**CUSTOMER CHURN PREDICTION**

*A project report submitted to ICT Academy of Kerala*

*in partial fulfillment of the requirements*

*for the certification of*

**CERTIFIED SPECIALIST**

**IN**

**DATA SCIENCE & ANALYTICS**

submitted by

**Radhika K J**

**Reshma Raghavan M V**

**Sreehari Kunnath**

**Tony Alex**



**ICT ACADEMY OF KERALA**

**THIRUVANANTHAPURAM, KERALA, INDIA**

**July 2023**

**List of Figures**

**1.Missing Matrix 12**

**2.Customer Segmentation by DaysSinceLastOrder 13**

**3.At Risk Customers Gender 13**

**4.At Risk Customers City Tier 14**

**5.Customers at risk Preferred Order Category 15**

**6.Customers at risk Marital Status 15**

**7.Complaints by customers at risk 16**

**8.Log in devices of customers at risk 16**

**9.Customer Tenure 17**

**10.Customer Satisfaction 17**

**11.Hours Spend on App 18**

**12.Order Count 18**

**13.WareHouse To Home 18**

**14.The Confusion Matrix 25**

**15.Hosting 28**

**List of Abbreviations**

* **ML: Machine Learning**
* **AI: Artificial Intelligence**
* **SVM: Support Vector Machine**
* **KNN: k-Nearest Neighbors**
* **RFC: Random Forest Classifier**
* **PCA: Principal Component Analysis**

**CV: Cross-Validation**

* **AUC: Area Under the Curve**
* **ROC: Receiver Operating Characteristic**
* **TP: True Positive**
* **FP: False Positive**
* **TN: True Negative**
* **FN: False Negative**
* **SMOTE: Synthetic Minority Oversampling Technique**

**Table of Contents**

**Abstract 1**

**Problem Definition 6**

**Introduction 7**

**Literature Survey 9**

**Exploratory Data Analysis 11**

**Data Preprocessing 19**

**Model Selection 21**

**Result 25**

**Hosting 27**

**Conclusion 29**

**Abstract**

The rise of online shopping and the global impact of the COVID-19 pandemic have transformed the retail landscape, leading to intensified competition among e-commerce companies. To gain a competitive edge, it has become crucial for businesses to focus on customer retention through competitive pricing and personalized services. Customer churn, representing the percentage of customers who cease using a company's products or services, is a critical metric to monitor and address. In this project, we aim to predict customer churn by analyzing various factors, including tenure, login device, warehouse-to-home distance, and more. Furthermore, we will examine the individual impact of these factors on churn. By understanding the drivers of customer churn, businesses can develop effective strategies to enhance customer loyalty and drive sustainable growth in the dynamic e-commerce landscape.

**1. Problem Definition**

**1.1 Overview**

In the competitive landscape of E-commerce, customer retention plays a crucial role in reducing acquisition costs and fostering customer loyalty. It is well-established that retaining existing customers is more cost-effective than attracting new ones (Saghir et al., 2019). To address this challenge, accurately predicting customer churn in advance can empower businesses to implement personalized retention strategies and build strong customer relationships.

This project focuses on E-commerce customer churn prediction. SMOTE will be used to address the data imbalance. Several different machine learning algorithms will be utilized for predictive modeling. Additionally, the project aims to identify key predictors' importance to assist decision-makers in making informed choices for the organization.

**1.2 Problem Statement**

Customer retention poses a significant challenge for e-commerce organizations, as retaining existing customers is more difficult than attracting new ones. Existing customers provide higher value, making churn prediction crucial for e-commerce businesses. This project aims to develop a predictive model for the e-commerce sector that correlates key attributes to customer churn. By identifying factors leading to churn, the model will help businesses proactively address customer attrition and improve customer retention strategies.

**2. Introduction**

In the rapidly evolving world of E-commerce, businesses face a critical challenge in retaining customers and fostering long-term loyalty. As the digital landscape expands and the impact of the novel coronavirus continues to shape consumer behavior, the shift towards online shopping has intensified the competition among E-commerce companies. The race to stay ahead and secure customer loyalty demands innovative strategies that go beyond competitive pricing and include personalized services tailored to individual needs.

Customer Churn, a pivotal metric in the E-commerce realm, quantifies the percentage of customers who discontinue using a company's services or products. Understanding and predicting churn is of paramount importance for businesses seeking to mitigate revenue losses and enhance customer retention. By identifying customers who are likely to churn in advance, E-commerce companies can proactively address their needs, offer targeted incentives, and foster enduring brand loyalty.

The aim of this project is to develop an advanced prediction model for E-commerce customer churn. We will leverage various customer attributes, such as tenure, login device, warehouse-to-home distance, satisfaction score, and more, to discern patterns that indicate potential churn behavior. This predictive model will not only enable companies to anticipate and address churn more effectively but also empower them to allocate resources wisely, reducing acquisition costs and bolstering customer retention efforts.

To achieve this, the research will focus on customer segmentation to gain deeper insights into distinct customer groups and their likelihood of churn. Additionally, we will employ an improved SMOTE (Synthetic Minority Over-sampling Technique) to balance the data, optimizing the predictive accuracy of our model. By leveraging the power of three different machine learning algorithms, we will create a comprehensive churn prediction framework that enables E-commerce businesses to make data-driven decisions for retention and customer engagement strategies.

The key objective of this project is to identify essential predictors that significantly influence churn behavior. These insights will guide decision-makers in making well-informed choices to foster customer loyalty, tailor personalized experiences, and ultimately strengthen the competitive position of E-commerce companies in the dynamic market landscape.

In summary, this research endeavor aims to contribute to the domain of E-commerce by providing a robust predictive model for customer churn. By combining sophisticated data analysis techniques and machine learning algorithms, we aspire to equip businesses with the tools and insights necessary to anticipate and address churn, foster customer loyalty, and thrive in the dynamic world of E-commerce.

