Black Friday Sales

Athira k, Department of Computer Science

Rajagiri College Of Social Sciences, Kochi, Kerala, msccs2210@rajagiri.edu

***Abstract* - Black Friday Sale is a colloquial term for the Friday after Thanksgiving in the United States. It marks the shopping season and many stores offer highly promoted sales at discounted prices and often open early. Also important to retailers since it is Cyber Monday, the first day back to work for many consumers after the long holiday, on which online retailers offer major discounts. This is an important day for sellers, so analyzing and predicting customer purchasing patterns can help them to achieve more profit and provide special offers to attract targeted customers and others too. This case study is to identify the customer purchase behavior against different product categories and aims to build a model to predict the same and create a personalized offer for customers against different products.**

***Keywords-model, black friday sale, gradient boosting, purchase.***

1. **INTRODUCTION**

The Black Friday Sales dataset is a publically available dataset that contains information on customer transactions during the Black Friday Sales event at a retail store. The dataset includes a range of variables such as customer demographics, product categories, purchase amounts, and product ratings. The dataset provides a valuable resource for studying customer behavior and purchase decisions during this significant shopping event

The dataset comprises 550,068 rows and 12 columns, making it a substantial dataset for analysis. This study, aim to perform exploratory data analysis on the Black Friday Sales dataset to gain insights into customer behavior and purchase patterns. Specifically, plan to investigate the following research questions:

1. Understanding the customerson the basis of their purchasing habits.
2. Understanding the purchasing habits according to Age groups, Occupation, City\_Categories.
3. The above segmented group of users can be then used to model the data and use to predict the purchase spend for each customer.
4. **LITERATURE REVIEW**

The Black Friday sale dataset on Kaggle contains data on purchases made during Black Friday in a retail store. The data in a retail store. The data includes customer demographics, purchase amount, product categories, and purchase history.

Several studies have used this dataset to analyze the shopping behavior of customers during Black Friday. In one study men spend more money than women , and that both men and women are more likely to buy electronics and appliances compared to other product categories. Another study analyzed the impact of demographic factors such as age, gender, and marital status on purchasing behavior, and found that younger shoppers are more likely to buy electronic products, while older shoppers are more likely to buy home appliances.

Another interesting finding from the dataset is the impact of customer loyalty on purchasing behavior. One study found that customers who have previously made a purchase during Black Friday are more likely to spend more money during subsequent Black Fridays, suggesting that loyalty programs can be effective in increasing sales.

The dataset has also been used to build predictive models for future Black Friday sales. One study used machine learning algorithms to predict customer purchase behavior based on their demographics and purchase history, and found that their model was able to accurately predict customer purchase behavior.

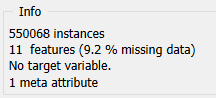
1. **IMPLEMENTATION**

**Tools Used-** Orange

Orange(3.31.0) is an open-source data mining and visualization toolkit. It is used for explorative rapid qualitative data analysis and interactive data visualization.

1. **Data Description**

The dataset used in this project is Black Friday Sales from Kaggle. The dataset is about a company, they wants to understand the customer purchase behavior against various products of different categories. They have shared purchase summaries of various customers for selected high-volume products from previous month. And, want to build a model to predict the purchase amount of customers against various products which will help to create a personalized offer for customers.

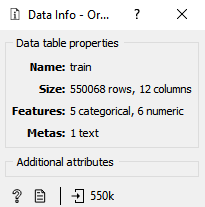
This dataset consists of 550068 and 11 features. Out of 12 variables in train data, there 6 numeric variables, 5 are categorical variables and 1 meta. The attributes in the datasets are:

|  |  |  |  |
| --- | --- | --- | --- |
| No | Attribute | Description | Type |
| 1 | User\_ID | Unique ID for each user | numeri |
| 2 | Product\_ID | Unique ID for each product | categorical |
| 3 | Gender | Gender of the user(M/F) | categorical |
| 4 | Age | Age of the users | categorical |
| 5 | Occupation | Occupation(masked) | numeric |
| 6 | City\_Category | Category of the city(Type B/Type C/Other) | categorical |
| 7 | Stay\_In\_Current\_City\_Years | Number of years stayed in the current city | numeric |
| 8 | Marital\_Status | Marital status(Married/Unmarried) | categorical |
| 9 | Product\_Category\_1 | Product category(masked) | numeric |
| 10 | Product\_Category\_2 | Product may belongs to other category also (masked) | numeric |
| 11 | Product\_Category\_3 | Product category may belongs to other category also (masked) | numeric |
| 12 | Purchase | Total Purchase amount | numeric |

