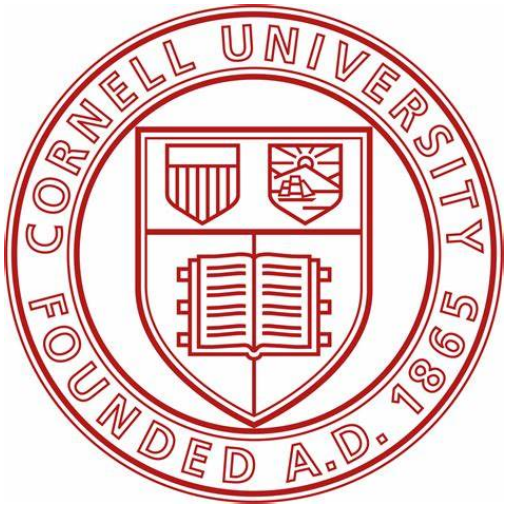
Interest Rate Prediction

Sam Shah (sns224), Kaitlin Tubbs (kt392), Kyle Walsh (kew96)



Cornell University
ORIE 4741
Ithaca, New York
10 December 2019

Interest Rate Prediction

Sam Shah (sns224), Kaitlin Tubbs (kt392), Kyle Walsh (kew96)

December 10th 2019

1 Introduction

The purpose of this project is to predict the correct interest rate applied to a loan requested by a given borrower with a certain history and characteristics. In the lending industry, interest rates are crucial in determining the riskiness of the borrower. A lender will apply a higher interest rate to a riskier borrower to account for the greater likelihood of default. The risk of default can be explained by a plethora of factors relating to the borrower, such as their annual income, history of default, the amount of current debt, etc. To predict the interest rate of a loan, we will train models using a dataset containing examples of accepted loan applications with the applied interest rate.

2 Data Analysis

2.1 Data Description

We will be working with the data set *All Lending Club Loan Data* from Kaggle, which contains 2007 through 2019 Lending Club accepted loan data. The dataset contains 2,260,701 data entries and 151 explanatory variables involved in assessing the interest rate for the loan. After filling expected missing values, there were very few missing entries in the raw dataset which allowed us to remove the remaining rows with missing values while training our models without sacrificing data integrity. We immediately cut out 112 columns that we knew were not significant, and we were left with 39 features. Of these remaining features that we used to train our models, 8 were nominal, 29 were categorical, 2 were ordinal.

2.2 Visualizations

Data visualization is a very useful tool to determine which features are important to predicting interest rates. The figure below graphs grade, loan amount, and interest rate and highlights the importance of the grade of the borrower in relation to interest rates. There is clear independence on loan amount as an increase or decrease has no effect on the interest rate and its associated loan grade; this was an early precursor to understanding the feature importance of loan grade.

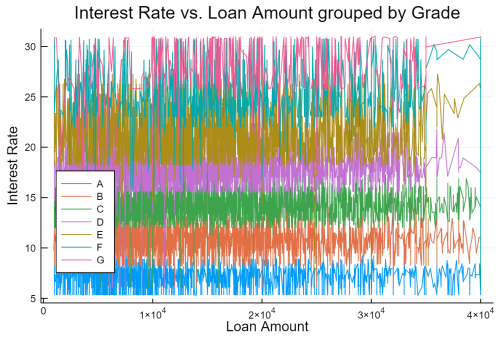


Figure 1: Interest Rate vs. Loan Amount

We then categorized the features by data type, and focused on numeric columns to create a correlation matrix shown in Figure 2 below. A clearer, more detailed graph can be found in the appendix. As we can see, the dependent variable fico range high and fico range low is highly negatively correlated with interest rate which makes sense because higher fico scores indicate low default rate and low missed payments hence a lower interest rate.