**3. Literature Survey**

**Introduction:**

Customer churn, a phenomenon where customers discontinue using a company's products or services in favor of competitors, poses a significant challenge for businesses, particularly in the e-commerce sector. As e-commerce customer churn occurs in non-contractual relationships, detecting potential churn in advance is crucial to retain high-value customers and study the purchasing behavior of non-churn customers. Accurate prediction of customer churn can aid in implementing targeted retention strategies and reducing customer acquisition costs, leading to improved customer relationship management and enhanced profitability.

**Literature Review:**

**1. Customer Churn Prediction Models in E-commerce:**

Researchers have extensively explored various prediction models to identify customer churn in the e-commerce domain. Machine learning algorithms, such as logistic regression, decision trees, random forests, support vector machines, and ensemble methods like AdaBoost, have been utilized to build accurate churn prediction models. These models leverage historical transaction and customer data to distinguish between churners and non-churners.

**2. Handling Imbalanced Data:**

E-commerce customer churn datasets often suffer from class imbalance, where churners are significantly outnumbered by non-churners. To address this issue, researchers have employed techniques like SMOTE (Synthetic Minority Over-sampling Technique) to oversample minority churn instances, balancing the dataset and improving model performance.

**3. Customer Segmentation for Targeted Marketing:**

Customer segmentation based on transaction behavior and characteristics has been a key approach to identify valuable customers and allocate marketing resources effectively. The Pareto principle, indicating that a substantial portion of profits is derived from a small segment of customers, emphasizes the importance of customer segmentation for targeted marketing efforts.

**4. Leveraging Ensemble Methods:**

Ensemble methods, such as Random Forest and AdaBoost, have demonstrated enhanced performance in predicting customer churn. The integration of multiple classifiers allows for better generalization and improved accuracy in churn identification.

**5. Utilizing Purchasing Behavior for Customer Value Assessment:**

Purchasing behavior, such as the amount of purchase and transaction frequency, has been widely adopted as a measure of customer value in e-commerce. The study of customer value helps businesses identify valuable customers and tailor retention strategies accordingly.

**6. The Role of SMOTE in Imbalance Handling:**

SMOTE, as a synthetic data generation technique, has proven effective in reducing class imbalance and improving churn prediction accuracy. By generating synthetic minority samples, SMOTE enhances the performance of machine learning algorithms on imbalanced datasets.

**Conclusion:**

The literature on prediction of customer churn in e-commerce highlights the significance of machine learning algorithms, data balancing techniques, customer segmentation, and customer value assessment for accurate churn prediction. Leveraging ensemble methods and handling class imbalance with SMOTE have demonstrated promising results. Further research can explore the integration of text analytics, NLP techniques, and deep learning approaches to enrich churn prediction models and enhance customer retention strategies in the ever-evolving e-commerce landscape.

**4.Exploratory Data Analytics**

Exploratory Data Analysis (EDA) plays a crucial role in understanding and gaining insights from the dataset related to customer churn in the e-commerce sector. EDA is a preliminary and essential step in the data analysis process, where we investigate the dataset, identify patterns, trends, and potential relationships among variables. The main objective of EDA is to prepare the data for further analysis and model building, as well as to extract meaningful information to make data-driven decisions.

In the context of customer churn prediction, EDA enables us to explore various features that might impact customer behavior. By visualizing and summarizing the data, we can identify key attributes and patterns related to churn. EDA also helps in identifying potential data quality issues, such as missing values or outliers, which may need to be addressed before building the predictive model.

In this documentation, we will leverage EDA techniques to gain insights into the dataset and answer critical questions such as:

1. What is the distribution of customer churn across different categories?

2. Are there any patterns or trends in customer churn based on demographic or behavioral attributes?

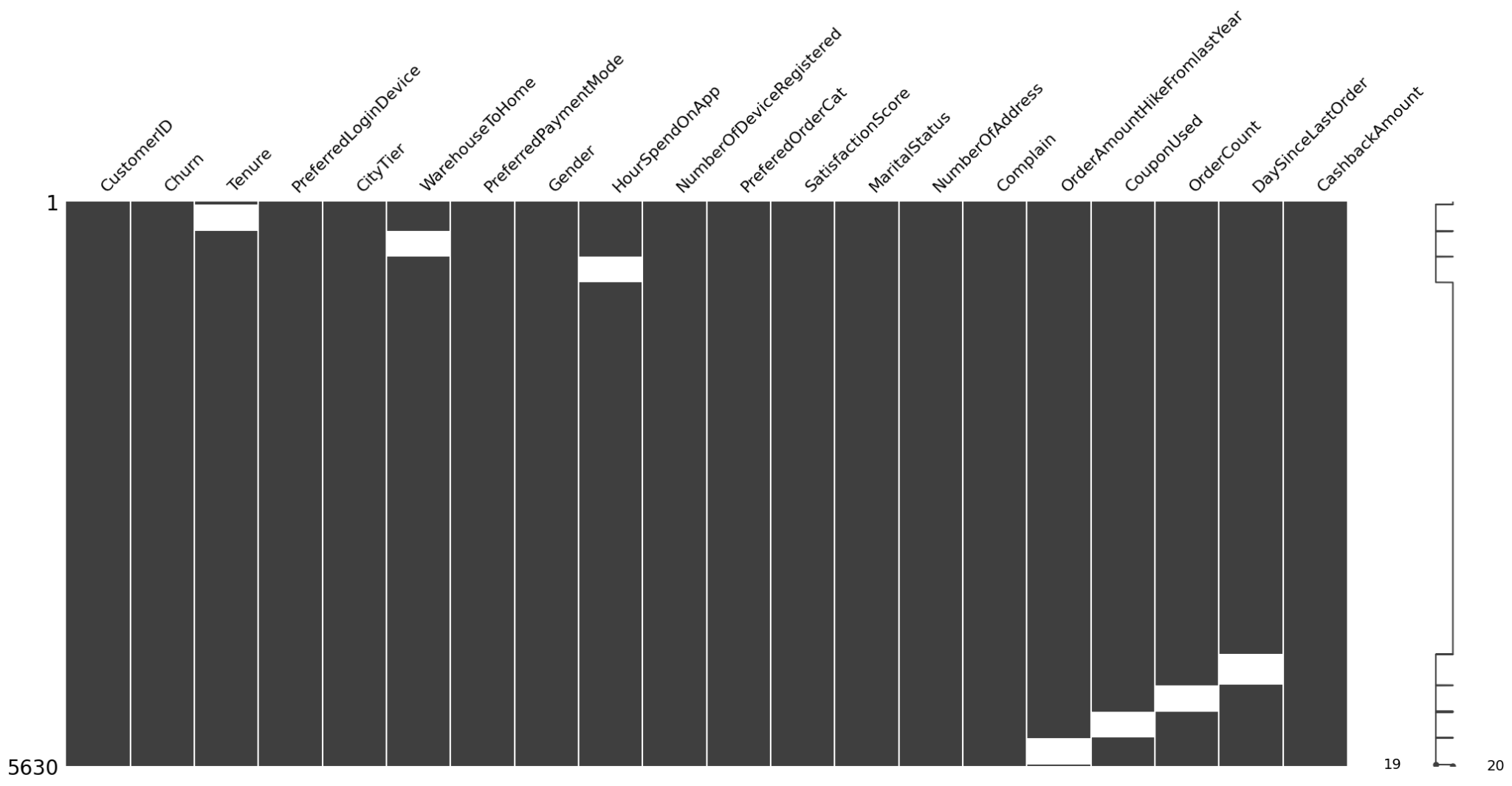
3. How do customer tenure and satisfaction score relate to churn?