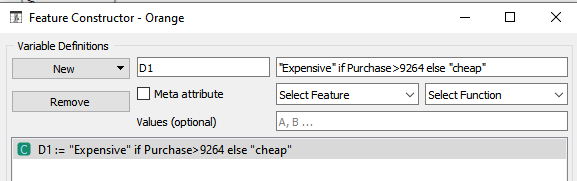
Masked in the description means already converted from categorical value to numerical column.

1. **Data Preprocessing**

The dataset contains 9.2% missing values. The target variable is Purchase which is numeric in nature so that it is converted into categorical by taking the mean from Feature Statistics of purchase, classified into cheap and expensive as a new column called D1 using Feature Constructor and the purchase column is removed. So D1 is set as target. Feature statistics gives the basic distribution of the data.



The missing values are imputed by average or most frequent value and a normalization is also applied to the dataset. Some special characters like + are also removed from columns ‘Age’ and ‘Stay\_In\_Current\_City\_Years’.Product\_ID which is meta , User\_ID are also removed.



The outliers in the dataset are removed using the Outlier widget,

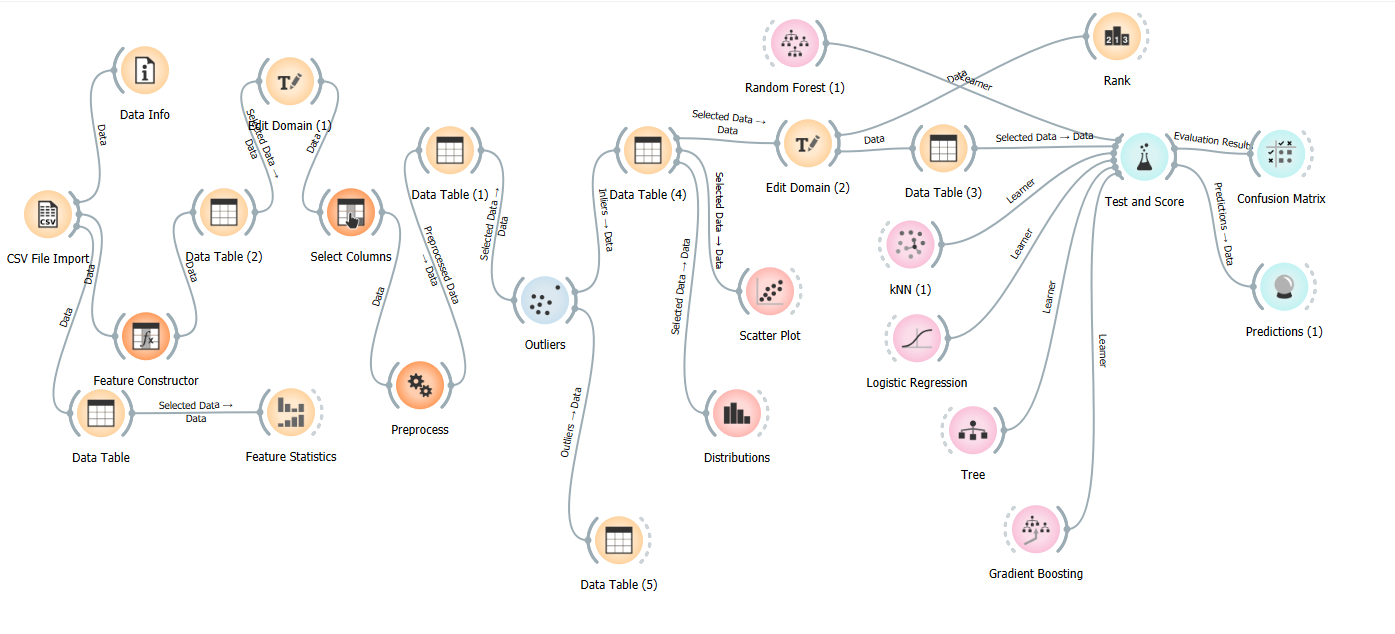
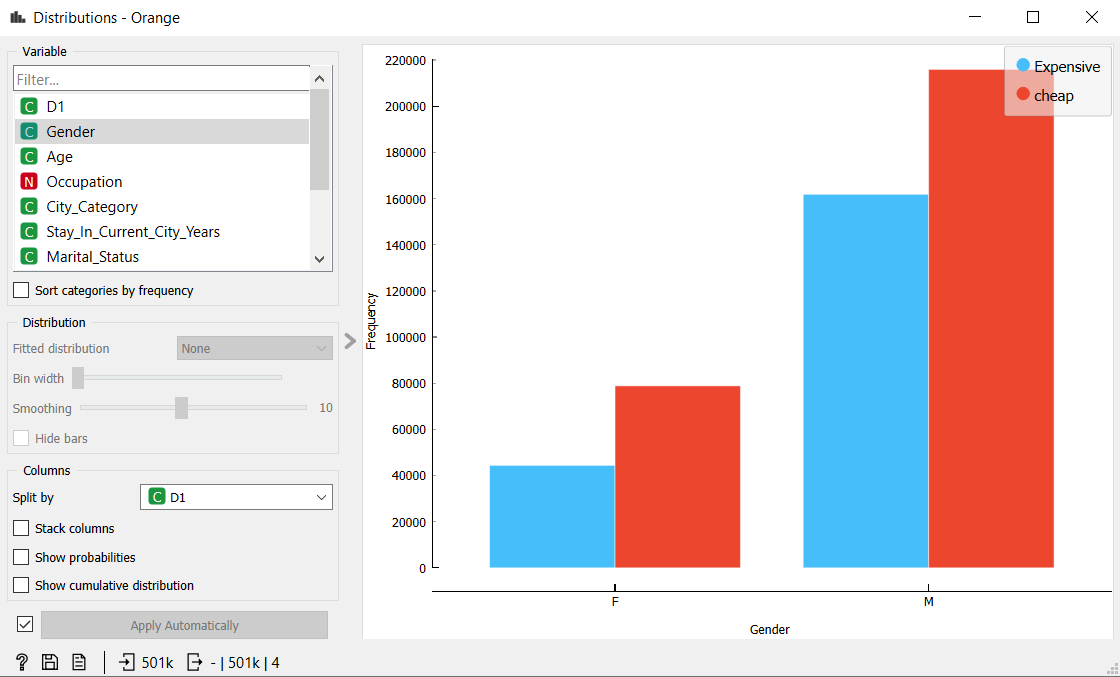


