Analysis and Prediction of House Sales in King County, USA

Sayoni Chatterjee, Shubham Puri, and Shikha Singh

August 15, 2018

Abstract

A precise prediction of the house price is vital for all the real estate participants, especially the seller and the buyer. Housing costs have a substantial impact on individuals, families, businesses and governments. It has been one of the major expenses for most of the people and one of the greatest investments of the life time! Yet, no outline that states evidently the factors to consider while buying a house was made till date and our conception of what drives the value of houses is restricted. In this paper, we use the data from Kaggle that is available for Analysis and Prediction of House Sales in King County, USA.

King County is the most populous county in Washington, and the 13th-most populous in the United States. This project uses 20 illustrative attributes against which the dependent variable price can be determined, and it also has 21,613 entries of housing sales in King County, USA. The important features have been extracted by Principal Component Analysis(PCA). A major proportion of our work also dealt with feature engineering and feature reduction for some models. We have then applied the models of Linear Regression and Support Vector Regression(SVR) with the reduced features to study which of the two models produce a lower mean square error(MSE). We have also applied Decision Trees, Random Forests and Gradient Boosting with all the variables to come up with a model that gives the best result and to interpret the outcomes and learn if they are in-line with the real-world scenarios.

1 Introduction

Predicting sale price of properties has always been an exciting problem. The real-estate market is on a roller-coaster ride and one can never single out any true measure for evaluating the property prices; no one will ever be able to list out the exact attributes that affect the sale prices of houses across the globe. The house prices are not only a concern for the players of the real estate market, but they also indicate the current economic situations. Hence, sale price prediction is a vast phenomenon that affects the economy, the government as well as an individual.

Real estate value prediction is a multi-dimensional problem and it involves estimating many facets of a property like it's neighborhood, the year it was built on, the structure of the property and the extra amenities

that come along with it. In the past, the prediction of value did not involve a lot of systematic discerning and would mostly account for major gambling or a decent experience in the real estate field. During the past decades, all valuations have been mostly based on surveys that are were carried out and the existing trends in the data were deduced to provide a more rational and concrete base to the value prediction structure. We are seeing the advent of machine learning and data analytic which are serving us even better to read the historical data that is present with us to make future predictions. Markets like Trulia,Zillow and Hotpads have emerged which estimate the property value of past listings very precisely. Some of the attributes that we are considering for our study are the condition of the house, the grade of the house, the infrastructure of the property etc. Pleasant neighborhoods, renovation, presence of a waterfront only add to the value of the property. Some characteristics like the number of bedrooms/floors, presence of parking lot etc add to the vitality of the property and are maintain a positive linearity with the sale price.

In our research, we study the main attributes that are affecting the property values. We are analyzing 20 attributes and 21613 entries in total. We start by inspecting the attributes and if there exists any skewness in the data. We perform some feature engineering such that the outcomes of our study are more insightful to us. We also try to implement some models and methods for predicting the prices of the houses. Linear Regression and Support Vector Regression have been implemented by constructing a model with the four most important attributes square foot, grade, living room area in 2015(assuming some renovation was done) and number of bathrooms that have been detected by principal component analysis(PCA). Later, we use all the variables to construct models on decision tree, random forest and gradient boosting and discover the top-most variables on which these models depend. In the first approach where we compute the important variables by PCA and implement the linear regression and support vector regression, we can clearly infer that support vector regression is a better model for price prediction as the mean square error (MSE) on the test model is almost 3 times lower than from the linear regression. In the second approach, random forest and gradient boosting both generated MSE that was quite less than that of the decision tree. Interesting thing to notice in our study was that the attribute grade remained constantly in the top two positions in the list of most important variables. One cause of this result can be that grade is a measure that summarizes the overall characteristics of the house. By this we mean that, a house with a good neighborhood, the apt structure and condition will always be priced more. We can offer this attribute a higher weight than the others.

The rest of the paper is organized as follows.

2 Data

We begin by understanding the data set and learning the importance of each and every variable in our dataset. The description of the 21 variables is given below.

2.1 Dataset Description

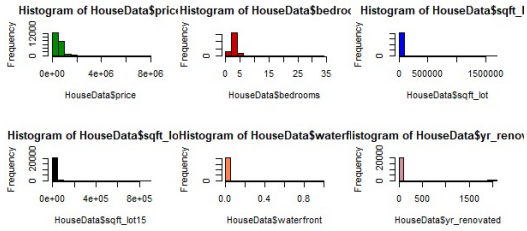
ID	Name	Description
1	id	unique numeric number assigned to each house being sold
2	date	date on which the house was sold out
3	price	price of house which we have to predict i.e. target variable
4	bedrooms	number of bedrooms in a house
5	bathrooms	number of bathrooms in a bedroom of a house
6	sqft_living	measurement of house in square foot
7	sqft_lot	square foot of the lot
8	floors	levels of house
9	waterfront	whether a house has a view to waterfront 0 is No 1 is Yes
10	view	whether a house has been viewed or not
11	condition	overall condition of a house on a scale of 1 to 5
12	grade	overall grade given to the housing unit,a scale of 1 to 11
13	sqft_above	square footage of house apart from basement
14	sqft_basement	square footage of the basement of the house
15	yr_built	date of building of the house
16	yr_renovated	year of renovation of house
17	zipcode	zipcode of the location of the house
18	lat	latitude of the location of the house
19	long	longitude of the location of the house
20	sqft_living15	Living room area in 2015(implies– some renovations)
21	sqft_lot15	lotSize area in 2015(implies– some renovations)

