PuriCloud

SMS SPAM DETECTION

AUTHOR: BRADLEY D. CASTLE

6/11/2023

CONTENTS / AGENDA

- Executive Summary
- Business Problem Overview
- Solution Approach & Design
- Data Definition
- Exploratory Data Analysis
- Data Preprocessing
- Model Results
- Model Performance Summary
- Insights
- Recommendations
- Appendix

EXECUTIVE SUMMARY

Problem Overview: SMS communication has evolved with various applications but comes with the challenge of spam that may include phishing links, posing a threat to cybersecurity.

Objective: Develop an optimal machine learning model that predicts SMS spam by creating a spam classifier using a sample dataset of labeled SMS texts as "spam" or "ham" (non-spam), to safeguard the organization from potential cyber threats.

Methodology:

- Data Preparation: Preprocessing of SMS text data.
- Exploratory Data Analysis (EDA): Assess data distribution and intervariable relationships.
- Sentiment Analysis: Derive sentiment scores using "Sentiword Net" algorithm.
- Insight Generation: Evaluate and visualize sentiment scores via word clouds.
- Solution Design: Construct Decision Tree and Random Forest classifiers for spam detection.

Key Insights:

- Dataset: 4,825 ham and 747 spam messages, no missing values.
- Frequent words in spam: "Call", "Free", "Text", "Mobile", "Stop", "Cash".
- Ham messages: Neutral sentiment (average score: 0.02)
- Spam messages: Slightly negative sentiment (average score: -0.18).

BUSINESS PROBLEM OVERVIEW & OBJECTIVE

Problem Overview: Short Message Services (SMS), a product of global mobile communication standards, has evolved beyond simple chatting. However, with new innovation comes new challenges, including spam, the misuse of electronic messaging for unsolicited bulk communications. Spam can be utilized for commercial ads and for distributing dangerous phishing links. These harmful traps prey on unsuspecting users, leading to devastating financial losses through hacking or theft. As a major contributor to global cybercrimes, spam's increasingly sophisticated techniques have prompted extensive research, giving rise to innovative applications on spam classifiers using Natural Language Processing (NLP).

Objective: Safeguard the organization from potential cyber threats via SMS messages to employee phones. The process consists of a sample dataset of labeled SMS texts, marked as either "spam" or "ham" (non-spam). Extract valuable insights from this data and utilize it to construct a classification model. Leverage machine learning algorithms on this preprocessed SMS data and accurately predict whether an SMS is "spam", thereby contributing to the greater safety of the digital environment.

