Predicting Home Prices in King County

How different methods of analysis yield different results

Alexander Bell-Towne, Jack Fox, Sean Ker, Spencer Pease

March 8, 2017

Quantitative Question: What are efficient predictors for home prices?

Abstract

A 150 word summary of your project and findings. This work develops a model based on a dataset found on Kaggle for House Sales in King County, USA in order predict house prices from a set of covariates. We created and compared the predictability of our linear models and decision trees by choosing a specific number of attributes that had a higher R^2 value. The best model that we were able to create had an RMSE value of 153.35. In our interactive resource, you will be able to compare and contrast these models, to evaluate as we did, the predictability of house prices given certain attributes.

Introduction

The problem we're trying to solve is using historical data of house prices and attributes of houses in King County, US area to determine the features (covariates) that predict prices of houses. By identifying these covariates, we wanted to explore applications of our model in predicting house prices in real-world settings that buyers or sellers of real estate can leverage. A quantitative question we hoped to answer with our model is "which attribute(s) of a home directly correlates with the price of that home?" With this answer, sellers can direct their attention on providing the most value through those top attributes which correlate with price, in order to maximize profit. Buyers can also use this information to determine the type of house given a select amount of attributes that they can afford.

Related Work

A description of previous papers or projects related to your project.

We found our dataset on Kaggle, which is a platform that engages a community of budding and experienced data scientists to compete in creating the best model that fits the dataset, usually with the intention to predict an outcome. As a result, there were many "kernels" (preloaded code uploaded by an individual in the community in a format similar to a Python Notebook or RMarkdown) that tried to accomplish the same goal as our team. A list of the kernels for our dataset submitted by the community are listed here.

A related paper was published by as a dissertation at NYU called Predicting the Market Value of Single-Family Residential Real Estate, where the authors developed linear models for predicting house prices based on data from 2003 to 2009 in Los Angeles County.

Methods

Since our mission was to identify the features (e.g number of bathrooms or square footage of the property) of a home that would efficiently predict its sale's price, we first needed to discover which features had the most positive impact on the price. Then, using these identified features, we could create a model that would be used to predict the sale price of a home. This approach required us to implement both linear models and machine learning tools.

But before applying any statistical methods, we randomly split the data into three portions: the training set (70% of the total data), the validation set (10% of the total data) and the testing set (20% of the total data). Each subset of the data was then transformed into a data frame where we created custom metrics (such as price in thousands) and ensured that factored features had the same levels across all three sets.

Using the training data set, we created linear models for each home feature to learn which of these attributes had the largest impact on the price of the home. From each model we collected and compared each R squared value. Of the twenty-five features included in the data set, we arbitrarily selected features whose R squared value was greater than or equal to ten percent. As such, these curated list of features, twelve in all, included R squared values that ranged from grade (51%) to bedrooms (10%).

Next, we ran multiple decision trees and used the root mean squared error (RMSE) to validate the accuracy of each tree. Our first tree included all the features of the data set which produced a RMSE of 153.35 (measured in price in thousands). Our next tree included features that had a R squared value greater than twelve percent which had an accuracy of 159.63. Next we split this group of features into two categories, features with a R squared value greater than twenty-five percent and those with a R squared value less than twenty-five percent. The former group resulted in a RMSE of 174.18 where the latter produced a RMSE of 241.81.