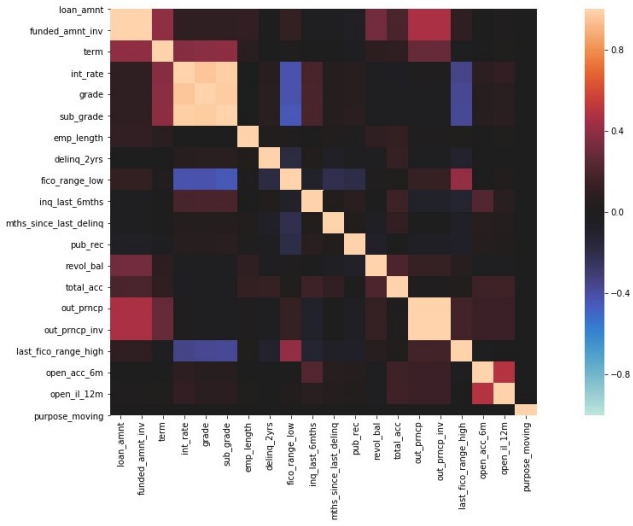


Figure 2: Feature Correlation

appendix.

2.3 Data Cleaning

In addition to removing the entries with missing values, we performed additional data cleaning procedures to ensure data quality and improve model performance. First, we converted some categorical features from strings to integers, including *employment length* and *loan terms.* After sorting through the original dataset and removing 112 features that were not applicable, we separated the remaining features into two groups based on their importance. Within the less important group, we looped through the features and removed those with over 20,000 missing entries. For missing values in the more important group, we identified an appropriate value to substitute for missing values of each feature. For example, for the feature *employment length*, we replaced missing values with -1 to signify that the individual is currently unemployed. Similar logic was applied to each of the remaining important features.

After obtaining our clean and functional data set, we dropped the interest rate column and stored it as a target vector. Then, we split the remaining data set into training and test set data frames to use for our models. The additional target vector was necessary for a few models like XGBoost and RidgeCV, but is not necessary for others. This allowed for easier visualization of a side by side comparison of the target values against the predicted values.

2.4 Feature Engineering

We considered various methods of feature engineering for our dataset, and this section will discuss the methods we decided to use and our motivations behind them. First, we used a label encoder for *grade* and *subgrade*. In doing so, lower grades were assigned lower numbers and higher grades higher numbers. The logic behind our label encoders was to ensure our regressions would correctly predict interest rates according to the magnitude of the encoded values and ultimately to incorporate into a regression model as numerical data. Next, for a subset of the categorical variables, we used one-hot-encoding in order to include these features in our regression models. This subset included features such as home ownership, application type, and payment plan. Finally, we added an offset to each of our models which will be discussed in greater detail in later sections.

2.5 Feature Selection

To determine the relative significance of a subset of the features, we used the Scikit-Learn package in Python to show how much the inclusion of a specific variable improves the model's prediction. We have an intuitive understanding of what explanatory variables should contribute to determining an interest rate for a borrower, and the resulting importance of each feature displayed by Scikit-Learn confirmed our assumptions. For example, strong features we believed were significant were *grade, sub-grade, installment, annual income, dti (debt to income ratio), delinquencies*, and *amount requested*, therefore we used these in our models and Scikit learned ranked the importance of each feature given a model's highest performance on the training and test sets. An important adjustment after realizing the high significance of grade and

especially sub-grade was to introduce a regularizer, as they are intrinsically calculated from the interest rate among other metrics and therefore would carry a significant feature weight.

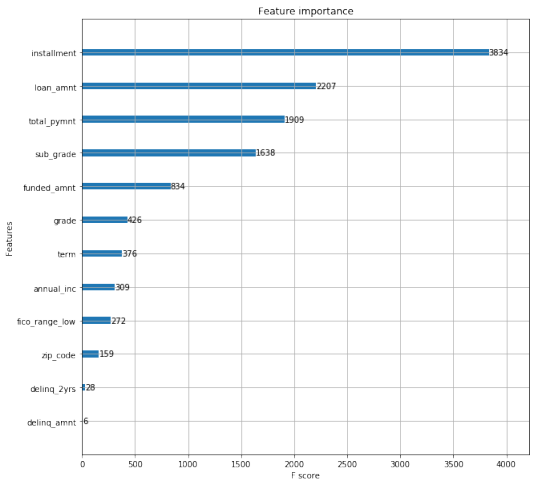


Figure 3: Feature Importance

appendix.

