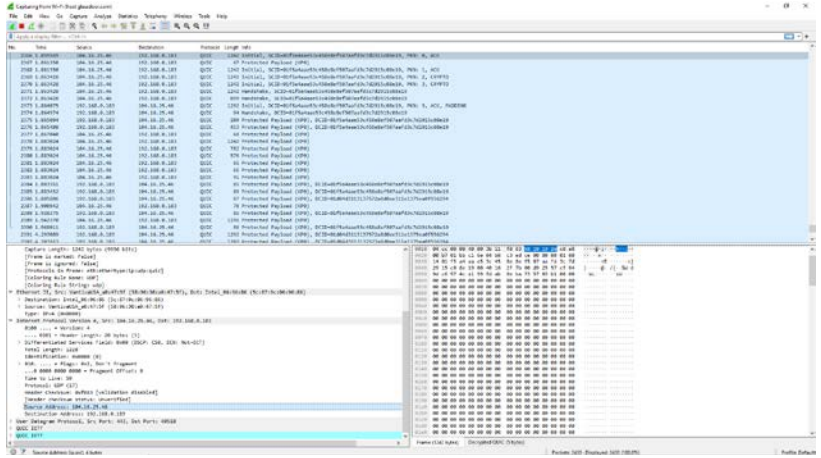
CSCI 434 NETWORK TRAFFIC PROJECT

Tyler Pringle

GOALS

1. Monitor network traffic from popular websites
2. Develop models that extract features from network traffic and automatically classify the origin website of the traffic



Monitor traffic using Wireshark…

…and classify website origin

DATASET

DATASET SAMPLES

Label	Protocol	Length
Glassdoor	UDP	1292
Glassdoor	R-GOOSE	1242
Glassdoor	TLSv1.2	227
X	TCP	66
X	TCP	54
X	TLSv1.3	1414
Stack Overflow	TLSv1.2	381
Stack Overflow	TCP	60
Stack Overflow	TLSv1.2	85
…	…	…

DATASET SIZE

Label	Size
Glassdoor	1834
X	2138
Stack Overflow	2041
Total	**6013**

EXTRACTED FEATURES

Feature	Details
Protocol	The protocol the packet used for communication. The protocol determines how data is formatted, exchanged, and processed.
Length	The total size, in bytes, of the packet.

DECISION TREE CLASSIFIER

A **decision tree** is a nonparametric supervised learning method for both classification and regression. The tree forms a hierarchical structure with branches where every internal node makes a decision based on the features of the dataset.

A decision tree is made through **recursive splitting of the dataset**: the tree is created by splitting the dataset into subsets based on the values of the features until a certain stopping criterion is met, like the maximum depth of the tree.

The baseline for the model will be a logistic regression classifier.

