TITLE : PREDICTIVE ANALYTICS FOR CUSTOMER CHURN IN E-COMMERCE

Bhakti Dwivedi
Usha Mittal Institute of Technology
Branch: Computer Science and
Techonolgy
Roll no. 21

Simran Gupta
Usha Mittal Institute of Technology
Branch: Computer Science and
Techonolgy
Roll no. 26

Tanvee Joshi
Usha Mittal Institute of Technology
Branch: Computer Science and
Techonolgy
Roll no. 31

Archana Kalathiya
Usha Mittal Institute of Technology
Branch: Computer Science and
Techonolgy
Roll no. 33

Abstract **—** This study delves into the realm of predictive analytics to address the challenge of customer churn in the e-commerce industry. The objective of this research is to develop a robust predictive model that can accurately forecast customer churn, enabling e-commerce businesses to proactively implement retention strategies and enhance customer lifetime value. The predictive analytics model developed in this study leverages a combination of supervised and unsupervised learning methods, including logistic regression, random forests, and neural networks, to create a robust framework. The study emphasizes the significance of feature engineering and selection, highlighting the crucial variables contributing to customer churn. Furthermore, temporal patterns and seasonality in e-commerce data are considered through time-series analysis, including Autoregressive Integrated Moving Average (ARIMA) models and Long Short-Term Memory (LSTM) networks.

I. INTRODUCTION (*HEADING 1*)

In the dynamic landscape of e-commerce, where competition is fierce and customer preferences are ever-evolving, businesses face the critical challenge of retaining their customer base. Customer churn, the phenomenon where customers cease their engagement with a brand, has a significant impact on a company's revenue and growth prospects. As such, the ability to predict and mitigate customer churn has become a strategic imperative for e-commerce enterprises. By analysing vast volumes of customer data encompassing behaviours, interactions, purchase histories, and more, predictive models can discern subtle patterns and indicators that signal potential churn. The introduction outlines the pivotal role that customer churn prediction plays in modern e-commerce operations. It highlights the economic implications of churn and underscores the potential of predictive analytics in minimising churn rates. This project encompasses various phases, including data collection, preprocessing, feature engineering, model selection, and evaluation. Each stage contributes to the overarching goal of creating accurate and reliable churn prediction models. Through this endeavour, we seek to demonstrate the potential of data-driven decision-making in mitigating churn rates and fostering customer loyalty.

PROBLEM STATEMENT

In the realm of e-commerce, the critical challenge of retaining customers has become increasingly complex due to shifting consumer preferences, intense competition, and the ease of switching between brands. Customer churn, the phenomenon where customers discontinue their engagement with a brand, poses a significant threat to the growth and profitability of e-commerce enterprises. The cost of acquiring new customers is substantially higher than retaining existing ones, making churn prevention a priority for businesses. The primary objective of this project is to explore the application of predictive analytics for customer churn in the e-commerce domain.

II. Tools and Software (*Heading 2*)

- *Hardware Requirements–*

  Processor: Quad-core or higher
  RAM: 8GB or higher
  Storage: At least 256GB SSD
  Internet Connection: Required for downloading libraries and datasets

- *Software Requirements–*

  Operating System: Windows
  Python Interpreter: Jupyter Notebook
  Programming Language: Python
  Libraries: Pandas, Numpy, Matplotlib, Seaborn,Scikit-learn, XGBoost. These libraries are used for data manipulation, visualization, machine learning, and analysis.