4. What are the differences in churn between different customer segments?

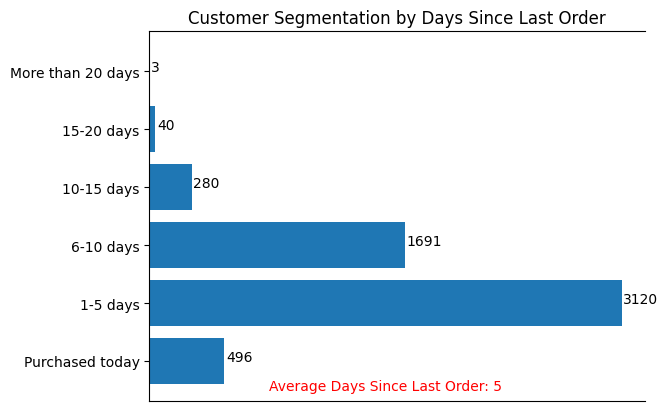
By performing EDA, we aim to enhance our understanding of customer churn in the e-commerce sector, leading to more informed business decisions and ultimately improving customer retention strategies.

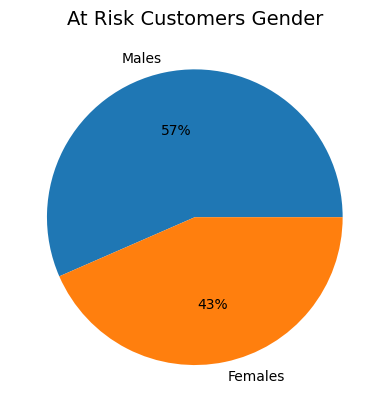
**Inferences from EDA for Customer Churn Documentation:**

**1. Data Imbalance**: The dataset exhibits a class imbalance, where the number of churned customers is significantly lower than non-churned customers. This imbalance may require special attention during model development to ensure accurate predictions for both classes.

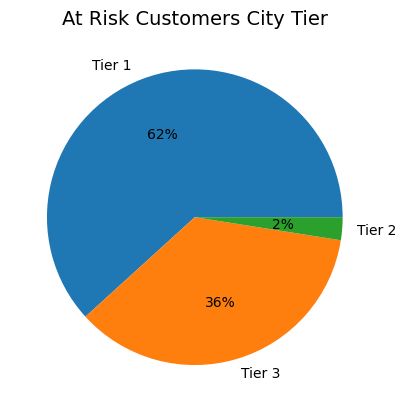
**2. Missing Values:** The dataset contains missing values, specifically in the 'Cashback Amount' feature. The missingness seems to follow a certain pattern, indicating that it is Missing At Random. Proper handling of missing data is essential to avoid bias in the analysis.

**3. At-Risk Customer Segment**: By grouping customer segments based on a five-day interval, we observe a substantial decline in the number of customers placing orders after ten days. We can consider these customers as at-risk, and out of the 481 at-risk customers, 34 eventually churned. Targeted retention strategies for at-risk customers might be beneficial.

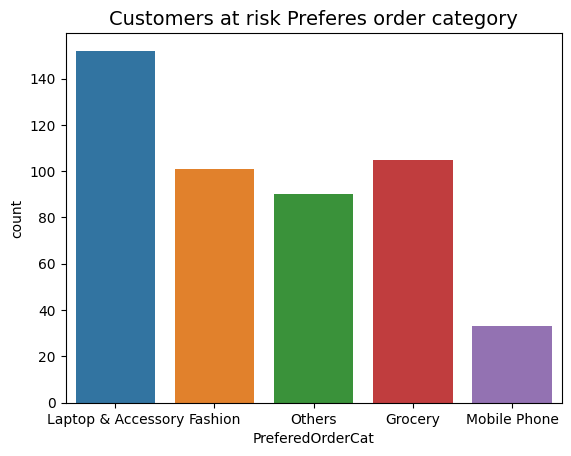


**4. Gender and Churn**: Approximately 57% of churned customers are male, suggesting that men are slightly more at risk of churning. This insight could guide gender-specific marketing strategies.