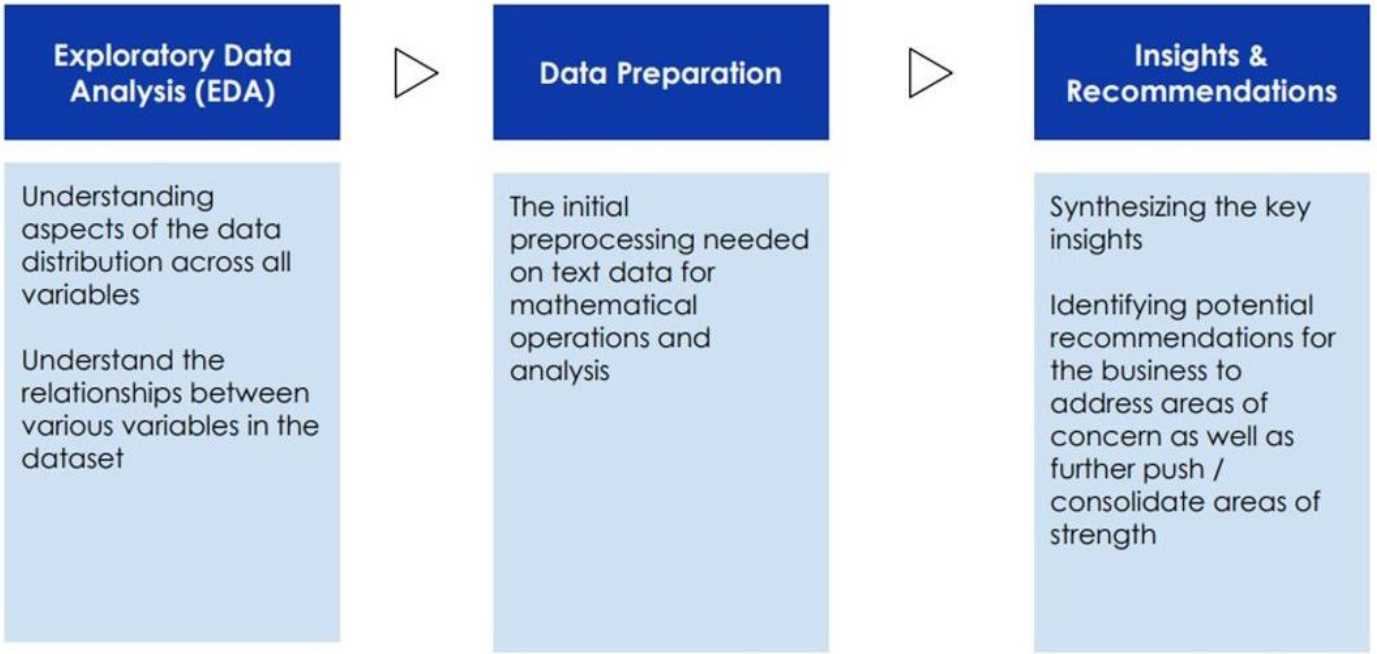
Link to Appendix: Dangers of SMS Spam

SOLUTION APPROACH

Methodology for extracting meaningful insights from the data to predict whether an SMS is "spam" or not

- **Data Preparation:** Preprocess SMS text data for statistical operations and analysis.
- **Exploratory Data Analysis (EDA):** Assess data distribution across all variables. Comprehend relationships within the SMS dataset.
- **Sentiment Analysis**: Implement the "Sentiword Net" algorithm for deriving sentiment scores from the SMS text data.
- **Insight Generation**: Evaluate sentiment scores across the SMS dataset. Visualize the connections between sentiment scores through "Word Clouds".
- **Solution Design:** Develop a Machine Learning classifier to predict whether an SMS is "spam" or "not spam" based on the content. Build two Decision Tree models (1 unpruned, 1 pruned), and two Random Forest models (1 unpruned, 1 pruned).
- **Insights and Recommendations:** Extract key insights from data analysis. Propose strategic suggestions to improve spam detection and handle potential concerns.

SOLUTION DESIGN



DATA DEFINITION

This data contains a collection of messages on different subjects. Based on the messages in the file, it appears to be a mix of personal and promotional messages. Some of the messages are friendly and casual, while others are spam or promotional in nature. The tone of the messages varies from playful to serious, and some contain jokes or wordplay. Overall, this dataset provides a snapshot of different types of communication that people have sent via text message or other messaging platforms. We have two columns:

Category	Contains the labels "spam" or "ham" for the corresponding text data
Message	Contains the SMS text data

Link to Appendix: Other Factors to Consider

EXPLORATORY DATA ANALYSIS

- 5.572 Rows: Each row in the dataset represents a message.
- 2 Columns: The columns / attributes in the dataset contain messages on various subjects.

