

Capstone Project Report

AI-Driven Eco-Friendly Email Management for Enhanced Productivity in the Automotive Industry

Submitted for the fulfilment of ‘Certification in Data Science & AI’

Student : Annamdas Venu Gopal Madhav,PhD

Location : Singapore

Email Id: [venuannamdas1976@gmail.com](mailto:venuannamdas1976@gmail.com)

Table of Contents

i. Title Page  
ii. Problem Statement   
iii. Business context, stakeholders and value

* Overall understanding of the business domain
* Explanation of the business context
* Formulation of the business question
* Understanding and engagement of stakeholders
* Estimation of the business value

iv. Data description, sources, quality

* Translation of the business question into a data question
* Defining what data is needed to answer the business question
* Understand how to source the data
* Understanding of how the data was generated
* Understanding of the quality of data and its limitations
* Understanding of how the data can be sourced in the future

v. Data exploration, analysis and visualisation

* Data exploration showing the key entities and their business significance
* Using effective visualisation to communicate key aspects of the data

Vi. Documentation: text document, presentation and Notebooks

● Using the appropriate level of details to document the problem,

stakeholders and solution

● Organisation and structure of documentation and code

vii. The project planning, effort allocation and next steps

● Understanding of the effort used to perform the design work and

remaining effort to complete the project

● Defining the next steps to bring the project to production

viii. Feature Engineering

● Using business domain knowledge to select appropriate features

● Using appropriate encoding for each feature

ix. Creation of an effective reproducible pipeline

● Creation of a reproducible pipeline to ingest and prepare data and to train and evaluate the Machine Learning model

● Separation of the modelling pipeline from code for exploring and

analysing the data

x. Machine Learning model algorithms and accuracy

● Selection of the appropriate Machine Learning algorithm

● Evaluation of the model performance

● Applying multiple algorithms and comparing results

● Using appropriate metrics to express model performance

xi. Overall end-to-end solution

● Showing the overall end-to-end solution (UI, model, data,

infrastructure).

● State tools, libraries and frameworks used in the development of the

model and planned for the delivery of the solution.

● Appreciation of the effort and skills required to implement the whole solution

xii. Delivery of the presentation, poise and audience engagement

● Ability to deliver a clear, concise and engaging presentation

● Ability to listen effectively and address questions

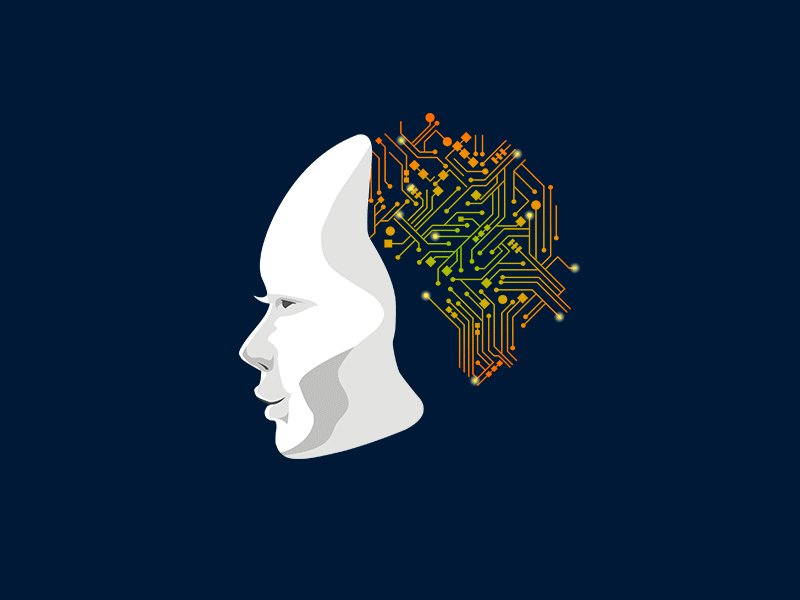
● Overall poise, confidence and rapport with the audience

● Keeping the time

i. Title Page

Title of the Project: "AI-Driven Eco-Friendly Email Management for Enhanced Productivity in the Automotive Industry"

(Motivation for the Project: Clean and Green Singapore https://www.cgs.gov.sg/)



ii. Problem Statement

Dr. David Woon, the Director at Continental Automotive Singapore, is grappling with email management challenges. He contends with a substantial daily influx of emails, creating difficulties in prioritization and timely responses. This inefficiency not only hampers his productivity but also presents a substantial cost concern, particularly if extrapolated to a larger employee base.

**Project Objectives:**

1. **Email Ranking Based on Urgency:** The project aims to develop a system that prioritizes emails in Dr. Woon's inbox based on their level of urgency.
2. **Multi-classification of Emails:** The project seeks to categorize the email set into distinct labels, providing a structured approach to email management.
   * **Required Response:** These are emails that demand a response from Dr. Woon, containing important information or requests requiring his attention.
   * **Required Response (BOSS):** This category includes emails of higher importance, originating from superiors. Dr. Woon should prioritize these.
   * **Required Response (SUB):** These are emails from Dr. Woon's subordinates or team members, and he should respond to them to maintain effective communication within his team.
   * **Job Seeker:** Dr. Woon may not necessarily need to respond to these emails personally; instead, he can forward them to the appropriate department or HR for follow-up.
   * **Time Waste:** While Dr. Woon's company filters most spam emails, there may still be a minority of emails that could potentially waste his time, such as old emails or those that have already been resolved.
   * **Time Waste (BOSS):** Similar to the "Time Waste" category, but these emails could originate from Dr. Woon's boss or superior employees within the company.