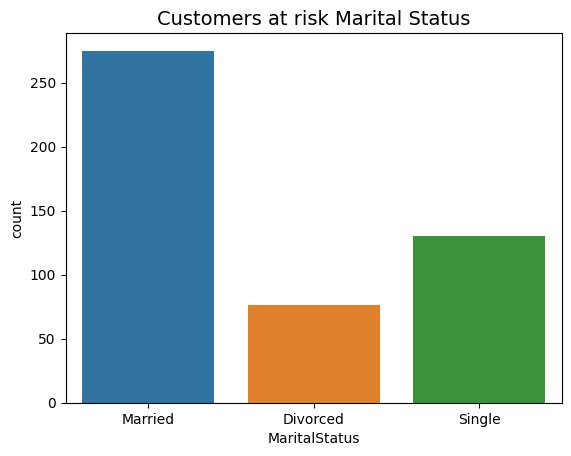
5**. City Tier and Churn**: Tier 1 customers have a higher churn rate of around 62%, indicating that they are more prone to churning. E-commerce businesses can focus on improving services and offers for Tier 1 customers to enhance retention.



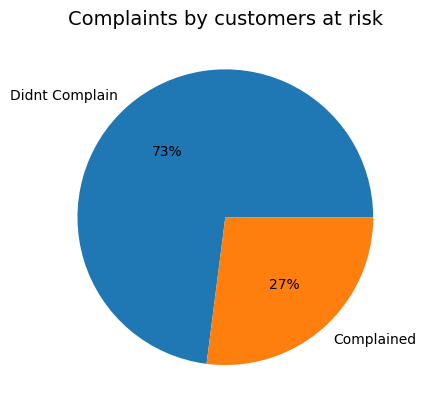
**6. Preferred Order Category**: Customers who prefer to purchase laptops and accessories show a higher likelihood of being at risk of churn. This insight can help tailor promotions or incentives to retain customers in this category.



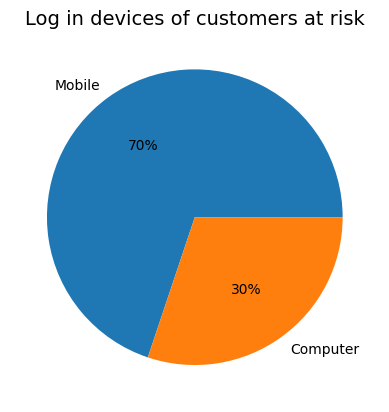
**7. Marital Status and Churn**: Married customers seem to be more at risk of churning. E-commerce companies may explore targeted campaigns to retain married customers.



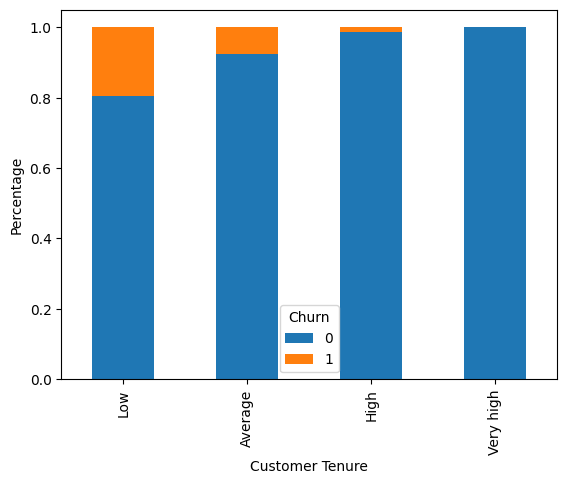
**8. Complaints and Churn**: Surprisingly, 73% of at-risk customers didn't have any complaints. This finding highlights the need to identify other factors that contribute to customer churn.



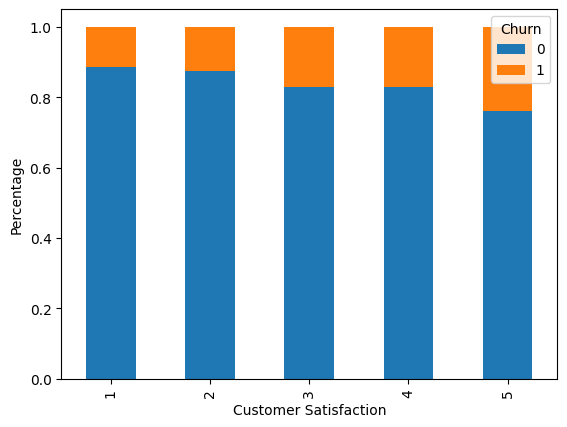
**9. Preferred Login Device**: Most at-risk customers use mobile as their login device. Implementing campaigns or optimizations for mobile users could positively impact retention.



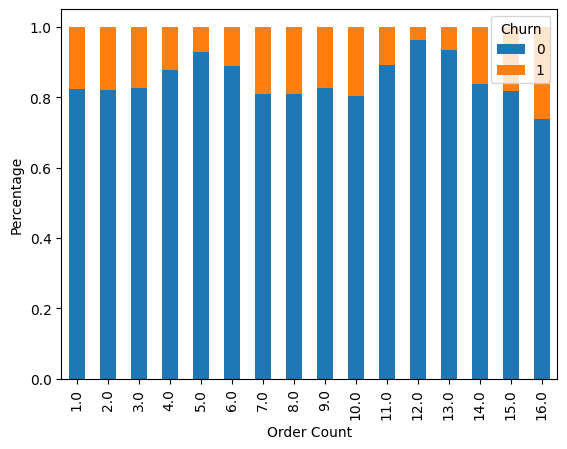
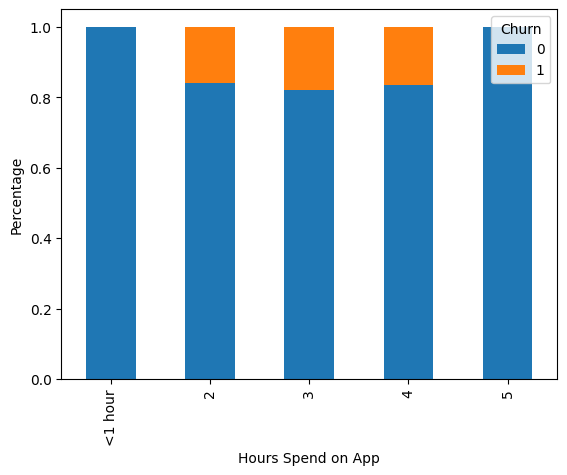
**10. Tenure**: Churn rates significantly reduce after 20 weeks of customer tenure. Retention efforts for long-term customers may prove effective in reducing churn.



