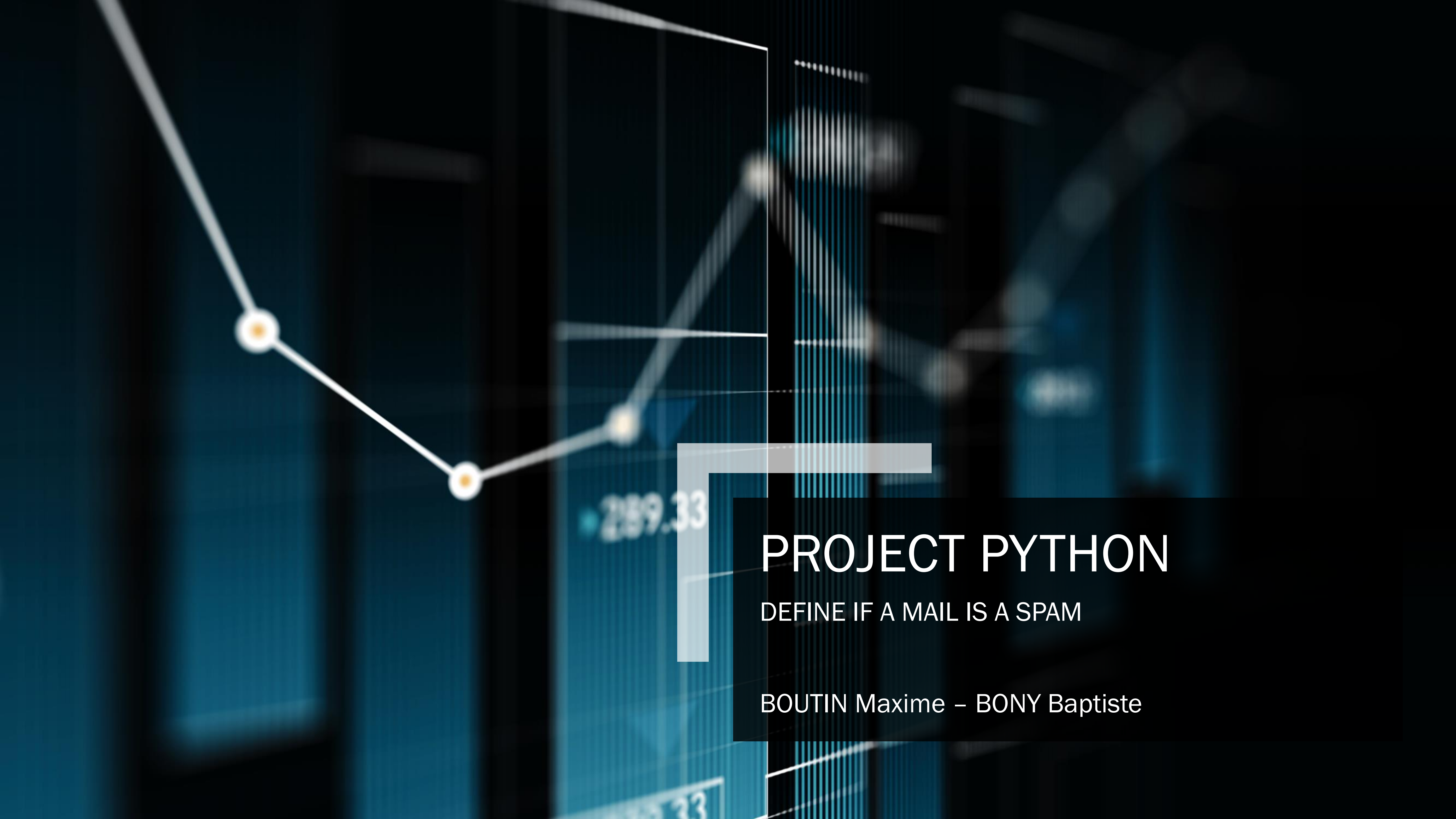
