Supervised Capstone:

Predicting Zillow Zestimate residual error (logerror).

Background information

- Zillow created “Zestimate” which gives customers a lot of information about homes and housing markets at no cost by using publicly available data.
- 7.5 million statistical and machine learning models that analyze hundreds of data points on each property are used by Zillow to create and improve “Zestimate”. They improved median margin of error from 14% to 5%. Zillow announced a Kaggle competition to improve the accuracy of “Zestimate” even further.
- This problem can be handled by a typical supervised machine learning, because supervised learning algorithms learns and analyzes labeled training data and then generates function to predict output.

Problem Statement

- Build a model to improve the Zillow Zestimate residual error (also know as the log-error)#(i.e. “Zestimates” are estimated home values based)

Research Questions:

- Is there a strong correlations between the property features and the logerror?
- What is the best model to use for predicting the logerror?
- What are the most important features used by the strongest performing model?

Solution statement

- The goal of this project is to predict the log error between Zestimate and actual price, the following steps will be followed to accomplish this goal:
- Clean the data – includes dealing with Null and missing values, as well as converting non-numerical data to numerical data and remove outliers.
- Explore the data and understand the correlations between our features and predictor variable.
- Split the dataset into training and testing set
- Develop four models and determine which model yields the best performance.

Evaluation metrics

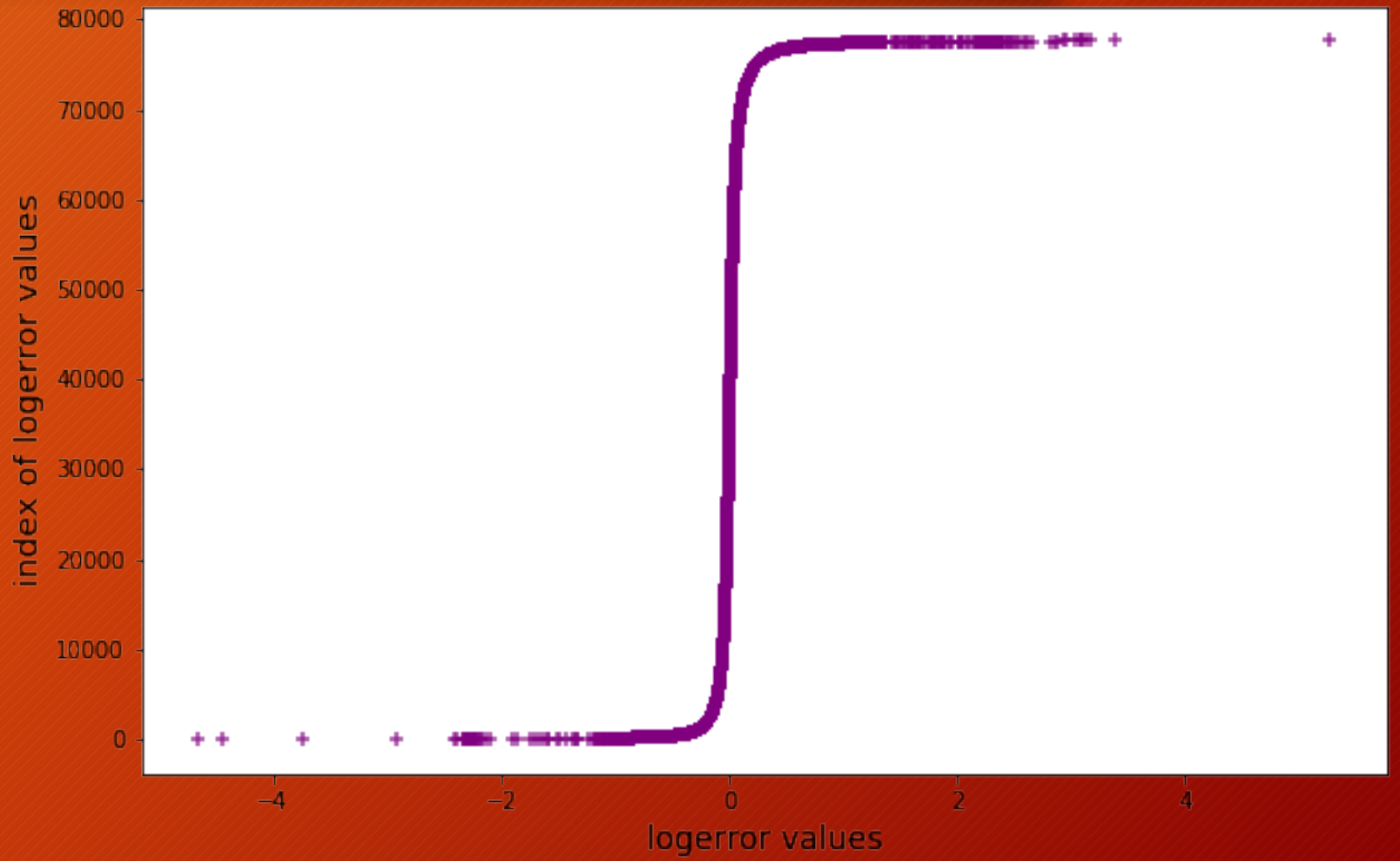
- The four models that will be used will produced metrics for
  - R2 – total variance explained by model / total variance. **Variance** Is a measure of how far observed values differ from the average of predicted values, i.e., their difference from the **predicted value mean**
  - MSE - is the average of the square of the errors. Error is the difference between the observed values and the predicted values.
  - RMES – Square root of MSE. It's the square root of the average of squared differences between prediction and actual observation.
- The best score based on the RMES will determine which model will be selected. RMSE is more useful when large errors are particularly undesirable.

