Seasonal Sales Prediction and Visualization for Walmart Retail Chain Using Time Series and Regression Analysis: A Comparative Study

Rut Vyas[1], Revathi AS[2]
[1] U.G. Student, Department of Electronics and Telecommunications
[2] Assistant Professor, Department of Electronics and Telecommunications
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India
[1] rutvyas@gmail.com, [2] Revathi.as@djsce.ac.in

***Abstract*— The forecasting of sales forms the foundation of maximizing profits and successful planning operation at the start of the fiscal year for Business to consumer (B2C) models like Walmart retail chain. It is paramount for such companies to understand the needs and buying pattern of their customers. This paper puts forth some crucial sales visualizations that can prove indispensable to a store manager and performs a comprehensive evaluation of the different forecasting models to predict sales of a retail chain. One of the chief goals of the paper is to determine prediction models that can accommodate the seasonality in sales during four holidays.**

Keywords—time series, regression, Python, sales prediction models, seasonality, information visualization, Tableau.

I. INTRODUCTION

Multinational retail stores like Amazon, Costco and Walmart make millions of dollars in sales every day. For such fast moving B2C (Business to Consumer) corporations, it is imperative for them to understand the needs of their customers. Each customer segment has its own defining characteristic and big conglomerates spend a large amount of capital to identify this pattern using a wide range of models like Recency Frequency and Monetary (RFM) model [1].

Different departments bring in varying amount of revenue depending on the time of the year. Sales is further affected by a multitude of factors like geographical location, financial background of the population, age demographics, competition and so on. Analyzing these factors and by drawing insights from it, the retail chains are able to discern patterns in their sales and maximize their profits. By calculating restocking needs accurately, companies are able to reduce overheads and minimize the risk of running out of inventory. However, there are several difficulties faced by analysts that make sales prediction difficult. One of the major challenges is the presence of seasonality in the sales data. There is always a hike in incoming revenue around holidays where people go out to buy presents during Christmas or splurge on different commodities due to the huge Black Friday Sales around the Thanksgiving holiday. It is very important that the model used for sales prediction takes into account such anomalies that take place during holidays as retail chains have an opportunity to earn exponentially more in this time period. The PowerReviews Holiday Consumer Survey 2021 concluded that inspite of Covid 19, 82% of their audience planned on spending the same or higher during this holiday season. This paper will focus specifically on four such holidays and implement the model which will be able to best accommodate for seasonality.

II. LITERATURE SURVEY

Sales forecasting lies at the heart of retail stores and is an instrumental element in determining their profits and success. Much work has been done in developing models to predict the sales of such stores but the fact is that each brand has their own unique historical data and corresponding attributes. Hence, different machine learning algorithms need to be explored to fit one's characteristics.

An integration of XGBoost and LightGBM framework was proposed for Walmart sales prediction by Jingru Wang [2]. This hybrid model gave a lower root mean square error (RMSE) than individual models displaying stronger predictive ability and lower run time. However, gradient boosting is highly sensitive with respect to outliers because all classifiers are forced that they fix the errors of its predecessor learners and thus does not work well on seasonal data.

In [3], Kumari Punam et al. created a two-level statistical model where the bottom layer models (Linear Regression, support vector regression and cubist) used the original attributes of the data as input and the top layer model (cubist) took the predictions of lower layer as input and produced final predictions as output. But, with such a system, as the number of predictor variables and sales time period increases, the complexity and rum time increases significantly.

Prof. Sunil K Punjabi et al. predicted the sales of new automobiles through sentiment analysis from social media and expert opinions on web [4]. They used a polynomial regression model to fit their data. However, there are several factors like marketing campaigns, celebrity endorsements and last-minute collaborations that could derail the predicted sales as new factors need to be considered for determining the degree of the regression model.

Tomislav Hlupić et al. proposed a sales data mart on top of which the ARIMA model would be used for forecasting purposes [5]. To deal with multiple univariate series, they used the 'auto-arima' function in R as adjusting the parameters

individually for each time series would be time-consuming. However, one of their reasons for choosing ARIMA was to reduce seasonality but this paper focuses on incorporating seasonality into the prediction model and hence further analysis is required.

