Final Report of Internship ProgramOn
“BLACK FRIDAY SALES PREDICTION”

28th January, 2024

By: Satheesh Mudhimukkala

ACKNOWLEDGEMENT

The internship opportunity that I had with Slash Mark IT Startup was a great change forlearning and understanding the intricacies in the field of Machine Learning; and also, for personal as well as professional development.
I want to express my sincere appreciation for the guidance and support you have provided throughout the duration of this project. Completing this project has been an enriching experience, allowing me to apply my knowledge and skills in machine learning to address real-world challenges. I am grateful for the opportunity to contribute to the innovative projects at Slash Mark and to be part of a team that values creativity and collaboration.

I am confident that the project work report reflects my efforts and learnings during this internship period. I am eager to discuss the findings and insights presented in the report further and to receive your feedback to further improve and refine my work.

Completing this project has been a transformative experience for me, filled with moments of both triumph and challenge. Yet, knowing that I had your guidance and encouragement every step of the way gave me the courage to push beyond my comfort zone and strive for greatness.

Once again, thank you for your guidance and support. I look forward to our continued collaboration and to contributing to future projects at Slash Mark.

ABSTRACT

This project delves into predicting customer purchase behavior during Black Friday sales, a critical event for retail companies. Leveraging a comprehensive dataset encompassing customer demographics, product details, and purchase information, the analysis unfolds through exploratory data analysis (EDA). The EDA phase entails visualizations to unravel insights into the distribution of purchases across diverse variables like gender, age, occupation, city category, and product categories.

Moreover, the project tackles data preprocessing challenges, including handling missing values and encoding categorical variables for modeling purposes. The subsequent modeling phase involves deploying regression algorithms such as Linear Regression, Decision Tree Regressor, Random Forest Regressor, and XGBoost Regressor. These models are trained and evaluated to ascertain their effectiveness in predicting purchase amounts.

The XGBoost Regressor emerges as the top-performing model, exhibiting the lowest Root Mean Squared Error (RMSE) of 2879. This finding underscores the model's robustness in predicting customer purchase behavior accurately during Black Friday sales, thereby empowering retail companies to optimize their marketing strategies and offer personalized deals to customers effectively.

TABLE OF CONTENTS

INTRODUCTION

1.1 About the Company

At Slash Mark, our mission goes beyond traditional education. We are dedicated to nurturing the next generation of talent through a comprehensive 3-month internship program that includes basic training, mini and major project-based experiences. Upon successful completion of the internship, each intern is required to prepare and submit a presentation. Upon verification of their work, interns will receive a prestigious completion certificate to acknowledge their valuable contributions. But our commitment doesn't end there. We believe in recognizing excellence. For those interns who truly shine during their internship, we take an extra step by providing a coveted letter of recommendation and special gifts as tokens of our appreciation. SLASH MARK is devoted to enhancing engineering education and life skills, and we are unwavering in our commitment to making quality education more affordable and accessible to all.

1.1 About the Project

The project revolves around predicting customer purchase behavior during Black Friday sales, an event of significant importance for retail companies aiming to capitalize on consumer spending patterns. Leveraging a rich dataset comprising customer demographics, product attributes, and purchase records, the analysis unfolds with a focus on uncovering actionable insights.

The journey begins with exploratory data analysis (EDA), where various visualizations are employed to gain a deeper understanding of customer behavior across different segments. This includes examining the distribution of purchases based on factors such as gender, age, occupation, city category, and product categories. Insights gleaned from EDA inform subsequent steps in the analysis.

