CUSTOMER PURCHASE PREDICTION

Using Random Forest Regression.

Developed by: Spandana J, Meghana M, Sanjana G P,

Lanchana Gudami , K S Meghana

Smart Bridge-Remote Summer Internship Program

1.INTRODUCTION

The purpose of this study was to observe and analyze the consumer behaviors of the Black Friday customer. Black Friday, the day after Thanksgiving, is a term used by the retail industry in the United States that signifies the start of the Christmas holiday shopping season. Thanksgiving Day is celebrated on the fourth Thursday of November; therefore, the holiday shopping season runs from the Friday after Thanksgiving Day and continues until December 24, the day before Christmas. Black Friday is not considered an official national holiday; however, many employees have Thanksgiving Day off along with the following day, which increases the number of potential shoppers on that Friday. The origin of Black Friday is based on an accounting term when records were kept in ink with red signifying a loss in profits and black signifying a profit. Retailers generally operate in the red (unprofitable) throughout the year and depend heavily on the holiday season sales to end the year in the black with a profit (Black Friday, n.d.).

With the “hype” of Black Friday, customers are exposed to a retail environment that can stimulate frustration and aggression. Black Friday is traditionally known for long lines with customers waiting outdoors in cold weather waiting for the store to open, confusion and chaos of customers once the retail doors are opened for business, heavily crowded stores, a limited amount of products available at a reduced price, long checkout lines, and the lack of availability of advertised sale products.

1.1 Overview

In addition, the number of purchase in sales sector is rapidly growing and huge data volumes are available which represent the customers behavior. Data Mining is one of the most motivating and vital area of research with the aim of extracting information from tremendous amount of accumulated data sets. Here a new model for classifying purchase in sales sector is obtained by using Machine Learning concepts. The model has been built using data form sales sector to predict the status of purchase. Four algorithms have been used to build the proposed model: Random Forest Regression, Linear Regression, Decision Tree, Multilinear regression. By using the algorithm, a Flask model has been implemented and tested. The results have been discussed and a full comparison between algorithms was conducted. Random forest regression was selected as best algorithm based on accuracy.

1.2 Purpose

Our aim from the project is to make use of pandas, matplotlib, & seaborn libraries from python to extract the libraries for machine learning for the purchase amount prediction.

Secondly, to learn how to hyper tune the parameters and search cross validation for the Random Forest Regression machine learning algorithm.

And in the end, to predict whether the purchase amount of applicant depends on occupation, gender etc using the above techniques of combining the predictions from multiple machine learning algorithms and withdrawing the conclusions.

2.LITERATURE SURVEY

Data mining is the process of analyzing data from different perspectives and extracting useful knowledge from it. It is the core of knowledge discovery process. The various steps involved in extracting knowledge from raw data as depicted in figure-1. Different data mining techniques include classification, clustering, association rule mining, prediction and sequential patterns, neural networks, regression etc. Classification is the most commonly applied data mining technique, which employs a set of preclassified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to classification technique. This approach frequently employs Decision tree based classification algorithm. In classification, a training set is used to build the model as the classifier which can classify the data items into its appropriate classes.

A test set is used to validate the model.

2.1Existing Problem

A categorical instrument developed by the researchers was divided into three sections: in-line observations prior to the store opening, store entry observations, and individual customer observations. Each instrument also provided an open-ended item for observers to record general consumer comments heard in-line and in the store. Operational definitions were developed by the researchers by adapting previous research by Ekman (1992), Gottman, et al. (1995), and Lee and Dubinsky (2003). After receiving university approval for the instrument, a pilot test was conducted.Nine items were developed for the in-line observation and measured the approximate number of customers waiting for the doors to open, the overall emotions and behaviors exhibited by the customers, the presence and absence of shopping companions, and customer demographics. Cronbach's alpha, a coefficient of reliability, was computed for the in-line section items. Results revealed an alpha coefficient of .81, indicating that the section had good reliability.

