Prediction of Drug Sales by Using Neural Network Algorithm

Ashish Kumari[1],Navdeep Bohra[2]
{ashishkumari@msit.in[1],navdeepbohra@msit.in[2]}

Maharaja Surajmal Insitiute of Technology,C-4,Janapuri,New Delhi[1,2]

Abstract. Expectation of sales analysis patterns has been aterritory of extraordinary premium both to scientists end eavoring reveal the data coverd up in the sales information and for the individuals who wish to benefit by predicting sales.The greatly nonlinear nature of the sales information makes it exceptionally hard to structure a framework that can foreseen the future bearing of the sales of articles with adequate exactness. Recommender systems are capable of coming across patterns in sales and generating future income parent based on the patterns as a consequence discovered can considerably supplement the selection-making process of an organization or a trader. Our work presents data analyzation & prediction on sales data by applying neural network algorithms, which produces highly accurate sales forecasts. We have a data that is categorized on various factors i.e. assortment, promotion period, school holiday and state holiday etc.There is also redundant data that is unnecessary as well as some outliers that will be removed too. The data also includes a csv file that contains all the information related to the stores. We feed the data into our trained classifier to give us the prediction of the sales in the coming future, which assist in building a better characteristics and support the sales prediction version.

Keywords: Sales Prediction, Random Forest, XGBoost, Gradient Boosting.

1 Introduction

Predicting the sales of an organization depends on various outer factors like Competition, climate, seasonal developments, and so on and internal movements like promotions, income activities, pricing, collection making plans and so forth, adding to the complexity of the hassle [1]. Previously, financial specialists built up various sales investigation strategies that could enable them to foresee the heading of future sales. Demonstrating and foreseeing of value future cost, in view of the current sales related data and news, is of huge use to the financial specialists. Financial specialists need to know whether sale will rise or fall over certain timeframe.

Our Aim is to analyze fundamental characteristics of the stores provided in the data, comparing these fundamentals to the real sales performance over time. Our objective is to see if we can use machine learning to identify future sales with solid fundamentals that matter so we can invest in them to increase the efficiency of the model and do the optimizations in the algorithm. We address the task of predicting future sales of a particular commodity or various stores of an organization. The algorithms used are random forest algorithm and the XGboost.

ICIDSSD 2020, February 27-28, New Delhi, India

DOI 10.4108/eai.27-2-2020.2303124

These are trees based algorithms that categorize the data and provides the best result. The data is in two parts that is Test data and the Training data.

2 Background

Sales prediction is a mixture of community analysis equipment and time series forecasting techniques. Due to loss of enough beyond income statistics of each drug, an explorative community primarily based evaluation is performed to locate clique sets and institution participants and to apply other member's sales information in their sales prediction to find the best results [3].

Various Machine-learning methods are used to make sales predictions on historical statistics for precise Sales time series within the case while a brand new product or shop is released [4].

Using statistical method to expect the future, by way of studying the relation between the total sales (structured variable) and a number of capability influential factors (unbiased variables). In addition, the importance of each capability thing was quantified using Random Forest set of rules [5].

Gradient Boosting algorithm is used to layout a prediction version to accurately estimate income for retail outlets ofa primary European Pharmacy agency. Sales is primarily based on a combination of temporal and low in cost capabilities along with earlier income records, save promotions, retail competition and country vacations, location and accessibility of the store as well as the time of year[1].

Manufacturing according to the future sales is a typical task for an organization. We here tried to analyze whether is there any certain possible way to guess or predict the future outcome of sales using some machine learning algorithms. Firstly we analyzed all the aspects of the sales that affects the sales and gathered the data and the algorithms information that may help us to perform the task.The sale of a store fluctuates due to various reasons or factors so we had to gather and analyze the data that is varying every day. The data that is decided to work on was Rossman Store Sales data. This is the major drug store all over the europe that consists of total 3000 stores. But the data here provided to us is of 1115 stores. Approximately there are 8 lakh columns that we have to work upon.

