ISSN No:-2456-2165

An Advanced Sales Forecasting Using Machine Learning Algorithm

B.Sri Sai Ramya[1], K. Vedavathi[2]
(1) GITAM (Deemed to be University), Visakhapatnam
(2) GITAM (Deemed to be University), Visakhapatnam

Abstract:- Sales forecasting is the manner of estimating future sales. Accurate income forecasts allow retailing market to make knowledgeable enterprise choices and predict non permanent and long-term performance. Sales forecasting helps a retailer to estimate its predicted future revenues for income made in a precise duration of time. Hence the time plays a significant role in Sales forecasting. Technically, a time sequence is a sequence of information factors indexed in time order. The time sequence evaluation makes the evaluation of observations as a series of statistics at precise intervals over a length of time, with the cause of figuring out trends, cycles, and seasonal variances this will useful resource in the forecasting of future events. With the development of records technology, giant shops have started out the use of statistical methods like Index numbers, time collection and a couple of regression evaluation for the cause of income forecasting. In this paper, XG Boost algorithm was implemented for prediction The algorithm utilizes Rossmann sales data that operates over 8000 drug stores in European country at 7 locations.

Keywords:- *Sales Forecasting, Time series, ARIMA Model, Gradient Descent and Random Forest Algorithm.*

I. INTRODUCTION

The retail market of sales business is one of the gravest and major business provinces of the statistical mining in data science. The accomplishment of the statistical mining due to its numerous optimization problems and prolific data such are most appropriate prices, discounts, tips and inventory stage can be solved with the aid of analyzing the existing data. The usual problems those are tackled by applying the data mining techniques includes response modeling, recommendation system, , demand prediction, rate discrimination, income match planning, class administration etc., The most accurate forecasting for dynamic business environment in today's competitive business market remains a challenge to satisfy the demands of the customer from the sales market. To make the minor improvements of the diversified retailers in-terms of higher demand for prediction is to lower the costs. It should be done while making the improvements of sales and customer satisfaction. In the current study, for constructing a mannequin for the income prediction of a retailing company, Germany's second-largest drug agency specifically Rossmann accumlation with 3,000 stockpiles in Europe was once chosen. The income reserve of products in stores Rossmann used to be challenged to predict 6 weeks of every day income for 1115 of their abundance positioned throughout Germany. It is a vital trouble for Rossmann as presently their shop managers are tasked with predicting their every day income for up to six weeks in advance. The Store income are influenced through many factors, which include promotions, competition, faculty and nation holidays, seasonality, and locality . As heaps of person managers predict income based totally on their special circumstances, the accuracy of outcomes may also pretty assorted

Time Series ARIMA Model

Algorithm is a time series collection optimized miniature used for forecasting of the non-stop values over the time. A time sequence miniature can predict developments primarily based solely on the statistics set that is used to create the model. The special spotlight of the time collection representation is that it can operate cross prediction. While making a prediction, any new information delivered to the model, the records is mechanically integrated into the representation to function fashion analysis. Mostly used prediction model is an Autoregressive Integrated Moving Average (ARIMA) model. One constituent of ARIMA is the autoregressive model, which fashions a expected estimate x(t) as a linear mixture of values at preceding times. This model can only depend upon the past errors (errors before time t).The ARIMA model also uses the time series as exponential smoothing model. Time sequence forecasting falls below the class of quantitative forecasting whereby statistical standards and principles are utilized to a given historic information of a variable to forecast the future values of the identical variable. Some time sequence forecasting fashions include:

1. Autoregressive Models (AR)
2. Moving Average Models (MA)
3. Seasonal Regression Models
4. Distributed Lags Models

II. LITERATURE REVIEW

A empirical evidence to records of mining in data commercial enterprise john wiley and sons : The author content of the business industry is to make the development of rapid technology to make more user friendly applications and websites. one of the most essential factors of today's choice making world is the forecasting macroeconomic and financial variables. This applies to many industries including finance, education and healthcare so on. However, not many business analysts or developers people know how to use machine learning approach and

technologies to build successful forecast applications. Time series prediction problems pose an important role in many domains and multi-series (more than one time series), multivariate (multiple predictors) and multi-step forecasting like stock price prediction of different symbols could help people to make better decisions. Viewing a problem as supervised machine learning problem by taking lags and calculating on moving averages both on target and features. I would like to examine the relative overall pursuance of the XG Boost to the time series prototypal. The client centered organization regards each regard of an interplay with a purchaser or prospect every name to patron guide ,whole factor of sale transaction, every catalog order and every visit to a organization internet site as a mastering opportunity. But gaining knowledge of requires extra than virtually gathering data. For the companies in the service sector information confers competitive advantage. Credit-card companies, lengthy distance providers, airlines and shops of all types as a great deal or have been on provider as on charge. The most aspect in the back of the success of XG Boost is the scalability in all the scenarios. The machine has greater than ten instances quicker than present famous options on a single system and scales to billions of instance in allotted or reminiscence restricted settings. The scalability of XG Boost is due to quite a few essential structures and algorithmic optimizations. This review is which mentioned about the authors in the reference list that they performed the fine related to what is have done the project. The significance of forecasting is discovering in a tremendous vary of planning and selection making circumstances. It is vital to point out these views that forecasting can come to be a beneficial device for administration in many departments of many organizations. In advertising and marketing a gorgeous amount of choices can be improved. Significantly by way of join them with reliable forecasts of market measurement and market characteristics.

