Lab 5 - Classification

DS3010 - Introduction to Machine Learning

Instructions

1. Provide commented, indented code. Variables should have meaningful names.
2. Submit one .ipynb file containing all answers. The name should be [student name][roll_number] assignment[number].ipynb
3. Read the questions carefully before answering. If a question asks to follow a particular approach or to use a specific data structure, then it must be followed.
4. Write questions in separate markdown blocks in Jupyter Notebook before the code block containing answers.
5. All plots should have appropriate axis labels, titles, and legend.
6. **Late submissions will not be accepted.**

Tasks for the Lab

Run this command before importing SMOTE
pip install imbalanced-learn

Dataset : link

1. Naive Bayes Classifier (12 points)

1.1 Data Preprocessing (6 points)

A. Load the given Customer Churn dataset (whether the customer will leave the service or not) for the classification task.
B. Check for null values. Drop the unnecessary column. (1)
C. Find out the columns with the categorical values and convert it into numerical data using appropriate technique. (1)
D. Find out the columns with the numerical values and apply the StandardScaler to the numerical columns. (1)
E. Separate features and target column(Churn). (1)
F. Check the class distribution of the target variable (Churn). Write your observation. (2)
G. Split the data into train-tests in a ratio 8:2.

1.2 Over sampling (2 points)

A. Initialize SMOTE.
B. Apply SMOTE to the training data. (1)

C. Check the class distribution after SMOTE. Write your observation. (1)

1.3 Model Building - Naive Bayes (4 points)

A. Initialize Naive Bayes with specified hyperparameter var_smoothing=1e-09.
B. Train the Naive Bayes model on the resampled training data. (1)
C. Predict on the test data. (1)
D. Evaluate the model using Confusion Matrix, Accuracy Score, Classification Report. (1)
E. Write your observation about the performance of your model. Which metric are you considering and why? (1)

2. Logistic Regression (8 points)

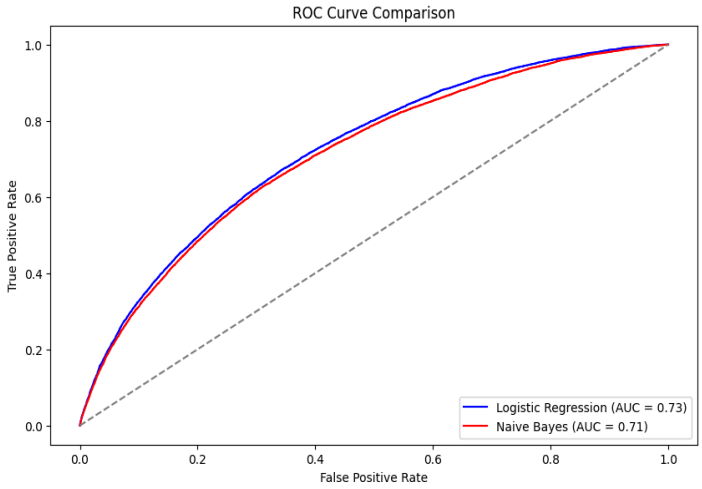
2.1 Model Building - Logistic Regression (4 points)

A. Initialize the Logistic Regression model with hyperparameters C=0.01, max_iter=100, penalty='l2', solver='liblinear'.
B. Train the model on the resampled training data. (1)
C. Predict on the test set. (1)
D. Evaluate the model using Confusion Matrix, Accuracy Score, Classification Report. (1)
E. Write your observation about the performance of your model. Which metric are you considering and why? (1)

2.2 Model Selection (4 points)

A. Predict probabilities for Logistic Regression.
B. Compute ROC curve and roc_auc_score for Logistic Regression b/w y_test and predicted probabilities. (1)
C. Predict probabilities for Naive Bayes.
D. Compute ROC curve and roc_auc_score for Naive Bayes b/w y_test and predicted probabilities. (1)
E. Plot the ROC curves with Labels and Title. (1)
F. Compare and choose the best model in the context of an imbalance dataset. (1)

The plot should look like this.


ROC Curve Comparison
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression (AUC = 0.73)
Naive Bayes (AUC = 0.71)