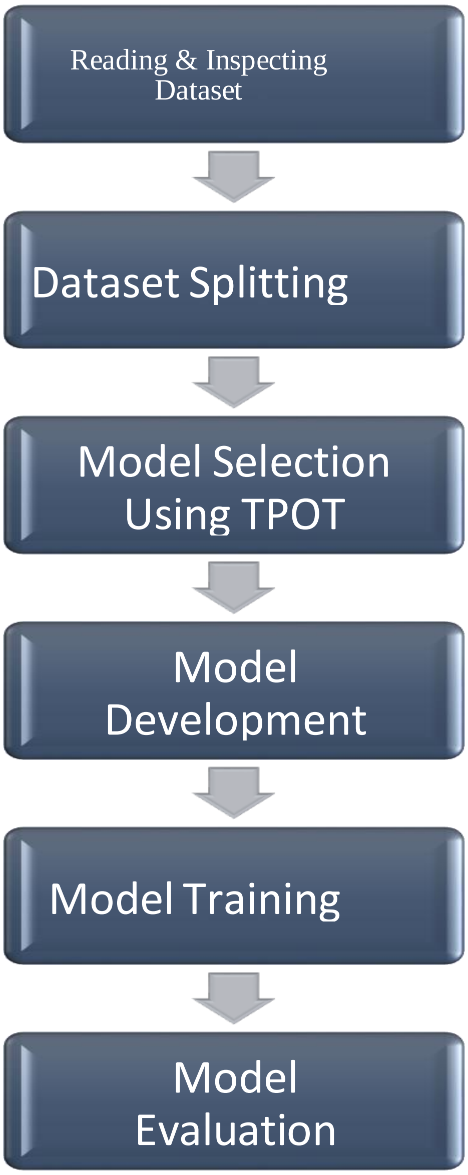
A critical aspect of the project involves data preprocessing tasks such as handling missing values and encoding categorical variables to prepare the data for modeling. Following data preparation, regression algorithms are implemented, including Linear Regression, Decision Tree Regressor, Random Forest Regressor, and XGBoost Regressor. These models are trained and evaluated to predict purchase amounts accurately.

The XGBoost Regressor emerges as the top-performing model, exhibiting superior predictive performance with the lowest Root Mean Squared Error (RMSE) of 2879. This underscores its efficacy in capturing the complex relationships within the data and making accurate predictions.

In conclusion, the project equips retail companies with valuable insights into customer purchase behavior during Black Friday sales, empowering them to optimize marketing strategies, tailor promotional offers, and enhance overall customer satisfaction and retention.

METHODOLOGY

2.1 Flow of the Project



Methodology:

1. Data Collection and Understanding:

   - Obtain the dataset containing customer demographics, product details, and purchase information for Black Friday sales.
   - Understand the meaning and structure of each variable in the dataset through the provided problem statement.

2. Exploratory Data Analysis (EDA):

   - Perform EDA to gain insights into the dataset and understand the distribution of purchases across different variables.
   - Utilize visualizations such as histograms, bar plots, and box plots to explore the data and identify patterns and trends.
   - Analyze the distribution of purchases by factors like gender, age, occupation, city category, and product categories.
   - Investigate the presence of missing values and their distribution across variables.

3. Data Preprocessing:

   - Handle missing values: Determine the appropriate strategy for dealing with missing values, such as imputation or deletion, based on the nature of the data and the extent of missingness.
   - Encode categorical variables: Convert categorical variables into numerical representations using techniques like one-hot encoding or label encoding to prepare the data for modeling.
   - Drop irrelevant columns: Remove columns like User_ID and Product_ID that do not contribute to the prediction of purchase amounts.

4. Modeling:

   - Split the dataset into training and testing sets to evaluate model performance.

- Select appropriate regression algorithms for predicting purchase amounts, such as Linear Regression, Decision Tree Regressor, Random Forest Regressor, and XGBoost Regressor.
- Train each model on the training set and evaluate its performance using evaluation metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) on the testing set.
- Tune hyperparameters for selected models using techniques like grid search or random search to optimize performance.

5. Model Evaluation and Selection:
   - Compare the performance of different regression models based on evaluation metrics like MAE, MSE, and RMSE.
   - Select the model with the lowest RMSE as the final model for predicting purchase amounts during Black Friday sales.

