SMS Spam Detection

No Code AI and Machine Learning

05/08/2024

Contents / Agenda

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

Data Dictionary

- **Category**: Contains the labels 'spam' or 'ham' for the corresponding text data
- **Message**: Contains the SMS text data

How to use this deck?

- This slide deck serves as a comprehensive template for your project submission
- Within this deck, you will come across various questions that are intended to test your ability to understand data visualizations, discover patterns / insights and postulate hypothesis. Think thoroughly and provide answers to these questions
- You are encouraged to modify this deck as required, by replacing the questions with suitable answers
- Please feel free to incorporate additional points if you deem necessary

Note: The data visualizations you see in this deck are obtained from RapidMiner

Business Problem Overview and Solution Approach

- Building a classification model to distinguish between spam and legitimate SMS messages can definitely help enhance cybersecurity for businesses. Here's a step-by-step approach you could take:
- **Data Exploration:** Begin by exploring the labeled SMS dataset. Understand its structure, the distribution of spam vs. ham messages, and any patterns or trends that might exist in the data.
- **Data Preprocessing:** Preprocess the SMS text data to prepare it for modeling. This may involve tasks such as removing punctuation, converting text to lowercase, handling stop words, and tokenizing the text into individual words or tokens.
- **Feature Engineering:** Extract meaningful features from the preprocessed text data that can be used to train the classification model. This could include techniques like TF-IDF (Term Frequency-Inverse Document Frequency) vectorization or word embeddings.
- **Model Selection:** Choose appropriate machine learning algorithms for text classification, such as Naive Bayes, Support Vector Machines (SVM), or more advanced techniques like Random Forests or Gradient Boosting Machines (GBM). Experiment with different algorithms to see which ones perform best on your dataset.
- **Model Training:** Split the labeled dataset into training and testing sets, and train the chosen classification models on the training data.

- **Model Evaluation:** Evaluate the performance of each trained model using appropriate evaluation metrics such as accuracy, precision, recall, and F1-score. This will help you assess how well the models are able to distinguish between spam and ham messages.
- **Hyperparameter Tuning:** Fine-tune the hyperparameters of the best-performing models to further improve their performance.
- **Model Deployment:** Once you have a well-performing classification model, deploy it into production within the organization's cybersecurity infrastructure. This could involve integrating the model into existing security systems or creating a standalone application for SMS classification.
- **Continuous Monitoring and Improvement:** Regularly monitor the performance of the deployed model and update it as needed to adapt to new spamming techniques or changes in the SMS landscape.

By following these steps and leveraging the power of machine learning, you can help Cyber Solutions effectively combat SMS spam and enhance the security of businesses against cyber attacks.

EDA

- **5,572 Rows:** Each row in the dataset represents a review about a SMS.
- **2 Columns:** The columns / attributes in the dataset contain important

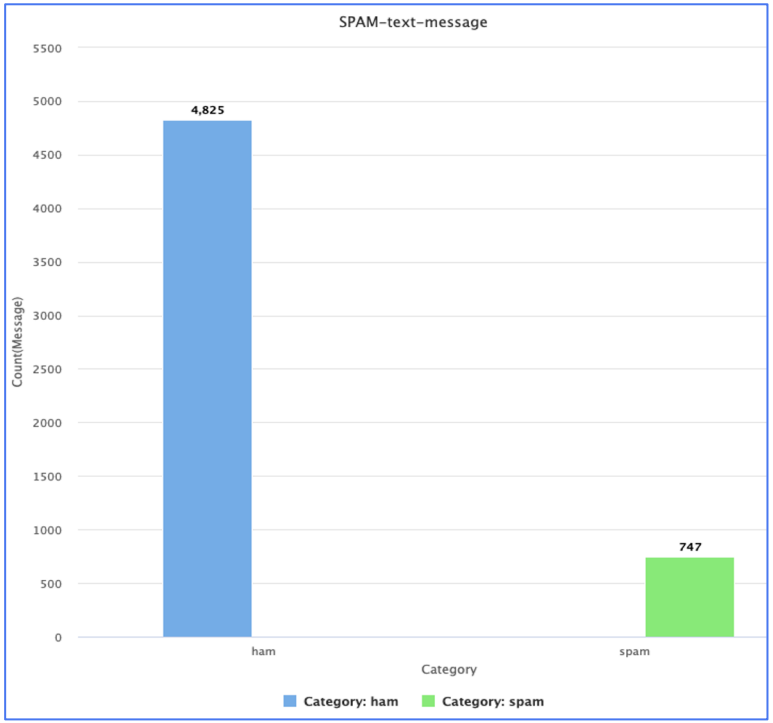
Open in Turbo Prep Auto Model Filter (5,572 / 5,572 examples): no_missing_attrib...

