International Research Journal of Computer Science (IRJCS) ISSN: 2393-9842
Issue 04, Volume 6 (April 2019) www.irjcs.com

SALES FORE CASTING USING ENSEMBLE METHODS

P.Subashini [1], Yash Kimtani [2], Yuvraj Talukdar [3], Ajit Kumar Shah [4]
[1][2][3][4] Department of Computer Science & Engineering,
SRM Institute of Science and Technology,
Ramapuram campus, Chennai, India

Manuscript History
Number: **IRJCS/RS/Vol.06/Issue04/MRCS10082**
Received: 25, March 2019
Final Correction: 05, April 2019
Final Accepted: 10, April 2019
Published: April 2019
Citation: Subashini, Yash, Talukdar, Ajit (2019).Sales Fore Casting Using Ensemble Methods. IRJCS:: International Research Journal of Computer Science, Volume VI, 84-89. **doi://10.26562/IRJCS.2019.APCS10080**
Editor: Dr.A.Arul L.S, Chief Editor, IRJCS, AM Publications, India

Abstract: Sales Forecasting provide information of how much sales of a specific product is likely to occur in a specified time period so that the organization can predict their targets and improve their strategy in order to improve their sales or productivity for the coming future. This paper use ensemble learning model which is a combination of different machine learning models like CATBOOST, XGBOOST etc., to forecast monthly sales for the one of large Russian software firms called 1C Company. The model building process was guided by the commonsense reasoning and by analytic knowledge discovered during data analysis and definitive conclusions.

Keywords: Cat boost, Xg boost; SVM; Light GBM; ARIMA;

I.INTRODUCTION

Business sales forecasting are becoming more sophisticated for both strategic and judicious business planning. Thus, there is still an outstanding question arise how to enhance the quality of sales forecasting. The Non-Transparent, top performing black-box machine learning models, such as boosting, random forests, SVM-(Support Vector Machine), and neural networks attain remarkable greater predictive performance in comparison to simple, interpretable models such as Naive Bayes, Decision tree. This is only the reason for low utilization and acceptance of predictive ML models in areas where transparency and comprehensibility of decisions are required. For data gathering trend and seasonal patterns, failure in describing for the trends may result in very poor predictions. In dealing with these problems of sales prediction from last few decades, time series forecasting technique, such as moving average, exponential smoothing, multivariate regressions, and ARIMA, have been introduced and commonly used in practice to solve for these patterns.

Artificial Neural Networks (ANNs) is a new opponent in forecasting sophisticated patterns and seasonal trends. Artificial neural networks has arisen as a technology with a huge potential for recognising and modeling data patterns in last few decades that are hard to observe by conventional statistical methods in variety fields of computer science, and finance. (ARIMA) stands for Auto Regressive Integrated Moving Average, it has been already used in forecasting commodity prices, such as natural gas or oil. In power systems, ARIMA model has also been used for forecasting load, with satisfying results. In present, that is being done in many countries with the restructuring process, simpler Auto Regressive (AR) models are also being used for predicting monthly prices, like in the Norwegian system.

In this paper, we will focus on the sales prediction using Ensemble Learning Model. That is, this paper provides sales forecasting by using the different machine learning models and combine the output of applied models and produce the accurate sales forecast. These models are based on the time series analysis and provide reliable and accurate forecasts of sales in the different businesses.

II. METHODOLOGY

This section describes various ensemble methods and prediction models for sale forecasting.

