JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

RETAIL SALES PREDICTION USING MACHINE LEARING AND DEPLOYEMENT WITH AWS SAGEMAKER

Mohammad Ahsan Khan
School of Computer Applications,
Lovely Professional University,
Phagwara, Punjab.
mahsank111@gmail.com

Aman Tiwari
School of Computer Applications,
Lovely Professional University,
Phagwara, Punjab.
amant2418@gmail.com

Dr. Kamal Nain Sharma
School of Computer Applications,
Lovely Professional University,
Phagwara, Punjab.
Kamalnain3@gmail.com

Danish Mushtaq Baig
School of Computer Applications,
Lovely Professional University,
Phagwara, Punjab.
danishmushtaq24900@gmail.com

Karine Vinay Kumar
School of Computer Applications,
Lovely Professional University,
Phagwara, Punjab.
Vinaykarine9@gmail.com

***Abstract*—Sales forecasting is a typical function which more often than not revolves around the process of asking about the number of items sold or popularity of a product in some period of time. Making such revenue forecasts enables companies to have more realistic budgets, as data analysis facilitates informed formulation of strategic plans for the future. Budget, revenue allocation, and future advents strategies are all affected by the revenue forecasts that are produced, as another reflection of its key role. Contrasting with the category of targets above, the expected sales are the assumptions about the business performance in the future which immigrants from the historic performance data, available trends, and improvement initiatives for a verification of accuracy. In this evaluation, two different kinds of forecasts: digital and shopper, is analyzed by applying machine learning algorithms in a European retail drug-store chain for a 6-weeks period. We implement machine learning techniques and data analysis to build a very useful instrument which predicts store sales to managers on time and fills their schedules with efficient employees to promote productivity. Libraries of Python programming and math machine implements — matplotlib, scikit-learn, NumPy and Pandas — were presented to a make the set of algorithms wide like Random Forest, Decision Tree and Linear Regression flexible. In this scheme, we used RMSE Index for assessing the performance of our model which is a tool that provides both the accuracy and also acts as a metric.**

I. INTRODUCTION

Sales forecasting is one of the most imperative attributes of any business which is attempting to grow financially. Such forecasting's depend on precise predictions of customers behaviors and expectations. There is a process of forecasting the sales within a specific timeframe which depends heavily on past records details and trends. These forecasts are often the basis that gourmet policies and corporate strategies are made upon. Our Retail Sales Prediction project features machine learning models anticipating sales for 1115 drug stores located in Europe basing on the key factors like more or less sales and further by using this information we plan on improvements. Nowadays forecasting is crucial to all business advances, still a lot of difficulty can be linked with its development in different spheres. In fast-paced and turbulent fashion retail industry, the deep learning technique proves to be indispensable for forecasting sales volume of a new item, which is of vital importance to the purchasing operations. Machine learning has the capability of being able to model sophisticated data patterns and this finds application is diverse areas such as Finance, Management, Marketing, and Retailing[4]. This project targets at predicting sales revenue for Rossmann chain, incorporating elements like holidays, promotions, competition, as well as location. Spot-on sales projections dictate the course of the whole strategic planning process, meaning that businesses can prepare to cater to consumers' changing needs and patterns. The emergence of machine learning technology has provided the accuracy in forecasting visible improvement all over since the different industry set-ups. In fashion retail in fierce competition between producers, where product lifecycles are diminishing and customer preferences are changing within a short period of time, an ability to accentuate sales forecasts can reinforce an advantage in competition. Our research makes use of some advanced learning methods to predict sales, therefore this forecasting helps in sharp purchasing decisions in clothing industry. Using the actual dataset for a retail fashion company as an example, we aspire to resolve these bottlenecks by giving operational performance and revenue-driving executives the data and insights they need to thrive.

II. LITERATURE REVIEW

It will presents an elaborate review of the already available literature on sales predictions in the retail area, majorly by the application of machine learning methods. In this section data from numerous scholarly articles, cases studies, and industry reviews are analyzed[2]. Traditional forecasting methods are shown to have both their own strengths and weaknesses, and multi-variable and non-parametric approaches are also consider. The review also elucidates the rising importance of machine learning as the latest technology that has been recognized to significantly enhance prediction accuracy, which is particularly relevant for business purposes. The review integrates main points from the existing literature In order to alleviate a clear picture of the current retail sales forecasting circumstances and outline areas of future research and study.