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 q...
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? ...
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 ...
9	spam	WINNER!! As a valued network customer you have been selected to receivea £900 prize reward! To clai...
10	spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for ...
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 to...
12	spam	SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, ...
13	spam	URGENT! You have won a 1 week FREE membership in our £100,000 Prize Jackpot! Txt the word: CLAIM...

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

EDA

- Around 86.5 percentage of ham and 13.5 spam messages in the dataset.
- Imbalanced datasets, where one class (in this case, spam) is significantly more prevalent than the other (ham), can lead to biased models that perform poorly on the minority class. Here are some strategies to handle imbalance:
    - **Resampling**: This involves either oversampling the minority class (e.g., duplicating samples) or under sampling the majority class (e.g., randomly removing samples) to balance the dataset.
    - **Synthetic Minority Over-sampling Technique (SMOTE)**: SMOTE generates synthetic samples for the minority class based on existing samples, effectively balancing the dataset.
    - **Evaluation Metrics**: Instead of relying solely on accuracy, consider using evaluation metrics like precision, recall, and F1-score, which provide a more comprehensive understanding of model performance on imbalanced datasets.
- By employing one or more of these strategies, you can mitigate the effects of class imbalance and build a more robust SMS spam classification model.



EDA: Word Cloud for Spam messages

Bag-of-Words (BoW):

In the Bag-of-Words model, a document is represented as a collection of words, disregarding grammar and word order. Each word is treated as a feature, and the frequency of each word in the document is counted. This results in a sparse matrix where each row represents a document and each column represents a unique word in the entire corpus. The value in each cell indicates the frequency of the corresponding word in the document.

Term Frequency-Inverse Document Frequency (TF-IDF):

TF-IDF is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents (corpus). It takes into account not only the frequency of a word in a document (term frequency) but also the inverse document frequency, which penalizes words that are common across the entire corpus. TF-IDF assigns higher weights to words that are frequent in a document but rare in the corpus, thus capturing their significance.

How it helps solve the problem statement:

These representations help solve the problem of text classification, such as distinguishing between spam and ham messages in SMS data, by converting textual data into numerical features that machine learning algorithms can understand and process. By representing text in this format, the models can learn patterns and relationships between words and use them to make predictions.

When is it used:

The Bag-of-Words model and TF-IDF representation are commonly used in natural language processing (NLP) tasks, including text classification, sentiment analysis, document clustering, and information retrieval. They are particularly useful when dealing with large volumes of textual data where understanding the semantics or context of the text is less important than identifying patterns based on word frequency or importance. These techniques are widely applied in various industries, including cybersecurity, marketing, finance, and healthcare, wherever textual data analysis is required.

Text Analysis

The text analysis technique used to find the frequencies of words occurring in documents is known as "Term Frequency (TF)."

In Term Frequency analysis, each document in a corpus is represented as a vector where each component corresponds to the frequency of a specific term (word) in that document. The frequency is typically calculated as the count of occurrences of each term normalized by the total number of terms in the document.

Terms like **Call, Get, Come** are **<u>most</u>** terms used in the dataset and **Love, Time, Want** are **<u>least</u>** used terms in the dataset.

