Project Presentation

Spam Ham Predictions

No Code AI & Machine Learning

10/05/2025

Amit Mangotra

Contents / Agenda

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

Executive Summary

- **Confidentiality policies will determine whether a NLP prediction model can be deployed or only word clouds can be used**
  - ❑ A prediction model can be deployed only by building an application that will screen and filter spam messages. It seems most messages are of a personal nature. Employees may have consented to sharing some messages for training data, but will employees consent to sharing all messages? Maybe one can mask phone number or use discreet attributes like message length, suspicious links for prediction. For employees who do not consent, word clouds would be a viable alternative.
  - ❑ Word Clouds can be communicated to employees warning them of keywords used in phishing texts
  - ❑ **Treatment of messages classified as spam:** Will spam classified messages be delivered with a warning, will they be stored somewhere for employee to view or will they will be completely blocked?
- **Choice of Performance Measure for Model: Recall primary with Precision as secondary**
  - ❑ Given the objective, **Recall seems** the most important measure as one does not want to receive text messages which are spams, assume they are genuine, and get scammed.
  - ❑ ***However,*** *Precision can not be ignored as* one does not want genuine and important messages, like a meeting invite or an urgent client request, to be blocked just because they were classified as spam

Executive Summary

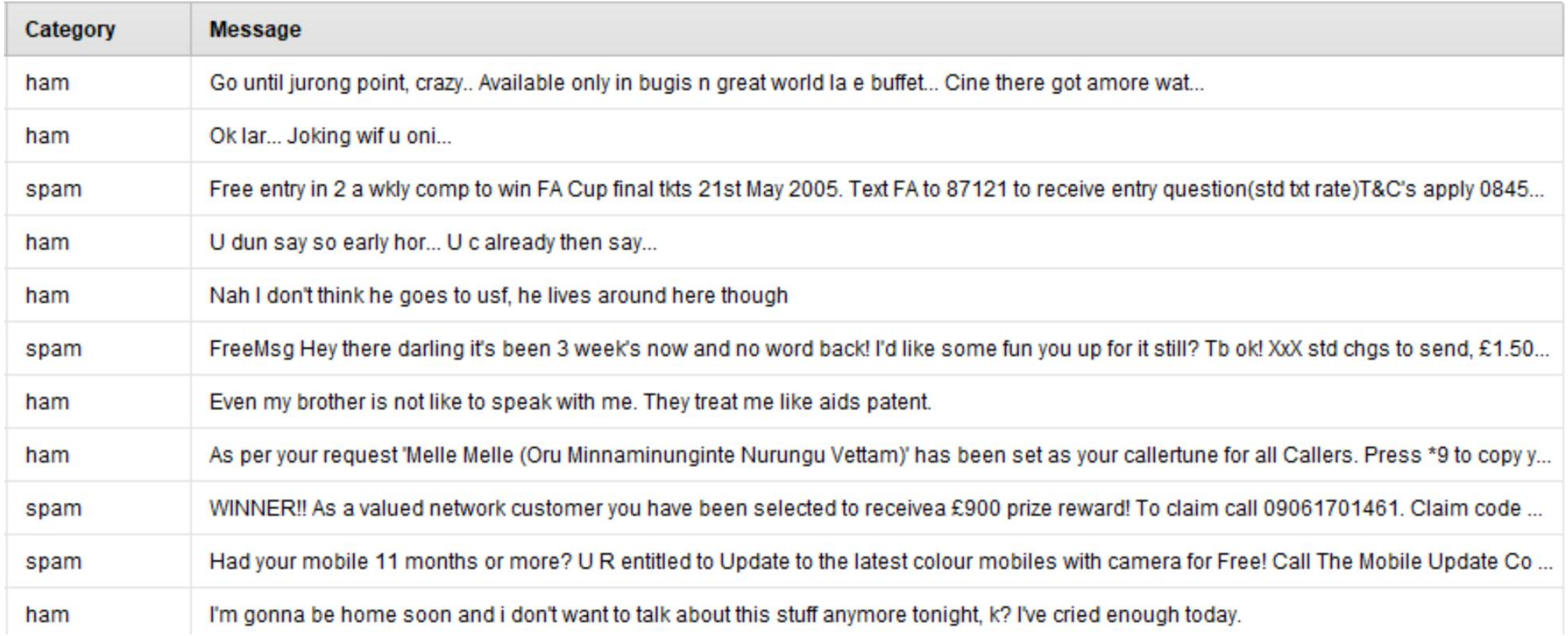
- **Random Forest No Pruning Model is the chosen NLP model although other three models have also performed well**
  - ❑ Chosen model generalizes well and has high test scores for both recall and precision, 90% and 98%.
  - ❑ Three out of four models, including the chosen model, have 100% precision score on training data. Normally, a 100% score is not advisable. Still, the model is chosen since Recall takes precedence over Precision even though both criteria are taken into account for selecting a model.
- **Use Word Cloud in conjunction with model and evaluate alternate model using Sentiment Score**
  - ❑ As highlighted earlier, if one is planning to completely block texts classified as spam, it is critical to be almost 100% sure. So, sentiment score model can be also be evaluated or can also be used to further validate the NLP model results: genuine messages are more likely to have extreme scores, either very low or very high
  - ❑ Word Clouds are an excellent means to alert employees of keywords used in phishing attacks. They are easiest to implement and take care of any confidentiality issues. Word clouds can be communicated easily in an email to all employees. However, for continuously updating word clouds, employees will have to consent to sharing text messages from time to time.