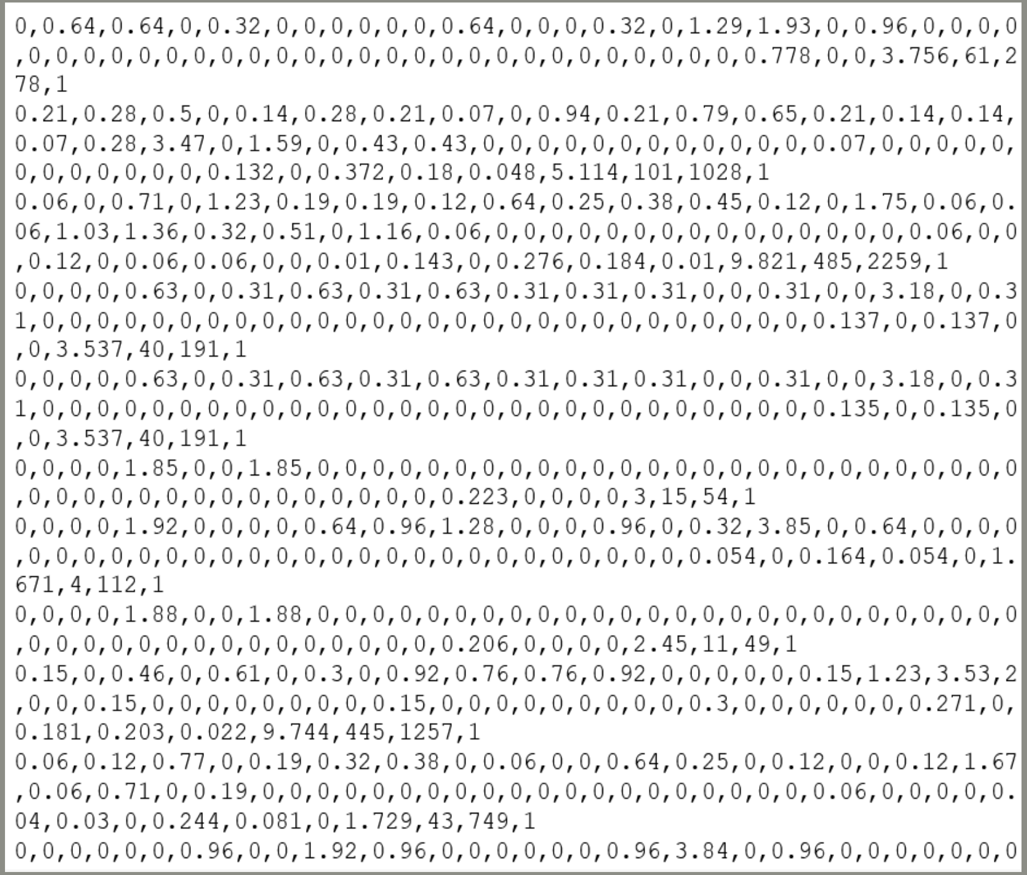