Seventeen individual items measured the emotions and behaviors of customers as a group upon store entry. Observers circled the approximate percentage of customers that exhibited each behavior under question (e.g., courteous, punching, tripping, cart bumping). Percentages were presented to the observers as categories; each category represented 10% of the consumers (e.g., 0-10%, 11-20%, 21-30%). Other items in the "store entry" section of the instrument assessed the number of customers who fell down or were injured, the length of time the chaos upon store opening lasted, and the products sought by customers once inside the store. Products sought by consumers were measured in the same manner that behaviors were, that is, using categories of percentages. Cronbach's alpha was .73 for the store entry observations section. Individual consumer behavior instrument items included gender, age, and the emotions and behaviors exhibited by the individual shopper. Observers circled emotions and behaviors that individual shoppers displayed from a list of 15 potential emotions and 20 potential behaviors. Cronbach's alpha was .70 for the individual customer behavior items. Undergraduate and graduate students serving on the observation team attended two three-hour training sessions. Each observer selected her or his own retail store as an observation site. Upon arrival at the store, observers documented in-line customer behavior. When the doors to the retail store were opened, observers documented the emotions and behavior of consumers upon store entry. After the "rush" during store entry, the observers selected one consumer and documented her or his emotions and behavior. Observers continued recording until ten observations of individual consumer emotions and behavior were completed.

2.2 Proposed Solution

Machine Learning (Random Forest Regression):

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of classes(classification) or mean prediction(regression) of the individual trees. Bagging in the random forest method, involves training each decision tree on a different data sample where sampling is done with replacement.

And also we have created an UI using the Flask for the purchase amount status prediction, this UI will allow the users to predict the purchase status very easily and the User interface is user friendly not at least one complication in using the interface, and it can be used just by entering some necessary details into the UI in real time it'll give the predicted value like how much the customer spends on the product ,do the occupations of the people have any impact on sales and which age group is the highest spender.

Basically this model will give the predicted value when a customer with details will purchase a product by just taking some necessary details of the customer in real time, and those details will be collected by sales person within minutes.

3.THEORETICAL ANALYSIS

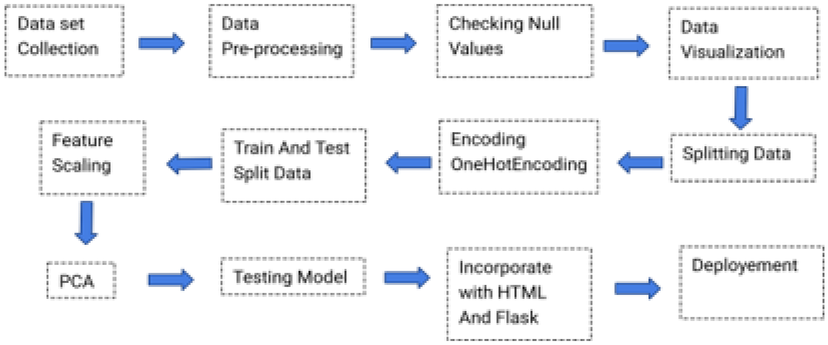
While selecting the algorithm that gives an accurate prediction we gone through lot of algorithms which gives the results abruptly accurate and from them we selected only one algorithm for the prediction problem that is Random Forest Regression, it assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature that is how the prediction work great with the Random Forest Regression Algorithm.

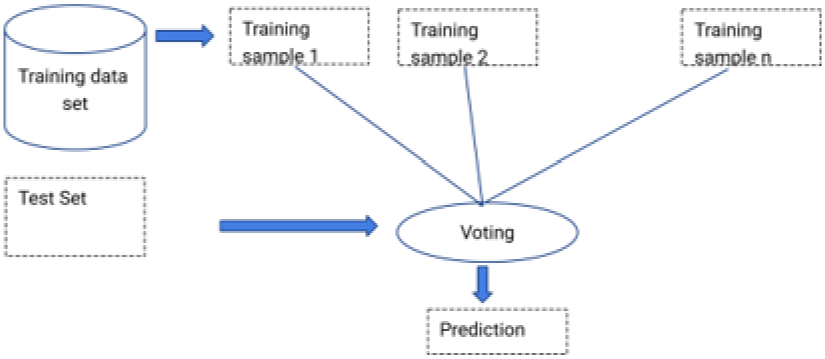
The peculiarity of this problem is collecting the customers details real time and working with the prediction at the same time, so we developed an user interface for the people who'll be accesssing for the purchase amount prediction. Accuracy is defined as the ratio of the number of samples correctly classified by the classifier to the total number of samples for a given test data set. The formula is as follows

$$\text{Accuracy}=TP+TN/TP+TN+FT+FN$$

At first we got like lot of worst accuracies because we tried lot of algorithms for the best accurate algorithm, finally after all of that we tried the best suitable algorithm which gives the prediction accurately is Random Forest Regression and developed it to use as a real time prediction probelm for the purchase amount prediction.

