A STUDY ON BIGMART SALES, BLACK FRIDAY SALES AND MALL CUSTOMER SEGMENTATION.

Masters in Science in Data Analytics (MSCDAD_B)
Data Mining and Machine Learning

ANNJOYS ROBERT
Data Mining and Machine Learning
National College of Ireland
X22137459@student.ncirl.ie

The Video Presentation for the project is available at the following link:
https://www.canva.com/design/DAFiDLcSbDs/llyLyOzD0xAv6AqMtavlcQ/view?utm_content=DAFiDLcSbDs&utm_campaign=designshare&utm_medium=link&utm_source=recording_view

I. ABSTRACT:

This project aims to explore data-driven insights for retail by analyzing sales data from Big Mart, Black Friday and customer segmentation data from a shopping mall. The project employs machine learning techniques to predict sales and customer behaviour and to segment customers based on their purchasing patterns. The study is conducted using publicly available datasets, which are preprocessed and analyzed using Python libraries such as pandas, NumPy, sci-kit-learn, and seaborn. The results of the analysis reveal insights into sales patterns and customer behaviour and provide useful information for retailers to optimize their sales strategies and marketing efforts. The study also showcases the potential of machine learning techniques in providing valuable insights for the retail industry.

II. INTRODUCTION:

1. MOTIVATION:

Retailers are always seeking ways to enhance their sales and marketing tactics to stay ahead of the competition because the retail sector is one of the biggest and most competitive in the world. Big data and machine learning have increased interest in employing data-driven methods to understand consumer behaviour, forecast sales patterns, and improve marketing initiatives.

The necessity for merchants to use data-driven insights to make wise business decisions is the driving force behind this project. A unique chance to investigate the possibilities of machine learning approaches in the retail sector is provided by the examination of sales data from Big Mart and Black Friday, as well as customer segmentation data from a shopping complex. We can glean insights and knowledge from these datasets by applying machine learning algorithms, which will enable retailers to enhance their customer satisfaction levels, sales and marketing tactics, and financial performance.

The outcomes of this project can also benefit the larger academic community by demonstrating how machine-learning approaches can produce insightful data for the retail sector. This may spur additional study and development of data-driven retail strategies and result in more effective and efficient sales and marketing plans for retailers all over the world.

2. OBJECTIVE:

This project's goal is to use machine learning algorithms to evaluate customer segmentation data from a mall, Big Mart, and Black Friday sales data in order to:

- Use historical data to forecast sales and customer behaviour.
- Create client segments based on the customers' shopping habits, and then pinpoint the essential traits of each segment.
- Draw conclusions and expertise from the data to offer information that will help merchants improve their marketing and sales strategy.

The project's objectives are to show how machine learning techniques may yield insightful data for the retail sector and to provide concrete advice for retailers on how to enhance their sales and marketing tactics in light of the research's findings.

III. METHODOLOGY:

DATASET 1: BIGMART SALES PREDICTION

Dataset Description:

The BigMart Sales Prediction dataset contains sales data for 1559 items that customers purchased and sold in ten different grocery chains. Using the provided attributes, this dataset seeks to forecast each product's sales at various retailers. The dataset has been used in numerous studies to assess various machine-learning techniques for regression problems and has acted as a baseline for sales prediction models. The availability of this dataset makes it a valuable resource for academics, data scientists, and anybody else interested in comprehending and applying predictive analytics in the retail industry.

1. IMPORTING LIBRARIES:

The Pandas, NumPy, Seaborn, Matplotlib, and Warnings libraries are the libraries that were imported into the program at its initial stage, later many other libraries were also imported and installed.

2. READING DATA:

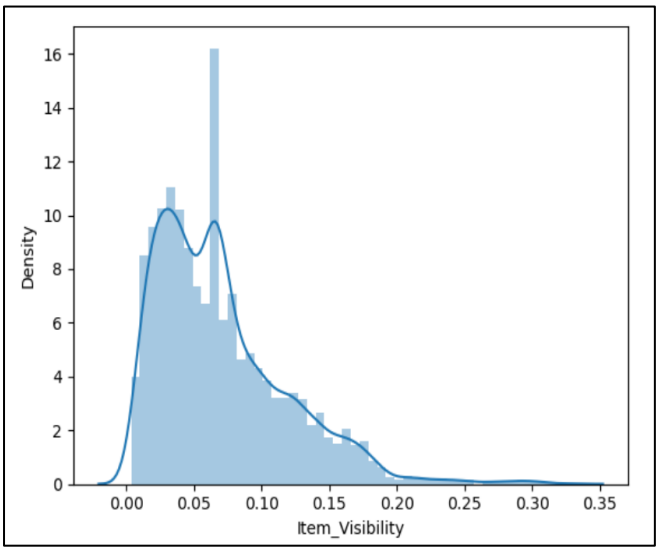
This section outlines the procedures used to read and explore the dataset, including reading the CSV file, displaying the first five rows of data, showing statistical data, displaying attribute data types, and performing unique value checks on the dataset as well as checks for null values and categorical attributes.

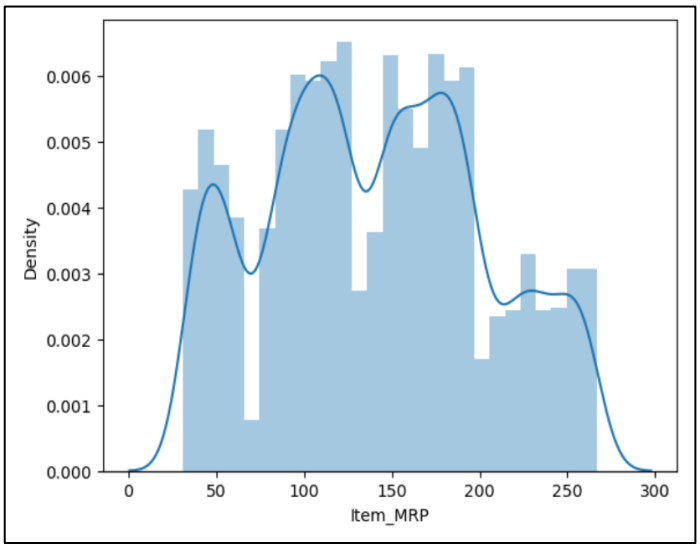
3. DATA PREPROCESSING:

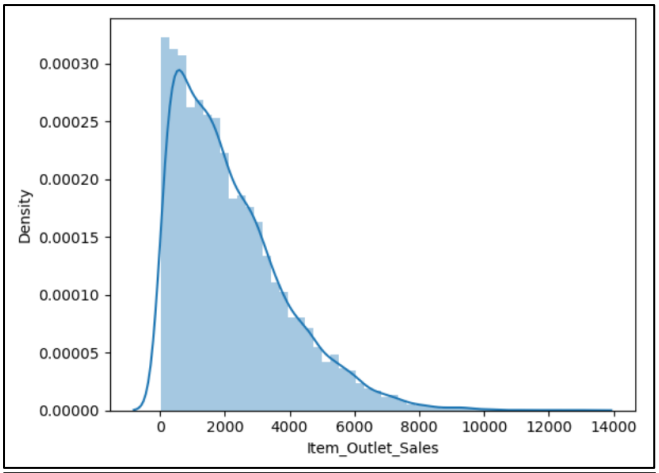
The procedures used to clean and get the data ready for analysis are described in this section. The missing data for item weight and outlet size were handled, zeros in item_visibility was replaced with mean, item fat content is combined, a new column is created for item kind, and a new column is created for outlet years.

