Homework1

RULES

Please upload to **Moodle** a **.zip** archive containing your Jupyter Notebook with solutions and all data required to reproduce your solutions.

Please also prepare a requirements.txt file which lists all the packages that you have used for your homework, one package per line. This will allow us to install all required packages in one go, by using "pip install -r requirements.txt".

Please name your .zip archive using your full name and student id as follows - \***Jessica\_Butler\_18206383\_COMP47350\_Homework1.zip**\*.

For your Notebook, please split the code and explanations into many little cells so it is easy to see and read the results of each step of your solution. Please remember to **name your variables and methods with self-explanatory names** NOT X!. Please remember to write comments and where needed, justifications, for the decisions you make and code you write. Feel free to revisit \***tips\_to\_keep\_your\_ipython\_notebook\_readable\_and\_easy\_to\_debug.html**\* provided on Moodle.

Your code and analysis is like a story that awaits to be read, make it a nice story please.

The accepted file formats for the homework are:

- .ipynb

- .zip

- .pdf

- .csv

Please keep the whole code in a single notebook. Usage of external tools/files is discouraged for portability reasons. Files in any other format but mentioned above can be used but will be ignored and not considered for the submission (including .doc, .rar, .7z, .pages, .xlsx, .tex etc.).

Any image format is allowed to be used as far as the images appear embedded in your report (.ipynb or .pdf or .html).

\*\*Deadline: Sunday, 10 March, 2019, midnight.\*\*

credit scoring company

concerned with reducing credit repayment risk. The company wants to

use the data collected about their customers to build a data analytics solution for credit risk prediction.

use the information about the applicant in their credit report to predict whether they will repay their credit within 2 years.

The target variable to predict is a binary variable called RiskPerformance.

* The value “Bad” indicates that a consumer was 90 days past due or worse at least once over a period of 24 months from when the credit account was opened.
* The value “Good” indicates that they have made their payments without ever being more than 90 days overdue.

The dataset we work with is a sample of the data used in this data challenge: https://community.fico.com/s/explainable-machine-learning-challenge?tabset-3158a=2

Each student will work with a different subset of the data. The CSV file is named using the format: \*\*CreditRisk-[your-student-number].csv\*\*, e.g., \*\*CreditRisk-12345678.csv\*\* is the data file for a student with number 12345678. You need to work with the CSV file corresponding to your student number. There are 4 parts for this homework. Each part has an indicative maximum percentage given in brackets, e.g., part (1) has a maximum of 40% shown as [40].

(1). [40] data quality report:

- Check how many rows and columns your CSV has.

- Print the first and the last 5 rows.

- Convert the features to their appropriate data types (e.g., decide which features are more appropriate as continuous and which ones as categorical types).

- Drop duplicate rows and columns, if any.

- Drop constant columns, if any.

- Save your updated/cleaned data frame to a new csv file.

For the updated CSV and data frame (after column/row removal):

- Prepare a table with descriptive statistics for all the continuous features.

- Prepare a table with descriptive statistics for all the categorical features.

- Plot histograms for all the continuous features.

- Plot box plots for all the continuous features.

- Plot bar plots for all the categorical features.

- Discuss your initial findings.

- Save the initial discussion of your findings into a single data quality report PDF file. The PDF report should focus on the key issues identified in the data and discuss potential strategies to handle them. Simple **listing of tables and plots** without discussion and justification **will not receive full mark**s.

(2). [30] data quality plan for the cleaned CSV file.

- Mark down all the features where there are potential problems or data quality issues.

- Propose solutions to deal with the problems identified.

Explain why did you choose one solution over potentially many other.

- Apply your solutions to obtain a new CSV file where the identified data quality issues were addressed.

- Save the new CSV file with a self explanatory name.

- Save the data quality plan to a single PDF file.

(3). [15] Exploring relationships between feature pairs:

- Choose a subset of features you find promising and plot pairwise feature interactions (e.g.,

continuous-continuous feature plot or continuous-categorical plots or correlation plots).

Explain your choices.

- Discuss your findings from the plots above. Do you find any features or feature combinations that are indicative of the target outcome? Explain in plain words (a short paragraph) the story of your findings so far.

(4). [15] Transform, extend or combine the existing features to create a few new features (at least 3) with the aim to better capture the problem domain and the target outcome. Justify the steps and choices you are making. Add these features to your clean dataset and save it as a CSV file with a self explanatory name \*\*CreditRisk-[your-student-number].csv\*\*