ISSN: 2278-4632
Vol-13, Issue-11, November 2023

BLACK FRIDAY SALES PREDICTION USING MACHINE LEARNING

Thalari Abhinav [1],P.Krishna Prasad[2]
[1]MCA Student, Chaitanya Bharathi Institute of Technology (A), Gandipet, Hyderabad, Telangana State, India
[2]Assistant Professor, Department of MCA, Chaitanya Bharathi Institute of Technology (A), Gandipet, Hyderabad, Telangana State, India

ABSTRACT

Understanding the buying patterns of diverse consumers (dependent variable) concerning various products using their demographic data (IS characteristics, mostly self-explanatory) is the goal of this study. The dataset presents challenges with redundant, unstructured, and null values. Leveraging machine learning, the retail industry frequently utilizes this dataset to aid shop owners in inventory management, financial planning, promotion, and marketing by developing a predictor with clear commercial value. The process involves pre-processing, modeling, training, testing, and evaluation. To streamline and simplify the approach, frameworks will be implemented to automate certain steps. Among the regressors tested, the XGB Regressor stood out as the best performer, achieving an RMSE value of 2529.3684, making it the ideal choice for the predictive model.

KEYWORDS:

Sales prediction, Regressor, Random Forest, Machine Learning,RMSE

I. INTRODCUTION

The day following Thanksgiving, known as "Black Friday," has become synonymous with shopping extravaganzas and great bargains, driving a massive influx of customers to retail stores across the United States. Originally named for the chaotic traffic and violence it caused, Black Friday has evolved into a carnivallike sales event. For retail companies, the volume of sales on this day can make the difference between profit and loss. Therefore, accurate sales forecasting becomes crucial for effective industry management.

To enhance sales predictions during Black Friday, companies are now turning to data-driven approaches. By carefully organizing and analyzing customer data, they aim to uncover relationships between independent variables and the target variable, which, in this context, refers to sales of various products. The creation of robust prediction models allows businesses to better anticipate and cater to customer demand, thereby maximizing their profits during this critical shopping event.

Efficient data organization and thorough analysis are essential to establish meaningful relationships between different variables, enabling accurate sales estimations for various products based on their independent variables. By structuring the data thoughtfully and conducting comprehensive analyses, a model can be trained to perform computations and make precise sales predictions. This process involves understanding how the independent variables impact sales and uncovering patterns and correlations in the data. Armed with valuable insights gained through rigorous examination, the model can make informed predictions, assisting businesses in optimizing sales strategies and achieving improved forecasting accuracy.

The Two goals are emphasized in this study are:

- Exploring all the relevant client data to understand how the independent variables influence the target variable.
- Projecting sales via testing and training

II. LITERATURE SURVEY

Beheshti-Kashi et al [1] presented the methods for sales forecasting in consumer-oriented markets, focusing on fashion and new product industries. Overcoming challenges of uncertain demands and limited historical data, their study explores innovative strategies leveraging user-generated content and search queries to enhance predictive accuracy. Their survey paper provides valuable insights for accurately predicting sales in dynamic markets.

Smith, Oliver at al[2] discussed the uncertainty surrounding the permanence of Black Friday as a shopping event in the UK. They observed that major retailers quickly updated their websites to promote the event, hinting at its potential continuation. They tentatively suggest that the data indicates a strong possibility of Black Friday's recurrence, with the event becoming a competitive arena where success is measured through shopping competence.

Majumder et al. [3] conducted research to explore the relationship between purchase behavior (dependent variable) and various products, utilizing customer demographic information (independent features). The dataset encountered challenges such as null values, redundancy, and unstructured data. To address these issues, the authors applied machine learning, specifically the Random Forest regressor algorithm, to create a predictive model with commercial value. This model proved beneficial for shop owners, as it facilitated inventory management, financial planning, advertising, and marketing decisions. The proposed approach achieved an average accuracy of 83.6% and a RMSE of 2829 on the Black Friday sales dataset. The researchers also developed frameworks to automate several stages of the process, thereby reducing complexity and streamlining the analysis.

