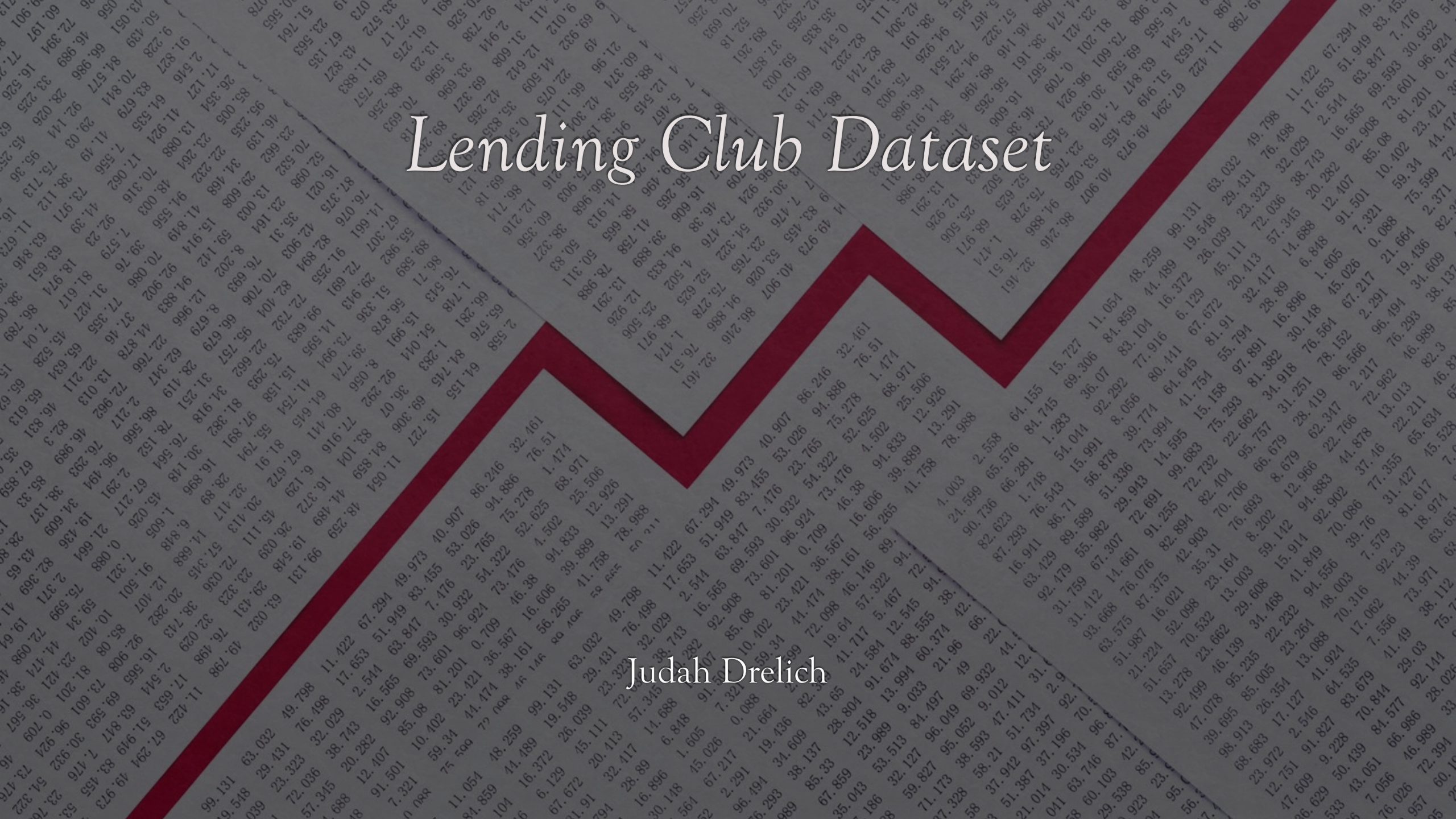


Lending Club Dataset

Judah Drelich

Target Variable

- Separate the features (X) from the target variable (y).
  - `*loan_status*`: supposed to be the target variable.
    - Too many of the features have information that was taken after the loans originated
      - `*next_pymnt_d*` records the date that the next loan payment is due
  - Shift the time of the question to the time that the dataset was taken.
    - Find the value of the remaining loans
      - Target variable is the fraction of the loan amount that the borrower will be able to pay back

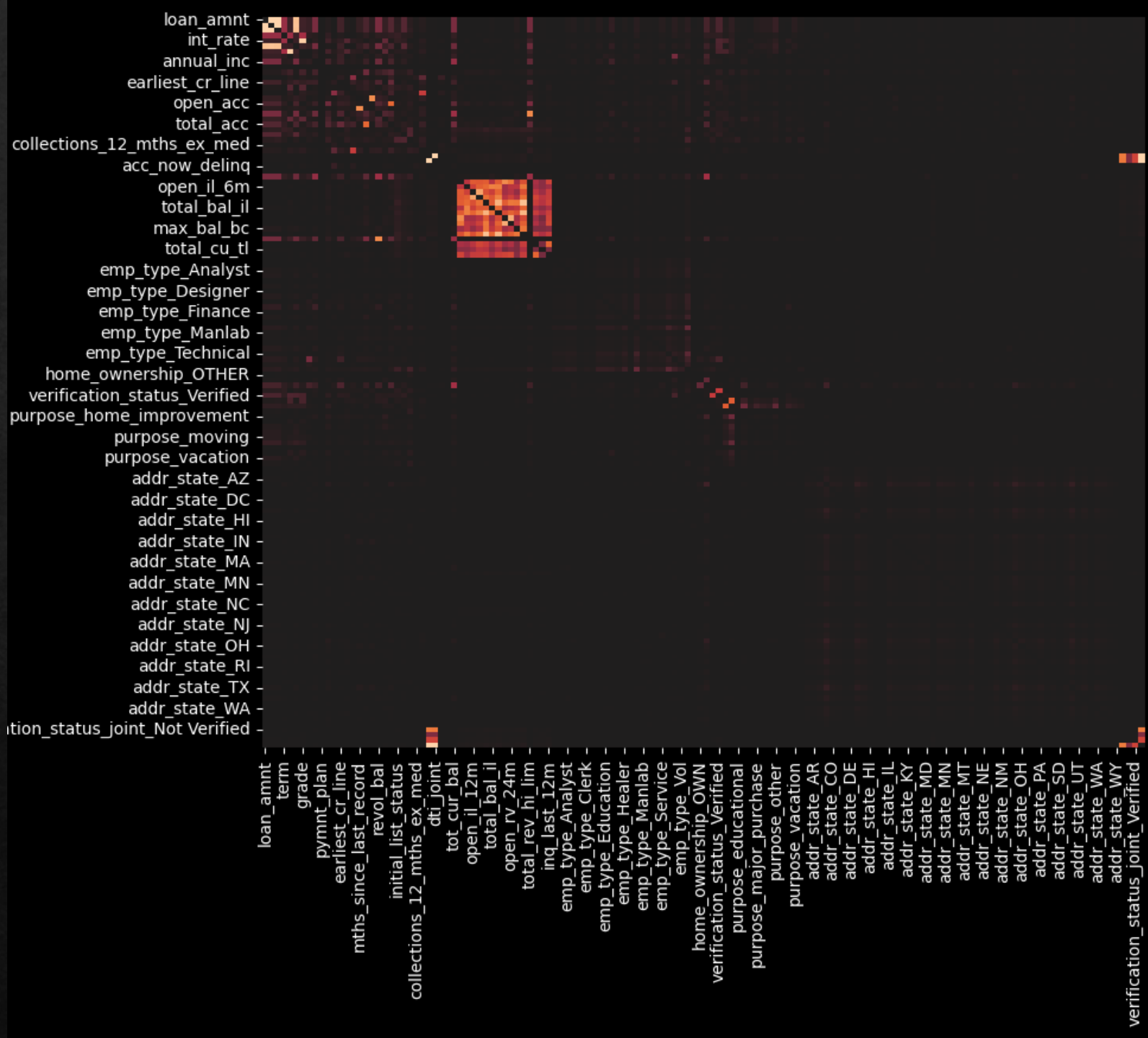
Data Wrangling

- All the features must have only numerical values
  - Some columns are numerical but need to be cleaned (Term: ‘ 36 months’)
  - Others are Categorical such as the type of employment that the borrower has.
- Imputation of values for this dataset is the most challenging aspect of this project.
  - Assumed that the event didn’t happen if the data was missing and assigned values accordingly.
    - For counting variables such as the number of accounts, I filled in with a zero
    - For Months since an event happened, I filled in a large number so that it the models would act like it never happened.