2.1 Ensemble method:

It help enhance machine learning results. It is done by combination of several models. This method delivers high quality prediction compared to a single model. Various ensemble methods include the following: -

a. Stacking: It combines multiple classifications models using a meta-model or a meta-classifier and is trained on the output from the base level model as the features.

b. Boosting: It makes reference to a group of algorithm that has the capability to convert less strong learners into strong learners. It runs on the principle of fitting a series of weak learning models which are slightly better compared to random guessing, example, small decision tree into weighted version, of the data. More weights are given to examples that were wrongly predicted by earlier rounds. The predictions then combined using a weight sum to produce final prediction.

c. Bagging: Bagging stands for bootstrap aggregation. The way to reduce the variance of an estimate is to average together the multiple estimates. For example, different subsets of the data chosen randomly can be used for training M different trees with replacement and we compute the ensemble result using:

$$f(x) = 1/M \sum_{m=1}^{M} f_m(x)$$

Bootstrap sampling method is used by bagging to get the data subsets for training the base learners. For the purpose of averaging the outputs of base learners bagging uses *voting for classification* and *averaging for regression*.

d. Blending: Blending uses the same method as that of stacking but it uses only a holdout set from the training dataset to make the required predictions. In different term, the predictions are only made on the holdout set. The prediction and holdout set are used to construct a model that is executed on the testing dataset.

2.2. Prediction Model

a. XG Boost: Xg boost refers to extreme Gradient Boosting. It is based on gradient boosting framework. Gradient boosting is the machine learning technique to deal with classification, regression and ranking problems. Xg boost has regularized model formalization to control over fitting that boost its performance. It provides a good result for most of the datasets involving linearity and nonlinearity. It is efficient as it supports parallel computation on a single machine. Model of Xg boost is based on the concept of Gradient boosting which believes that single trees does not provide enough strength to give accurate prediction. Hence, the ensemble of decision trees are used, where trees are added in such a way that they optimizes the current error. The tree ensemble model uses classification and regression trees (CART). It optimizes the result by using objective functions with certain set of parameters.

b. Light GBM: It is a distributed, fast as well as high-performance gradient boosting (GBRT,GBDT, GBM or MAR) framework that uses a learning algorithm that is tree-based, and is used for classification, ranking as well as other different ML tasks. It is a gradient boosting framework which makes use of tree-based algorithm and also follow leaf-wise method which is different from other algorithms that work in a level wise approach pattern.

c. Cat Boost: It is developed by Yandex and is recently open-sourced. It is a good machine learning algorithm. It can easily work with deep learning framework like Tensor Flow and Core ML. It can work with different data types to help solving a variety of problems that businesses deals in today. To boost it up, it provides good accuracy.

It has two major speciality:

- It delivers good results without large scale data training which is normally required by other ML methods.
- Provides powerful ready-made supports for the more expressive data formats that many business faces today. “Cat Boost” name comes from two words “Category” and “Boosting”.

The library works well with multiple Categories of data, such as text, audio, image and historical data.“Boost” derived from gradient boosting machine learning algorithm as this library is founded on gradient boosting library. Gradient boosting is a useful ML algorithm widely applied to different types of new business problems like recommendation and fraud detection and forecasting, it performs quiet well too. It can also delivers good result with very less data, unlike DL models that need to be trained on massive amount of data.

d. Random forests: It is also called random decision forests, is a ensemble method that is used in building predictive models for solving regression and classification problems. Ensemble methods uses numerous learning models to assist in better predictive results, for random forest, the model generate a forest of random unrelated decision trees to reach at the best feasible output. It has been used to predict time series because they accept past values as input and predicts the future values.

III. PROPOSED MODEL

3.1 Data Description

We obtained the dataset from Kaggle and it's a time series dataset containing daily sales data, published by large Russian company firms 1C Company. The dataset is split into separate files which contain different columns as follows

Attribute	Description
ID	an id which acts as identifier for a (Shop, Item) tuple within the test set
shop_id	unique identifier for a shop
item_id	unique identifier of a product
item_category_id	unique identifier of item category
item_cnt_day	no of products sold-out. Weneed to predict a monthly amount of this measurement
item_price	current price of an item
Date	date in format dd/mm/yyyy
date_block_num	a consecutive month number, used for convenience. January 2013 is 0, February 2013 is 1,..., October 2015 is 33
item_name	name of item
shop_name	Name of shop
item_category_name	name of item category

The different files are item.csv, item_category.csv, shops.csv, sales_train.csv, test.csv.
items.csv:- It contain information about the items/products. It has 22170 unique items.
item_category.csv:- It contain information about items categories with 84 unique item categories.
Shops.csv:- It contains information about the shops. The total no of data is 60 unique shop names.
Sales_train.csv:- It is a training set which contain daily historical data from Jan 20 2013 to Oct 2013. It has 2935849 data points.
Test.csv:-It's the test set. On which sales prediction is to be made for these shops and products for November 2015. It has 214200 individual data points.
As the sales prediction should be in monthly basis we transform the data from daily wise transaction to monthly wise.

