Predicting LendingClub's Loan Interest Rates

Caroline Ryu (jr894), Corban Chiu (cpc75), Tyler Carreja (tcc86)
ORIE 4741 Final Project

Abstract

Loans are an essential part of our economy. The economics of loans help people to fund projects which otherwise they would not have been able to do. Through LendingClub's loan application database, our goal is to replicate their interest rate model, uncovering the factors shaping algorithmic loan rates. Utilizing regression and decision tree models, we aim to forecast rates and gain insights into potential algorithmic biases. Our regression models highlight the substantial impact of loan terms and purposes on interest rates. The significance of loan terms aligns with industry norms, reflecting their crucial role in rate determination. However, the unexpected significance of loan purposes as determinants, outweighing applicant financial profiles, raises concerns about algorithmic fairness. By delving into these findings, we aim to shed light on the interplay between loan rates, loan characteristics, and potential biases, ultimately contributing to a more comprehensive understanding of LendingClub's interest rate model.

I. Introduction

LendingClub was established in 2006 as a peer-to-peer lending platform as a Facebook application to facilitate the connection between borrowers and investors who fund their loans. Its primary aim was to enable the matching of investors and borrowers, thereby eliminating traditional banks from the process. While LendingClub assesses and services loans that are approved, the decision to fund a loan is made by investors based on loan details and the debtor's credit history. LendingClub benefits the borrowers by allowing them to obtain loan rates without affecting their credit score, while also offering fixed rates and transparent terms.

Despite its shift towards the neo-bank sector and discontinuation of its peer-to-peer platform due to increasing regulatory scrutiny, LendingClub's dataset from loans serviced between 2007 and 2018 remains a valuable resource for predicting the interest rate on new loan applications. The dataset comprises financial and demographic information that LendingClub used to assess the creditworthiness of potential borrowers and determine the interest rate. A detailed methodology for analyzing the data could provide valuable insights into the lending process and aid in identifying factors that contribute to the interest rate assignment.

II. Data

1. Data Description

The dataset used in this data analysis project is sourced from Kaggle, where fragmented datasets from LendingClub were compiled. The dataset consists of 2,260,701 rows and 107 columns, providing comprehensive information about accepted loan applications from 2007 to 2018. Each row represents a loan application, encompassing a wide array of demographic and financial data related to the debtor. The dataset includes crucial information about the loan funding, loan terms, and various other attributes.

The primary focus of this analysis is to explore the factors influencing the interest rate, which serves as the predictor variable. By examining the relationship between the interest rate and the available attributes, we aim to identify the key factors that influence the interest rate assigned to loan applications. This dataset provides valuable insights into the loan application process and allows us to investigate the relationships between various demographic, financial, and loan-related attributes and the interest rate.

2. Data Processing

Thorough data cleaning and normalization techniques were applied to ensure the quality and integrity of the dataset. The following steps were performed to prepare the data for analysis:

1. Column Selection: A subset of relevant columns was extracted from the original dataset.The process we used for this selection was determining which features would have been important at the time of applying for the loan. The selected columns include variables such as interest rate, loan amount, employment length, home ownership, annual income, verification status, purpose, zip code, debt-to-income ratio (DTI), credit history, and various other attributes related to loan applications.
2. Handling Missing Data: Rows with missing values were dropped from the dataset using the dropna() function. Some rows contained a significant number of missing values, making it difficult to derive meaningful insights from those records. Considering the size of the dataset, with over 2 million rows, the removal of these rows was deemed acceptable as it did not significantly impact the overall dataset size and did not compromise the integrity of the analysis.
3. Credit History and Estimated FICO Score Calculation: The credit length in months was computed by taking the difference between the loan issue date (issue_d) and the earliest credit line date (earliest_cr_line). This metric provides valuable information about the length of an applicant's credit history, which can significantly impact the interest rate assigned. Additionally, the fico_score was calculated as the average of the FICO range low and high values, providing an estimated credit score for each applicant. To streamline the dataset and eliminate redundant information, certain columns were identified as unnecessary for the analysis. As a result, the columns earliest_cr_line, issue_d, fico_range_low, and fico_range_high were dropped from the dataset.
4. Encoding Binary Variables: We needed to modify the 'verification status' column because it contained string values of 'Not Verified', 'Verified', and 'Source Verified'. We replaced these values using the transformation function 'Not Verified' = 0 and 'Verified' or 'Source Verified' = 1. Similarly, the 'initial_list_status' column was converted to boolean values, with 'w' representing True (1) and 'f' representing False (0). The final feature that needed a binary transformation was the `application_type` column which was encoded as 1 for 'Individual' and 2 for 'Joint App'.
5. One-Hot Encoding: Two features: 'purpose' and 'home_ownership' contained categorical variables. We used one-hot encoding to transform each unique category into its own feature . This process transformed these variables into multiple binary columns representing different

