2023 International Conference on Artificial Intelligence and Smart Communication (AISC) | 979-8-3503-2230-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/AISC56616.2023.10084959

Black Friday Sales Prediction using Supervised Machine Learning

Shambhavi Patil
Department of CSE
Symbiosis Institute of Technology, Symbiosis, International (Deemed University) Pune, Maharashtra India
shambhavi.patil.btech2019@sitpune.edu.in

Om Nankar
Department of CSE
Symbiosis Institute of Technology, Symbiosis, International (Deemed University) Pune, Maharashtra India
om.nankar.btech2019@sitpune.edu.in

Renuka Agrawal
Department of CSE
Symbiosis Institute of Technology, Symbiosis, International (Deemed University) Pune, Maharashtra India
renuka.agrawal@sitpune.edu.in

Kanhaiya Sharma
Department of CSE
Symbiosis Institute of Technology, Symbiosis, International (Deemed University) Pune, Maharashtra India
Kanhaiya.sharma@sitpune.edu.in

Shashank Awasthi
Department of CSE
GL Bajaj Institute of Engineering and Management Greater Noida UP
shashankglbitm@gmail.com

Neha Jha
Department of CSE
GL Bajaj Institute of Engineering and Management, Greater Noida UP
neha.jha@glbitm.ac.in

***Abstract*—Machine learning has developed as one of the most influential research domains in the last decade with reasonable doubt. The emphasis on "learning" in machine learning enables computers to judge better. Based on previous experiences, machine learning models are able to judge better and predict future outcomes precisely. Recent advancements in machine learning have promoted efficient intelligence in business decisions and have further made systems capable of a wide range of applications from facial recognition to natural language processing. Prediction models are put to use in businesses in order to determine the most likely outcomes based on the data that is presented. Understanding and predicting the future purchase pattern of discrete customers against different products based on their demographic information of the features is the motive behind the work. The ideology discussed in this work helps to design and develop a predictor model which will be of much assistance to sales administration at the time of Black Friday. The developed model before implementation is tested with different classification techniques. Random Forest regression-based approach used to predict black Friday sales.**

Keywords—Machine Learning, Classification, Regression, Black Friday.

I. Introduction

The Friday just after Thanksgiving, otherwise coined as Friday, is one amongst the largest spending days of the year in the US. Very next day after Thanksgiving Day, being a holiday traditionally is referred as Black Friday. Since. It's a holiday so most of the office going people spend on shopping on this day. To use this opportunity for maximizing sales, shopping stores throughout US offers special deals and big discounts on a wide range of products [1]. Black Friday is also considered as the beginning of shopping season. National chain stores often give a small number of money-saving discounts on a range of goods while simultaneously giving equivalent deals online in an effort to entice shoppers into their physical locations.

The sales done on Black Friday are time and again considered as a confirmation for predicting the complete commercial assessment of the nation and also considered as a way for financial analysts to predict the extent of discretionary spending of the regular American citizen [2]. The finance analysts who believe that spending money or purchasing items strengthen nation's economy interpret poorer Black Friday sales figures as a forerunner of dawdling economic growth. Many people trust that the phrase "Black Friday" originated from the belief that companies run at a loss, or are "in the red," until the day following Thanksgiving, when great sales ultimately enable the companies either to increase sales and make good profit or put them "in the black" in case the sales graph didn't rise up [3]. A successful Black Friday is essential for many businesses, notably toy and game shops. The NRF estimates that from 2017 through 2021, the Christmas shopping season will account for around 19% of many stores' annual revenues. The NRF forecasts that Black Friday sales will be robust once again this year despite inflation. Their forecast that total holiday spending will rise by 6% to 8% over last year is an early clue that the day after Thanksgiving may have a substantial economic impact [4]. Rest of the paper is organized as follows: Section 2 discusses the architecture of the proposed system. Besides this, the section also focuses on discrete techniques of machine learning for regression and classification. Whereas section 3 pays attention to the data set used in its pre-processing et. Al, section 4 is all about exploratory data analytics, and section 5 is about data processing. Section 6 focuses on model implementation. Section 7 takes into consideration results obtained after model implementation and finally, section 8 is the conclusion and future work.