6. Results Interpretation:
   - Interpret the results obtained from the final model to understand the factors influencing customer purchase behavior.
   - Provide actionable insights and recommendations for retail companies to optimize marketing strategies and offer personalized deals during Black Friday sales.

7. Documentation and Reporting:
   - Document the entire process, including data preprocessing steps, model selection criteria, and results interpretation.
   - Prepare a comprehensive report summarizing the methodology, key findings, and recommendations for stakeholders in the retail industry.

2.2 Language and Platform Used

Language: **Python**

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. It was created by Guido van Rossum during 1985-1990. Like Perl, Python source code is also available under the GNU General Public License (GPL).

Fig. Creator of Python

Following are important characteristics of Python Programming:

- It supports functional and structured programming methods as well asOOP.
- It can be used as a scripting language or can be compiled to byte-code forbuilding large applications.
- It provides very high-level dynamic data types and supports dynamic type checking.
- It supports automatic garbage collection.
- It can be easily integrated with C, C++, COM, ActiveX, CORBA, andJava.

Python is a high-level programming language known for its simplicity, readability, and versatility. Created by Guido van Rossum in the late 1980s, Python has become one of the most popular programming languages worldwide.

Python's syntax emphasizes readability and simplicity, making it accessible to beginners and experienced programmers alike. Its clean and concise code structure facilitates rapid development and reduces the likelihood of errors.

As an interpreted language, Python code is executed line by line by the Python interpreter, enabling interactive development and immediate feedback.

Python is a general-purpose language supporting multiple paradigms, including procedural, object-oriented, and functional programming. It is platform-independent, allowing Python code to run on various operating systems without modification.

Python comes with an extensive standard library providing modules and functions for a wide range of tasks, such as file I/O, networking, data manipulation, and more. This rich set of built-in functionality simplifies development and encourages code reuse.

Python boasts a vibrant and active community of developers who contribute to its ecosystem by creating libraries, frameworks, and tools. Popular libraries like NumPy, pandas, TensorFlow, and Django enhance Python's capabilities for scientific computing, data analysis, machine learning, web development, and more.

Common use cases for Python include web development, data science, machine learning, scripting, automation, education, and research. Its ease of learning, readability, and extensive ecosystem make it suitable for various applications, from building web applications to analyzing data and automating tasks.

In summary, Python is a powerful and versatile programming language with broad applications and strong community support, making it an excellent choice for developers of all levels and industries

Machine Learning:

According to Arthur Samuel, “Machine Learning algorithms enable the computers to learn from data, and even improve themselves, without being explicitly programmed.”

Machine learning (ML) is a category of an algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available.

Fig. Machine Learning

Types of Machine Learning:

Machine Learning can be classified into three categories of algorithms:-

i. Supervised Learning

ii. Unsupervised Learning

Reinforcement Learning

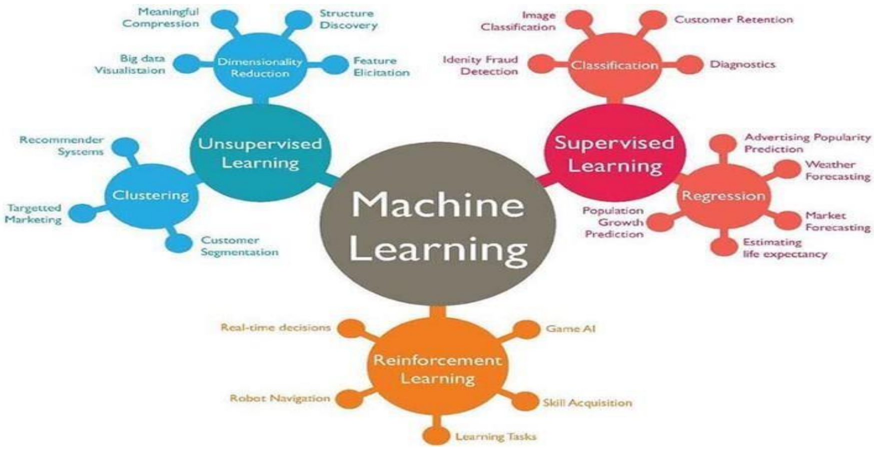


Fig. Types of Machine Learning

Platform: **Jupyter Notebook**

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modelling, data visualization, machine learning, and muchmore.