Data Wrangling

Multicollinearity

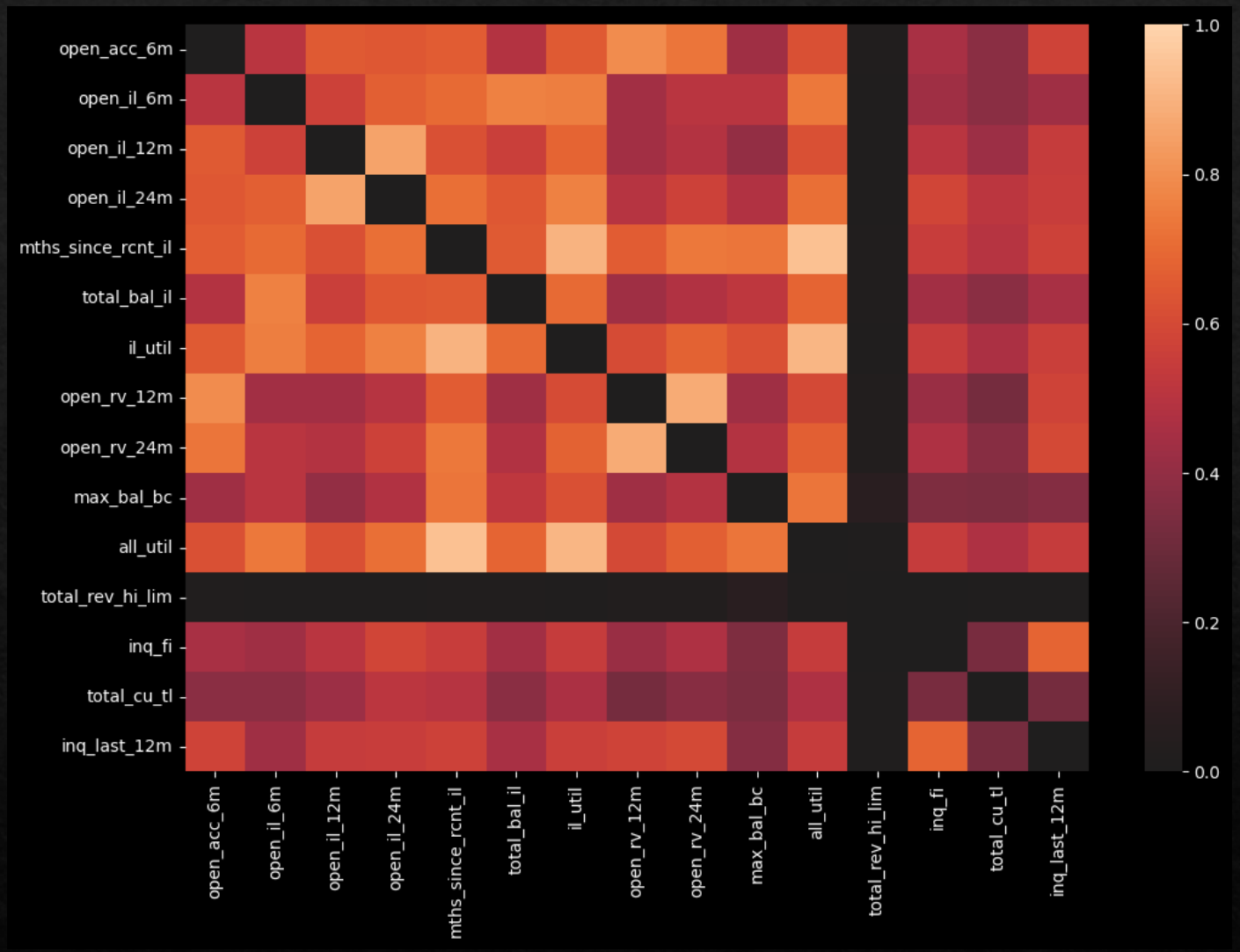
Initial heatmap for the correlations between the features



Data Wrangling

Multicollinearity

Focusing on the bright square of the previous heatmap



Data Wrangling

- Lasso Regularization is a technique for variable selection that uses linear regression to evaluate the effect that each features has on a target variable.

Lasso Regularization

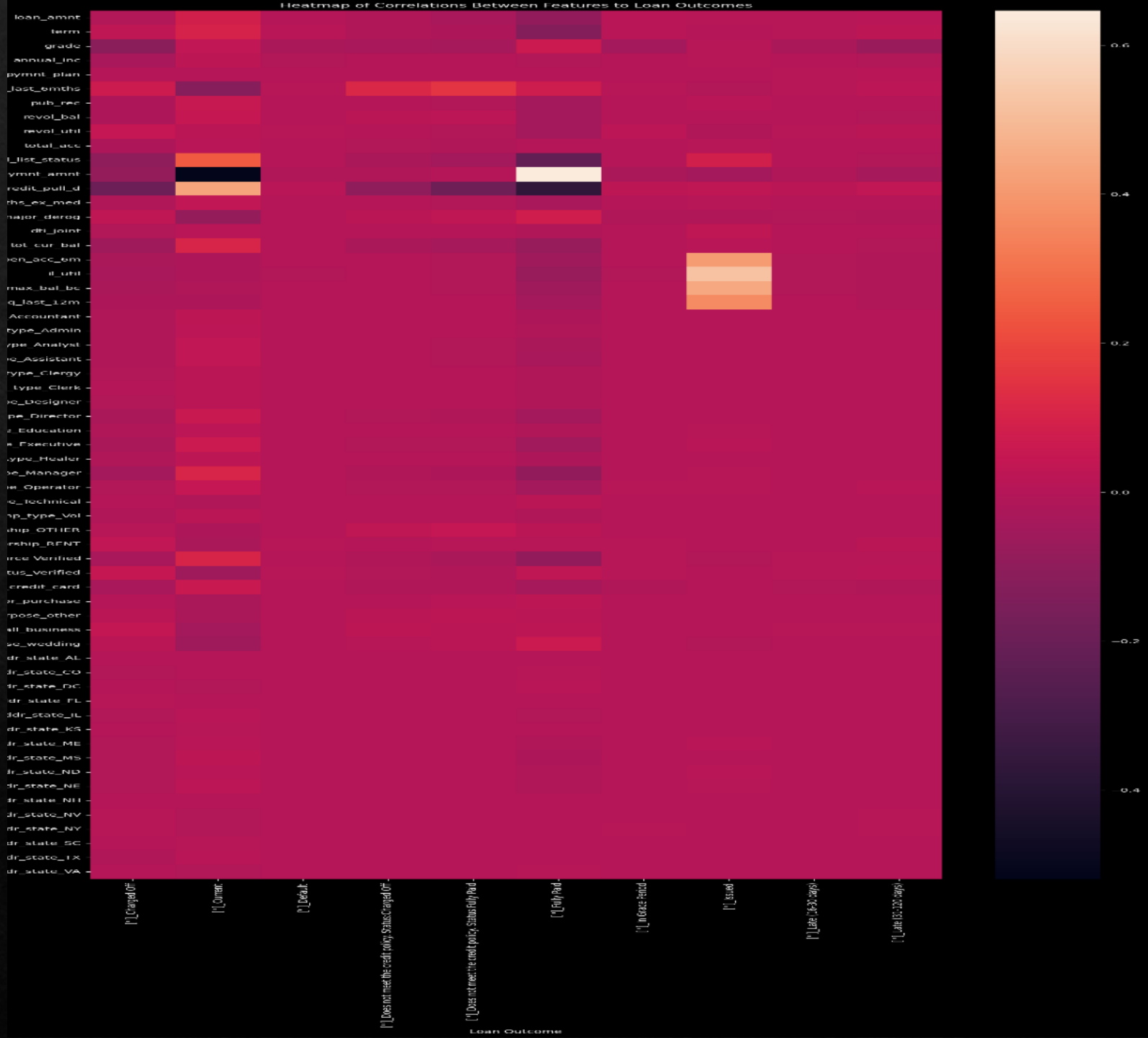
$$Loss(\beta_1, \dots \beta_n) = SSD + \alpha \sum_{i=1}^{n} |\beta_i|$$

- I used Lasso Regularization to create to datasets
  - Small: 12 features, binary intrusion or not
  - Big: 61 features, multi class model, determines intrusion types

