**Milestone 3: Final Draft Capstone**

**Predicting E-Commerce Customers' Behavior Using Machine Learning Algorithms**

**Assessment Description**

**The goal of Milestone 3 is to develop/implement a solution for a known problem, issue, or business case.**

**Complete the final draft of Milestone 3 as outlined within the "Capstone Project Handbook: Masters of Science in Data Science," located on the College of Science, Engineering and Technology page in the Student Success Center.**

**Assignment deliverables will depend on the project. All elements in need of revision, as decided in the Capstone Completion Plan, should be updated. The instructor will provide additional feedback for implementation prior to the submission of the completed Capstone.**

**This assignment uses a rubric. Please review the rubric prior to beginning the assignment to become familiar with the expectations for successful completion.**

**Rubric Criteria**

**Application Functionality and Execution**

**Source Code Listing**

**Industry Terminology**

**Code Review Functional**

**Requirements**

**System Entities**

**User Guide**

**Implementation Plan**

**Detailed Model Pipeline Design: Dataset Types and Formatting Industry Terminology**

**=======================================================**

**Milestone 3: Final Draft Capstone**

Predicting E-Commerce Customers' Behavior Using Machine Learning Algorithms

Jean-Marie Nfundiko Katula

MS Data Science - Grand Canyon University

DSC-590: Data Science Capstone Project

Dr. Dwight Farris

Jun 19, 2024

**Predicting E-Commerce Customers' Behavior Using Machine Learning Algorithms**

**Introduction**

This document presents the final draft of Milestone 3 for the Master of Science in Data Science Capstone Project. The project aims to develop and implement a solution for predicting e-commerce customer behavior using machine learning algorithms. This solution will help businesses understand customer preferences, optimize marketing strategies, and improve overall customer retention.

**Project Description**

The e-commerce industry continues to grow rapidly, making it crucial for businesses to understand and predict customer behavior. By implementing advanced machine learning algorithms to gather customers’ online purchase behaviors and perform analytics, e-commerce firms will improve their market targeting strategies, resulting to greater ROI. Leveraging the machine learning algorithms will also enable e-commerce businesses to personalize the customer experience and maximize revenue.

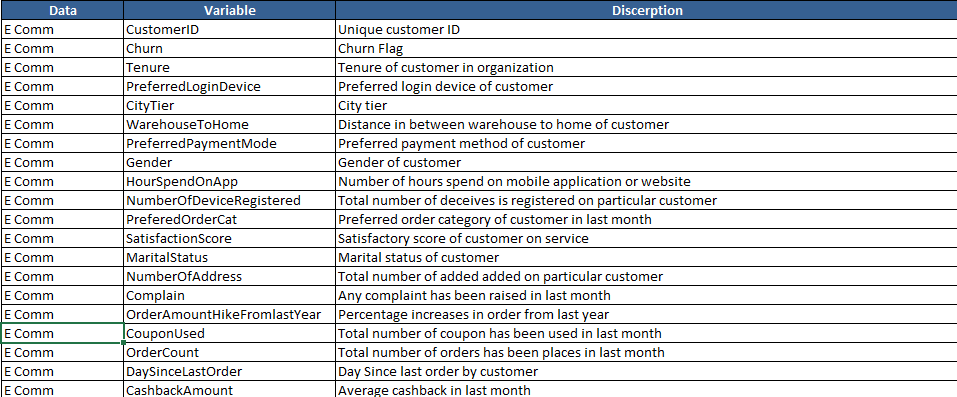
The demographic, purchase history and product information data are essential for predicting the future customer behaviors accurately. Acquiring and analyzing the consumers behavior data using the machine learning algorithms enables e-commerce businesses to predict accurately their future purchase patterns and preferences. As a result, these firms gain crucial insights about their customers’ expectations and preferences, thus enhancing the operational and marketing strategies to cater to their needs and quickly adapt to the dynamic marketplace. In 2022, global e-commerce sales surpassed $5.2 trillion. This was attributed by the growing number of online shoppers, with mobile shoppers attributing to 41.8% of all global retail e-commerce sales in 2022. This shows that many consumers are adopting the mobile approach for their online purchase.

