



ISSN: 2454-9940



**INTERNATIONAL JOURNAL OF APPLIED
SCIENCE ENGINEERING AND MANAGEMENT**

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

Using Machine Learning to Forecast Black Friday Sales

¹ B. Kethana, ² K. Shruthi,

¹Assistant Professor, Megha Institute of Engineering & Technology for Women, Ghatkesar.

² MCA Student, Megha Institute of Engineering & Technology for Women, Ghatkesar.

Abstract—

Making sense of how different types of consumers use their demographic data (IS characteristics), the majority of which are self-explanatory, to make purchases of various items (dependent variable). Unstructured data, duplicate entries, and null values make up this dataset. The retail sector makes extensive use of machine learning. With this idea, we can create a predictor that will be of great use to store owners in terms of advertising, marketing, financial planning, and inventory management. Everything from preprocessing to modeling to training to testing and evaluation is a part of constructing a model. Therefore, in order to automate some of this process and make it simpler, frameworks will be created. Our suggested Random Forest regressor approach achieved an average accuracy of 83.6% and a minimal Root Mean Squared Error (RMSE) score of 2829 on the tire Black Friday sales dataset.

Analysis of Data, Black Friday, Sales Forecasting, Random Forest Regressor, Testing, and Training are Index Terms.

I. INTRODUCTION

"Black Friday" is the moniker given to the post-Thanksgiving shopping day. Many shoppers were involved in car accidents and even acts of violence on this day, leading some to dub it "Black Friday" [1, 2]. The chaos caused by the heavy foot and vehicle traffic in downtown retail areas is what the police used to coin the term. A company's ability to turn a profit or loss is heavily dependent on the volume of sales in the retail sector. Efficient industrial management is made possible with precise sales forecasting. In the United States, Black Friday is akin to a carnival sale. Today, there is a massive sale on highly sought-after items, and the prices are really low. In order to generate sales, a prediction model is developed to focus on the product type that sells the most. The goal of analyzing a client's behavior is to forecast how much money the customer will spend on a given day. The purpose of this article is to forecast a business's sales on "Black Friday" [3]. We need to examine the link between many factors and arrange the information properly if we want to forecast product sales using their independent variables. In order for a model to calculate well and provide reliable sales predictions. Section A: Goal The focus of this work is on two goals. This is a list of them: 1. Examining all client data to determine the link between independent factors and the target variable. 2. Using testing and training to forecast sales. Algorithm II An example of a predictive modeling approach, regression analysis looks at the connection between an independent variable (the target) and a dependent variable (the predictor) [4]. Time series discovery, prediction, and modeling of the variables' cause effect connection are all possible with this method. It all depends on three metrics. First, the kind of dependent variable, second, the total number of independent variables, and third, the form of the regression line. To forecast the quantity to be bought. We have tested and evaluated many machine learning algorithms based on performance metrics and accuracy. Since this is a regression issue, the RMSE loss function is appropriate [5]. The initial section of our research included data pre-processing estimations; this provided the structured data that was then split into a testing set and a training set for the purpose of checking the correctness. Fig. 1 depicts the data flow diagram. Here we detail some of the methods and equipment that were crucial in the completion of the project. Keep in mind that for machine learning algorithms to work, the dataset has to be balanced, meaning that there should be an equal number of characteristics and samples for each class. Otherwise, when training the prediction model, it will be slanted toward any such category, and the forecast would reflect that.