By considering these points, we can ensure that the email response recommendation and classification system is robust, customizable, and capable of meeting the specific needs of different organizations and users.

Top of Form

iii. Business context, stakeholders and value

**Business Context, Stakeholders, and Value**

1. **Overall Understanding of the Business Domain**

The project demonstrates a robust understanding of the automotive industry and its intricacies. It recognizes the challenges faced by professionals in this sector, especially in email management.

1. **Explanation of the Business Context**

The business context is articulated clearly, highlighting the specific challenges faced by Dr. David Woon, the Director at Continental Automotive Singapore. The excessive volume of daily emails and its impact on productivity and cost efficiency are well-detailed.

1. **Formulation of the Business Question**

The project successfully formulates the business question: How can Dr. David Woon effectively manage his email inbox to enhance productivity and reduce associated costs?

1. **Understanding and Engagement of Stakeholders**

Stakeholders in the automotive industry, particularly Dr. David Woon, are identified and engaged effectively. A Figma diagram provides additional details regarding stakeholders and their roles in the project.

1. **Estimation of the Business Value**

The project underscores the significance of this email management challenge by quantifying it. It is estimated that there is a wastage of approximately 31.25% of salary per day, which could be saved. For a company with over 200,000 employees globally, this translates to a substantial amount of time and cost savings daily.

**Stakeholders**

* **Who are the stakeholders?** The primary stakeholders are within the automotive industry, with a specific focus on Dr. David Woon, Director at Continental Automotive Singapore.
* **Why do they care about this problem?** The stakeholders, including Dr. Woon, are deeply concerned about reducing the time spent on email management and saving costs. The project emphasizes that the inefficiencies in email handling not only affect individual productivity but also have significant financial implications for the company.
* **What are the stakeholders’ expectations?** Given that this project is academic and serves as a capstone, the stakeholders' expectations are met by the project's ability to rank emails and provide multi-classification for Dr. Woon's email inbox. The project highlights the potential for substantial cost savings, making it a valuable endeavor for the stakeholders.
* **Stakeholders background and issues**

 Figure: Smart Continental Automotive Presence across the globe (fictitious location)

DashBoard Info: [raw.githack.com/venuannamdas/E\_clean\_AI/master/employee\_wastage\_map.html](https://raw.githack.com/venuannamdas/E_clean_AI/master/employee_wastage_map.html)

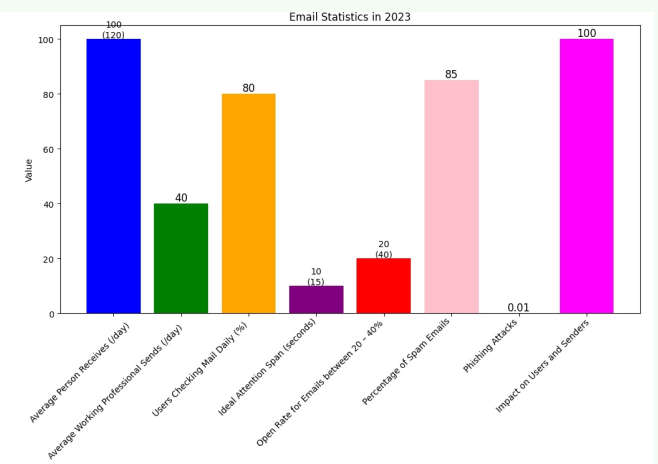


Figure : Company’s employee’s email usage

David gets a high volume of emails which is higher than the global average. He aims to reduce attention span, lower than the global average. His company provided a basic email organizer. Which is effective in reducing spam by 85%. It provides a generic Ranking based on emails sent and received. The average usage is around 150 minutes a day, which is very valuable for him. He wants to reduce this time and there is huge potential for the present project to be a pilot project (initial proposal without worrying about accuracies) in proposing a reduction in time and saving money for him and his company.

Dr. David Woon, is currently grappling with several email management challenges that require immediate attention. The baseline conditions for his email-box status needs improvement especially:, the project will concentrate on the following key

objectives: 1) Email Ranking Based on Urgency: Develop on the top of Dr. Woon's inbox.