Table 1: Variable Description Table

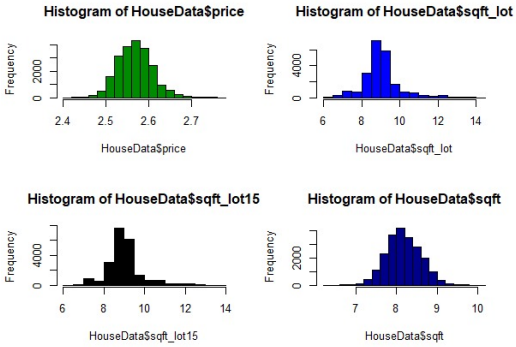
2.2 Analyzing the data for potential outliers and performing Feature Engineering

Summarizing the statistics of each column of the dataset, we see unusual behavior in price,bedrooms and sqftlot. Each of them has a maximum value which is distant from the mean value of the same. We also check the data set for any NA values. To check for skewness exhibited by the data, we apply the skewness function under the package "moments". Any value outside the range of (-2,2) will be considered an outlier. According to this method, the variables that demonstrated skewness were price,sqftlot, waterfront, yrrenovated, sqftlot15. We plotted the histograms for these variables to check if the skewness in the dataset.

We should not work with skewed data as it may result into incorrect predictions by our machine learning model. We need to perform feature transformation on the skewed variables to get the better data distribution. To remove the right skewness, we apply log transformation on the variables plotted above and get bell curved results as below.



(a) Before Transfromation



(b) After Transformations

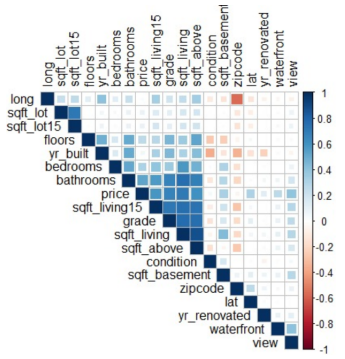
Figure 1: Histograms before and after Feature Engineering

2.3 Co relation between variables

We calculate the correlation between each pairs of attribute. From the corrplot, we see that sqftliving is correlated with sqftabove and grade.

For better results, we can either drop the variable, or apply transformation. We decided to combine the sqftabove and sqftliving into another variable sqft.

Figure 2: Predicted vs. Observed value



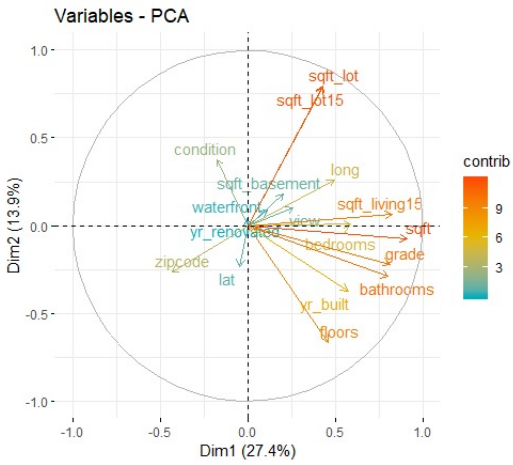
3 Our Approach

We tried implementing ek do cheeze. First we identified important components and then we tried to run different models such as aa bb an cc we also tried to analyse and compare the results of these different models.

3.1 Performing Principal Component Analysis

Principal components are the underlying structures in data. They are the directions where there is most variance.Hence, as our data set contains many variables, we simplify it by turning the original variables into smaller number of "principal components". When we add more principal components to our dataset, it increases the accuracy of the prediction. We performed a PCA on our dataset and obtained 17 principal components PC1-PC17. From the plot in ¡figure¿ we see that sqft,grade,sqftlot,sqftlot15, bathrooms and floors are the greatest contributors to the PC1.