During online purchase, factors such as quality of services, user experience, and company relationship with customers influence the customer churn rate. Using machine learning algorithms to gather and analyze the customer shopping behaviors enable e-commerce firms to establish specific factors that cause customers to stop using their products or services during specific periods. For instance, about 70.19% of online shopping carts are usually abandoned annually. This illustrates the need for e-commerce companies to optimize user experience to drive conversions and minimize customer churn rate. Therefore, this machine learning algorithm project focuses on accurately predicting future purchases, customer churn, and other actions by analyzing historical customer data and demographics.

**Methodology**

**Data Acquisition and Preprocessing:**

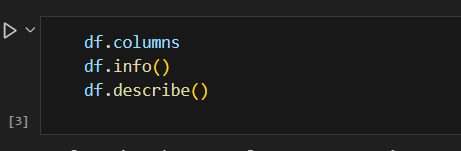
§ Collect historical e-commerce customer data, including purchase history, demographics, and product information.

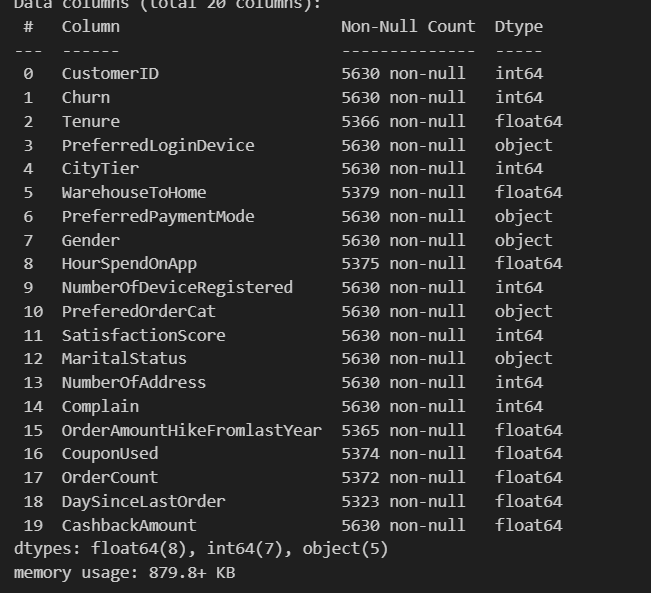


§ Preprocess the data to handle missing values, outliers, and format inconsistencies.

Importance

Preprocessing data is crucial because it improves model accuracy, enhances data quality, facilitates better insights, and increases efficiency.





**Feature Engineering / EDA:**

§ Perform EDA.

Importance:

Exploratory Data Analysis (EDA) is essential for understanding data, detecting anomalies, assessing quality, selecting features, and informing further analysis. It involves summarizing data, visualizing distributions and relationships, and ensuring model assumptions are met. By uncovering patterns and insights, EDA sets the foundation for effective modeling.

**Smooth Transitions**

1. **Data Collection to EDA**:
   * **Summary**: Highlight data sources and volume.
   * **Link**: "Understanding the collected data through EDA is crucial before meaningful analysis."
2. **Feature Selection to Model Training**:
   * **Summary**: Emphasize criteria for feature selection.
   * **Link**: "Selected features directly impact model choice and performance."

**Question 1:**

***\*\**\*\*1.1 Does the gender of a customer have a significant impact on their likelihood to churn?\*\*\*\***

**Answer 1:**

Understanding the impact of gender on churn can provide insights into the factors that influence customer satisfaction and loyalty. This information can help the company tailor their marketing and customer service strategies to better meet the needs of male and female customers. **A screen shot of a computer program

Description automatically generated**

**A graph with blue and orange bars

Description automatically generated**

**The chart above shows that the likelihood of churn by gender is very similar.**

**Question 2:**

***\*\**\*\*1.2 Does a customer's tenure have a significant impact on their likelihood to churn?\*\*\*\***

**Answer 2:**

Knowing the relationship between tenure and churn can help the company identify which customers are most at risk of leaving, so they can take proactive measures to retain these customers. This information can also help the company to understand which customer segments are most valuable to their business and why.