The Data was collected from Rossman Store Sales that has the all the fluctuating data about their stores. The raw data required the cleaning and filtration. The row and columns included time and the data that is required for the process. Also there were several other factors we had to consider.

3 Data Engineering

Data evaluation on records amassed via a couple of statistics resources and stored in warehouses, an organization should recognize what numbers of cases are possible on data, variable, missing values and standard hypotheses to guide. Once these entire things are covered then we can start data mining process by creating and deploying data models to do so at the insights gained.

3.1 Data Insights

Plotting the average sales per months revealed that the sales were high during the months of December, March and July, perhaps due to Christmas, Easter festivals and public holidays, **Figure.1.**

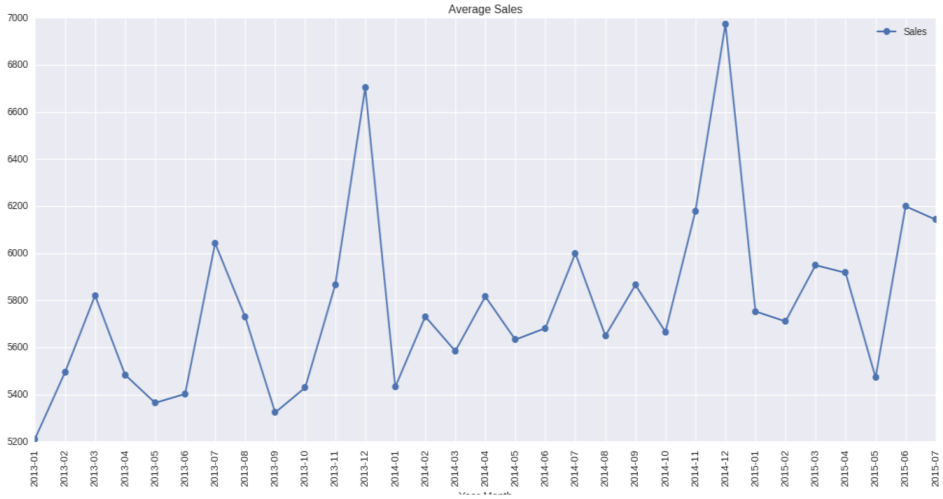


Fig. 1. Month wise Average Sales

The average sales per day revealed that the average sales of the stores that were open on Sundays were higher than the ones that were not, **Figure. 2**. While the stores A, C, D reported the same average sales, the store type B reported much higher sales. Added to this, the stores with extended assortment supplies reported larger sales, **Figure. 3(a)** and **Figure. 3(b)**. In sales area Promo played an important role for the store. The average sales of the stores almost doubled in the promotion weeks. Similarly, sales were higher on school holidays, **Figure. 4**.

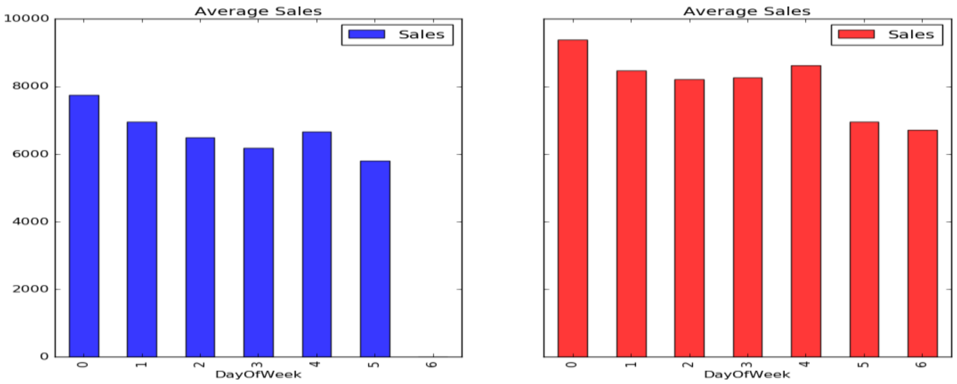


Fig. 2. Sales increased on Sunday

Plotting the sales of stores with time exposed anomalies in the data. Sales data for some months were missing for certain stores, whereas some stores reported spike in sales just before they were closed as shown in above **Figure 5.** These sets were one in events and considered anomalies as they had a negative impact on the average sales being predicted. Similarly, some stores which were open on Sundays never reported zero sales. Out of the 1115 stores some were open on Sundays, whereas the others were not. But plotting the sales time graph were realized that some stores were open on Sundays for certain weeks and closed for other.

