Spam Base Project

Ryan ASSAAD – Felix HAMEL

TD DIA4 - Alternant

What is Spam Email ? AND Context

✓ **Definition :**

- Spam email is unsolicited and unwanted junk email sent out in bulk to an indiscriminate recipient list. Typically, spam is sent for commercial purposes. It can be sent in massive volume by botnets, networks of infected computers

✓ **Why do people send out spam email ?**

- Spam email is sent for commercial purposes. While some people view it as unethical, many businesses still use spam.

✓ **Is Spam Email dangerous ?**

- Spam email can be dangerous. It can include malicious links that can infect your computer with malware (see What is malware?). Do not click links in spam. Dangerous spam emails often sound urgent, so you feel the need to act.

✓ **Spam Statistics :**

- The average number of legitimate email messages sent overt the internet each day : 22.43 billion
- Nearly 85% of all emails are spam
- Advertising makes up 36% of all world spam content

How can we reduce Spam ?

✓ **Identify different types of spam : Comon types**

- Commercial advertisements
- Antivirus warnings
- Email spoofing
- Sweepstakes winners
- Money scams

✓ **Read the content**

✓ **Use a spam filter**

Goal of this project :

The aim of our project is to predict the nature of an email : is it a spam or not ?

In this way, we are going to analyse different variables of mails and try different machine learnings algorithms with changing hyper paramters to have the best prediction solution.

Data Set Characteristics:	Multivariate	**Number of Instances:**	4601	**Area:**	Computer
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	57	**Date Donated**	1999-07-01
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	632744

Data in which analysis are based on more than two variables

It is interesting to see that all the characteristics are numbers here

Here the problem is a classifciation problem

Here the problem is a classifciation problem : the target can have only two possibilities, spam or not spam

Each individual is described by 57 attributes

According to this résumé, there are some missing values that we will have to treat in the data exploration

According to this résumé, there are some missing values that we will have to treat in the data exploration

Date donated of the data

A hit is often a request made to a website for a particular file to be downloaded from the server.

Data Set Information

The last column of 'spambase.data' denotes whether the e-mail was considered spam (1) or not (0), i.e. unsolicited commercial e-mail. Most of the attributes indicate whether a particular word or character was frequently occuring in the e-mail. The run-length attributes (55-57) measure the length of sequences of consecutive capital letters. For the statistical measures of each attribute, see the end of this file. Here are the definitions of the attributes:

48 continuous real [0,100] attributes of type word_freq_WORD
= percentage of words in the e-mail that match WORD, i.e. 100 * (number of times the WORD appears in the e-mail) / total number of words in e-mail. A "word" in this case is any string of alphanumeric characters bounded by non-alphanumeric characters or end-of-string.

6 continuous real [0,100] attributes of type char_freq_CHAR]
= percentage of characters in the e-mail that match CHAR, i.e. 100 * (number of CHAR occurences) / total characters in e-mail

1 continuous real [1,...] attribute of type capital_run_length_average
= average length of uninterrupted sequences of capital letters

1 continuous integer [1,...] attribute of type capital_run_length_longest
= length of longest uninterrupted sequence of capital letters

1 continuous integer [1,...] attribute of type capital_run_length_total
= sum of length of uninterrupted sequences of capital letters
= total number of capital letters in the e-mail

1 nominal {0,1} class attribute of type spam
= denotes whether the e-mail was considered spam (1) or not (0), i.e. unsolicited commercial e-mail.