Row No.	word	in documents	total ↓
1	call	480	542
2	get	399	429
3	come	264	283
	dont	244	270
	know	238	253
	free	197	241
	day	222	238
	love	199	237
	time	222	235
	want	224	235

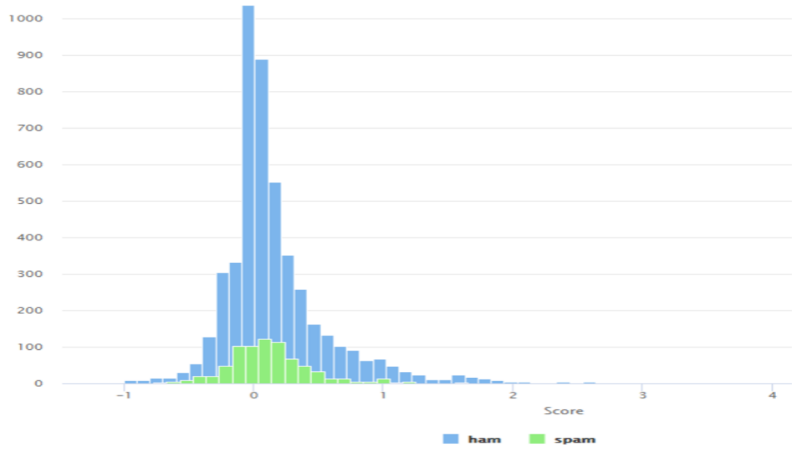
Text Analysis

- The table below is an example of the results we would get on applying SentiWordNet to every SMS Dataset.
- Each review has a positivity and negativity component score, and the final score is the sum of each component and this led to an overall score (sum) of **+0.173.**
- We can also see the words which contributed the most to the classifying the review as **positive**
- With this new Score column giving us a sentiment score for every SMS, we can now understand the distribution of the average sentiment score across various dimensions in this dataset to extract business insights from text.

Row No.	Score	Message	Category
1	0.173	Go until jurong point crazy Available only in bugis n great world la e bu...	ham
2	0.121	Ok lar Joking wif u oni	ham
3	0.243	Free entry in a wkly comp to win FA Cup final tkts st May Text FA to t...	spam
4	0.173	U dun say so early hor U c already then say	ham
5	0.295	Nah I don t think he goes to usf he lives around here though	ham
6	0.156	FreeMsg Hey there darling it s been week s now and no word back I ...	spam
7	0.763	Even my brother is not like to speak with me They treat me like aids pa...	ham
8	0.156	As per your request Melle Melle Oru Minnaminunginte Nurungu Vetta...	ham
9	0.173	WINNER As a valued network customer you have been selected to rec...	spam
10	-0.139	Had your mobile months or more U R entitled to Update to the latest ...	spam

Sentiment Scores for Spam and Ham texts

- Overall, analyzing sentiment scores of both spam and ham messages in SMS communication provides valuable insights for preventing cyber attacks, safeguarding customer interests, and maintaining trust in business communications. By leveraging sentiment analysis as part of their cybersecurity strategy, businesses can effectively mitigate the risks associated with SMS-based threats and strengthen their overall security posture



Model Performance Evaluation Decision Tree

- These metrics indicate that the Decision Tree model performs quite well on both the training and test sets. It achieves high accuracy, recall, and precision on both sets, which suggests that the model effectively distinguishes between spam and ham messages.
- However, it's also essential to consider other factors such as the dataset's balance, the potential presence of class imbalance, and the complexity of the model in interpreting these results accurately. Overall, these metrics demonstrate the model's strong performance in classifying SMS messages, with a slight drop in performance on the test set compared to the training set, which is expected.

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

Comparing the two:

- The decision tree before pruning has slightly higher accuracy, recall, and precision values compared to the pruned decision tree on both the training and test sets.
- After pruning, there's a slight reduction in accuracy, recall, and precision, indicating that the pruned decision tree may not generalize as well to unseen data as the unpruned version.
- However, pruning typically helps in reducing overfitting and simplifying the model, which can improve its interpretability and efficiency, especially when dealing with very large or complex decision trees.
- Overall, the decision to prune or not depends on the trade-off between model complexity and performance on unseen data. In this case, while pruning slightly reduces model performance, it may still be beneficial if it helps improve the model's generalization and interpretability.