Business Problem Overview and Solution Approach

- The business objective is to prevent SMS cyber attacks on employee phones. These SMSs can scam employees and make them share confidential information.
- **Solution approach / methodology**
  - ❑ EDA, Text Analysis using Word Cloud and Sentiment Scoring
  - ❑ Data Preprocessing (not required here)
  - ❑ Model Building using Decision Trees, Random Forests
  - ❑ Model Performance Evaluation
  - ❑ Insights and Recommendations
- **Implementation Challenges:** Will an application be built that will screen and filter spam messages? Will employees consent to such a system? How to prevent attacks for employees who do not consent? How to prevent attacks while application is being built? Will spam messages be completely blocked?

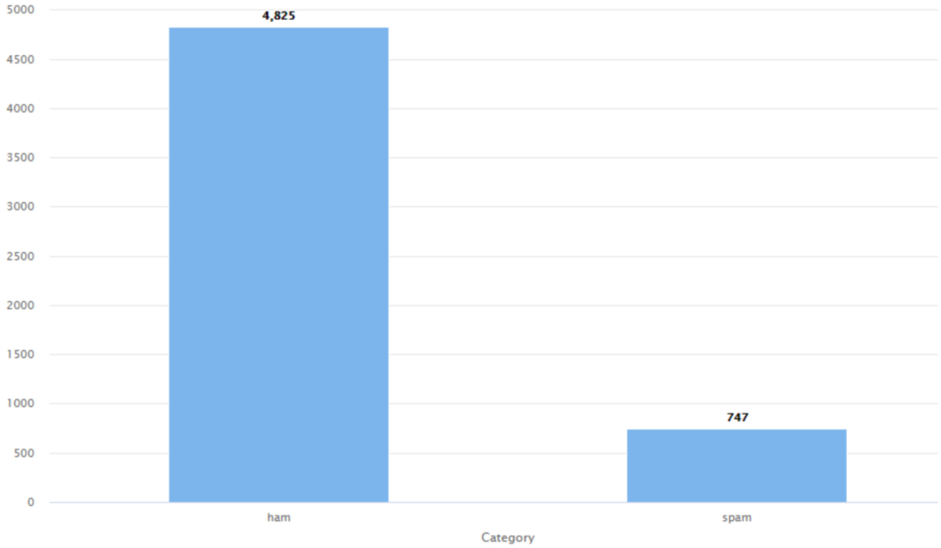
Data Dictionary

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



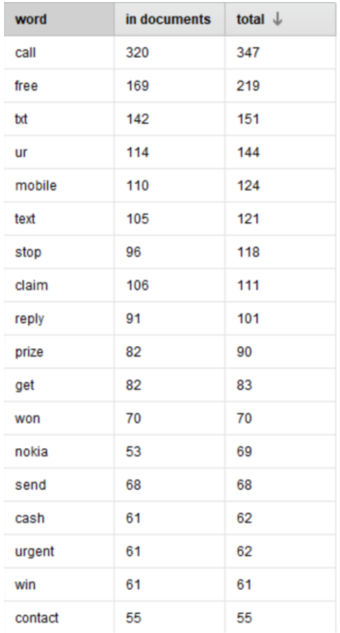
Category	Message
ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
ham	Ok lar... Joking wif u oni...
spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 0845...
ham	U dun say so early hor... U c already then say...
