**Introduction:**

Imagine yourself as a data scientist working for a real estate agency. You are going to predict the log-error between the price Zillow estimates, *Zestimate*, and the actual sale price, given all the features of a home. The log error is defined as

When log-error is <0, it means that Zillow underestimated and the house was sold at the higher price than what Zillow suggested. The goal is to make a model that predicts the log-error whose average of |log-error| is lower than what Zillow’s average of |log-error|:

**Data Description:**

1. You are provided with a full list of real estate properties in three counties (Los Angeles, Orange, and Ventura, California) data in 2016.

2. The train data has logerror values by parcelid.

* Properties\_2016.csv - all the properties with their home features for 2016.
* Train\_2016.csv - the training set with transactions from 1/1/2016 to 12/31/2016.
* Zillow\_data\_dictionary.xlsx - Refer for more information.

**Grading Scheme:**

|  |  |
| --- | --- |
| Effort + Modeling | 25% |
| Cleanness and Explanation | 20% |
| EDA | 35% |
| Top 5 features | 10% |
| Better than Zillow? | 5% |
| \*Ranking Score | 10% |

This project is a competition. We have a total of 35 students. The ranking score will be graded as the following table.

|  |  |
| --- | --- |
| Ranking | Points |
| 1 – 5 | 10% |
| 6 – 10 | 9% |
| 11 – 15 | 8% |
| 16 – 20 | 7% |
| 21 - 25 | 6% |
| 26 – 30 | 5% |
| 31 - | 4% |

**Two Major Questions:**

1. What properties of house overprices the house?
2. What is the best algorithm to predict logerror?

**Tasks:**

1. In the properties data set, there are 58 features. Do some EDA, feature engineering, and feature selection. You learn about training data any ways you want to but must answer the followings.

* Determine the number of missing values on each feature. Determine the parceled that has the greatest number of missing features. **You do not drop the feature unless a particular feature has missing value ratio more than 90%.**
* Determine the number of cities in each county.
* Determine which county has the highest and lowest logerrors in each city.
* Determine any outliers. **You can drop up to 20 outliers.**
* There are total of 17 columns containing column names with ‘id’. Referring to the data dictionary, which id’s can be eliminated? Justify your reasons.
* Determine the correlations between the size and the room numbers.
* Split the dataset to **80-20%** for train and test sets. Show the distributions of log-error by counties.

1. Use a random forest classifier to determine top 5 most weighted features dominated in overpricing houses. Evaluate your result.
2. Use linear regression, decision tree, random forest, XGBoost, and ensembled models to predict the logerror. Which method provide the best result? Report the result.
   1. Linear regression – can use all features – computationally cheaper than other methods but model can be relatively complex.
   2. Decision tree, random forest, and XGBoost – may take several hours if all features are used. Feature extraction is very necessary – even 30 features can take up to an hour.
   3. Ensembled models – you are going to aggregate models together by weighting algorithms:

where is the weight that is less than 1 and , is the RMSE of each model. You can choose any combination of models. Report the best one.

1. Report the results using the table below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model # | Model | RMSE (R) | Number of Features (n) | Score | Beat Zillow? (Yes/No) |
| 1 | Linear Regression (+Regularization) |  |  |  |  |
| 2 | Decision Tree |  |  |  |  |
| 3 | Random Forest |  |  |  |  |
| 4 | XGBoost |  |  |  |  |
| 5 | Modelw |  |  |  |  |
| 6 | Modelavg |  |  |  |  |

The score can be calculated as follow:

where is the total features in the given data set. Present the score to 5th digits after the decimal.

1. Among 6 models, choose the best model and evaluate the model. Report the score value you want to enroll the competition.
2. Determine if you beat the Zillow model by comparing the average of |log-error|.