Feature Selection

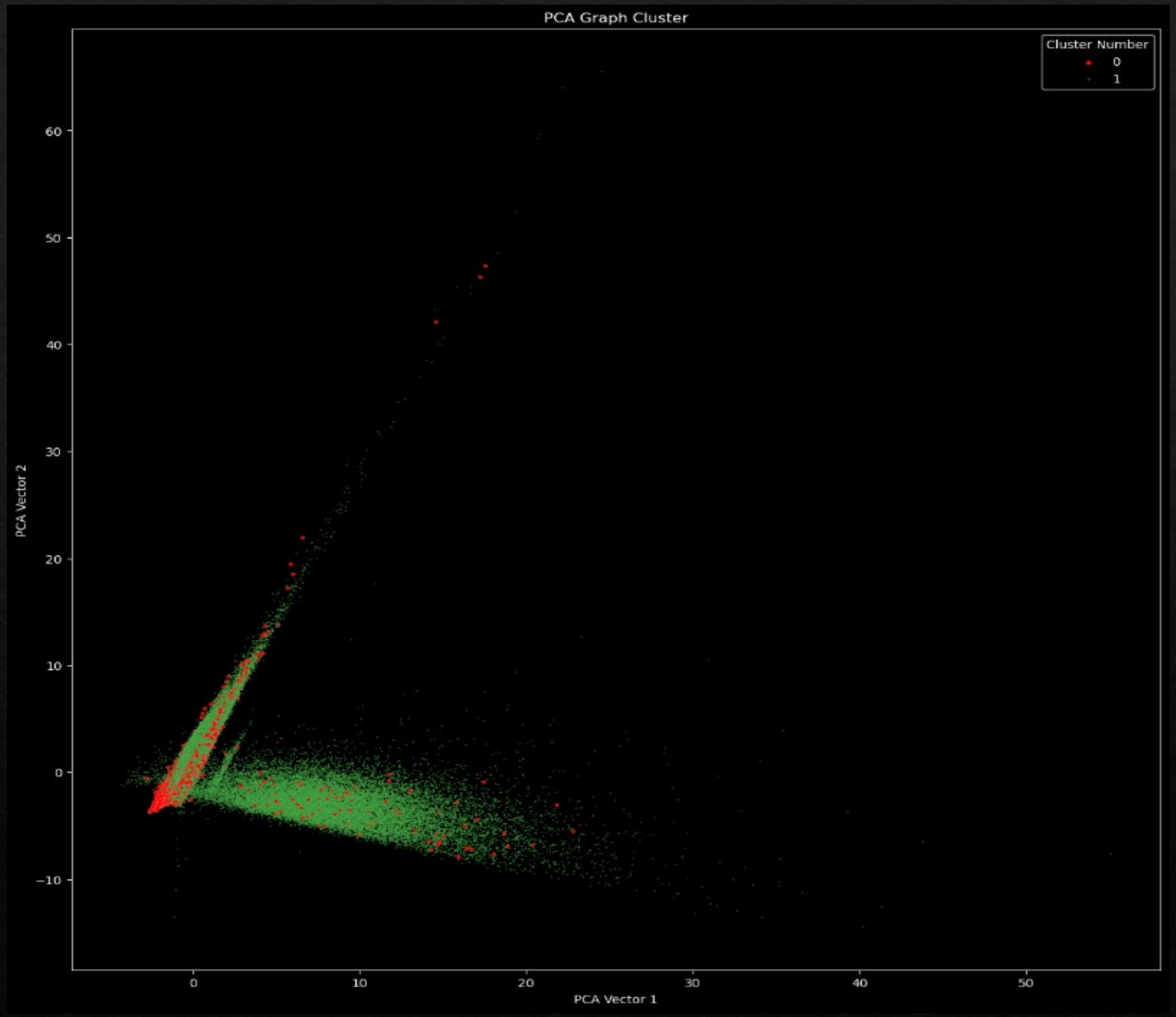
EDA

Correlation Heatmap between features and loan statuses

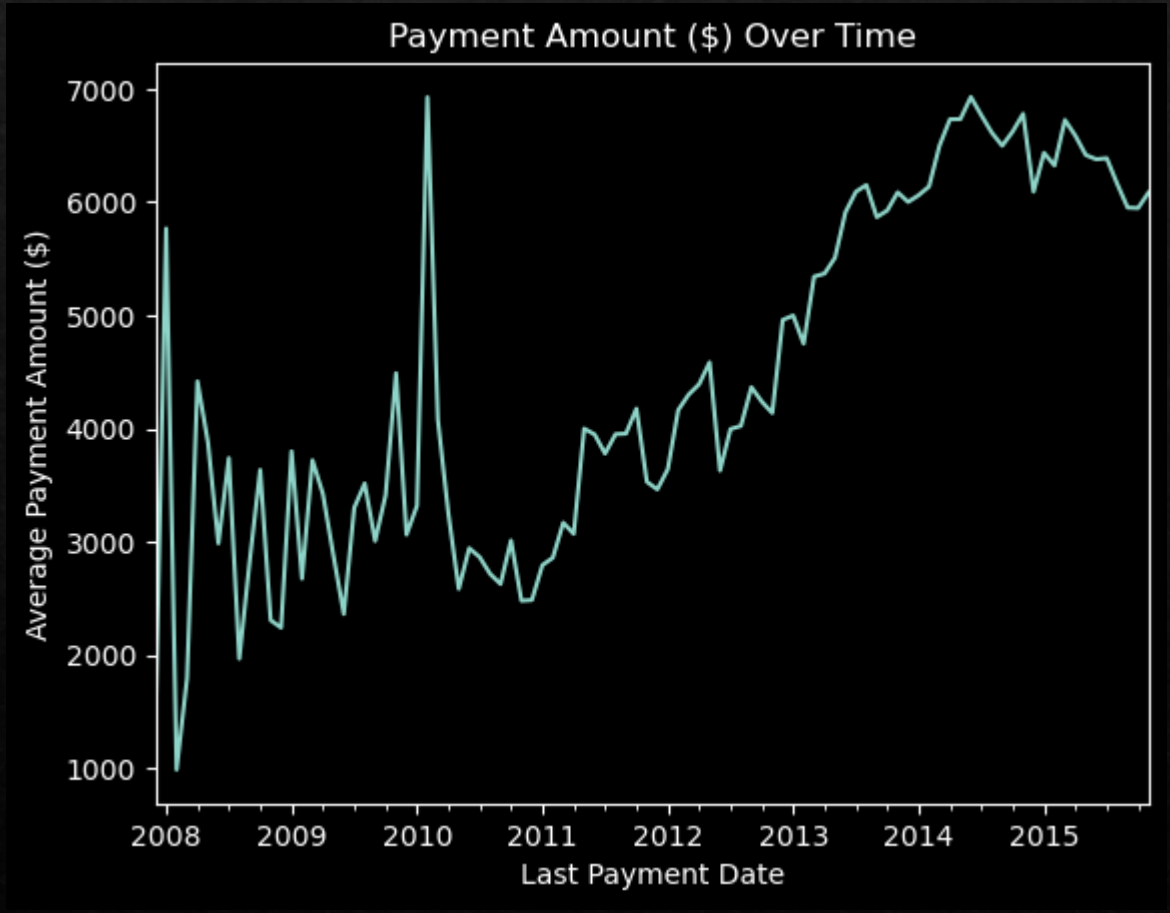


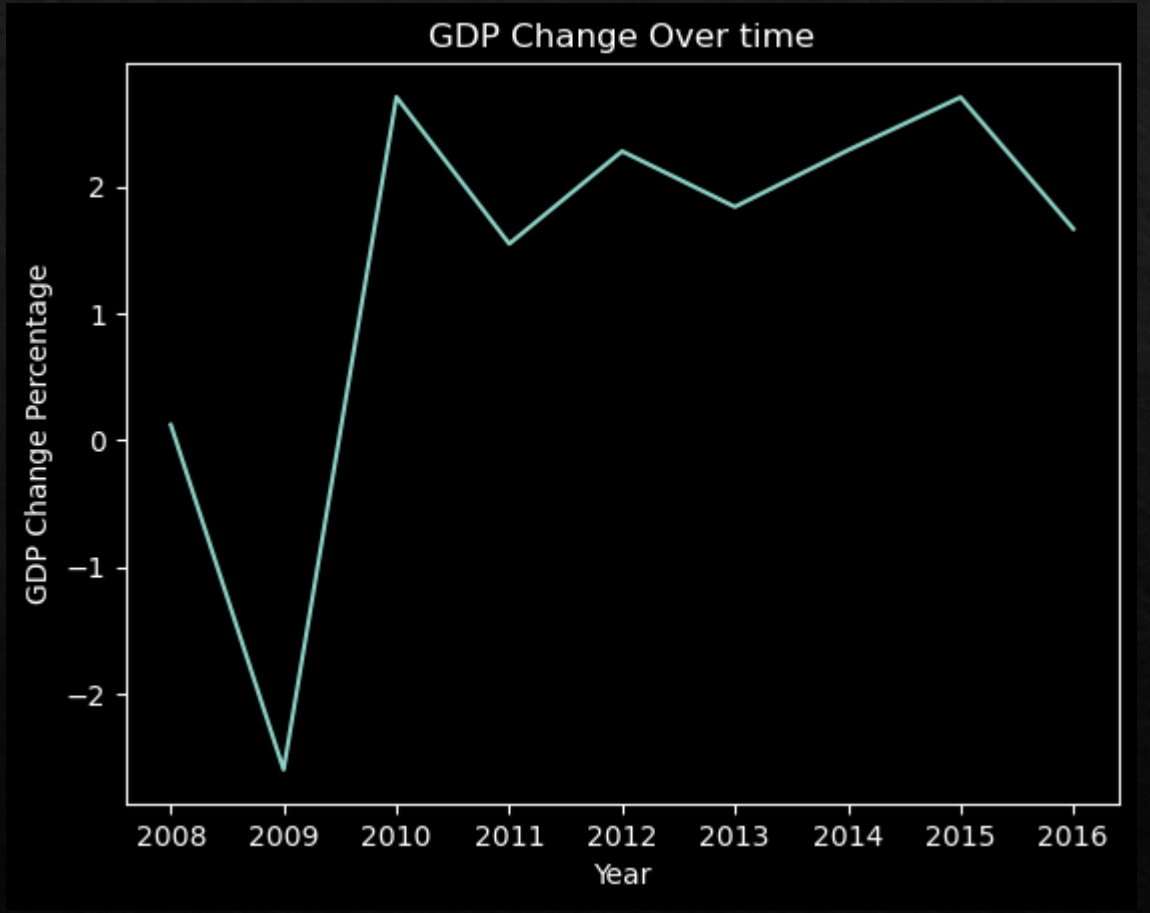
EDA

PCA graph of the features with bad outcomes colored in red and good outcomes colored in green