II. System Architecture

Machine learning is defined as the learning of different algorithms that can develop and improve their own performance by means of previous experience & old data [5]. A study of such types of models is called as Machine Learning, where models improve their performance by learning from their previous errors. The steps followed to complete this study make up the system architecture. The steps include collection of Data, Exploratory Data Analysis, data preprocessing, comparison of different techniques of classification and regression, evaluation of results and finally model implementation as shown in figure 1.

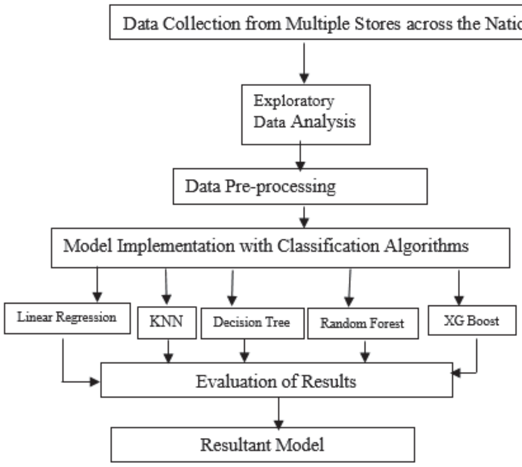

Data Collection from Multiple Stores across the Nation
Exploratory Data Analysis
Data Pre-processing
Model Implementation with Classification Algorithms
Linear Regression
KNN
Decision Tree
Random Forest
XG Boost
Evaluation of Results
Resultant Model

Fig. 1. System Architecture

The different models used for analysis and comparison include linear regression, KNN, Decision Tree, Random Forest and XG Boost (extreme gradient boost), the working of each of these models has been explained in this section.

Regression and Classification are basic machine learning techniques for Supervised Learning. Both the algorithms, after being trained on the training data set, are used for prediction of the test data set. These algorithms of machine learning work fine with labeled dataset. While classification is used to classify or segregate data sets in one or other category, regression techniques are effectively utilized for future sales prediction. Another difference between regression and classification techniques in machine learning is that regression works very well to predict continuous variables like cost, expenses, salary, age etc., whereas classification reflects good results for discrete variables like whether certain attributes classify the customer fit for loan granting or not, whether it will rain or not.

A. Linear regression

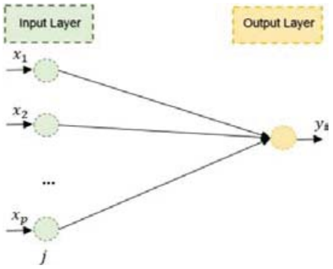


Fig. 2. Linear regression

Figure 2 shows the working of Linear regression. Since regression searches for correlation between dependent and independent attributes, the task of the Regression models is to find the mapping function that maps the input variable(independent) to the continuous output variable(dependent). [6]

Example: Suppose we want to predict sales in a particular season, so for this, we can very well utilize the Regression algorithm. Firstly, the model needs to be trained from the past data, and after completion of training the model can be deployed for future prediction. Regression models are finding wide usage in weather forecast prediction, stock market predictions, game outcome predictions, polls predictions etc. Besides this, many other realistic domains find wide usage of regression models. Variations of regression are logistic regression and root mean square regression which finds major applications.

B. KNN

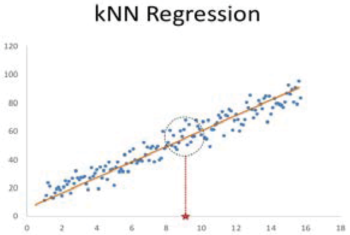


Fig. 3. KNN

Figure 3 shows the plot drawn between dependent and independent variable in case of KNN. Another supervised machine learning model which is widely used for classification and regression is K-Nearest Neighbor. The logic behind classifying data in K NN is the similarity of features in test data with the available data set. [7]

K-NN algorithm is a lazy learner parametric algorithm because its didn't initiate training upon arrival of training data set, instead it stores the data set and when classification of test data is needed, it performs the action of training the data set used for same. It can be used both for classification as well as for regression, but it mostly finds its applications in classification problems.