III. PROPOSED SYSTEM (*Heading 3*)

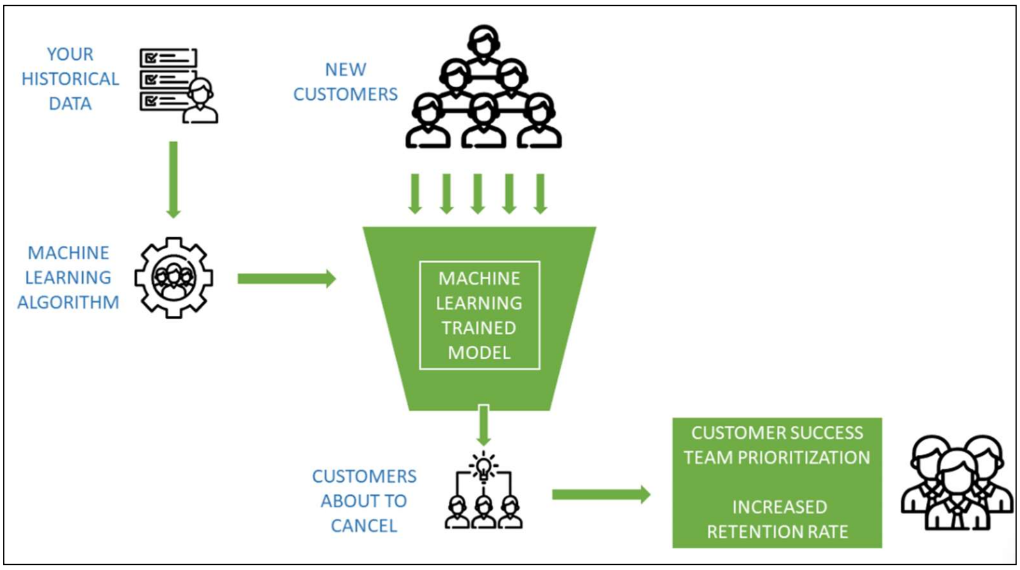

YOUR HISTORICAL DATA
NEW CUSTOMERS
MACHINE LEARNING ALGORITHM
MACHINE LEARNING TRAINED MODEL
CUSTOMERS ABOUT TO CANCEL
CUSTOMER SUCCESS TEAM PRIORITIZATION
INCREASED RETENTION RATE

Figure 1: Proposed System

Through this project we are trying to predict the type of customer that has the potential to churn by identifying features to minimize customer churn rates and get the right business decisions.

Main Objectives of this system include:

1. To create a model to predict users who have the potential to churn and understand what features cause users to churn and features that spur potential to minimize churn rates.

2. Optimizing the company's revenue by grouping users who have the potential to churn so that they can be given different treatments so that the churn rate decreases.

IV. METHODOLOGY (*HEADING 4*)

- **Data Import**:
  - Loaded the dataset using Pandas and perform initial data exploration to understand its structure and features.
  - Loaded the dataset using pd.read_excel().

- **Data Cleaning:**
  - Handled missing values in the dataset using appropriate imputation methods.
  - Addressed outliers by applying quantile-based flooring techniques.
  - Used df.describe() to get basic statistics.To check for missing values with df.info().
  - Dropped irrelevant columns like 'customerID'.
  - To handle outliers we used Boxplot and it displayed this plot. A box plot summarises the distribution of data and highlights potential outliers. Outliers are data points that significantly differ from the majority of the data. They are located outside the "whiskers" of the box plot. The code creates a box plot for each numerical feature in the dataset.

