International Journal of Scientific Research and Engineering Development-– Volume 8 Issue 2, Mar-Apr 2025
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Prediction of Black Friday Sales using Machine Learning Algorithms

S. Varsha[1*], M. Sharath[1*], M. Shashank[1*], S. Prashanth[1*], P.Raghavendra Prasad[2]
[*1] Student, Department of IT, MREC(A), Maisammaguda, Hyderabad-500100
2.Asst.Professor, Deartment of IT Department of IT, MREC(A), Maisammaguda, Hyderabad-500100

Abstract: Black Friday marks the beginning of the Christmas shopping festival across the US. On Black Friday big shopping giants like Amazon, Flipkart, etc. lure customers by offering discounts and deals on different product categories. The product categories range from electronic items, Clothing, kitchen appliances, Décor. Research has been carried out to predict sales by various researchers. The analysis of this data serves as a basis to provide discounts on various product items. With the purpose of analyzing and predicting the sales, we have used three models. The dataset Black Friday Sales Dataset available on Kaggle has been used for analysis and prediction purposes. The models used for prediction are linear regression, lasso regression, ridge regression, Decision Tree Regressor, and Random Forest Regressor. Mean Squared Error (MSE) is used as a performance evaluation measure.

Keywords: *Regression, Linear Regression, Ridge Regression, Lasso Regression, Decision Tree Regressor, Random Forest Regressor, Mean Squared Error.*

I. INTRODUCTION

Black Friday has become one of the most significant shopping events worldwide, marking the beginning of the holiday shopping season. Originating in the United States, it takes place on the Friday following Thanksgiving and has evolved into a global phenomenon. Retailers offer substantial discounts on various products, attracting millions of customers eager to take advantage of the limited-time deals. Over the years, Black Friday has transformed from a single-day event into an extended shopping period, often lasting several days, including Cyber Monday, which focuses on online deals. The rise of e-commerce and digital platforms has further amplified the impact of Black Friday, allowing consumers to shop conveniently from their homes while businesses expand their reach beyond physical stores.

The evolution of Black Friday sales can be attributed to changing consumer behavior, technological advancements, and strategic marketing efforts by retailers. Initially, the day was associated with physical stores opening early and long queues of shoppers eager to secure the best deals. However, with the increasing popularity of online shopping, retailers have adapted by offering exclusive discounts on their websites, mobile apps, and digital platforms. This shift has led to significant competition among businesses, compelling them to leverage data-driven strategies to enhance customer engagement and maximize sales. Predictive analytics, artificial intelligence, and machine learning have become essential tools in forecasting demand, optimizing inventory, and personalizing promotions for different customer segments. The role of machine learning in Black Friday sales prediction has gained prominence due to its ability to process vast amounts of data and uncover hidden patterns. Traditional forecasting methods, such as time series analysis and regression models, have limitations in handling the complexity of modern retail environments. In contrast, machine learning algorithms, including, Linear Regression, and Ridge Regression, offer superior accuracy and scalability. These models can analyze historical sales data, customer demographics, and product attributes to predict future demand with higher precision.

II. LITERATURE SURVEY

A great deal of work having been gotten really intended to date the territory of deals foreseeing. A concise audit of the important work in the field of big mart deals is depicted in this part. Numerous other Measurable methodologies, for example, with regression, (ARIMA) Auto-Regressive Integrated Moving Average, (ARMA) Auto-Regressive Moving Average, have been utilized to develop a few deals forecast standards. Be that as it may, deals anticipating is a refined issue and is influenced by both outer and inside factors, and there are two significant detriments to the measurable technique as set out in A. S. Weigend et A mixture occasional quantum relapse approach and (ARIMA) Auto-Regressive Integrated Moving Average way to deal with every day food deals anticipating were recommend by N. S. Arunraj and furthermore found that the exhibition of the individual model was moderately lower than that of the crossover model.

C.M.Wu et alhave[1] proposed a prediction model to analyze the customer's past spending and predict the future spending of the customer. The

dataset referred is Black Friday Sales Dataset from analyticsvidhya. They have machine learning models such as Linear Regression, MLK classifier, Deep learning model using Keras, Decision Tree, and Decision Tree with bagging. The performance evaluation measure Root Mean Squared Error (RMSE) is used to evaluate the models used. Simple problems like regression can be solved by the use of simple models like linear regression instead of complex neural network models. Odegua, [2] Rising have proposed a sales forecasting model. The machine learning models used for implementation are Linear regression and Lasso regression. The dataset used for the experimentation is provided by Data Science Nigeria, as a part of competitions based on Machine Learning. The performance evaluation measures used are Mean Absolute Error (MAE). Random Forest outperformed the other algorithms with a MAE rate of 0.409178.