0	word_freq_make	4601 non-null	float64
1	word_freq_address	4601 non-null	float64
2	word_freq_all	4601 non-null	float64
3	word_freq_3d	4601 non-null	float64
4	word_freq_our	4601 non-null	float64
5	word_freq_over	4601 non-null	float64
6	word_freq_remove	4601 non-null	float64
7	word_freq_internet	4601 non-null	float64
8	word_freq_order	4601 non-null	float64
9	word_freq_mail	4601 non-null	float64
10	word_freq_receive	4601 non-null	float64
11	word_freq_will	4601 non-null	float64
12	word_freq_people	4601 non-null	float64
13	word_freq_report	4601 non-null	float64
14	word_freq_addresses	4601 non-null	float64
15	word_freq_free	4601 non-null	float64
16	word_freq_business	4601 non-null	float64
17	word_freq_email	4601 non-null	float64
18	word_freq_you	4601 non-null	float64
19	word_freq_credit	4601 non-null	float64
20	word_freq_your	4601 non-null	float64
21	word_freq_font	4601 non-null	float64
22	word_freq_000	4601 non-null	float64
23	word_freq_money	4601 non-null	float64
24	word_freq_hp	4601 non-null	float64
25	word_freq_hpl	4601 non-null	float64
26	word_freq_george	4601 non-null	float64
27	word_freq_650	4601 non-null	float64
28	word_freq_lab	4601 non-null	float64
29	word_freq_labs	4601 non-null	float64
30	word_freq_telnet	4601 non-null	float64
31	word_freq_857	4601 non-null	float64
32	word_freq_data	4601 non-null	float64
33	word_freq_415	4601 non-null	float64
34	word_freq_85	4601 non-null	float64
35	word_freq_technology	4601 non-null	float64
36	word_freq_1999	4601 non-null	float64
37	word_freq_parts	4601 non-null	float64
38	word_freq_pm	4601 non-null	float64
39	word_freq_direct	4601 non-null	float64
40	word_freq_cs	4601 non-null	float64
41	word_freq_meeting	4601 non-null	float64
42	word_freq_original	4601 non-null	float64
43	word_freq_project	4601 non-null	float64
44	word_freq_re	4601 non-null	float64
45	word_freq_edu	4601 non-null	float64
46	word_freq_table	4601 non-null	float64
47	word_freq_conference	4601 non-null	float64
48	char_freq_;	4601 non-null	float64
49	char_freq_(	4601 non-null	float64
50	char_freq_[	4601 non-null	float64
51	char_freq_!	4601 non-null	float64
52	char_freq_$	4601 non-null	float64
53	char_freq_#	4601 non-null	float64
54	capital_run_length_average	4601 non-null	float64
55	capital_run_length_longest	4601 non-null	int64
56	capital_run_length_total	4601 non-null	int64
57	spam	4601 non-null	int64

Data Pre-Processing :

We separate the features into two categories : features_frequency and features_numbers

- **Features_Frequency : Contains all the features that are frequencies, it means their value is between 0 and 100 %**

'word_freq_make', 'word_freq_address', 'word_freq_all', 'word_freq_3d', 'word_freq_our', 'word_freq_over', 'word_freq_remove', 'word_freq_internet', 'word_freq_order', 'word_freq_mail', 'word_freq_receive', 'word_freq_will', 'word_freq_people', 'word_freq_report', 'word_freq_addresses', 'word_freq_free', 'word_freq_business', 'word_freq_email', 'word_freq_you', 'word_freq_credit', 'word_freq_your', 'word_freq_font', 'word_freq_000', 'word_freq_money', 'word_freq_hp', 'word_freq_hpl', 'word_freq_george', 'word_freq_650', 'word_freq_lab', 'word_freq_labs', 'word_freq_telnet', 'word_freq_857', 'word_freq_data', 'word_freq_415', 'word_freq_85', 'word_freq_technology', 'word_freq_1999', 'word_freq_parts', 'word_freq_pm', 'word_freq_direct', 'word_freq_cs', 'word_freq_meeting', 'word_freq_original', 'word_freq_project', 'word_freq_re', 'word_freq_edu', 'word_freq_table', 'word_freq_conference', 'char_freq_;', 'char_freq_(', 'char_freq_[', 'char_freq_!', 'char_freq_$', 'char_freq_#'

- **Features_Numbers : Concern the capital_run_length, that is the total number of capital letters in the mail, we have differnt informations regarding it :**

'capital_run_length_average', 'capital_run_length_longest', 'capital_run_length_total'

Capital_run_length_average : it is the mean length of uninterrupted sequences of capital letters

Capital_run_length_longest : it is the longest sequence of capital letters

Data Pre-Processing :

