USA HOUSE DATA

Build a Machine Learning Regression Model and Hypothesis testing from king county USA house sales data

JUNE 3, 2018

SYED MEHEDI HASAN
Student Id: 480255897, Uni key: shas5428

Setup

Data set:

The data was collected from Kaggle https://www.kaggle.com/harlfoxem/housesalesprediction. It has only one csv file showing 21613 houses sold in king county, USA between May 2014 and May 2015 including house description, location and sale price(US$). The data comes in clean csv format I just load it to Jupyter notebook using pandas.

Our goal is to build a Machine learning model to predict the price of any unseen house. First, we have built a regression model using Multiple Linear regression method (Ordinary Least Squares Method) from the training data, and used the test data to test the accuracy and standard error of the model. We calculated its accuracy and standard error. Then we did the same work using Random Forest Regression method and compare its accuracy and standard error with the previous model. The method that gave best accuracy and less standard error used as our final model.

Hypotheses testing and the Reliability:

Problem: Build a regression model to predict house price from given set of house sales data. Compare this model with other related model about its accuracy and error.
Research Question: Is Random Forest is better than Multiple Linear Regression?
Null Hypothesis(H0): No**,** both method will give same standard error.
Alternative Hypothesis (H1): Both method will not give same standard error i.e. error means are different.
Can we reject the null hypothesis using the Kruskall-Wallis H-test?
To qualify the hypothesis, we used the standard error from Multiple Linear regression and Random Forest regression model using K-fold (K=20) cross-validation and used the significance level of 0.01, with Kruskall-Wallis H-test. While we were using python program, we got p-value=0.009023 from the H-test, so we can reject null hypothesis, and alternative hypothesis stand here.

Evaluation:

As it is a regression problem, we calculated R-squared to get accuracy of the model. We got average r-squared .6973 using Multiple Linear Regression (ordinary least squares method), so model is 69.73% accurate. But when we used Random forest regression, we got R-squared .8613, so here it is 86.13% accurate.

Approach

Machine learning model:

USA HOUSE DATA

After analysing the data, we have seen that we are able build supervised learning method, so first we have used Multiple Linear Regression method (OLS) to build the model to predict house price using the features provided with the data. Then we did the same job using Random forest regression method.

To do that we shall find out correlation between different features with the price we need to predict. Draw scatter plots of all important features we got from correlation matrix. In our Multiple Linear regression model our target variable is price column, from correlation matrix we can see that some features have very poor correlation with price, that means if we remove those from our feature matrix the result is not affecting significantly, so we can remove those from our features list.

To get good accuracy and avoid overfitting we have used K-fold cross validation. We have divided the data into K subsets. Then using each subset k= {1..K}, we take k^{th} (k=1..K) set as test set and remaining (K-1) sets as training set. Use training set to train the model and use test set to check accuracy of that model. In each iteration we calculated coefficients of features, R-squared and standard error to choose the best model. Then we get coefficients of the best one to build our final model.

Our proposed machine learning model flowchart is given as followed:

