DS605 Fundamentals of Machine Learning
Academic Year 2025-26 (Autumn)

LAB - 5

End-to-End Classification Pipeline

The Challenge: Predicting Telecom Customer Churn

A major telecom provider is experiencing significant customer churn (customers leaving their service). Your task is to build and evaluate a machine learning model that can predict which customers are likely to churn based on their account information and usage patterns. Identifying these customers proactively allows the company to offer targeted incentives to retain them.

The Dataset

You will use a synthetic dataset (telecom_churn.csv) that mimics real-world data. It contains messy, un-preprocessed data with the following characteristics:

- Mixed Data Types: Numerical (Tenure, MonthlyCharges) and Categorical (Contract, PaymentMethod).
- Missing Values: np.nan values are present in several columns.
- Irrelevant Features: Some columns (CustomerID, Timestamp) are not useful for prediction and must be handled.
- Class Imbalance: The 'Churn' target variable is imbalanced—far fewer customers churn than stay.

Dataset Schema (Partial):

- CustomerID: Unique identifier.
- Gender: Customer's gender.

- SeniorCitizen: Whether the customer is a senior citizen (0 or 1).
- Partner: Whether the customer has a partner.
- Dependents: Whether the customer has dependents.
- Tenure: Number of months the customer has stayed with the company.
- PhoneService: Whether the customer has phone service.
- InternetService: Type of internet service (DSL, Fiber optic, No).
- MonthlyCharges: The amount charged to the customer monthly.
- TotalCharges: The total amount charged to the customer (contains missing values and is an object type).
- Churn: (Target Variable) Whether the customer churned (Yes or No).

---

3. Assignment Tasks

You must complete the following tasks in a single Jupyter Notebook. Provide detailed explanations and justifications for your choices in markdown cells.

Task 1: Exploratory Data Analysis (EDA) and Initial Cleaning

1. Load and Inspect: Load the telecom_churn.csv dataset.
2. Initial Cleaning:
   - The TotalCharges column is loaded as an object type due to some non-numeric entries and has missing values. Investigate and convert it to a numeric type, deciding on a strategy to handle any errors that arise during conversion.
   - Identify all columns with missing values and analyze the extent of the missing data.
3. Data Visualization:

- ○ Create insightful visualizations to understand the relationships between features and the Churn target.
- ○ Generate at least one plot showing the distribution of a numerical feature for churned vs. non-churned customers (e.g., a histogram or KDE plot).
- ○ Generate at least one plot showing the relationship between a categorical feature and churn (e.g., a count plot).
- ○ Summarize your key findings from the EDA in a markdown cell.

Task 2: Feature Engineering

Create at least two new, meaningful features from the existing data. Justify why you believe these features might improve model performance.

- *Example Idea (do not use this one):* Create a TenureInYears feature from the Tenure (in months) feature.
- *Example Idea (do not use this one):* Create a binary feature HasHighMonthlyCharges for customers with charges above the 75th percentile.

Task 3: Building the End-to-End Pipeline

This is the most critical part of the assignment. You must create a single Scikit-learn Pipeline that encapsulates all your preprocessing and modeling steps. This pipeline will be fed raw data and will output predictions.

1. Define Preprocessing Steps:
    - ○ Create a preprocessor using ColumnTransformer to apply different transformations to different columns.
    - ○ Numerical Features: Impute missing values (e.g., using SimpleImputer with a median strategy) and then scale the

data (e.g., using StandardScaler).

   - Categorical Features: Impute missing values (e.g., using SimpleImputer with the most frequent strategy) and then encode the features (e.g., using OneHotEncoder). Make sure to handle unknown categories that might appear in test data.
   - Identify and drop any irrelevant columns from the original dataset.

2. Integrate with Pipeline:
   - Combine your ColumnTransformer preprocessor with a classifier (e.g., LogisticRegression) inside a Pipeline object.
   - Demonstrate that you can successfully train this pipeline on your training data.

Task 4: Model Selection and Hyperparameter Tuning

1. Choose Models: Select three different classification algorithms suitable for this task (e.g., Logistic Regression, Random Forest, Gradient Boosting Classifier). Justify your choices briefly.
2. Hyperparameter Tuning: For your best-performing model type, use GridSearchCV to find the optimal hyperparameters.
   - The GridSearchCV should be performed on your entire pipeline, not just the classifier. This ensures that preprocessing steps are not leaking information from the validation folds.
   - Define a hyperparameter grid to search over. For the classifier, tune at least two hyperparameters (e.g., n_estimators and max_depth for a Random Forest).
   - Use an appropriate scoring metric for imbalanced data, such as f1_weighted or roc_auc. Do not use accuracy.

Task 5: Handling Class Imbalance

The Churn column is imbalanced. Your task is to implement and evaluate a strategy to handle this.

1. Implement a Strategy: Choose one of the following methods and integrate it into your modeling process:
   - Class Weighting: Use the class_weight='balanced' parameter in your chosen classifier.
   - Resampling: Use a library like imbalanced-learn to apply a resampling technique (e.g., SMOTE for oversampling or RandomUnderSampler for undersampling) to the training data. This should be added as a step in your final Pipeline.
2. Compare and Conclude: Compare the model's performance with and without the imbalance-handling strategy. Does it improve the results? Explain why or why not.

Task 6: Final Evaluation and Interpretation

1. Test Set Evaluation: Evaluate your final, tuned pipeline (with the imbalance handling) on the held-out test set.
2. Metrics Report: Report the following metrics: Precision, Recall, F1-Score, and the ROC AUC score. Also, display the Classification Report and Confusion Matrix.
3. Interpretation:
   - Analyze the confusion matrix. In the context of customer churn, what is the business cost of a False Positive versus a False Negative? Which one is worse?
   - If your final model was a tree-based ensemble (like Random Forest or Gradient Boosting), extract and visualize the top 10 most important features. Discuss what these features tell you about the main drivers of customer churn.