A REPORT

ON

Daily retail demand forecasting using machine learning

Submitted in

The course Business Analytics using R (OE4M23)

For the Degree of

Bachelor of Technology (B. Tech)

Under Guidence of:

Dr. Sunil Agrawal

(Associate Professor, IIITDMJ)

Submitted by:

Gopalji Yadav (2018094)

Harsh Singh Bais (2018101)

Ayush Kumar Gupta (2018055)

PDPM IIITDM JABALPUR

INDEX

1. INTRODUCTION

Demand forecasting is one of the major challenges for retailers as it is the input for many operational decisions. In particular, for perishable goods with a high rate of deterioration, it is important to provide the correct quantities every day. Such goods have higher average sales and higher order frequencies than nonperishable goods. Their freshness decreases rapidly, which makes daily replenishment inevitable. Unsold items are waste, and discarded goods are a major cost factor that can be reduced by more accurate demand forecasts as these enable the reduction of safety stocks. Retailers typically run a large number of stores and offer a broad assortment of goods. Consequently, numerous daily decisions need to be supported by predictions. A competitive advantage can be gained by automating the prediction process. Moreover, retailers accumulate very large datasets (e.g., sales history) over years that can also be enhanced by external information such as calendar events.

One key challenge is to forecast demand in future days that are subject to vastly different demand patterns than on regular days. We present the case of a store for which we address the problem of forecasting the daily demand for different product categories at the store level. Such forecasts are an input for production and ordering decisions. We treat the forecasting problem as a supervised machine learning task and provide an evaluation of different methods, including artificial neural networks and gradient-boosted decision trees. In particular, we outline and discuss the possibility of formulating a classification instead of a regression problem. An empirical comparison with established approaches reveals the superiority of machine learning methods, while classification-based approaches outperform regression-based approaches. We also found that machine learning methods not only provide more accurate forecasts but are also more suitable for

applications in a large-scale demand forecasting scenario that often occurs in the retail industry.

