CSE3505
Foundations of Data Analytics
J Component Report

A project report titled

BLACK FRIDAY SALES PREDICTION

By

19BEC1367 M. S. L. CHANDAN ABHISHEK
19BEC1129 AKASH GOWDA K. R.

BACHELOR OF TECHNOLOGY
IN
ELECTRONICS AND COMMUNICATION ENGENEERING

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Submitted to

Dr. R. KARTHIK

NOV 2021

School of Electronics Engineering

DECLARATION BY THE CANDIDATES

I hereby declare that the Report entitled "**BLACK FRIDAY SALES PREDICTION"** submitted by me to VIT Chennai is a record of bonafide work undertaken by us under the supervision of **Dr. R. Karthik, Senior Assistant Professor, SENSE, VIT Chennai.**

Signature of the Candidates

Akash Gowda

Chandan Abhishek

Chennai

30/11/2020.

ACKNOWLEDGEMENT

We wish to express our sincere thanks and deep sense of gratitude to our project guide, **Dr. R. Karthik,** School of Electronics Engineering for his consistent encouragement and valuable guidance offered to us in a pleasant manner throughout the course of the project work.

We are extremely grateful to **Dr. Sivasubramanian. A,** Dean of the School of Electronics Engineering (SENSE), VIT University Chennai, for extending the facilities of the School towards our project and for his unstinting support.

We express our thanks to our **Head of The Department Dr. Vetrivelan. P (for B.Tech-ECE)** for his support throughout the course of this project.

We also take this opportunity to thank all the faculty of the School for their support and their wisdom imparted to us throughout the courses till date.

We thank our parents, family, and friends for bearing with us throughout the course of our project and for the opportunity they provided us in undergoing this course in such a prestigious institution.

BONAFIDE CERTIFICATE

Certified that this project report entitled **"BLACK FRIDAY SALES PREDICTION"** is a bonafide work of **M.S.L. CHANDAN ABHISHEK (19BEC1367)** and **AKASH GOWDA K.R (19BEC1129)** carried out the "J"-Project work under my supervision and guidance for CSE 3505 Fundamentals of Data Analysis

Dr.R.Karthik

School of Electronics Engineering

VIT University, Chennai

Chennai – 600 127.

TABLE OF CONTENTS

ABSTRACT

Understanding the purchase behavior of various customers (dependent variable) against different products using their demographic information (IS features where most of the features are self -explanatory). Data analysis is the most common applications in the domain retail industry. This concept helps to develop a predictor that has a distinct commercial value to the shop owners as it will help with their inventory management, financial planning, advertising and marketing. This entire process of developing a model includes preprocessing, modelling, fitting the model and evaluating. Hence, frameworks will be developed to automate few of this process and its complexity will be reduced. Ability to recognize and track patterns in data help businesses shift through the layers of seemingly unrelated data for meaningful relationships. Through this analysis it becomes easy for the online retailers to determine the dimensions that influence the uptake of online shopping and plan effective marketing strategies.

The largest shopping day of the year in America is the Friday following the Thanksgiving holiday. It is recognized as the ignition of one of the busiest shopping seasons in a year. From the computer science point of view, one of the most interesting applications of machine learning in the retail industry is to effectively predict how much a customer is probably to spend at a store based on historical purchasing patterns. If retailers comprehensively understand their customers in terms of characteristics, behaviors and motivations in the previous shopping seasons, they can implement and develop more effective marketing strategies for specific customers categories. This study proposes an empirical implementation of extreme gradient boosted trees algorithm for addressing an interesting challenge in the retail industry. From the experimental results, the authors can conclude that the applications of bagging and boosting techniques can achieve great performance and be further improved by a proper combination of models' hyperparameters tuning and feature engineering.

CHAPTER-1

INTRODUCTION

The purpose of this study is to observe and analyze the consumer behaviors of the Black Friday customer. The day after Thanksgiving, Black Friday is a term used by the retail industry in the United States that signifies the Christmas holiday shopping season.

Fig 1.1

Thanks giving Day is on the last Thursday of November; therefore, the holiday shopping season runs from the Friday after Thanksgiving Day and continues until Christmas eve, the day before Christmas. Black Friday is not a national holiday. However, many employees have Thanksgiving Day holidays and the following day, increasing the number of potential shoppers on that Friday. Black Friday is famously known for long lines with customers waiting outdoors in cold weather waiting for the store to open, confusion, and customers' chaos. Once the retail doors open for business, the challenges faced are Heavily crowded stores, limited products available at a reduced price, long lines, and the lack of advertised sale products.

Black Friday was originated in Philadelphia around the year 1961. The police described it as the day where there would be heavy pedestrians and vehicular traffic after Thanksgiving Day. Since the 21st century, retailers have been attempts in the US to introduce Black Friday in other countries. Retailers outside the US have promoted black Friday sales to compete with the US retailers in online sales. Black Friday is considered to be the most significant sale that happens in the United States. Black Friday and Cyber Monday combined have a revolutionary history in the shopping industry. However, there have been many changes in

the trends and the shoppers. There are many 7 advancements and changes in shopping approaches, and there have also been many factors that influenced the traditional shopping methods. Since this is a digital era, there have been many online shopping activities compared to in-store shopping. There were predictions that the sales might go down during the recent pandemic since the in-store shopping was canceled for safety reasons. However, to our surprise, the end sales figures were nowhere near a loss.

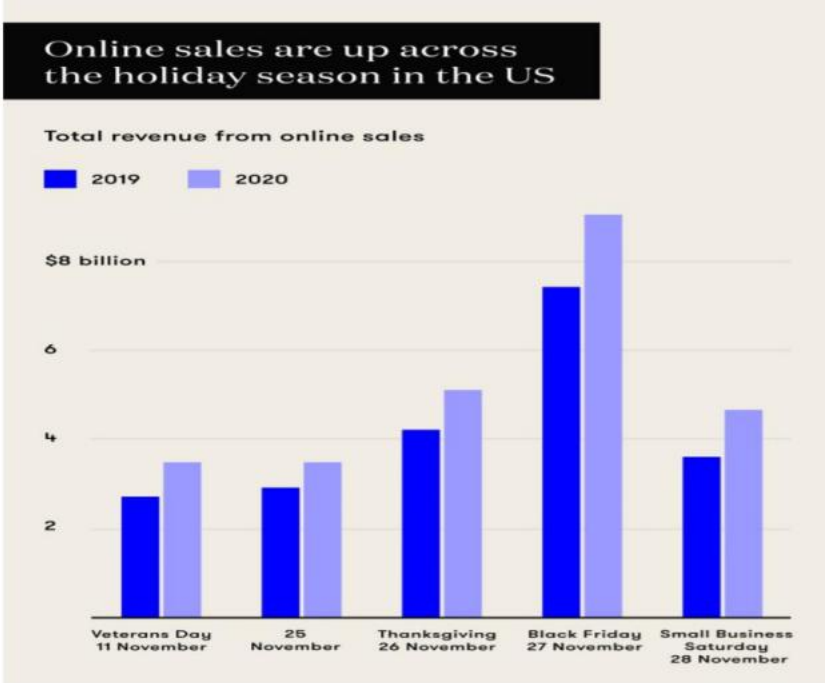

Fig 1.2

All this was possible by online shopping. Shoppers could sit at home and have the fun of Black Friday shopping and get the products delivered to their doorstep. Identifying all the data from this sale gives us a better opportunity to formulate a research topic and study it. It enables us to understand the customer perspective and the retailer's perspective and identify the shopping patterns of various age groups and their interest in shopping. It helps us categorize the products sold at a fast pace and the age groups that are purchasing these products. It can also help the retail industry understand what products are more in demand during the sale, predict future sales, and have their inventory ready to cater to their customers efficiently. Purchasing patterns and sales patterns can be formulated and help us analyze and identify the factors influencing the retail industry

CHAPTER-2

REQUIREMENTS AND PROPOSED SYSTEM

Proposed System

We are using Black Friday Sales Dataset publicly available on Kaggle. The dataset consists of sales transaction data. The dataset consists of 5,50,069 rows. The dataset consists of attributes such as user_id, product_id, martial_status, city_category, occupation, etc. The dataset definition is mentioned in below Table 2.1. The Black Friday Sales dataset is used for training various machine learning models and also for predicting the purchase amount 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. The Purchase Variable will be the predictor variable. The Purchase Variable will predict the amount of purchase made by a customer on the occasion of Black Friday sales.

