FISHING FOR PHISHING SITES

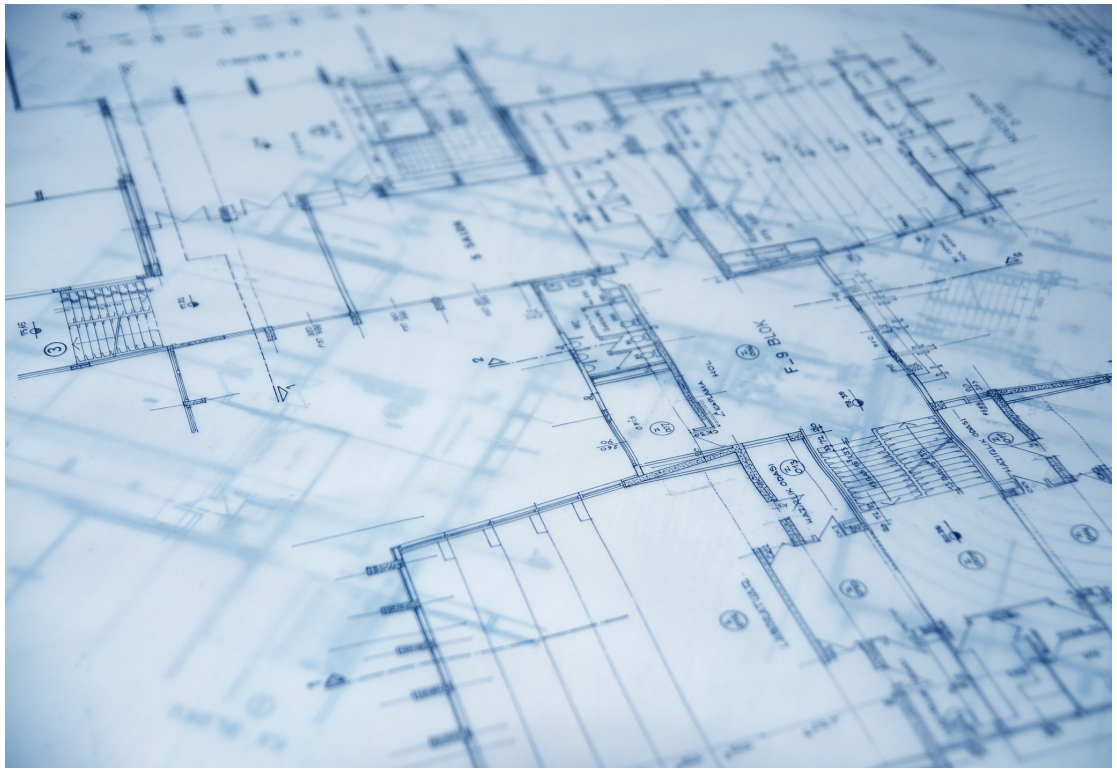
DIL DHALIWAL, EVAN IMTIAZ

INTRODUCTION

- In the age of the internet that we live in today, phishing scams have become increasingly common and problematic
- Thousands of Canadians are victim to phishing attacks yearly, resulting in millions of dollars in losses
- The problem will only continue to worsen unless better measures are put in place to detect and prevent these attacks