Fig: Orange connection diagram

1. Data Visualization

Visualization of dataset helps us to get a better idea about the dataset, helps to discover patterns, graphic representation of data and information. It is an efficient way of communicating when the information and data is numerous. Distributions is one of the visualization widget used here and scatter plot is also used. Using Scatter plot misclassified instances are understood that are close to class-bordering regions in the scatter plot projrction.

Fig: Distributions -Gender with respect to D1(target variable)

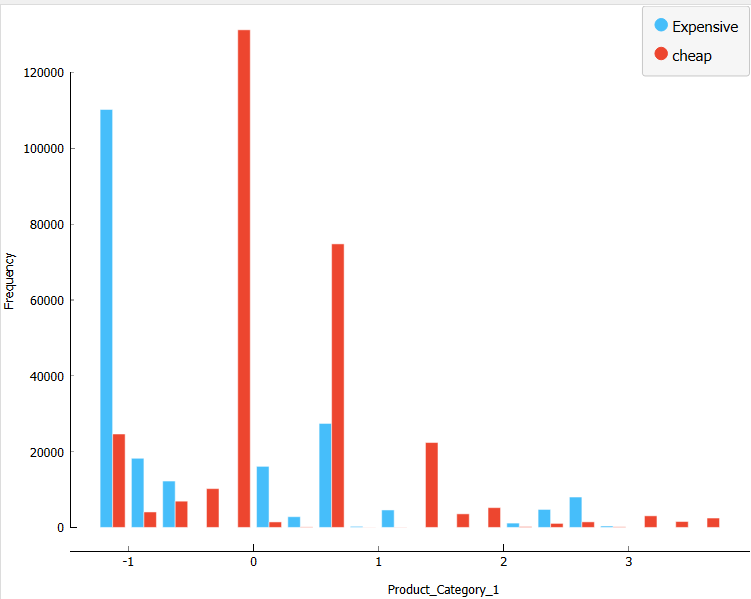


Majority of the purchase is made by male compared to females whether the product is expensive or cheap. According to this dataset, females buy less products.