3.2 Data Analysis:-

In this section we have done exploratory data analysis where we find the answers of certain questions such as what is the seasonal trend? Top selling category? Or year in which more sale occurred?

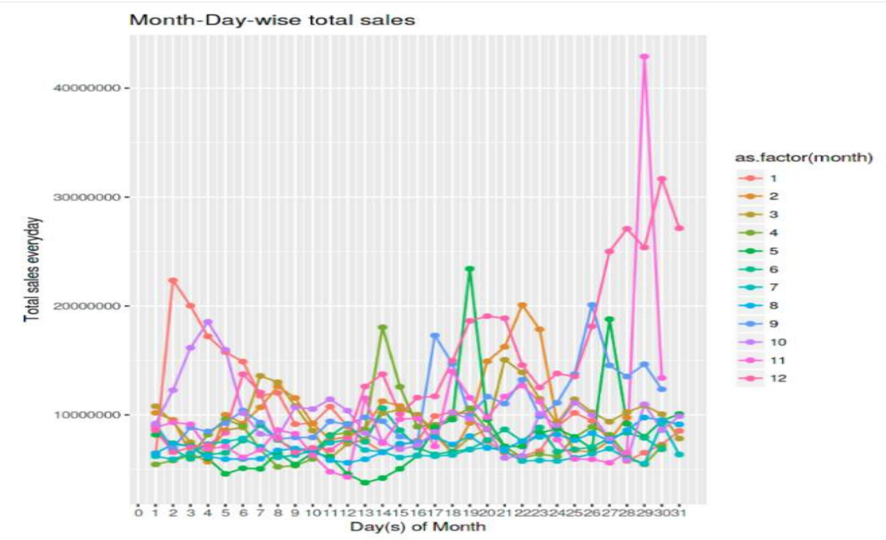

Figure 1

In figure 1 it shows what is the monthly sales. From this graph we observe that the sales are generally high during the beginning and end of a month and the sales are a lot high during the month of November compared to rest of the year.

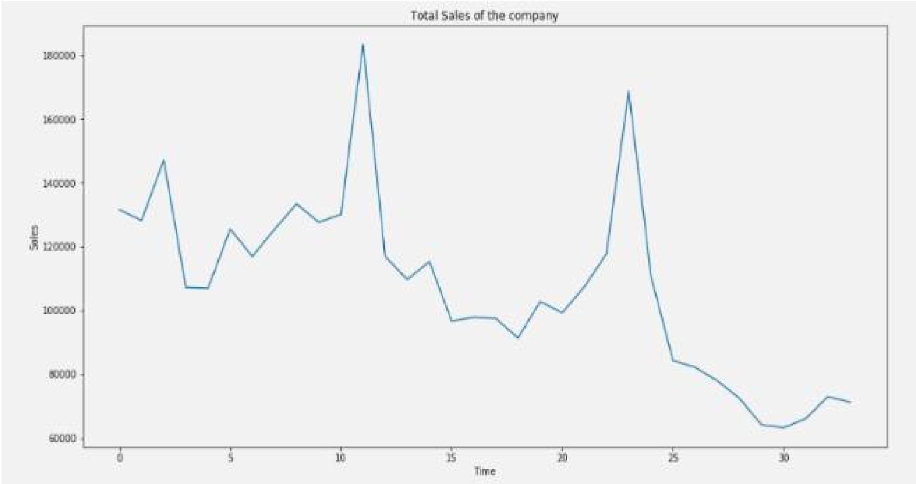

Figure 2

In Figure 2 we see what is the sales pattern of the company in an interval of time.