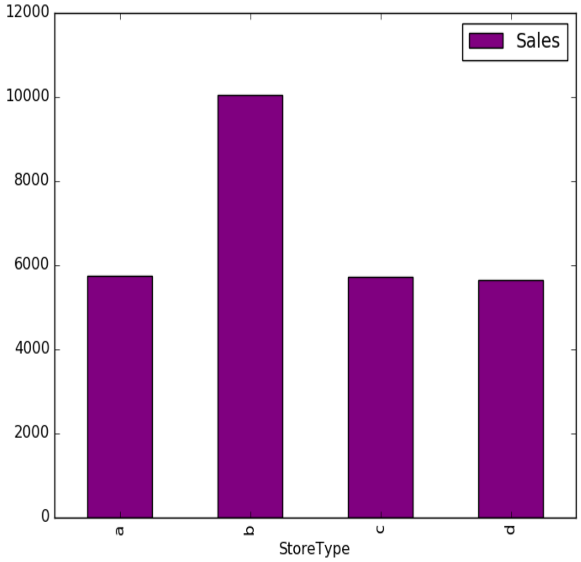


Fig. 3(a). Average sales by Store type

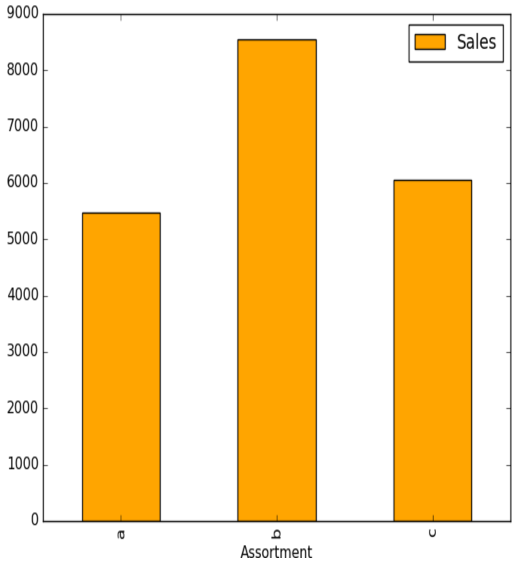


Fig. 3(b). Average Sales by Assortment

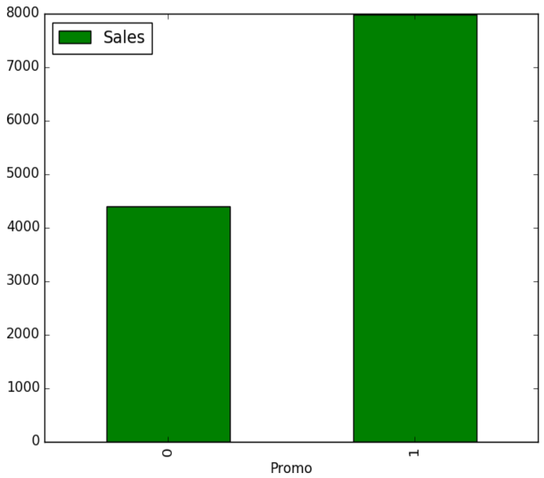


Fig. 4. Average sales by Promotion period

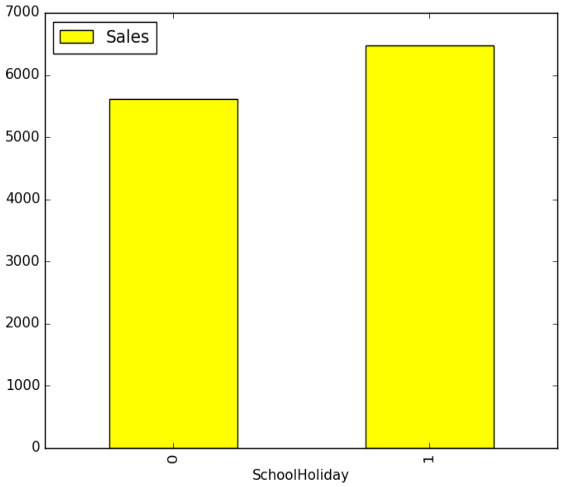


Fig. 5. Average Sales by School Holiday

3.2 Data Transformation

One Hot Encoding. A massive part of the preprocessing is encoding, through the usage of this we can represent such statistics in a way that the computer can understand easily. In other words we can say “convert to computer understandable code”. Label Encoding and One Hot Encoding both are types of encoding, which are mostly used.Label Encoding is used to transform the labels into numeric values so as to convert it into the machine-readable form. It assigns a unique value 'starting from zero' to each data. But this may leadto the precedence issue at the time of data training because high value of a label may be considered to have high precedence than a label having lower value. So, for avoiding this labeling priority problem we have used one hot encoding. One hot encoding is a method which is used to convert variables into a form that would be supplied toML algorithms for better prediction. So, for encoding we have applied this to sales prediction. For predication, many categorical variables were provided in the raw feature list such as – Day Of Week, Store Type, Assortment, Month Of Year. By one-hot encoding, we bloated the feature space as Day of Week got converted to 6 features (Sunday represented as all zeroes). This resulted in RMSPE score improvement by 0.15.

Removing data outliers. An outlier is a statement factor that is distant from different observations in statistics. Sometimes outliers are terrible facts, and need to be excluded. Two types of analysis i.e. uni-variable and multi-variable are used to find the outliers. Mainly Z-score and IQR score (mathematical functions) are used to discover outliers, that can remove easily. So that it will keep away from unusual sales and out-of-stocks situations [7]. Here unexpectedly high sales (four times the mean value) were reported on few days for limitedstores. These outliers and anomalies were clipped and limited to mean ± 3* standard deviation.

