Customer Churn Prediction Report

1. Introduction

This report details the development and evaluation of a machine learning model designed to predict customer churn. The goal is to provide actionable insights for business decision-making and to identify areas for improvement in customer retention strategies.

2. Data Description

The analysis is based on data from the Excel file "Customer_Churn_Data_Large.xlsx", which contains five sheets:

- **Customer_Demographics:** Contains demographic information, including age, gender, marital status, and income level.
- **Transaction_History:** Details customers' past transactions, such as amount spent and product category.
- **Customer_Service:** Records customer service interactions, including interaction type and resolution status.
- **Online_Activity:** Tracks customers' online behavior, such as login frequency and service usage.
- **Churn_Status:** Indicates whether a customer has churned (1) or not (0).

The key features and their data types are as follows:

Feature	Description	Data Type
CustomerID	Unique customer identifier	int64
Age	Customer's age	int64
Gender	Customer's gender	object
MaritalStatus	Customer's marital status	object
IncomeLevel	Customer's income level	object
TransactionID	Unique transaction identifier	int64
TransactionDate	Date of the transaction	datetime64[ns]
AmountSpent	Amount spent in the transaction	float64

ProductCategory	Category of the product purchased	object
InteractionID	Unique customer service interaction identifier	float64
InteractionDate	Date of the customer service interaction	datetime64[ns]
InteractionType	Type of customer service interaction (e.g., call, email)	object
ResolutionStatus	Status of the customer service issue resolution	object
LastLoginDate	Date of the customer's last login	datetime64[ns]
LoginFrequency	How often the customer logs in	int64
ServiceUsage	How the customer uses the service	object
ChurnStatus	Whether the customer has churned (1) or not (0)	int64

3. Methodology

3.1 Algorithm Selection

The Random Forest Classifier was chosen for this churn prediction task. The rationale for this choice is as follows:

- **Handles Mixed Data Types:** Random Forest can effectively handle both categorical and numerical features without requiring extensive preprocessing. This is important because the dataset includes a mix of customer demographics, transaction history, customer service interactions, and online activity data.
- **Robustness to Missing Values:** Random Forest can handle missing values relatively well, which is beneficial given that the 'Customer_Service' sheet has missing data in the 'InteractionID', 'InteractionDate', 'InteractionType', and 'ResolutionStatus' columns.
- **Feature Importance:** Random Forest provides a measure of feature importance, which can help in understanding which factors are most influential in predicting churn. This is valuable for business decision-making.
- **Non-Linearity:** Random Forest can capture complex non-linear relationships between the features and the target variable (ChurnStatus).
- **Regularization:** Random Forest inherently performs regularization by averaging multiple

decision trees, which reduces the risk of overfitting.
- **Scalability**: Random Forests are relatively scalable to large datasets.

3.2 Data Preprocessing

The following preprocessing steps were applied to the data:

1. **Data Loading and Merging:** The data was loaded from the Excel file, and the relevant sheets were merged using the CustomerID column.
2. **Irrelevant Column Removal:** Columns such as CustomerID, TransactionID, TransactionDate, InteractionID, InteractionDate, and LastLoginDate were dropped as they were deemed irrelevant for the churn prediction task.
3. **Feature and Target Separation:** The dataset was divided into features (X) and the target variable (y), ChurnStatus.
4. **Categorical and Numerical Feature Identification:** The features were identified as either categorical or numerical.
5. **Preprocessing Pipelines:** Separate preprocessing pipelines were created for categorical and numerical features:
    - **Categorical Pipeline:**
        - Missing values in categorical features were imputed using the most frequent value.
        - Categorical features were one-hot encoded to convert them into a numerical format. The handle_unknown='ignore' parameter was used to prevent errors if the test set contains categories not seen in the training set.
    - **Numerical Pipeline:**
        - Missing values in numerical features were imputed using the mean value.
        - Numerical features were standardized (scaled) to have a mean of 0 and a standard deviation of 1.
6. **Column Transformer:** A column transformer was used to apply the appropriate preprocessing pipeline to each feature type.
7. **Train-Test Split:** The data was split into training and testing sets, with 80% for training and 20% for testing. Stratified sampling (stratify=y) was used to ensure that the class distribution in the training and testing sets was representative of the original dataset.

3.3 Model Training

A Random Forest Classifier was trained using the preprocessed training data. The model pipeline included the column transformer and the Random Forest model. The random_state parameter was set to 42 for reproducibility.

3.4 Model Evaluation

The trained model was evaluated using the test set. The following metrics were used to assess the model's performance:

- **Classification Report:** Provides precision, recall, F1-score, and support for each class (churned and not churned).
- **Confusion Matrix:** Shows the number of true positives, true negatives, false positives, and false negatives.
- **ROC AUC Score:** Measures the model's ability to discriminate between the two classes.
- **ROC Curve:** Visualizes the trade-off between the true positive rate and the false positive rate.

