Exercise 0 - Dataset description

October 20, 2023

Team members: Ahmed Sabanovic (12330648), Ivan Babiy (12142160), Yat Hin Chan (12331419)

1 Choice of data sets

The datasets Spambase and Nursery were chosen due to the contrast in their data structures. Spambase has has higher dimensions (4601x57) while Nursery has a taller shape (12960x8); and Spamebase has continuous numeric features while Nursery has categorical features.

2 Spambase

This dataset incldues 4601 emails, each with 57 features and 1 target variable. 54 of the features contain the relative frequencies of different words and characters, and 3 features are some summary statistics of the email content. The target attribute, `Class`, denotes whether an email is a spam.

2.1 Attribute details

Feature	Feature type	Data type	Range	Description
word_freq_WORD (x48)	Ratio	Real	[0, 42.81] (Max: 100)	Percentage of words in the e-mail that match WORD
char_freq_CHAR (x6)	Ratio	Real	[0, 32.478] (Max: 100)	Percentage of characters in the e-mail that match CHAR
capital_run_length_average	Ratio	Real	[1, 1102.5]	Average length of uninterrupted sequences of capital letters
capital_run_length_longest	Ratio	Integer	[1, 9989]	Length of longest uninterrupted sequence of capital letters
capital_run_length_total	Ratio	Integer	[1, 15841]	Total number of capital letters in the e-mail
Class (target)	Nominal	Binary	{0, 1}	Whether the e-mail was considered spam

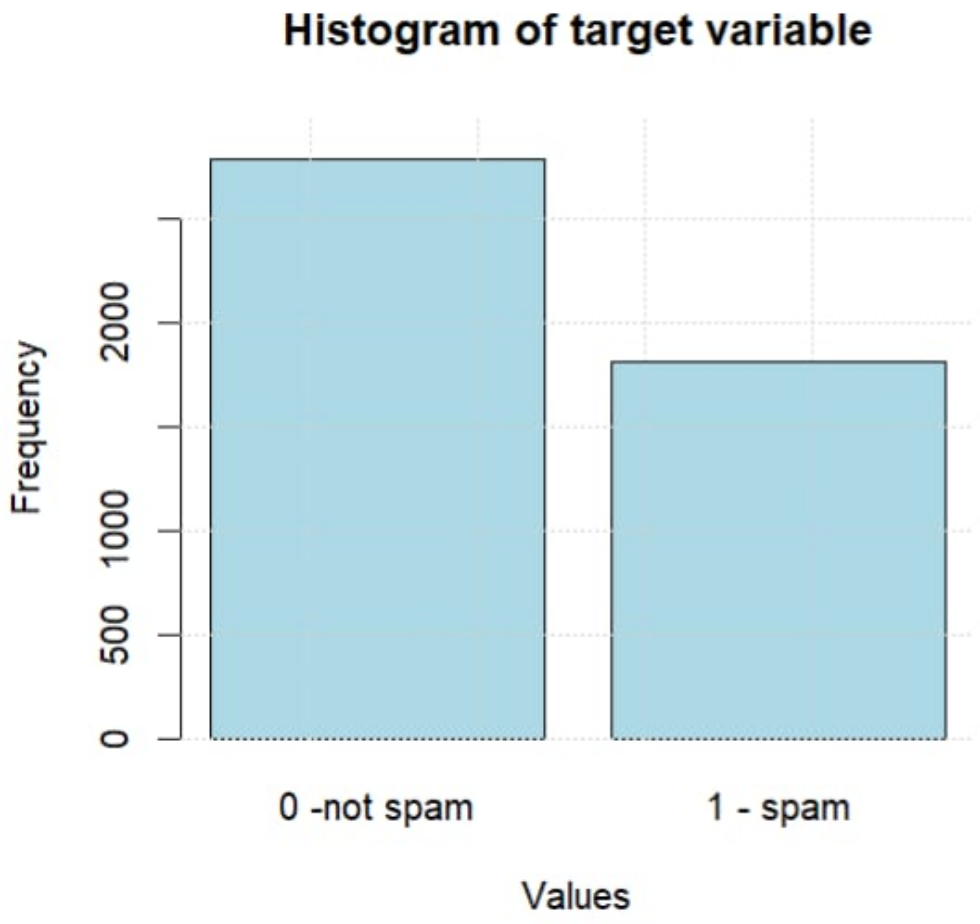
2.2 Potential pre-precossing steps

The dataset has no missing values, so there is no need for missing values handling. All frequency features are also pre-normalised into relative frequencies in each letter, and have desirable distributions.

However, based on the summary statistics of the capital_run_length attributes, they appear to be heavily skewed. This may warrant the application of log transformation.

2.3 Distributions

This is the distribution of the target class



3 Nursery

The nursery dataset was derived from a hierarchical decision model that ranks nursery-school applications for nursery schools. It has 12960 rows, and 9 columns.

3.1 Attribute details

Feature	Feature type	Data type	Range	Description
parents	Ordinal	string	[usual, pretentious, great_pret]	Parents' occupation
has_nurs	Ordinal	string	[proper, less_proper, improper, critical, very_crit]	Child's nursery
form	Ordinal	string	[complete, completed, incomplete, foster]	Form of the family
children	Ordinal	string	[1, 2, 3, more]	Number of children
housing	Ordinal	string	[convenient, less_conv, critical]	Housing conditions
finance	Ordinal	string	[convenient, inconv]	Financial standing of the family
social	Ordinal	string	[non-prob, slightly_prob, problematic]	Social conditions
health	Ordinal	string	[recommended, priority, not_recom]	Health conditions
class (Target)	Ordinal	string	[not_recom, recommend, very_recom, priority, spec_prior]	Evaluation of applications for nursery schools

There are 8 feature attributes, they are parents, has_nurs, form, children, housing, finance, social, and health. All of the features are of ordinal type.

The target feature is class, it is of ordinal type and takes the following values (in ascending order): not_recom, recommend, very_recom, priority, spec_prior.

3.2 Potential pre-precossing steps

There are no missing values, nor outliers in the dataset. However, there are only 2 instances (out of 12960) of the value *recommend* for the target attribute.

In order to satisfy the pre conditions of some of the models that we expect to use, and since all of the features are of categorical type, it will be necessary to transform them into dummy/binary variables. However, this may lead to a dataset with too sparse data, which may lead us to merge categories.

3.3 Distribution

This is the distribution of the target class. As mentioned, there are only 2 **recommended** instances, which may make this class problematic.

class	N	N[%]
not_recom	4320	33.333 %
recommend	2	0.015 %
very_recom	328	2.531 %
priority	4266	32.917 %
spec_prior	4044	31.204 %