PROJECT PYTHON

DEFINE IF A MAIL IS A SPAM

BOUTIN Maxime – BONY Baptiste

Understanding of the data



```
0,0.64,0.64,0,0.32,0,0,0,0,0,0,0.64,0,0,0,0.32,0,1.29,1.93,0,0.96,0,0,0,0
,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.778,0,0,3.756,61,2
78,1
0.21,0.28,0.5,0,0.14,0.28,0.21,0.07,0,0.94,0.21,0.79,0.65,0.21,0.14,0.14,
0.07,0.28,3.47,0,1.59,0,0.43,0.43,0,0,0,0,0,0,0,0,0,0,0,0,0.07,0,0,0,0,0,
0,0,0,0,0,0,0,0.132,0,0.372,0.18,0.048,5.114,101,1028,1
0.06,0,0.71,0,1.23,0.19,0.19,0.12,0.64,0.25,0.38,0.45,0.12,0,1.75,0.06,0.
06,1.03,1.36,0.32,0.51,0,1.16,0.06,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.06,0,0
,0.12,0,0.06,0.06,0,0,0.01,0.143,0,0.276,0.184,0.01,9.821,485,2259,1
0,0,0,0,0.63,0,0.31,0.63,0.31,0.63,0.31,0.31,0.31,0,0,0.31,0,0,3.18,0,0.3
1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.137,0,0.137,0
,0,3.537,40,191,1
0,0,0,0,0.63,0,0.31,0.63,0.31,0.63,0.31,0.31,0.31,0,0,0.31,0,0,3.18,0,0.3
1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.135,0,0.135,0
,0,3.537,40,191,1
0,0,0,0,1.85,0,0,1.85,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.223,0,0,0,0,3,15,54,1
0,0,0,0,1.92,0,0,0,0,0.64,0.96,1.28,0,0,0,0.96,0,0.32,3.85,0,0.64,0,0,0,0
,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.054,0,0.164,0.054,0,1.
671,4,112,1
0,0,0,0,1.88,0,0,1.88,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.206,0,0,0,0,2.45,11,49,1
0.15,0,0.46,0,0.61,0,0.3,0,0.92,0.76,0.76,0.92,0,0,0,0,0,0.15,1.23,3.53,2
,0,0,0.15,0,0,0,0,0,0,0,0,0.15,0,0,0,0,0,0,0,0,0,0.3,0,0,0,0,0,0,0.271,0,
0.181,0.203,0.022,9.744,445,1257,1
0.06,0.12,0.77,0,0.19,0.32,0.38,0,0.06,0,0,0.64,0.25,0,0.12,0,0,0.12,1.67
,0.06,0.71,0,0.19,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.06,0,0,0,0,0.
04,0.03,0,0.244,0.081,0,1.729,43,749,1
0,0,0,0,0,0,0.96,0,0,1.92,0.96,0,0,0,0,0,0,0.96,3.84,0,0.96,0,0,0,0,0,0,0
```

- 58 different features

There are 57 numerical values, 1 nominal class label :