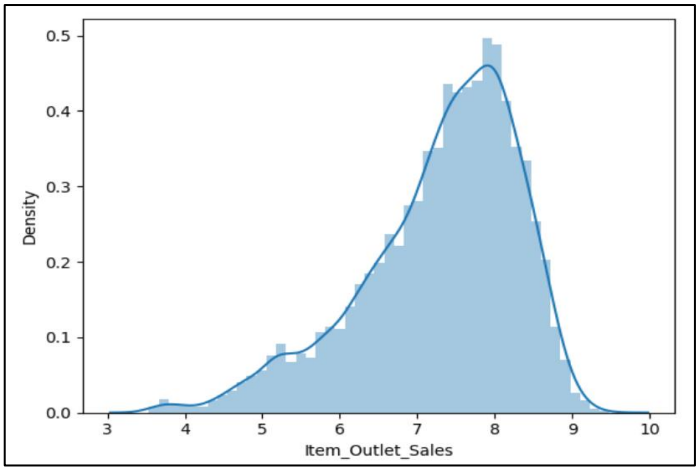
4. EXPLORATORY DATA ANALYSIS:

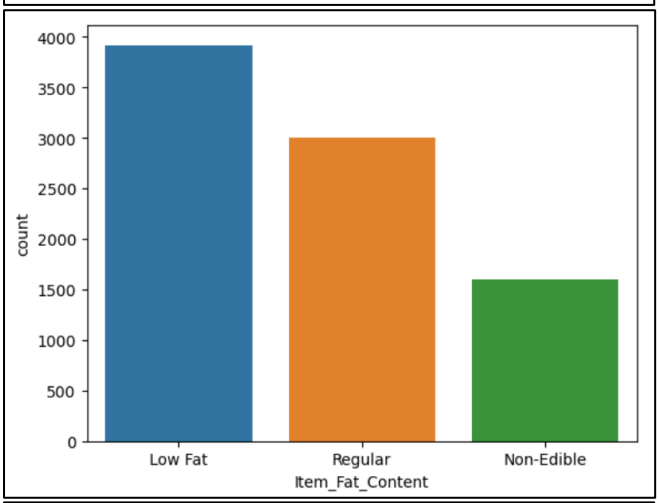
The visualizations made to examine the connections between the various variables in the dataset are covered in this section. The visualizations include a correlation heatmap, item fat content distribution, item type distribution, outlet establishment year distribution, outlet size distribution, outlet location type distribution, outlet type distribution, item weight distribution, item visibility distribution, item MRP distribution, item outlet sales distribution, and performing log transformation on item outlet sales.

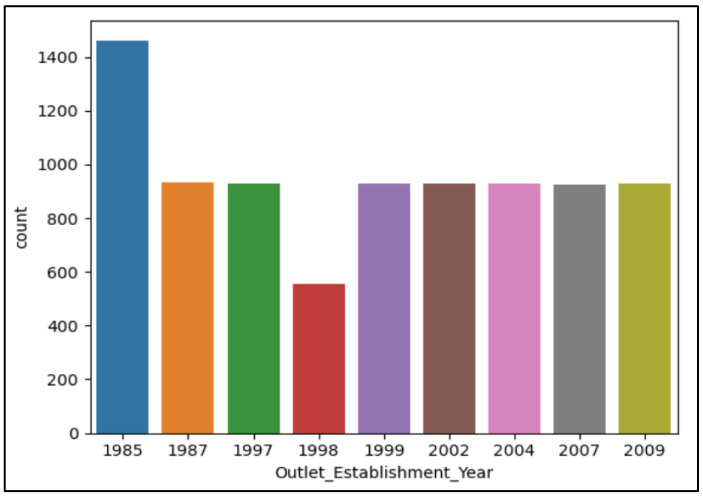


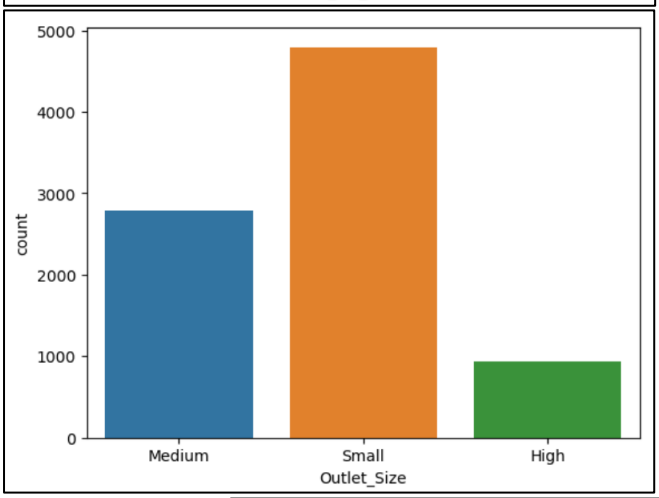


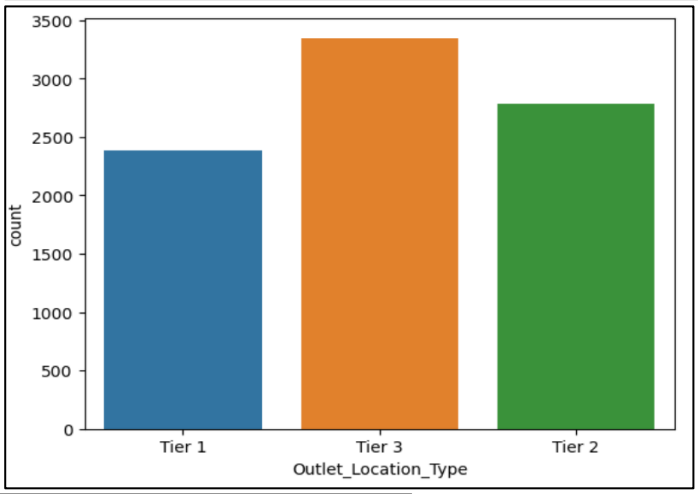

Density
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0
2000
4000
6000
8000
10000
12000
14000
Item_Outlet_Sales