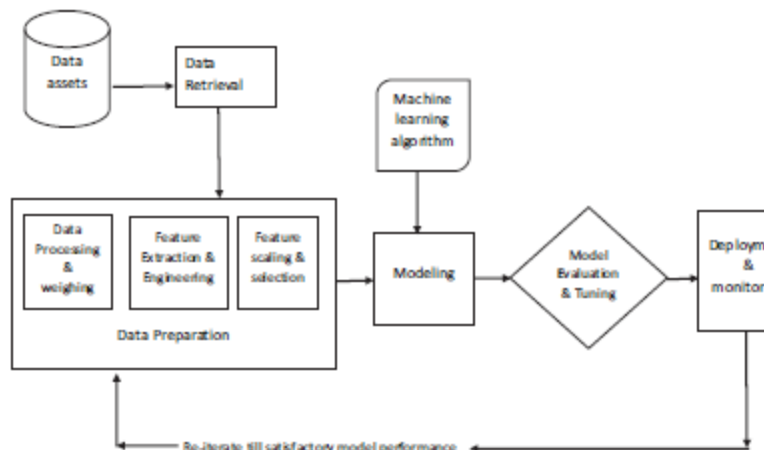


Fig. 1. Data Flow Architecture

Part A. Instruments and Programs Here we show the machine learning algorithms [6] and methods that were employed in this study. 1) A Regressor With a Random Forest A Random Forest is a hybrid approach that combines the features of several decision trees with the Bootstrap Aggregation methodology, often referred to as bagging, to accomplish both classification and regression tasks [7], [14]. The basic idea is to use a combination of decision trees to determine the final outputs rather than relying on just one. The following are some of the characteristics of Extreme gradient boosting: i. XGBoost is Sparse Aware, therefore it automatically handles missing data values. (part two). The building of trees may be supported in a parallel fashion. subject (iii). The fitted model is the result of ongoing training and may be further improved with more datasets. There are two sets of parameters used by the models: training and test. Here, we shall forecast sales using the test set. One additive model that uses judgments from many base models to forecast sales is the Random Forest (RF) model. Eq. (1) is the last equation in the RF model:

$$P(x)=f_0(x)+f_1(x)+f_3(x)+..... \quad (1)$$

Every model type has its own set of benefits. When dealing with numerical characteristics that have less than few categories or tabular data that include categorical information, the random forest model performs the best. Random forests are able to detect non-linear feature-target cooperation, in contrast to linear models. In Random Forest, trees run simultaneously. During construction, there is zero contact between the trees. The following is the code for the random forest algorithm: The first step is to choose some random numbers from the database. The second step in getting the prediction results is to build the decision trees for each sample. Third, cast your vote for each of the anticipated outcomes. Step 4: The ultimate forecast result is the one with the most votes. The data set is partitioned into two halves, with 80% used for testing and 20% for training. To standardize the value of each field, the usual scalar techniques are used. We make use of the maximum leaf node 900, which has 20 estimators. 2) Root-Mean-Squared Defect It represents the dispersion of the remaining forecasts. The residuals are used for the purpose of estimating the distance between the data points that make up the regression line. To get a sense of the dispersion of these residuals, one may use the root mean squared error [8]. We can learn how concentrated the data will be from this. Climate scientists, meteorologists, and forecasters often use root mean squared error.

$$RMSE_{r0}=[\sum_{i=1}^N(Z_{fi}-Z_{oi})^2/N]^{1/2} \quad (2)$$

Where \sum = Summation

$(Z_{fi}-Z_{oi})$ = differences squared

N= Sample size

3) Gradient Boosting Machines (GBM) have received an upgrade with the release of XGBoost Extreme Gradient Boost (XGBoost) [9], [13]. By optimizing the system using differentiable loss functions, it achieves efficient performance over the GBM framework.

TABLE I. MSE AND MAE VALUES COMPARISON OF ML ALGORITHMS

Model	MAE	MSE
Linear Regression	86.1	127
Decision Tree	41.6	068
Rihdge Regression	86	129
XGBoost	89.03	052

When it comes to structured data, XGBoost is an algorithm that produces better results than applied ML and kaggle contests. A performance- and speed-oriented implementation of a gradient-boosted decision tree. The Mean Squared Error (MSE) and Mean Absolute Error (MAE) of several ML algorithms are shown in Table 1.

