Case Study

SMS Spam Detection

9/18/2024

Togzhan Nurmukanova

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

The advancement of the technology era gave a significant rise to cyber crimes. The most popular form of cybersecurity crime is known as **“smishing**”, which refers to communicating scam messages by text (SMS). According to the Proofpoints 2024, state of Phish report, **nearly 75% of the organizations have experienced smishing attacks in 2023.** The people trust to these messages as they appear to look legitimate and create an impression that originated from legitimate senders such as governmental, financial institutions, big trusted brands, or delivery shipping services. Also, it can be misleading as these messages could be customized information about the person or organization, therefore making individuals believe they are required to take certain actions. The consequences of the smishing attacks are serious for both individual and business levels including **financial losses due to fraudulent transactions, corporate/personal information breaches, compromised security, malware spread, and reputational damage.**

To resolve the problem, the business needs consulting services on the cybersecurity measures that help to protect business from the smishing attacks. As the data scientist of the Cybersecurity solutions consulting would have a key role in working on the project to **develop a solution to predict accurately whether the incoming email is SPAM or not using ML algorithms and pre-processed data.**

Executive Summary

Collection of the **SMS text labeled as SPAM and not spam "HAM", 5,572 in total**. A brief look at the data showed that **13% of the text messages were SPAM**. The **Decision tree** approach was chosen as this project requires classification prediction, as it only needed to determine the type of the incoming text message, and there is no number prediction involved. Also, the dataset is imbalanced, therefore algorithm methods such as Decision Tree and Random forest are well used for this problem.

For the text analysis part, the **TF-IDF technique** has been used for the following tasks text representation (the transformation of the text data into the numerical features), extracting features for the **sentiment analysis**, and term frequency calculation for generating **word clouds**. TF-IDF was chosen as the powerful and efficient method to provide valuable insights finding main characteristics as well as studying the correlation of the sentiment score with the type of the SMS text being SPAM or HAM.

The word cloud for the SPAM messages has shown that the most frequent words were **call, claim, free, txt, reply, and prize**. These words aim to push you to take actions that cause you to become a victim of scam.

Executive Summary

It was proved that the **sentiment score does exist for the SPAM and HAM** text messages. However, text analysis has shown that **HAM and SPAM text messages do not depend on the positive or negative sentiment scores.**

The obtained results show that the **Decision Tree (Unpruned)** demonstrates overall good performance for three parameters, as **Accuracy, Precision, and Recall**. The most important parameter for the project is determined to be Precision for the **False Positive (FP) class**, as it allows us to determine the number of the SMS text messages being TRUE SPAM, but misclassified as HAM messages. These messages create risk for the users as they could become victims of smishing scams. The FP for the Decision tree (Unpruned) is **96.88%**

After applying hyperparameter tuning for the Decision Tree model, it was found that overall performance for **Accuracy and Precision has increased**. However, the difference between Recall values for training and test datasets has slightly increased to 8.18% for the Decision Tree (Pruned) compared with 4.51% for the Decision Tree (Unpruned). The following change is considered insignificant and does not cause overfitting problems.

Executive Summary

The application of the **Random Forest (Unpruned)** model demonstrated good performance for **Accuracy and Precision** but performed poorly on **Recall with 71.21% for the Test dataset**. After the **application of the pruning techniques, it was observed balanced improvement in all three parameters**. Also, this approach allowed to solve the problem of the low **Recall, which increased to 90.16% for the test dataset.**

Comparing the experimental values for the Decision Tree (Pruned) and Random Forest (Unpruned) have similar values for all three parameters. However, a closer look at the **Accuracy and Precision for the Training and Test dataset was slightly higher for the Decision tree (Unpruned).** The Precision for the **FP class was slightly higher for the Decision tree (Pruned) model at 97.68%** vs Random Forest (Pruned) at 97.24%. Choosing between two models, the **Decision tree** is known to be a **simpler model** and, therefore **more time efficient** as it does not have complexity related to the parameter as a number of trees.

Based on the findings of the case study, we can make several recommendations to minimize the risk of the smishing attacks. First, we recommend **using the anti-malware app**, that will protect users from malicious apps, including smishing attacks. Another measure is use of the **multi-factor authentication** would provide an additional key for verification in case if the account was breached by a smishing attacker. Finally, the business could provide **educational training on the measures against phishing attacks**, which would also include defensive measures against smishing.