Singh, K et al[3] have analyzed and visually represented the sales data provided in the complex dataset from which we ample clarity about how it works, which helps the investors and owners of an organization to analyze and visualize the sales data, which will outcome in the form of a proper decision and generate revenue. The data visualization is based on different parameters and dimensions. The result of which will enable the end-user to make better decisions, ability to predict future sales, increase the production dependencies on the demand, and also regional sales can be calculated. S.Yadav et al [4]have analyzed and compared the performance of K-Fold cross-validation and hold-out validation method. The result of the experimentations where k-fold cross-validation gives more accurate results. The accuracy results of K - Fold cross-validation were around 0.1 - 3% more accurate as compared to hold-out validation for the same set of algorithms.

Aaditi Narkhede et.al[5] has applied machine learning algorithm in tracking sales at places like shopping center big mart to anticipate the demand of customers and handle the management of inventory accordingly the methods presented here are an effective method for data shaping and decision making. New ways that can better identify consumer needs and calculate marketing plans which will improve sales. M.Sahaya Vennila et al[6] have analyzed, preprocessed, and applied machine learning techniques to predict sales. The dataset used for the analysis and experimentation purpose is Black Friday Sales Dataset from Kaggle. The dataset is preprocessed. K - Fold method is used for the purpose of splitting the dataset into training and testing datasets. The prediction model is implemented using Linear Regression Decision Tree, Random Forest and Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are used as the accuracy evaluation measures. As a result of experimentation, the Random Forest performed significantly with an accuracy of 77%, with an RMSE value of 2730 and MAE value.

III. METHODOLOGY

The Black Friday Sales Prediction project is divided into several core modules that collectively contribute to building a robust and accurate prediction system. Each module plays a critical role in ensuring that the system can process raw data efficiently and generate meaningful and actionable insights. Fig 1 shows the demonstration of proposed system.

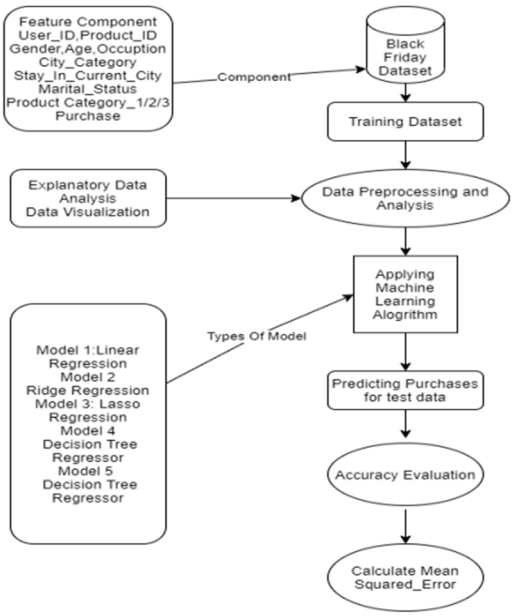

Figure 1: Demonstration of Proposed System

3.1. FEATURE EXTRACTION

Chi-Square test identifies the top 4 features highly correlated with the target variable then fed into the classification models. Features are reduced to 4 key features after feature extraction and each one is denoted by

$X = [[f1, f2, \ldots\ldots, fn], Y \in \{0,1\}]$----------(1)

Where X and Y represents feature vector and target variable and Y=0, poor performance and y=1, successful performance.

$X' = [f1', f2', f3', f4']$----------(2)

Chi-Square, $\chi 2 = \sum_{k=1}^{K} (O_k\text{-}E_k)^2/E^k$----------(3)
Where, O_k and E_k are observed frequency and expected frequency.
The extracted features help the models to differentiate between students likely to perform well and those at risk of dropping out.

3.2. LINEAR REGRESSION

Linear Regression is one of the supervised machine learning algorithms. A regression problem can be stated as a case when the output variable is continuous [10]. Linear regression predicts a dependent variable (y) based on a given independent variable (x). The model depicts a linear relation among the variables. Function for linear regression is: Y = Θ1 + Θ2 .x Here, the input variable is x, the output value is y and Θ1 represents intercept and Θ2 represents the coefficient of x. This algorithm aims to calculate and find the best fit line to target variable and independent variable.

3.3. RANDOM FOREST

In training stage several decision trees are constructed then merge their results to get better accuracy and then reducing over-fitting. In the implementation, parameters such as n_estimators=1 and max_depth=0.9 restrict the model's ensemble capability, causing it to behave similarly to a single tree. Prediction Aggregation for classification

$y\hat{} = Mode\{T1(x), T2(x), \dots, TK(x)\}$-----(4)

Where $T_k(x)$ is the prediction from the k-th tree.