Challagulla et al. conducted a study aimed to model the effectiveness of various kinds of machine learning methods in predicting the software defects. Through their empirical analysis, they sought to enhance software quality and reliability by leveraging advanced machine learning approaches for defect

ISSN: 2278-4632
Vol-13, Issue-11, November 2023

prediction. Their findings contribute valuable insights to the field of software development and quality assurance.

In their research, Chu et al. (citation [5]) conducted a comparative analysis of different linear and nonlinear models for forecasting aggregate retail sales. Acknowledging the substantial seasonal fluctuations in retail sales, the study explored conventional seasonal forecasting methods, such as time series and regression approaches, alongside their nonlinear counterparts using neural networks. The researchers also delved into issues concerning seasonal time series modeling, including deseasonalization techniques. The results revealed that the nonlinear models exhibited superior out-of-sample forecasting performance compared to linear models. Notably, the neural network model's predictive accuracy showed significant improvement when the data underwent prior seasonal adjustment. Ultimately, the neural network built on deseasonalized time series data emerged as the most effective model overall. However, the study emphasized the limitations of dummy regression models and found that trigonometric models were not suitable for accurate aggregate retail sales forecasting.

Makridakis at al [6] in their study covered diverse techniques and their practical uses. Emphasizing data-driven decision-making, it provides insights into time series analysis, statistical methods, and machine learning. They highlighted the significance of accurate forecasting for optimizing inventory, finance, and resource allocation.

Correia et al[7] in their study focused on exploring the concept of joints in Random Forests, a popular machine learning algorithm. They investigated the integration of these joints within the Random Forest framework, aiming to enhance the algorithm's predictive capabilities and understanding of data dependencies. Their findings contribute to advancing the effectiveness and interpretability of Random Forests for various applications in the field of machine learning.

Kvalheim et al. [8] performed a study titled "Determination of Optimum Number of Components in Partial Least Squares Regression from Distributions of the Root-Mean-Squared Error Obtained by Monte Carlo Resampling." Their objective was to identify the optimal number of components for partial least squares regression using Monte Carlo Resampling. Through their research, they sought to determine the number of components that yielded the most accurate predictions and improved the performance of the partial least squares regression model. They analyzed the rootmean-squared error distributions to improve the model's accuracy and predictive performance. Their findings provide valuable insights for enhancing the efficiency of partial least squares regression and optimizing the selection of components for various applications in chemometrics and related fields.

Sheridan et al [9] explored the application of extreme gradient boosting (XGBoost) in the field of quantitative structure-activity relationships (QSAR). They focused on leveraging XGBoost, a powerful machine learning algorithm, to develop robust QSAR models. The findings demonstrate the effectiveness of XGBoost in predicting molecular properties and activity, providing a valuable and efficient approach for analyzing chemical data and supporting drug discovery and other molecular design tasks.

III. METHODOLOGY

A. SALES DATA

The dataset contains sales transactions recorded at a retail store, offering an excellent opportunity to delve into feature engineering and gain valuable insights from diverse shopping experiences. This dataset represents a regression problem, where the goal is to predict sales figures based on various input features. Derived from AnalyticsVidhya[10] and featured in a hackathon project, this dataset challenges participants to explore and unleash their data analysis and modeling skills. Through this project, participants can refine their understanding of the retail domain, uncover patterns in customer behavior, and develop effective predictive models to optimize sales forecasting.

B. DATA PREPROCESSING

During the data preparation step, two datasets are merged into a single dataset named "combined." The "test" dataset's "Purchase" column is added to match the structure of the "sales" dataset, with the new column containing NaN values for test data. A new column named "data" is introduced to differentiate between training and testing data. Missing data in the "Product_Category_2" column is imputed with random values based on the existing distribution of categories in the dataset. These steps ensure that the "combined" dataset is ready for further analysis and modeling in the study.

