

Data Mining project - part I

Katarzyna Macioszek, Ada Majchrzak

November 10, 2023

1 Introduction

In this project, we will use data mining methods to perform analysis of *Spambase* dataset. The considered dataset was created from two collection of e-mails, the first being a set of non-spam personal and professional e-mails, and the second one one coming from individuals who had filed spam. The e-mails were analyzed in terms of frequency of appearance of 54 selected words and characters, and also in terms of the number of capital letters present. Those statistics were then collected into table and each record was labeled as spam or non-spam. Our goal is to first extract interesting and insightful data characteristics by Exploratory Data Analysis procedures, and then to perform a classification task, distinguishing spam from non-spam. For the second part, we will compare the performance of several methods, namely KNN, Random Forest, **DOPISAĆ PÓŹNIEJ JAKIE KLASYFIKATORY**. The benefit from this kind of study could be obtaining a model effective in terms of spam filtering. We know that spam e-mails are a very common problem in modern world, and although every mailbox nowadays has built-in filter for that kind of content, non of them are completely accurate, so it is certainly a good idea to keep studying this topic.

2 Exploratory Data Analysis

The dataset consist of 4601 records, none of which contain any missing values. There are 57 numerical features and one categorical, *type*, being a target class. The target class takes two values - spam and nonspam. As we can see from Figure 1, the classes are quite well balanced. Around 60% of the records fall into nonspam category, and about 40% of them are spam. As for the numerical features, their summary statistics can be found in Table 1

