Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Predicting Online Sales: A Machine Learning Approach for Sales Forecasting in Online Platform

V Rajeshkumar Pitani[1], Dr.Harsh Lohiya[2]

[1]Research Scholar, Dept. of Computer Science, Sri Satya Sai University of Technology and Medical Sciences, Sehore Bhopal-Indore Road, Madhya Pradesh, India.

[2]Research Guide, Dept. of Computer Science, Sri Satya Sai University of Technology and Medical Sciences, Sehore Bhopal-Indore Road, Madhya Pradesh, India

Article Info
Page Number: 12227-12237
Publication Issue:
Vol. 71 No. 4 (2022)

Article History
Article Received: 15 September 2022
Revised: 24 October 2022
Accepted: 18 November 2022
Publication: 21 December 2022

Abstract
There are very few opportunities left for traditional merchants to grow their revenue through increasing sales as a result of increased sales because online shopping has become such a significant sector in the modern period. It is possible to utilise an algorithm that makes use of machine learning to create predictions regarding the kind of products that ought to be offered during a particular month in order to increase sales overall. When the forecast has been finished, a dashboard will be developed to illustrate which products should have been sold in order to obtain high amounts of revenue. This will be done so that the prediction can be validated. It has been identified how to bill for the sales, and an expert's support was utilised in conducting the analysis. However, in this predicament, not everyone possesses the resources necessary to consult with professionals who are able to aid them. For sellers, experience is an essential qualification to have. People who have only been operating their businesses for a couple of years have very little to no experience and are looking for support. The process of making accurate projections regarding future product sales is a crucial part of effective purchase management. The unpredictability, global scope, and ever-changing nature of the commercial environment in which businesses must compete is one of the most critical challenges that companies must face in today's world. Because customers' expectations about pricing and quality are always expanding, modern manufacturers “can no longer rely only on the cost advantage that they have over their competitors. This is because customers' expectations are consistently becoming more demanding. It is vital to forecast sales in order to maintain suitable inventory stock levels. Estimating the exact future demand for goods has been a constant challenge for firms in all different types of industries”. There is a chance that the overall profit will be put in jeopardy if the commodities are difficult to obtain or if there is an excess of goods available compared to the amount of demand for them.

Keywords: sales, marketing, machine learning

Introduction

In the past, companies would manufacture items without taking into account the total number of sales or the amount of demand for their products. A manufacturer needs data regarding the

demand for the products that are currently on the market in order to make a decision regarding whether or not to increase or decrease the production of a certain number of units. If a business competes in the market without taking these principles into consideration, the business runs the risk of incurring financial losses. When it comes to figuring out their demand and sales, numerous companies make use of a wide range of distinct factors [1]. In today's highly competitive environment and constantly shifting consumer landscape, doing accurate and timely revenue forecasting, also known as sales forecasting, can provide important information to businesses that are engaged in the manufacturing, distribution, or retailing of goods [2]. This type of forecasting is also known as sales forecasting. Long-term predictions can deal with concerns connected to the expansion of a firm as well as decision-making, but short-term forecasts can assist greatly with production planning and stock management [1]. In industries like manufacturing and retail, where a large percentage of the products have a limited shelf life, accurate sales forecasting is even more important than usual. Depending on the conditions, this can lead to a decrease in revenue, either because there is a scarcity of inventory or because there is an excess of inventory. If there are too many orders, there won't be enough products, but if there aren't enough orders, there won't be enough opportunities. If there aren't enough orders, there won't be enough opportunities. Because of this, the level of competition in the food industry is always changing as a result of a variety of reasons including changes in pricing and advertisement as well as an increase in demand from customers [3]. The vast majority of the time, managers will make their sales predictions without giving any thought to the process. On the other hand, professional managers are getting harder and harder to find, and they aren't always reachable (e.g., they can get sick or leave). Forecasting potential sales through the use of computer systems is something that can be done. When qualified managers are unavailable, these computer systems can either take their place or assist managers in making the best decision feasible by providing potential sales predictions. [4] One approach to put such a method into practise is to attempt to replicate the skills of professional managers within a computer programme. This is also one of the ways that such a method can be applied. Another way to put such a method into practise is to use it. On the other hand, the availability of sales data and information related to it can be put to use through the application of techniques related to machine learning in order to automatically generate accurate sales forecasting models. This can be accomplished by applying the machine learning techniques. This tactic is a lot less difficult to implement. It is not in any way affected by the particulars of a single sales manager, and it is adaptable, which means that it can react to changes in the data. Additionally, it is not influenced in any way by the particulars of a single customer.

