Predicting Housing Prices Based on Features of the Home prepared by G1

By Carlos Alves Pereira, Tommy Ge and Jinghong Li

Abstract

There are many factors which contribute to the real monetary value of a House. Houses have many varied features to them, which can include total area, number of bathrooms, number of kitchens and over thirty other features, which are covered in this project. This project makes use of classification methods, to determine some of the most important features, which directly affect price. Specific classification methods used and discussed included, K-means clustering as well as K-prototype clustering. From this information, we created a new dataset and used three linear regression methods to determine whether or not the new dataset improved the data, which according to our R-squared analysis did include small improvements, due to using the updated dataset, compared to our original dataset. From both the classification and regression methods used, we determined that the overall quality and finish of the materials used to build the house had the largest impact on housing prices, as well as having additional floors, which included upper floors and basements. The best model that predicted the housing prices was the multiple linear regression model combined with the dataset created through the clustering methods.

Introduction

The housing market is a trending topic in current events as the Covid-19 Pandemic caused housing prices to fall. Currently housing prices are rising again due to increased demand. Our project addresses this event by testing models to predict housing prices to try to gain insights into what types of amenities or features are desirable in a home. The end goal is to compare different models used in this project and find the best model for predicting house prices based on amenities, while also attempting to determine the factors, which are most important to increasing housing values. Our project and problem could be of importance to homeowners and renters who would like to evaluate the price of their current homes.

Related Work

Approaches, which have already been used in the past to predict housing prices include mostly predictive modeling methods. In terms of the specific methods used, multiple papers reference various types of multivariate regression and clustering. A paper by Chakraborty [4] made use of multiple linear regression techniques, such as ridge regression and lasso regression, which are used in this report. Another method by Donghui Shi, Jian Guan [2], which is employed in this project is clustering data based on the submarkets they come from. Another interesting method, which has been used for predicting housing prices included K-nearest neighbors, which according to Weikun Zhao, Cao Sun and Ji Wang [7] performed better than Bayesian classifiers, Decision Trees and Support Vector Machines for numerical price predictions. Other mentions of clustering large datasets with categorical values through the use of k-means clustering are mentioned by Zhexue Huang [8].

Methodology

Why do we add the new variable AverPrice to the dataset?

A main obstacle to efforts to obtain an accurate valuation of real estate properties is the heterogeneous nature of real estate data. An increasingly common approach to reduce data variability and improve accuracy is to cluster the data set into submarkets that are more homogeneous. Kmeans helps us group properties into more homogeneous clusters.

According to Donghui Shi, Jian Guan [2], the accuracy of real estate price prediction can be noticeably improved by some of the clusters, they offer FCM clustering method to cluster data and introduce an additional input in price prediction. We came up with the idea to try using KMeans to generate a new variable -- the average prices of each property's k nearest cluster centers, i.e., AverPrice, and to see if it can noticeably help improve prediction accuracy.

Permutation feature importance

Donghui Shi, Jian Guan [2] use the longitudes, latitudes, and sale prices of properties to obtain new variables, while our dataset doesn't contain the specific longitudes and latitudes of each property, we then use the most important k variables instead. Random Forest algorithm and permutation importance calculated by R squared score, are applied to decide the most important variables in our training dataset. The permutation feature importance is the decrease in a model score when a single feature value is randomly shuffled, which is broadly applicable as it doesn't rely on internal model parameters and is recommended for almost any model.

K-means clustering

After extracting the most important variables – OverallQual, GrLivArea, BsmtFinSF1, TotalBsmtSF, and the target variable SalePrice, Kmeans clustering model generates ten cluster centers based on our training dataset. We use Euclidean distance, the square root of the sum of the square differences, to compare their similarity, because they are all continuous variables, in addition, Euclidean distance metrics have less distortion and converge more quickly. [K-means with Three different Distance Metrics] The number of clusters is determined by the Elbow Method, the "elbow" i.e., the point after which the inertia starts decreasing in a linear fashion. Then we average the sale prices of the nearest k cluster centers adjacent to each property, insert these values as the new variable AverPrice to our original dataset.

