**BigMart Sales Prediction**

Predicting sales volume for products with Regression

I worked on this project during the Analytics Vidhya DataHack Hour. The test/train data is provided by them and the goal was learning.

Step 1: What is the problem?

I want to find what drives the sales amount for a certain product in different stores and try to predict where and how I can maximize the sales for this particular product. The task is to predict the sales of a certain product at a particular store, part of a chain of stores and find out what influences that sale. I have access to 2013 collected data, for 1559 products across 10 stores in different cities. We will evaluate the model for the predictive accuracy using Root Mean Square Error. Assumptions:

* we are using only grocery type products, with few features; this is a small model
* the category of the product might have an impact on sales (like dairy sells more than canned food beacause is used more often)
* the type of sore and it’s location is important for sales
* the size of the store might be important (people go to big stores to shop all they need at once)

Step 2: Why does the the problem need to be solved?

I’m building this model for my own learning purposes. It should provide a good insight in what drives the sales for a grocery product. This is an easily scalable model to provide detailed info and accurate predictions for sales volume for different type of products as there is a lot of data out there. This solution can be used for projects, start-ups and sales forecast.

Step 3: How would I solve the problem?

I would find the sales data for a product as detailed as possible (with as many features as possible). Select all the features with no NaN or missing data. Select the obviously important features for the model. All the other put them aside as we will be experimenting with them. Visualize the data (read through it and build some scatter, history plots for linearity and dimensionality and box plots for outliers). Build the model with different algorithms starting with the simplest and moving up to more complicated. Evaluate the performance of each algorithm. Try combining 2-3 of them and evaluate the new performance. Choose the best model and deploy it on all the test data you can find.

What i did:

1. Replaced the Nans, identified outliers, feature selection and normalization - for both train and test data.
2. Built the regression models: linear and decision tree. Predicted the sales, cross validated the scores, calculated the R^2 (coefficient of determination - better when using decision tree regression)
3. Classified the train data with a decision tree and a random forest and calculated the accuracy score and the R^2. (the clear winner is the decision tree classifier)