Example: Suppose, we have a picture of a fruit that resembles in features with that of an apple and pear, but we need to recognize the fruit as an apple or pear. One of the algorithm which can be used for this is KNN algorithm, since its working is dependent on similarity measure. KNN model works by correlating features of the new data set to that of apples and pears images and based on the most frequent attributes it will classify the test data in one of the categories.

C. Decision Tree

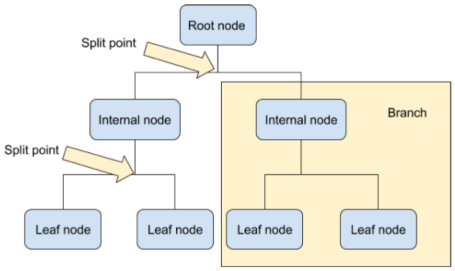


Fig. 4. Decision Tree

Figure 4 describes how a tree is drawn depending on the results obtained at points of splitting in a node. A decision Tree is again a Supervised machine learning technique that finds its applications for both classification and Regression problems, but generally, it is chosen for solving Classification problems [8]. The structure of the classifier developed in the decision tree is that of a tree as the name suggests. Internal nodes of the trees represent different features in the dataset, decision rules are represented as branches of the decision tree, and the resulting outcome of the rules are represented as leaf nodes in the tree. Decision nodes represented as branches of the tree can take multiple values but the outcome or result of the decision represented by Leaf didn't further continue with any more branches. The decision for moving to a particular branch or Leaf is dependent on features present in the dataset. It uses two methodologies Iterative Dichotomiser (ID3) the CART algorithm, which stands for Classification and Regression Tree algorithm.ID3 basically calculates information gain of features for decision while CART calculates Gini Index of features present in data set for decision. The answer to the question being asked while constructing a decision tree results in either yes or No. The decision whether to split the tree further or not is based on the binary answer received.

D. Random Forest

Random Forest is another famous widely used supervised machine learning algorithm. It finds its usage for both Classification and Regression techniques in ML [9]. Ensemble learning is the principle upon which Random Forest works. Ensemble learning solves complex classification problems by mixing multiple classifiers. This further improves the efficiency of the overall model built. Figure 5 reflect that Random forest is combination of multiple decision trees.

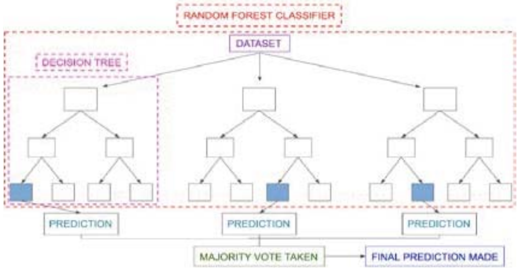


Fig. 5. Random forest

Random Forest is a unique type of classifier that consists of a combination of multiple decision trees designed on discrete subsets of the original dataset. To improve predictive accuracy of the test dataset, it takes help of the average of results of each decision tree.

E. XG Boost

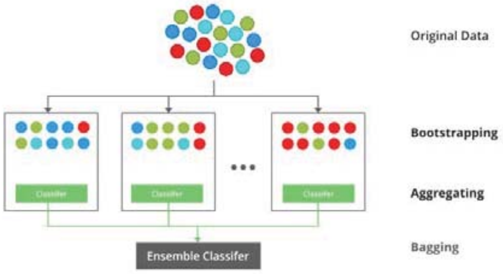


Fig. 6. XG boost

Figure 6 shows XG Boost as a combination of Bagging, Aggregating and Bootstrapping [10]. After having discussed different machine learning models and algorithms that we will be utilizing in our data set. A comparative study of the results obtained by these models will be done before selecting the final model for the dataset used in this work. Among all the models discussed, the supervised machine learning whose performance is at par with its peers will be selected for final design.