How is our method applied to a new observation?

Steps applied to a new observation included the following:

1.Extract OverallQual, GrLivArea, BsmtFinSF1, TotalBsmtSF of the new observation
2.Calculate the distance to above generated cluster centers based on these variables
3.Select the nearest k centers, average their sale prices
4.Add this new variable AverPrice to the observation and predict it's sale price through regression models

Other attempt on the clustering models: K-prototype clustering

Although the most important variables are all numerical, we can use K-means smoothly, we notice that the dataset contains a lot of categorical variables and hope to utilize the dataset

as complete as possible. K-Prototype clustering extends K-Means and K-Modes and is particularly adapted to handle mixed datasets that contain both continuous and categorical variables. The distance between a point and to its cluster center(its prototype) that is to be minimized is $D(x,p) = E(x,p) + \lambda C(x,p)$, where E is the euclidean distance between the continuous variables and C is the count of dissimilar categorical variables. We selected “Huang” as the init, the model will select the k distinct objects from the data set as initial k-modes and then assign the most frequent categories equally to the initial k-modes. The Elbow method based on the sum distance of all points to their respective cluster centroids helps us find the optimal k that is 15.

But the limitations of the K-prototype clustering is that it uses a custom dissimilarity metric to handle both categorical and continuous variables, which makes it difficult to specify the difference of observations’ distances. But it can still be used for statistical data analysis.

Data Preparation

The data was prepared by removing any nan values as well as dealing primarily with numerical values, because of issues relating to difficulties with being able to translate some of the categorical variables with more than 2 possible values to a numerical value, which would accurately represent the variable. This data was then split randomly multiple times into Train-Validation-Test sets as described below.

Train-Validation-Test split

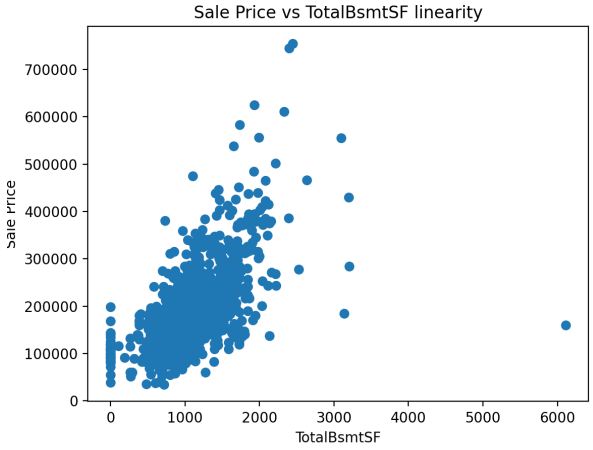
For testing all the regression methods, the training data with our newly chosen features was randomly split into train and validation sets for multiple iterations. The specific dataset taken from kaggle [1] included the training data and testing data in separate files The training data set was split into a training set with 75% of the data and a validation set with the remaining 25% of the data. The data chosen for each training set was randomly chosen for every iteration. The test was also repeated multiple times and averages were taken for all the error calculations over all the iterations chosen. Note that the test data did not come with the “SalePrice” data which was the variable we want to predict. This means we cannot compute an r-squared value for our final model chosen at the end. Instead we decided to look at the r-squared values from our validation splits.