External Factors

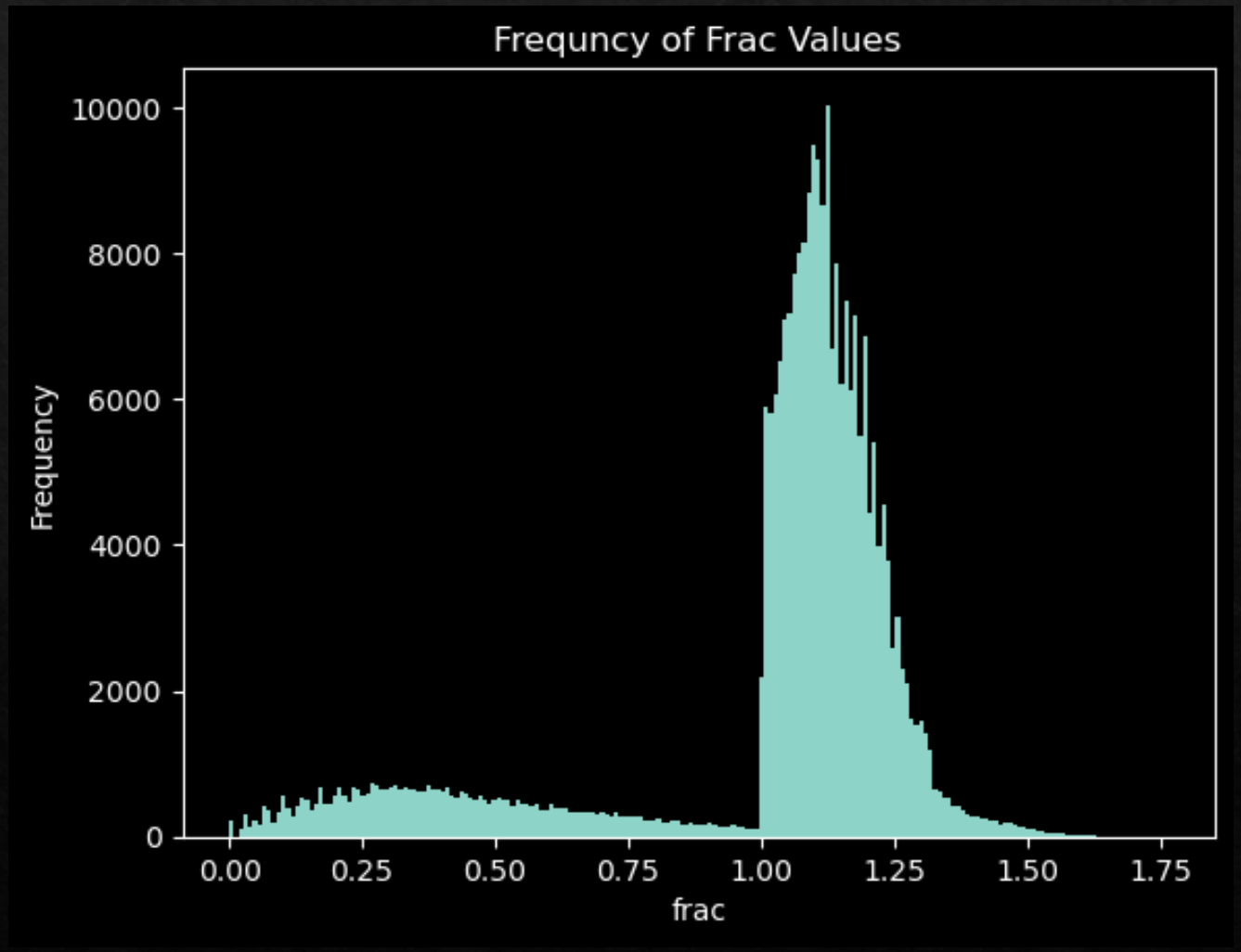




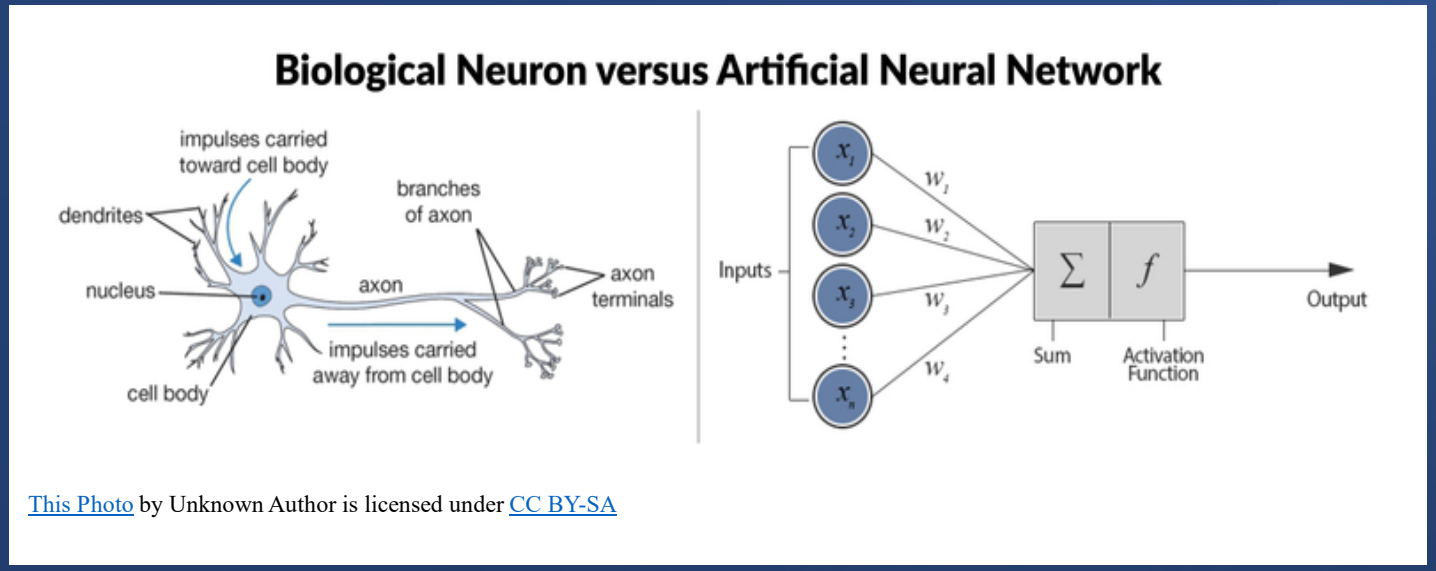
Preprocessing

- $Expected\ Payment = Principal * (1 + Interest\ Rate) * Years$
  - Where $Interest\ Rate$ is in decimal form
- $Actual\ Payment = Total\ Payment + Total\ Late\ Fees + Recoveries$
- $frac = \frac{Actual\ Payment}{Expected\ Payment}$
- $Expected\ Payment = Principal * (1 + Interest\ Rate) * Years - Actual\ Payment$

EDA - *frac*


Frequncy of Frac Values
Frequency
10000
8000
6000
4000
2000
0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
frac

Modelling



Neural Networks

Diagram of the similarity between biological neuron and a node in a neural network