On the other hand, there is a possibility that it will exaggerate the accuracy of the prognosis that was made by a human expert, which is normally simply an estimate. This is because forecasts are typically merely estimates. For instance, earlier in the day, it was common practise for enterprises to produce items without taking into account the total number of sales or the degree of demand, despite the fact that this practise led to a variety of problems. It is hard for any producer to decide whether the quantity of units should be increased or lowered without data indicating the amount of demand for things coming from customers. This is because they do not know how much they will sell, and they cannot predict how much they

will sell. When competing in the market, companies will wind up incurring monetary losses if they do not take into account the fundamentals that have been discussed here. Numerous companies rely on a diverse set of qualities when it comes to establishing their target audience and tracking their sales. There are a variety of approaches to forecasting sales, and in the past, businesses have focused their efforts on a wide range of statistical models, such as time series and linear regression, feature engineering, and random forest models, in order to obtain a forecast of future sales and demand. There are a number of methods for forecasting sales, and there are also a variety of approaches to forecasting sales. These techniques are broken down in greater detail further on.

A time series is a collection of data points that are preserved for a given amount of time and are used to create predictions about the future of the system. A time series is a collection of data points that are obtained over the course of a period at successive points that are evenly spaced apart from one another. These data points are collected over the course of a time period. Some of the most important factors to research and look at are things like patterns, seasonality, irregularity, and cyclicity. [Clarification needed] When making predictions about future values based on historical data, the mathematical technique known as linear regression can be utilised as an effective tool. It may be helpful in discovering the underlying trends and solving problems that include inflated rate estimates [5, 6].

The process of using data on domain expertise and the production of features with the goal of making predictive models produced with machine learning more accurate is referred to as "feature engineering." Feature engineering makes use of data. It makes it possible to conduct a thorough examination of the facts and to do it from a viewpoint that is more illuminating [7]. The usage of a decision tree as the foundational structure of a random forest model is essential to the success of the model. The decision tree technique is a methodology that is utilised in data mining for the aim of anticipating and classifying the data that has been collected. The decision tree method does not provide any conceptual understanding of the topic that is being discussed at this time. The more sophisticated method known as random forest allows for the combination of the results of numerous separate trees in order to get a conclusion.

The random forest model's strategy of taking an average of the predictions provided by each individual tree results in more accurate forecasts being generated as a result of the model's application of this method. In order to properly organise the information, the whole data set is often divided into two portions, which are referred to as the training data section and the test data section. The term "training data" refers to the information that is utilised in the process of instructing a model, whereas the term "test data" describes the information that is utilised in the process of evaluating a model that has been instructed. A common approach to data analysis is known as the 80-20 split. According to this approach, “80 percent of the data is used to train the model, and the remaining 20 percent of the data is utilised to test the model. On the other hand, it is well known that certain techniques, such as stratified K-fold cross-validation, give adequate results. There were many more variations of cross-validation, such as simple k-folds, leave one out, stratified k-fold cross-validation and a great deal of others [8, 9]”.

Objectives

1. The process of transforming data into a format that is suitable for the application of machine learning algorithms by making use of a variety of pre-processing methods.
2. To discover which Machine Learning algorithm is most suited for sales forecasting.