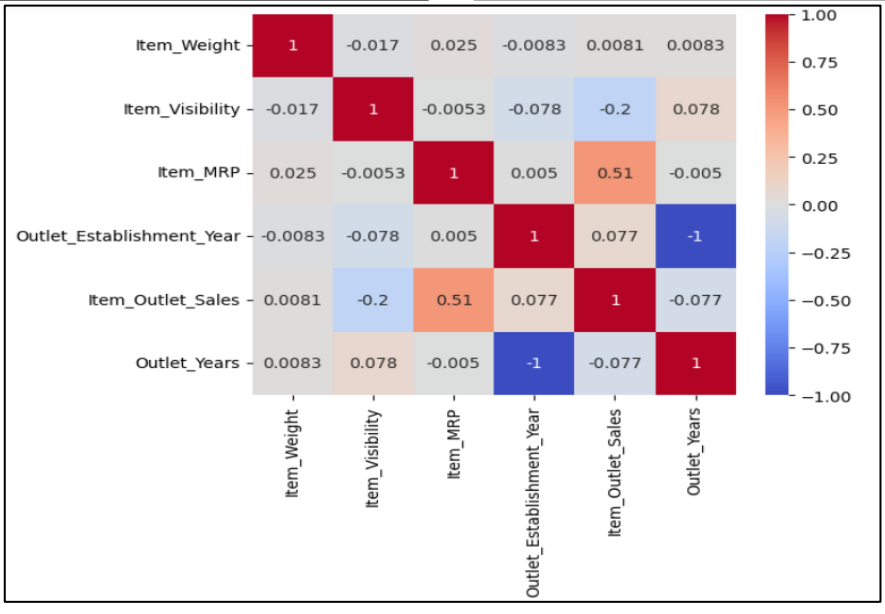

Density
0.5
0.4
0.3
0.2
0.1
0.0
3
4
5
6
7
8
9
10
Item_Outlet_Sales


count
4000
3500
3000
2500
2000
1500
1000
500
0
Low Fat
Regular
Non-Edible
Item_Fat_Content


count
1400
1200
1000
800
600
400
200
0
1985
1987
1997
1998
1999
2002
2004
2007
2009
Outlet_Establishment_Year


count
5000
4000
3000
2000
1000
0
Medium
Small
High
Outlet_Size


count
3500
3000
2500
2000
1500
1000
500
0
Tier 1
Tier 3
Tier 2
Outlet_Location_Type


Item_Weight
Item_Visibility
Item_MRP
Outlet_Establishment_Year
Item_Outlet_Sales
Outlet_Years
1
-0.017
0.025
-0.0083
0.0081
0.0083
-0.017
1
-0.0053
-0.078
-0.2
0.078
0.025
-0.0053
1
0.005
0.51
-0.005
-0.0083
-0.078
0.005
1
0.077
-1
0.0081
-0.2
0.51
0.077
1
-0.077
0.0083
0.078
-0.005
-1
-0.077
1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

5. MODEL CONSTRUCTION AND EVALUATION:

The construction and assessment of the model are covered in this section. The process involves choosing the model's columns, building dummy variables for categorical features, dividing the dataset into training and testing sets, training the model, predicting the testing set, performing cross-validation, displaying the model report, visualizing residuals, assessing the model using Statsmodels. In general, this part seeks to shed light on the effectiveness and precision of the model's sales prediction.

6. CONCLUSION:

Analysis for Bigmart Sales Prediction using Linear Regression and Random Forest Regression:

Linear Regression:

Scikit Learn and Statsmodel are the two libraries implemented for the Linear Regression Model for the Bigmart Sales Prediction Dataset. With an MAE of 1.4132715851298768e-15, RMSE of 1.777983987162317e-15, and an R-squared value of 1.0, the scikit-learn model exhibits extremely low error metrics. Additionally, the statsmodel model's R-squared value is 1.0, demonstrating that it completely accounts for the variance in the response variable.

Random Forest Regression:

In terms of prediction accuracy, the XGBoost model surpasses the sklearn Random Forest Regression model due to lower MSE, RMSE, and MAE values and a higher R-squared value. The sklearn Random Forest Regression model, on the other hand, has a higher cross-validation score, implying more generalization and perhaps better performance on data.

DATASET 2: BLACK FRIDAY SALES PREDICTION

Data Description:

The Black Friday sales projection dataset contains information on the purchases made on Black Friday at a retail store. The dataset, which consists of 550,000 observations, contains 12 variables, including User_ID, Product_ID, Gender, Age, Occupation, City_Category, Stay_In_Current_City_Years, Marital_Status, Product_Category_1, Product_Category_2, and Purchase. The dataset seeks to forecast the customer's expected spending on goods. The dataset has been extensively used to examine patterns of customer behaviour, analyze consumer preferences, and project sales for prospective Black Friday events.

1. IMPORTING LIBRARIES:

The Pandas, NumPy, Seaborn, Matplotlib, and Warnings libraries are the libraries that were imported into the program at its initial stage, later many other libraries were also imported and installed.

2. DATA EXPLORATION:

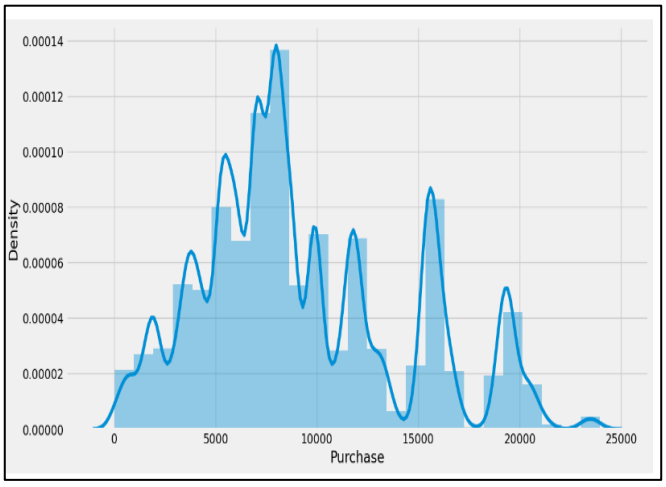
After loading the dataset Pandas were used to perform several operations on the data, such as obtaining statistical data, identifying distinctive values, and learning about the data types. In order to see the purchase amounts visually, a distribution plot was created. By creating countplots for Gender, Age, Marital Status, Occupation, Product Category 1, Product Category 2, Product Category 3, City Category, and Stay In Current City Years, it was able to run univariate analysis. In order to investigate the link between purchase quantity and the attributes of Age, Occupation, and Gender, a bivariate analysis was also conducted. Later the association between the features using Seaborn's heatmap was determined.