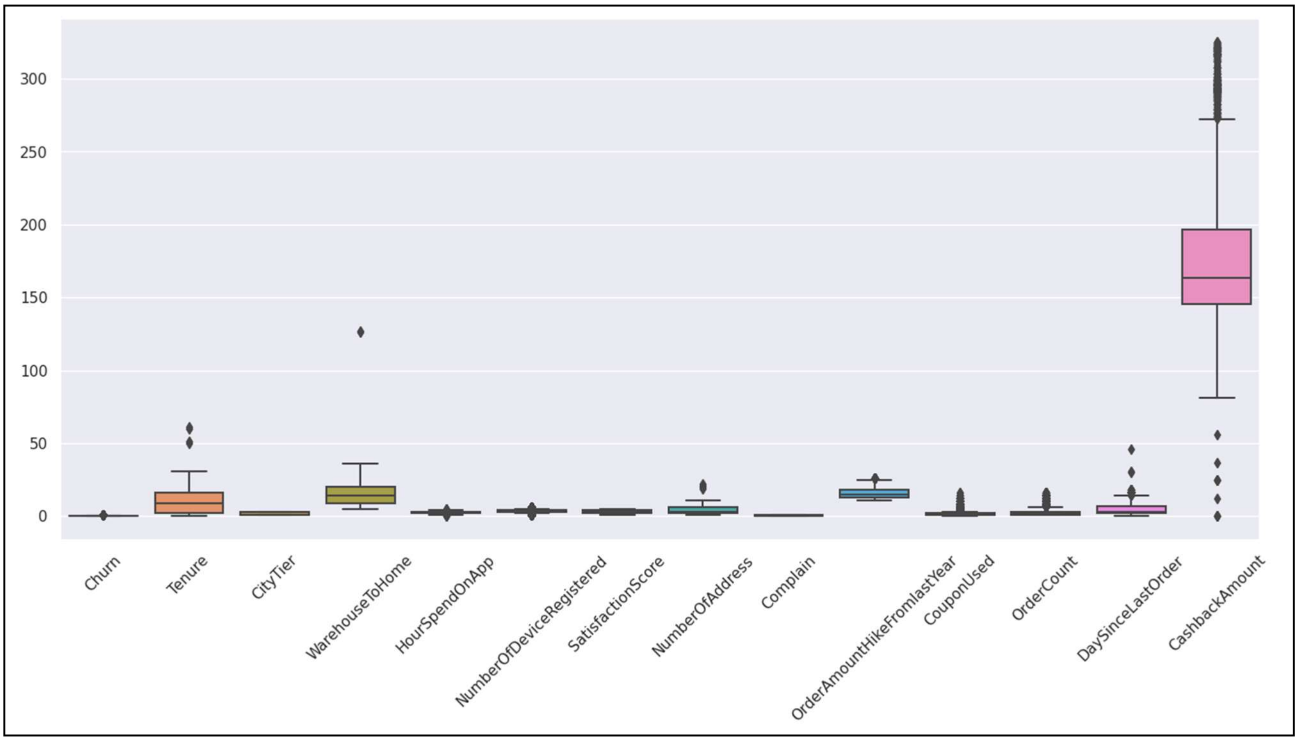


Figure 2: Handling Outliers

- There are quite a lot of features with outliers. We have used a quantile-based flooring method to treat outliers. As per the method, any value beyond 1.5*Q1 and 1.5*Q3 will be regarded as outliers, and all the outlier values will be replaced by Q1-1.5*Q1 and