# Predict customer purchases on Black Friday sales prediction using Machine Learning

Dr.K.Lakshminadh, Ph.D
Computer Science And Engineering
Narasaraopeta Engineering College
Narasaraopet, India
drlakshimnadh@gmail.com

Venkata Gopi Krishna Bobbepalli
Computer Science And Engineering
Narasaraopeta Engineering College
Narasaraopet, India
bvgk.2003@gmail.com

Imran Shaik
Computer Science And Engineering
Narasaraopeta Engineering College
Narasaraopet, India
skimran1704@gmail.com

Prasanth pallepogu
Computer Science And Engineering
Narasaraopeta Engineering College
Narasaraopet, India
prasanthpallepogu81@gmail.com

*Abstract*— **Machine learning has emerged as a transformative research field in recent years, offering powerful tools for decision-making. By learning from past data, machine learning models can make informed predictions about future outcomes, enhancing their accuracy over time. This study aims to design a predictive model specifically tailored for predicting Black Friday sales. The model's development involves testing with various regression techniques to ensure its effectiveness. Ultimately, a XGB regression-based approach is adopted for its predictive capabilities. Implementing such a model could significantly benefit sales administration, particularly during high-traffic events like Black Friday. The algorithm we proposed was XG Boost regressor based approach used to predict Black Fridaysales.**

**Keywords—Black Friday, Sales Prediction, Machine Learning, Regression, Decision Tree, Gradient Boosting, RMSE and R2 Score**

## I. INTRODUCTION

Black Friday, traditionally the day following thanksgiving, has evolved into a major shopping event characterized by deep discounts and high consumer spending. Originally dubbed "Black Friday" due to the chaos it caused, this day now represents a critical opportunity for retailers to boost sales and attract customers with attractive deals [1]. As such, accurately predicting sales on Black Friday is of utmost importance for retailers seeking to maximize their profits and optimize their operations.

This Paper aims to develop a predictive model for forecasting sales on Black Friday using machine learning techniques [2]. By analyzing historical sales data, demographic information, and other relevant factors, the model seeks to provide insights into consumer behavior and preferences . These insights can help retailers make informed decisions about inventory management, marketing strategies, and product offerings.

The significance of Black Friday extends beyond individual retailers, impacting the economy at large. Black Friday sales are often seen as an indicator of consumer sentiment and economic health [1]. Analysts and policymakers closely monitor Black Friday sales data to gauge consumer confidence and assess the overall economic outlook [4].

In this paper present a detailed analysis of our approach to predicting sales on Black Friday. We discuss the methodology used to develop the predictive model, including data preprocessing, feature selection, and model training. We also present the results of our model and discuss its implications for retailers and the economy as a whole.

## II. LITERATURE SURVEY

Ramachandra et al. [2] used a regression model to determine the association between input variables and predictions. Initially, they identified the dependent association between variables in order to forecast client purchases. The regression model was used to analyse the time and cost characteristics of products, with an 86% success rate in discovering connections between input and target variables.

Chen et al. [3] used neural networks to anticipate client demand for specific products. Their hybrid technique discovered similar qualities among customer-favored products. The approach used linear relationships to find hidden connections between customers and products. Machine learning techniques were used to classify products based on the relevance of attributes. The approach also calculated the divergence between the input and bias variables.

Javed et al. [5] used a variety of data mining approaches to forecast sales of different products. These strategies were useful in revealing hidden correlations between the variables required for prediction. Sales forecasting for a variety of products was determined by analysing past performance and reliability. The study also revealed relationships between numerous product features, resulting in an amazing 89% accuracy in sales forecast.

Gurnani et al. [6] used neural networks to anticipate sales by taking past values as input and anticipating future values, basically conducting time series forecasting. Time series forecasting has long been the accepted method for analysing the behaviour of any process. These estimates are based on previous demand patterns and other probable future considerations. Continued refinement can improve forecast accuracy.

Cheriyan et al. [9] approach emphasises the need of employing advanced tools for accurate sales forecasts. They provide a robust strategy that can be utilised across diverse organisations, ensuring reliable predictions that can lead to informed decision-making.

## III. PROPOSED SYSTEM FOR BLACK FRIDAY SALES PREDICTION

Our research for Black Friday Sales prediction is designed with a structured approach comprising several phases and steps, each aimed at enhancing accuracy, efficiency, interpretability, robustness, and fairness. The key criteria for our model include.