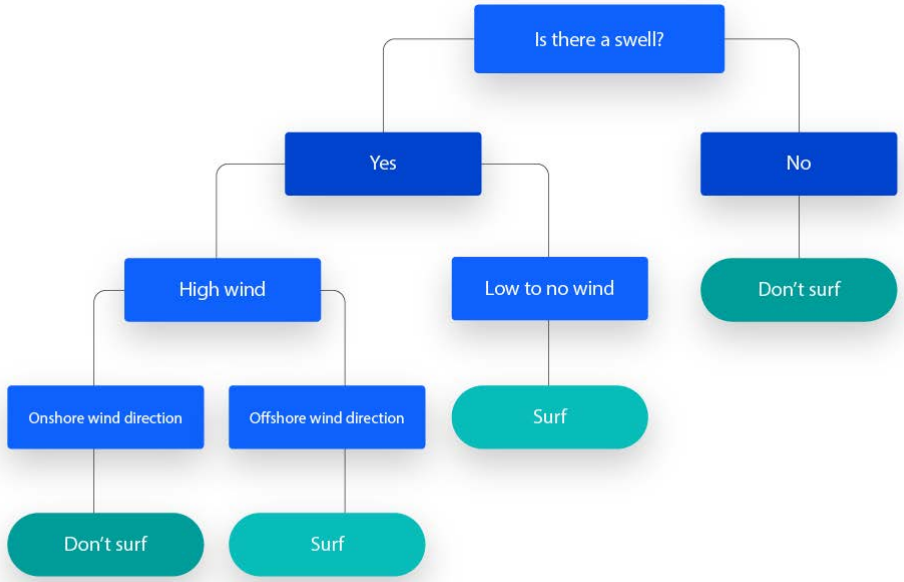


Image Source: https://www.ibm.com/topics/decision-trees

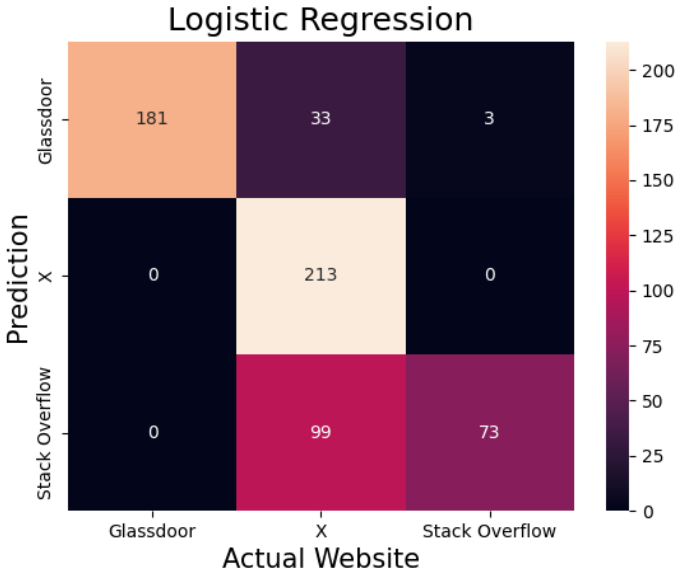
RESULTS

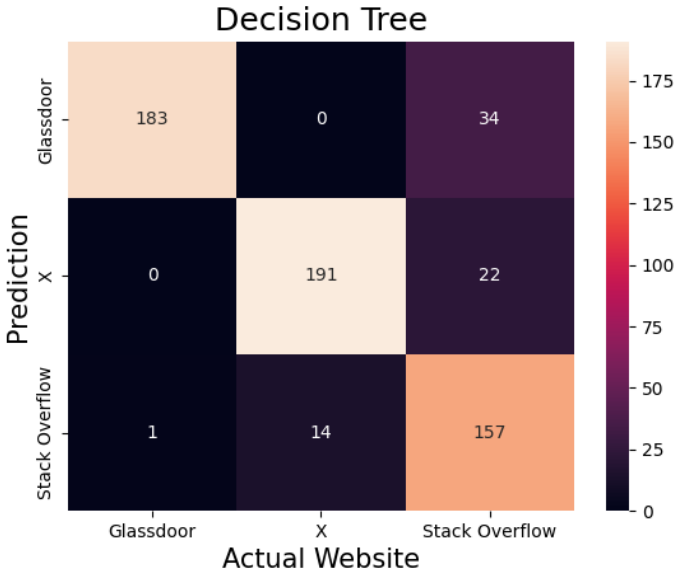
EVALUATION RESULTS ON TESTING DATA
(Train:Validation:Test – 8:1:1)

Model	Accuracy	Macro F1 Score
Logistic Regression	0.78	0.75
Decision Tree	**0.88**	**0.88**

The decision tree model outperformed the baseline model in both accuracy and F1 score, by **0.1** and **0.13** respectively.

Based on the confusion matrices to the right, we can see that the decision tree's predictions tend to align with the actual values. The logistic regression has more trouble, especially with properly identifying traffic from X.





REFLECTIONS

- The decision tree model sometimes misidentified Stack Overflow traffic as coming from elsewhere, and in rarer cases, identified other traffic as Stack Overflow traffic. Essentially, the issues in the model seem to be related to the Stack Overflow part of the dataset.
- My attempts to hypertune the parameters for the decision tree model did not seem to significantly improve the model's accuracy. More work could be done on that front, though.