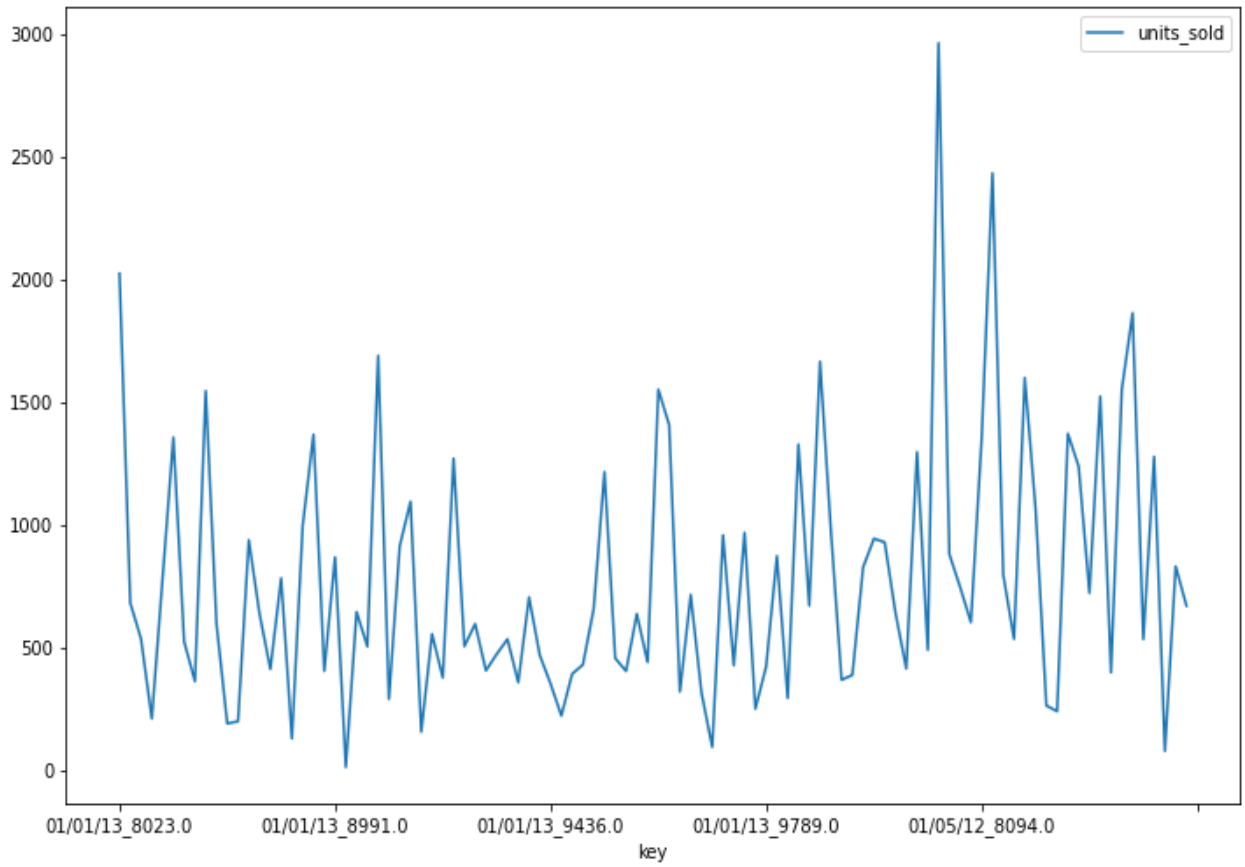
2. PROBLEM DESCRIPTION

Our Problem description is motivated by the requirements of a real-world problem for a large retail store chain that shares many characteristics with other mid-size stores. The company runs over 100 stores in a geographically restricted area. Pastries are produced in a centralized production facility and delivered daily to the stores. Daily delivery is necessary as the freshness of baked goods decreases rapidly, which only allows them to be sold on the day of production. The supply chain is quite agile as all of its important parts are operated by the company; i.e., production and distribution, as well as the stores. Daily demand forecasts are required for many operational decisions, including the determination of production and delivery quantities, and staffing decisions. In other way of view, we can also consider that there is a single large store which is having different number of products, and on the basis of past data of total products sold we have to predict number of products will be sold on upcoming weeks with most efficiency. From the existing literature, it is unclear whether machine learning methods are able to outperform established approaches for retail demand forecasting given the characteristics of the present case, which play to the strengths of machine learning. With this study, we address, Are machine learning methods a viable alternative to established approaches for the given forecasting scenario? And Which machine learning method provides the most accurate predictions? Which modeling approach works best? We also extend the existing literature in multiple directions. First, we present a real-world daily demand forecasting application. Second, we elaborate on the possibilities of formulating time series forecasting as a machine learning problem. This includes a rather uncommon transformation of the regression problem to a classification problem. Third, we provide an empirical evaluation of state-of-the-art machine learning methods for large-scale demand forecasting. Based on a comparison with established approaches, we illustrate the viability of using machine learning in a productive setting.

3. METHODOLOGY OF SOLUTIONS

3.1. Datasets

We have dataset which includes sales history of large store chain which is having many stores and total products sold. Our dataset has more than fifteen lakh sales records from which we have divided it into two parts: one is for training the model which contains 85% of the total data and other is for testing the trained model which contains the remaining 15% of data.



3.1.1. Train data

In train dataset, we have four attributes like record_id which stores number of records as indexes, another one is week which stores the different date of sale, another one is store_id which reflects from which store on a particular day how many products are get sold and the last attribute is units_sold which give the total

number of products get sold from a particular store on a particular date. In train data after splitting, we have 8395 rows and each row is having 4 columns.

	A	B	C	D
1	record_ID	week	store_id	units_sold
2	1	17/01/11	8091	20
3	2	17/01/11	8091	28
4	3	17/01/11	8091	19
5	4	17/01/11	8091	44
6	5	17/01/11	8091	52
7	9	17/01/11	8091	18
8	10	17/01/11	8091	47
9	13	17/01/11	8091	50
10	14	17/01/11	8091	82
11	17	17/01/11	8095	99
12	18	17/01/11	8095	120
13	19	17/01/11	8095	40
14	22	17/01/11	8095	68
15	23	17/01/11	8095	87
16	24	17/01/11	8095	186
17	27	17/01/11	8095	54
18	28	17/01/11	8095	74
19	29	17/01/11	8095	102
20	30	17/01/11	8095	214

3.1.2. Test data

The test data has only three attributes: one is record_id which stores number of records as indexes, another one is week which stores the different date of sale, another one is store_id which reflects from which store on a particular day how many products are get sold. Based on the trained model we have to predict demand forecast. In test data we have 1481 rows and each row has one column.