III. Dataset Used

The Black Friday Discounts dataset [11] is used to train a variety of machine learning models as well as to forecast the amount of people who will make purchases during Black Friday sales. Retailers will be able to study and tailor offers for more customers' favorite products using the purchase prediction provided. The dataset mentioned is the Analytics Vidhya Black Friday Sales Dataset. They have machine learning models like XGBoost, Linear Regression, MLK classifier, Decision Tree, Decision Tree with bagging, and Deep Learning model using Keras. The models utilized are assessed using the performance evaluation metric Root Mean Squared Error (RMSE). The dataset used is the Black Friday sales dataset, it is a popular open-source dataset that can be found on Kaggle as well. The dataset contains 2 CSV files namely, test.csv and train.csv.

- The training data has 474330 entries and 12 columns.
- The test data has 233599 rows and 11 columns (the purchase column is missing).

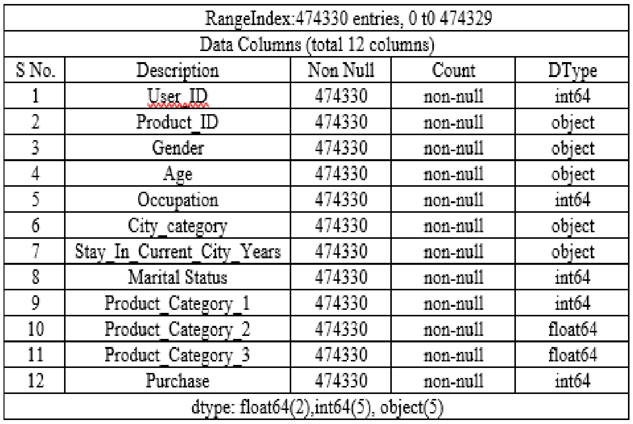
Before proceeding with the stated dataset in discrete models of machine learning, an analysis is done for the removal of redundant or trivial data [12]. This is needed to maintain uniformity in the dataset. Besides this removal of redundant or trivial data reduces system complexity.

IV. Exploratory Data Analysis

For summarizing the primary characteristics of a dataset, for analyzing and investigating datasets exploratory data analytics is needed. This is basically done by means of visualization methods. Exploratory data analytics helps to manipulate or make modifications in data sets to get the requisite results. [13]

The first step is to go through the data and understand the datatypes(DType) in the dataset as shown in Table 1.

TABLE I. TABLE1. INFORMATION OF COLUMNS IN TRAINING DATA



RangeIndex:474330 entries, 0 t0 474329				
Data Columns (total 12 columns)				
S No.	Description	Non Null	Count	DType
1	User_ID	474330	non-null	int64
2	Product_ID	474330	non-null	object
3	Gender	474330	non-null	object
4	Age	474330	non-null	object
5	Occupation	474330	non-null	int64
6	City_category	474330	non-null	object
7	Stay_In_Current_City_Years	474330	non-null	object
8	Marital Status	474330	non-null	int64
9	Product_Category_1	474330	non-null	int64
10	Product_Category_2	474330	non-null	float64
11	Product_Category_3	474330	non-null	float64
12	Purchase	474330	non-null	int64
dtype: float64(2),int64(5), object(5)				

The above table gives information about the data types of the columns in the training dataset. To further explore the dataset, finding null values and unique values in the data are common approaches. The observations made from these two operations are:

- Product_Category_2 comprises 31.57% null values which can be dropped before moving ahead.
- Product_Category_3 comprises 69.67% null values so this feature can be dropped.

Table2 returns a list as the percentage of null values in each column using the. isnull () operation.

TABLE II. TABLE 2. CHECKING NULL VALUES

User_ID	0.0%
Product_ID	0.0%
Gender	0.0%
Age	0.0%
Occupation	0.0%
City Category	0.0%
Stay_In_Current_City_Years	0.0%
Marital Status	0.0%
Product_Category_1	0.0%
Product_Category_2	31.04%
Product_Category_3	69.42%
Purchase	0.0%
dtype: object	

The next steps include performing univariate, bivariate, and multivariate analyses to better understand the important features in the dataset and the main factors affecting the purchase.

