Detection of Spam Emails

1. Which topic did you choose to apply the data science methodology to?
I had chosen the topic of emails to apply the data science methodology.

2. Next, you will play the role of the client and the data scientist. Using the topic that you selected, complete the Business Understanding stage by coming up with a problem that you would like to solve and phrasing it in the form of a question that you will use data to answer. You are required to:

1. **Describe the problem, related to the topic you selected.**
2. **Phrase the problem as a question to be answered using data.**

For example, using the food recipes use case discussed in the labs, the question that we defined was, "Can we automatically determine the cuisine of a given dish based on its ingredients?".

The problem is to detect the spam and fraud emails and distinguish them to the spam folder of the user. As we are known that there are many fraudulent websites all over the world, which keeps on sending the spam emails to the users. The phishers gather email addresses, through web scraping, data breaches, etc. and craft them to look similar to the official emails. These emails create threats to the people.
Question: **Can we automatically determine which emails are spam and which are not?**

3. Briefly explain how you would complete each of the following stages for the problem that you described in the Business Understanding stage, so that you are ultimately able to answer the question that you came up with.

1. **Analytic Approach**
2. **Data Requirements**
3. **Data Collection**
4. **Data Understanding and Preparation**
5. **Modeling and Evaluation**

1. Analytic approach:
My topic needs a yes/no answer, like if the email is spam or not? So, to use a classification model for this problem will be an appropriate decision. I need to determine the evaluation metrics such as accuracy, F1 score, Recall, ROC curve, etc.

2. Data Requirements:
In this stage, I need to understand what kind of data is required to fulfil my task. I need to revise the requirement based on the availability, quality and content.
For my topic, I need the data of all the emails of a particular person. I have divided into following types:
a. Data of the email: This data includes the subject line, body text, etc.
b. Labeling: Each of the mail should be labeled as spam or ham.

c. Quantity and Quality: The data should be balanced, and volume needs to be large. The quality of data should be clean, consistent, well-labeled data.

3. Data Collection:
This stage requires knowing the source to find the needed data elements, data acquisition strategies and deferred decisions on unavailable data.
Here, my task would be to collect the data of the emails.
I can gather the data from various repositories like UC Irvine ML repository, Aws datasets, Kaggle datasets.
I can also gather diverse public datasets labeled as ham or spam from Enron dataset or SpamAssassin dataset.

4. Data Understanding and Preparation:
The data understanding is the stage were analysis and exploration of collected data is done to ensure that the data is representative of the problem to be solved.
I need to analyze the dataset of emails thoroughly. The following analysis need to be performed for understanding the data.
Text Content Analysis:
Frequency of words: I need to analyze the frequency of words and phrases in spam and non-spam emails. Basically, my attempt would be to find common patterns in the emails.
N-grams: I need to look at sequences of words (such as bigrams, trigrams) to capture contextual information.
Visualization: I would use word clouds or bar charts to visualize frequently occurring words or phrases in both spam and non-spam emails.
Example Exploration
Sample Emails: I would examine a subset of emails from each class to understand their content and structure. And thereby look for common characteristics of spam versus non-spam emails.
Metadata: If available, I would analyze metadata such as sender addresses, subject lines, and timestamps. These might provide additional clues for classification.
The data preparation stage revolves around the concept of cleaning, transforming and organizing the data to facilitate effective analysis and modelling, including feature engineering.
Label Encoding: I will convert categorical labels into numerical format if necessary. For instance, mapping "spam" to 1 and "ham" to 0.
Text Preprocessing: Perform text cleaning and preprocessing, such as:

- Tokenization: Split text into individual words or tokens.
- Normalization: Lowercase text, remove punctuation, and handle special characters.
- Stop Words Removal: Remove common words that may not be useful for classification.
- Stemming/Lemmatization: Reduce words to their base or root form.

Feature Engineering: Here I would need to create additional features if needed. For example, extract features from email headers, subject lines, or the presence of certain keywords.

5. Modeling and Evaluation:
In modeling stage, I would prefer to use predictive analysis, as it will provide me with yes/no answer for the question whether the email is spam or ham? Can use ROC curve for finding the false positive rate and true positive rate.
Following are the steps in brief:

Selection of Models:

- Baseline Models: I would start with models like Naive Bayes or Logistic Regression for initial performance benchmarks.
- Advanced Models: I would jump to experiment with more complex models such as Support Vector Machines (SVM), Random Forests, Gradient Boosting, or deep learning models like Recurrent Neural Networks (RNNs) or Transformers for potentially better performance.
- Ensemble Methods: Make sure to combine predictions from multiple models to improve accuracy and robustness.

Model Training and Evaluation:

- Split Data: Need to divide the dataset into training, validation, and test sets to evaluate model performance.
- Cross-Validation: Make use of techniques like k-fold cross-validation to ensure the model generalizes well to unseen data.
- Hyperparameter Tuning: Optimize model parameters to improve performance using techniques like grid search or random search.
- Accuracy: I would measure the proportion of correctly classified emails.
- Precision and Recall: These will be used to assess the model's ability to correctly identify spam and non-spam emails, especially in imbalanced datasets.
- F1 Score: It will compute the harmonic mean of precision and recall to balance both metrics.
- ROC Curve and AUC: Analyze the trade-offs between true positive and false positive rates.