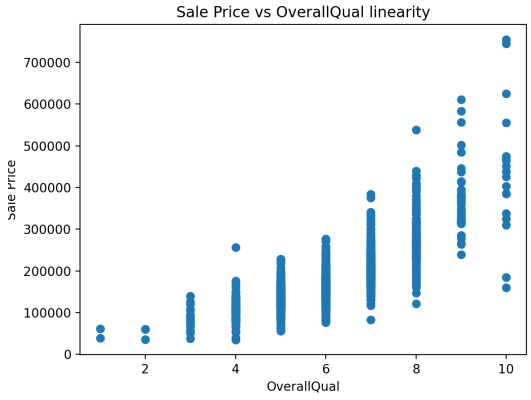
Linear Regression

Our first idea we had to create a model to predict house prices was to use a linear regression model. For this model we wanted to predict the sale price (“SalePrice”), so that was the dependent variable. We chose to include 37 features to train on for this model which were MSSubClass,LotFrontage,LotArea,OverallQual,OverallCond,YearBuilt,YearRemodAdd,MasVnrArea,BsmtFinSF1,BsmtFinSF2,BsmtUnfSF,TotalBsmtSF,1stFlrSF,2ndFlrSF,LowQualFinSF,GrLivArea,BsmtFullBath,BsmtHalfBath,FullBath,HalfBath,BedroomAbvGr,KitchenAbvGr,TotRmsAbvGrd,Fireplaces,GarageYrBlt,GarageCars,GarageArea,WoodDeckSF,OpenPorchSF,EnclosedPorch,3SsnPorch,ScreenPorch,PoolArea,MiscVal,MoSold,YrSold,SalePrice,AverPrice. The descriptions of these variables can be found at the source of the data on Kaggle[1].
First, we tested assumptions of the data to see if there is a linear relationship between the independent variables and the dependent variable, and if the variables are normally distributed.

With our assumptions checked for a better understanding of the data and to understand what features we should use for linear regression, our next step was to train the linear regression model on our data. After training, we validated this model with a validation dataset and calculated an r squared value. Our final step was to compare this model with our other models considered in this project.

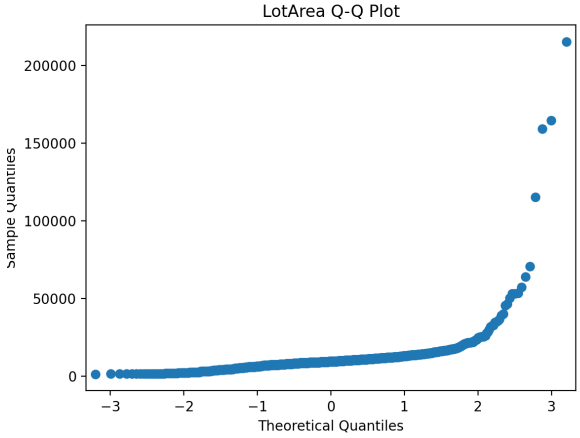
For the linear regression model we first checked for linearity between the sale price (“SalePrice”) and other features in the dataset:

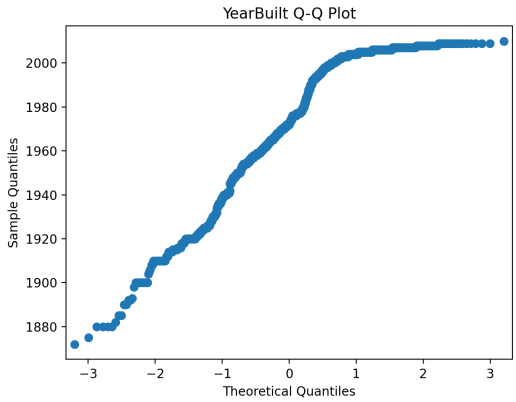




The two graphs shown above are examples of how the features were plotted against the value we wanted to predict (“SalePrice”) to check for linearity in the data. As we can see the features here seem to have some linear correlation with sale price just based on an eye examination. Generally all the features showed some slight linear correlation with our dependent variable. The number of features used to train the linear regression model was 37.

Next, the features were checked to see if they were normally distributed. To to this we made quantile-quantile plots of each feature.





We can see from “LotArea Q-Q Plot” and “YearBuilt Q-Q Plot” that these plots do not indicate that these variables are normally distributed as they do not form a straight line. This may be of

interest for optimizing or finding a different model. We still decided to continue with linear regression even though we had troubles with the normal distribution of variables.

