E-COMMERCE CHURN ANALYSIS

INTERNSHIP – A PROJECT REPORT

Submitted by

MANIARASAN J

22127028

BACHELOR OF COMPUTER SCIENCE

WITH DATA ANALYTICS

SRI RAMAKRISHNA COLLEGE OF ARTS & SCIENCE

NAVA INDIA, COIMBATORE – 641 006

JUNE – 2024

SRI RAMAKRISHNA COLLEGE OF ARTS & SCIENCE

BONAFIDE CERTIFICATE

Certified that this project report **"E-COMMERCE CHURN ANALYSIS"** is the bonafide work of **"MANIARASAN J (22127028)"** Who carried out the project work under my supervision.

SIGNATURE OF GUIDE

Mr.R.JANARTHANAN MCA., M.Phil., (Ph.D)
ASSISTANT PROFESSOR
BSc. CS WITH CYBER SECURITY
SRI RAMAKRISHNA COLLEGE OF
ARTS & SCIENCE
NAVA INDIA, COIMBATORE-06

SIGNATURE OF HOD

DR.V.VIJAYAKUMAR MCA.,M.Phil.,Ph.D.,
HEAD OF THE DEPARTMENT
BSc. CS WITH DATA ANALYTICS
SRI RAMAKRISHNA COLLEGE OF
ARTS & SCIENCE
NAVA INDIA, COIMBATORE-06

TABLE OF CONTENTS

I.INTRODUCTION TO DATA ANALYTICS

1. Introduction to Data Analysis / Analytics

Data analysis, also known as analytics, is the process of examining and interpreting data to derive meaningful insights, patterns, and trends. It involves applying various techniques and tools to transform raw data into valuable information that can drive informed decision-making.

The goal of data analysis is to uncover valuable insights and make data-driven decisions that can lead to improvements in various areas, such as business operations, marketing strategies, scientific research, and more. By analyzing data, organizations can gain a deeper understanding of their customers, markets, and internal processes, enabling them to identify opportunities, detect problems, and optimize performance.

To conduct data analysis, analysts use a variety of tools and technologies, such as spreadsheet software (e.g., Microsoft Excel), statistical programming languages (e.g., R or Python), data visualization tools.

Churn analysis is the process of identifying customers who are likely to stop using a product or service. It involves analyzing customer behavior and characteristics to predict churn. This type of analysis is critical for businesses because acquiring new customers is often more costly than retaining existing ones. By understanding the factors that contribute to churn, companies can implement targeted strategies to improve customer retention. Common methods include tracking customer activity, engagement levels, and satisfaction scores. Predictive models use historical data to forecast which customers are at risk. These insights help businesses proactively address issues, enhance customer satisfaction, and reduce churn rates.

2. Data Analytics Approaches

Descriptive Analytics: This approach focuses on summarizing and describing historical data to gain insights into past events and trends. Descriptive analytics techniques include basic statistical measures, data visualization, and summary reports. It helps to understand what has happened and provides a foundation for further analysis.

Diagnostic Analytics: Diagnostic analytics aims to identify the reasons behind past events or trends. It involves analyzing data to uncover patterns, correlations, or anomalies that can explain specific outcomes. Diagnostic analytics often utilizes statistical analysis, root cause analysis, and exploratory data analysis techniques.

Predictive Analytics: Predictive analytics involves using historical data to make predictions or forecasts about future events or outcomes. It leverages statistical modeling, machine learning algorithms, and data mining techniques to identify patterns and build predictive models. Predictive analytics helps organizations anticipate future trends, make informed decisions, and take proactive actions.

Prescriptive Analytics: Prescriptive analytics goes beyond prediction and provides recommendations on the best course of action to achieve desired outcomes. It uses optimization algorithms, simulation models, and decision analysis to generate actionable insights. Prescriptive analytics helps organizations optimize resources, streamline processes, and make data-driven decisions.

Exploratory Analytics: Exploratory analytics is an open-ended approach used to discover hidden patterns, relationships, or insights in data. Exploratory analytics is often used in research or when dealing with unstructured or large datasets.

Churn Analytics: Churn analytics refers to the process of measuring customer attrition—the rate at which customers stop using a product, service, or website. By analyzing why customers churn, businesses can take proactive steps to retain them and enhance overall loyalty.

3. Steps of Churn Analytics

Define the Business Problem: Clearly articulate the business problem and objectives. In this case, the goal is to predict customer churn for an e-commerce platform. Understanding the problem helps in selecting the right data and methods. Define what constitutes churn. For example, churn might be defined as a customer not making a purchase for six months.

Collect Relevant Data: Gather data from various sources such as transactional records, user interactions, customer feedback, and demographic information. Ensure the data collected is relevant to the churn problem. This may include purchase history, website visits, customer service interactions, and subscription details.

Data Cleaning and Preprocessing: Clean the data by removing duplicates, handling missing values, and correcting errors. This step ensures data quality and reliability. Transform categorical variables into numerical values using techniques like one-hot encoding or label encoding. Normalize numerical features to ensure they are on a similar scale, which improves model performance.