3.4. DECISION TREE

Based on feature threshold, decision tree split the dataset recursively. The splits are used to maximize class separation using Gini Impurity or Information Gain. The model trained using parameters such as n_estimators=10, learning_rate=0.2, and a max_depth=2.

a. Gini Impurity

$G = \sum_{i=0}^{C} p^2{}_i$----------(5)

Where, P_i is Proportion of samples belonging to class iii and C is Total number of classes.

b. Information Gain

$Ig(S,A) = H(S) - \sum v \in A \ | Sv |/| S | \ H(Sv)$-(6)

Where, H(S) is Entropy of set S and $H(S_v)$ is Entropy of subset S_v

c. Entropy

$H(S) = -\sum_{i=1}^{C} p_i \log_2 p_i$----------(7)

This process repeats recursively, creating branches until all data points are classified or a stopping criterion is met (e.g., max depth).

- Input Features: $X' = [f_1', f_2', f_3', f_4']$
- Output: Classification as good performer (y=1) or poor performer (y=0).

Initialize the model with a constant value:

$F_0(x) = \arg\{\min \{c \sum_{i=0}^{N} L(y_i, c)\}\}$----------(8)

Additive model:

$Fm(x) = Fm - 1\ (x) + \eta \cdot hm(x)$---------(9)

Where, $F_m(x)$: Model at iteration mmm, η: Learning rate and $h_m(x)$: A weak learner Gradient descent minimizes the loss function.

$g(i) = -[\partial L(yi, F(xi))/\partial F(xi)]$---------(10)

The weak learner $h_m(x)$ g_i.

Loss function: Cross-entropy loss for classification:

$L(y,y\hat{}) = -\sum_{i=1}^{N} y_i \log(y\hat{}_i) + (1-y_i)\log(1-y\hat{}_i)$(11)

3.5. RIDGE REGRESSION

Multiple regression data can be analyzed using Ridge Regression. Least Square estimates are unbiased when multicollinearity occurs. Based on the degree of bias it reduces the standard errors that is added to the regression estimates.

$\beta = (X^T X + \lambda * I) X^T y$----------(12)

3.6. LASSO REGRESSION

Lasso Regression provides both variable selection and regularization. It makes use of soft thresholding. Only a subset of the covariates provided is select for use in the final model in the case of Lasso regression.

$N^{-1} \sum_{i=1}^{N} f(xi, y1, \boldsymbol{\alpha}, \beta)$ ----------(13)

3.7 EVALUATION METRICS

The performance of the classification model has been evaluated using various evaluation metrics like accuracy, sensitivity, specificity, precision, recall, f1-measure, MSE, RMSE, MAE and ROC curve (AUC).

Table 1. The performance metrics used for classification and regression

Metric	Formula
Precision (P)	$\frac{TP}{TP + FP}$
Recall (R)	$\frac{TP}{TP + FN}$
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
F1-score	$2 * \frac{R * P}{R + P}$
MSE	$\frac{1}{m}\sum_{i=1}^{m}(y - y\hat{}i)^2$

RMSE	$\frac{1}{m}\sum_{i=1}^{m}\sqrt{(y-y\hat{}i)2}$
MAE	$\frac{1}{m}\sum_{i=1}^{m}\lvert(y-y\hat{}i)^2\rvert$

IV. RESULT

With this particular Black Friday sale analysis, we are more interested in figuring out how much will a customer spend based on certain attributes such as their Age group, City Category, etc.

	A	B	C	D	E	F	G	H	I	J	K	L
1	User_ID	Product_ID	Gender	Age	Occupatio	City_Categ	Stay_In_Cu	Marital_St	Product_C	Product_C	Product_C	Purchase
2	1000001	P0006904	F	0-17	10	A	2	0	3			8370
3	1000001	P0024894	F	0-17	10	A	2	0	1	6	14	15200
4	1000001	P0008784	F	0-17	10	A	2	0	12			1422
5	1000001	P0008544	F	0-17	10	A	2	0	12	14		1057
6	1000002	P0028544	M	55+	16	C	4+	0	8			7969
7	1000003	P0019354	M	26-35	15	A	3	0	1	2		15227
8	1000004	P0018494	M	46-50	7	B	2	1	1	8	17	19215
9	1000004	P0034614	M	46-50	7	B	2	1	1	15		15854
10	1000004	P0097242	M	46-50	7	B	2	1	1	16		15686
11	1000005	P0027494	M	26-35	20	A	1	1	8			7871
12	1000005	P0025124	M	26-35	20	A	1	1	5	11		5254
13	1000005	P0001454	M	26-35	20	A	1	1	8			3957
14	1000005	P0003134	M	26-35	20	A	1	1	8			6073
15	1000005	P0014504	M	26-35	20	A	1	1	1	2	5	15665
16	1000006	P0023134	F	51-55	9	A	1	0	5	8	14	5378
17	1000006	P0019024	F	51-55	9	A	1	0	4	5		2079
18	1000006	P0096642	F	51-55	9	A	1	0	2	3	4	13055
19	1000006	P0005844	F	51-55	9	A	1	0	5	14		8851
20	1000007	P0003684	M	36-45	1	B	1	1	1	14	16	11788
21	1000008	P0024954	M	26-35	12	C	4+	1	1	5	15	19614
22	1000008	P0022044	M	26-35	12	C	4+	1	5	14		8584
23	1000008	P0015644	M	26-35	12	C	4+	1	8			9872
24	1000008	P0021374	M	26-35	12	C	4+	1	8			9743
25	1000008	P0021444	M	26-35	12	C	4+	1	8			5982