ham	Nah I don't think he goes to usf, he lives around here though
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...
ham	Even my brother is not like to speak with me. They treat me like aids patent.
ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy y...
spam	WINNER!! As a valued network customer you have been selected to receivea £900 prize reward! To claim call 09061701461. Claim code ...
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 ...
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.

EDA – Univariate analysis

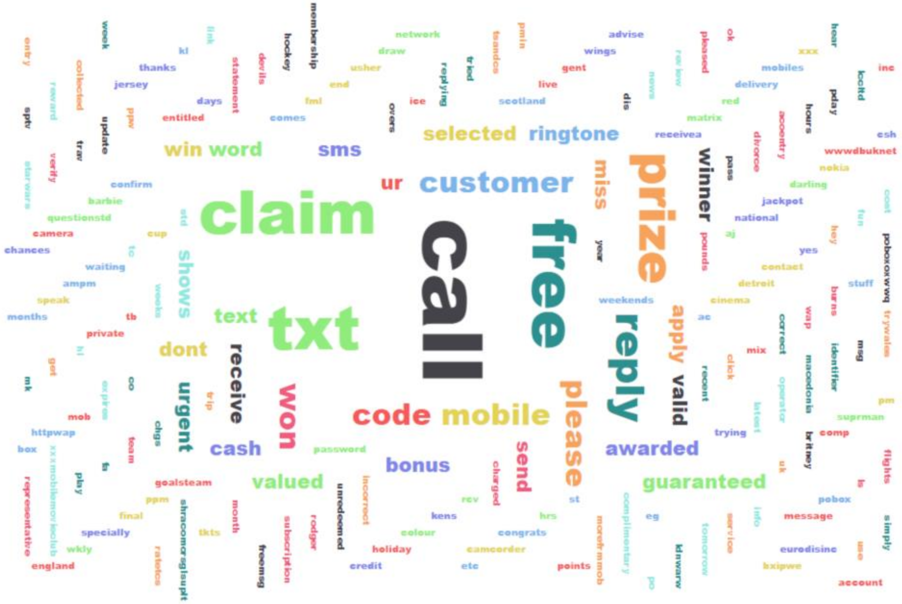


- The dataset contains two attributes: messages and category
- **Class Distribution:** About 13% of messages are spams. To handle this uneven class distribution, stratified sampling would be used for splitting train and test data. Moreover, accuracy would NOT be used as a performance measure

EDA - Text Visualization - Word Cloud for Spams



word	in documents	total ↓
call	320	347
free	169	219
txt	142	151
ur	114	144
mobile	110	124
text	105	121
stop	96	118
claim	106	111
reply	91	101
prize	82	90
get	82	83
won	70	70
nokia	53	69
send	68	68
cash	61	62
urgent	61	62
win	61	61
contact	55	55



