E-COMMERCE SYSTEM FOR SALE PREDICTION USING MACHINE LEARNING TECHNIQUE

Pranay Kumar 2420090065 2420090065@klh.edu.in

Nandini. J _ 2420030262 2420030262@klh.ed.in

K.Hemanth Sai 2420030593 2420030593@klh.edu.in

Abstract. E-commerce is a platform where people are able to buy and sell goods. The main purpose of e-commerce is to provide convenience to the customers where they do not have to go to a physical store to make a purchase. As the will be able to make the purchase online and the item will be in their door step in the following days. In 2019, a total of $603 billion worth of sales were done via e-commerce in the United States compared to 3.17 billion in retail sales in the United States. The purpose of this study was to build machine learning algorithms which are able to forecast the sales of the e-commerce platform. A research was being done to understand the literature reviews based on similar systems and similar studies that relates to the researcher project. The purpose of doing this literature review is to understand which machine learning model was being used by other studies so the researcher will be able to select some of the best machine learning models for this study. Once the researcher has selected the models, he will them build the models and test their accuracy, error and performance. At the end, the researcher will compare all of the model's accuracy and errors to get the best model which have low error and high accuracy for forecasting sales. The model which have been fulfil the criteria, will be integrated into the system which is being built by the researcher. The system will give a view of the current and forecasted sales.

Index Term. E-commerce, machine learning, sales forecasting, ARIMA, SARIMA

1. Introduction

The domain for this research is E-commerce. Starting a new method to gather and analyze data could be a huge impact to an organization, as the outcome can be positive or it can go the other way. E-Commerce Platforms collect a large amount of data and store it in their data centers. They fail to look at this as an advantage for their business opportunity such and analyzing the data and its pattern trough out the years. For example, All the customer data from registration, search history, sales, chats are being stored in their server and will be only be used when there is a problem with existing data. It's understandable why they won't want other company to analyze their data is due to privacy issues but they are also able to create their own team to analyze the data which can be profitable for them.

Its shown that E-Commerce are one of the top 10 business which have a very large amount of data storage. With the access to the amount of data they will be able to create a game changing environment for the E-Commerce industry. This industry has been spending hundreds of millions of dollars in advertising, social media, sorting secured data and much more to generate more sales but

they didn't realize that with machine learning they will be able to step up they game against their competitors. Machine learning is a large tree branch which have many specializations such as data mining, artificial intelligence, augmented reality and prediction. For this research, will be only focusing on the prediction using machine learning.

With the ability of prediction using machine learning algorithm for e-commerce, we will be able to identify any hidden patterns, outliers, point of interest (POI) and much more. This will allow e commerce to be able to properly identify the important details in each and every aspect. They will be able to use all their data such as amount of product purchased, product categories, payment method, interest rate, duration of delivery and customer location to have a better understanding on how to improve and manage their sales.

If the e-commerce platform is able to forecast its sales for the upcoming month or day, they will be able to make better business decisions. They will also be able to track and trends in their sales if any festival or event happens yearly. They will also be able to keep track of their inventory so all items will be stock sufficiently which will avoid any overstocking and understocking of a product as they will be able to get a rough estimate of purchases which are likely to happen. Not only that, they will also be able to keep better track at their finance and make reasonable purchases option and have a proper budget throughout the business operation.

2. Literature Review

Sales prediction is an essential task which have to be done by the e-commerce and the prediction will be able to provide crucial impact to towards the business decision making process. Not only that by having sales prediction for the e-commerce platform, they can have a better understanding about their financial status to manage the workforce, and further improving their supply chain management system. Based on [1] and [2], a sales prediction allows the e-commerce platform to have a better accuracy and reliable prediction which will help them with inventory planning, competitive price, and timely promotions strategies. According to [3], The prediction of e-commerce sales allows to understand the lifecycle of the e-commerce platform as its sales and growth, stability, decline and how are the sales being affected by short terms product goals such as promotion, pricing, season and ranking online.