**A computer screen shot of text

Description automatically generated**

**A graph with numbers and lines

Description automatically generated**

**It is notable that the possibility of churn occurs more easily with a lower tenure.**

**Question 3:**

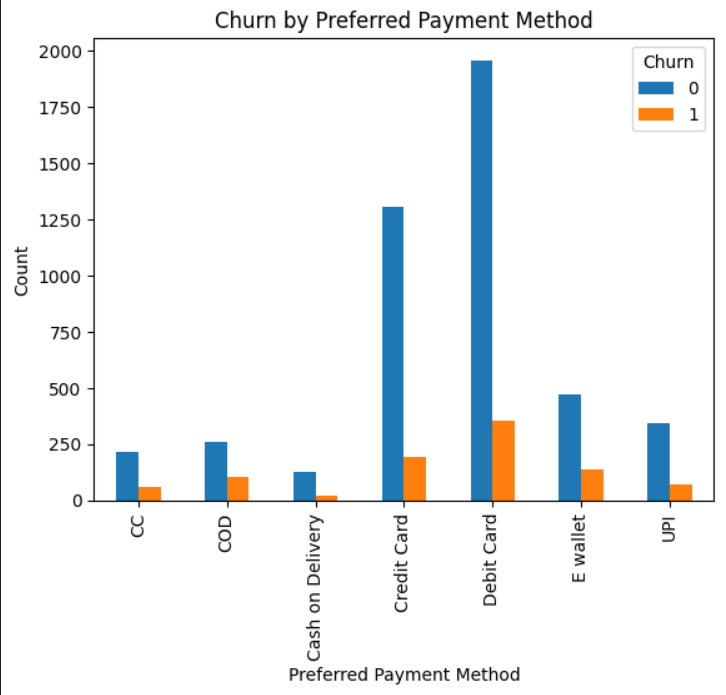
***\*\**\*\*1.3 Is there a significant relationship between a customer's preferred payment method and their likelihood to churn?\*\*\*\***

**Answer 3:**

Understanding the relationship between preferred payment method and churn can help the company identify any barriers to customer retention that may be related to the payment process. This information can be used to make improvements to the payment process and increase customer satisfaction, leading to reduced churn.

**A computer screen with text on it

Description automatically generated**

****

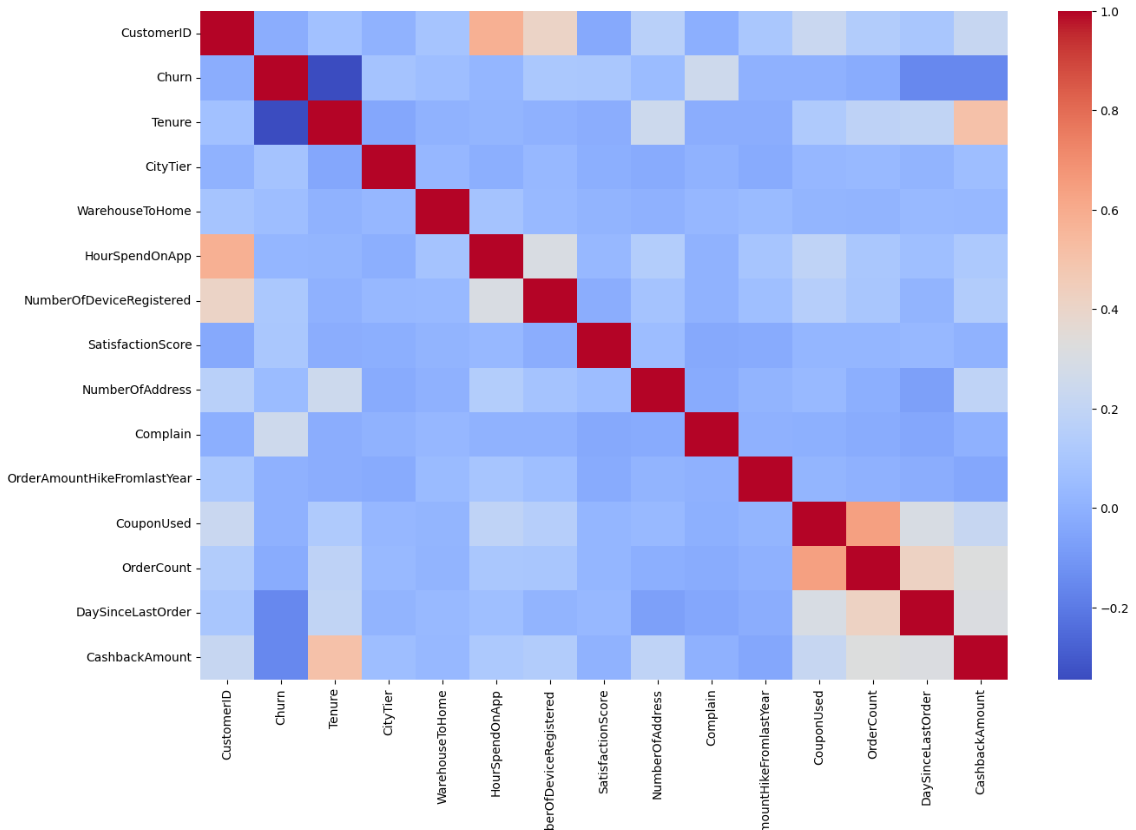
Customers who use credit cards and debit cards tend not to churn