The Role of AI and Machine Learning: AI and machine learning solutions transform data-driven retail sales prediction into a trend by giving intelligent insights, predictive analytics, and other tools. This segment describe the main ML approaches used in the retail sales prediction model – linear and nonlinear regressions, time series forecasting and classification algorithms. Moreover, the paper delves on the use of AI for feature engineering, anomaly detection, and the customizing of recommendations to raise accuracy.

Challenges and Limitations: AI technology for forecasting sales in the retail sector is undoubtedly gifts and challenges too. The part here deals with those impediments which arise as like the quality of data, accuracy of models, scalability of this technology and the ethics issue for private citizen in privacy. Comprehendin these issues is necessarily a vital step for an appropriate AI technologies installation in retail[4].

Deployment with AWS SageMaker: AWS SageMaker turns out to be the best option when the model required to be applied directly in the real time scenario, which is lightweight and low-cost service for predicting store sales within platform itself [5]. The next section examines the deployment process using Sagemaker, which enables the execution of training (data), optimisation (model), hosting (an environment for deployment) and monitoring. Case studies and examples combined with practical illustrations explain the benefits of SageMaker which is in shortening the deployment pipeline as well as reducing the time for launching solutions based AI retail.

Future Directions and Conclusions: AI has already laid the way for new sales prediction in the retail market, consequently, possible research directions and emerging trends are worth consideration [7]. The last part of the section uncovers innovative options like the implementation of deep learning structures, reinforcement learning techniques, and to make comprehensive predictive analysis with the help of IoT data. Summing up, the analysis defines the progressive nature of AI recognition power and highlights directions for future improvements that will definitely be implemented in this very field.

III. Related Work

Catal et al. discovered that the application of lineal regression and random forest regression and the use of time series techniques like ARIMA, seasonal ARIMA, and ARIMA without a season lead to corrects sales predictions[4]. Employing regression algorithms along with the use of Azure ML Studio and sales data by Walmart that was posted online, the researchers were able to achieve this. The Spark Streaming technology together with R programming language helped them to achieve the goal of developing time series analytical techniques. Following this experimental evaluation, it was observed that regression approaches would yield better results than the time series analysis approaches. (After) Multivariate Adaptive Regression Splines (MARS) was rated as an approach that fits for solving nonlinear regression problems by the standard. It has a large tailor-made functionality area for different applications including credit scoring and energy price forecasting.

The aim of secondary study by Lu was to propose a unified model paired with MARS and Support Vector Regression (SVR) which guarantees the accuracy of sales prediction[7]. This strategy intended to remove the drawbacks brought by the earlier implemented methods. Through analyzing weekly sales data for a Taiwan chain store for IT, the practical use of the hybrid forecasting method was observed considering particular product categories like motherboards, LCD monitors and notebooks.

In their endeavors to improve the accuracy of sales forecasting, Omar and Liu worked on a BPNN-based model that employs magazines' search popularity data brought from the online Google search engine. To improve the BJT predictability, the model concentrated on the nonlinear in the historical data analysis and the use of usual headline words of the celebrities. We framed the effectiveness of suggested model through page number of Chinese published magazines. The Feng et al. create a novel algorithm ELM (Extreme Learning Machine) which computed with and statistics are effectively used to the book sales forecasting by e-commerce company in China. Stepwise Recurrent Neural Networks (PRNN) were utilized as an appropriate statistical method for sales forecasting, which gives more importance to their ability to cope with nonlinearity found in the actual sales data, the approach applied by Müller-Navarra and al[9].

Holt reflected on the role of the seasons in shaping the sales pattern through the use of the Exponential Weighted Moving Averages (EWMA) model, coupled with feature clustering to improve the forecasting efficiency. Past techniques of seasonal sales mapping were significantly outperformed by the proposed model, shown how well it modeled the observed sales patterns[9].