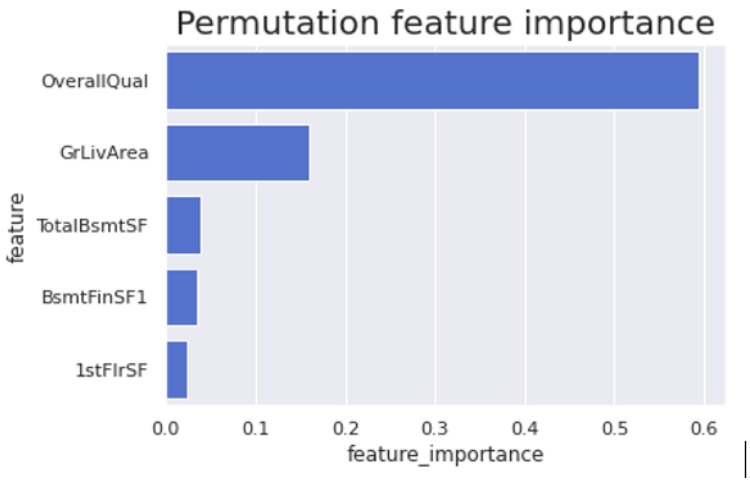
Ridge Regression and Lasso Regression

For the Ridge and Lasso regression tests were done on both the original dataset and the improved dataset chosen through clustering. The way that Ridge regression works is it makes use of the Linear Least Squares L2-Regularized Regression algorithm to calculate the coefficients for each variable. [5] What makes it different to the linear regression model above is that it applies a penalty on the size of the coefficients. The effectiveness of these methods was tested using the R squared validation method, which basically measures how far away the predicted data is from the line of best fit. For both the Ridge and Lasso Regression, the regularization parameter of alpha was chosen through brute force, where the parameter was chosen between 0 and 50 with increments of 0.1, to find the alpha, which led to the largest average R squared value. Through this testing, we determined that we would use an alpha of 4.1, as that led to the largest R squared value. For Lasso Regression, the same method was used to obtain the optimal regularization parameter, and we found that the best R squared value was received when alpha was set to 20.2. Both these methods were tested over 100 iterations of randomly generated Train-Validation-Test methods as shown above and average values for the parameters were used to determine the most impactful parameters when valuing a house.

Empirical Results

The key question we wanted to answer from this project was creating a model that best predicts the sale price of a house from our dataset. The results here indicate how we chose the features for our models, the r-squared values after training our models, and the model we chose based on the r-squared value of our models.

Permutation feature importance



According to the index of the decrease in r2 score in shuffled dataset: the top5 important numerical features are: OverallQual, GrLiveArea, TotalBsmtSF, BsmtFinSF1,1stFlrSF. We can see that the overall material and finish of the house plays a remarkably important role in evaluating the sale price. Above grade (ground) living area square feet is the second highest importance feature, which is also a common sense, but it was surprisingly rated after the overall material, we infer that the locals would prefer to live in a safe and comfortable house rather than live in a big house.

K-prototype clustering

Cluster 9&10&11 have the highest sale prices, their masonry veneer type are all Brick Face, and their ratings of basement finished area are all the top level (Good Living Quarters), they have apparent distinction in these two categorical variables with other lower prices' clusters.

Cluster 6&7 have the lowest sale prices, their exterior covering on house are all wood siding, this verifies our previous inference that the locals attach more importance to houses' security, in addition, cluster 6 is the only cluster with medium residential density, the others are all low, which indicates that high residential density may affect house prices in this region.

As the graph shows that clusters 9&10&11 gather in high OverallQual and high GrLivArea area, and there is a slight linear relationship between these two variables.