3 Model Selection

3.1 Linear Regression

After determining the 20 most optimal features for our learning algorithms via parameter tuning, we ran a least squares linear regression on the highest 8 as our base model for comparison. We used the Scikit learn built-in label encoder to map ordinal features such as grade and sub_grade to numerical values in order to perform linear regression. We recognize this model as very fundamental and acts to serve as a baseline for continuous improvement. Since our features are both categorical and numerical, it wouldn't make sense to stop with linear regression; however, it is always a good start to understand our data and observe the impacts of certain feature transformations in a linear context. **Our test MAE = 1.19 and train MAE = 1.03 .**

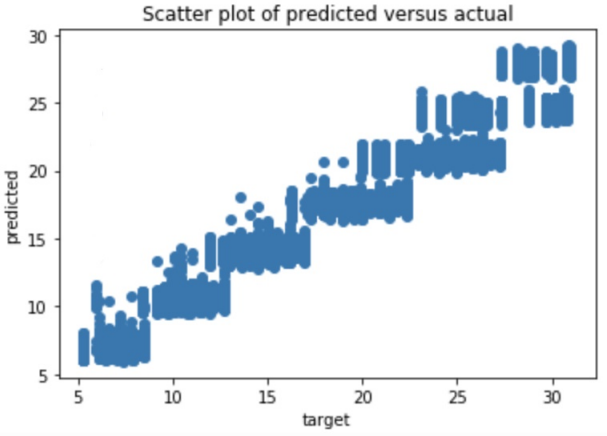


Figure 4: Linear Regression

3.2 Ridge Regression

By using Ridge Regression, we were able to account for multicollinearity among the features and add regularization to penalize outliers. An advantage over Lasso regression is that we do not set unimportant features to zero, and since we have a real world understanding of some of the feature importances, we found it necessary to still place emphasis on all of our features. An important assumption we made was the inclusion of Gaussian noise. If we had more data to work with, we could compare results for Lasso versus Ridge on new data and see if these less important features were actually insignificant as Lasso assumes. Furthermore, we used RidgeCV (a built in model ScikitLearn) to implement a 10-fold cross validation to fine tune lambda, i.e., the penalty term. After manually tuning based on smallest MAE and by considering RidgeCV's suggested lambda, we arrived at a lambda of 10. **Our training MAE was 0.66877 and our test MAE was 0.78801.** Again, these error estimates are higher, as expected due to the exclusion of the features

3.3 XGBoost

$$\sum_{i=1}^{n} l(y_i, \hat{y}^t) + \sum_{i=1}^{t} \Omega(f_i) \tag{1}$$

We discovered a very powerful tool in the world

of machine learning called XGBoost. The objective function being minimized is displayed in the equation above, where omega is the regularizer. It is an implementation of gradient boosted decision trees designed for speed and performance. Essentially, new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It uses gradient descent to minimize loss when adding new models. XGBoost solves categorical and regression problems, which is perfect in our case since we aim to predict interest rates; furthermore, our features are categorical and numerical, and XGBoost combines several decision trees to incorporate the two and improve prediction accuracy.

Our implementation used a squared error loss function and a mean absolute error metric. We further optimized our model using parameter tuning on Max Depth, Min Child Weight, Gamma, Subsample, Col Sample by Tree, lambda, and alpha. Parameter tuning involves several sub-models and considers possible ranges of each parameter. We decided to override the values assigned to lambda and alpha, (L2 and L1 regularization factors) to compensate for the large importance placed on sub-grades.

4 Results

Model	Test Set MAE	Train Set MAE
Linear Regression, Least Squares	1.19	1.03
Ridge Regression	0.78801	0.66877
XGBoost(default)	0.50928	0.498106
Final XGBoost	0.41694	0.3921

Figure 5: Test Results

4.1 Final Model

After parameter tuning, we selected the optimal 20 features and arrived at a final XGBoost model with **test set error of 0.41694**, the lowest error yet.