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

- Overall, the Random Forest model performs well on both the training and test sets, with high accuracy and precision. However, there's a slight drop in recall on the test set compared to the training set, indicating that the model may struggle to correctly identify spam messages in the test data. This could be due to factors such as class imbalance or the complexity of the dataset. Evaluating additional metrics and considering potential adjustments to the model may help improve its performance further.

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

Comparing the two:

- After pruning, there's a slight improvement in train accuracy, but a slight reduction in test accuracy, suggesting that pruning may not significantly impact generalization performance.
- There's a noticeable improvement in train recall after pruning, indicating that the pruned model is better at identifying positive instances (spam) in the training data. However, there's a decrease in test recall, suggesting that the pruned model may not generalize as well to unseen data.
- Train precision remains relatively stable after pruning, but there's a slight decrease in test precision, indicating that the pruned model may be less precise in classifying positive instances (spam) in the test data.
- Overall, the decision to prune or not depends on the trade-off between model complexity and generalization performance. In this case, while pruning slightly improves performance on the training set, it may not necessarily lead to significant improvements on the test set. It's essential to carefully evaluate the impact of pruning on both train and test performance before making a decision.

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 Summary

Based on the provided metrics, both Decision Tree and Random Forest models have been evaluated before and after pruning. Let's analyze which model best fits the company's objective and which evaluation metric is most important to achieve that objective:

Company Objective: The company aims to prevent cyber attacks on businesses through SMS messages by accurately classifying them as spam or ham.

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

Model Evaluation:

- **Decision Tree:**
  - Before Pruning: Achieves high accuracy, recall, and precision on both training and test sets.
  - After Pruning: Slight decrease in performance metrics, especially in test precision and recall.
- **Random Forest:**
  - Before Pruning: Achieves slightly lower recall compared to Decision Tree, but similar precision and accuracy.
  - After Pruning: Improves train accuracy and recall but shows a decrease in test recall and precision.
- **Evaluation Metric:** Given the company's objective to prevent cyber attacks, Precision is the most important evaluation metric. Precision measures the percentage of correctly identified spam messages out of all messages classified as spam by the model. Maximizing precision is crucial because misclassifying legitimate messages as spam (false positives) can inconvenience users, but allowing spam messages through (false negatives) can potentially lead to security breaches or cyber attacks.
- **Model Recommendation:** Based on the importance of precision and considering both before and after pruning, the Decision Tree model before pruning appears to be the best fit for the company's objective. It achieves high precision on the test set (93.50%) while maintaining strong overall performance in accuracy and recall. The Random Forest model, although competitive, does not outperform the Decision Tree in terms of precision, especially after pruning.

Suggestions and Advice:

- **Deployment and Monitoring:** Deploy the selected Decision Tree model before pruning into production for real-time classification of SMS messages. Continuously monitor its performance and retrain the model periodically to adapt to evolving spamming techniques.
- **User Feedback Loop:** Implement a feedback mechanism where users can report misclassified messages. Incorporate this feedback into model updates to improve its accuracy and precision over time.
- **Ensemble Approaches:** Consider ensemble methods like Random Forest for further experimentation and model improvement. Ensemble methods combine multiple models to achieve better performance and robustness, which could be beneficial in this context.
- **Feature Engineering:** Explore additional features such as message metadata (sender information, timestamps) or linguistic features (message length, language patterns) to enhance model performance and detection capabilities.
- **Human-in-the-Loop:** Although machine learning models can automate much of the classification process, consider incorporating human review or intervention for critical decisions or cases where the model's confidence is low. Human oversight can provide an additional layer of security and accuracy in detecting potential threats.

Insights and Recommendations

Certainly! Here are some actionable insights and recommendations based on the analysis of the models and evaluation metrics:

- Implement **"Decision Tree Model Before Pruning"**
- **Actionable Insight:** Deploy the Decision Tree model before pruning into the company's cybersecurity infrastructure for real-time classification of SMS messages.
- **Recommendation:** Utilize the Decision Tree model's high precision (93.50%) to accurately identify spam messages and prevent potential cyber attacks.
- **Action:** Integrate the model into the existing SMS filtering system to automatically flag and block suspicious messages.