Figure 3: Principal Component Analysis of Attributes



3.2 Linear Model Regression

Linear model is the first model that we are building on our processed data. It establishes a relationship between independent and dependent variables by fitting a best line which can be represented by :

$$y = b_0 + b_1 x$$

We first considered generating a regression model with all the attributes against price. The results received were quite interesting as all the variables except the sqftlot, were marked as significant in the model. The mean square error (MSE) established on the test data by applying our learned model was 26.2. The model seemed to be good, but there was an error in it as we see that the number of bedrooms had a negative slope value in the model. Ideally, this cannot be the case as number of bedrooms should have a positive relation with price. Next model that we created by considering the significant variables from the corrplot. Bedrooms again showed the same abnormality and the MSE increased to 35.73. As number of bedrooms is an important variable in determining the value of the property, we considered on building a model with bedrooms as the only attribute. Though the relationship between price and bedrooms was amended, the MSE was 49.31! We then chose to consider all the significant variables from the PCA, i.e sqft,grade,sqft_living15 and bathrooms and saw that the model considered all the variables as significant, and the MSE was 35.73. According to our analysis, this model evidenced to be the best out of the ones that were discussed before.

3.3 Support Vector Regression

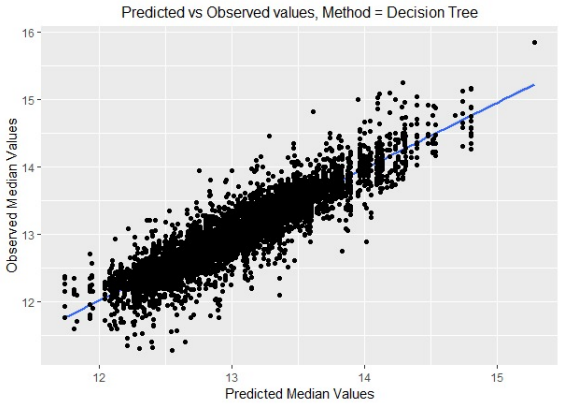
Support Vector Regression is a useful and flexible technique, helping users to deal with the limitations pertaining to distributional properties of underlying variables, geometry of the data and the common problem of model overfitting We implemented SVR on the training dataset, once by considering all the variables and then by considering only the significant variables from the PCA, i.e sqft,grade,sqft_living15 and bathrooms such that the results could be compared with those of LMR. While considering all the features in the model, an MSE of 13.02 was established and when considering only the prominent features of PCA, an MSE of 6.36 was established.

3.4 Decision Trees

To generate a decision tree, we used the Gini purity measure We also pre-pruned the tree: a node could not be split further if it would result in fewer than 10 instances in one of the child nodes. To improve our model, we used the best CP value which was 0.0001330312.

To calculate the MSE of the model on the test data, we used the model to predict the median value of the houses in the test data set. We then took the difference between the observed value and the predicted value, squared it, and took the average across all instances in the test data. The MSE on the testing data is 7.07. We plotted the predicted vs. observed values, which is given in Figure 4. We could also list out the top variables used for the splitting of the decision tree. They were Grade,Latitude,Longitude and sqft.

Figure 4: Predicted vs. Observed value

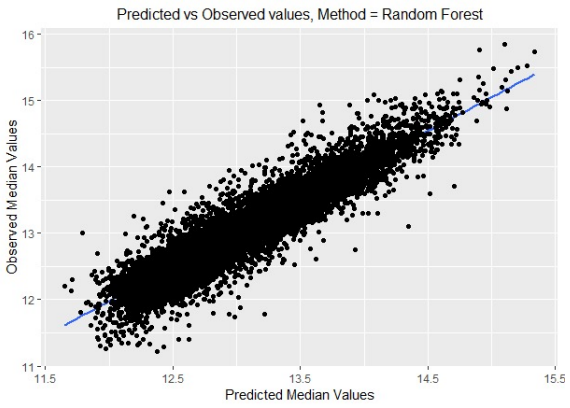


3.5 Random Forests

For the random forest model, we used the R function randomForest(), and set the model to produce 100 trees, and a minimum node size of 10 (for comparability with our decision tree model). To calculate the MSE, we pulled the predicted values from the predict function, and compared them to the observed values in the data. The MSE for this model is 3.29, and the plot of predicted versus observed values is shown in Figure 5.

By using varImpPlot() function, we concluded that latitude,grade,sqft and sqftliving_15 are the top variables used in this model.

Figure 5: Predicted vs. Observed value



3.6 Gradient Boosting Model

Gradient boosting is a machine learning technique for regression and classification problems. It produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. We generated a model of the same using all the attributes and established a MSE of 3.97. The 4 most important variables were Grade,sqft,sqftlot,bathrooms. Gradient Boosting trees build trees one at a time, where each new tree helps to correct errors made by previously trained tree. Although it may seem this algorithm is better than random forests, it is more prone to overfitting.