II. EXPERIEMENTS

Using data pre-processing, univariate analysis, and bivariate analysis—all derived from machine learning and data mining methods [10], [16]—a series of experiments were carried out on the dataset. Afterwards, Random Forest Regressor was used for testing and training. This precise data set classification was achieved by means of all these model trials. A. Initial Steps Before we can apply any ML technique to our dataset, the data has to be pre-processed [10]. In addition, the data has to be transformed so that a machine learning algorithm can use consumer information to forecast the purchase variable's value. It deals with NaN values, ignores fields that aren't relevant for data analysis or prediction, and may substitute a numerical value with a category one. When some data in the dataset isn't suitable for prediction, it's important to change it so that the prediction analysis can proceed. In this case, it is necessary to modify the column age that contains values from multiple ranges [15]. B. Investigating Data In order to identify which qualities are present in the dataset, this step involves viewing the data prior to training. We are currently exploring the data to see whether the dataset is appropriate for training and to go on with the next stages. Part C: Cleaning the Data It may be necessary to remove missing values from a dataset. The dataset should be updated to include all missing values or removed if any of the values are not there. Inconsistencies in the outcome could be caused by the missing values. D. Analysis with One Variable There is no more basic method of statistical analysis than this. Because it just takes into account one variable, univariate analysis might lead to inaccurate conclusions. E. Non-Destructive Analysis To start, we looked at a few of the existing features separately. Next, we learn how each predictor relates to the target variable and how each predictor relates to each other. F. Analyzing Outliers Any data point in a dataset that deviates significantly from the norm is called an outlier. We use the boxplot technique to classify the outliers into the pre-existing class interval values. The G. Dataset Model has been trained to perfection by being fed exhaustive datasets for supervised learning. With 8523 observations and 12 attributes such as Product_category_1, Product_category_2, Product_category_3, Age, Gender, Occupation, Stay in Current City Years, Marital Status, and Purchase, this dataset is a treasure trove of information. We can see how different machine learning algorithms fare when we compare their accuracy and performance [11], [12]. This work's algorithms achieve a high level of accuracy. With a mean squared error score of 2829.09 and an accuracy of almost 81%, the Random Forest Algorithm was the most effective of the studies conducted in predicting Black Friday sales. You can get algorithms that are around 10% more accurate.

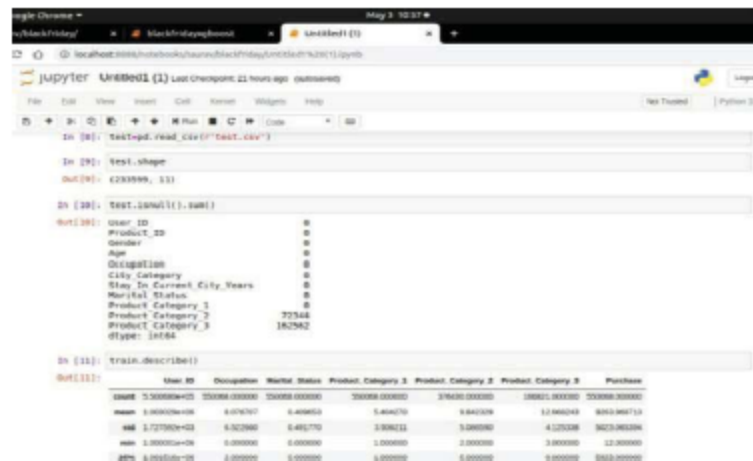
Name	Type	Subtype	Description	Segment	Expectation
User_ID	Numeric	Discrete	User ID	Customer	Low_Impact
Product_ID	Numeric	Discrete	Product ID	Product	Low_Impact
Gender	Categorical	Nominal	Sex of User	Customer	High_Impact
Age	Categorical	Ordinal	Age in bins	Customer	High_Impact
Occupation	Categorical	Nominal	Occupation (Masked)	Customer	Medium_Impact
City_Category	Categorical	Ordinal	Category of the city (A,B,C)	City	High_Impact
Stay_in_Current_City_Years	Categorical	Ordinal	Number of years stay in current city	City	Low_Impact
Marital_Status	Categorical	Ordinal	Marital Status	Customer	Low_Impact
Product_Category_1	Categorical	Nominal	Product Category (Masked)	Product	High_Impact
Product_Category_2	Categorical	Nominal	Product may belongs to other category also (Masked)	Product	Low_Impact
Product_Category_3	Categorical	Nominal	Product may belongs to other category also (Masked)	Product	Low_Impact
Purchase	Numeric	Continuous	Purchase Amount (Target Variable)	Product	NAN

Fig. 2. Schema of dataset

	User_ID	Occupation	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3	Purchase
count	5.375778e+05	537577.00000	537577.00000	537577.00000	370991.00000	164278.00000	537577.00000
mean	1.002992e+06	8.08271	0.408797	5.295546	9.842144	12.669840	9333.859853
std	1.714389e+03	6.52412	0.491612	3.750701	5.087259	4.324341	4981.022138
min	1.000001e+06	0.000000	0.000000	1.000000	2.000000	3.000000	185.000000
25%	1.001405e+06	2.000000	0.000000	1.000000	5.000000	9.000000	5866.000000
50%	1.003031e+06	7.000000	0.000000	5.000000	9.000000	14.000000	8062.000000
75%	1.004417e+06	14.000000	1.000000	8.000000	15.000000	16.000000	12071.000000
max	1.006040e+06	20.000000	1.000000	18.000000	18.000000	18.000000	23961.000000