3.1 Block Diagram





3.2 Software Designing

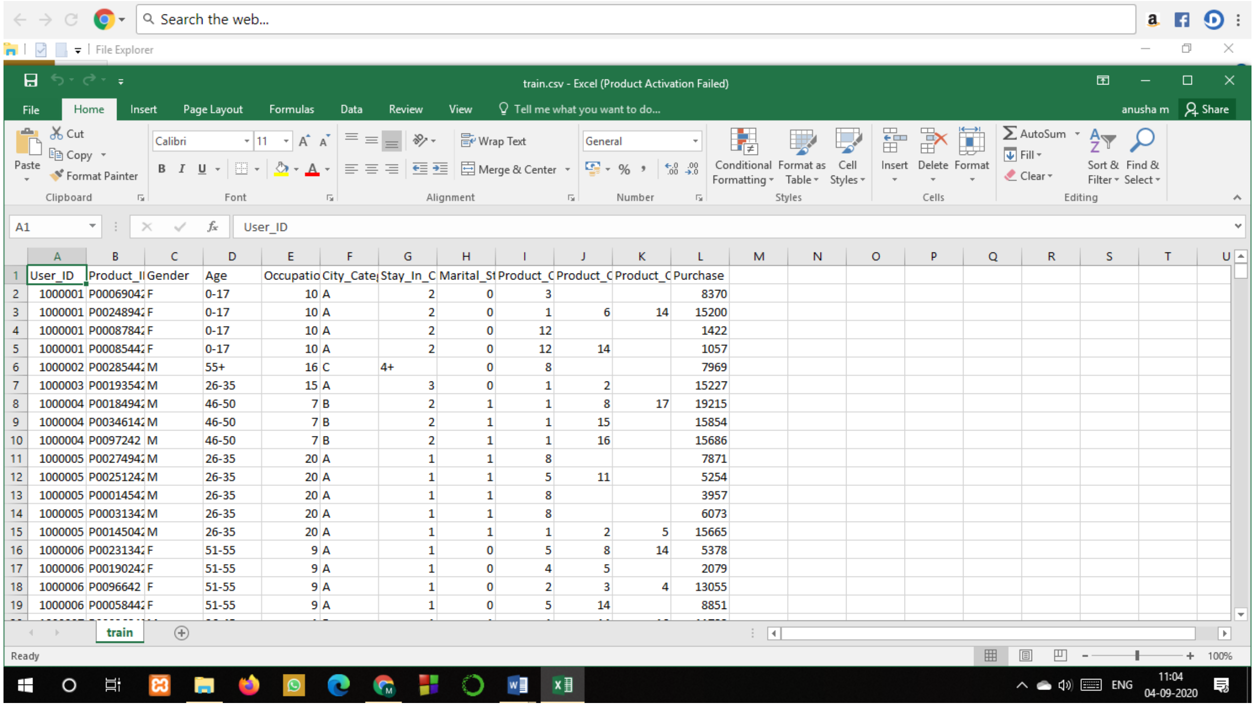
1. Jupyter Notebook Environment
2. Spyder IDE
3. Machine Learning Algorithms
4. Python (pandas, numpy, matplotlib, seaborn, sklearn)
5. HTML
6. Flask

We developed this purchase amount prediction by using the Python language which is a interpreted and high level programming language and using the Machine Learning algorithms for coding we used the Jupyter Notebook environment of the Anaconda distributions and the Spyder, it is an integrated scientific programming in the python language.

For creating an user interface for the prediction we used the Flask. It is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions, and a scripting language to create a webpage is HTML by creating the templates to use in the functions of the Flask and HTML.

