**Stat 502 Class Project Weekly Updates**

**Competition:** Kaggle 502 House Prices

**Team Kaggle Name (if Relevant):** tom\_and\_jerry

**Team Members (Non-502ers in Parentheses):**

Xiyuan Sun, Zhaoxin Lin, Dinesh Poddaturi

**Week 0 (March 19-23) Summary of Activity and Progress:**

We understood the dataset. Started doing data engineering part and shared our ideas of how to approach the problem. We also finished setting up a github repository for the project and started using github application on our machines to push and pull our work.

**Week 1 (March 26-30) Summary of Activity and Progress**

This week we used techniques to find the best predictors that explains the price of the house. We used correlation matrix to find good predictors. We changed the datatype of some variables to better fit the model. In our first attempt we tried GBM to train and predict the house prices.

**Week2 (April 2-April 6)**

We tried knn(9 175890.4 0.7960958 90762.21), linear regression and Enet( 1.0e-04 15000 48000.35 0.9880368 28569.60). Knn has lower RMSE and scores.

Given that sqft\_living = sqft\_above + sqft\_basement, we dropped 3 variables from the trainning and test data (sqft\_above,sqft\_basement, view), but added two additional variables (recently\_built (yr\_built>2000), recently\_renovated(yr\_renovated>1990)). Our training dataset ends up with 18 variables: price, bedrooms, bathrooms, sqft\_living, sqft\_lot, floors, waterfront, condition, grade, yr\_built, yr\_renovated, zipcode, lat, long, sqft\_living15, sqft\_lot15, rec\_built, rec\_renovated. It gives best RMSE using knn with ( 5 173149.0 0.7962411 89539.54).

Then we tried xgboost with RMSE=128900.9 and the final tuned parameters are: nrounds=2000, max\_depth=4, eta=0.05, gamma=0, colsample\_bytree=0.5, min\_child\_weight=1, subsample=1.

Week 3

We summarized all our tuning parameters from each individual method. Then select the top methods and stack them.

Feature matrix used: (by Dinesh’s work)

> house\_train%>%glimpse()

Observations: 10,000

Variables: 13

|  |  |
| --- | --- |
| Variables | Type |
| housePrice | dbl |
| bedrooms | int |
| bathrooms | dbl |
| Sqft\_living | int |
| Sqft\_lot | int |
| floors | dbl |
| waterfront | fct |
| view | int |
| condition | fct |
| grade | int |
| Yr\_built | int |
| Yr\_renovated | int |
| zipcode | fct |

> house\_test %>% glimpse()

Observations: 11,613

Variables: 12

|  |  |
| --- | --- |
| Variables | Type |
| bedrooms | int |
| bathrooms | dbl |
| Sqft\_living | int |
| Sqft\_lot | int |
| floors | dbl |
| waterfront | fct |
| view | int |
| condition | fct |
| grade | int |
| Yr\_built | int |
| Yr\_renovated | int |
| zipcode | fct |

|  |  |  |
| --- | --- | --- |
| Method | Parameters | RMSE |
| tree | Cp=0.0001 | 183945.6 |
| gbm | Shrinkage=0.05, interaction.depth = 5, n.minobsinnode = 7, n.trees = 160 | 16841.08 |
| nnet | Size = 10, decay=0.38 | 280029.6 |
| lm | Intercept=TRUE | 169968.5 |
| knn | K=4 | 193332.3 |
| xgboost | Eta=0.3, max\_depth=3, gamma=0, colsample\_bytree=0.8, min\_child\_weight=1, subsample=0.75, nrounds=150 | 154796.7 |
| Random forest | Mtry=26 | 161072.8 |
| Elastic Net | Alpha= 0.07142857, lambda=325 | 170114.9 |
| PCR | Ncomp=49 | 211394.4 |
| PLS | Ncomp=41 | 170069.3 |

Feature matrix used: (by Zhaoxin’s work)

> retrain %>%glimpse()

Observations: 10,000

Variables: 18

|  |  |
| --- | --- |
| Variables | Type |
| price | dbl |
| bedrooms | int |
| bathrooms | dbl |
| Sqft\_living | int |
| Sqft\_lot | int |
| floors | dbl |
| waterfront | int |
| condition | dbl |
| grade | dbl |
| Yr\_built | int |
| Yr\_renovated | int |
| zipcode | int |
| lat | dbl |
| long | dbl |
| Sqft\_living15 | int |
| Sqft\_lat15 | int |
| indirenovated | dbl |
| Recent\_built | dbl |

|  |  |  |
| --- | --- | --- |
| Method | Parameters | RMSE |
| knn | K=5 | 173157.5 |
| Elastic net | alpha = 7.5e-05, lambda = 15000 | 209482.7 |
| xgb | Nrounds=2000, max\_depth=4, eta=0.05, gamma=0, colsample\_bytree = 0.5, min\_child\_weight=1, subsample=1 | 128900.9 |

Feature used: (by Xiyuan’s work)

Variables: 18

|  |  |
| --- | --- |
| Variables | Type |
| price | dbl |
| bedrooms | int |
| bathrooms | dbl |
| Sqft\_living | int |
| Sqft\_lot | int |
| floors | dbl |
| waterfront | int |
| condition | int |
| grade | int |
| Yr\_built | int |
| Yr\_renovated | int |
| zipcode | fct |
| lat | dbl |
| long | dbl |
| Sqft\_living15 | int |
| Sqft\_lat15 | int |
| Rec\_built | dbl |
| Rec\_renovated | dbl |

|  |  |  |
| --- | --- | --- |
| Method | Parameters | RMSE |
| knn | K=4 | 189015.2 |
| lm | Intercept = true | 172687.4 |
| PCR | Ncomp=14 | 210618.2 |
| PLS | Ncomp=14 | 179715.3 |
| tree | Cp=0.001 | 192556.0 |
| Random forest | Mtry=9 | 141561.0 |