Q3+1.5*Q3. Now, to check if we are free from the outliers we displayed boxplot again. To Handle missing values we did df.isnull().sum() and it had quite a lot of missing values indeed. We imputed these features with means and medians wherever appropriate.

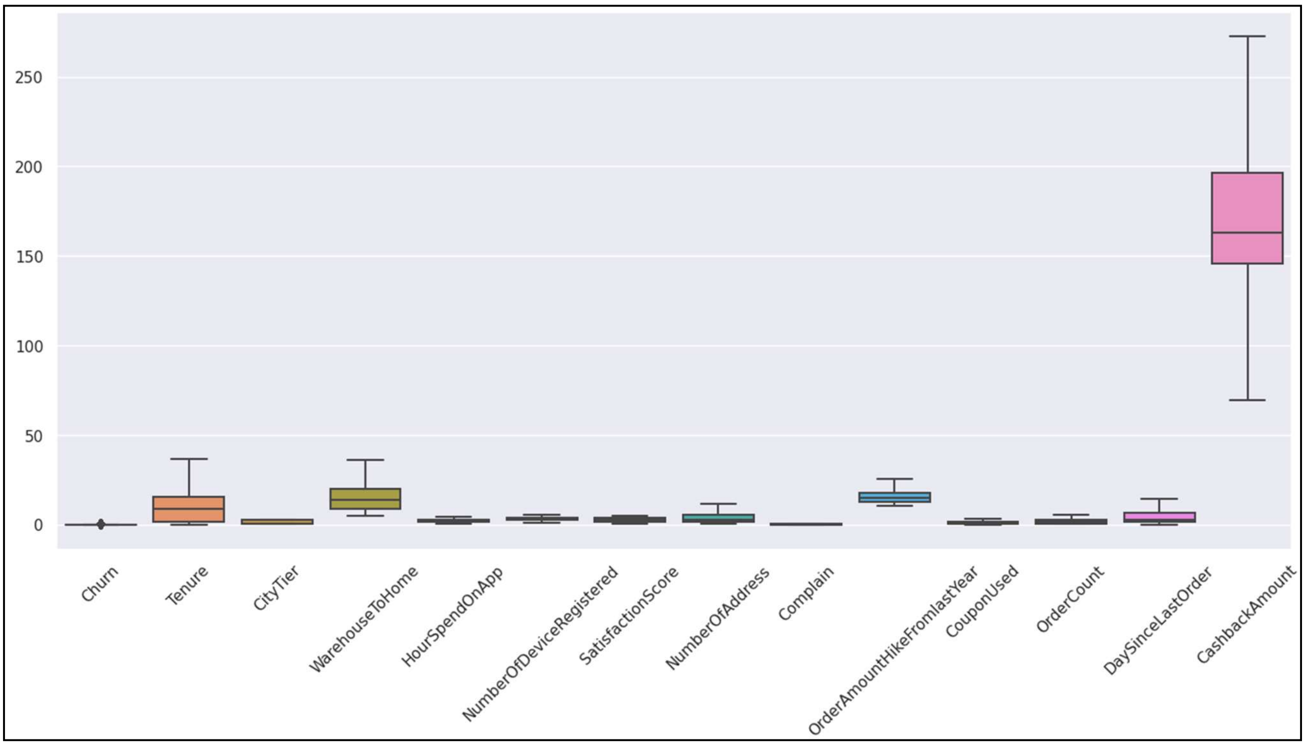
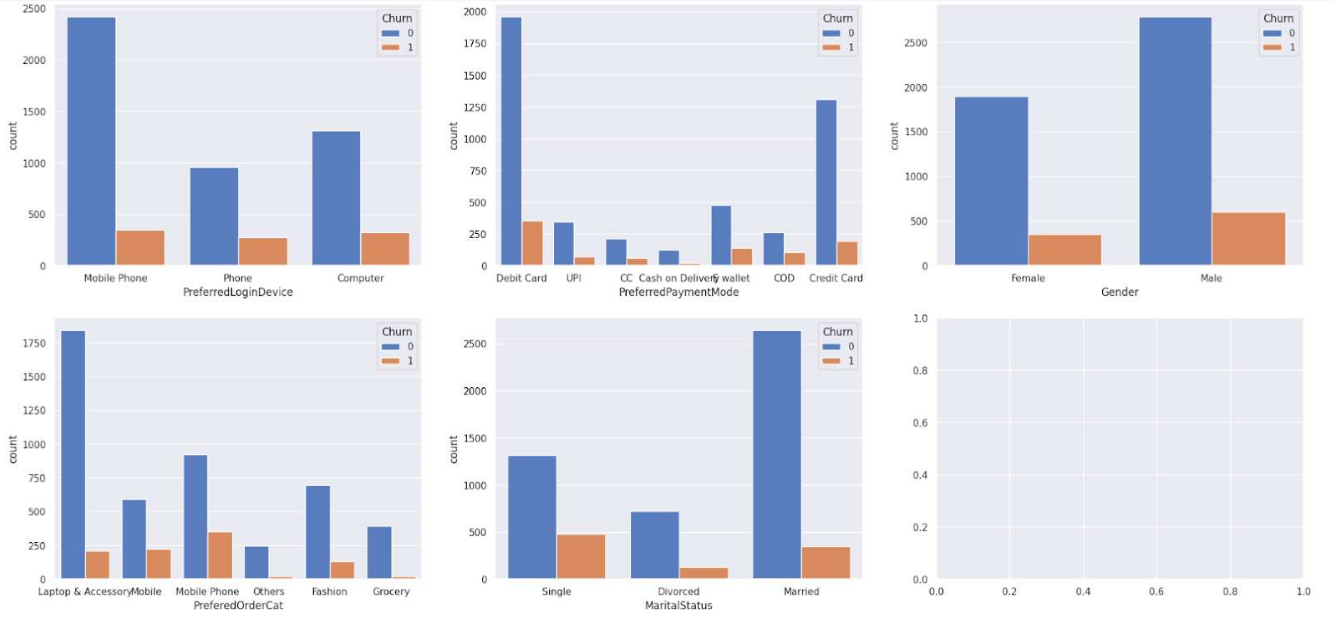


