Predicting Spam from SMS Message Board

Group member: Winnie Shao, Yihui Zhang, Wenjia Dou

Summary of research problems

1. What information can we get to distinguish the difference between ham and spam messages if we only perform basic exploratory data analysis?

   In this part, we want to perform the Exploratory Data Analysis on our dataset. Because the dataset we used only has 2 columns, we want to find more information from those text messages. By doing that, we want to compare the length of two categories and also intend to explore the top 10-20 words used in spam and ham messages. We may plot a word cloud to visualize different word choices in spam and ham SMS.
2. How can we process the text message to make them machine-readable?

   In order to predict spam from the message board, we want our data machine-readable so that we need to transform our raw SMS messages into numerical. We plan to develop 5 features (word count / average text length / URLs count / special punctuation count / TF-IDF) that can define these messages. By converting the raw SMS messages into numerical language features, the messages will be machine-readable so that we can use the dataset to make predictions.
3. How can we find a proper model to predict spam messages?

   For this problem, we will try some different models, such as Decision Tree Classifier, Naive Bayes Classifier, SVM Classifier, and XGBoost Classifier, to find the most optimized one. To do this, we divide the language feature generated data into training and testing sets and then try to find the minimum test mean squared error and the most accurate train method. Finally, we can plot out the internal decision tree of our model so that we can easily find out the most proper model and most important language features in defining spam messages.

Dataset

URL: https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection

We obtained this dataset from the UCI Machine Learning Archive. In this dataset, we have 5574 text messages. The dataset only has 2 columns: the first column indicates if the text is spam by using string value ‘spam’ and ‘ham’, and the second column contains the text content.

Challenge Goals

Two of the Challenge Goals we plan to use are **Machine Learning** and **New Library**. Our project will meet the goal of Machine Learning because we will

look through various model types and the different settings of hyperparameters to decide which model possesses higher accuracy in predicting spam. In addition, we will use a new library of advanced visualizations to perform the analysis of the difference between spam and ham messages and use a new library of natural language processing to process the large volume of the text in our dataset.