Homework 3
CSE 519 (Data Science Fundamentals)

Zafar Ahmad, 111195037
Ubaid Ullah Hafeez, 111195374
Submission Date: 10-01-2017

We tried a bunch of things which are as follows

- Clustering based on Desirability and Regression on Features
- Predicting logerror just using the desirability
- Regression on Desirability
- Regression Using some Selected Features
- XGBoost on Features
- Taking care of Outliers while training
- Different combination of features
- Polynomial feature engineering
- Polynomial Regression and Boosting
- Clustering based on Desirability and regression using XGBoost on selected features of properties already given to us.

Best Model

Our best model was based on Clustering and using multiple predictors based on XGBoost. What we did is as follows:

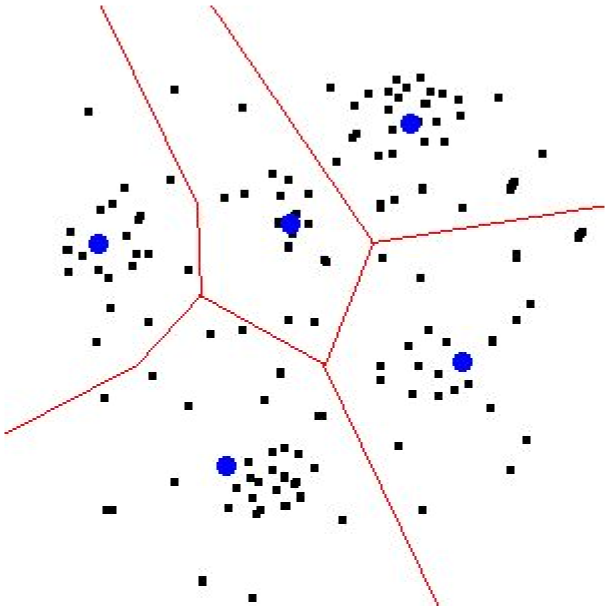
- Created clusters using desirability scores of training data using KMeans clustering and to decide on the parameters.
- For each cluster train a predictor based on XGBoost using all the points that lie in that cluster. This predictor was based on all the actual features of the data not just desirability scores
- Predict the clusters of all data in the sample_submission file on Zillow using KMeans predictor
- For each data point, use the predictor of the predicted cluster to predict the log_error.

Why this Model is Best?

We think this model works best because our desirability score is based mostly on the features provided already in the data-set and clustering based on desirability, clusters the similar houses together. So, if we use the desirability of a house to find the houses that are similar and then use the regressor that is trained only based on similar houses is more likely to predict a better log-error as compared to a regressor that is trained using all the houses. The reason behind this is that the predictor used by zillow is most likely dealing the houses with similar desirability in a similar way.

How does it work:

K-means Clustering:



Given a set of observations (x_1, x_2, …, x_n), where each observation is a *d*-dimensional real vector, *k*-means clustering aims to partition the *n* observations into *k* (≤ *n*) sets S = {S_1, S_2, …, S_k} so as to minimize the within-cluster sum of squares. Formally, the objective is to find:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}\in S_i} \|\mathbf{x}-\boldsymbol{\mu}_i\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i$$

where μ_i is the mean of points in S_i. This is equivalent to minimizing the pairwise squared deviations of points in the same cluster:

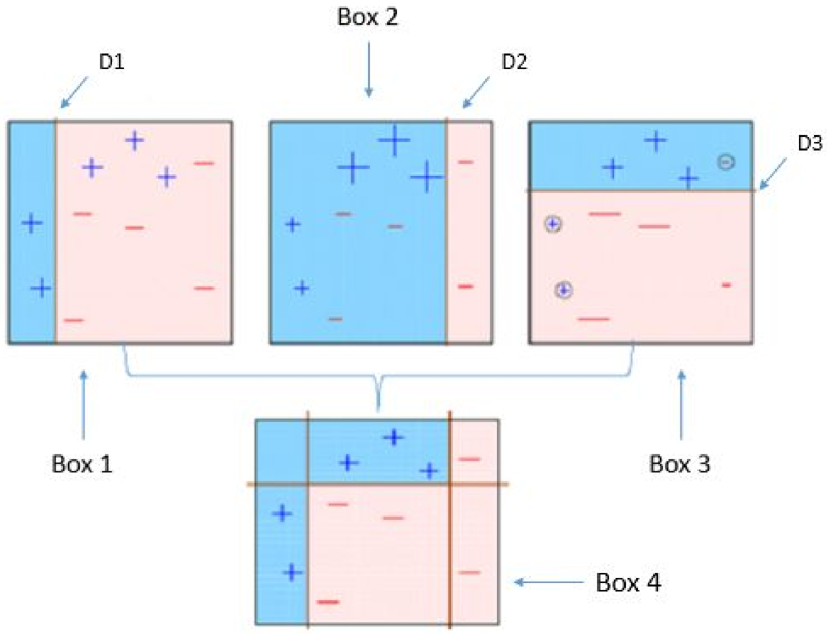
$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \frac{1}{2|S_i|} \sum_{\mathbf{x},\mathbf{y}\in S_i} \|\mathbf{x}-\mathbf{y}\|^2$$

The Equivalence can be deduced from identity

$$\sum_{\mathbf{x}\in S_i} \|\mathbf{x}-\boldsymbol{\mu}_i\|^2 = \sum_{\mathbf{x}\neq\mathbf{y}\in S_i} (\mathbf{x}-\boldsymbol{\mu}_i)(\boldsymbol{\mu}_i-\mathbf{y})$$

Because the total variance is constant, this is also equivalent to maximizing the squared deviations between points in *different* clusters.

Boosting:



In boosting we create weak learners in each step and finally combine all decision stump for the prediction. In each step we increase the weight of misclassified data points so that in next step the learner correctly classify them. Individually these weak learners are not that powerful but combining them we can get a strong prediction model.

Evaluation:

We have submitted our prediction in Kaggle. Result are given below:
MAE: 0.0648731
Rank: 2425
Our permutation test value is 0.0

Interesting Findings:

In our first assignment we ignored the geographical features. It is said that, three most important feature of real estate business are Location, Location and Location. After analyzing the geographical features and creating scores using zip code clustering we see that certain zip codes desirability changes over time. And this change has some good effect in the Zestimate/Logerr as well. We incorporated these values and make cluster. Instead of a global prediction clustered prediction gives us a better prediction. We have a strong feeling that if we refine our desirability score a bit more and get more fine grain clustering and spend some more time exploring the best parameters for the tools we used in this assignment, it's likely that we can get a score in top 10 or at least top 100, but I think that's out of the scope of this assignment considering the time allotted.

Yes Professor, You were Right!

If we don't use clustering based on desirability and use only one XGBoost regressor to predict logerrors for all the properties accuracy of prediction is less, which means since clustering based on desirability makes us more familiar with the data. It in a sense, enables us to guess logerrors for rest of data in a more educated manner.

Acknowledgement: Wikipedia, analyticsvidhya.com and python library documentation.