Sl. No	Variable	Definition	Masked
1	User_ID	Unique ID of the User	False
2	Product_ID	Unique Product ID	False
3	Gender	Sex of User	False
4	Age	Age in bins	False
5	Occupation	Occupation	True
6	City_Category	Category of the City (A,B,C)	True
7	Stay_In_Current_City_Years	Number of years stay in current city	False
8	Marital_Status	Marital Status	False
9	Product_Category_1	Product Category	True
10	Product_Category_3	Product Category	True
11	Product_Category_3	Product Category	True
12	Purchase	Purchase Amount	False

Table 2.1

Our proposed method tries to implement the machine learning models such as Ridge Regression, Elastic Net Regression, Lasso Regression, and Gradient Boosting Regression to forecast sales. Figure 2.1 depicts the flow of data through the proposed model. Exploratory Data Analysis has been performed on the dataset. The tools used for the data analysis are python, pandas, matplotlib, NumPy, array, seaborn and jupyter notebook.

Fig. 2.1 Flowchart of Proposed System The Black Friday Sales Dataset is the input dataset. Data visualization of the various attributes of this dataset is performed. Data preprocessing which mainly includes filling missing values is performed. The categorical values are label encoded to numeric form. The categories such as Gender where F represents female and M represents Male is converted to numerical form as 0 and 1 also other categorical values such as City_Category, Stay_In-Current_City, Age are converted to numerical form by applying Label Encoding. The attributes such as User_id and Product_id are removed to train the model with no bias based on user_id or product_id and to achieve better performance. The algorithms used for implementing the system are Ridge Regression, Elastic Net Regression, Lasso Regression, and Gradient Boosting Regression. The performance evaluation measure used is Root Mean Squared Error (RMSE).

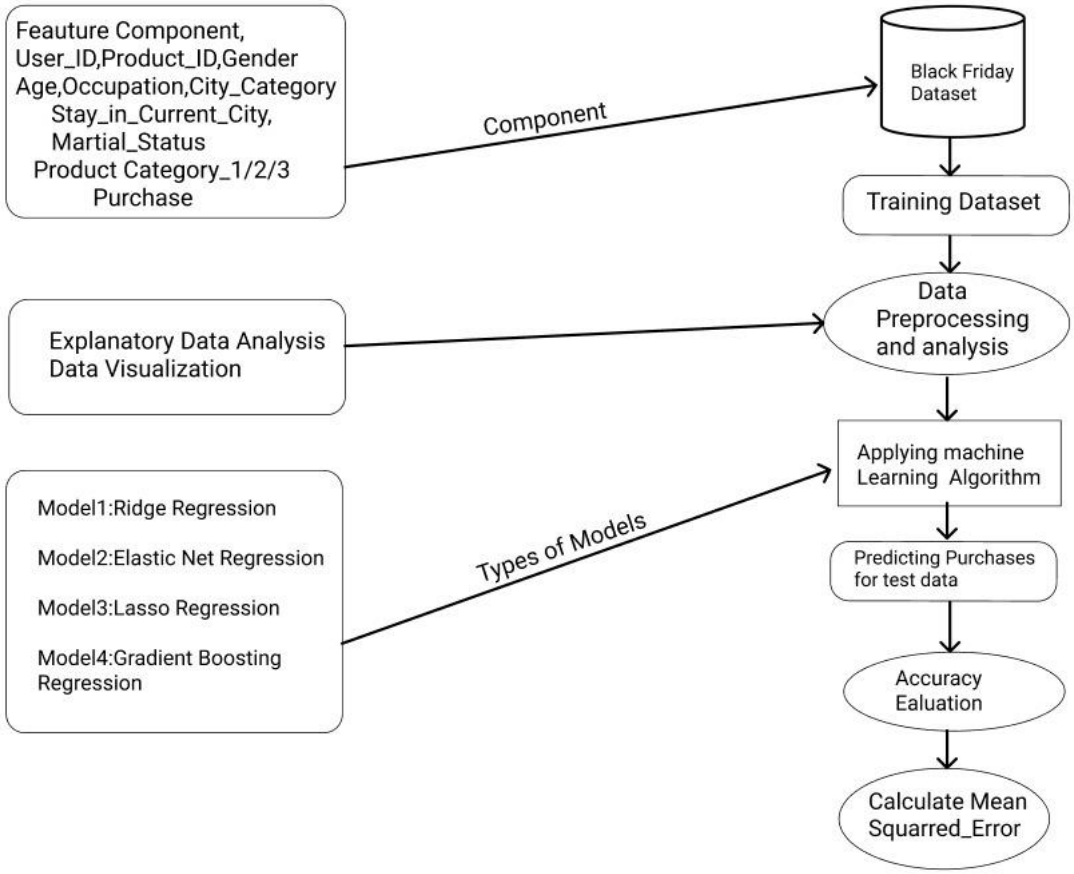

Fig 2.1

Requirements

The proposed Regression algorithms are implemented using Google colab notebook. Colab is a product from Google Research which allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education.

Google Colab adds collaboration, free GPU and TPU, cloud features, and additional pre-installed ML libraries and also Google Colab gives us three types of runtime for our notebooks namely CPUs(Central Processing Units), GPUs(Graphics Processing Units), and TPUs(Tensor Processing Units).

Fig 2.2

CHAPTER-3

MODULE DESCRIPTION

Data Visualization

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3	Purchase
0	1000001	P00069042	F	0-17	10	A	2	0	3	NaN	NaN	8370
1	1000001	P00248942	F	0-17	10	A	2	0	1	6.0	14.0	15200
2	1000001	P00087842	F	0-17	10	A	2	0	12	NaN	NaN	1422
3	1000001	P00085442	F	0-17	10	A	2	0	12	14.0	NaN	1057
4	1000002	P00285442	M	55+	16	C	4+	0	8	NaN	NaN	7969

Fig 3.1

In our training dataset we have 5,50,069 rows and 12 colums and in our testing dataset we have 2,30,300 rows and 11 columns and we have to determine the “Purchase” column.