III. METHODOLOGY

Three datasets provided by Walmart spanning a time period of 3 years were used for this paper. First dataset consisted details of the sales made by a particular department of a specific store on that date, aggregated weekly. Second dataset comprised of the features like fuel price, temperature, markdowns, consumer price index (CPI), unemployment rate and if it was a holiday at that particular date in the specific store. Last dataset provided the type and area of the different stores of Walmart.

The implementation is divided into three sections. First is visualization insights that mangers can use to make preliminary conclusions about their store sales. Second has data cleaning and feature extraction steps to make the data model ready. Third section consists of a comparison of different prediction models to determine the most accurate one for this scenario.

A. Data Exploration

The data comprises of the weekly sales information for 45 different Walmart stores that are further divided into three categories: A, B and C. Managers need to know which of their departments are performing the best so they can allocate man power and other resources accordingly. A top-n department filter as shown in Fig. 1 will help them do exactly that for their store.

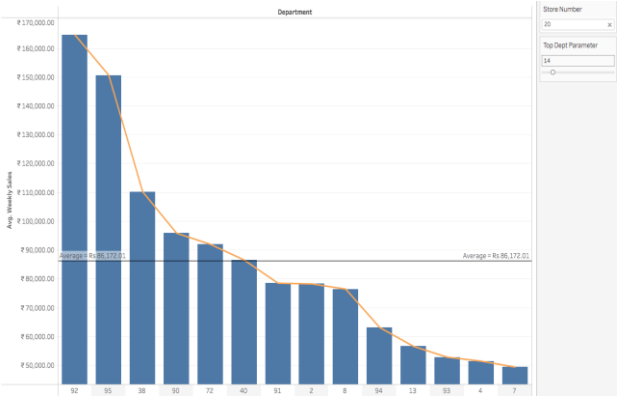

Fig. 1. Top-n Department Sales by Store

Opening a new store consumes a mammoth proportion of resources and is a huge commitment. Therefore, the companies always need to perform a preliminary analysis based on their existing store data to get an approximate idea whether the return on investment (ROI) is substantial. This depends on factors like geographical location, consumer background, store size and so on. Here, in Fig. 2, it is seen that a polynomial trend model of degree 2 can approximately predict the average weekly sales based on its relationship with the size of the store. The coefficient of determination (R^2), also known as the 'goodness of fit', helps in determining the variability of one factor due to its association with another factor. Here, it is 0.66 which states that the model explains 66% of variation in data and the p-value is less than 0.0001 through which we can conclude that the model is significant enough to be considered.

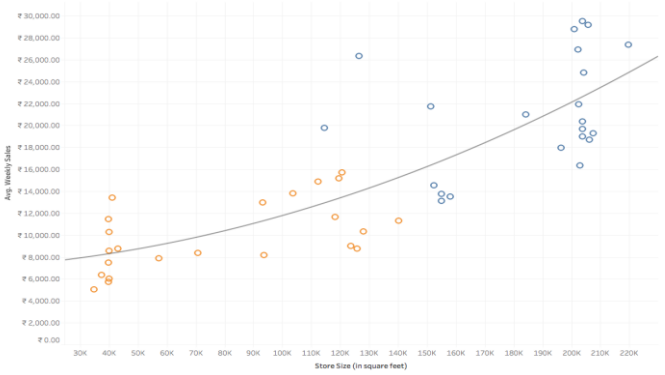

Fig. 2. Store size vs. Average Weekly Sales

On comparison of median and average weekly sales of holiday and non-holiday dates in Fig. 3, we can see that there is huge gap between the two. A conclusion is drawn from this that some departments are doing substantially better than other departments due to which the median sales is almost 50% less than the average sales.

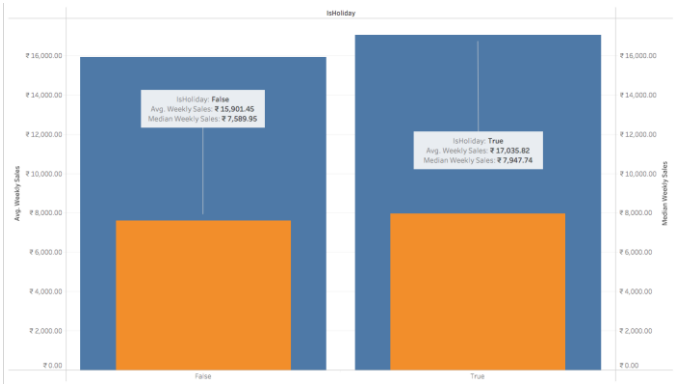

Fig. 3. Bar in Bar Chart of Median and Average Weekly Sales

Furter analysis yields the fact that over the years sales is at its peak during the Thanksgiving as compared to other holidays. This can be attributed to the huge discounts provided by the 'Black Friday Sales' where consumers splurge on all types of products. A tree map created for each store manager helps them see how their departments performed on specific holidays and determine their inventory needs accordingly for future holidays. One such tree map can be seen in Fig. 4.