Business Problem Overview and Solution Approach

Short Message Service (SMS) is acknowledged as the fundamental communication technology accepted by mobile communication standards throughout the globe. Nowadays, with technological advancement, the major concern of the messaging system arises from the spread of **unsolicited messages, known as Spam.** These messages are classified by nature into two types "**non-harmful"** representing the commercial advertisement sent to numerous potential customers and "**dangerous"** containing phishing links that cause people to fall victim to fraud. The spread of phishing links through SMS causes serious harm to individual users as well as business organizations. The most common implications **include financial damage, reputational damage, exposure of sensitive information, and spread of the malware.**

The phishing attacks, also known as "**smishing" is considered the most common type of cybersecurity crime,** therefore businesses are taking serious measures to protect the business from cyberattacks coming through the SMS of the employees of the organization. To resolve the following issue, businesses need the professional consulting services of Cyber Solutions.

Business Problem Overview and Solution Approach

Considering the essence of the problem, **the Data Science Manager** stands as the frontier of the project that helps to develop solutions to accurately identify the Spam messages. As the basis for the project work, there is a database of the labeled SMS texts with two labels Spam and Ham (not Spam). The successful accomplishment of the project considers reaching the following goals:

1. To bring actionable insights from data, that help the organization e**stablish effective security measures against cyberattacks.**
2. Application the tree-based such as **Decision Tree and Random Forest Models** to predict accurately characteristics of the SPAM message.

EDA and Text Visualization

Row No.	Category	Message
1	ham	Go until jurong point, crazy.. Available only in...
2	ham	Ok lar... Joking wif u oni...
3	spam	Free entry in 2 a wkly comp to win FA Cup fi...
4	ham	U dun say so early hor... U c already then sa...
5	ham	Nah I don't think he goes to usf, he lives aro...
6	spam	FreeMsg Hey there darling it's been 3 week'...
7	ham	Even my brother is not like to speak with me...
8	ham	As per your request 'Melle Melle (Oru Minna...
9	spam	WINNER!! As a valued network customer yo...
10	spam	Had your mobile 11 months or more? U R e...
11	ham	I'm gonna be home soon and i don't want to...
12	spam	SIX chances to win CASH! From 100 to 20,0...
13	spam	URGENT! You have won a 1 week FREE m...

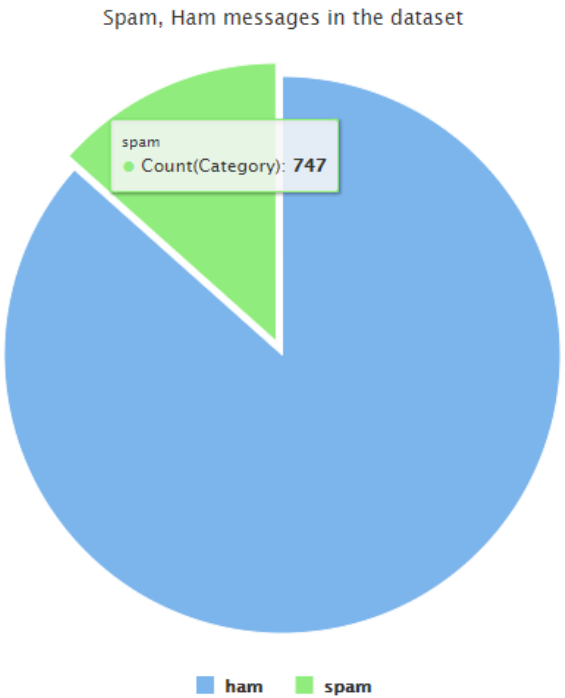
ExampleSet (5,572 examples, 0 special attributes, 2 regular attributes)

- **5,572** Rows: each row represents the individual sms text.

- **2 columns**: Category and Message.

- Category: the label for corresponding SMS. There are two labels as Spam and Ham for the text data.
- Message: contains text data for corresponding SMS.

EDA and Text Visualization



- **13%** of the SMS text has been labeled as **Spam**.
- The most of the SMS text are non-Spam and contain relevant information, that should not be misclassified.
- According to the chart, there is an imbalanced dataset for the Spam to Ham ratio. The use of the algorithmic methods as **Decision Tree and Random forest** are particularly suited for this type of the problem.

