IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11662

SALES PREDICTION USING MACHINE LEARNING

Vimala P[1],Rajesh Kumar Y[2], Sowndarya Lakshmi R[3], Thabasum Mohaseena S[4]

Assistant Professor, Department of Computer Science and Engineering, Dhirajlal Gandhi College of Technology, Salem, Tamil Nadu, India[1]

Student, Department of Computer Science and Engineering, Dhirajlal Gandhi College of Technology, Salem, Tamil Nadu, India[2-4]

Abstract: Connected devices, sensors, and mobile apps make the retail sector relevant tested for big data tools and applications.We investigate how big data is, and can be used in retail operations. Based on our state-of-the-art literature review, we identify four themes for big data applications in retail logistics: availability, assortment, pricing, and layout planning. Our semi-structured interviews with retailers and academics suggest that historical sales data and loyalty schemes can be used to obtain customer insights for operational planning, but granular sales data can also benefit availability and assortment decisions can be used for demand forecasting and pricing. However, the path to exploiting big data is not a bed of roses. Challenges include shortages of people with the right set of skills, the lack of support from suppliers, issues in IT integration, managerial concerns including information sharing and process ntegration, and physical capability of the supply chain to respond to real-time changes captured by big data. We propose a data maturity profile for retail businesses and highlight future research directions. Association Rules is one of the data mining techniques which is used for identifying the relation between one item to another. Creating the rule to generate the new knowledge is a must to determine the frequency of the appearance of the data on the item set so that it is easier to recognize the value of the percentage from each of the datum by using certain algorithms, for example apriori. This research discussed the comparison between market basket analysis by using apriori algorithm and market basket analysis without using algorithm in creating rule to generate the new knowledge. The indicator of comparison included concept, the process of creating the rule, and the achieved rule. The comparison revealed that both methods have the same concept, the different process of creating the rule, but the rule itself remains the same.

Keywords: Big data; retail operations; maturity;availability; assortment; replenishment; pricing; layout; logistics.

I. INTRODUCTION

Earlier companies used to produce goods without considering the number of sales and demand. For any manufacturer to determine whether to increase or decrease the production of several units, data regarding the demand for products on the market is required. Companies can face losses if they fail to consider these values while competing on the market. Different companies choose specific criteria to determine their demand and sales .

In today's highly competitive environment and ever-changing consumer landscape, accurate and timely forecasting of future revenue, also known as revenue forecasting, or sales forecasting, can offer valuable insight to companies engaged in the manufacture, distribution or retail of goods. Shortterm forecasts primarily help with production planning and stock management, while long-term forecasts can deal with business growth and decision-making. Sales forecasting is particularly important in the industries because of the limited shelf-life of many of the goods, which leads to a loss of income in both shortage and surplus situations. Too many orders lead to a shortage of products and still too few orders lead to a lack of opportunity. Therefore, competition in the food market is continuously fluctuating due to factors such as pricing, advertisement, increasing demand from the customers. Managers usually make sales predictions randomly. Professional managers, however, become hard to find and not always available (e.g., they can get sick or leave). Sales predictions can be assisted by computer systems that can play the qualified managers' role when they are not available or allow them to make the right decision by providing potential sales predictions. One way of implementing such a method is to try and model the professional managers' skills inside a computer program. Alternatively, the abundance of sales data and related information can be used through Machine Learning techniques to automatically develop accurate sales predictive models. This approach is much simpler. It is not prejudiced by a single sales manager's particularities and is flexible, which means it can adapt to data changes. It has, however, the potential to overestimate the accuracy of the prediction of a human expert, which is normally incomplete. For example, once companies used to produce the products without taking into consideration the number of sales and demand as they faced several problems. Since they don't know

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022
DOI: 10.17148/IJARCCE.2022.11662

how much to sell, for any manufacturer to decide whether to increase or decrease the number of units, data regarding the consumer demand for products is essential. If companies do not consider these principles when competing in the market, they will face losses. Different companies choose different parameters to determine their market and sales. There are several ways of forecasting sales in which companies have 2 previously focused on various statistical models such as time series and linear regression, feature engineering and random forest models to obtain future sales and demand prediction. Time series contains data points that are stored over a fixed period and are used to forecast the future. Time series is a collection of data points which are collected in period at sequential, evenly spaced points. The most important components to analyze are patterns, seasonality, irregularity, cyclicity. Linear regression is a mathematical tool used to forecast past values. It can help to determine the underlying trends and address cases involving overstated rates. Feature engineering is the use of data on domain knowledge and the development of features to make predictive Machine Learning models more accurate. It makes for deeper data analysis and a more useful perspective. A decision tree is a fundamental principle behind a model of random forests. The decision tree approach is a technique used in data mining to forecast and classify data. The decision tree approach does not provide any conceptual understanding of the issue itself. Random forest is the more sophisticated method that allows and merges many trees to make decisions. The random forest model results in more accurate forecasts by taking out an average of all individual tree decision predictions. The entire data set is usually divided into two parts, namely the training data and the test data. Training data is a data that is used to train the model, and test data is the data used to evaluate the trained model. A classical approach is 80-20 split, stating that 80 percent of the data is used to train the model, and the remaining 20 percent of the data is used to test the model. But approaches like stratified K-fold cross-validation are known to provide good results. There were many cross-validation variants, such as simple kfolds, leave one out, stratified k-fold cross-validation, and so on.