48 attributes are the percentage of some words in the e-mail

6 attributes are the percentage of some characters in the e-mail

1 is the length of longest uninterrupted sequence of capital letters

1 is the total number of capital letters in the e-mail

1 is the average length of uninterrupted sequences of capital letters

The last attributes denotes if the mail is a spam or not

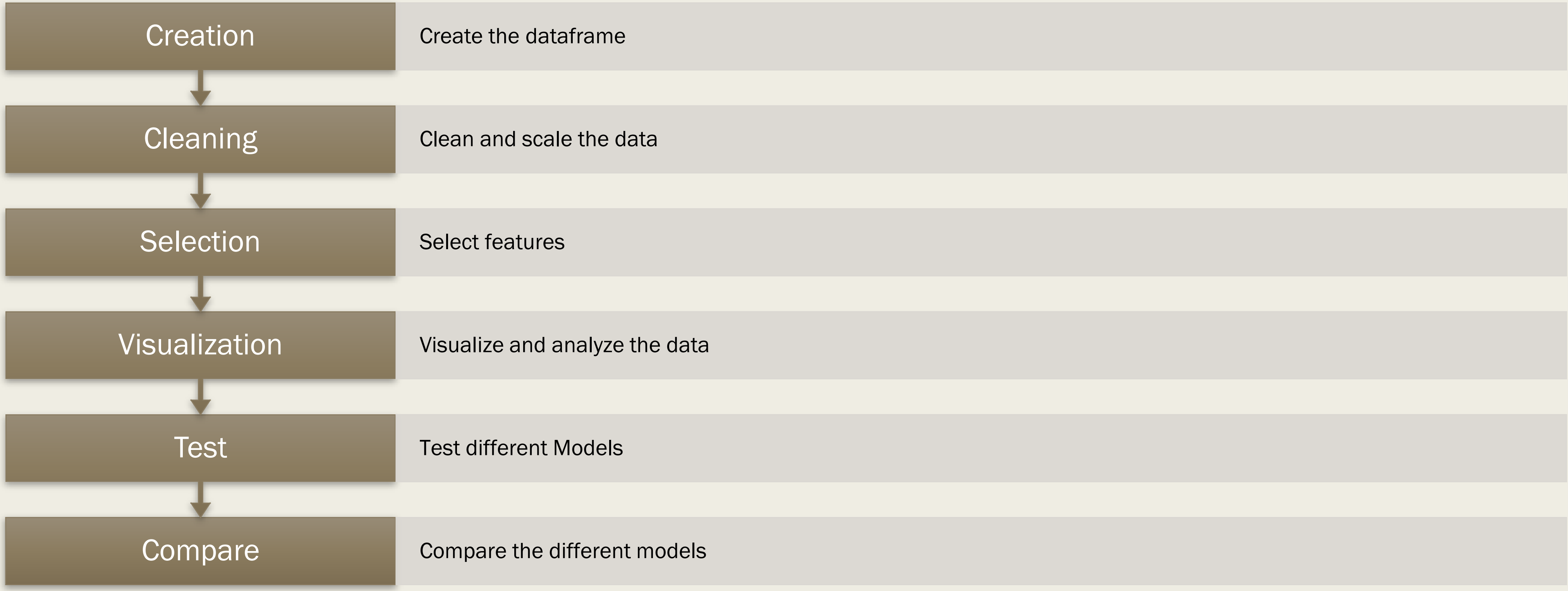
- 39.4% of instances are spam, and we have 4601 instances

The goal of this project

This dataset enables us to build a model which will be able to determine whether a mail is a spam or not.

It is useful in our everyday life, since we receive many e-mails every day.

