Ali Malenchik

Milestone 3 - Audience Questions

September 26, 2021

1. **Are there any commonalities between the content of the spam and ham messages?**

   Some of the messages utilize the same words, though not at the same frequency. Additional exploratory analysis could evaluate and compare the frequency of words shared between the spam and ham messages.

2. **Does having imbalanced classes impact the project in any way?**

   Because the spam data is the minority class, it is more challenging for the models to learn the characteristics of spam. This may increase the instances of false negatives in the classification. One enhancement to this project would be to collect more examples of spam SMS messages in order to balance out the classes and increase the model accuracy.

3. **How were the algorithms for training selected over other algorithms?**

   The five algorithms selected were done so because the root of this project was a binary classification problem. They were chosen over other binary classifiers simply due to time constraints. Another enhancement of this project would be to include K-Nearest Neighbors and other neural networks.

4. **Was punctuation considered when training the models? Should it be?**

   Punctuation was not considered, as non-alphanumeric characters were removed in the data cleansing step. Given more time, punctuation could be considered as features to determine if it impacts the performance of the models.

5. **Why was Python chosen as a language over R?**

Python was chosen because it was more familiar and I have experience with binary classifiers using Python. I wanted to focus this project on Natural Language Processing (NLP), which meant I needed to save time in other areas such as modeling.

6. **Does the removal of stop words make any impact to the performance of the models?**

   The removal of stop words should not have a negative impact to the performance since stop words generally don't change the semantics of the message and the removal of unnecessary features actually improves model efficiency.

7. **Why was the evaluation metric chosen over other metrics?**

   Precision was chosen since it is critical in spam classification to ensure valid messages are not falsely identified as spam.

8. **What, if anything, could be explored to improve the performance of the model(s)?**

   Hyperparameter tuning should be explored, since all models were trained with default parameters.

9. **How will you identify and/or prevent overfitting?**

   Overfitting can be identified in the model evaluation step. Since the models performed well on unseen data, it is unlikely that they are overfitted. The models should be continually trained on new data at a large scale to prevent overfitting.

10. **Can the same algorithm be applied to email messages?**

    Because the content style of spam and non-spam messages may differ between SMS and email, I would not recommend applying the same models to email. For instance, text slang is not often utilized in emails, and emails may contain images whereas texts may not; these differences would impact the performance of the models.