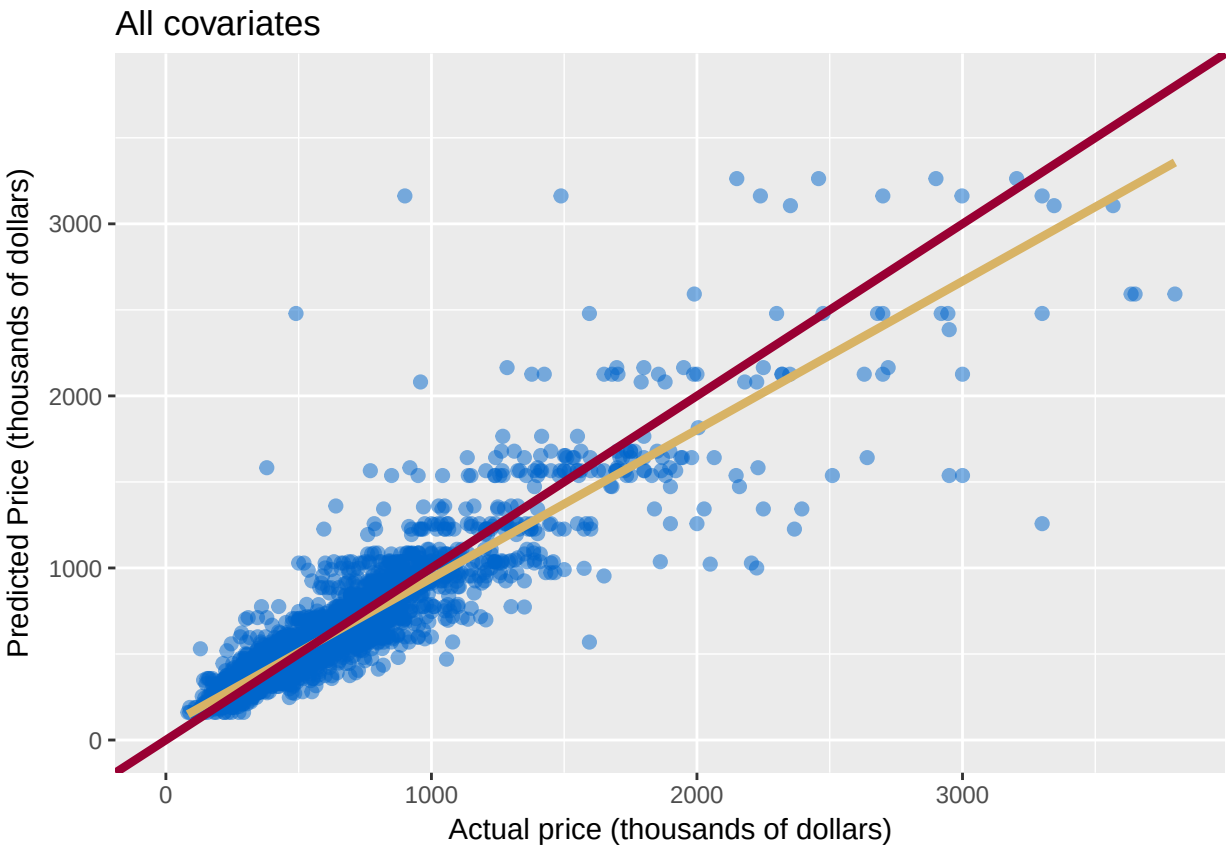
Finally, for each tree we plotted the observed values against the predicted values to visualize the error rates.

