Lab 2: Classification and Prediction

**Submission Instructions:**

1. **You will submit a single script (in either Matlab, Python or R).**
2. **Your script will clearly mark out answer and code for each question.**
3. **Your script will print out results that support your direct answer.**

Example: lab2.R

# Answer to question 0:

# The number of attributes in the file is 100. There are 2 classes.

# Below are codes to print out the numbers in answer above.

XXXXXXXXX

**Part 1:**

**Data Description**:

The data we will use is from “Porto Seguro’s Safe Driver Prediction” challenge in Kaggle (same as lab 1). The dataset was used to predict the probability that an auto insurance policy holder files a claim (the “target” column).

We have already preprocessed the train.csv and the test.csv data by removing any attributes whose missing values are over 2%. Next we replaced the missing values of remaining attributes with majority for qualitative attributes and median for quantitative attributes. In the train.csv (and test.csv) data, features that belong to similar groupings are tagged as such in the feature names (e.g., ind, reg, car, calc). In addition, feature names include the postfix **bin** to indicate binary features and **cat** to indicate categorical features. Features without these designations are either continuous or ordinal. The **target** columns signifies whether or not a claim was filed for that policy holder.

In Lab 2, you will evaluated the classification performance of decision tree and random forest. You will use train.csv for training.

1. Train a Decision Tree model using the train.csv. Please answer the following questions:
   1. Provide confusion matrix for predictions on both train.csv and test.csv.
   2. Calculate precision, recall and f-score on claim class based on confusion matrix for train data.
   3. Calculate precision, recall and f-score on claim class based on confusion matrix for test data.
   4. Compare the results of ii and iii. How do precision & recall change between training data and test data? Why do the changes happen?
   5. Calculate True Positive Rate (TPR) and False Positive Rate (FPR) based on confusion matrix for train data.
   6. Calculate True Positive Rate (TPR) and False Positive Rate (FPR) based on confusion matrix for test data.
   7. Compare the results of v and vi. How do TRP & FPR change between training data and test data? Why do the changes happen?
   8. You would have observed different behavior of measures (precision&recall v.s. TPR&FPR) based on your answers in iv and vii. Which measures among precision&recall and TPR&FPR would you suggest to evaluate a classification model? (**Hint**: the number of samples in each class may vary dramatically in our population).
2. Train a Random Forest model using the train data. Please answer the following questions:
   1. Provide confusion matrix for predictions on both train.csv and test.csv.
   2. Calculate precision, recall and f-score on claim class based on confusion matrix for train data.
   3. Calculate precision, recall and f-score on claim class based on confusion matrix for test data.
   4. Compare the results of ii and iii. How do precision & recall change between training data and test data? Why do the changes happen?
   5. Calculate True Positive Rate (TPR) and False Positive Rate (FPR) based on confusion matrix for train data.
   6. Calculate True Positive Rate (TPR) and False Positive Rate (FPR) based on confusion matrix for test data.
   7. Compare the results of v and vi. How do TRP & FPR change between training data and test data? Why do the changes happen?

**Part 2:**

**Dataset Description:**

Each row contains a pixel’s spectral information as well as its land cover class.

|  |  |
| --- | --- |
| Train Data | Part2Train.csv |
| Test Data | Part2Test.csv |

|  |  |
| --- | --- |
| Column | Type |
| 1, 2,3,4 | Feature (continuous value for spectral response on red, green, blue, near infrared respectively) |
| 5 | Land cover classes (0 for dry land, 1 for wetland) |

Note: Data contains no header information.

1. Using the Part2Train.csv data train the following models:
   1. Decision Tree
   2. SVM
   3. Logistic Regression
   4. Artificial Neural Network

For each model, please answer to the following questions:

1. Provide confusion matrix for test data.
2. Calculate precision, recall and f-score on claim class for test data.
3. How would different models’ performance compare with each other?