Fig. 3. Dataset

Additionally, "if we analyze the User_ID columns using the unique method and it can be concluded that about 5891 customers have purchased something from that particular retail store on Black Friday," The extract information reveals that 3623 distinct goods were sold in the Product_ID category. Figure 3 illustrates this.



```

In [8]: test=pd.read_csv('test.csv')

In [9]: test.shape
Out[9]: (239999, 8)

In [10]: test.isnull().sum()
Out[10]: User_ID      0
Product_ID      0
Gender          0
Age            0
Occupation     0
City_Category  0
Stay_in_Current_City_Years  0
Marital_Status  0
Product_Category_1      72548
Product_Category_2      0
Product_Category_3     162582
dtype: int64

In [11]: train.describe()
Out[11]:
```

	User_ID	Occupation	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3	Purchase
count	5.375778e+05	537577.00000	537577.00000	537577.00000	370991.00000	164278.00000	537577.00000
mean	1.002992e+06	8.08271	0.408797	5.295546	9.842144	12.669840	9333.859853
std	1.714389e+03	6.52412	0.491612	3.750701	5.087259	4.324341	4981.022138
min	1.000001e+06	0.000000	0.000000	1.000000	2.000000	3.000000	185.000000
25%	1.001405e+06	2.000000	0.000000	1.000000	5.000000	9.000000	5866.000000

Fig. 4. Displaying the Null values for each attributes

Figure 4 shows the null values for all variables. Figure 5 displays the corresponding mathematical procedures performed on the data that is provided.