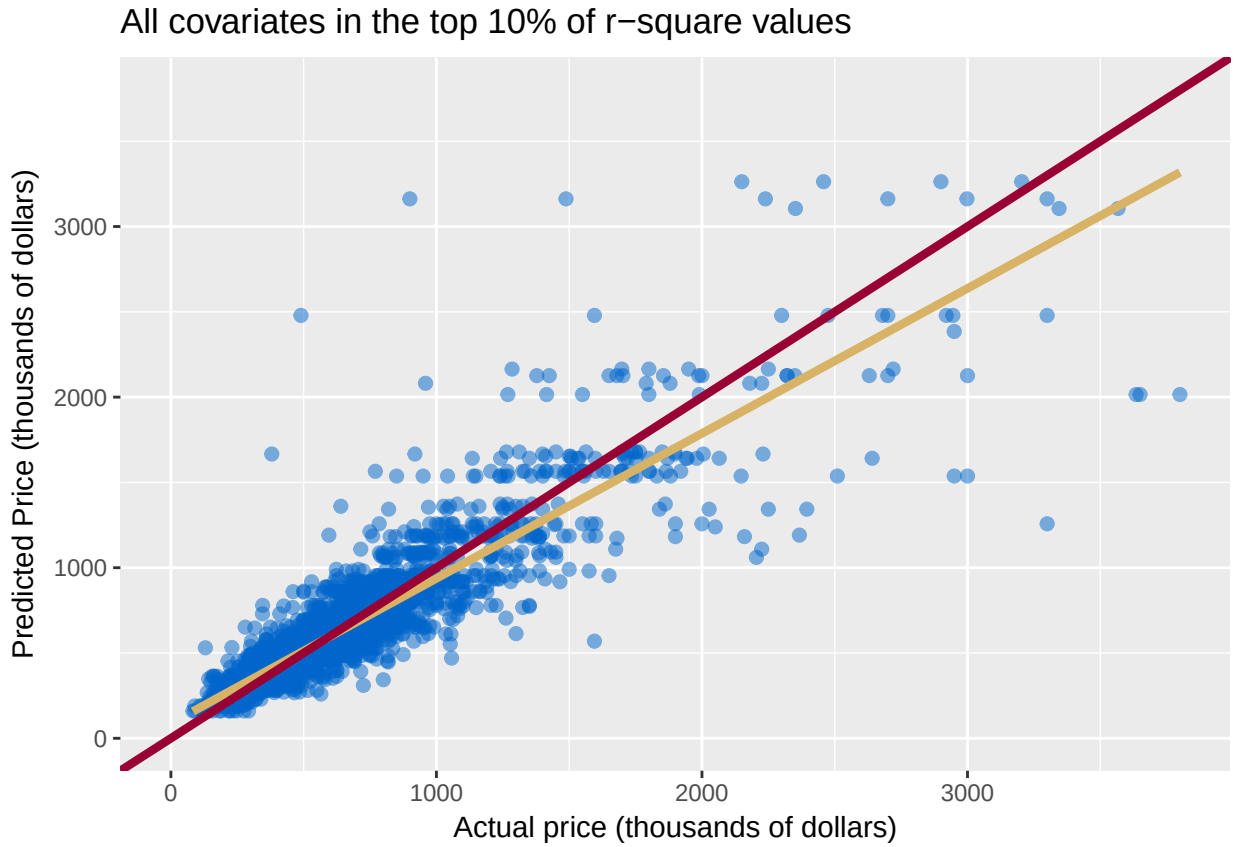
Results

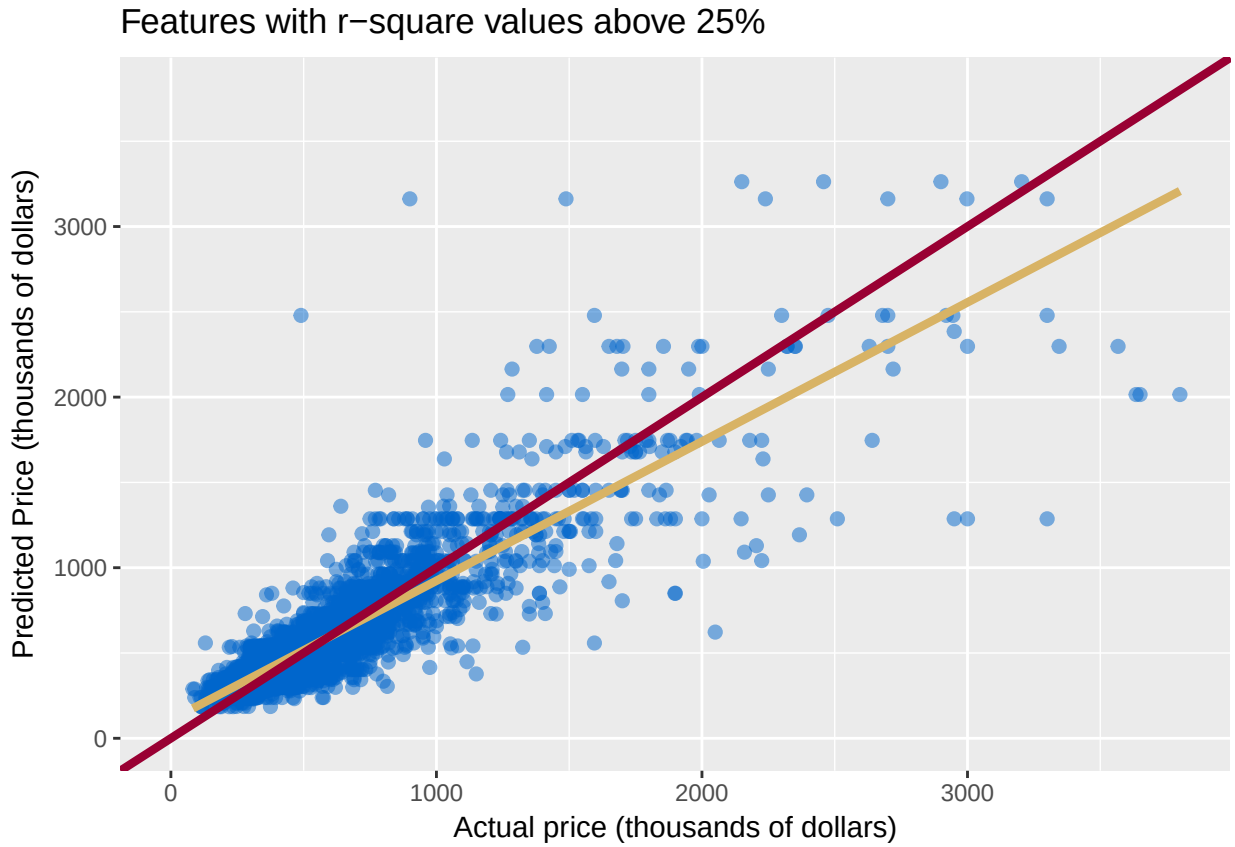
One major insight we drew from our research is that by using all the attributes in our dataset, we were able to obtain the lowest RMSE value and a linear model that had the best fit for our test data. In addition, when comparing the the group of covariates with R^2 value of over 25% and those that were below 25%, we discovered that the former set had a better fitting linear model.