Data Description

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3	Purchase
count	5.500680e+05	550068	550068	550068	550068.000000	550068	550068	550068.000000	550068.000000	376430.000000	166821.000000	550068.000000
unique	NaN	3631	2	7	NaN	3	5	NaN	NaN	NaN	NaN	NaN
top	NaN	P00265242	M	26-35	NaN	B	1	NaN	NaN	NaN	NaN	NaN
freq	NaN	1880	414259	219587	NaN	231173	193821	NaN	NaN	NaN	NaN	NaN
mean	1.003029e+06	NaN	NaN	NaN	8.076707	NaN	NaN	0.409653	5.404270	9.842329	12.668243	9263.968713
std	1.727592e+03	NaN	NaN	NaN	6.522660	NaN	NaN	0.491770	3.936211	5.086590	4.125338	5023.065394
min	1.000001e+06	NaN	NaN	NaN	0.000000	NaN	NaN	0.000000	1.000000	2.000000	3.000000	12.000000
25%	1.001516e+06	NaN	NaN	NaN	2.000000	NaN	NaN	0.000000	1.000000	5.000000	9.000000	5823.000000
50%	1.003077e+06	NaN	NaN	NaN	7.000000	NaN	NaN	0.000000	5.000000	9.000000	14.000000	8047.000000
75%	1.004478e+06	NaN	NaN	NaN	14.000000	NaN	NaN	1.000000	8.000000	15.000000	16.000000	12054.000000
max	1.006040e+06	NaN	NaN	NaN	20.000000	NaN	NaN	1.000000	20.000000	18.000000	18.000000	23961.000000

Fig 3.2

As depicted in the Fig 3.2 we can infer all the necessary statistical details of the dataset i.e., maximum value of the column, minimum value, mean, quantile divisons, standard deviations etc,.

A basic observation from the above description is that:

- Product P00265242 is the most popular product.
- Most of the transactions were made by men.
- Age group with most transactions was 26-35.
- City Category with most transactions was B.

We will explore more with data analysis.

Data types of variables

```
User_ID                       int64
Product_ID                   object
Gender                       object
Age                          object
Occupation                    int64
City_Category                object
Stay_In_Current_City_Years   object
Marital_Status                int64
Product_Category_1            int64
Product_Category_2          float64
Product_Category_3          float64
Purchase                      int64
dtype: object
```

Fig 3.3

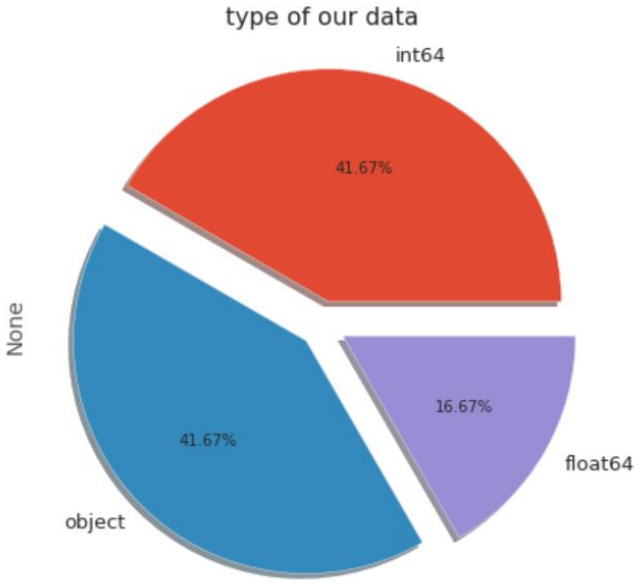

Fig 3.4

From the Fig 3.3 and 3.4 we can observe that in the dataset we have variables of object and integer type (each having 5 out of 12) and we have float variables at low count (2 out of 12).

Correleation Matrix

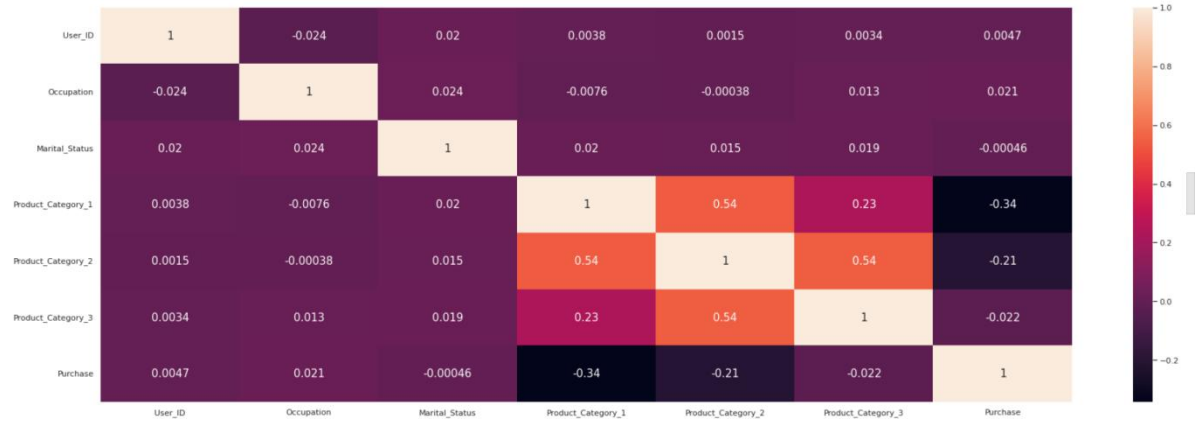

Fig 3.5

From the above Correlation matrix we can infer correlation between each variables. Surprisingly we can observe that most of variables have weak correlations between them except in case of Product categories.

Missing Values

	missing_values	percent_missing
User_ID	0	0.000000
Product_ID	0	0.000000
Gender	0	0.000000
Age	0	0.000000
Occupation	0	0.000000
City_Category	0	0.000000
Stay_In_Current_City_Years	0	0.000000
Marital_Status	0	0.000000
Product_Category_1	0	0.000000
Product_Category_2	173638	31.566643
Product_Category_3	383247	69.672659
Purchase	0	0.000000

Fig 3.6

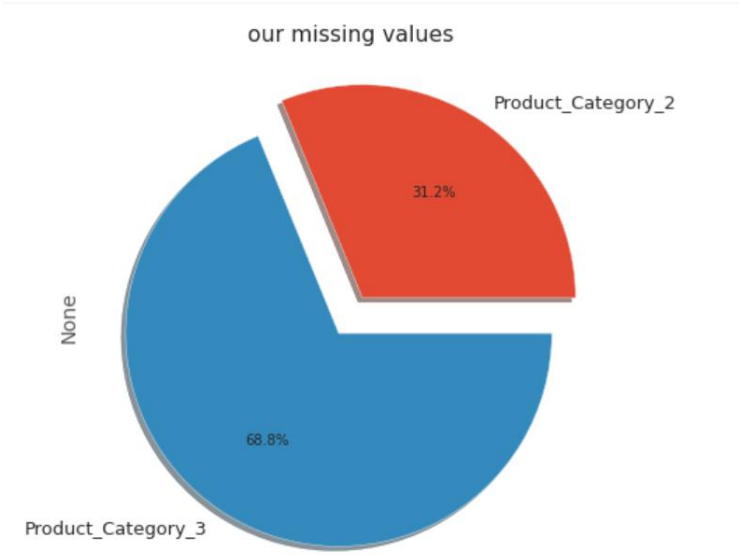

Fig 3.7

In our dataset missing values are only in Product_Category 2 and 3, but it's notable that in Product_Category_3 most of the values are missing (68.8%) so we won't get much information from that column so we are dropping that column. And we are filling Product_Category_2 column with median as it is less affected by outliers than mean.

Data imputation

	missing_values	percent_missing
User_ID	0	0.0
Product_ID	0	0.0
Gender	0	0.0
Age	0	0.0
Occupation	0	0.0
City_Category	0	0.0
Stay_In_Current_City_Years	0	0.0
Marital_Status	0	0.0
Product_Category_1	0	0.0
Product_Category_2	0	0.0
Purchase	0	0.0

Fig 3.8

After removing Product_Category_3 column and filling missing values of Product_Category_2 column with it's median we can observe that we have completed data imputation.