- Using TF-IDF, key words in spam messages can be identified and a word cloud can be created. TF-IDF gives weights not only to occurrences but also to proportion of occurrences in all documents

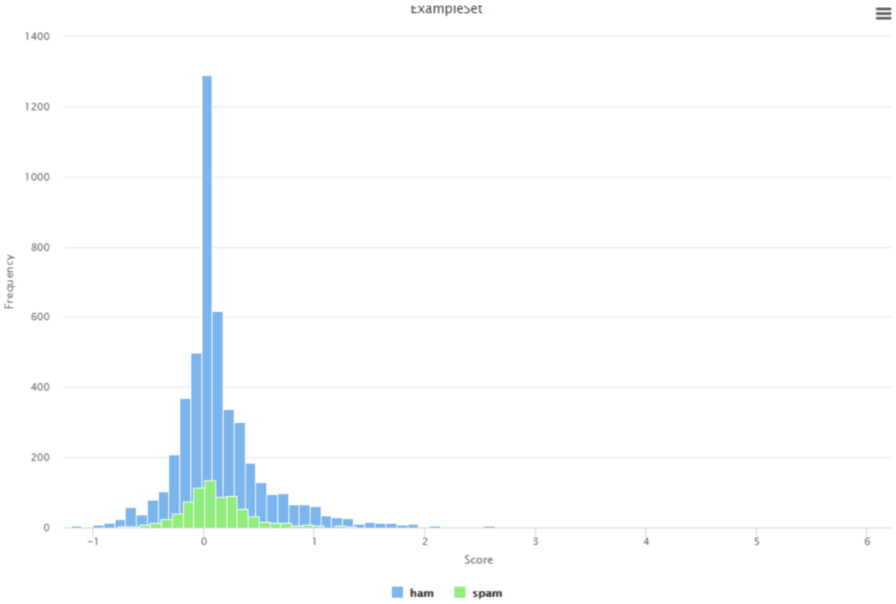
- The word cloud can be communicated to employees alerting them of possible phishing. The employees can also voluntarily share such messages with Data Science team to fine-tune the model

EDA - Sentiment Scores for Spam and Ham texts

- SentiWordNet has been used for scoring sentiment after removing all special characters (except £) and whitespaces. The '£' sign features in many spam messages. Hence it is retained.

- Ham and Spam Sentiment Scores have a similar distribution except one difference: at higher and lower score values, likelihood of hams is much higher.

- In addition to NLP model spam predictions, these extreme sentiment score values can also be used for spam detection using Decision Trees/Random Forest.



Model Building – using Stratified Sampling

Since our prediction class is not evenly distributed i.e. number of spam records is much less than 50% of total records, use stratified sampling for train and test data

Accuracy would not be used as a performance measurement parameter due to this imbalance in class distribution

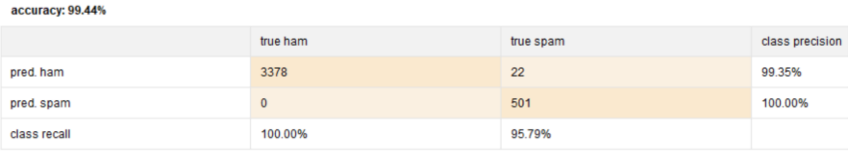
Model Performance - Decision Tree / No Pruning

- Interestingly, using Information Gain and Gain Ratio give different recall results with same *Max Tree Depth* of 20.
- **Information Gain model is a better model as test Recall and Precision values are around 90% and 87%.**
- **Info Gain model is also good at generalizing** as about 5pp difference in Recall scores for test and train data. However, it is not generalizing as well for Precision. But generalizes well on F score.

