Data Mining and Predictive Analysis Lab

PROJECT REPORT

Forecasting Product Demand

By
Devanshi Agarwal – 20095198
Ranjan Satish – 200953202
Gaurav M Gowda - 20095182

Introduction

A product demand forecasting system is an ML-based approach to make accurate forecast for thousands of product demand, based on the previous sales and production records. To implement various deep learning & machine learning algorithms to improve the accuracy of the system. Can be used by companies for forecasting for products for the month after next, and prepare production plan and manage the inventory accordingly, which in turn results in efficient resource utilization and maximizing profits. This is a Time-series problems and it looks at a set of values discovered consecutive overtime and is employed to perform time-based predictions. Forward that past information patterns like level, trend, and seasonality repeat this will produce models mistreatment solely of the info being forecasted to predict future patterns. We have used Random Forest, Linear Regression and Moving average algorithms to predict the order demand for the previous available data and in doing so, you'll be able to anticipate demand fluctuations more effectively.
When done right, anticipating demand can assist you tweak your processes to extend potency right along the provision chain. as a result of you're higher able to predict what customers can wish and once they'll wish it, you will even be able to decrease excess inventory levels, therefore increasing overall gain.

Software Requirements

Jupyter Notebook/Visual studio, Talend, Rapid Miner

Literature review- Forecasts for Product Demand

1) Cao et al [1] made a study that makes the case that, while people do not always know their preferences, they can attempt to do so when given the right incentives. The method we propose controls people's preference-learning incentives using a single parameter, the realization probability, which is the likelihood that a person would really make the purchase for which she has expressed an interest. They get to a theoretical conclusion on the connection between elicited preferences and realization probability. As a result, we can predict demand in actual buying circumstances. The statistics in the research show that the hypothesis is true and that the suggested method is both reliable and affordable. The main disadvantage is more parameters should be considered besides realization probability

2) Ren et al [2] have used a real-world case study to investigate how a fashion store approaches the problem of demand forecasting. The arrival of big data has caused a revolution in demand forecasting for trendy products and poses a significant threat to conventional forecasting techniques and inventory planning. The fashion business is of particular relevance when evaluating product demand projection because it is recognized short product life cycles, variable, and unexpected demand. The report draws the following conclusions: a) It is worthwhile to research ways to better incorporate customer needs into demand forecasting performance. b) Two step demand forecasting has the drawbacks of inconsistent results and time delay

3) Zhang et al [3] his study uses the Chinese e-commerce site VIPS (Weipinhui) as a case study to examine how sentimental aspects of customer reviews impact product demand on flash sale platforms. The emergence of e-commerce flash discounts as a marketing tactic has changed customers' typical buying behaviours as the digital economy has grown. For businesses that engage in flash sales, this presents fresh obstacles for making decisions. Building an effective product demand forecast study that focuses sales & behaviours is crucial for businesses. The paper uses two emotive dictionaries-based sentiment analysis algorithms. The trials show that the autoregressive model that incorporates sentiment variables outperforms models without sentiment components in terms of forecasting. The trials further demonstrate that demand forecast effects are most accurate when product demand for the preceding two weeks and customer review sentiment elements for the prior week are taken into account. Not taken into consideration is quality of reviews.

4) Fildes et al [4] In this essay, has researched literature on predicting retail demand is reviewed. Strategic decisions regarding location are supported by aggregate forecasting. Forecasts at the product level typically have an impact on choices made at the shop level. Forecasters may have the problem of having too many variables and less data due to the demand-influencing elements, particularly promotional information. The next section of the article assesses the data on comparative forecasting accuracy. Despite the fact that causal models perform better than straightforward benchmarks, sufficient data on machine learning techniques is still lacking. There are many ways of predicting new products are looked at independently, but little data on their efficacy is discovered.

5) Moon et al [5] analysed charging demand by analysing consumers 'charging patterns. With the rise of EV usage, demand for electricity generation also increased. To meet the demand for electricity, there was need to predict the charging demands for EV accurately. Here are some observations noted: people prefer charging during evening hours, People prefer electric vehicle which is faster supply during load time(peak), trade-offs between price and charging time. The model used to solve id divided into two steps, step1: forecasting total electricity demanded by EV adoption-which considers constructing scenarios for technological improvements of EV, estimating customers preference for new vehicle, assuming consumers driving pattern and fuel efficiency of EV, step2: estimating consumers preference for types of public EVSEs and analysing consumers charging pattern. Limitations include we assumed that all cunsumers and vehicle travel the same distance daily.

6) Steenbergen et al.[6] proposed Machine learning based approach for Forecasting demand for new products, by combining data(sales) of recently introduced products, product features. Forecasting demand for new products is quite challenging because like other regular products they do not have a track record of sales in the past, which is a key indicator for predicting future sales. A poor forecast has a huge impact on a company's profitability and change of customer perception of the product(brand). The methods include Random Forest and Quantile Regression Forest (QRF) algorithms. K-means is employed for clustering the demand and it is a greedy algorithm. Performance and advantage: Predictions that are accurate and inventory saving. Short item life cycle, new item forecasting gains and describes the concept of feature importance. Providesa higher forecasting quality, high reliable prediction calculations and service.

7) Mostard et al [7] used generic methods for forecasting demand of single period-products. Reducing life-cycles of the product, high variety of products, sourcing and manufacturing are the challenges in apparel industry. For single-period products, companies have to place order pre-season. Data can be obtained from the customers for the pre-order and judgements can be obtained by company experts. The methods are generic and can be used for any single-period product with certain alterations. A priori methods is used to tackle this problem, these methods use either historical data or expert judgment. Company frequently has significant leftovers of individual SKUs which cannot be carried over to the next season and need to be sold at high markdowns. Apriori algorithm is based on the use of items by a consumers for calculating the demand. It’s generic method & SKUs are very well categorized into proper and correct groups

8) Chen et al [8] In his research, he concentrated on a optimizing and computing the framework that incorporates discrete event demand forecasting, and decentralized planning models that can pool demand profiles with many probabilities The cost and time involved in bringing a new medicine to market are significant in the pharmaceutical sector. A critical and relatively expensive step in the development process are clinical trials. To determine sequence, volume, production, they are required. This strategy aids in solving the clinical trial supply chain management issue. There are Three valid case studies with different types of demand described and that are compared to assess the effectiveness of aforementioned method. There is a trade-off between costs of supply chain and customer service level is a major disadvantage.

9) . Bedi et al [9] carried out the best forecasting of power consumption, using a variety of simulation tools methodologies based on artificial intelligence are being applied. The demand for the Energy consumption has increased to a higher level as a result of the expansion and rise of the current population combined with the exponential increase in technological advancements in recent decades. The usage of power is at an all-time high in the modern world. The sector has received very little interest or relevance in comparison to other fields, despite the fact that energy is so important and in high demand. In this paper we will examine a, multi-output demand forecasting model. Without any prior assumptions, the model learns previous dependencies of data. .By contrasting the aforementioned method with several models, including recurrent networks, vector machines, we will assess its effectiveness and performance.

10) Glerum et al[10] his study delves into the market and offers a thorough technique for predicting demand for a technology that is still in its infancy, like electric cars. Electric vehicles have been around for a while, but in recent years, their retail sales have sharply increased by three times. When opposed to gasoline or diesel, electric vehicles have significant advantages because they produce no greenhouse pollutants like carbon dioxide. The purpose of the study is to outline a comprehensive technique for predicting the market for electric vehicles. Stated preference, latent variable, and logit models are the ones in use. Disadvantages include Limited range, Long charging time with inefficient battery life, Few charging stations.