IV. METHODOLOGY

The approaches adopted in this investigation are begun with a careful literature review in which those studies are included with the same field as ours in sales prediction by making use of machine learning that contribute in formulating an extensive understanding of the core information[2]. The outcome of this literature review will provide the basis upon which our inquiry will be made concerning the retail sales data trending. Our main purpose is getting acquainted with the performance of machine learning algorithms including Linear Regression, Random Forest Regression and Xtreme Boosting Regression on sales data from point-of-sales.

Furthermore, the literature review extends to exploring recent advancements in machine learning algorithms and their applications in sales forecasting, providing valuable insights into emerging methodologies and best practices.

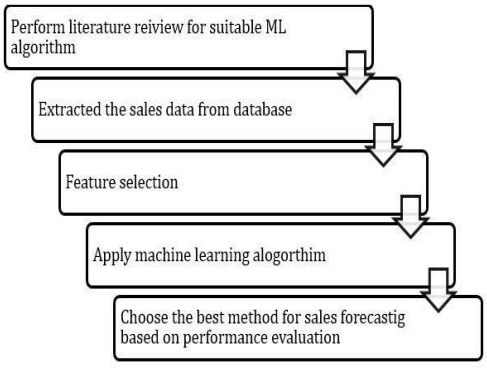

Figure 1 Forthcoming methodology for the sales forecasting is depicted in the subsequent.

Python is a programming language utilized during the research while the tools like pandas, NumPy, Matplotlib, seaborn, and scikit-learn are as well used during the

research which are Python functionalities. By supplying us with essential tools for data manipulation, visualization, and model implementation, these libraries become the avenue for getting down to selling innovations and the application of resulting forecast approaches.

a. Dataset Description

This article provides specific direction for undertaking the retail Point of Sale system during that time in test

set(s) |S| = 32|S| = 32 l

ocations around January 2007. The acquisition of sale histories through the Citadel Point of Sale system was conducted by writing SQL queries from the several tables. This data records are representations of five different stations in the stores. The dataset is a client history data, consisting of bilingual records with 228 invoices and an average of 5 products per invoice. The period of data collecting was from 2013 to 2018 when testing we were using data for the year 2020. The training set enumerates 87,847 rows in which every row contains an item ID, store ID, item sales, and total selling amount respectively[6].

b. Data Pre-processing

Different standardized prediction mechanisms are different (in nature like size and degree of accuracy). This diversity exists on different levels: some of the methods are very qualitative in nature, while the remaining ones (at different degrees) are quantitative. This variety coexists in the level of accuracy (from the very high to the lowest) and the source including the academic background (a higher education background and others). In the algorithm, the procedure that I wrote was about determining these values by the subsequent days, weeks and months of the PSD data. Indeed, in the first place, we have validated the abnormal data and have taken out any of the senseless, absent or unjustifiable data. Initially cleaning data followed by edit and arrangement procedures were performed as the first step, and then I presented the prepared unit to a randomly picked representative staff to proceed to final test. These methodologies are roughly divided into three categories: In front of me come up now the letters H & P, Q & P, and C & P. (Ballon,)[4].

c. Feature Selection

Machine learning models depend largely on feature selection since the latter prevents overfitting, reduces the amount of redundancy in the data, and consequently, also improves the model accuracy. The research included the investigation of model performance using various feature selection techniques in this study. Within these methods a particular method of correlation that proved immensely vital.

So in this case the correlation method analysis will be done that the correlation coefficients will be evaluated between each feature and target variable. Items successfully feature correlations (positive correlations; negative correlations) with the target variable are retained while those not correlating with the target variable or just have weak correlating tend to be discarded. This operation reveals the most characteristic features, the ones which correlate in the most with the machine learning model's predictive abilities.

Additionally in this process of feature selection, only features detected to be positively correlated with the dependent variable were retained as the negative correlations were eliminated. Therefore, this proactive work makes sure that the most dominant and instructive features are kept, implicating this to lead to the enhanced predictive score of the machine learning models.

This study will exploit the row correlation method and the one-hot encoding machine learning technique to build a more robust and precise model that will help optimize the training process's length, complexity, and consequently, delivery of more precise and reliable sales forecasts in retail.

V. System Design

a. System Architecture

System architecture is a fundamental roadmap that defines the system setup and how it operates. It presents structure, which is one of key elements that help in understanding and analyze the project in the beginning[5].