categories, enabling their inclusion in the analysis. Each one-hot encoded column is named following the prefix of the original feature and an underscore. (i.e. purpose_debt_consolidation)

6. Normalization: As part of the data preprocessing phase, the dataset underwent feature scaling using the RobustScaler from the sklearn.preprocessing module. The RobustScaler is a method that accounts for outliers during scaling. It subtracts the median and divides by the interquartile range, making it suitable for non-normal distributions.

After completing this stage, the dataset consisted of 2,045,432 rows and 34 columns. The first 17 columns included: 'int_rate', 'term', 'loan_amnt', 'emp_length', 'annual_inc', 'verification_status', 'zip_code', 'dti', 'initial_list_status', 'application_type', 'pub_rec_bankruptcies', 'tax_liens', 'tot_hi_cred_lim', 'total_bal_ex_mort', 'credit_length_months', 'fico_score', along with 12 one-hot encoded columns for different purposes and 5 one-hot encoded columns for homeownership. A more detailed description of our dataset can be found in our appendix.

3. Data Visualization

Feature Correlation

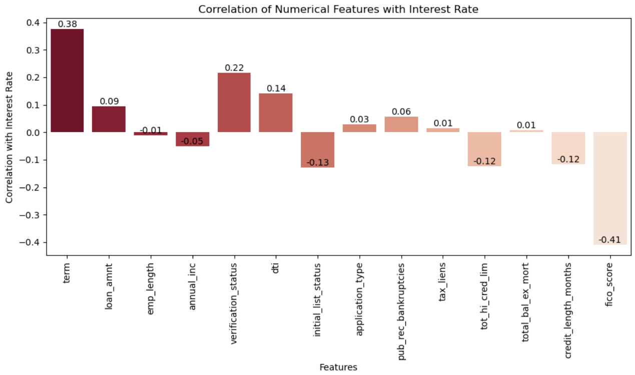

Figure 1a: Correlation of Numerical Features with Interest Rate

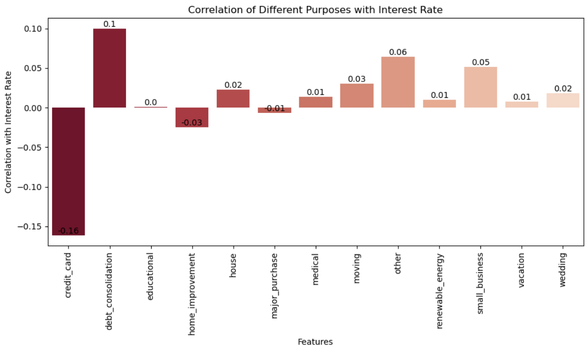

Figure 1b: Correlation of Different Purposes with Interest Rate

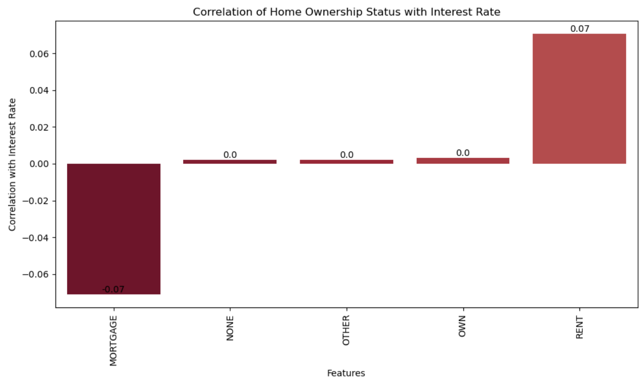

Figure 1c: Correlation of Home Ownership Status with Interest Rate

Before delving into the process of running the models, it is important to examine the correlation of various features with the target variable, interest rate. The correlation coefficients indicate the strength and direction of the relationship between each feature and the interest rate. These correlations were visualized through a heatmap, revealing the following insights:

