CSCI4155 – Fall 2022 – Assignment 1

Topics Covered: Data Pre-processing, Numpy and Pandas.

Assessment Objective

The purpose of this assignment is to assess your understanding of the concepts taught in the Data pre-processing learning modules and lectures.

We can assess these modules easily by:

- Asking you to work on a given case study and do step-step process of data pre-processing on a given data set

Introduction to the given problem

The main aim is to construct a Credit Risk Model that measures probability of "default" for every personal account. Here in this assignment, you are not supposed to use any predictive models. Here the main objective is to pre-process the data so that further predictions can be made by applying certain data model in the later stage. So, you're working as a data analyst in the data science team of a central bank in Europe. Being in the data analyst role means your goal to obtain a clean and preprocessed loan data set, we can hand over to the next people in the analytical chain. You have to prepare one document and write down all the changes you're making to the original data set where you describe what each column of the new data set represents.
To help you out, they've provided details on what data is stored in every column, as well as a set of rules on how to clean and pre-process the values in each one.

Following **key points** are given as per the scenario:

1. You've been given is a sample from a larger data set that belongs to an affiliate bank based in the United States. Therefore, all the values are in dollars, so you need to provide their euro equivalents.
2. The loan information is stored in the CSV file called loan data, which you can download from the resources.
3. Then every categorical variable must be quantified. So, we need to change any text columns into numbers based on the information they contain. For some, like the issue date on each loan, the transformation is extremely straightforward, since we can split the accounts by months. For example, if all the data is related to a single year only, there is no need to deal with whole date. Rather we can just retain the month information.
4. Similarly for other columns, we only care if they provide positive or negative connotations. So, we'll be turning them into dummy variables that hold either zero or one. Furthermore, when we're measuring credit worthiness, we need to be extremely risk averse and distrustful of any unavailable data. That's why the consensus in the field is that missing information suggests foul play because loan applications are self reported to elaborate since

candidates fill out their loan applications manually. There is an incentive to withhold information which can lower their chances of getting a loan. Of course, we prefer to give out loans to applicants who can repay them so that the information isn't available will just assume the worst. However, what is worst varies from one column to the next.

5. The management team has provided us the casting directions for the data set. They are not in the favour of deleting any missing data. Instead, they prefer to fill the missing data by using minimum, maximum or some other value depending upon the context. Obviously, you can discard the columns which do not play any significant role or you can also delete those columns which are highly co-related or can replace each other.

Submission Checklist:

An assignment 1 folder is already created on Brightspace (in Assessment section)

While submitting the assignment, the name of the submission items must be "Firstname_Lastname_Assignment1"

A. Create a document (in the pdf format) that includes the following:
   I. Your name, CSID and Banner #
   II. Mention the link of the jupyter notebook
   III. If necessary, credit for any project resources you have used as starting point for your assignment
   IV. In this document, write down the description of what you have implemented and how you have implemented (step by step) with proper reasoning.
   (for detailed description and format, please refer the file named "Assignment Format.pdf"
B. Upload the jupyter notebook with the same name.

Academic Integrity:

All the submission documents will be crosschecked via Moss and Turnitin. Do not share your code and report with anyone. In case of academic integrity violation, a strict action will be taken.

Suggested steps:

1. Import the required packages/libraries
2. Check if there is any problem of missing data. Do some statistical checks to analyze the given features
3. Split the dataset into numeric and non-numeric sub-sets. Make sure that both newly created data sets have proper headers for understanding.
4. Start with non-numeric data pre-processing
   Some of the tasks you can do is:
   a. There are different types of loan-status: You can categorize these status into good or bad and can assign some 0,1 values
   b. Analyze the "term" column: Analyze and assign some values .
   c. Analyze the relationship between grade and sub-grade column. Are these columns

related? If yes, what should be the next step in terms of pre-processing? Can we assign some values to these columns instead of textual information?

d. Work on "verification status" column.
e. Analyze the "url" column and think about its significance and next step.
f. Analyze the "state address" column. Check how many number of entries are there for each state. Is there any problem of missing data? The other thing we can implement here is to assign values to these states based on their location into four regions as per this document https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf and can replace the missing entries by 0.

5. Analyze and work on numeric data columns:
   Some of the things which you can do:
   a. Fill the missing values of Funded amount, Loaned Amount, Interest Rate, Total Payment, Installment, interest rate with some fillers like min, max, average depending upon the context.
   b. Convert the currency USD to Euro wherever applicable in the data set
   c. Sort and store the new refined data set.

Note: Please follow the above given steps as hints only. Using your judgement instead of blindly following instructions is an important trait. Creativity with justification is encouraged. Bonus marks will be provided in case of extra efforts or steps implemented apart from the above given steps.

Files provided along with Assignment:

1. EUR-USD: In this file conversion rates are given for every month. There are 5 columns: Open, High, Low, Close and Volume. High and low represent the highest and lowest value of conversion. Open and Close represent the value of conversion when the market is opened and closed.
2. Loan-data-dictionary: In this file, description of every column in the data set is given.
3. Loan-data: This includes the complete dataset.