Information Gain

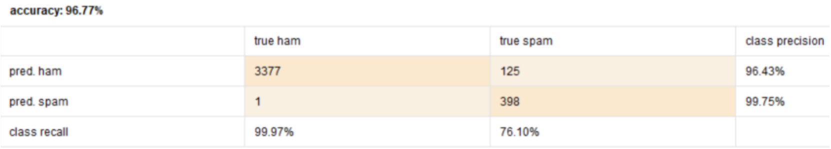
Gain Ratio

Train



accuracy: 99.44%

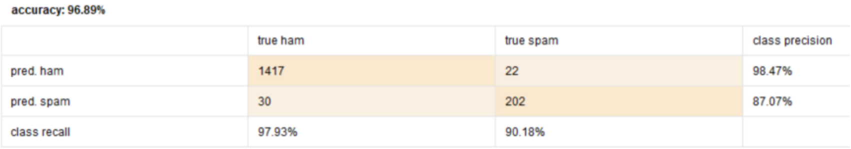
	true ham	true spam	class precision
pred. ham	3378	22	99.35%
pred. spam	0	501	100.00%
class recall	100.00%	95.79%	



accuracy: 96.77%

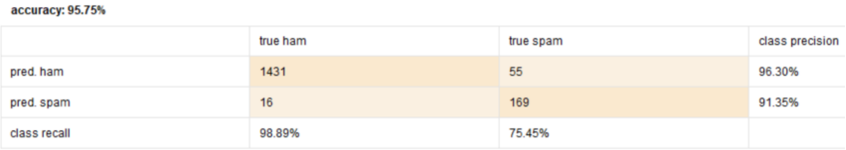
	true ham	true spam	class precision
pred. ham	3377	125	96.43%
pred. spam	1	398	99.75%
class recall	99.97%	76.10%	

Test



accuracy: 96.89%

	true ham	true spam	class precision
pred. ham	1417	22	98.47%
pred. spam	30	202	87.07%
class recall	97.93%	90.18%	



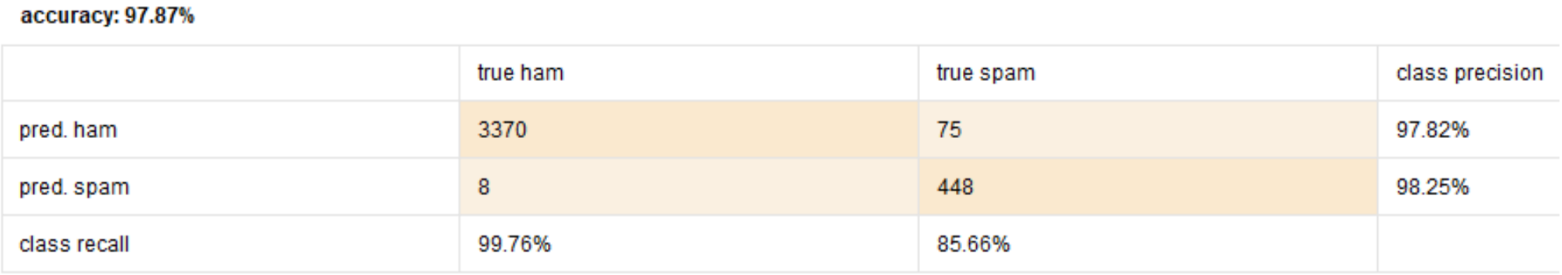
accuracy: 95.75%

	true ham	true spam	class precision
pred. ham	1431	55	96.30%
pred. spam	16	169	91.35%
class recall	98.89%	75.45%	

Model Performance - Decision Tree / Pruned

- Criterion: Information Gain, Maximal Tree Depth=15, Confidence = 0.1
- Pruned Model has test recall and precision values of around 78% and 89%
- Model generalizes well as Train vs Test Recall and Precision do not differ by more than 10 pp although a difference of less than 5pp would be ideal