Stages of the project



Creation — Create the dataframe

Cleaning — Clean and scale the data

Selection — Select features

Visualization — Visualize and analyze the data

Test — Test different Models

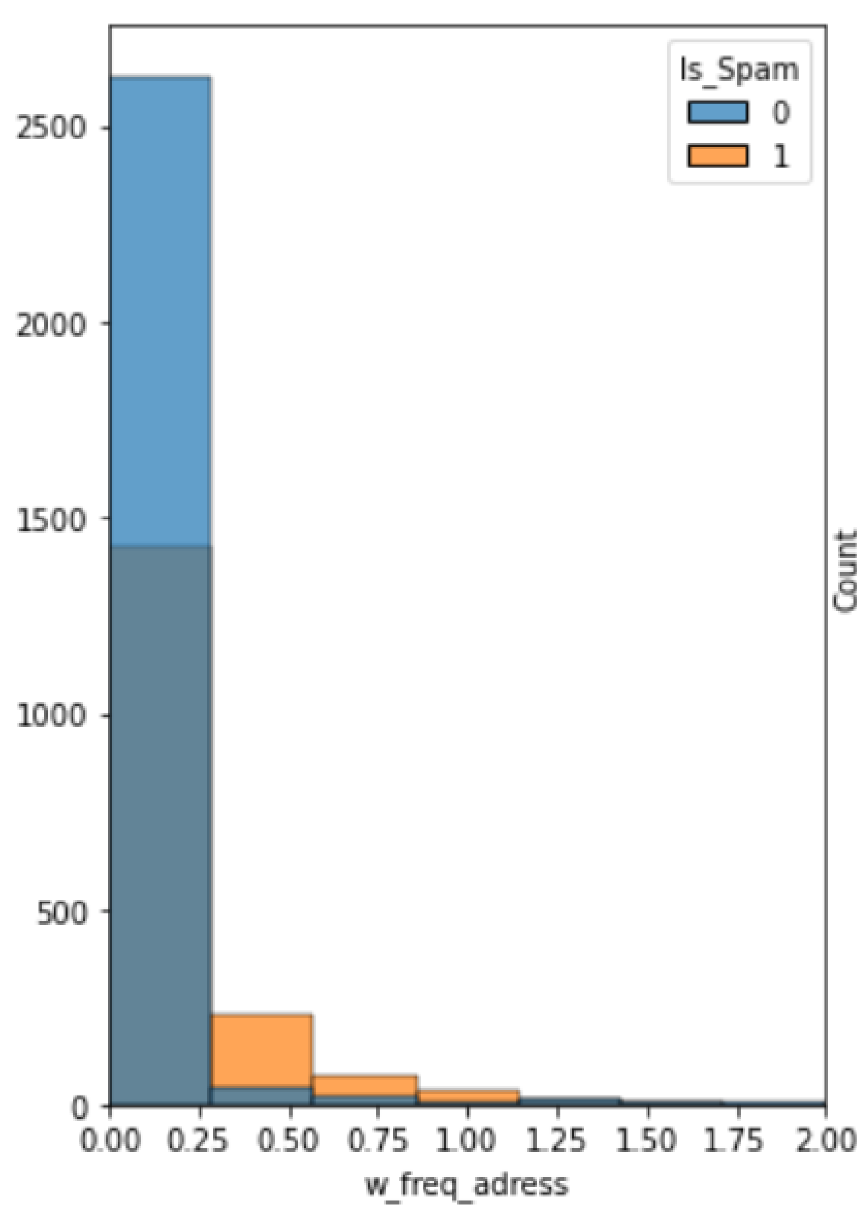
Compare — Compare the different models

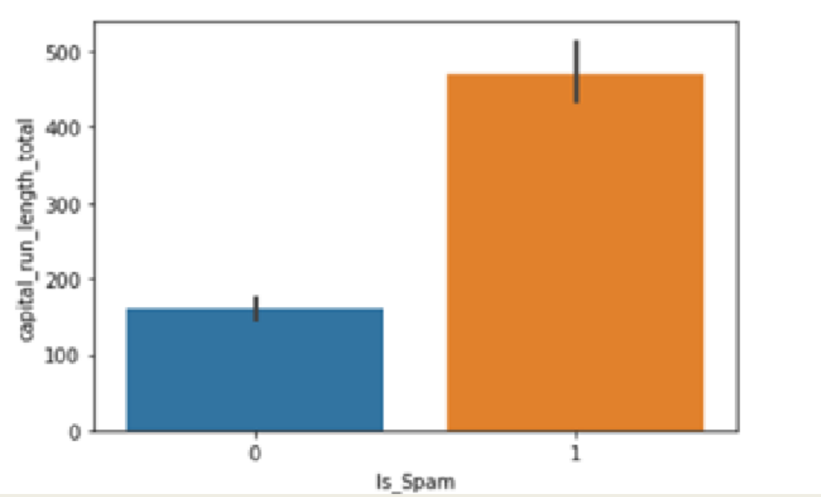
Our first impressions about the data

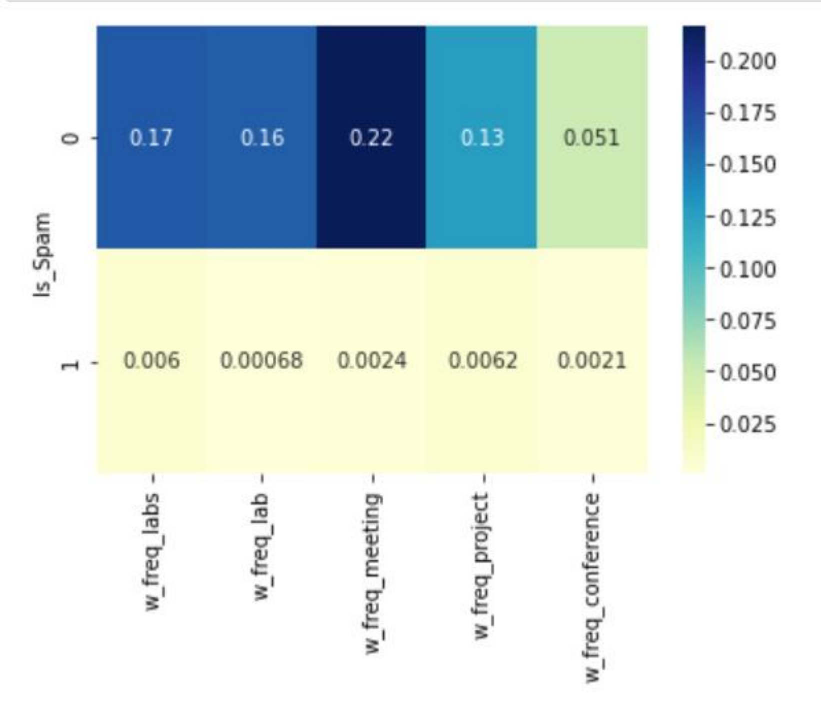
- At first, we assume that the frequency of some words used in the professional world will be greater in non-spam e-mails.
- We also assume that on the other hand, some characters like « ! » for example are used a lot in spams.
- We also guess that in spams, the number of capital letters is greater than in non-spam e-mails.

- To be sure of our thoughts, we will analyze and visualize the data.