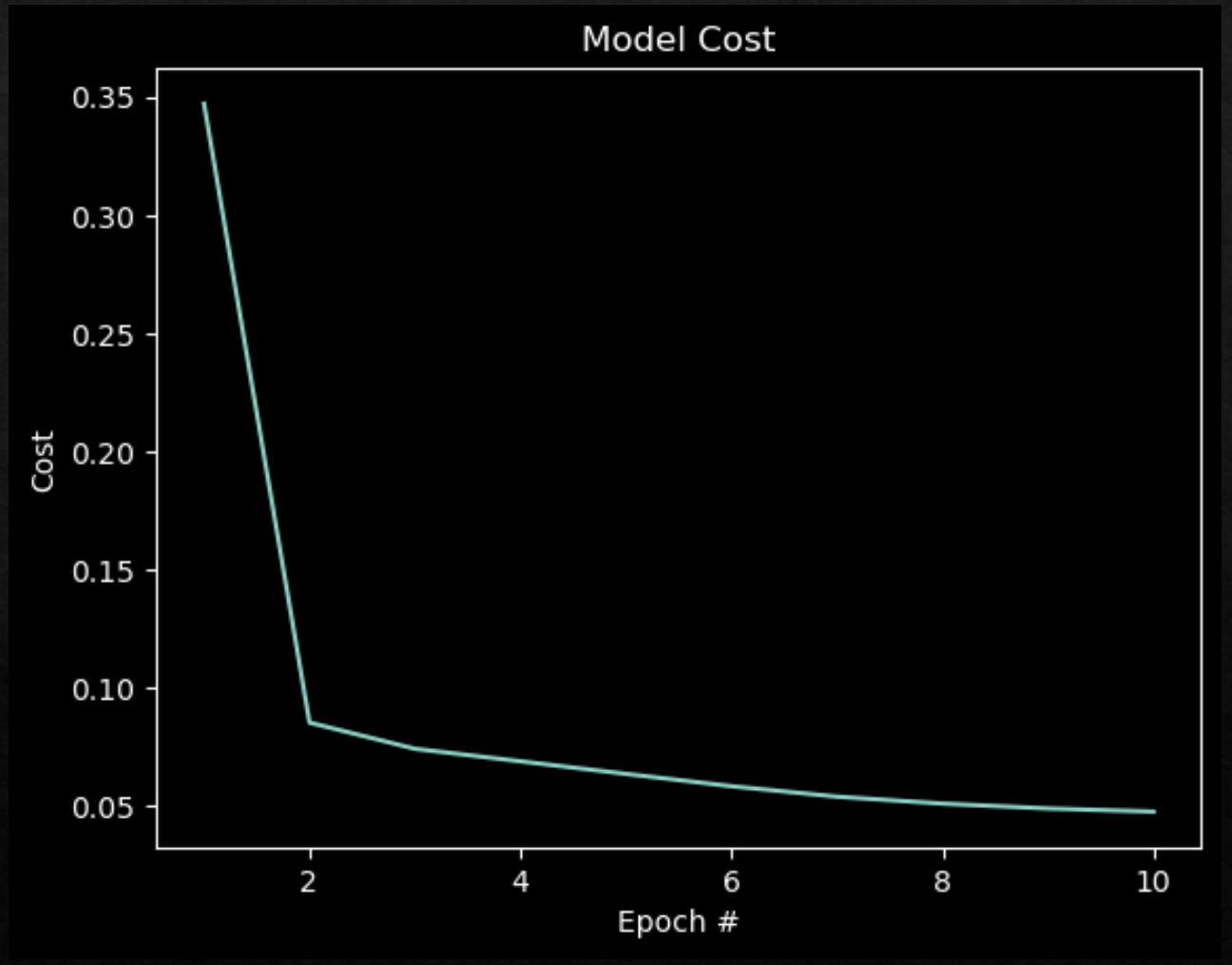
$$Output = f(\sum_1^n w_i x_i) + b$$

Formula for a node in a neural network

Metrics

- MAE
  - $MAE = \frac{\sum_{i=1}^{n} |error_i|}{n}$
- MSE
  - $MSE = \frac{\sum_{i=1}^{n} error_i^2}{n}$
- RMSE
  - $RMSE = \sqrt[2]{\frac{\sum_{i=1}^{n} error_i^2}{n}}$
- Percent Error
  - $Percent\ Error = \frac{\sum_{i=1}^{n} \frac{error_i}{pred_i}}{n}$

Training



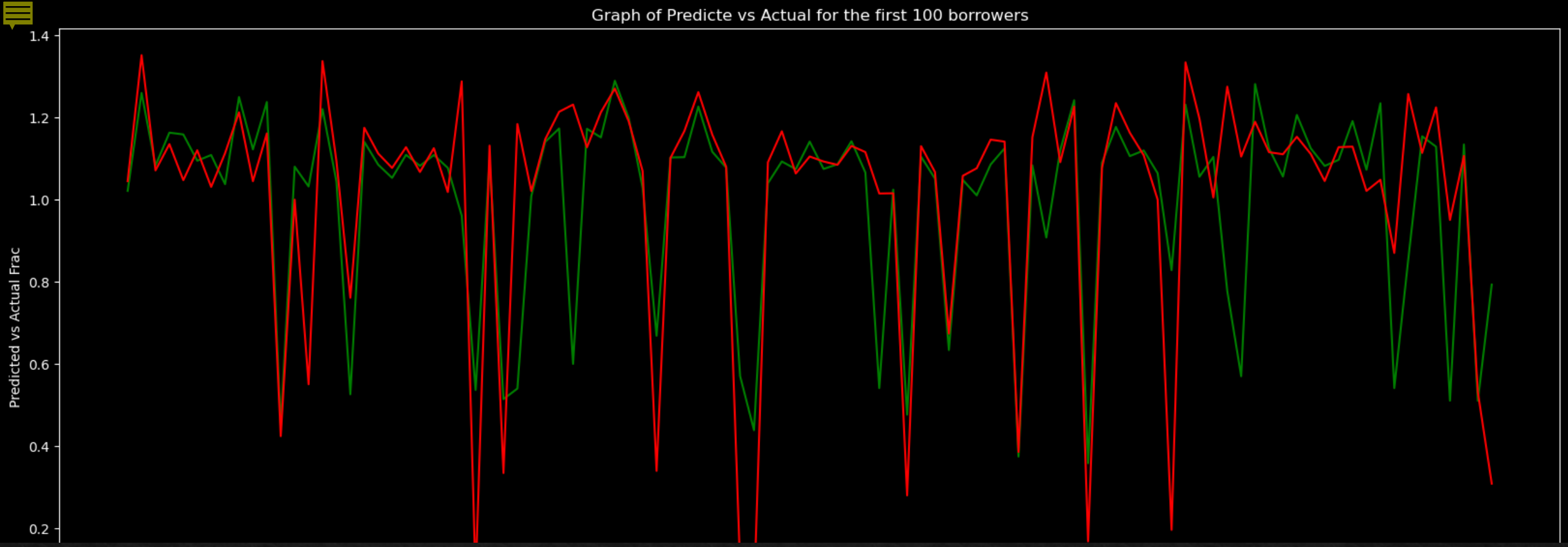
Testing (Neural Network)

MSE = 0.0490,

RMSE = 0.2213,

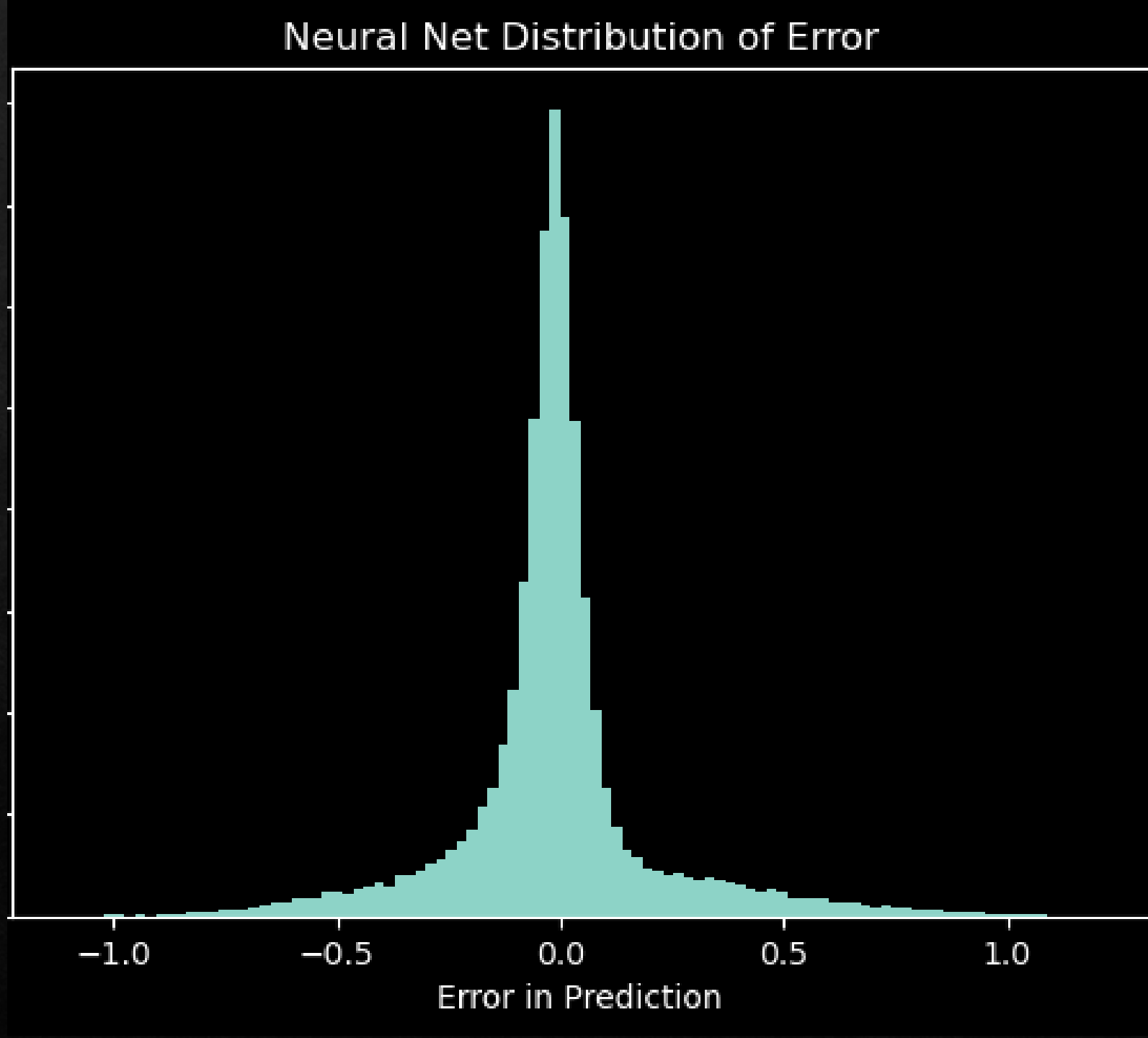
MAE = 0.1386,

Mean Percent Error of 14.6275%.