The next step after data preparation is Exploratory Data Analysis (EDA). During the univariate analysis of the "train" data, I visualized the distribution of customers based on their genders, ages, occupations, city categories, duration of stay in their current city, and marital status using countplots. This allows to gain insights into the gender composition, age demographics, occupational backgrounds, geographical representation, residential stability, and marital demographics of the customer base.Figure 1 hows that the majority of the customers that purchase things during the sales season mainly belong to the age group of 26-35 and 3645.

ISSN: 2278-4632
Vol-13, Issue-11, November 2023

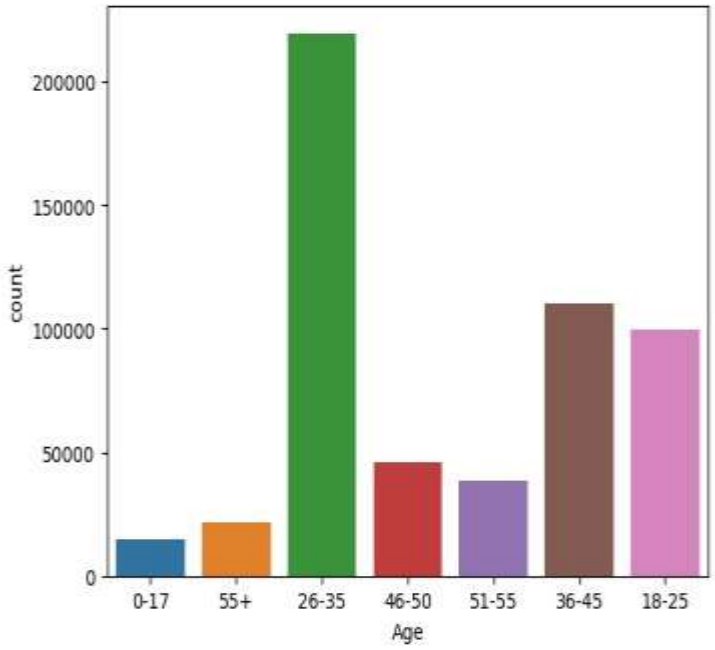

Figure 1: The graphical representation illustrates the dominance of the 26-35 and 36-45 age groups during the sales season.

After conducting the univariate analysis, I proceeded with bivariate analysis to explore the relationships between various variables. For instance, I examined the Average purchase amount against age groups to reveal spending patterns by different age demographics. Additionally, I analyzed the relationship between Average purchase amount and the duration of stay in the current city to understand how residency duration impacts purchasing behavior (as shown in Figure 2). Moreover, I used bar plots to investigate how marital status influences the average purchase amount, providing insights into potential spending differences between married and unmarried customers. To gain an understanding of popular and revenuegenerating products, I identified the Top 10 products with the highest sales. Additionally, count plots with gender as a hue were utilized to compare customer distribution based on marital status, offering insights into gender-specific trends. Furthermore, count plots with city category as a hue were employed to examine customer occupations across different geographical locations. These analyses contribute to a comprehensive understanding of the data and offer valuable insights into customer behavior and preferences.

By conducting a thorough univariate and bivariate analysis, I have uncovered valuable insights within the dataset, revealing crucial patterns, trends, and potential relationships between various variables. These findings will serve as a solid foundation for informing the modelling strategies and guiding informed decisionmaking for addressing the regression problem effectively. Armed with these comprehensive insights, we are better equipped to optimize sales forecasting, enhance industry management, and achieve successful outcomes during Black Friday, the bustling carnival-like sales event.