All 45 stores are divided into three types. A box and whisper plot was created and each data point in Fig. 5 represents the weekly sales on that date. It is clear from the plot that stores of type A almost always perform better than stores of type B and stores of type B almost always perform better than stores of type C. This information can be instrumental in determining the sales of new stores of a particular type and it is worth exploring what this store type stands for. Most of the outliers represent sales on dates surrounding Thanksgiving and

Christmas. This is proof that there is seasonality present in the given data.

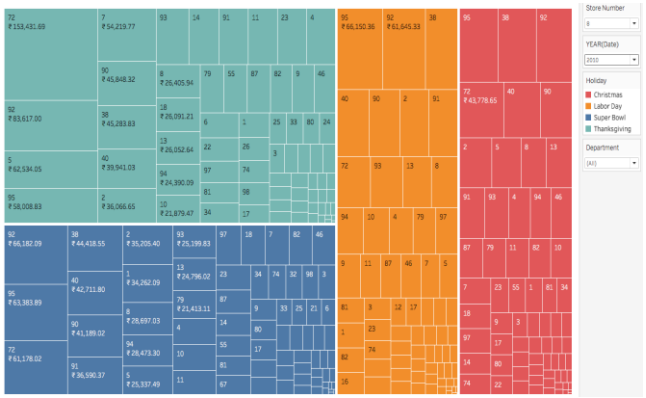

Fig. 4. Tree map of Average Weekly Sales by Holiday and Store

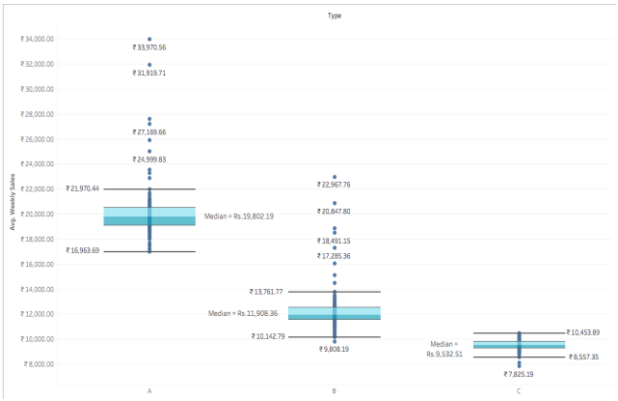

Fig. 5. Store Type Effect

Lastly, the different features provided in the dataset were compared to the average weekly sales to see if they have any correlation ship. It was observed from Fig.6 that temperature and fuel price have no effect on sales whatsoever. Average weekly sales remained relatively constant once the unemployment rate crossed 11 but more data needs to be mined to get a definitive answer. Though, the departments clearly have an effect on sales and it is probable that higher earning departments have highly marked products like electronics or regularly bought items like food products.

Fig. 6. Features vs. Average Weekly Sales

B. *Data Pre-Processing*

The three datasets were combined using an inner join on the 'Store' and 'Date' column which provided details of the store number and weekly sale date respectively. The data had 4,21,570 rows and 17 columns before cleaning. Multinational retail chains carry out a huge number of transactions in a day and there is a possibility that errors may creep in here due to incorrect reporting by employees or incompatibility of old software used by some store. Hence, it is very important to clean the data and remove incorrect entries to the best of one's ability so as to improve the predictive power of the model.