5 Implications

5.1 Weapons of Math Destruction

A Weapon of Math Destruction (WMD) is a predictive model whose results can have negative implications on large numbers of people[1]. There are a couple of factors that suggest that our project is a WMD, notably the financial issues that can be raised from high interest rates. The assignment of uncharacteristically high interest rates can discourage consumer spending and negatively affect the economy. On the other hand, our model exhibits many qualities that are uncharacteristic of a WMD. First, we have a large amount of data and have continuous access to more as time goes on. In addition, we have multiple experiments (models) that predict the same outcome (interest rate) using varying tools and algorithms, which also suggests the project is not a WMD.

5.2 Fairness

In the context of loans, neglecting fairness in algorithms can have large and often negative impacts. In our dataset, the feature *addr_state* is a potential source of bias in predicting interest rates for individuals. According to the Equal Credit Opportunity Act[2], discrimination based on national origin is illegal, and an individual's geographic location is inevitably linked to their national origin. To measure fairness in our application, we first applied a group by to group each state, then calculated the mean interest rate for each state. We then sorted the interest rates and identified the states with the highest and lowest interest rates: Utah = 12.285 and West Vir-

[1] https://people.orie.cornell.edu/mru8/orie4741/lectures/limits.pdf
[2] https://people.orie.cornell.edu/mru8/orie4741/lectures/fairness.pdf

ginia = 13.747. These rates may not seem drastically different, but you can see the differences in the interest rate distributions on the graph below. Although the means are not too different in the short-term, depending on the loan amount and installment plan, the payment discrepancies can be quite substantial.

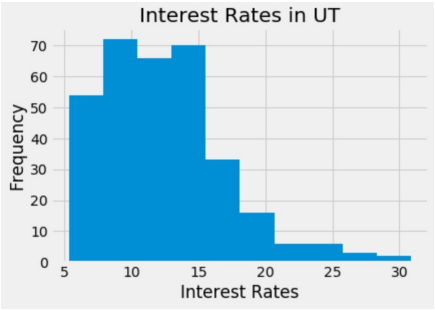


Figure 6: Interest Rates by State

Despite the larger inconsistencies between the states with the maximum and minimum interest rates, our models do generally satisfy demographic parity when looking at the dataset as a whole. If we separate the examples into groups by states, the distributions of the interest rates are approximately equal for most of the locations, excluding the extreme values. Although the rates were similarly distributed for most geographic groups, we decided to remove zip codes and states from the data set entirely. This was done in an attempt to more completely satisfy predictive rate parity, and we ran each of our final models excluding the state and zip code predictors.

6 Future Improvements

6.1 Reducing Bias

An interesting possible expansion of our analysis is the reduction of demographic bias in interest rate prediction and assignments. Although this demographic data was not available to us, features such as age, sex, race, etc. could help identity crucial discriminatory issues in the lending industry. It would be very interesting to explore the correlations between these factors and interest rates, and ultimately develop new models that predict independently of these attributes.

6.2 Classification Problem

Another expansion of our project is into the classification category. In our preliminary analysis, we considered the classification of loans into "good" and "bad" groups. This idea is interesting, because the classification of loans would be highly valuable to individuals. Many people lack the knowledge and experience necessary to determine if a loan is fair or not, so such a classification algorithm could directly impact the financial decisions of a large number of individuals.

We could potentially solve this classification problem using a Support Vector Machine Algorithm. SVM aims to maximize the distance between the two classes and returns the corresponding optimal hyperplane. Maximizing the margin makes sense in the context of interest rates, as small differences in percentage can have large financial impacts on individuals.