EDA and Text Visualization

- The most frequent words for the word cloud of Spam messages are **call, free, txt, prize, won, urgent, guarantee**. These words are more important as this allows us to detect particularly the Phishing links and allows us take preventive measures to maintain security. The Phishing links aimed to allure people to take certain actions, that causes them to fall victim of cyber crime

Text Analysis - how is it important?

Row No.	Score	Message	Category
1	0.173	Go until jurong point crazy Available onl...	ham
2	0.121	Ok lar Joking wif u oni	ham
3	0.243	Free entry in a wkly comp to win FA Cup...	spam
4	0.173	U dun say so early hor U c already then ...	ham
5	0.295	Nah I don t think he goes to usf he lives ...	ham
6	0.156	FreeMsg Hey there darling it s been we...	spam
7	0.763	Even my brother is not like to speak with ...	ham
8	0.156	As per your request Melle Melle Oru Min...	ham
9	0.173	WINNER As a valued network customer ...	spam
10	-0.139	Had your mobile months or more U R ...	spam
11	0.052	I m gonna be home soon and i don t wa...	ham
12	0.191	SIX chances to win CASH From to po...	spam
13	-0.520	URGENT You have won a week FREE ...	spam

To obtain the sentiment score the following procedure is used:

1. Apply **text preprocessing**, that includes tokenization, stop words removal, stemming/lemmization;

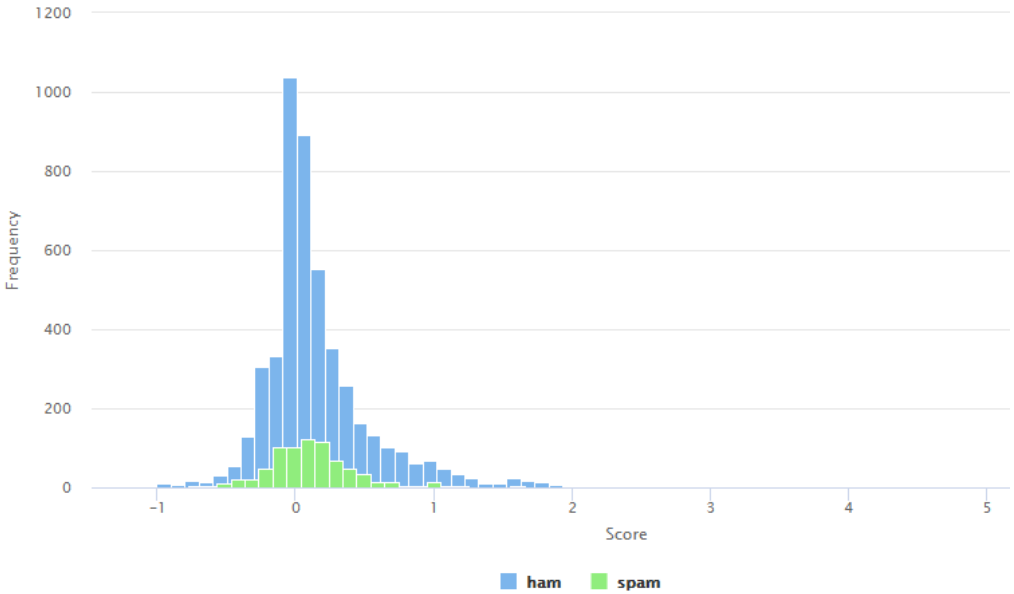
2. **TF-IDF calculation**, measuring the word frequences within document and measuring the importance of the word in corpus, and multiplication of these values.

3. Create **feature matrix**, where the rows are documents (text messages) and columns are terms (words)

4. Perform **Model Training** on the Feature Matrix, using the labeled Sentiment Dataset, in our case SentiWordNet Model.

5. **Predict sentiment labels** for the new document (message)

Text Analysis - how is it important?



- As can be seen from the graph, sentiment score does exist for both SPAM and HAM text messages. **Positive or negative score does not have correlation** with the type of the text being **SPAM or HAM**.

- Th overall sentiment scores are normally distributed between -1 to 1. However, the number of the **positive reviews are higher than negative.**

Text Analysis - how is it important?