Exploratory Data Analysis (EDA): Perform exploratory data analysis to understand data distributions, identify patterns, and detect anomalies. Use visualization techniques like histograms, scatter plots, and correlation matrices to gain insights into the data. Identify key features that might influence churn, such as average purchase value, frequency of visits, and customer tenure.

Feature Engineering: Create new features from existing data to enhance model performance. For example, compute the average purchase value per month or the time since the last purchase. Select features that are highly correlated with churn and drop irrelevant or redundant features.

Select and Train Models: Choose appropriate machine learning algorithms based on the problem and data characteristics. Common models for churn prediction include logistic regression, decision trees, random forests, and gradient boosting machines (e.g., XGBoost). Split the data into training and testing sets to evaluate model performance. Typically, 70-80% of the data is used for training, and 20-30% for testing.

Model Evaluation: Evaluate the models using metrics like accuracy, precision, recall, and F1-score. These metrics help assess the model's ability to correctly identify churn and non-churn cases. Use cross-validation techniques to ensure the model is robust and performs well on unseen data.

Interpret Model Results: Interpret the model results to understand the factors influencing churn. Feature importance scores and coefficients provide insights into the most significant predictors.

Deploy the Model: Deploy the predictive model in a production environment to make real-time churn predictions. Integrate the model with the existing system to automate churn prediction and enable proactive retention strategies.

Monitor and Maintain the Model: Continuously monitor model performance to ensure it remains accurate and reliable over time. Update the model with new data to adapt to changing customer behavior and improve predictions.

Implement Retention Strategies: Use the insights gained from the churn analysis to develop targeted retention strategies. These might include personalized offers, loyalty programs, and proactive customer support.

4. Applications of Churn Analytics

Telecommunications: Telecom companies use churn analysis to identify customers likely to switch to competitors. By understanding factors like poor service quality or better offers from competitors, they can develop targeted retention campaigns. Personalized retention offers, such as discounts or service upgrades, can be created for at-risk customers. Additionally, improving network coverage and customer service based on churn insights can reduce dissatisfaction.

Financial Services: Banks and financial institutions predict which customers might close their accounts due to high fees or poor service. Churn analysis helps design loyalty programs that keep customers engaged. It can also identify unusual behavior that might indicate fraud, allowing for early intervention.

Retail: Retailers use churn analysis to segment customers based on their likelihood to churn, enabling more targeted marketing efforts. Personalized promotional strategies, like exclusive discounts, can re-engage at-risk customers. Understanding customer preferences and purchasing patterns helps in better inventory management, reducing overstocking or stockouts.

Subscription-Based Services: Streaming platforms and other subscription services use churn analysis to predict which subscribers might cancel. By understanding viewing habits, they can recommend content to keep subscribers engaged. Churn insights also inform flexible pricing strategies, such as offering discounts for long-term commitments.

Healthcare: Healthcare providers predict which patients might discontinue services due to dissatisfaction or high costs. Proactive care management strategies, like follow-up appointments and personalized care plans, can improve patient retention. Churn analysis also helps optimize resource allocation by understanding patient needs**.**

Insurance: Insurance companies predict which policyholders might cancel or not renew their policies. Personalized communication, offers, and incentives can encourage policy renewal. Churn insights guide the development of new insurance products that better meet customer needs.

Travel and Hospitality: Travel and hospitality companies use churn analysis to design loyalty programs that encourage repeat business. By understanding factors like poor service or high prices, they can improve the overall customer experience. Targeted marketing campaigns address the specific needs of at-risk customers, increasing retention.

Education: Educational institutions predict which students might drop out due to academic difficulties or financial issues. Targeted support services, such as tutoring or financial aid, can help at-risk students stay enrolled. Insights from churn analysis inform curriculum development, ensuring programs meet student needs and interests.

II. OVERVIEW OF THE PROBLEM

1.Problem Study

The problem under study is predicting customer churn in an e-commerce platform. Churn prediction involves identifying customers who are likely to stop using the platform. This is crucial for the business as it directly impacts revenue and growth. The study aims to build a predictive model that can accurately forecast churn. Understanding the factors leading to churn helps in developing targeted retention strategies. The analysis considers various customer behaviors, such as purchase frequency, browsing patterns, and interaction with the platform. By addressing the churn problem, the platform can improve customer satisfaction and retention.

2. Existing & Proposed System

Existing System: Several approaches have been employed to predict customer churn in the past. Customer churn analysis has been previously conducted using a variety of techniques and models. Traditional methods include basic statistical analysis and simple machine learning models, logistic regression, decision trees, random forest and basic clustering techniques.

Demerits of Existing System:

- **Overfitting:** Models often perform well on training data but poorly on new data.
- **Scalability:** Struggles with large datasets.
- **Accuracy:** Basic models may not capture complex patterns.
- **Interpretability:** Complex models can be hard to understand.