III. MACHINE LEARNING ALGORITHM: EXTREME GRADIENT BOOSTING

The extreme gradient boosting is analyzing through sales forecasting predictions through time series approach and the second one is based on the independent and identically distributed variables which denotes the store sales. This algorithm is a machine learning technique of approach used for constructing predictive tree-based models. Boosting is an ensemble approach in which new fashions are brought to right the mistakes made by means of current models. Models are delivered sequentially until no in addition enhancements can be made. The ensemble approach makes use of the tree ensemble representation which is a set of classification and regression tress (CART). The ensemble method is used due to the fact a single CART, usually, does no longer have a robust predictive power. By the usage of a set of CART (i.e. a tree ensemble model) a sum of the predictions of more than one timber is considered. It is an method the place new fashions are created that predict the residuals or blunders of prior fashions and then delivered collectively to make the closing prediction. We can use cross-validation feature for XG Boost. The cross-validation procedure is then repeated nrounds times, with every of the nfold subsamples used precisely as soon as the validation data.

IV. ABOUT DATASET AND ATTRIBUTES

I have taken dataset from the Rossmann shops which are of 3000. I had amassed facts from extraordinary records warehouse. I had chosen Rossmann drug store's income statistics which is the second greatest drug save in Germany. We carried out Extreme Gradient Boost to predict the everyday income for 6 weeks into the future for extra than 1,000 stores. The Rossmann Data includes facts about 1115 shops from 1st Jan 2013 to thirty first July 2015 (942 days). In complete we have 1017209 entries.

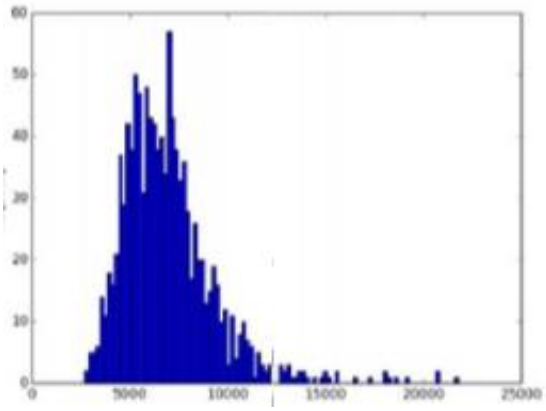

Fig 1:- Histogram of the mean sales per store

The most shops have very comparable income and that there is a small share of outliers. Average income have a tendency to be greater on the Sunday as in contrast to different days. However, the magnitude of clients journeying a keep on weekends tends to be much less due to the truth that most shops are closed on Sunday. Some of the attributes we used are store, store id, sales, customer open, kingdom holiday, faculty holiday, keep type, assortment, opposition distance, promo, promo intervals.

A	B	C	D	E	F	G
Store	DayOfWe	Date	Sales	Customer	Open	Promo
1	5	7/31/2015	5263	555	1	1
2	5	7/31/2015	6064	625	1	1
3	5	7/31/2015	8314	821	1	1
4	5	7/31/2015	13995	1498	1	1
5	5	7/31/2015	4822	559	1	1
6	5	7/31/2015	5651	589	1	1
7	5	7/31/2015	15344	1414	1	1
8	5	7/31/2015	8492	833	1	1
9	5	7/31/2015	8565	687	1	1

Fig 2:- Train data set

V. RESULTS AND ANALYSIS

```
params = {"objective": "reg:linear",
          "booster" : "gbtree",
          "eta": 0.3,
          "max_depth": 10,
          "subsample": 0.9,
          "colsample_bytree": 0.7,
          "silent": 1,
          "seed": 1301
          }
num_boost_round = 200

print("Train a XGBoost model")
X_train, X_valid = train_test_split(train, test_size=0.012, random_state=10)
y_train = np.log1p(X_train.Sales)
y_valid = np.log1p(X_valid.Sales)
dtrain = xgb.DMatrix(X_train[features], y_train)
dvalid = xgb.DMatrix(X_valid[features], y_valid)

watchlist = [(dtrain, 'train'), (dvalid, 'eval')]
gbm = xgb.train(params, dtrain, num_boost_round, evals=watchlist, early_stopping_rounds=100,
```

