MARTINEZ, **Paula Joy B.**
IE 198
Capstone

PRE-PROCESSING

One of the pre-processing steps employed was to check for and remove the duplicate values in the dataset. The classes “legit” and “spam” were also converted into binary variables 0 and 1, respectively. For the text, I used regular expressions to replace the “http”, phone numbers, email addresses, currency signs, punctuations. The purpose of this is to eliminate any potential link between spam messages and email addresses or phone numbers. Numbers and leading and trailing white spaces were also replaced, as well as the words that are irrelevant to the classification, such as prepositions and conjunctions. All regular expressions were based on this source: http://regexlib.com/Search.aspx. Data tokenization was also done for the machine to be able to process the text data and use them as inputs to estimators.

PARAMETERS USED AND OBTAINED VALUES

For this classification analysis, the parameter of interest is the area under the curve (AUC) of the ROC, since we want the model that can best distinguish between the two classes. In this case, the higher the AUC score, the better the performance of the model at distinguishing between legit and spam messages.

Four models were used: SVM, Random Forest, Logistic Regression, and Naïve-Bayes. The results for each model are as follows:

SVM

```
Test AUC score for SVM is 0.9285291606434869
              precision    recall  f1-score   support

           0       0.98      0.99      0.99       903
           1       0.96      0.86      0.91       131

    accuracy                           0.98      1034
   macro avg       0.97      0.93      0.95      1034
weighted avg       0.98      0.98      0.98      1034
```

Random Forest

```
Test AUC score for RF is 0.9225567024253336
              precision    recall  f1-score   support

           0       0.98      1.00      0.99       903
           1       0.98      0.85      0.91       131

    accuracy                           0.98      1034
   macro avg       0.98      0.92      0.95      1034
weighted avg       0.98      0.98      0.98      1034
```

MARTINEZ, **Paula Joy B.**
IE 198
Capstone

Naïve-Bayes

```
Test AUC score for Logistic Regression is 0.8511450381679388
              precision    recall  f1-score   support

           0       0.96      1.00      0.98       903
           1       1.00      0.70      0.83       131

    accuracy                           0.96      1034
   macro avg       0.98      0.85      0.90      1034
weighted avg       0.96      0.96      0.96      1034
```

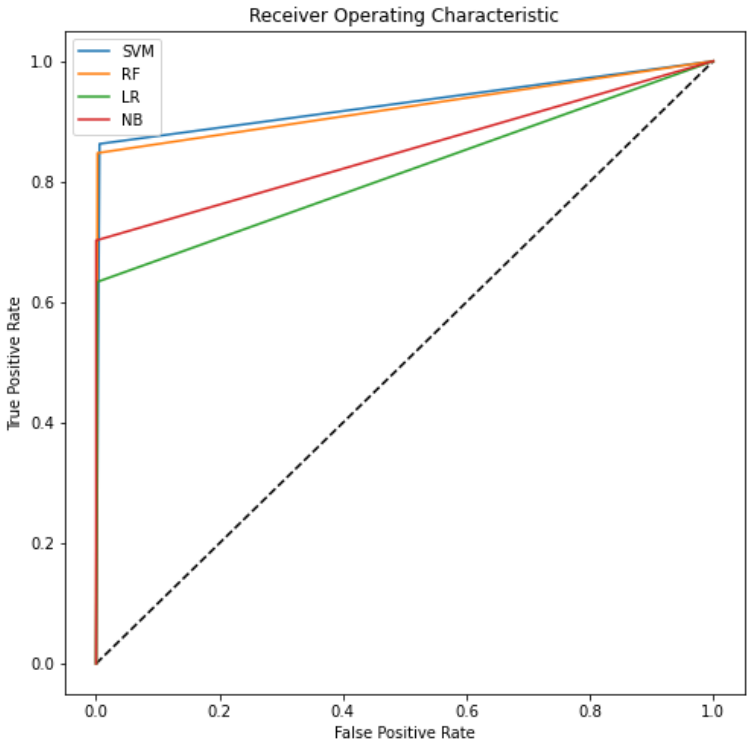
Logistic Regression

```
Test AUC score for Logistic Regression is 0.8156864734177
              precision    recall  f1-score   support

           0       0.95      1.00      0.97       903
           1       0.98      0.63      0.77       131

    accuracy                           0.95      1034
   macro avg       0.96      0.82      0.87      1034
weighted avg       0.95      0.95      0.95      1034
```

To visualize the AUC scores of each model, the ROC curve was plotted.



From the plot and figures above, the best model to classify legit and spam messages is the SVM model.

MARTINEZ, **Paula Joy B.**
IE 198
Capstone

References

3.2. Tuning the hyper-parameters of an estimator. scikit. (n.d.). Retrieved from https://scikit-learn.org/stable/modules/grid_search.html

Brownlee, J. (2019, August 21). *How to Tune Algorithm Parameters with Scikit-Learn*. Machine Learning Mastery. Retrieved from https://machinelearningmastery.com/how-to-tune-algorithm-parameters-with-scikit-learn/

Defining a Search Space. Defining search spaces - Hyperopt Documentation. (n.d.). Retrieved from http://hyperopt.github.io/hyperopt/getting-started/search_spaces/

Galstyan, L. (2020, July 25). *Predicting Spam Messages*. Medium. Retrieved from https://towardsdatascience.com/predicting-spam-messages-17b3ca6699f0

Gilde, K. (2021, January 17). *Faster Hyperparameter Tuning with Scikit-Learn's New HalvingGridSearchCV*. Medium. Retrieved from https://towardsdatascience.com/faster-hyperparameter-tuning-with-scikit-learn-71aa76d06f12

Hunner, T. (2018, October 11). Asterisks in Python: what they are and how to use them. Retrieved from https://treyhunner.com/2018/10/asterisks-in-python-what-they-are-and-how-to-use-them/

Hyperopt - Alternative Hyperparameter Optimization Technique. Analytics Vidhya. (2020, December 29). Retrieved from https://www.analyticsvidhya.com/blog/2020/09/alternative-hyperparameter-optimization-technique-you-need-to-know-hyperopt/

Labs, D. D. (2017, December 23). *Parameter Tuning with Hyperopt*. Medium. Retrieved from https://medium.com/district-data-labs/parameter-tuning-with-hyperopt-faa86acdfdce

N-gram range. MonkeyLearn. (n.d.). Retrieved from https://help.monkeylearn.com/en/articles/2174105-n-gram-range

The first Regular Expression Library on the Web! Regular Expression Library. (n.d.). Retrieved from https://regexlib.com/Search.aspx