Northeastern University

ALY6050 Introduction to Enterprise Analytics

2019 Spring Quarter

Final Project Report

Huiwen Li

Completion Date: June 29, 2019

1. About the Data

In this final project, we have a data set consists of the sales record of house sold in Sammamish WA from 2013 to 2017. There are in total 5031 observations (Training + Test) and 109 features. It will be difficult if we use all the features in model building. Selecting the right subset of features will help improve the prediction accuracy. Therefore, data pre-processing and feature engineering are extremely important.

1. Data Processing

Out of 109 features, we will first use our domain knowledge to remove features that are not important in predicting the Sold\_Price. Features such as “Virtual Tour URL”, “Selling Agent ID”, “Remarks” are not relevant at all in the prediction they will be removed first. Secondly, features that are only available after house is sold (such as “Sold Date”) should not be included in the model building because these information are considered information leakage and they will not be available when we perform the prediction. Thirdly, we need to consider the correlation among different features and avoid selecting two or more that are highly correlated with one another.

There are many missing values and extreme values in the data set. I have used different approaches to handle these values in numerical and categorical variables. For numerical variable, the missing values are imputed by mean of that specific feature since it only consists of house sale data in Sammamish city. As for the extreme values, it is necessary to examine the distribution and if possible identify the cause of such extreme values. For example, in the column “Year\_Built”, we observe an extreme value of “2106”. By intuition, this was caused by data entry error so the correct value should be 2016. Other than this there are many “0” values in the numeric variables “Bathrooms”, “Bedrooms”, “Lot Size”, “Fireplaces”. When we filter out these observations with “0” value, they are mostly 1 story and 2 story houses which means it is impossible that these houses do not have any bathrooms, bedrooms or fireplaces. Those “0” values are due to mistakes hence they are replaced by the column mean.

For each categorical variable, examining the proportion distribution of classes is important to help us decide if re-grouping necessary. At the same time, it will give us a better idea how to group the missing values too. For example, in “House Style” feature, there are 16 classes in the original dataset. When we take a closer look at the frequency distribution of all 16 classes, we realize it is possible to group all three Condo classes “Condo(1 level)”, “Condo(2 levels)”, “Condo(3 levels)” into 1 class “Condo”. Similarly, we can group “Tri-level”, “Split Entry”, “Multi-Level” to a new class. In this way we can reduce the 16 classes to only 6. As for missing values in categorical variable, if the proportion of NAs is small, we can simply group NAs with other low percentage classes to “Others”. Otherwise if NA percentage is high, they should have an individual class called “No Information”. Similar techniques apply to other categorical features too.

1. Optimized Metric

In this project we are performing multinomial linear regression analysis for predicting house sold price which is a continuous target variable. For linear regression, the common metrics used to compare different models are Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) etc. Here we use both RMSE and MAPE for comparing models and the metric chosen to be optimized is RMSE, which is the square root of mean squared error. RMSE is a very commonly used indicator of model performance because of its advantages over other metrics. RMSE takes square of the individual error terms to avoid cancelling out of positive and negative errors, at the same time it assigns heavier weight to extreme values which gives high error. It is adjusted for the degrees of freedom for error through averaging all error terms (sample size minus number of model coefficients). By taking square root of the mean square error, it removes the unit and makes the comparison easier. The objective is to minimize RMSE, while we still look at MAPE which is expressed in generic percentage terms.

1. Summary of Findings

|  |  |  |
| --- | --- | --- |
| Models | RMSE | MAPE |
| fit\_lm1 (with List\_Price) | 26253 | 4.67% |
| fit\_lm2 (w/o List\_Price) | 85544 | 13.6% |
| fit\_lasso (with LASSO regularization) | 25121 | 4.51% |

Based on the results shown in the above table, the model with LASSO regularization has the lowest RMSE and its mean absolute percentage error is 4.51%. If time permits, we could explore other features and include them in the model building to check if that will improve RMSE and MAPE.