Proposed System: The proposed system leverages advanced machine learning algorithms like XGBoost, known for its high performance and ability to handle complex data structures. This approach aims to improve the accuracy and robustness of churn predictions, providing actionable insights for customer retention.

The proposed system aims to overcome the limitations of existing systems by:

- **Feature Selection:** Utilizing advanced feature selection methods like Recursive Feature Elimination (RFE) to identify the most influential factors.
- **Model Optimization:** Implementing and optimizing advanced machine learning models, including logistic regression and XGBoost.
- **Scalability:** Ensuring the system can handle large datasets efficiently.
- **Interpretability and Accuracy:** Balancing model interpretability with high predictive accuracy.

3. Challenges / Need of the study

Churn analysis faces several challenges. Handling imbalanced datasets is a primary issue, as the number of churned customers is often much smaller than retained ones. Feature selection is critical, as irrelevant features can reduce model accuracy. Ensuring data quality and handling missing values is essential. Another challenge is interpreting the model results to derive actionable insights. Despite challenges, the need for churn analysis is significant. High churn rates can lead to substantial revenue loss. By predicting churn, businesses can implement proactive measures to retain customers, enhance loyalty, and optimize marketing.

4. Hardware / System Requirements

System requirements are the configuration that a system must have in order for a hardware or software application to run smoothly and efficiently. The system used in this project is Windows 11. It has 11th Gen Intel(R) Core (TM) i3-1115G4 @ 3.00GHz processor with 8.00GB RAM and 64-bit operating system, x64-based processor system type. GPU is Optional for large-scale data processing.

5. Software, Tools and Libraries Requirements

- **Operating System**: Windows 11
- **Programming Language**: Python 3.11.4
- **Libraries**:
  - pandas: For data manipulation and analysis.
  - numpy: For numerical computations.
  - scikit-learn: For machine learning algorithms and evaluation.
  - XGBoost: For the implementation of the gradient boosting algorithm.
  - Seaborn: For data visualization.
  - Matplotlib: For graphical plotting.
- **Development Environment**: Jupyter Notebook or any Python IDE.
- **Data Source:** Excel, CSV files.

PYTHON:

Python is a high-level programming language that is widely used for various purposes, such as web development, data analysis, machine learning, scientific computing, and more.

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. It is commonly used for data analysis, scientific computing, machine learning, and more.

III. DATA PREPARATION

1.Data Collection Approaches

The dataset used for this project is sourced from Kaggle. It comprises 5630 entries and 20 variables. It provides various features that are significant for predicting customer churn in the E-Commerce sector.

Dataset link:

https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction

Field Description:

• **CustomerID:** Unique identifier for each customer

• **Churn:** Binary variable indicating churn status

• **Tenure:** Duration with the company

• **PreferredLoginDevice:** Device used for login

• **CityTier:** Customer's city tier

• **WarehouseToHome:** Distance from warehouse to home

• **PreferredPaymentMode:** Preferred payment method

• **Gender:** Customer's gender

• **HourSpendOnApp:** Hours spent on the app

• **NumberOfDeviceRegistered:** Number of devices registered

• **PreferedOrderCat:** Preferred order category

• **SatisfactionScore:** Customer satisfaction score

• **MaritalStatus:** Marital status

• **NumberOfAddress:** Number of addresses

• **Complain:** Indicates if the customer has made a complaint

• **OrderAmountHikeFromLastYear:** Increase in order amount from the previous year

• **CouponUsed:** Number of coupons used

• **OrderCount:** Number of orders

• **DaySinceLastOrder:** Days since the last order

• **CashbackAmount:** Cashback amount received

2. Data Method

Exploratory Data Analysis:

Exploratory Data Analysis is a method of evaluating or comprehending data in order to derive insights or key characteristics. EDA can be divided into two categories: graphical analysis and non-graphical analysis. EDA is a critical component of any data science or machine learning process. You must explore the data, understand the relationships between variables, and the underlying structure of the data in order to build a reliable and valuable output based on it.

Prediction Using Machine Learning:

Machine learning (ML) prediction is a powerful technique used to forecast future events or behaviors based on historical data. In the context of churn analysis, ML models are used to predict which customers are at risk of leaving a service or product.

Data Loading and Exploration:

The dataset was loaded, and initial exploration was performed to understand the data structure and distributions.

Loading data

In [ ]:
```
df = pd.read_excel("e_commerce_data.xlsx",
                   sheet_name = 'E Comm')

df.head()
```

Out[33]:

	CustomerID	Churn	Tenure	PreferredLoginDevice	CityTier	WarehouseToHome	PreferredPaymentMode	Gender	HourSpendOnApp	NumberOfDeviceRegistere
0	50001	1	4.0	Mobile Phone	3	6.0	Debit Card	Female	3.0	
1	50002	1	NaN	Phone	1	8.0	UPI	Male	3.0	
2	50003	1	NaN	Phone	1	30.0	Debit Card	Male	2.0	
3	50004	1	0.0	Phone	3	15.0	Debit Card	Male	2.0	
4	50005	1	0.0	Phone	1	12.0	CC	Male	NaN	

Data Cleaning:

Missing or irrelevant information was removed to ensure data quality.