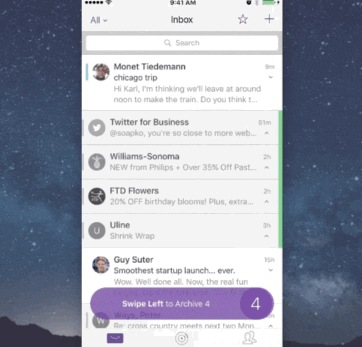
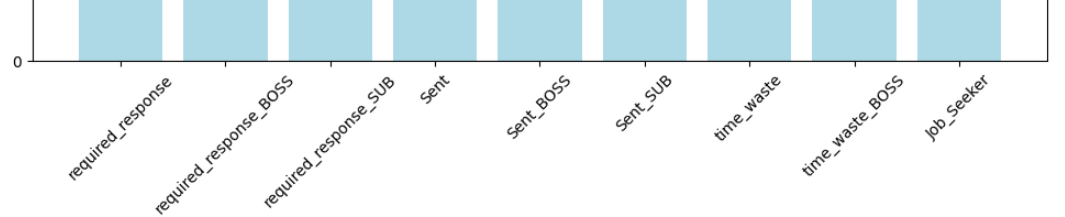


Figure: David’s email replications

2) Multi-Classification of Emails: Implement a multi-classification system for emails, categorizing them into distinct labels. This classification will provide a structured approach to email handling.



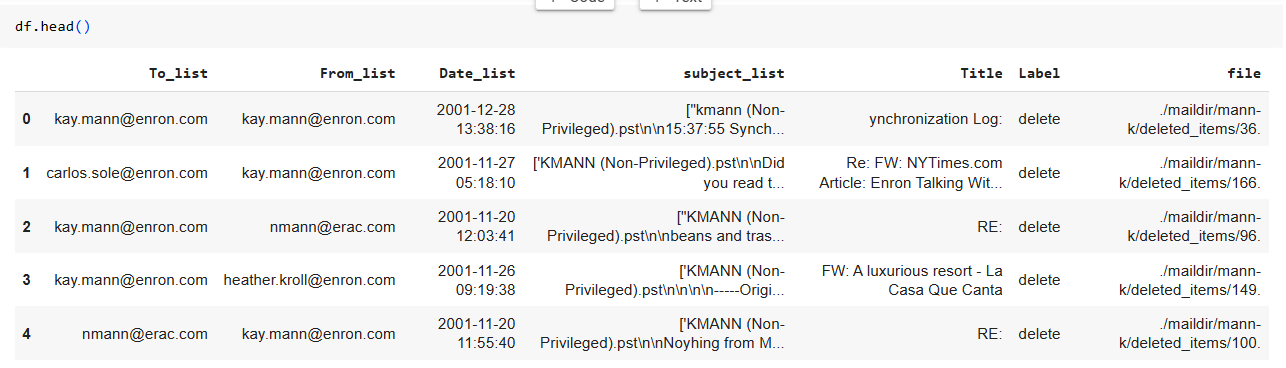
Overall, the project effectively addresses the business context, engages relevant stakeholders, and quantifies the business value, aligning with the stakeholders' concerns and expectations.

iv. Data description, sources, quality

Data is Enron email dataset of 600k observations **(**Reference 1: [The Enron Email Dataset | Kaggle](https://www.kaggle.com/datasets/wcukierski/enron-email-dataset)). The original data is enormous and only around 9571 observations are considered for the project for the first part. The data is now available at 'https://raw.githubusercontent.com/venuannamdas/E\_clean\_AI/master/CAPSTONE.csv'

The data is not of good quality, it is far from ideal for my study. My client is Dr. David Woon, and due to his organization’s rule he cannot provide me his own email ID. So, I had to take initial data from Enron and modify it.

Example of data.



The features are To\_list, From\_list, Date\_list, subject\_list, Title are the variable that are used as input. While Label is considered as output/target variable.

The data was modified, for example the email id is changed from  [kay.mann@enron.com](mailto:kay.mann@enron.com) to [david.woon@smartcontinental.com](mailto:david.woon@smartcontinental.com) to suit my client.

The details of initial preprocessing of data is given in preparatory.ipynb (it is in my github folder [venuannamdas/AI\_Clean\_Email\_Modeling: Clean and Green Email : Ranking and multi-classification Problem -AI solutions (github.com)](https://github.com/venuannamdas/AI_Clean_Email_Modeling/tree/master)

* Translation of the business question into a data question

The business question is how to provide proof of concept related to a customizable email tool kit to optimize the time and money for the end user via AI based machine learning tools.

* Defining what data is needed to answer the business question

Due to company’s policy I cannot get email data set from the company directly so I need to provide proof of concept to my client so that he can trust my vision. Thus I had to take initial data from Enron company.

* Understand how to source the data

Yes, all the aspects of data are presented in preparatory.ipynb and capstone\_Final.ipynb.

|  |
| --- |
| Loading Data from Preparatory File (cleaned initial data) Data questions & Replies  ● What is the data question that needs to be answered?  Reply: Emailset or Email box of the client  ● What is the data required to answer the question?  Reply: Customized and Cleaned Data is required  Data ● Where was the data sourced? Reply: Enron (basic) [https://www.kaggle.com/datasets/wcukierski/enron-email-dataset](https://colab.research.google.com/corgiredirector?site=https%3A%2F%2Fwww.kaggle.com%2Fdatasets%2Fwcukierski%2Fenron-email-dataset). (which is multi-user email dataset)  ● What is the volume and attributes of the data? Reply: (600K Observations)- TO, FROM, SUBJECT\_LINE, TITLE, LABEL, FILE  ● How reliable is the data?  Reply: INITIALLY LESS, as it belongs to enron company and the user dont want to share his dataset as i am not part of his company.  ● What is the quality of the raw data?  Reply: poor, I HAD TO SPEND LOT OF TIME MODIFYING to suit my client  ● How was this data generated? Reply: Initially took it from Enron, i needed something around 10k observations. so i first modified the data in another notebook (<https://github.com/venuannamdas/E_clean_AI/blob/master/preparatory.ipynb>). then did remaining steps like cleaning or addressing missing observations, etc.  ● Is this data available on an ongoing basis? Reply: it is not a continuous data but there is some dependency between columns such as Title and Subject\_line. Thus deep learning(Neural Network) related modelling can be carried out for modelling. |