A. Challenges in sales of products

E-commerce is the practise of conducting commercial transactions through the use of the internet and purchasing items from any location in the world. E-commerce also refers to the practise of selling things to customers all over the world. People are increasingly purchasing their day-to-day requirements online, especially during this COVID-19 time period, so that they can do it without having to leave the convenience of their own homes. This trend is especially prevalent in the United States. Using the expertise of people trained in business management and a wide range of information technology (IT) solutions, countless businesses are working toward the goal of selling their commodities through websites that specialise in online commerce. There are many IT solution providers who are currently working on e-commerce technologies in order to assist their customers in the process of developing e-commerce websites and selling their products online. These IT solution providers are working on e-commerce technologies in order to help their customers.

Emerging technologies make it possible for online shoppers to personalise the product they buy, such as by trying on different virtual versions of an article of clothing, customising the décor of a room, or purchasing tickets via virtual tours. For example, buyers can try on different virtual versions of an article of clothing; buyers can also personalise the décor of a room. Because of these qualities, consumers will have a higher level of happiness and excitement towards their online transactions. When designing a website for e-commerce, the designer needs to consider what features need to be included in the website, ensure that the website is easy to use, pick how the products will be displayed, and a variety of other considerations. The fact that everything that is sold through e-commerce is contemporary, current, and available at prices that are cheaper than those that are offered locally confronts the sector with its most major challenge. The task of selling and purchasing the merchandise is handled by large teams of people who are solely dedicated to that responsibility. They generate original strategies, which are then put into action, after which all of the strategies are developed, reviewed, and then eventually presented to the market.

B. Machine learning algorithms

Machine Learning is a subset of Artificial Intelligence (AI) that has the ability to learn from data and assess the knowledge it has obtained. It is also known as ML. The purpose of the field of study known as machine learning (ML) is to instruct computers on how to learn so that they can independently solve difficult issues without requiring any assistance from a person of any kind. There are essentially four classes that are utilised within the realm of machine learning. These classes are:

- Learning with Supervision
- Learning without Supervision
- Learning with Only Partial Supervision
- Learning through Reinforcement

Each of these subfields within machine learning can be used into a wide variety of applications to facilitate the resolution of intricate issues. The customer will be able to store the products in advance and plan how to sell them with the assistance of ML Algorithms, which will assist the client in predicting what should be sold in the future month. There are several distinct categories of machine learning algorithms, each of which can be broken down into the following categories:

- Algorithms for Regression
- Algorithms for Decision Trees
- Bayesian Algorithms
- Clustering Algorithms
- Predictive Algorithms
- "Artificial Neural Network Algorithms"
- Deep Learning Algorithms

"There are so many where machine learning algorithms can be used to build models and some of the applications are"

- Natural language understanding
- Online advertising
- Handwriting recognition
- Economics
- DNA sequence classification
- Fraud detection
- Bioinformatics
- Time series forecasting
- Recommender systems
- Natural language understanding

The majority of us already make use of at least one of machine learning's applications on a day-to-day basis, which has resulted in a simplification of a variety of complex business activities that were previously difficult to do. Learning based on algorithmic processes that are driven by data gives them the ability to hazard accurate judgments about the future world. The company's employees have profited from having features like these so that they may better comprehend the company's sales values. In addition, the level of understanding that exists between the company's clients and its owners has significantly advanced as a direct result of the application of machine learning algorithms. In the 1970s and 1960s, machine learning was already in its infancy, but very few people paid attention to it since they either did not understand it or did not care about it. At this moment, however, it is impossible to

simplify challenging jobs without the assistance of machine learning. Forecasting sales will be accomplished through the application of machine learning algorithms as part of the scope of work for this project.