1. Term: Longer-term loans exhibit a moderate positive correlation (0.38) with higher interest rates, likely due to the increased risk associated with extended repayment periods.
2. FICO Score: Higher credit scores demonstrate a strong negative correlation (-0.41) with lower interest rates, suggesting that lenders tend to offer more favorable rates to borrowers with better creditworthiness.
3. Verification Status: Loans with verified income sources show a moderate positive correlation (0.22) with higher interest rates. This correlation may arise because verified loans provide greater assurance to lenders, which can influence interest rate decisions.
4. Debt-to-Income Ratio: There is a moderate positive correlation (0.14) between the debt-to-income ratio and interest rate. This correlation suggests that higher ratios tend to be associated with higher interest rates, potentially due to the increased credit risk.
5. Initial List Status: Loans initially listed as 'Fractional' exhibit a moderate negative correlation (-0.13) with interest rates. This correlation may be influenced by market preferences or liquidity concerns, which can affect interest rate variations.
6. Total High Credit Limit: There is a moderate negative correlation (-0.12) between the credit limit and interest rate. This correlation indicates that higher credit limits tend to be associated with lower interest rates, potentially reflecting responsible credit usage.
7. Different Purposes: Among various loan purposes, the only noteworthy correlation is observed with loans intended for paying off credit card debt (-0.16). This correlation suggests that individuals seeking loans specifically for credit card consolidation tend to receive lower interest rates. However, it is important to note that correlation does not imply a causal relationship between loan purpose and interest rates.

Box Plots of Key Numerical Variables

In order to gain insights into the overall distribution of the dataset, we generated box plots for key numerical variables that we deemed significant. Specifically, we focused on loan amount, FICO score, interest rate, and annual income. From the boxplots, we see that:

1. The distribution of loan amounts ranges from 1,000 to 40,000. The mean loan amount of 15,358 indicates the average borrowing size in the dataset.
2. The distribution of FICO scores in the dataset ranges from 662 to 847. The mean score of 700 indicates that the majority of borrowers have a relatively good credit standing.
3. The distribution of interest rates spans from 5.31% to 30.99%. The mean interest rate of 13.1% represents the average borrowing cost in the dataset. The standard deviation of 4.85 suggests some variability in interest rates, indicating that some loans carry higher or lower rates compared to the mean.
4. For the annual income variable, we noticed that there were two outliers—income of $61,000,000 and $110,000,000— that affected the scale of the box plot, making it challenging to visualize the distribution of other values. These outliers, being significantly higher than the majority of the data, caused the axis to extend to accommodate their values. To address this challenge, we employed a symlog scale provided by matplotlib. This scale offers a logarithmic transformation that allows for a more balanced representation of the data, even in the presence of extreme values.

