

# Black Friday Sale Prediction

CSCI 6409 Process of Data Science Project

## **Instructor**

Dr. Evangelos Milios

## **Group 3 Members**

B00923820 - Anmol Sidhu

B00888136 - Amankumar Manojkumar Patel

B00895684 - Fenil Milankumar Parmar

B00911364 - Kalpit Machhi

B00942240 - Omid Amini

# Table of Content

- Abstract
- Introduction
- Literature Review
- Methodology
- Experiments
- Results
- Conclusion
- Reference

# Abstract

This research project aims to use data science approaches to precisely forecast shop sales on Black Friday. Data from previous Black Friday sales, including the number of consumers, the average purchase amount, and other variables that influence the outcomes, are collected and cleaned for the project. We covered several machine learning methods in this project to assess the data and accurately estimate future Black Friday sales. To comprehend the dataset better, we also include exploratory data analysis of the issue. We utilized XGBoost, Decision Tree Regressor, Random Forest Regressor, and Linear Regression as our base model. The findings of this study may be valuable information for merchants to maximize their profits.

# Introduction

- Nowadays, festive sales have picked up a lot of popularity. It is human nature to get attracted to good deals, and many customers wait for sales to buy products. It is also observed that customers linger on buying high-priced products until the upcoming sales. The Black Friday sale is one of the most awaited sales of the year. Since there is so much customer inflow, it is an excellent opportunity for a business to attract customers and make a profit.
- Therefore, if a business knows what deals to offer to a specific client, it can make a lot of money. This is what we are targeting through our project. We aim to predict the amount a customer will spend during the sale through previous data and buying trends of the individual with features like age, marital status, etc. This will allow the company to give specific offers to the customer according to their budget. If the model predicts that the customer will be buying products worth a significant amount, the business can give them offers for those kinds of products and vice versa.
- With the boom in e-commerce and online shopping, customers don't tend to think a lot before buying; therefore, if they get a good deal, they will go for it. The main attraction lies in providing them deals they want, which our proposed model will help in. Therefore, we believe this will be an excellent addition to a business and can bring a lot of profit during the Black Friday sale if used efficiently.



# Introduction - Problem Statement

- A retail company, 'More For Less,' wants to understand the customer behavior of purchasing certain products during the Black Friday sale. So that they can provide customized Black Friday offers to different sets of customers to increase their profit during the sale. They want to predict a customer's purchase amount using the Black Friday data gathered over the last five years.
- The company wants to build a model that can predict a customer's purchase amount. Such a model can help them create customized offers that attract new customers.
- Customizing Black Friday offers to different sets of customers to increase profit during the sale.
- Predict the purchase amount of a customer to create perfect deals according to the customer's budget.
- Restock certain items more likely to be sold in a sale.

# Introduction - Objective

We are solving a problem for supermarket leaders, eCommerce leaders, and various stores which have a vast range of customers in terms of occupation, age, gender, etc. There are various benefits which our proposed solution can offer. Some of them are listed below:

- A better and more accurate prediction technique can be found by analyzing various Machine Learning algorithms for predicting Black Friday sales.
- A better understanding of the behaviour of customers on Black Friday based on their purchases in different product categories.
- Accurate prediction of the purchasing capabilities of potential customers so that business owners can better understand their customers.
- Business owners can stock up on the most needful items by better understanding the demand for particular products from our exploratory data analysis.
- Retain more profitable customers by providing them special benefits or offers on such occasions as Black Friday.

With all these mentioned benefits, businesses can expect significant growth by selling the most demanded items to potential buyers.

# Literature Review

- It is a basic human nature to get attracted to good deals and a lot of customers wait for sales to buy products.
- Sales play a key role in the building of loyalty and trust between customer and business.
- Understanding the buyer is the foundation of effective selling.
- Identifying the experience, the buyer wants to have.

# Literature Review (Cont.)

Various research has been done to predict the total purchase based on different parameters such as age, gender, occupation, and marital status. After having a thorough understanding of such research, we have concluded that Machine Learning algorithms are best to work with this statistical data prediction, therefore, we are aiming to find a better approach to analyze the data and then predict the data with more accuracy using the best machine learning algorithms [1].

The use of deep neural networks can result in overfitting, and it is not the best fit for predicting purchases of users based on a few features of black Friday sales. There are multiple types of research regarding our problem statement, and a few of them are briefly described below:



# Literature Review (Cont.)

- Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data [1]
- Summary of above paper: Researchers have done a thorough analysis of different types of algorithms consisting of Machine Learning, Deep Learning methods, and concluded that Machine Learning algorithm such as XGBoost is performing best in comparison to Neural Networks, Decision Tree, etc [1].

# Literature Review (Cont.)

- Machine Learning Application for Black Friday Sales Prediction Framework [2]
- Summary of above paper: Authors have analyzed the performance of various Machine Learning algorithms to find the best results and concluded that Random Forest Regressor gave more precise predictions than using techniques such as Ridge Regression, and Decision Trees [2].