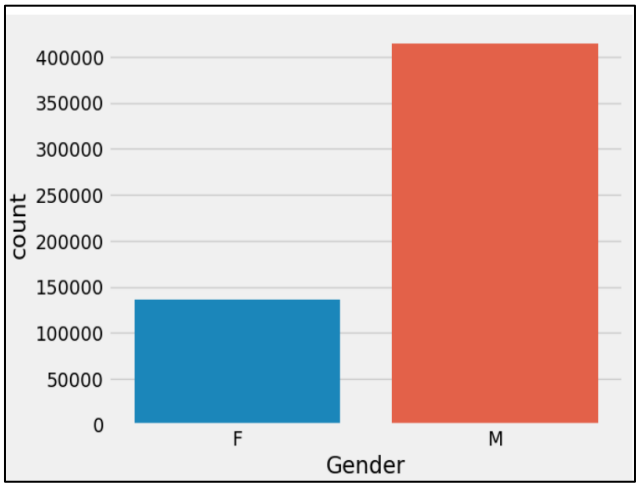
3. DATA PREPROCESSING:

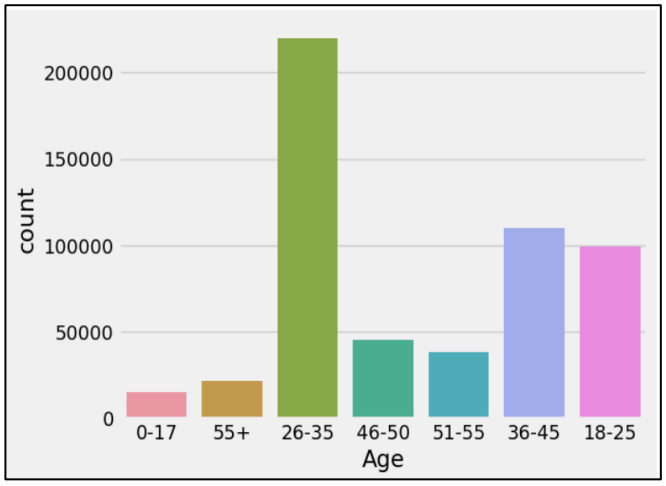
Looking for missing data, it was discovered that Product_Category_2 and Product_Category_3 both had them. After changing the data type to float32 and filling in the missing values with -2.0. Label encoding for Age, City Category, and Stay In Current City Years while employing a dictionary to encode Gender. The dataset was subsequently divided into training and testing sets.

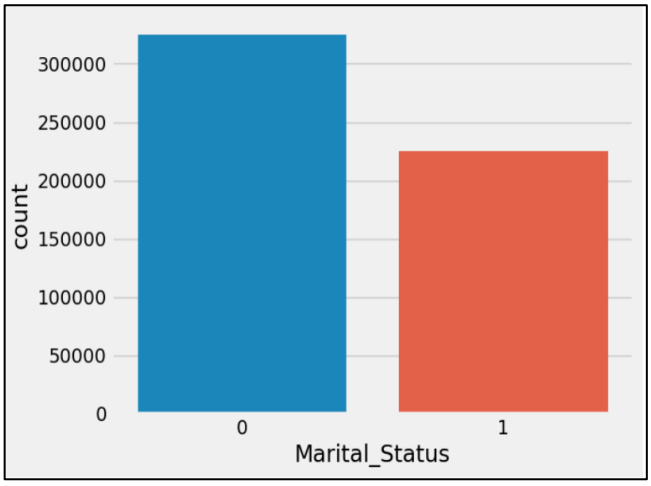
4. EXPLORATORY DATA ANALYSIS:

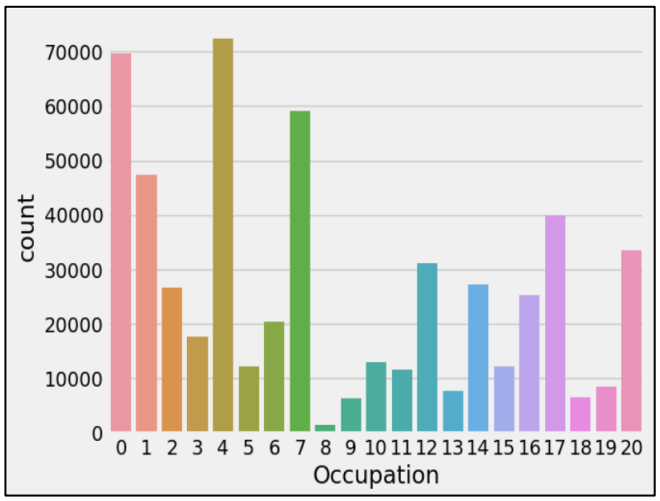
The visualizations were made to examine the connections between the various variables in the dataset covered in this section. The distribution of the target variable "Purchase" is depicted in the graphic. The bins on the histogram visualization were all set to 25. The count of the distinct values for each Categorical Variable in the dataset is plotted using a Countplot. Gender, age, marital status, occupation, product categories 1, 2, and 3, city category, and stay in current city years are among the categorical variables in the dataset. Bivariate analysis approach was used to compare how the target variable varies with each categorical variable by taking the mean of "Purchase" for each distinct value of a categorical variable. This analysis takes into account three variables: occupation, age, and gender. To show the relationship between various numerical variables in the dataset, a heatmap is produced using the correlation matrix.