Row No.	word	in documents	total ↓
23	call	13	14
34	claim	9	10
84	free	8	10
229	txt	9	10
182	repli	7	9
139	mobil	6	8
169	prize	8	8
161	pleas	7	7
142	msg	6	6
194	servic	5	6
36	code	5	5
53	custom	5	5
178	receiv	5	5

ExampleSet (256 examples, 0 special attributes, 3 regular attributes)

- how the **TF-IDF** technique based on calculating frquencies of the word in the document and measures its importance.
- In this case, it is used for generating **word clouds**, as the more frequently used words appear to have larger font size on the diagram.
- used as the **feature extraction** method to perform sentiment analysis.
- Among the most important words for the SPAM messages were **call,claim, free, text, prize etc.**

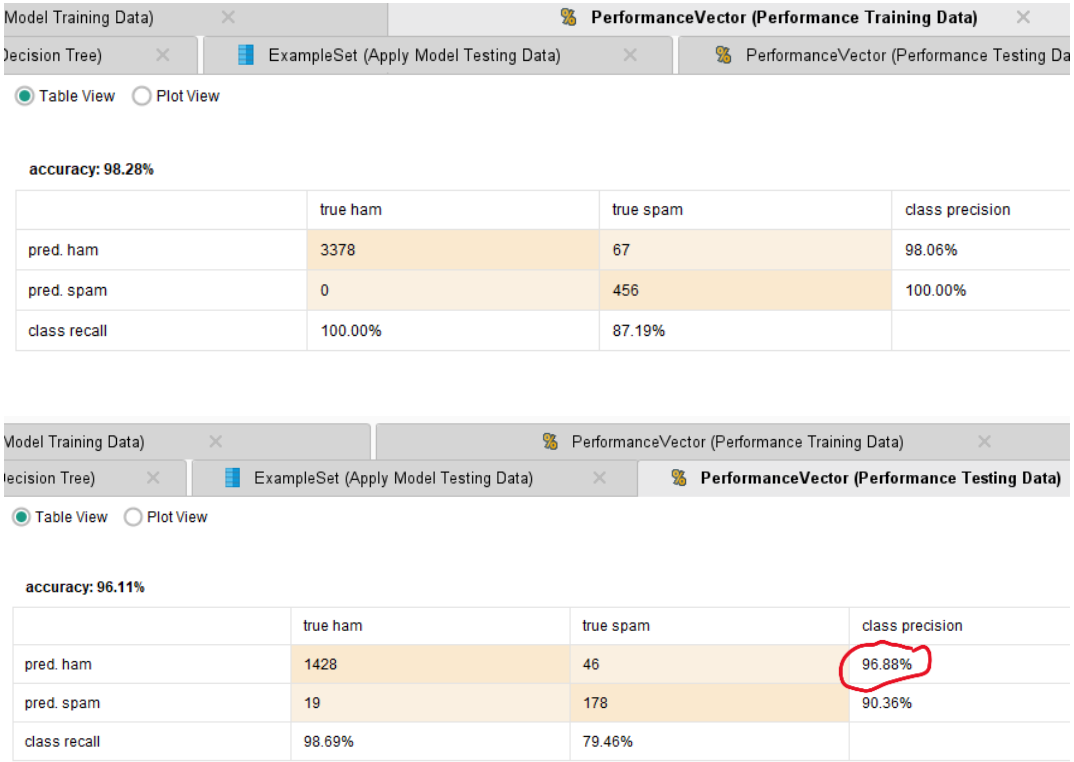
Text Analysis - how is it important?

Show the data in a table

rd	in documents	total ↓
1764	get	339	368
646	call	271	300
860	come	279	300
2575	love	212	256
1037	dai	232	254
2396	know	235	249
1811	good	223	236
4500	time	222	235
1827	got	223	234
4822	want	203	212
2557	lor	145	162
4084	sorri	148	157
4352	take	154	155

- According to the information retrieved for the HAM messages, the more frequently used words are **get, call, come, love, good, know**. The following words expressed personal feelings rather than gave request to take an actions.