Let’s have a look at some of our plots







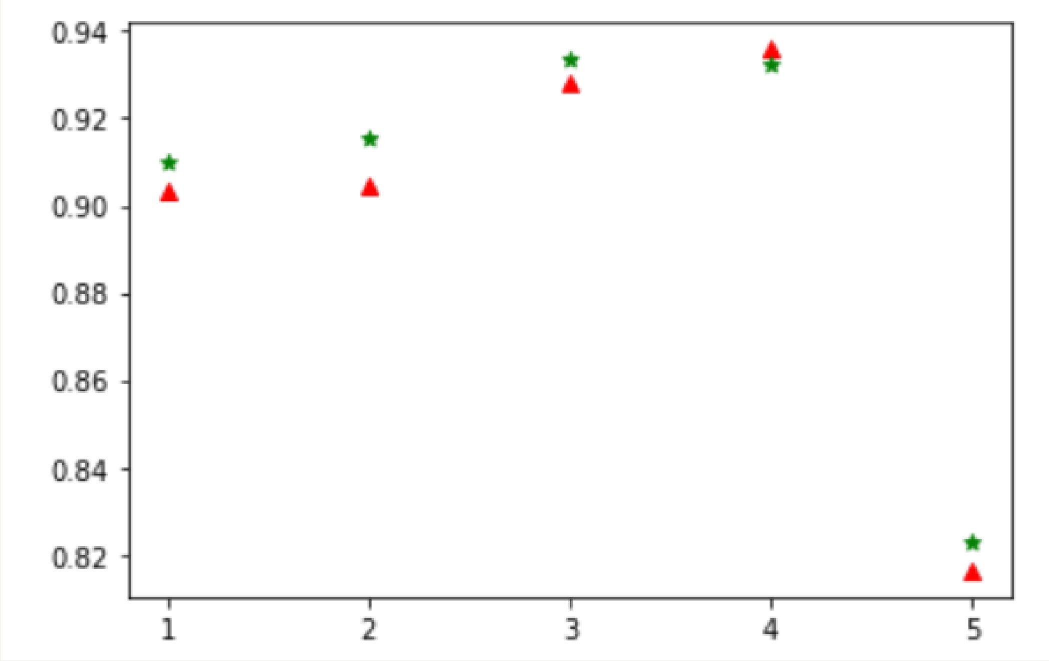
With these plots,

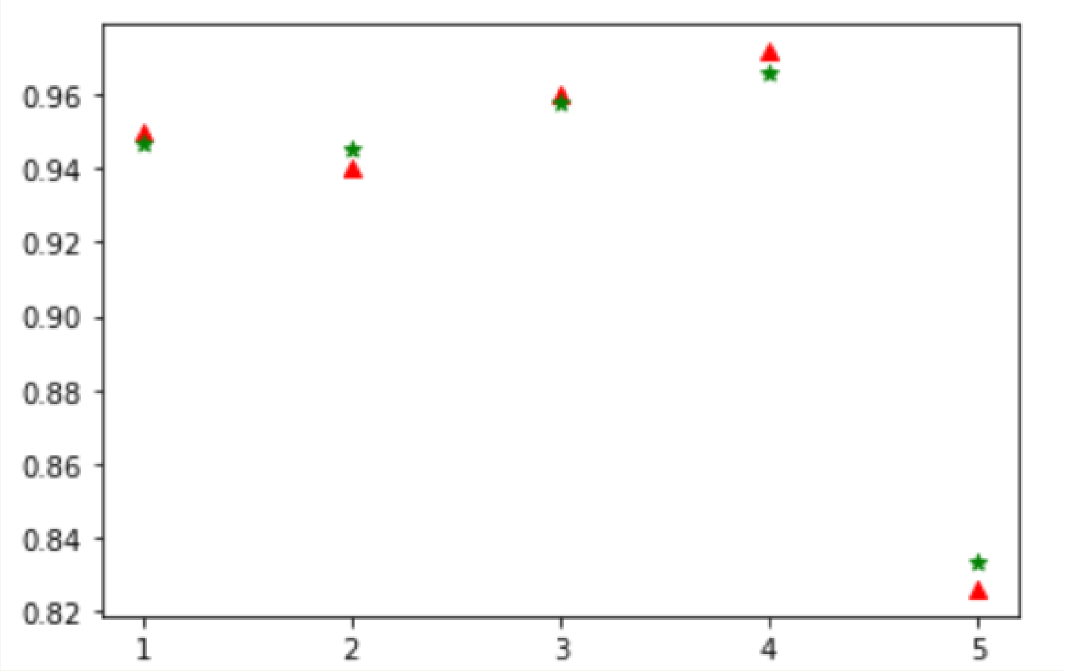
we can say that

our assumptions

were correct.

We can find out if our models are better when using all the features or only the best features