Unique Values

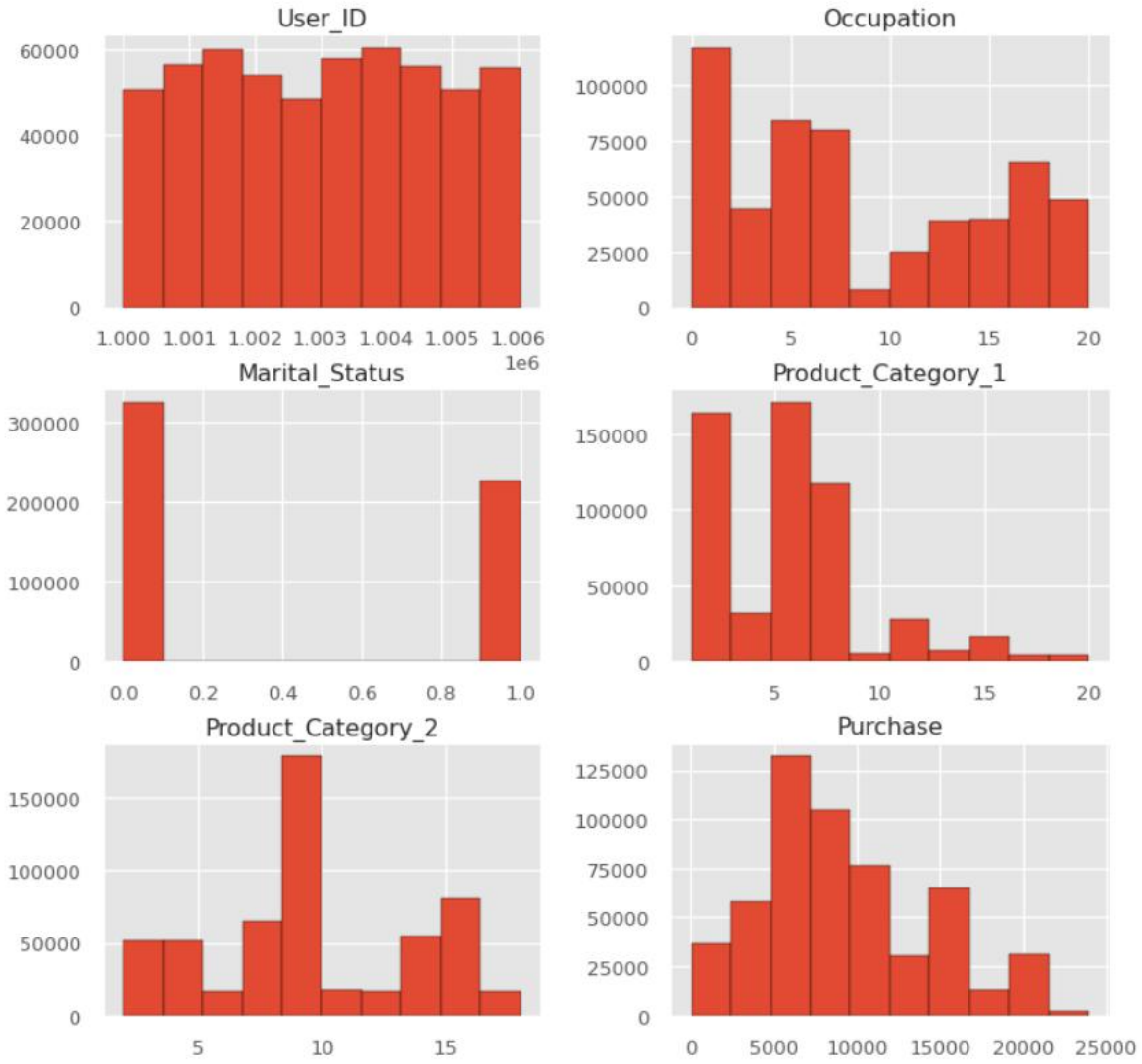

Fig 3.9

In the Fig 3.9 we can notice the unique values in each variable/column.

Factors Affecting the Purchase

- **Gender**

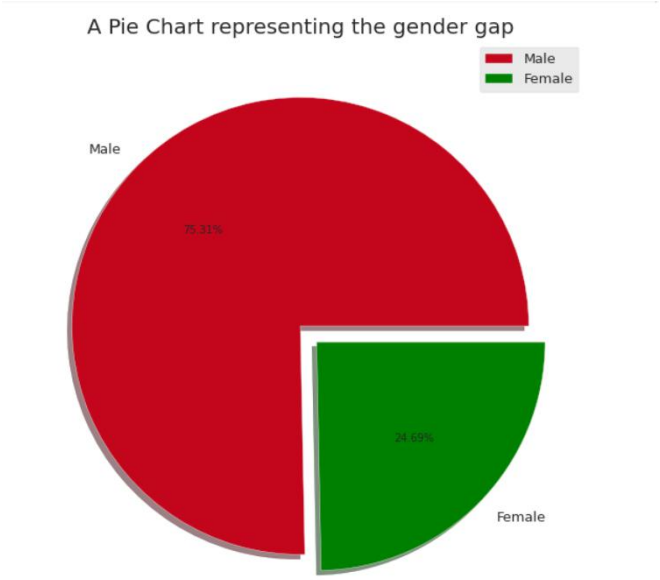

Fig 3.10

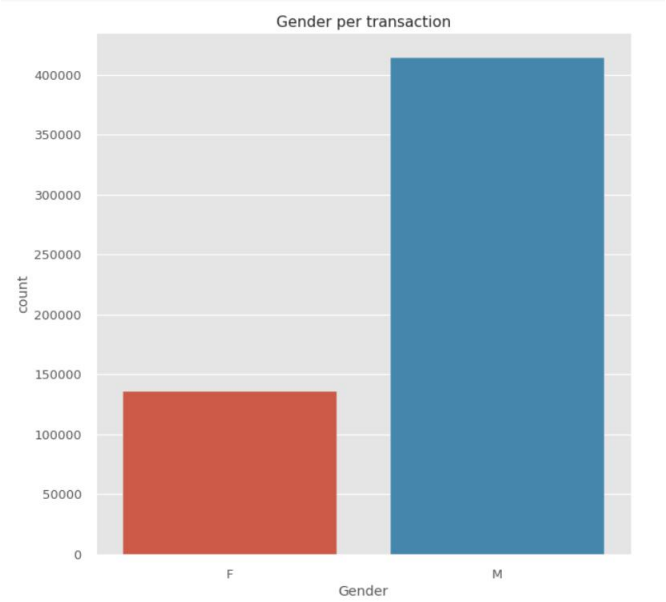

Fig 3.11

From the Fig 3.10 and Fig 3.11 we can clearly infer that Males are doing more purchase than Females.