Gathering Data

- Today I will be presenting my analysis and models based on 2017 property dataset, which contains features that can be used to determine the value of a house. Understanding the features and it's relationship to the logerror, can be used to to help a model increase the accuracy of the logerror prediction. The following links below are sites I used to gather more information about the dataset.

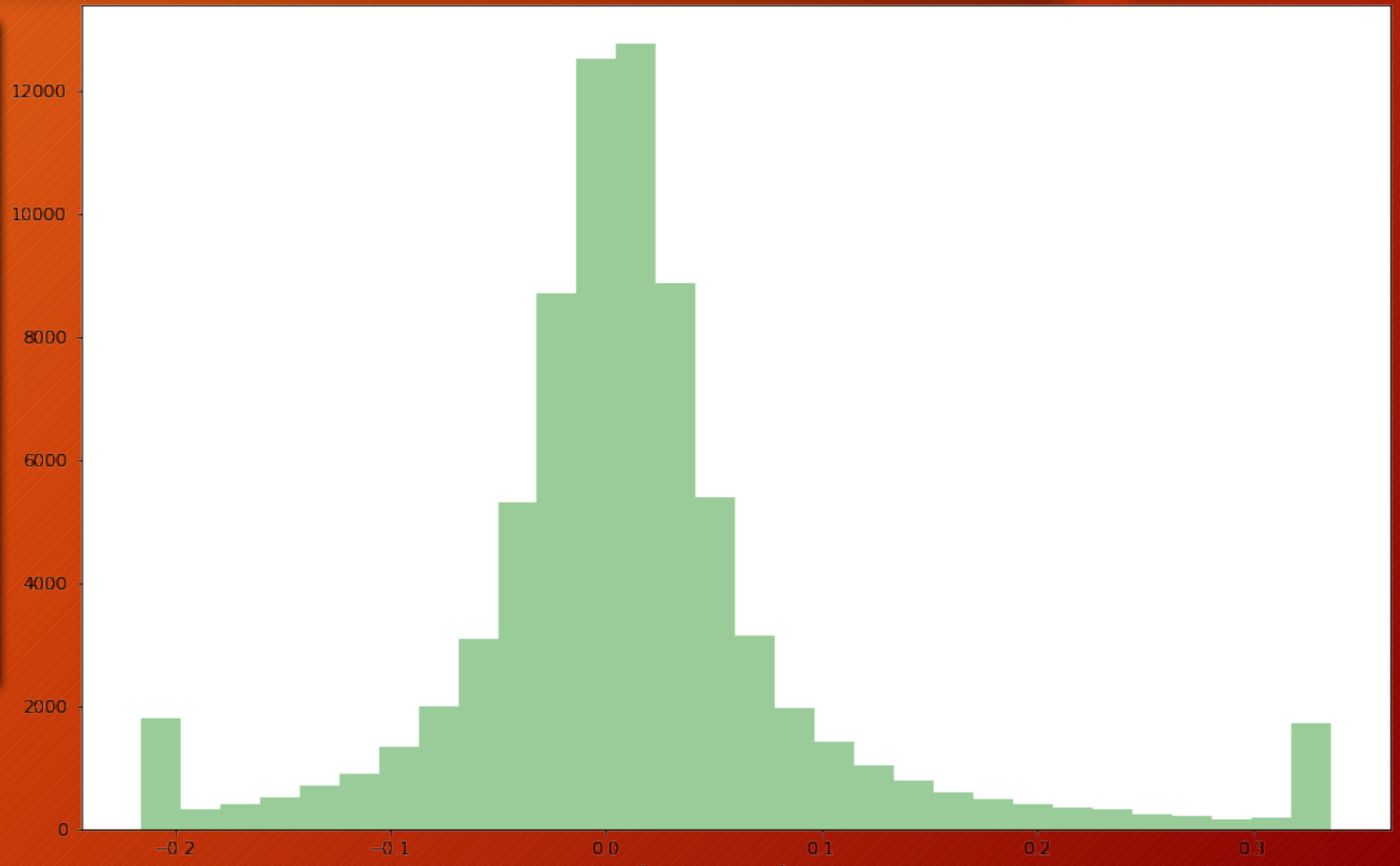
Analysis

Looking at the data specifically the logerror and determine if we have outliers.



