Assignment 3. Data Preparation

Evaluation - 10% of total course mark

Instructions

In this assignment you will use datasets collected and analyzed in Assignment 2. The goal is to prepare them for predictive modeling. Data preparation is one of the most important and often time-consuming aspects of data mining. In fact, it is estimated that data preparation usually takes 50-70% of a project's time and effort. Devoting adequate energy to the earlier business understanding and data understanding phases can minimize this overhead, but you still need to expend a good amount of effort preparing and packaging the data for mining. Perform the following tasks:

- Merge data sets and/or records if you have several of them. Document entity relationship.
- Select a sample subset of data if you think it is necessary. Explain why it is necessary.
- Remove or replacing blank or missing values. Explain methods selected.
- Perform feature engineering.

Use Jupyter notebook for this assignment and include code, text, your comments, observations, and visualizations. Please make sure that the submitted notebooks have been run and the cell outputs are visible. Once created, submit it in pdf format.

Scoring Rubric

Category	Criteria	Maximum (Points)
Merging Datasets	Provide entity relationship diagram for your datasets and explanation on merging.	**1**
Dealing with Missing Values	Explain methods used to deal with missing data.	**3**
Data Preparation	Derive at least five new attributes to your dataset. Explain logic used for every single feature.	**4**
Code	Python code is exceptionally well organized and very easy to follow.	**2**
Total		**10**