HOMEWORK 3

DENNIS MULUMBI KYALO

22nd October 2021

1. INTRODUCTION

We begin by loading the dataset and skimming through it to have a better understanding of it.

Table 1: Data summary

Name	spam_set
Number of rows	4601
Number of columns	58
Column type frequency:	
factor	1
numeric	57
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
type	0	1	FALSE	2	non: 2788, spa: 1813

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
make	0	1	0.10	0.31	0	0.00	0.00	0.00	4.54
address	0	1	0.21	1.29	0	0.00	0.00	0.00	14.28
all	0	1	0.28	0.50	0	0.00	0.00	0.42	5.10
num3d	0	1	0.07	1.40	0	0.00	0.00	0.00	42.81
our	0	1	0.31	0.67	0	0.00	0.00	0.38	10.00
over	0	1	0.10	0.27	0	0.00	0.00	0.00	5.88
remove	0	1	0.11	0.39	0	0.00	0.00	0.00	7.27
internet	0	1	0.11	0.40	0	0.00	0.00	0.00	11.11
order	0	1	0.09	0.28	0	0.00	0.00	0.00	5.26
mail	0	1	0.24	0.64	0	0.00	0.00	0.16	18.18
receive	0	1	0.06	0.20	0	0.00	0.00	0.00	2.61
will	0	1	0.54	0.86	0	0.00	0.10	0.80	9.67
people	0	1	0.09	0.30	0	0.00	0.00	0.00	5.55
report	0	1	0.06	0.34	0	0.00	0.00	0.00	10.00
addresses	0	1	0.05	0.26	0	0.00	0.00	0.00	4.41
free	0	1	0.25	0.83	0	0.00	0.00	0.10	20.00
business	0	1	0.14	0.44	0	0.00	0.00	0.00	7.14

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
email	0	1	0.18	0.53	0	0.00	0.00	0.00	9.09
you	0	1	1.66	1.78	0	0.00	1.31	2.64	18.75
credit	0	1	0.09	0.51	0	0.00	0.00	0.00	18.18
your	0	1	0.81	1.20	0	0.00	0.22	1.27	11.11
font	0	1	0.12	1.03	0	0.00	0.00	0.00	17.10
num000	0	1	0.10	0.35	0	0.00	0.00	0.00	5.45
money	0	1	0.09	0.44	0	0.00	0.00	0.00	12.50
hp	0	1	0.55	1.67	0	0.00	0.00	0.00	20.83
hpl	0	1	0.27	0.89	0	0.00	0.00	0.00	16.66
george	0	1	0.77	3.37	0	0.00	0.00	0.00	33.33
num650	0	1	0.12	0.54	0	0.00	0.00	0.00	9.09
lab	0	1	0.10	0.59	0	0.00	0.00	0.00	14.28
labs	0	1	0.10	0.46	0	0.00	0.00	0.00	5.88
telnet	0	1	0.06	0.40	0	0.00	0.00	0.00	12.50
num857	0	1	0.05	0.33	0	0.00	0.00	0.00	4.76
data	0	1	0.10	0.56	0	0.00	0.00	0.00	18.18
num415	0	1	0.05	0.33	0	0.00	0.00	0.00	4.76
num85	0	1	0.11	0.53	0	0.00	0.00	0.00	20.00
technology	0	1	0.10	0.40	0	0.00	0.00	0.00	7.69
num1999	0	1	0.14	0.42	0	0.00	0.00	0.00	6.89
parts	0	1	0.01	0.22	0	0.00	0.00	0.00	8.33
pm	0	1	0.08	0.43	0	0.00	0.00	0.00	11.11
direct	0	1	0.06	0.35	0	0.00	0.00	0.00	4.76
cs	0	1	0.04	0.36	0	0.00	0.00	0.00	7.14
meeting	0	1	0.13	0.77	0	0.00	0.00	0.00	14.28
original	0	1	0.05	0.22	0	0.00	0.00	0.00	3.57
project	0	1	0.08	0.62	0	0.00	0.00	0.00	20.00
re	0	1	0.30	1.01	0	0.00	0.00	0.11	21.42
edu	0	1	0.18	0.91	0	0.00	0.00	0.00	22.05
table	0	1	0.01	0.08	0	0.00	0.00	0.00	2.17
conference	0	1	0.03	0.29	0	0.00	0.00	0.00	10.00
charSemicolon	0	1	0.04	0.24	0	0.00	0.00	0.00	4.38
charRoundbracket	0	1	0.14	0.27	0	0.00	0.06	0.19	9.75
charSquarebracket	0	1	0.02	0.11	0	0.00	0.00	0.00	4.08
charExclamation	0	1	0.27	0.82	0	0.00	0.00	0.32	32.48
charDollar	0	1	0.08	0.25	0	0.00	0.00	0.05	6.00
charHash	0	1	0.04	0.43	0	0.00	0.00	0.00	19.83
capitalAve	0	1	5.19	31.73	1	1.59	2.28	3.71	1102.50
capitalLong	0	1	52.17	194.89	1	6.00	15.00	43.00	9989.00
capitalTotal	0	1	283.29	606.35	1	35.00	95.00	266.00	15841.00

From the above spam dataset, there are 4601 samples and 58 variables. Out of the 58 variables, all but type is a factor variable with spam and non-spam samples. The rest 57 predictors are all numeric.