* Understanding of how the data was generated

The features such as to, from, title, subject/content are the most important aspects of my data. So I later created additional email data to include David’s boss Vincent Wong and his subordinate Dr Shravan Kumar’s fictitious email conversations with my client Dr. David.



Figure: Network of client with his boss and subordinate

* Understanding of the quality of data and its limitations

There are missing data, and the data is not directly related to my client. This itself poses a challenge for me to modify the data to give meaning insights for a business goals of Dr. David Woon.

Click the Video to know my client



<https://youtu.be/eMWdeP17StA?si=2Ol_jclqmtz6Gj_l>

* Understanding of how the data can be sourced in the future

The project is aimed to show proof of concept to automotive company. I will have to convince them to source me their data for effective application of proof of concept.

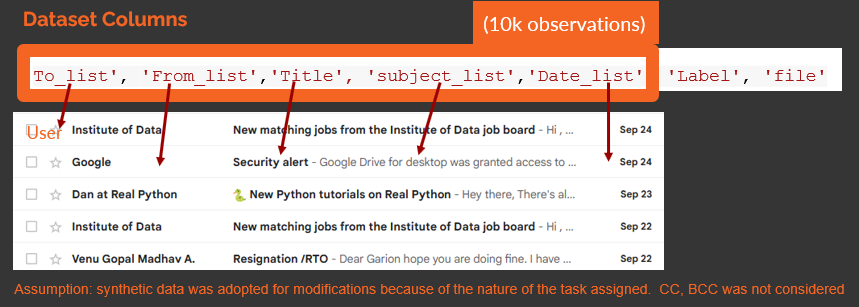


Figure: Data template.

'https://raw.githubusercontent.com/venuannamdas/E\_clean\_AI/master/smartcontinental.csv'

v. Data exploration, analysis and visualisation

* Data exploration showing the key entities and their business significance
* Using effective visualisation to communicate key aspects of the data

The baseline conditions for David’s email-box status needs improvement especially: the present project will concentrate on the following key objectives: 1) Email Ranking Based on Urgency: To develop on the top of Dr. Woon's inbox. 2) Multi-Classification of Emails: Implement a multi-classification system for emails, categorizing them into distinct labels. This classification will provide a structured approach to email handling.

Ranking prioritizes his emails thus saves his time; the segregation helps in organizing his email folders. These two aspects are related to impact his daily business affairs. Let us look at some (business significance) insights (refer to capstone\_Final.ipynb).

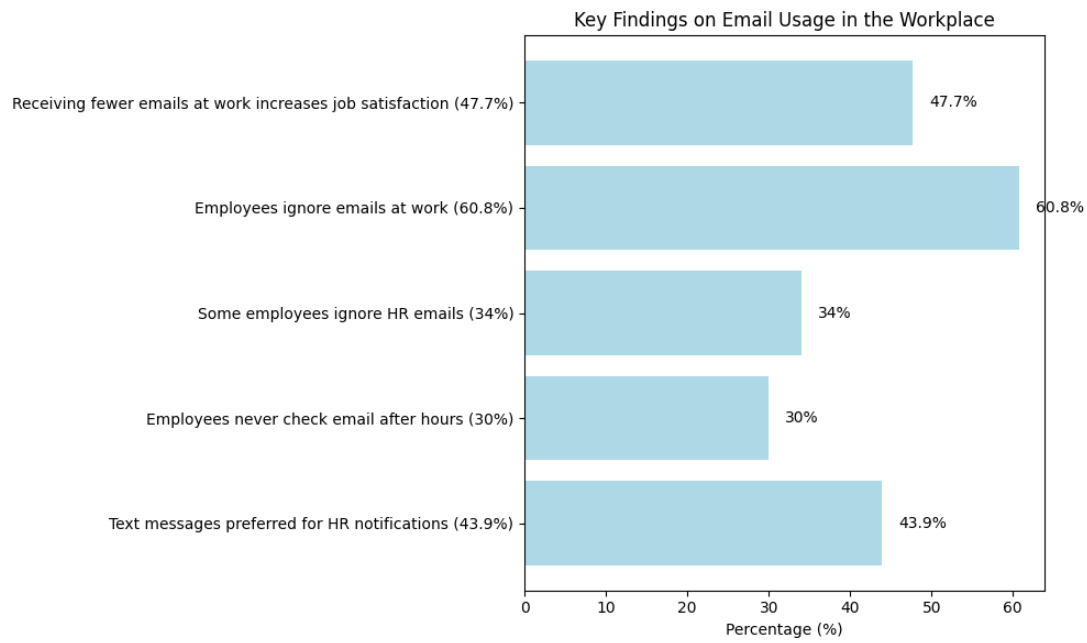


Figure: statistics of email usage

Let us also understand the wastage at each location as shown below.



