MS4610 – Introduction to Data Analytics (July-Nov 2020)

Project - Loan Default Prediction

The dataset consists of the following details on loans taken by different customers:

- ID: A unique identifier for every financial loan that is being considered.
- Loan type: Type of loan taken (Two types, 'A' or 'B').
- Occupation type: Occupation of the customer (Three occupation types, 'X', 'Y', 'Z').
- Income: A continuous variable that is indicative of the annual income of the customer. This is not the exact income value.
- Expense: A continuous variable that is indicative of the annual expense of the customer. This is not the exact expense value.
- Age: Age of customer – Value of '0' is considered as below 50, and value of '1' is considered as above 50.
- Score1, Score2, Score3, Score4, Score5: Represents five different metrics calculated by the organization, about the customer and the loan that is being considered.
- Label: '0' means non-default, and '1' means default on that loan.

Using the above information and the data, build a model to predict whether a loan will go default or not, and to understand which of the features are important and helpful in the prediction.

The dataset has been split into training and test sets. The ZIP file contains the CSV files for train_x, train_y, and test_x. The train_y file consists of the "Label" that is mentioned above, and train_x consists of the other features.

Make predictions on the test_x data that is given. The predicted labels of the test set from your model should be provided as a CSV file, in the same format as train_y. Name this file as 'pred_y.csv'.

Use any method for imputing missing values.

Note: The test set is **not** expected to have been a random sample of the entire dataset.

Submissions required:

1. Report on the model(s) used.
2. Code that was used to get the predictions.
3. Predicted labels of the test set, as mentioned above (pred_y.csv).

We will be using accuracy to measure a group's performance on test_y, in addition to testing the code and report. Plagiarism testing software will be applied on the code and report.