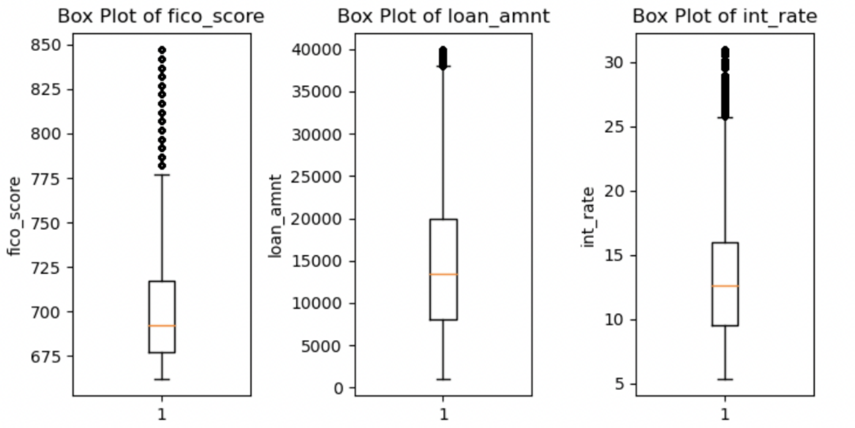

Figure 2a: Box Plot of FICO score, Loan Amount, and Interest Rate

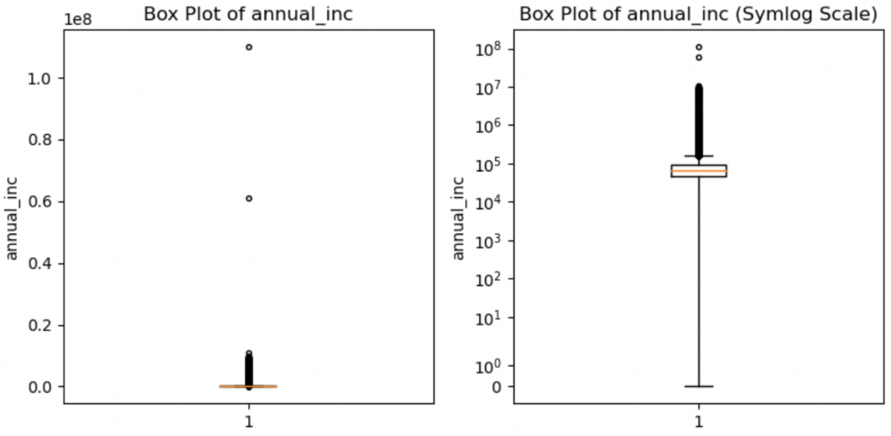

Figure 2b: Box Plot of Annual Income and Annual Income on a Symlog Scale

III. Models

1. Model Selection

We implemented 5 models and varied their hyperparameters and regularization settings to find one with the least mean squared error. We used a 75/25 split of train and test data, and decided that k-fold cross validation was unnecessary due to the size of our dataset. We started with three variations of linear regression, stochastic gradient descent (SGD), ridge, and lasso.

Our SGD regressor used a least squares loss function with an L2 regularization term. We varied the alpha parameter from 1e-5 to 2e-4 with a step size of 1e-5 and the max number of iterations from 700 to 1200 with a step size of 100. We found that an SGD regressor with alpha and max_iter set to 1.7e-4 and 1200, respectively, was best with an MSE of 14.0657. For the ridge regressor, we varied the alpha parameter from 1 to 20 with a step size of 1 and found that a ridge regressor with alpha set to 7 was best with an MSE of 14.0517. We noticed that the MSE largely remained the same when alpha changed so we chose not to test on a smaller step size. For the lasso regressor, we varied the alpha parameter from 0.001 to 0.009 with a step size of 0.001 and found that a lasso regressor with alpha set to 0.001 was best with an MSE of 14.0578.

Across our regression models, we found relatively similar errors so we tried decision trees and random forests because we thought that feature splitting might be better suited to one-hot encoded features.

Our decision tree regressor evaluated splits using mean squared error and prioritized the best ones. We varied the maximum tree depth from 6 to 17 with a step size of 1 and found that a decision tree with a max_depth of 12 was best with an MSE of 13.6814. Our random forest regressor similarly evaluated splits using mean squared error and used bootstrap samples to build trees. We varied the number of trees in a forest from 8 to 14 with a step size of 1 and the maximum tree depth from 12 to 19 with a step size of 1. We found that a random forest with 14 estimators each with a max_depth of 15 was best with an MSE of 13.1004.

2. Analysis

We chose to use mean squared error as our measure of accuracy because it accounts for both positive and negative errors, and penalizes larger mistakes in our predicted value more so than the mean absolute error.