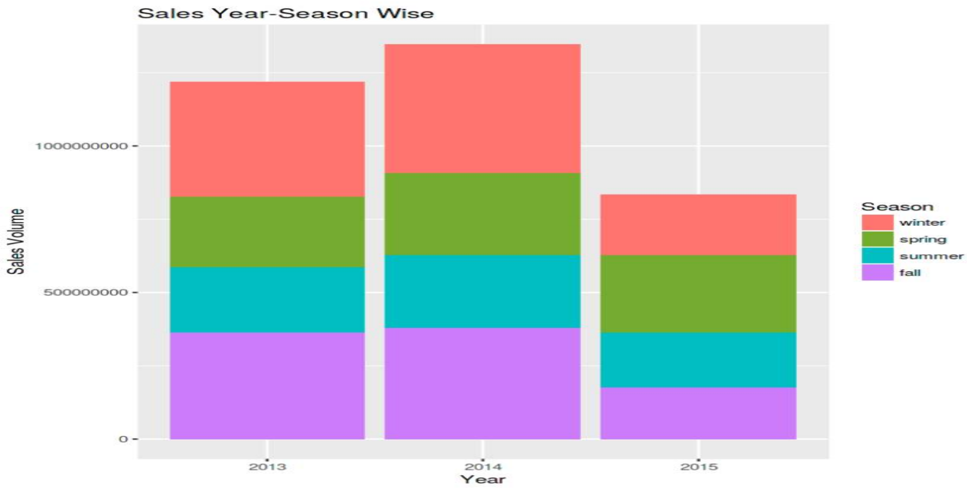

Figure 3

In figure 3 we observe that winter season sales are higher compared to other season. Sales of the year 2014 was higher compared to other years. In the year 2015 sales were down.

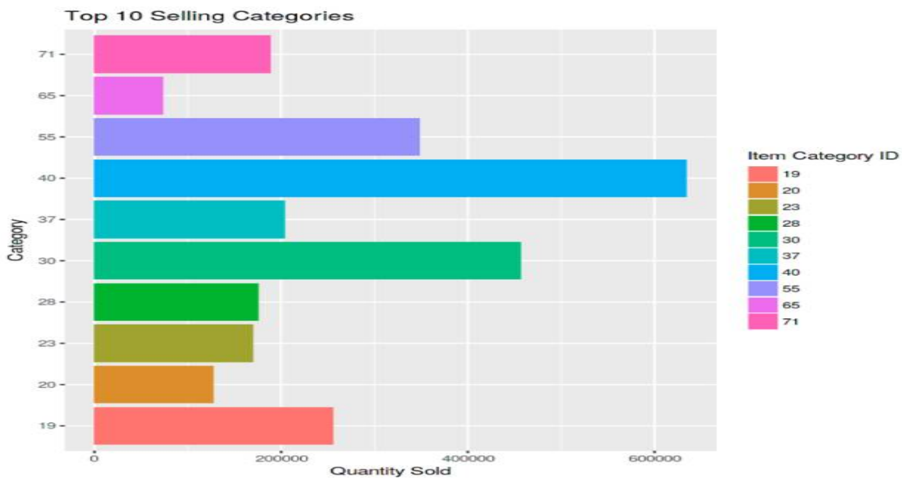

Figure 4

In figure4 we observe the top 10 selling categories of the company. Here we also observe that only few categories holds most of the sales count. From above analysis, we are able to generate new features which will help the model in delivering better prediction. It also help us to understand the dataset better and help us remove the outliers.