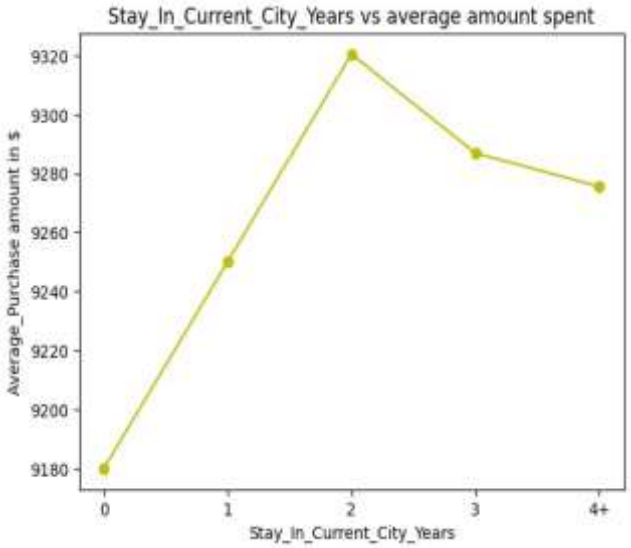

Figure 2: Representing Long-term residents spend more during Black Friday sales.

During data preprocessing, various transformations are applied to prepare the dataset for analysis and modelling. Numeric representations are assigned to certain values, and specific prefixes are removed from others. Data types are converted to ensure numeric compatibility where necessary. Adjustments are made to remove certain notations in one column. Values in another column are mapped to integers. Additionally, one-hot encoding is performed on a column, creating dummy variables to represent different categories without introducing ordinality. These preprocessing steps optimize the dataset for further analysis, facilitating the regression problem.

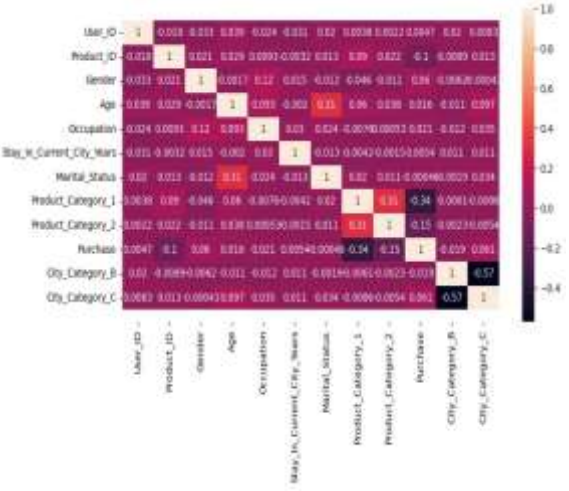

Figure 3: Heatmap to show the correlation between various variables of the train data set

Figure 3 represents an intensity-map between various variables in the "train" dataset and the correlation between them. Correlation values range from -1 to 1, where positive values indicate a positive correlation, negative values represent a negative correlation, and values close to 0 imply a weak or no correlation. The intensity-map reveals significant positive correlations between "Marital_status" and "Age," as well as "Product_Category_1" and "Purchase,"

ISSN: 2278-4632
Vol-13, Issue-11, November 2023

suggesting potential relationships between these variables. Additionally, a positive correlation is observed between "City_Category_B" and "City_category_A," indicating a connection between these city categories. The heatmap serves as a useful tool for identifying interdependencies and associations among variables, aiding in data analysis and model development for the regression problem.

IV. MACHINE LEARNING (ML) MODELS

In this study, ML Regressors were applied, namely Linear Regression(LR), Decision Tree Regressor(DT), Random Forest (RF)Regressor, XGBoost Regressor, and Extra Trees(ET) Regressor, to predict and model the relationship between variables for the regression problem. The entire workflow is shown in the below figure 4

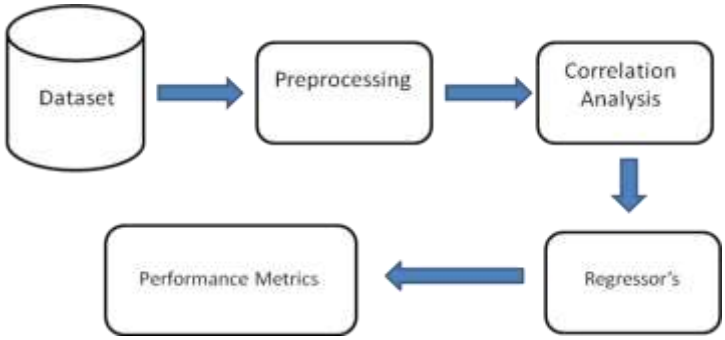

Figure 4: Black Friday sales prediction Architecture

The significance of each algorithm is explained below.