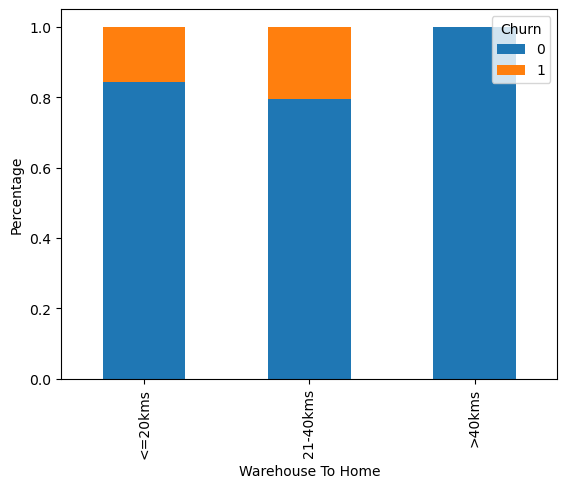
**11. Customer Satisfaction Score**: Strangely, customer satisfaction scores don't seem to be a strong indicator of churn. Churn is distributed across different satisfaction levels, suggesting the presence of other influential factors.



**12. Hours Spent on App and Order Count:** Hours spent on the app and order count do not appear to have a substantial effect on churn. Focusing on other aspects may be more critical in predicting churn.



**13. Warehouse to Home Distance**: Warehouse to home distance doesn't significantly impact churn. Other factors may be more influential in determining customer churn.



These insights from EDA provide a foundation for further analysis and model development to predict customer churn accurately. By leveraging these findings, e-commerce companies can devise targeted retention strategies, enhance customer satisfaction, and reduce customer churn, ultimately leading to improved business performance.

**5. Data Preprocessing**

We have performed the following preprocessing steps for the customer churn project. Here is a detailed explanation of each step and how it can improve the analysis**.**

**1. Data cleaning.**

We have replaced the word "phone" with "mobile phone" in the Preferred Login Device and Preferred Order Category columns, to avoid confusion with landline phones. We have also replaced the abbreviations "CC" and "COD" with "Credit card" and "Cash on Delivery" in the Preferred Payment Mode column, to make it more clear and consistent. We Have converted the City Tier, Satisfaction score and complaints columns from numeric to object data types, so that they are treated as categorical variables and not as continuous ones. This way, each value in these columns will have equal importance and not be affected by the order or magnitude. We have also rounded the order category column to a whole number, since it does not make sense to have fractional categories.

**2. Handling missing values.**

We have filled the missing values using KNN imputation, as the data is Missing at Random (MAR). This means that the missingness of a value depends on some other observed variables, but not on the value itself. KNN imputation is a method that uses the k nearest neighbors of a record to estimate its missing values, based on the similarity of the observed variables. This method can preserve the distribution and relationships of the data better than using mean or median imputation.

**3. Handling outliers.**

We have used Winsorization on the Tenure and Number of Addresses columns, to reduce the effect of extreme values on the analysis. Winsorization is a technique that replaces the values above or below a certain percentile with the values at that percentile. For example, if we use 5% Winsorization, we replace the values above the 95th percentile with the value at the 95th percentile, and the values below the 5th percentile with the value at the 5th percentile. This way, we can reduce the skewness and variance of the data without losing too much information.

**4. Feature selection.**

We have conducted Two Sample T-Test to see which numerical columns to select for the analysis. The Two Sample T-Test is a statistical test that compares the means of two groups (in this case, churned and non-churned customers) and determines if they are significantly different from each other. We have selected only those columns that have a p-value less than 0.05, which means that there is less than 5% chance that the difference in means is due to random variation. We have conducted chi-square test of independence on categorical columns, to see which ones are associated with churn. The chi-square test of independence is a statistical test that compares the frequencies of different categories in two groups (in this case, churned and non-churned customers) and determines if they are independent of each other. We have selected only those columns that have a p-value less than 0.05, which means that there is less than 5% chance that the frequencies are independent of each other. After selecting these columns, Wehave rejected five columns: Customer ID, Hours Spend on App, Order Count, Order Hike From Last Year and Coupon Used. The Null Hypothesis for both tests was that The columns do not have an effect on churn.

**5. Encoding categorical variables.**

We have used one hot encoding on the categorical variables, to convert them into binary variables that can be used by machine learning algorithms. One hot encoding is a technique that creates a new column for each unique value in a categorical variable, and assigns 1 or 0 depending on whether that value is present or not in a record. For example, we have a column called City Tier with values 1, 2 and 3, we can create three new columns called City Tier\_1, City Tier\_2 and City Tier\_3, and assign 1 or 0 accordingly. This way, we can avoid assigning any order or weight to categorical values that are not ordinal.

**6. Handling Data Imbalance.**

As the data was highly imbalanced, with more non-churned customers than churned customers, We have used SMOTE to balance it. SMOTE is a technique that generates synthetic samples of the minority class (in this case, churned customers) by using a combination of nearest neighbors and random perturbation. This way, we can increase the size of the minority class without creating exact duplicates or losing information from the original data.

**7. Data Scaling**

We have used Standard Scaling to scale the numerical variables, to make them comparable and reduce the effect of different units or ranges on the analysis. Standard Scaling is a technique that subtracts the mean and divides by the standard deviation of each variable, resulting in a standardized distribution with zero mean and unit variance. This way, we can avoid any bias or distortion caused by variables with large or small values.

**6. Model Selection**

Classification model building is a pivotal aspect of machine learning that involves creating algorithms capable of categorizing data points into distinct classes or labels based on their features. Once the data is prepared, the choice of classification algorithm becomes paramount. Decision trees, for instance, create a hierarchical structure of decisions based on feature values, while ensemble methods like random forests combine multiple decision trees to enhance accuracy and reduce overfitting. Support Vector Machines (SVMs) aim to find hyperplanes that optimally separate different classes, while k-Nearest Neighbors (k-NN) classify data points based on the labels of their closest neighbors. Logistic regression, despite its name, is a fundamental classification algorithm used to predict binary outcomes by modeling the probability of an instance belonging to a particular class.