Figure 2: Sample Dataset

Model	RootMeanSquareError	Accuracy of the model
Linear Regression	4703.5238	0.1269
Lasso Regression	4703.5994	0.1269
Ridge Regression	4703.4937	0.1269
Decision Tree Regressor	3083.3668	0.7083
Random Forest Regressor	3015.6041	0.7058

Figure 3: Results of i) Linear regression, ii) Random Forest, iii) Lasso regression iv) Decision tree and v) Ridge regression

Table .2. Evaluation Results

Algorithm	AUC	CA	Precision	Recall	F1-score
Decision Tree	0.94	70.8	0.85	0.88	0.86
Lasso Regression	0.98	12.6	0.80	0.82	0.81
Random Forest	0.40	70.5	0.87	0.89	0.88
Ridge Regression	0.97	12.6	0.81	0.83	0.82
Linear	0.90	12.6	0.79	0.80	0.79
Regression					

The Decision Tree algorithm achieved the highest accuracy of 70.8%, and a weighted-average F1-score of 91%. The Random Forest also performed well, achieving an accuracy of 70.5% the weighted-average F1-score is 88%. Lasso Regression with an accuracy of only 12.6%. The Ridge Regression algorithm achieved an accuracy of 12.6%. Linear Regression with an accuracy of 12.6%.

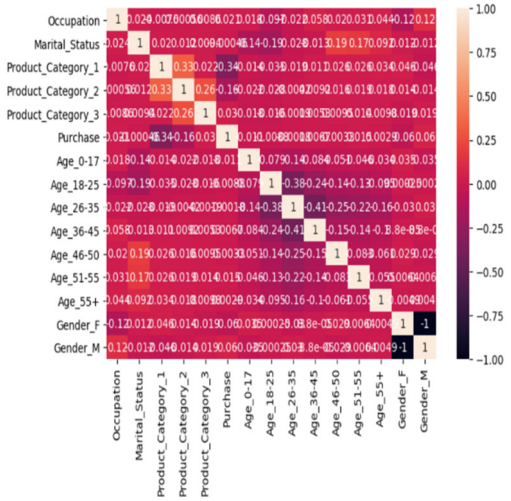

Figure .4. Correlation Matrix

V. CONCLUSION

Ample research is carried out on the analysis and prediction of sales using various techniques. There are many methods proposed to do so by various researchers. In this section, we will summarize a few of the machine learning approaches. The Black Friday Sales dataset is used for training various machine learning models and also for predicting the purchase number of customers on black Friday sales. The purchase prediction made will provide an insight to retailers to analyze and personalize offers for more customer's preferred products. With traditional methods not being of much help to business growth in terms of revenue, the use of Machine learning approaches proves to be an important point for the shaping of the business plan taking into consideration the shopping pattern of consumers. Projection of sales concerning several factors including the sale of last year helps businesses take on suitable strategies for increasing the sales of goods that are in demand. Thus, the dataset is used for the experimentation, Black Friday Sales Dataset from Kaggle [9]. The models used are Linear Regression, Lasso Regression, Ridge

Regression, Decision Tree Regressor, and Random Forest Regressor. The evaluation measure used is Mean Squared Error (MSE). Based on Table II Random Forest Regressor is best suitable for the prediction of sales based on a given dataset. Thus, the proposed model will predict the customer purchase on Black Friday and give the retailer insight into customer choice of products. This will result in a discount based on customer-centric choices thus increasing the profit to the retailer as well as the customer.

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. K.Singh and R. Wajgi, "Data analysis and visualization of sales data," 2016 World Conference on Futuristic Trends in Research and 10.1109/STARTUP.2016.7583967.
4. S.Yadav and S. Shukla, "Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification," 2016 IEEE 6th International Conference on Advanced Computing (IACC),
5. 2016, pp. 78-83, doi: 10.1109/IACC.2016.25.
6. Aaditi Narkhede, Mitali Awari, Suvarna Gawali, Prof.Amrapal Mhaisgawali " Black Friday Sales Prediction Using Machine Learning Techniques" International Journal of Scientific Research and Engineering Development (IJSRED) Vol3-Issue4 | 693-697.
7. M.Sahaya Vennila; Holy Cross College, Nagercoil. Affiliated to Manonmaniam Sundaranar University, Tirunelveli – 627 012.