Linear Regression (LR)
LR[12] is a widely-used regression algorithm that models the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the input features and the target variable. The objective of linear regression is to identify the optimal straight line (or hyperplane in higher dimensions) that minimizes the difference between the actual and predicted values. The formula for simple linear
regression with one independent variable can be expressed as

$y=\beta_0+ \beta_1*x$.....(1)

Where:
y denotes predicted value . x
denotes input feature . β_0 denotes
y-intercept (bias term).
$\beta_1$is the coefficient (slope) associated with the input feature x.

Decision Tree Regressor (DT):
A DT[13] Regressor is a tree-based algorithm specially designed for regression tasks. It operates by recursively partitioning the data based on the input feature values, creating a hierarchical tree structure. Each internal node in the tree corresponds to a decision based on a particular feature, while the leaf nodes store the predicted values for the target variable. The prediction formula for a sample x in a decision tree can be represented as:

Prediction=Value of leaf node corresponding to the path taken by sample x.

Random Forest Regressor (RF):
RF[11] is an ensemble method that enhances prediction accuracy and mitigates overfitting by combining multiple decision trees. By constructing numerous decision trees using random subsets of the data and features, this approach achieves its objective. The final prediction is obtained by averaging the predictions of all the trees in the forest. The prediction formula for a random forest regressor is:

$P= (1/N)\sum P_i$.....(2)

Where:
P be the final prediction of the RF model.
P_i be the prediction of the i^{th} DT in the forest.
N be the total number of decision trees in the RF.

XGBoost Regressor:
XGBoost[14] (Extreme Gradient Boosting) is widely recognized as a powerful gradient boosting algorithm that excels in diverse machine learning tasks, particularly regression. It sequentially adds decision trees to the model, with each tree correcting the errors made by the previous ones. It uses a regularization term to control overfitting and employs efficient optimization techniques for faster training. The prediction formula for XGBoost is the sum of predictions from individual weak learners (decision trees), each multiplied by a corresponding weight:

$Prediction=\sum Weight_i.Prediction_i$.....(3)

Where:
Prediction is the final prediction of the XGBoost model, $Prediction_i$ is the prediction of the i^{th} decision tree in the XGBoost model,
$Weight_i$ is the weight assigned to the i^{th} decision tree based on its contribution to the overall performance.

Extra Trees Regressor (ET):
Extra Trees[15] is another ensemble algorithm that extends the idea of RF's. Like RF's, it builds multiple decision trees and aggregates their predictions. However, Extra Trees further randomizes the tree-building process by considering random splits for each feature, which increases diversity and reduces variance. The efficiency of ET in training is attributed to its randomness, which makes it faster compared to RF but might require more trees to achieve the same performance. The prediction formula for Extra Trees is similar to that of Random Forests:

$Prediction= (1/N)\sum Prediction_i$.....(4)

Where:
The prediction refers to the final prediction made by the ET model.
$Prediction_i$ represents the prediction of the i^{th} DT within the ET model.
The total number of decision trees in the Extra Trees model is denoted by N.

In this study, various regression models have been trained and evaluated using multiple tuning parameter settings. The models employed include LR, DT, RF, XGBoost Regressor, and ET. For each model, different combinations of hyperparameters have been explored to find the optimal settings that result in the best predictive performance.

ISSN: 2278-4632
Vol-13, Issue-11, November 2023

For the DT, two different instances have been trained with distinct tuning parameter values. One instance, DT, was configured with a maximum depth of 15 and a minimum number of samples required in a leaf node set to 100. The other instance, DT2, had a maximum depth of 8 and a minimum samples leaf constraint of 150. This variation in tuning parameters allows the DT models to consider different levels of complexity and granularity in their splits, potentially affecting their predictive capabilities. Similarly various different instances have been trained with distinct hyperparameter values for other Regressor's too.