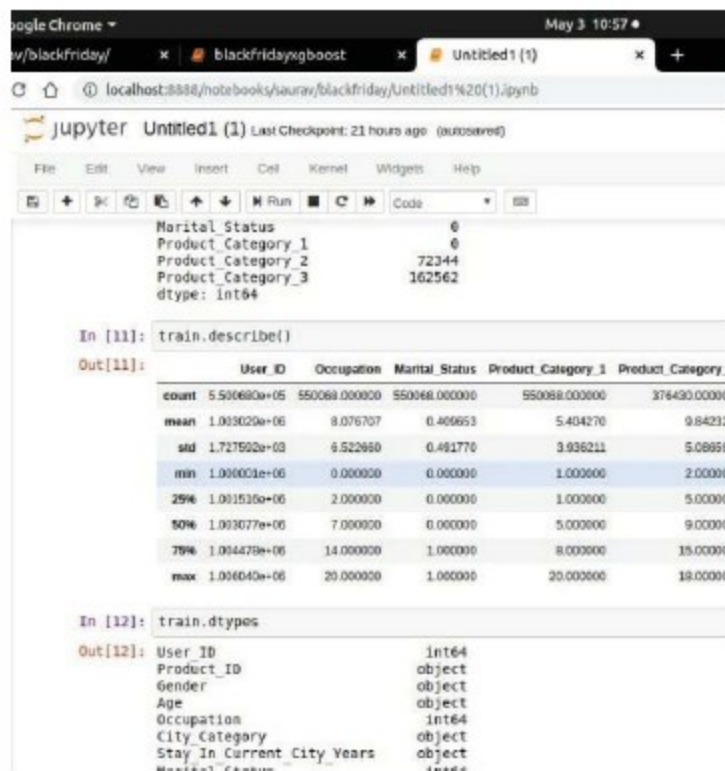


Fig. 5. Performing the mathematical operations on the data available

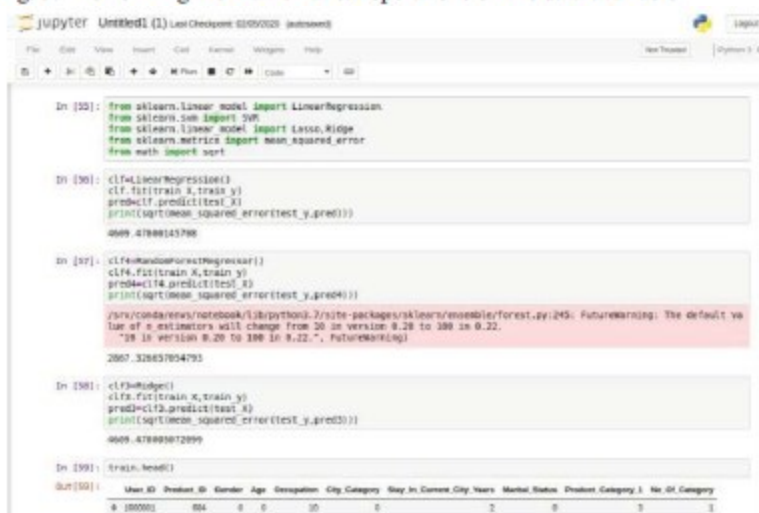
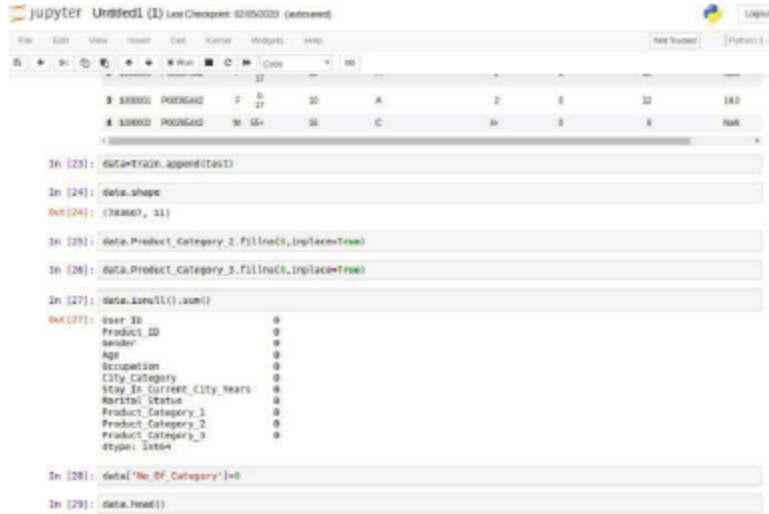


Fig. 6. Applying Models on the dataset



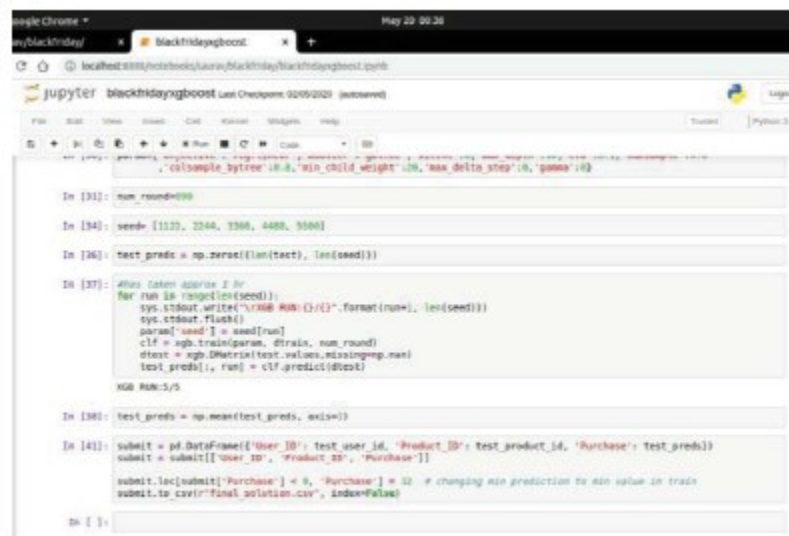
```

jupyter Untitled (1) Last Checkpoint: 02/05/2025 (autosaved)
File Edit View Insert Cell Kernel Help Python 3.7
In [23]: data = Train.append(test)
In [24]: data.shape
Out[24]: (788667, 11)
In [25]: data.Product_Category_2.fillna(0, inplace=True)
In [26]: data.Product_Category_3.fillna(0, inplace=True)
In [27]: data.isnull().sum()
Out[27]:
user_id      0
product_id    0
gender        0
age           0
occupation    0
city_category 0
stay_in_current_city_years 0
marital_status 0
product_category_1 0
product_category_2 0
product_category_3 0
dtype: int64
In [28]: data['No. Of_Category'] = 0
In [29]: data.head()

