**RAW INPUT DATA:**

For applying our statistical learning techniques, we have 21 features in our data which make up each data point in the dataset. All the data are in a csv file. The features are as follows

***Property***: Unique identity for each house sold in King County

***Date***: the date house was sold

***Price***: Price of each house sold

***Bedrooms***: Number of bedrooms in the house

***Bathrooms***: Number of bathrooms in the house. Here in the data .5 means a bathroom with only toilet but no shower

***Sqft\_living***: Square footage of the interior house living

***Sqft\_lot***: Square footage of the lot

***Floors***: Number of floors in the house

***Waterfront***: An indicator variable whether the house is overlooking waterfront or not

***View***: Whether the house has been viewed

***Condition***: An overall condition rating of the house

***Grade***: An overall rating given to the house according to the King County grading system

***Sqft\_above***: The square footage of the interior house spacing above ground level

***Sqft\_basement***: The square footage of the interior of the house spacing below ground level

***Yr\_built***: The year the house was built initially

***Yr\_renovated***: The year the house was last renovated (if renovated)

***Zipcode***: The Zip code area the house is located

***Latitude***: The latitude of the house

***Longitude***: The longitude of the house

***Sqft\_living15***: Square footage of the interior house living in 2015

***Sqft\_lot15***: Square footage of the lot in 2015

*Example data point:*

[1999700045, 20140502, 3, 1.5, 1340, 7912, 1.5, 0, 0, 3, 7, 1340, 0, 1955, 0, 98133, 47.7658, -122.339, 1480, 7940]

**EXPLORATORY DATA ANALYSIS:**

First, we do simple techniques to check which features are highly correlated to the price and correlated to the other features in the data by plotting a correlation matrix.

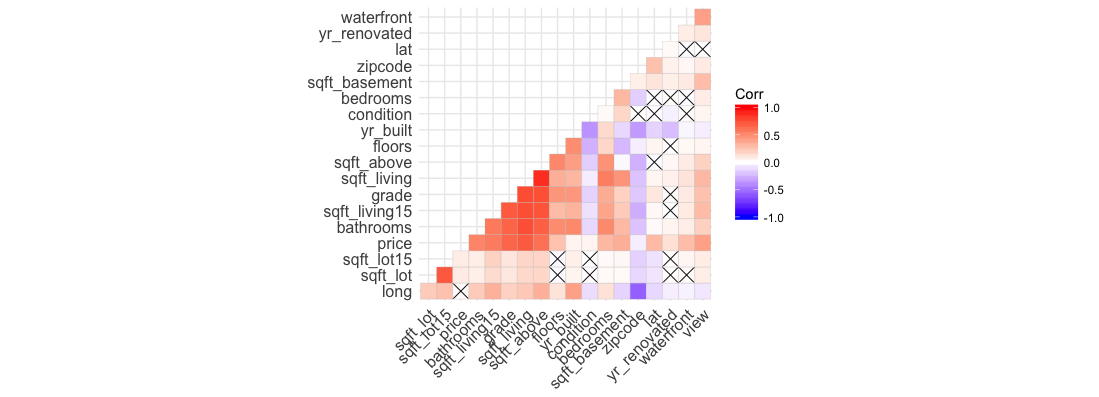


Figure 1: Correlation plot of the house price and features

By looking at the plot above it’s clear that *Sqft\_living* is highly correlated to the price of the house and *zipcode* is least correlated to the price. In our analysis we will keep all the features since the houses in certain area (*zipcode*) might different from other areas.

By initial glance of the data we found out that some of the features in the data are not helpful. For example, we removed property, and date features from the data. The feature, ‘*property*’ doesn’t add anything to the learning process since it’s the unique identity of the house. All the houses are sold in a period of two years. So, keeping the feature ‘*date*’ doesn’t add anything important to our learning process. In the raw data all the features are integer/double/float datatypes. We changed the datatype of some of the features to factor datatype. ‘*waterfront*’ is an integer variable (0 or 1), so we changed it to factor datatype. We changed the data type of the feature ‘*condition*’ to factor. Finally, we changed the data type of ‘*zipcode*’ to factor since it is a location.

In our analysis we introduced two additional features called ‘*recently\_built*’ and ‘*recently\_renovated*’. The feature ‘*recently\_built*’ is an indicator variable with 1 being the house constructed after the year 2000 and 0 being the house constructed before 2000. The feature ‘*recently\_renovated*’ is also an indicator variable with 1 being the house renovated after 1990 and 0 being the house renovated before 1990.

**METHODS USED:**

We used different methods to train our data. In each method the data are trained using 10-fold repeated cross validation with 100 repetitions. We used Root Mean Squared Error to quantify our error, because it is more representative of how much our prediction is incorrect by. We trained our data using each method individually noting, Root Mean Squared Error and corresponding best tune values. Finally, we did stacking using different combination of our methods. In each iteration of our stacking we trained that data over reasonable tuning parameters and used it to get our predictions. And then used those predictions to stack (Random Forest/OLS) and tuned the model to get our final prediction model.

All the models are trained using *train()* function in *caret* package in R over a range of tuning parameters.

*Random Forest*:

We trained the data using Random forest method and found our cross-validation estimate of error based on “out-of-bag” samples with a range of 1 to 100 number of dimensions considered at each split and with ntree=500.

The resulting best tune and the RMSE values are mtry=47 and 163809.1 respectively.

*K- Nearest Neighbor*:

The data are trained using knn with nearest neighbors ranging from 1 to 20. Our best tune for this model is k=4 with the RMSE 192743.7.

*Elastic Net*:

*GBM*:

*PCR*:

*PLS*:

*XGBOOST*:

*TREE*:

**LESSONS LEARNED**:

The RMSE values of the individual methods are very high compared to the stacking RMSE value. We tried different combinations to stack the predictions using Random Forest method.

**CONCLUSIONS**:

It’s very difficult to predict the exact house prices for this data. There are many factors to be considered apart from the given features. Some of them could include information about the potential buyers, climatic condition of the location and economic stability of the potential buyer. Adding to that, the value of the house can be heavily influenced by the neighborhood and features of the house, which are extremely important. The artistic quality of the house, architecture and the style of the house (like French, Colonial, Victorian, Contemporary etcetera) can appeal to one buyer but not the another. However, with all the methods we tried, we came up with an ensembling method that beat all other methods and also gave us the best score in the leaderboard.

Since the data comes from specific location, in this case King County we can only generalize the predictions to this countyor some other similar/identical counties in the state. Working on this project we developed a deeper understanding of the housing market and the features that are highly correlated with the price of the house, which could help us in developing a more general framework for predicting housing prices in different locations. Some of the significant features (like *bedrooms*, *bathrooms*, *sqft\_living* etcetera) can be used in predicting different datasets. By applying different statistical learning methods on this data, we expanded our knowledge on how the methods work and how we might use them our future projects.

**REFERENCES**:

1. Kaggle: ISU Stat 502 2018 RMSLE Prediction (Price prediction based on characteristics of homes in King County, Washington, USA.) <https://www.kaggle.com/c/isu-stat-502-2018-a>
2. Course Material – Statistics 502: Applied Modern Multivariate Statistical Learning, Spring 2018 by Dr. Stephen Vardeman.