After selecting an appropriate algorithm, the model is trained using labeled data, which allows it to learn the relationships between features and classes. This training process entails fine-tuning algorithm parameters to achieve optimal performance. Model evaluation comes next, where metrics like accuracy, precision, recall, and F1-score are used to assess the model's effectiveness. It's important to choose the evaluation metrics that align with the specific goals of the project, which in this case is predicting churn.

We used the following algorithms to test our models.

**Logistic Regression:**

A supervised learning algorithm that is used to predict the probability of a categorical outcome. It is based on the logistic function, which maps any real value to a value between 0 and 1. Logistic regression can be used for binary classification (such as spam detection) or multiclass classification (such as digit recognition).

**KNN:**

A non-parametric algorithm that is used to classify new data points based on the similarity to the existing data points. It is also known as the k-nearest neighbors algorithm, because it finds the k closest data points to the new data point and assigns it the most common class among them. KNN can be used for both classification and regression problems.

**SVM:**

A supervised learning algorithm that is used to find a hyperplane that separates the data points of different classes with maximum margin. It is also known as the support vector machine, because it only depends on the data points that are closest to the hyperplane, called the support vectors. SVM can be used for both linear and non-linear classification problems.

**Decision Trees:**

A supervised learning algorithm that is used to create a tree-like structure that splits the data into subsets based on some criteria. Each node in the tree represents a feature, each branch represents a decision rule, and each leaf represents an outcome. Decision trees can be used for both classification and regression problems.

**Random Forest:**

An ensemble learning algorithm that is used to improve the performance of decision trees by creating multiple trees from different subsets of the data and averaging their predictions. It is also known as the random forest classifier, because it reduces the overfitting and variance of decision trees by introducing randomness and diversity.

**Linear Discriminant Analysis:**

A dimensionality reduction technique that is used to find a linear combination of features that best separates the data points of different classes. It is also known as Fisher's linear discriminant, because it maximizes the ratio of between-class variance to within-class variance. LDA can be used for both classification and dimensionality reduction problems.

**XGBoost Classifier:**

A gradient boosting algorithm that is used to create an ensemble of weak learners (usually decision trees) that are sequentially trained to correct the errors of the previous learners. It is also known as extreme gradient boosting, because it uses a more regularized model formalization and a more efficient tree learning algorithm than other gradient boosting algorithms.

**Adaboost Classifier:**

Another gradient boosting algorithm that is used to create an ensemble of weak learners (usually decision stumps) that are sequentially trained to focus more on the misclassified data points by assigning them higher weights. It is also known as adaptive boosting, because it adapts the weights of the data points according to their importance.

To evaluate the performance of machine learning models for customer churn prediction, some common metrics are accuracy, precision, recall, and F1-score. These metrics can be calculated from a confusion matrix, which is a table that shows the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for a binary classification problem.

A **confusion matrix** is a table that summarizes the performance of a machine learning model on a set of test data for which the true labels are known. It shows how many instances of each class were correctly predicted by the model (true positives and true negatives) and how many were incorrectly predicted (false positives and false negatives). A confusion matrix can be used to calculate various metrics that measure the accuracy, precision, recall, specificity, and F1-score of a model. These metrics can help to evaluate the strengths and weaknesses of a model and compare it with other models.

**Accuracy** is the proportion of correct predictions among all predictions. It is calculated as (TP + TN) / (TP + FP + TN + FN). Accuracy measures how well the model predicts both positive and negative classes, but it can be misleading if the classes are imbalanced.

**Precision** is the proportion of correct positive predictions among all positive predictions. It is calculated as TP / (TP + FP). Precision measures how reliable the model is when it predicts a positive class, but it does not account for how many positive instances it misses.

**Recall** is the proportion of correct positive predictions among all positive instances. It is calculated as TP / (TP + FN). Recall measures how complete the model is when it predicts a positive class, but it does not account for how many negative instances it misclassifies.

**F1-score** is the harmonic mean of precision and recall. It is calculated as 2 \* (precision \* recall) / (precision + recall). F1-score balances both precision and recall and gives more weight to low values. F1-score is a good metric to use when the classes are imbalanced or when both precision and recall are important.

After the models were run, we chose Linear Regression, SVM and Decision Tree as the weak classifiers for Adaboost Classifier, as Adaboost performed the best in the test. We used both Grid Search CV and Optuna for hyperparameter tuning.

**Grid Search CV:**

A hyperparameter tuning technique that is used to find the optimal combination of hyperparameters for a machine learning model by exhaustively searching over a predefined grid of values. It is also known as grid search cross-validation, because it uses cross-validation to evaluate the performance of each combination.

**Optuna:**

A hyperparameter optimization framework that is used to find the optimal combination of hyperparameters for a machine learning model by using various search algorithms (such as random search, TPE, CMA-ES, etc.) and pruning strategies (such as median pruning, hyperband, etc.). It is also known as Optuna optimization, because it automates and simplifies the hyperparameter tuning process.

After both tests were run, Optuna gave the best results. It selected Decision Tree as the best weak classifier for Adaboost Classifier.

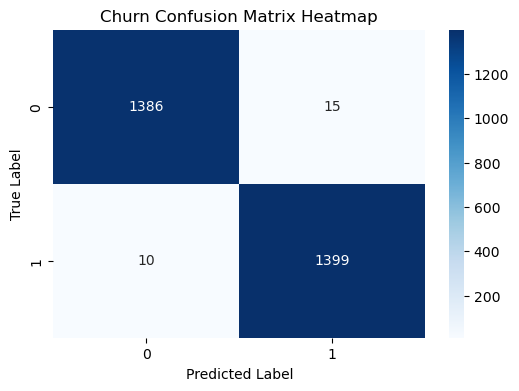
The parameters it selected were 'n\_estimators': 147, 'learning\_rate': 0.42329791924956006, 'algorithm': 'SAMME', 'base\_estimator': 'decision\_tree', 'criterion': 'gini', 'max\_depth': 27, 'min\_samples\_split': 7.