```

Fig. 7. Data Refinement

The dataset was processed by using the ML Models, as shown in Figure 6. The procedure for data refining is shown in figure 7.



```

Google Chrome
blackfriday
blackfridayxgboost
localhost:8888/notebooks/karavay/blackfriday/blackfridayxgboost.ipynb
jupyter blackfridayxgboost Last Checkpoint: 02/05/2025 (autosaved)
File Edit View Insert Cell Kernel Help Python 3.7
In [31]: num_round=100
In [32]: seed= [1122, 2244, 3366, 4488, 5500]
In [33]: test_preds = np.zeros((len(test), len(seed)))
In [34]:
for run in range(len(seed)):
    sys.stdout.write("\rXGB RUN: (%i)" % (run+1, len(seed)))
    sys.stdout.flush()
    param['seed'] = seed[run]
    clf = xgb.train(param, dtrain, num_round)
    dtest = xgb.DMatrix(test.values, missingnp.nan)
    test_preds[:, run] = clf.predict(dtest)
XGB RUN: 5/5
In [35]: test_preds = np.mean(test_preds, axis=1)
In [36]:
submit = pd.DataFrame({'user_id': test.user_id, 'product_id': test.product_id, 'purchase': test_preds})
submit = submit[['user_id', 'product_id', 'purchase']]
submit.loc[submit['purchase'] < 0, 'purchase'] = 0 # changing min prediction to min value in train
submit.to_csv('final_solution.csv', index=False)
In [37]:

```

Fig. 8. Using XGBoost

Figure 8 displays the results of using the XGBoost algorithm to analyze the provided data.

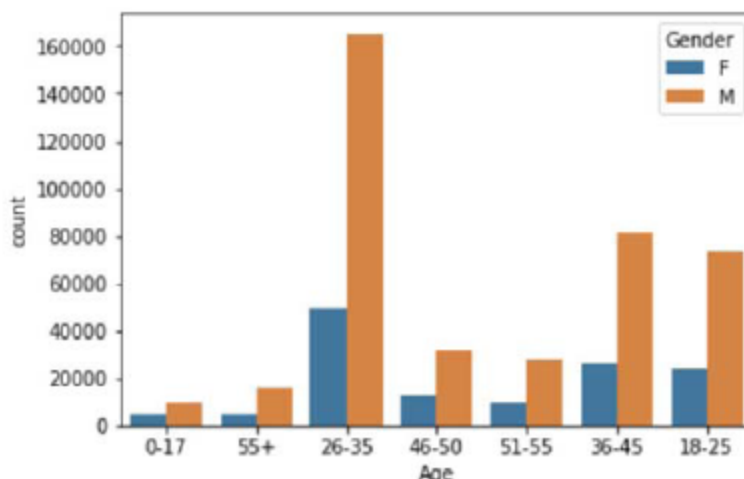


Fig. 9. Graphical analysis of no. of customers vs Age group

Figure 9 shows that the majority of clients fall within the age bracket of 26 to 35 for both sexes. On Black Friday, you won't see nearly as many individuals of any age. The majority of the items marketed to consumers in their late twenties and early thirties should be carried by retail establishments, according to the findings. Increasing the quantity of items aimed at persons in their 30s and decreasing the number of products aimed at those older or younger will lead to higher profits. Black Friday sales were the most lucrative for "people with occupations 0 and 4," as seen in Figure 10. In contrast, the individuals who have spent the least amount of money are those whose vocations include ID 18, 19, and, more precisely, occupation 8. The wealthy do not frequent these kind of establishments, hence it follows that these communities are the poorest.