Figure 3: Handling Outliers

- **Data Exploration:**
  - Generated descriptive statistics for the dataset.
  - Visualize data distributions using histograms and box plots.
  - Explored categorical features with count plots and pie charts.

- **Univariate Analysis:**
  - Analyzed the distribution and relationship of each feature with the target variable (churn).
  - Created count plots and pie charts to visualize the distribution of categorical features based on churn.



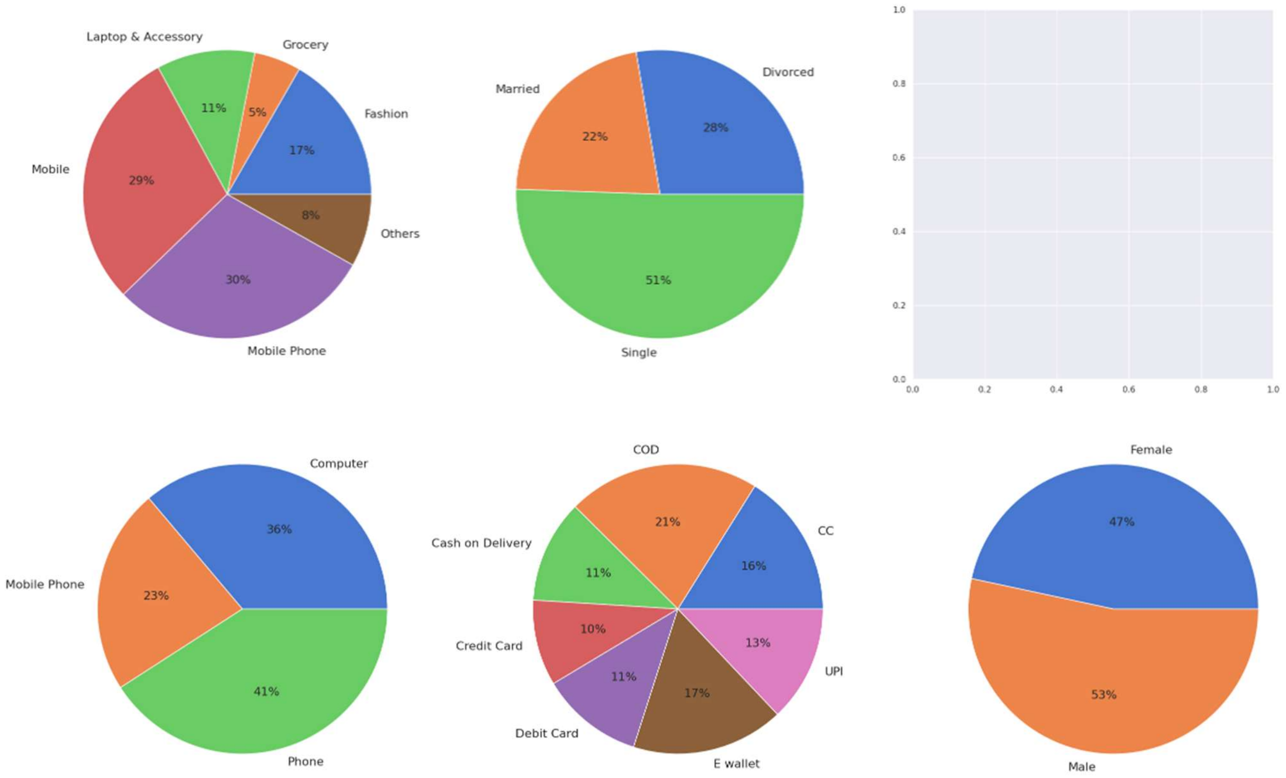


Figure 4: Categorical Analysis

- Plotted line charts to show the distribution of numerical features based on churn.

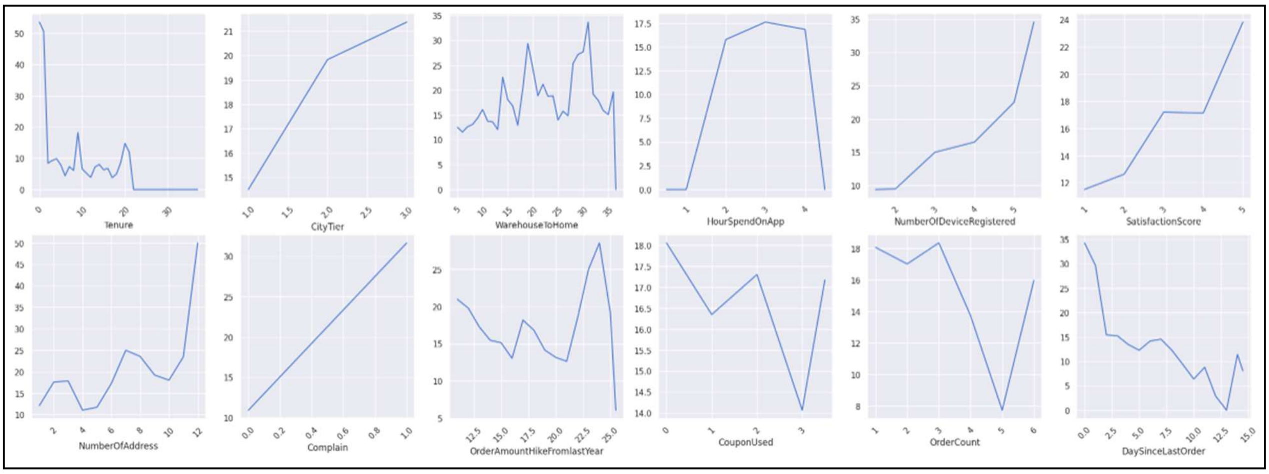


Figure 5: Line Charts

- **Bivariate Analysis:**

Calculate the correlation matrix to find relationships between numerical features. Visualize the correlation matrix using a heatmap.

