# ***Sales Prediction in E-Commerce Sites***

Abhishek Thakur (120050008)
Syamantak Naskar (120050016)
Rajesh Roshan Behera (120050079)
B Soma Naik (120050080)

Guide : P*rof. Ganesh Ramakrishnan*

*Indian Institute of Technology Bombay*

## ***Proposed Idea :***

The problem we would like to address through this project is that predict the amount of sales of a product on given day and in given time interval using the previous sales data for that particular product.

## ***Data and Description:***

The data on previous sales is taken from machine learning repository of UCI. The link to data set is **http://archive.ics.uci.edu/ml/datasets/Online+Retail**. The data contains *product description, quantity, price, product id, transaction id* and *date and time.*The data set contain more than 500,000 data points.

## ***Initial Attempt :***

We considered Bag of words(BOW) which contaned all product name, all possible colors, all possible size, all possible brand name for all products. In this case our feature vector was *price, quantity, month, day of week, time interval and BOW.* This resulted in the dataset becoming too large to handle and it also contained a lot of redundant data as some of features of one product does not affect the sales of other products.So, we therefore decided to curb on the number of products and take the most important ones and train the model(s) independently for each product.

## ***Final Attempt:***

### ***Preprocessing and Product extraction:***

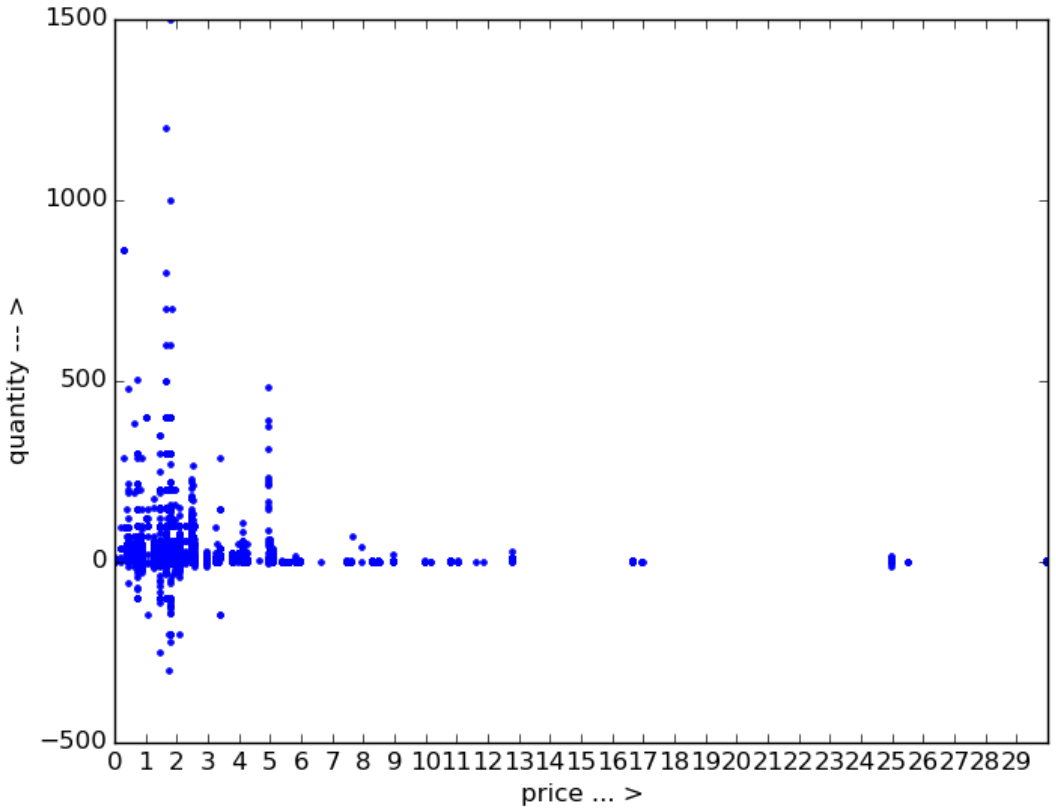
The data contains lot of products with small number of data points. So, in order to reduce the size we have considered only those products with at least 1000 data points.We mapped each product description with its frequency and omitted all those products with data points less than 1000. After getting the product description of the data the product name is extracted manually.There are only 18 products we have considered and data for these products is seperated from entire data using key word search in the product description i.e search for product name in the product description of each data points and group them.The data for each particular product is divide into 20% test and 80% training data.The division of data is done randomly.