The figure ***1*** above outlines the system architecture that includes user interaction, model deployment, and evaluation modules for model deployment. First, it launches code from AWS SageMaker server which triggers execution of the code on a dedicated server. This procedure starts with the ingestion of the data, its preprocessing including feature extraction and the use of the classification algorithms including linear regression and Random Forest among others. Furthermore, scores on performance evaluation metrics are then calculated to determine the showcase of the model in encompassing the upshots in analyzing the plays of sales in the retail sector[4].

b. Flow Chart: A flowchart is a graphic representation of what is happening in a process or how a workflow is to be executed. In this, the steps are shown in sequential positions making it easier for one to have a clear visualization of what happens. It can be compared to like a chart that illustrates the protocol of an algorithm which is the main method of obtaining a solution or carrying a procedure out. Generally speaking, a flowchart contains distinct shapes to represent distinct actions or decisions which can be connected by arrows to show the sequences of the given steps[6].

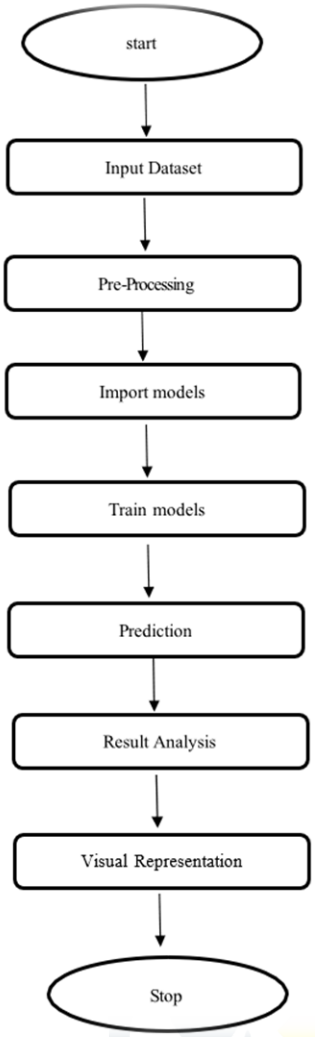

Fig: 2. Flow Chart

Flowchart depicting the streamlined process of retail sales prediction using machine learning: input, preprocessing, modeling, prediction, analysis, and visualization

VI. USE CASE DIAGRAM:

While the analysis phase, use cases obtain special significance to show how to achieve required functionality of the system - it separates actors and aims. In this workshop, actors play simultaneously a set of different system user roles which are diverse and can range from humans to other computational entities or software systems. The use case diagrams are excellent in gathering data on systems requirements and they highly take into account the impacts of internal and external issues on the system[8]. The design criteria are what majorly dealing with this aspect of the instructions. Although, the case is that there are use cases and the actors, but the most important thing is to analyse the system which helps in the clear picture and design phase of the system functionality.

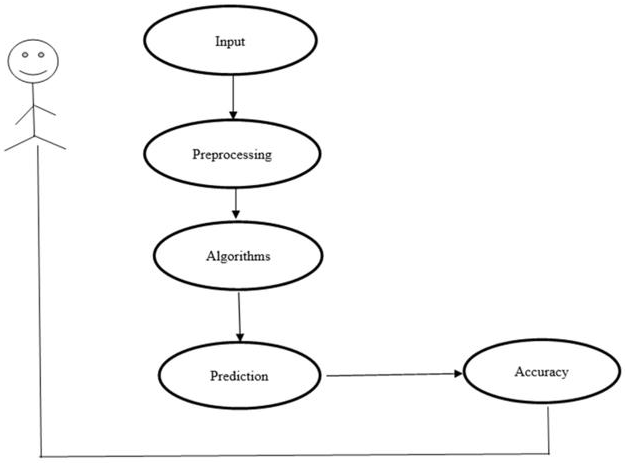

Fig. 3. Use Case Diagram.

Use case diagram depicting system functionality and actor interactions, crucial for requirements gathering and system analysis during the project's analysis phase

VII.IMPLEMENTATION

In our Project we have used three algorithms to predict the future sales:

- Linear Regression
- Random Forest
- Decision Tree

1. Linear Regression:

The linear regression is a base machine learning algorithm we use to determine functional relations between the dependent and the independent variable by means of a regression line. In this context, when we define the Rossman retail sales prediction problem, we assume that the dependent variable of this model represents sales figures, while the independent variables account for several factors such as store dimensions, geographical location, temporal aspects like week-day and time of year, promotional campaigns, and competitive environment.