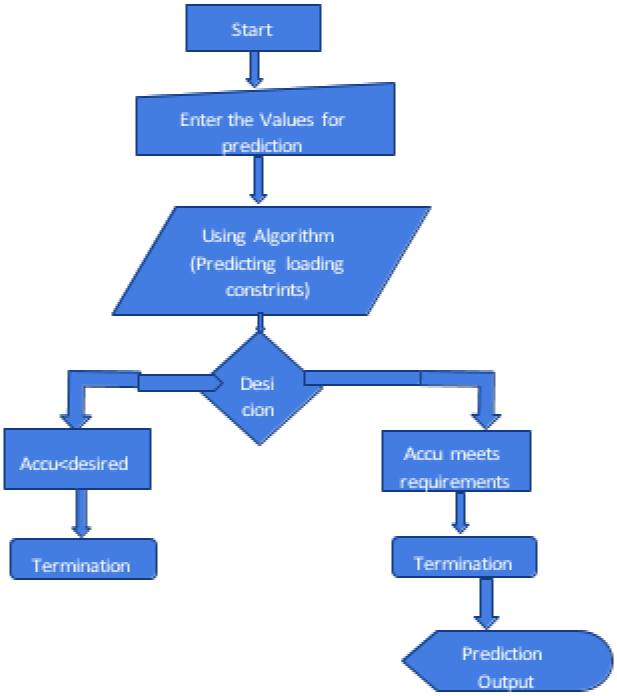
4. EXPERIMENTAL INVESTIGATION

In this paper, the dataset we used is derived from https://github.com/shwetachandel/Black-Friday-Dataset It contains more than 550070 users with 12 attributes. After that, the missing values are filled in by means of mean interpolation, and the duplicate or meaningless attributes are deleted, finally we have retained to 11 attributes. Those attributes were shown below in the screenshot of the data set we used.



	A	B	C	D	E	F	G	H	I	J	K	L
1	User_ID	Product_I	Gender	Age	Occupatio	City_Cate	Stay_In_C	Marital_St	Product_C	Product_C	Product_C	Purchase
2	1000001	P00069042	F	0-17	10	A	2	0	3			8370
3	1000001	P00248942	F	0-17	10	A	2	0	1	6	14	15200
4	1000001	P00087842	F	0-17	10	A	2	0	12			1422
5	1000001	P00085442	F	0-17	10	A	2	0	12	14		1057
6	1000002	P00285442	M	55+	16	C	4+	0	8			7969
7	1000003	P00193542	M	26-35	15	A	3	0	1	2		15227
8	1000004	P00184942	M	46-50	7	B	2	1	1	8	17	19215
9	1000004	P00346142	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	P00274942	M	26-35	20	A	1	1	8			7871
12	1000005	P00251242	M	26-35	20	A	1	1	5	11		5254
13	1000005	P00014542	M	26-35	20	A	1	1	8			3957
14	1000005	P00031342	M	26-35	20	A	1	1	8			6073
15	1000005	P00145042	M	26-35	20	A	1	1	1	2	5	15665
16	1000006	P00231342	F	51-55	9	A	1	0	5	8	14	5378
17	1000006	P00190242	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	P00058442	F	51-55	9	A	1	0	5	14		8851

5. FLOWCHART



6. RESULT

In this paper, the Random Forest Regression algorithm is used to predict its performance, and compared with another three machine learning methods namely the decision tree, simple linear regression, multi linear regression. The obtained results are displayed in Table below. The results show that, the performance of decision tree and Random forest regression have comparable performance than that of linear regression and multi linear regression , but random forest regression still performs the best, with an accuracy of 70%, higher than the decision tree with an accuracy of 43%.

ALGORITHM	ACCURACY
1.Simple linear regression	13.84 %
2.Multi linear regression	13.84 %
3.Decision tree	43.26 %
4.Random forest regression	70.14 %

7. ADVANTAGES AND DISADVANTAGES

Advantages:

- **Average Session Time** will show how long consumers spend on our website. This is an effective metric for analysing customer behavior for measuring the number of pages (or content) being viewed by shoppers in every visit. Based on this metric, you can identify the most (or least) viewed website pages and work on their respective strengths and weaknesses.
- **Traffic flow** is an efficient metric in monitoring how consumers move (or navigate) through your store pages. It indicates the online store pages that are most attractive to shoppers. Through this data, you can design the best navigation path for shoppers to reach your most popular products or product categories.
- **Customer Loyalty** is the one from which it is through free shipping or freebies, it is an excellent yardstick for observing customer behavior. And it can help you to track the buying habits of each shopper and understand the merchandise preferred by each demographic group.
- And also one is able to know correctly the factors which influence buying decisions of the consumer one can promote sales of existing or new product. The scheme of buying back old items has helped a lot in pushing sales.
- **Helps in Development of New Products** that means before launching a new product proper study of consumer tastes i.e. behaviour avoids later failure and loss. In certain cases, if a product is reintroduced after a long gap this type of study helps.