3.3 Stacked ensemble model

In this section, we are proposing a stacked ensemble architecture in which there is two levels. In level 1 there are four models cat boost xg boost, random forest and knn. At level 2 we are using linear regression. We are using simple hold out scheme for training and validation purpose in which we split training data into 3 parts, lets call them part A,B,C where we fit part a to each model of first level and predict on part B, C and test data. Getting the meta features let's call it part_B_ meta part_C_ meta and test_ meta respectively. Than we train level 2 model on part_B_meta and valiadate on part_C_ meta and when we are satisfied with the performance we train the level 2 model with combined part_b_meta and part_C_meta and predict for test meta. We train the level 1 model cat boost, xgboost, random forest and knn on simple holdout scheme next we form a new dataset which contain the output of the level 1 models and train the level 2 model on this dataset.

IV. PERFORMANCE ANALYSIS

In this section we will discuss the performance of the model we have proposed above. We used root mean squared error

$$RMSErrors = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

Where the output is value for theithy$_i$ observation and is they$_i$ hat predicted value.

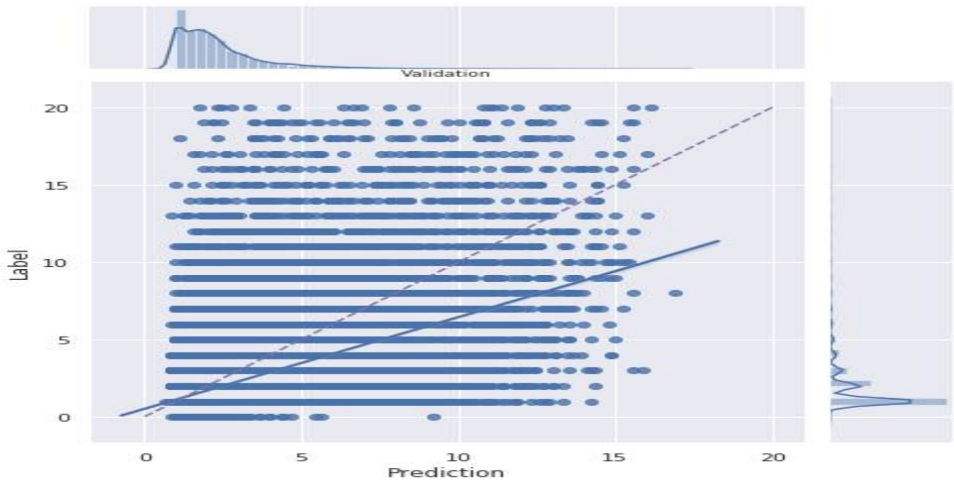

Figure 5

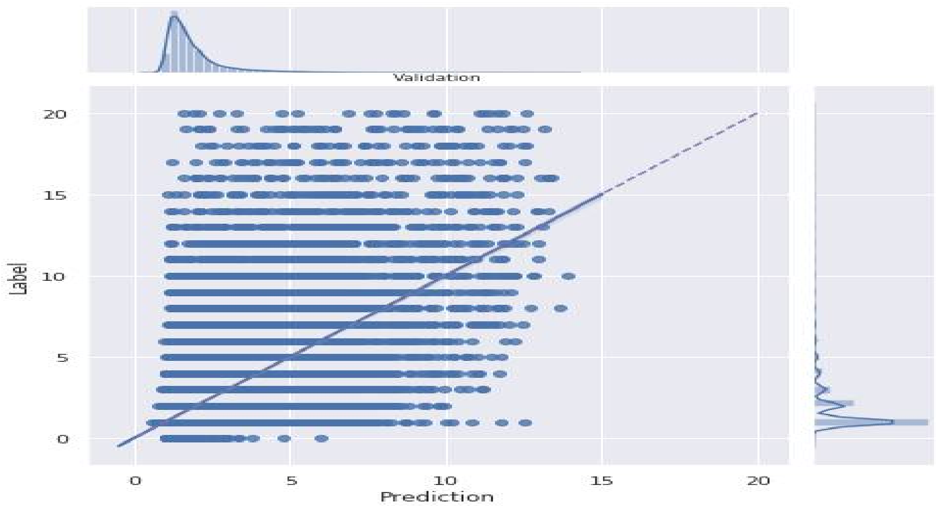

Figure 6

So, we are using root mean squared error for calculating the performance of the proposed model and comparing with the other model on validation set. We compare results of the ensemble model with a single model like KNN, CATBOOST etc. and found that the proposed ensemble model was able to recognize pattern of the data in a decent way compared to other models.