Methodology

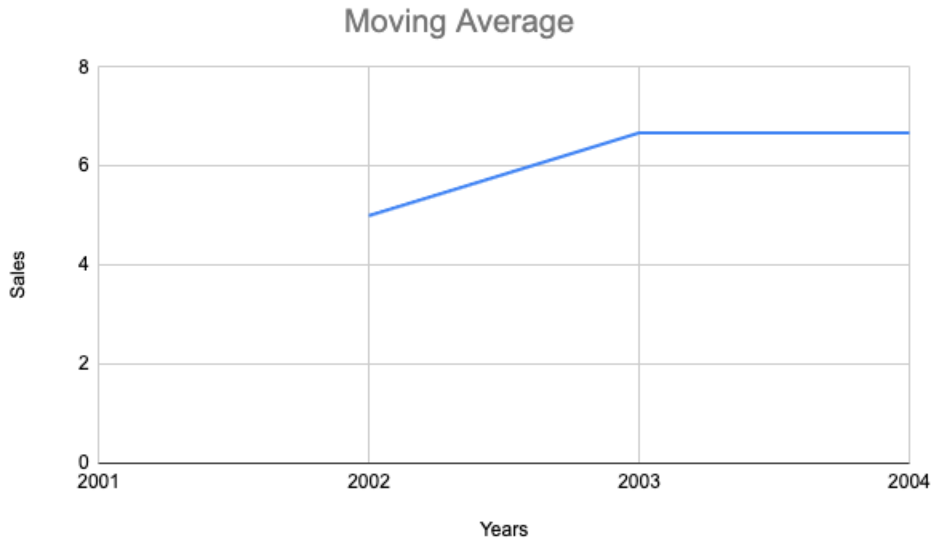
1). Simple Moving Average

A technique based on statistics for predicting trends is moving average. The method entails shifting the range while averaging a group of numbers from a particular range. Suppose that the moving average must be calculated using the sales data for the six years between 2000 and 2005. To construct the moving average, one would average the years 2000 through 2002, 2001 through 2003 and so on.

For example consider the dataset given below.

Year	**Sales (M$)**	**Moving Average (MR)**
2000	4	NaN
2001	7	NaN
2002	4	5
2003	9	6.67
2004	7	6.67
2005	10	

When the simple moving average for above table is plotted, it looks like:



Although alternatives are available, for simple moving average techniques, we will be using SMA for this project.

Simple moving average (SMA): As demonstrated in the example above, for a specified period(3 years interval) of time SMA employed the sliding window. It is an n-record average that is equally weighted. We average all of the previously recorded values and use the result as the subsequent value. It won't be accurate, of course, but it will be near. There are circumstances where this strategy actually performs best as a forecasting method. The formula for calculating the average is as follows, where 'y' is the attribute to be forecasted and 'p' is the finite number of the fixed last values. Therefore, for all $i > p$:

$$\widehat{y_i} = \frac{1}{p}(y_{i-1} + y_{i-2} + y_{i-3} \ldots\ldots + y_{i-p})$$

With time-series data, the this approach is used to smoothen short-term volatility. Long-term trends and forecasting are also made possible thanks to this technique. By using it on the Historical Product Demand dataset, one may estimate future product demand and improve business decisions on production, inventories, and other areas, to name a few.

Steps of the Algorithm

1. Time frame has to be established, for the one we wanted to review
2. Find high price points for the given interval of time.
3. Price points should be summed up as a whole.
4. The previous result has to be divided by number of time intervals.

Advantages of simple moving average forecasts

- Simplifies the analysis of commonly bought goods: When deciding whether to buy a stock for a product that consumers frequently buy, this kind of moving average works well. Food, toiletries, and personal care products are a few examples of this.

- Insights of high and low price points for the product can be derived from the forecast, which is beneficial for stock market traders. They also show support and resistance levels. By using the present resistance or support from buyers and sellers, they can decide whether to buy or sell a stock.

- Provides more accurate data: Since a smooth line is used to represent data points in simple moving average forecasting, there is less room for error during analysis.

Disadvantages of simple moving average forecasts

- Simple moving average forecasts examine long-term stock price trends, making them less desirable for sudden price surges. Because of this, experts often avoid using this kind of moving average when examining price patterns in sectors with significant price swings.

- Using simple moving average, if is difficult to predict the prices in the non-trending or narrow segments of a market. We get a whipsaw-like pattern, cannot derive the prices out of it.

Evaluating Model Performance

By calculating the square root of MSE, RMSE is calculated.
Also known as the Root Mean Square Deviation, RMSE.
The average error magnitude is measured, and the differences from the true value are of concern.
A model has a perfect fit when the RMSE value is 0.
The model and its predictions are better the smaller the RMSE.
A greater RMSE denotes a significant departure between the residual and the ground truth.
The RMSE can be utilised with a variety of features because it aids in determining whether or not the feature enhances the model's prediction.

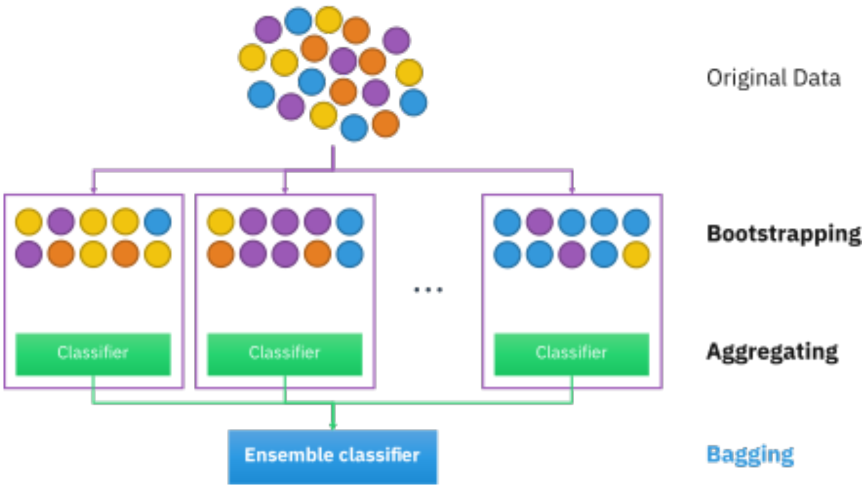
```
RMSE = 46.7284072511
```

2). Random Forest Regression

Random forest algorithm is a supervised ML algorithmic rule that's employed wide in Regression problems. It constructs decision trees for different samples collected and majority vote is taken into consideration for averaging. One of the necessary and important option of the Random Forest is that it can take care of continuous variables of the info set. Better results are expected for issues related to classification.

The Ensemble technique used by Random Forest is *Bagging*, which is also known as *Bootstrap Aggregation.* It creates a special coaching set from sample training dataset which is a subset of actual dataset with replacement & the ultimate output relies on majority vote.

Hence every model is created from the samples of initial dataset with replacement known as row sampling. Row sampling along with the replacement is called as bootstrap. Every model of the samples are trained separately and individually. The output is decided on voting majority after combining the results of all models.