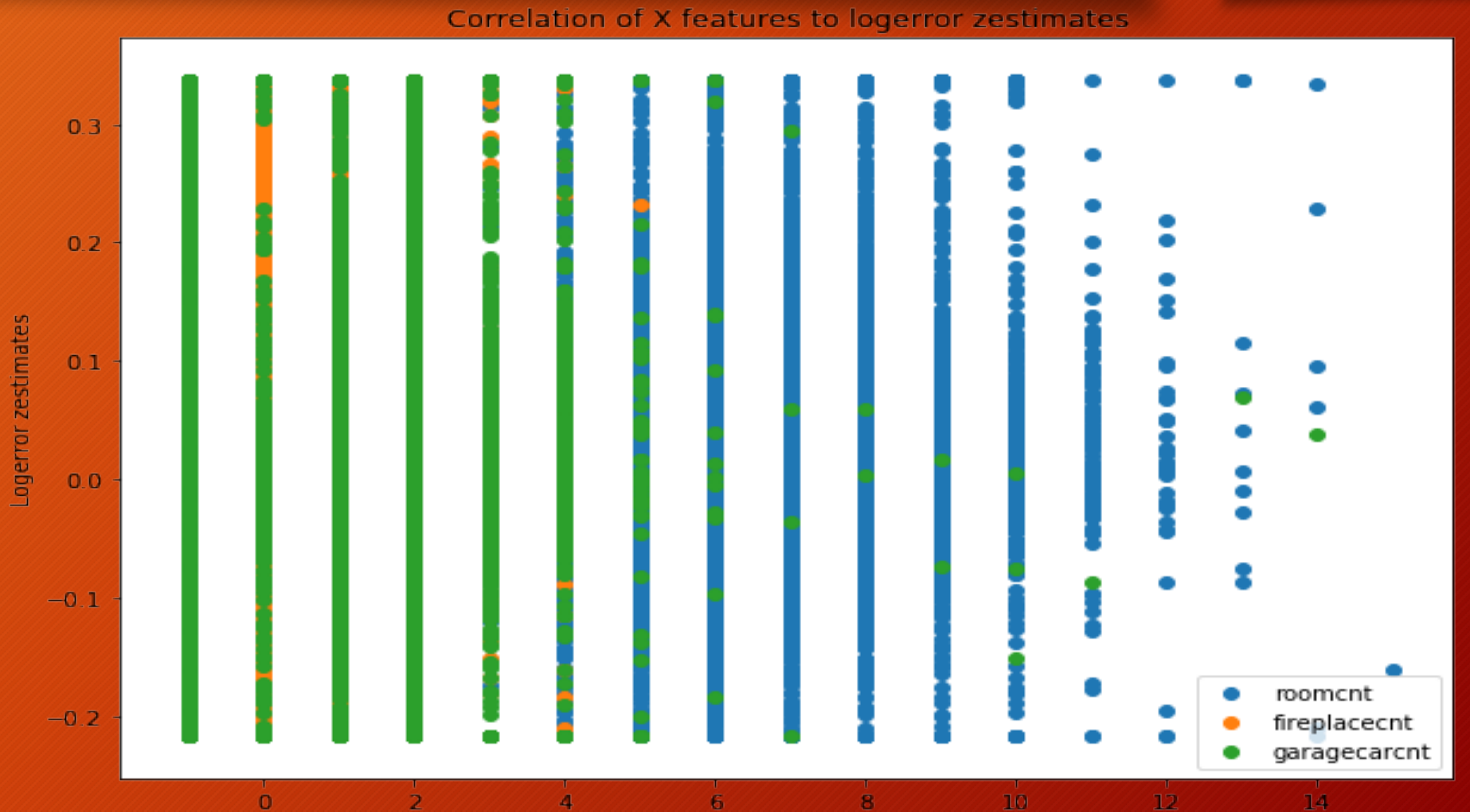
Analysis continuation

We can see from graph above, it has a good normal distribution. Since logerror = log(Zestimate) - log(actual price), negative logerror value means the Zestimate is underestimated, positive value means the Zestimate is overestimated. We can see from this normal distribution, it estimates pretty accurate most of the time. We need to find out when Zesimate does well and when it doesn't. We need to find out correlations between different features and logerror.