**Feature Engineering:**

**Correlation Matrix: (Used to extract and drop features):**

A screen shot of a computer code

Description automatically generated



**Explanation:**

The correlation matrix shows that coupon used, order count, days since last order, and cashback amount are positively correlated. This means that as the value of one variable increases, the value of the other variables is likely to increase as well. This could indicate that customers who use coupons, place more orders, have made their last order more recently, and receive a higher cashback amount are more likely to be active and engaged with the company.

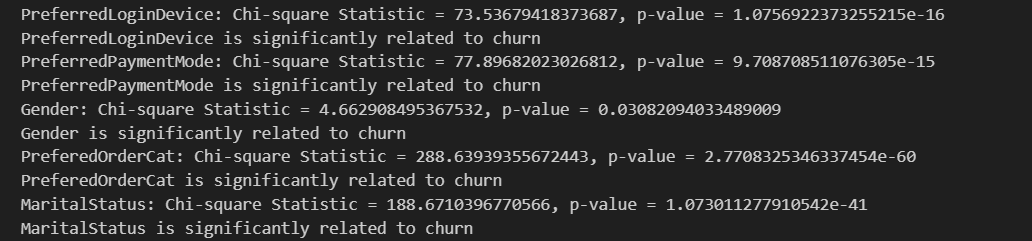
On the other hand, tenure is negatively correlated with churn. This suggests that customers who have been with the company for a longer period of time are less likely to leave compared to those with shorter tenures. This could be due to increased loyalty and satisfaction with the company's products and services over time.

*\*\****\*\*Categorical Features\*\***\*\*

The chi-square test can be used in feature selection to check the dependence between two categorical variables.

A screen shot of a computer program

Description automatically generated



**Explanation:**

The results of the chi-square test show that the PreferredLoginDevice, PreferredPaymentMode, Gender, PreferedOrderCat, and MaritalStatus features are all significantly related to churn. The p-values for each feature are well below the threshold of 0.05, indicating that there is a statistically significant relationship between each of these features and churn.

The PreferredLoginDevice feature has the highest chi-square statistic value, suggesting that the choice of login device is highly related to churn. Similarly, the PreferredPaymentMode feature also has a high chi-square statistic value, indicating that the choice of payment mode is also significantly related to churn.

Gender, PreferedOrderCat, and MaritalStatus also have significant chi-square statistic values and p-values, suggesting that these features also play a role in determining churn. These results highlight the importance of considering the effects of demographic factors, such as gender and marital status, as well as customer preferences, such as login device and payment mode, when analyzing churn.

**Model Selection and Evaluation:**

§ Implement various machine learning algorithms such as Logistic Regression, Random Forest, Naive Bayes or Decision Tree.

A screen shot of a computer program

Description automatically generated

§ Split the data into training and testing sets.

A computer screen with text

Description automatically generated

§ Train each model on the training data and evaluate their performance on the testing set using metrics like accuracy, precision, recall, and F1-score.