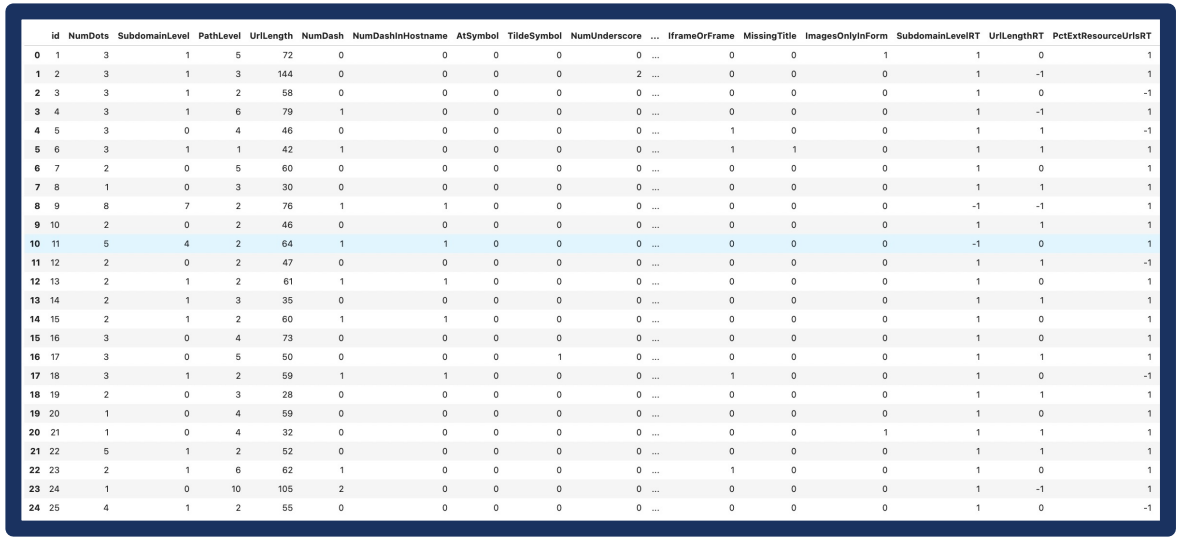
PROJECT DESCRIPTION

- Machine Learning offers a solution to phishing through the creation of classification models
- Parameters can be used to find the best indicators to ensure accuracy in the models
- Multiple models will be used and compared to find the best model for the problem
- Model accuracy will be closely watched to ensure the model is not overfitted to the training data



DATA COLLECTION

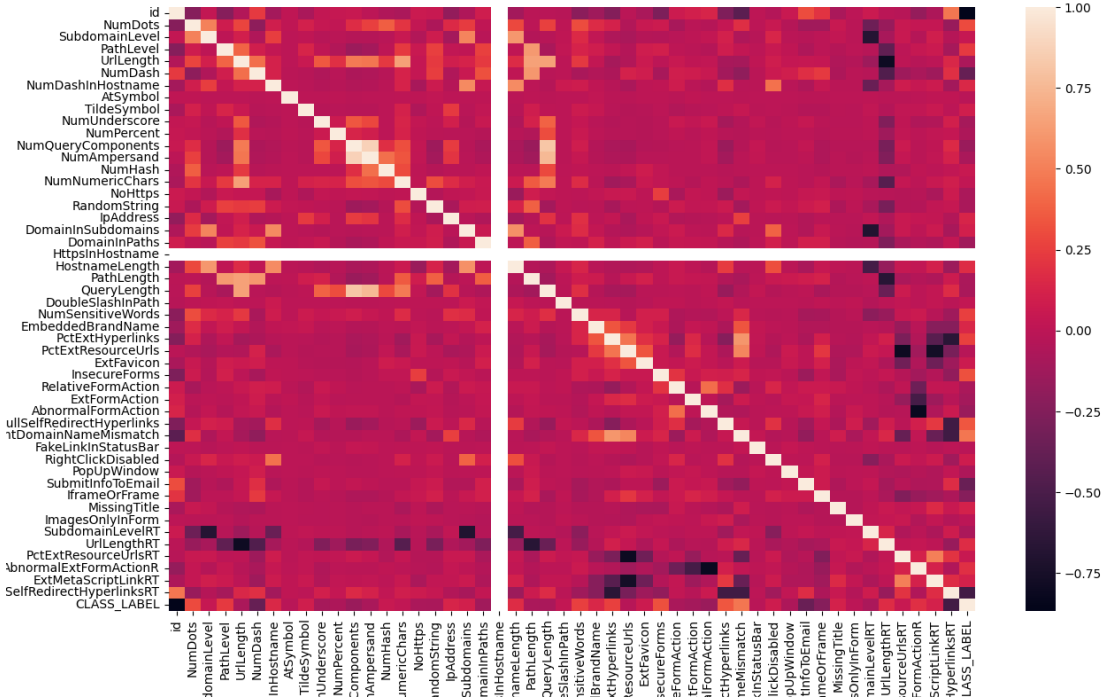
- 10,000 webpages both phishing and legitimate are used in the dataset
- Each webpage comes with 50 features that are used to better predict phishing webpages
- Features list includes webpage information such as if an IP address is provided, path length and number of dashes and dots
- The dataset was retrieved from Kaggle https://www.kaggle.com/datasets/shashwatwork/phishing-dataset-for-machine-learning



