**Comparison on the predictions of housing prices in Ames, Iowa between OLS, Random Forest, and Ridge**

HARVEY WENHAO JIANG

*University of Southern California, Economics Department, Kaprielian Hall, 3620 S Vermont Ave, Los Angeles, CA 90089*

*Corresponding author: wenhaoji@usc.edu*

*May 5, 2021*

**Abstract**

**This paper aims to compare different results rendered by the OLS, Random Forest, and Ridge models on the housing data from Ames, Iowa, and provides explanations and insights into the difference among the three predictions. The paper discusses how the dataset is processed and cleaned for model training. Data visualizations are provided to gain insights on the key features and characteristics on the data. Out of the three different models, Ridge regression was the best performing with the highest accuracy and lowest RMSE score, and Random Forest was the worst performing model. With the increasing popularity of the Random Forest method in modern data science research, this paper intends to show that Random Forest may not always be the ‘go-to’ method, and that the model selection process can be methodical.**

1. INTRODUCTION

Selection of machine learning models and algorithms still is an arbitrary process in data science today (Kirasich et al. 2018). With the abundant amount of machine learning algorithm packages on coding languages such as Python, data science practitioners tend to include and implement every available package during the data modeling process. This way, practitioners will be able to select the best performing model by comparing model accuracy and errors among the long list of machine learning models that they implement to the datasets. The ad-hoc process of machine learning model selection is the main motivation for this study.

Every dataset is different with its unique characteristics and dimensions. This study has chosen a comprehensive dataset on housing from Ames, Iowa. This dataset contains various types of data, which makes it difficult for practitioners to know which specific model to implement. More detailed descriptions about the features of the housing dataset will be discussed below in data processing. Because of this, analysts would resort to the arbitrary machine learning model selection as mentioned above. The goal of this paper is not only to generate predictions for future housing sale prices in Ames, but also compare the prediction results rendered by three commonly used machine learning models: Linear regression, Random Forest, and Ridge.

1. DATA PROCESSING

Data processing and cleaning is a very essential procedure in any data science project. Real-world datasets are filled with human errors and missing values during the collection process. Data science researchers need to clean and process the raw datasets for anything that may jeopardize the integrity of machine learning algorithms training. For the housing dataset that is being analyzed in this study, I have applied the same standard procedure used by practitioners.

* Loading data

Two sets of data were provided by the source: train set and test set. For beginners, train set is the dataset that is used to train the learning algorithms in the research, and test set is the dataset that is used to test the accuracy of the models that are previously trained. Here I import the Pandas package from the Python libraries to load and read the two sets of data. Two sets of data, train and test, then respectively displayed via output to further study their features and characteristics. The training set contains numerous features including ID, street, lot shape, different squared footage measurements, style descriptions of the house. We can see both categorical and numerical variables in the train dataset, which I will process under encoding.

* Dealing with dataset shape and missing values

In dealing with missing values in the datasets, I first look at the shapes of individual dataset. From the code, we see that the shape of training data has 1460 rows and 81 columns, and testing data has 1459 rows and 80 columns. The one missing column in the test set is the target variable, house sale price. Now we know the shape of each data set, I will then look for missing or null values in the data. By compiling a list of variables with the most null values in descending order, I would drop the top three columns from the data sets. For variables with homogenous values, I would also drop them from data. Then a code is generated to fill the categorical variables with the value ‘None’ and the numerical variables with the median of all rows in order to avoid anomaly in values.

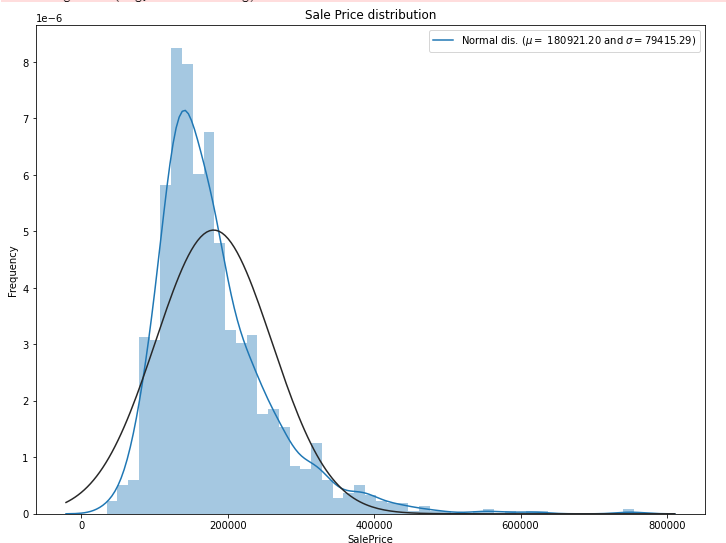
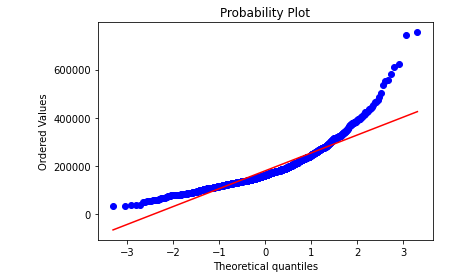
The next step is to generate a heatmap to ensure that there are no missing values left in both data sets. With the completion of processing missing values, I proceed to the last step of data preprocessing, encoding.