Moreover, linear regression enables the quantification of the impact of independent variables on sales figures, providing insights into the relative importance of different factors in driving sales outcomes. By capturing both linear and non-linear trends in the data, linear regression facilitates the identification of underlying patterns and trends, enabling more accurate sales forecasts. Additionally, the interpretability of linear regression models makes them valuable tools for stakeholders seeking actionable insights into sales dynamics and performance drivers[9]. Overall, linear regression plays a pivotal role in informing strategic decision-making and optimizing sales strategies in the retail industry.

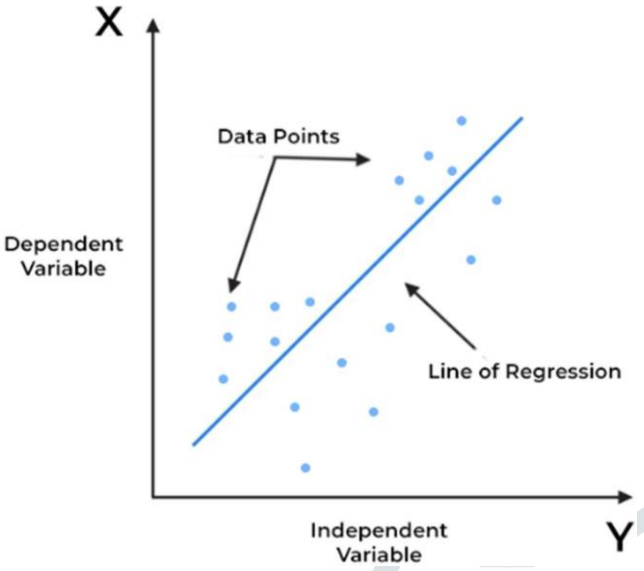

Fig. 4. - Linear Regression

Applying linear regression to predict retail sales for Rossman stores using historical data and relevant variables

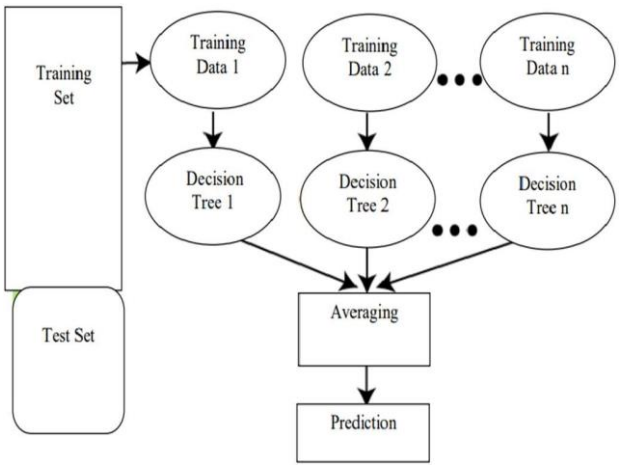

Fig. 5 - Random Forest

Unlocking Random Forest: Powerful ensemble learning for robust performance, handling imbalanced data, and simplifying web data scraping

The primary stages in model development consist of the assembling of historical data (independent variables) and the sales figures by using linear regression. These data generate the linear regression model and then use previously identified variables to estimate future sales from this model. Which is the goal of the linear regression is to find a linear equation that perfectly fit the data, implying the minimum error margins between the predicted and actual figures. Finally, this equation provides for predicting future sales utilizing new independent variables though it lacks much depth.

2. Random Forest:

In this study, studies on Random Forest are conducted and their performance are monitored by adjusting the number of trees used as another performance variable. Such method is used by a machine for its verification.

Random Forest technique operates with a high speed, generalization to irregular and missing data is its main feature. It is achieved by tying together due outcomes of numerous decision trees and thus the problem of overfitting is solved. What is more, it requires no data prep phases, e. g. scaling and neither. It turns out to be an appropriate solution in our study because of its capacity to work with diverse datasets of the retail sales forecasting nature. Through the use of its ensemble based approach, it does more than reducing the outlier effect and missing data; necessary factors for perfect predictions in complex market circumstances.