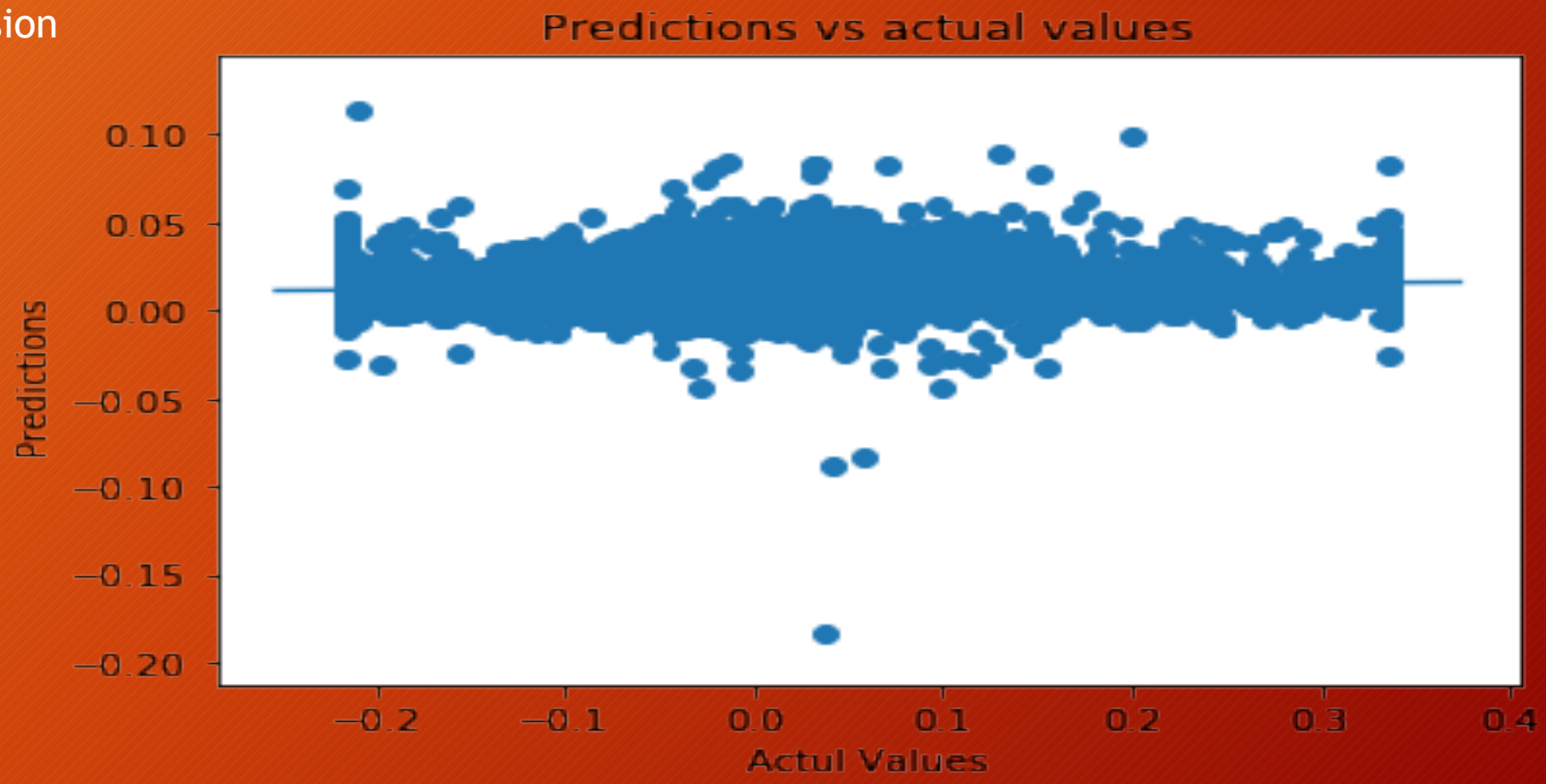
Analysis continuation

The graph above tells us that roomcnt and garagecarcnt have a correlation with the logerror but fireplacecnt doesn't have a strong correlation. This also tells me that as the Zestimate estimates performs well for bigger size houses but fireplacecnt does not have much impact on the Zestimate.