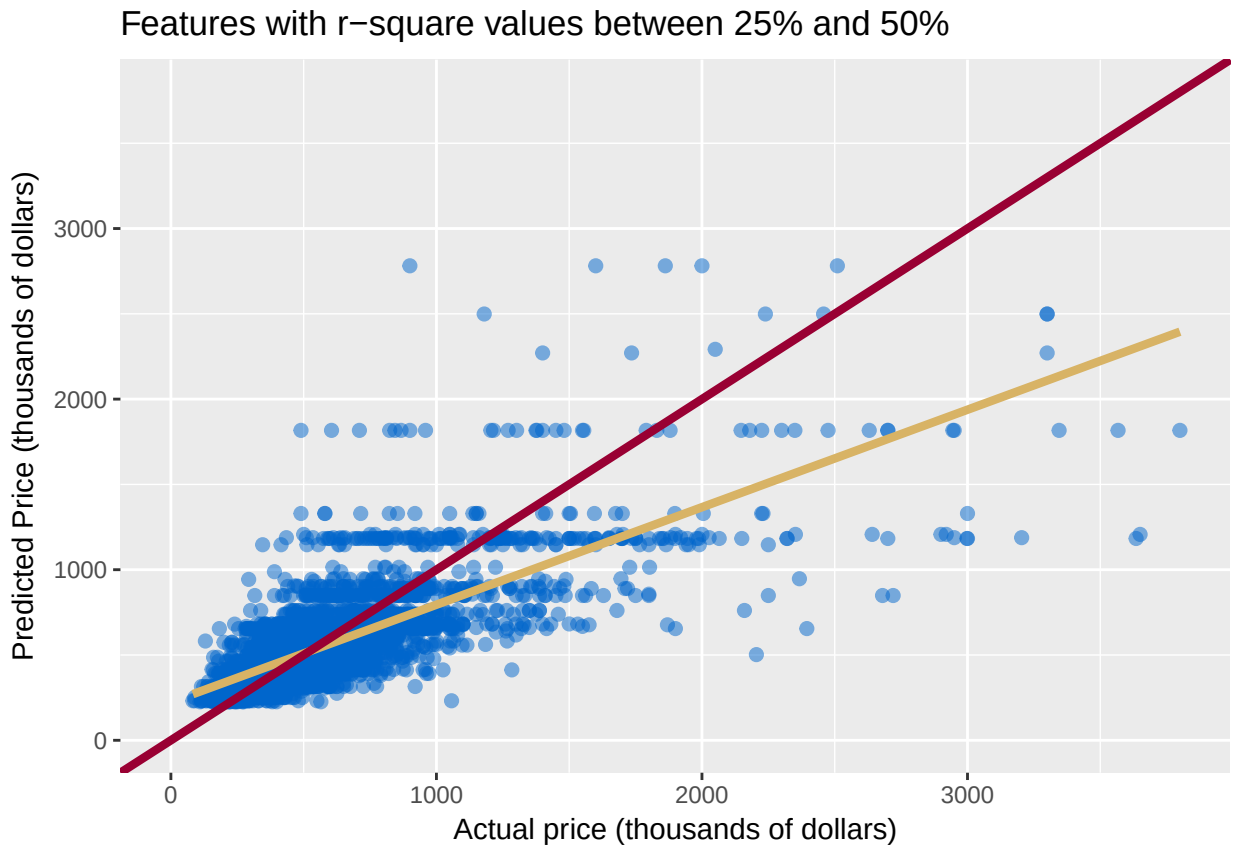
Models:

(Note: The gold line shows the best fit line of the data, and the red line shows the "ideal fit line" (ie, perfect predictions).)


All covariates
Predicted Price (thousands of dollars)
0
1000
2000
3000
0
1000
2000
3000
Actual price (thousands of dollars)


All covariates in the top 10% of r−square values
Predicted Price (thousands of dollars)
3000
2000
1000
0
0
1000
2000
3000
Actual price (thousands of dollars)


Features with r−square values above 25%
Predicted Price (thousands of dollars)
3000
2000
1000
0
0
1000
2000
3000
Actual price (thousands of dollars)



Discussion

As a result of our findings, we are able to report with significant evidence that although the group of covariates had an R^2 value of less than 25%, using these covariates in our final model reduced our RMSE value and produced an overall best fitting model. One important difference in our method that we could make to improve the predictability of our model is to evaluate the outcome of removing the outliers. Perhaps our model could have a better fit for the data if we had done this.

Future Work

We had started our project with the intention of exploring possible applications of our model in real-world settings. Our system could be further developed for the use of individuals in the real estate industry. We are envisioning a system that allows real estate buyers to be able to modularly select desired features of a house and determine the price given their selected conditions. The only way that this system is possible is if the pricing is accurate and we are confident that our predictive model, based on historical data of prices given a set of features, will enable this future.