Model Performance Summary – Decision Tree (Unpruned)



The following changes to the Decision Tree (Unpruned) model:

- Addition of the operator filter **token by length (3-15 characters)**
- Applied parameters: Splitting Criterion- **Gini Index**, **Max. Depth 50, no pruning**.
- The model performs well both on training and test dataset, and indicates **no overfitting problem exists.**

Model Performance Summary – Decision Tree (Unpruned)

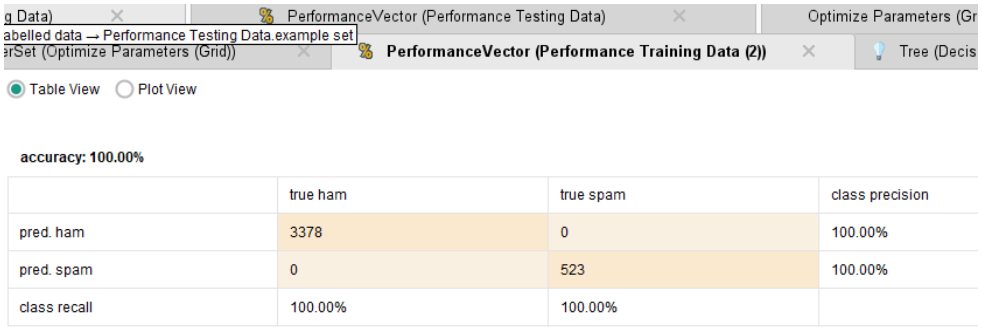
Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
Decision Tree	98.28	96.11	93.59	89.08	99.03	93.62

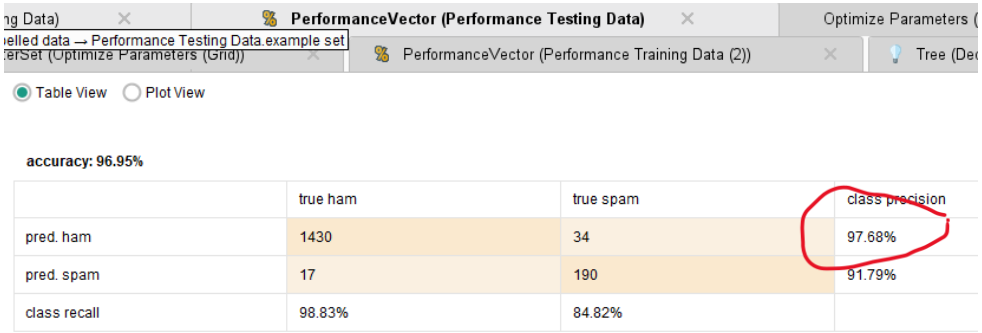
- The Decision Tree (Unpruned) model demonstrated **good overall performance throughout all three performance parameters**. This result was achieved by increasing **max depth tree to 50**, however this caused increase of the processing time.

- the case study aim to minimize number of Spam messages that being missclassified as Ham. The current model provides good results by achieving the **Precision of False positive class of 96.88%**

***Note**: You can use more than one slide if needed*

Model Performance Summary – Decision Tree (Pruned)





Optimizing parameters:

- Criterion: Information Gain, Gini Index, Gini Index
- Maximal Depth: 10 to 40
- Apply Pruning: true, false
- Apply prepruning: true, false
- Min Gain 0.01; Min leaf size 2; Min size for split 4, No of prepruning alternatives 3

The optimized parameters for the model with best performance:

- Criterion: **Information Gain,**
- Iteration:**40**
- Maximal Depth: **40**
- Apply pruning: **False**
- Apply Prepruning: **False**

The model performed well on both training and testing dataset, **no overfitting problem.**

Model Performance Summary – Decision Tree (Pruned)

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
Decision Tree-PRUNED	100	96.95	100	91.82	100	94.73

- After Hyperparameter tuning the **overall performance improved for Accuracy and Precision values** for both Training and Test Dataset. Opposite to this improvement, the difference between Recall values for Training and Testing Dataset has increased to 8.18%

- The target value for the **Precision of the False Positive class has increased to 97.68%**, that proved improved performance with fine-tuning of parameters.