Our regression models ended up with nearly the same MSEs surprisingly converging to very similar weight vectors. We inferred that regression models might be able to predict the interest rate effectively, but not with particularly high accuracy because most of our features were very slightly linearly correlated to the interest rate. We thought this lack of correlation in features would make lasso regression better than ridge regression. Additionally, l1 regularization can lead to feature weights approaching 0 which was important given that we selected the set of trainable features primarily based on intuition, so we could test if features were actually relevant or not. All three models gave loan_term_length a large positive weight and purpose_credit_card a large negative weight. Lasso regression diverged from the other regression models by treating purpose_debt_consolidation as a strong predictor while completely ignoring purpose_home_improvement, purpose_medical, purpose_small business, home_ownership_ none, home_ownership_other, and home_ownership_rent. In contrast, SGD and ridge regression both gave purpose_small_business and purpose_wedding large positive weights.

All three models having a large positive weight for loan term makes sense because it had the highest linear correlation to interest rate. The purpose of the loan being an important factor in the regression models could be a source of bias within the algorithm which generates interest rate. This bias in our models leads to people with the same financials having different rates based on their purpose for the loan. Intuitively, the purpose of the loan should have some bearing on the ability to pay back the loan, but a hike or decrease of almost 2% depending on the purpose of the loan seems to be a bit high. Because categorical features seemed to have a big impact on the predicted loan rate, we thought decision trees might be a better model of Lending Club's interest rate algorithm.

Our decision tree and random forest models had reasonably better MSEs compared to our regression models. One of the primary concerns with decision trees is their tendency to overfit on data. As discussed earlier, we tried to reduce this with the max_depth parameter. By preemptively limiting the size of each tree, we could stop it from growing too large and becoming too good of a predictor of the training data. Figures 3a and 3b show plots of the max_depth parameter against MSE for the decision tree regressor and random forest regressor, respectively. In both figures, we can observe clear regions of underfitting and overfitting as forcing trees to be too small or allowing trees to be too large will cause MSE to rise. Finding the "sweet" spot gives us confidence that the lower MSEs of the random forest and decision tree regressors are fair assessments of their quality and that they are in fact better models.