C. AI for sales

Artificial intelligence (AI)-based methods will make it simpler for the organisation to conduct its operations and engage with its clients and consumers in a way that is both efficient and successful. Businesses have been able to raise their profits while simultaneously cutting their costs thanks to the advent of artificial intelligence (AI). This has been accomplished by replacing certain portions of their operations with technology that is powered by AI. The ability to run analysis on the data, which can then be used to construct marketing plans and make decisions, is made possible by the process of data extraction. AI technologies are utilised so that accurate estimates of future sales can be made. The proprietors of the businesses were able to track and examine the sales of their products for each individual month, as well as check the sales projections for the months that were to follow. Because of this skill, marketing teams are able to determine the cause of a problem if there is a decrease in sales and create a variety of marketing plans with the goal of increasing those sales. In addition, they are able to pinpoint the source of the problem. AI will make it simpler to analyse datasets, which will in turn contribute to the development of predictions and product suggestions. AI will also play a role in the development of product ideas. You have access to a wide variety of cutting-edge AI tactics that you can implement in order to increase the amount of money you make from sales and quicken the expansion of your business. The following is a list of some of the sales techniques that are based on AI:

- Predictive forecasting
- Price optimization
- Upselling and cross-selling ˆ
- Performance management

The use of artificial intelligence could be beneficial to relationships not only between businesses but also between businesses and customers. Recent years have seen a substantial increase in the quantity of research that has been conducted in the field of artificial intelligence. Because AI can aid in doing so and make the process much more efficient overall, it is not required to rely on any software that requires human labour to input the data. This is because AI can do so and makes it possible.

FORECAST FOR BIG MART SALES

the individual who advised making use of this tactic. The first step in this methodology was to pre-process the raw data that was collected from Big Mart in order to search for any anomalies or outliers that might have been overlooked. After that, these data were fed into an algorithm so that it could perform research on them in order to construct a model. The two kinds of algorithms that were put to use are known as random forests and multiple linear regression. The technology known as ETL, which stands for "Extract, Change, and Load," “was utilised in this methodology to extract data from one database and then transform it into

a format that was suitable". ETL is an acronym that stands for "Extract, Change, and Load." A format that allowed for the data to be comprehended was developed from the raw data samples that were initially used. In order to achieve the desired outcomes, the model was applied throughout the analysis process.

B.SALES TIME SERIES FORECASTING

It was Bohdan M. Pavlyshenko who proposed employing this tactic in the first place. During this stage of the procedure, a stacking approach is utilised in order to construct the regression. Research and analysis were done on the process of putting together a collection of individual models. Random forest and regression were the two different kinds of algorithms that were utilised during this process. The findings showed that by utilising stacking approaches, we are able to improve the performance of the predictive model. This was proved by the fact that.

ANALYSIS OF STUDIED SYSTEMS

Sales was based on old dataset and not on user generated data

In previous studies, making predictions with preserved datasets was found to be less accurate than making predictions with new data. The dataset that was utilised in the study was at least two to three years old, and the current sales projections were based on the information that was contained within that dataset. These days, data is being generated at such a breakneck speed, and when compared with data that was recorded in the past, newly generated data is certain to have a great many discrepancies.

"More samples did not improve the accuracy"

The Random Forest algorithm is utilised in a few of the systems. When applied to tiny datasets, random forest provides reliable predictions. On the other hand, the accuracy does not improve with increasing the dataset size if the project is employing a larger one.

"Appropriate parameters were not considered"

There is a certain correlation between the several factors that impact sales. In order to make more accurate predictions, it is necessary to take into account the parameters that have a higher degree of correlation with one another.