Removing missing values

```
In [ ]: df.dropna(inplace = True, axis = 0)
        df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 3774 entries, 0 to 5629
Data columns (total 20 columns):
 #   Column                       Non-Null Count  Dtype  
---  ------                       --------------  -----  
 0   CustomerID                   3774 non-null   int64  
 1   Churn                        3774 non-null   int64  
 2   Tenure                       3774 non-null   float64
 3   PreferredLoginDevice         3774 non-null   object 
 4   CityTier                     3774 non-null   int64  
 5   WarehouseToHome              3774 non-null   float64
 6   PreferredPaymentMode         3774 non-null   object 
 7   Gender                       3774 non-null   object 
 8   HourSpendOnApp               3774 non-null   float64
 9   NumberOfDeviceRegistered     3774 non-null   int64  
 10  PreferedOrderCat             3774 non-null   object 
 11  SatisfactionScore            3774 non-null   int64  
 12  MaritalStatus                3774 non-null   object 
 13  NumberOfAddress              3774 non-null   int64  
 14  Complain                     3774 non-null   int64  
 15  OrderAmountHikeFromlastYear  3774 non-null   float64
 16  CouponUsed                   3774 non-null   float64
 17  OrderCount                   3774 non-null   float64
 18  DaySinceLastOrder            3774 non-null   float64
 19  CashbackAmount               3774 non-null   float64
dtypes: float64(8), int64(7), object(5)
memory usage: 619.2+ KB
```

Data Preprocessing:

Categorical variables were transformed into dummy variables for modeling purposes.

Transforming categorical variables into dummy variables

```
In [ ]: # Identify categorical columns in the DataFrame
        categorical_columns = df.select_dtypes(include=['object']).columns.tolist()

        # Create dummie variables for categorical columns
        df_dummies = pd.get_dummies(df, columns=categorical_columns, drop_first=True)

        # Show the first rows of the new DataFrame with dummies variables
        df_dummies.head()
```

Out[39]:

	CustomerID	Churn	Tenure	CityTier	WarehouseToHome	HourSpendOnApp	NumberOfDeviceRegistered	SatisfactionScore	NumberOfAddress	Complain	...
0	50001	1	4.0	3	6.0	3.0	3	2	9	1	...
3	50004	1	0.0	3	15.0	2.0	4	5	8	0	...
5	50006	1	0.0	1	22.0	3.0	5	5	2	1	...
11	50012	1	11.0	1	6.0	3.0	4	3	10	1	...
12	50013	1	0.0	1	11.0	2.0	3	3	2	1	...

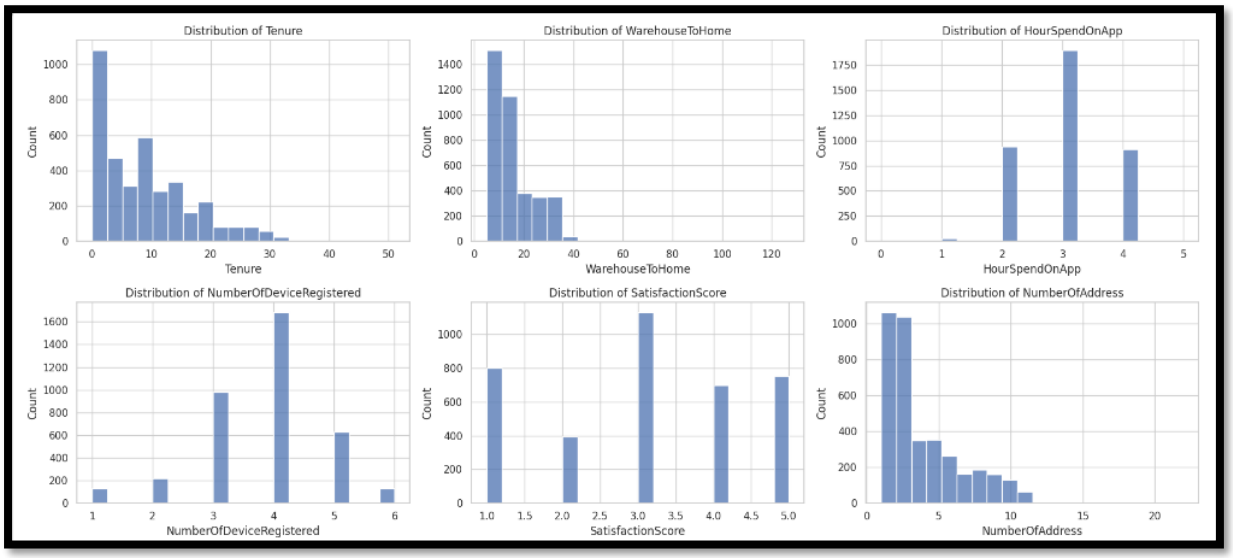
5 rows × 31 columns

Data Visualization:

Data visualization is the graphical representation of information and data. It involves creating visual elements like charts, graphs, and maps to make complex data more accessible, understandable, and usable.