* Encoding

Data encoding is one of the most important process in data cleaning. It simply encodes the variables in the datasets with appropriate labels in order to machine learning algorithms to better identify the nature of exogenous variables. I label the exogenous variables into numerical, and categorical nominals and ordinals. With LabelEncoder, I added dummy variables to avoid multicollinearity in the datasets. The final step would be to drop the initial categorical nominal columns and replace them with new ‘one hot encoded’ columns. Encoding helps the machine to learn the features of a certain exogenous variable, and this would be the last step for data processing and cleaning.

1. DATA VISUALIZATION

Before I begin the most important data modeling process, I want to analyze the target variable “Sale Price” for trends and correlations via tables and graphs. First, a graph of the distribution of house sale prices is generated as we can see in the codes’ output. The sale price distribution is skewed to the left, and we see the probability plot below is showing a log shape function, as shown below in figure 1.

Figure 1: Distribution of Sale Price for Houses in Ames, Iowa, and its probability plot

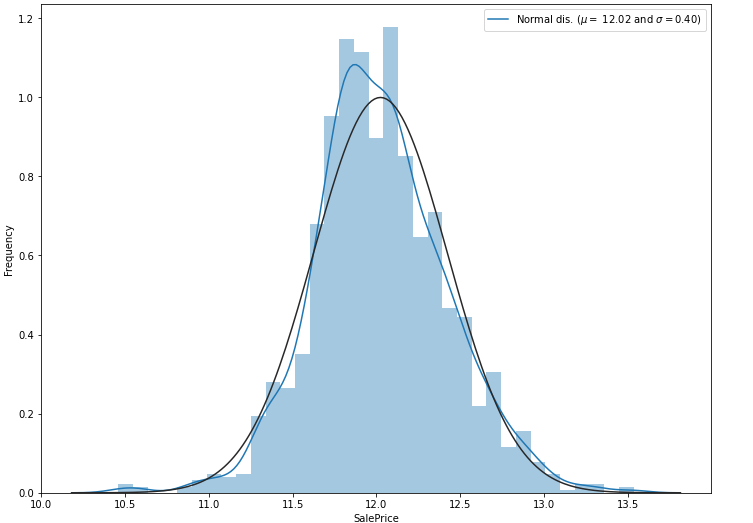
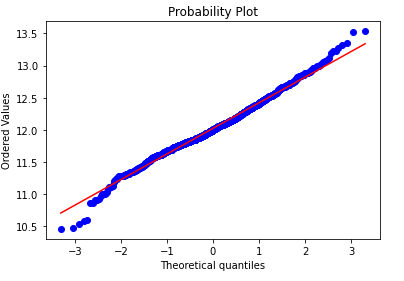
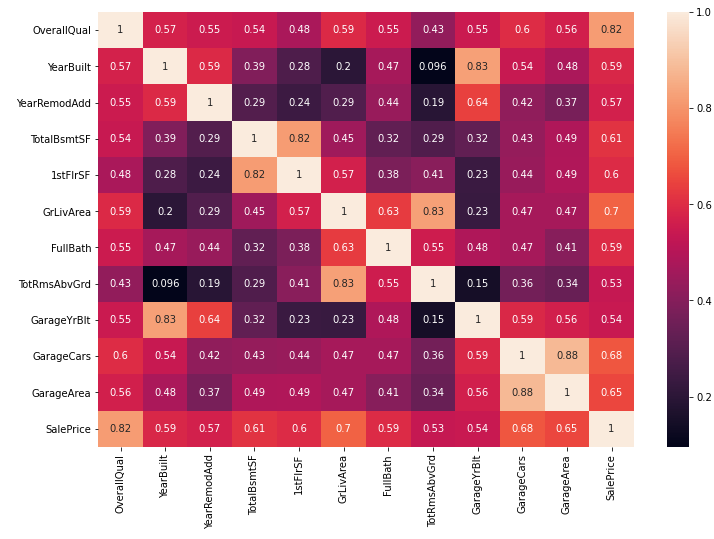
Because of this discernable feature, I implemented a log transformation to the target variable, sale price, which would render to approximate normal distribution.

Figure 2: Log Distribution of Sale Price for Houses in Ames, Iowa, and its probability plot

After the log transformation, we see that the sale price distribution is close to being normalized, and that its probability plot is showing a linear relationship. From the log transformation, we can more precisely identify which machine learning models can perform adeptly in this scenario. The data visualization really helps data science researchers to better understand the data for further analysis.

Furthermore, we want to gain some insights on variable correlations within the datasets, and a heatmap is again generated to show the series of correlation. This is shown in figure 3 below.