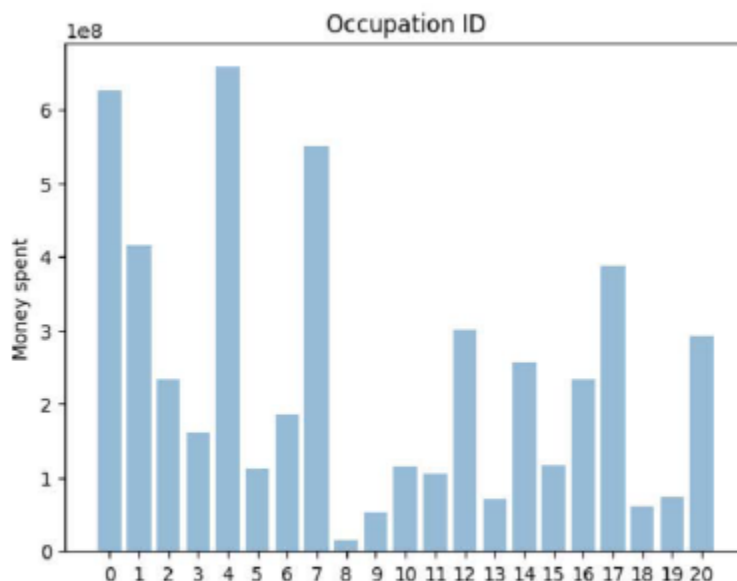


Fig. 10. "Money spent" analysis with respect to the the people's occupation ID's

IV. CONCLUSION AND FUTURE WORK

For any of these jobs, you can utilize Machine Learning (ML). The purpose of this study is to demonstrate how to utilize ML algorithms to forecast how much money customers will spend during the upcoming "Black Friday" sale. In order to uncover intriguing patterns in the dataset, exploratory data analysis has been carried out. According to this study's findings, users should take demographic information like age, gender, and profession into account when trying to guess which product a consumer will buy. When compared to other methods, such as decision trees and ridge regression, our approach is able to provide more accurate predictions, according to the experiments. The

methodologies are summarized after a comparison. Our model with the lowest RMSE also outperforms the existing models, as we have discovered. It is possible to enhance the model's accuracy by using more machine learning algorithms and doing better data cleaning and analysis. Better forecasts will be produced by an expanded dataset. We need to update the dataset with the kind of characteristics contained in it if we want to improve the data or achieve a better outcome. We need to utilize a completely balanced dataset with unique values for every field if we want to improve the outcome. Simply updating the dataset is all that is required to use the same algorithms. To get the most out of it, use a big dataset.

REFERENCES

- [1] Beheshti-Kashi, S., Karimi, H.R., Thoben, K.D., Lutjen, M., Teucke, M.: " A survey on retail sales forecasting and prediction in fashion markets, " *Systems Science & Control Engineering* 3(1), 154, 161(2015)
- [2] Smith, Oliver, and Thomas Raymen. " Shopping with violence: Black Friday sales in the British context. " *Journal of Consumer Culture* 17.3 (2017): 677-694.
- [3] Majumder, Goutam. " ANALYSIS AND PREDICTION OF CONSUMER BEHAVIOUR ON BLACK FRIDAY SALES. " *Journal of the Gujarat Research Society* 21.10s (2019): 235-242.
- [4] Challagulla, Venkata Udaya B., et al. " Empirical assessment of machine learning based software defect prediction techniques. " *International Journal on Artificial Intelligence Tools* 17.02 (2008): 389-400.
- [5] Chu, C.W., Zhang, G.P.: " A comparative study of linear and nonlinear models for aggregate retail sales forecasting, " *International Journal of production economics* 86(3), 217{231(2003) }
- [6] Makridakis, S., Wheelwright, S.C., Hyndman, R.J.: " Forecasting methods and applications, " John wiley & sons(2008)
- [7] Correia, Alvaro, Robert Peharz, and Cassio P. de Campos. " Joints in Random Forests. " *Advances in Neural Information Processing Systems* 33 (2020).
- [8] Kvalheim, Olav Martin, et al. " Determination of optimum number of components in partial least squares regression from distributions of the root mean squared error obtained by Monte Carlo resampling. " *Journal of Chemometrics* 32.4 (2018): e2993.
- [9] Sheridan, Robert P., et al. " Extreme gradient boosting as a method for quantitative structure–activity relationships. " *Journal of chemical information and modeling* 56.12 (2016): 2353-2360
- [10] Ngiam, Kee Yuan, and Wei Khor. " Big data and machine learning algorithms for health-care delivery. " *The Lancet Oncology* 20.5 (2019): e262-e273.
- [11] Domingos, P.M.: A few useful things to know about machine learning. *Communacm* 55(10), 78{87(2012)
- [12] Langley, P., Simon, H.A.: Applications of machine learning and ruleinduction.*Communications of the ACM* 38(11), 54{64(1995)
- [13] Website url: <https://machinelearningmastery.com/gentle-introductionxgboost-> applied-machine-learning, accessed on 20th Sept, 2020
- [14] Xiang Gao, Junhao Wen, Cheng Zhang, " An Improved Random Forest Algorithm for Predicting Employee Turnover", *Mathematical Problems in Engineering*, vol. 2019, Article ID 4140707, 12 pages, 2019. <https://doi.org/10.1155/2019/4140707>
- [15] Das, P., Chaudhury, S.: " Prediction of retail sales of footwear using feedforward and recurrent neural networks, " *Neural Computing and Applications* 16(4-5), 491 {502 (2007)}
- [16] Loh, W.Y.: " Classification and regression trees, " *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1(1), 14{23 (2011) }