- Overall, sentiment analysis provides valuable insights into customer opinions, market trends, brand perception, and public sentiment across diverse industries. By leveraging sentiment analysis techniques, organizations can make informed decisions, improve customer experiences, mitigate risks, and drive business growth.

Here's how sentiment analysis can be utilized and applied across different sectors:

- **Customer Feedback Analysis:**
  - Retail: Analyze product reviews and customer feedback to understand customer sentiment towards products and improve product offerings.
  - Hospitality: Monitor guest reviews and sentiment on platforms like TripAdvisor to identify areas for improvement in service quality.
  - E-commerce: Analyze customer sentiment on social media and review platforms to identify trends and inform marketing strategies.
- **Risk Management and Fraud Detection:**
  - Insurance: Analyze sentiment in customer communications and claims data to identify potential fraudulent activities and mitigate risks.
  - Banking: Monitor sentiment in financial news and social media to detect market sentiment shifts and potential financial risks.

Overall, sentiment analysis provides valuable insights into customer opinions, market trends, brand perception, and public sentiment across diverse industries. By leveraging sentiment analysis techniques, organizations can make informed decisions, improve customer experiences, mitigate risks, and drive business growth.

- Let's consider the healthcare industry and focus on the issue of patient satisfaction and experience. One prospective issue that could be resolved through text processing is the analysis of patient feedback from various sources such as surveys, reviews, and comments.

Problem Statement:

- The healthcare provider wants to improve patient satisfaction and experience but faces challenges in analyzing and extracting insights from the large volume of unstructured textual data, including patient surveys, feedback forms, and online reviews.

Approach to Solving the Problem:

- 1. Data Collection: Gather patient feedback data from various sources, including surveys conducted after medical visits, feedback forms provided at healthcare facilities, and online reviews on platforms like Google Reviews or Healthgrades.

Text Preprocessing:

- Clean the text data by removing punctuation, special characters, and irrelevant information.
- Tokenize the text into words or phrases and convert them to lowercase for consistency.
- Remove stop words (commonly used words like "the," "and," "is") that do not contribute much to the sentiment analysis.
- Apply techniques like stemming or lemmatization to reduce words to their root form.

Sentiment Analysis:

- Utilize sentiment analysis techniques to assign sentiment scores (positive, negative, neutral) to patient feedback.
- Choose a sentiment analysis approach suitable for healthcare text data, considering the specific terminology and language used in medical contexts.
- Use pre-trained sentiment analysis models or train custom models on labeled healthcare text data to accurately classify sentiment.

Topic Modeling:

- Apply topic modeling techniques such as Latent Dirichlet Allocation (LDA) or Non-negative Matrix Factorization (NMF) to identify common themes or topics in patient feedback.
- Extract topics such as staff behavior, wait times, facility cleanliness, communication, and treatment effectiveness from the feedback data.

Insights and Actionable Recommendations:

- Analyze the sentiment distribution and identify areas of improvement based on sentiment scores.
- Identify recurring topics or themes in patient feedback to prioritize improvement initiatives.
- Generate actionable recommendations for healthcare providers to address specific issues and enhance patient satisfaction and experience.
- Monitor changes in sentiment and topic distribution over time to evaluate the effectiveness of improvement efforts.

Integration and Feedback Loop:

- Integrate the text processing and sentiment analysis pipeline into the healthcare provider's feedback management system.
- Establish a feedback loop where insights from text analysis inform ongoing improvement efforts and guide decision-making in patient experience initiatives.

By implementing this approach, healthcare providers can gain valuable insights from patient feedback, identify areas for improvement, and take proactive steps to enhance patient satisfaction and experience. This, in turn, can lead to improved patient outcomes, increased patient loyalty, and a positive reputation for the healthcare organization.

APPENDIX

greatlearning
Power Ahead

Happy Learning !