➢ Age

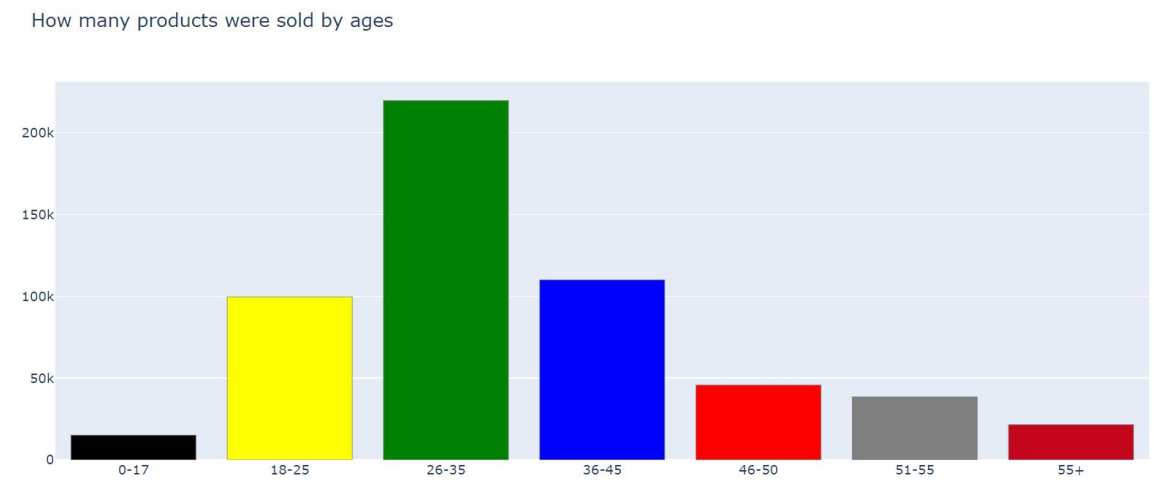

Fig 3.12

From the resuts obatined in Fig 3.12 we can observe that more purchase is done by people in age group of 26 to 35.

➢ Occupation

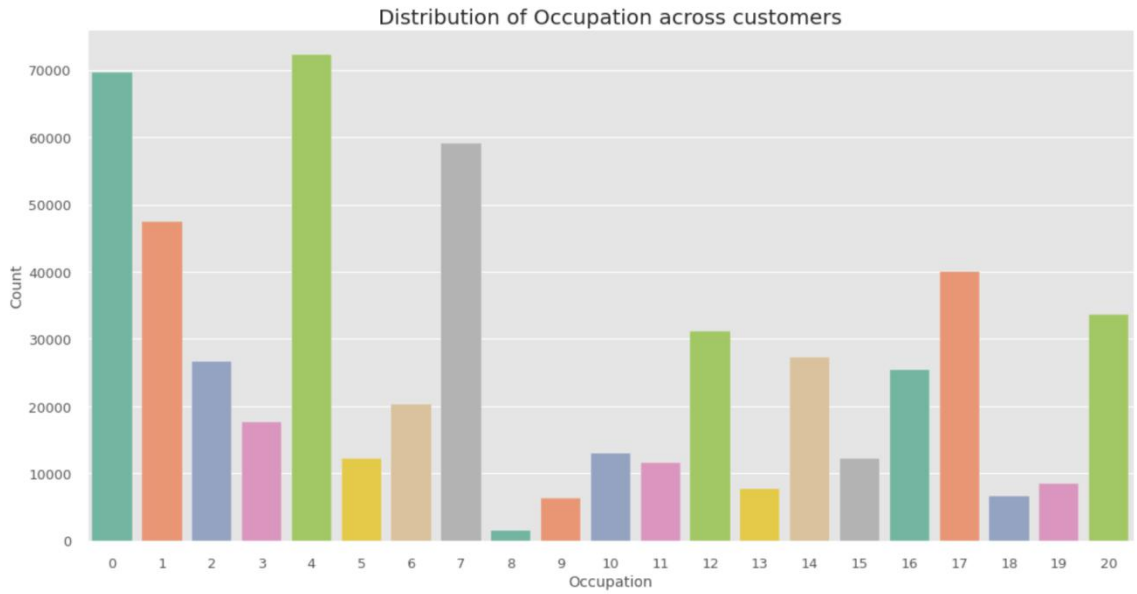

Fig 3.13

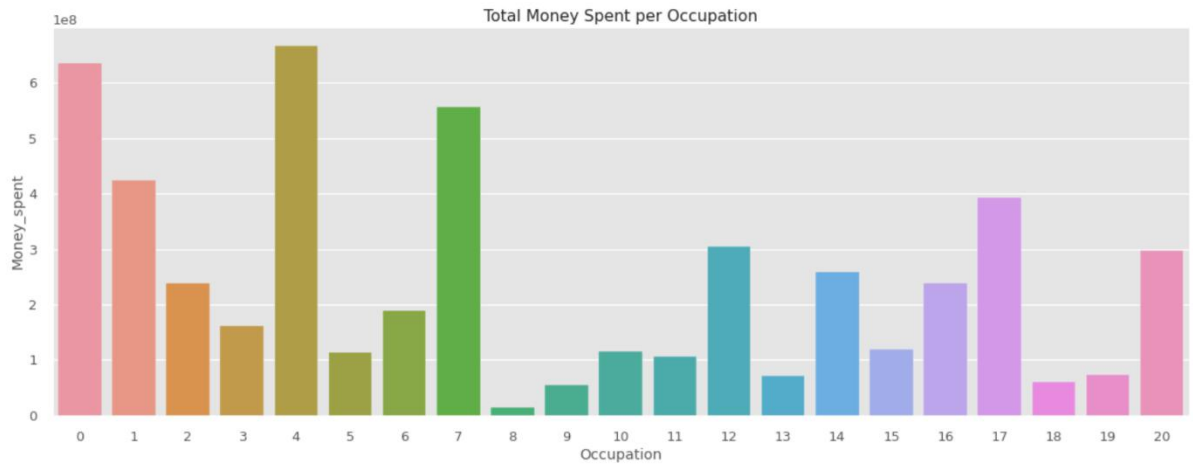

Fig 3.14

From Fig 3.13 and Fig 3.14 we can spot that, the distribution of the mean amount spent within each occupation appears to mirror the distribution of the amount of people within each occupation. This is fortunate from a data science perspective, as we are not working with odd or outstanding features. Our data, in terms of age and occupation seems to simply make sense.

➢ Cities

Fig 3.15

More purchase in done by people of city ‘C’

➢ **Stay in current years**

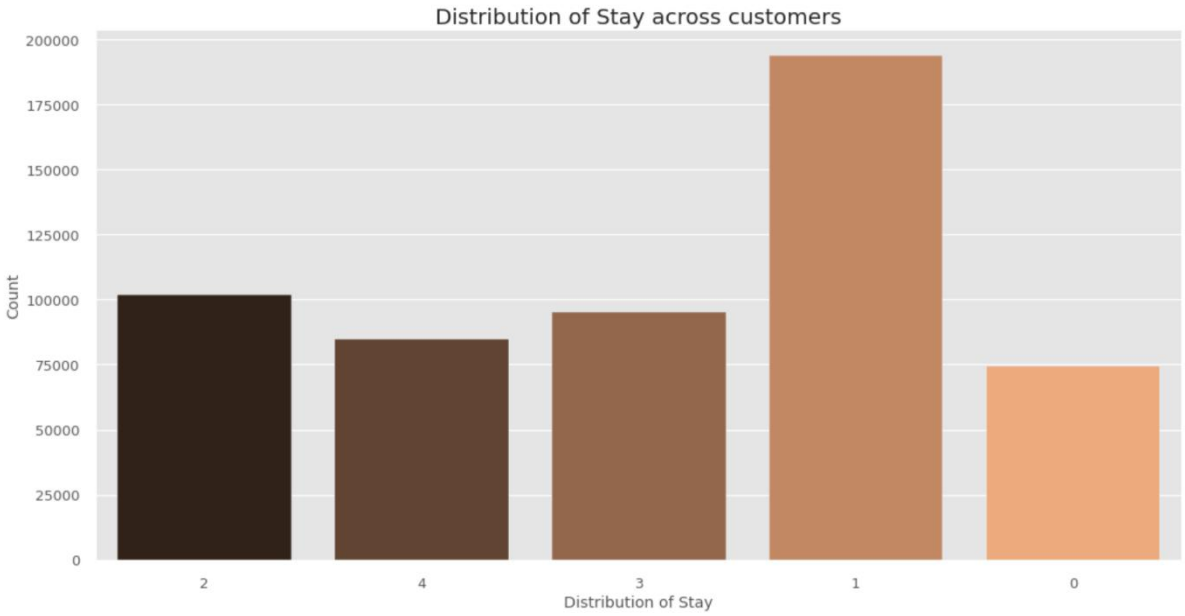

Fig 3.16

From the Fig 3.16 we can perceive that people do more purchase in the 1st year when they move to new city as they have to buy more items compared to later upcoming years