Figure 3: heatmap for correlations between exogenous variables in the datasets

1. MODELING METHODOLOGY

After data processing and visualizations, we arrive at the key part of this study which is training of the models. This is the segment where data science researchers implement their machine learning algorithms and arrive at prediction results. Since I combined the train set and test set into one dataset, I now need to split it back to train set and test set. Now the data is ready to be trained. Linear regression, Random Forest, and Ridge are the three models that are selected and implemented.

* 1. LINEAR REGRESSION

Linear regression, also known as OLS regression, is the least complex and most common regression in statistics and data science. The regression tends to perform well when the data is continuous, and often can render a more accurate prediction than highly complex machine learning models. Under the normal theory for OLS regression model, there are three assumptions that must hold: homoscedasticity, normally distributed error, and no correlation between the errors (Washington and Wolf 1997). In examination of the datasets in this study, the three assumptions are likely to hold. The overall accuracy of the OLS model for the prediction of house sale prices is about 89%. The root mean square error (RMSE) is about 0.0173. This result is promising, considering the complicated nature of the dataset. The reason behind the high level of accuracy would potentially be contributed by the contiguous nature of all the exogenous variables, since we know that OLS model does not perform well when data points are clustered or do not have a linear trend.

* 1. RANDOM FOREST

Random Forest, or tree-based method, has been the most population machine learning algorithms lately in data science research. There are two types of Random Forest methods. The first one is the classification tree, and the second one is the regression tree. Classification tree is used to partition data that is discrete, such as categorical and class data. Regression tree is to partition data that is continuous such as numerical variables (Washington and Wolf 1997). Random forest method is composed of finite number of decision trees, with each node assigned with a probability and error distribution for the occurrence of event. At every node, entropy which measures the homogeneity of the subset data is computed in order to determine the split (Kirasich et al. 2018). Theoretically, Random Forest is capable of handling both discrete and continuous data. However, this method has the tendency to overfit the model and therefore generating a high level of noise (Kirasich et al. 2018). The result from Random Forest on this dataset has an accuracy of 86%, with RMSE score of 0.02197. This result is worse than the prediction by the OLS regression. The suspected reason would be the overfitted random forest model.

* 1. RIDGE

One common problem from the two methods above is the overfitting of the model. Ridge regression properly addresses that problem by shrinkage or regularization. The regression is obtained by adding a penalizing parameter for overfitting to all coefficients except the intercept (Pereira et al. 2016). The model is represented by the function below:

where λ ≥ 0 is the tuning parameter.

To our expectation, the Ridge regression renders a 90% accuracy with 0.0159 RMSE, the highest accuracy rate and the smallest error of all three models. The main explanation is that there is a lot of noise in the housing dataset, and Ridge regression successfully reduces the noise from penalizing overfitting. The prediction from Ridge is therefore the best among all three learning models in this study.

* 1. FURTHER EVALUATION

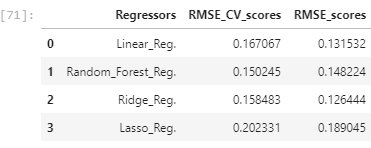
This paper implemented cross validation using “K Fold” method to assess and evaluate the results rendered by each model. The cross-validation scores for RMSE are displayed in table 1 below.

Table 1: Regressors, RMSE cross validation scores and RMSE

Under cross-validation, Random Forest regression has a lower RMSE score than Ridge, but not under normal conditions. This is an interesting case which can be further pursued in detail in future work.

1. ETHICS

Ethics concerns should be a big part of every data science research. Today, any type of data can be sourced or purchased from the internet. The foundation of data science is built on the availability of real-life data from different individuals. These mega data would certainly highlight the concern over individual privacy. It is the accessibility to data that has helped data science to progress in recent years. If practitioners and researchers are limited to access these data in order to protect individual privacy, then data science would certainly not be where it is today. Therefore, the ethics concern for data privacy will always be a double edge sword in the field of data science.

1. CONCLUSION

In order to adequately implement machine learning models to the housing data in Ames, Iowa, one must conduct data processing and cleaning, visualization, and model training. By analyzing the data via visualization, researchers can detect key features about the dataset and have a general sense of which machine learning algorithms are best suited for the data. From the housing dataset from Ames, we noticed that the data shows a continuous nature and that there could potentially have a lot of noise in the data. The results from three different models, OLS, Random Forest, and Ridge, corroborate this finding, as Ridge regression is the best performing model. This implied that the selection of machine learning models does not have to be an ad-hoc process. Researchers should apply more targeted model selection process.

**REFERENCES**

1. Kirasich, Kaitlin, T. Smith and Bivin Sadler. “Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets.” (2018).
2. Pereira, Jose Manuel, Mario Basto, and Amelia Ferreira Silva. “The Logistic Lasso and Ridge Regression in Predicting Corporate Failure.” *Procedia Economics and Finance* 39 (2016): 634–41.
3. Washington, Simon, and Jean Wolf. “Hierarchical Tree-Based Versus Ordinary Least Squares Linear Regression Models: Theory and Example Applied to Trip Generation.” *Transportation Research Record* 1581, no. 1 (January 1997): 82–88.