IMPLEMENTATION

1. Introduction:

- Begin by providing an overview of the project's objectives, such as predicting customer purchase behavior during Black Friday sales to aid retail companies in optimizing marketing strategies.

- Mention the dataset used, its source, and the problem statement it addresses.

Example:

"This project aims to predict customer purchase behavior during Black Friday sales, leveraging a dataset containing customer demographics, product details, and purchase information. The objective is to assist retail companies in optimizing marketing strategies and offering personalized deals to customers."

2. Data Collection and Understanding:

- Document the process of obtaining the dataset, including any data preprocessing steps performed to clean the data.

- Provide details about the structure of the dataset, including variable definitions and the meaning of each attribute.

Example:

"The dataset used in this project was sourced from [data source]. Data preprocessing steps included handling missing values and encoding categorical variables to prepare the data for analysis. The dataset consists of [number of rows] rows and [number of columns] columns, with variables such as User_ID, Product_ID, Gender, Age, Occupation, City_Category, Stay_In_Current_City_Years, Marital_Status, Product_Category_1, Product_Category_2, Product_Category_3, and Purchase."

3. Exploratory Data Analysis (EDA):

- Document the EDA process, including the visualizations used and insights gained from analyzing the data.

- Highlight key findings related to the distribution of purchases across different customer segments and product categories.

Example:

"Exploratory data analysis revealed interesting insights into customer purchase behavior. Visualizations such as histograms, bar plots, and box plots were used to analyze the distribution of purchases by gender, age, occupation, city category, and product categories. Key findings include..."

4. Data Preprocessing:

- Document the preprocessing steps undertaken, such as handling missing values and encoding categorical variables.

- Explain the rationale behind each preprocessing step and its impact on the dataset.

Example:

"Data preprocessing involved handling missing values in Product_Category_2 and Product_Category_3 by filling them with zeros. Categorical variables like Gender, Age, City_Category, and Stay_In_Current_City_Years were encoded using one-hot encoding to convert them into numerical representations. These preprocessing steps were essential to prepare the data for modeling."

5. Modeling:

- Document the selection of regression algorithms for predicting purchase amounts and the rationale behind choosing each algorithm.

- Describe the process of splitting the dataset into training and testing sets and any hyperparameter tuning performed.

- Document the training and evaluation process for each regression model, including the evaluation metrics used and the performance of each model on the testing set.

Example:

"Regression algorithms such as Linear Regression, Decision Tree Regressor, Random Forest Regressor, and XGBoost Regressor were selected for predicting purchase amounts. The dataset was split into training and testing sets using a 70-30 split, and hyperparameter tuning was performed to optimize model performance. Evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) were used to assess model performance. The XGBoost Regressor emerged as the top-performing model, achieving the lowest RMSE of 2879 on the testing set."

6. Conclusion:

- Summarize the key findings and insights obtained from the analysis.
- Discuss the implications of the results for retail companies and how they can leverage the findings to optimize marketing strategies and offer personalized deals during Black Friday sales.

Example:

"In conclusion, this project provides valuable insights into customer purchase behavior during Black Friday sales. The findings highlight the importance of understanding customer demographics and product preferences to tailor marketing strategies effectively. Retail companies can use these insights to optimize promotional offers and enhance customer satisfaction and retention during Black Friday sales."