➢ Product Categories

Here we are going to explore the products themselves. This is the most important factor affecting the output, as we do not have labeled items in this dataset. Theoretically, a customer could be spending $5,000 on 4 new TVs, or 10,000 pens. This difference matters for stores, as their profits are affected. Since we do not know what the items are, let's explore the categories of the items.

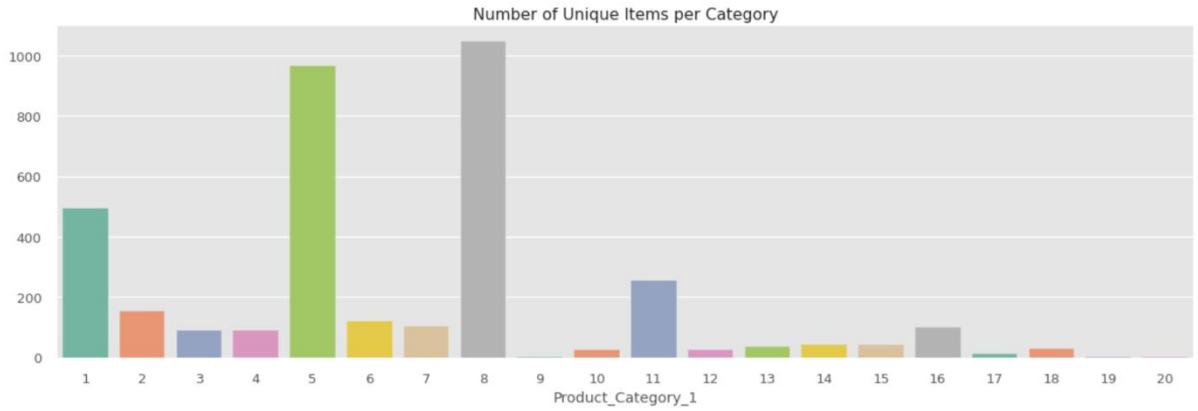

Fig 3.17

From the above Fig 3.17 we can notice that Category labels 1, 5, and 8 clearly have the most items within them. This could mean the store is known for that item, or that the category is a broad one.

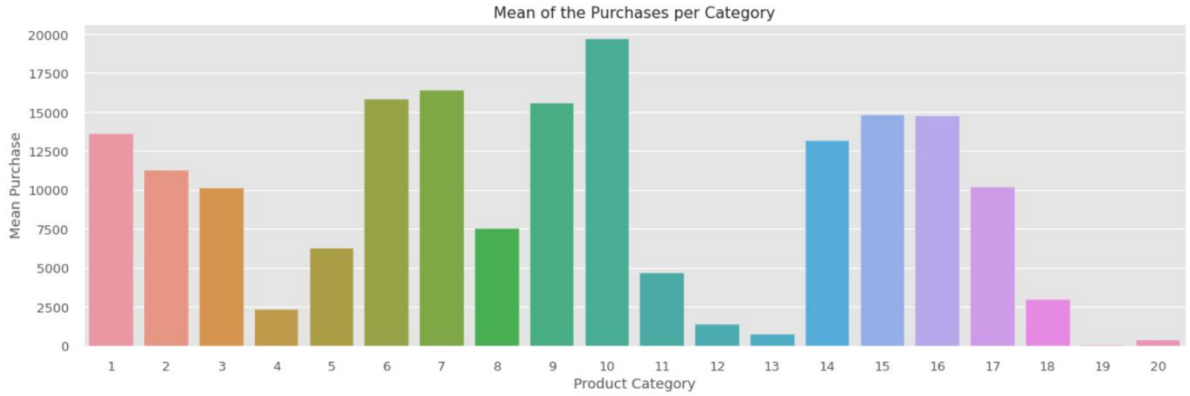

Fig 3.18

From the Fig 3.18 we can note that, Category labels 6, 7, 9, and 10 though not having much unique items in them have got high mean purchase. This could mean that they have got products of high cost or whatever they have are the most important ones.

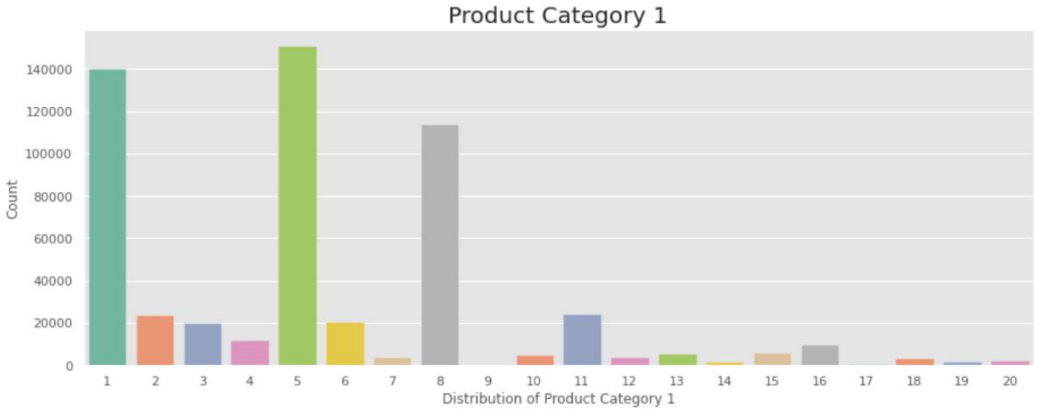

Fig 3.19

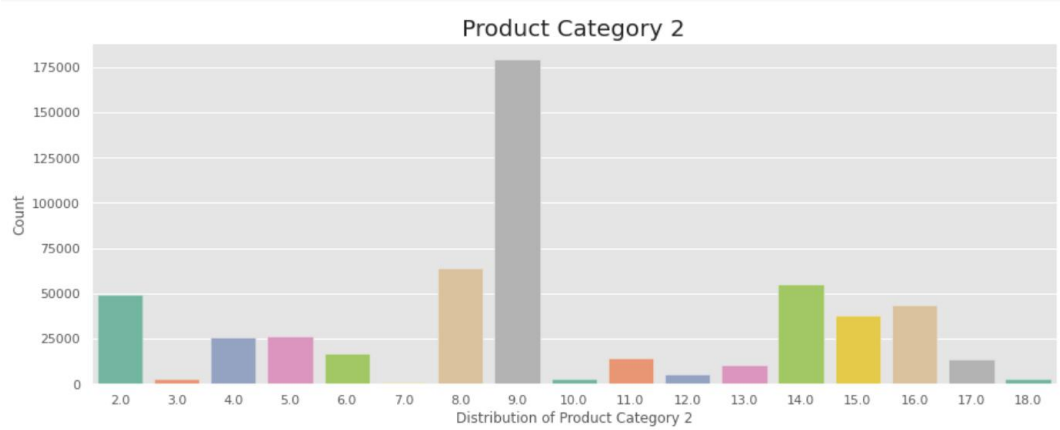

Fig 3.20

As it clear from the Fig 3.19 and Fig 3.20 that Product Category 1 have got more products of labels 1, 5, and 9 and Product Category 2 have got more products of label 9.