A screen shot of a computer program

Description automatically generated

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Metric | Logistic Regression (Train) | Logistic Regression (Test) | Naive Bayes (Train) | Naive Bayes (Test) | Random Forest (Train) | Random Forest (Test) | Decision Tree (Train) | Decision Tree (Test) |
| Accuracy | 0.8988 | 0.8854 | 0.7507 | 0.7469 | 0.9980 | 0.9458 | 1.0000 | 0.9547 |
| Precision | 0.8923 | 0.8759 | 0.8356 | 0.8189 | 0.9980 | 0.9460 | 1.0000 | 0.9544 |
| Recall | 0.8988 | 0.8854 | 0.7507 | 0.7469 | 0.9980 | 0.9458 | 1.0000 | 0.9547 |
| F1 Score | 0.8915 | 0.8745 | 0.7769 | 0.7717 | 0.9980 | 0.9426 | 1.0000 | 0.9546 |

**\*\* Best Model is Decision Tree\*\***

**Advantages and disadvantages of Decision Tree:**

**Advantages:**

Decision Trees are easy to interpret and explain.

They can handle non-linear relationships between variables and can capture complex interactions.

They are computationally fast, making them an efficient algorithm.

**Disadvantages:**

Decision Trees can easily overfit the data and may not generalize well to new data.They are not good for continuous variables, and can be sensitive to small changes in the data.

Based on the accuracy results, the Decision Tree algorithm performed well on the test set, with an accuracy of 0.9547.

**Model Selection and Deployment:**

§ Select the model with the best performance on the testing set.

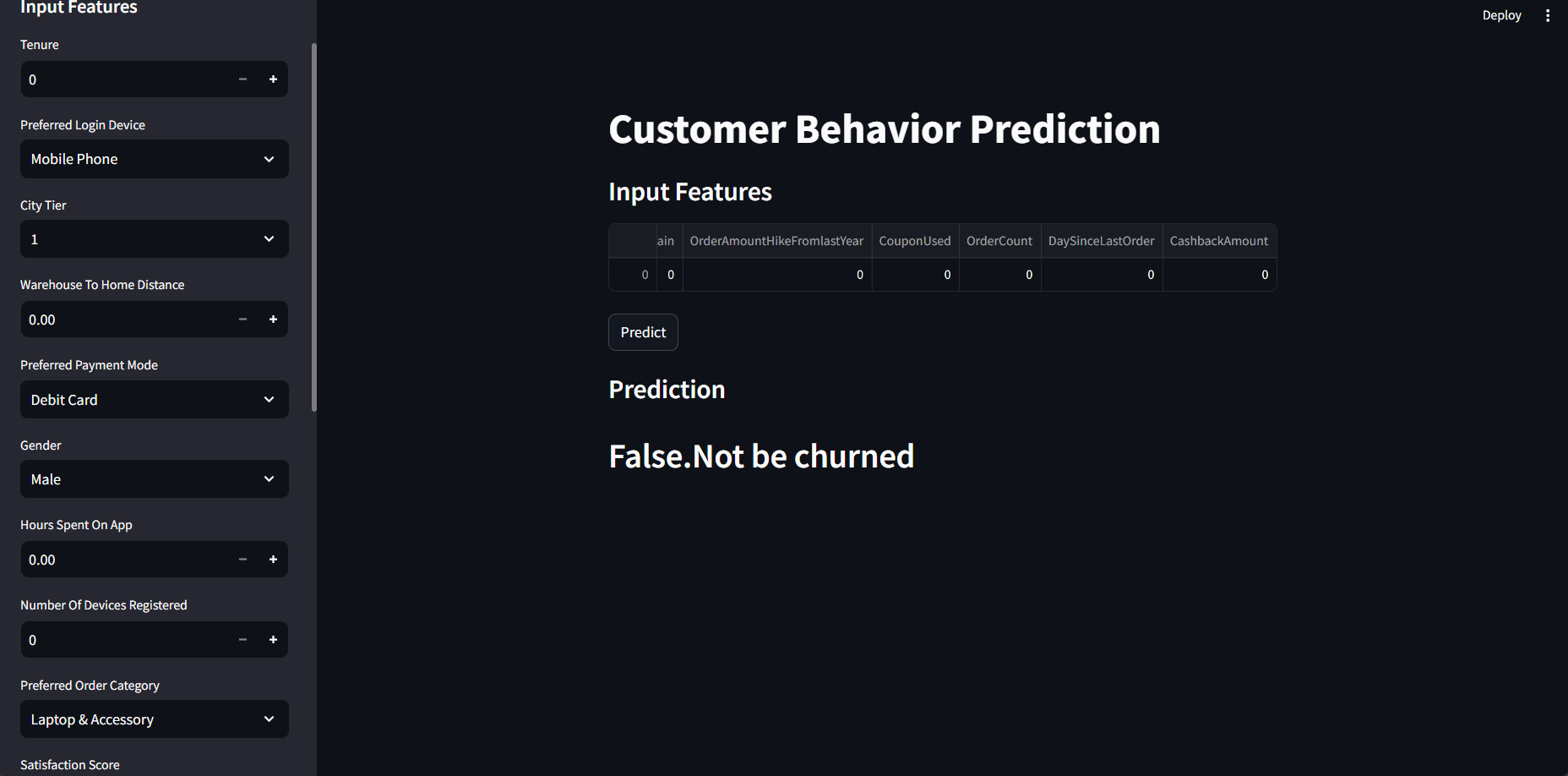
§ Consider deploying the chosen model into a production environment for real-time or batch-based predictions.