Row No.	Category	Message
1	ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
2	ham	Ok lar... Joking wif u oni...
3	spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
4	ham	U dun say so early hor... U c already then say...
5	ham	Nah I don't think he goes to usf, he lives around here though
6	spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, £1.50 to rcv
7	ham	Even my brother is not like to speak with me. They treat me like aids patent.
8	ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune
9	spam	WINNER!! As a valued network customer you have been selected to receivea £900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.
10	spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 08002986030
11	ham	I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough today.
12	spam	SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info
13	spam	URGENT! You have won a 1 week FREE membership in our £100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX 4403LDNW1A7RW18
14	ham	I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise. You have been wonderful and a blessing at all times.
15	ham	I HAVE A DATE ON SUNDAY WITH WILL!!
16	spam	XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap. xxxmobilemovieclub.com?n=QJKGIGHJJGCBL
17	ham	Oh k...i'm watching here:)
18	ham	Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.

ExampleSet (5,572 examples, 0 special attributes, 2 regular attributes)

EXPLORATORY DATA ANALYSIS

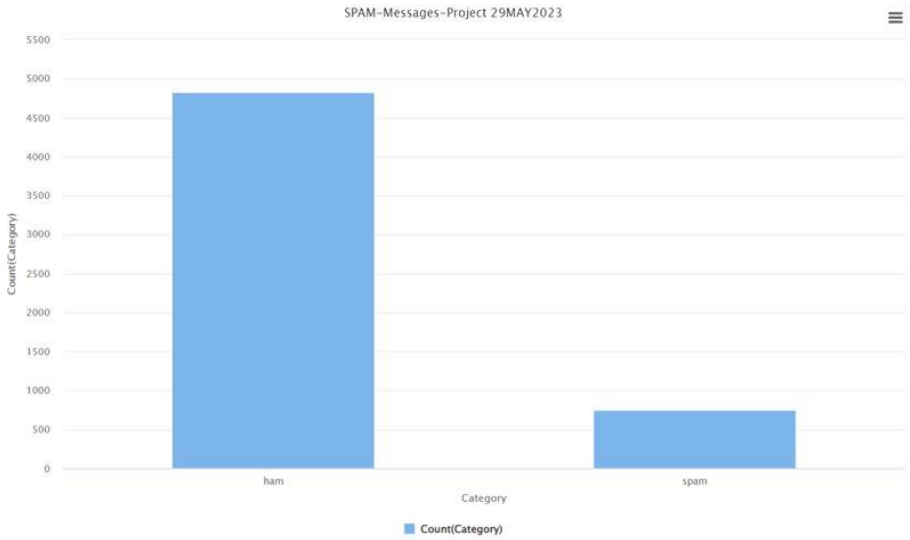
- Both attributes are of the categorical type.
- There are no missing values.
- According to the subject statistics, ham (4,825) has the highest number of messages and spam (747) has the lowest number of messages, which makes sense because most email communication is legitimate (ham), thus outnumbering spam in typical datasets.

Name	Type	Missing	Statistics		Filter (2 / 2 attributes): Search for Attributes
			Least	Most	Values
Category	Polynominal	0	spam (747)	ham (4825)	ham (4825), spam (747)
Message	Polynominal	0	... we r s [...] oon " (1)	Sorry, I [...] ater (30)	Sorry, I'll call later (30), I cant p [...] a message (12), ...[5155 more]

Link to Appendix: Data Background and Contents

EXPLORATORY DATA ANALYSIS

- Ham and spam are the two primary forms of messaging presented, which is understandable given the dataset is collectively about global communication through SMS.



DATA PREPROCESSING

- **Text Cleaning:** Remove irrelevant characters such as punctuation and numbers as well as convert all text to lowercase for uniformity.
- **Stopword Removal:** Filter out common words (i.e. "and", "the") that do not contribute significantly to the meaning of messages.
- **Tokenization:** Break down the text into individual words or tokens, which serve as the basis for feature extraction.
- **Sentiment Analysis:** Sentiment scores were derived using the Sentiword Net algorithm to capture the underlying sentiment of each message. This information is used as an additional feature for the model.
- **Vectorization:** Convert the tokenized text data into numerical vectors using techniques such as TF-IDF (Term Frequency-Inverse Document Frequency) to represent the importance of terms within messages.
- **Dataset Splitting:** Divide the dataset into training and testing sets to evaluate the model's performance on unseen data.