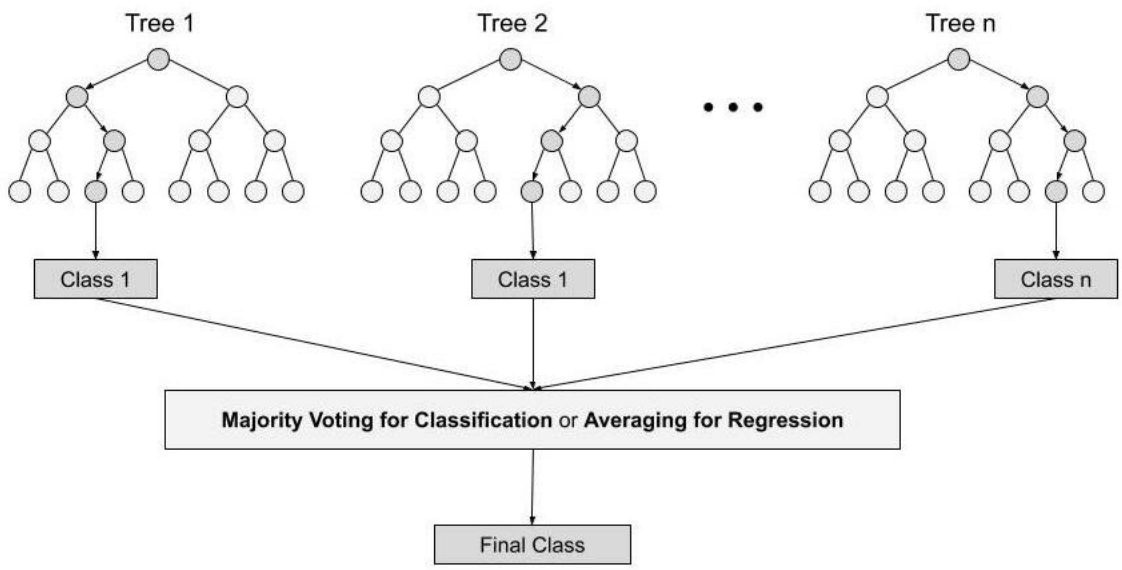
Steps of the Algorithm

1). In Random Forest, sample sets are created from the dataset chosen.

2). Decision tree for each sample generated are created.

3). Each decision tree formed gives a associative output.

4). Final output is based on voting Majority or Averaging.



Features of the algorithm

1). Every tree is different, because not all attributes are taken into consideration while constructing the tree.

2). The feature space is reduced tremendously, as all attributes are not considered and making it immune to the Curse of Dimensionality.

3). Efficient utilization of CPU to build random trees because each tree is independently generated (Parallelization)

4). Very stable because the result is not biased and is outcome of majority vote.

Hyperparameter Tuning for Random Forest

They are used for either improving the performance or power of models or the speed

For increasing Prediction power.

1). n_estimators : Before the averaging is done, the number of trees that are created by the program.

2). max_features: Options that are available while the node is getting split.

3). mini_samples_leaf: Number of leaves required while node(internal) is getting separated.

For increasing the speed.

1). n_jobs: it will use only 1 processor however if the result is -1 there's no limit. It limits the usage of processor to the engine. Default 1 is used, if the result is negative(-1), then there is no limit.

2). random_state– The randomness of the sample taken is controlled by this parameter.

Setting Hyperparameters:

```python
from sklearn.model_selection import RandomizedSearchCV

# Number of trees in random forest
n_estimators = [int(x) for x in np.linspace(start = 50, stop = 250, num = 10)]
# Number of features to consider at every split
max_features = ['auto', 'sqrt']
# Maximum number of levels in tree
max_depth = [int(x) for x in np.linspace(0, 120, num = 20)]
max_depth.append(None)
# Minimum number of samples required to split a node
min_samples_split = [2, 5, 10]
# Minimum number of samples required at each leaf node
min_samples_leaf = [1, 2, 4]
# Method of selecting samples for training each tree
bootstrap = [True, False]
# Create the random grid
random_grid = {'n_estimators': n_estimators,
               'max_features': max_features,
               'max_depth': max_depth,
               'min_samples_split': min_samples_split,
               'min_samples_leaf': min_samples_leaf,
               'bootstrap': bootstrap}
print(random_grid)
```
Python

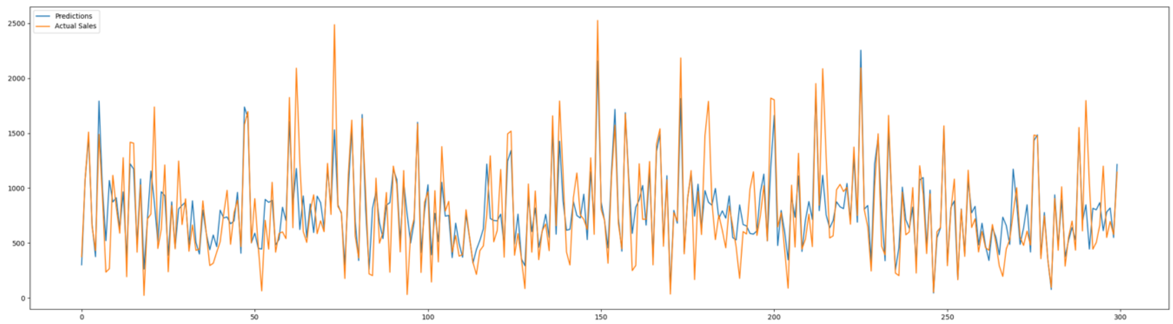
Easy preprocessing is a library used for common ML related pre-processing activities and handling NULL values.
We create a new attribute called 'key' for unique identification. This is done in order to handle the duplicate data in 'Date' column.

Drawbacks of using Random Forest

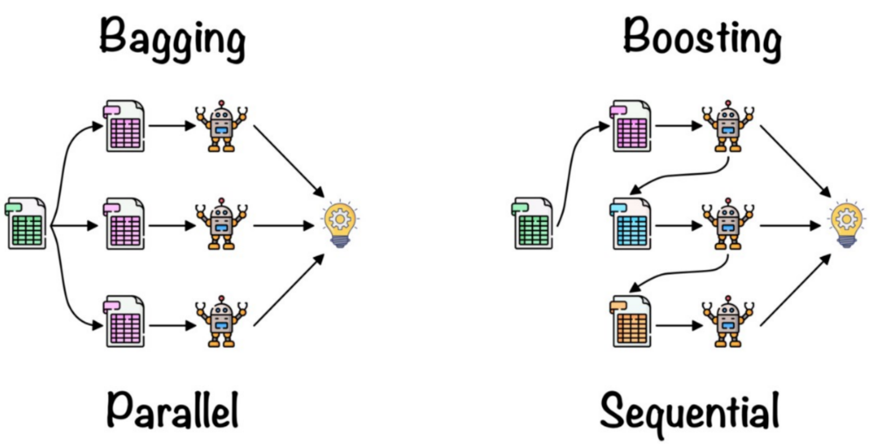
1). Highly complex because the decisions are not made by following the tree path like it's done in decision tree.

2). Due to complexity, the training time is also increased compared to other models.

Expected Output:



How is Random Forest different from XG Boost ?



XG Boost implements Boosted decision trees. Boosting can be a standard boosting formula. In boosting, every predictor corrects its predecessor's error.

Decision trees are successively created in XG boost algorithm. Weights has a important role in XG Boost. Weights are allocated to variables that are then considered by the tree that to make predictions and get results. The variables that are seen as incorrect are considered into the second tree.
This algorithm works on ranking, regression based and classification based problems.

Evaluating Model Performance

To evaluate the model performance, Mean Square Error (MSE) can be used.
A model has a perfect fit when RMSE value is 0.
The validity of the model/The accuracy of the model is greater when the RMSE value is lower I, e. the predictions of the model are better when the RMSE value is lower, and worse when the RMSE value is larger.

$$MSE = \sum_{i=1}^{n} \frac{(Y_{actual} - Y_{predicted})^2}{n}$$

3). Linear Regression

Linear regression is a machine learning algorithm which is used to predict continuous/real or quantitative variables such as sales, salary, age, product price, and so on. It is a technique used for modelling the connection between independent and dependent variables.