Data visualization ¶

In [ ]:
```
# Configure Seaborn's style
sns.set(style="whitegrid")

# Create a figure with several subplots
fig, axes = plt.subplots(nrows=3, ncols=3, figsize=(18, 12))

# List of numeric columns for visualization
numeric_columns = ['Tenure', 'WarehouseToHome', 'HourSpendOnApp', 'NumberOfDeviceRegistered',
                   'SatisfactionScore', 'NumberOfAddress', 'OrderAmountHikeFromlastYear',
                   'CouponUsed', 'DaySinceLastOrder']

# Loop to create the histograms
for idx, col in enumerate(numeric_columns):
    row, col_idx = divmod(idx, 3)
    sns.histplot(df[col], bins=20, kde=False, ax=axes[row, col_idx])
    axes[row, col_idx].set_title(f'Distribution of {col}')

# Adjust the layout
plt.tight_layout()

# Show graphs
plt.show()
```



Feature Selection:

Recursive Feature Elimination (RFE) was used to identify key factors influencing churn.

Setting up the model

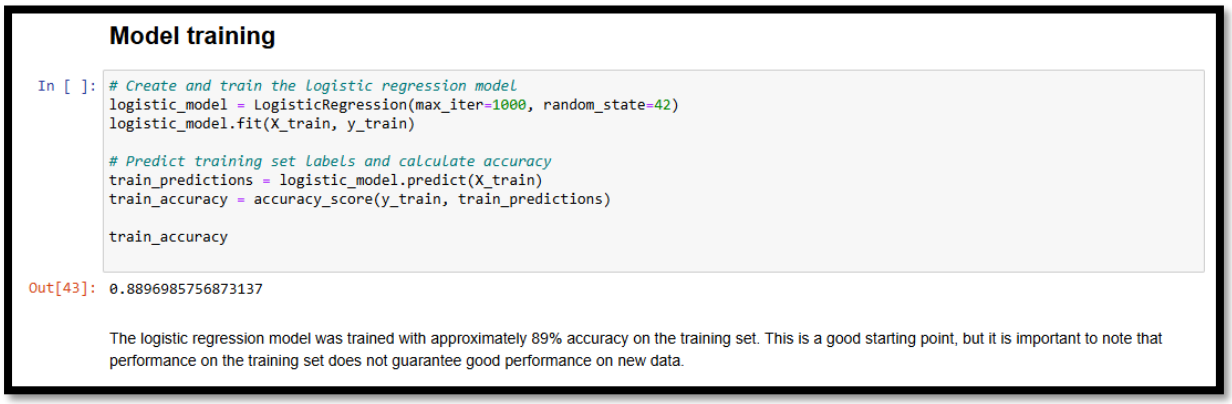
Data separation

In [ ]:
```
# Select only important variables for X
important_features = ranking_scaled[ranking_scaled['Ranking'] == 1]['Feature'].tolist()
X_important = df_dummies[important_features]

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X_important, y, test_size=0.2, random_state=42)

# Show dimensions of training and test sets
X_train.shape, X_test.shape, y_train.shape, y_test.shape
```

Out[42]:
```
((3019, 10), (755, 10), (3019,), (755,))
```

Model Training:

- **Logistic Regression:** Used for initial model training, achieving an accuracy of 90.2% on the test set.
- **Decision Tree:** Trained to capture non-linear relationships but did not perform as well as XGBoost.
- **XGBoost:** Used for ensemble learning, significantly improving model accuracy to 97.6%.



Model training

In [ ]:
```
# Create and train the logistic regression model
logistic_model = LogisticRegression(max_iter=1000, random_state=42)
logistic_model.fit(X_train, y_train)

# Predict training set labels and calculate accuracy
train_predictions = logistic_model.predict(X_train)
train_accuracy = accuracy_score(y_train, train_predictions)

train_accuracy
```

Out[43]:
```
0.8896985756873137
```

The logistic regression model was trained with approximately 89% accuracy on the training set. This is a good starting point, but it is important to note that performance on the training set does not guarantee good performance on new data.

Model Optimization:

GridSearchCV was applied for hyperparameter tuning, optimizing the performance of the XGBoost model.