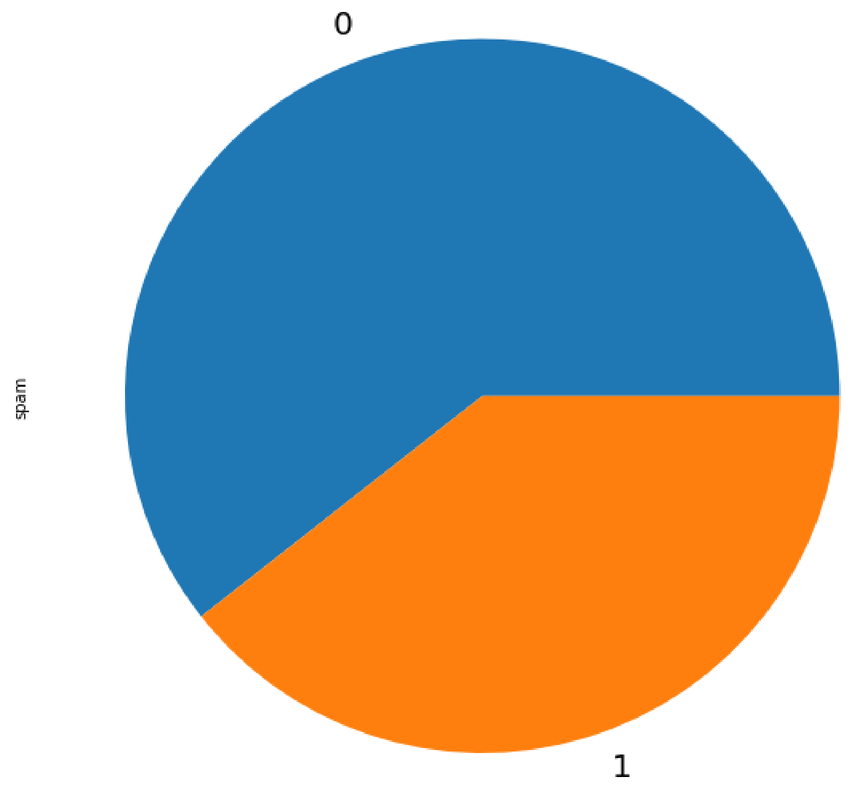
- df.isna().sum() result : 0
- Our data set is CLEAN ! We have no NaN values
- No work to complete the ptential lack of data
- All lines are completed with proper informations so we can exploit the data.

Data Pre-Processing : Descriptive Statistics
Some interesting informations

- 51 frequency features have **75% of data values equal to 0**
- Complexity of the language : **470 000 english words** ➔ One word could not be present in every mails and for most of them it is not present !
- We are going to deal with small values and it adds **complexity** to our analyze …

	word_freq_make	word_freq_address	word_frec
count	4601.000000	4601.000000	4601.0(
mean	0.104553	0.213015	0.2٤
std	0.305358	1.290575	0.5(
min	0.000000	0.000000	0.0(
25%	0.000000	0.000000	0.0(
50%	0.000000	0.000000	0.0(
75%	0.000000	0.000000	0.42
max	4.540000	14.280000	5.1(

