HarvardX PH125.9x

Kaggle Housing Prices Project Submission

Chelsea Gorius

Contents

March X, 2019

1. Overview

1.1. Introduction

This project aims to use some of the machine learning and data analyzation skills taught in the HarvardX PH125.9x course. The project includes data analysis followed by manipulation. Regression models are then used to train machine learning algorithms to predict the scenario's output variable. The final model decision has been supported based on a set of compared statistical analysis values.

-The Root Mean Squared Error is the standard deviation of the prediction errors. This is a goodness of fit test meaning it represents the regression model's average deviation from the data it was modeled after.

-The R squared is a scaled representation of RMSE. Meaning it to is only a goodness of fit test, representing how well the regression model fits the data it was modeled after.

-The Average Validation Error is the mean of the residuals between the predicted output values and the actual observed output values. This metric observes a regression models' ability to predict a value from an observation that was not used to create the model.

-The Accuracy value is a relative metric to demonstrate a regression models' ability to make predictions of an output value based on an observation not used to create the model. In this case it an be considered a scaled measurement of the average validation error.

After developing the optimal variable list and comparing the outlined statistical metrics a preferred regression model will be selected based on its ability to accurately make a prediction.

1.2. Project Description

The aim of this project is to develop a list of chosen independent variables from the Housing training data provided by Kaggle in order to produce a regression model with a high accuracy and a low validation error. The regression model will use the Housing data's independent variables in order to predict a house's Sale Price. The project aims to select a maximum of 20 from the original 79 independent variables in order to predict Sale Price. Data cleansing and manipulation are required considering the numerous independent variables of containing both numerical and categorical values. Once a set of optimal independent variables are determined, various regression models will be trained and tested in order to determine the most accurate when determining a houses' Sale Price.

1.3. Dataset

The Housing dataset provided by Kaggle includes both a training set, *train.csv*, and a test set, *test.csv* . The training set contains 1460 observations of 80 variables including the output variable for this project, the house sale price. The test set however does not include the house sale price. In this project since accuracy and validation error are being used as primary factors for creating the model, we will need to test it against data that already has the sale price listed. Therefore in this project only the training set is used. More information on the data manipulation is provided below. The training data set contains 43 categorical variables, 36 numerical variables, and the output value of sale price.

2. Method and Analysis

2.1. Data Analysis and Manipulation

In this project only the training dataset provided by Kaggle is used. This project divides the training dataset into a sub training set containing 80% of the data and a sub validation set containing 20% of the data. This is done in order to calculate the regression model's accuracy and validation error. This cannot be done with the original test dataset provided by Kaggle because it does not include the sale price of the observations so there would be no data to calculate the accuracy of the model's predictions with.

After some data exploration it is clear that not all the data is filled within the data set. Using the missForest package a randomForest process was run repeatedly in order to determine the missing observation values and predict the NA values.

In order to evaluate both numerical and categorical variables simultaneously the process of one hot encoding was used to alter the categorical variables into a numerical form that can be analyzed simultaneously with the original numerical variables. Encoding the categorical variables into numerical changes the data set from 79 independent variables to 288. Meaning that the original 43 categorical variables were expanded into 252 numerical variables. From each original categorical variable a new variable is create with the category option type and is labeled as followed; *"Category.Option"*.

After the optimal set of independent variables was determined, containing 17 specifically, an optimal sub training set was created from the sub training set of both categorical and numerical variables. Therefore the one hot encoding process was performed again to expand the categorical variables in the optimal set in order to develop the regression model using all the variables simultaneously.

2.2. Modeling Approaches

After all data was transformed into numerical variables the following training models were used on the sub training set; glm, lm, knn, kknn, svmRadial, svmRadialCost, svmRadialSigma, and svmLinear.

After training each model method the top 50 most important or influential variables are listed. Recall that the categorical variables were expanded into each individual possibility existing as a binary. It is because of this that when the names of top 50 variables from each model were recorded the categorical variable names were altered to represent the entire variable and not just that option of that category. Recall that after the one hot encoding process the categorical variables were expanded into specific *Category.Option* form. By removing the part of the name after and including the period the model will maintain the effects of the whole category.

After determining the similar variables between the top 50 important variables for each model used an optimal list was determined. After which a new data set was created including only the optimally selected independent variables. The dataset includes the original variables, categorical and numerical. The same one hot encoding process was used on the optimal set's categorical variables. The optimal data set was then used to train the final model. Three final models were designed using the optimal variables dataset. Each model was analyzed and compared to the validation set to determine the one with the greatest accuracy and lowest validation error. It should be noted that the sub validation set contains all 288 variables, numerical and expanded categorical.

3. Results

After performing each of the 8 regression model methods on the sub training set of data the following list of 17 independent variables was developed based on variable importance in each of the models. An optimal data set was then created by developing a subset including only the optimal variables from the original sub training set of data. Since 8 of the 17 variables are categorical, the one hot encoding process was used again to convert them into numerical values.

```
> top_list
 [1] "X2ndFlrSF"    "X1stFlrSF"    "KitchenQual"  "BsmtFinSF1"   "BsmtExposure" "OverallQual"
 [7] "LotArea"      "YearBuilt"    "MasVnrArea"   "MasVnrType"   "BsmtUnfSF"    "BsmtQual"
[13] "Alley"        "WoodDeckSF"   "ExterQual"    "Neighborhood" "GarageType"
```

The three best models from the first set of regression models were; glm, lm, and svmLinear. After retraining each model with the optimal data set, each was used to predict the Sale Price of the sub validation set of data. Below the RMSE, R squared, Mean validation error, and Accuracy displayed.

```
> opt_svmL$bestTune
    C
8 0.7
```

```
> opt_res
# A tibble: 34 x 5
   method      RMSE     Rsq val_er   Acc
   <chr>      <dbl>   <dbl>  <dbl> <dbl>
 1 glm       31960.   0.838 19349. 0.895
 2 lm        34790.   0.814 19349. 0.895
 3 svmLinear 33279.   0.833 18453. 0.900
 4 svmLinear   NaN  NaN     18450. 0.900
 5 svmLinear 31125.   0.847 18450. 0.900
 6 svmLinear 31130.   0.847 18450. 0.900
 7 svmLinear 31125.   0.847 18450. 0.900
 8 svmLinear 31123.   0.847 18450. 0.900
```

It can be seen that the svmLinear regression model contains the best of each defined statistical metric at the C = 0.7 level. This means the model not only had minimal residuals within between the data used to produce it but also with the predicted variables it produced for comparison to the sub validation set. The svmLinear final model is depicted below.

```
> opt_svmL$finalModel
Support Vector Machine object of class "ksvm"

SV type: eps-svr  (regression)
 parameter : epsilon = 0.1  cost C = 0.7

Linear (vanilla) kernel function.

Number of Support Vectors : 981

Objective Function Value : -142.8493
Training error : 0.160049
```

4. Conclusion

Since the one hot encoding process was used on optimal variable set, it's categorical variables were expanded. This explains the numerous coefficients for the regression model. Further observations into the coefficients reveals the most influential independent variables for predicting the Sale Price. Though most of the statistical metrics were close between the regression models in the optimal results table, the svmLinear model was selected as the best for predicting house Sale Price using limited variables because of its' superior accuracy in predictions. The purpose of the project was to develop the best model for estimating house Sale Prices with limited variables. The svmLinear model did so the best in this scenario.