The number of null values were calculated in the data set and they were replaced the digit '0'. In the 'Weekly_Sales' column, all the negative sales tuple were dropped just in case there was some error while reporting. During the joining process, two extra columns namely 'IsHoliday_x' and 'IsHoliday_y' were formed. One of them was retained and the other recurring column was dropped. Further, Spearman's rank correlation coefficient was used to find out what kind of relationship does the weekly sales share with the other features as seen in Fig. 7. This will enable one to drop the features that are irrelevant to improve the accuracy of forecasting models

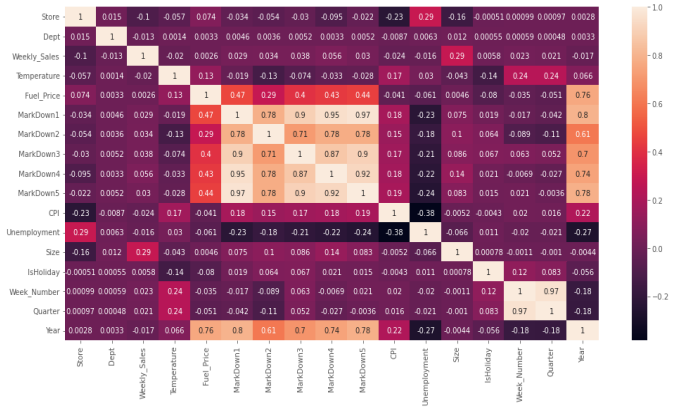

Fig. 7. Spearman's rank correlation coefficient

The date column is converted into a datetime datatype to work with date objects. To fit the model parameters, categorical variables need to be converted into numerical variables. Hence, the store types A, B and C are transformed to the values 1,2 and 3 respectively while the values True and False in Holiday column are converted into binary digits 1 and 0 respectively. Based on Fig. 7, we remove the unnecessary columns of Unemployment, CPI and Markdown to help the model provide more accurate results. Finally, the data is converted into train and test data with 70% values given to the training set and the remaining 30% to the testing set. The train data is further split into train and cross validation set with a 70:30 split ratio. Cross-validation is used to examine a model's capacity to forecast new data which was not used in its estimation, in order to identify issues such as overfitting or selection bias [6], as well as to provide insight into how the model will generalize to an independent dataset. In the end, we are left with 15 columns and 2,05,939 rows in training set, 88,260 rows in cross validation set and 1,26,086 rows in testing set.

Two kinds of performance parameters are used to determine the accuracy of the prediction models. First one is weighted mean absolute error (WMAE) and second parameter used is root mean square error (RMSE) which are given by:

$$WMAE = \frac{1}{\Sigma w_i}\sum_{i=1}^{n} w_1 |y_i - \hat{y}_i| \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \quad (2)$$

where,

- n is the number of rows
- $\hat{y}_i$ is the predicted sales
- y_i is the actual sales
- w_i are the weights. It is 5 on holiday and 1 on non-holiday to focus more on holiday sales prediction.

C. Prediction Models

1) Time Series Models: It refers to an ordered series of data. Features of the dataset are not taken into account in these models. Decomposition of time series data using additive model was performed to see whether trend, seasonality and residual are present. It was observed from Fig. 8 that:

- Trend of weekly sales is decreasing till March and then increasing again.
- It was already known that data contains seasonality around holidays which is corroborated by the symmetric seasonal graph.

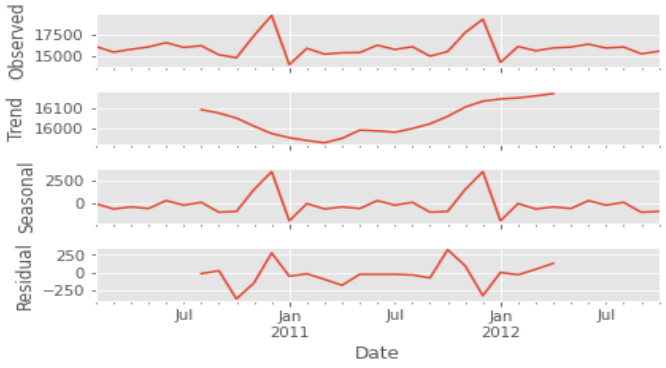

Fig. 8. Decomposition graph

a) ARIMA: It stands for Autoregressive Integrated Moving Average. This model is used when the past data itself is sufficient and creates patterns that are helpful in future prediction while not being dependant on exogenous variables. As it takes univariate data, we drop all columns except the weekly sales. The model has three parameter namely auto regressive (p), moving average (q) and integrated (d) components that need to be determined. To check whether the time series is stationary, the Augemented Dickey Fuller test was performed. This is a procedure that tests the null hypothesis which states that a unit root is present in the time series data. A p-value of 0.0007 was obtained which is way less than 0.05 threshold. Thus, the null hypothesis can be rejected since the data does not have a unit root and the time series is declared stationary. Additionally, since the ADF Statistic value of -4.17 is smaller than the 1% critical value of -3.76, the decision of rejecting the null hypothesis can be further verified. Thus, no differncing needs to be performed and the 'd' value becomes 0. After running several iterations, the 'p' value was chosen as 2 and the 'q' value as '1' because the p-value of constants were were less than 0.05 and the Akaike information criterion (AIC) was least with this combination of parameters.

