Prediction of King County House Prices

Xiyuan Sun, Dinesh Poddaturi, Zhaoxin Lin

**Introduction**

The ISU Stat502 2018 RMSLE prediction competition involves predicting house prices based on house characteristics and dates of sale.

For the purpose of this competition, the original dataset has been broken at random into a training set of size 10000 and a test set of size 11613. A randomly selected 15% of the test set is used to provide public scores and rankings on the leaderboard.

We used some statistical machine learning(ML) techniques discussed in stat502 course to make the predictions. Public Kaggle score is used to assess our predictors’ performance.

In this report, we described our attempts to select important features, implement different ML algorithms, and stack predictors in an effective way. We ended up with some reflection on what we learned through this competition.

**Feature Engineering**

The whole dataset included 10000 training cases and 11613 test cases. All the twenty covariates are numeric.

According to our exploratory analysis, …

We considered ways to make the variables contribute more to the prediction by changing their type, creating new features. Recently\_built, recently\_renovated…

Appendix:

> train %>% glimpse()

Observations: 10,000

Variables: 21

$ property <dbl>

$ date <int>

$ price <dbl>

$ bedrooms <int>

$ bathrooms <dbl>

$ sqft\_living <int>

$ sqft\_lot <int>

$ floors <dbl>

$ waterfront <int>

$ view <int>

$ condition <int>

$ grade <int>

$ sqft\_above <int>

$ sqft\_basement <int>

$ yr\_built <int>

$ yr\_renovated <int>

$ zipcode <int>

$ lat <dbl>

$ long <dbl>

$ sqft\_living15 <int>

$ sqft\_lot15 <int>

> test %>% glimpse()

Observations: 11,613

Variables: 21

$ id <int>

$ property <dbl>

$ date <int>

$ bedrooms <int>

$ bathrooms <dbl>

$ sqft\_living <int>

$ sqft\_lot <int>

$ floors <dbl>

$ waterfront <int>

$ view <int>

$ condition <int>

$ grade <int>

$ sqft\_above <int>

$ sqft\_basement <int>

$ yr\_built <int>

$ yr\_renovated <int>

$ zipcode <int>

$ lat <dbl>

$ long <dbl>

$ sqft\_living15 <int>

$ sqft\_lot15 <int>