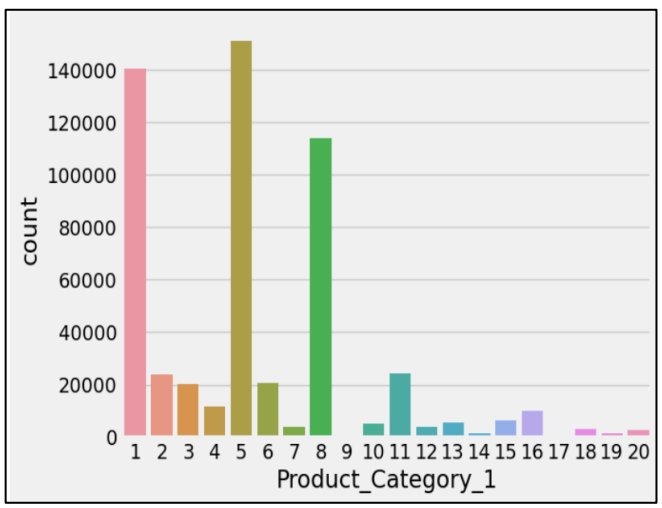


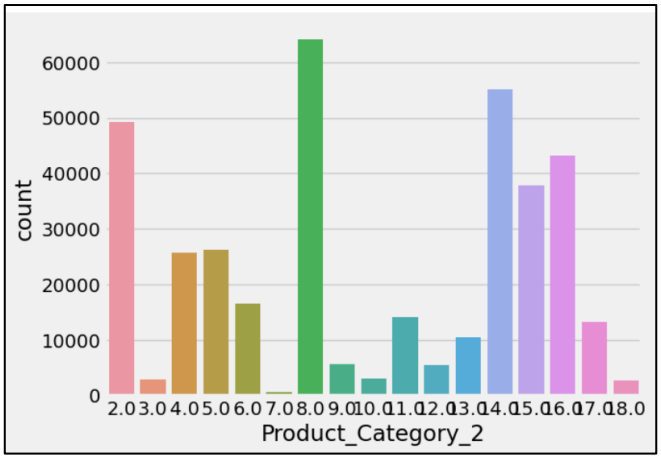


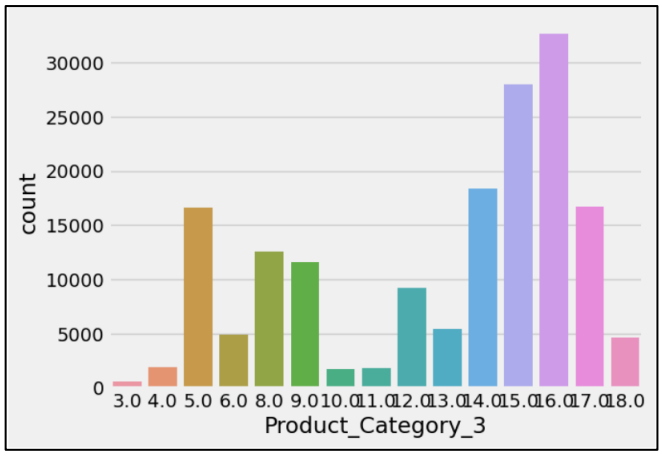


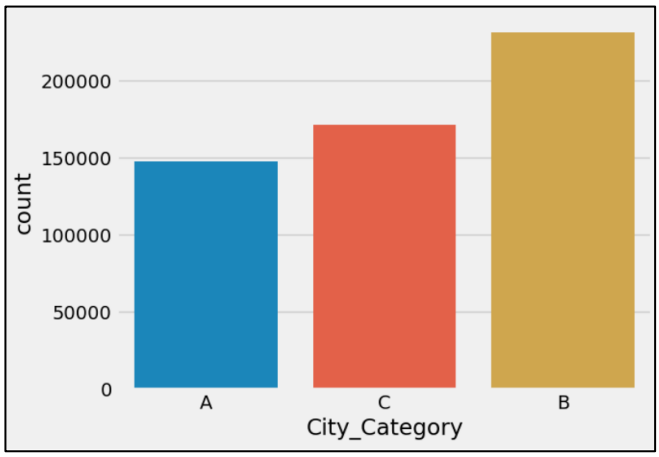


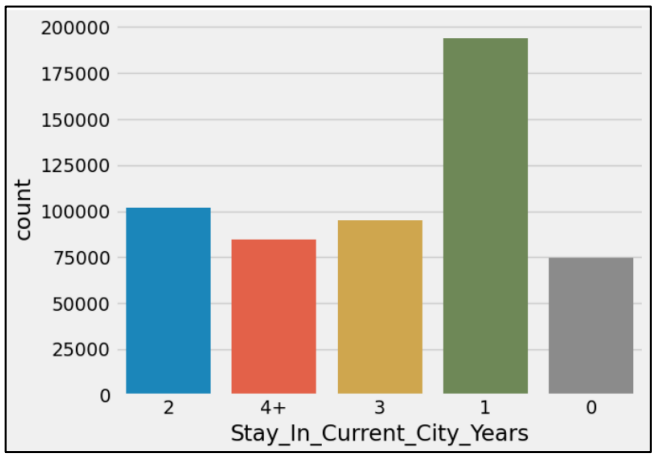


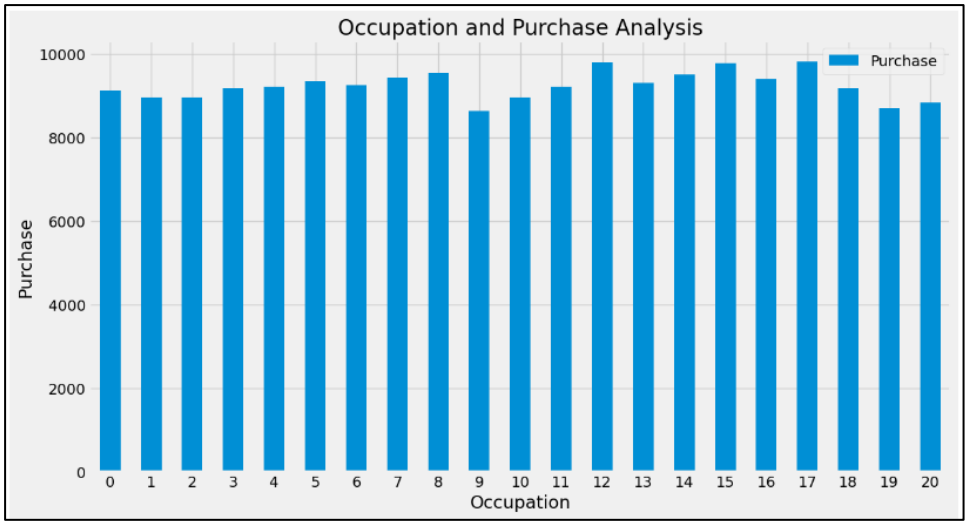


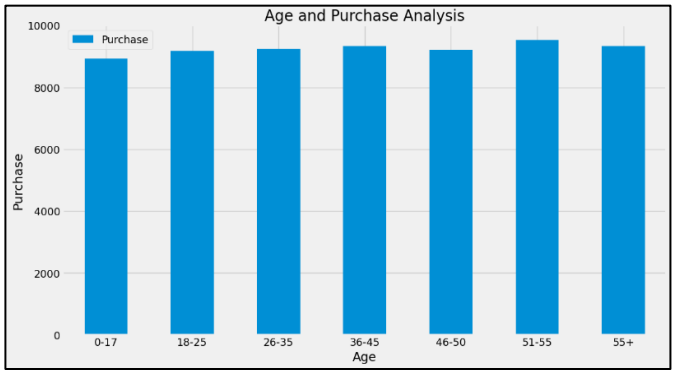

count
Product_Category_2
2.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0 10.0 11.0 12.0 13.0 14.0 15.0 16.0 17.0 18.0

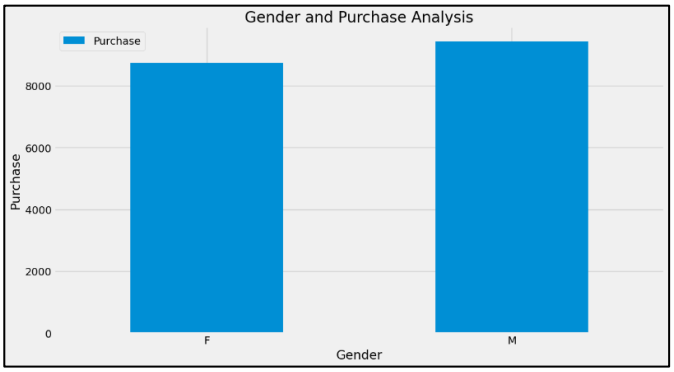

count
Product_Category_3
3.0 4.0 5.0 6.0 8.0 9.0 10.0 11.0 12.0 13.0 14.0 15.0 16.0 17.0 18.0