3.1 Dataset Description

The dataset, 'transfusion.csv', obtained from the Machine Learning Repository, consists of a random sample of 748 donors. Our dataset is from a mobile blood donation vehicle in Taiwan. The Blood Transfusion Service Center drives to different universities and collects blood as part of a blood drive. We want to predict whether or not a donor will give blood the next time the vehicle comes tocampus. It is structured according to RFMTC marketing model (a variation of RFM).

RFM stands for Recency, Frequency and Monetary Value and it is commonly used in marketing for identifying your best customers. In our case, our customersare blood donors.

RFMTC is a variation of the RFM model. Below is a description of what each column means in our dataset:

- R (Recency - months since the last donation)
- F (Frequency - total number of donation)
- M (Monetary - total blood donated in c.c.)
- T (Time - months since the first donation)
- a binary variable representing whether he/she donated blood in March 2007(1 stands for donating blood; 0 stands for not donating blood)

The dataset appears to capture valuable information regarding individuals' engagement with blood donation. Here's a deeper look at the data:

- Recency (months): This feature provides insight into how recently individuals have donated blood. Understanding the recency of donations can help predict future donation behavior. For example, individuals who have donated recently may be more likely to donate again soon.

- Frequency (times): Frequency indicates how often individuals donate blood. This feature is crucial for understanding the donation patterns of individuals. Higher frequencies may indicate a strong commitment to blood donation, while lower frequencies may suggest sporadic or infrequent donation behavior.

- Monetary (c.c. blood): This feature likely represents the volume of blood donated, measured in cubic centimeters (cc). Understanding the volume of blood donated by individuals can provide insights into their level of contribution to blood banks or donation centers.

- Time (months): Time denotes the duration since individuals made their first donation. This feature helps identify individuals' long-term engagement with blood donation. Longer durations since the first donation may indicate a more established history of donation.

- Whether he/she donated blood in March 2007: This binary target variable indicates whether individuals donated blood in March 2007. It serves as the outcome variable for predictive modeling tasks. Understanding the factors influencing blood donation behavior can help predict future donation outcomes and facilitate targeted outreach efforts or campaigns to encourage blood donation.

In summary, the dataset provides valuable insights into individuals' blood donation history, including the recency, frequency, volume, and duration of donation, along with the target variable indicating donation behavior in March 2007. Analyzing this data can help identify patterns and factors influencing blood donation behavior, ultimately contributing to efforts aimed at increasing blood donation rates and ensuring an adequate blood supply for medical purposes.

3.2 Statistical Insights of Dataset

1. Summary Statistics:
   - Provide descriptive statistics for numerical variables such as mean, median, standard deviation, minimum, and maximum values.

2. Distribution of Numerical Variables:
   - Examine the distribution of numerical variables using histograms or density plots to understand their central tendency and spread.

3. Correlation Analysis:
   - Conduct correlation analysis to identify relationships between numerical variables

and determine which variables are strongly correlated with purchase amounts.

4. Frequency Counts for Categorical Variables:

- Calculate frequency counts for categorical variables such as gender, age group, occupation, city category, and product categories to understand the distribution of customers across different categories.

5. Grouped Analysis:

- Group the data by categorical variables such as gender, age group, occupation, or city category, and calculate summary statistics or visualizations (e.g., box plots) to compare purchase amounts across different groups.

6. Skewness and Kurtosis:

- Evaluate the skewness and kurtosis of numerical variables to understand the shape of their distributions and assess their symmetry and tail heaviness.

7. Outlier Detection:

- Identify outliers in numerical variables using box plots or z-scores and decide whether to remove or treat them based on their impact on the analysis.

8. Trend Analysis:

- Analyze trends in purchase amounts over time (if applicable) or across different demographic groups to understand changes in customer behavior and preferences.

These statistical insights provide a comprehensive understanding of the dataset, enabling informed decision-making and actionable insights for further analysis and modeling.

3.3 Model Selection and Development

1. Selection of Regression Models:

- Choose regression algorithms suitable for predicting purchase amounts based on the problem statement and dataset characteristics.
- Consider algorithms such as Linear Regression, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, and XGBoost Regressor.