Linear Regression

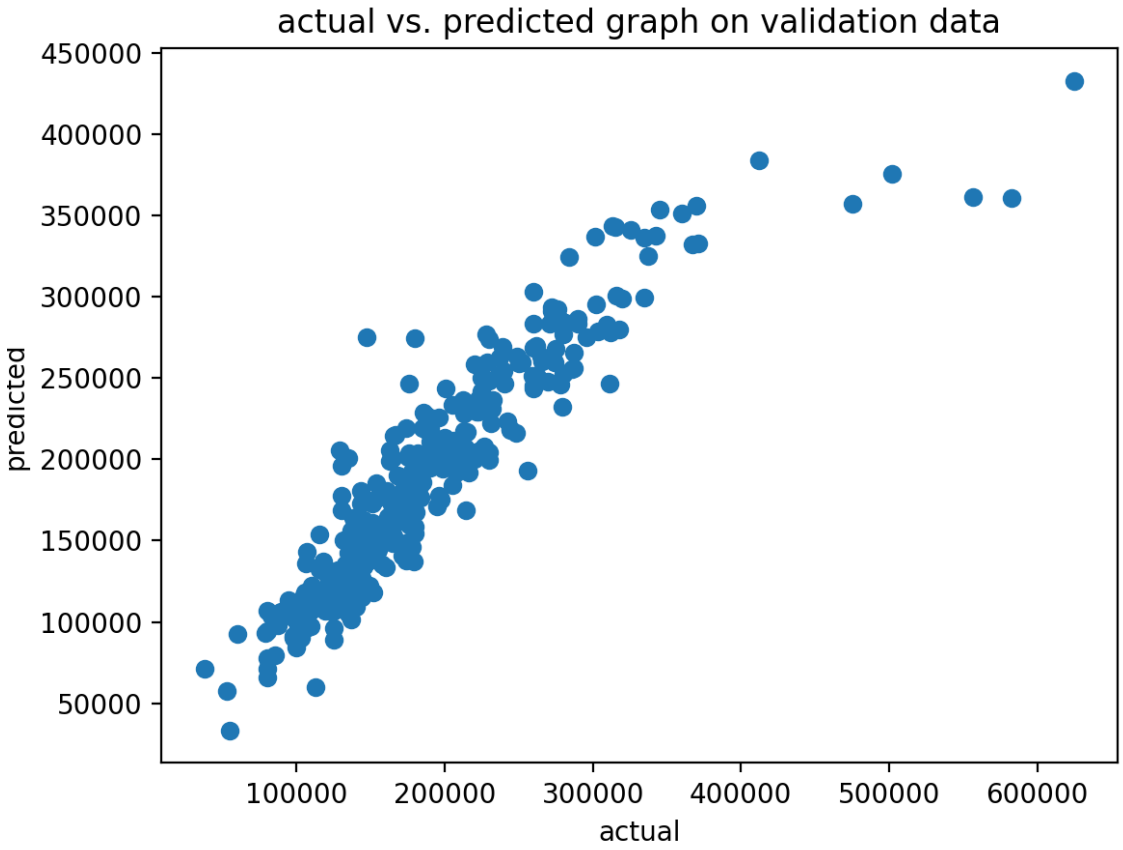
Over 100 iterations, all with random train-validation-test sets, the R squared values were as follows:

The original dataset means that all the features from the original dataset were used in this model without doing any feature selection.

Dataset	**Regression Type**	**R-Squared Value**
Original Dataset	Multiple Linear Regression	0.831
Updated With Clustering	Multiple Linear Regression	0.845

We can see from the R-squared values in the above table that the r-squared value for this model increased once we did feature selection. The r-squared value seems relatively high for

this model as this indicates that 84.5% of the variance in the data is accounted for by our multiple linear regression model.



We see from the actual vs. predicted graph above for our linear regression model that generally our predictions fall in a linear line with the actual values. This graph was created from validation data as our test data from Kaggle did not come with the actual values. This graph indicates that this model could be good for predicting the sale price of a house because of the linear relationship shown.

