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

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 creating 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. 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, ridge regression as well as Support Vector Machines and varying classification methods, which have been or will be tested in this project.

Change in Scope of Proposal

The focus of this project will still be to find the best regression and clustering models in order to predict housing prices. The chosen methods used in this paper have been changed to more fit the data provided in the Kaggle dataset. The focus will be on choosing the best regression and classification methods, which have been initially tested below and doing further analysis on the best four methods. These models will be compared using various error analysis metrics, such as R-Squared and Residual Mean Square Error.

Methodology

Linear Regression

Our baseline model we thought of for this problem was to create a linear regression model as this is the classic method for creating a predictive model. The variable we wanted to predict was the ‘Sale Price’ of a house. The variables chosen for our preliminary linear regression model were 'LotFrontage', 'LotArea', 'OverallQual', 'OverallCond', 'YearBuilt', 'YearRemodAdd', 'TotalBsmtSF', '1stFlrSF', '2ndFlrSF', 'GrLivArea', 'FullBath', 'TotRmsAbvGrd', 'GarageArea', 'YrSold' which are all numerical and described in the description of our dataset. In addition to fitting the regression model, we also checked the linearity, homoscedasticity, independence, and normality assumptions to note any possible problems with this model. With our multiple linear regression model, we hope to compare this to other predictive models with an r-squared or MSE value. For the next iteration of our project we are planning on adding in categorical variables and trying all variables in the data set to see how the fit changes and what variables affect the model.

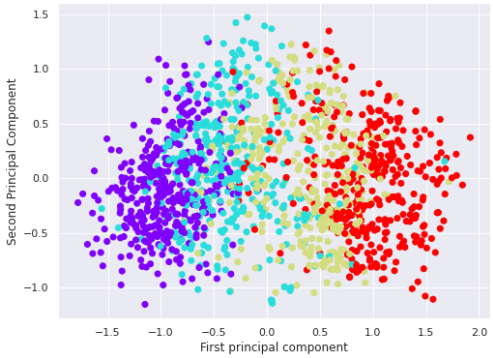
Ridge, Lasso, Elastic-Net Regression

For Ridge, Lasso and Elastic-net regression the analysis was done solely on numerical variables, such as LotArea, OverallQual, OverallCond and YearBuilt. 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. What makes it different to the linear regression model above is that it applies a penalty on the size of the coefficients. For the Lasso, Ridge and Elastic-Net Regression models, the variables chosen were LotArea, OverallQual, OverallCond and YearBuilt and comparisons between the methods were done using 'R-Squared' Coefficient of determination. For the preliminary results, all the models used a default value of alpha of 1 for the regression. The Elastic-Net regression had an l1-ratio of 0.5.

K-means Clustering

Clustering can help us identify which large category of properties a given property belongs to, and which features most affect the final categories. In our study, we plan to obtain the potential relationship above by using K-means. For the database of our choice, the clustering process can be explained in the following steps:
(1) Clean the data and do some pre-processing. We deleted variables irrelevant with sale prices, like Id, as well as some variables containing too many missing values, such as MSZoning, LotShape, in addition we use mean value to fill with sparse missing values, and scale the data to range between 0 and 1. (2) Reduce dimensions while preserving data characteristics as much as possible, as k-means operates only on distances, the right distance matrix to use should be the one preserved by the dimensionality reduction. We respectively use PCA to reduce dimensions to two and the number with the explained variance to be 95%, the former helps us visualize the data in general, the latter is used to obtain the final result. (3) Cluster the processed data by the k-means method from the sklearn package, the appropriate number of clusters is determined using the so-called Elbow method. (4) Evaluate the performance of the k-means clustering model by SSE, and discuss its feasibility.



The price interval clusters after the PCA

K-nearest Neighbors (KNN) Classification

Compared to Bayesian classifiers, decision tree, KNN classification is more usually used in numerical price prediction. The classification preparation is mostly the same as K-means clustering, while "house price" is set as the label to do supervised learning. The elbow method still helps to select the optimal number of clusters, and the KPI is accuracy.

Hierarchical Clustering

Hierarchical clustering is a data mining technique used for grouping similar objects into clusters. Clusters are created based on the distance between points. Clusters are also merged based on the distance between clusters in this algorithm. The end visualization is a dendrogram where you can see each cluster and how they are merged. We chose to try hierarchical clustering to see if we can identify features which are important in a house. This will also help us view data that are similar in our dataset.

