of the loans issued through Lending Club go into late status or default.
- Before creating the models a new column was created with a simple 1 or 0 binary depending on if the loan stayed current (1) or went into some kind of unwanted status (0)
- This created a heavily imbalanced data set which would have created complications during modeling so the results were under sampled to create an even balance of good and bad loans to train the models on.

Random Forest Results

- The Random Forest Classifier created a model with a overall 95.9% accuracy.
- Number of Estimators: 100
- Misclassification Rate: 4.03%
- True Positive Rate: 99.14%
- False Positive Rate 7.21%
- True Negative Rate: 92.79%
- This model has the best specificity of the models tested.

Random Forest Confusion Matrix

Confusion Matrix	Predicted Bad	Predicted Good
Actual Bad	46,482	3,499
Actual Good	3,097	349,682

XGB Classifier

- 98.44% Overall accuracy
- 99.99% True Positive Rate
- 12.57% False Positive Rate
- Worse specificity than the random forest model but fantastic ability to identify good loans.

XBG Confusion Matrix

	Predicted Bad	Predicted Good
Actual Bad	43,939	6,561
Actual Good	5	352,255

Gradient Boosting Model

- Multiple learning rates were tested: (.05, .1, .25, .5, .75, 1)
- The best results used a learning rate of .75 and 1
- Accuracy: 95%
- Misclassification Rate: 4.48%
- True positive rate: 98.99%
- False positive rate: 7.96%
- True negative rate: 92.03%

Gradient Boosted Confusion Matrix

	Predicted Bad	Predicted Good
Actual Bad	44,845	5,655
Actual Good	206	352,054

Gaussian Naive Bayes

- Worst results of the group on both accuracy and specificity.
- Confusion Matrix.
- Specificity of 60%

	Predicted Bad	**Predicted Good**
Actual Bad	29,930	20,051
Actual Good	1,068	351,711

Further Questions

- A deeper dive would be able to compare the annual returns of using different models
- What are the characteristics of the loans by region and by state? Which states have the most Lending Club loans per capita and which states have the highest average default rate.
- Lenders who have a loan go into default typically receive at least a portion of their original investment back. A further consideration would be to create models that include how much of the loan was paid back before the loan went into default.
- An examination of the loans by each individual status could prove useful. According to Lending Club many of their loans that go into late status can be turned back into performing notes. What effect would this have on a hypothetical portfolio?