# Literature Review (Cont.)

- Analysing and Predicting the purchases done on the day of Black Friday [3]
- Summary of above paper: Contributors of this paper have approached several techniques to analyze the performance of models XGBoost, TFidTransformer, XGBoost + TFidTransformer, and ExtraTreeRegressor. They have developed 4 types of cases where they trained models using various quantity combinations of training and testing sets [3].

# Customer Behaviors of purchasing certain products during the Black Friday sale

- Customizing Black Friday offers to different sets of customers to increase profit during the sale.
- Predict the purchase amount of a customer to create perfect deals according to the customer budgets.
- Restock certain items that are more likely to be sold in a sale.

# Methodology

We developed an accurate and efficient algorithm to analyze customer spending in the past and output the customers' future spending with the same features. In order to achieve this goal, we have implemented different machine learning algorithms, including Linear Regression, Decision Trees, Random Forest Regression and XGBoost for regression.



# Methodology Execution Plan

- In order to have better and more reliable results, we tend to balance the dataset by removing unusable and unwanted data from the dataset and trying to fill the missing values with other available information to have a clean dataset. Moreover, to have a reliable dataset for the next steps. Only the columns `Product_Category_2` and `Product_Category_3` have null values because some products only belong to one category. For almost 70% of transactions, the `Product_Category_3` feature is null and unusable, so we replace the nulls with a '-' negative value so that the model understands that they were empty, because deleting such hefty records may harm our model training.
- In the EDA step, we use statistical graphics and other data visualization methods to explore our dataset and identify potential relationships between variables to summarize the main characteristics of our dataset. For example, Men bought twice as much as women, and The age group with the most transactions was 26-35.
- After exploring the dataset, we start with the preprocessing steps that transform raw data into features that can be used in our machine learning algorithms. New features consist of an outcome variable (purchases) and predictor variables; the most valuable predictor variables are created from our raw data and selected for the predictive model during our feature engineering process.

# Methodology Execution Plan

- Machine Learning models are of two types: Supervised and Unsupervised. For our problem, we will build a supervised machine learning model. Our target value is "Purchases," which is correlated with other feature columns such as "Age", "Occupation", and "Gender" therefore, we can train our model using XGBoost, Decision Trees, and some other regression algorithms as well. To train our Page 4 model accurately, we have to set training and testing datasets balanced so that our model performs efficiently. We also need to normalize or standardize the dataset to make it correctly fit the model [4].
- The primary goal of supervised machine learning is an accurate prediction. Evaluation metrics for regression models are quite different from those we discussed for classification models because we now predict in a continuous range instead of a discrete number of classes. We are going to use The R2 coefficient, Root Mean squared error, and Explained variance [6]

# Methodology (Cont.) - Dataset

We will use a Kaggle dataset comprising sales transactions captured at a retail store from multiple shopping experiences [5]. Dataset has 537577 rows (transactions) and 12 columns (features) as described below [5]:

- User\_ID: Unique ID of the user. There are a total of 5891 users in the dataset.
- Product\_ID: Unique ID of the product. There are a total of 3623 products in the dataset.
- Gender: indicates the gender of the person making the transaction.
- Age: indicates the age group of the person making the transaction.
- Occupation: shows the user's occupation labelled by numbers from 0 to 20.
- City\_Category: User is living city category. categorized into three categories: 'A', 'B', and 'C.'
- Stay\_In\_Current\_City\_Years: This shows how long the users have lived in this city.
- Marital\_Status: if the user is married, the value is one; otherwise, it is zero.
- Product\_Category\_1 to Product\_Category\_3: Category of the product.
- Purchase: The purchase amount is our targeted value.

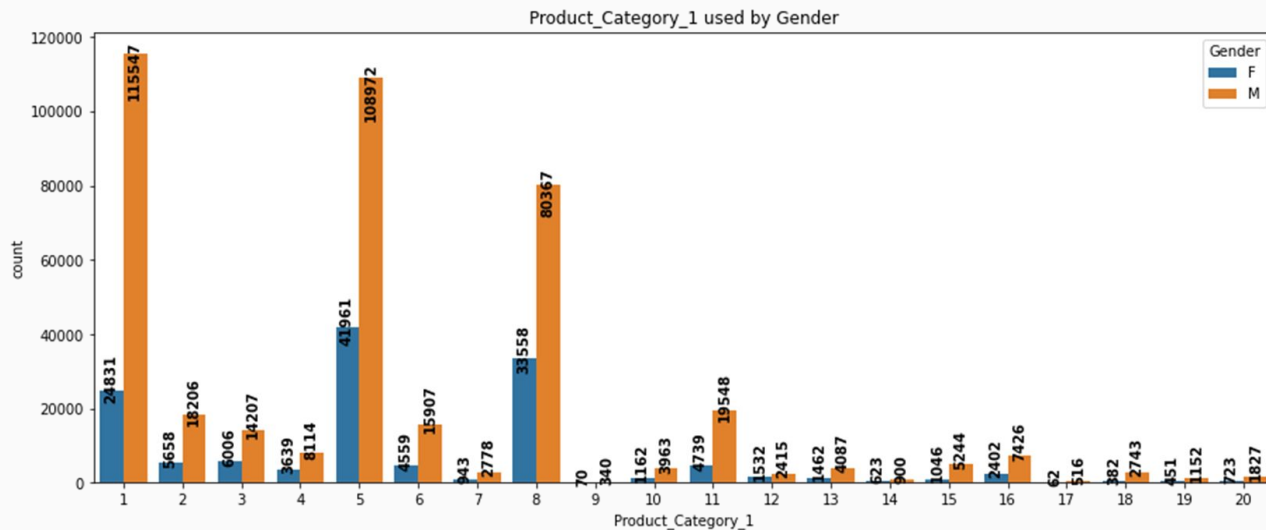
In addition, we also created some new features to improve the model training, we will discuss about these features in Experiments.