- Dataset Analysis
- Visualization of Data
- Preprocessing Techniques
- Model Development

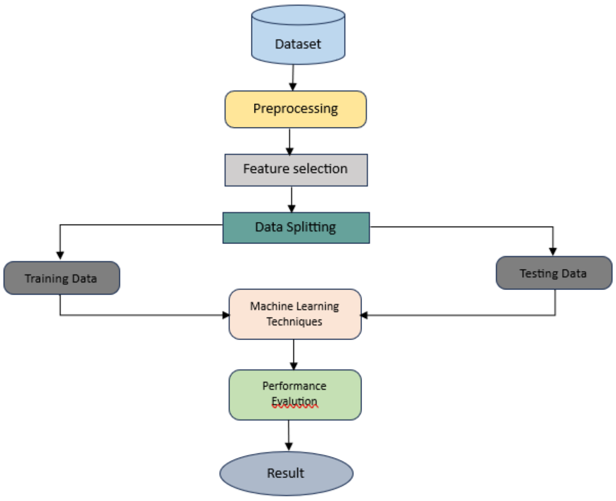
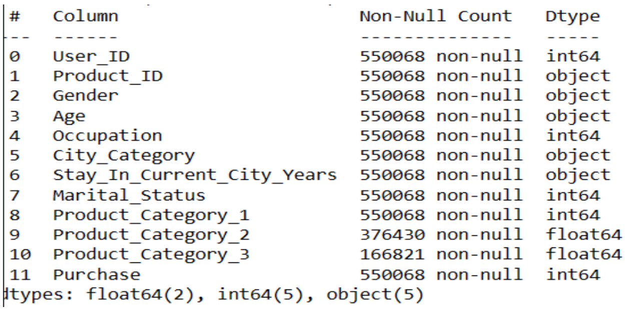


Fig.1 Steps involved in a Model

### A. Dataset Analysis

The dataset used for this study is the Black Friday Sales Dataset from Kaggle web. This dataset is widely used for training machine learning models and forecasting the number of people making purchases during Black Friday sales. Retailers can utilize the purchase prediction to tailor offers for customers' favorite products. The set consist of 12 attributes and 550068 rows. The information as follows.



```
#   Column                       Non-Null Count   Dtype
--  ------                       --------------   -----
0   User_ID                      550068 non-null  int64
1   Product_ID                   550068 non-null  object
2   Gender                       550068 non-null  object
3   Age                          550068 non-null  object
4   Occupation                   550068 non-null  int64
5   City_Category                550068 non-null  object
6   Stay_In_Current_City_Years   550068 non-null  object
7   Marital_Status               550068 non-null  int64
8   Product_Category_1           550068 non-null  int64
9   Product_Category_2           376430 non-null  float64
10  Product_Category_3           166821 non-null  float64
11  Purchase                     550068 non-null  int64
dtypes: float64(2), int64(5), object(5)
```

Fig.2 Information about columns in training data

Analysing the dataset to detect null and unique values is a frequent approach for better understanding the data. Null values indicate missing or incomplete data that may need to be addressed before analysis. Unique values give information about the diversity and distribution of data, which is useful for feature engineering and model building.

The examination of the dataset reveals that the column "Product_Category_2" includes 31.57% null values, while the column "Product_Category_3" contains 69.67% null values. Columns with a high percentage of null values are frequently removed from the dataset to ease subsequent analysis and modelling.

### B. Visualiztion of Data

#### a) Univariant Analysis

Univariate analysis is a fundamental data analysis technique that examines only one variable at a time. It entails gathering, summarising, and portraying data with graphical and numerical tools like bar graphs, pie charts, and histograms [14]. Unlike other types of data analysis, which investigate linkages or causes, univariate analysis aims to find patterns and comprehend the distribution of a single variable.