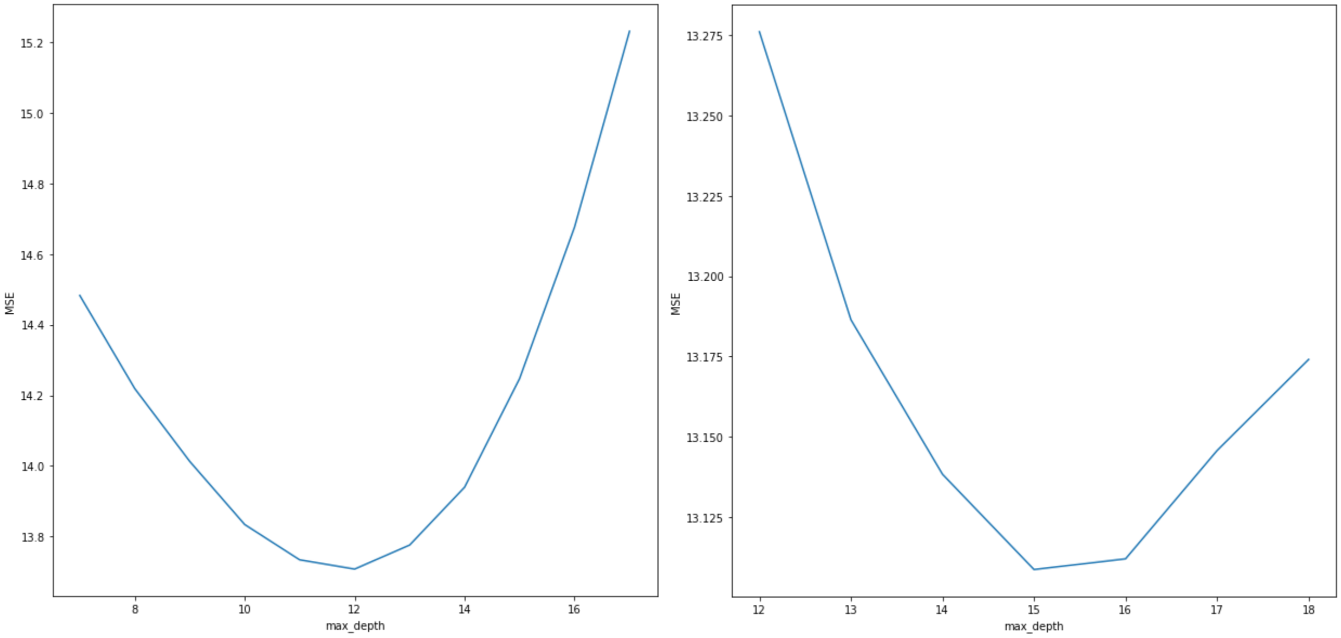

Figure 3a (left): Decision tree regressor max_depth vs MSE, **Figure 3b(right):** Random forest regressor max_depth vs MSE

IV. Conclusion

In our attempt to replicate Lending Club's interest rate algorithm, we found that a random forest model with 14 trees each with a max depth of 15 performed best. Some of the most relevant features influencing a loan's interest rate are unbiased such as the term of the loan, but others such as the purpose of the loan reflect potential unfairness in the algorithm. Through this analysis, we can potentially see how Lending Club's interest rate algorithm could be a "Weapon of Math Destruction." Loans can be an important aspect of people's livelihoods affecting lifelong decisions such as taking on a mortgage for a home or purchasing a car. It is essential that an algorithm which could have such a profound impact on someone's life should be carefully constructed to avoid biases which could hurt someone for negligible reasons.

Appendix

Feature Table

Data Element	**Type**	**Description**
term	object	The number of payments on the loan. Values are in months and can either be 36 or 60
emp_length	object	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years
home_ownership	object	The home ownership status provided by the borrower during registration or obtained from the credit report. Possible values are RENT, OWN, MORTGAGE, OTHER, or NONE
annual_inc	float64	The self-reported annual income provided by the borrower during registration
verification_status	object	Indicates if income was verified by LC, not verified, or if the income source was verified
issue_d	object	The month which the loan was funded
purpose	object	A category provided by the borrower for the loan request. Possible values are credit card, debt consolidation, educational, home improvement, house, major purchase, medical, moving, other, renewable energy, small business, vacation, and wedding
zip_code	object	The first 3 numbers of the zip code provided by the borrower in the loan application
dti	float64	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income
earliest_cr_line	object	The month the borrower's earliest reported credit line was opened

fico_range_low	float64	The lower boundary range the borrower's FICO at loan origination belongs to
fico_range_high	float64	The upper boundary range the borrower's FICO at loan origination belongs to
initial_list_status	object	The initial listing status of the loan. Possible values are: W, F
application_type	object	Indicates whether the loan is an individual application or a joint application with two co-borrowers
pub_rec_bankruptcies	float64	Number of public record bankruptcies
tax_liens	float64	Number of tax liens
tot_hi_cred_lim	float64	Total high credit / credit limit
total_bal_ex_mort	float64	Total credit balance excluding mortgage

Regression Weights and Corresponding Columns

```
Columns: Index(['loan_amnt', 'term', 'emp_length', 'annual_inc', 'verification_status',
       'zip_code', 'dti', 'initial_list_status', 'application_type',
       'pub_rec_bankruptcies', 'tax_liens', 'tot_hi_cred_lim',
       'total_bal_ex_mort', 'credit_length_months', 'fico_score',
       'purpose_credit_card', 'purpose_debt_consolidation',
       'purpose_educational', 'purpose_home_improvement', 'purpose_house',
       'purpose_major_purchase', 'purpose_medical', 'purpose_moving',
       'purpose_other', 'purpose_renewable_energy', 'purpose_small_business',
       'purpose_vacation', 'purpose_wedding', 'home_ownership_MORTGAGE',
       'home_ownership_NONE', 'home_ownership_OTHER', 'home_ownership_OWN',
       'home_ownership_RENT'],
      dtype='object')
```

```
Ridge: [ 0.20635513,  4.13410218,  0.01724488, -0.03407924,  0.94798392, -0.05060383,
0.54665321, -1.46788704,  0.57680518, -0.20103045, -0.06080995, -0.3616412 ,  0.01553468,
-0.44990502, -2.22116314, -1.6196855 , -0.19059107, -0.49459829,  0.11629467,  1.57392637,
0.37141757,  1.0463583 ,  1.74140185,  1.45409679,  1.96423419, 2.44545114,  1.11314349,
3.96161604, -0.68654403,  0.98578627, 1.43793363, -0.29809305, -0.36980604]
```