Linear Regression Model Formula:

Linear Regression shows the relationship between two variables which may be dependent or independent. The equation of Linear Regression is given by:

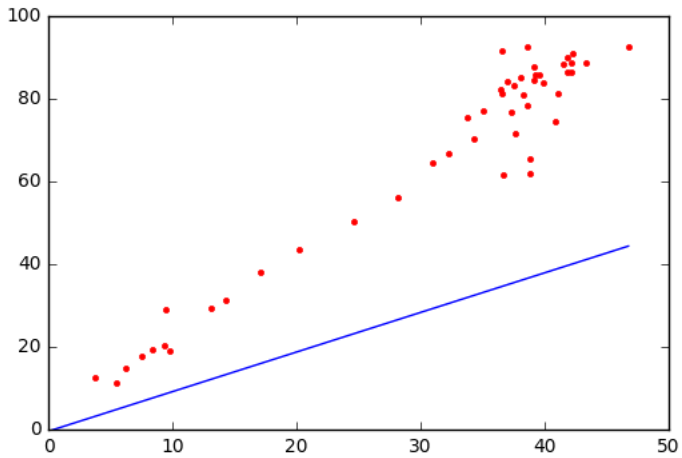
y = a + bx
a and b can be found out using:

$$a = \frac{[(\sum y)(\sum x^2) - (\sum x)(\sum xy)]}{[n(\sum x^2) - (\sum x)^2]}$$

$$b = \frac{[n(\sum xy) - (\sum x)(\sum y)]}{[n(\sum x^2) - (\sum x^2)]}$$

Where X and Y are the variables from the dataset and a, b are the Y-intercept and X-intercept, respectively

Linear regression is widely used in time series forecasting, it is used to predict trends and result of the time series as a trend. The chosen dataset is a time series dataset and hence we will be performing linear regression analysis on it to predict future order demand.



Steps involved in of the Algorithm

- Develop new attributes for sales of previous months.

- Pre-process Data and categorise input and output attributes.

- Store the necessary attributes as NumPy arrays and reshape them.

- Split data set into test and training datasets.

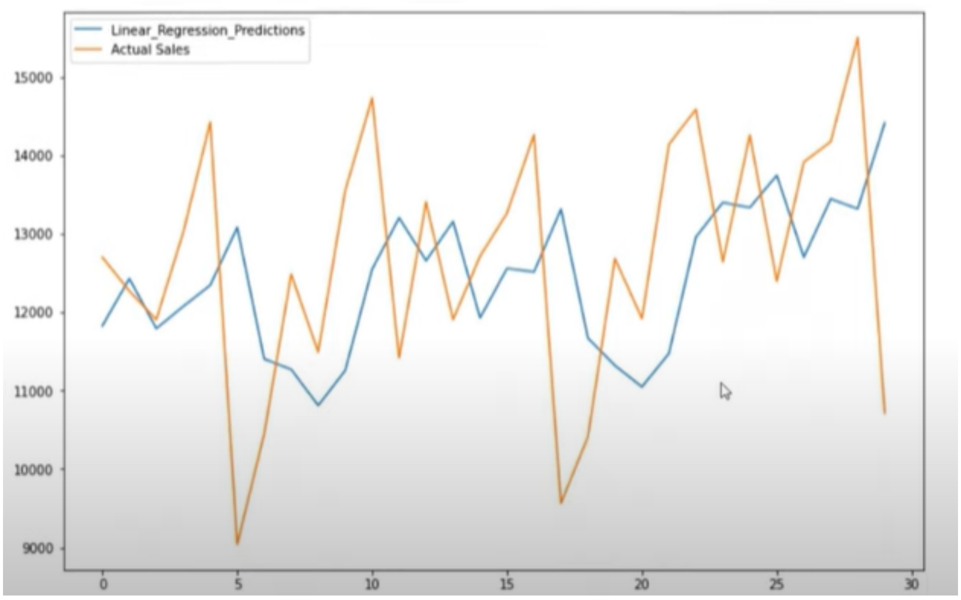
- Fit the model onto the data.

- Make predictions.

Evaluating Model Performance:

To evaluate the model performance, Root mean square error (RMSE) and mean Square Error (MSE) can be used.
A model has a perfect fit when RMSE value is 0.
The validity of the model/The accuracy of the model is greater when the RMSE value is lower I, e. the predictions of the model are better when the RMSE value is lower, and worse when the RMSE value is larger.

Expected Output:



Result

As per the standards ,RMSE values can range between anywhere between 0.3 and 1 which indicates that predicted values of the model is reasonable and can be taken into consideration. Moreover, an adjusted R-squared greater than 0.75 is a very good indication of accuracy. In some cases, an adjusted R-squared of 0.4 or greater is also acceptable.

The RSME value of Moving Average model is: 46.7284

The RSME value of Random Forest Regression model is: 1259.52

The RSME value of Linear Regression model is: 1100.32

Note: The RSME values of the above created models vary drastically from the range of good indication values of accuracy because of the dataset being corrupted and a subset of the dataset was considered for training the model, which cannot yield accurate results as all instances are not considered (less data).

Conclusion

Sure, there must be situations where linear regression outperforms random forest, but I think the most important thing to consider is model complexity. Linear models and Moving average have few parameters, but random forests have many more. This means that random forests overfit more easily than linear regression or moving averages. Averaging makes random forests better than single decision trees, thus improving accuracy and reducing overfitting. The predictions from the random forest regressor are the average of the predictions made by the trees in the forest. But the key advantage of linear or moving average over random forest is they could be used in for anomaly detection because of the extrapolation and needs less data for good results and finally better interpretability.

References for Literature survey

1) Cao, X., & Zhang, J. (2021). Preference learning and demand forecast. *Marketing Science*, *40*(1), 62-79.
2) Ren, S., Chan, H. L., & Siqin, T. (2020). Demand forecasting in retail operations for fashionable products: methods, practices, and real case study. *Annals of Operations Research*, *291*(1), 761-777.
3) Zhang, M., Wang, Y., & Wu, Z. (2021). Data mining algorithm for demand forecast analysis on flash sales platform. *Complexity*, *2021*.
4) Fildes, R., Ma, S., & Kolassa, S. (2019). Retail forecasting: Research and practice. *International Journal of Forecasting*.
5) Moon, H., Park, S. Y., Jeong, C., & Lee, J. (2018). Forecasting electricity demand of electric vehicles by analyzing consumers' charging patterns. *Transportation Research Part D: Transport and Environment*, *62*, 64-79.
6) Van Steenbergen, R. M., & Mes, M. R. (2020). Forecasting demand profiles of new products. *Decision support systems*, *139*, 113401.
7) Mostard, J., Teunter, R., & De Koster, R. (2011). Forecasting demand for single-period products: A case study in the apparel industry. *European Journal of Operational Research*, *211*(1), 139-147.
8) 8) Chen, Y., Mockus, L., Orcun, S., & Reklaitis, G. V. (2012). Simulation-optimization approach to clinical trial supply chain management with demand scenario forecast. *Computers & Chemical Engineering*, *40*, 82-96.
9) Bedi, J., & Toshniwal, D. (2019). Deep learning framework to forecast electricity demand. *Applied energy*, *238*, 1312-1326.
10) Glerum, A., Stankovikj, L., Thémans, M., & Bierlaire, M. (2014). Forecasting the demand for electric vehicles: accounting for attitudes and perceptions. *Transportation science*, *48*(4), 483-499.