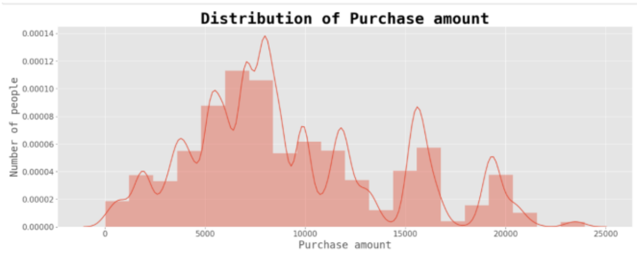


Fig 3. Distribution Of Purchase Amount

Figure 3 : The distribution plot depicts the distribution of purchase amounts vs. number of persons. This visualisation sheds light on the distribution of purchase quantities among various categories of people, highlighting any patterns or trends in purchasing behaviour.

#### b) Bivariant Analysis

Bivariate analysis investigates the relationship between two variables or attributes to evaluate whether or not there is a connection and how strong that correlation is. This study examines the relationship between the two variables and measures the strength of the correlation in order to uncover any differences and probable causes for them. Bivariate analysis allows researchers to understand the link between two variables and how changes in one may impact the other.

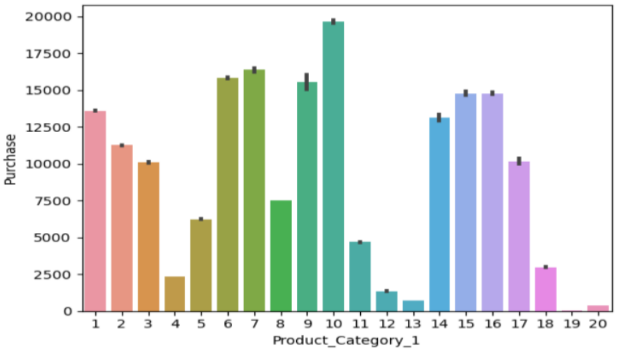


Fig.4. Purchase amount across product category 1

Figure 4 :The bar plot shows the link between product category 1 and the average purchase price during Black Friday discounts. This visualisation provides insights on the purchasing behaviour associated with various product categories, assisting retailers in determining which categories are more popular or result in higher purchases.

### c) *Multivariant Analysis*

Multivariate analysis is the simultaneous evaluation of numerous variables to understand their linkages and combined impact on the outcomes of interest. This form of analysis is critical for acquiring a thorough grasp of complicated datasets and detecting hidden patterns or trends that may not be visible in univariate or bivariate analyses

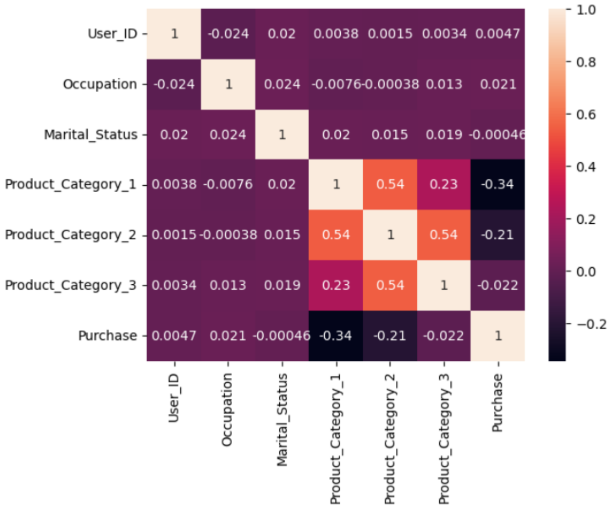


Fig.5. Heatmap to Display Correlation Between Variables

Figure 5 : A heatmap is a graphical representation of data in which the individual values of a matrix are represented as colours. The colours in the heatmap represent the magnitude of the numbers, allowing you to easily see patterns and relationships in the data. In multivariate analysis, a heatmap can be used to depict the correlation matrix between features in a dataset.

## C. *Data Processing*

Data processing was a vital step in getting the data ready for model training. This procedure entailed making appropriate changes to the data based on insights gained via exploratory data analysis [14].