TEST ANALYSIS - HAM

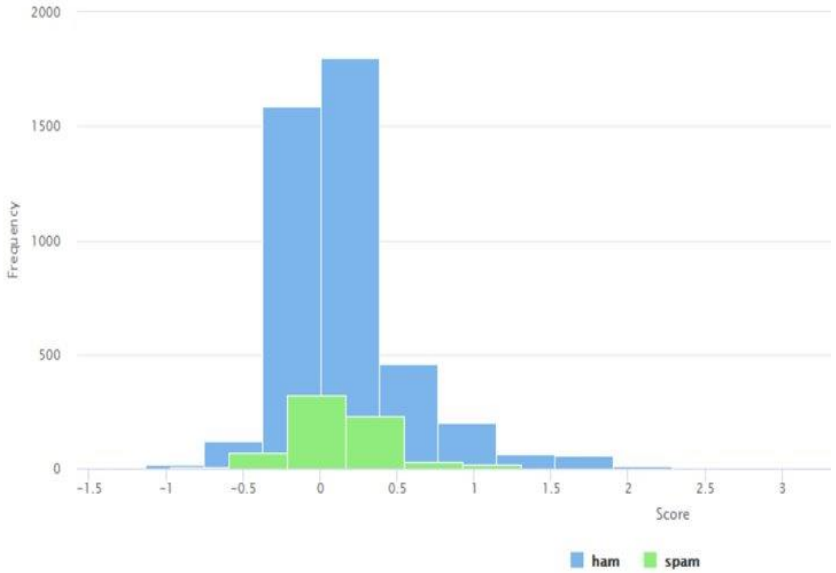
- The following word cloud was created from the **ham** text, highlighting prominent terms. This aids in identifying prevalent words for data preprocessing, enabling us to filter out noise such as common stop words, and discern the most pertinent terms.
- The following words appear to be the more frequently used words: Call, Come, Time, Good, Want, Love.

TEST ANALYSIS - SPAM

- The following word cloud was created from the **spam** text, highlighting prominent terms. This aids in identifying prevalent words for data preprocessing, enabling us to filter out noise such as common stop words, and discern the most pertinent terms.
- The following words appear to be the more frequently used words: Call, Free, Claim, Prize, Stop, Text.

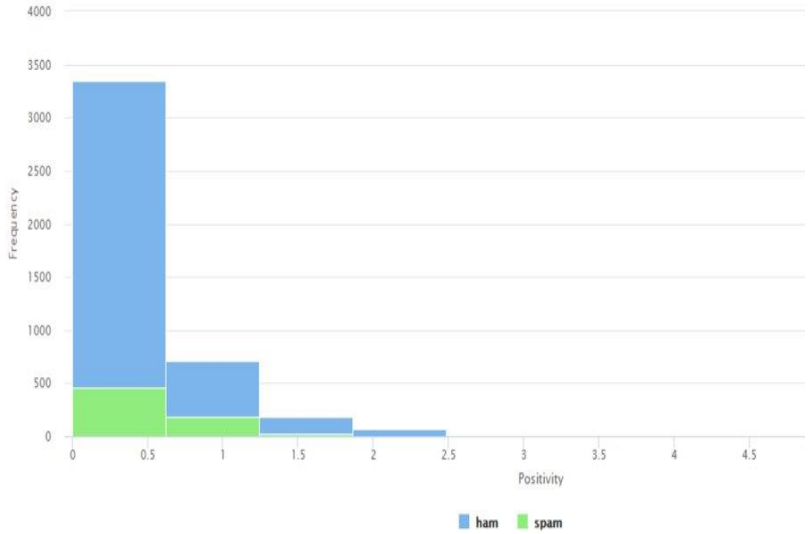
SENTIMENT ANALYSIS

- The sentiment scores for all the **ham** messages range from –1.50 to 2.28, with an average score of 0.02. This suggests that the messages in this class are generally neutral, with some variation in positivity and negativity.
- The sentiment scores for all the **spam** messages are generally more negative, with scores ranging from –.97 to 1.30, with an average score of –0.18. This is not surprising, as spam messages are often unsolicited and can be perceived as annoying or intrusive.