Addition of New features . Stores open on sundays throughout the year reported higher average sales than the other stores which were closed on sundays. A new feature was added to report such stores. Also, promo interval was introduced by merging few features from the training set.

Imputation of Missing Data . Few stores were closed for more than six months due to refurbishment. This data was affecting the final model and had to be excluded from the final training set.

Normalizing and Data Binning . Competition Distance and Competition Open had a high cardinality in their data and ranged from 0 to 35000. This had to be normalized and binned to reduce its effect on sales.

4 Implementataion

Two famous strategies are generated from the machine learning research Society are boosting (e.g. Shapiree 1998) and bagging Breiman (1996) of category trees [2]. Here bagging and boosting are applied to a consumer database of Rossman Store Sales data.

4.1 Bagging-Random Forest algorithm

Bagging is the handiest method to Improve or boost, the overall performance of a classifier. Bagging involves training multiple models using bootstrap data sets and aggregating the result of the individual models. We used a particular type of bagging technique called Random Forest. Random Forest constructs a collection of decision trees during training phase and outputs the mean prediction of the individual trees. A random set of features are used to train the decision tree over random training data samples. This is done to make sure the individual models are not correlated.

Random Forest is trained on the data and the performance of the algorithm is recorded by varying the number of trees used. Cross validation error is used to record the performance. As evident from the results , the performance of the algorithm saturates after the number of trees is around 100 which is used to calculate the error on the test data on the Kaggle Competition. Random Forests are biased in favor of categorical variables with different number of levels while calculating the feature importance. One-hot encoding, described in Section Data Transformation, helped us overcome this problem.