Model Performance Summary – Decision Tree (Pruned)

ParameterSet

```
Parameter set:

Performance:
PerformanceVector [
-----accuracy: 100.00%
ConfusionMatrix:
True:   ham     spam
ham:    3378    0
spam:   0       523
-----weighted_mean_recall: 100.00%, weights: 1, 1
ConfusionMatrix:
True:   ham     spam
ham:    3378    0
spam:   0       523
-----weighted_mean_precision: 100.00%, weights: 1, 1
ConfusionMatrix:
True:   ham     spam
ham:    3378    0
spam:   0       523
]
Decision Tree.maximal_depth     = 40
Decision Tree.apply_pruning     = false
Decision Tree.criterion = information_gain
Decision Tree.apply_prepruning  = false
```

Optimize Parameters (Grid) (48 rows, 6 columns)

iteration	Decisio...	Decision Tree.apply_pruning	Decisio...	Decision Tree.apply_...	acc... ↓
40	40	false	informati...	false	1
39	30	false	informati...	false	0.998
48	40	false	gini_index	false	0.997
38	20	false	informati...	false	0.996
36	40	true	informati...	false	0.993
47	30	false	gini_index	false	0.993
35	30	true	informati...	false	0.993
44	40	true	gini_index	false	0.992
34	20	true	informati...	false	0.991
43	30	true	gini_index	false	0.990
42	20	true	gini_index	false	0.989
11	30	true	informati...	true	0.988
46	20	false	gini_index	false	0.988

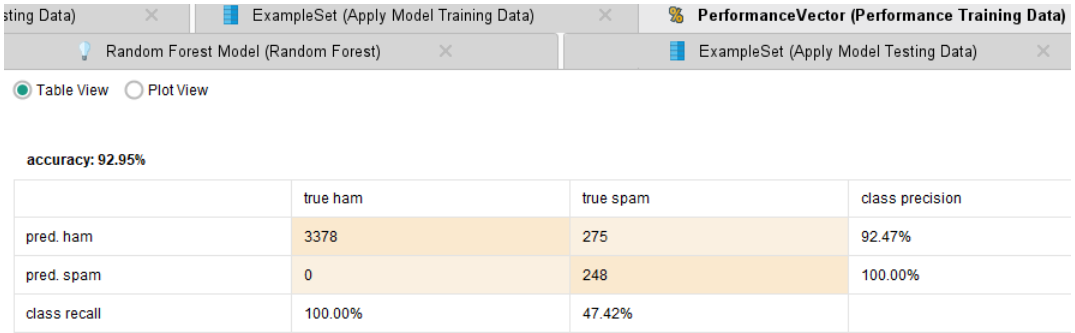
The Optimized Grid Results, show that the Accuracy improvements achieved mainly through the increase of the Max Tree Depth and use of Splitting Criterion as Information Gain and Gini Index. Other criterions as application **of pruning and pre-pruning does not cause significant increase in accuracy values.**

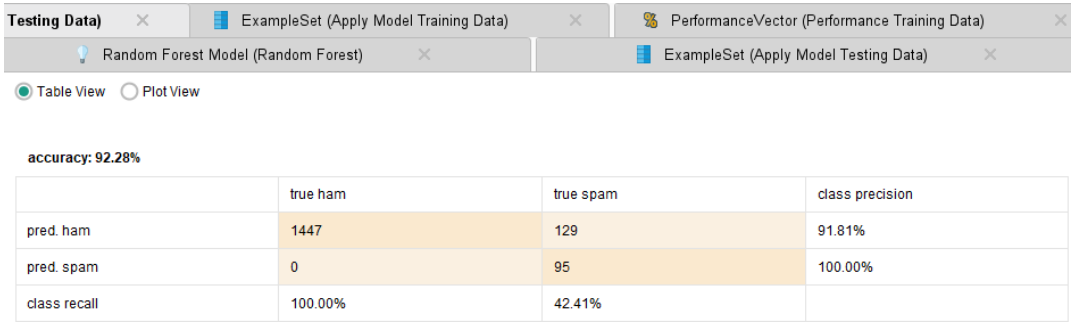
The Performance Comparison Decision Tree (Unpruned) vs Decision Tree (Pruned)

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
Decision Tree	98.28	96.11	93.59	89.08	99.03	93.62
Decision Tree-PRUNED	100	96.95	100	91.82	100	94.73

The Decision Tree (Pruned) **demonstrated higher values for all three parameters for both Training and Testing dataset** compared with Decision Tree (Unpruned) with default parameters.
The difference in Precision of Training and Test Dataset was slightly less for the Decision Tree (Pruned)

Model Performance Summary – Random Forest (Unpruned)





The following changes to the default Random Forest (Unpruned) model:

- Addition of the operator filter **token** by length **(3-15 characters)**
- Applied parameters: Splitting Criterion- **Information Gain**, N of trees **10**, Max. Depth **50**, no pruning, **no prepruning, confidence vote**
- The model shows **moderate performance on both training and test dataset,** and does not posses overfitting problem,