b) Holt-Winters: Holt-Winters is a time-series model that can be used to simulate three components of a time series: the average value, the slope (trend) across time, and the cyclical repeating pattern (seasonality). It can handle the seasonality in the data set by just calculating the central value and then adding or multiplying it to the slope and seasonality Holt-Winters employs exponential smoothing to encode a large number of historical values and use them to forecast "typical" values for the present and future. Seasonal periods are the number of periods in a complete seasonal cycle and here it is 7 as this reopistory has daily data with weekly cycle. Through experimentation, additive trend and additive seasonality are chosen as they provide the lowest AIC value.

2) Regression Models: It can be used on ordered and non-ordered series in which the values of a target variable are influenced by the values of other variables. These additional variables are referred to as features. New values of features are provided when constructing a prediction, and regression analysis delivers an answer for the target variable. Regression is essentially an intrapolation technique.

a) Linear Regression: The equation used in regression analysis is

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_r x_r + \varepsilon \quad (3)$$

where $\beta_0, \beta_1, \ldots, \beta_r$ are the regression coefficients, and ε is the random error. The training set is split into two components namely, dependant (weekly sales) and independant (features) values. Date column needs to be removed so that the model can fit the data. We notice 15 slopes as slope is calculated for each indepndant variable. It was observed that the coefficient of determination (R^2) was 0.087 which means a very low amount of variation in y can be explained by the dependance on x using this model [7]. This was already observed by the poor values of the Spearman's rank correlation coefficient in Fig. 7.

b) Decision Tree Regression: This prediction model determines a targeted value using a collection of binary rules [8]. The mean squared error (MSE) is commonly used in decision trees regression to determine if a node should be split into two or more sub-nodes. A number of itertions were carried out by changing two hyperparameters: maximum depth of the tree and minimum number of samples for a terminal node. By considering the model that gives least training WMAE and cross validation WMAE, depth of 25 to 35 was shortlisted. Decision trees are prone to overfitting and hence the tree is pruned to a maximum depth of 25 and minimum sample leaf of 6. Fig. 9 shows the final decision tree.

c) Random Forest Regression: It is a supervised learning algorithm for regression that use the ensemble learning method. It fits a number of classifying decision trees on various sub-samples of the dataset and employs averaging to increase the prediction accuracy.It aslo controls the problem of over-fitting observed in the Decision Tree regression model due to the Law of Large Numbers [9]. A number of itertions were carried out by changing two hyperparameters: maximum depth of the tree and number of trees in the forest. By considering the model that gives least training WMAE and cross validation WMAE, 90 trees each having a maximum depth of 35 were fed as parameters into the model.

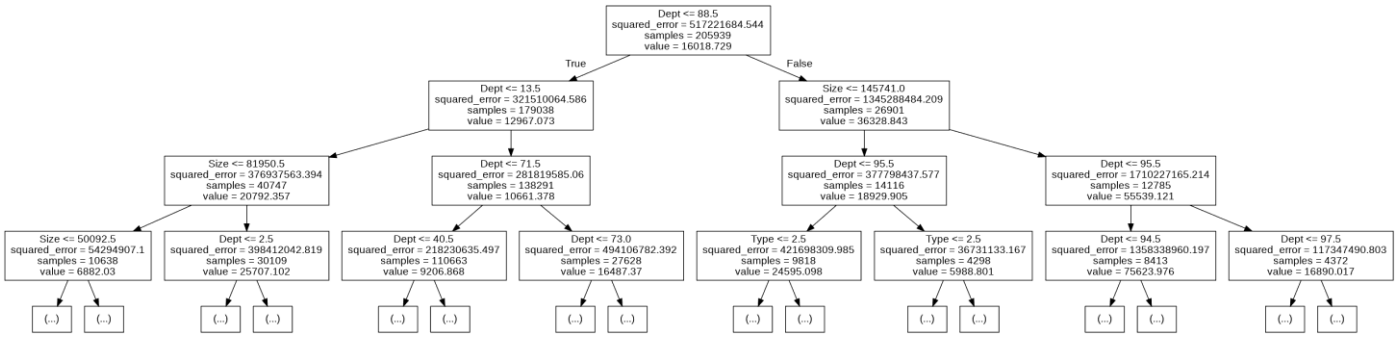

Fig. 9. Decision tree displaying a depth of 3

IV. Result

The time series models and regression models are compared separately as the former does not take into account the features of the dataset. Based on a number of performance parameters, the observations are as seen in Table 1 and Table 2.

