# Data Exploration – Unstructured Data

## No code AI/ML - MIT

06/08/2024

Presented by : Nehal Naik

# 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

# Executive Summary

**Spam via SMS**: While use of SMS is inevitable as an effective communication tool, its misuse is also on rise. Spam is the abuse of electronic messaging systems to send unsolicited messages in bulk indiscriminately and spreading phishing links. This has resulted in many people's bank account and identity being hacked.

**Effective warning**: If there exist an automated system, which can detect incoming spam and warn user about the possibility of received message being spam, that can help lot of people save time, reduce anxiety and avoid hacking risk.

**Machine Learning Solution**: Data scientist at Cyber Solutions are tasked with finding a solution to process incoming SMS and flag potential harmful SMS as spam using Machine Learning model. As SMS are considered unstructured data (Text), it requires special treatment to work with it. Text messages needs to be somehow converted into the format on which mathematical classification models can be applied.

# Business Problem Overview and Solution Approach

**Problem statement:**

- **Surge in Spam SMS:** SMS technology is being misused for unethical purposes like bulk marketing blasts and phishing scams, putting personal information and accounts at risk.

- **Disrupted Communication:** While SMS remains a vital communication channel for both personal and business use, spam messages are causing frustration and a decline in user trust.

- **Response:** Recognizing the urgency of the situation, there's a pressing need to find a solution that can flag potential spam messages and establish more confidence in users of SMS.

- **Solution:** There are many tools available in ML area to solve this problem. Looking at the need for decision making flagging SMS, one of the classification model can be used.

# Business Problem Overview and Solution Approach

**Proposed solution**

- Basic Exploratory Data Analysis (EDA) needs to be performed to find obvious connection and correlations among captured attributes to identify patterns.
- Before we apply any ML model, unstructured SMS data needs to be prepared using embedding technique to be processed by classification model.
- Looking at the problem statement and available data points, variation of Decision Tree methodology will provide much needed solution to predict likelihood of booking cancellation.
- From available data set, 70% of the records will be used as training set, while 30% will be used at testing set.
- Based on the performance of the methodology, final solution will be picked to be used.
- Will measure success of the final solution for next 6 months to decided if further improvement is needed.

# EDA and Text Visualization

Open in Turbo Prep Auto Model

| Row No. | word | in documents | total |
|---|---|---|---|
| 1 | A | 1 | 1 |
| 2 | As | 1 | 1 |
| 3 | Burns | 1 | 1 |
| 4 | C | 2 | 2 |
| 5 | CASH | 1 | 1 |
| 6 | CLAIM | 1 | 1 |
| 7 | CSH | 1 | 1 |
| 8 | Call | 1 | 1 |
| 9 | Claim | 1 | 1 |
| 10 | Co | 1 | 1 |
| 11 | Cost | 1 | 1 |
| 12 | Cup | 1 | 1 |
| 13 | ENGLAND | 1 | 1 |
| 14 | England | 1 | 1 |
| 15 | FA | 1 | 2 |
| 16 | FREE | 2 | 2 |

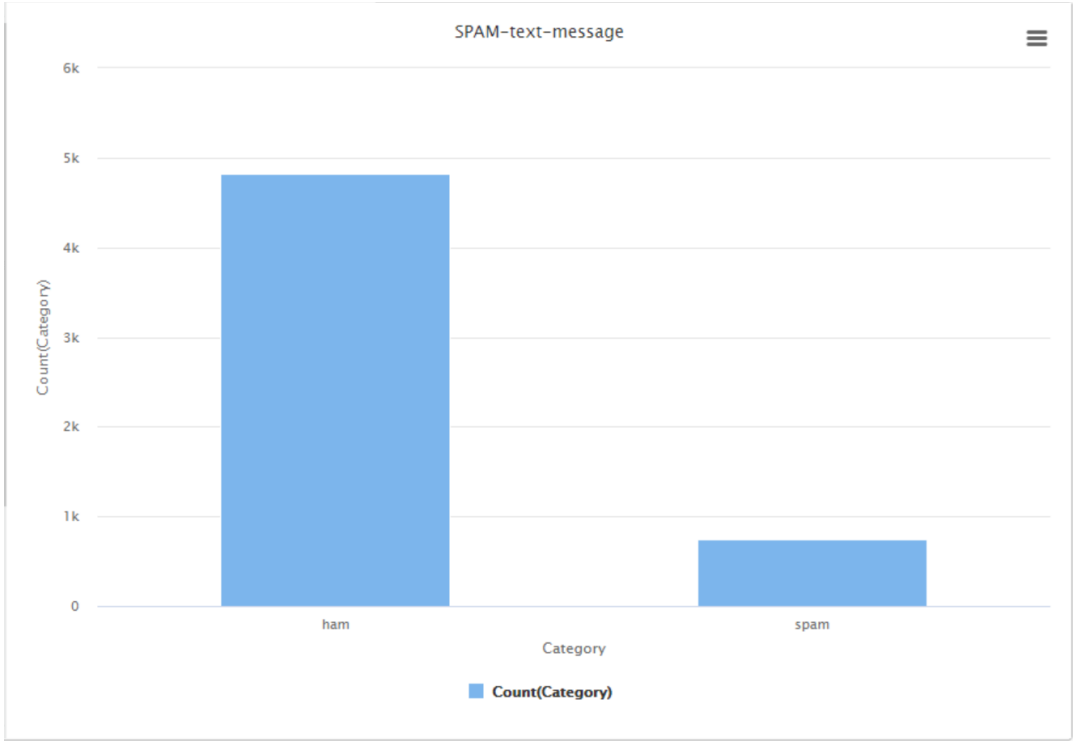
ExampleSet (187 examples, 0 special attributes, 3 regular attributes)