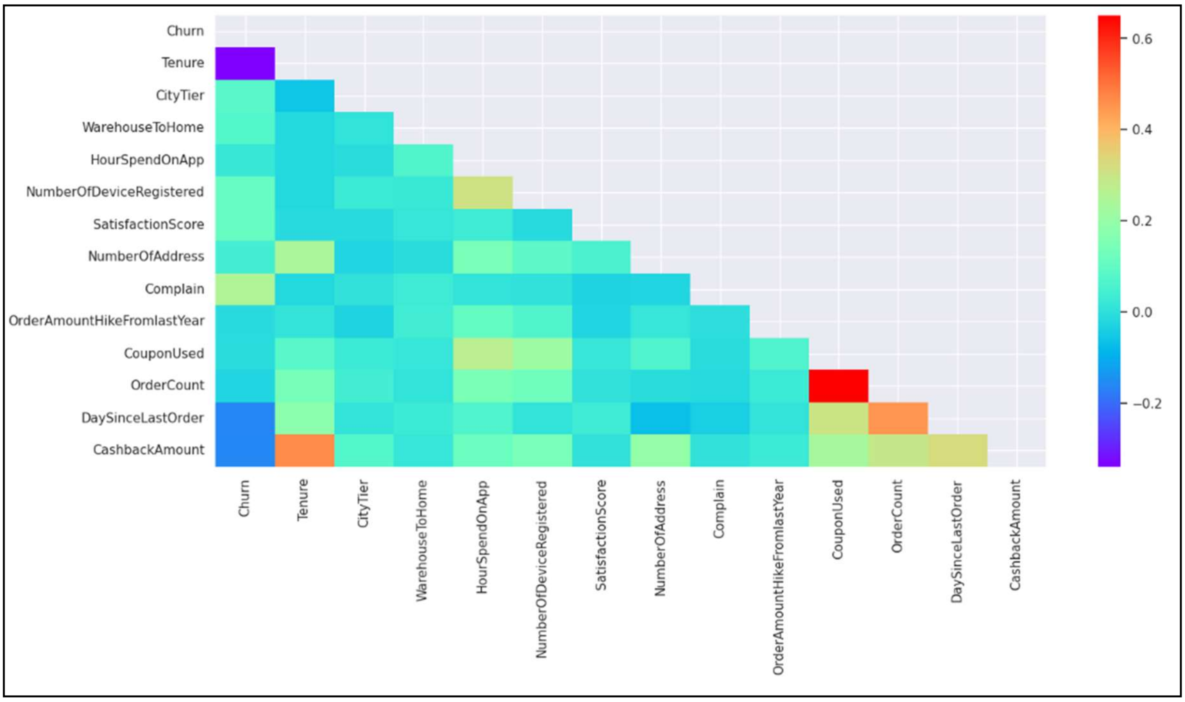
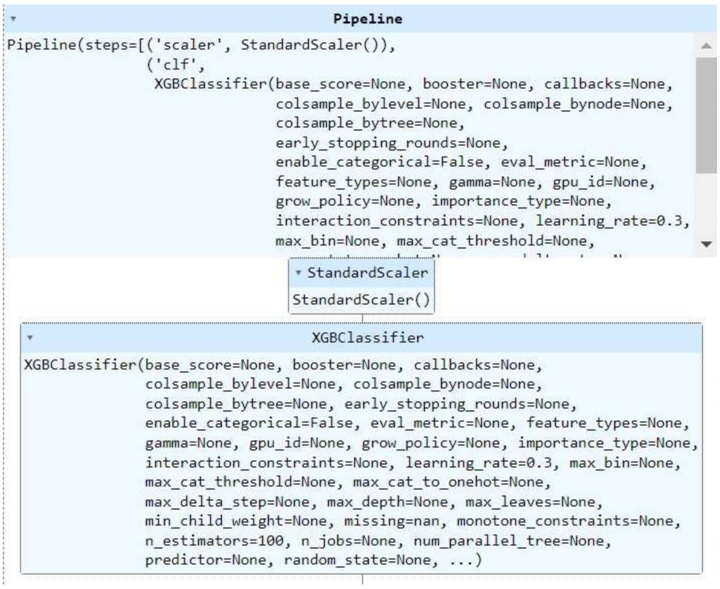


Figure 6: Bivariate Analysis

- **Data Preprocessing:**
  - Encoded categorical variables using LabelEncoder**.**
  - Splitted the data into training and testing sets using train_test_split**.**

- **Customer Classifier and Pipeline:**
  - Created a custom classifier class (inheriting from BaseEstimator and ClassifierMixin).
  - Defined a pipeline with a scaler and the custom classifier.

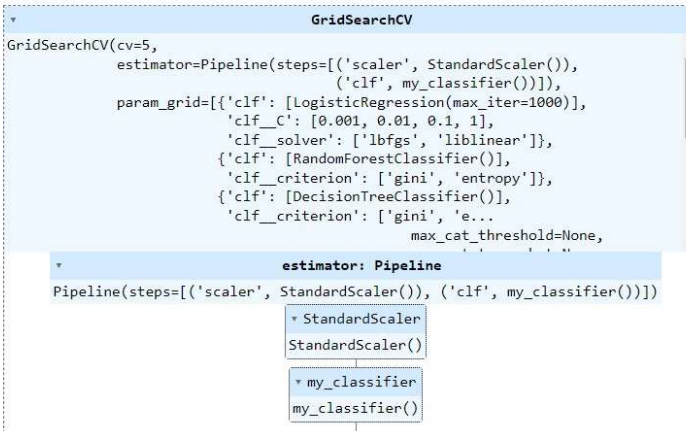
- **GridSearchCV for Model Selection**
  - Define a parameter grid with different classifiers and hyperparameters.
  - Apply GridSearchCV on the pipeline to find the best classifier.



```
Pipeline(steps=[('scaler', StandardScaler()),
                ('clf',
                 XGBClassifier(base_score=None, booster=None, callbacks=None,
                               colsample_bylevel=None, colsample_bynode=None,
                               colsample_bytree=None,
                               early_stopping_rounds=None,
                               enable_categorical=False, eval_metric=None,
                               feature_types=None, gamma=None, gpu_id=None,
                               grow_policy=None, importance_type=None,
                               interaction_constraints=None, learning_rate=0.3,
                               max_bin=None, max_cat_threshold=None,
```

```
StandardScaler()
```

```
XGBClassifier(base_score=None, booster=None, callbacks=None,
              colsample_bylevel=None, colsample_bynode=None,
              colsample_bytree=None, early_stopping_rounds=None,
              enable_categorical=False, eval_metric=None, feature_types=None,
              gamma=None, gpu_id=None, grow_policy=None, importance_type=None,
              interaction_constraints=None, learning_rate=0.3, max_bin=None,
              max_cat_threshold=None, max_cat_to_onehot=None,
              max_delta_step=None, max_depth=None, max_leaves=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              n_estimators=100, n_jobs=None, num_parallel_tree=None,
              predictor=None, random_state=None, ...)
```



```
GridSearchCV(cv=5,
             estimator=Pipeline(steps=[('scaler', StandardScaler()),
                                       ('clf', my_classifier())]),
             param_grid=[{'clf': [LogisticRegression(max_iter=1000)],
                          'clf__C': [0.001, 0.01, 0.1, 1],
                          'clf__solver': ['lbfgs', 'liblinear']},
                         {'clf': [RandomForestClassifier()],
                          'clf__criterion': ['gini', 'entropy']},
                         {'clf': [DecisionTreeClassifier()],
                          'clf__criterion': ['gini', 'e...
                                                 max_cat_threshold=None,
```

```
Pipeline(steps=[('scaler', StandardScaler()), ('clf', my_classifier())])
```

```
StandardScaler()
```

```
my_classifier()
```

