Exercise – 7: DS203-2024-S2

Submissions are due by: March 31, 2025, 11:55 pm

You are given a time-series Rainfall Prediction dataset spanning 6 years (E7_data.csv). Your goal is to train a robust classifier by performing data preprocessing, handling missing values, removing outliers, addressing multicollinearity, and tackling class imbalance.

Follow the steps below carefully and complete the tasks in each section.

1. Data Analysis and preprocessing

- Load the dataset and print descriptive statistics using `df.describe()`.
- Record observations on missing values in the dataset.
- As you are given time series data for 6 years, plot all the numerical features against the 'ID' column (total 8 plots excluding recorded_day column) to observe the pattern and comment on the pattern for all the features.
- Based on the trend observed from the above plots, handle missing values appropriately and justify your method.
- Generate a 2D t-SNE plot and a 2D PCA plot (top 2 principal components) of the dataset. Mention which has better class separation in low dimension.

2. VIF and PCA

- Fit a Logistic regression model on the data and perform 5-fold cross-validation. Report the average ROC-AUC score for your logistic regression model.

Now let's analyze whether similar performance can be achieved with less number of features using the VIF method

- Standardize features by making mean 0 and scaling to unit variance. Compute VIF for all columns
- Based on VIF values, selectively drop:
    - 3 columns
    - 6 columns
    - 8 columns
- Fit Logistic regression models on three of the reduced datasets above and perform **5-fold cross-validation**. Plot the average ROC-AUC score for your logistic regression models for the above 3 cases.

Now let's analyze whether similar performance can be achieved with fewer features using PCA.

- Standardize features by making mean 0 and scaling to unit variance. Perform PCA on the dataset. Select the top
    - 1 Principal component

    - 3 Principal components
    - 6 Principal components
- Fit Logistic regression models on the above principal component data and perform 5-fold cross-validation. Plot the average ROC-AUC score for your logistic regression models for the above 3 cases.

Tabulate the ROC-AUC scores of both VIF and PCA-based approaches and record your observations.

3. Class Imbalance:

Based on the VIF analysis in the previous section, determine the optimal number of features to retain before proceeding with the next steps and mention the same.

Note: Since the dataset exhibits high multicollinearity, you must drop some features. But choose the number of features to retain based on the above analysis in Part 2.

The given dataset is imbalanced. Choose one of the methods from below which best handle class imbalance based on the ROC-AUC score of the logistic regression.

- **Assigning Weights**: Use the `class_weight='balanced'` parameter in logistic regression.
- **SMOTE (Synthetic Minority Over-sampling Technique)**: Perform oversampling to balance the dataset. (You can use `SMOTE` function from `imblearn` library)

Comment on the ROC-AUC score of the above model and the ROC-AUC score of the model before handling class imbalance.

Your submissions should include the following:

1. A PDF document with all the above analyses and comments. Ensure that you include the required figures and tables (i.e. metrics data) in your report, along with the explanations and analysis.
2. Your Python source file (.py file). Please DO NOT upload Jupyter Notebooks – they get bulky!
3. Name of the PDF should be E7-your-roll-number.pdf and the name of the Python source file should be E7-your-roll-number.py
4. Upload the PDF and the source file to the assignment submission point E7.