Table: wastage in USD for each country (42 countries)

Thus, there is a business scope which can be addressed through this project.

Capstone\_Final.ipynb provides the details of the email data. Only few are presented below.

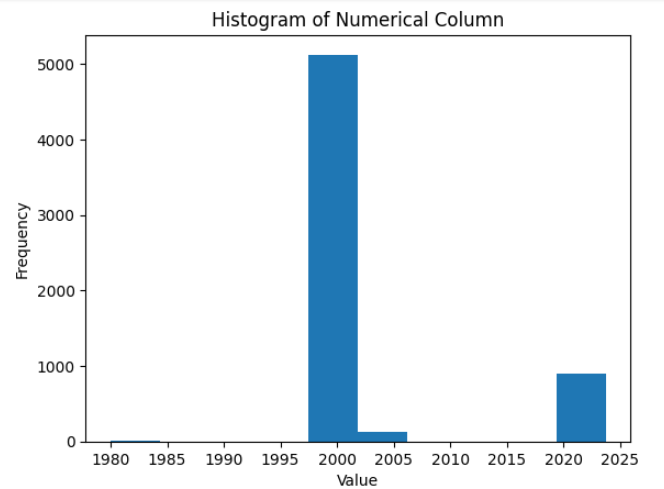
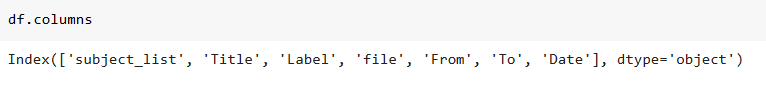


Figure: Column ‘Date’

the column Date shows that the emails taken for the present study is centred between 2000-2005, and 2020-2023

the dataframe has the following columns.



Vi. Documentation: text document, presentation and Notebooks

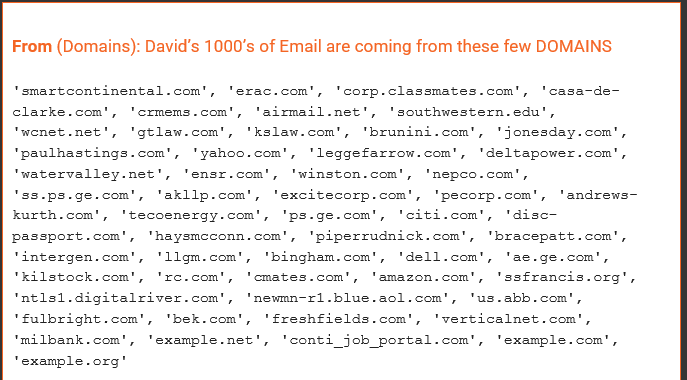
● Using the appropriate level of details to document the problem,

stakeholders and solution

● Organisation and structure of documentation and code

The notebooks (preparatory and capstone\_Final) provide more information of the data.

The modified data is processed well and several insights were found. For example,



Ranking : NextorkX based python package was used to obtain ranking (generic) and after considering customized options by using urgency list

urgency\_list = ['soon', 'shortly', 'urgent', 'pressing', 'speedy', 'prompt', 'straightaway', 'fast', 'promptly', 'quick', 'immediate', 'quickly', 'agile', 'ready', 'Vincent', 'vincent']

the following Figure shows a comparison between a simple (pre-existing baseline) and after implementing ranking algorithm using urgency list.