Disadvantages:

- **Inconsistency** One of the biggest drawbacks of relying too heavily on consumer buying behavior is that consumers rarely apply the same steps in the same way for every product and service purchase. This makes it more difficult for marketers trying to stimulate a need or to offer messages that enhance the likelihood of a purchase for their brand. Thus, most companies have to perform more research into their particular market segments and how they approach their brand.

- **Limited Buyer Interest** Another primary limitation for marketers using the consumer buying behavior model is that consumers sometimes are much less involved in a purchase decision. For instance, someone buying laundry detergent is generally less involved in the purchase than someone buying a car or washer and dryer. Thus, the ability of marketers to affect consumers by analyzing buyer behavior is limited. Consumers that are less involved spend less time seeking or viewing information about the purchase.

8. APPLICATION

1. ANALYSING MARKET OPPORTUNITY:

Consumer behaviour study helps in identifying the unfulfilled needs and wants of consumers. This requires examining the trends and conditions operating in the marketplace, consumers lifestyles, income levels and emerging influences.

2. SELECTING TARGET MARKET :

A review of market opportunities often helps in identifying distinct consumer segments with very distinct and unique wants and need. Identifying these groups, learning how they behave and how they make purchase decisions enables the marketer to design and market products or services particularly suited to their wants and needs.

3. MARKETING MIX :

Once unsatisfied needs and wants are identified, the marketer has to determine the right mix of product, price, distribution and promotion. Here too, consumer behaviour study is very helpful in finding answers to many perplexing questions.

4. USE IN SOCIAL AND NON-PROFITS MARKETING :

Consumer behaviour studies are useful to design marketing strategies by social, governmental a not-for-profit organisations to make their programmes such as family planning, awareness about AIDS, crime against women, safe driving, environmental concerns and other more effective.

9. CONCLUSION:

The study of consumer behaviour basically is to mould consumer behaviour and decisions by marketing and to avoid failure of their product, promote new products and for sales promotion. The science at times is misused and to protect consumers there are a number of enactments both in India and other countries. Consumer behavior analysis has emerged as an important tool to understand customers. By looking into consumer psychology and the forces behind customer buying behavior, companies can craft new products, marketing campaigns and increase profitability. Companies should talk to consumers,and identify their needs and expectations.

10. FUTURE SCOPE:

- **Demand Forecasting:** Estimating the demand for products and services.
- **Marketing:** Understanding the needs, expectations, problems of consumers, Formulating Marketing Mix Strategies.
- **Advertising:** Understanding human behaviour towards different advertising appeals and message, selecting the type of media.
- **Human Behaviour:** Understanding the various motives that influence behaviour of a consumer
- **Operations:** Formulating production, pricing and distribution policies