We ran the Adaboost Classifier with Decision Tree as the weak classifier and the above parameters. The accuracy for the above model was 99.1%.

**7. Result**

In this section, we present a comprehensive evaluation of the classification model's performance using an AdaBoost classifier. We utilized the dataset for customer churn prediction to assess the model's efficacy in accurately predicting whether customers are likely to churn or not.

**The confusion matrix.**

****

The confusion matrix shows that out of the 1409 cases of churn, our model correctly predicted 1399 cases and incorrectly classified 10 as not being at risk of getting churned. So this model is very good at detecting churn.

**1. Accuracy:** The model achieved an impressive accuracy score of approximately 99.11%. This demonstrates our model's ability to make accurate predictions in a vast majority of cases. High accuracy indicates that we were successful in capturing the underlying patterns and nuances within the customer data.

**2. Precision:** Our precision score stood at approximately 98.94%. This signifies our model's ability to correctly classify positive instances while minimizing false positives. It shows that we were cautious in labeling instances as potential churn cases, leading to minimal misclassification.

**3. Recall:** The recall score, also known as sensitivity, was calculated to be about 99.29%. This metric underscores our model's capability to identify a significant proportion of actual positive churn cases. It reflects our success in capturing potential churn scenarios effectively.

**4. F1 Score:** The F1 score, which balances precision and recall, was computed at around 99.11%. This score showcases our model's ability to strike a balance between accurate positive predictions and identifying a large portion of actual churn cases.

**8. Hosting**

**FLASK:** Flask is a micro web framework written in Python that allows you to build web applications quickly and with minimal boilerplate code. It's designed to be lightweight, flexible, and easy to use, making it an excellent choice for building small to medium-sized web applications, APIs, and prototypes.

**Key Features of Flask:**

**1. Micro Framework:** Flask is often referred to as a "micro" framework because it provides the essential tools for building web applications without imposing too much structure. This gives developers the freedom to choose the components they need and integrate third-party libraries as desired.

**2. Routing:** Flask allows you to define URL routes, which map URLs to Python functions. This makes it easy to handle different requests and execute the corresponding code.

**3. Templates:** Flask includes a templating engine that helps you generate dynamic HTML content. We can create templates with placeholders for dynamic data and then render these templates with actual data when generating responses.

**4. View Functions:** View functions are Python functions that handle requests and return responses. Flask simplifies this process, allowing us to focus on writing the logic for our application.

**5. Jinja2 Templating:** Flask uses the Jinja2 templating engine, which provides advanced template inheritance, filters, and macros. This allows us to create reusable templates and maintain a clean separation of concerns between application logic and presentation.

**6. Web Forms:** Flask includes tools for handling HTML forms and validating user input. It makes it easy to build forms, validate data, and display validation errors to users.

**7. Static Files:** We can serve static files like images, stylesheets, and JavaScript directly from Flask. This simplifies the process of including static assets in your application.

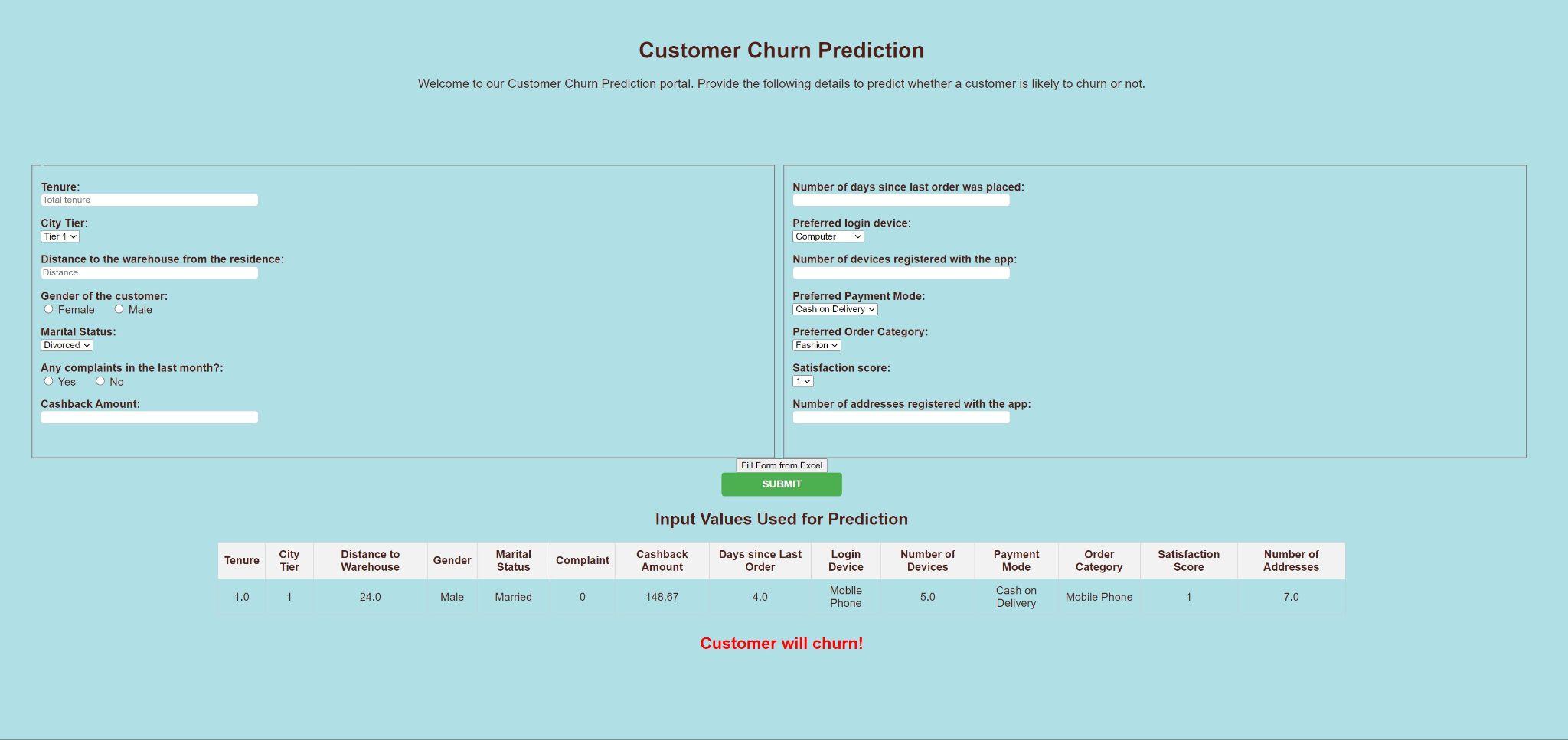
**8. Extension Support:** Flask has a modular architecture that allows you to add extensions for various features, such as database integration, user authentication, caching, and more. These extensions enhance the functionality of our application.