A. Univariate analysis

A univariate analysis is the most basic kind of data analysis. Your data only has one variable since Uni stands for "one". Instead of dealing with causes or correlations as other types of data analysis does, it finds patterns in the data by collecting it, summarizing it, and expressing it as bar graphs, pie charts, histograms, etc. [14].

The results of the univariate analysis helped in finding which features are actually relevant and which are not.

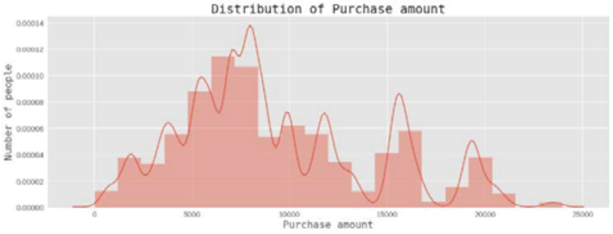


Fig. 7. Distribution of purchase amount

Figure 7 shows the bar plot for the distribution of purchase amount vs number of people.

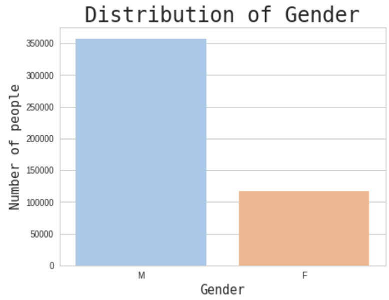


Fig. 8. Distribution of gender

In figure 8, the number of males purchasing against the number of females purchasing during black Friday is different and displayed.

Another interesting trend can be observed in figure 9, the age group distribution of the people purchasing.

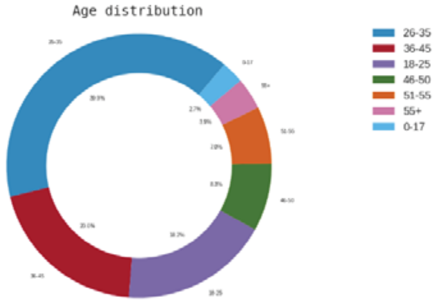


Fig. 9. Age distribution

B. Bivariate analysis

Bivariate analysis, by definition, evaluates any current connection between two variables or attributes. This study evaluates the link between the two variables as well as the strength of this correlation in order to ascertain whether there are any differences between the two variables and probable explanations of these discrepancies.

How the purchase quantity fluctuates across the product category 1 (figure 10) is an example of bivariate analysis employed in the project.

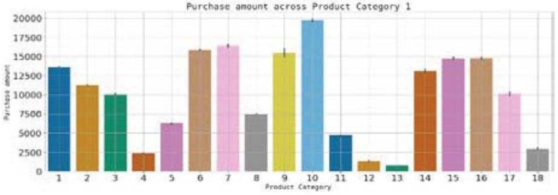


Fig. 10. Purchase amount across Product Category 1

C. Multivariate analysis

Each experimental unit is subjected to several measurements throughout the statistical analysis of data, and the relationships and organization of the multivariate measurements are important.

Using a heatmap, the relationship, correlation between the features is displayed (figure 11).

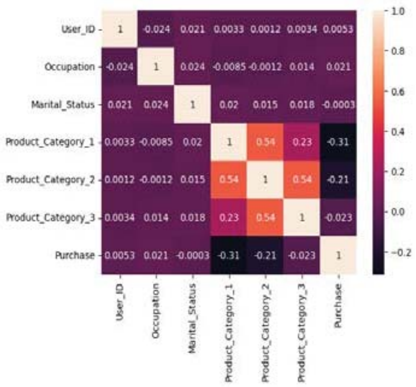


Fig. 11. Heatmap to Display Correlation Between Variables

A pair plot is also plotted using the seaborn library inbuilt function but as it can be seen in figure 12, the pair plot is difficult to deduce and the multiple features make it difficult to draw observation from it, in comparison the heatmap is much easier to observe and infer from.