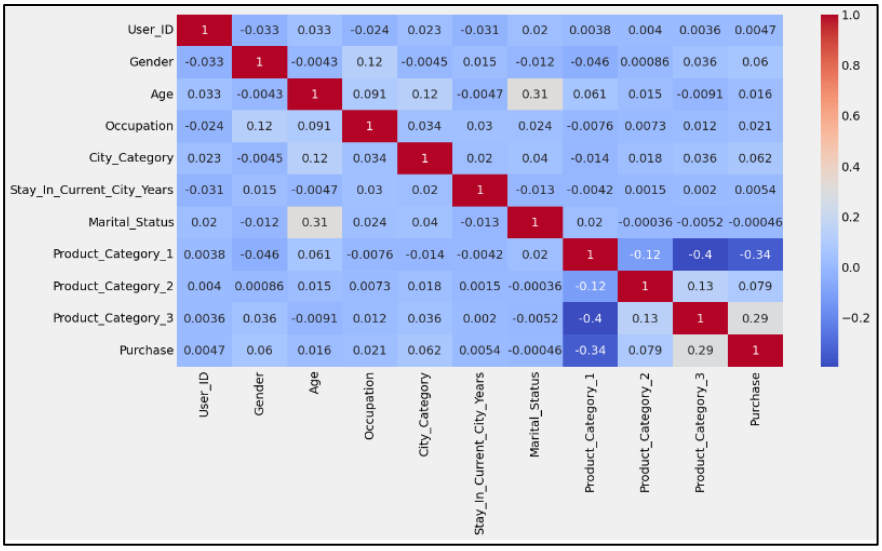

count
City_Category
A
C
B


count
Stay_In_Current_City_Years
2
4+
3
1
0


Occupation and Purchase Analysis
Purchase
Occupation


Age and Purchase Analysis
Purchase
Age
0-17
18-25
26-35
36-45
46-50
51-55
55+


Gender and Purchase Analysis
Purchase
Gender
F
M



5. MODEL CONSTRUCTION AND EVALUATION:

To train the dataset, we used the DecisionTreeRegressor model, and to assess the model's performance, we used metrics like MSE, RMSE, MAE, R-Squared, and Cross-Validation Score. The dataset was then trained using the XGBoost model, and the model's performance was assessed using the same metrics. Later two popular Python libraries were used to demonstrate linear regression and evaluate model performance.

6. CONCLUSION:

Analysis for Black Friday Sales Prediction using Linear Regression and Decision Tree Regression

The MSE, RMSE, MAE, and R-squared values for the decision tree regression using sklearn were 11251121.84, 3354.27, 2369.93, and 0.55 respectively. The RMSE scores for cross-validation were between 3330.44 and 3345.49. The MSE, RMSE, MAE, and R-squared values for the XGBoost multiple decision tree regression were 22634159.68, 4757.54, 3787.86, and 0.10, respectively. 2935.13 was the cross-validation RMSE score.

Similarly for Linear Regression, MSE, RMSE, and MAE values of about 21303239.68 and 3522.47 were obtained from the linear regression using the sklearn and statsmodels, respectively. The R-squared score, however, was low (0.15), indicating that the model did not adequately account for the data's variability. Additionally, negative and relatively high cross-validation scores showed the model's poor performance.

Overall, the cross-validation RMSE score for the decision tree regression using XGBoost was the lowest, indicating that it outperformed the other model.

DATASET 3: MALL CUSTOMER SEGMENTATION

Dataset Description:

The Mall Customer Segmentation dataset is a collection of data on mall patron behaviour. Customer ID, Gender, Age, Annual Income, and Spending Score are the five attributes shared by each of the 200 records in the collection, each of which belongs to a separate customer. The customer's gender is indicated by the Gender attribute. The Age attribute displays the customer's age in years. The Annual Income feature displays the customer's annual income in dollars and thousands. Based on a customer's purchasing habits and buying behaviour, the mall assigns a score, which is represented by the feature known as purchasing Score.

The dataset is commonly employed for consumer segmentation and clustering analysis. The process of grouping consumers based on their behaviour, demographics, and other relevant criteria is known as customer segmentation. Marketing techniques can be improved along with customer involvement when clustering research reveals patterns and linkages between consumer groupings. Utilizing the Mall consumer Segmentation dataset can help businesses better understand their target market and their marketing strategies.

1. IMPORTING LIBRARIES:

Importing the required libraries for the analysis is done in this stage. These include widely used libraries like Pandas, Numpy, Matplotlib, Scikit-learn, Scipy, and OpenCV.

2. DATA EXPLORATION:

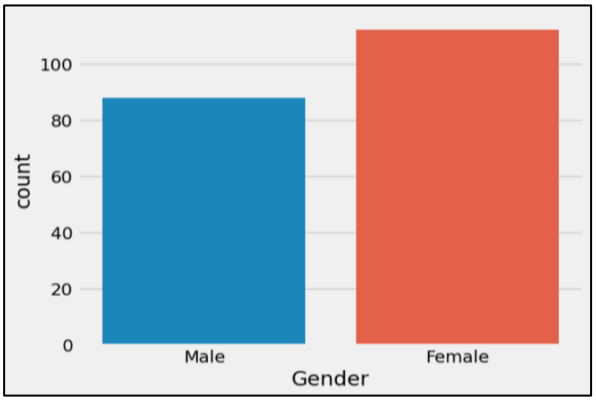
The dataset is loaded into the analysis code. The dataset is then put through data exploration, which entails looking at the dataset's dimensions, looking at the different variable kinds, and looking for missing data.

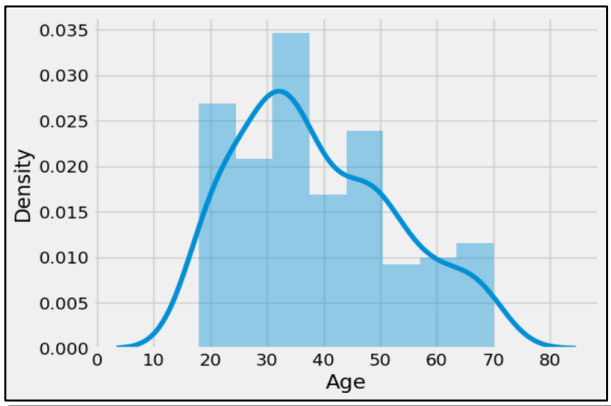
3. DATA PREPROCESSING:

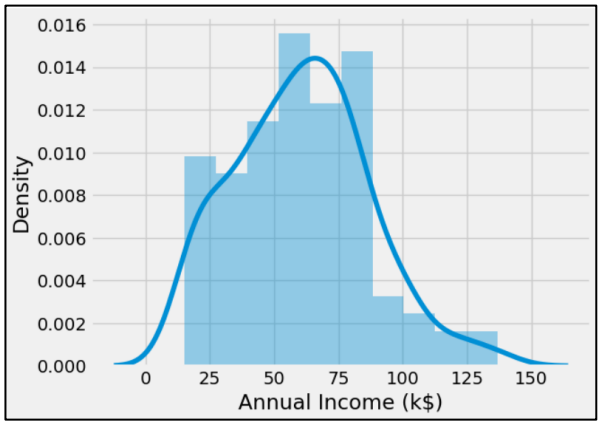
Data preprocessing includes feature selection, data cleaning, and data transformation, which is applied to the dataset. Duplicate values and missing values from the dataset are removed during data cleaning. While data transformation entails scaling the features to enhance the performance of the clustering algorithms, feature selection involves choosing the pertinent features for analysis.