According to the research conducted by [1], they have used the convolution neural network (CNN) algorithm to do sales forecasting in e-commerce. This research was being done to solve the identified limitation which was method require case-by-case manual feature engineering for specific scenarios which is difficult, time-consuming and requires a lot of expert knowledge. However, the goal for this research was to identify if this approach can automatically extract the effective features and provide the sales forecasting based on extracted features was mention by [1]. The main algorithm which was used for this research was the CNN algorithm to perform the sales prediction. However, for comparison purpose the research have chosen ARIMA, DNN, TL and WD algorithm to find the most accurate results for the sales prediction. The researcher also has used sample weight decay and transfer learning technique to further improve the forecasting accuracy further, which have been proved to be highly effective in the experiments. Based on the MST boxplot, ARIMA model have the highest average value, however the CNN algorithm achieved its goal where it can automatically extract the effective features and do a sales forecasting using the extracted features.

Based on the researches which was conducted by [2] and [3], they have both chosen neural network algorithm. But both of this neural network algorithm have their own approach where Nonlinear Autoregressive Neural Network (NARNN) is used by the 2018 research and the 2019 research have conducted the research using Recurrent Neural Networks (RNN) and Long Short-Term Memory Networks (LSTM) algorithm which is a special neural network. These researches have used this approached algorithm to find sales prediction and demand of e-commerce. The problem that the researches have stated is similar which is difficulty in identifying the different cross-product demand/sales pattern and the correlations which are available. The goal for both of the research paper was to purpose a systematic pre-processing framework to overcome challenges in e-commerce settings

and also purposed a forecasting framework. The algorithm which have been used by to compare both of these researches was ARIMA (time series analysis). The results discussion for the 2018 research have shown that the prediction error for NARNN is at 0.1016 and ARIMA was at 0.1389, which shows that NARNN have a lower error rate compared to the ARIMA. For the research in 2019, the results also show that LSTM has a lower mean and median compared to ARIMA.

Sales prediction is usually done by using the most common method, time series analysis. Time series analysis involve the Autoregressive function which helps which any type of prediction analysis. According to [4] study on the machine learning model for sales times series forecasting, it has mention that sales prediction is a modern business intelligence method. Also mention by [3], ARIMA model has a better approach for the performance in prediction in the time series analysis. This main problem stated in this research by [4] is that for time series data, the data required is large to capture the seasonality and the large transactional sales data can have many missing data and outliers. These data will then need to be take into account a lot of different factors which can impact the sales. The goal for this time series analysis it to combine different time series algorithm in order to improve this prediction accuracy. There were five algorithm which have been selected in the research which are ExtraTree, ARIMA, RandomForest, Lasso and Neural Network which are all time series algorithm and supervised. Based on the results for the forecasting error testing, ExtraTree has the highest validation error compared to the rest and Neural Network has the lowest validation error making it one of the best algorithms for prediction.

Based on the other researcher [5], have conducted a research on forecasting of Walmart sales using machine learning algorithms. The key for this research was done by implementing several different classification algorithms in the sales data from all different Walmart locations all over the united states. The problem which was highlighted in this research was creating a competitive comparative analysis to find the best algorithm. The researcher had selected 3 different algorithms for the comparison and test it using the MAE evaluation R^2 Score. The goal of this research is to find accuracy of algorithm using different hyperparameters of each model to obtain the best Mean Absolute Error (MAE) and R^2 score. The algorithm which was used for this research was Random Forest, Gradient Boosting and Extremely Randomized Tree (Extra Tree). The results of this research indicate that the Random Forest is the best algorithm which have scored the minimum amount in MAE evaluation (1979.4) and a high R^2 (0.94) score which have shown a high accuracy compared with the others.