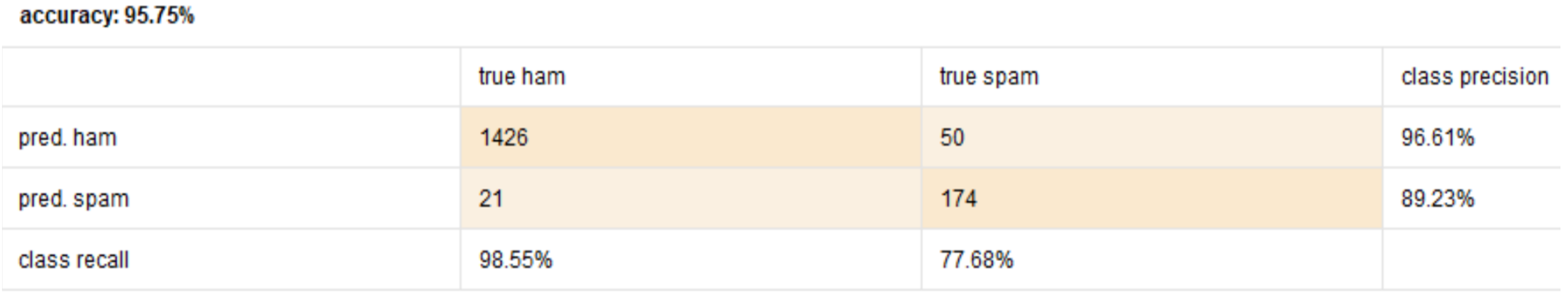
Train



accuracy: 97.87%

	true ham	true spam	class precision
pred. ham	3370	75	97.82%
pred. spam	8	448	98.25%
class recall	99.76%	85.66%	

Test



accuracy: 95.75%

	true ham	true spam	class precision
pred. ham	1426	50	96.61%
pred. spam	21	174	89.23%
class recall	98.55%	77.68%	

Model Performance Summary – Decision Tree

- **Choice of Performance Measure: Recall, Precision or Recall with Precision as secondary**

  ❑ From Problem statement, the objective is to identify reduce cyber attacks on employees' phones by identifying spams. So, **Recall seems** the most important measure as one does not want to receive text messages which are spams, assume they are genuine, and get scammed,

  ❑ ***However,*** *Precision can not be ignored as* one does not want to miss genuine and important messages, like a meeting invite or an urgent client request, just because they were classified as spam. Unlike emails which can still be viewed in a separate folder, spam texts may never be seen.

- **It is not a straightforward choice between the two but model without pruning is selected**

  ❑ Model without pruning has higher Recall values and generalize well on recall but has over 10pp difference in precision values for test and train data and does not generalize as well on precision

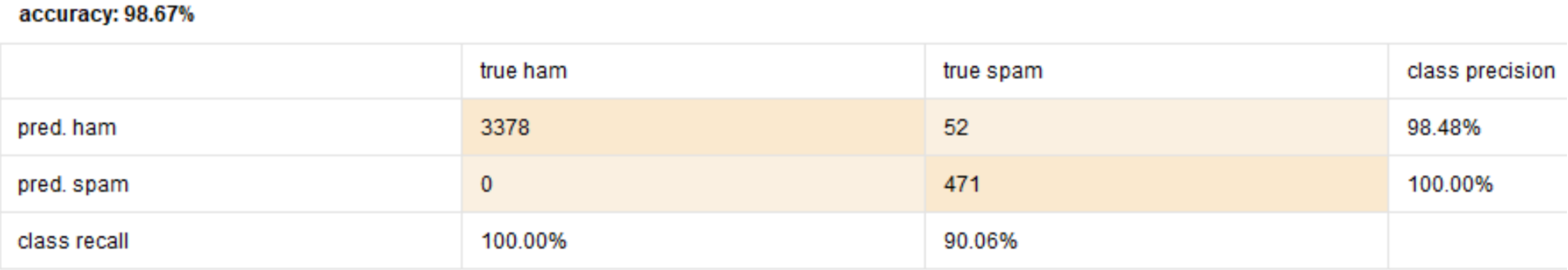
  ❑ On the other hand, model with pruning generalizes better on precision as it has lower difference in test vs train values

  ❑ Since Recall is more important than Precision, and model without pruning has higher recall test value while precision test value is also 87%, it is chosen among the two models.

