MSSC 5780 Final Project Proposal
Sylvester Mensah, Navid Mohseni, Adam Reeson

Introduction

Logistic regression is a special type of regression which is used when the response variable is categorical. Binomial logistic regression is used when the response has only two categories but when there are more than two categories, then it becomes a multinomial logistic regression.
In this project the dataset used has “yesno” to be the dependent variable which is categorical with two two categories “y” and “n” hence we will employ binomial logistic regression here.

There has been a tremendous increase in the use of technology, hence the increase in cyber fraud. Cyber Fraud has different forms and one of the means this occurs is through spam emails. This is a typical classification problem where we intend to fit a logistic regression model to predict if an email is spam or not based on the regressors.

Methods

In our proposed analysis, we plan to use logistic regression to identify spam emails. The model will be built using various predictors from our dataset, such as occurrences of signs like the Dollar signs, exclamation signs, words “make” and “money,” and occurrences of the string ‘000’. To ensure our model is effective and not just tailored to our specific dataset, we’ll use a portion of the data to train the model and another to test it. This split into training and testing sets helps us understand how well our model performs in real-world scenarios. Our main goal is to create a model that can accurately distinguish between spam and non-spam emails, making it a valuable tool for email filtering.

Conclusion

Once we have developed and tested an effective logistic regression model, the idea is that this can be used in the future to classify suspicious emails as “spam” or “not spam,” given the proposed features are extracted from the email text. In doing this, one could have the spam emails automatically deleted from their inbox or moved to a different folder, creating a more efficient sorting process and thus minimizing the annoyance of having to filter these emails manually and even potentially preventing cyber fraud that was previously mentioned. Ultimately, we hope to demonstrate the usefulness of logistic regression in everyday binary classification problems.

Dataset:
https://github.com/rfordatascience/tidytuesday/blob/master/data/2023/2023-08-15/readme.md