Ridge and Lasso Regression

Over 100 iterations, all with random train-validation-test sets, the R squared values were as follows:

Dataset	**Regression Type**	**R-Squared Value**
Original Dataset	Ridge	0.726
Updated With Clustering	Ridge	0.738
Original	Lasso	0.734
Updated With Clustering	Lasso	0.741

Both Ridge Regression and Lasso Regression improved with the updated dataset, which included clustering and average price as a parameter. Also, Lasso regression seemed to perform better than Ridge regression albeit by a small amount.

As for the parameters, which had the largest coefficients according to Ridge and Lasso regression, it was found that the top 3 were OverallQual, TotRmsAbvGrd and KitchenAbvGr. These relate directly to Overall material and finish quality, the number rooms above the ground and and number of kitchens above the ground. What we could infer from this data is that the quality of the materials that the house is made of has a significant impact on the price of a house. Also, since two of the parameters that were most effective refer to houses, which have more than one floor, it is also possible that having a larger sized house would lead to an increase in house valuation, which logically would make sense.

Conclusion/Discussion

Overall the use of the updated dataset, which was created through clustering with the best features did lead to an improvement in R-Squared values for the multiple linear regression, ridge regression and lasso regression, however the improvements were fairly small but consistent. Out of the regression methods, multiple linear regression with the updated data had the best R squared value overall. An interesting observation, which came from both the clustering and the regression methods, was that OverallQual was the most significant factor when determining housing prices. This is a numerical value, which relates directly to defining the overall material and finish quality of a house. This is very interesting as it shows that the factors used to determine the value of OverallQual correctly determine housing pricing. Interesting observations, which we can take from combining the clustering results is that grouping houses by whether they have basement rooms or upper floor rooms seemed to have a large impact on pricing. The k-means clustering results found that the parameters related to basement rooms quality and area had some of the most impacts on price, while the regression results showed that the number of rooms on upper floors had some of the most significant impacts to price. This seems to suggest that having multiple floors to a house leads to an increase in house valuation, as when purchasing a house, the information generally shown summarising houses includes attributes, such as number of bathrooms, bedrooms and so on. Overall, the values, which most people would expect to increase housing prices, such as number of rooms and floors do increase housing prices, however the overall quality of a house seems to be determined by the materials, which make up the houses themselves.

Bibliography

1. https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data?select=test.csv
2. Shi, Donghui; Guan, Jian; Zurada, Jozef; and Levitan, Alan S. (2015) "An Innovative Clustering Approach to Market Segmentation for Improved Price Prediction," *Journal of International Technology and Information Management*: Vol. 24 : Iss. 1 , Article 2. Available at: https://scholarworks.lib.csusb.edu/jitim/vol24/iss1/2
3. Okmyung Bin, A prediction comparison of housing sales prices by parametric versus semi-parametric regressions, Journal of Housing Economics, Volume 13, Issue 1, 2004, Pages 68-84, ISSN 1051-1377, https://doi.org/10.1016/j.jhe.2004.01.001
4. Chakraborty, D., Elhegazy, H., Elzarka, H., & Gutierrez, L. (2020). A novel construction cost prediction model using hybrid natural and light gradient boosting. *Advanced Engineering Informatics*, *46*, 101201. doi:10.1016/j.aei.2020.101201
5. scikit-learn developers. (n.d.). *1.1. Linear Models*. scikit. Retrieved November 2, 2021, from https://scikit-learn.org/stable/modules/linear_model.html#ridge-regression
6. Data4help. (2020, November 2). Clustering real estate data. Medium. https://becominghuman.ai/clustering-real-estate-data-594894e24484
7. Zhao, W., Sun, C., & Wang, J. (2014). The research on price prediction of second-hand houses based on KNN and stimulated annealing algorithm. International Journal of Smart Home, 8(2), 191-200. https://doi.org/10.14257/ijsh.2014.8.2.19
8. Huang, Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery 2, 283–304 (1998). https://doi.org/10.1023/A:1009769707641
9. Singh, A., Yadav, A., & Rana, A. (2013). K-means with Three different Distance Metrics. International Journal of Computer Applications, 67, 13-17.