Model optimization

In [ ]:
```
# Define the hyperparameters to test
param_grid = {'C': [0.001, 0.01, 0.1, 1, 10, 100],
              'penalty': ['l1', 'l2', 'elasticnet'],
              'solver': ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga']}

# Create a StratifiedKFold object for cross-validation
cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)

# Initialize GridSearchCV
grid_search = GridSearchCV(estimator=logistic_model, param_grid=param_grid, cv=cv, scoring='accuracy')

# Run grid search on training set
grid_result = grid_search.fit(X_train, y_train)

# Get the best hyperparameters
best_params = grid_result.best_params_
best_score = grid_result.best_score_
best_params, best_score
```

```
Out[53]: ({'C': 0.1, 'penalty': 'l1', 'solver': 'saga'}, 0.8877055121742282)
```

Evaluation and Interpretation:

The models were evaluated using metrics such as accuracy, precision, recall, and F1-score. Key factors influencing churn included customer complaints, preferred order category, marital status, tenure, and city tier.

Training and evaluation of the new model

In [ ]:
```
# Create and train the logistic regression model with the best hyperparameters
best_logistic_model = LogisticRegression(C=0.1, penalty='l1', solver='saga', max_iter=1000, random_state=42)
best_logistic_model.fit(X_train, y_train)

# Predict test set labels and calculate accuracy
best_test_predictions = best_logistic_model.predict(X_test)
best_test_accuracy = accuracy_score(y_test, best_test_predictions)

# Create the classification report for the optimized model
best_class_report = classification_report(y_test, best_test_predictions)

print(best_test_accuracy)
print(best_class_report)
```

```
0.904635761589404
              precision    recall  f1-score   support

           0       0.92      0.97      0.95       648
           1       0.75      0.49      0.59       107

    accuracy                           0.90       755
   macro avg       0.84      0.73      0.77       755
weighted avg       0.90      0.90      0.90       755
```

Results interpretation

In [ ]:
```
# Get the coefficients of the logistic regression model
coefficients = logistic_model.coef_[0]

# Create a DataFrame to store the variables and their coefficients
coeff_df = pd.DataFrame({'Feature': important_features, 'Coefficient': coefficients})

# Sort the DataFrame by the absolute coefficients
coeff_df['abs_coefficient'] = coeff_df['Coefficient'].abs()
coeff_df = coeff_df.sort_values(by='abs_coefficient', ascending=False)

coeff_df
```

Out[45]:

	Feature	Coefficient	abs_coefficient
7	Complain	1.574266	1.574266
3	PreferedOrderCat_Laptop & Accessory	-1.190099	1.190099
0	MaritalStatus_Single	0.890078	0.890078
2	CityTier	0.435605	0.435605
4	NumberOfDeviceRegistered	0.355862	0.355862
5	SatisfactionScore	0.271667	0.271667
1	Tenure	-0.228813	0.228813
6	NumberOfAddress	0.204998	0.204998
9	OrderCount	0.158776	0.158776
8	DaySinceLastOrder	-0.087837	0.087837

3. Purpose of Data

The purpose of data in churn analysis is to build a predictive model that identifies customers at risk of churning. By analyzing customer behavior and characteristics, the model can forecast future churn. This helps businesses implement targeted retention strategies, such as personalized offers and proactive support. The data also provides insights into the factors driving churn, enabling companies to address underlying issues. Additionally, understanding customer segments with higher churn risk allows for better resource allocation and marketing efforts. Ultimately, the data-driven approach helps improve customer satisfaction, retention rates, and overall business performance.

IV.METHODOLOGY

The methodology for churn analysis involves a series of structured steps:

Data Preprocessing: Cleaning and preparing the data by handling missing values, encoding categorical variables, and normalizing numerical features.

Feature Selection: Identifying important features that influence churn, such as customer demographics, purchase history, and engagement metrics.

Model Selection: Choosing appropriate machine learning algorithms based on the problem and data characteristics. Common models include logistic regression, decision trees, random forests, and XGBoost.

Model Training: Training the selected models using the training dataset. This involves fitting the model to the data and optimizing parameters.

Model Evaluation: Evaluating model performance using metrics like accuracy, precision, recall, and F1-score. Cross-validation techniques ensure robustness.

Hyperparameter Tuning: Optimizing model parameters to improve performance. Techniques like grid search and random search are commonly used.

Model Interpretation: Interpreting the model results to understand the factors influencing churn. Feature importance scores help identify key drivers.

Deployment: Deploying the model in a production environment for real-time churn prediction. This involves integrating the model with the existing system.

Monitoring and Maintenance: Continuously monitoring model performance and updating it with new data to ensure accuracy over time.

------This structured methodology outlines the complete process of developing a robust predictive model for customer churn, from data collection to deployment and result interpretation.

V.RESULTS & FINDING INSIGHTS

The XGBoost model demonstrates robust performance with an accuracy of 97.62% and strong classification metrics across both classes. The use of ensemble learning with XGBoost has proven beneficial, capturing complex patterns in the data and delivering high predictive accuracy. The slightly lower recall for churn customers suggests an area for potential improvement, but overall, the model is well-suited for predicting customer churn in the dataset.

