SMS Spam Detection

Data Exploration - Unstructured Data / MIT-No Code AI and Machine Learning-February 2025-A

04/05/2025

Begum SOZER

Contents / Agenda

- Data Dictionary
- Business Problem Overview and Solution Approach
- Exploratory Data Analysis
- Model Performance Summary
- Insights & Recommendations
- Appendix

Data Dictionary

COLUMN NAME	DESCRIPTION
Category	Label that shows if a message is ‘spam’ or ‘ham’ (not spam)
Message	The actual SMS text content

Business Problem Overview and Solution Approach

Business Problem

Cyber criminals often send spam SMS messages to trick people into clicking phishing links, which can lead to financial loss or data breaches. Cyber Solutions wants to protect employees by detecting and blocking these spam messages.

Solution Approach

- Understand the data: Analyze SMS messages labeled as ‘spam’ or ‘ham’.
- Explore patterns: Use visual tools to find trends and common spam words.
- Text processing: Convert messages into numbers using TF-IDF and analyze the message sentiment.
- Build models: Train and test machine learning models like Decision Tree and Random Forest.
- Evaluate results: Compare model performance using accuracy, precision, and recall.
- Recommend actions: Suggest the best model and ways to reduce spam risks.

EDA - A brief overview of the raw dataset

COLUMN NAME	DESCRIPTION
Category	Label that shows if a message is ‘spam’ or ‘ham’ (not spam)
Message	The actual SMS text content

EDA

Spam vs Ham Messages

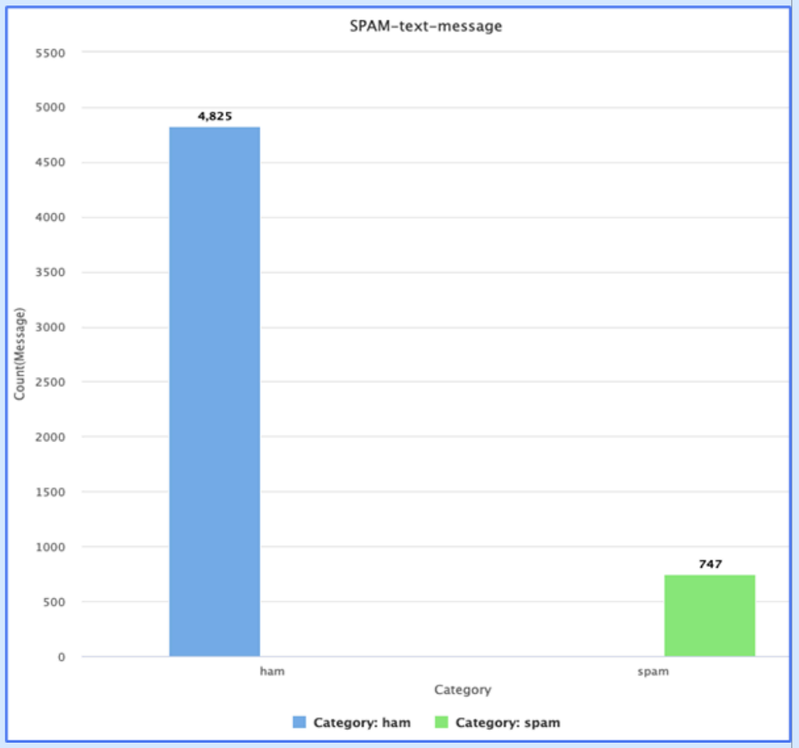
- The dataset contains a significant imbalance between ham (non-spam) messages and spam messages.
- There are 4,825 ham messages and only 747 spam messages.

Percentage of Ham and Spam Messages

- Ham messages: ~ 86.7% of the dataset
- Spam messages: ~ 13.3% of the dataset

Handling Dataset Imbalance

- Resampling Techniques
- Over-sampling: Increase the number of spam messages by duplicating or generating synthetic examples.
- Under-sampling: Reduce the number of ham messages to match the spam messages count.
- Algorithmic Adjustments
- Use models that can handle imbalanced datasets (like tree-based methods or boosting algorithms).
- Adjust the decision threshold to focus more on predicting the minority class (spam).



EDA: Word Cloud for Spam messages

What is this illustration called?

- This is called a Word Cloud.

When is it used?

- A word cloud is used during text analysis to visually highlight the most frequent words in a set of text data. It is commonly used in the early stages of data exploration.

How does it help solve the problem?