### ***Features and their extraction:***

The feature vector for each product contains certain features which are common among all the products of interest which are *price, day of week([0,...,6]), month([1,..,12]), cancellation of order* and *time interval([0,...,23])* and features based on the properies of the product like, size, color, brand, type and etc. The features based on the properties are basically bag of words(BOW). These BOW are extracted by considering all the words in the product description after removing words like, *the, is, as, of, an, and,* and product name.The features *day of week, month, time interval* are nominal features.

***Hypothesis:***

The dependency of dependent variable *quantity* on various independent variable *price* is decides by considering various functions. The functions we considered are *quanity α price* , *quantity α* $e^{-price}$ , *quantity α* $e^{-price}$ + *price* , *quantity α* $price^{-k}$ and *quantity α* $price^{k}$ + $price^{n}$. We have considered all these functions beacuse from *price vs quantity* graph it is clear that *quantity* does not linearly dependent on *price.* Dependency of *quantity* over other features is linear as these are nominal.



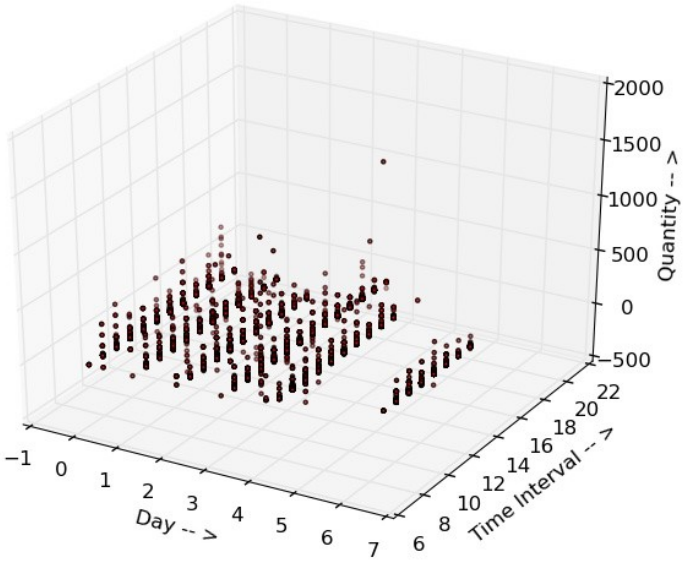

*price* VS *quantity*

***Models and Analysis:***

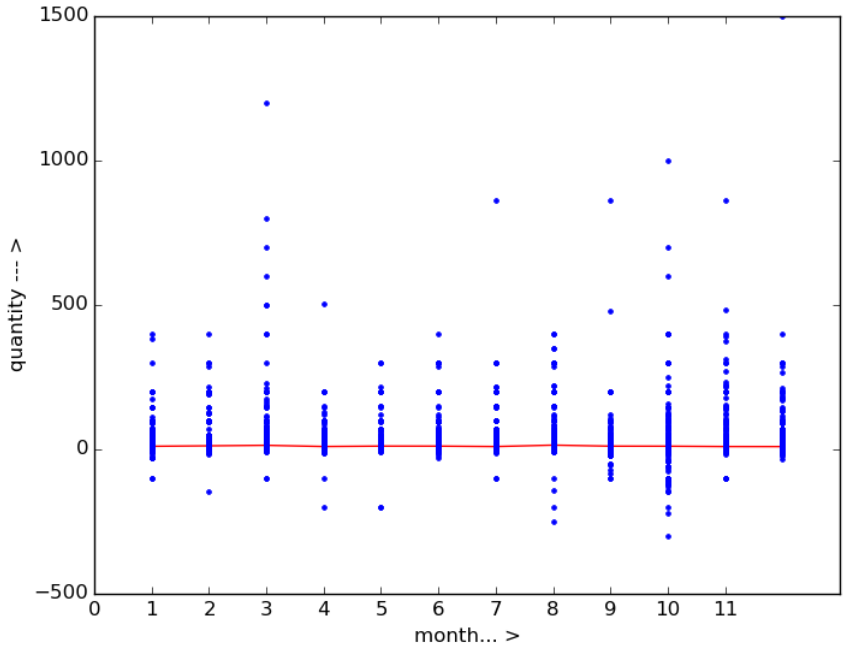
