CSCN8000 – Artificial Intelligence Algorithms and Mathematics

Lab 3: Logistic Regression

Dataset file: ***Credit Card Fraud Detection on Kaggle***

Data Preprocessing Tasks (9 Points):

1. (7 points) Detect and **handle** outliers for all the numerical features in the dataset using either the Whiskers Approach, or Z-score approach.
2. (2 points) If you feel it's needed, normalize numerical features using appropriate method based on feature characteristics.

Descriptive Analytics Tasks (10 Points):

1. (2 points) Analyze the distribution of fraud vs non-fraud transactions and comment on it.
2. (4 points) Analyze the correlation between fraud/non-fraud transactions and all the other numerical features.
    a. Comment on which features have the highest correlation with the target variable.
3. (4 points) From the previous step, choose the 10 features with highest correlation with the target variable, and plot their distributions against the fraud/non-fraud transactions.

ML Model Training and Testing Tasks (27 Points):

1. (2 point) Split the **cleaned** data into training and testing sets (e.g., 80% training, 20% testing).
2. (3 points) Use 5-fold cross-validation to train and validate the performance of all the models in this section.
3. (3 points) Train a logistic regression model as a baseline model.
4. (2 points) Print the learned coefficients (weights) of the model.
    a. Comment on which feature the model gave higher weight to in the weight vector.
5. (4 Points) Evaluate the model's performance on the ***test set*** using the following metrics:
    a. Accuracy
    b. Precision
    c. Recall
    d. F1-Score
6. (4 Points) Plot the ROC Curve and print the AUC of the model on the test set. Comment on the quality of the model performance as seen in the ROC curve.
7. (6 Points) Apply a well-known technique to handle class imbalance and compare the model performance on the test set with and without this approach. Comment on your outcome.
8. (3 Points) Try the KNN and SVM classification compare their performance to the logistic regression baseline.

Organization Criteria (4 Points)

1. (4 points) Provide an organized notebook at the end with clear section, markdown comments and printed outputs. Make sure no errors are printed in the final submission.

Deliverables

1. Include all your findings and task solutions in one Jupyter notebook (.ipynb) that shows all the printed cell outputs. Prepare a html version (.html) of the notebook file. Both files should be named as follows: [Full Name]_[Student ID]_[Section Number]_Lab3.[html/ipynb].
2. Submit both .html and .ipynb files on eConestoga in the Lab 3 under the Assignments section.