# Methodology (Cont.) - What can we do with Dataset?

- **Analyzing**
  - Who is more likely to spend more in a black Friday sale?
    - Men / Women
    - Married / Single
    - Old Residents / New residents
  - Which type of products are more likely to be sold in a sale like black Friday?
  - Which type of products are common among men and which among women?
- **Prediction**
  - Predict age of a customer.
  - Predict gender of a customer.
  - Predict amount of purchase.
- **Recommendation**
  - Suggest products based on age.
  - Suggest products based on customers location.
  - Suggest products based on gender.

# Methodology (Cont.) - Basic Observation from Dataset

- Product **P00265242** is the **most popular** product.
- **Most of the transactions** were made by **men**.
- **Male** customers tend to **spend more** than female customers
- **Age** group with most transactions was **26 - 35**.
- **Most** of our **customers** come from **City B**, but customers come from **City C** spent more.
- Most **users** come from **City Category C**, but **more people** from **City Category B** tend to **purchase**

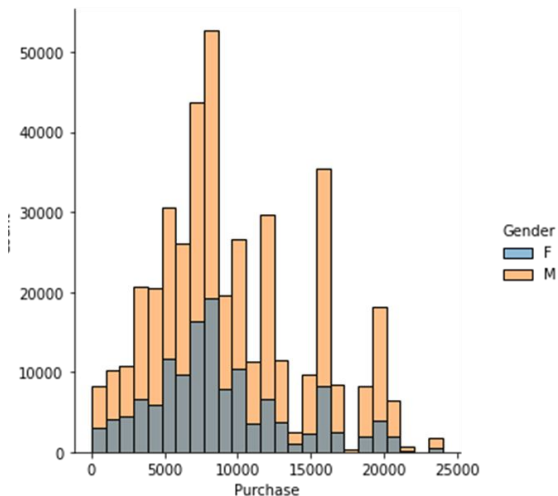




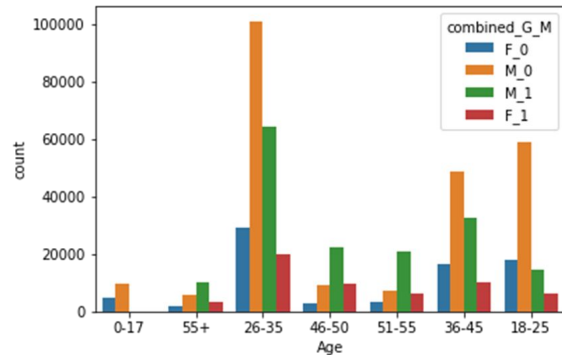
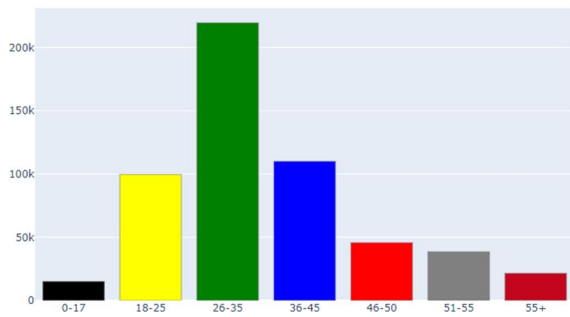
# Methodology (Cont'd) - Basic Observation from Dataset

- In Prodcut\_Category\_1 , **category 10** adds up to highest amount of purchases.
- In Prodcut\_Category\_2 , **category 10** adds up to highest amount of purchases.
- In Prodcut\_Category\_3 , **category 3** adds up to highest amount of purchases.
- If we see the initial correlation matrix, **Occupation** looks to be the most correlated feature with our target variable Purchase. Although , we will see this correlation matrix again when we create new features.
- If we see the Distplot, the target data looks to be in a **normal distribution**.