Overall, by systematically exploring various hyperparameter settings for each regression model, this study aims to identify the optimal configurations that lead to the highest predictive performance. This thorough evaluation process ensures that the selected models are well-suited for the specific problem at hand and can make accurate predictions on unseen data.

V. RESULT ANALYSIS

The results of the regression models were analyzed using two commonly used performance metrics: R-squared (R2) score and Root Mean Squared Error (RMSE). They provide valuable understandings into the accuracy and goodness-of-fit of the models, allowing for a comprehensive evaluation of their predictive capabilities.

R-squared (R2) Score: The R-squared score is a statistical measure representing the proportion of variance in the dependent variable (target) explained by the independent variables (features) in the model. It ranges from 0 to 1, where 0 indicates no explanation of variance in the target variable, and 1 signifies a perfect fit, where the model explains all the variance.

$$R2=1-(\text{Unexplained variance}/\text{Total variance}) \quad \text{.....(5)}$$

Root Mean Squared Error (RMSE): The RMSE is a metric used to assess the accuracy of a model's predictions. It is calculated as the square root of the average of the squared differences between the predicted values and the actual target values. The RMSE is expressed in the same units as the target variable, and lower values indicate better model performance.

By using both the R-squared score and RMSE, the analysis provides a comprehensive assessment of the regression models' performance. The R-squared score helps in understanding how well the models explain the variability in the target variable, while RMSE quantifies the accuracy of the predictions in real-world units.

Table 1: Performance metrics

Regressor	R2 Score	RMSE
Linear Regression	0.1318	4685.9198
Decision Tree	0.1327	2734.3359
Random Forest Regressor	0.7126	2695.9834
ExtraTreesRegressor	0.6817	2837.0745
XGB Regressor	0.7470	2529.3684

In the evaluation of various regression models shown in table 1, the RMSE was used as a key performance metric to assess their predictive accuracy. Among the models tested, the XGB Regressor emerged as the most accurate, displaying the lowest RMSE value of 2529.3684. This result indicates that the XGB Regressor is effective in minimizing prediction errors, making it a strong candidate for applications where precision is crucial. Following closely, the RF Regressor and ET Regressor demonstrated competitive performance, achieving RMSE values of 2695.9834 and 2837.0745, respectively. These ensemble-based methods proved their ability to make accurate predictions with relatively low error rates.

In contrast, the DT model exhibited an RMSE of 2734.3359, falling within the same range as the ensemble methods but slightly higher. While Decision Trees are capable of capturing complex relationships in the data, they may not match the accuracy of ensemble models due to their susceptibility to overfitting. Whereas the LR model displayed the highest RMSE of 4685.9198, indicating that it may not be the most suitable choice for this particular regression task. LR assumes a linear relationship among the different varibles, which might not fully capture the underlying complexity in the dataset.

Overall, the XGB Regressor stands out as the top performer, showcasing its effectiveness in minimizing prediction errors and providing superior predictive power. The ensemble-based methods, including the RF Regressor and ET Regressor, also demonstrated promising results. However, careful consideration should be given to the selection of the appropriate model based on the various characteristics of the given data and the overall goal of the regression task. By using RMSE as the evaluation metric, these findings offer valuable insights into the relative strengths and weaknesses of each model, aiding in the informed choice of the most suitable regression model for future predictions.