SGD: [0.23344309, 4.06047874, 0.08519983, -0.27432571, 0.91267225, -0.08708933, 0.62016557, -1.44149429, 0.52760468, -0.18866595, -0.00466262, -0.22963701, 0.07869182, -0.43524851, -2.26111432, -1.91524621, -0.42791531, -0.00778951, -0.1439429 , 1.24863912, 0.11033827, 0.7861369 , 1.40902661, 1.26828141, 1.38746765, 2.14972342, 0.87349075, 2.36840357, 1.0698519 , 0.18963541, 0.21351023, 1.47878379, 1.36697262]

Lasso: [0.20482162, 4.12865273, 0.01080862, -0.03396767, 0.94908949, -0.04744085, 0.54736953, -1.46528722, 0.55395443, -0.19364172, -0.05392757, -0.35920373, 0.01427472, -0.44849698, -2.2193754 , -2.02638606, -0.59930169, 0, -0.27873158, 1.00048645, 0, 0.54939814, 1.17837054, 1.0258201 , 0, 1.9384701 , 0.55022276, 1.28134966, -0.31894102, 0, 0, 0.05844721, 0]

Additional Graphs

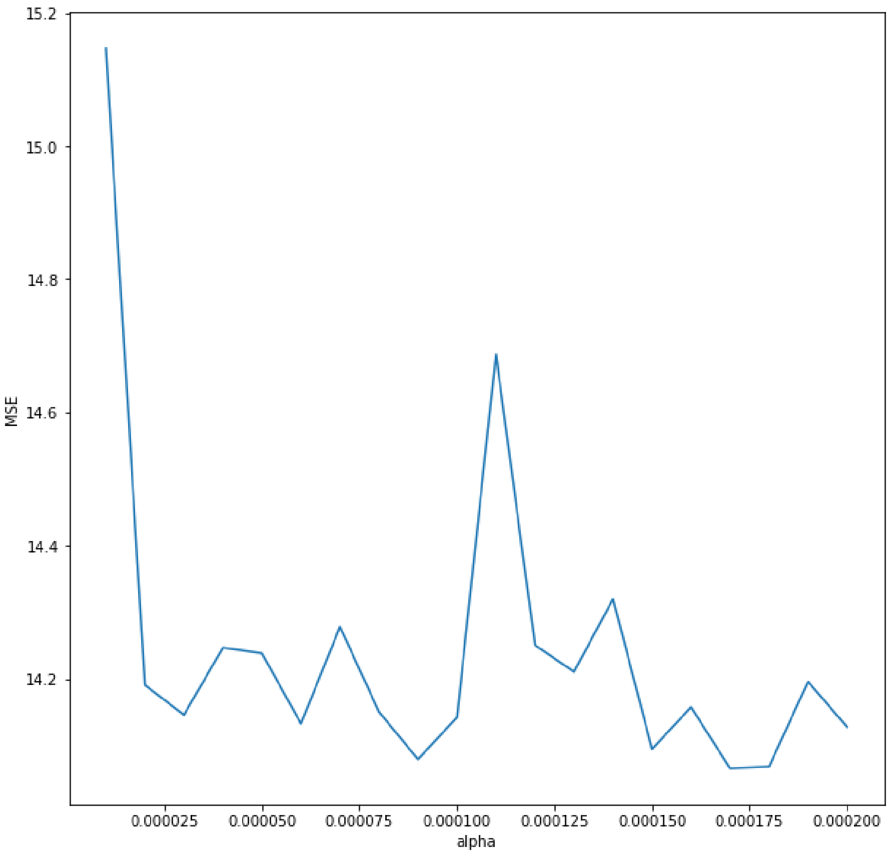

Figure 4: SGD Regressor alpha vs MSE