COMP 562 Final Paper: Predicting Sale Prices of Houses

Maxwell Peng, Emma Jia, Hua Jiang, Teresa Pan

Introduction

Unlock the secrets of your dream home's value and discover the price of your next abode today. The real estate market can be intimidating as it is a complex and dynamic domain. There can be so many factors influencing the prices of properties, so it's not as simple as looking at the square footage and lot size. In this paper, we will analyze the various factors that go into determining a house price and 79 potential variables. We will use different models and machine learning algorithms such as Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regressor, and K nearest neighbors to help aid us in this examination. The motivation behind this project is to help create the most accurate model that predicts house prices after looking at all significant factors which are beneficial to any house buyer.

The Dataset and Data Cleaning

Our original dataset containing 2930 observations and 79 predictors was obtained from Kaggle and comprises actual housing data from Ames, Iowa between 2006 and 2010. Even though the dataset contains a plentiful amount of data on house price predictors, we had to do some cleaning on it before we could use it to run tests and algorithms.

Our steps are outlined below:

- We deleted the variables that had more than ten percent null values so that the missing data wouldn't skew our results.
- For the remaining variables, we filled in the null values with the mean of the values of that column.
- We also dropped the variables that had clearly no influence on the results, such as "id".
- We chose to keep only numeric variables because we can't run regressions on categorical values.

We saved these new datasets as new_train.csv and new_test.csv.

The Plan

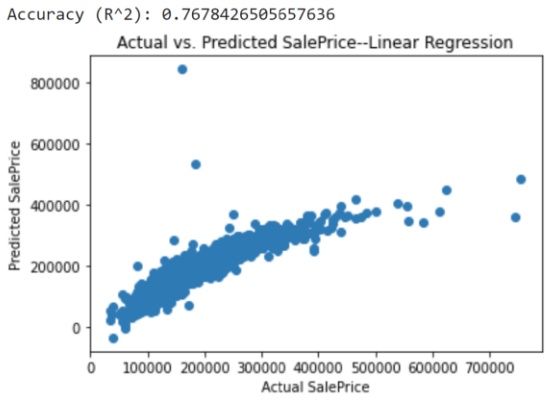
We split our data into train and test with the training dataset used to train the model and the test used to evaluate the performance and accuracy of the trained model. These two datasets were given by Kaggle. We will then apply machine learning algorithms to develop models that can predict housing prices. We will then examine the results and examine the accuracy of each model.

Regression and Models

1. Linear Regression

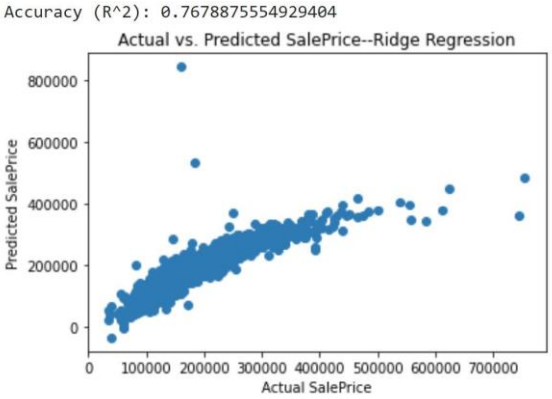
Our first approach was to use a Linear Regression Model to predict the sale price of

homes. We implemented this by using the Scikit-learn library in Python. The input variables for the model were made by dropping the SalePrice from the training dataset and the output variable was set to be SalePrice. We used the *fit()* method from the *LinearRegression* class to fit the model onto the training dataset. Furthermore, we used cross-validation with a 6-fold validation scheme to evaluate the performance of our model. The R^2 values generated from the validation were used to assess the validity of the model. To create the predicted values for each point in the data in the training set that we plotted against the actual values with Matplotlib, we used the *cross_val_predict()* function. Lastly, we used the R^2 *score()* function to calculate the R^2 score of the model on the training data.



2. Ridge Regression

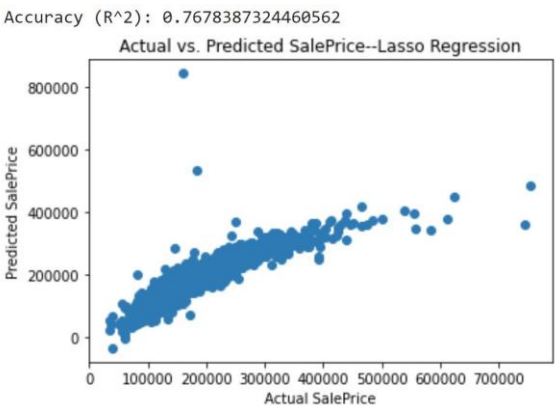
Ridge Regression is a regularized linear regression model used to mitigate problems of multicollinearity.