	A	B	C
1	record_ID	week	store_id
2	1	17/01/11	8091
3	2	17/01/11	8091
4	3	17/01/11	8091
5	4	17/01/11	8091
6	5	17/01/11	8091
7	9	17/01/11	8091
8	10	17/01/11	8091
9	13	17/01/11	8091
10	14	17/01/11	8091
11	17	17/01/11	8095
12	18	17/01/11	8095
13	19	17/01/11	8095
14	22	17/01/11	8095
15	23	17/01/11	8095
16	24	17/01/11	8095
17	27	17/01/11	8095
18	28	17/01/11	8095
19	29	17/01/11	8095
20	30	17/01/11	8095

3.2. Algorithms

Statistical time series methods (e.g., exponential smoothing, ARIMA models) have been successfully applied to many forecasting problems, and there is no definite evidence that they are inferior to machine learning methods. The results of the most recent competition suggest that (combinations of) statistical methods outperform pure machine learning methods, while a hybrid approach performed best at forecasting univariate time series. Machine learning methods offer features that are well suited for the present forecasting scenario. They are designed to learn patterns from data and are naturally able to process large datasets. Therefore, they do not impose assumptions on the data. This holds for the data generating process as well as the scope of the model. Moreover, the methods do not explicitly distinguish between information that directly stems from the time series and external data. The characteristics and flexibility of machine learning methods make them, in principle, a viable alternative to the established approaches (e.g., multivariate linear regression models and univariate statistical time series models). We are implementing four algorithms which are Time series (Arima), Linear regression, Random Forest and one gradient boosting algorithm using XG Boost.

3.2.1. ARIMA Model

An ARIMA model is a class of statistical models for analyzing and forecasting time series data. It explicitly caters to a suite of standard structures in time series data, and as such provides a simple yet powerful method for making skillful time series forecasts. ARIMA is an acronym that stands for Autoregressive Integrated Moving Average. It is a generalization of the simpler Autoregressive Moving Average and adds the notion of integration. We faced difficulty while implementing ARIMA model.

3.2.2. Linear Regression

Linear regression is a model that captures the linear relationship between variables, one labeled as the dependent variable and the other(s) labeled as the independent variable(s). In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data, such models are called linear models. Learning a linear regression model

means estimating the values of the coefficients used in the representation with the data that we have available.

After fitting the model, we got the **R square score for the train dataset is 0.01930 which is nearly 2%** and for the test dataset we got R square score even worse which is in negative that means **this model fitted our dataset extremely bad.**

3.2.3. Random Forest

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. The forest it builds, is an ensemble of decision trees, usually trained with the bagging method. The general idea of the bagging method is that a combination of learning models increases the overall result. One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems. Let's look at random forest in classification, since classification is sometimes considered the building block of machine learning. Random forest has nearly the same hyperparameters as a decision tree or a bagging classifier. Fortunately, there's no need to combine a decision tree with a bagging classifier because you can easily use the classifier-class of random forest. With random forest, you can also deal with regression tasks by using the algorithm's regressor.

For our dataset this model has performed good. **The R square score for this model we are getting is 0.6855** that means this model is fitted with our dataset with 69% accuracy.

3.2.4. XG-Boost

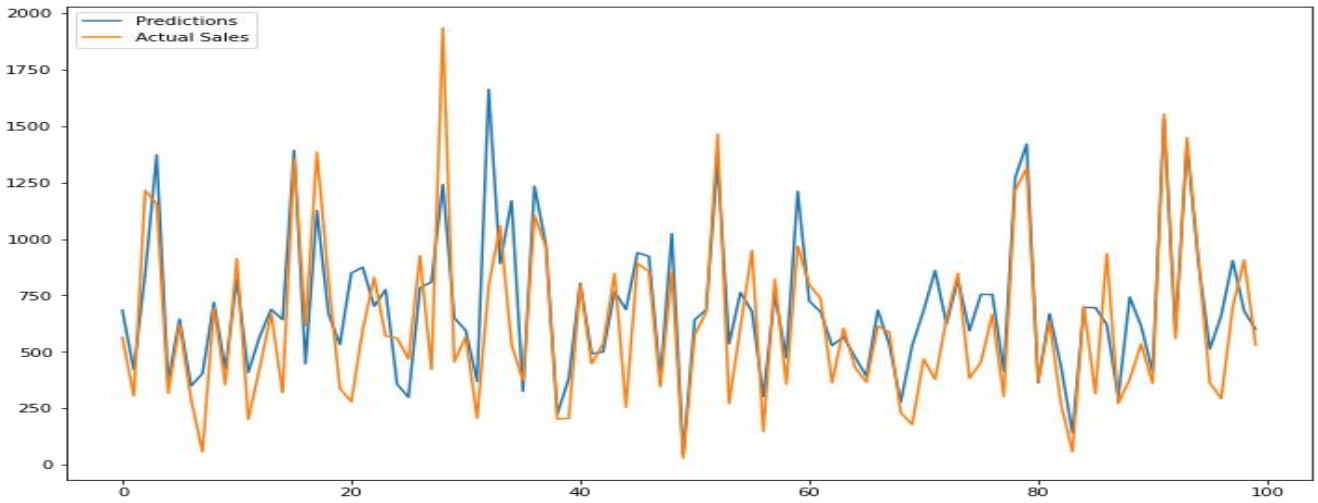
Extreme Gradient Boosting (XGBoost) is an open-source library that provides an efficient and effective implementation of the gradient boosting algorithm. It is an efficient implementation of gradient boosting for classification and regression problems. It can also be used for time series forecasting, although it requires that the time series dataset be transformed into a supervised learning problem first. XGBoost provides a highly efficient implementation of the stochastic gradient boosting algorithm and access to a suite of model hyperparameters designed to provide control over the model training process.