Furthermore, Random Forest emerges as an ideal solution for our study owing to its adaptability to diverse datasets inherent in retail sales forecasting. Leveraging its ensemble-based approach, Random Forest not only addresses outliers and missing data but also excels in navigating complex market dynamics. Consequently, it stands as a pivotal tool for achieving accurate predictions and insightful analyses within the retail industry's intricate landscape[5].

3. Decision Tree:

Decision trees, the core method of machine learning, effectively deal with the representation of decision-making steps in a tree-like form, where decisions are made at each fork. At each node you have to follow specific criteria, due to which there are branches you may continue the search. Within our universe, the Indecision forest is applied to estimate retail sales for Rossman stores. Training the model with the data covering store information such as size and location, seasonal patterns taken into account like weekday or holiday, communications and rivalry are included. Next, the model utilizes this data to forecast future sales figures when new store elements are entered, which will be used for strategic planning for the current model[6].

Subsequently, leveraging this rich dataset, the Decision Tree model extrapolates future sales figures when presented with new store attributes, thereby facilitating strategic planning and informed decision-making processes. By providing actionable insights into the drivers of sales performance, Decision Trees play a pivotal role in optimizing retail strategies and enhancing business outcomes.

Decision trees excel in their ability to provide transparent and interpretable decision-making frameworks, making them invaluable tools for stakeholders seeking actionable insights. Their intuitive representation of complex decision paths allows for easy understanding and communication of model outcomes. Additionally, decision trees are robust against outliers and missing data, ensuring reliable predictions even in imperfect datasets.

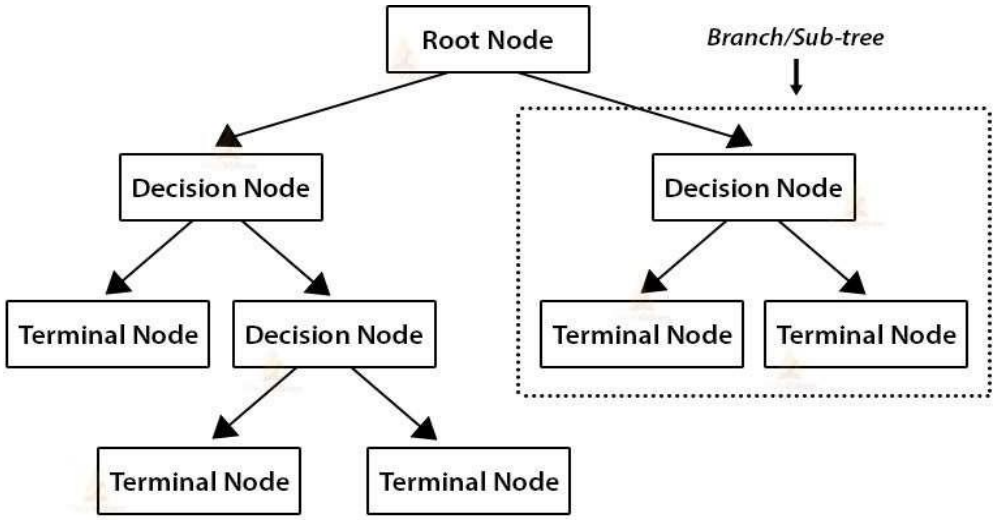

Fig 6. - Decision Tree

Utilizing Decision Trees to forecast Rossman store sales based on key factors, visualized through a concise flowchart

4. Testing:

- During our project the run of validating the models we used were tested rigorously to make sure they worked, were reliable and increase our usability to the customers.
- This test classification leads to the effective application of functional unit tests, integration tests, and user-centered tests while considering retail sales application for Rossman outlets[8].
- Close-to-the-last-word, we practice deep-down testing and after that fixing any bugs within our models. That allows to improve performance and increase accuracy by making our models more responsive and thus correspondent with the retail market trends.