- Table on the left shows Term Frequency and Inverse Document Frequency (TF-IDF) which reflects importance of a word in a document relative to a collection of documents.
- This information helps in feature extraction, used by search engine to rank documents by relevance and identify most significant terms in set of documents.
- It will also use to identify Stop

# EDA and Text Visualization

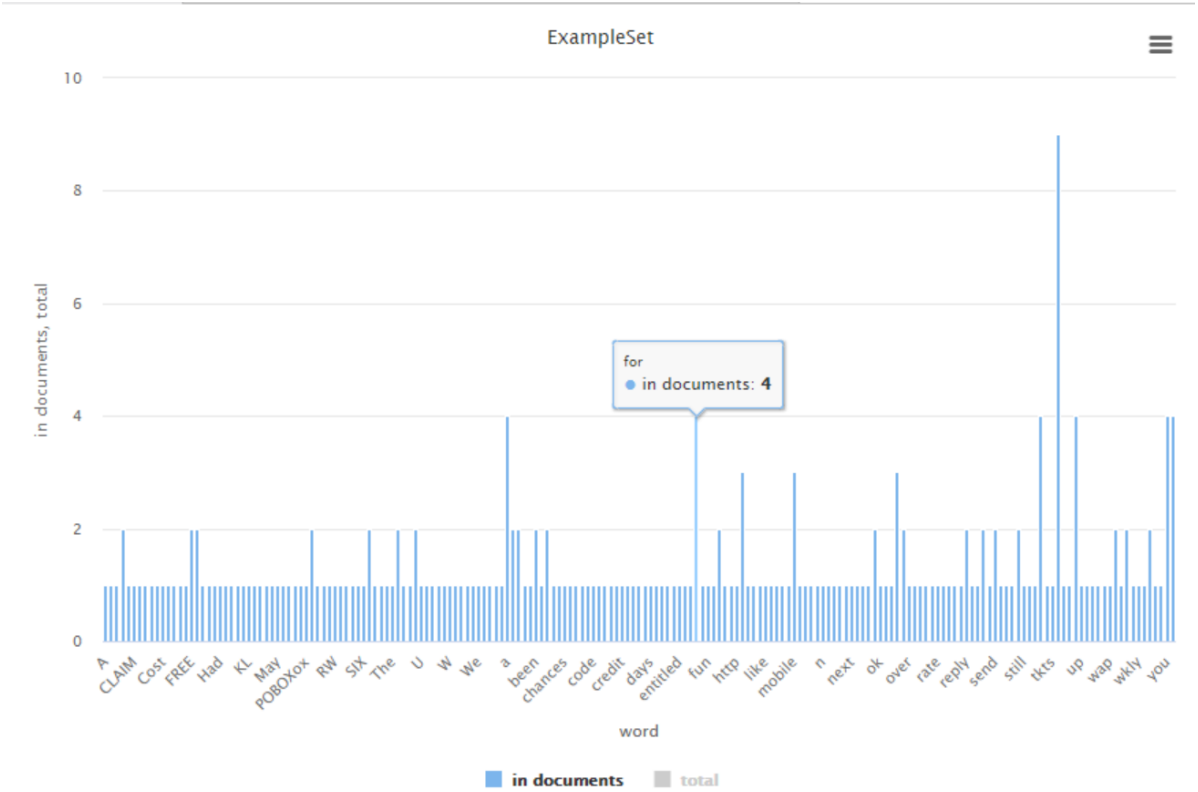

- Word cloud is prepared after tokenizing and stemming of SMS.
- This provides set of important words which will then be converted into vector for further processing.

