ETC5242 - Report

13th October 2016

Purpose

The purpose of this report is to analyse the relationships of various factors against the housing price. Our task on hand here is to make the best model to predict the sale price of houses, with the least errors. This report will outline the progress and decisions that we made to reach the best model that we could make.

Introduction

We have conducted some research on factors that affect the housing market in the U.S. in order to have some knowledge of the relationship between the variables and the house prices. Several variables (LotArea, HouseStyle, TtlBsmtSF) are related to the size of the property and we expect that they have a positive relationship with SalePrice. Furthermore, variables relating to timing (YrBuilt, YrSold YYYY, MoSold MM, SaleDate) have important impact on SalePrice due to factors such as condition of the house, inflation rate, materials used, etc. Generally, we would expect a newer house to have a higher price.

Method and Progress

Data Sorting:

We found some extreme values among the sale prices. We discovered some of the sale prices are under 10000, and consequently we deleted 7 variables. However, the train data still consists of a lot of errors. For example, one of the observation is 0 for LotArea, and also two observations might have extremely different sale prices although their other variables are same. We did not delete any of these observations as we want to have more observations to predict the model.

Data Adjustment:

We have decided to make some changes and create new variables from the variables given. We created a new variable "houseyear" to represent the age of the house, and a new variable denoted as the day of sales from "SaleDate", we log the GarageArea, and TtlBsmtSF, factor YrSold YYYY, MoSold MM and SaleDate as they are recognised as numeric. Also we cubic TotRmsAbvGrd, Cars, and squared NmbrBRs.

Relationship between variables:

In order to find out the distribution of SalePrice, a histogram of SalePrice was generated. In the histogram, it is found that SalePrice is not evenly distributed, and therefore we chose to use log(SalePrice) as y in our future predictions. Then, to find out the relationship between variables and log(SalePrice), we generated a series of ggplots with x(variables) and y(log(SalePrice)), in order to determine what variables should be included in the model.

It is logical to assume that there must be some dependence between variables. Therefore based on some research information, we generated a series of graphs between variables to observe their dependence.

Model Selection:

We decided to divide into 4 models as we observed a linear relationship between log(SalePrice), log(GarageArea) and log(TtlBsmtSF). All the single variables are included, even though some of them are not very significant. We added some interaction terms and are found to have dependence between variables. The models are split depend on whether observations have garage or basement, and we obtained a score of 36000 when we first started with glm. We decided to change the model and use the function "randomforest", and it did improve our score.

- Model 1 has garage and basement, therefore we include all the variables above. **Formula:**

*randomForest(log(SalePrice) ~ Neighborhood + LotArea + houseyear + HouseStyle + Foundation + RoofMatl + Ext1 + Heating + Central.Air + TtlBsmtSF + TotRmsAbvGrd + GarageType+ GarageArea+ YrSold_YYYY*MoSold_MM + Cars + NmbrBRs + YrBuilt*TtlBsmtSF + YrBuilt*HouseStyle + YrBuilt*Foundation + YrBuilt*GarageArea + YrBuilt*LotArea + LotArea*TotRmsAbvGrd + LotArea*TtlBsmtSF + YrBuiltExt1 + TotRmsAbvGrd*GarageArea+YrBuilt*Central.Air*

- Model 2 does not have basement (TtlBsmtSF = 0), we deleted all the variables involve with TtlBsmtSF.
- Model 3 does not have a garage(GarageArea = 0), we deleted all the variables involve with GarageArea.
- Model 4 has neither basement nor garage (TtlBsmtSF and GarageArea = 0), we deleted all the variables involve with those two variables.

Discussion and Conclusion:

Interesting point:

We observed that the predicted values are commonly lower than the SalePrice values in the train data. We also found that YrBuilt does not influence SalePrice by itself, but it can influence SalePrice by influencing other variables that are related to SalePrice. Furthermore, throughout many trials on Kaggle, we found that changes in SalePrice equation could vary the scores but it does not necessary improve the score significantly. The way to improve the error is to design a better model(glm vs randomforest).

Conclusion:

We obtained a final score 29971 for our model of prediction and close to the top score compared to other group, it illustrates that our model have improved and included the most important point but still there are some variables that we are missing.The model might be improved by performing a cookdi test on glm of our model, to delete observations that are high influential point, then use the new dataset to re-do randomforest model.

Appendix

Some graphs for Relationship between Variables