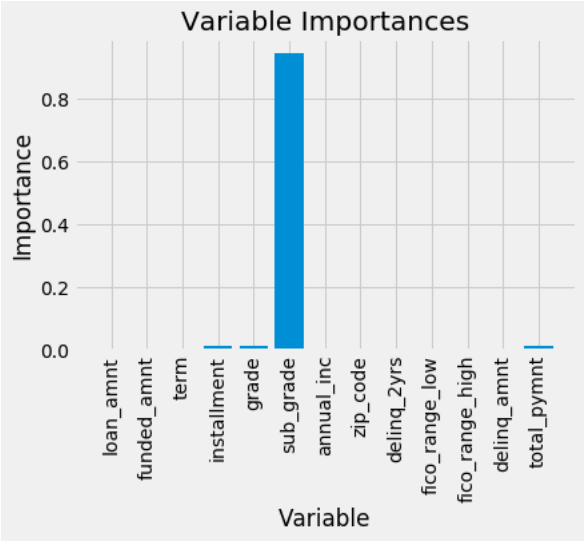
6.3 Computational Speed

Throughout the project, our computers did not have enough RAM to run each model within a reasonable time frame. Next time, either working on more powerful computers or purchasing external "computers" such as through AWS to increase the number of

cores would allow us to increase the number of boosting rounds (as in XGBoost) to minimize our error and push the limits of the boosted decision tree. Moreover, we could have applied a more extensive grid search to fine tune our parameters but due to the computational cost of iterating thousands of times over different combinations of parameters we could only achieve so much. Our project shed light on the importance of model selection while training on a large data set,i.e., the time complexity, and how limited our ability is to determine the best parameters of built in python packages for machine learning.

6.4 *Grade* Inclusion

An important realization was the heavy significance of grade and especially sub-grade to all of our models. We discovered that these features and their values were calculated through a combination of borrower characteristics and credit reports, according to Lending Club's website. Since these hold heavy weight and their calculations are unknown, we may consider removing these in the future and create our own "grade" and "subgrade" columns to potentially improve the accuracy of our models. Since these were not original features, we may rationalize why it may be unfair to include them because we do not know how they were determined, although their inclusion in our models drastically reduces our mean absolute error by almost 2 across all models. Therefore, we found it advantageous to utilize and acceptable since our goal was to predict the proper interest rates to apply to loans, and the consequences of providing accurate predictions of interest rates are far more significant than a small chance of poor model quality.



7 Conclusion

Accurate interest rate prediction is crucial in our economic system and influences a wide range of individuals and their financial expenditures. Our results suggest that an individual's credit and risk portfolio can be used to accurately predict the interest rate for a given loan. We started with a baseline linear regression model, and used parameter tuning to determine which features are the most optimal predictors. Then, after selecting a subset of 20 of the most optimal features, we ran a ridge regression model and reduced the error by 0.4. Finally, we ran an XGBoost model and tuned the parameters to arrive at a **final error of 0.501.**

In addition to accuracy, fairness plays an influential role in this industry as well, and we focused on reducing bias throughout our analysis. After removing all features relating to an individual's geographic location and employment titles, the largest source of discrimination was removed from our dataset. Because of these fairness considerations, we would approve our results for use in companies to change how interest rate decisions are made. Although more work can be done to eliminate discrimination in interest rate predictions, this analysis is a step in the right direction towards demographic parity.

