RANDOM FOREST NOTES

* Read the introduction and take notes of the factors left unused and the reason why
* The Table of Contents lists the steps taken to build the model
* Use a 3:1 train:test ratio
* The features used are 'bedrooms', 'bathrooms', 'sqft\_lot', 'waterfront', 'view', 'condition', 'grade', 'sqft\_above', 'sqft\_basement', 'yr\_built'
* r^2: Coefficient of determination or the amount of variance in the response variable (price) that can be explained by variance in one or multiple independent variable(s)
* Results in “Test & Measure Model” section:

**Train: Random Forest r^2 Accounted: 0.9634**

**Comment:** Total r^2 is close to 1, so when using that many factors, the

model is very accurate at predicting the price

**Test: Random Forest r^2 Accounted: 0.7259**

**Comment:** On test data (that it has not already seen), unsurprisingly, the model predicts the price less accurately. But a coefficient of determination of 0.7 is moderately high. For that coefficient, the model is useable in real industry, especially if more features were to be input for the coefficient to be even higher.

Graph:

**Comment:** The grade or construction quality of the home is the most significant influencer of the price, accounting for almost half of the variance. In contrast, the size of the property accounts for less than a measly 10%. It appears for this region, ignoring potential biases in how the dataset was collected, quality matters more than quantity. The least significant feature is the number of bedrooms, which mattered less than the number of bathrooms.

Suggestion for Improvement: Add more features to the model. As this dataset covers only a single year, the influence of the economy and market are unlikely to have affected the price (response variable).

Team Members

Alexandre Makhmudyantsev

Brandon Nguyen

Ryan Martin Elli Cervantes

Triet Lieu

Predicting Sale Price of Homes in King County, WA

Abstract: Buyers want to purchase a home for lower than its market value and sellers want to sell their homes for higher than its market value. By knowing what a home’s baseline price within a range should be based on its attributes, each negotiator can better balance between offering the best deal for themself and a deal the other negotiator considers fair and would accept. Real estate transaction companies like Redfin or Zillow offer assessment tools to home prospectors so when prospectors complete transactions through Redfin or Zillow, these companies collect fees for their consultant and recommendation services. Thus, the better real estate transaction companies can assess home prices, the more home prospector use their services, the more transactions they facilitate, and the greater their income from fees.

In this project, we build a multilinear regression model and a random forest model to predict the price of homes in King County between 2014 and 2015 based on 15 factors. We split the dataset into training and testing sets in an 80-20 ratio. We calculate a squared-difference prediction error, normalized to target price. We sum all prediction costs for a model to get its total error and contrast the performances of our multilinear regression model vs random forest model. We hypothesize causes for their differencing performances and identify the most predictive factors for sell price.

Evaluation: Calculate each prediction error as the squared difference of the predicted and target prices divided by the target price. Dividing each squared difference by the target price normalizes and scales the error and makes comparing prediction errors possible. Sum all prediction errors for a model to get its total error. The lower its total error, the more accurate the model.

Dataset: kc\_house\_data.csv

21614 [rows/homes] x 10 [columns/attributes]

|  |  |  |
| --- | --- | --- |
| Col Factor | Comment | Data Type |
|  |  |  |
| id | Use to index | INT64 |
| date | Unused factor | ISO Datetime |
| price | Target value | INT |
| bedrooms | # bedrooms (integer) | INT |
| bathooms | # bathrooms (fraction: .25) | FLOAT |
| sqft\_living | Size of home | INT |
| sqft\_lot | Size of parking lot | INT |
| floors | # of floors | INT |
| waterfront | True or False: Has view | BOOL |
| View | # of times home was viewed | INT |

References:

Kaggle project and dataset

<https://www.kaggle.com/datasets/harlfoxem/housesalesprediction>

Linear Regression Tutorial in Scikit-Learn

<https://www.geeksforgeeks.org/multiple-linear-regression-with-scikit-learn/>

<https://tung-dn.github.io/prog_ML_reg1.html>

<https://codeburst.io/multiple-linear-regression-sklearn-and-statsmodels-798750747755>

Random Forest Tutorial in Scikit-Learn

<https://www.datacamp.com/tutorial/random-forests-classifier-python>

<https://scikit-learn.org/stable/modules/ensemble.html>