Customer Churn Prediction

Task 2: Predictive Modeling and Evaluation Report

1. Introduction
The objective of Task 2 is to develop, evaluate, and interpret a predictive model for customer churn using the cleaned and preprocessed dataset produced in Task 1. The focus is on establishing a reliable baseline model, assessing performance using appropriate metrics, and identifying strengths and limitations.

2. Modeling Objective
Customer churn prediction is formulated as a binary classification problem. The target variable ChurnStatus indicates whether a customer has churned (1) or remained active (0). The business objective is to identify customers at risk of churning in order to enable proactive retention strategies.

3. Data Preparation for Modeling
Identifier variables were removed prior to modeling to prevent noise and overfitting. Raw date variables were initially found to cause data leakage and were therefore excluded. Recency-based features were engineered to capture temporal behaviour in a valid and interpretable manner, including DaysSinceLastLogin, DaysSinceLastInteraction, and DaysSinceLastTransaction.

4. Feature Types
The final dataset consisted of numeric features (Age, AmountSpent, LoginFrequency, and recency variables) and categorical features (Gender, MaritalStatus, IncomeLevel, ProductCategory, InteractionType, ResolutionStatus, and ServiceUsage).

5. Preprocessing Pipeline
Numeric features were imputed using the median and standardised using z-score normalisation. Categorical features were imputed using the most frequent category and encoded using one-hot encoding. All preprocessing steps were applied within a unified pipeline to prevent data leakage.

6. Model Selection and Training
A logistic regression model was selected as the baseline classifier due to its interpretability and efficiency. Class imbalance was addressed using class-weight balancing. The dataset was split into training and testing sets using an 80/20 stratified split.

7. Model Evaluation
Model performance was evaluated using precision, recall, F1-score, ROC–AUC, and a confusion matrix. The model achieved a realistic ROC–AUC of approximately 0.60, reflecting limited but meaningful predictive power given the available data. Recall for churned customers was prioritised due to business considerations.

8. Feature Importance and Interpretation
Feature coefficients revealed that customer inactivity, reduced engagement, and unresolved service issues were the strongest drivers of churn. Conversely, frequent logins, resolved

interactions, and higher income levels were associated with customer retention.

9. Model Limitations
Model performance was constrained by the absence of longitudinal behavioural data, customer tenure information, and usage trends. These limitations highlight opportunities for future data collection and model improvement.

10. Conclusion
This task successfully delivered a transparent and methodologically sound churn prediction model. While predictive performance was moderate, the model learned interpretable and business-relevant patterns, providing a strong foundation for future enhancements.