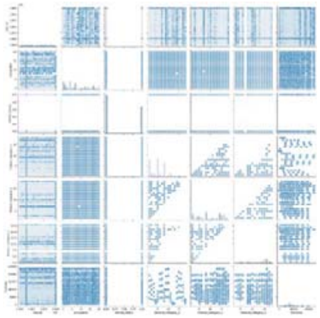


Fig. 12. Pairplot of features

The main observations made from these analyses are:

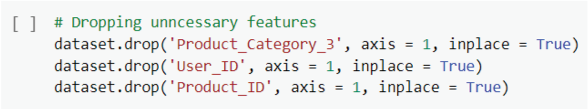
The number of females who shopped on Black Friday was less than the number of men

According to the heatmap, the dependent feature 'Purchase' is highly correlated with 'Product Category 1' and 'Product Category 2'.

V. Data Processing

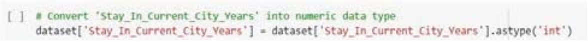
Data processing mainly involved getting the data ready for training. The changes to the data had to be made based on the observations made in the exploratory data analysis, this included, the code snippets for some important steps are added below the point itself:

- Dropping unnecessary features: User_ID and Product_ID, these features are better for database management. Product Category 3 as most of the values are null.



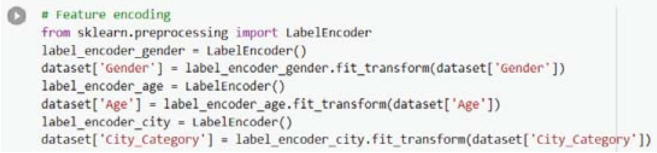
```
[ ]  # Dropping unncessary features
     dataset.drop('Product_Category_3', axis = 1, inplace = True)
     dataset.drop('User_ID', axis = 1, inplace = True)
     dataset.drop('Product_ID', axis = 1, inplace = True)
```

- Fixing null values: based on observations in the exploratory data analysis.
- Replacing values
- Merging train and test data
- Converting ‘Stay in current years’ into numeric data type



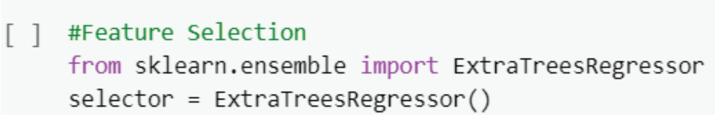
```
[ ]  # Convert 'Stay_In_Current_City_Years' into numeric data type
     dataset['Stay_In_Current_City_Years'] = dataset['Stay_In_Current_City_Years'].astype('int')
```

- Feature encoding: to convert the remaining data types into numerical data type for model training.



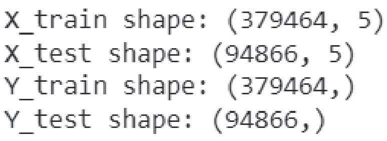
```
# Feature encoding
from sklearn.preprocessing import LabelEncoder
label_encoder_gender = LabelEncoder()
dataset['Gender'] = label_encoder_gender.fit_transform(dataset['Gender'])
label_encoder_age = LabelEncoder()
dataset['Age'] = label_encoder_age.fit_transform(dataset['Age'])
label_encoder_city = LabelEncoder()
dataset['City_Category'] = label_encoder_city.fit_transform(dataset['City_Category'])
```

- Feature scaling: in order to standardise all independent features in given range.
- Separating the data back in train and test
- Feature selection: ExtraTreeRegressor from the scikit learn library.



```
[ ]  #Feature Selection
     from sklearn.ensemble import ExtraTreesRegressor
     selector = ExtraTreesRegressor()
```

- Splitting the data into test and train sets: the shape of each subset can be viewed in figure 13.

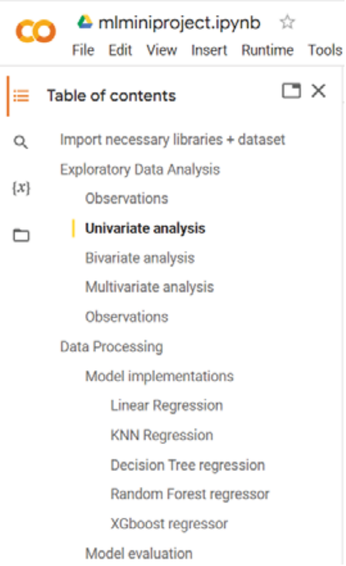


```
X_train shape: (379464, 5)
X_test shape: (94866, 5)
Y_train shape: (379464,)
Y_test shape: (94866,)
```