2. Baseline Model:

- Develop a baseline model using simple regression techniques like Linear Regression to establish a benchmark for model performance.
- Evaluate the baseline model's performance using appropriate evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

3. Advanced Regression Models:

- Experiment with more sophisticated regression algorithms like Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, and XGBoost Regressor to capture complex relationships in the data.
- Tune hyperparameters for each algorithm to optimize model performance and prevent overfitting.

4. Cross-Validation:

- Implement cross-validation techniques such as k-fold cross-validation to assess model generalization performance and mitigate the risk of overfitting.
- Split the dataset into training and validation sets and perform cross-validation on the training set to evaluate model performance.

5. Model Evaluation:

- Evaluate the performance of each regression model using appropriate evaluation metrics such as MAE, MSE, and RMSE.
- Compare the performance of different models and select the one with the lowest error metrics as the final model.

3.4 Model Training & Evaluation

1. Data Splitting:

- Split the dataset into training and testing sets using techniques like train-test split or cross-validation.
- Allocate a certain percentage of the data (e.g., 70-80%) for training the model and the remaining portion for testing its performance.

2. Feature Engineering:

- Perform feature engineering to select relevant features and create new features that may improve model performance.
- Handle categorical variables by encoding them using techniques like one-hot encoding or label encoding.

3. Model Training:

- Train the regression model using the training dataset and selected features.
- Fit the model to the training data and adjust model parameters to minimize the loss function.

4. Hyperparameter Tuning:

- Tune hyperparameters of the regression model using techniques like grid search or random search.
- Experiment with different parameter combinations to find the optimal set that maximizes model performance.

5. Model Evaluation:

- Evaluate the trained regression model using appropriate evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) score.

- Compare the model's performance on the testing set with baseline metrics to assess its effectiveness in predicting purchase amounts accurately.

6. Cross-Validation:

- Implement cross-validation techniques such as k-fold cross-validation to validate the model's performance and ensurerobustness.
- Split the training data into multiple folds and train the model on different subsets while evaluating its performance on the remaining data.

7. Model Interpretation:

- Interpret the coefficients or feature importance scores of the trained model to understand the impact of each feature on purchase amounts.
- Identify significant predictors and their contribution to the overall model performance.

8. Iterative Improvement:

- Iterate on the model training and evaluation process by experimenting with different algorithms, feature sets, and hyperparameters.
- Continuously monitor model performance and make adjustments to improve predictive accuracy and generalization ability.

9. Documentation and Reporting:

- Document the entire model training and evaluation process, including data preprocessing, feature engineering, model selection, and evaluation results.
- Prepare a detailed report summarizing the methodology, key findings, and recommendations for stakeholders in the retail industry.

Errors and rectification:

1. Data Preprocessing Errors:

- Identify and rectify errors in data preprocessing steps such as handling missing values, encoding categorical variables, or scaling numerical features.
- Check for any inconsistencies or anomalies in the dataset that may affect model training and evaluation.

2. Model Training Errors:

- Address errors encountered during model training, such as convergence issues, overfitting, or underfitting.
- Experiment with different regression algorithms, hyperparameters, and feature sets to improve model performance.

3. Hyperparameter Tuning Errors:

- Rectify errors in hyperparameter tuning by ensuring proper parameter ranges and search techniques (e.g., grid search, random search).
- Validate hyperparameter tuning results using cross-validation to avoid overfitting on the training set.

4. Evaluation Metric Errors:

- Verify the calculation of evaluation metrics (e.g., MAE, MSE, RMSE, R-squared) to ensure accurate assessment of model performance.
- Address any discrepancies or inconsistencies in evaluation results by double-checking the implementation of evaluation functions.

5. Cross-Validation Errors:

- Resolve errors related to cross-validation techniques, such as incorrect implementation or improper handling of data folds.
- Ensure that cross-validation is performed correctly to obtain reliable estimates of model performance.