8 Appendix

8.1 Enlarged Graphs

Figure 8: Enlarged Feature Correlation

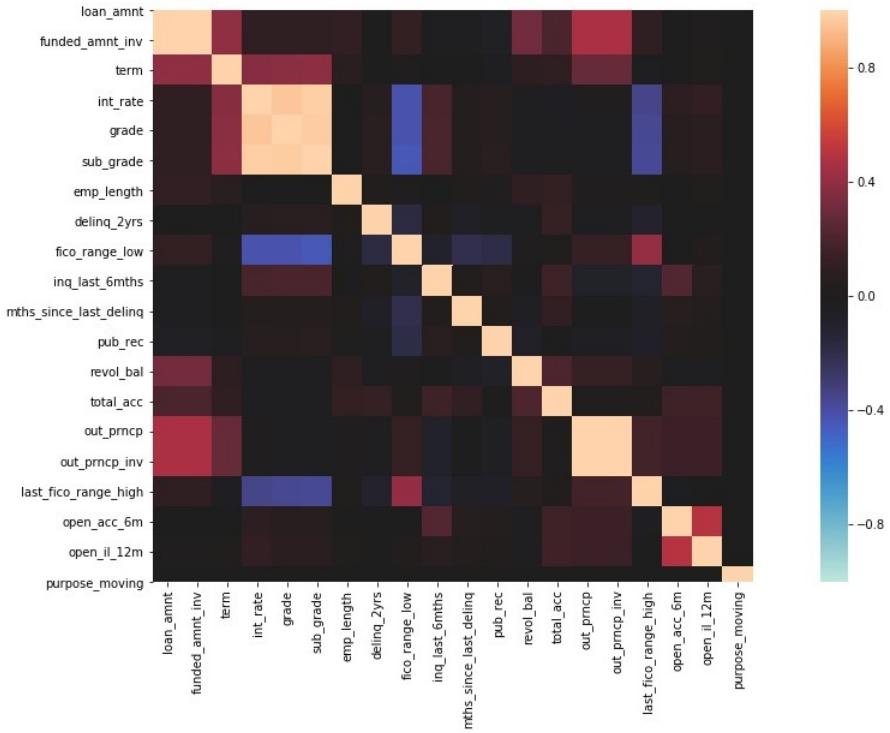
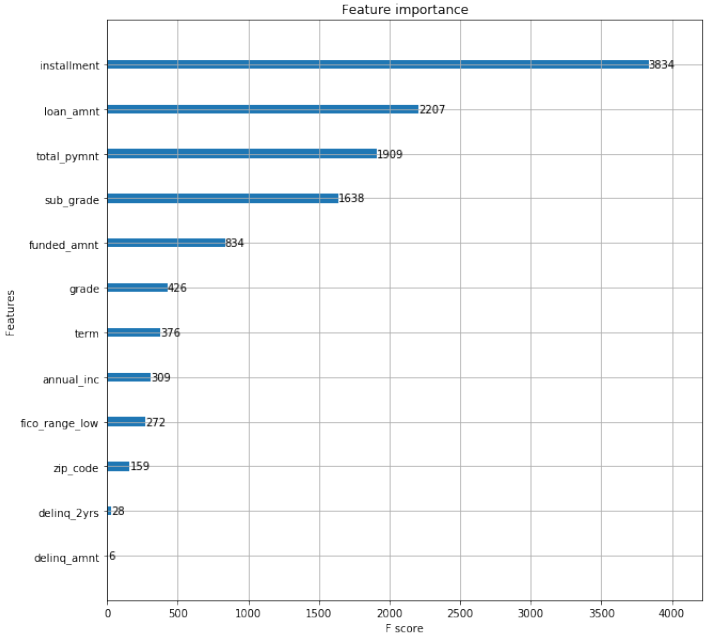


Figure 9: Enlarged Feature Importance



8.2 Tables

Figure 10: Parameter Tuning Table

Tuning	Max Depth	Min Child Weight	Gamma	Subsample	Col Sample by Tree	Lambda	Alpha
Test 1	[5, 7, 9]	[1, 3, 5]	0.2	0.8	0.8	1	1
Test 2	[4, 5, 6]	[2, 3, 4]	0.2	0.8	0.8	1	1
Test 3	5	2	[0, .., 0.9]	0.8	0.8	1	1
Test 4	5	2	0.5	[0.6, …, 0.9]	[0.6, …, 0.9]	1	1
Test 4b	5	2	0.5	[0.85, 0.8, 0.95]	[0.85, 0.9, 0.95]	1	1
Test 5	5	2	0.5	0.85	0.9	[1e-5, 1e-3, 1, 10, 100]	[1e-5, 1e-3, 1, 10, 100]
Final	5	2	0.5	0.85	0.9	1e-5	1e-5