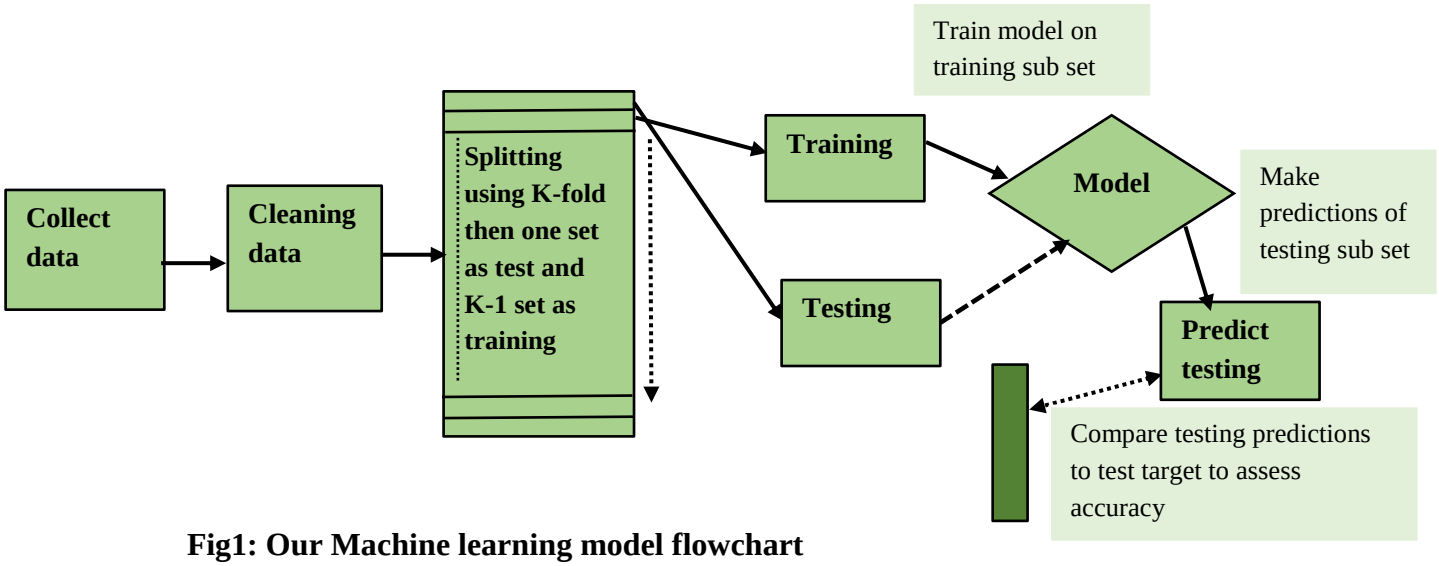


Fig1: Our Machine learning model flowchart

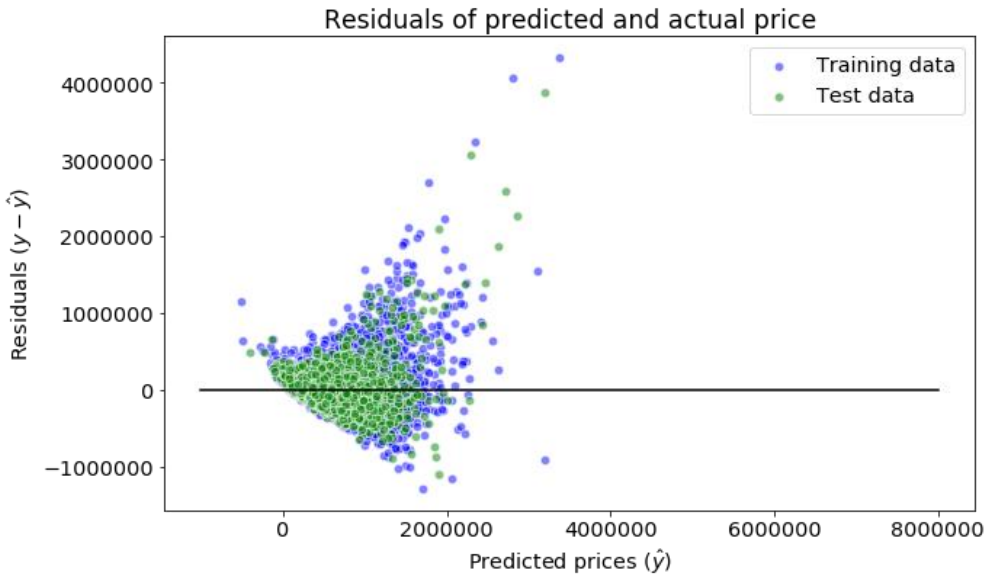
Results

Result presentation: Below is the result achieved applying Multiple Linear regression and Random Forest regression method on our house sale data:

Method	R-squared	Standard error	Mean absolute error
Multiple Linear regression (OLS)	.6973	200969.73	126304.57
Random Forest regression	0.8613	136325.69	73972.09

USA HOUSE DATA

We plotted residual of predicted prices and actual prices to check how our model is working as below.



Results Analysis:

Using our model, we predicted the test data, then we calculated different types of error including standard error and mean absolute error. Using Multiple linear regression, we got standard error 200969.73 and mean absolute error 126304.57 when using random forest regression, we got average standard error 136325.69 and average mean absolute error 73972.09. The results we got here is good enough to apply in any real-world problem. But mathematically still there has some notable gap between predicted and actual price. Predicting accurate house price is really a tough problem, because in many cases, price does not follow any specific rule. It depends on buyers' special choice to any specific area or specific design, his financial condition and in auction when two or more buyers try to beat one another. A complex model like neural network can give better accuracy, if there is any non-linearity exists in this data set.

Conclusion:

We have learned a lot from this project ranging from choosing suitable dataset, cleaning and analysing data, find out the business problem and how to predict from data as a data scientist what we supposed to do.

We build model using two types ML method compare those results according to model accuracy and standard error, so that we can choose best model according to business needs. It will be very helpful to solve business problem in real-life industries, and give us confidence about data driven problem solving.

References:

1. Data https://www.kaggle.com/harlfoxem/housesalesprediction.
2. Data Science from Scratch - O'Reilly Media.

Appendix:

1. Python code of the project in jupyter note book.