	id	NumDots	SubdomainLevel	PathLevel	UrlLength	NumDash	NumDashInHostname	AtSymbol	TildeSymbol	NumUnderscore	...	IframeOrFrame	MissingTitle	ImagesOnlyInForm	SubdomainLevelRT	UrlLengthRT	PctExtResourceUrlsRT
0	1	3	1	5	72	0	0	0	0	0	...	0	0	1	1	0	1
1	2	3	1	3	144	0	0	0	0	2	...	0	0	0	1	-1	1
2	3	3	1	2	58	0	0	0	0	0	...	0	0	0	1	0	-1
3	4	3	1	6	79	1	0	0	0	0	...	0	0	0	1	-1	1
4	5	3	0	4	46	0	0	0	0	0	...	1	0	0	1	1	-1
5	6	3	1	1	42	1	0	0	0	0	...	1	1	0	1	1	1
6	7	2	0	5	60	0	0	0	0	0	...	0	0	0	1	0	1
7	8	1	0	3	30	0	0	0	0	0	...	0	0	0	1	1	1
8	9	8	7	2	76	1	1	0	0	0	...	0	0	0	-1	-1	1
9	10	2	0	2	46	0	0	0	0	0	...	0	0	0	1	1	1
10	11	5	4	2	64	1	1	0	0	0	...	0	0	0	-1	0	1
11	12	2	0	2	47	0	0	0	0	0	...	0	0	0	1	1	-1
12	13	2	1	2	61	1	1	0	0	0	...	0	0	0	1	0	1
13	14	2	1	3	35	0	0	0	0	0	...	0	0	0	1	1	1
14	15	2	1	2	60	1	1	0	0	0	...	0	0	0	1	0	1
15	16	3	0	4	73	0	0	0	0	0	...	0	0	0	1	0	1
16	17	3	0	5	50	0	0	0	1	0	...	0	0	0	1	1	1
17	18	3	1	2	59	1	1	0	0	0	...	1	0	0	1	0	-1
18	19	2	0	3	28	0	0	0	0	0	...	0	0	0	1	1	1
19	20	1	0	4	59	0	0	0	0	0	...	0	0	0	1	0	1
20	21	1	0	4	32	0	0	0	0	0	...	0	0	1	1	1	1
21	22	5	1	2	52	0	0	0	0	0	...	0	0	0	1	1	1
22	23	2	1	6	62	1	0	0	0	0	...	1	0	0	1	0	1
23	24	1	0	10	105	2	0	0	0	0	...	0	0	0	1	-1	1
24	25	4	1	2	55	0	0	0	0	0	...	0	0	0	1	0	-1

EXPLORATORY DATA ANALYSIS

- Finding the target variable in our case it was Class_Label which represents if it is a phishing or legitimate webpage
- Class_Label is set to 0 if it is a legitimate webpage and a 1 if it is a phishing webpage
- A correlation matrix was created to view the correlation between the target variable and predictors
- Many predictors were found to have a weak correlation with only a few showed a strong correlation to the target variable



PRE-PROCESSING

- Removed all null values to ensure data consistency and quality
- Checked for any out of place data values and proper formatting of all values
- Another overview was done to ensure the data integrity
- Vital for machine learning to ensure that the data is properly processed

FEATURE SELECTION

- Only the id variable was removed as it is a redundant variable and only served as an index
- Variables left in even with less correlation to see how models would perform with a complete dataset
- In future renditions of our models' changes will be made to see how it will affect the target variable
- The choice of features is crucial to ensuring the model can produce accurate results and achieve our end goal of classifying webpages

FEATURE SCALING

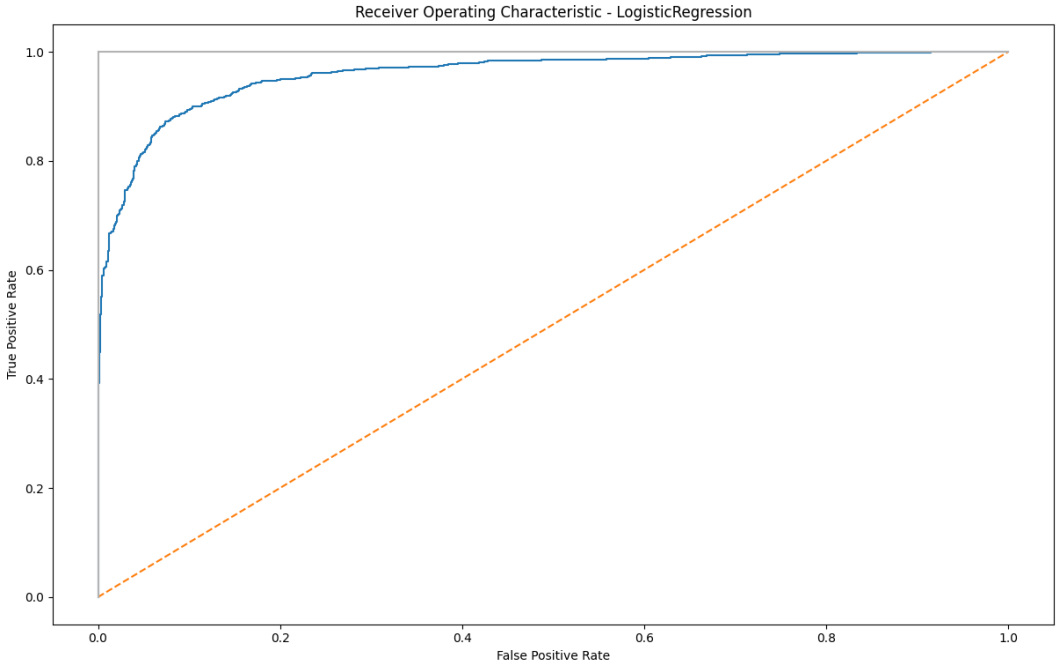
- Due to a lack of regularization of data types values must be scaled and regularized to ensure data quality
- Normalization of data is crucial for working with it to ensure no underlying problems occur later in the development process
- If feature scaling is done incorrectly it leads to problems with model due to the values leading to unexpected side effects in the future