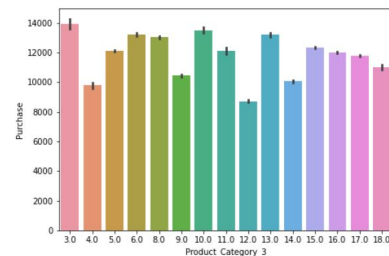
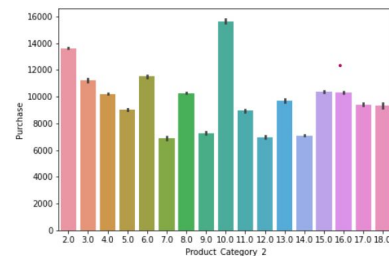
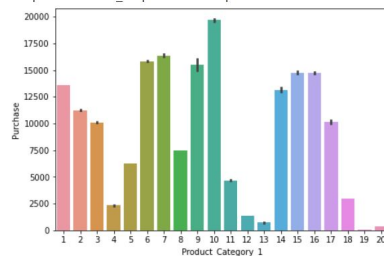
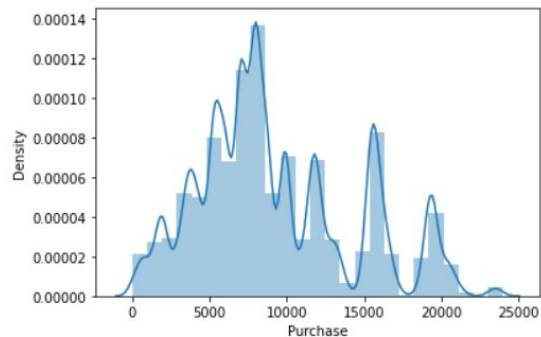
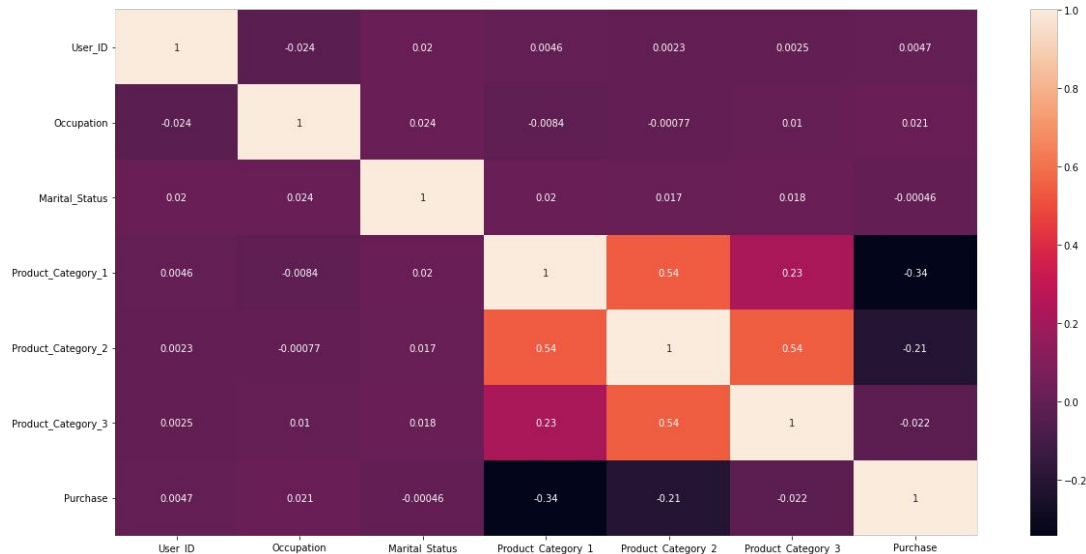
# What have we obtained from dataset?



How many products were sold by ages



# What have we obtained from dataset?



# Experiments

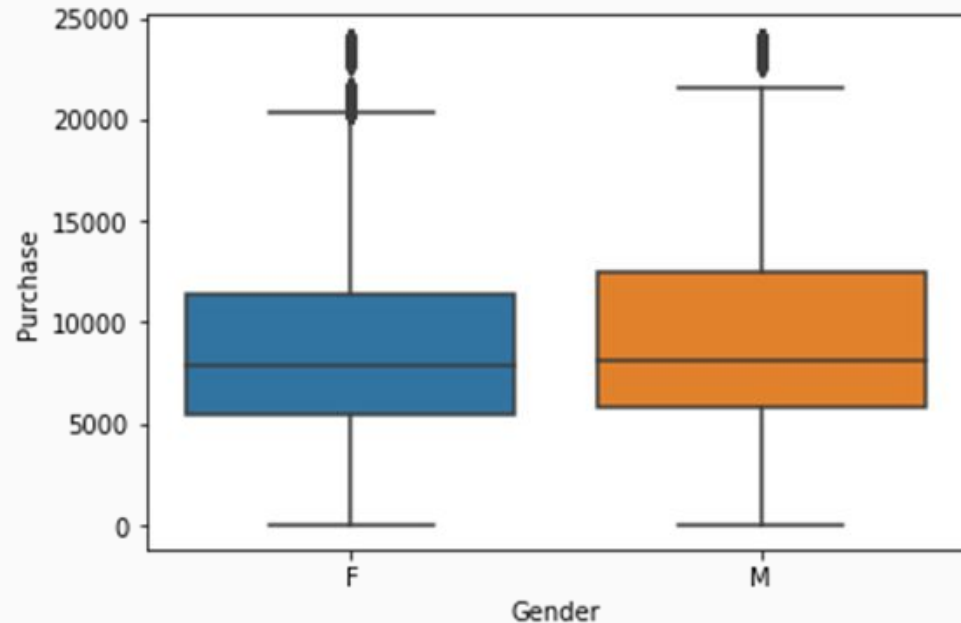
We have done various experiments with our dataset which includes below steps:

- Preprocessing - Data Cleaning
- Modeling
- Hyperparameter tuning

# Experiments (Cont'd)

## Pre-processing - Cleaning Data

- Outliers - As we can see from three boxplots of Gender, Age and Marital Status vs Purchase, they are not having any significant effect.
- Null Values - Only Product\_Category\_2 and Product\_Category\_3 have null values, but those null values are additional category for each product (some items has more than one category).

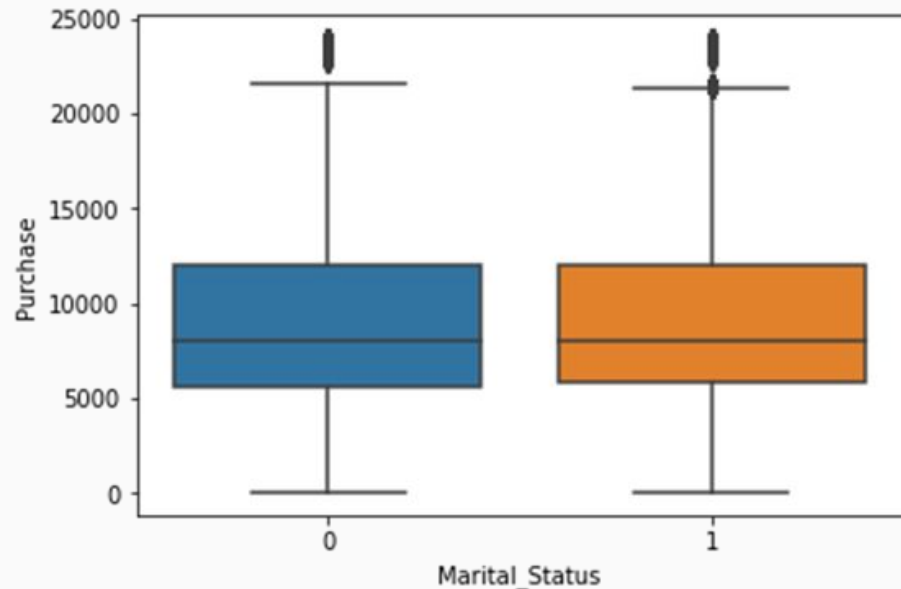




# Experiments (Cont'd)

## Pre-processing - Cleaning Data

- Create feature
  - Product Popularity count of Products for Product ID
  - Buying Power (total amount spent) for User ID.
  - The number of categories in which the product is present . [1,3]
- Remove Useless Features
  - User Id
  - Product ID

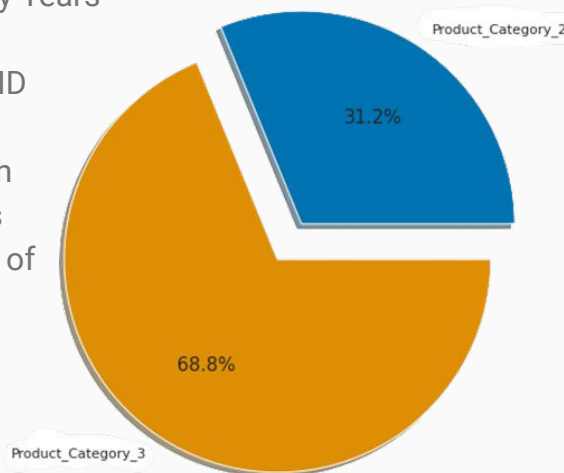


# Experiments (Cont'd)

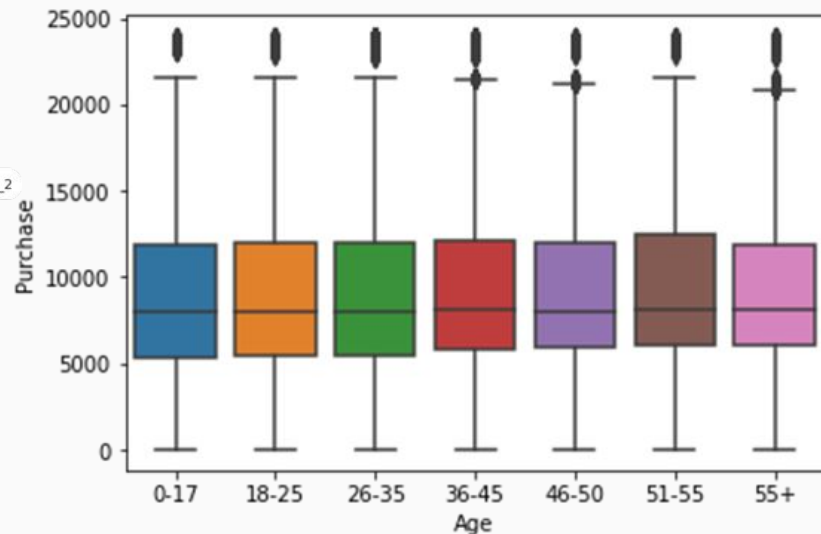
## Pre-processing - Cleaning Data

- Converting Features
  - City
  - Age
  - Stay In Current City Years
  - Gender

- As we have to drop User ID and Product ID columns, created new features with approximately resembles them. Product Popularity of products for ProductID And Buying power for UserID.