In figure 5 displays the performance of level 1 model, the middle line is drawn value of max and min prediction. The blue line represents how the model behaves the closure the line to the dotted line the better is the prediction. By the above graph that alone level 1 model is not capable for better prediction. Figure 6 shows the performance of ensemble model. This model delivers a better performance compared to level 1 model.

CONCLUSION

Sales prediction is a widely studied topic both in business. Due to the non-static nature of the data it is difficult to forecast the sales. The sales depend on numerous factors some of which may be unexpected line economy of the country weather patterns etc. Still based on the historic data we tried to design a model for this purpose. We can collect more data on the locations of the shop and the festival seasons in the locality so model can make more accurate predictions. Ensemble learning models can be applied on variety of problems such as stock prediction weather forecast etc.

REFERENCES

1. Cat Boost: gradient boosting with categorical features support Anna Veronika Dorogush, Vasily Ershov, Andrey GulinYandex
2. XG Boost: A Scalable Tree Boosting System Tianqi Chen University of Washingtontqchen@cs.washington.edu Carlos Guestrin University of Washingtonguestrin@cs.washington.edu
3. RANDOM FORESTS Leo Breiman Statistics Department University of California Berkeley, CA 94720
4. A KNN Research Paper Classification Method Based on Shared Nearest Neighbor Yun-lei Cai, Duo Ji ,Dong-feng Cai Natural Language Processing Research Laboratory, Shenyang Institute of Aeronautical Engineering, Shenyang, China, 110034
5. Journal of Artificial Intelligence Research 11 (1999) 169-198Submitted 1/99; published 8/99Popular Ensemble Methods: An Empirical Study David Opitzopitz@cs.umt.eduDepartment of Computer Science University of Montana Missoula, MT 59812 USA Richard Maclinrmaclin@d.umn.eduComputer Science Department University of Minnesota Duluth, MN 55812 USA
6. OnlineARIMAAlgorithmsforTimeSeriesPredictionChenghaoLiu1,2,StevenC.H.Hoi 2,PeilinZhao3,JianlingSun1 1School of Computer Science and Technology, Zhejiang University, China 2School of Information Systems, Singapore Management University, Singapore 3Institute for Infocomm Research, A*STAR, Singapore twinsken@zju.edu.cn, chhoi@smu.edu.sg, zhaop@i2r.a-star.edu.sg, sunjl@zju.edu.cn
7. SUPPORT VECTOR MACHINE FOR REGRESSION AND APPLICATIONS TO FINANCIAL FORECASTING Theodore B. Trafalis* and Huseyin Ince** School of Industrial Engineering, University of Oklahoma 220 W. Boyd, Suite 124, Norman, Oklahoma 73019
8. Light GBM: A Highly Efficient Gradient Boosting Decision Tree Guolin Ke1, Qi Meng2, Thomas Finley3, Taifeng Wang1,Wei Chen1, Weidong Ma1, Qiwei Ye1, Tie-Yan Liu11Microsoft Research2Peking University3Microsoft Redmond1{guolin.ke, taifengw, wche, weima, qiwye, tie-yan.liu
9. LEARNING THE NAIVE BAYES CLASSIFIER WITH OPTIMIZATION MODELSSONATAHERI ,MUSAMAMMADOV,Centre for Informatics and Applied Optimization School of Science, Information Technology and Engineering, University of Ballarat, Victoria 3353,Victoria Research Laboratory National ICT Australia, Victoria 3010,
10. Akhter Mohiuddin Rather, "A prediction based approach for stock returns using autoregressive neural networks," Information and Communication Technologies (WICT), 2011 World Congress on, Mumbai, 2011, pp. 1271-1275.

