Data Engineering W24 - Milestone 1 Description

1 Introduction

The goal of this milestone is to load the provided CSV files, perform exploratory data analysis (EDA) with visualization, extract additional data, perform feature engineering, and pre-process the data for downstream cases such as ML and data analysis.

You will be working with a fintech dataset. It contains records about customers, loans, and more and their corresponding loans.

Note: Each student has their assigned dataset so you can find the dataset assigned to you via this link and download your assigned dataset from here.

You can also find weight distribution and marking rubric for the milestone on the same CMS, which will help you understand what is precisely required from you.

IMPORTANT NOTES

- Each task should be written as a function. Organizing the code as functions will help you in milestone 2, so it's highly recommended to make all actions function-based.
- Your solutions and functions should be compatible with any data with the same schema, i.e. if we supply you with more data from the same dataset, you need to make sure that your functions are compatible with it.
- **VERY IMPORTANT:** Before submission, you need to make sure that you run all the notebook and submit it with the output. As we won't have time to run each and every notebook on its own.

2 Requirements

Note that, you don't need to follow the same flow or order of the tasks mentioned, you need to do them in a logical order.

2.1 EDA

1. Explore the dataset and ask at least 5 questions to give you a better understanding of the data provided to you.
2. Visualize the answer to these 5 questions. Make sure to mention what the visualization shows and what observations you have found in these insights.

2.2 Data Cleaning

1. Tidy up the column names, make sure there are no spaces
2. Choose a suitable column index

3. Observe, comment on, and handle inconsistent data. (i.e. duplicates, irrelevant or incorrect data, different spelling with the same meaning eg. INDIVIDUAL & Individual, etc)
4. Observe missing data and comment on why you believe it is missing **(MCAR, MAR, or MNAR)**.
5. Handle missing data (imputation)
   -*Hint*: When trying to impute with mean values, try to find patterns from other features, maybe the missing values are in certain categories from other features. eg. In the titanic dataset, missing values in age may be in certain pclasses in this case, we can take the average for the age in each class and impute with it.
6. Observe and comment on outliers
7. Handle outliers

IMPORTANT NOTE: With every change you make to the data you need to comment on why you used this technique and how has it affected the data (by both showing the change in the data i.e change in a number of rows/columns, change in distribution, etc. and commenting on it).

2.3 Data Transformation and Feature Engineering

1. Add the following 4 columns
   - **Month number**: *integer* - create a month number column (1-12) to be able to discretize the data by month number.
     - *Tip*: Change the datatype of the date feature to `datetime` type instead of an object.
   - **Salary Can Cover:** *boolean* - this column should contain a boolean value (true = 1 & false = 0) that shows if the annual income can cover the loan amount or not.
   - **Letter Grade:** *categorical* - encode the grade column using the categories in the dataset description (A-G).
   - **Installment per month**: *float* - use the below formula to calculate the monthly installments:

$$M = P \times \frac{r \times (1+r)^n}{(1+r)^n - 1}$$

     where:
     - M is the monthly installment
     - P is the loan principle/amount
     - r is the monthly interest rate , $r = int\_rate/12$
     - n is the number of payments/months
2. Encode any categorical feature(s) and comment on why you used this technique and how the data has changed (Make sure to include this in the lookup table, more on this later in section 2.4).
3. If exists, Identify feature(s) that need normalization and show your reasoning. Then choose a technique to normalize the feature(s) and comment on why you chose this technique.

2.4 Lookup Table(s)

You will need to create a lookup table (CSV) which will contain info about the original values for any feature/values you have imputed. This lookup table must be created programmatically and not hard-coded (not done by hand) as much as possible.

For any imputation with arbitrary values or encoding done, you have to store what the value imputed or encoded represents in a new csv file. I.e if you impute a missing value with -1 or 100 you must have a

Column	Original	Imputed
grade	1	A
grade	4	A
int_rate	missing	0.2 (mean)
loan_status	Current	1
loan_status	Fully Paid	2
letter_grade	A	1
letter_grade	B	2

Table 1: Example of a Lookup table

csv file illustrating what -1 and 100 means. Or for instance, if you encode cities with 1,2,3,4,etc what each number represents must be shown in the new csv file. See table 1 for an example.

Note: You can use the values in the lookup table later to reverse all of the imputed values to their original values.

2.5 Bonus Task

Using the state column, create another column named *state_name*, this column should contain the full state name, eg. CA full name is California, and TX is Texas. To be able to create this new column you need to extract data from the internet via an API or Web Scrapping.

- **API**: You can use https://freetestapi.com/apis/us-states, but take care that the output generated by the search API is multiple states.
- **Web Scrapping**: You can use selenium library (or any other library) to web scrape this website https://www23.statcan.gc.ca/imdb/p3VD.pl?Function=getVD&TVD=53971 and get the state names from state code.

2.6 Saving the Output Dataset

After finishing all the above steps you need to save the output of the cleaned dataframe as well as the lookup table.

1. Save the cleaned dataset into a new csv/parquet file named:

   ```
   fintech_data_{MAJOR}_{GROUP}_{ID}_clean.csv/parquet
   ```

   Replace the group with your group number, and the id with your id (please don't use - (dash/hyphen), instead use _ (underscore)). Example:

   ```
   fintech_data_MET_P1_52_1111_clean.parquet
   ```

3 Submission Guidelines

Create a Google Drive folder named " *DE W24 FirstName LastName ID GROUP MAJOR* " and include your milestones in it.

- **Important:** The Project is Individual
- The only accepted format is *ipynb*.
- Name of the notebook MUST be M1_MAJOR_GroupNo_ID.ipynb, replace the MAJOR, GroupNo, and id. i.e. *M1_MET_P1_52_1111.ipynb*
- Upload your notebook, cleaned csv/parquet file, and lookup table csv/parquet to your Google Drive folder in another folder called M1.

Upload your google drive folder here https://forms.gle/vWU4cVScxH394hQv6

Important notes regarding submission(read carefully):

- Each and every task should be written as a function. Organizing the code as functions is extremely important. Your functions should be as dynamic as possible and able to handle different datasets with the same schema and format.
- DO NOT under any circumstance edit your drive folder after the deadline. Here it is again, DO NOT under any circumstance edit your drive folder after the deadline
- Notebook and csv file names must be as instructed, not doing so will result in marks deduction
- **VERY IMPORTANT**: The notebook will not be run, your notebook MUST have the output of any function or task already shown.

Good luck with your first milestone ;)
Feel free to reach out if you hav any questions or need help.