Another research was conducted by [6] which was to study the sales forecast for Amazon sales based on the different statistic methodology. This research has primarily focused on the amazon data and forecast the future sales using the historical data by using statistic algorithms. The problem that's identified by the lecturer in this research is the how a statistical methodology for sales methodology can can help in sales forecasting. Statistical methodology algorithm are a part of the time series analysis models [4]. The goal for this research is to conduct sensitivity analysis on the three methods, and identify which is the most reliable, accurate and suitable approach. The better the accuracy for a method, the better will the prediction will be for the sales forecasting. There were three different approaches used for this research which includes Winters' exponential smoothing, time-series decomposition and ARIMA. The results of this research were done by measuring the forecasting error (RMSE). All of the method has a very low amount of forecasting error, therefore all of the method can be implemented to conduct sales forecasting for the Amazon sales.

The research which was conducted by [7], regarding the car sales prediction using machine learning algorithm. This research emphasizes on the data about car sales and how they are derived from various sources. The main issues which is identified by the researcher was that getting varied idea about how well the various criteria's in our dataset works and identifying the appropriate algorithm which can be used. The outcome of this research needs to apply multiple different machine algorithm on the sales dataset and provide proper analysis of algorithm used [8]. Sales of car doesn't contain independent variable as most of the factor such as car size, petrol capacity, price, height and tire are some of the features which influence the sales of the cars. The algorithm which is being used

for this research is Random Forest. The results for the random forest determine that the price is the main attribute that will make large impact on the sale of the car sales value. The random forest is also high accuracy percentage (above 85%).

Another research which was done by [9], which discuss about Explaining Machine Learning models in sales prediction. This research mainly discusses about all the main models of machine leaning which is commonly being used in sales prediction and also will show analysis of the best machine learning model available. The problem which is identified by this research paper was that how to identify the appropriate model based on the business understanding by using the intelligence and data driven models. The goal of this research was to demonstrate how effective each of the model is and its usability. This is being done to ensure that the correct method have been selected to the selected business environment was mentioned by [10] and [4]. The method (algorithm) which was been used by this research was the decision tree, neural network, naïve bayes, random forest and support vector machine (SVM). Based on the chosen algorithm, the results are tabulating against the accuracy. The random forest is at 85%, naïve bayes is at 83%, decision tree at 76%, neural network at 70% and finally the SVM is at the lowest 59%. Therefore, the best method to be chosen is the random forest as it has a high accuracy model at 85%.

According to the research which was conducted by [11] which discuss about the Machine Learning for Restaurant Sales Forecast. In this research, it explains on how restaurant can be implementing machine learning to improve and understand the sales. The problem identified is that many restaurants do not have solid forecast of their daily sales. This is because they don't have proper education about calculating the sales prediction. The goal for this research is to investigate the possibility to create a forecasting solution based on the supervised learning. This will help the restaurant business to record and analyses the sales and can make better decision in relation to the financing. The algorithm which have been used by this research is Extreme Gradient Boosted and Long Short-Term Memory Boosted. The results of this study were that the Extreme gradient boosted algorithm works perfectly in this testing approach while the LSTM has some limited support on these problems.

3. Knowledge Discovery in Databases (KDD)

For this project and research, the KDD methodology have been selected as it mostly fits the requirement of this project. This methodology has been widely used in this machine learning field for pattern recognition, statistic, databases, artificial intelligence (AI), and data visualization (usually the outcome) (DBD, 2019). There are 5 steps in the KDD methodology which will be discusses. For this sales prediction in E-commerce using machine learning, first step will be finding a suitable E-commerce dataset which is free and opensource then pre-process the selected the dataset, followed by transforming the dataset (eliminating what is not required or combining from multiple datasets), modelling using different algorithm to find higher accuracy and finally the researcher will evaluate the results from the modelling steps.