Data Exploration:
Target distribution

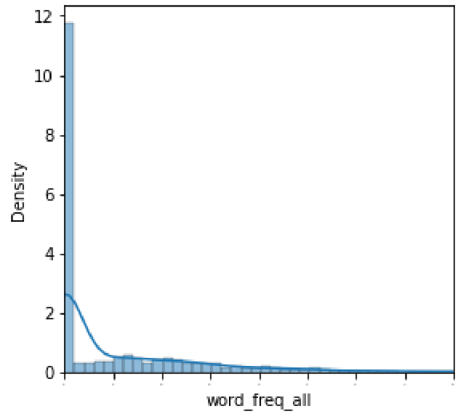


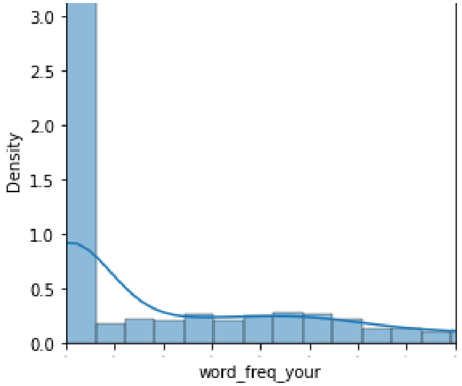
61% of the population

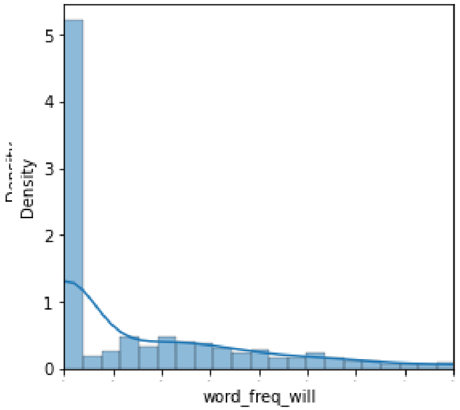
non spam

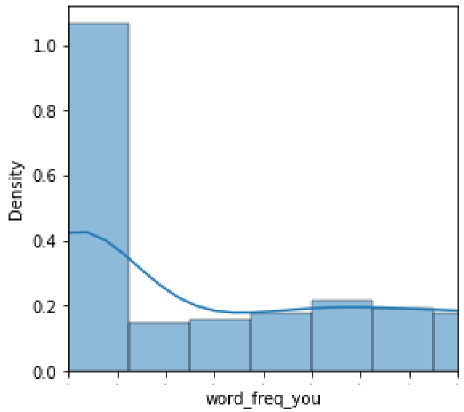
39 % of the population spam

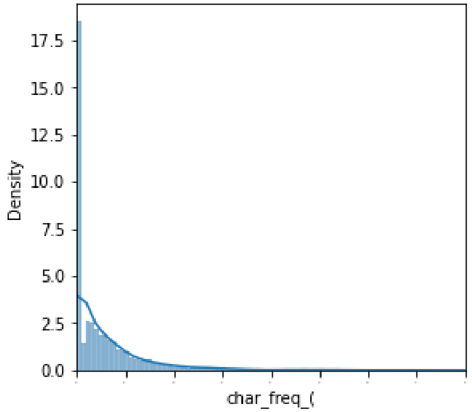
Data Exploration:
Frequency Features with Particular Distributions





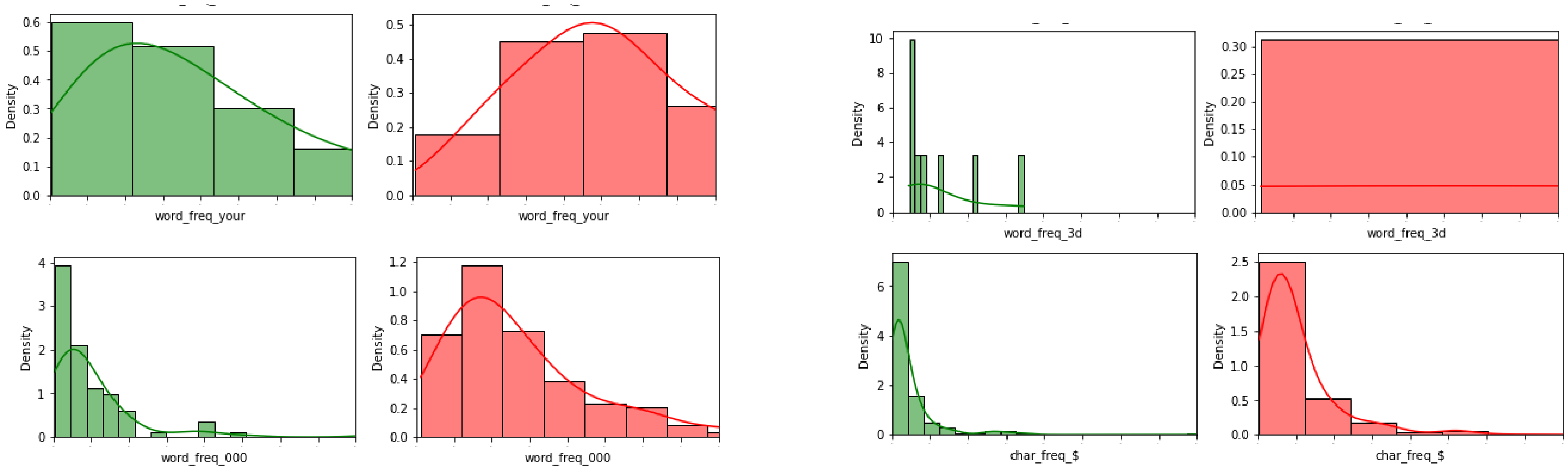






When we see these distribution graphes we can guess « normal laws » on these intervals. It is not very precise but we can guess.

