CCS 247 – MACHINE LEARNING
FINALS PROJECT

Case Study: SmartBank—Customer Satisfaction Analysis

Background:
SmartBank is a large financial institution with over 10 million customers and branches across the country. The bank offers various services including savings accounts, checking accounts, loans, and credit cards. Despite its large customer base and extensive service offerings, SmartBank has recently faced increasing customer dissatisfaction and a higher-than-average customer attrition rate.

Scenario:
Jordan, the head of customer experience at SmartBank, is tasked with investigating the rise in customer dissatisfaction and finding ways to improve customer retention. Jordan has received anecdotal feedback from branch managers and customer support teams about specific complaints, such as issues with online banking services, long wait times at branches, and unresponsive customer service.

SmartBank wants to proactively address these issues and reduce customer attrition rates. Jordan believes that a classification model can help identify customers who are at risk of leaving the bank based on their interactions with the bank and service usage patterns.

Objective:
Your goal is to develop machine learning models to predict whether a customer is likely to leave the bank (churn) in the next six months. By identifying customers at risk of churning, SmartBank can take targeted actions to improve customer satisfaction and retain these customers.

Task 1: Data Preprocessing and Exploratory Data Analysis
1. Clean the data, handle missing values, and encode categorical features as needed.
2. Perform EDA on the dataset to understand its structure and key insights. Provide a discussion of the things that you did for EDA that includes the answers to the following questions:
   - What is the distribution of the target variable (churn)?
   - Are there any correlations between features and the target variable?
   - What are the distribution and summary statistics (e.g., mean, median, standard deviation) for each numerical feature?
   - Are there any categorical features with imbalanced classes?
   - Are there any outliers in the data that need to be addressed?
   - What trends or patterns can be observed in customer complaints or branch visits?
   - How do different account types or demographics (age, gender) relate to customer churn?

Task 2: Model Training and Evaluation
1. Choose two machine learning algorithms to train on the data. Suggested algorithms include:
   - Logistic Regression
   - Decision Tree
   - Naïve Bayes
   - k-Nearest Neighbors (k-NN)
   - Support Vector Machines (SVM)
2. Model Evaluation: Evaluate the models using a variety of metrics such as accuracy, precision, recall, and F1-score.
3. Compare the performance of the two models and provide a conclusion on which model performs better for predicting customer churn.
4. Provide a report detailing the algorithms chosen, methods used, performance evaluation, and conclusion.

Submission Guidelines:

1. Submit the following:
    a. Notebook should be submitted in HTML format (Jupyter notebook has a *"download as HTML"* feature).
    b. Answer to discussions should be submitted in PDF format.
        i. Format: Arial, 11pt, Single-spaced
2. Include screenshots of figures in your discussions.
3. A Google Drive link will be provided for submission.
4. Deadline: May 24, 2024 | 11:59PM

CCS 247 – MACHINE LEARNING
FINAL PROJECT
Rubrics

Criteria	Excellent (5)	Good (4)	Fair (3)	Needs Improvement (2)	Poor (1)
Data Preprocessing	Thoroughly handles missing values and encodes categorical features. Uses effective strategies to deal with outliers and ensures data quality.	Handles missing values and encodes categorical features adequately. Manages outliers reasonably well.	Addresses missing values and encodes categorical features. May overlook some outliers.	Lacks consistent data cleaning approach. Significant missing values and/or outliers remain.	Little to no data preprocessing performed.
Exploratory Data Analysis (EDA)	Provides deep insights into data structure, trends, and patterns. Effectively visualizes data and answers EDA questions comprehensively.	Offers good insights into data structure and trends. Answers EDA questions well with sufficient visualizations.	Answers EDA questions and presents some visualizations but lacks depth or variety in analysis.	Answers EDA questions with minimal insights and limited visualizations.	Inadequate EDA with little or no visualizations.
Model Training	Trains two models effectively, using hyperparameter tuning and validation methods to optimize performance.	Trains two models adequately, with some tuning and validation for better performance.	Trains two models with basic parameter settings; little to no tuning.	Trains models but lacks tuning or validation, impacting performance.	Trains models with little regard for proper parameter settings.
Model Evaluation	Thoroughly evaluates models using various metrics (accuracy, precision, recall, F1-score) and interprets results accurately.	Evaluates models using appropriate metrics and interprets results well.	Evaluates models but lacks some metrics or interpretation depth.	Evaluates models with limited metrics or poor interpretation of results.	Inadequate or incorrect model evaluation and interpretation.
Performance Comparison	Provides comprehensive and insightful comparison of model performance, explaining strengths and weaknesses of each.	Provides solid comparison of model performance, explaining strengths and weaknesses clearly.	Offers comparison of model performance but lacks depth in explanations.	Compares model performance with limited explanations.	Minimal or no performance comparison provided.
Final Report	Submits a clear, well-structured report detailing the approach, methods, evaluation, and conclusions.	Submits a clear report detailing approach, methods, evaluation, and conclusions.	Submits a report with some structure but may lack clarity or depth in explaining methods and evaluation.	Submits a report, but it lacks clarity or depth in key areas.	Submits a disorganized or unclear report lacking key components.