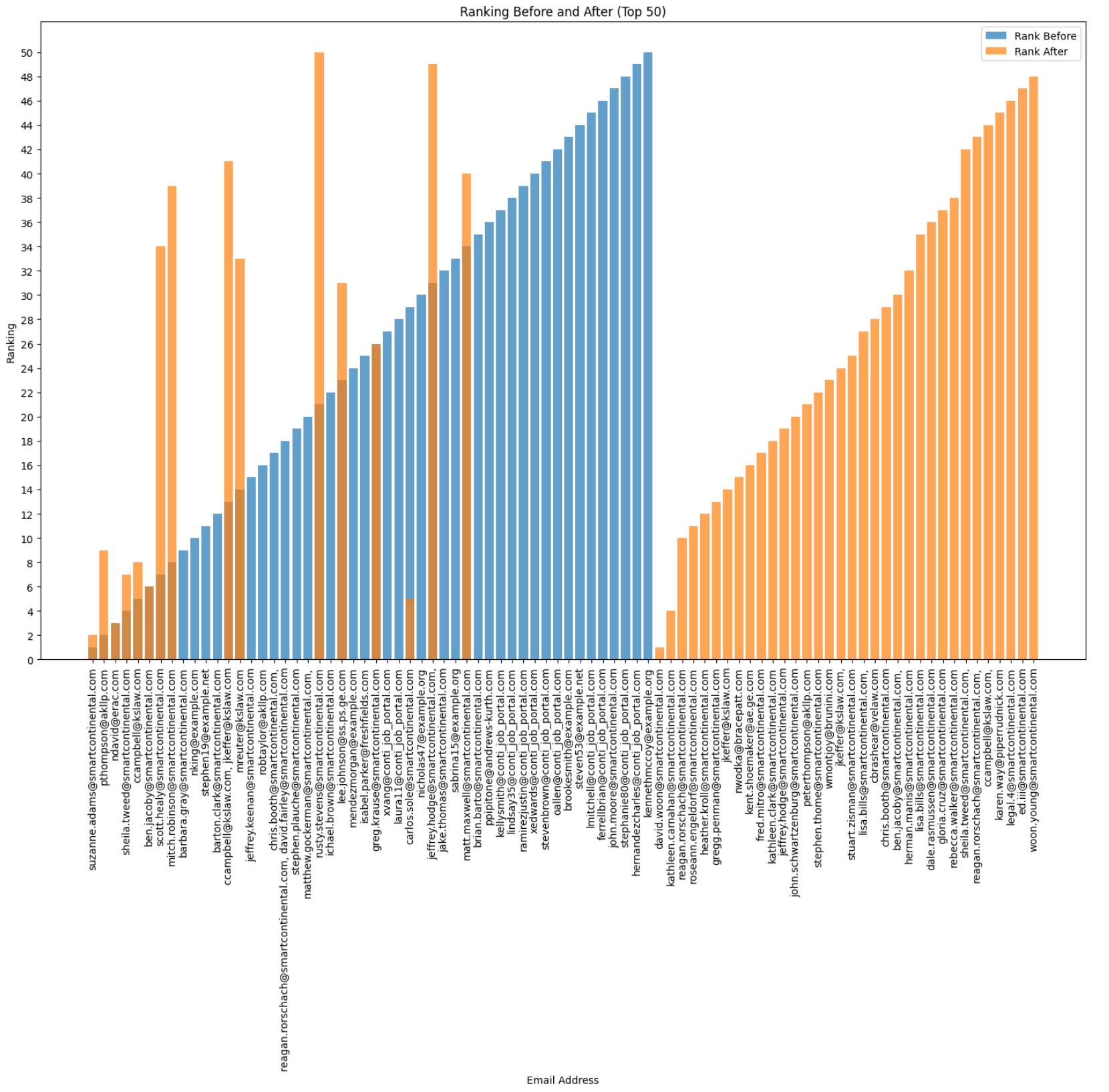


Figure: ranking comparison.

It should be noted that the colour blue and saffron show completely different email id (from) who have emailed Dr David woon. Just image the time involved in going through a wrong ranking system vs. a correct ranking system. It will be enormous, this it is important to customize any product to client based on his needs rather making it generic for all.

For details refer to my notebooks.

vii. The project planning, effort allocation and next steps

● Understanding of the effort used to perform the design work and

remaining effort to complete the project

As the student of datascience and AI, only one person is involved in designing the project. My client has requested wire framed design work to present to his team. I am presently working on wire based design to show this to the client in the near future.

Link: [Email\_project – FigJam (figma.com)](https://www.figma.com/file/6CuLnyoTRoE13pmuHykP7E/Email_project?type=whiteboard&node-id=875-1085&t=zoi1aoTqF643EAOa-0)

● Defining the next steps to bring the project to production

The model i.e the project will be showed to my client and will first get his feedback. Then will request him to give me original email set for showing the true potential of python/machine learning modelling.

Two weaks of time to finish this project from start to finish is really challenging and thus I took it more like proof of concept rather than ready to go project.

However, in the near future, this project will go for production for at least for one user.

viii. Feature Engineering

● Using business domain knowledge to select appropriate features

● Using appropriate encoding for each feature

Both the above are discussed in my notebooks and presentation slides. The most import input features in any text based email is TO, CC, BCC, DATE, SUBJECT\_LINE, TITLE. However, I assumed that there is no CC or BCC.

The output for modelling was Label, which comprised of several categories as explained in problem statement.

|  |  |
| --- | --- |
|  | **required\_response:** Emails that require a response from David, important information/requests that need his attention.  **required\_response\_BOSS**: These have higher importance, these are from superiors. David should prioritize  **required\_response\_SUB:** Emails from David's subordinates or team members. David should respond to maintain effective communication within his team.  **Job\_Seeker:** David may not necessarily need to respond. Instead, he can forward them to the appropriate department/person/HR for follow-up. |

To simplify the modelling, few labels were dropped (refer notebook for more information)

Let us revisit a Figure (which was seen earlier)

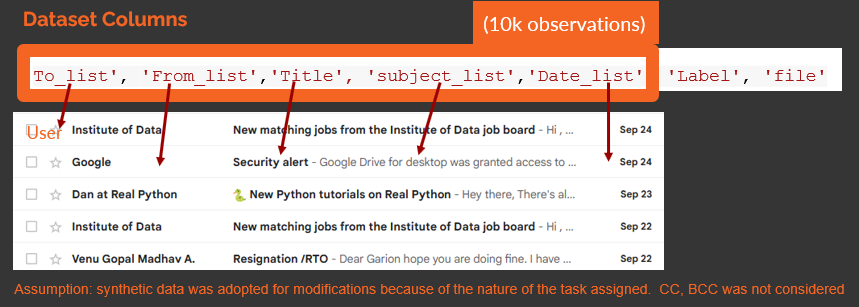


Figure: Data template.