SPLITTING THE DATA

- We decided a 70/30 split for training testing to ensure the data is not overfit
- If the training part of the split is too large, then it will lead to an overfitted dataset as it will be too accustomed to the train part
- The 70/30 split was a compromise between an aggressive train set of 80/20 and a more underfitted model of 60/40

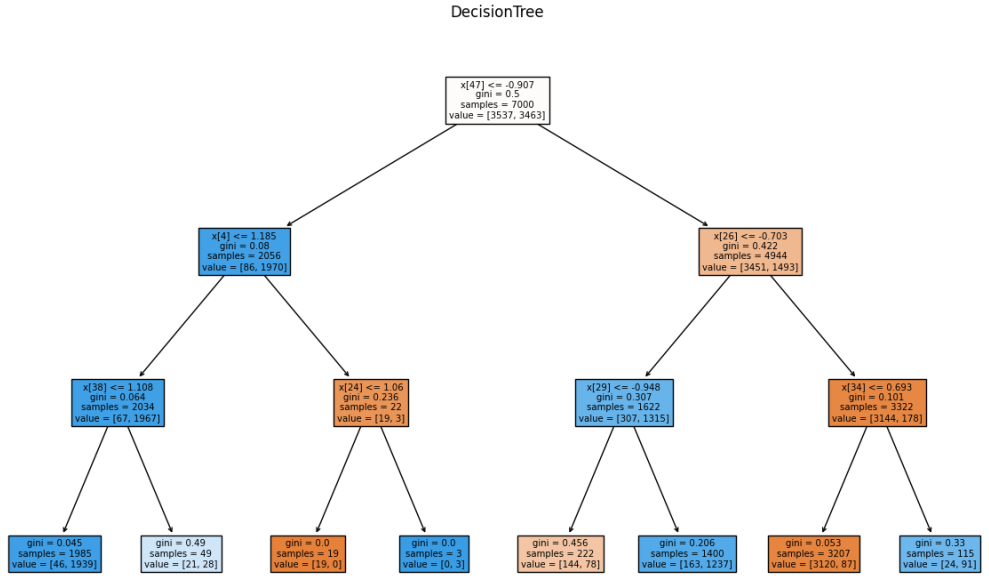
LOGISTIC REGRESSION

- Methods In the class are initialize, fit, predict and metrics
- Fit and predict are used to fit the model and then retrieve the models' predictions.
- For regularization, the choice was made to apply L2 regularization
- To evaluate the model accuracy, predication, recall and f1 were retrieved along with K-fold with 8 splits and 4 repeats



DECISION TREE

- Methods In the class are initialize, fit, predict and metrics
- Fit and predict are used to fit the model and then retrieve the models' predictions.
- For regularization, the choice was made to apply L2 regularization
- To evaluate the model accuracy, predication, recall and f1 were retrieved along with K-fold with 8 splits and 4 repeats



RESULTS

- Model evaluations show that the decision tree was incredibly overfit an accuracy score of over 90%
- Logistic regression was appropriately fit with it having an accuracy of around 85%
- From this we can see that in future iterations of this project logistic regression is a more useful model for the data

	Logistic Regression	Decision Tree
Initial Accuracy	0.836	0.947
Precision	0.864	0.947
Recall	0.839	0.947
F1-Score	0.883	0.947
K-Fold Accuracy	0.884	0.942

CONCLUSION

- After running the models most features are shared between both phishing and legitimate websites
- We rely heavily on very few features to distinguish between the two types of webpages due to their similarities
- Improvements to be made include implementing new models such as Random Forest and comparing against current models to see effectiveness
- Changes to feature selection include dropping weak indicators to see the effect on the overall dataset

REFERENCES

- http://www.citethisforme.com. https://publications.gc.ca/collections/collection_2022/grc-rcmp/PS61-46-2021-eng.pdf

- https://www.kaggle.com/datasets/shashwatwork/phishing-dataset-for-machine-learning

THANK YOU