- **Model Evaluation**
    - Access the best estimator from GridSearchCV**.**
    - Fit the model on the training data and predict on the test data.
    - Evaluate performance using confusion matrix and F1 score.

- **Feature Importance:**
    - Extracted feature importance scores from the trained XGBoost model and visualized the feature importance using a bar plot.

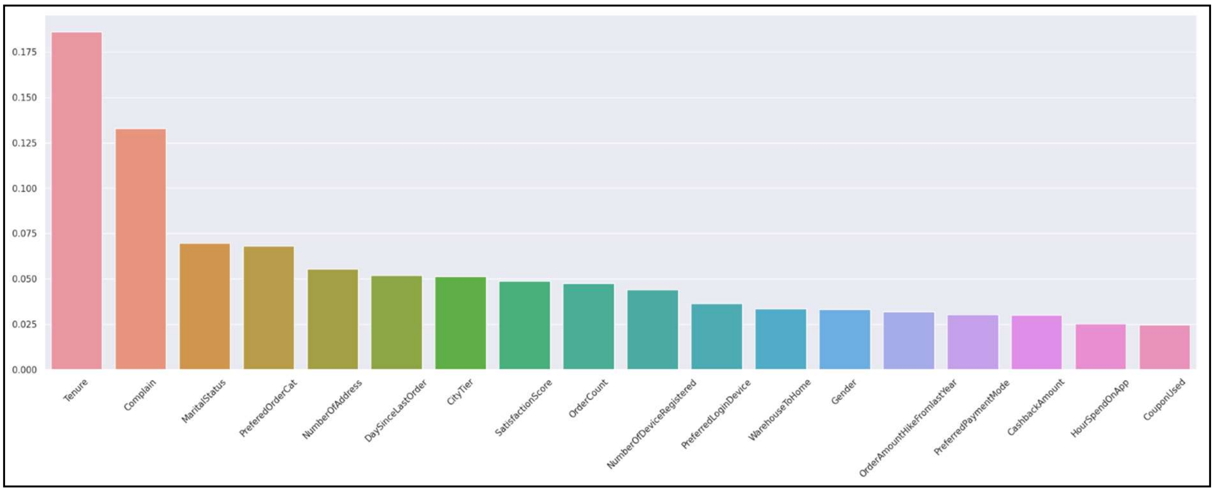


Figure 7: Bar Graph

V. RESULTS (*HEADING 5*)

1. **Confusion Matrix:** A confusion matrix is a useful tool for understanding the performance of a classification model. It provides a breakdown of predicted and actual class labels.

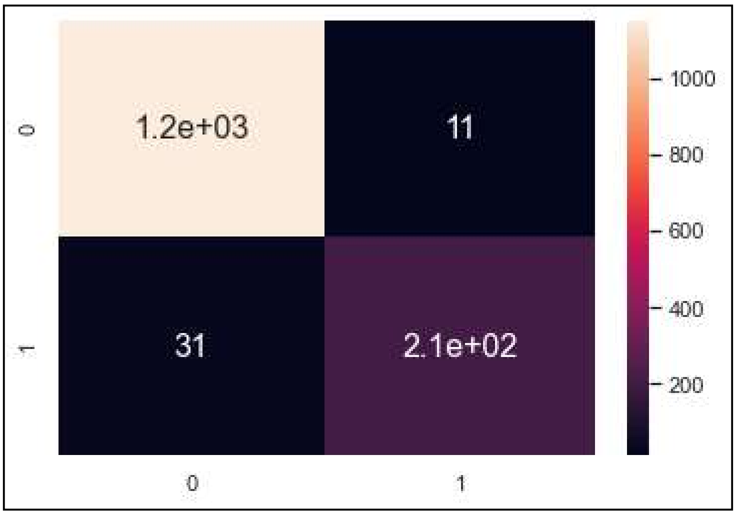


Figure 8: Confusion Matrix

In our analysis:
True Positives (TP): 1.2e+03
False Positives (FP): 11
False Negatives (FN): 31
True Negatives (TN): 2.1e+02

2. F1 Score: The F1 score is a metric that combines precision and recall. It is especially useful when dealing with imbalanced datasets.