PROPOSED SYSTEM

The findings of the study helped to construct a model that would be helpful in the conduct of future business studies for the goal of "predicting product sales in an online environment. The primary objective of the project is to demonstrate that product demands can be anticipated through the comparative influence of promotional marketing strategies such as discounts and the provision of free delivery choices, user-generated content such as the volume and valence of online reviews, and the sentiments of the web reviews. This will be accomplished by analysing the data collected from the project's participants. Asynchronous input and output is utilised by the algorithms so that data can be requested, retrieved, and pre-processed in real-

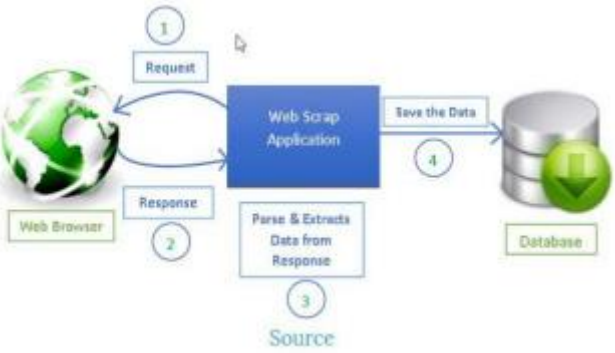
time from Amazon.com utilising a web crawler. After the data have been collected, the text of the testimonials is processed by a natural language algorithm. This step comes after the gathering of the data (NLP). For the purpose of doing further research and analysis, the resultant feeling is categorised as good, negative, or neutral. This study will then use a Multiple Linear Regression to predict product sales, as well as to predict the effects of online sentiments on the same, in order to design effective promotional strategies and sales tactics. This study's purpose is to design these strategies and tactics in order to design effective promotional strategies and sales tactics. In addition, a Multiple Linear Regression will be utilised so that the results of this study can accurately forecast the effects that online sentiments have on the same".

Parameters considered for prediction.

VARIABLES	DESCRIPTION
Discount Value	The monetary value of price deduction from usual price
Discount Rate	The percentage of price deduction from usual price
Current Price	The price of the product
Free Delivery	Whether the product is delivered without delivery fee
Customer Review Rating (Valence)	The accumulated average numeric rating of inline review
Number of customer Reviews (Volume)	The number of all online reviews
Percentage of Negative Review	The proportion of 1-star and 2-star reviews in total reviews
Percentage of Neutral Review	The proportion of 3-star reviews in total reviews
Percentage of Positive Review	The proportion of 4-star and 5-star reviews in total reviews
Review text Sentiment (Sentiment)	The sentiment of most helpful reviews
Number of Answered Questions	The number of answered questions in Customer Question & Answers
Manufacturer	The name of the manufacturer of the product
Sales Rank	The Best Sellers Rank of the product

A. Extraction

In order to complete the process of web scraping, web crawlers will be necessary. It would be necessary to make use of a wrapper programme in order to ascertain the specific locations of the templates within the source. After gathering and copying the relevant real-time data from the web, the information is then saved in a file and subjected to further processing.



"B. Classification"

Algorithm used for Classification:

Natural Language processing algorithms

It is concerned with the interactions between computers and human languages.

-The major goal of this system "is to read, comprehend, and make sense of human language in a way that is advantageous to the user". -[T]his system's primary goal. The Natural Language Toolkit, which is sometimes referred to by its acronym NLTK at times, provides libraries that can be utilised for classification purposes. The following are examples of parameters that are used by the algorithm: Feedback Provided by Our Clientele.

C. Prediction

Algorithm used for Prediction:

Multiple Linear Regressions

"It is a statistical technique that uses various explanatory variables to predict the outcome of response variable. Formula is y=b0+b1*x1+b2*x2+.....bn*xn Where y=dependent variable and x=independent variables Parameters that algo uses":

1. An understanding and compassionate investigation of the Reviews
2. Online review Volume
3. No Charge for Delivery
4. The total number of inquiries that were posed by clients and were given answers
5. Discount Value
6. Online Ratings