# EDA and Text Visualization



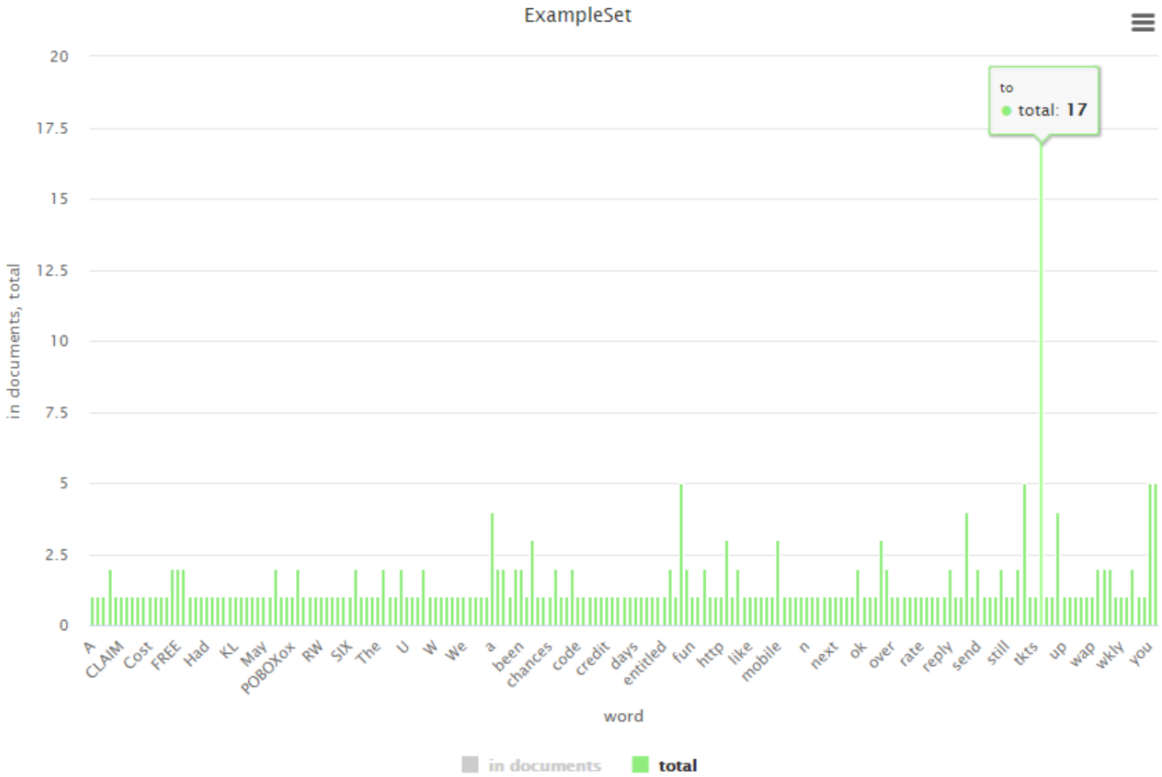

- This graph shows difference in number of spam vs ham in a given data set.

# EDA and Text Visualization




- This graph shows frequency of each word in given set of spammed SMS.
- This is used to perform StopWord removal, where common words like “in, for, a, the” is removed to reduce vector size.

# EDA and Text Visualization




- This graph shows total words frequency in both spam and ham messages.

# Text Analysis : Sentiment Analysis

| Row No. | Score | Category | Scoring String | Negativity | Positivity | Unco... | Total ... | Message |
|---|---|---|---|---|---|---|---|---|
| 38 | 0.139 | ham | see (0.12) on (0.02) | 0 | 0.139 | 6 | 8 | I see the l... |
| 39 | 0 | ham | | 0 | 0 | 5 | 5 | Anything l... |
| 40 | 0.260 | ham | go (0.10) just (0.26) see (0.1... | 0.243 | 0.503 | 30 | 36 | Hello Ho... |
| 41 | 0.401 | ham | go (0.10) ahead (0.03) just (0... | 0.511 | 0.911 | 12 | 20 | Pls go ah... |
| 42 | 0.815 | ham | forget (-0.07) tell (0.07) want ... | 0.433 | 1.248 | 32 | 41 | Did I forg... |
| 43 | -0.156 | spam | call (0.05) free (-0.19) mobile... | 1.023 | 0.867 | 33 | 40 | Rodger... |
| 44 | -0.035 | ham | seeing (-0.03) | 0.035 | 0 | 3 | 4 | WHO AR... |
| 45 | 1.538 | ham | great (0.02) hope (0.10) like (... | 0.202 | 1.740 | 14 | 20 | Great I h... |
| 46 | -0.260 | ham | no (-0.19) missed (-0.07) | 0.277 | 0.017 | 3 | 5 | No calls ... |
| 47 | 0.104 | ham | get (0.10) | 0.052 | 0.156 | 8 | 9 | Didn t you... |

- Sentiment analysis is performed using SentiWordNet 3.0 sentiment dictionary (https://sentiwordnet.isti.cnr.it/) to score the text.
- This table shows scoring of each word in a sentence and total positivity score.
- Lower score can result in message begin flagged as Spam.