4.2 Gradient Boosting Algorithm

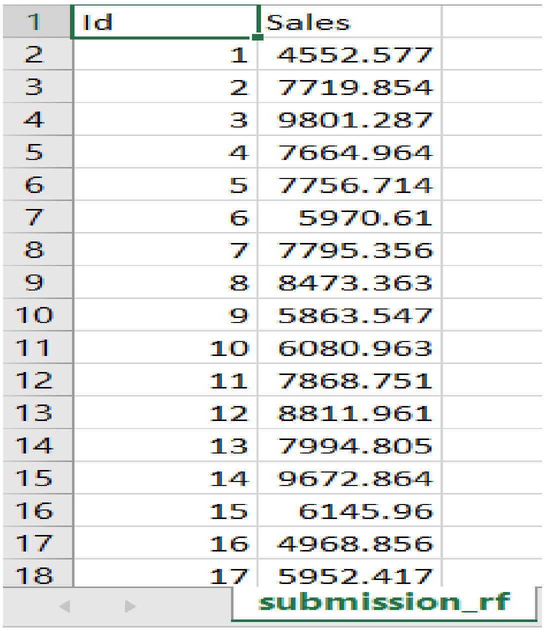
Gradient Boosting algorithm is a machine learning technique used for Regression and classification issues, that generate a prediction model, usually decision trees. While searching at higher strategies for records analysis and forecasting on-line, then we came on the XGBoost which gives plenty higher performance effects than Random Forest Regression [1]. Boosting involves training multiple models in sequence where the error function used to train a specific model depends on the performance of previous models. Gradient Boosting is a technique which constructs a collection of decision trees during training phase and outputs the weighted average of prediction of the individual trees. Gradient Boosting is known to perform better than Random Forest in the case of correlated features. The performance of the algorithm is recorded by varying the number of boosting iterations. As indicated by the results, 800 iterations are used to calculate the error on the test data.

5 Result & Analysis

The current set of features provided a strong model to play with and we were successful to an extent to get the best out of them. Nevertheless, addition of more features that could potentially influence the prediction is to be considered. The information about weather, for instance, can play a vital role in determining the customer‘s willingness to visit the store which is directly related to sales. Similarly, clustering the stores based on the geographical regions and modeling each region separately might help in improving the accuracy. Since the stores are located in majorly Germany, the local events such as festivals, football matches can heavily influence the sales figures. Most importantly, given time-series nature of data, we strongly feel that running time-series modeling techniques such as ARIMA can significantly boost our accuracy[1] [3].

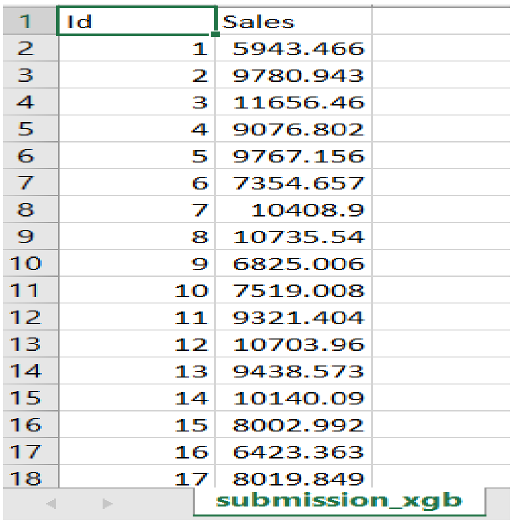
The predicted output of the Random forest algorithm was not similar to that of XGboost algorithm. There was a whole lot difference in the results. There were 3 .cvv or excel files were created. First result was of random forest algorithm, shown in **Table 1**, Second was of XGboost algorithm , **Table 2** and the third was the final prediction. The final prediction file was the weighted average of both the algorithm outputs. Here are the few insights of predicted data in **Table 3**.