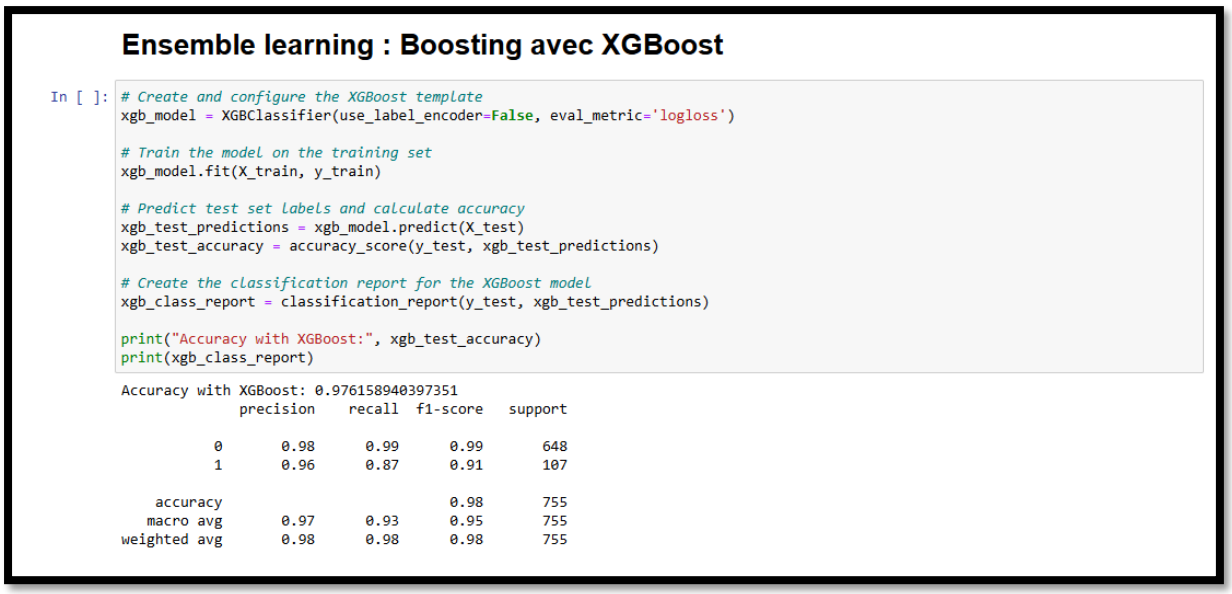
Accuracy with XGBoost: 97.6%

Classification Report:

- **Precision:** High precision for both churn and non-churn classes, indicating that the model makes accurate positive predictions.
    - Class 0: 0.98
    - Class 1: 0.96

- **Recall:** High recall for the non-churn class and reasonable recall for the churn class, indicating the model effectively identifies most non-churners while reasonably identifying churners.
    - Class 0: 0.99
    - Class 1: 0.87
- **F1 Score:** Overall high F1 scores, indicating a balanced performance between precision and recall.
    - Class 0: 0.99
    - Class 1: 0.91
- **Support:** Support refers to the number of actual occurrences of each class in the dataset. It indicates how many instances belong to each class.
    - Class 0: 648 instances
    - Class 1: 107 instances

The high precision, recall, and F1-score, especially for class 0, indicate that the model is highly effective in identifying non-churn customers. The performance for class 1 (churn customers) is also strong, though slightly lower in recall, suggesting that there are some churn customers that the model does not identify.



Ensemble learning : Boosting avec XGBoost

In [ ]:

```
# Create and configure the XGBoost template
xgb_model = XGBClassifier(use_label_encoder=False, eval_metric='logloss')

# Train the model on the training set
xgb_model.fit(X_train, y_train)

# Predict test set labels and calculate accuracy
xgb_test_predictions = xgb_model.predict(X_test)
xgb_test_accuracy = accuracy_score(y_test, xgb_test_predictions)

# Create the classification report for the XGBoost model
xgb_class_report = classification_report(y_test, xgb_test_predictions)

print("Accuracy with XGBoost:", xgb_test_accuracy)
print(xgb_class_report)
```

```
Accuracy with XGBoost: 0.976158940397351
              precision    recall  f1-score   support

           0       0.98      0.99      0.99       648
           1       0.96      0.87      0.91       107

    accuracy                           0.98       755
   macro avg       0.97      0.93      0.95       755
weighted avg       0.98      0.98      0.98       755
```

Key Findings:

Targeted Retention Strategies: Accurate churn predictions can inform targeted retention campaigns, helping reduce churn rates and improve customer loyalty.

Focus on Key Features: Insights into important predictors of churn (e.g., Tenure, SatisfactionScore) can guide business decisions to enhance customer satisfaction and retention.

Implementation of XGBoost: Deploying the XGBoost model in a production environment can provide real-time churn predictions, enabling proactive customer engagement.

Ongoing Model Improvement: Continuous refinement and retraining with new data can maintain and improve predictive accuracy, adapting to changing customer behaviors and market conditions.