Reference:

https://study.com/academy/lesson/demand-forecasting-techniques-moving-average-exponential-smoothing.html - :~:text=A moving average is a,select set of time periods.

https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/moving-average/

https://byjus.com/maths/linear-regression/

https://www.vedantu.com/formula/linear-regression-formula

https://www.bing.com/search?q=linear+regression+time+series+forecasting&FORM=HDRSC1

https://www.geeksforgeeks.org/calculate-mse-for-random-forest-in-r-using-package-randomforest/

https://www.geeksforgeeks.org/xgboost/

https://stackoverflow.com/questions/51037363/linear-regression-vs-random-forest-performance-accuracy

Forecasting Product Demand

by Ranjan Satish

Submission date: 05-Nov-2022 04:28PM (UTC+0800)
Submission ID: 1945234215
File name: DMPA_FinalReport.docx (1.37M)
Word count: 3647
Character count: 20193

Data Mining and Predictive Analysis Lab

PROJECT REPORT

Forecasting Product Demand

By
Devanshi Agarwal – 20095198
Ranjan Satish – 200953202
Gaurav M Gowda - 20095182

Introduction

A product demand forecasting system is an ML-based approach to make accurate forecast for thousands of product demand, based on the previous sales and production records. To implement various deep learning & machine learning algorithms to improve the accuracy of the system. Can be used by companies for forecasting for products for the month after next, and prepare production plan and manage the inventory accordingly, which in turn results in efficient resource utilization and maximizing profits. This is a Time-series problems and it looks at a set of values discovered consecutive overtime and is employed to perform time-based predictions. Forward that past information patterns like level, trend, and seasonality repeat this will produce models mistreatment solely of the info being forecasted to predict future patterns. We have used Random Forest, Linear Regression and Moving average algorithms to predict the order demand for the previous available data and in doing so, you'll be able to anticipate demand fluctuations more effectively.
When done right, anticipating demand can assist you tweak your processes to extend potency right along the provision chain. as a result of you're higher able to predict what customers can wish and once they'll wish it, you will even be able to decrease excess inventory levels, therefore increasing overall gain.

Software Requirements

Jupyter Notebook/Visual studio, Talend, Rapid Miner

Literature review- Forecasts for Product Demand

1) Cao et al [1] made a study that makes the case that, while people do not always know their preferences, they can attempt to do so when given the right incentives. The method we propose controls people's preference-learning incentives using a single parameter, the realization probability, which is the likelihood that a person would really make the purchase for which she has expressed an interest. They get to a theoretical conclusion on the connection between elicited preferences and realization probability. As a result, we can predict demand in actual buying circumstances. The statistics in the research show that the hypothesis is true and that the suggested method is both reliable and affordable. The main disadvantage is more parameters should be considered besides realization probability

2) Ren et al [2] have used a real-world case study to investigate how a fashion store approaches the problem of demand forecasting. The arrival of big data has caused a revolution in demand forecasting for trendy products and poses a significant threat to conventional forecasting techniques and inventory planning. The fashion business is of particular relevance when evaluating product demand projection because it is recognized short product life cycles, variable, and unexpected demand. The report draws the following conclusions: a) It is worthwhile to research ways to better incorporate customer needs into demand forecasting performance. b) Two step demand forecasting has the drawbacks of inconsistent results and time delay

3) Zhang et al [3] his study uses the Chinese e-commerce site VIPS (Weipinhui) as a case study to examine how sentimental aspects of customer reviews impact product demand on flash sale platforms. The emergence of e-commerce flash discounts as a marketing tactic has changed customers' typical buying behaviours as the digital economy has grown. For businesses that engage in flash sales, this presents fresh obstacles for making decisions. Building an effective product demand forecast study that focuses sales & behaviours is crucial for businesses. The paper uses two emotive dictionaries-based sentiment analysis algorithms. The trials show that the autoregressive model that incorporates sentiment variables outperforms models without sentiment components in terms of forecasting. The trials further demonstrate that demand forecast effects are most accurate when product demand for the preceding two weeks and customer review sentiment elements for the prior week are taken into account. Not taken into consideration is quality of reviews.

4) Fildes et al [4] In this essay, has researched literature on predicting retail demand is reviewed. Strategic decisions regarding location are supported by aggregate forecasting. Forecasts at the product level typically have an impact on choices made at the shop level. Forecasters may have the problem of having too many variables and less data due to the demand-influencing elements, particularly promotional information. The next section of the article assesses the data on comparative forecasting accuracy. Despite the fact that causal models perform better than straightforward benchmarks, sufficient data on machine learning techniques is still lacking. There are many ways of predicting new products are looked at independently, but little data on their efficacy is discovered.

5) Moon et al [5] analysed charging demand by analysing consumers ‘charging patterns. With the rise of EV usage, demand for electricity generation also increased. To meet the demand for electricity, there was need to predict the charging demands for EV accurately. Here are some observations noted: people prefer charging during evening hours, People prefer electric vehicle which is faster supply during load time(peak), trade-offs between price and charging time. The model used to solve id divided into two steps, step1: forecasting total electricity demanded by EV adoption-which considers constructing scenarios for technological improvements of EV, estimating customers preference for new vehicle, assuming consumers driving pattern and fuel efficiency of EV, step2: estimating consumers preference for types of public EVSEs and analysing consumers charging pattern. Limitations include we assumed that all cunsumers and vehicle travel the same distance daily.

6) Steenbergen et al.[6] proposed Machine learning based approach for Forecasting demand for new products, by combining data(sales) of recently introduced products, product features. Forecasting demand for new products is quite challenging because like other regular products they do not have a track record of sales in the past, which is a key indicator for predicting future sales. A poor forecast has a huge impact on a company’s profitability and change of customer perception of the product(brand). The methods include Random Forest and Quantile Regression Forest (QRF) algorithms. K-means is employed for clustering the demand and it is a greedy algorithm. Performance and advantage: Predictions that are accurate and inventory saving. Short item life cycle, new item forecasting gains and describes the concept of feature importance. Providesa higher forecasting quality, high reliable prediction calculations and service.

7) Mostard et al [7] used generic methods for forecasting demand of single period-products. Reducing life-cycles of the product, high variety of products, sourcing and manufacturing are the challenges in apparel industry. For single-period products, companies have to place order pre-season. Data can be obtained from the customers for the pre-order and judgements can be obtained by company experts. The methods are generic and can be used for any single-period product with certain alterations. A priori methods is used to tackle this problem, these methods use either historical data or expert judgment. Company frequently has significant leftovers of individual SKUs which cannot be carried over to the next season and need to be sold at high markdowns. Apriori algorithm is based on the use of items by a consumers for calculating the demand. It's generic method & SKUs are very well categorized into proper and correct groups

8) Chen et al [8] In his research, he concentrated on a optimizing and computing the framework that incorporates discrete event demand forecasting, and decentralized planning models that can pool demand profiles with many probabilities The cost and time involved in bringing a new medicine to market are significant in the pharmaceutical sector. A critical and relatively expensive step in the development process are clinical trials. To determine sequence, volume, production, they are required. This strategy aids in solving the clinical trial supply chain management issue. There are Three valid case studies with different types of demand described and that are compared to assess the effectiveness of aforementioned method. There is a trade-off between costs of supply chain and customer service level is a major disadvantage.