# Text Analysis - how is it important?

- Provide appropriate details on which technique is used to find the sentiment scores
- Comment on the sentiment scores of the Spam and Ham texts
- Comment on how the TF-IDF technique is used and derive valuable insights

# Model Performance Summary (Decision Tree)

- Following tables clearly shows that Precision and Class Recall remain in line when model is run on Test data set.
- Overall accuracy on test data is reported at 97% which is very good.
- As Training and Test data both are reporting 97% accuracy, this model performs well.

**Decission Tree - Training Set**

| | true ham | true spam | **class precision** |
|---|---|---|---|
| pred. ham | 3826 | 90 | 97.70% |
| pred. spam | 34 | 508 | 93.73% |
| **class recall** | 99.12% | 84.95% | 97% |

**Decission Tree - Testing Set**

| | true ham | true spam | **class precision** |
|---|---|---|---|
| pred. ham | 950 | 21 | 97.84% |
| pred. spam | 15 | 128 | 89.51% |
| **class recall** | 98.45% | 85.91% | 97% |

# Model Performance Summary (Decision Tree – Pruned)

- Following tables clearly shows that Precision and Class Recall for spam drops around 15% with both training and test data.
- Overall accuracy on test data is reported at 95% which is less than non-pruned version of decision tree provided.
- As we are focusing on flagging Spam, having 95% recall and precision on test data does provide confidence in the model, but at the same time is less that non-pruned version.

**Decission Tree Pruned- Training Set**

| | true ham | true spam | **class precision** |
|---|---|---|---|
| pred. ham | 3814 | 90 | 97.69% |
| pred. spam | 46 | 508 | 91.70% |
| **class recall** | 98.81% | 84.95% | 97% |

**Decission Tree Pruned - Testing Set**

| | true ham | true spam | **class precision** |
|---|---|---|---|
| pred. ham | 944 | 34 | 96.52% |
| pred. spam | 21 | 115 | 84.56% |
| **class recall** | 97.82% | 77.18% | 95% |

# Model Performance Summary (Random Forest)

- Following tables shows that Precision and Class Recall remained very strong for flagging Spam when running model on test data.
- Overall accuracy on test data is reported at 98% which is the best among the four models ran so far. Difference between Training and Test is nonexistence, which indicates good performance on future unseen data.

**Random Forest - Training Set**

| | true ham | true spam | **class precision** |
|---|---|---|---|
| pred. ham | 3860 | 81 | 97.94% |
| pred. spam | 0 | 517 | 100.00% |
| **class recall** | 100.00% | 86.45% | 98% |

**Random Forest - Testing Set**

| | true ham | true spam | **class precision** |
|---|---|---|---|
| pred. ham | 961 | 23 | 97.66% |
| pred. spam | 4 | 126 | 96.92% |
| **class recall** | 99.59% | 84.56% | 98% |

# Model Performance Summary (Random Forest – Pruned)

- Following tables shows that Recall for flagging Spam is very low (75.84%)while running model on test data. This may not provide much comfort.
- Overall accuracy on test data is reported at 96% which is very good, and it is also not much different than training set.

**Random Forest - Pruned - Training Set**

| | true ham | true spam | **class precision** |
|---|---|---|---|
| pred. ham | 3838 | 88 | 97.76% |
| pred. spam | 22 | 510 | 95.86% |
| **class recall** | 99.43% | 85.28% | 98% |

**Random Forest - Pruned - Testing Set**

| | true ham | true spam | **class precision** |
|---|---|---|---|
| pred. ham | 953 | 36 | 96.36% |
| pred. spam | 12 | 113 | 90.40% |
| **class recall** | 98.76% | 75.84% | 96% |

# Executive Summary

**Model development and selection results:**

- Upon evaluating available data and problem statement, classification technique is used to flag SMS as potential Spam or Ham as a part of the solution.

- Based on availbale historical data, 4 different classification models were tested and one was finalzed with better outcome in terms of better overall accuracy and other parameteres.

- **Random forest** based ML model was finalized based on the its accuracy to predict likelyhood of cancelation and for having 96% of overall accracy and minimal difference between results fo training set and testing set.

- Cyber Solution should try out this solution for few months and see how it performs. Since other models are also providing good accuracy, those can be tries if Random Forest does not provide required accuracy.

# APPENDIX

# Model Building - Decision Tree / Random Forest

Based on the test conducted, Random forest and Decision Tree both models behaved better. But Random Forest provided more accurate classification.

## Overall Accuracy

| | Decision Tree | Decision Tree - Prune | Random Forest | Random Forest - Prune |
|---|---|---|---|---|
| Training | 97.22% | 96.95% | 98.18% | 97.53% |
| Test | 96.77% | 95.06% | 97.58% | 95.69% |

Great Learning

# Happy Learning !