- Dropping redundant characteristics, such as User_ID and Product_ID, is a typical data preparation technique. These features are better suited to database maintenance and add little to the model's predictive power. simplifies the dataset and improves model performance by minimising noise and unnecessary data.
- To fix null values, missing values are replaced with acceptable substitutes based on observations from exploratory data analysis. This prepares the dataset for analysis and model training.
- To convert 'Stay in current years' to a numeric data type, change the column's format from categorical or text to numerical.
- Feature Encoding: is the process of converting categorical data into a numerical representation that is useful for training machine learning models. This phase guarantees that all features in the dataset are in a format that the model can comprehend and analyse efficiently.
- Outlier Removal: To ensure reliable predictions, we must first remove outliers from our dataset. Outliers, or data points that are markedly different from others, might skew our results. We use statistical approaches such as the z-score to find outliers, which we then eliminate or adjust to increase the predictive model's dependability.
- Feature scaling: standardises the range of independent features, ensuring that each contributes equally to the analysis
- Feature Selection: Extra Tree Regressor detects the most relevant features, simplifying the model and boosting performance.
- Separating the data into Train and test

## D. *Model Development*

In our Black Friday sales prediction project, we've used machine learning models like Linear Regression, KNN, Decision Tree, Random Forest, and XG Boost. These models help us predict sales based on different features.

These models enable us to predict sales amounts based on different features in the dataset. While regression is ideal for forecasting continuous variables like sales amounts, classification is better suited for discrete variables.

### a) *Linear Regression*

Linear regression is a basic machine learning approach for examining the relationship between independent and dependent variables. In our project, it aids in predicting sales based on historical data. After training on historical data, the model can be used to anticipate future sales. Linear regression is commonly used in weather forecasting, stock market projections, and a variety of other industries because of its ability to predict continuous variables accurately [12].

### b) *KNN Regression*

K-Nearest Neighbours (KNN) is a supervised machine learning model used in classification and regression. It works on the premise of determining the similarity between the attributes of test data and the existing dataset. KNN is classified as a lazy learner and non-parametric algorithm because it does not train on the training dataset when it arrives. Instead, it maintains the dataset and only trains when test data has to be classified [13].

*c) Decision Tree*

A decision tree is a supervised machine learning technique used for classification and regression, with a primary emphasis on classification. Its structure mimics a tree, with central nodes representing features, branches representing decision rules, and leaf nodes representing outcomes. The decision to go to a branch or leaf is based on dataset characteristics [12]. The decision-making process in a decision tree is binary, resulting in either a "yes" or "no" answer, which determines subsequent tree splitting.

*d) Random Forest*

Random Forest is a popular supervised machine learning technique that may be used for classification as well as regression [5]. It works on the basis of ensemble learning, which entails integrating many classifiers to solve difficult classification tasks. This ensemble approach improves the model's efficiency and accuracy.The integrated decision-making process of several trees yields a more robust and accurate model for forecasting results.

*e) XG Boost*

XG Boost is a machine learning technique that combines the concepts of bagging, aggregation, and bootstrapping. It is well-known for its efficiency and performance when working with huge datasets, and it is frequently used for classification and regression applications.

*f) Ridge Regression*

Ridge regression is a linear regression technique that is especially beneficial for dealing with multicollinearity, which occurs when the independent variables in a regression model are highly linked. It works by including a penalty component into the conventional linear regression equation, which helps to minimise the coefficients of the linked variables.

*g) Polynomial Regression*

Polynomial regression analyses the relationship between an independent variable (x) and a dependent variable (y) using an nth degree polynomial. It is utilised when the relationship between variables is non-linear and a curve can better explain the situation.

## IV. Results

The model's performance is evaluated using RMSE (root mean square error) and R2 scores.

*a) Root Mean Square Error*

The Root Mean Square Error (RMSE) is a metric that measures the difference between a model's projected values and its actual values. It is calculated by taking the square root of the average of the squared discrepancies between expected and actual values. The following is the RMSE formula shown in Equation (1):

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(Ypi - Yi\ )\hat{}2}{n}} \qquad ....(1)$$

*b) $R^2$ Score*

The coefficient of determination, or R^2 score, assesses a regression model's goodness of fit. It runs between 0% and 100%, with higher numbers suggesting a better fit. A score of 100% indicates that the model fully explains the variance in the dependent variable, whereas a score of 0% shows that the model does not explain any variance. The formula for R2 score is given by equation (2):

$$R^2 = 1 - \frac{RSS}{TSS} \qquad ......(2)$$

Table.1. Tabular representation of the regression model's performance.