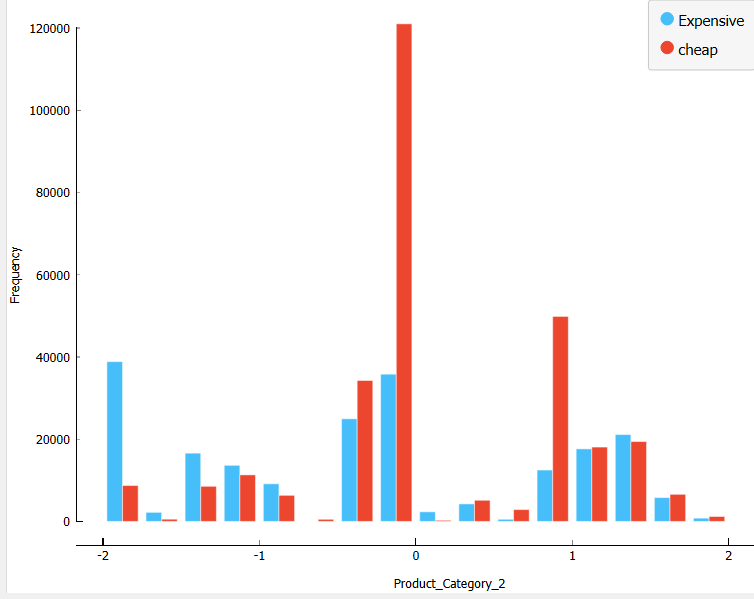
Fig:City\_Category

People staying in B category of the city are mostly buying expensive products compared to all other categories. In each category the number of people who buys cheap products are higher. To look at purchase of expensive products category B comes first and category C.

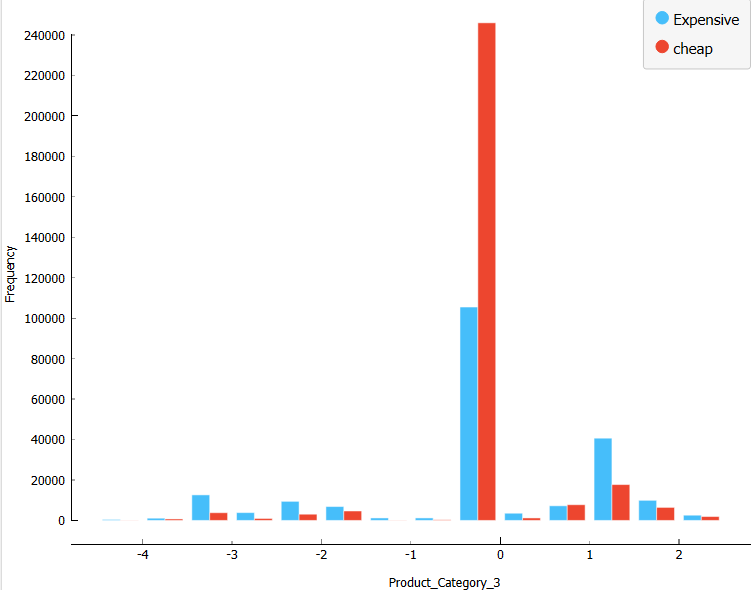
Fig: Product category 1



In product\_category\_1 products that closer to 0 are cheapest product bought by people and below -1 are the most expensive products being bought.



In product\_category\_2 is same as regarding the cheapest product being bought is closer to 0 category. and also expensive product is below -1 category and that is closer to -2 and also the second comes products closer to 0.



In Product\_Category3 the cheapest and expensive product being bought is between 0 and -1, that is products closer to 0.

1. Model Creation

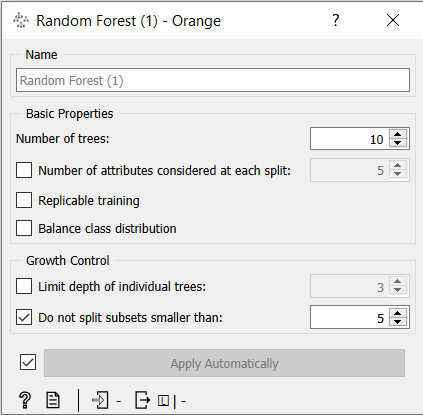
Data prediction is done with analysis, statistics and machine learning techniques. Here Random Forest, Logistic Regression, K-Nearest Neighbor, Gradient Boosting, Tree are used for prediction.