9) . Bedi et al [9] carried out the best forecasting of power consumption, using a variety of simulation tools methodologies based on artificial intelligence are being applied. The demand for the Energy consumption has increased to a higher level as a result of the expansion and rise of the current population combined with the exponential increase in technological advancements in recent decades. The usage of power is at an all-time high in the modern world. The sector has received very little interest or relevance in comparison to other fields, despite the fact that energy is so important and in high demand. In this paper we will examine a, multi-output demand forecasting model. Without any prior assumptions, the model learns previous dependencies of data. .By contrasting the aforementioned method with several models, including recurrent networks, vector machines, we will assess its effectiveness and performance.

10) Glerum et al[10] his study delves into the market and offers a thorough technique for predicting demand for a technology that is still in its infancy, like electric cars. Electric vehicles have been around for a while, but in recent years, their retail sales have sharply increased by three times. When opposed to gasoline or diesel, electric vehicles have significant advantages because they produce no greenhouse pollutants like carbon dioxide. The purpose of the study is to outline a comprehensive technique for predicting the market for electric vehicles. Stated preference, latent variable, and logit models are the ones in use. Disadvantages include Limited range, Long charging time with inefficient battery life, Few charging stations.

References

1) Cao, X., & Zhang, J. (2021). Preference learning and demand forecast. *Marketing Science*, *40*(1), 62-79.
2) Ren, S., Chan, H. L., & Siqin, T. (2020). Demand forecasting in retail operations for fashionable products: methods, practices, and real case study. *Annals of Operations Research*, *291*(1), 761-777.
3) Zhang, M., Wang, Y., & Wu, Z. (2021). Data mining algorithm for demand forecast analysis on flash sales platform. *Complexity*, *2021*.
4) Fildes, R., Ma, S., & Kolassa, S. (2019). Retail forecasting: Research and practice. *International Journal of Forecasting*.
5) Moon, H., Park, S. Y., Jeong, C., & Lee, J. (2018). Forecasting electricity demand of electric vehicles by analyzing consumers' charging patterns. *Transportation Research Part D: Transport and Environment*, *62*, 64-79.
6) Van Steenbergen, R. M., & Mes, M. R. (2020). Forecasting demand profiles of new products. *Decision support systems*, *139*, 113401.
7) Mostard, J., Teunter, R., & De Koster, R. (2011). Forecasting demand for single-period products: A case study in the apparel industry. *European Journal of Operational Research*, *211*(1), 139-147.
8) 8) Chen, Y., Mockus, L., Orcun, S., & Reklaitis, G. V. (2012). Simulation-optimization approach to clinical trial supply chain management with demand scenario forecast. *Computers & Chemical Engineering*, *40*, 82-96.
9) Bedi, J., & Toshniwal, D. (2019). Deep learning framework to forecast electricity demand. *Applied energy*, *238*, 1312-1326.
10) Glerum, A., Stankovikj, L., Thémans, M., & Bierlaire, M. (2014). Forecasting the demand for electric vehicles: accounting for attitudes and perceptions. *Transportation science*, *48*(4), 483-499.

Methodology

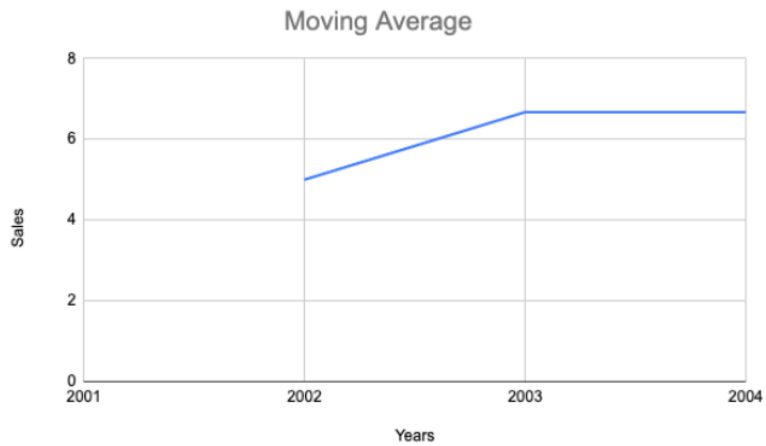
1). Simple Moving Average

A technique based on statistics for predicting trends is moving average. The method entails shifting the range while averaging a group of numbers from a particular range. Suppose that the moving average must be calculated using the sales data for the six years between 2000 and 2005. To construct the moving average, one would average the years 2000 through 2002, 2001 through 2003 and so on.

For example consider the dataset given below.

Year	**Sales (M$)**	**Moving Average (MR)**
2000	4	NaN
2001	7	NaN
2002	4	5
2003	9	6.67
2004	7	6.67
2005	10	

When the simple moving average for above table is plotted, it looks like:



Although alternatives are available, for simple moving average techniques, we will be using SMA for this project.

Simple moving average (SMA): As demonstrated in the example above, for a specified period(3 years interval) of time SMA employed the sliding window. It is an n-record average that is equally weighted. We average all of the previously recorded values and use the result as the subsequent value. It won't be accurate, of course, but it will be near. There are circumstances where this strategy actually performs best as a forecasting method. The formula for calculating the average is as follows, where 'y' is the attribute to be forecasted and 'p' is the finite number of the fixed last values. Therefore, for all i > p:

$$\hat{y}_i = \frac{1}{p}(y_{i-1} + y_{i-2} + y_{i-3} \ldots\ldots + y_{i-p})$$

With time-series data, the this approach is used to smoothen short-term volatility. Long-term trends and forecasting are also made possible thanks to this technique. By using it on the Historical Product Demand dataset, one may estimate future product demand and improve business decisions on production, inventories, and other areas, to name a few.

Steps of the Algorithm

1. Time frame has to be established, for the one we wanted to review
2. Find high price points for the given interval of time.
3. Price points should be summed up as a whole.
4. The previous result has to be divided by number of time intervals.

Advantages of simple moving average forecasts

- Simplifies the analysis of commonly bought goods: When deciding whether to buy a stock for a product that consumers frequently buy, this kind of moving average works well. Food, toiletries, and personal care products are a few examples of this.

- Insights of high and low price points for the product can be derived from the forecast, which is beneficial for stock market traders. They also show support and resistance levels. By using the present resistance or support from buyers and sellers, they can decide whether to buy or sell a stock.

- Provides more accurate data: Since a smooth line is used to represent data points in simple moving average forecasting, there is less room for error during analysis.

Disadvantages of simple moving average forecasts

- Simple moving average forecasts examine long-term stock price trends, making them less desirable for sudden price surges. Because of this, experts often avoid using this kind of moving average when examining price patterns in sectors with significant price swings.

- Using simple moving average, if is difficult to predict the prices in the non-trending or narrow segments of a market. We get a whipsaw-like pattern, cannot derive the prices out of it.

Evaluating Model Performance

By calculating the square root of MSE, RMSE is calculated.
Also known as the Root Mean Square Deviation, RMSE.
The average error magnitude is measured, and the differences from the true value are of concern.
A model has a perfect fit when the RMSE value is 0.
The model and its predictions are better the smaller the RMSE.
A greater RMSE denotes a significant departure between the residual and the ground truth.
The RMSE can be utilised with a variety of features because it aids in determining whether or not the feature enhances the model's prediction.