By leveraging these findings, the e-commerce company can enhance its customer retention efforts, ultimately leading to increased customer satisfaction and business growth.

VI.DICUSSION & FUTURE WORK

Discussion

Reducing churn should be a strategic focus for businesses because retaining customers is more cost-effective than acquiring new ones. High retention rates lead to increased revenue and higher customer lifetime value (CLV). This analysis shows which customers are at risk of leaving, allowing targeted retention strategies.

The analysis also highlights key predictors of churn, such as frequent complaints or low engagement levels. By addressing these issues, businesses can improve the customer experience and reduce churn rates. Additionally, the insights from the analysis can help businesses allocate resources more efficiently, focusing on segments most at risk.

However, the analysis has limitations. The quality and completeness of data are crucial. Missing or outdated information can impact model accuracy. The models themselves have limitations, such as overfitting to historical data or underfitting if too simplistic. External factors like market trends and economic conditions, which might not be captured in the data, also influence churn. Therefore, models need regular updates to reflect changing customer behaviors.

Future Work

- **Incorporating Additional Features:** Expand the model by integrating diverse data sources such as social media activity, customer reviews, and behavioral data to enhance predictive capabilities.

- **Real-time Prediction:** Develop a real-time prediction system to offer immediate insights, enabling timely intervention and proactive retention strategies.
- **Model Interpretability:** Improve the interpretability of the model to better understand the underlying factors influencing churn predictions, facilitating more transparent and effective retention strategies.
- **A/B Testing:** Implement A/B testing methodologies to validate the efficacy of different retention strategies suggested by the model, ensuring practical applicability and performance improvement.
- **Feature Engineering:** Continuously explore new features and interactions between existing features to refine the model's predictive accuracy and robustness

VII.SUMMARY & CONCLUSION

Summary:

The churn analysis project aims to understand and predict customer churn, providing businesses with actionable insights to enhance customer retention and satisfaction. By leveraging data analytics and machine learning techniques, the project seeks to identify customers at risk of leaving, understand the factors driving churn, and help businesses implement effective retention strategies. The dataset sourced from Kaggle. We began by cleaning the data, removing rows with missing values to ensure completeness. During exploratory data analysis, we used descriptive statistics and visualizations to understand the distribution and relationships of the features. For feature engineering, we employed Recursive Feature Elimination (RFE) to select the most relevant features and standardized numerical features for uniform contribution to model performance.

We developed and compared models, including Logistic Regression and XGBoost, by splitting the data into training and testing sets and evaluating their performance using metrics such as accuracy, precision, recall, and F1-score.

The results revealed that XGBoost outperformed Logistic Regression, achieving higher accuracy and a better precision-recall balance, and highlighted key predictors of churn. These predictions can be used for targeted retention campaigns and improving customer satisfaction by addressing the factors driving churn. Future enhancements of the project may include adding more features, exploring advanced machine learning models, developing real-time prediction systems, and ensuring the model is fair and unbiased.

Conclusion:

The churn analysis project effectively demonstrates the power and potential of data analytics and machine learning in understanding and predicting customer behavior. By identifying customers at risk of churning and uncovering the underlying factors driving their decisions, businesses can implement targeted strategies to enhance customer retention and satisfaction. The churn analysis project on e-commerce underscores the critical role of data analytics and machine learning in modern business strategies. By understanding customer behavior and predicting churn, businesses can implement proactive measures to retain valuable customers, ultimately improving overall performance. The insights gained from this analysis provide a solid foundation for developing effective retention strategies, fostering a customer-centric culture, and achieving long-term business success.

VIII.REFERENCES:

Web links:

1. Understanding churn analysis

https://www.paddle.com/resources/customer-churn-analysis

2. Introduction to data/churn analytics

https://www.simplilearn.com/churn-analysis-article

3. Techniques and importance of churn analysis

hhttps://www.netsuite.com/portal/resource/articles/human-resources/customer-churn-analysis.html

4. Approaches of data analytics

https://www.kdnuggets.com/2023/04/data-analytics-four-approaches-analyzing-data-effectively.html

5. Exploratory data analytics

https://www.analyticsvidhya.com/blog/2022/04/exploratory-data-analysis-eda-in-python/

6. Applications of churn analytics

https://userpilot.com/blog/churn-analytics/

7. Step by step guide to data analytics process

https://careerfoundry.com/en/blog/data-analytics/the-data-analysis

process-step-by-step/

8. Customer Churn analysis in E-Commerce

https://www.returnlogic.com/blog/what-is-customer-churn-in-ecommerce/

9. Prediction using machine learning

https://www.javatpoint.com/machine-learning-prediction

10. Jupyter notebook tutorial

https://www.dataquest.io/blog/jupyter-notebook-tutorial/

Books:

"Hands-On Machine Learning with Scikit-Learn" by Aurélien Géron

"Python for Data Analysis" by Wes McKinney

"Python Data Visualization Cookbook" by Igor Milovanović, Dimitry Foures, and Giuseppe Vettigli