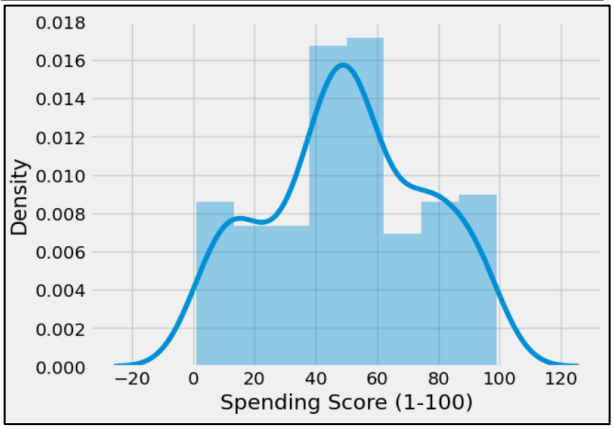
4. EXPLORATORY DATA ANALYSIS:

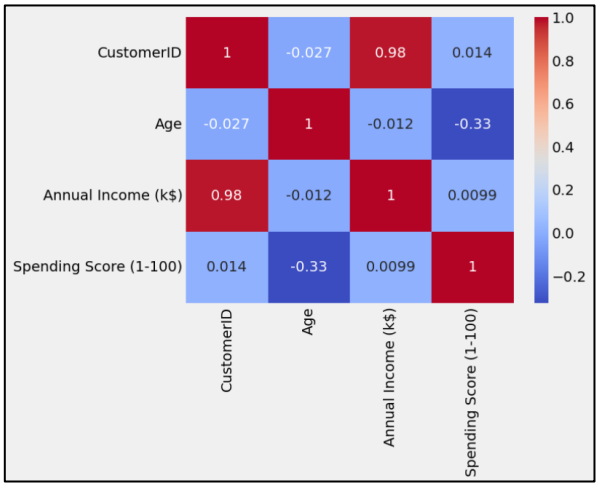
To learn more about the dataset, exploratory data analysis (EDA) was used. To comprehend the distribution pattern of the variables and spot any kind of trends in the data, descriptive statistics and data visualization techniques are utilized. Scattered plots, dendrograms, and silhouette plots are some of the visualizations.





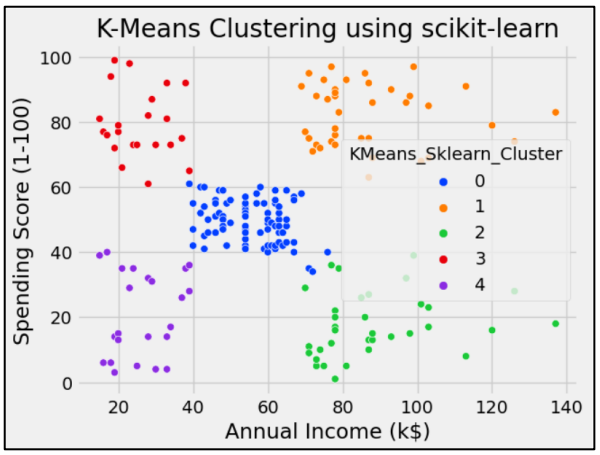


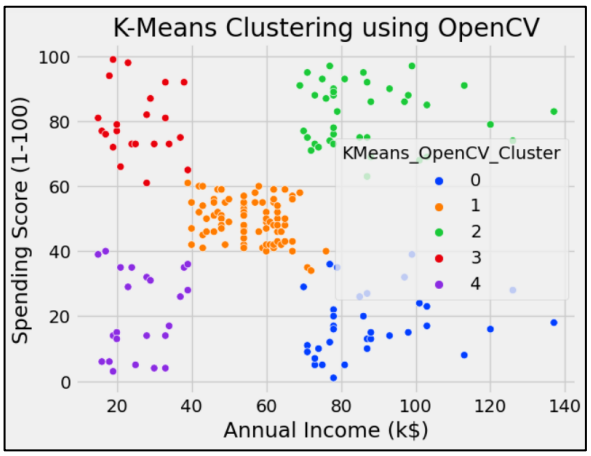


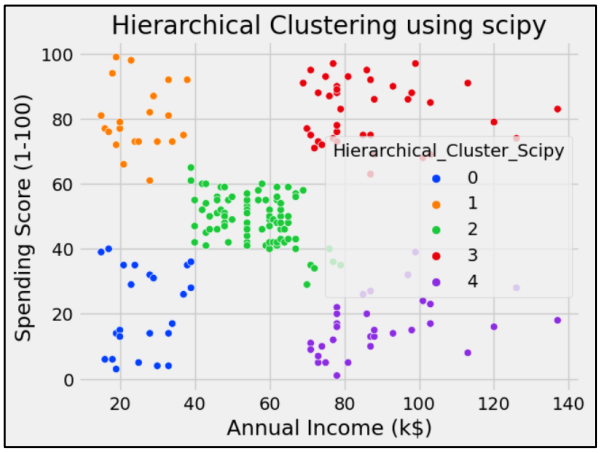


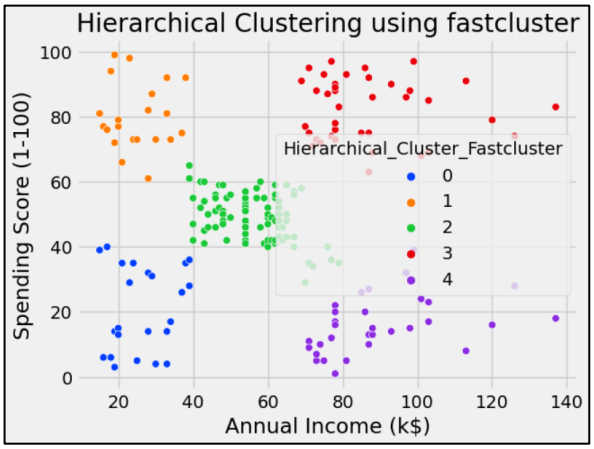
5. MODEL CONSTRUCTION AND EVALUATION:

The creation and evaluation of models are then carried out using three clustering algorithms: Hierarchical clustering with Scipy and Fastcluster, K-Means clustering with OpenCV, and K-Means clustering with Scikit-Learn. Each algorithm begins with the preparation of the data, the construction of the model, and an evaluation of the model's performance using a variety of metrics, including the silhouette score, the Calinski-Harabasz score, and the Davies-Bouldin score. To help understand how each model performed, the results are each visually represented.









6. CONCLUSION:

Analysis for Mall Customer Segmentation using K Mean and Hierarchical Clustering.

Segmenting customers depending on their spending habits is the aim, customers are clustered together using K-Means Clustering. K-Means clustering is carried out using the OpenCV and scikit-learn tools. With a Silhouette Score of 0.555 and a Calinski-Harabasz Index of around 248, both libraries generated comparable results.

