House Prices: Advanced Regression

Aditya Lahiri and Ching Yung Chong

*Abstract*-These instructions give you basic guidelines for preparing reports in IEEE format.

I. Introduction

In this project, we analyzed a dataset provided by Kaggle [1] as a part of a competition. The dataset provided is based off houses in Ames, Iowa. It consist of 79 explanatory variables describing (almost) every aspect of residential homes in Ames and the goal of the project is to predict the final sales prices of residential homes. The dataset is divided into training and testing sets. The training set contains final sale prices for the homes, whereas, the testing set does not contain any information regarding sale prices. The training and testing set contain 1460 and 1461 instances of data points respectively. These datasets contain missing data and various explanatory variables which are non-numeric in nature. Thus, extensive data preprocessing steps are carried out on both the training and testing set to build a viable model to predict and predict the final house prices. Further information regarding this dataset can be found in the journal paper written by De Cock et.al [2]. The entire analysis of this dataset was done in the R programming language.

II. Measure of Success

Since this dataset is a part of an ongoing competition hosted by Kaggle the measure of success is defined using the metric Root Mean Squared Logarithmic Error (RMSLE). The RMSLE is defined as follow:

In the above equation represents the predicted sale prices of a houses and represents the actual sale price of the same house. The total number of data points is represented by n. The lower the value of RMSLE the more accurate is the prediction of the model.

II. Data Preprocessing

A.. Missing Data

There were several independent features in both the training and testing set where instances of data points were found to be missing. Explain about the way you removed some of the features and then imputation via mice. (add plots)

B. Correlation Analysis

Explain the features removed via correlation. Add the corr heatmap. Add a line on multicollinearity as a reason for removing correlated features.

C. Feature Engineering

The correlation analysis revealed that the feature “First Floor Surface Area” and “Total Basement Surface Area” were highly correlated. Using domain knowledge of house prices, we know that the total square footage or area of houses is an important factor in its pricing. So we created a feature called the “Total Surface Area” by adding the features “First Floor Surface Area”, “Total Basement Surface Area”, “Second Floor Surface Area”, and “Total Ground Floor Living Surface Area”. Since the feature “Total Surface Area” is a deterministic function of the rest of the four features, those features were removed to avoid the effects of multi-collinearity in the final predictions.

D. Skewed Feature Analysis

In the figure below we plot the sale prices in the training set (left) and observe that it resembles a normal distribution skewed to the left. Furthermore we use the RMSLE metric to evaluate the models so we log transform the training sale prices to center them. With log transformation the sale prices appear to be more centered (right).

INSERT FIGURES HERE

The independent features in the dataset were adjusted for skew using the Box Cox transformation. If any of the independent features had a skewness larger than 0.75 they were selected for transformation. The Box Cox transformation is a more generalized version of the log transform and is given by:

In the equation above X represents the independent feature being transformed, T(X) represents the transformed feature, and λ represents the transform parameter. The value of λ was set to 0.15 for our analysis.

E. Categorical Feature Analysis

The dataset had multiple features which were non numeric categorical in nature. Some of these variables were nominal and the rest were ordinal. These features were converted to numeric values using Label Encoder and then converted into dummy variables using One-Hot Encoder. Label Encoder and One-Hot Encoder are commonly used categorical data preprocessing techniques. We used them through the CARET package [4] in the R programming language.

F. Dataset Partition

The testing set was preprocessed the same way as mention in sub-sections a-e and then set aside. After the training set was preprocessed as outlined above, it was divided into two parts. 25% of the training data was designated as validation dataset and the rest 75% of the training set constituted the training set used for model building. In the sections below we refer to this 75% subsection of training dataset as the training set.

III. Modeling

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g.” Try to avoid the stilted expression, “

IV. Parameter Tuning & Model Selection

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g.” Try to avoid the stilted expression, “One of us (R. B. G.) thanks …” Instead, try “R.B.G. thanks …” Put sponsor acknowledgments in the unnumbered footnote on the first page.

V. Results

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g.” Try to avoid the stilted expression, “One of us (R. B. G.) thanks …” Instead, try “R.B.G. thanks …” Put sponsor acknowledgments in the unnumbered footnote on the first page.

VI. Cross Validation & Tuning Timing

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g.” Try to avoid the stilted expression, “One of us (R.

n the first page.

VII. Feature Importance

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g.” Try to avoid the stilted expression, “One of us (R.

n the first page.

VIII. Availability of data and script

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g.” Try to avoid the stilted expression, “One of us (R.

References

[1] Kaggle.com. (2019). House Prices: Advanced Regression Techniques | Kaggle. [online] Available at: https://www.kaggle.com/c/house-prices-advanced-regression-techniques [Accessed 27 Apr. 2019].

[2] De Cock, D. (2011). Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project. Journal of Statistics Education, 19(3).

[3] Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations inR. Journal of Statistical Software, 45(3).

[4] Kuhn, M. (2008). Caret package. Journal of Statistical Software, 28(5)

[5] R. Nicole, “Title of paper with only first word capitalized,” *J. Name Stand. Abbrev.,* in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” *IEEE Transl. J. Magn. Japan*, vol. 2, pp. 740-741, August 1987 [*Digests 9th Annual Conf. Magnetics Japan,* p. 301, 1982].

[7] M. Young, *The Technical Writer’s Handbook.* Mill Valley, CA: University Science, 1989.