11. BIBLIOGRAPHY:

1. Industrial and economic planning division of TPCO.

2. www.rncos.com/Report/IM03.htm visited on 16/12/2011

3. Prabhakar Sinha, The Times of India, Jul 24, 2013.

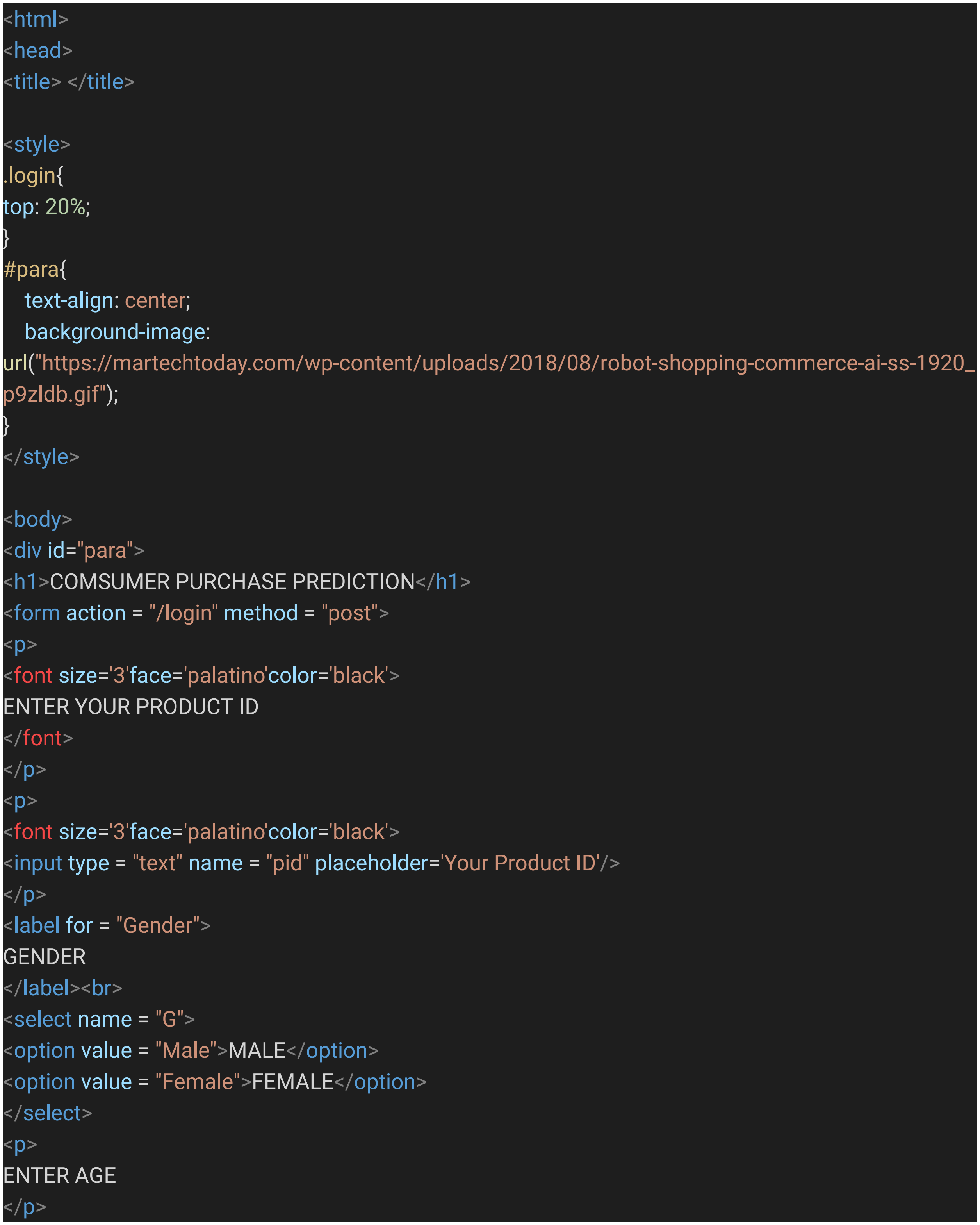
4. Moslehuddin Chowdhury Khaled, Tasnim Sultana, Sujan Kanti Biswas, Rana Karan, (2012). Real Estate Industry in Chittagong (Bangladesh): A Survey on Customer Perception and Expectation, Developing Country Studies ISSN 2224-607X (Paper) ISSN 2225-0565 (Online) Vol 2, No.2.

5. Tan Teck-Hong, (2012). Housing satisfaction in medium- and high-cost housing: The case of Greater Kuala Lumpur, Malaysia, Habitat International 36 108-116.

6. James Kottarapalli, (2012). Easy Living, Parpidam- Malayala Manorama

APPENDIX

HTML: index1.html



```
<html>
<head>
<title> </title>

<style>
.login{
top: 20%;
}
#para{
   text-align: center;
   background-image:
url("https://martechtoday.com/wp-content/uploads/2018/08/robot-shopping-commerce-ai-ss-1920_p9zldb.gif");
}
</style>

<body>
<div id="para">
<h1>COMSUMER PURCHASE PREDICTION</h1>
<form action = "/login" method = "post">
<p>
<font size='3'face='palatino'color='black'>
ENTER YOUR PRODUCT ID
</font>
</p>
<p>
<font size='3'face='palatino'color='black'>
<input type = "text" name = "pid" placeholder='Your Product ID'/>
</p>
<label for = "Gender">
GENDER
</label><br>
<select name = "G">
<option value = "Male">MALE</option>
<option value = "Female">FEMALE</option>
</select>
<p>
ENTER AGE
</p>
```