***Note**: You can use more than one slide if needed*

Model Performance Summary – Random Forest (Unpruned)

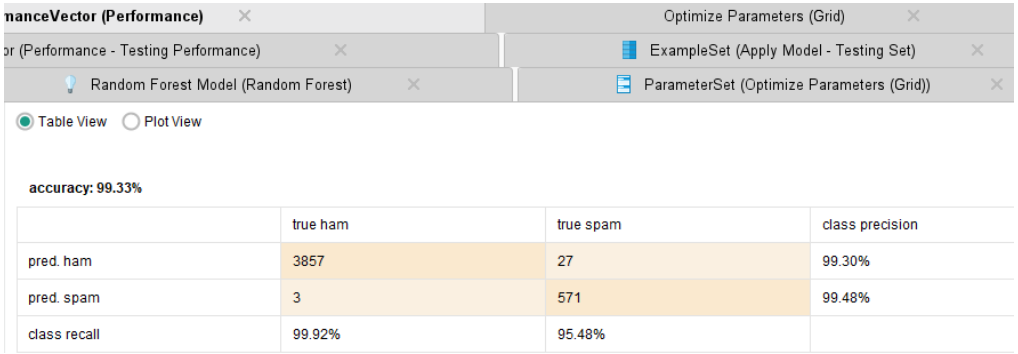
Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
Random Forest	92.95	92.28	73.71	71.21	96.24	95.91

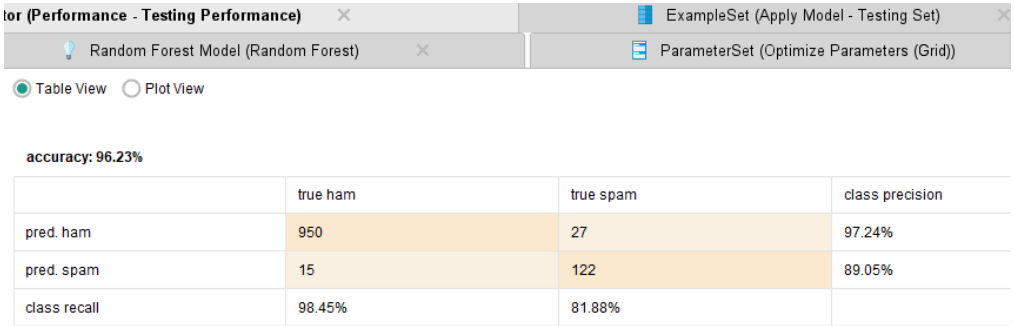
Random Forest (Unpruned) demonstrated **good performance in Accuracy and Precision**, but there is a **significant drop in the Recall value** for both Training and Test Dataset.

The model helps fits to the purpose of the case study by achieving increased value of the **Precision of the False Positive Class to 91.81%**

There is **a problem with classification of the True Negative class**, that possesses low values of 42.41%

Model Performance Summary – Random Forest (Pruned)





Optimizing parameters:

- Criterion: Information Gain, Gini Index, Gini Index
- No of trees: 1 to 20
- Maximal Depth: 1 to 50
- Apply Pruning: true, false
- Apply prepruning: true, false
- Min Gain 0.01; Min leaf size 2; Min size for split 4, No of prepruning alternatives 3

The optimized parameters for the model with best performance:

- Criterion: **Gini Index, No of trees: 18**,
- Iteration:**1055**
- Maximal Depth: **32**
- Apply pruning: **False**
- Apply Prepruning: **False**

The model demonstrated the **outstanding performance** for both training and test dataset

Model Performance Summary – Random Forest (Pruned)

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
Random Forest Pruned	99.33	96.23	97.7	90.16	99.39	93.14

The Random Forest model (Pruned) demonstrated **balanced improvement throughout 3 parameters for** both Training and Test dataset.

The model achieved the target of the case study and demonstrated **Precision of 97.24% for the False Positive Class.**

Resolved problem with low values for the True Negative class compared with the Random Forest (Unpruned) model.