Table 1. Output from Random Forest Algorithm



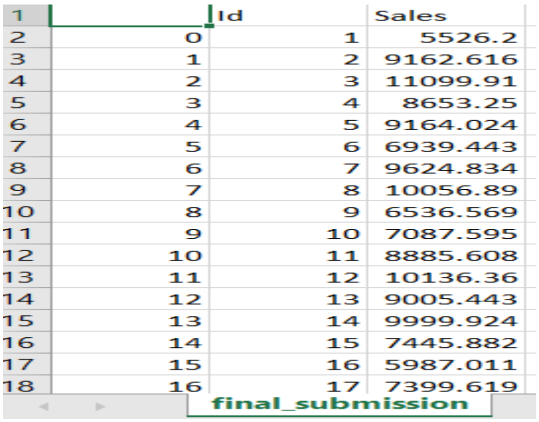
1	Id	Sales
2	1	4552.577
3	2	7719.854
4	3	9801.287
5	4	7664.964
6	5	7756.714
7	6	5970.61
8	7	7795.356
9	8	8473.363
10	9	5863.547
11	10	6080.963
12	11	7868.751
13	12	8811.961
14	13	7994.805
15	14	9672.864
16	15	6145.96
17	16	4968.856
18	17	5952.417

Table 2. Predicted output from XGboost Algorithm



1	Id	Sales
2	1	5943.466
3	2	9780.943
4	3	11656.46
5	4	9076.802
6	5	9767.156
7	6	7354.657
8	7	10408.9
9	8	10735.54
10	9	6825.006
11	10	7519.008
12	11	9321.404
13	12	10703.96
14	13	9438.573
15	14	10140.09
16	15	8002.992
17	16	6423.363
18	17	8019.849

Table 3. The final Output from Data Implemented on both Random Forest and XGboost Datasets.



1		Id	Sales
2	0	1	5526.2
3	1	2	9162.616
4	2	3	11099.91
5	3	4	8653.25
6	4	5	9164.024
7	5	6	6939.443
8	6	7	9624.834
9	7	8	10056.89
10	8	9	6536.569
11	9	10	7087.595
12	10	11	8885.608
13	11	12	10136.36
14	12	13	9005.443
15	13	14	9999.924
16	14	15	7445.882
17	15	16	5987.011
18	16	17	7399.619

6. Conclusion &Future Work

The project helped in solidifying our understanding of the different machine learning techniques and algorithms. We started by directly applying machine learning algorithms on raw data and then organically evolved by polishing the data and using robust algorithms. The exercise emphasized the importance of data analysis and data engineering before running the algorithms on them. We learnt that each technique and model comes with its own set of advantages and disadvantages and no one model can accommodate the data. Accordingly, after the initial attempts of running various algorithms such as Support Vector Machines and SGD on Linear Regression, we obtained better results with ensemble regression techniques: Boosting (Gradient Boosting) and Bagging (Random Forest). Further, to prevent over- fitting and be robust to correlated features, we used a weighted average of the above techniques and obtained even better results.

References

[1] Chandra, S. Jain,A. : Sales Forecasting for Retail Chains. (2015)

[2] Liawand, A. Wiener, M..: Classification and Regression by Random Forest. Vol. 2/3, (2002)

[3] Farvaresh, H. Sepehri, M.: Intelligent Sales Prediction for Pharmaceutical Distribution Companies : A data Mining Based Approach.vol 2/3,(2014)

[4] Pavlyshenko, B.: Machine LearningModels for Sales Time Series Forecasting, mdpi journal (data 2019)

[5] JaeLe, H. Park, J.: A Study of Predict Sales Based on Random Forest Classification,Vol 10, (2017)

[6] Chow, H. Chen, J.: Stochastic Gradient Boosted Distributed Decision Trees, Proceeding 8th ACM Conference on Information and KnowledgeManagement,CIKM'09

[7] Serucab, I. Ribeiroa, A.: Improving Organizational Decision Support : Detection of outliers and Sales Prediction for a Pharmaceutical Distribution Company,(2017).