TABLE I. Time Series Model Performance Metric

Model	RMSE	Training Time
ARIMA	366.829	1.35 seconds
Holt-Winters	535.898	0.054 seconds

It is clear that among the time series models, ARIMA performs better than the Holt-Winters model. However, the time taken for training the dataset in ARIMA model is about 25 times the time taken by Holt-Winters model.

TABLE II. Regression model Performance Metric

Model	R^2	RMSE	WMAE	Training Time
Linear Regression	0.087	21737.6	14832.4	0.042 seconds
Decision Tree Regression	0.947	5248.31	667.94	1.562 seconds
Random Forest Regression	0.96	3804.77	520.69	130.827 seconds

According to the coefficient of determination (R2), Random Forest model can explain 96% of variation in data closely followed by the Decision Tree model. However, from RMSE and WMAE parameters, it is clear that Random Forest model is the best fit for prediction purposes for this Walmart dataset. However, the training time for the Random Forest is approximately 84 times the time taken by Decision Tree model. Linear regression model may have a very low run time but its prediction capabilities are very low. This was to be expected as linear models are mostly bound to fail when used on seasonal data.

One may expect the regression models to work better than the time series model as several features were considered within its framework. But for this dataset, it was observed from Fig. 7 that the weekly sales are very weakly correlated to some features and has almost no correlation to the rest. Thus, it can be hypothesized that the regression models in their attempt to use features to make prediction actually ended up performing worse than the time series model.

V. Conclusion

Every model has its own set of characteristics that they bring to the table. It is upon the businesses to decide whether they want to sacrifice accuracy for lower run time or vice-versa depending on their operations. However, large conglomerates conduct preliminary predictions at the beginning of the fiscal year since they need to not only provide a target report to their investors and board of directors but also make commitments to suppliers and store managers. Thus, it may be beneficial to focus on accuracy during such preliminary investigations.

It is worth exploring XGBoost, Gradient Tree Boosting or a hybrid model which will combine weak learners to produce a strong prediction rule especially to account for holiday seasonality. Future work will be concentrated on creating strong learners that can eventually improve the prediction ability through boosting algorithms.

REFERENCES

[1] I. Maryani and D. Riana, "Clustering and profiling of customers using RFM for customer relationship management recommendations," 2017 5th International Conference on Cyber and IT Service Management (CITSM), 2017, pp. 1-6, doi: 10.1109/CITSM.2017.8089258.

[2] J. Wang, "A hybrid machine learning model for sales prediction," 2020 International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI), 2020, pp. 363-366, doi: 10.1109/ICHCI51889.2020.00083.

[3] K. Punam, R. Pamula and P. K. Jain, "A Two-Level Statistical Model for Big Mart Sales Prediction," 2018 International Conference on Computing, Power and Communication Technologies (GUCON), 2018, pp. 617-620, doi: 10.1109/GUCON.2018.8675060.

[4] S. K. Punjabi, V. Shetty, S. Pranav and A. Yadav, "Sales Prediction using Online Sentiment with Regression Model," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 2020, pp. 209-212, doi: 10.1109/ICICCS48265.2020.9120936.

[5] T. Hlupić, D. Oreščanin and A. -M. Petric, "Time series model for sales predictions in the wholesale industry," 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO), 2020, pp. 1263-1267, doi: 10.23919/MIPRO48935.2020.9245255.

[6] Cawley, Gavin C.; Talbot, Nicola L. C. (2010). "On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation" (PDF). 11. Journal of Machine Learning Research: 2079–2107.

[7] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

[8] L. Breiman, J. Friedman, R. Olshen, and C. Stone, “Classification and Regression Trees”, Wadsworth, Belmont, CA, 1984.

[9] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001).