II. LITERATURE SURVEY

- ✓ Intelligent Sales Prediction Using Machine Learning Techniques
- ✓ MachineLearning Models for Sales Forecasting
- ✓ Walmart's Sales Data Analysis- A Big Data Analytics Perspective
- ✓ Forecast of Sales of Walmart Store Using Big Data Applications

III. EXISTING SYSTEM

Sales forecasting is usually done by collecting the sales data of a shop of a time period and make predictions using various prediction techniques. There are many factors which affects the sales forecasting which includes direct and indirect competition, state and city holidays, population changes, sales promotions etc. The above factors create a great deviation in sales prediction in existing system which is not providing accurate results as expected. The confidence level has not taken for all algorithms. The holiday factors which is important in sales prediction is no considered. Thus, the sales varies on using different machine learning algorithms.

IV. PROPOSED SYSTEM

Proposed Method:

An experiment is chosen for the first research question i.e. correlation. Each data attribute can be selected by applying feature selection methods like data correlation and which will make the predictable attributes more accurate. This will reduce a lot of strain on the Machine Learning model during pre-processing and cleansing the data. For the second research question an experiment is chosen because the experiments provide control over factors and a deeper understanding of many common research techniques such as a case study or survey[29]. One can describe the procedure followed in this experiment as follows:

- Extracting the data required for the sales.
- Applying specified Machine Learning (supervised) algorithms.
- The performance of the output can be enhanced by comparing metrics such as accuracy score, mean absolute error and max error.
- Based on assessment tests, the best suitable algorithm can be selected

V. CONCLUSION

Sales forecasting plays a vital role in the business sector in every field. With the help of the sales forecasts, sales revenue analysis will help to get the details needed to estimate both the revenue and the income. Different types of Machine.Learning techniques such as Support Vector Regression, Gradient Boosting Regression, Simple Linear Regression, and Random Forest Regression have been evaluated on food sales data to find the critical factors that influence sales to provide a solution for forecasting sales. After performing metrics such as accuracy, mean absolute error,

DOI: 10.17148/IJARCCE.2022.11662

and max error, the Random Forest Regression is found to be the appropriate algorithm according to the collected data and thus fulfilling the aim of this thesis.

Future Work

In future work one can attempt performance metrics such as time while predicting the sales. These metrics can play a crucial role in evaluating multiple Machine Learning algorithms. And also one can attempt to implement more accurate data in the continued study. Machine Learning has the advantage of analyzing data and key variables so that you can aim to develop a systematic approach using a variety of Machine Learning techniques.

REFERENCES

[1] Patrick Bajari, Denis Nekipelov, Stephen P Ryan, and Miaoyu Yang. Machine learning methods for demand estimation. *American Economic Review*, 105(5):481–85, 2015.
[2] Kris Johnson Ferreira, Bin Hong Alex Lee, and David Simchi-Levi. Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management*, 18(1):69–88, 2016.
[3] Ankur Jain, Manghat Nitish Menon, and Saurabh Chandra. Sales forecasting for retail chains, 2015.
[4] Grigorios Tsoumakas. A survey of machine learning techniques for food sales prediction. *Artificial Intelligence Review*, 52(1):441–447, 2019.
[5] Xiaogang Su, Xin Yan, and Chih-Ling Tsai. Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(3):275–294, 2012.
[6] Toby J Mitchell and John J Beauchamp. Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032, 1988.
[7] Zheng Li, Xianfeng Ma, and Hongliang Xin. Feature engineering of machinelearning chemisorption models for catalyst design. *Catalysis today*, 280:232–238, 2017.
[8] Xinchuan Zeng and Tony R Martinez. Distribution-balanced stratified crossvalidation for accuracy estimation. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(1):1–12, 2000.
[9] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. On the stratification of multi-label data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 145–158. Springer, 2011.
[10] Chris Rygielski, Jyun-Cheng Wang, and David C Yen. Data mining techniques for customer relationship management. *Technology in society*, 24(4):483–502, 2002.
[11] Krzysztof J Cios, Witold Pedrycz, Roman W Swiniarski, and Lukasz Andrzej Kurgan. *Data mining: a knowledge discovery approach*. Springer Science & Business Media, 2007.
[12] Maike Krause-Traudes, Simon Scheider, Stefan Rüping, and Harald Meßner. Spatial data mining for retail sales forecasting. In *11th AGILE International Conference on Geographic Information Science*, pages 1–11, 2008.
[13] Stephen Marsland. *Machine learning: an algorithmic perspective*. CRC press, 2015.
[14] ML documentation. https://www.mathworks.com/discovery/ machine-learning.html). Accessed: 2020-04-22.
[15] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2020.
[16] Arvin Wen Tsui, Yu-Hsiang Chuang, and Hao-Hua Chu. Unsupervised learning for solving rss hardware variance problem in wifi localization. *Mobile Networks and Applications*, 14(5):677–691, 2009.
[17] Bohdan M Pavlyshenko. Machine-learning models for sales time series forecasting. *Data*, 4(1):15, 2019.
[18] Taiwo Oladipupo Ayodele. Types of machine learning algorithms. *New advances in machine learning*, pages 19–48, 2010.
[19] Sanford Weisberg. *Applied linear regression*, volume 528. John Wiley & Sons, 2005.
[20] Gradient Boosting documentation. https://turi.com/learn/userguide/ supervised learning/boosted_trees_regression.html). Accessed: 2020-05-19.
[21] JN Hu, JJ Hu, HB Lin, XP Li, CL Jiang, XH Qiu, and WS Li. State-of charge estimation for battery management system using optimized support vector machine for regression. *Journal of Power Sources*, 269:682–693, 2014.
[22] Wangchao Lou, Xiaoqing Wang, Fan Chen, Yixiao Chen, Bo Jiang, and Hua Zhang. Sequence based prediction of dna-binding proteins based on hybrid feature selection using random forest and gaussian naive bayes. *PloS one*, 9(1), 2014.
[23] İrem İşlek and Şule Gündüz Öğüdücü. A retail demand forecasting model based on data mining techniques. In *2015 IEEE 24th International*
Symposium on Industrial Electronics (ISIE), pages 55–60. IEEE, 2015.
[24] Takashi Tanizaki, Tomohiro Hoshino, Takeshi Shimmura, and Takeshi Take-naka. Demand forecasting in restaurants using machine learning and statistical analysis. *Procedia CIRP*, 79:679–683, 2019.
[25] Xu Ma, Yanshan Tian, Chu Luo, and Yuehui Zhang. Predicting future visitors of restaurants using big data. In *2018 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 1, pages 269–274. IEEE, 2018.