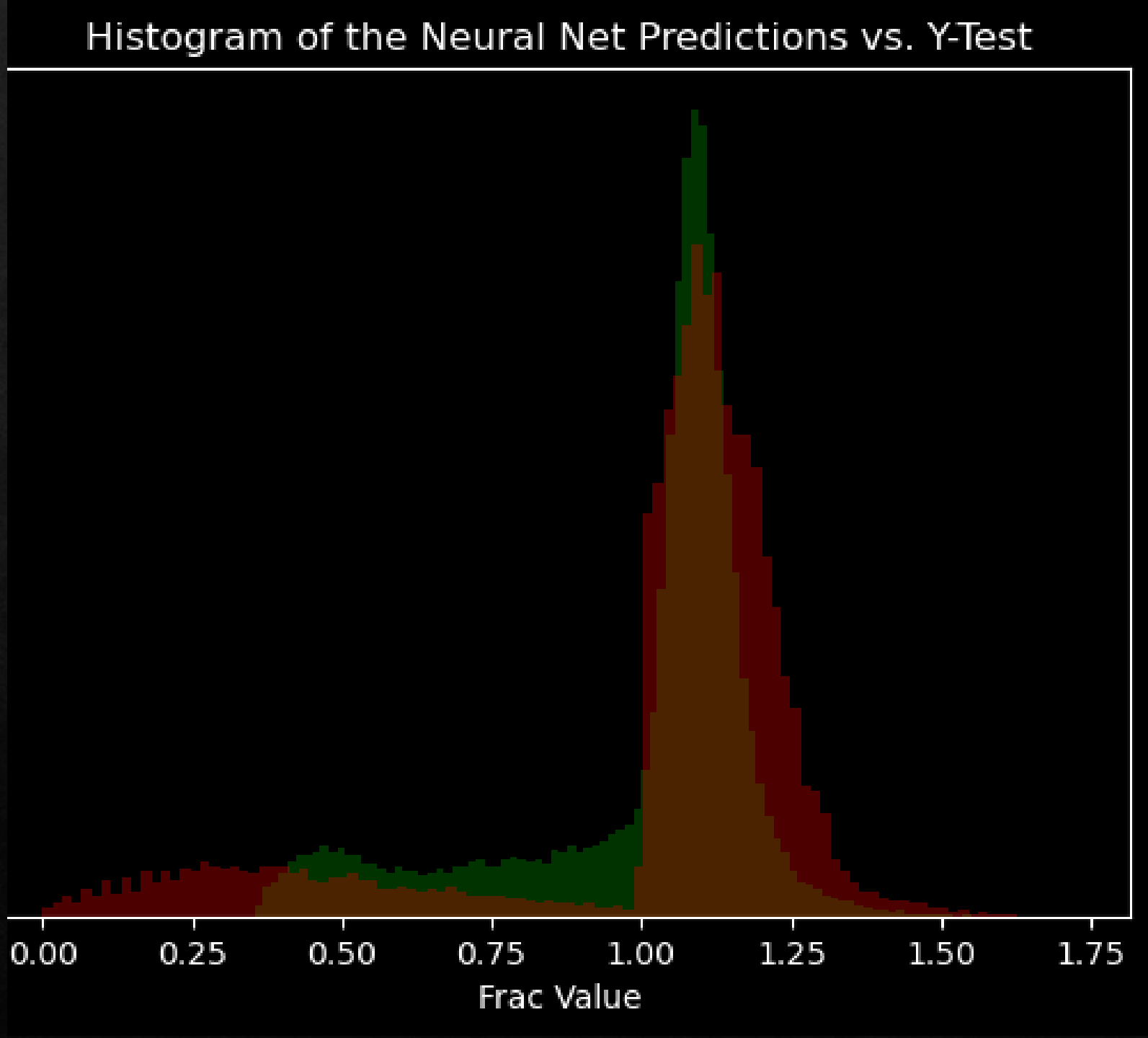


Testing NN cont.

Testing NN cont.



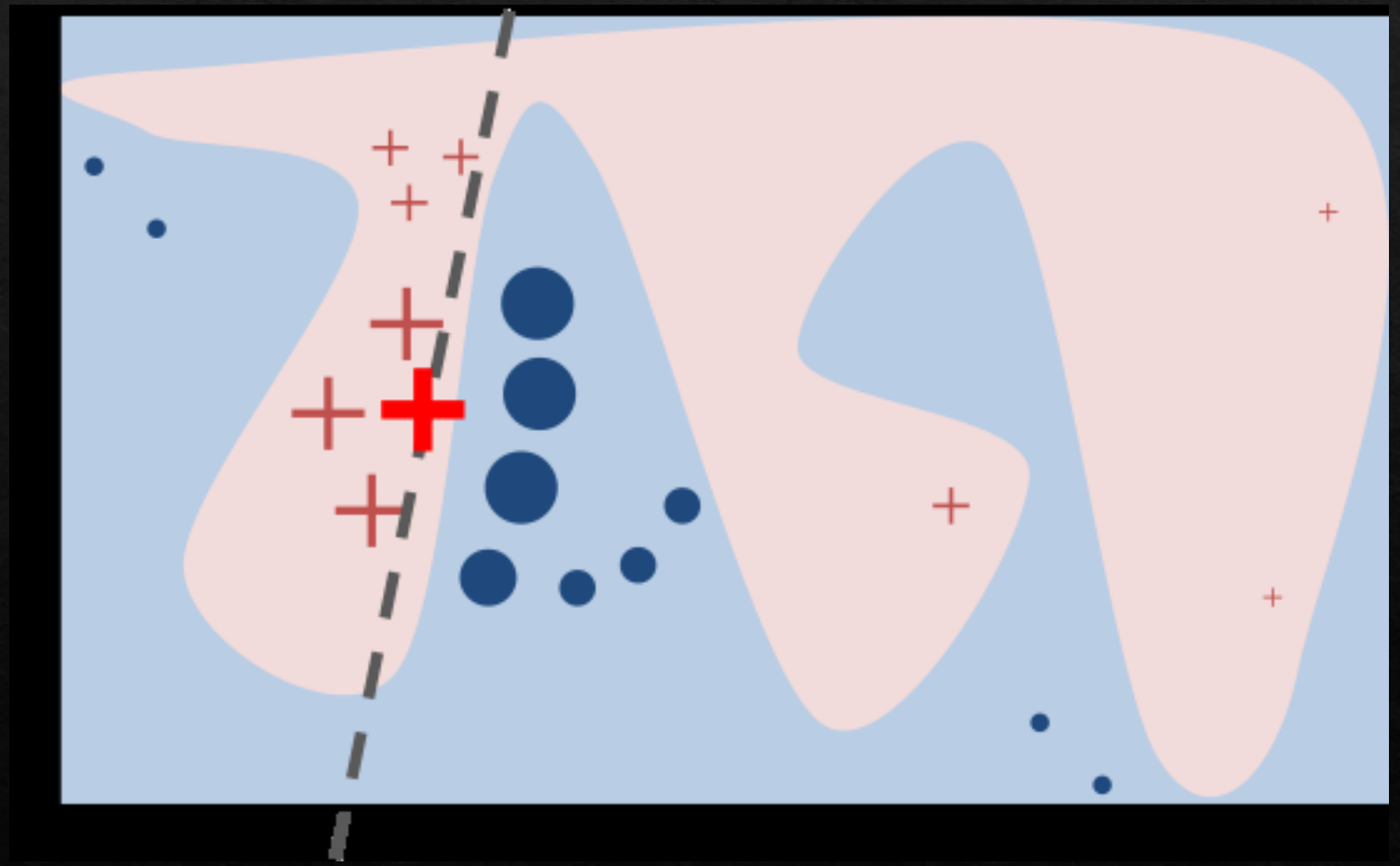
Testing NN cont.