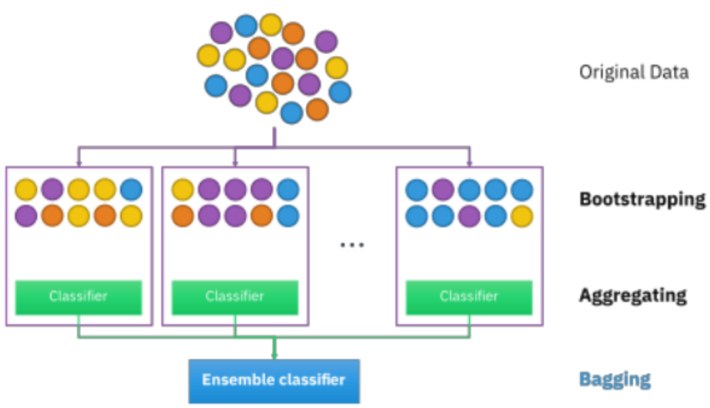
```
RMSE = 46.7284072511
```

2). Random Forest Regression

Random forest algorithm is a supervised ML algorithmic rule that's employed wide in Regression problems. It constructs decision trees for different samples collected and majority vote is taken into consideration for averaging. One of the necessary and important option of the Random Forest is that it can take care of continuous variables of the info set. Better results are expected for issues related to classification.

The Ensemble technique used by Random Forest is *Bagging*, which is also known as *Bootstrap Aggregation*. It creates a special coaching set from sample training dataset which is a subset of actual dataset with replacement & the ultimate output relies on majority vote.

Hence every model is created from the samples of initial dataset with replacement known as row sampling. Row sampling along with the replacement is called as bootstrap. Every model of the samples are trained separately and individually. The output is decided on voting majority after combining the results of all models.



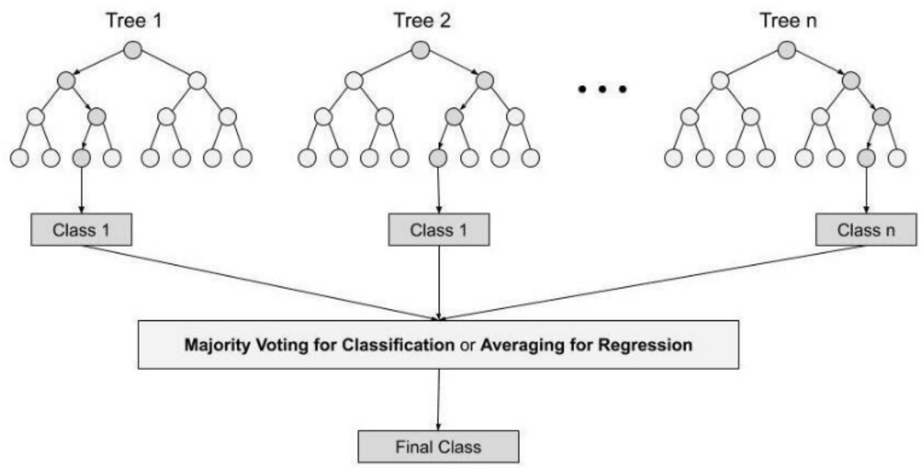
Steps of the Algorithm

1). In Random Forest, sample sets are created from the dataset chosen.

2). Decision tree for each sample generated are created.

3). Each decision tree formed gives a associative output.

4). Final output is based on voting Majority or Averaging.



Features of the algorithm

1). Every tree is different, because not all attributes are taken into consideration while constructing the tree.

2). The feature space is reduced tremendously, as all attributes are not considered and making it immune to the Curse of Dimensionality.

3). Efficient utilization of CPU to build random trees because each tree is independently generated (Parallelization)

4). Very stable because the result is not biased and is outcome of majority vote.

Hyperparameter Tuning for Random Forest

They are used for either improving the performance or power of models or the speed

For increasing Prediction power.

1). n_estimators : Before the averaging is done, the number of trees that are created by the program.

2). max_features: Options that are available while the node is getting split.

3). mini_samples_leaf: Number of leaves required while node(internal) is getting separated.

For increasing the speed.

1). n_jobs: it will use only 1 processor however if the result is -1 there's no limit. It limits the usage of processor to the engine. Default 1 is used, if the result is negative(-1), then there is no limit.

2). random_state– The randomness of the sample taken is controlled by this parameter.

Setting Hyperparameters:

```python
from sklearn.model_selection import RandomizedSearchCV

# Number of trees in random forest
n_estimators = [int(x) for x in np.linspace(start = 50, stop = 250, num = 10)]
# Number of features to consider at every split
max_features = ['auto', 'sqrt']
# Maximum number of levels in tree
max_depth = [int(x) for x in np.linspace(0, 120, num = 20)]
max_depth.append(None)
# Minimum number of samples required to split a node
min_samples_split = [2, 5, 10]
# Minimum number of samples required at each leaf node
min_samples_leaf = [1, 2, 4]
# Method of selecting samples for training each tree
bootstrap = [True, False]
# Create the random grid
random_grid = {'n_estimators': n_estimators,
               'max_features': max_features,
               'max_depth': max_depth,
               'min_samples_split': min_samples_split,
               'min_samples_leaf': min_samples_leaf,
               'bootstrap': bootstrap}
print(random_grid)
```
Python

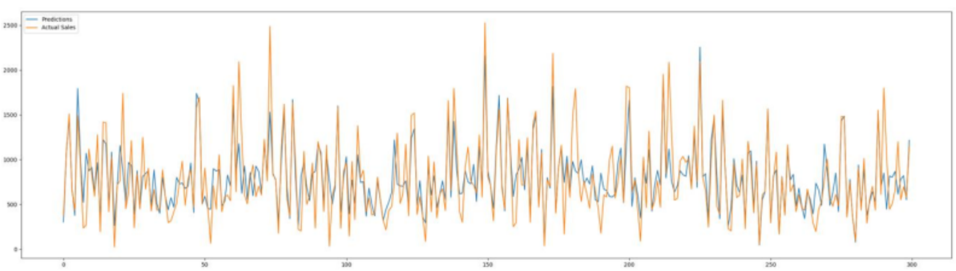
Easy preprocessing is a library used for common ML related pre-processing activities and handling NULL values.
We create a new attribute called 'key' for unique identification. This is done in order to handle the duplicate data in 'Date' column.

Drawbacks of using Random Forest

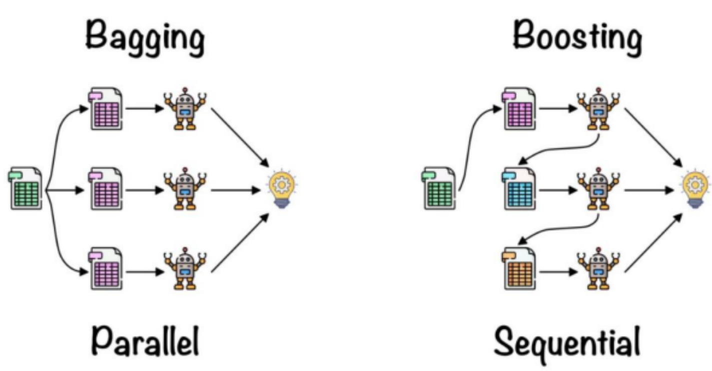
1). Highly complex because the decisions are not made by following the tree path like it's done in decision tree.

2). Due to complexity, the training time is also increased compared to other models.

Expected Output:



How is Random Forest different from XG Boost ?



XG Boost implements Boosted decision trees. Boosting can be a standard boosting formula. In boosting, every predictor corrects its predecessor's error.

Decision trees are successively created in XG boost algorithm. Weights has a important role in XG Boost. Weights are allocated to variables that are then considered by the tree that to make predictions and get results. The variables that are seen as incorrect are considered into the second tree.
This algorithm works on ranking, regression based and classification based problems.