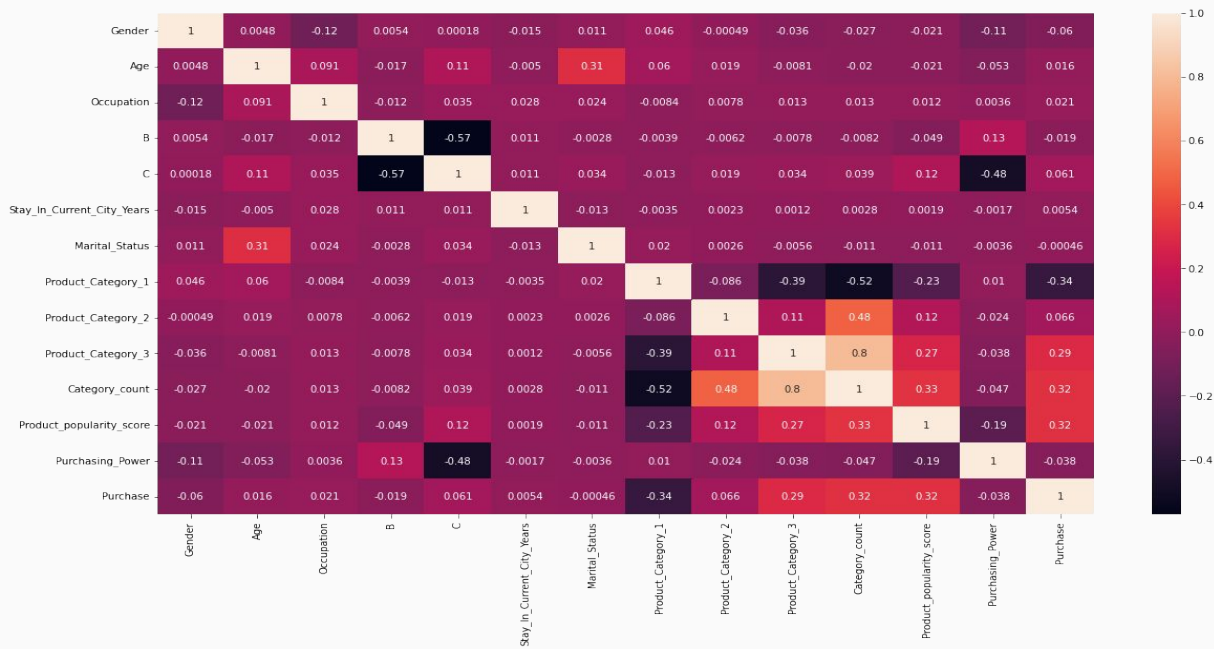
Null Values



# Experiments (Cont'd)

## Pre-processing - Cleaning Data

Final Correlation  
Matrix including  
all the final  
features



# Experiments (Cont'd)

## Models

This is the feature importance score calculated via selectKBest.

Our new features which we have created get a good ranking.

```
Product_Category_1      1669.641452
Product_popularity_score 1070.421238
Product_Category_2       529.360141
Product_Category_3       234.620595
Category_count          171.956953
Purchasing_Power        70.651576
Occupation               9.624324
Gender                   6.898495
Age                     5.438610
C                        3.390445
B                        1.198837
Stay_In_Current_City_Years 0.750075
Marital_Status           0.000000
dtype: float64
```

# Experiments (Cont'd)

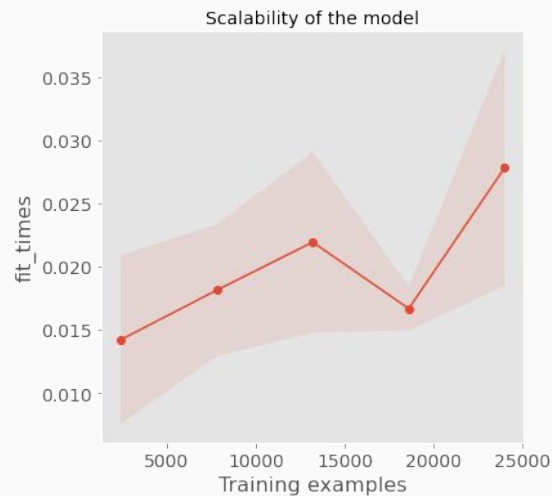
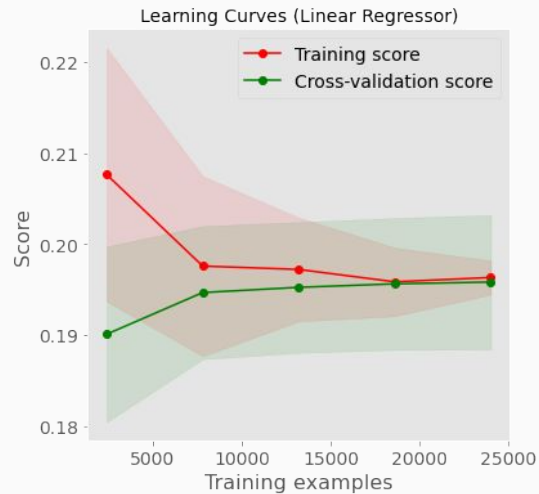
## Models