3.1. Data Selection

The dataset for this research will be a transactional data set. This is because we will be doing a sales prediction which require all the past transactional data to predict the future sales. The dataset transactional will be from one of the e-commerce which is open source and can be used without any restriction. The dataset which have been obtained is from Kaggle.com which have listed a Brazilian E-Commerce Public Dataset by Olist Store (E-Commerce Site). There are about 100,000 transactional order history data provided. All of these provided data are from 2016 to 2018. They also have provided eight different dataset which contains different datasets such as product dataset, order dataset, customer dataset and order item datasets. Below figure 1 shows the connection between all of the provided dataset for this e-commerce site.

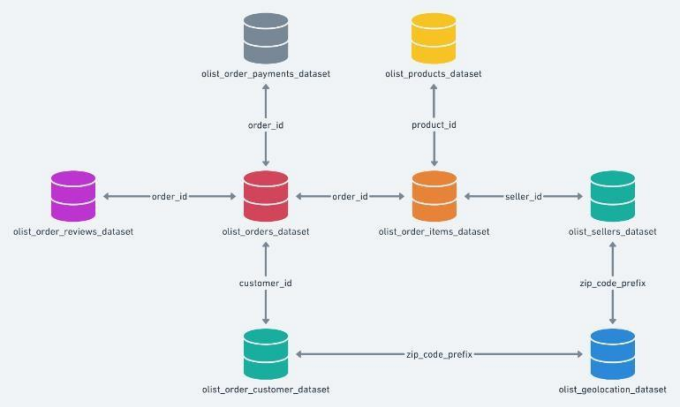

Figure 1. The connection between acquired datasets

3.2. *Data Preprocessing*

Data preprocessing is a process which explain how the selected data will be cleaned from all the noise or outliers. This means that cleaning up the data which have a huge amount of additional meaning which is meaning less and not required. For example, in this dataset there is product review and there is no need of product review in sales prediction, therefore the noise is the product review and it needs to be removed. Not only that, if the dataset has missing values for the sales and price values, we will need to handle it appropriately by replacing the missing value with the average value or use the mean or median imputation to keep the data consistent. At this stage we also can account the time sequence of the data and the known changes.

3.3. Data Transformation

Data transformation is a process of converting data from one format to another different format to satisfy the needs. This process also can be referred to the ETL process which means Extract, Transform and Load. Transformation have become a really important task as the data volume have increasing tremendously. Therefore, robust data transformation will allow user to focus on data which satisfy the business needs. Same goes for this project, all of the dataset will undergo robust transformation and only the important data will be combined into a new data format. This will help to increase the researcher focus towards the important data and will be able to build a better prediction model which will provide better results for the accuracy. The researcher will have to analyze all the eight different data set and identify all of the important values and information and transform them in to one or more different file format. This will make the research work much more easier in the modelling stage as they won't have to look at the irrelevant information.

3.4. Modelling

Modelling is a process where you identify the algorithms which you are going to be using for the project research purpose. For this research, we will be using two different algorithms which are Gradient Boosting and Random Forest. The algorithms are being selected is because there are commonly related in prediction analysis. Gradient boosting is a machine learning technique which involves classification and regression to product based on weak prediction models such as decision tree. Random Forest contain a huge amount of decision tree that works in a group and each of the individual tree will provide a class prediction.

The models which are going to be tested out will be based on the time series analysis. There are two modelling method which are going to be tested for this research which are Autoregressive Integrated Moving Average (ARIMA) and Seasonal Autoregressive Integrated Moving Average (SARIMA). These models are chosen as they are some of the best models out there which can provide a proper accurate accuracy for the prediction.

3.5. Evaluation Matrix

At the end of this research, researcher will be doing an overall evaluation matrix to see which algorithm have the highest prediction accuracy. The evaluation will include the error percentage and prediction accuracy for each and every algorithm. After that, researcher can do an overall analysis based on the methods and algorithm which we have used and see which is the best method to be used. The researcher also can understand the any hidden pattern which are interesting and can help to improve the accuracy.