Model Performance – Random Forest/ No Pruning

- Model built using No of trees =100, Criterion = Information Gain, Max Tree Depth = 40
- Recall values are 90%. Model Generalizes well because train vs test score recall values have a difference of less than 5pp. Moreover, test recall value is higher than train value.
- Precision values are 100% and 98% indicating that the model rarely misclassifies a genuine text

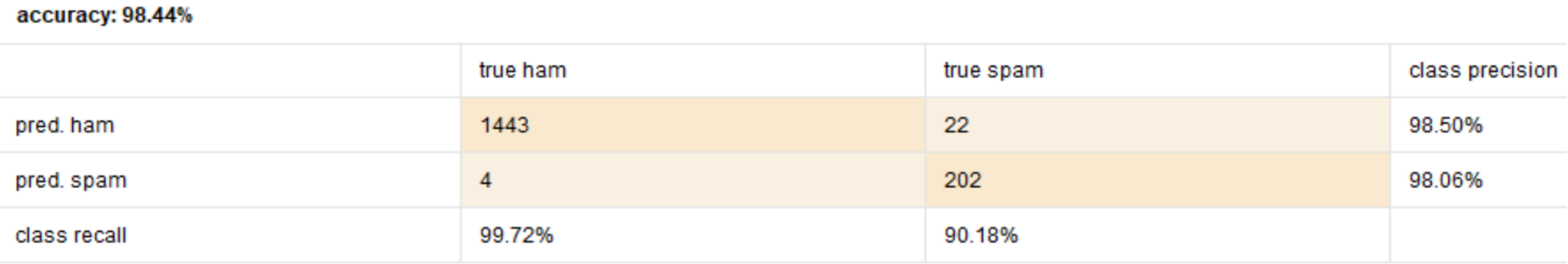
Train



accuracy: 98.67%

	true ham	true spam	class precision
pred. ham	3378	52	98.48%
pred. spam	0	471	100.00%
class recall	100.00%	90.06%	

Test



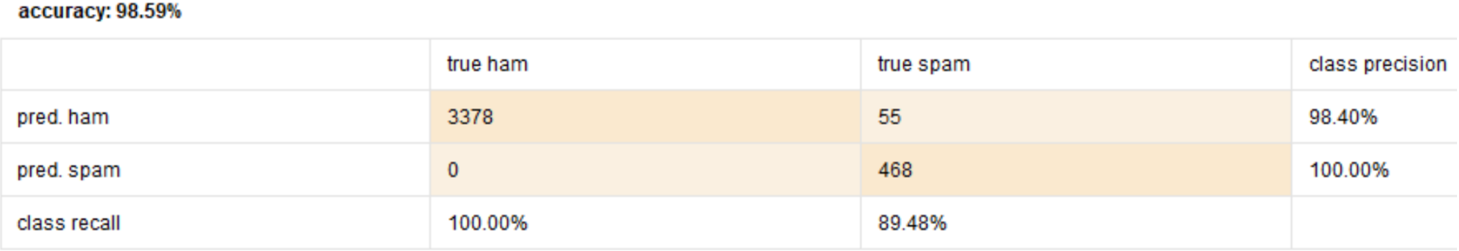
accuracy: 98.44%

	true ham	true spam	class precision
pred. ham	1443	22	98.50%
pred. spam	4	202	98.06%
class recall	99.72%	90.18%	