- We used different models to predict the **amount spent** by the customer on Black Friday.
- **Linear Regression**
  - **Base model**
  - **simplest machine learning models.**
- **Decision Trees for Regression**
- **XGBoost for Regression**
  - powerful approach for building supervised regression models.
- **Random Forest Regression**
  - Ensemble learning method
  - Based on decision trees and runs various decision trees in parallel.



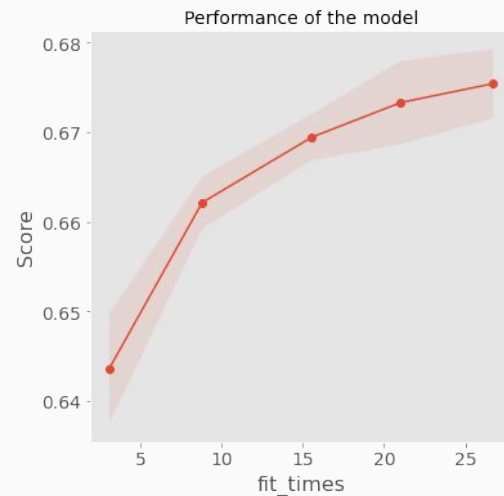
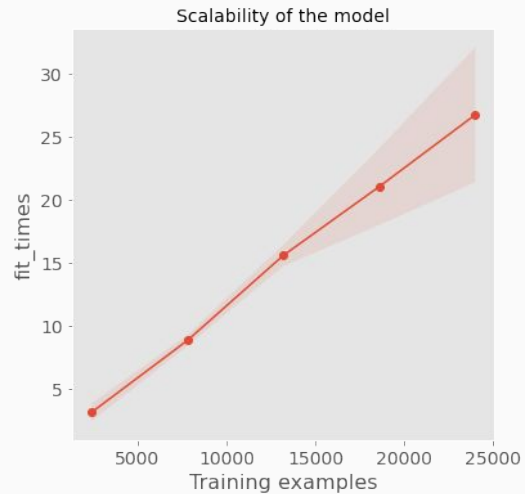
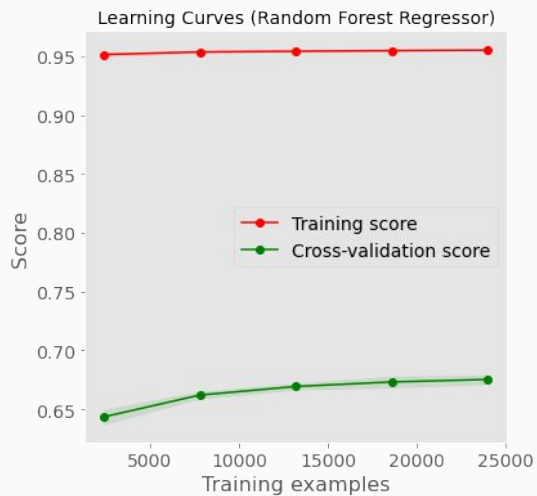
# Linear Regression