Fig. 13. Train, Test set shapes

VI. Model Implementation

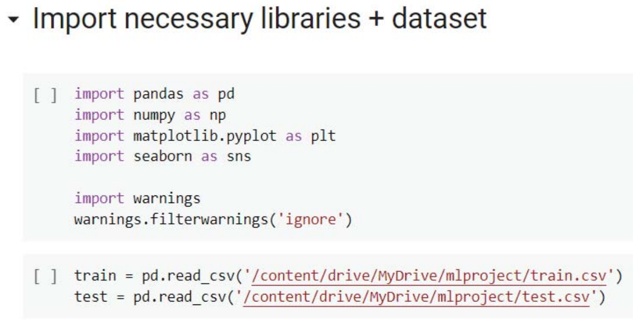
One of the most important aspects of model implementations is figuring out its hardware specifications and the environment for training. The models trained as a part of this implementation were trained on an intel i7 10th gen laptop with 16 GB RAM. The environment chosen for training is the Google Colab, the freely available platform for machine learning and data science. The GPU hardware accelerator was picked for better training, the Tesla K80 with 12 GB Ram.



Import necessary libraries + dataset
Exploratory Data Analysis
Observations
Univariate analysis
Bivariate analysis
Multivariate analysis
Observations
Data Processing
Model implementations
Linear Regression
KNN Regression
Decision Tree regression
Random Forest regressor
XGboost regressor
Model evaluation

Fig. 14. Google Colab notebook

Figure 14 shows the sections of Google Colab notebook used in the study. Importing necessary libraries and the dataset is the first step before starting any other process, Figure 15.



▾ Import necessary libraries + dataset

```
[ ] import pandas as pd
    import numpy as np
    import matplotlib.pyplot as plt
    import seaborn as sns

    import warnings
    warnings.filterwarnings('ignore')
```

```
[ ] train = pd.read_csv('/content/drive/MyDrive/mlproject/train.csv')
    test = pd.read_csv('/content/drive/MyDrive/mlproject/test.csv')
```

Fig. 15. Importing libraries and dataset

After a thorough data analysis and processing the data is ready to be trained. All models are imported from sklearn, numerous effective methods for machine learning and statistical modeling, such as classification, regression, clustering, and dimensionality reduction, are included in the Sklearn package.

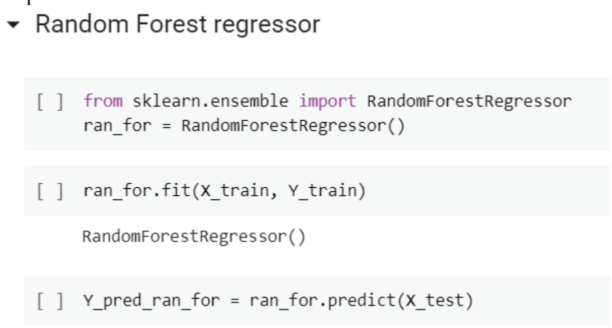
The models implemented are:

- Linear regression
- KNN
- Decision tree
- Random forest
- XGboost

The procedure for model training is using the same protocol and syntax.

- First, import the needed model
- Fit the model on X_train and Y_train dataset
- Use the predict function on X_test and save the result in Y_pred

Figure 16 correctly describes the process used for model implementation.



▾ Random Forest regressor

```
[ ] from sklearn.ensemble import RandomForestRegressor
    ran_for = RandomForestRegressor()
```

```
[ ] ran_for.fit(X_train, Y_train)

    RandomForestRegressor()
```

```
[ ] Y_pred_ran_for = ran_for.predict(X_test)
```

Fig. 16. Model Implementation Snippet

VII. Results

The metrics used for evaluating the model performance are RMSE (root mean squared error) and R^2 score.

