Introduction

Hello and welcome to HIT401 Capstone project, we are group 5 Deep Learning Applications for cyber security. Today’s world is very centric towards technology and its advancements. With the ever-growing influence of technology in our daily lives, creating a safe and secure environment is absolutely vital.

This is where cyber security comes in place ensuing safety from illicit malicious activities of all sort. . Explainable artificial intelligence poses immense importance in the field of cyber security. The XAI tools critically analyse the decisions made by the cyber security application, especially emphasizing on how it uses datasets to make decisions in order to combat cyber threats. Once the ideas and functionalities behind the predictions are found, cyber security analysts and engineers further improves their applications to increase efficiency within their cyber security applications ultimately enhancing their defensive combat mechanism.

In this project we explore how LIME, one of the proven most tools of explainable AI makes critical decision in analysing crucial cyber security threats. Local interpretable model agnostic explanations, commonly referred to as LIME. explains comprehensible predictions and decision-making abilities of the Artificial intelligence or machine learning in general. LIME illuminates us as to how a blackbox takes decisions and what steps it takes to make a prediction. LIME also explains how certain anomalies contribute towards making a comprehensive prediction.

In this project we had taken a dataset ‘finance.csv’ and ran the explainable AI model, LIME to produce explanations and reasonings behind the predictions made by the artificial intelligence system. With this dataset our main aim is to locate Fake reviews which is a pressing issue in the cyber-crime arena, fake reviews heavily diminishes a business image and reputation ultimately affecting its financials and longevity. We analyse this by determining the number of reviews posted by paid subscribers and other non-subscriber. The non-subscribed reviews are the primary source of falsified reviews. In this dataset we have used random forest which is a classification algorithm commonly used in machine learning to solve regression problems. It constructs multiple decision-making tress which aid in solving the classification-based tasks. In this algorithm, we used a confusion matrix which evaluate the prediction results and performance within the classification of the problem. With the aid of confusion matrix, we were able to determine accuracy score of 85%, recall value of 73% and a precision of 62%.

These algorithms helps us determine the LIME model where we can use the explainable AI tools to understand the crucial findings. We can conclude that there are a large number of reviews being posted by non-subscribers. This makes its very troublesome for the business as potential clienteles would come under false influence which ultimately creates an inappropriate impression over the business. In such scenarios organisations could take the results into consideration and evaluate how accurate their reviews are or to what extent are they not real. If they have a good precision rating of the reviews, they could seriously consider the review contents and improve their services. These findings would also assist cyber security professionals to streamline a process aimed towards combatting falsified reviews.