| MODEL | RMSE | $R^2$ Score |
|---|---|---|
| Linear Regression | 4088.26 | 0.2350 |
| KNN Regression | 3183.52 | 0.5361 |
| Decision Tree | 3098.21 | 0.5604 |
| Random Forest | 3023.99 | 0.5817 |
| XG Boost | 2918.94 | 0.6100 |
| ExtraTree Regressor | 3023.27 | 0.5704 |
| Ridge Regression | 4088.25 | 0.2350 |
| Lasso Regression | 4088.26 | 0.2350 |
| Polynomial Regression | 3776.7 | 0.3471 |

In contrast to the existing paper, our research paper focuses on refining the dataset by removing outliers, a step that leads to a more robust model. By implementing this preprocessing step, our model achieves a lower Root Mean Square Error (RMSE) value compared to the baseline paper. This improvement signifies a higher level of accuracy in predicting Black Friday sales, demonstrating the effectiveness of our approach in enhancing model performance.

## V. Conclusion

Utilizing a machine learning algorithm to predict a customer's "Black Friay" spending is demonstrated in this study. To find interesting patterns in the dataset, it has been decided to use exploratory data analysis. According to this study, the user's gender, age, and occupation all play a role when they try to predict which product a client is more likely to buy based on their gender, age, and occupation. Tests show that our system, when contrasted with methods like choice trees and edge relapse, can yield more precise expectations.

In this research, we attempted to either construct a model employing various algorithms like linear regression, KNN regression, decision tree regression, random forest regression, and XGB regress, or to make the most accurate prediction. The hyper boundary tuned XGB gives us the best times esteem and r2 score for this issue. Machine learning can be used for a wide range of tasks.

## REFERENCES

[1]. Simpson, L., Taylor, L., O'Rourke, K., & Shaw, K. "An analysis of consumer behavior on Black Friday". American International Journal of Contemporary Research.2011.

[2]. H. V. Ramachandra, G. Balaraju, A. Rajashekar and H. Patil, "Machine Learning Application for Black Friday Sales Prediction Framework," 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), 2021, pp. 57-61.

[3]. J. Chen, W. Koju, S. Xu and Z. Liu, "Sales Forecasting Using Deep Neural Network And SHAP techniques," 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), 2021, pp. 135-138.

[4]. Aaditi Narkhede, Mitali Awari, Suvarna Gawali, Amrapal Mhaisgawali "Big Mart Sales Prediction Using Machine Learning Techniques" International Journal of Scientific Research and Engineering Development Vol3-Issue4 | 693-697,2020.

[5]. M Javed Awan, MS Mohd Rahim," A Big Data Approach to Black Friday Sale", Intelligent Automation and Soft Computing, Vol. 27, no.3,pp.785–797, 2021.

[6]. M. Gurnani, Y. Korke, P. Shah, S. Udmale, V. Sambhe and S. Bhirud, "Forecasting of sales by using fusion of machine learning techniques," 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI), 2017, pp. 93-101.

[7]. Pasumpon. "Review Of Machine Learning Techniques for Voluminous Information Management." Journal of Soft Computing Paradigm 1, no.2: 103-112.

[8]. R. P and S. M, "Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 1416-1421.

[9]. S. Cheriyan, S. Ibrahim, S. Mohanan and S. Treesa, "Intelligent Sales Prediction Using Machine Learning Techniques," 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE), 2018, pp. 53-58

[10]. S. H. Lye and P. L. Teh, "Customer Intent Prediction using Sentiment Analysis Techniques," 2021 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2021, pp. 185-190.

[11]. W. Liao, G. Ye, Y. Yin, W. Yan, Y. Ma and D. Zuo, "Auto Parts Sales Prediction based on Machine Learning for Small Data and a Long Replacement Cycle," 2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA), 2020, pp. 1-5.

[12]. Singh, A., Thakur, N., & Sharma, "A review of supervised machine learning algorithms",3rd International Conference on Computing for Sustainable Global Development pp. 1310-1315, 2016.

[13]. Kohli, S., Godwin, G. T., & Urolagin, S., "Sales prediction using linear and KNN regression". In Advances in machine learning and computational intelligence (pp. 321-329). Springer, Singapore.2021

[14]. Somula Ramasubbareddy, AdityaSai, AdityaSaiKharisma, Govinda Kharisma Govinda, E. Swetha, "Sales Analysis on Back Friday Using Machine Learning Techniques", DOI: 10.1007/978-981-15-5400-1_32, In book: Intelligent System Design, January 2021.