2. ANALYSIS AND RESULTS

2.1 FEATURE SELECTION

Next, we set up a spam detection classification model, which should only use five predictor variables from the current 57. We first set the following null and alternative hypotheses to select the top fifteen most statistically significant predictors that highly predict the type variable.

H0: There is no relationship between the independent variable and the dependent variable.
H1: There is a relationship between the independent variable and the dependent variable.

We shall test this hypothesis using a 1% significance level, in which we shall reject the null hypothesis when the p-value is less than the level of significance.Now we apply the logistic regression algorithm to predict the type variable (dependent variable) as either spam or non-spam based on the 57 predictors (independent predictors).

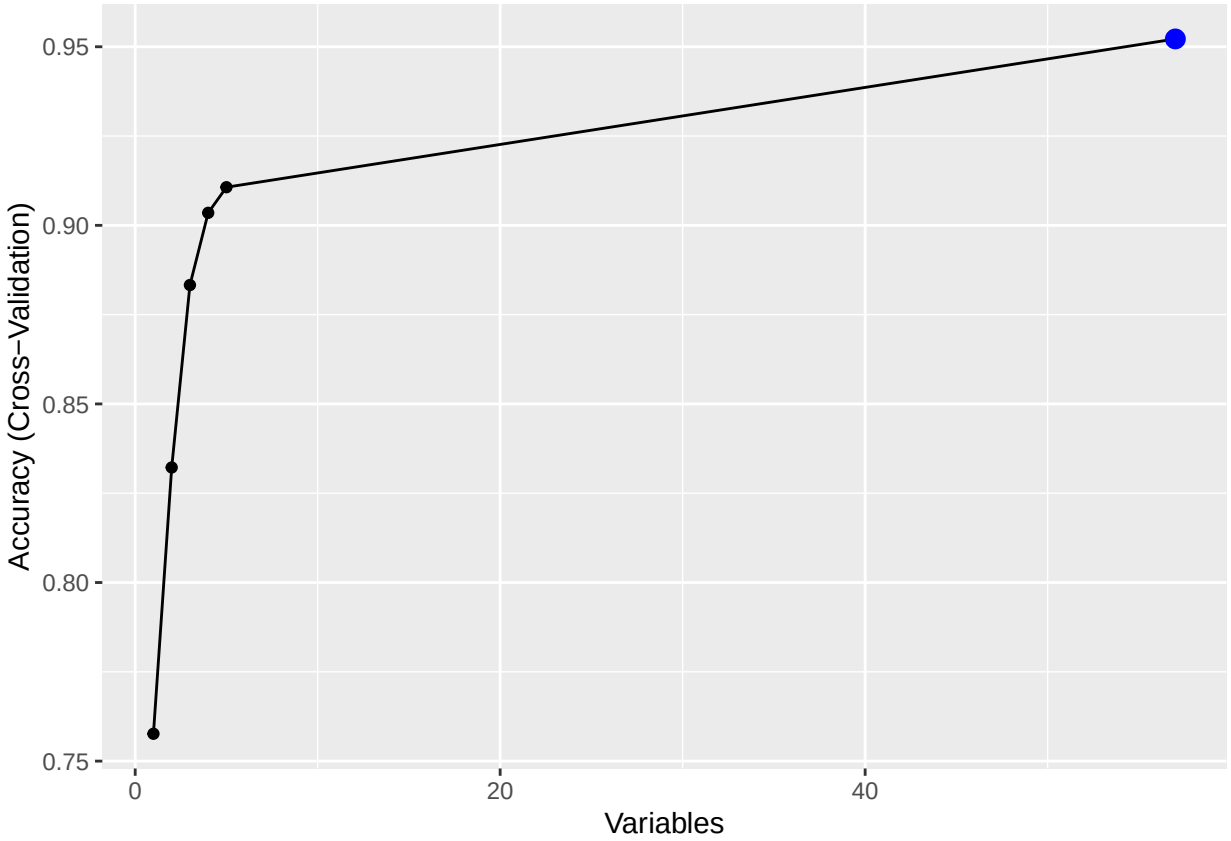
Table 4: Coefficients

	Estimate	Std. Error	z value	p-value
(Intercept)	-1.5686144	0.1420362	-11.043767	2.349719e-28
charDollar	5.3360174	0.7064366	7.553427	4.239529e-14
free	1.0385899	0.1456954	7.128500	1.014690e-12
remove	2.2785173	0.3328051	6.846401	7.573083e-12
hp	-1.9204165	0.3128278	-6.138894	8.309803e-10
george	-11.7671895	2.1131288	-5.568610	2.567799e-08
our	0.5623844	0.1017997	5.524423	3.305708e-08
edu	-1.4592476	0.2685620	-5.433559	5.524113e-08
re	-0.7923358	0.1556316	-5.091100	3.559920e-07
num000	2.2452367	0.4714440	4.762468	1.912399e-06
your	0.2419341	0.0524285	4.614557	3.939349e-06
business	0.9598609	0.2250906	4.264332	2.005014e-05
charExclamation	0.3471968	0.0892645	3.889528	1.004393e-04
capitalTotal	0.0008437	0.0002251	3.747440	1.786489e-04
capitalLong	0.0091185	0.0025206	3.617610	2.973358e-04

In table 4, above, we have chosen the best fifteen statistically significant variables in a decreasing order based on the p-value that meet the elimination criteria. Next, we subject these variables to recursive feature selection to further narrow the variables by identifying the five best predictors significant in our model.

```
Recursive feature selection

Outer resampling method: Cross-Validated (5 fold)

Resampling performance over subset size:

 Variables Accuracy  Kappa AccuracySD KappaSD Selected
         1   0.7577 0.4749   0.013312 0.03091
         2   0.8322 0.6411   0.007051 0.01706
         3   0.8833 0.7511   0.014284 0.03276
         4   0.9035 0.7953   0.012313 0.02645
         5   0.9107 0.8097   0.012385 0.02682
        57   0.9522 0.8992   0.009741 0.02081        *

The top 5 variables (out of 57):
   charExclamation, remove, charDollar, capitalAve, hp
```

From the above recursive feature selection analysis, we narrow down to the top five variables selected, which are: charExclamation, remove, capitalAve, charDollar, and hp.



