**Rossmann Sales Prediction Report**

*Name: Girish Mallya, Babagouda Patil, Ganesh Gajakosh, Ganesh Dubey*

*Roll No: 62,37,59,58*

*KLETECH*

*October 22, 2017*

**Abstract**

There are many machine learning models which can be used to monitor, analyze and predict the sales of a particular network of shops. The particular dataset in our interest contains the data of hundred thousand shops across Europe containing the records for sales, customers and other attributes for that network of shops. The data which is a raw data after performing series of preprocessing techniques to get a clean data which contained the relevant and sound attributes, after extracting the features which were relevant to our problem from the sample data we built a linear regression model and also used a python module ‘statsmodels’ . In addition, we tried to boost the model using xgboost in order to obtain a better prediction score. The best accuracy we obtained is 0.4.

**Introduction**

Prediction of sales is very important for inventory management in the retail industry. Of course, the stores can use accurate prediction to maintain their shops in good business while also satisfying the needs of the customers therefore keeping a good margin of profit. However, there are many small factors like holidays, promos and opening of competitors which can effect significantly on the sales of the store on any given day. Hence, we seek to study the impact of such factors with the help of Data mining and machine learning, and predict the sales of the store based on all these factors.

The data used in this particular study is from the Kaggle competition Rossmann Store Sales – Forecast the sales using store, promo and competitor data ([https://www.kaggle.com/c/rossmann-store-sales](https://www.kaggle.com/c/rossmann-store-sales/data)). Rossmann is a major drug store chain with over 3,000 stores across Europe, including stores in Poland, Hungary, Czech Republic, Albania and Turkey. The data we are provided contains record of 1115 stores located across Germany including the sales of record for each store over a course of period giving use a total of about 1 million data points, the data contains 15 attributes, including both Boolean and continues variables. There are 3 sets of data. Train dataset containing the historical data including sales, Test dataset containing the historical data excluding the sales and Store dataset containing the supplemental data about the stores.

The aim of our study is to use the various prediction models and techniques to accurately determine the sales taking into the consideration the various factors mentioned above which can affect the business of the store. And give relation between these entities and the sales of the store. We are using a linear regression model for this purpose; linear regression is a very powerful statistical technique and can be sued to generate insights on consumer behavior, understanding business and factors influencing profitability. It can be used to make forecasts.

Using this model we found that the sales variation is depended very much on the factor such as mentioned above, we believe that we can obtain the near accurate sales using this model by boosting it.