- It helps identify common words used in spam messages (like “win”, “free”, “cash”, “claim”), which are often used in phishing attempts.
- These patterns help us understand the language and keywords commonly found in spam, which supports:
- Feature extraction (like using frequent words as indicators)
- Model training for spam classification
- It can also help in detecting trends or updates in how spam messages are crafted over time.

Text Analysis

Which technique is used to find word frequencies?

- We use TF-IDF (Term Frequency – Inverse Document Frequency).
- How does it help solve the problem?
- TF-IDF helps identify important words in messages by considering:
- How often a word appears in a message (Term Frequency)
- How unique that word is across all messages (Inverse Document Frequency)
- It reduces the weight of common words (like “the”, “is”, “and”) and highlights words that are more meaningful, especially for identifying spam keywords (like “win”, “offer”, “urgent”).

Why is it useful for this problem?

- Helps transform SMS text into numerical features that machine learning models can understand.
- Makes it easier to detect patterns in spam messages compared to regular ones.
- Improves the model's ability to accurately classify spam vs. ham.

Text Analysis

What technique is used to find sentiment?
We use Sentiment Analysis techniques, often based on pre-trained NLP models or sentiment lexicons, to assign a sentiment score to each message (positive, negative, or neutral).

How can sentiment scores be used?

- Each message is analyzed for its emotional tone.
  - Spam messages often have urgent or manipulative language (e.g., “Congratulations!”, “Act now!”, “Limited offer!”).
  - Ham messages tend to be more neutral or friendly.
- By comparing sentiment scores, we can differentiate the tone used in spam vs. ham messages.

Why is this helpful for solving business problems?

- Helps detect emotionally charged spam, which often tries to trick users.
- Adds another feature that improves spam classification accuracy.
- Businesses can use these insights to develop smarter filters and reduce the risk of phishing attacks.

Sentiment Scores for Spam and Ham texts

Why analyze sentiment scores from a business perspective?

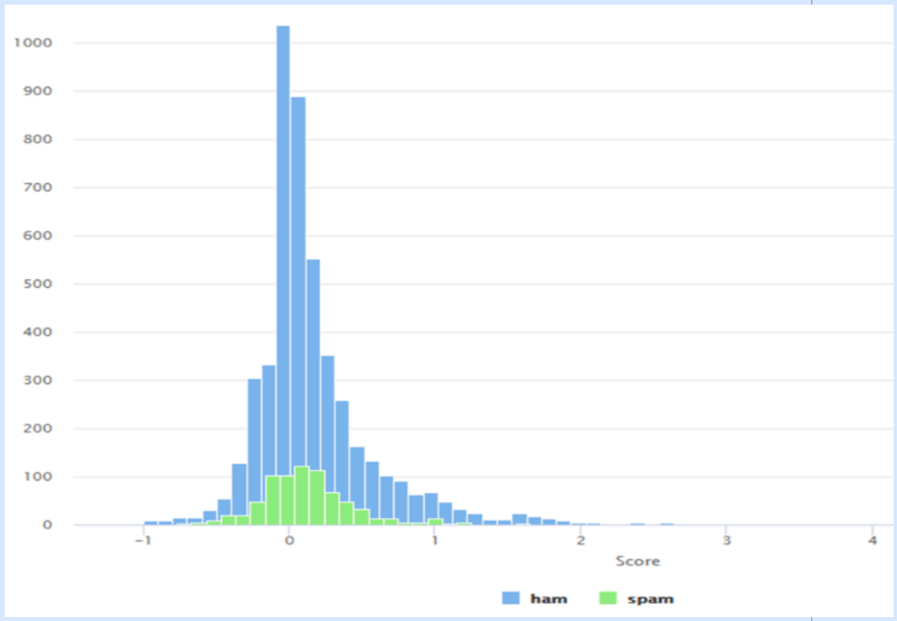
- Spam messages often use highly emotional or urgent language to create pressure or excitement (like “Win now!”, “Exclusive offer!”, “Urgent action required!”).
- Ham messages, being regular communication, usually have a neutral or positive tone.

How does this help prevent cyber attacks?

- By analyzing the difference in sentiment between spam and ham, we can:
  - Better detect suspicious, emotionally charged messages.
  - Train models to flag messages with negative or manipulative sentiment.
  - Alert employees to be cautious when receiving such messages.

Business Value

- Enhances spam filters with sentiment-based rules.
- Reduces the risk of employees falling for phishing and scam texts.
- Helps protect sensitive data and financial resources.



Model Performance Evaluation Decision Tree

Model Evaluation

- The Decision Tree model is a simple and easy-to-understand classification model.
- It works by splitting the data based on important features (like certain spam keywords or sentiment).