4 Experimental Results

We have implemented PCA as the first step after data processing and saw that the same set of attributes that displayed collinearity from the corrplot, associated with each other in this model as well. Each principal component sums up a certain percentage of the total variation in the data. The variables in the principal component 1 can be used to approximate the complexity of the dataset. Adding more principal components adds to the accuracy and hence we have accounted the least number of principal component for our analysis on linear regression and support vector regression.
Our prediction analysis on different models has been summarized below:

Model	Linear Regression	Support Vector Regression	Decision Tree	Random Forest	Gradient Boosting
MSE on Test Data	0.357	0.1281	0.0707	0.0329	0.0397
Top 4 variables					
1	Sqft	Sqft	Grade	Latitude	Grade
2	Grade	Grade	Latitude	Grade	Sqft
3	Sqft_living15	Sqft_living15	Longitude	Sqft	Latitude
4	Bathrooms	Bathrooms	Sqft	Sqft_living15	Bathrooms

Table 2: MSE values across different models

Our analysis shows that LMR(model with 4 variables) has a MSE of 37.5 whereas the SVR(model with 4 variables) has a MSE of 12.81. SVR is superior to LMR because it considers the non-linearity in the dataset. As per our study, there existed a non-linear relationship between price and some variables like sqftbasement and sqftlot.
Later, we considered all the attributes in our dataset and modelled decision tree, random forest and gradient boosting. We concluded that random forest has the least MSE on our test data and it was unsurprising. Both random forests and gradient boosting are improvement over the simple decision tree. The MSE and the top 4 variables can be seen in the table below. Random forest has the least MSE because it randomly selects the

variables used in tress construction and hence breaks any correlation amongst trees.
Moving on from the models and studying the variables, we saw that grade remained constantly in the top two positions in the list of most important variables. One cause of this result can be that grade is a measure that summarizes the overall characteristics of the house. By this we mean that, a house with a good neighborhood, apt structure and condition will always be priced more. The transformed variable sqft was also one of the important variables that was considered for price prediction. These variables can be considered more intently for further analysis on this subject.

5 Empirical Strategy

5.1 Data

5.1.1 Description

(Ruggles et al. 2018). I use the years 2005-2016, and consider all individuals aged 17-40. By keeping such a large range of individuals, I am able to test for heterogeneity in the responsiveness to Pell Grant eligibility by dependency status.[1] According to the US Dep't of Education (2004-2016b), a student is a dependent student if all of the following criteria are met:

1. The student will be less than 24 years old on January 1 of the academic year.
2. The student is single (never married) or divorced at the time they file the FAFSA.
3. The student is not enrolled in a graduate/professional degree program.
4. The student is not on active military duty and is not a US Armed Forces veteran.
5. The student has no dependents (either own children or other dependents, not including a spouse) who receive at least half their support from the student.
6. The student is not homeless, a ward of the state, or an emancipated minor.

Individuals who fail to meet at least one of these criteria are considered independent for financial aid purposes.

1. I chose the range 17-40, rather than 17-24 (the "traditional" college age), for a number of reasons. First, non-traditional students today make up a much larger proportion of the student population than they did in the past (Pascarella and Terenzini 1998; Denning 2017; Seftor and Turner 2002). These students also make up a large share of Pell Grant recipients: individuals between the ages of 25 and 40 made up between 30 and 35% of all Pell Grant recipients during the sample period (US Dep't of Education 2004-2016a). This age range is also less likely to have children who are college-aged than extending the maximum age to, say, 45 or 50, meaning that there will be fewer instances of both parents and children being Pell-eligible, since in such a case it is unlikely that the parent would be considering attending college at the same time as their child.

5.1.2 Limitations

6 Results

6.1 Main Results

7 Conclusion

References

Denning, Jeffrey T. 2017. "Born Under a Lucky Star: Financial Aid, College Completion, Labor Supply, and Credit Constraints." W.E. Upjohn Institute for Employment Research, *Upjohn Institute Working Paper* (Kalamazoo, MI), no. 17-267.

Pascarella, Ernest T., and Patrick T. Terenzini. 1998. "Studying College Students in the 21st Century: Meeting New Challenges." *Review of Higher Education* 21:151–165.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. 2018. *Integrated Public Use Microdata Series: Version 8.0 [dataset].* University of Minnesota, Minneapolis, MN. doi:https://doi.org/10.18128/D010.V8.0.

Seftor, Neil S., and Sarah E. Turner. 2002. "Back to School: Federal Student Aid Policy and Adult College Enrollment." *The Journal of Human Resources* 37 (2): 336–352.

US Dep't of Education. 2004-2016a. *Federal Pell Grant Program End of Year Reports.* https://www2.ed.gov/finaid/prof/resources/data/pell-data.html. [Accessed Dec 2, 2017]. US Department of Education.

———. 2004-2016b. *The EFC Formula.* https://ifap.ed.gov/ifap/byAwardYear.jsp?type=efcinformation&set=archive. [Accessed Dec 2, 2017]. US Department of Education.