Customers were also divided into clusters using hierarchical clustering. Hierarchical clustering is carried out using the scipy and fastcluster packages. With a Silhouette Score of roughly 0.555 and a Calinski-Harabasz Index of roughly 244, both libraries generated comparable outcomes.

K-Means and Hierarchical Clustering both yielded generally comparable outcomes. Both the clustering is deemed to be reasonably strong based on the Silhouette Score and Calinski-Harabasz Index.

IV. RELATED WORKS:

1. G. Gopinath and K. Shanthi's "BigMart Sales Prediction using Machine Learning Techniques" This study investigates how to estimate sales in BigMart stores using a variety of machine learning algorithms, including linear regression, decision trees, and random forests.

2. "BigMart Sales Prediction with Neural Networks" by S. S. Suresh and S. S. Jagadeesh. In this study, BigMart sales are predicted using neural networks, and the outcomes are compared to those from other machine learning models.

3. The article "BigMart Sales Prediction: A Comparative Study of Machine Learning Models" by P. N. Karthik, V. Chandrasekaran, and P. Balakrishnan. In order to anticipate BigMart sales, this article evaluates the effectiveness of various machine learning algorithms, such as gradient boosting and support vector regression.

4. "Black Friday Sales Prediction Using Ensemble Techniques" by Dr. K. Duraiswamy and V. Dhivya In order to anticipate Black Friday sales, this research compares many machine learning algorithms, including decision trees, k-nearest neighbors, and support vector regression.

5. R. Sivakumar and Dr. K. Thangadurai's "Black Friday Sales Prediction Using Machine Learning Techniques": In order to estimate Black Friday sales, this research suggests a hybrid approach combining decision trees and artificial neural networks.

6. K. Wang, Y. Wu, and C. Lin's article "A Novel Method for Black Friday Sales Prediction Based on User Behavior Analysis": This study presents a novel approach to predicting Black Friday sales that uses user behavior analysis and elements including clickstream data and browsing history.

7. Pattarawan Prasertwattana and Chatchai Ratanachai, "Customer Segmentation of Shopping Mall Visitors Using Hierarchical Clustering Analysis: A Case Study of the Bangkok Central Business District" (2019).

8. Srinivas R, H. S. Mohana, and J. Anuradha's article "Customer Segmentation for Mall Marketing Using RFM Analysis" was published in 2019.

9. Shubhanshu Mishra and Gaurav Kumar's "Mall Customer Segmentation Using Machine Learning Techniques" (2020).

10. A. Hossain, M. A. Hasan, and M. H. Kabir, "Predicting Black Friday Sales Using Multiple Linear Regression and Decision Tree": The efficacy of multiple linear regression and decision tree algorithms for predicting Black Friday sales is examined in this research along with that of other machine learning techniques.

V. FUTURE ENHANCEMENT:

1. Incorporate external data: Projects can be made better by using external data sources, such as weather information, economic indicators, and social media data, to boost the predictability of the results.

2. Investigate alternate machine-learning models: The projects can be enhanced by investigating alternative machine-learning models, such as Random Forest, Gradient Boosting, or Neural Networks, to see if they outperform the current models.

3. Feature engineering: This method involves transforming raw data into characteristics that can more accurately depict the underlying problem. By testing different feature engineering strategies to determine whether they can increase prediction accuracy, the projects can be made better.

4. Hyperparameter tweaking: The projects can be enhanced by hyperparameter tuning in order to find the optimal set of hyperparameters for the machine learning models.

5. Ensemble methods: These methods combine different machine learning models to boost the accuracy of the predictions. The projects can be enhanced by experimenting with ensemble techniques like stacking or bagging.

6. Online learning: Online learning is a method of machine learning that changes the model as fresh data are made available. By regularly updating the models and raising the forecast accuracy over time, online learning can be used to enhance the projects.

7. Explainability and Interpretability: By integrating explainability and interpretability approaches, projects can be made better by providing a clearer understanding of the factors affecting projections.

8. Real-time predictions: The projects can be improved by including real-time prediction skills to give clients prompt advice or to enhance inventory management.

VI. CONCLUSION:

Finally, machine learning initiatives on BigMart sales forecast, Black Friday sales prediction, and mall customer segmentation have demonstrated the promise of utilizing machine learning algorithms for solving challenging business challenges. These initiatives have shown how well machine learning methods like regression, clustering, and classification work at predicting sales, identifying client segments, and giving organizations insights.

Machine learning is used to predict the sales of various products based on numerous factors such as item weight, visibility, and store location, as demonstrated by the BigMart sales prediction project. The experiment has also made clear how crucial data preparation, preprocessing and model choice are for producing precise sales forecasts.

The research to predict Black Friday sales has shown how machine learning may be used to forecast the sales of various goods over the holiday season. In order to provide accurate sales estimates, the project has also underlined the value of feature selection, model evaluation, and ensemble approaches.

Machine learning may be used to identify multiple client categories based on numerous criteria such as age, income, and shopping habits, as demonstrated by the mall customer segmentation project. The initiative has also brought attention to the significance of feature scaling, data visualization, and model choice in attaining precise consumer segmentation.

Overall, this machine-learning project has shown how machine-learning approaches have the ability to solve complicated business problems and offer insightful data to enterprises. To increase the precision and scalability of these machine learning models, more study and development are still needed.

VII. BIBLIOGRAPHY:

1. Khan, N., Rahman, A., & Ullah, A. (2021). Big Mart Sales Prediction Using Machine Learning Techniques. International Journal of Emerging Trends in Engineering Research, 9(3), 1759-1763.

2. Verma, P., & Singh, R. (2020). Black Friday Sales Prediction using Machine Learning Algorithms. International Journal of Computer Applications, 179(33), 10-14.

3. Kumar, A., & Sharma, N. (2020). Customer Segmentation for Mall using Machine Learning. International Journal of Computer Science and Mobile Computing, 9(2), 32-39.

4. Jain, A., & Jain, A. (2020). Mall Customer Segmentation using Machine Learning. International Journal of Advanced Research in Computer Science and Software Engineering, 10(9), 325-329.

5. Patel, K., & Dave, D. (2019). Sales Prediction using Machine Learning Algorithm for BigMart. International Journal of Computer Applications, 181(22), 18-22.

6. Vignesh, M., Kumar, S. B., & Ramachandran, S. (2020). Predictive Modelling for Black Friday Sales using Machine Learning Techniques. Journal of Advanced Research in Dynamical and Control Systems, 12(Special Issue 7), 2567-2576.

7. Tripathi, R., Kumar, A., & Singh, A. (2020). Mall Customer Segmentation: A Review. International Journal of Computer Science and Mobile Computing, 9(2), 119-124.

The Project Code File is available at the following link:
Machine-Learning/BIGMART, BLACK FRIDAY AND MALL CUSTOMER SEGMENTATION.ipynb at main · annjoys-ncirl/Machine-Learning (github.com)