Programming Assignment 3

Goal: You are hired as a data scientist at a bank, say CoinBank. Let's first imagine that. How cool would that be! Interestingly, on your very first day at CoinBank, your are assigned a gruesome task to detect fraudulent e-commerce transactions happening at the bank by analyzing the transaction log files.

Dataset:

Set A

Set-A.X.csv : contains information regarding the 94682 transactions.
Set-A.y.csv : target labels for each of the transaction found in the corresponding Set-A.X.csv dataset, 0 for legitimate and 1 for fraudulent.

Set-B

Set-B.X.csv : contains information regarding the 100,000 transactions, with slightly modified set of variables (custAttr1, and custAttr2).
Set-B.y.csv : target labels for each of the transaction found in the corresponding Set-B.X.csv dataset, 0 for legitimate and 1 for fraudulent.

Some note on the datasets

- Please treat fraudulent transaction as "positive" and legitimate as "negative" sample.
- Due to privacy reasons, the provider of the dataset renamed the data fields to some abstract names. Therefore, you may treat each field equally without giving any preference to any field in the dataset.
- An exploratory data analysis can be found in the provided Fraud-Detection-EDA.ipynb (jupyter notebook)
- **If needed, take fractions of the dataset to accommodate into the memory capacity of your workstation.**

Tasks

- PLEASE SAVE ALL YOUR WORKS (Codes+Results+Texts) IN A JUPYTER NOTEBOOK, and SUBMIT IN CANVAS.
- Randomly split the SET-A dataset into 80% training and 20% test (holdout set).

Task 1 : Logistic Regression based Classifier (LR)

1. Logistic Regression based classifier construction. (Mini-batch gradient descent). Once again, do not call library function for logistic regression fit method.
1.1 Use a 10-fold cross validation strategy to obtain the best set of hyper-parameter values of the logistic regression. Please report a discussion why you would choose the values. (On the training set)
1.2 Train the classifier on the training set, and report the training confusion matrix, accuracy, precision, recall, F1-score, the ROC curve and the corresponding AUC score.
1.3 Using the model, test the test set, and report the testing confusion matrix, accuracy, precision, recall, F1-score, the ROC curve and the corresponding AUC score.
1.4 Looking at results obtained in Tasks 1.2 and 1.3: Is your classifier showing any sign of overfitting, or underfitting? And explain why do you think that. If yes, can you suggest a solution, and utilize it to solve the issue (if present).
1.5 (Graduate Student Requirement) Dataset follows a highly skewed distribution in terms of the target class label. Can you devise a workaround to this? Please explain and redo tasks 1.1, 1.2, 1.3, and 1.4.

Task 2 : Naive Bayes Classifier (NB)

2. Naive Bayes Classifier construction considering ***the log likelihood strategy with Laplace based smoothing for discrete variables, and Gaussian smoothing for the continuous variables to estimate the corresponding conditional probabilities.*** Once again, do not call library function for Naive Bayes classification's fit method.
2.1 We do not have any hyper-parameters to train here. Then, simply report the 5-fold cross validation performance in terms of accuracy, precision, recall, F1-score. (On the training set)
2.2 Train the classifier on the training set, and report the training confusion matrix, accuracy, precision, recall, F1-score, the ROC curve and the corresponding AUC score.
2.3 Using the model, test the test set, and report the testing confusion matrix, accuracy, precision, recall, F1-score, the ROC curve and the corresponding AUC score.
2.4 Looking at results obtained in Tasks 2.1, 2.2 and 2.3: Is your classifier showing any sign of overfitting, or underfitting? And explain why do you think that. If yes, can you suggest a solution, and utilize it to solve the issue (if present).
2.5 (Graduate Student Requirement)) Dataset follows a highly skewed distribution in terms of the target class label. Can you devise a workaround to this? Please explain and redo tasks 2.1, 2.2, 2.3, and 2.4.

Task 3 : kNN Classifier (kNN)

3. k-NN Classifier construction (considering the normalize to unit vector strategy and Euclidean distance measure as discussed in class). Once again, do not call library function for kNN fit method.
3.1 Use a 10-fold cross validation strategy to obtain the best hyper-parameter value of the k-NN classifier (i.e., the value of k). Please report a discussion why you would choose the value. (On the training set)
3.2 Train the classifier on the training set, and report the training confusion matrix, accuracy, precision, recall, F1-score, the ROC point on the ROC space.
3.3 Using the model, test the test set, and report the testing confusion matrix, accuracy, precision, recall, F1-score, the ROC point on the ROC space.
3.4 Looking at results obtained in Tasks 3.2 and 3.3: Is your classifier showing any sign of overfitting, or underfitting? And explain why do you think that. If yes, can you suggest a solution, and utilize it to solve the issue (if present).
3.5 (Graduate Student Requirement) Dataset follows a highly skewed distribution in terms of the target class label. Can you devise a workaround to this? Please explain and redo tasks 3.1, 3.2, 3.3, and 3.4.

Task 4 : Comparing performances of LR, NB and kNN classifiers

4.1 The CoinBank manager asks you to choose one classifier for her which is less likely to miss any fraudulent transaction. With the help of a cost matrix, can you suggest her the best classifier from the three you developed in Task 1, 2 and 3? Which one would you pick, and why?

4.2. CoinBank manager is also looking for an alternate classifier which will not make the legitimate customers angry due to the fact that the classifier you chose in Task 4.1 is going to raise a lot of false alarms. With the help of a cost matrix, can you suggest the best classifier from the three you developed? Which one would you pick, and why?

4.3 Which of the three classifiers do you think would be a balanced choice for CoinBank? And why?

Graduate Students' Additional Tasks

- Repeat Tasks 1, 2, 3 and 4 with Set B.