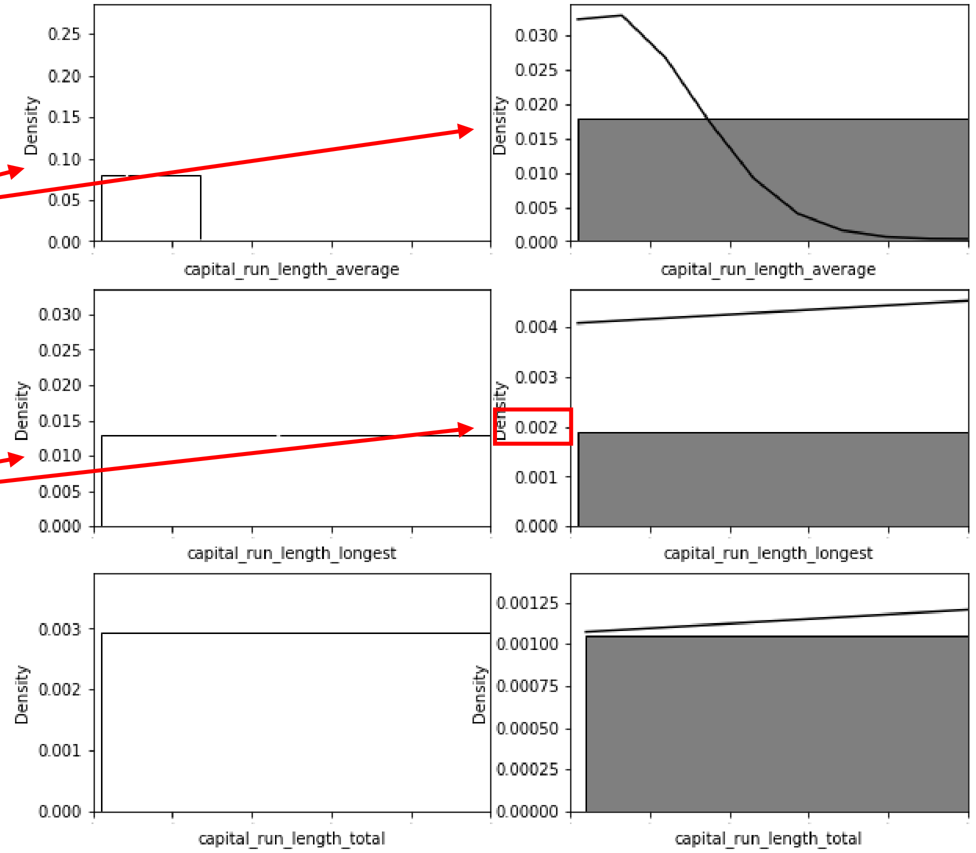
Data Exploration:
Frequency Features Distributions – Non Spam (Green) VS Spam (Red)
AND without zeros



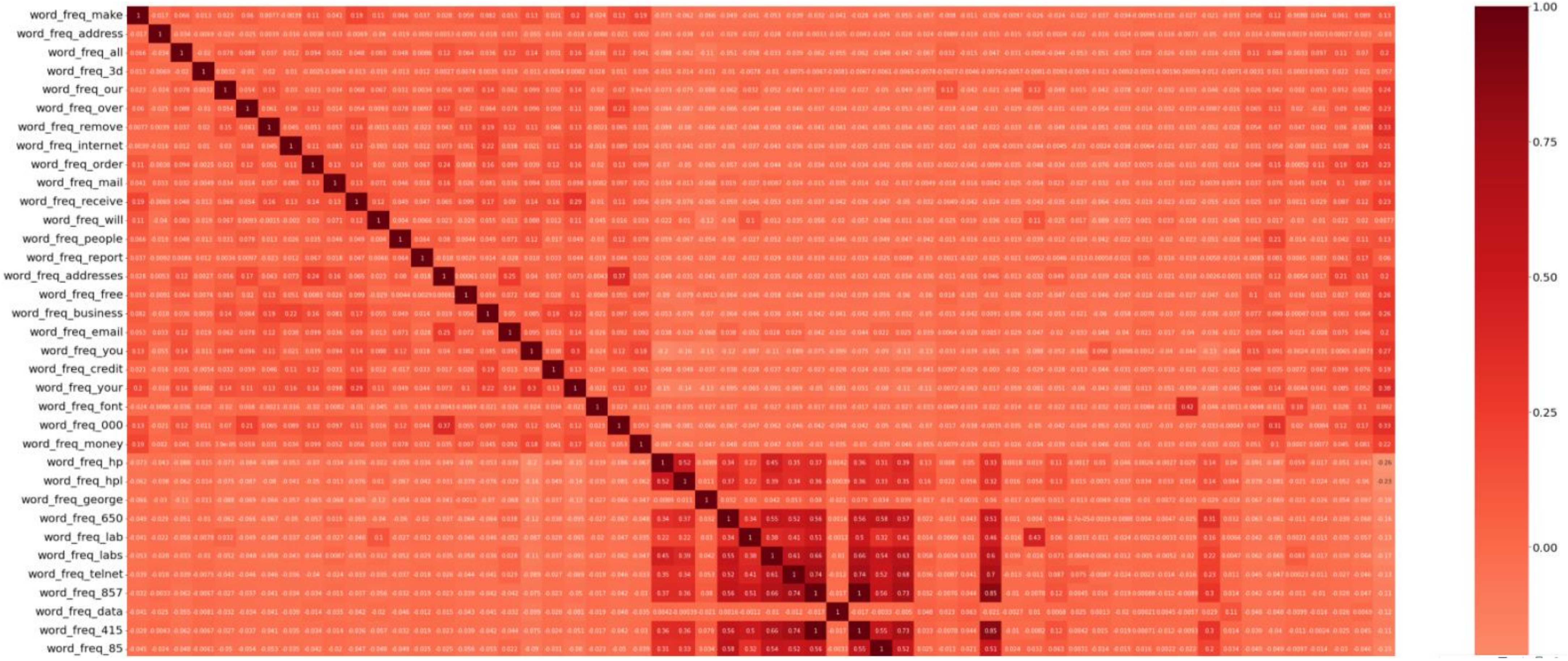
We can see differences in terms of distribution between non_spam and spam.