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree-based algorithms are considered best-in-class right now. Please see the chart below for the evolution of tree-based algorithms over the years.

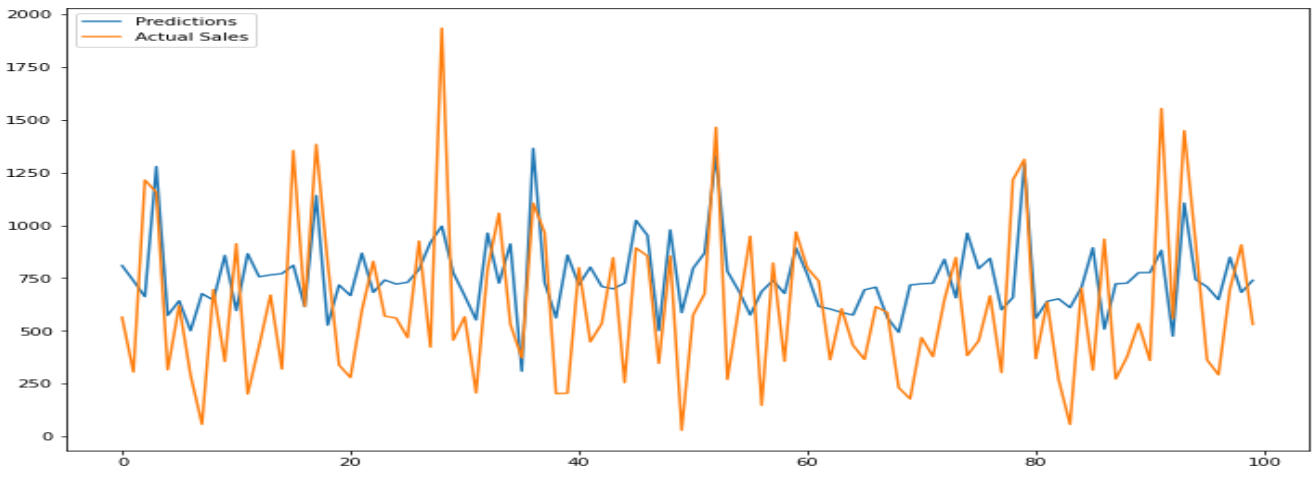
For our dataset this model has performed average. **The R square score for this model we are getting is 0.3168** that means this model is fitted with our dataset with nearly 32% accuracy.