**Deployment:**

**A computer screen shot of a program

Description automatically generated**

**A simple front end that takes features as input and give results.**

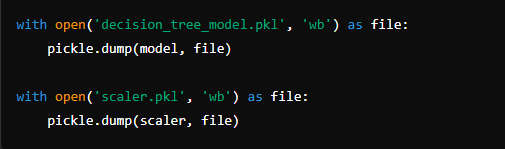
****

**Steps taken to deploy a model:**

### Deployment Process Overview

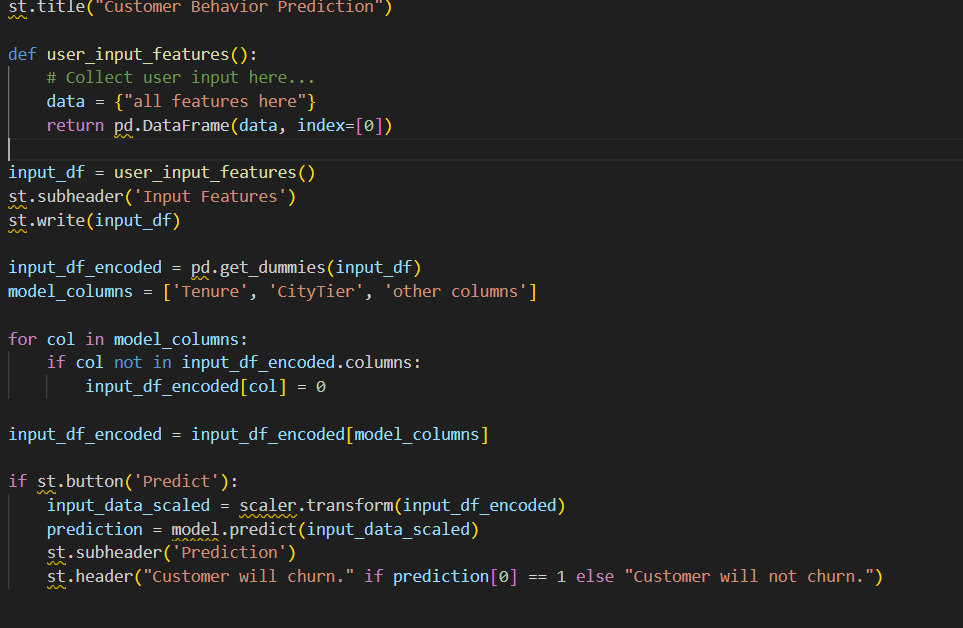
#### 1. Model Training and Serialization

1. **Model Training:**
   * Train the decision tree model with preprocessed data.
   * Use the StandardScaler for feature scaling.
2. **Serialization:**
   * Save the trained model and scaler using pickle.