DOI: 10.17148/IJARCCE.2022.11662

[26] Mikael Holmberg and Pontus Halldén. Machine learning for restaurant sales forecast, 2018.
[27] I-Fei Chen and Chi-Jie Lu. Sales forecasting by combining clustering and machine-learning techniques for computer retailing. *Neural Computing and Applications*, 28(9):2633–2647, 2017.
[28] Malek Sarhani and Abdellatif El Afia. Intelligent system based support vector regression for supply chain demand forecasting. In *2014 Second World Conference on Complex Systems (WCCS)*, pages 79–83. IEEE, 2014.
[29] Jason Brownlee. *Introduction to time series forecasting with python: how to prepare data and develop models to predict the future*. Machine Learning Mastery, 2017.
[30] Python history. https://en.wikipedia.org/wiki/Python_(programming_ language). Accessed: 2020-04-29.
[31] Guido Van Rossum et al. Python programming language. In *USENIX annual technical conference*, volume 41, page 36, 2007.
[32] Travis E Oliphant. *A guide to NumPy*, volume 1. Trelgol Publishing USA, 2006.
[33] Wes McKinney. Pandas, python data analysis library. *see http://pandas. pydata. org*, 2015.
[34] Niyazi Ari and Makhamadsulton Ustazhanov. Matplotlib in python. In *2014 11th International Conference on Electronics, Computer and Computation (ICECCO)*, pages 1–6. IEEE, 2014.
[35] Raul Garreta and Guillermo Moncecchi. *Learning scikit-learn: machine learning in python*. Packt Publishing Ltd, 2013.
[36] Seaborn documentation. https://seaborn.pydata.org/introduction. html). Accessed: 2020-04-26.
[37] Chung-Jui Tu, Li-Yeh Chuang, Jun-Yang Chang, Cheng-Hong Yang, et al. Feature selection using pso-svm. *International Journal of Computer Science*, 2007.
[38] Tao Zhang, Tianqing Zhu, Ping Xiong, Huan Huo, Zahir Tari, and Wanlei Zhou. Correlated differential privacy: Feature selection in machine learning. *IEEE Transactions on Industrial Informatics*, 2019.
[39] Pearson documentation. https://en.wikipedia.org/wiki/Pearson_ correlation_coefficient). Accessed: 2020-04-25.
[40] Kedar Potdar, Taher S Pardawala, and Chinmay D Pai. A comparative study of categorical variable encoding techniques for neural network classifiers. *International journal of computer applications*, 175(4):7–9, 2017.
[41] Cross validation documentation. https://towardsdatascience.com/ cross-validation-explained-evaluating-estimator-performance e51e5430ff85). Accessed: 2020-04-28.
[42] Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*, 2018.
[43] scilearn max error. https://scikit learn.org/stable/modules/model_ evaluation.html#max-error. Accessed: 2020-05-10. scilearn mean absolute error. https://scikit learn.org/stable/modules/ model_evaluation.html#mean-absolute error. Accessed: 2020-05-10.