VIII. CONCLUSION

The Rossman sales store dataset with ML methods plays as a promising tool for Mars sales forecast using ML algorithms. The utilization of historical sales data together with features like promotions and store size along the competition can provide ML models with the necessary ingredients to be able to efficiently forecast trend sales. In order to optimize the performance of the algorithm, it is advisable to utilize ensemble models. However, this may not provide desirable accuracy if data quality, feature selection, and algorithm selection and tuning are not taken into consideration. Besides, this approach creates practical recommendations ‘for storefronts to better tailor sales processes and improve profitability[9].

On the other hand, the cognitive variety of ML data algorithms allows users to connect varied data sources not only from Rossman dataset but also the services giving model endurance and powerful forecasting. As soon as data become out-dated, new models will be created as a result of which the forecast will remain adaptive to the changes in the market dynamics and consumer behaviors therefore, leading to higher accuracy over the period of time. Moreover, boosting the working of advanced tools like feature engineering and ensemble learning enhances the performance of model significantly and hence it helps companies to predict sales trends much more accurately[7].

In essence, the consumption of ML-driven sales forecasting will enable the companies with the preemptive abilities of foreseeing the market uncertainties and grabbing emerging possibilities. In this way, such utilization is inevitable for a company in a dynamic retail market. Organizations can both effectively fine-tune the resource allocation and enhance the decision-making processes by applying the capabilities of ML algorithms in providing predictable insights. Such an approach can become a foundation for a sustainable business growth.

IX References

[1]. The original dataset can be found on the Kaggle platform: https://www.kaggle.com/c/rossmann-store-sales/data

[2]. G. Gouttefangeas, "Sales Forecasting with Machine Learning: Understanding the Rossman Stores Dataset," Towards Data Science, 2020. [Online]. Available: https://towardsdatascience.com/sales-forecasting-with-machine-learning-understanding-the- rossman-stores-dataset-6f7f9c2f4ee7

[3]. A. Abualkishik, "Retail Sales Prediction using Machine Learning," Medium, 2021. [Online]. Available: https://a-abualkishik.medium.com/retail-sales-prediction-using-machine-learning-6aa2cb6d1db9

[4]. Y. Yang, S. Wang, J. Zhang, and J. Huang, "Sales prediction based on machine learning algorithms: A case study of the Rossman store sales dataset," IEEE Access, vol. 9, pp. 14080-14090, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9373559

[5]. H. Liu and C. Zhang, "A comparison of machine learning models for retail sales prediction: a case study of the Rossman store sales dataset," Journal of Intelligent & Fuzzy Systems, vol. 40, no. 6, pp. 11877-11889, 2021. [Online]. Available: https://content.iospress.com/articles/journal-of-intelligent-and-fuzzy-systems/ifs2021048

[6]. A. Makarychev, "Sales Forecasting for Rossmann Store Chain Using Machine Learning," Towards Data Science, 2020. [Online]. Available: https://towardsdatascience.com/sales-forecasting-for-rossmann-store-chain-using-machine- learning-a9eb7a83e526

[7]. M. Y. Al-Tahat, S. Alsmadi, and O. Al-Tarawneh, "Predicting Retail Store Sales Using Machine Learning Techniques: Case Study of Rossmann Store Sales Data," International Journal of Advanced Computer Science and Applications, vol. 10, no. 11, pp. 37-43, 2019. [Online]. Available: https://thesai.org/Downloads/Volume10No11/Paper_5-Predicting_Retail_Store_Sales_Using_Machine_Learning_Techniques.pdf

[8]. S. S. Hady and M. M. Rady, "Sales forecasting using machine learning algorithms: A case study for Rossmann stores," in 2019 4th International Conference on Advanced Technology & Sciences (ICAT), 2019, pp. 155-160. [Online]. Available: https://ieeexplore.ieee.org/document/8968486

[9]. S. Kim, J. Kim, J. Hwang, and Y. Seo, "Sales forecasting for retail stores using machine learning: a case study of the Rossmann store sales dataset," Journal of Open Innovation: Technology, Market, and Complexity, vol. 6, no. 4, p. 92, 2020. [Online]. Available: https://www.mdpi.com/2199-8531/6/4/92

[10]. R. Manikandan and R. P. Swaminathan, "Sales prediction in retail stores using machine learning: a case study of Rossmann store sales dataset," in 2021 International Conference on Inventive Research in Computing Applications (ICIRCA), 2021, pp.