#### 3. Developing the Streamlit App

1. **User Interface:**
   * Use Streamlit's sidebar for input collection.
2. **Input Handling:**
   * Capture user inputs and create a DataFrame.
   * Encode categorical features.
3. **Prediction Logic:**
   * Load the model and scaler.
   * Preprocess inputs, make predictions, and display results.

****

#### 4. Deployment

**Local Deployment:**

**`streamlit run app/streamlit\_app.py`**

**Deliverables**

**1. Jupyter Notebook or Python Script:**

§ Include the complete code for data preprocessing, feature engineering, model training, and evaluation.

**2. Model Outputs:**

§ Save the final model in a format compatible with deployment, such as a pickle file.

**3. Performance Report:**

§ Prepare a comprehensive report documenting the performance of different models using various evaluation metrics.

**4. User Guide (Optional):**

§ Include a user guide for individuals unfamiliar with machine learning, explaining how to use the developed model.

**main.ipynb :** Main code, EDA, model training .

**App.py: Include model deployment with frontend.**

**Dataset.xlsx: Include data**

**Future Work**

§ Explore more advanced models, such as deep learning architectures, for potentially better prediction accuracy.

§ Integrate the model with existing e-commerce platforms for real-time customer behavior insights.

§ Analyze the impact of predicted customer behavior on marketing campaigns and overall business performance.

**Conclusion**

In conclusion, the results of this project show that various factors, such as coupon usage, order count, days since last order, cashback amount, tenure, login device, payment mode, gender, order category, and marital status, are related to customer churn. The findings of the correlation matrix and chi-square test indicate that these factors can play a significant role in determining churn and that they should be considered when analyzing customer behavior.

The RandomForestClassifier and decision tree models showed good accuracy results, with the decision tree model achieving a higher accuracy score on the test set. These results suggest that using machine learning models can be an effective method for predicting customer churn.

The results of this project have practical implications for the company's financial performance. By better understanding the factors that contribute to customer churn, the company can take proactive measures to reduce churn and retain valuable customers. This can lead to increased customer loyalty, repeat business, and ultimately, improved financial performance.

The findings of this project can be translated into financial results in the following ways:

* **Reducing Churn**: By identifying the factors that contribute to customer churn, the company can take steps to reduce it. This can lead to increased customer retention and repeat business, resulting in increased revenue for the company.
* **Improving Customer Loyalty**: By understanding the factors that influence customer churn, the company can take steps to improve customer loyalty. This can result in higher customer lifetime value, as loyal customers are more likely to make repeat purchases.
* **Increased Efficiency**: By using machine learning models to predict customer churn, the company can target its retention efforts more effectively. This can lead to increased efficiency in retention efforts and a more focused use of resources, resulting in cost savings for the company.

In conclusion, the final draft of Milestone 3 presents a comprehensive solution for predicting e-commerce customer behavior using machine learning algorithms. By leveraging historical customer data and demographics, businesses can gain valuable insights into customer preferences, optimize marketing strategies, and improve customer retention. The methodology outlined in this document covers data acquisition, preprocessing, feature engineering, model selection, evaluation, and deployment.

The deliverables, including the code, model outputs, performance report, and optional user guide, provide a complete package for implementing and utilizing the developed model. Future work suggestions include exploring advanced models, integrating the model with existing e-commerce platforms, and analyzing the impact of predicted customer behavior on marketing campaigns and overall business performance.

With this solution, businesses can enhance their understanding of customers and make data-driven decisions to drive success in the e-commerce industry.

**References**

Ascarza, E ., Netzer, O., & Hardie, B. G. (2018). Some customers would rather leave than use your app: Leveraging customer heterogeneity to target and retain app users. Marketing Science, 37(2), 326-351.

Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. Tourism Management, 59,467-483.

Leeflang, P. S., Verhoef, P. C., Dahlström, P., & Freundt, T. (2014). Challenges and solutions for marketing in a digital era. European management journal, 32(1), 1-12.

Statista. (2023). Global retail e-commerce sales 2014-2023. Retrieved from<https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/>