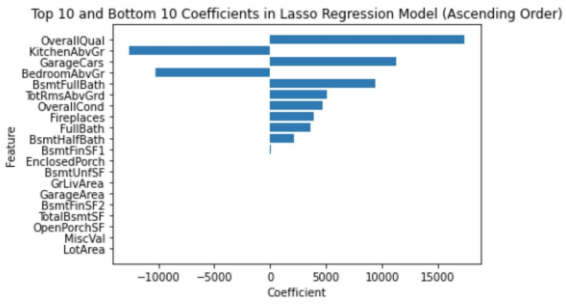


We initiated the model with an alpha value of 1.0 and prepared the training data by splitting the dataset into input and output variables and then making predictions of the test data. We use 6-fold cross-validation and R^2 to evaluate the model's accuracy. We created a scatter plot of actual vs. predicted sale prices using Matplotlib.

3. Lasso Regression

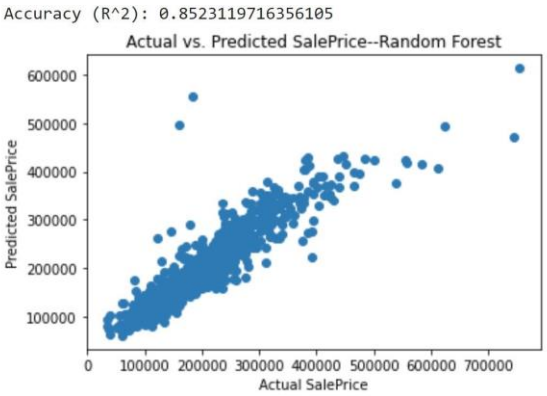
A Lasso Regression is a type of regression that uses shrinkage, which is where data points are shrunk to a central point such as the mean. It's useful for models with high multicollinearity. The model is created using the Lasso function from the *sklearn.linear_model* module. We used an alpha value of 1.0 for regularization strength, which is used to regularize coefficients that become exactly zero. We fit the model onto the training data and performed 6-fold cross-validation using the *cross_val_score* and then the *cross_val_predict* functions to analyze the accuracy of the model. The R^2 score is used to measure the accuracy of the results. We also created a horizontal bar graph that displays the top and bottom ten coefficients in the Lasso Regression model.





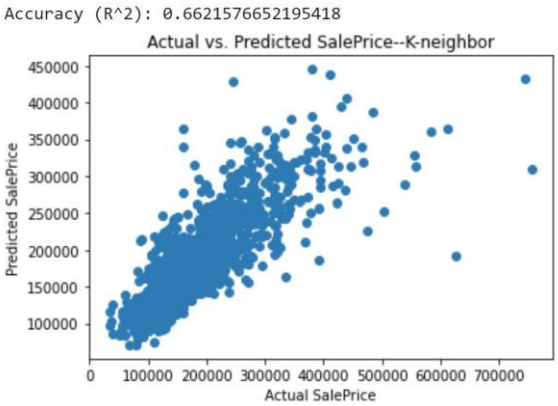
4. Random Forest Model

Using the scikit-learn library in Python, we built a model to predict housing prices. We instantiated the model with 100 decision trees and a random state of 0. We created the training data by dropping the SalePrice column and setting it as the target variable that we are predicting. We fit the model onto the training data using the *fit()* method. A 6-fold cross-validation was performed utilizing the *cross_val_predict* method. We once again found the R^2 score and plotted an actual vs. predicted sale prices graph.



5. K Nearest Neighbors Regression

The K Nearest Neighbors attempts to predict the correct class for the test data by calculating the distance between the test data and all training points. We created a KNN regression model with five neighbors and fit it onto the training data. A 6-fold cross-validation was performed to address the performance of the model and the predicted sale prices were plotted against the actual sale prices. We calculated the accuracy of the model using the R^2 score.



Conclusion

In this study, we used Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regressor, and K Nearest Neighbors to examine the performance of these models in predicting housing prices.

The accuracy of each model was evaluated using R^2 and we found out that the Random Forest Regressor had the highest accuracy with an R^2 score of 0.85, indicating that it is the most accurate choice for predicting housing prices. On the other hand, the K Nearest Neighbors model had the lowest R^2 score of 0.66 so it is not a good choice for making this prediction. Overall, our examinations and results indicate that out of the five models we chose, the Random Forest Regressor is the most reliable model for predicting housing prices in our dataset. This model would help people estimate the price of their dream house using their ideal standards and requirements.

References

[1] Joby, "What is K-nearest neighbor? an ML algorithm to classify data," *Learn Hub*. [Online]. Available: $https://learn.g2.com/k-nearest-neighbor$. [Accessed: 26-Apr-2023].

[2] G. L. Team, "What is ridge regression?," *Great Learning Blog: Free Resources what Matters to shape your Career!*, 16-Nov-2022. [Online]. Available: $https://www.mygreatlearning.com/blog/what-is-ridge-regression/$. [Accessed: 26-Apr-2023].

[3] "House prices - advanced regression techniques," *Kaggle*. [Online]. Available: $https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data$. [Accessed: 26-Apr-2023].

[4] "Sklearn.linear_model.linearregression," *scikit*. [Online]. Available: $https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html$. [Accessed: 26-Apr-2023].

[5] Stephanie, "Lasso regression: Simple definition," *Statistics How To*, 27-Apr-2021. [Online]. Available: $https://www.statisticshowto.com/lasso-regression/$. [Accessed: 26-Apr-2023].