4. RESULTS AND ANALYSIS

4.1. Random Forest Prediction

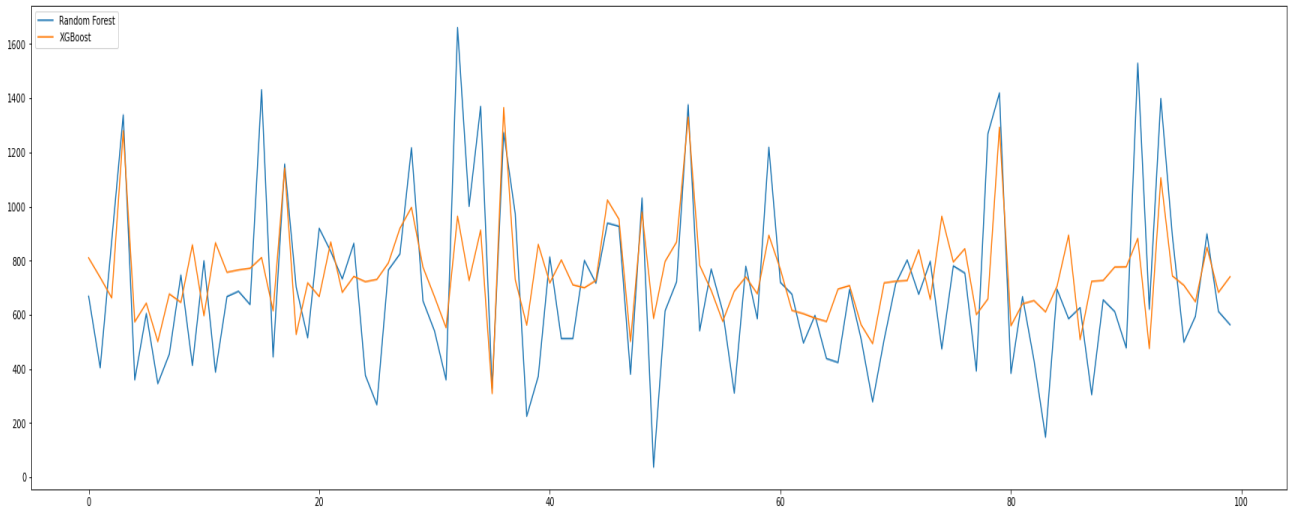


4.2. XGBoost Prediction

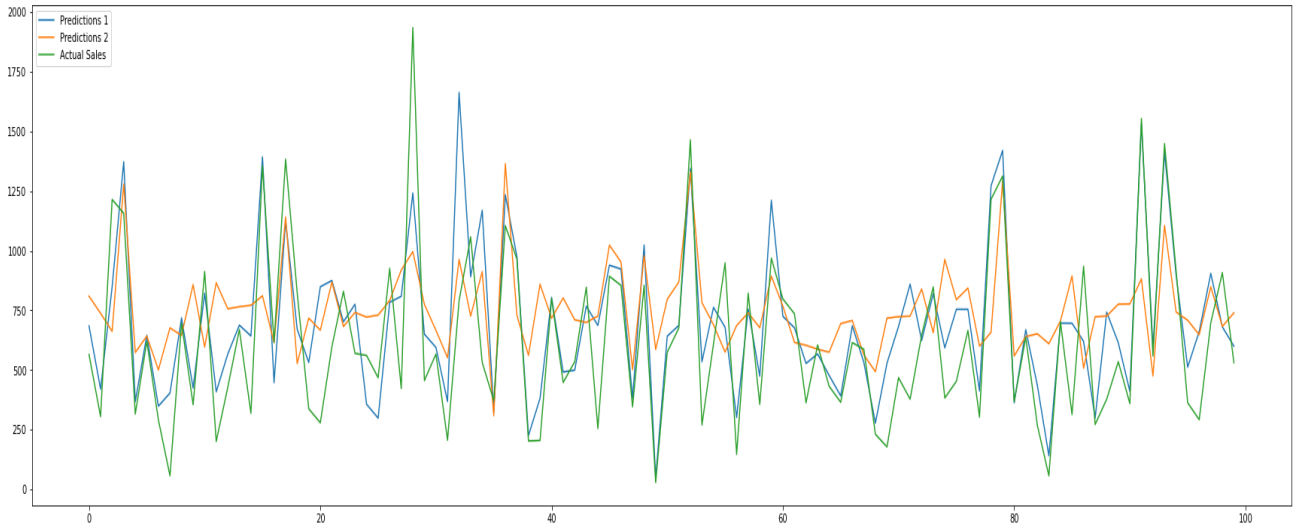


4.3. Comparison

4.3.1 Random forest Vs XGBoost



4.3.2 Overall Comparison



5. CONCLUSION

From above all the fitted model we get the R square scores are:

Model	R Square Score	Accuracy
Linear Regression	0.019	2%
Random Forest	0.6855	69%
XGBoost	0.3168	32%

From above table we conclude that, Random Forest model is best fitted model for our dataset.

6. REFERENCES

6.1. https://doi.org/10.1016/j.ijforecast.2020.02.005 0169-2070/© 2020 International Institute of Forecasters. Published by Elsevier B.V.

6.2. https://builtin.com/data-science/random-forest-algorithm [accessed 10 oct 2021].

6.3. https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d [accessed 12 oct 2021].