1. Random Forest

It consists of a number of trees that work as a whole.

Each tree in the random forest gives a prediction in which the one has high votes will be the model prediction. The greatest number in the of trees in the forest leads to higher accuracy and prevents the problem of over fitting.

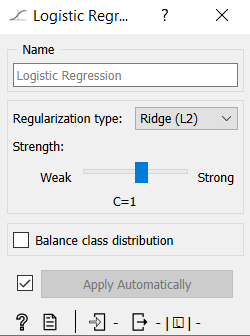
Fig:1



1. Logistic Regression

Logistic Regression is a statistical classification method that fits data into a logistic function.

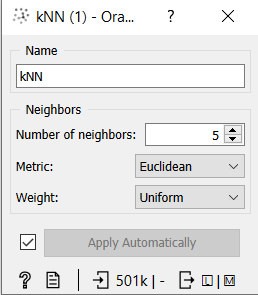
Fig:2



1. K-Nearest Neighbor

It is a supervised machine learning algorithms that provides simple solution to both regression and classification. When a new sample comes, the algorithm begins by calculating the distance between the new sample and the existing data points and it is assigned to the group common to the k nearest selected neighbors.

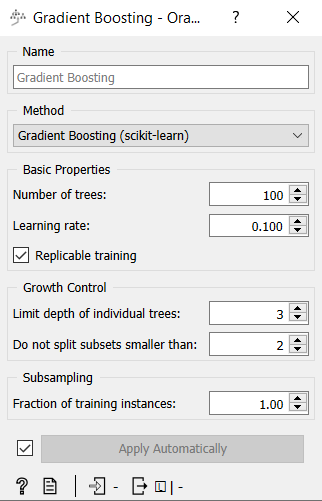
Fig:3



1. Gradient Boosting (GB)

It is a boosting technique in which every predictor corrects its predecessor’s error. The base of the gradient boosting is the CART (Classification and Regression Trees). Each tree predicts a label and final prediction is done by formulas:

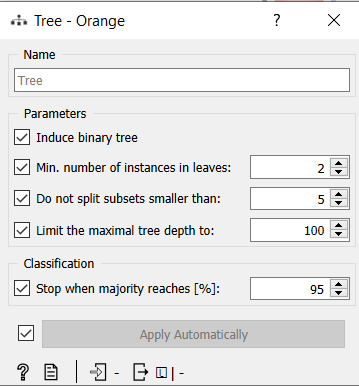
Fig:4



1. Tree

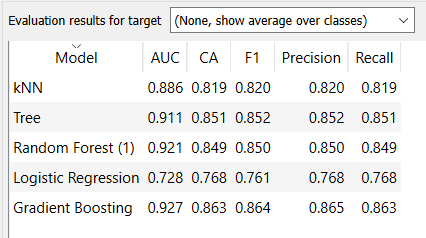
Tree is a simple algorithm that splits data into nodes by class purity.

Fig:5

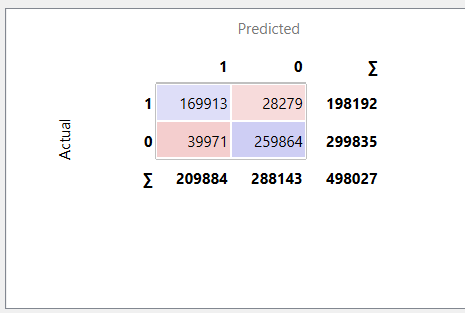
****

1. **RESULT**

The value obtained from each model is shown below:



Among the applied classification models, Gradient boosting has the highest accuracy. The confusion matrix of Gradient Boosting is given below:

****

1. **CONCLUSION**

The model created using Gradient Boosting algorithm can predict the purchase value whether it is expensive or cheap with an accuracy of 86.3%. The ultimate goal of the sale is to find the purchase amount against different product categories and provide customized offers. There are several factors that affect the purchase amount. From the analysis it is more men are buying products and in each of the product categories the cheapest products are being purchased at a greater level with respect to purchase(D1).

REFERENCES

<https://www.kaggle.com/datasets/rajeshrampure/black-friday-sale?datasetld=2758244>

<https://www.kaggle.com/code/ishanvardhan/black-friday-sales-analysis-and-prediction>

<https://www.investopedia.com/terms/b/blackfriday.asp>