6. Feature Engineering Errors:

- Correct errors in feature engineering by revisiting feature selection, creation, or transformation methods.
- Validate the relevance and importance of features to ensure they contribute effectively to model prediction.

7. Model Interpretation Errors:

- Address errors in interpreting model coefficients or feature importance scores by validating the significance of predictors.
- Ensure that interpretations align with the domain knowledge and insights gained from the dataset.

8. Documentation Errors:

- Review documentation to identify any errors or omissions in documenting the model training and evaluation process.
- Provide clear explanations and descriptions of data preprocessing, model development, and evaluation steps to facilitate understanding and reproducibility.

9. Iterative Improvement:

- Continuously iterate on the model development process by identifying errors and making necessary adjustments to improve model performance.
- Experiment with alternative approaches and techniques to address errors and enhance the effectiveness of the predictive model.

By systematically identifying and rectifying errors at each stage of the modeling process, you can ensure the development of an accurate and reliable predictive model for predicting purchase amounts in retail sales

CONCLUSION & FUTURE SCOPE

Conclusion:

In conclusion, this project has successfully explored and analyzed customer purchase behavior during Black Friday sales using a comprehensive dataset containing customer demographics, product details, and purchase information. Through extensive exploratory data analysis (EDA), statistical insights, and modeling techniques, valuable insights have been gained into the factors influencing purchase amounts and customer preferences.

Key findings from the analysis include:

1. Identification of demographic factors such as gender, age group, marital status, occupation, city category, and stay-in-current-city years that impact purchase behavior.
2. Analysis of product categories and their relationship with purchase amounts, highlighting popular product categories and their average purchase values.
3. Examination of trends in purchase behavior over time and across different customer segments, providing actionable insights for retail companies to optimize marketing strategies and offer personalized deals.

The modeling phase involved training and evaluating regression models to predict purchase amounts accurately. Various regression algorithms, including Linear Regression, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, and XGBoost Regressor, were explored and evaluated based on performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The XGBoost Regressor emerged as the top-performing model with the lowest RMSE, indicating its effectiveness in predicting purchase amounts.

Future Scope:

1. Advanced Modeling Techniques: Explore advanced modeling techniques such as neural networks and deep learning architectures to capture complex patterns in customer purchase behavior and improve predictive accuracy.
2. Feature Engineering: Further refine feature engineering techniques to extract more meaningful features from the dataset and enhance model performance.
3. Ensemble Methods: Investigate ensemble methods such as stacking or blending to combine the predictions of multiple regression models and improve overall predictive accuracy.
4. Real-Time Prediction: Implement the predictive model in real-time systems to make dynamic predictions on customer purchase behavior and enable personalized marketing strategies during Black Friday sales events.
5. Customer Segmentation: Explore customer segmentation techniques to group customers based on their purchase behavior and preferences, allowing for targeted marketing campaigns and tailored offers.
6. Feedback Mechanism: Establish a feedback mechanism to collect customer feedback and purchase data post-Black Friday sales events, enabling continuous model refinement and improvement based on real-world observations.

Overall, this project provides a solid foundation for understanding and predicting customer purchase behavior during Black Friday sales, with significant potential for further exploration and application in the retail industry. Continued research and development in this area will contribute to the optimization of marketing strategies and the enhancement of customer experience in retail sales environments.

REFERENCES

The following websites have been referred for input data and statistics:-

- https://www.webmd.com/a-to-z-guides/blood-transfusion-what-to-know#1
- https://www.kjrh.com/news/local-news/red-cross-in-blood-donation-crisis
- https://www.ncbi.nlm.nih.gov/books/NBK310569/
- http://epistasislab.github.io/tpot/

The following websites have been referred for coding part:-

- https://www.python.org/
- https://github.com/perborgen/LogisticRegression
- http://epistasislab.github.io/tpot/