All the features seems to be independent and only the Label is dependent on other variables

ix. Creation of an effective reproducible pipeline

● Creation of a reproducible pipeline to ingest and prepare data and to train and evaluate the Machine Learning model

In the capstone\_Final file, I considerd joblib and pickle, in order to use them for production.

● Separation of the modelling pipeline from code for exploring and

analysing the data

The code is provided below the conclusions in the file. It can be used separately and executed.

x. Machine Learning model algorithms and accuracy

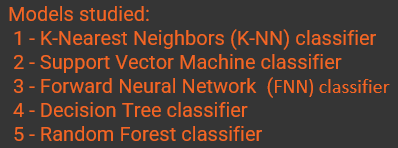
● Selection of the appropriate Machine Learning algorithm

● Evaluation of the model performance

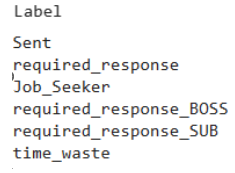
● Applying multiple algorithms and comparing results

● Using appropriate metrics to express model performance

Five models are used to study the present project.



The output required are reduced, i.e the Labels are now 6 as shown below

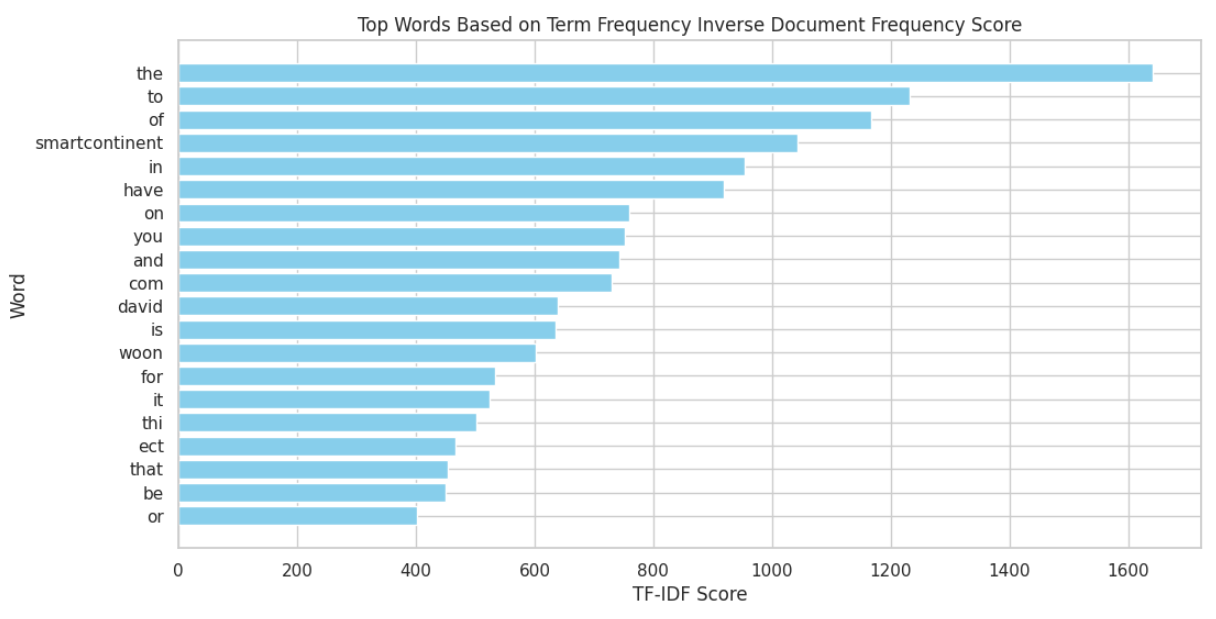


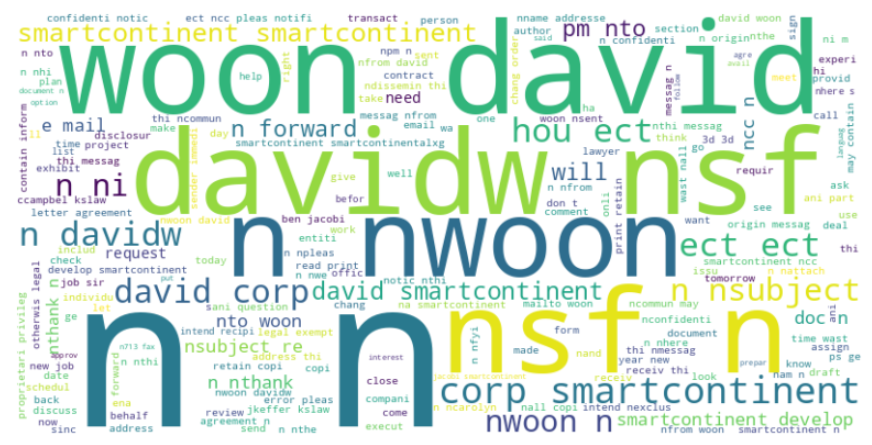
First preprocessing was carriedout were punctuations are removed . Preprocessing is an essential step in natural language processing (NLP) tasks, and one common preprocessing task is removing punctuation from text data. Punctuation includes characters like periods, commas, question marks, exclamation points, and other symbols used to structure sentences and convey meaning. Removing punctuation can help clean and simplify text data, making it easier to analyze and process.