Model Performance - Random Forest/ Pruned

- Model built using No of trees =100, Criterion = Information Gain, Max Tree Depth = 40, Confidence = 0.1
- Recall values are around 90%. Model Generalizes well because train vs test score recall values have a difference of less than 5pp. Moreover, test recall value is higher than train value.
- Precision values are 100% and 98% indicating that the model rarely misclassifies a genuine text

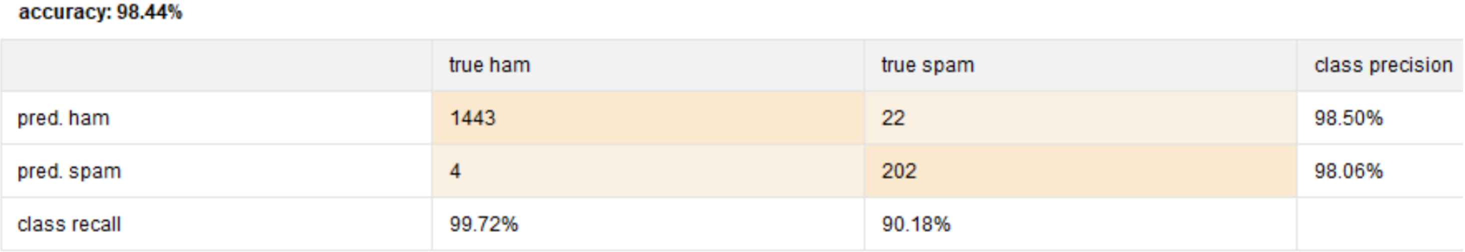
Train



accuracy: 98.59%

	true ham	true spam	class precision
pred. ham	3378	55	98.40%
pred. spam	0	468	100.00%
class recall	100.00%	89.48%	

Test



accuracy: 98.44%

	true ham	true spam	class precision
pred. ham	1443	22	98.50%
pred. spam	4	202	98.06%
class recall	99.72%	90.18%	

Model Performance Summary – Random Forest

- **Choice of Performance Measure: Recall, Precision or Recall with Precision as secondary**
  - ❑ From Problem statement, the objective is to identify reduce cyber attacks on employees' phones by identifying spams. So, **Recall seems** the most important measure as one does not want to receive text messages which are spams, assume they are genuine, and get scammed,
  - ❑ ***However,*** *Precision can not be ignored as* one does not want to miss genuine and important messages, like a meeting invite or an urgent client request, just because they were classified as spam. Unlike emails which can still be viewed in a separate folder, spam texts may never be seen.
- **It is not a straightforward choice between the two but model without pruning is selected**
  - ❑ Both models have recall test and train values around 90%, and Precision values of 98%+. Both Models Generalize well.
  - ❑ Since model without pruning has slightly higher test recall values, it is chosen

Overall Performance Summary – Choosing Best Model

	Recall Train	Recall Test	Precision Train	Precision Test
Decision Tree/No Pruning	95.8%	90.2%	100%	87.1%
Decision Tree/ Pruning	85.7%	77.7%	98.2%	89.2
Random Forest/ No Pruning	90.1%	90.2%	100%	98.1%
Random Forest/ Pruning	90.2%	89.5%	100%	98.1%

- **Choice of Performance Measure: Recall, Precision or Recall with Precision as secondary**
  - ❑ From Problem statement, the objective is to identify reduce cyber attacks on employees' phones by identifying spams. So, **Recall seems** the most important measure as one does not want to receive text messages which are spams, assume they are genuine, and get scammed,
  - ❑ ***However,*** *Precision can not be ignored* as one does not want to miss genuine and important messages, like a meeting invite or an urgent client request, just because they were classified as spam. Unlike emails which can still be viewed in a separate folder, spam texts may never be seen.
- **Random Forest - no pruning** model is chosen model since it has higher test scores on both Recall and Precision and also generalizes well on both. However, 100% scores on training data are not advisable and perhaps one can deploy this model initially and evaluate the results.

Insights and Recommendations

Please refer Executive Summary

APPENDIX