Title

## presented by

Team Member 1

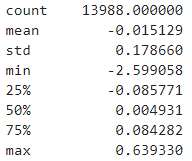
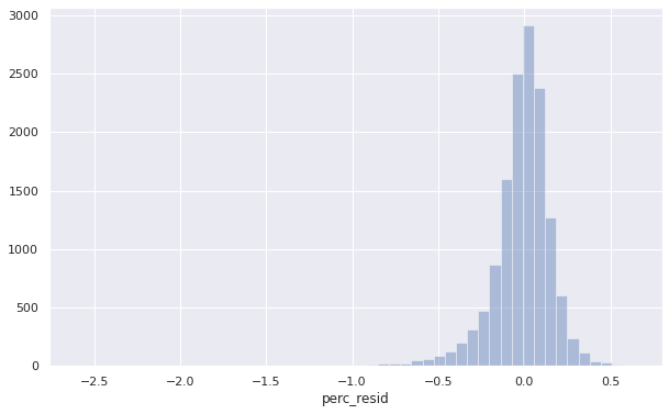
Team Member 2

Team Member 3

1. Background

Using on the data provided by the client, our goal was to predict \_\_\_y\_\_\_ as accurately as possible. We used a Gradient Boosted Regression Tree. Our model was able to achieve an RMSE of about ####. This means that on average our predicted \_\_\_\_y\_\_\_ was #### away from the actual price. The R2 value for our model is 0.##.

Perhaps a better measure of model performance is how close to the true value of \_\_\_y\_\_\_ the model predicts, as a percentage of the true value of \_\_\_y\_\_\_. Below is a chart and a summary table to describe how well our prediction measures up to the actual value as a percent of the actual value:



You can see that over 50% of the predictions where within 10% of their true value. There were a few instances where the model groslly overpredicted the value of \_\_\_y\_\_\_. These overpredictions tended to occurr for observations where \_\_\_x\_\_\_ was low or \_\_\_z\_\_ was equal to [CATEGORY].

1. Methodology

Key restrictions were applied to the data, and properties with zero \_\_\_w\_\_ were excluded. One observation had a value of ## for \_\_\_a\_\_\_, it will excluded; as were all observations greater than ##### on \_\_\_\_\_b\_\_\_.

A gradient boosted regression model with 600 trees was used to make the predictions. The model was actually built to predict log of \_\_\_\_x\_\_\_. In this way, the model attempts to minimize the error as a proportion of the \_\_\_y\_\_\_, rather than seeking to minimize the amount of raw errors in the predictions. In other words, the model does not give extra weight to large \_\_\_y\_\_\_ values, but rather to the low values of \_\_\_y\_\_. Thus hoping that the large \_\_\_y\_\_\_ values do not overwhelm the model.

1. Individual Attributes

Many, many features were included in the model. However, a handful of features made the biggest difference. Here you can see that knowing whether an observation belonged to \_\_\_category p\_\_\_ or \_\_\_category q\_\_\_ had the largest impact on the model. \_\_\_g\_\_\_ was the second most important factor. The variable was represented with bins, defined as 0-10, 11-20, 21-30, 31-40, 41-50. As shown below, \_\_\_\_f\_\_\_ , \_\_\_h\_\_\_ and whether it was \_\_\_k\_\_ were all also very important in making an accurate prediction.

[GRAPH PLACEHOLDER]

1. Python Notebooks