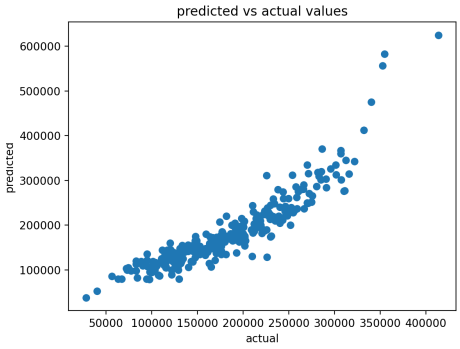
Error Analysis

The methods that will be used for error analysis and comparisons between different methods include R-Squared, Residual Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE). Currently, R-Squared has only been used to calculate errors for regression models.

Preliminary Results

Linear Regression

The r-squared value of our linear regression was 0.782. This means that 78.2% of the data can be explained by our model. For our future work we will add more variables as well as categorical variables to try to get a larger r-squared value.



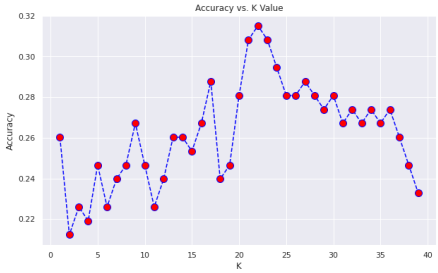
As we see from our predicted vs actual graph, our current model trends in a curve, especially when the values get higher. This indicates this model is over-estimating values as the prices get larger. We will try to correct this in our future iteration by doing outlier analysis and possibly changing what variables we use in the regression.

Ridge, Lasso, Elastic-Net Regression

For the variables of LotArea, OverallQual, OverallCond and YearBuilt, Ridge Regression, Lasso Regression and Elastic-net Regression had the following R-Squared values:
Ridge: 0.669, Lasso: 0.669, Elastic-Net: 0.643
The R-squared for both Ridge and Lasso was almost the same, while the Elastic-net performed worse, in terms of the R-Squared distance. There isn't enough evidence to determine whether one method is necessarily better than the other. Further testing will be done using larger varieties of variables and using other previously discussed methods of calculating error.

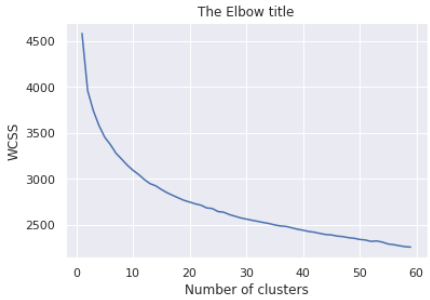
Clustering

K-nearest Neighbors (KNN) Classification



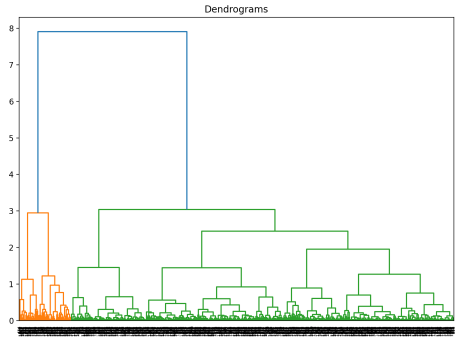
The optimal value of K is 22. The error rate is at the lowest point in this position, i.e., about 0.32. Commonly the classification accuracy above 80% is acceptable, so the original KNN classification model does not work well on this database.

K-means Clustering



The optimal value of K is at the “elbow”. Thus for the given data, we concluded that the optimal number of clusters is approximately 15. Sum of squared distances(SSE) of samples to their closest cluster center is 2883.585, for formula $\sqrt{SSE/(N-1)} = RMSE$, RMSE is 1.385 which can not be considered as a good R-value.

Hierarchical Clustering



The dendrogram plot shows how the clustering joined the objects in our dataset. The tree structure shows where each object was joined, so the structures that are closer together suggest that those objects are closely related and should be in a cluster. To analyze the features, we need to use the dendrogram to further break our data into clusters. This is our next task with this method, though this may prove difficult because of the amount of features that need to be analyzed in a cluster.

Plan of Work

What still has to be done includes further testing of regression and classification methods using the updated error analysis methods. The classification methods used different error comparisons, so these need to be updated to use the same error comparisons. Another step is iterating on all the models, with the features, which our models have found to be the best for determining the housing prices, in order to minimize the error of our final model.

Conclusion

In terms of the work, which has been, the regression model with the best outcome so far is only in terms of the R-Squared value is Linear Regression. The linear regression model used more variables than the other regression models, which could have led to this better R-Squared values, so the other models will be updated and tested to see if that was the main reason. In terms of the clustering models, varying error methods were used on each, due to the nature of how the clustering methods were done. These have to be updated to use the same error methods, so that comparisons can be done. Some papers pointed out that weighted-KNN predictions are more precise than KNN in terms of estate market data. We will apply this method in our database to see if we can get a prediction promotion. In addition, we will try to use the appropriate clustering model to derive more desirable features.

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