Kaggle Competition: House Price - Project Report

Team Golden Squirrels: Lingjie Qiao, Minsu Kim, Kevin Liao, Cheng Peng

December 3, 2016

Abstract

This paper summarizes the background, problem, methodology and results of our team's final project in the course Stats 159 Reproducible and Collaborative Statistical Data Science. To make full use of statistical models and predictive tools we have learned from the class and challenge ourselves to the next level, our team choose to complete **"House Price: Advanced Regression Techniques"** from Kaggle Competition and enter the competition with our work.

Competition Link: https://www.kaggle.com/c/housepricesadvancedregression techniques

The goal of this project is to present the use of predictive modeling process and utilize software tools that effectively communicate the results. While the competition only emphasizes the accuracy of predicted values, our team at the same time are dedicated to maintain project reproducibility and provide both objective and personal reflections upon regression analysis.

1 Introduction

The House Price project thoroughly explores the predictive modeling process and advanced regression techniques. From previous study, in order to understand the relationship of one dependent variable with several independent variables, we fit a multiple linear regession with Ordinary Least Squares. However, since OLS may have high variance and include irrelevant variables, Predictive Modeling Process can improve the results in terms of ***Prediction Accuracy*** and ***Model Interpretability***.

The competition sets the background of the project: Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition requires participants to predict the final price of each home. Our team therefore follows the idea of model prediction and tries to use different techniques in order to most accurately predict the final sales price of each house.

2 Data

The datasets are obtained from the Kaggle Competition website. We have access to four files:
data description, which gives the official definition for fields
train.csv, which provides 1459 real observations that can be used for model construction
test.csv, which is used to fit the predictive model and create submission entry for the final sales price of 1460 observations
sample submission, which gives an example of how the fitted values should be submitted.

The train dataset has in total 80 variables, 79 potential predictors and 1 dependent variable called *SalePrice*. We observe both categorial predictors, such as *FireplaceQu*, *GarageCond* and *MasVnrType* as well as numerical predictors, such as *PoolArea*, *EnclosedPorch* and *YrSold*. Since we can potentially create a lot of different new variables, our goal is to understand the relationship between *SalePrice* and these predictors with statistical fitting procedures that minimizes Mean Square Error.

3 Methodology

4 Analysis

With the least square estimators, we compute the regression coefficients.

Table 1: Information about Regression Coefficients

Coefficients	Estimate	Std. Error	t-statistics	Pr Value
Intercept	7.0325	0.4578	15.36	<0.00
TV	0.0475	0.0027	17.67	<0.00

Here is the scatterplot

More information about the least squares model is given in the table below:

Table 2: Regression Quality Indices

Quantity	Value
Residual Standard Error	3.259
R-squared	0.612
F-statistic	312.14

5 Results

6 Conclusion

In conclusions, while participating in the Kaggle competition - ”House Price: Advanced Regression Techniques”, we get a chance to dig down into real-life dataset that have complicated variables, missing values and large number of data entries. Based our analysis on the train dataset, we try to find the regression model and advanced techniques that most accurately predict the housing price, and use the exact model to predict new entries from test file. Exploring and creating visualization of data, we compare the usage of different regression models to understand the relationship between dependent variable *SalePrice* and 79 potential predictors. Fortunately, we ranked No.10 among all 2000 participating teams with our prediction, which not only shows the precision of our own predictive model, but also suggests room of improvement and growth.

This course gives us great opportunity to practice **project reproducibility**, **advanced statistical model learning** and **teamwork delegation**. By deligating work to each team mate and implementing the model to generate results, we aim to maintain project reproducibility and publicity so that more people can learn from our model and benefit for future researches. *If you have any questions or concerns regarding any part of our analysis, please feel free to contact the authors for further clarification.*