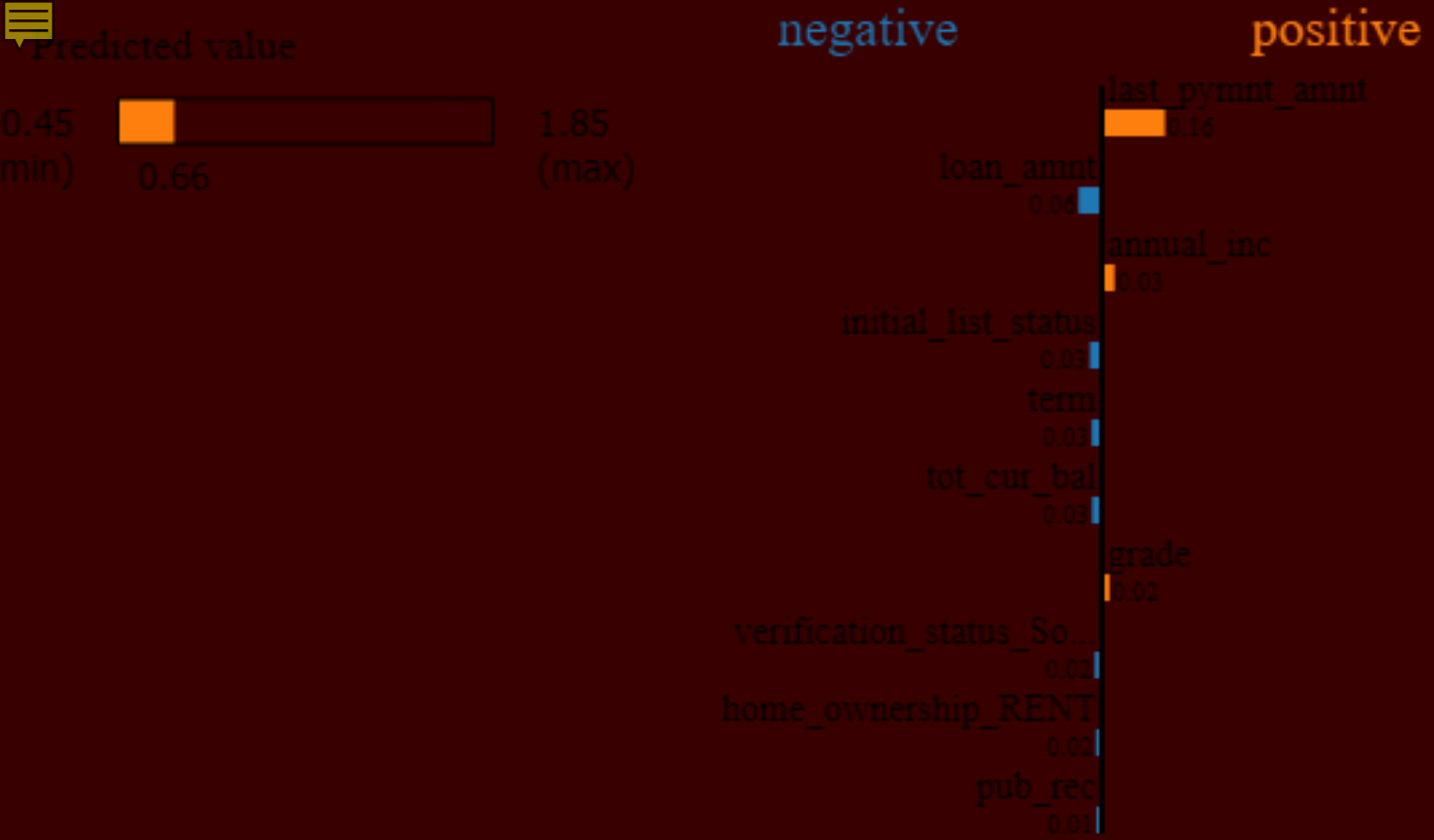


LIME

- LIME stands for Locally Interpretable Model Explanation and gives simpler explanations to how complex machine learning models arrive at a prediction
- $\xi(x) = \underset{g \in G}{\operatorname{argmin}}(\mathcal{L}(f, g, \pi_x) + \Omega(g))$
- I created a function that could take in the number of any loan and create a LIME explanation for that loan

LIME cont.

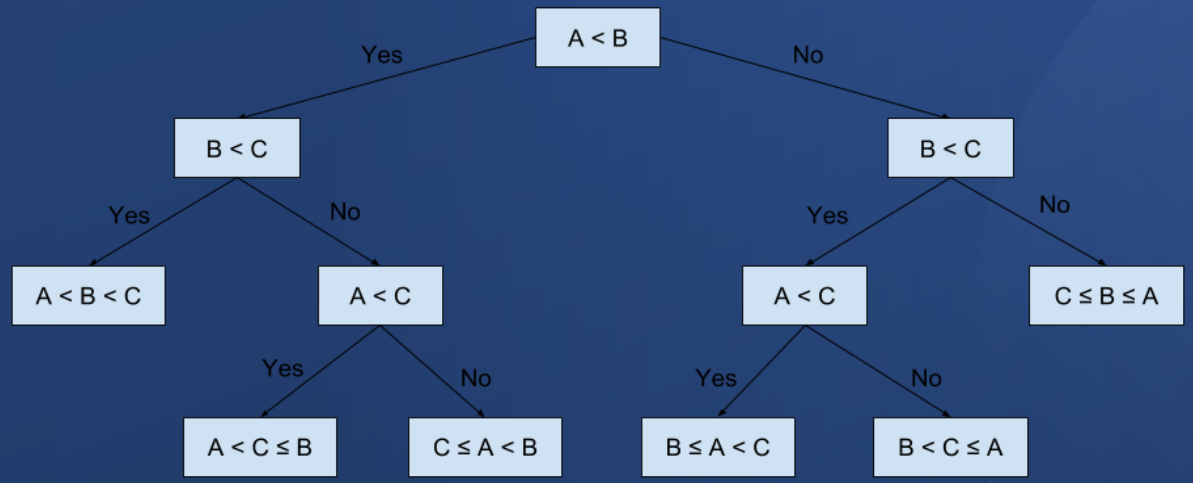




Feature	Value
last_pymnt_amnt	-0.33
loan_amnt	0.20
annual_inc	-0.40
initial_list_status	-0.97
term	-0.65
tot_cur_bal	-0.45
grade	-0.47
verification_status_Source Verified	-0.77
home_ownership_RENT	1.22
pub_rec	1.38

LIME cont.

Modelling



Random Forest

Model of a single decision tree in a random forest

$$Gini = 1 - \sum_{i=1}^{n} {p_i}^2$$

Purity measurement formula

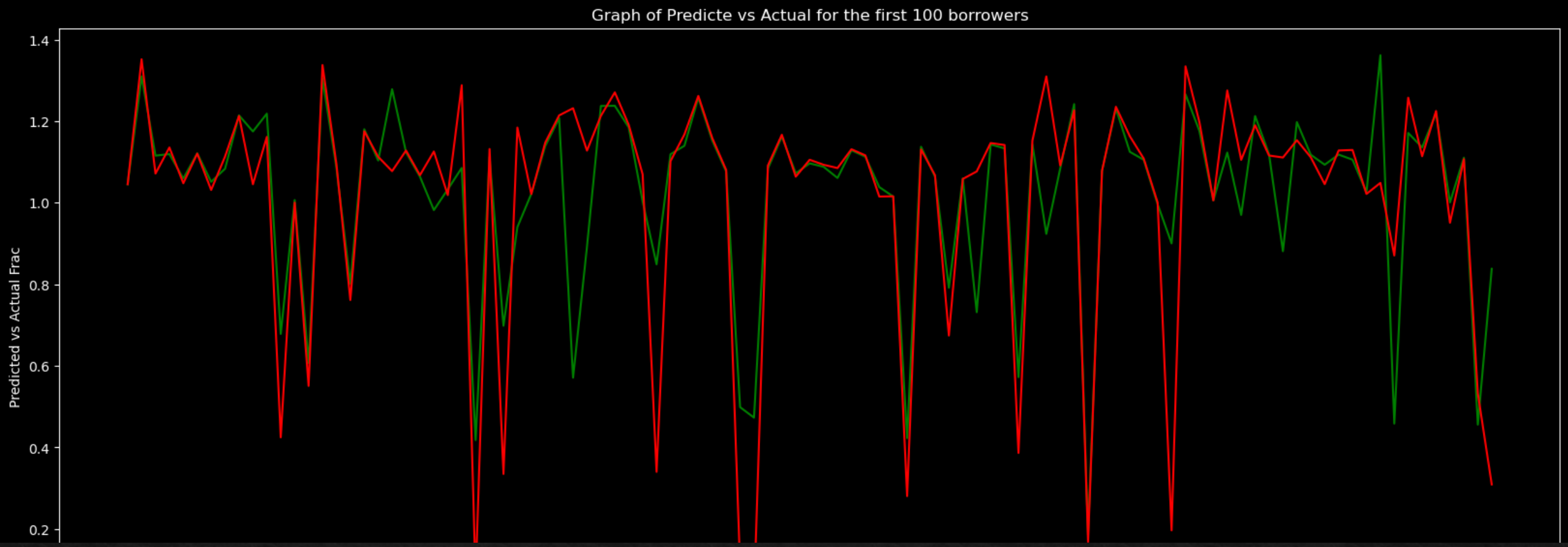
Testing (Random Forest)

