Cribs

Regression modeling for house sales in King County

Jun Tae Son, Omer Saif Cheema, Yusheng Zhu

Introduction:

The dataset we used in the analysis had historical data on house sales in King County, Washington from May 2014 to May 2015. The goal of the analysis is to generate the best model for predicting home sale price in King county. We included the following features in the analysis; date of house sales, location of the property, the number of bedrooms and bathrooms, the size of the property, floors, waterfront, view index, conditions of house, levels of construction and design, renovation history and basement. We analyzed how influential these features are on house price and conducted prediction test using our final model. We assumed that the prices of the houses are only affected by the variables used in the data set and factors like interest rates, property taxes other political and economic factors stay constant.

Methodology:

Instead of using all 21, 613 rows of data, each team member selected a random sample of 1500 records from the whole dataset. Then, we preprocessed the data by selecting the relevant variables; creating dummy variables; transforming the dependent and some independent variables; and finally we created interactive variables that seemed to have a joint effect on price. Later we cleaned the random sample by taking out outliers and influential points because outliers and influential points negatively affect our regression model and reduce the model's prediction accuracy. Furthermore, we split our sample into a training and testing set with 75/25 split ratio. The training set is used to create the regression model and testing set is used to test the model's prediction performance. Afterward, we conducted different model selection methods to determine the best model for each member and finally, tested performance of the final model using the testing set. After predicting the values for house prices for test set by using cross validation and comparing the prediction parameters such as Adj-R-square, and RMSE, we chose the model which had the best performance as our final model. Adj-R-square shows the percentage of variance in house prices explained by our model. RMSE gives the measurement of error in our model

Conclusion:

All three of us reached the same conclusion that house prices in King County are most influenced by interior area of a house and the house location. We hypothesized in our technical report that location will be one of the most influential predictors of house prices and it is understandable why interior area of a house is an important predictor. Based on the analysis, we concluded that 75.36% of variability in house price in King County can be explained by our final model. Goodness of fit test and cross-validated R-square results also supported that our final model is well fitted in the dataset and can be used for prediction