Evaluating Model Performance

10
To evaluate the model performance, Mean Square Error (MSE) can be used.
A m5del has a perfect fit when RMSE value is 0.
The validity of the model/The accuracy of the model is greater when the RMSE value is lower I, e. the predictions of the model are better when the RMSE value is lower, and worse when the RMSE value is larger.

$$MSE = \sum_{i=1}^{n} \frac{(Y_{actual} - Y_{predicted})^2}{n}$$

3). Linear Regression

Linear regression is a machine learning algorithm which is used to predict continuous/real or quantitative variables such as sales, salary, age, product price, and so on. It is a technique used for modelling the connection between independent and dependent variables.

Linear Regression Model Formula:

Linear Regression shows the relationship between two variables which may be dependent or independent. The equation of Linear Regression is given by:

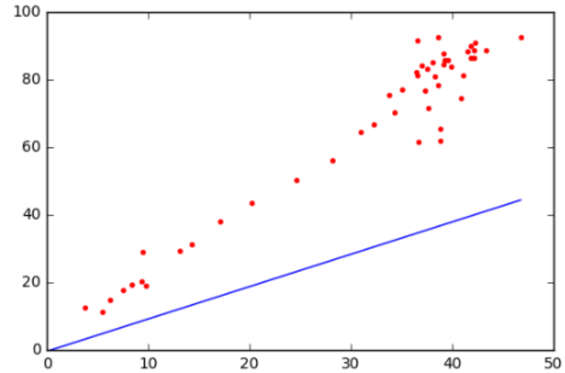
y = a + bx
a and b can be found out using:

$$a = \frac{[(\sum y)(\sum x^2) - (\sum x)(\sum xy)]}{[n(\sum x^2) - (\sum x)^2]}$$

$$b = \frac{[n(\sum xy) - (\sum x)(\sum y)]}{[n(\sum x^2) - (\sum x^2)]}$$

Where X and Y are the variables from the dataset and a, b are the Y-intercept and X-intercept, respectively

Linear regression is widely used in time series forecasting, it is used to predict trends and result of the time series as a trend. The chosen dataset is a time series dataset and hence we will be performing linear regression analysis on it to predict future order demand.



Steps involved in of the Algorithm

- Develop new attributes for sales of previous months.

- Pre-process Data and categorise input and output attributes.

- Store the necessary attributes as NumPy arrays and reshape them.

- Split data set into test and training datasets.

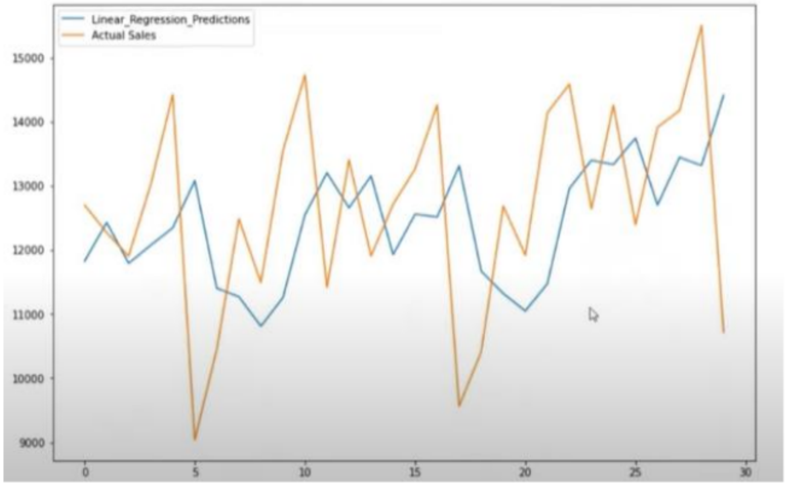
- Fit the model onto the data.

- Make predictions.

Evaluating Model Performance:

To evaluate the model performance, Root mean square error (RMSE) and mean Square Error (MSE) can be used.
A model has a perfect fit when RMSE value is 0.
The validity of the model/The accuracy of the model is greater when the RMSE value is lower I, e. the predictions of the model are better when the RMSE value is lower, and worse when the RMSE value is larger.

Expected Output:



Result

As per the standards ,RMSE values can range between anywhere between 0.3 and 1 which indicates that predicted values of the model is reasonable and can be taken into consideration. Moreover, an adjusted R-squared greater than 0.75 is a very good indication of accuracy. In some cases, an adjusted R-squared of 0.4 or greater is also acceptable.

The RSME value of Moving Average model is: 46.7284

The RSME value of Random Forest Regression model is: 1259.52

The RSME value of Linear Regression model is: 1100.32

Note: The RSME values of the above created models vary drastically from the range of good indication values of accuracy because of the dataset being corrupted and a subset of the dataset was considered for training the model, which cannot yield accurate results as all instances are not considered (less data).

Conclusion

Sure, there must be situations where linear regression outperforms random forest, but I think the most important thing to consider is model complexity. Linear models and Moving average have few parameters, but random forests have many more. This means that random forests overfit more easily than linear regression or moving averages. Averaging makes random forests better than single decision trees, thus improving accuracy and reducing overfitting. The predictions from the random forest regressor are the average of the predictions made by the trees in the forest. But the key advantage of linear or moving average over random forest is they could be used in for anomaly detection because of the extrapolation and needs less data for good results and finally better interpretability.

Reference:

https://study.com/academy/lesson/demand-forecasting-techniques-moving-average-exponential-smoothing.html - :~:text=A moving average is a,select set of time periods.

https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/moving-average/

https://byjus.com/maths/linear-regression/

https://www.vedantu.com/formula/linear-regression-formula

https://www.bing.com/search?q=linear+regression+time+series+forecasting&FORM=HDRSC1

https://www.geeksforgeeks.org/calculate-mse-for-random-forest-in-r-using-package-randomforest/

https://www.geeksforgeeks.org/xgboost/

https://stackoverflow.com/questions/51037363/linear-regression-vs-random-forest-performance-accuracy

Forecasting Product Demand

ORIGINALITY REPORT

6%	3%	4%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Julien Mostard, Ruud Teunter, René de Koster. "Forecasting demand for single-period products: A case study in the apparel industry", European Journal of Operational Research, 2011 Publication	1%
2	vitalflux.com Internet Source	1%
3	philpapers.org Internet Source	1%
4	HyungBin Moon, Stephen Youngjun Park, Changhyun Jeong, Jongsu Lee. "Forecasting electricity demand of electric vehicles by analyzing consumers' charging patterns", Transportation Research Part D: Transport and Environment, 2018 Publication	<1%
5	Mayukh Dass, Masoud Moradi, Fereshteh Zihagh. "Forecasting purchase rates of new products introduced in existing categories", Journal of Marketing Analytics, 2022	<1%

Publication

6	R.M. van Steenbergen, M.R.K. Mes. "Forecasting demand profiles of new products", Decision Support Systems, 2020 Publication	<1 %
7	mdpi-res.com Internet Source	<1 %
8	"Deep Learning: Algorithms and Applications", Springer Science and Business Media LLC, 2020 Publication	<1 %
9	www.voxco.com Internet Source	<1 %
10	archives.njit.edu Internet Source	<1 %
11	answeregy.com Internet Source	<1 %
12	research.utwente.nl Internet Source	<1 %
13	www.researchgate.net Internet Source	<1 %
14	Mingyang Zhang, Yixin Wang, Zhiguo Wu. "Data Mining Algorithm for Demand Forecast Analysis on Flash Sales Platform", Complexity, 2021 Publication	<1 %