The following processes was carriedout before modelling.

Term frequency - Inverse Document frequency

1. Prepossing’
2. Stemming using porter stemmer
3. POS tagging





The important words are printed as shown in the Figure. You can notice my Client Name, David Woon, his company smart continental, need, contract etc.

The results of the modelling are as follows

# K-Nearest Neighbors (K-NN) Classifier for TF-IDF transformed data and evaluate its performance

# 

# Support Vector Machine classifier for TF-IDF transformed data and evaluate its performance.

# 

# Forward Neural Network (Deep Learning)for TF-IDF transformed data and evaluate its performance

# 

# Decision Tree classifier for TF-IDF transformed data and evaluate its performance

# 

# Random Forest classifier for TF-IDF transformed data and evaluate its performance

# 

# -

# 

# Conclusion:

Supervised Support Vector Machine modeling and Feedforward Neural Network methods exhibit promising performance. While the former treats input variables as independent features, the latter considers a more interconnected relationship among input variables, particularly Title and Subject. Therefore, further investigation and exploration of these models could lead to potential enhancements.

# Files for deployment

SVM PICKLE FILE - svm\_model.pkl

fNN -JOBLIB

label\_encoder.joblib

onehot\_encoder.joblib

----------------------

References

[https://www.demandsage.com/how-many-emails-are-sent-per-day/](https://colab.research.google.com/corgiredirector?site=https%3A%2F%2Fwww.demandsage.com%2Fhow-many-emails-are-sent-per-day%2F)

[https://earthweb.com/how-many-emails-does-the-average-person-receive-per-day/#:~:text=billion%20by%202023.-,2.,by%20an%20average%20office%20worker](https://colab.research.google.com/corgiredirector?site=https%3A%2F%2Fearthweb.com%2Fhow-many-emails-does-the-average-person-receive-per-day%2F%23%3A%7E%3Atext%3Dbillion%2520by%25202023.-%2C2.%2Cby%2520an%2520average%2520office%2520worker)

[https://www.slicktext.com/blog/2019/06/survey-workplace-communication-statistics/#:~:text=Almost%20half%20(45.6%25)%20of,some%20emails%20in%20their%20inbox](https://colab.research.google.com/corgiredirector?site=https%3A%2F%2Fwww.slicktext.com%2Fblog%2F2019%2F06%2Fsurvey-workplace-communication-statistics%2F%23%3A%7E%3Atext%3DAlmost%2520half%2520%2845.6%2525%29%2520of%2Csome%2520emails%2520in%2520their%2520inbox)

[https://www.digitalinformationworld.com/2023/01/people-spend-33-less-time-reading.html#:~:text=Back%20in%202018%2C%20studies%20showed,considered%20and%20taken%20into%20account](https://colab.research.google.com/corgiredirector?site=https%3A%2F%2Fwww.digitalinformationworld.com%2F2023%2F01%2Fpeople-spend-33-less-time-reading.html%23%3A%7E%3Atext%3DBack%2520in%25202018%252C%2520studies%2520showed%2Cconsidered%2520and%2520taken%2520into%2520account)

[https://hbr.org/2019/01/how-to-spend-way-less-time-on-email-every-day#:~:text=The%20average%20professional%20spends%2028,120%20messages%20received%20per%20day](https://colab.research.google.com/corgiredirector?site=https%3A%2F%2Fhbr.org%2F2019%2F01%2Fhow-to-spend-way-less-time-on-email-every-day%23%3A%7E%3Atext%3DThe%2520average%2520professional%2520spends%252028%2C120%2520messages%2520received%2520per%2520day)

[https://hbr.org/2019/01/how-to-spend-way-less-time-on-email-every-day#:~:text=The%20average%20professional%20spends%2028,120%20messages%20received%20per%20day](https://colab.research.google.com/corgiredirector?site=https%3A%2F%2Fhbr.org%2F2019%2F01%2Fhow-to-spend-way-less-time-on-email-every-day%23%3A%7E%3Atext%3DThe%2520average%2520professional%2520spends%252028%2C120%2520messages%2520received%2520per%2520day)

[https://financesonline.com/email-statistics/](https://colab.research.google.com/corgiredirector?site=https%3A%2F%2Ffinancesonline.com%2Femail-statistics%2F)

xi. Overall end-to-end solution

● Showing the overall end-to-end solution (UI, model, data,

infrastructure).

● State tools, libraries and frameworks used in the development of the

model and planned for the delivery of the solution.

● Appreciation of the effort and skills required to implement the whole solution

* Reference 1: [The Enron Email Dataset | Kaggle](https://www.kaggle.com/datasets/wcukierski/enron-email-dataset)

[E-mail marketing worldwide - statistics & facts | Statista](https://www.statista.com/topics/1446/e-mail-marketing/#topicOverview)

[Number of Email Users Worldwide 2022/2023: Demographics & Predictions - Financesonline.com](https://financesonline.com/number-of-email-users/)