4. Results and Discussion

For this results and discussion, the researcher will be showing all of the evaluation matrix which have been used according to their algorithms. Since this project was about sales forecasting, the researcher has used two regression models which are Random Forest and Gradient Boosting. Another two models were used from the time series analysis which were ARIMA and SARIMA.

For the regression models, there are many different ways to evaluate the model using different evaluation matrix. The researcher has chosen Accuracy, Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and R Squared method to evaluate the models. The results obtain from each of the model will be compared with the other model to see which model has the least error.

For the time series models, the number of ways to do evaluation is limited. Therefore, the researcher has chosen Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) to evaluate time series models. Both of the models will then be comparing the evaluation matrix to see which is the better model

4.1. Model

	Train Accuracy	Test Accuracy	RSME	MAE	R^2
Random Forest	90.99%	87.39%	76,669.68	55,133.85	0.87
Gradient Boosting	99.99%	86.88%	82,230.19	62,188.42	0.85

Figure 1. Results of sales prediction using proposed models

Based on the figure 2 above, the evaluation matrix has been stated for both Random Forest and Gradient Boosting. It was observed that the train accuracy achieved by gradient boosting was 99.99% which was higher compared to the random forest. However, the test accuracy received by the random forest is 87.39% which is higher compared to the gradient boosting. This already shows that random forest model has a better fit of the data because the train and test score are very close compare to gradient boosting. There were about ~13% difference of train and test accuracy for gradient boosting which suggested that the test data wasn't really accepted by the model therefore it had a lower test accuracy. Moving on to the other evaluation matrix, it was notices that the Root Mean Square Error (RSME) obtain by the random forest is smaller which means that random forest has a lower residual (prediction error) compared to the gradient boosting. The Mean Absolute Error obtained by gradient search is higher which means that the gradient boosting has more mean error. Finally, the R^2 which was obtained by the random forest is higher by 0.02 compared to gradient boosting.

Based on that, the random forest and gradient boosting model both did great in fitting the data into the model. However, the best model for regression model is the random forest as it has the lower error accuracy compared gradient boosting.

4.2. Time Series Model

	RSME	MAE	MAPE
ARIMA	1806.35	1528.94	8.62%
SARIMA	1812.38	1521.32	6.69%

Figure 2. Sales prediction results using time series models

Based on the figure 2 above, all the evaluation matrix was tabulated according to the time series model, ARIMA and SARIMA. First and foremost, the Root Mean Square Error (RSME) for both of the models are almost similar but SARIMA has a higher RSME value which means that it has a small amount residual that ARIMA. Next is the Mean Absolute Error (MAE), this was also a close call as the difference was '7.0', which means ARIMA model has a slightly higher error rate compared to SARIMA. Finally, MAPE is one of the most trusted method when doing time series forecasting. The SARIMA MAPE score is smaller that the ARIMA model which means that SARIMA only have 6.69% of error in the model while ARIMA has the error of 8.72%.

After going through all of the evaluation matrix for the time series model, the researcher has chosen SARIMA to be the best model as it has a lower MAPE score which means that it has lower error compared to ARIMA model. SARIMA also allows to see the seasonal trends in the data which gives a greater view.

5. Conclusions

By doing this project of using machine learning for forecasting the ecommerce sales, it was noticed that the in this project, there are many different method of forecasting the sales of the ecommerce platform but the researcher was only able to focus on only four algorithms which are commonly being used when forecasting the sales of the future. The researcher was able to build and test all of the selected machine learning models which have been selected. The model which have the best prediction range, where the predicted value and the actual value are almost similar is chosen as the best algorithm. This best algorithm will then be integrated into a web application which will also built by the researcher.

5.1. Future Enhancement

For future enhancement, the researcher suggests that different time series model to be used to see the pattern and results. Next, obtain dataset from other e-commerce to identify pattern differential and how sales are being affected. Finally, the size of the dataset to be larger as 20 months of data is not really enough to get better forecasting and trend comparison.

