repon link: https://github.com/RaivoPoisu/SpamMail5.git

Task 2. Business understanding (0.5 point)

Developing a business understanding within CRISP-DM consists of four tasks: identifying your business goals, assessing your situation, defining your data-analysis, data-mining or machine learning goals and producing your project plan. For this exercise, please develop a business understanding of your project. According to CRISP-DM, you should report the following:

Identifying your business goals

- Background

When it comes to sharing information, emails are still considered one of the most important communication channels, whether it be for family-related occasions or professional matters, a person with responsibilities has to check their emails multiple times a day for new information. In today's world unfortunately, those important email inboxes seem to be filled with spam emails, a lot of the time for advertising purposes, and some times for malicious scams. In order to both increase the efficiency of checking mail and save users from scam, we have conducted research to predict whether an email is spam or real based on the words used in mail.

- Business goals
    - spam-check users inbox and flag spam to speed up combing trough mail
    - spam check user inbox and delete/flag scam mail to protect the business and user data from harmful links and attacks

- Business success criteria

The results will be measured by if we can decrease our users exposure to spam by a noticeable margin, while not receiving reports of important mail being incorrectly flagged as spam/scam

Assessing your situation

- Inventory of resources

    - Researchers (us)
    - kaggle datasets, which are later linked
    - Delta-issued laptops
    - python and jupyter
    - panda, numpy, sklearn libraries

- Requirements, assumptions, and constraints
    - Need datasets of emails with labels that say if its spam or real and have them contain enough data (later specified)
    - Need the models to not incorrectly label real mail as spam
    - Should be completed by 9th December 2024

- Risks and contingencies
    - faulty datasets - search the web for new datasets, which work for us
    - model's faulty predictions - optimize the predicting algorithm, adjust sampling sizes

- Terminology
    - dataset - structured collection of data organized and stored together for analysis or processing
    - kagglehub - forum which contains a collection of datasets and models
    - spam - irrelevant message sent over the internet, typically to a large number of users, for the purposes of advertising, phishing, spreading malware
    - machine learning model - program that can find patterns or make decisions from a previously unseen dataset

- Costs and benefits
    - Costs:
        - Datasets are free
        - Wages for the researchers building and maintaining the model
        - Electricity bills
    - Benefits:
        - More efficient workforce (increase productivity)
        - Less successful malicious attacks (they can cost millions)

Defining your data-mining goals

- Data-mining goals
    - Most efficient model from testing
    - dataset with a processed outcome column
- Data-mining success criteria
    - Model accuracy of at least 90%
    - Model precision of at least 95%

Task 3. Data understanding (1 points)

- Data-mining goals
- Data-mining success criteria

Data understanding within CRISP-DM consists of performing four tasks: gathering, describing, exploring, and verifying data quality. For this exercise, please develop a data understanding of your project. Report the results of the tasks according to the following structure:

Gathering data

- Outline data requirements
  - at Least 100000 rows
  - data has about half rows data that are spam and half that isn't
  - data has label column which shows whether row is spam or not
  - data has text column which sais what messages was sent
- Verify data availability
  - Data is available on Kaggle, links:
    - Dataset 1: https://www.kaggle.com/datasets/abdallahwagih/spam-emails
    - Dataset 2: https://www.kaggle.com/datasets/purusinghvi/email-spam-classification-dataset
    - Dataset 3: https://www.kaggle.com/datasets/meruvulikith/190k-spam-ham-email-dataset-for-classification
- Define selection criteria
  - The dataset has a label and text column. Text is is in and utf-8 encoding
  - dataset has similar number of spam and not spam rows if it has substantial number of rows.

Describing data

- Datasets:
  - 1. has 5157 rows. 87% not spam, 13% spam
  - 2. has 83446 rows. 47 % not spam, 53% spam
  - 3. has 193849 rows. 53 % not spam 47% spam
  - Total: 282872 rows 52% not spam and 48% spam

- Data Columns:
  - **Label**: This column indicates whether the message is spam or not (binary classification: 0 for non-spam, 1 for spam).
  - **Text**: This column contains the content of the email message, usually in textual format.

Exploring data

Most frequent words in mail messages are “escapenumber”, “the” and “to” which makes sense. Data has 83863 of duplicate rows. there are no missing values or empty strings.

Word count in spam messages: 28568449
Word count in regular messages: 48733382

Looked for words that are more than 10 times more common in one than the other (took into account how many words were in Spam messages and not spam messages)

Words that are common spam messages but not common in regular messages:

1. cescapenumber spam: 116167, not spam: 4068
2. aescapenumber spam: 86593, not spam: 3279
3. eescapenumber span: 46980 , not spam: 2719

Common words that are in not spam messages but not in spam messages:

1. enron spam: 9872, not spam: 224707
2. org spam: 3066, not spam: 89643
3. pm spam: 1899, not spam: 66351

There are some special characters in mail messages which what common words are escapenumber and cescapenumber.

Verifying data quality

- Data consistency encoding is same in all datasets and has label column and column for mail message.
- Data integrity,
    i. 3. dataset label values has to be changed from string values to 0 and
    ii. Data has no missing values.
    iii. Data has a lot of duplicate rows. Only using dataset 3. there are 416 duplicate rows. no rows have mails with empty string.
    iv. There are some special characters in mails.

Task 4. Planning your project (0.25 points)

1. Make a detailed plan of your project with a list of tasks. There should be at least five tasks. Specify how many hours each team member will contribute to each task.

    1.1. Acquire a dataset of emails that fits our requirements
    1.2. Clean the data to make the words easily workable for the model

1.3. Analyze the data to find patterns
1.4. Make a machine learning model/models that can predict the email
1.5. Review and publish our findings and working model
1.6. Design and make the poster

2. List the methods and tools that you plan to use. Add any comments about the tasks that you think are important to clarify.

2.1. We wish to use kaggle to find the dataset
2.2. We'll try different machine learning models from sklearn, like:
    2.2.1. LVM
    2.2.2. Random forest
    2.2.3. Neural Net
2.3. Vectorize text using sklearn