We considered ***Linear Ridge Regression(LRR)*** and ***SVR*** for our prediction. When *quantity α price* the difference in **root-mean-square error (RMSE)** for both the models was not very different. As we changed the dependecy of *quqntity* over *price* from linear to non-linear the RMSE for SVR is mush less than RMSE for LRR and better prediction.So SVR is choosen for our prediction and *quantity* is non-linearly dependent on *price.*

*The other graphs used to get some intuition on the distribution of data points*

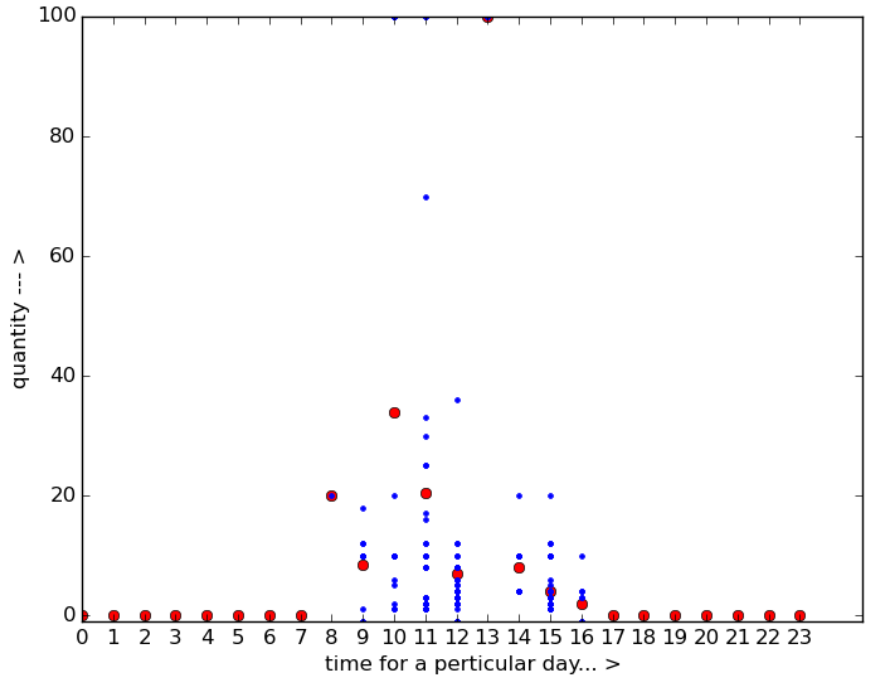
The following graphs are for product bag




*day of week* VS *time interval* VS *quantity*




*month* VS *quantity*




*time of interval(on particular day)* VS *quantity*

## ***Future Work :***

We can improve this model more if we have better training data. If we have more fields like ratings, reviews, competitive brands we can improve this model further. This would help in better prediction of the data. In future, we can automate the detection of products using NLP and figuring out what that particular product is. In case of a new product, we can categorise that new product into one of the 18 existing main products by comparing which features match that product the most or create a seperate entity for that product if possible.

## ***Conclusion:***

This project helps to have better idea about the number of sales likely to happen in the near future on a given day in given time interval. This helps the e-commerce sites to plan to stock their inventory.