MSE = 0.0309,

RMSE = 0.1758,

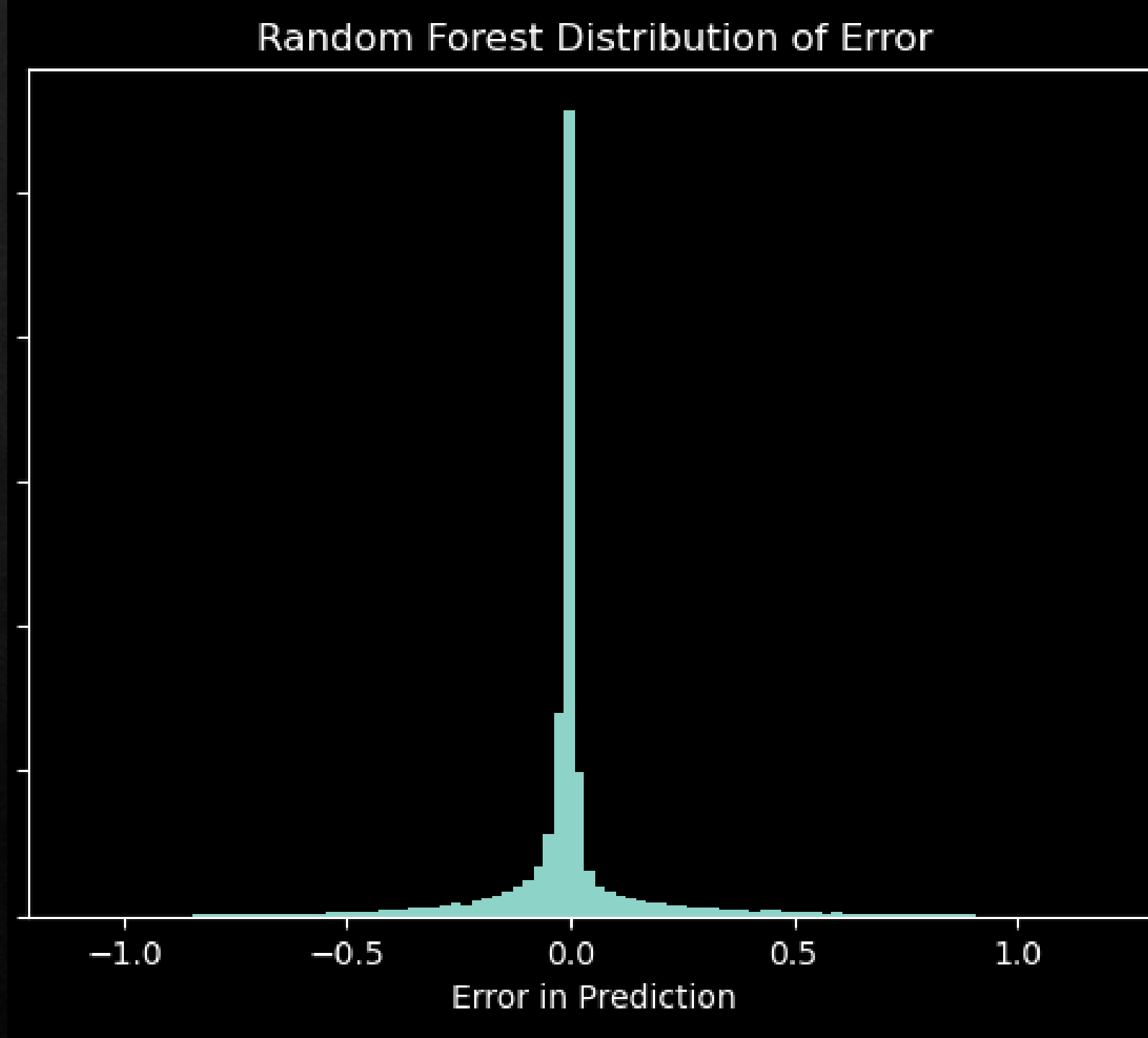
MAE = 0.0872

Mean Percent Error of 13.6849%.

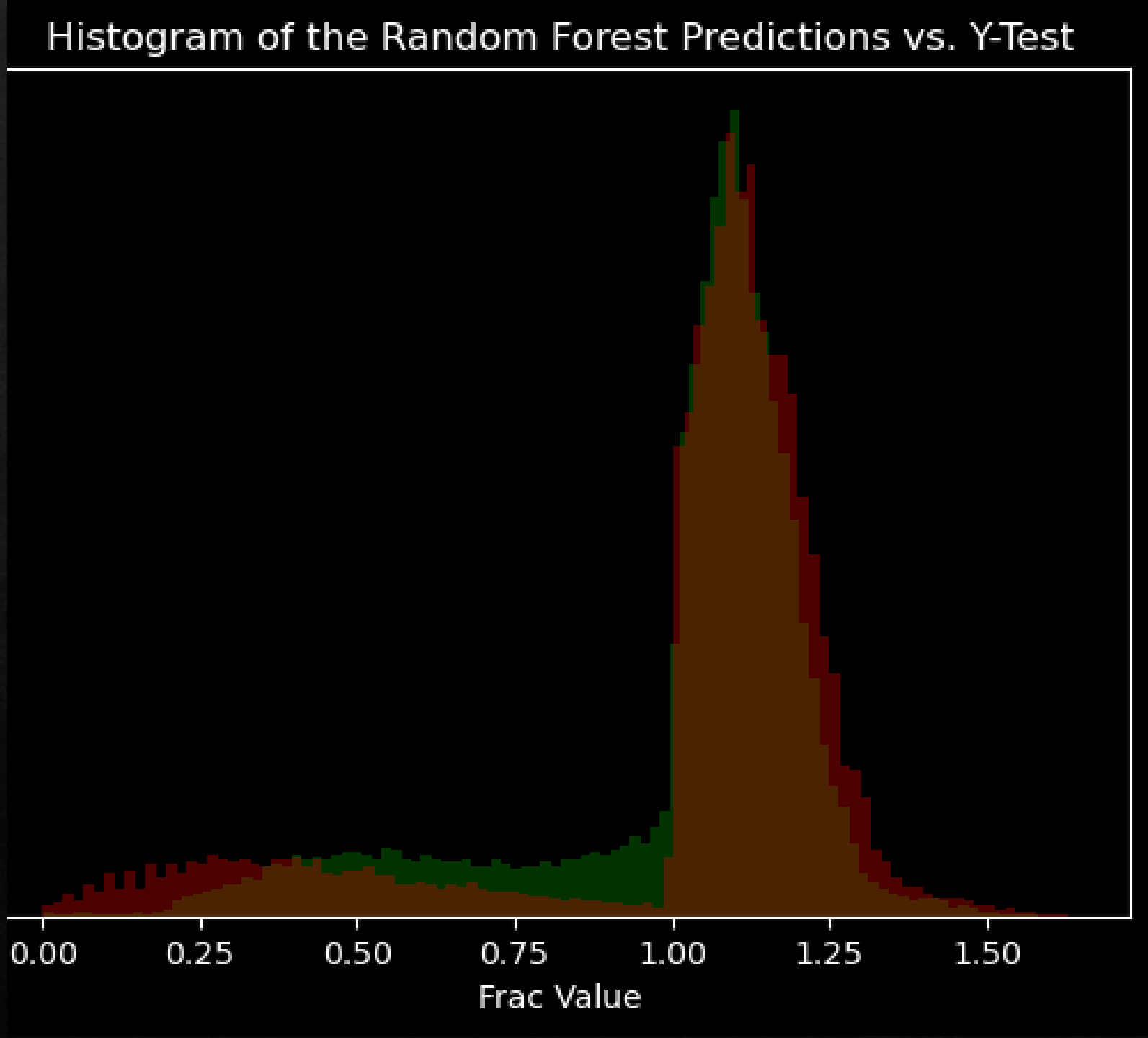


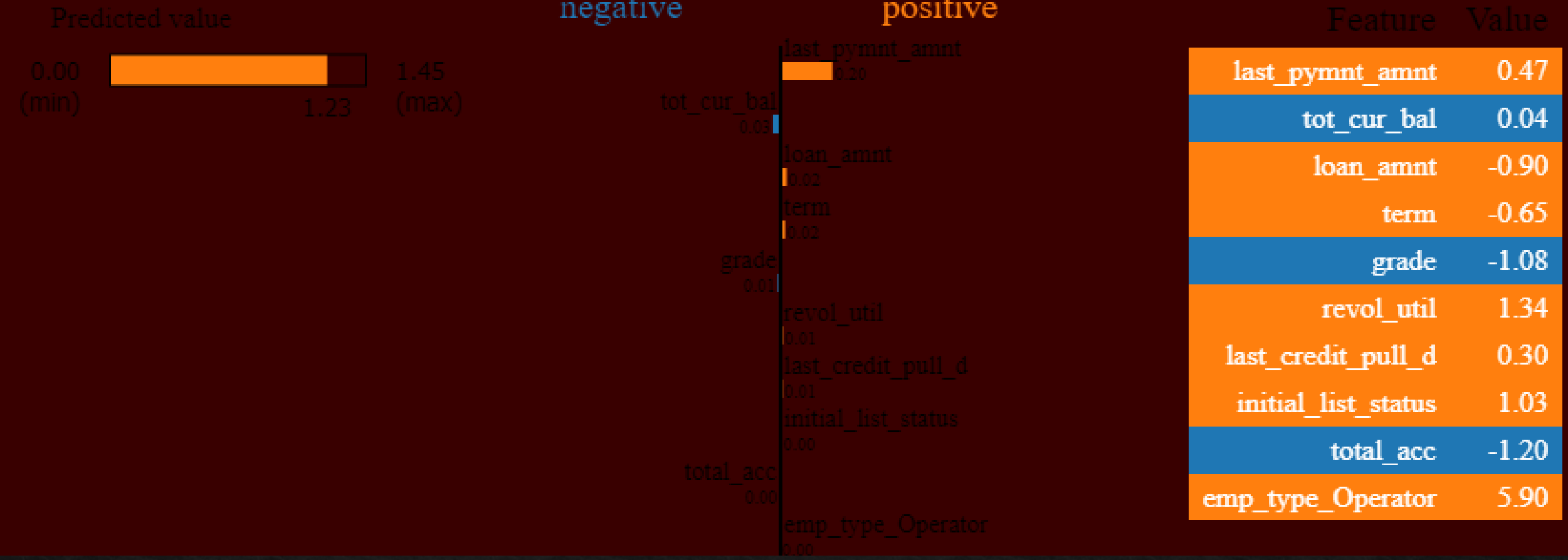
Testing (RF) cont.

Testing (RF) cont.



Testing (RF) cont.




Predicted value
0.00
(min)
1.23
1.45
(max)
negative
positive
last_pymnt_amnt
0.20
tot_cur_bal
0.03
loan_amnt
0.02
term
0.02
grade
0.01
revol_util
0.01
last_credit_pull_d
0.01
initial_list_status
0.00
total_acc
0.00
emp_type_Operator
0.00
Feature
Value
last_pymnt_amnt 0.47
tot_cur_bal 0.04
loan_amnt -0.90
term -0.65
grade -1.08
revol_util 1.34
last_credit_pull_d 0.30
initial_list_status 1.03
total_acc -1.20
emp_type_Operator 5.90

LIME (RF)

Conclusion

- The most important methods from this paper are:
  - Random Forest Modelling
  - Locally Interpretable Modelling Explanation (LIME)
- With most of the dataset comprised of current loans, it is impossible to evaluate the accuracy and precision of supervised learning methods.