Data Science Lab Assignment

General Data Science Process

1. Introduction
2. Data Description
3. Assignment Requirements

How can we predict the household poverty level in Costa Rica?

1. Introduction and Problem Definition

Many social programmes have a hard time making sure the right people are given enough aid. It is especially tricky when a programme is focusing on the poorest segment of the population. The world's poorest typically cannot provide the necessary income and expense records to prove the necessary income and expense records to prove that they qualify for such programmes.

Currently in Latin America, one popular method to determine who is eligible for such programmes is Proxy Means Test (PMT). Agencies use a model that considers a family's observable household attributes, such as the material of their walls and ceilings, or the assets found in the house to classify them and to predict their level of need.

Even though this is a big improvement, the accuracy of it still remains a big problem as the region's population grows and poverty declines. Therefore, to improve on PMT, we have to create new methods beyond traditional econometrics, based on the dataset of Costa Rican household characteristics. We need to **create a data science model that can predict the household poverty level using the household dataset.**

Goal: Predict the household poverty level using descriptive features provided.

Metric: Since the task is a classification problem, accuracy will be the main metric to be used for evaluation.

2. Data Description

There are 3 files given:

1. train.csv.zip - the zip file containing the training set
2. test.csv.zip - the zip file containing the test set
3. codebook.xlsx - full descriptions of each feature

Important note: One row (record) in our data represents one person. *Multiple people can be part of a single household*. You only need the predictions for the HEADS of households.

Data fields (features)

This data contains 142 total columns. This guide will not be explaining every single one of them, but it is up to you to find out the meanings of each feature by looking at the codebook excel file.

Here are the descriptions of some of the features being used in this problem:

- Id: unique identifier for each row
- Target: the target is an ordinal variable indicating groups of income levels, total of 4 categories
  - 1 = extreme poverty
  - 2 = moderate poverty
  - 3 = vulnerable households
  - 4 = non-vulnerable households
- idhogar: a unique identifier for each household. This can be used to create household-wide features. All rows in a given household will have a matching value for this identifier
- parentesco1: indicates if this person is the head of the household

3. Assignment Requirements

You have to create a data science model to predict the poverty level of each household. As you may already know, data science process mainly consists of these parts:

1. Understanding and defining the problem
2. Data exploration
3. Data preprocessing
4. Machine learning implementation
5. Performance evaluation
6. Conclusion and discussion

Therefore, in your submission, you have to perform each of these parts, and elaborate and explain your methods according to the problems stated below.

Part 1: Understanding and defining the problem

The goal of this task is to predict the poverty level for each household.

Q1: Define the machine learning problem of this problem (supervised, unsupervised, binary classification, etc.)

Q2: Explain the meaning of the target and the categories of the target.

In the dataset, we are given data on the individual level with each individual having unique features, but also information about their household. We have to make a prediction for every HOUSEHOLD ONLY, not the individuals.

Q3: Determine what methods can be done to get the prediction for each household, and not for each individual.

Q4: Some individuals belong to the label "no head of household". Determine what should be done with these individuals.

Q5: Identify the features that seem important just by reading their definitions.

Part 2: Data exploration

After reading the train and test data, we have to explore the dataset.

Q1: Identify the shapes of the train and test dataset.

Q2: Identify, print, and explain the distribution of the target in both train and test datasets.

Q3: Using the important features you selected in the previous section, determine the various distributions and statistics of those selected features. Determine if the results strengthen your selection.

Q4: Create a correlation matrix and determine which features are most correlated with the target. Explain and discuss the findings.

Q5: Determine if there are there any possible outliers? (You might have to take a look at all features)

Part 3: Data preprocessing

This step has to be in continuation with the data exploration step.

Q1: Determine how the outliers should be handled.

Q2: Explore if there are any missing values in the data.

Q3: Determine how these missing values should be handled.

Q4: Determine if some features need some labelling. Are there any other conversion of representation needed?

Q5: Determine if some feature engineering can be done and perform them as determined.

Part 4: Machine learning implementation

Once the data has been satisfactorily preprocessed, implement it onto machine learning models.

Q1: Scale the preprocessed data if needed.

Q2: Split the data into train and validation sets

Q3: Determine the machine learning models to be used and perform them. (Apply K-Fold cross validation or any other methods to improve the results.)

Part 5: Performance evaluation

Q1: Determine the results.

Q2: Explain the implementations of the results.

Q3: Determine if the results are satisfactory, and determine if improvements need to be made.

Q4: If improvements need to be made, state the sections that improvements will be made and explain why.

Q5: Determine and explain the limitations of your implementation.

Part 6: Conclusion

Q1: Conclude the data science model that you have created.

Q2: Give thorough discussions.

In a Jupyter Notebook, perform the tasks above with the data provided and explain your process. **Submit your notebook file that contains the codes, the results, and the explanations.** This is not a competition, so the accuracy level will not be compared, but try your best to get at least 60% accuracy.