A. *Root mean square error:*

The standard deviation of the residuals is defined as the Root Mean Square Error (RMSE). The residuals measure the distance between the data points and the regression line, and the RMSE measures the spread of these residuals. The following is the RMSE formula shown in Equation (1):

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

Where Y_{p1}, Y_{p2}, Y_{p3}……. are predicted Values
Y_1, Y_2, Y_3…… are observed Values
And n is the number of Observations

B. *R^2 score:*

The r^2 score is between 0% and 100%. It is closely connected to the mean square error even if they are not the same. The proportion of the dependent variable's variation that can be predicted using the independent variable is known as r2.

If the correlation coefficient is 100%, the two variables are perfectly associated, that is, there is absolutely no variation. Though not necessarily, a low number would suggest a poor correlation, which would suggest that the regression model is flawed. The formula for R^2 score is given by equation (2):

$$R^2 = 1 - \frac{RSS}{TSS} \quad \ldots\ldots(2)$$

Where R^2 = Coefficient of Determination
RSS = Sum of Square of Residuals
TSS = Total Sum of Residuals

TABLE III. Tabular Representation of Regression Model's Performance

Results		
Model	***RMSE***	***R^2***
Linear regression	4722.69	0.1029
KNN regression	3295.77	0.5631
Decision tree regression	3085.12	0.6171
Random forest regression	3047.12	0.6265
XGB regression	3023.49	0.9323

Table 3 shows a comparative study of different regression techniques based on RMSE and R^2 results obtained. Root mean square error RMSE, denotes the error. This should be as low as possible for any machine learning models. When different models are compared, the model which is having lesser RMSE is selected for deployment, since the lesser the error, the greater the model efficacy. Another parameter considered while evaluating model performance is R^2 Score, which denotes the correlation between variables in the model. It lies between 0 and 1. The greater its value, the higher correlation between different attributes used while designing model. So a value near 1 is the desirable R^2 value in a good machine learning model.

VIII. Conclusion and Future Work

After analyzing the scores obtained by different classification algorithms of machine learning based on RMSE and R^2 it is found that the machine learning model best suitable for dataset used in this study is XGB regression XGB regression which is an ensemble model in machine learning and theoretically outperforms the other models. This is verified in the obtained results as well.

The work can be carried forward as a customized business solution in the form of an application where product category wise numeric values can be returned to view the predicted price. We have considered only 3 products in the study, the results obtained can be further tested for more no of products. Also an analysis of customer behavior or shopping pattern can be done based on time of the year and currents needs of the customer as well. In machine learning, more models can be trained and visualized using different graphs. Other metrics can be used for comparison of performance amongst models and specific metrics for models (for e.g., information gain for decision tree) can be used to evaluate individual performance of machine learning models.

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] Abdulvahap Baydas, Serhat Ata,, Nesimi Kok, "An Empirical Study to Determine the Impact of Black Friday Days on Consumer Purchasing Behavior". Journal of Current Marketıng Approaches and Researches, Vol:1 Issue: 2.,2021.

[3] 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.

[4] Montgomery, D. C., Peck, E. A., & Vining, G. G. Introduction to linear regression analysis. John Wiley & Sons.2021.

[5] Song, Y. Y., & Ying, L. U.," Decision tree methods: applications for classification and prediction". Shanghai archives of psychiatry, .2015

[6] 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.

[7] Majumder, G. "Analysis and prediction of consumer behaviour on black friday sales". Journal of the Gujarat Research Society, 2019.

[8] 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.

[9] Holý, Vladimír, Ondřej Sokol, and Michal Černý. "Clustering retail products based on customer behaviour." Applied Soft Computing Journal PP 752-762, 2017

[10] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., & Chen, K. Xgboost: extreme gradient boosting. R package version 0.4-2015.

[11] https://www.kaggle.com/datasets/sdolezel/black-friday

[12] Martinez, W. L., Martinez, A. R., & Solka, J. Exploratory data analysis with MATLAB. Chapman and Hall/CRC. (2017).

[13] 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.

[14] M. J. Awan, M. Shafry, H. Nobanee, A. Yasin, O. I. Khalaf et al., "A big data approach to black friday sales," Intelligent Automation & Soft Computing, vol. 27, no.3, pp. 785–797, 2021.