VI. CONCLUSION AND FUTURE SCOPE

After a thorough evaluation of all the regression models, it is evident that the XGBRegressor model stands out as the best performer for predicting the purchase amount from our dataset. With tuning parameters set at n_estimators=500, max_depth=10, and learning_rate=0.05, the XGBRegressor achieved an impressive r2_score of 0.7492 and a low RMSE value of 2518.2849. The high R2 score score predicts that approximately 74.92% of the variance in the purchase amount can be explained by the model's features, while the low RMSE demonstrates its superior predictive accuracy. These results suggest that the XGBRegressor model is capable of providing precise and reliable estimates for purchase amounts, making it a valuable tool for real-world applications and decisionmaking processes related to purchase forecasting and optimization.

In conclusion, the XGBRegressor model, with its combination of high R2 score and low RMSE, outperforms other regression models and proves to be the optimal choice for predicting purchase amounts in our dataset. Its robust performance and accurate predictions make it a powerful tool for businesses and researchers seeking to gain valuable insights and make informed decisions based on purchase data analysis.

In the future, the application of deep learning techniques for predicting purchase amounts holds promising potential. With

ISSN: 2278-4632
Vol-13, Issue-11, November 2023

advancements in deep learning research and algorithms, these techniques can offer improved prediction accuracy and the ability to handle unstructured data, such as text descriptions or images of products. As businesses collect more data, there are opportunities for enriching the dataset with additional relevant information, like customer demographics and transaction history. Utilizing these advancements and data enrichment can result in a deeper understanding of customer behavior, empowering businesses to customize their strategies and enhance decision-making processes for purchase forecasting.

REFERENCES

[1] C. M. Wu, P. Patil and S. Gunaseelan, "Comparison of Different
Machine Learning Algorithms for Multiple Regression on Black Friday
Sales Data," 2018 IEEE 9th International Conference on Software
Engineering and Service Science (ICSESS), 2018, pp. 16-20, doi:
10.1109/ICSESS.2018.8663760.

[2] Odegua, Rising. (2020). Applied Machine Learning for Supermarket
Sales Prediction.

[3]Beheshti-Kashi, S., Karimi, H.R., Thoben, K.D., Lutjen, M., Teucke, M.: " A survey on retail sales forecasting and prediction in fashion markets, " Systems Science & Control Engineering 3(1), 154, 161(2015)

[4]Smith, Oliver, and Thomas Raymen. " Shopping with violence: Black Friday sales in the British context. " Journal of Consumer Culture 17.3 (2017): 677-694.

[6]Briana Milavec, "An Analys is of Consumer Misbehavior On Black Friday",2012

[7] Swilley, Esther & Goldsmith, Ronald, "Black Friday and Cyber Monday: Understanding consumerintentionson two major shopping days", Journal of Retailing and Consumer Services,2013

[8] Kvalheim, Olav Martin, et al. " Determination of optimum number of components in partial least squares regression from distributions of the rootĞmeanĞsquared error obtained by Monte Carlo resampling. " Journal of Chemometrics 32.4 (2018): e2993.

[9] Sheridan, Robert P., et al. " Extreme gradient boosting as a method for quantitative structure–activity relationships. " Journal of chemical information and modeling 56.12 (2016): 2353-2360

[10] Analytics Vidhya. (n.d.). Black Friday. Retrieved from https://datahack.analyticsvidhya.com/contest/black-friday/

[11] SamruddhiK ,Dr Ashok Kumar R, "Applying Different Machine Learning Techniques for Sales Forecasting", ISSN:1001-1749, Volume- 16, Issue-5, May2020

[12]Potturi, Keerthan, "Black Friday A study of consumer behavior and sales predictions" (2021). Creative Components. 784.

[13]Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. Mach Learn 63, 3–42 (2006). https://doi.org/10.1007/s10994-006-6226-1

[14]Figueiredo, Dalson & Júnior, Silva, & Rocha, Enivaldo. (2011). What is R2 all about?. Leviathan-Cadernos de Pesquisa Polútica. 3. 60-68. 10.11606/issn.2237-4485.lev.2011.132282.

[15] Chai, Tianfeng & Draxler, R.R.. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?– Arguments against avoiding RMSE in the literature. Geoscientific Model Development. 7. 1247-1250. 10.5194/gmd-7-1247-2014.