Purchase

The purchase attribute which is our target variable

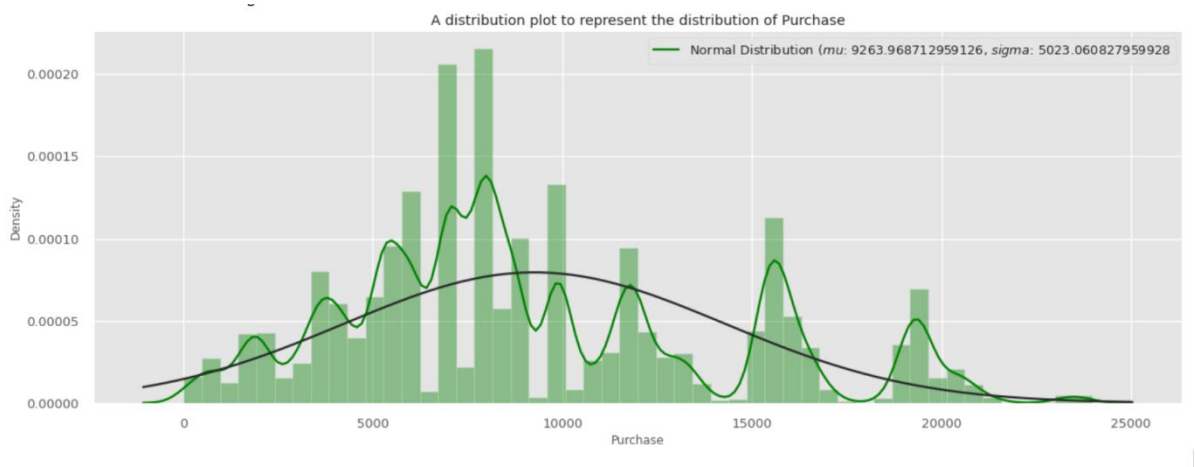

Fig 3.21

Above Fig 3.21 depicts the normal distibution of the purchase variable

CHAPTER-4

RESULTS AND DISCUSSION

As mentioned in the Proposed system we are implementing the machine learning models such as Ridge Regression, Elastic Net Regression, Lasso Regression, and Gradient Boosting Regression.

➢ Ridge Regression

Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values.

$$\text{Error}_{(m,b)} = \frac{1}{N}\sum_{i=1}^{N}(y_i - (mx_i + z))^2$$

$$\text{Subject to} \quad \sum_{i=1}^{p}(mx_i + z)^2 \leq s$$

Where ‘s’ is constrained value, it penalizes the bigger coefficient and therefore manages to shrink the biases accordingly in order to make proportionate with variance. This regularization is also known as L2 regularization.

$$\text{Ridge} = \frac{1}{N}\sum_{i=1}^{N}(y_i - (mx_i + z))^2 + \lambda \sum_{i=1}^{p}(mx_i + z)^2$$

```
from sklearn.linear_model import Ridge
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error

from math import *

model = Ridge()
model.fit(x_train, y_train)

y_pred = model.predict(x_test)


# finding the mean_squared error
mse = mean_squared_error(y_test, y_pred)
print("RMSE Error:", np.sqrt(mse))

# finding the r2 score or the variance
r2 = r2_score(y_test, y_pred)
print("R2 Score:", r2)
```

```
RMSE Error: 4703.496079726382
R2 Score: 0.12736140409143537
```

Fig 4.1

➢ Elastic Net Regression

Linear regression is the standard algorithm for regression that assumes a linear relationship between inputs and the target variable. An extension to linear regression involves adding penalties to the loss function during training that encourage simpler models that have smaller coefficient values. These extensions are referred to as regularized linear regression or penalized linear regression.

Elastic net is a popular type of regularized linear regression that combines two popular penalties, specifically the L1 and L2 penalty functions.

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - (mx_i + z))^2 + \lambda \sum_{i=1}^{p}(mx_i + z)^2 + \lambda \sum_{i=1}^{p}(mx_i + z)$$

```
from sklearn.linear_model import ElasticNet
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error

from math import *

model = ElasticNet()
model.fit(x_train, y_train)

y_pred = model.predict(x_test)


# finding the mean_squared error
mse = mean_squared_error(y_test, y_pred)
print("RMSE Error:", np.sqrt(mse))

# finding the r2 score or the variance
r2 = r2_score(y_test, y_pred)
print("R2 Score:", r2)
```

```
RMSE Error: 4737.961939753273
R2 Score: 0.11452566055708235
```

Fig 4.2

➢ Lasso Regression

Lasso stands for Least Absolute Shrinkage Selector Operator, Lasso regression performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients. This type of regularization can result in sparse models with few coefficients; Some coefficients can become zero and eliminated from the model. Larger penalties result in coefficient values closer to zero, which is the ideal for producing simpler models. On the other hand, L2 regularization (e.g. Ridge regression) doesn't result in elimination of coefficients or sparse models. This makes the Lasso far easier to interpret than the Ridge.

$$\text{Lasso} = \frac{1}{N}\sum_{i=1}^{N}(y_i - (mx_i + z))^2 + \lambda \sum_{i=1}^{p}(mx_i + z)$$

```
from sklearn.linear_model import Lasso
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error

from math import *

model = Lasso()
model.fit(x_train, y_train)

y_pred = model.predict(x_test)


# finding the mean_squared error
mse = mean_squared_error(y_test, y_pred)
print("RMSE Error:", np.sqrt(mse))

# finding the r2 score or the variance
r2 = r2_score(y_test, y_pred)
print("R2 Score:", r2)
```

```
RMSE Error: 4703.504740643469
R2 Score: 0.12735819037238216
```

Fig 4.3

➢ Gradient Boosting Regression

Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees. When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest. A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods, but it generalizes the other methods by allowing optimization of an arbitrary differentiable loss function.

```
y(pred) = y1 + (eta *  r1) + (eta * r2) + ..... + (eta * rN)
```

```
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error

from math import *

model = GradientBoostingRegressor(n_estimators = 100, max_depth = 5, min_samples_split = 2, learning_rate = 0.1)
model.fit(x_train, y_train)

y_pred = model.predict(x_test)

# finding the mean_squared error
mse = mean_squared_error(y_test, y_pred)
print("RMSE Error:", np.sqrt(mse))

# finding the r2 score or the variance
r2 = r2_score(y_test, y_pred)
print("R2 Score:", r2)
```

```
RMSE Error: 2965.186885509144
R2 Score: 0.653185708142229
```