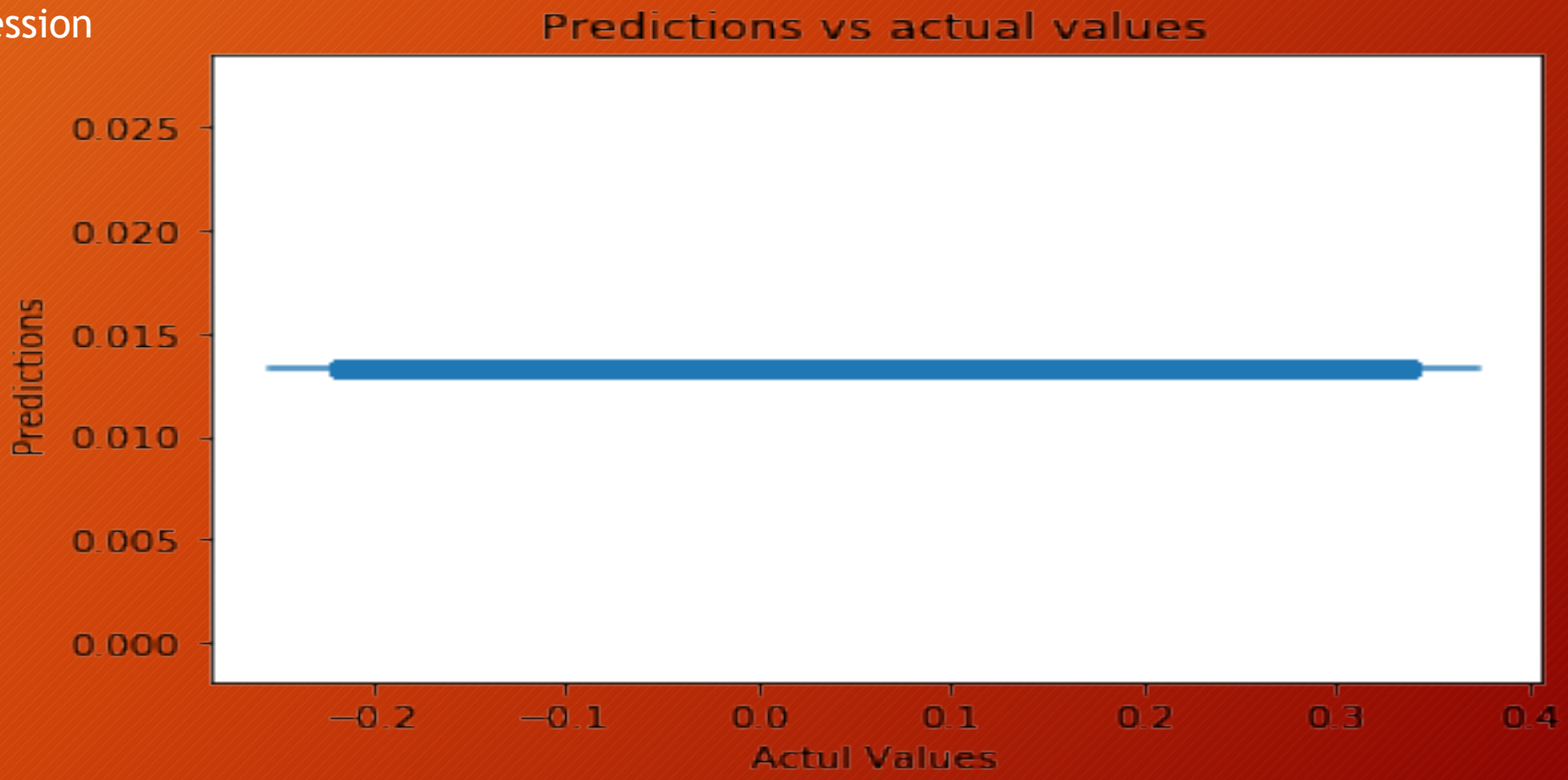
Analysis continuation

Prediction of Ridge Regression



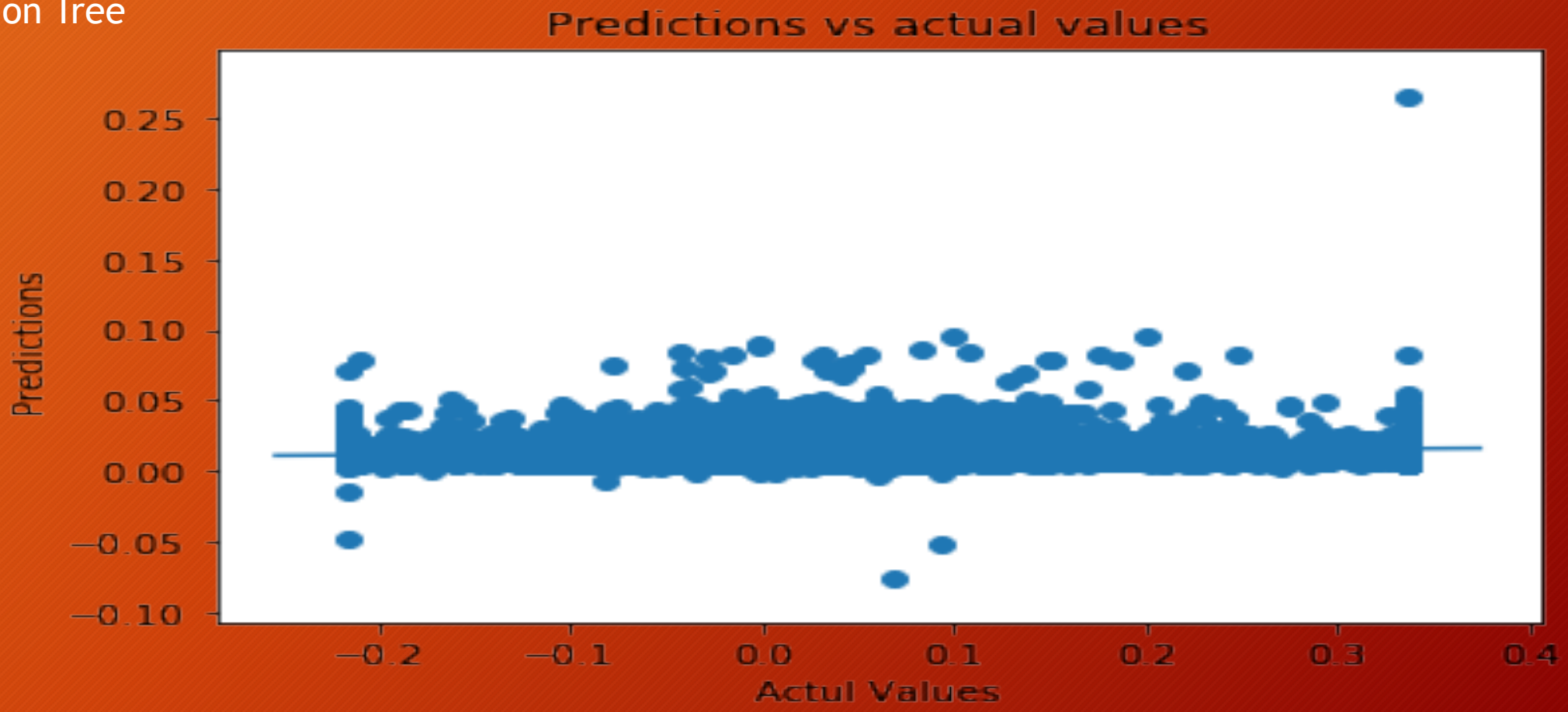
Analysis continuation

Predictions of Lasso Regression



Analysis continuation

Prediction of Decision Tree



Model results

Overall Random Forest performed the best out of all the models.

	Model	Best Parameter	CV mean train	CV mean test	R2	MSE	RMSE
0	Ridge	{'ridge__alpha': 0.01}	0.011170	0.007842	0.004940	0.007981	0.089338
0	Lasso	{'lasso__alpha': 0.01}	0.000000	-0.000042	-0.000127	0.008022	0.089565
0	Random Forest	{'randomforestregressor__max_depth': 5, 'rando...	0.215921	-0.008733	0.008549	0.007952	0.089176
0	Decision Tree	{'decisiontreeregressor__max_depth': 5, 'decis...	0.007819	0.003343	0.003590	0.007992	0.089398

Summary

- My benchmark model results are not good. The goal of the project was to build a model to improve the Zillow Zestimate residual error (also know as the log-error)
- The current Zestimate is at 0.06, therefore my model hasn't made an improvement.
- However, the next set of steps would be to perform feature selection using only the most important features and determine if the model improves. Based on our analysis above there are few columns we can be further investigated to see if they can provide further insight and more of a correlation with the output variable. I would also look to make more changes with the hyperparameters. If all of the above still does not improve performance then I would look to evaluate other models.