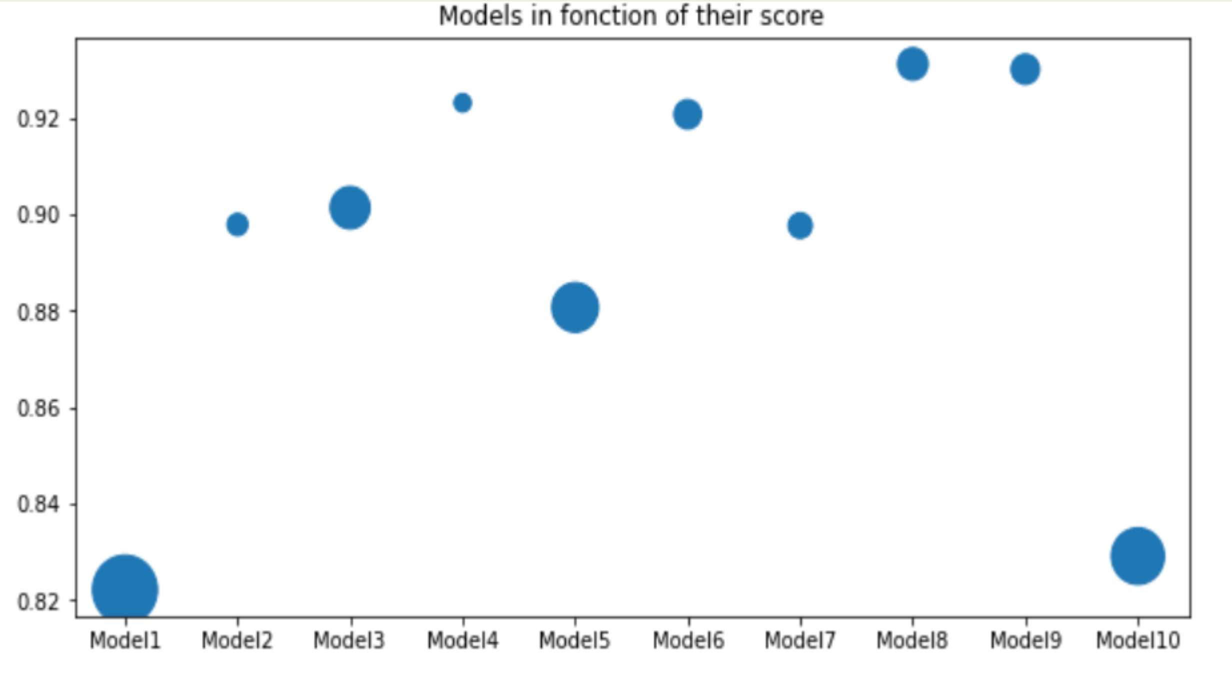




Now let’s compare the different models

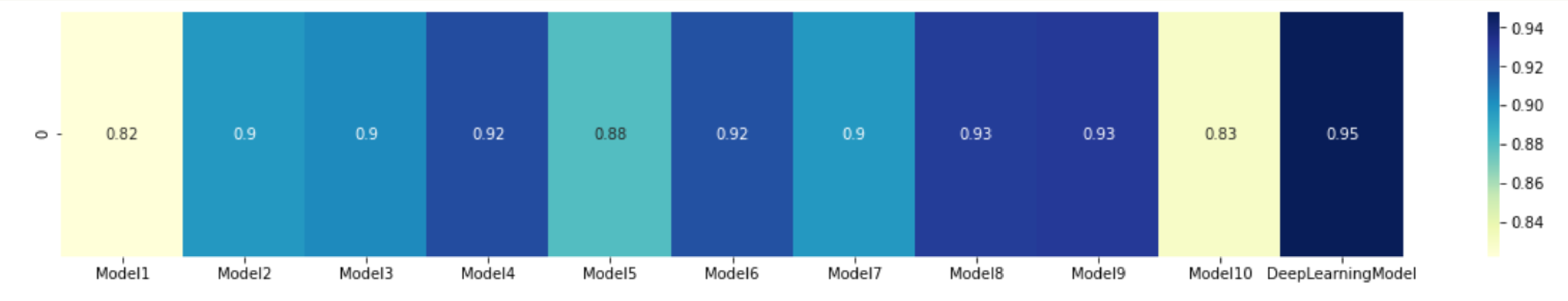
- **Model 1 : Gaussian Naive Bayes :** 0.8222
- **Model 2 : Benouilli Naive Bayes :** 0.8978
- **Model 3 : DecisionTreeClassifier :** 0.9028
- **Model 4 : Support Vector Machine Algorithm :** 0.9230
- **Model 5 : KNN :** 0.8806
- **Model 6 : Multi-layer perceptron :** 0.9213
- **Model 7 : Gaussian Process Classifier :** 0.8976
- **Model 8 : Random Forest Classifier :** 0.9284
- **Model 9 : Ada Boost Classifier :** 0.9300
- **Model 10 : Quadratic Discriminant analysis :** 0.8291
- Model 11 : **Deep Learning :** 0.9480

These percentages are obtained using the cross-validation method.



We can see that our best model is the deep learning model.

Among the other models, the two best models are the Ada Boos Classifier (a tree) and the Quadratic Discriminant Analysis.


Model1
Model2
Model3
Model4
Model5
Model6
Model7
Model8
Model9
Model10
DeepLearningModel