SENTIMENT ANALYSIS

- Overall, it seems that the "ham" class has higher sentiment scores than the "spam" class. This could be due to a number of factors, such as the content of the messages, the relationship between the sender and receiver, or the context in which the messages were sent.



MODEL RESULTS – DECISION TREE

- We built a Decision Tree classification model where Gini Index was used as the splitting criterion.
- The model performs very well on the training data, in terms of both the Precision and the Recall.
- After Pruning, the model indicates 98.44% accuracy.

Confusion Matrix

accuracy: 98.44%

	true ham	true spam	class precision
pred. ham	3370	56	98.37%
pred. spam	5	469	98.95%
class recall	99.85%	89.33%	

Performance Metrics	**Accuracy**	**98.44%**
	Precision	**98.66%**
	Recall	**94.59%**

MODEL RESULTS - DECISION TREE

- The model performs very well on the test data, with more than 95% score for Accuracy. This indicates that the model generalized well and can be used to make predictions on spam. The organization can use this model to predict spam as a security measure against cyberattacks.

Confusion Matrix

accuracy: 95.51%

	true ham	true spam	class precision
pred. ham	1432	57	96.17%
pred. spam	18	165	90.16%
class recall	98.76%	74.32%	

Performance Metrics	**Accuracy**	**95.51%**
	Precision	**93.17%**
	Recall	**86.54%**

PuriCloud

MODEL RESULTS – RANDOM FOREST

- The Random Forest model, an ensemble model, trained very well with an accuracy of 95%. The Gini Index was used as a split criterion for fine tuning hyperparameters.
- The training model also performs well in terms of precision and recall.

Confusion Matrix

accuracy: 95.00%

	true spam	true ham	class precision
pred. spam	449	6	98.68%
pred. ham	30	235	88.68%
class recall	93.74%	97.51%	

Performance Metrics	**Accuracy**	**95%**
	Precision	**93.68%**
	Recall	**95.62%**

PuriCloud

MODEL RESULTS – RANDOM FOREST

- The Random Forest model performs very well with the Test Data, indicating an accuracy of 93.51%.
- This ensemble model also performs over 92% in both precision and recall.
- The organization can certainly use this model with a high accuracy rate to predict whether email is spam in order to protect against cyberattacks or phishing.

Confusion Matrix

accuracy: 93.51%

	true spam	true ham	class precision
pred. spam	192	7	96.48%
pred. ham	13	96	88.07%
class recall	93.66%	93.20%	

Performance Metrics	**Accuracy**	**93.51%**
	Precision	**92.28%**
	Recall	**93.43%**

MODEL PERFORMANCE

- **Classifier Selection:** Two classifiers were considered, Decision Tree and Random Forest. Decision Trees are simple and interpretable, while Random Forests, being ensembles of Decision Trees, provide robustness against overfitting.
- **Model Training:** Both models were trained using the Gini Index as the splitting criterion. Random Forest was finetuned with hyperparameter optimization for better performance.
- **Model Evaluation:** Models were evaluated based on accuracy, precision, and recall. Decision Tree yielded 98.44% training accuracy and 95.51% test accuracy. Random Forest had a training accuracy of 95% and test accuracy of 93.51%.
- **Insight into Results:** Though Decision Tree had slightly higher accuracy, **Random Forest is recommended** due to its better generalization and ability to handle diverse data patterns.
- **Future Prospects:** Regularly update the dataset with new message patterns and periodically retrain the model. Explore other ensemble methods and deep learning approaches for further improvements.