## Learning Curve



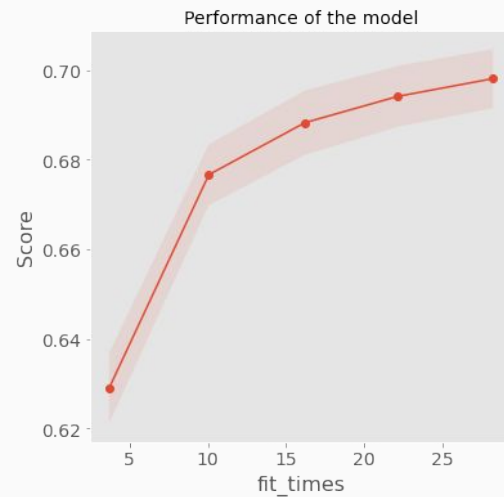
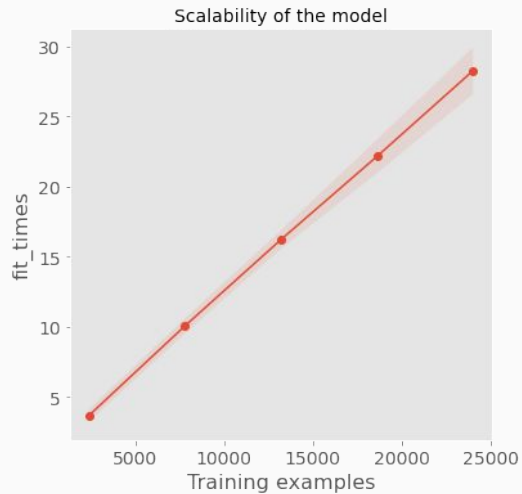
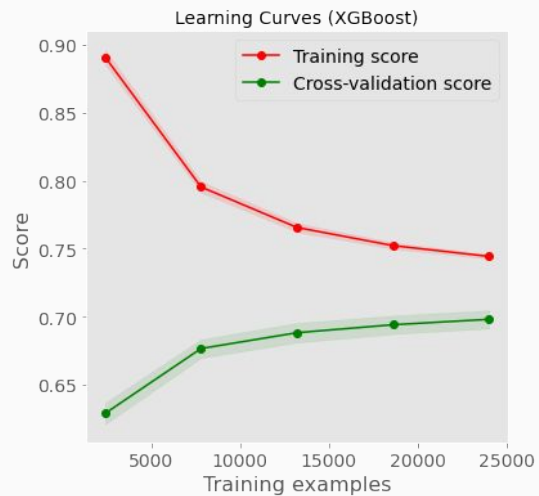
# Random Forest Regressor

## Learning Curve



# XGBOOST

## Learning Curve



# Learning Curves

In the above graphs, the learning curves, the scalability and the performance of the models are plotted.

We use Linear Regression as our baseline model and compare other models with its curve.

The Training learning curve for the XGBoost algorithm is going down as well as the Validation scores goes up shows us the model is getting better. The XGBoost algorithm can score more in terms of predicting the purchase amount. Therefore, it turns out to be the best algorithm if we compare the learning curves.

The RandomForestRegressor , shows it may be able to learn with the increase of training data. Its training score does not change much.

# Results

Model	RMSE	R2 Score
Linear Regression	0.1873319191162088	0.19892741942728787
Decision Tree	0.11737424108918598	0.6855193729461644
Random Forest	0.11342680238837338	0.7063164099661337
XGBoost	<b>0.10646594277076929</b>	<b>0.7412563438312136</b>

XGBoost Achieved higher R2 score

# Results

From the table we again come to the conclusion that XGBoost Performs the best for our problem. If we compare the R2 scores and Root Mean Squared errors of all the models used , XgBoost gives us the least RMSE and highest R2 score .

This also matches our conclusion from the learning curves as XGBoost showed the best learning curve out of all the models.

# Conclusions


We were successfully able to solve the problem for the business enterprise “More For Less”. Using our model they can predict the buying patterns of the buyer during Black Friday Sale and make better profits by providing correct deals to correct buyers.

We preprocessed the dataset and create useful features such as Categories counts, Product Popularity, and Buying Power of customers.

We trained several models such as Linear Regression, Random Forest, and XGBoost, and tested them in different aspects.

# Conclusions

Based on our models learning of the training data , we predicted the purchase value in the given dataset by the enterprise which had no values for the purchase column.



	Predicted Purchase	Purchase
0	0.628084	15053.994141
1	0.432533	10370.739258
2	0.271535	6514.979980
3	0.118789	2856.878174
4	0.104091	2504.869385
...	...	...
233594	0.285133	6840.656250
233595	0.237353	5696.363281
233596	0.396020	9496.286133
233597	0.742370	17791.007812
233598	0.122676	2949.972168

233599 rows × 2 columns



# Conclusions

Based on the problem , our model performed well with a high  $R^2$  score and a lower RMSE . Also, the features created by us came out to have good feature importance as well.

If we were equipped with a higher computing power , we could have performed hyperparameter tuning on whole of the dataset rather than a chunk of it which would yield in more optimized parameters and better scores.

# References

- [1] C.-S. M. Wu, P. Patil, and S. Gunaseelan, "Comparison of different machine learning algorithms for multiple regression on black Friday sales data," in 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), 2018, pp. 16–20.
- [2] H. V. Ramachandra, G. Balaraju, A. Rajashekar, and H. Patil, "Machine learning application for black Friday sales prediction framework," in 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), 2021, pp. 57–61.
- [3] S. Kalra, B. Perumal, S. Yadav, and S. J. Narayanan, "Analysing and Predicting the purchases done on the day of Black Friday," in 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020, pp. 1–8.
- [4] A. Bharadwa, "7 steps to a successful data science project," Towards Data Science, 06-Feb-2021. [Online]. Available: <https://towardsdatascience.com/7-steps-to-a-successful-data-science-projectb452a9b57149>. [Accessed: 20-Dec-2022].
- [5] StefanDolezel, "Black Friday." 21-Jan-2018.
- [6] J. Jordan, "Evaluating a machine learning model," Jeremy Jordan, 21-Jul-2017. [Online]. Available: <https://www.jeremyjordan.me/evaluating-a-machine-learning-model/>. [Accessed: 20-Dec-2022].

Thank you!