References

[1] Zhao, K. and Wang, C. (2017) 'Sales Forecast in E-commerce using Convolutional Neural Network', (August 2017). Available at: http://arxiv.org/abs/1708.07946.

[2] Bandara, K. et al. (2019) 'Sales Demand Forecast in E-commerce using a Long Short-Term Memory Neural Network Methodology'. Available at: http://arxiv.org/abs/1901.04028.

[3] Li, M., Ji, S. and Liu, G. (2018) 'Forecasting of Chinese E-Commerce Sales: An Empirical Comparison of ARIMA, Nonlinear Autoregressive Neural Network, and a Combined ARIMA-NARNN Model', Mathematical Problems in Engineering, 2018, pp. 1–12. doi: 10.1155/2018/6924960.

[4] Pavlyshenko, B. (2019) 'Machine-Learning Models for Sales Time Series Forecasting', Data, 4(1), p. 15. doi: 10.3390/data4010015.

[5] Elias, S. and Singh, S. (2018) 'FORECASTING of WALMART SALES using MACHINE LEARNING ALGORITHMS'

[6] YU, J. and LE, X. (2017) 'Sales Forecast for Amazon Sales Based on Different Statistics Methodologies', DEStech Transactions on Economics and Management, (iceme-ebm). doi: 10.12783/dtem/iceme-ebm2016/4132.

[7] Madhuvanthi, K. et al. (2019) 'Car sales prediction using machine learning algorithmns', International Journal of Innovative Technology and Exploring Engineering, 8(5), pp. 1039–1050.

[8] Xia, G. and He, Q. (2018) 'The Research of Online Shopping Customer Churn Prediction Based on Integrated Learning', 149(Mecae), pp. 756–764. doi: 10.2991/mecae-18.2018.133.

[9] Bohanec, M., Kljajić Borštnar, M. and Robnik-Šikonja, M. (2017) 'Explaining machine learning models in sales predictions', Expert Systems with Applications, 71(April), pp. 416–428. doi: 10.1016/j.eswa.2016.11.010.

[10] Mohammed, M., Khan, M. B. and Bashie, E. B. M. (2017) Machine learning: Algorithms and applications, Machine Learning: Algorithms and Applications. doi: 10.1201/9781315371658.

[11] Holmberg, M. and Halldén, P. (2018) 'Examensarbete 30 hp Maj 2018 Machine Learning for Restaurant Sales Forecast'. Available at: http://www.teknat.uu.se/student.

[12] Brownlee, J. (2019). 11 Classical Time Series Forecasting Methods in Python (Cheat Sheet). [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/time-series-forecasting-methods-in-python-cheat-sheet/ [Accessed 14 Oct. 2019].

[13] Ceriotti, M. (2019) 'Unsupervised machine learning in atomistic simulations, between predictions and understanding', Journal of Chemical Physics, 150(15). doi: 10.1063/1.5091842.

[14] Data-Driven-Science (2018). Python vs R for Data Science: And the winner is... [online] Medium. Available at: https://medium.com/@data_driven/python-vs-r-for-data-science-and-the-winner-is-3ebb1a968197 [Accessed 2 Oct. 2019].

[15] DBD, U. (2019). KDD Process/Overview. [online] Www2.cs.uregina.ca. Available at: http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html [Accessed 6 Oct. 2019].

[16] Hyde, K. K. et al. (2019) 'Applications of Supervised Machine Learning in Autism Spectrum Disorder Research: a Review', Review Journal of Autism and Developmental Disorders. Review Journal of Autism and Developmental Disorders, 6(2), pp. 128–146. doi: 10.1007/s40489-019-00158-x.

[17] Klassen, S., Weed, J. and Evans, D. (2018) 'Semi-supervised machine learning approaches for predicting the chronology of archaeological sites: A case study of temples from medieval angkor, Cambodia', PLoS ONE, 13(11), pp. 1–17. doi: 10.1371/journal.pone.0205649.