INSIGHTS

Word Cloud Analysis: Spam messages often contain words like "Call", "Free", "Text", "Mobile", and "Stop". These were considered as key features for classification.

Sentiment Scores:

- Ham messages: Generally neutral, scores ranging from -1.50 to 2.28.
- Spam messages: Slightly negative, scores ranging from -0.97 to 1.30.

Model Performance:

- Decision Tree: Training accuracy 98.44%, Test accuracy 95.51%.
- Random Forest: Training accuracy 95%, Test accuracy 93.51%.

RECOMMENDATIONS

- **Model Selection:** The Random Forest model is selected in this use case. While the Decision Tree performs slightly better in accuracy, Random Forest ensures better generalization due to its ensemble nature, reducing overfitting and handling outliers more efficiently.
- **Data Set for Prediction:** The test data set is optimal to evaluate model performance as it reflects the model's ability to generalize new, unseen data which is critical for practical applications.
- **Enhance Security:** Continuously update your SMS dataset to include new patterns in spam messages, and periodically retrain the model for better performance.
- **Customize Alert System:** Implement an alert system for employees in the case of spam detection, educating them on the potential risks and steps to take.

APPENDIX

OTHER FACTORS TO CONSIDER

The following factors of SMS spam detection should be taken into considerationsfor building a classification model to predict whether an SMS is spam or not:

- **Message length:** The length of the message can be a useful attribute for distinguishing between ham and spam messages, as spam messages are often longer than ham messages.
- **Presence of certain keywords or phrases:** Certain keywords or phrases are commonly used in spam messages, such as "free", "win", "prize", "offer", "discount".
- **Frequency of certain characters or words:** Spam messages often contain repeated characters or words, such as "!!!!" or "buy buy buy".
- **Time of day:** Spam messages are often sent at unusual times of the day or night, such as early in the morning or late at night.
- **Sender information:** Spam messages are often sent from unknown or suspicious senders, while ham messages are more likely to be sent from known contacts or reputable sources.
- **Capital letters:** Spam messages often use excessive capitalization to grab the reader's attention.
- **Special characters:** Spam messages may use special characters, such as %, $, and #, to create a sense of urgency or importance.
- **Emojis:** Emojis are often used in spam messages to make them more visually appealing.
- **URLs:** Spam messages often contain URLs that lead to phishing sites or other malicious content.
- **Phone numbers:** Spam messages may contain phone numbers that lead to scams or other fraudulent activities.
- **Misspellings or grammatical errors:** Spam messages may contain misspellings or grammatical errors, which can be a useful attribute for identifying spam messages.
- **Multiple languages:** Spam messages may contain text in multiple languages.
- **Specific message formats**: Spam messages may follow specific message formats, such as "You have won a prize!" or "Your account has been compromised".
- **Specific message topics:** Spam messages may focus on specific topics, such as weight loss, dating, or financial opportunities.

DANGERS OF SMS SPAM

Understanding the Risks of SMS Spam is Paramount in Safeguarding Data

- **Phishing Attacks:** Through SMS, hackers deploy deceptive messages, often mimicking legitimate entities, to dupe recipients into revealing sensitive information like passwords and credit card details.
- **Scams & Fraudulent Offers:** Spam messages may contain offers that are too good to be true, leading recipients into financial scams or fraudulent transactions. **Malware Distribution:** Harmful links embedded in spam messages can download malware onto a user's device, giving hackers unauthorized access to sensitive data and system resources.
- **Smishing:** A subset of phishing, this involves hackers using SMS to exploit the trust of recipients. This often results in identity theft or financial loss.
- **Service Disruption:** Excessive spam can clutter and overwhelm messaging services, impacting their efficiency and reliability. This may be a Denial of Service (DoS) attack, SMS flooding or SMS bomb.
- **Social Engineering Attacks:** Hackers manipulate recipients psychologically through spam, coercing them into executing actions detrimental to their security.