Data Exploration:
Number Features Distributions – Non Spam (White) VS Spam (Black)
AND without zeros

We can see differences in the distribution. Indeed, when we look at the scale, it is not the same, and for capital_run_length_average, the distribution is below 0.020 for spam, and for capital_run_length_longest it is even more obvious with a distribution below 0.002 for spam.



Data Exploration: Correlation Matrix



Data Exploration: Features Selection

- After studiyng this correlation matrix, we decided to select only the features with a correla tion to spam superior to 0.2 or inferior to -0.2.
  WE ORDER THEM IN ORDER OF CORRELATION :

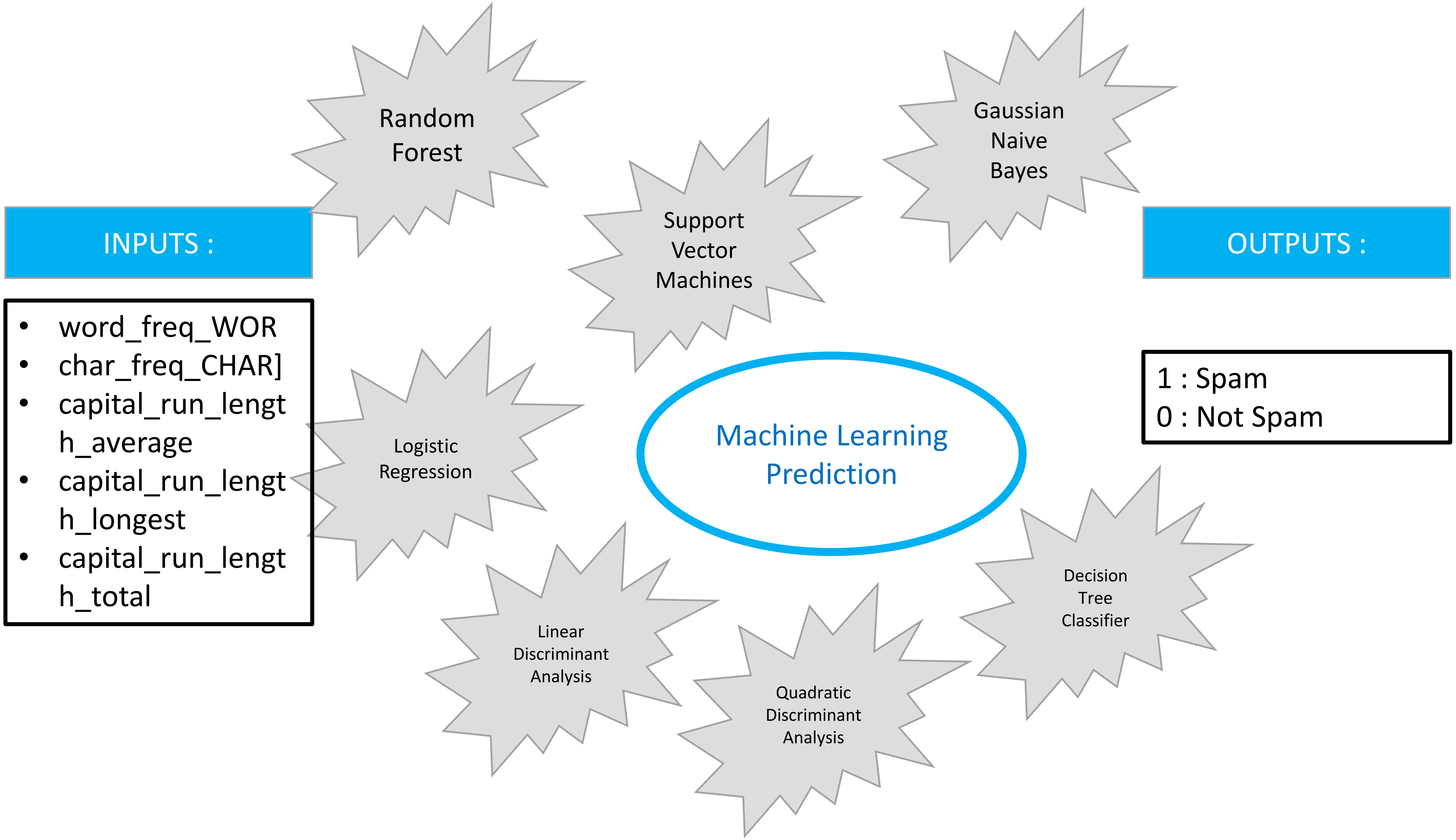
word_freq_your, word_freq_000, word_freq_remove, char__freq_$, word_freq_you, word_freq_business, word_freq_free, capital_run_length_total, word_freq_our, capital_run_length_longest, char_freq_!, word_freq_over, word_freq_order, word_freq_receive, word_freq_money, word_freq_internet, word_freq_all, word_freq_addresses, word_freq_email

Top_features = word_freq_your, word_freq_000, word_freq_remove, char__freq_$, word_freq_you is the tope 5 correlated features.