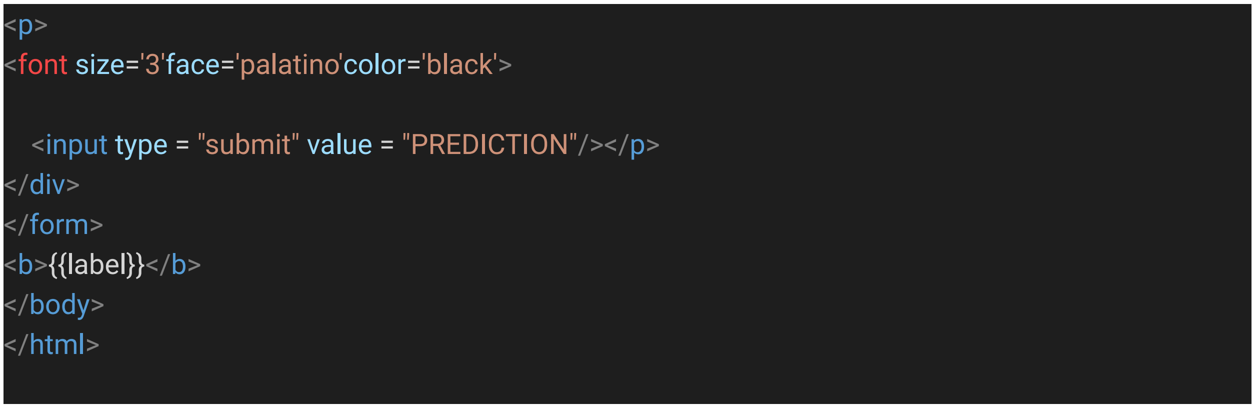
```html
<p>
<input type = "text" name = "age" placeholder='Your Age'/>
</p>
<p>
<font size='3'face='palatino'color='black'>
ENTER YOUR OCCUPATION CATEGORY IN RANGE
</p>
<p><input type = "text" name = "oc" placeholder='Enter Value'/></p>
<label for = "City Category">CITY CATEGORY</label><br>
<select name = "cc">
<option value = "1">1</option>
<option value = "2">2</option>
<option value = "3">3</option>
</select>

<p>
<font size='3'face='palatino'color='black'>
ENTER YOUR STAY IN CURRENT CITY IN YEARS
</p>
<p><input type = "text" name = "st" placeholder='No of years'/>
</p>
<label for = "Marital Status">MARITAL STATUS</label><br>
<select name = "ms">
<option value = "married">MARRIED</option>
<option value = "unmarried/divorced">UNMARRIED</option>
</select>

<p>
<font size='3'face='palatino'color='black'>
ENTER YOUR PRODUCT_CATEGORY_1
</p>
<p><input type = "text" name = "pca" placeholder='Enter category'/></p>

<p>
<font size='3'face='palatino'color='black'>
ENTER YOUR PRODUCT_CATEGORY_ 2 </p>
<p><input type = "text" name = "pcb" placeholder='Enter Category'/></p>

<p>
<font size='3'face='palatino'color='black'>
ENTER YOUR PRODUCT_CATEGORY_3</p>
<p><input type = "text" name = "pcc" placeholder='Enter Category'/></p>
```



```
<p>
<font size='3'face='palatino'color='black'>

  <input type = "submit" value = "PREDICTION"/></p>
</div>
</form>
<b>{{label}}</b>
</body>
</html>
```

APP.PY:

```
from flask import Flask , render_template , request
import pickle
app = Flask(__name__)
model = pickle.load(open('project.pkl','rb'))
@app.route('/')
def hello_world():
    return render_template('index1.html')
@app.route('/login', methods = ["POST"])
def login():
    pid = request.form["pid"]
    st = request.form["st"]
    age = request.form["age"]
    oc = request.form["oc"]
    pca = request.form["pca"]
    pcb = request.form["pcb"]
    pcc = request.form["pcc"]
    G = request.form["G"]
    if(G == "Male"):
        s2 = 0
    if(G == "Female"):
        s2 = 1
    cc = request.form["cc"]
    if(cc == "1"):
        c2,c3 = 0,0
    if(cc == "2"):
        c2,c3 = 1,0
    if(cc == "3"):
        c2,c3 = 0,1
    ms = request.form["ms"]
    if(ms == "married"):
        m2 = 0
    if(ms == "unmarried/divorced"):
        m2 = 1

total=[[int(pid),s2,int(age),int(oc),c2,c3,int(float(st)),m2,int(float(pca)),int(float(pcb)),int(float(pcc))]]
    p = model.predict(total)
    p = p[0][0]
    c = "THE PURCHASE IS = "+str(p)
    return render_template('index1.html',label = c)
if __name__=='__main__':
    app.run(debug = True)
```