The recursive feature selection plot shows the selected variables and their accuracies.

2.2. CLASSIFICATION

We now use the selected predictors in our spam detection algorithm by applying the following classification models: Logistic Regression, CART, Bagging, Boosting, Random Forest, K-NN, and Naive Bayes. After that, we compare them using the classification metrics; accuracy, sensitivity, and specificity when deployed on both the train and unseen test set.

Table 5: Classification Metrics on the Train Set

	Accuracy	AccuracyLower	AccuracyUpper	Sensitivity	Specificity
Boosting	0.9243478	0.9075182	0.9389650	0.9569584	0.8741722
K-NN	0.8982609	0.8793228	0.9151317	0.9397418	0.8344371
Logistic Regression	0.8930435	0.8737305	0.9103176	0.9583931	0.7924945
Naive Bayes	0.9147826	0.8971296	0.9302771	0.9713056	0.8278146
CART	0.8547826	0.8330792	0.8746525	0.9440459	0.7174393
Bagging	0.9886957	0.9807468	0.9939676	0.9956958	0.9779249
Random Forest	0.9704348	0.9589290	0.9794402	0.9971306	0.9293598

Table 5 shows the classification metrics when the algorithms are deployed on the train set. This helps us understand how well the models performed on the train set. The table also adds the 95% confidence interval of the accuracies to help us understand how generalizable the models performed on the unseen test set. Bagging performed really well on the train set with an accuracy of 98.86957%, while the CART method performed the least with an accuracy of 85.47826%.

Table 6: Classification Metrics on the Test Set

	Accuracy	Sensitivity	Specificity
Boosting	0.9156522	0.9526542	0.8587196
K-NN	0.8921739	0.9368723	0.8233996
Logistic Regression	0.8747826	0.9698709	0.7284768
Naive Bayes	0.8765217	0.9827834	0.7130243
CART	0.8147826	0.9397418	0.6225166
Bagging	0.8913043	0.9225251	0.8432671
Random Forest	0.9104348	0.9555237	0.8410596

Finally, we tested the models on the unseen test set, and the above table shows a clear summary of the results. Based on the 95% confidence interval of the accuracy from the train set, we can confidently conclude that Boosting, K-NN, and Logistic regression are generalizable as their test accuracies fall within the 95% confidence interval. Naive Bayes, CART, Bagging, and Random Forest did well on the train set, but their test accuracies do not lie within the 95% confidence interval of the train set; thus, we categorize them as non-generalizable models.

In summary, from the above-tested models, Boosting model was generalizable and had the highest prediction accuracy of 91.56522%.

3. REFERENCE

An Introduction to Statistical Learning 1st edition, by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, Springer, New York.