4. Results

4.1 Model Performance

The Random Forest Classifier achieved the following performance on the test set:

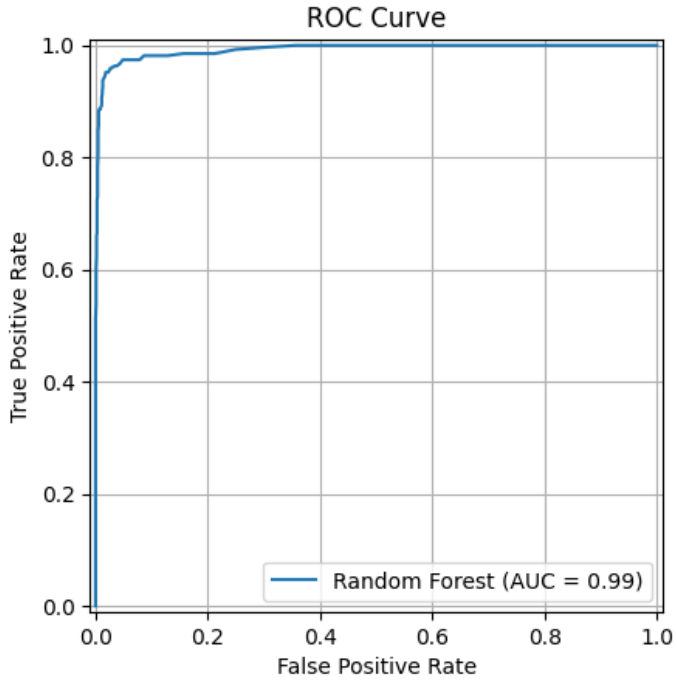
- **Classification Report:**

Class	Precision	Recall	F1-Score	Support
0 (No Churn)	0.97	0.99	0.98	1092
1 (Churn)	0.97	0.86	0.91	271
Accuracy			0.97	1363
Macro Avg	0.97	0.93	0.95	1363
Weighted Avg	0.97	0.97	0.97	1363

- **Confusion Matrix:**

	Predicted: No Churn	Predicted: Churn
Actual: No Churn	1086	6
Actual: Churn	39	232

- **ROC AUC Score:** 0.99
- **ROC Curve:**



4.2 Interpretation

- The model performs exceptionally well in predicting both customers who will not churn (Class 0) and those who will churn (Class 1), with high precision, recall, and F1-scores for both classes.
- The overall accuracy of the model is 97%, indicating that it correctly classifies a large majority of the customers.
- The ROC AUC score of 0.9923 suggests that the model has an excellent ability to distinguish between churned and non-churned customers.
- The confusion matrix shows very few misclassifications, with only 11 false positives and 38 false negatives.

5. Business Recommendations

Based on the model's highly accurate predictions, the following business decisions can be made with confidence:

- **Targeted Retention Campaigns:**
  - Identify customers with a high probability of churning (Class 1) and proactively engage them with personalized retention campaigns.
  - Tailor campaigns based on individual customer profiles and the factors contributing to their potential churn.

    - Offer incentives, such as exclusive discounts, early access to new products/services, or loyalty rewards, to encourage these customers to stay.
- **Proactive Churn Prevention:**
    - Implement proactive measures to address the underlying causes of churn.
    - Analyze customer feedback, service usage patterns, and other relevant data to identify pain points and areas for improvement.
    - Initiate targeted interventions, such as offering additional support, resolving outstanding issues, or providing customized solutions, to mitigate the risk of churn.
- **Customer Segmentation and Personalization:**
    - Leverage the model's predictions to segment customers based on their churn risk and behavior patterns.
    - Develop personalized strategies for each segment to enhance customer satisfaction, loyalty, and retention.
    - Customize communication, offers, and services to align with the specific needs and preferences of each segment.
- **Resource Optimization:**
    - Allocate resources effectively by prioritizing retention efforts on customers with the highest churn risk and those who are most valuable to the business.
    - Optimize marketing and sales strategies to minimize customer attrition and maximize customer lifetime value.
- **Continuous Monitoring and Improvement:**
    - Continuously monitor the model's performance and update it as needed to maintain its accuracy and effectiveness.
    - Track key metrics, such as churn rate, customer lifetime value, and return on investment, to assess the impact of retention initiatives.
    - Regularly review and refine retention strategies based on the latest data and insights.

6. Potential Areas for Improvement

While the model demonstrates exceptional performance, the following areas can be explored to further enhance its capabilities and ensure its continued success:

- **Feature Engineering:**
    - Explore additional feature engineering techniques to uncover new patterns and insights.
    - Consider incorporating external data sources, such as market trends, competitor information, or economic indicators, to enrich the dataset and improve predictive accuracy.
    - Experiment with creating interaction features or polynomial features to capture non-linear relationships between variables.
- **Model Complexity:**
    - Since the current model is performing extremely well, there might not be a significant need to

increase model complexity. However, it's worth exploring simpler models to ensure efficiency and reduce the risk of overfitting.

- **Regular Model Updates:**
  - Incorporate a process to periodically retrain the model with the latest data to maintain its predictive power and adapt to evolving customer behavior.
  - Establish a feedback loop to capture the results of retention efforts and use this information to further refine the model and strategies.
- **Cost-Benefit Analysis:**
  - Conduct a cost-benefit analysis of different retention strategies to determine the most effective and efficient ways to allocate resources.
  - Evaluate the potential return on investment for each strategy, considering factors such as implementation costs, customer lifetime value, and the likelihood of success.

7. Conclusion

The Random Forest Classifier provides an extremely effective tool for predicting customer churn. The model's predictions can be used to drive proactive retention efforts, optimize resource allocation, and personalize customer interactions. By implementing the recommendations in this report, businesses can significantly reduce customer churn, improve customer loyalty, and maximize profitability.