Comments

- The model achieves high accuracy, meaning it correctly predicts most messages.
- Precision is strong, showing that most messages predicted as spam truly are spam.
- Recall indicates that the model misses some actual spam messages, which could be risky from a security perspective.
- While effective, the model might be overfitting, as it builds a complex tree with high depth.

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
Decision Tree	97.22	96.59	92.03	91.50	95.71	93.50

Model Performance Evaluation
Pruned Decision Tree

What is Pruning?

- Pruning is the process of removing unnecessary branches from the decision tree to avoid overfitting.
- It helps the model generalize better to new, unseen data.

Comments

- Slight drop in accuracy and recall, but the model is now simpler and more generalizable.
- Precision (90.14%) is almost the same, meaning it's still very reliable in predicting spam.
- The small trade-off in performance is worth it because pruning reduces overfitting and makes the model more suitable for real-world use.

Business Relevance

- A pruned tree is better for deployment: faster, more stable, and less prone to mistakes when new types of spam appear.

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
Decision Tree Before Pruning	97.22	96.59	92.03	91.50	95.71	93.50
Decision Tree - Pruned	96.75	94.79	91.76	89.90	93.96	88.18

Model Performance Evaluation Random Forest

Comments on Model Performance

- The Random Forest model performs very well with high test accuracy (95.42%), showing it generalizes well to unseen data.
- Test Precision (97.49%) is excellent, meaning most messages predicted as spam are truly spam.
- Test Recall (82.89%) is slightly lower, indicating that some actual spam messages are still missed.
- There is a small drop from training to test scores, which is normal and healthy, suggesting the model is not overfitting.
- Overall, Random Forest shows a good balance between accuracy and generalization, making it suitable for deployment.

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
Random Forest	95.83	95.42	84.45	82.89	97.70	97.49

Model Performance Evaluation
Pruned Random Forest

Comments

- Pruned Random Forest shows improved recall, especially on test data (82.89% → 85.31%), meaning it captures more actual spam messages.
- However, test precision dropped slightly (97.49% → 90.78%), which means it made a few more false positives (ham marked as spam).
- Test accuracy is slightly lower (94.70%) compared to the unpruned model (95.42%), but still very high.
- The trade-off is acceptable if the business prefers catching more spam even if it occasionally flags a regular message.
- Overall, pruning improves generalization and increases spam detection coverage, which is important for security-focused use cases.

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
Random Forest before pruning	95.83	95.42	84.45	82.89	97.70	97.49
Random Forest - Pruned	97.49	94.70	91.70	85.31	97.32	90.78

Model Performance Evaluation
Pruned Random Forest

Best-Fit Model for the Business

Based on the comparison of all four models, the Decision Tree (before pruning) appears to best meet the company's objective.

Why?

- It provides the highest test accuracy (96.59%) and test recall (91.50%), meaning it is very effective at detecting actual spam messages, which is critical for preventing cyber attacks.
- While Random Forest models offer higher precision, recall is more important for security – it's better to catch more spam, even if it means a few false positives.

Key Evaluation Metric: Recall

- Recall is prioritized because:
  - The business goal is to detect as many spam messages as possible to protect employees and sensitive information.
  - Missing a spam message (false negative) could lead to phishing or financial fraud.

Recommendations for the Company

- Deploy the Decision Tree (before pruning) for maximum spam detection.
- Monitor false positives and apply filters or manual review where necessary.
- Regularly retrain the model with updated SMS data to keep up with new spam tactics.
- Explore combining models (like ensemble techniques) to balance precision and recall further.

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
Decision Tree before pruning	97.22	96.59	92.03	91.50	95.71	93.50
Decision Tree - Pruned	96.75	94.79	91.76	89.90	93.96	88.18
Random Forest before pruning	95.83	95.42	84.45	82.89	97.70	97.49
Random Forest - Pruned	97.49	94.70	91.70	85.31	97.32	90.78

Insights and Recommendations

Insights:

- Spam messages often use urgent or emotional language to trick users.
- Sentiment analysis helps detect such manipulative content.
- The Decision Tree model before pruning achieved the highest test recall and accuracy, making it suitable for spam detection.

Recommendations:

- Deploy the Decision Tree model to improve spam filtering.
- Retrain the model regularly with new data.
- Use sentiment scores as an additional spam indicator.

Application in Another Industry:

- Industry: E-commerce
- Use Case: Analyze customer reviews to detect dissatisfaction.
- Approach: Apply sentiment analysis to reviews, identify negative trends, and improve service based on feedback.