F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

In our analysis:

- F1 Score: 0.915

3. Training and Testing Scores: Training and testing scores help assess the model's performance on the data it was trained on and unseen data, respectively.

- Training Score (Accuracy): 0.96

Interpretation:

- The confusion matrix shows that our model correctly predicted 1.2e+03 instances as positive and 2.1e+02 instances as negative, but it made 11 false positive and 31 false negative predictions.
- The F1 score of 0.915 indicates a good balance between precision and recall.
- Our model achieved a high training accuracy of 0.96, indicating it fits the training data well.

VI. SCOPE (*Heading 6*)

- **Personalized Customer Engagement:** As predictive analytics and machine learning models continue to advance; e-commerce businesses can expect to personalize customer engagement to an unprecedented level. Tailored recommendations, pricing, and marketing campaigns will become even more finely tuned, increasing customer satisfaction and loyalty.

- **Real-time Churn Prediction:** The future of customer churn prediction lies in real-time analytics. Businesses will increasingly leverage streaming data and real-time monitoring to identify and address churn risks as they emerge, allowing for immediate intervention and retention efforts.

- **Enhanced Customer Feedback Analysis:** The integration of natural language processing (NLP) will become more sophisticated, enabling businesses to extract deeper insights from customer feedback. Sentiment analysis, emotion detection, and context-aware responses will enable companies to respond more effectively to customer concerns.

- **Cross-Channel Insights:** E-commerce companies will need to integrate data from multiple customer touchpoints, including websites, mobile apps, social media, and customer support. Analyzing cross-channel data will provide a more holistic view of customer behavior, helping businesses understand the customer journey better.

- **Ethical Considerations and Data Privacy:** As ecommerce companies collect and analyze more customer data, ethical considerations and data privacy will be at the forefront. Future advancements should be accompanied by robust data protection measures and transparent policies to maintain customer trust.

- **Augmented Reality (AR) and Virtual Reality (VR):** AR and VR technologies will enable immersive shopping experiences. Customers will be able to virtually try on clothing, visualize products in their homes, or explore virtual showrooms, potentially reducing the likelihood of churn due to product dissatisfaction.

- **Predictive Inventory Management:** E-commerce businesses can use predictive analytics not only for customer-related activities but also for inventory management. Predictive models will optimize stock levels, reducing overstock and stockouts, which can lead to customer dissatisfaction.

CONCLUSION

This comprehensive analysis serves as a guiding light for e-commerce businesses aiming to strengthen customer retention efforts. Leveraging advanced analytics, machine learning, and NLP, companies can not only predict customer churn but also devise personalized strategies to combat it. By proactively addressing churn drivers and enhancing the customer experience, e-commerce businesses can position themselves for sustained growth and success in a highly competitive market. As data continues to play a pivotal role in shaping business strategies, the insights presented in this analysis offer a valuable blueprint for e-commerce practitioners seeking to navigate the complex terrain of customer churn prediction and management. This proactive approach not only preserves revenue but also enhances customer satisfaction, fostering long-term success in the competitive e-commerce landscape.

ACKNOWLEDGEMENT

Projects have always been given to make students learn outside the books. This Project has really been helpful for all of us in understanding about new software and applications that are existing and are in the process of getting launched. We would sincerely like to Thank Dr. Anita Jadhav Mam for giving us this opportunity to understand what Data Analytics is and how we can use it in companies and solve bigger problems. Thank you to all the teammates for their active participation and cooperation.

REFERENCES

[1] Kamil Matuszela ´nski and Katarzyna Kopczewska, ”Customer Churn in Retail E-Commerce Business: Spatial and Machine Learning Approach” 2022 Journal of Theoretical and Applied Electronic Commerce Research.

[2] E-commerce Customer Churn Analysis and Prediction/Medium.

[3] Dash, S.K (2022, June 4) E-Commerce Customer Churn Prediction. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2022/06/e-commerce-customer-churn-prediction/

[4] Gogineni, C. (2017, March 31). How Predictive Analytics helps in reducing churn for e-retailers – https://www.comtecinfo.com/rpa/predictive-analytics-helps-reducing-churn-e-retailers/