Conclusion

This study is based on the hypothesis that social connections and promotional marketing methods, such as online reviews and questions and answers to those reviews, are both significant factors in determining sales. Online reviews and questions and answers to those reviews are examples of promotional marketing methods. This research demonstrates that opinions have a strong interaction with the amount and valence of online reviews, which may significantly affect product sales and allow for their prediction. The research that was

presented in this study demonstrates that opinions have a strong interaction with online reviews. In conclusion, “we have shown that when sentiments interact with volume and valence, a” factor's importance as a predictor of product sales grows, as does its predictive power. This was proved by the fact that their interactions increased the predictive power of the factor. The market basket analysis results in the production of the frequent item set, which is sometimes referred to as association rules, and which may easily represent the purchasing behaviour of consumers. A retailer can rapidly establish his or her retail shop with the assistance of these guidelines, and then continue to grow the enterprise in the years to come. During the process of market basket analysis, the Apriori algorithm is the one that is utilised most frequently as the principal tool. It is feasible for it to be a very effective device for analysing the shopping patterns of people, which is something that needs to be done. As its three statistical indicators, the market basket analysis makes use of support, confidence, and confidence. The predictive ability or accuracy of an algorithm is what's meant to be measured by confidence, while the frequency with which an item appears in a certain transactional data set is what's meant to be measured by support. Both of these metrics can be measured.

References

[1] East, R.,Hammond, K. and Lomax,W.(2008), “Measuring the impact of positive and negative word of mount on brand purchase probability ”,International Journal of research in Marketing ,Vol.25 No. 3,pp.215-224

[2] Cui, G., Lui, H. and Guo, X.(2012), “The effect of online consumer reviews on new product sales”, International Journal of Electronics , available at:http://www.tandfonline.com/doi/abs/10.2753/JEC1086- 4415170102(accessed 10 March 2015)

[3] Bohdan M. Pavlyshenko (2018),” Machine Learning models for sales time series forecasting”

[4] Professor Deven Ketkar(2018). “A Forecast for Big Data Sales based on Random Forests and Multiple Linear Regression”, IJEDR 2018, VOL. 6, ISSUE 4. ISSN: 2321-9939

[5] T.F. Cootes, M.C. Ionita, C. Lindner and P. Sauer, Robust and Accurate Shape Model Fitting Using Random Forest Regression Voting, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2012.

[6] U. Gro¨omping, Variable importance assessment in regression: Linear regression versus random forest, Amer. Statistic. 63 (2009) 308–319.

[7] J. Han, Y. Liu and X. Sun, A scalable random forest algorithm based on MapReduce, 2013 IEEE 4th Int. Conf. Software Engin. Serv. Sci. Beijing (2013) 849–852.

[8] Q. He, T. Shang, F. Zhuang and Zh. Shi, Parallel extreme learning machine for regression based on MapReduce, Neurocomput. 102 (2013) 52–58.

[9] V. Svetnik, A. Liaw, Ch. Tong, J. Christopher Culberson, R.P. Sheridan and B.P. Feuston, Random forest: A classification and regression tools for compound classification and QSAR modeling, J. Chem. Inf. Comput. Sci. 43(6) (2003) 1947–1958.

[10] S. Srivastava and R. Kumar, "Indirect method to measure software quality using CK-OO suite," 2013 International Conference on Intelligent Systems and Signal Processing (ISSP), 2013, pp. 47-51, doi: 10.1109/ISSP.2013.6526872.

[11] Ram Kumar, Gunja Varshney , Tourism Crisis Evaluation Using Fuzzy Artificial Neural network, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-1, Issue-NCAI2011, June 2011

[12] Ram Kumar, Jasvinder Pal Singh, Gaurav Srivastava, "A Survey Paper on Altered Fingerprint Identification & Classification" International Journal of Electronics Communication and Computer Engineering Volume 3, Issue 5, ISSN (Online): 2249–071X, ISSN (Print): 2278– 4209

[13] Kumar, R., Singh, J.P., Srivastava, G. (2014). Altered Fingerprint Identification and Classification Using SP Detection and Fuzzy Classification. In: , et al. Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012. Advances in Intelligent Systems and Computing, vol 236. Springer, New Delhi. https://doi.org/10.1007/978-81-322-1602-5_139

[14] W. Zhao, H. Ma and Q. He, Parallel K-Means Clustering Based on MapReduce, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2009.