CSE 519 -- Data Science (Fall 2017)
Homework 2: Exploratory Data Analysis in IPython

Name : Hemant Kumar Pandey
ID : 110828730

Report

27th September 2017

OVERVIEW

This homework is based on Kaggle challenge which involves a dataset by Zillow and revolves around predicting the price that a particular real estate property (usually a home) will sell for. The goal is to explore the data and uncover interesting observations about the Zillow data, model the data and submitting the results on Kaggle.

DATA PREPROCESSING

1. Dropping all the columns which have string values and does not contribute much to the correlation.
2. Conversion of the data types of all columns from object to numeric so that they can be manipulated faster and gives a better correlation.
3. Filling all the NAN values with the mean of respective column.
4. Standard scalar in the improvised model.

RANDOM FOREST REGRESSOR MODEL

How it works ?

It can be used for both classification and regression kind of problems. As the name suggest, this algorithm creates the forest with a number of trees. In general, the more rees in the forest the more robust the forest will be. In the same way in the random forest classifier, the higher the number of trees in the forest, we get more accurate results.

Random Forest is based on the decision tree concept which is a rule based system. If a training dataset is given with features and targets, the decision tree algorithm will come up with a set of rules. These rules can be used to make prediction on test dataset.

Suppose we want to see if an apartment will be sole for more than 200,000$ or not ? To model the decision tree you will use the training dataset like the list of apartments in the nearby regions that were sold with the comparative price in the past.

So once you pass the dataset with the target as the property will be sold for more than 200,000$ or not to the decision tree classifier. The decision tree will start building the rules with the apartments that were sold for good price as **nodes** and the targets which are not as the leaf nodes. By considering the path from the root node to the leaf node, we can get the rules. The simple rule could be if some apartment has more than 2 bedrooms and a pool, it will be sold for more than 200,000$. We just need to check the rules which are created by the decision tree to predict the result.

In random forest algorithm, Instead of using information gain or gini index for calculating the root node, the process of finding the root node and splitting the feature nodes will happen randomly.

How well it works ?

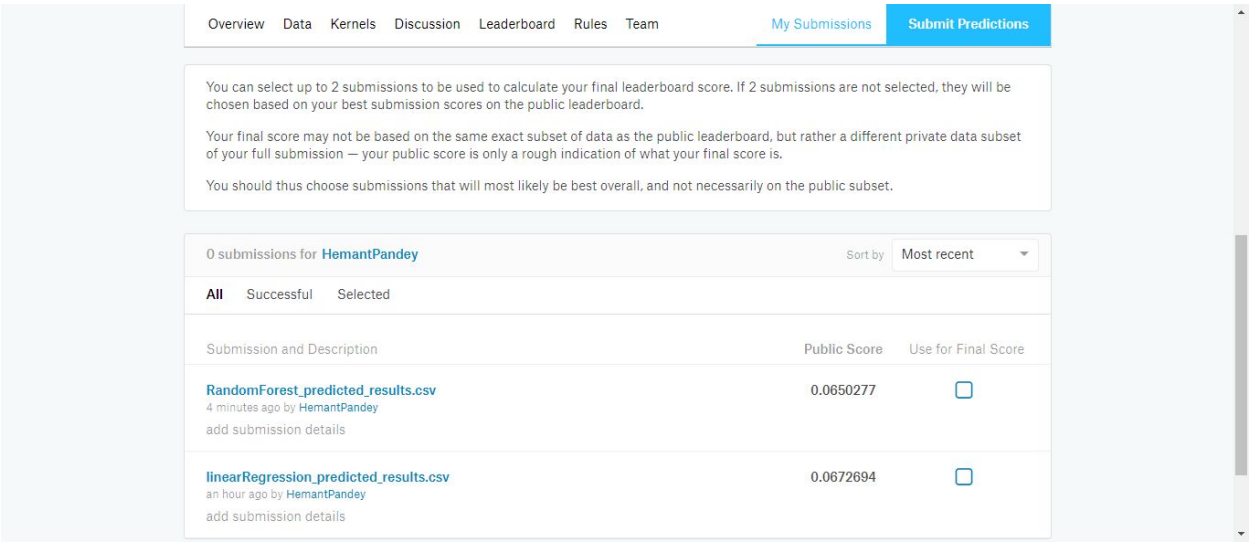
- Handles the missing values.
- When there are more trees in forest, it would not overfit the model.
- Can be used for both classification and regression tasks.

INTERESTING THINGS I LEARNED

- First experience with modelling and Jupyter.
- How selecting different variables can change the result which can vary much.
- The power of pandas and sklearn. It was very easy to use.

KAGGLE SCORE FROM BOTH MODELS

Score from both models



Rank On Kaggle