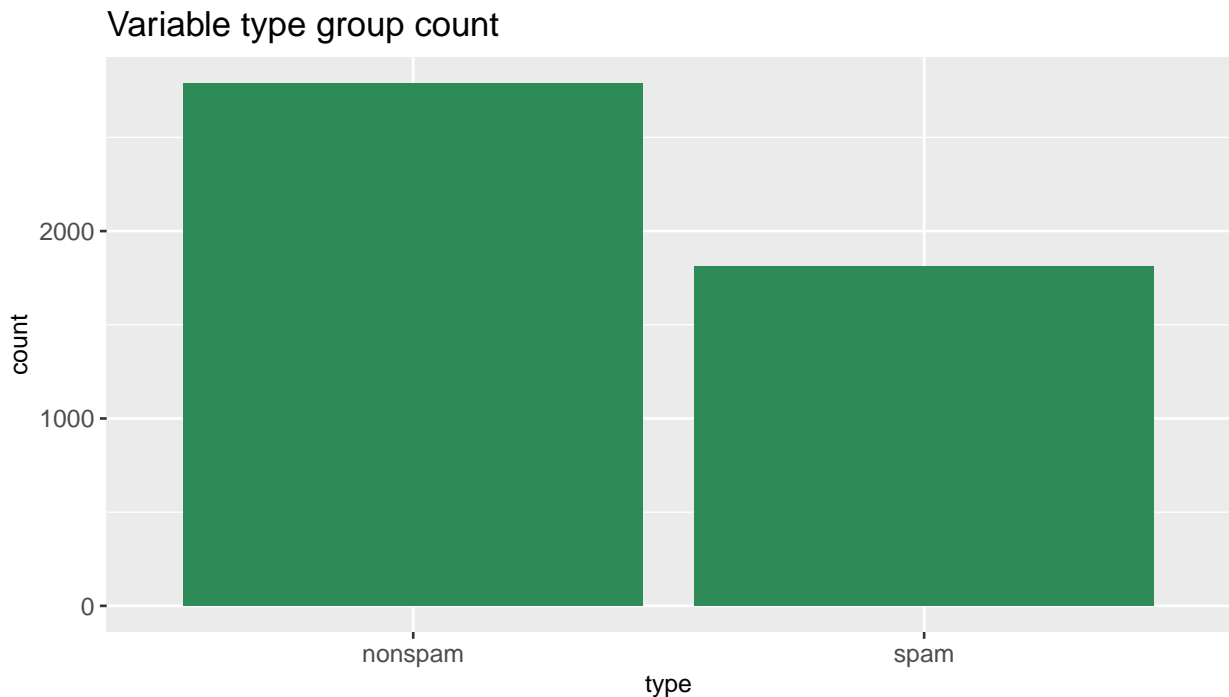


Figure 1: Target class distribution

	Mean	Std.Dev	Min	Max	Q1	Median	Q3
address	0.21	1.29	0.00	14.28	0.00	0.00	0.00
addresses	0.05	0.26	0.00	4.41	0.00	0.00	0.00
all	0.28	0.50	0.00	5.10	0.00	0.00	0.42
business	0.14	0.44	0.00	7.14	0.00	0.00	0.00
capitalAve	5.19	31.73	1.00	1102.50	1.59	2.28	3.71
capitalLong	52.17	194.89	1.00	9989.00	6.00	15.00	43.00
capitalTotal	283.29	606.35	1.00	15841.00	35.00	95.00	266.00
charDollar	0.08	0.25	0.00	6.00	0.00	0.00	0.05
charExclamation	0.27	0.82	0.00	32.48	0.00	0.00	0.32
charHash	0.04	0.43	0.00	19.83	0.00	0.00	0.00
charRoundbracket	0.14	0.27	0.00	9.75	0.00	0.06	0.19
charSemicolon	0.04	0.24	0.00	4.38	0.00	0.00	0.00
charSquarebracket	0.02	0.11	0.00	4.08	0.00	0.00	0.00
conference	0.03	0.29	0.00	10.00	0.00	0.00	0.00
credit	0.09	0.51	0.00	18.18	0.00	0.00	0.00
cs	0.04	0.36	0.00	7.14	0.00	0.00	0.00
data	0.10	0.56	0.00	18.18	0.00	0.00	0.00
direct	0.06	0.35	0.00	4.76	0.00	0.00	0.00
edu	0.18	0.91	0.00	22.05	0.00	0.00	0.00
email	0.18	0.53	0.00	9.09	0.00	0.00	0.00
font	0.12	1.03	0.00	17.10	0.00	0.00	0.00
free	0.25	0.83	0.00	20.00	0.00	0.00	0.10
george	0.77	3.37	0.00	33.33	0.00	0.00	0.00
hp	0.55	1.67	0.00	20.83	0.00	0.00	0.00
hpl	0.27	0.89	0.00	16.66	0.00	0.00	0.00
internet	0.11	0.40	0.00	11.11	0.00	0.00	0.00
lab	0.10	0.59	0.00	14.28	0.00	0.00	0.00
labs	0.10	0.46	0.00	5.88	0.00	0.00	0.00
mail	0.24	0.64	0.00	18.18	0.00	0.00	0.16
make	0.10	0.31	0.00	4.54	0.00	0.00	0.00
meeting	0.13	0.77	0.00	14.28	0.00	0.00	0.00
money	0.09	0.44	0.00	12.50	0.00	0.00	0.00
num000	0.10	0.35	0.00	5.45	0.00	0.00	0.00
num1999	0.14	0.42	0.00	6.89	0.00	0.00	0.00
num3d	0.07	1.40	0.00	42.81	0.00	0.00	0.00
num415	0.05	0.33	0.00	4.76	0.00	0.00	0.00
num650	0.12	0.54	0.00	9.09	0.00	0.00	0.00
num85	0.11	0.53	0.00	20.00	0.00	0.00	0.00
num857	0.05	0.33	0.00	4.76	0.00	0.00	0.00
order	0.09	0.28	0.00	5.26	0.00	0.00	0.00
original	0.05	0.22	0.00	3.57	0.00	0.00	0.00
our	0.31	0.67	0.00	10.00	0.00	0.00	0.38
over	0.10	0.27	0.00	5.88	0.00	0.00	0.00
parts	0.01	0.22	0.00	8.33	0.00	0.00	0.00
people	0.09	0.30	0.00	5.55	0.00	0.00	0.00
pm	0.08	0.43	0.00	11.11	0.00	0.00	0.00
project	0.08	0.62	0.00	20.00	0.00	0.00	0.00
re	0.30	1.01	0.00	21.42	0.00	0.00	0.11
receive	0.06	0.20	0.00	2.61	0.00	0.00	0.00
remove	0.11	0.39	0.00	7.27	0.00	0.00	0.00
report	0.06	0.34	0.00	10.00	0.00	0.00	0.00
table	0.01	0.08	0.00	2.17	0.00	0.00	0.00
technology	0.10	0.40	0.00	7.69	0.00	0.00	0.00
telnet	0.06	0.40	0.00	12.50	0.00	0.00	0.00
will	0.54	0.86	0.00	9.67	0.00	0.10	0.80
you	1.66	1.78	0.00	18.75	0.00	1.31	2.64
your	0.81	1.20	0.00	11.11	0.00	0.22	1.27

Table 1: Summary statistics for numerical features