Fig 3:- Input code for sales prediction using XG Boost

```
[43]    train-rmse:0.15534      eval-rmse:0.15610       train-rmspe:0.19617     eval-rmspe:0.16692
[44]    train-rmse:0.15383      eval-rmse:0.15470       train-rmspe:0.19473     eval-rmspe:0.16541
[45]    train-rmse:0.15290      eval-rmse:0.15402       train-rmspe:0.19296     eval-rmspe:0.16464
[46]    train-rmse:0.15076      eval-rmse:0.15194       train-rmspe:0.19101     eval-rmspe:0.16219
[47]    train-rmse:0.14793      eval-rmse:0.14914       train-rmspe:0.18792     eval-rmspe:0.15897
[48]    train-rmse:0.14592      eval-rmse:0.14718       train-rmspe:0.18602     eval-rmspe:0.15696
[49]    train-rmse:0.14528      eval-rmse:0.14663       train-rmspe:0.18525     eval-rmspe:0.15635
Validating
RMSPE: 0.156350
Make predictions on the test set
```

Fig 4:- Output code for sales prediction

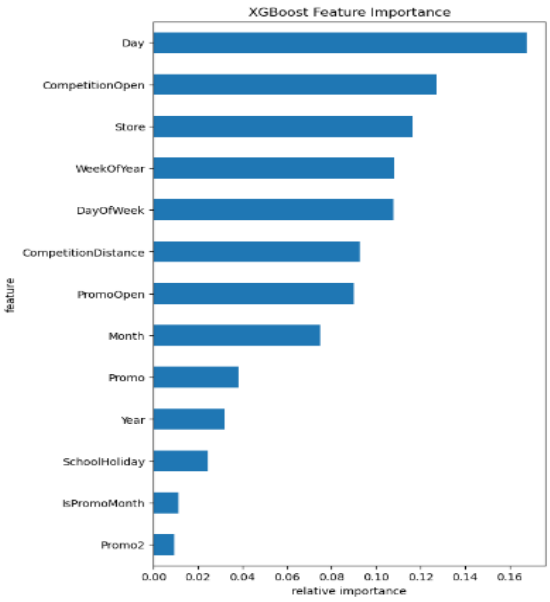

Fig 5:- Histograph for sales prediction in the market

➢ *Extreme gradient boosting algorithm steps:*

The ideas of gradient boosting framework and designed to push the severe of the computation limits of machines to furnish a scalable, transportable and correct library

XG Boost algorithm is that it can calculate function significance which is to exhibit which predictor columns (or variables) are greater influential on the prediction outcome

XG Boost internally produces envisioned likelihood values which are ranging between zero and 1instead of the envisioned label values such as real or false

VI. CONCLUSION AND FUTURE SCOPE

In this present study we used some data mining techniques for sales forecasting such are ARIMA models and XG Boost algorithms which get better efficiency to manipulate the trending sales analysis. A lot of evaluation was once carried out on the information to become aware of patterns and outliers which would booster obstruct the prediction algorithm. The points used ranged from keep data to purchaser data as properly associate-geographical information. Data mining techniques like Linear Regression, Random Forest Regression and XGBoost had been carried out and the outcomes compared. XGBoost which is an expanded gradient boosting algorithm was once found to function the excellent at prediction. With satisfactory effectively to amplify our answer to assist shops enhances productiveness and expands income by using taking benefit of information analysis. Sales prediction performs a necessary function in growing the effectively with which shops can function as it presents important points on the visitors a save can count on to get hold of on a given day. In addition to simply predicting the contemplated sales, there are different facts which can be mined to spotlight essential tendencies and additionally enhance planning. Such are advertisement, recommendation, predicting demand, consumer based totally pricing, holiday/extended sale planning and product classification.

REFERENCES

[1]. Shirley Coleman AhlemeyerStubbe, Andrea. A practical guide to data mining for business and industry. John Wiley and Sons, 2014.
[2]. P.Mekala, B.Srinivasan. Time series data prediction on shopping mall.In International Journal of Research inComputer Application and Robotics, Aug 2014.
[3]. Gordon S. Linoff Berry, Michael JA. Data mining techniques: for marketing, sales, and customer relationship management. John Wiley and Sons, 2004.
[4]. L. Breiman. Random forest. Mach. Learn., 4:5–32, 2001.
[5]. Tianqi Chen. Xgboost: Link: https://github.com/tqchen/xgboost.
[6]. Jerome H. Friedman. Stochastic gradient boosting.Computational Statistics and Data Analysis, pages 367–378, 2002.
[7]. Romana J. Khan and Dipak C. Jain.An empirical analysis of price discrimination mechanisms and retailer profitability. Journal of Marketing Research, 42.4:516–524, 2005.