Plagiarism Scan Report

20%
Plagiarism

4%
Exact Match

16%
Partial Match

80%
Unique

Words	1000
Characters	7202
Sentences	59
Paragraphs	104
Read Time	5 minute(s)
Speak Time	7 minute(s)

Content Checked For Plagiarism

1. Exploratory Data Analysis (EDA) and Business Implication

Univariate and Bivariate Analysis:

- Target variable (Churn): Binary variable indicating if a customer has churned (1) or not (0).
- Demographic Factors: Gender, Marital Status, and City Tier distributions were analysed to see if they correlate with churn rates.
- Behavioural Indicators: Features such as Tenure, rev_per_month, coupon_used_for_payment, and CC_Contacted_LY were examined both visually (histograms, boxplots) and statistically (correlation analysis).
- Multi-variable Interactions: Relationships between variables like service scores, payment methods, and login device types were explored to determine their joint effect on churn probability.
- Business Implications: Shorter tenure, lower revenue per month, frequent complaints, and high contact centre interactions appeared to be associated with higher churn rates, suggesting pain points in customer experience.

2. Data Cleaning and Preprocessing

- Handling Missing Values: Several columns like Tenure, cashback, and Day_Since_CC_connect had missing values. These were either imputed using median/mode values or removed based on their distribution and importance.
- Variable Transformation: Categorical variables (e.g., Payment, Gender, Login_device) were encoded using one-hot encoding. Skewed numeric features were normalized or log-transformed if required.
- Feature Engineering: Some derived variables (e.g., complaint frequency per tenure, churn per device type) were created to enhance model interpretability.
- Variable Elimination: Redundant or highly correlated variables were removed to reduce multicollinearity and model overfitting.

3. Model Building and Validation

- Model Selection: Multiple models were evaluated including Logistic Regression, Decision Tree, and Random Forest. Random Forest was chosen due to its superior performance and feature importance clarity.
- Model Performance: Evaluation was done using cross-validation and tested on a hold-out set. Metrics used included accuracy, precision, recall, F1-score, and AUC.
- Validation Strategy: Beyond accuracy, the model was assessed for recall (to capture actual churners) and AUC (to measure separation ability). Feature importance analysis was conducted to interpret model predictions.

Utility of the Research

This research provides both strategic and operational value for organizations seeking to improve customer retention in subscription-based or service-oriented industries. The developed churn prediction model, along with its analytical insights, serves multiple purposes:

1. Proactive Retention Strategy

By identifying customers at risk of churning before they leave, the organization can implement timely interventions such as personalized offers, improved support, or loyalty incentives to retain valuable users.

2. Resource Optimization

The model allows businesses to allocate retention budgets more efficiently. Instead of broadly targeting all users, interventions can be focused on high-risk segments, improving ROI on marketing and support efforts.

3. Service Improvement

Insights into churn drivers—such as poor customer service, complaints, or low engagement—highlight specific operational weaknesses. Addressing these issues can improve overall customer satisfaction and reduce churn organically.

4. Business Decision Support

The findings support data-driven decisions in areas such as pricing, promotion design, feature prioritization, and customer support policies. For example, if cashback reduces churn for new users, onboarding strategies can be adjusted accordingly.

5. Model Deployment in Real-Time Systems

The trained model can be integrated into the organization's customer relationship management (CRM) system, enabling live churn monitoring and automated alerts for customer service or marketing teams.

6. Foundation for Future Research

This work creates a robust framework that can be expanded upon with additional data, such as user activity logs, sentiment analysis, and time-to-event modeling, leading to even more accurate predictions and granular insights.

5. Results and Findings

Data Analysis and interpretation

1. Understanding of Attributes

Information of the data: df.info ()

```
NO  Column Non-Null Count   Dtype
5   AccountID    11260 non-null     int64
1   Churn    11260 non-null     int64
2   Tenure   11158 non-null     object
3   City_Tier      11148 non-null      float64
4   CC_Contacted_LY     11158 non-null      float64
5   Payment     11151 non-null      object
6   Gender  11152 non-null      object
7   Service_Score     11162 non-null      float64
8   Account_user_count       11148 non-null      object
9   account_segment     11163 non-null      object
10  CC_Agent_Score      11144 non-null      float64
11  Marital_Status    11048 non-null     object
12  rev_per_month  11158 non-null      object
13  Complain_ly       10903 non-null     float64
14  rev_growth_yoy 11260 non-null      object
15  coupon_used_for_payment  11260 non-null      object
16  Day_Since_CC_connect 10903 non-null     object
17  cashback     10789 non-null     object
18  Login_device       11039 non-null      object
```

Explanation:

- Displays column names, non-null counts, and data types.
- Helps detect: Columns with missing values
- Incorrect data types (e.g., numeric data stored as objects)
- Several columns like 'Tenure', 'Payment', and 'rev_per_month' are store as object types but might be numerical.
- Missing values exist in most columns (non-null < 11260).</mark>

2. Checking all the missing values

```
missing values = df.isna().sum()
print (missing: =missing_values[missing_values > 0])
Tenure   102
City_Tier       112
CC_Contacted_LY     102
Payment       109
Gender  108
Service_Score     98
Account_user_count       112
account_segment     97
CC_Agent_Score       116
Marital_Status     212
rev_per_month  102
Complain_ly       357
Day_Since_CC_connect 357
cashback     471
Login_device      221
```

Explanation:

- Calculates the number of missing entries per column.
- Filters to show only columns with at least one missing value.
- Most features have missing values, especially cashback and Day_Since_CC_connect.
- Handling these is critical before modeling (via imputation or dropping).

3.Checking stat understanding of data: df.describe()

The describe () function is used to generate descriptive statistics of the numerical columns in a Data Frame (df in this case). It provides a quick summary of the central tendency, dispersion, and shape of the distribution of each numeric feature.

Output Explanation:

For each numeric column in the dataset, it provides:

- count: Number of non-null entries
- mean: Average value
- std: Standard deviation
- min: Minimum value
- 25%: 25th percentile (1st quartile)
- 50%: 50th percentile (median)
- 75%: 75th percentile (3rd quartile)
- max: Maximum value

Columns in the Output:

These are the columns for which the statistics are displayed:

- AccountID: Unique identifier for accounts (no real analytical value in modeling)
- Churn: Binary column indicating whether a customer churned (0 or 1)
- City_Tier: Tier level of the customer’s city (1, 2, or 3)
- CC_Contacted_LY: Number of times customer contacted customer care last year
- Service_Score: Rating of the service quality (likely a score from 0 to 5)
- CC_Agent_Score: Rating of the customer care agent (score from 1 to 5)
- Complain_ly: Binary variable indicating whether the customer complained last year

Matched Source

Similarity 2%

Title:Multi-Identity Recognition of Darknet Vendors Based on Metric ...

Feb 17, 2024 � The evaluation metrics used included accuracy, precision, recall, F1-score, and AUC. It can be seen that the AUC values of the ten models�...

https://www.mdpi.com/2076-3417/14/4/1619

Similarity 2%

Title:How to Interpret a Machine Learning Model with Feature Importance

Sep 28, 2023 � Feature importance helps you identify which features most impact your model's predictions and which ones to focus on. For instance, if analysis�...

https://www.linkedin.com/advice/0/how-can-you-interpret-machine-learning-model-using

Similarity 2%

Title:The Importance of Customer Retention in Business Growth - Thinkific

Feb 26, 2025 � Improved revenue predictability allows businesses to create more accurate budgets and allocate resources effectively. When income streams�...

https://www.thinkific.com/blog/importance-of-customer-retention

Similarity 2%

Title:Customer Churn: 5 Types & Strategies to Boost Retention - DevRev

Feb 18, 2025 � Addressing these root causes proactively can enhance customer satisfaction, reduce churn, and increase overall retention rates. Predicting�...

https://devrev.ai/blog/customer-churn

Similarity 2%

Title:Predictive Analytics in Business | Process & Practical Applications

Nov 4, 2022 � 5. Model Deployment in Real-Time Systems. After model validation, the predictive model is implemented into actual business systems. This�...

https://serigor.com/process-of-predictive-analytics-and-its-use-in-the-business-sector

Similarity 2%

Title:Solved needing help : can someone help me with the codes i - Chegg

Dec 10, 2023 � Write a script to find the percentage of the missing values for each column in your dataset; Impute any numeric feature that has missing values�...

https://www.chegg.com/homework-help/questions-and-answers/needing-help-someone-help-codes-would-run-order-happen-dataset-excel-bring--read-dataset-p-q126650156

Similarity 2%

Title:Unveiling Customer Churn with Churn Guard - LinkedIn

Apr 28, 2024 � CHURN: Target variable indicating if the customer has churned (1) or not (0). Hypothesis: Null Hypothesis (H0):. There is no significant�...

https://www.linkedin.com/pulse/unveiling-customer-churn-guard-efosa-dave-omosigho-oiqzf

Similarity 2%

Title:How to filter columns containing missing values - Stack Overflow

Jul 23, 2022 � To select all columns with at least one missing value, use: df[df.columns[df.isna().any()]] Alternatively, you could use .sum() and then choose some threshold.

https://stackoverflow.com/questions/73094053/how-to-filter-columns-containing-missing-values