**9. URL Building:** Flask provides a convenient way to generate URLs within your application using the `url\_for` function. This helps ensure that our URLs are consistent and easy to manage.

**10. Lightweight:** Flask is minimalistic and doesn't impose a rigid structure on your project. This makes it well-suited for small to medium-sized projects and for developers who prefer more control over their application's components.

Flask's simplicity and flexibility make it an excellent choice for developers who want to build web applications quickly without the overhead of a larger framework. Its ease of use, combined with the availability of extensions and a strong community, make it a popular option for a wide range of web development projects.

The website is hosted using pythonanywhere.com. The website can either accept values from the user or can get values from a stored dataset. Once the values are entered the website uses the model to predict if the customer will churn or not.



**9. Conclusion**

Based on the detailed analysis and evaluation of our churn prediction model's performance, we can draw a comprehensive conclusion that highlights the model's effectiveness and potential implications:

Our customer churn prediction model has exhibited outstanding capabilities in identifying customers who are likely to churn. The analysis of the confusion matrix reveals that out of the total 1409 cases of churn, the model correctly predicted 1399 cases, while only misclassifying 10 cases. This remarkable accuracy underscores the model's ability to make highly precise predictions, which can significantly impact business strategies aimed at customer retention.

1. Accuracy: The achieved accuracy score of approximately 99.11% is a strong indicator of our model's proficiency in making accurate predictions. Such a high accuracy percentage reflects its skill in capturing intricate patterns within the customer data. This accuracy assures us that our model is making predictions that closely align with the actual outcomes.

2. Precision: Our precision score of about 98.94% highlights the model's discernment in labeling positive churn instances. This is crucial from a business standpoint as it means that when the model predicts a customer as likely to churn, the probability of that prediction being correct is extremely high. The low number of false positives showcases the model's prudence in its predictions.

3. Recall: The recall score of around 99.29% emphasizes our model's capability to identify a significant majority of actual positive churn cases. This high recall signifies that the model is not overlooking potential churn scenarios. Its sensitivity to actual churn instances is essential for businesses seeking to proactively address customer retention strategies.

4. F1 Score: The F1 score, standing at approximately 99.11%, demonstrates our model's ability to balance precision and recall effectively. This score indicates that our model can make precise positive predictions while still capturing a considerable number of actual churn cases. It's a comprehensive metric that showcases the model's overall robustness.

In conclusion, our churn prediction model presents a robust and highly accurate solution for anticipating customer churn. The precision, recall, and F1 score collectively reinforce its reliability and practical utility for businesses. By leveraging the insights from this model, companies can proactively identify customers at risk of churning and implement targeted strategies to retain their valuable clientele. The model's exceptional performance promises to have a substantial positive impact on businesses' bottom lines by optimizing customer retention efforts.

**References**

* Rajasekharan, B., & Srinivasan, S. (2021). E-commerce customer churn prevention using machine learning-based business intelligence strategy. International Journal of Business Intelligence and Data Mining, 18(1), 1-18
* Alshomrani, S., & Qamar, S. (2018). E-commerce customer churn prediction using ensemble learning and K-means clustering. International Journal of Advanced Computer Science and Applications, 9(5), 1-9
* Alshomrani, S., & Qamar, S. (2019). E-commerce customer churn prediction using machine learning and deep learning techniques. International Journal of Advanced Computer Science and Applications, 10(2), 1-10
* Chen, Y., & Xie, J. (2005). Third-party product review and firm marketing strategy. Marketing Science, 24(2), 218-240.
* Dahiya, R., & Bhatia, K. K. (2018). Customer churn analysis in the e-commerce industry using machine learning. In 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) (pp. 1-6). IEEE.
* Gao, Y., Liang, C., & Liang, Y. (2019). E-commerce customer churn prediction based on improved random forest algorithm. In Proceedings of the 2019 4th International Conference on Multimedia Systems and Signal Processing (pp. 1-5).
* Gupta, A., & Singh, A. (2020). Customer churn prediction in e-commerce using machine learning techniques: A comparative analysis. In Proceedings of the International Conference on Innovative Computing & Communications (ICICC) 2020 (pp. 1-7). Springer, Singapore.
* Liang, C., Gao, Y., & Liang, Y. (2020). E-commerce customer churn prediction based on improved XGBoost algorithm. In Proceedings of the 2020 International Conference on Big Data Engineering and Technology (pp. 1-5)
* Wu, Z., Jing, L., Wu, B., & Jin, L. (2022). A PCA-AdaBoost model for E-commerce customer churn prediction. Annals of Operations Research
* Flask Documentation. (2023). Flask - A Python Web Framework. <https://flask.palletsprojects.com/>
* Scikit-learn Documentation. (2023). Scikit-learn: Machine Learning in Python.<https://scikit-learn.org/>
* Optuna Documentation. (2021). Optuna: A Hyperparameter Optimization Framework.<https://optuna.readthedocs.io/>
* Understanding Classification Evaluation Metrics. (2021). Machine Learning Mastery.<https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/>
* Tour of Evaluation Metrics for Imbalanced Classification <https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/>