We put the other features in other_features and we will use them to check if there are second degree correlations with top_features.

Other Features Correlated with Top Features ?

We try to identify second degree correlations between top_features and other_features, with heatmap, but we don't find any.

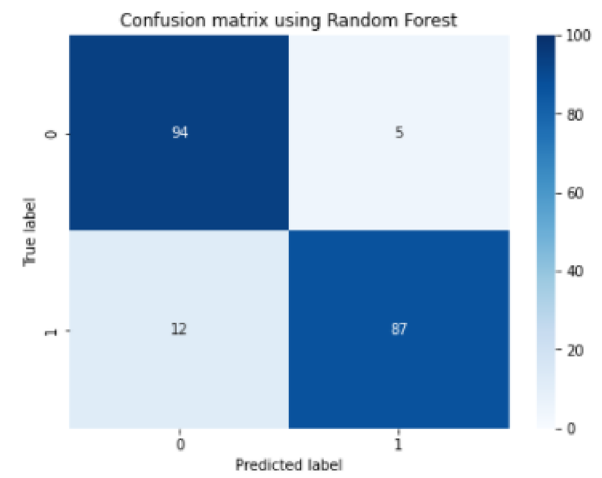


INPUTS :
word_freq_WOR
char_freq_CHAR]
capital_run_length_average
capital_run_length_longest
capital_run_length_total
Random Forest
Support Vector Machines
Gaussian Naive Bayes
Logistic Regression
Machine Learning Prediction
Linear Discriminant Analysis
Quadratic Discriminant Analysis
Decision Tree Classifier
OUTPUTS :
1 : Spam
0 : Not Spam

Data Prediction : Choice of Random Forest Classifier :

- Confusion Matrix :

In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as an error matrix is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one. Each row of the matrix represents the instances in an actual class while each column represents the instances in a predicted class



Data Prediction :
Choice of Random Forest Classifier :

- ROC Curve : An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

- True Positive Rate
- False Positive Rate