1 A Comparative Study of Big Mart Sales Prediction

The paper presents a model which employs Xgboost technique for predicting the sales of a company like Big Mart and found that the model produces better performance as compared to existing models. The study uses the 2013 Sales data of Big Mart as dataset. The latter consists of 8523 products, and contains 12 attributes among which $ItemOutletSales$ is the response variable. The dataset is based on hypothesis of store level and product level and is split into training set, and test set with a ratio of 80:20.

1.1 Data preparation and Model building

The first stage of the research is data exploration. This phase aims at extracting useful information about the data and fix the problems. Variables OutletSize and ItemWeight present missing values, and the minimum value of ItemVisibility appears to be zero which is unreasonable. The age of each Outlet is expressed with the year of establishment, which is irrelevant. Moreover, some of the variables are misspelled. A log operation is then performed on ItemOutletSales, since it is positively skewed. For the missing values data cleaning was implemented: Outlet Size has been replaced by the mode of the attribute, ItemWeight with its mean. For the model, it is assumed that there is no relationship between the measured attribute and imputed attribute. In addition to missing values, also some nuances were founded and solved through the featuring engineering phase. Since ItemVisibility attribute had a zero value (which is unreasonable), its mean value has been used to overcome the issue. All categorical attributes discrepancies were resolved by modifying those variables into appropriate ones. A third category "none" was created for ItemFat when non-consumables and fat content property were not specified. A new attribute ItemTypeNew with three categories was created: Foods, Drinks and Non-consumables to identify these products. One additional attribute Year was included in the dataset to determine how old a specific outlet is. Consequently, the dataset is ready for the building of the model. The proposed model was created using Xgboost technique, with the aim of forecasting sales of Big Mart, and it was then compared with other machine learning techniques like Linear regression, Decision tree etc… Xgboost has the following useful features: sparse aware (missing data automatically handled), supports parallelism of tree construction, and continued training so that the fitted model can further boost new data. All models received features as inputs, which are then segregated into training and test set. The test dataset is used for sales prediction whereas, the trained model is used to predict the future sales.

1.1.1 Implementation and Results

In the cross-validation stage, the dataset was divided randomly into 20 subsets with roughly equal sizes. Models are first trained by using 19 subsets as training data, and then used to predict accuracy by using the remaining subset as test data. This process continues until each subset is tested once. Data visualization showed that the lowest sales were produced in the smallest locations. In some cases, medium size locations produced higher sales though it was type-3 super market compared to larger size locations. Indeed, more locations should be switched to type 3 in order to increase sales.

ItemOutletSales was found to be highly correlated with ItemMRP, and affected by ItemType. Highest sales are made by OUT027 which is a medium size outlet in the super market type-3. Less visible products are sold more compared to the higher visibility products, which seems impossible, since outlet contains daily used items. Thus, the null hypothesis H0 “the visibility does not effect the sales” was rejected. Also, high size outlets are less than medium size ones in terms of count.

1.1.2 Conclusion

Accuracy in the predictions of sales is crucial for a company in order to not suffer losses or to minimize them. Experiments support that the used Xgboost technique is more accurate compared to decision trees, ridge regression, etc… It is also concluded that this model has the lowest MAE and RMSE and therefore performs better compared to other existing models.