Fig 4.4

Comparative Analysis

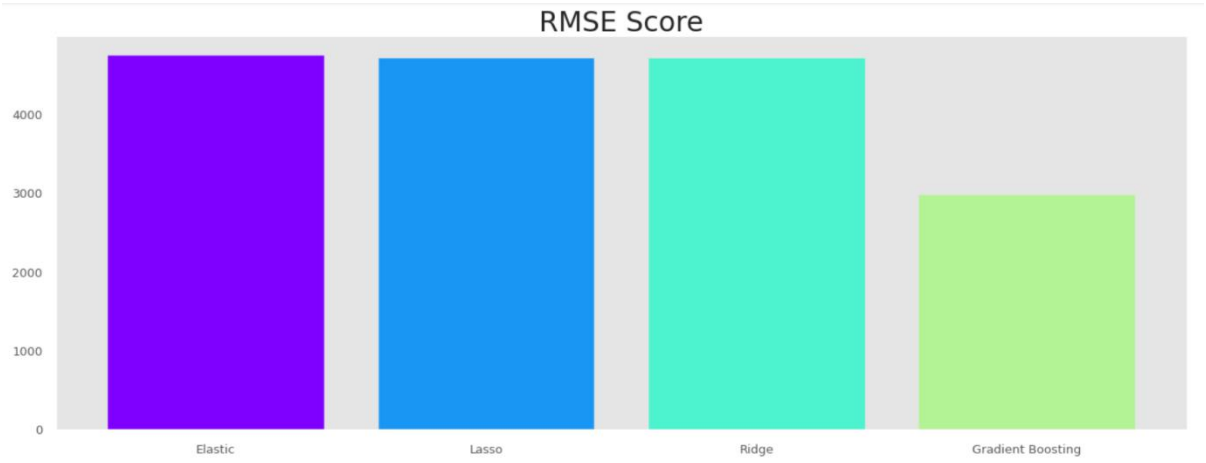

Fig 4.5

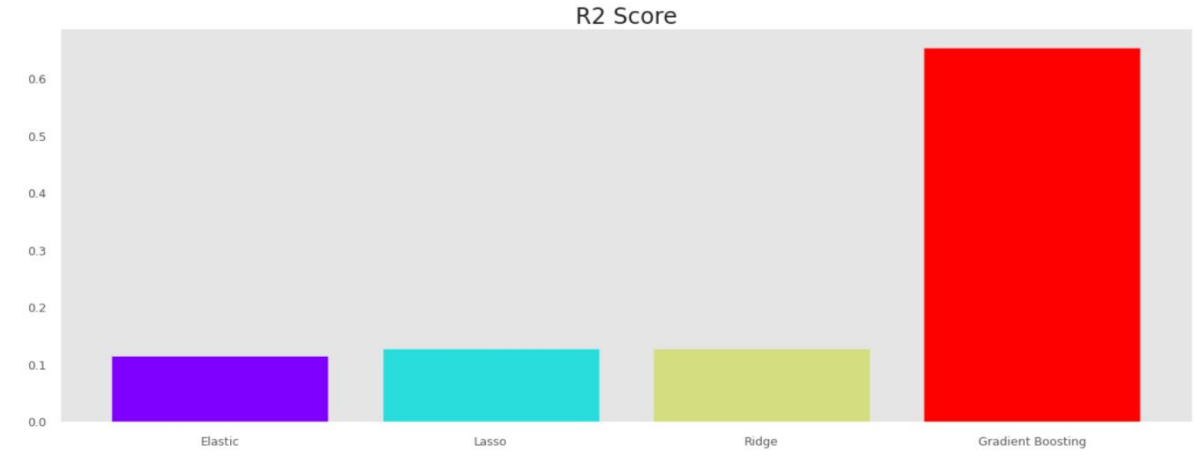

Fig 4.6

Above Fig 4.5 and Fig 4.6 depicts the comparative analysis of RMSE score and R2 score of various Regression models respectively.

The comparison between the RMSE scores of all algorithms is depicted in Table 4.1 below. Based on Table it can be observed that Gradient boosting regression gives better performance with comparison to other machine learning models namely Ridge regression, Elastic net regression and Lasso regression. The RMSE score of Gradient boosting regression is 2965.18 and it has more R2 score than other models and hence it is more suitable for the prediction model to be implemented.

Model	**RMSE Score**	**R2 Score**
Ridge Regression	4703.49	0.127
Elastic Net Regression	4737.96	0.114
Lasso Regression	4703.50	0.127
Gradient Boosting Regression	**2965.18**	**0.653**

Table 4.1

So we predict Purchase attribute for testing dataset using Gradient Boosting Regression algorithm as shown in fig 4.7

	A	B	C
1	Purchase	User_ID	Product_ID
2	15763.131	1000004	P00128942
3	11025.337	1000009	P00113442
4	6586.844	1000010	P00288442
5	2768.9932	1000010	P00145342
6	2822.8062	1000011	P00053842
7	11625.834	1000013	P00350442
8	13061.571	1000013	P00155442
9	10370.924	1000013	P0094542
10	18630.652	1000015	P00161842
11	5666.052	1000022	P00067942
12	12671.411	1000026	P00046742
13	5729.067	1000026	P00040042
14	5997.07	1000026	P00196542
15	6166.648	1000026	P00004542
16	17631.525	1000028	P00159542
17	14322.795	1000029	P00111542
18	14893.044	1000033	P00121042
19	6363.9604	1000033	P00344442
20	6657.9697	1000034	P00265242
21	10575.561	1000035	P0096642
22	6720.474	1000036	P00303042
23	14688.489	1000036	P00059642
24	14059.728	1000042	P00030842

Fig 4.7

CHAPTER-5

CONCLUSION AND FUTURE WORK

Conclusion

Indeed, it will not be quite the same as in earlier years. Covid pandemic shopping patterns power retailers to reexamine Christmas shopping. The day after Thanksgiving 2021 should, in any case, carry a lot of promotions and offers. We do expect that there to be a couple of changes in shopping trends because of the pandemic. We may see numerous stores are forced to dodge in-store traffic because of pandemics, clearing a path for some arrangements to be accessible on the web. We may see numerous pre–Black Friday deals to begin in October. Retailers are quick to make a sprinkle this occasion period to compensate for lost income over the pandemic. Stores are protracting the shopping season, stretching out Black Friday offers to online customers.

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.

Future Work

As future research, we can perform hyperparameter tuning and apply different machine learning algorithms.

CHAPTER-6

REFERENCES

[1] 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), 2016, pp. 78-83, doi: 10.1109/IACC.2016.25.

[2] Purvika Bajaj1, Renesa Ray2, Shivani Shedge3, Shravani Vidhate4, Prof. Dr. Nikhilkumar Shardoor5,"SALES PREDICTION USING MACHINE LEARNING ALGORITHMS",International Research Journal of Engineering and Technology (IRJET) ,Vol 7 Issue 6,2020,eISSN: 2395-0056 | p-ISSN: 2395-0072

[3] Ramasubbareddy S., Srinivas T.A.S., Govinda K., Swetha E. (2021) Sales Analysis on Back Friday Using Machine Learning Techniques. In: Satapathy S., Bhateja V., Janakiramaiah B., Chen YW. (eds) Intelligent System Design. Advances in Intelligent Systems and Computing, vol 1171. Springer, Singapore. https://doi.org/10.1007/978- 981-15-5400-1_32

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

[5] M.Sahaya Vennila; Holy Cross College, Nagercoil. Affiliated to Manonmaniam Sundaranar University, Tirunelveli – 627 012,Page No:133-136,doi.org/10.37896/whjj16.05/037

[6] Black Friday Sales Dataset Kaggle https://www.kaggle.com/kkartik93/black-friday-salesprediction?select=train.csv