Model Performance Summary – Random Forest (Pruned)

```
Parameter set:

Performance:
PerformanceVector [
-----accuracy: 99.33%
ConfusionMatrix:
True:   ham     spam
ham:    3857    27
spam:   3       571
-----weighted_mean_recall: 97.70%, weights: 1, 1
ConfusionMatrix:
True:   ham     spam
ham:    3857    27
spam:   3       571
-----weighted_mean_precision: 99.39%, weights: 1, 1
ConfusionMatrix:
True:   ham     spam
ham:    3857    27
spam:   3       571
]
Random Forest.number_of_trees    = 18
Random Forest.criterion = gini_index
Random Forest.apply_pruning      = false
Random Forest.apply_prepruning   = false
Random Forest.maximal depth      = 32
```

Optimize Parameters (Grid) (1584 rows, 7 columns)

iteration	Rando...	Rando...	Rando...	Rando...	Rando...	acc... ↓
781	20	informati...	false	false	23	0.990
878	16	informati...	true	false	28	0.990
912	18	informati...	false	false	28	0.990
924	20	gini_index	false	false	28	0.990
1180	5	gini_index	false	false	37	0.990
1297	18	gain_ratio	false	false	41	0.990
1184	12	gini_index	false	false	37	0.990
1446	9	gini_index	false	false	46	0.990
774	7	informati...	false	false	23	0.989
1048	5	gini_index	false	false	32	0.989
923	18	gini_index	false	false	28	0.989
1301	5	informati...	false	false	41	0.989
1415	12	gini_index	true	false	46	0.989

Similarly to the Decision Tree (Pruned) model, high Accuracy values are achieved using the **Information Gain and Gini Index** as Splitting criterion, whereas the Pruning and pre-pruning criterion did not have significant affect.

The Performance Comparison Random Forest (Unpruned) vs Random Forest (Pruned)

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
Random Forest	92.95	92.28	73.71	71.21	96.24	95.91
Random Forest Pruned	99.33	96.23	97.7	90.16	99.39	93.14

Random Forest (Pruned) demonstrated enhanced values for all three parameters for **Training ang Test dataset compared with Random Forest (Unpruned)**

The difference between Training and Testing parameters is lower for the Random Forest(Unpruned), although the difference between Random Forest (Pruned) **does not cause an overfitting problem.**

Recall for both Training and Test dataset is **higher** for the Random Forest (Pruned).

Overall Summary:
Choosing and commenting on best model

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
Decision Tree	98.28	96.11	93.59	89.08	99.03	93.62
Decision Tree- PRUNED	100	96.95	100	91.82	100	94.73
Random Forest	92.95	92.28	73.71	71.21	96.24	95.91
Random Forest Pruned	99.33	96.23	97.7	90.16	99.39	93.14

Decision Tree (Pruned) and Random Forest (Pruned) models have almost similar improvements for 3 parameters. However, **Accuracy and Precision of Training and Test dataset were slightly better for Decision Tree model (Pruned).**

Also, Decision Tree (Pruned) model is simpler than Random Forest (Pruned), therefore **time-efficient to run the algorithm.** It is easier to interpret results, as there is no complexity associated parameter as number of trees.

Overall Summary:
Choosing and commenting on best model

For the business objective, we are most interested in **increasing of precision parameter**. With increase of this parameter, would contribute **to minimizing the number of SPAM text messages being misclassified as HAM.** It is important, as missing out the false positives, can cause serious implications for the people, as they can cause them to be scammed by criminals as losing money, exposing sensitive information and experiencing personal reputation damage.

However, we also want not to sacrifice much on the recall values, as the HAM text messages that are wrongly classified as SPAM, would be completely ignored by users. These text messages containing important information would not be read, therefore this would cause miscommunication problems.

Accuracy is an important parameter that assesses the overall performance of the problem. In this case, we need to make sure that the **difference between the training and test datasets does not exceed 10%.** It is needed to confirm that the model does not experience an overfitting problem, i.e. performs well on both training and test datasets.

Overall Summary:
Choosing and commenting on best model

The following measures can be recommended to minimize risk of the phishing attacks through SMS text messages for the employees of the company:

1. Use an **Anti-Malware Application,** that could help to identify the SPAM text messages and prevent people from receiving numerous SPAM text messages.

2. **Multi-factor Authentication**, known as an additional layer of verification, can be used if the account password has been corrupted.

3. Organize **the trainings educating employees** on their actions against the phishing attacks. Develop the instruction for the employees on their actions how to verify contact information, store and protect sensitive information, as well as the further measures if they came across the phishing attempt.

APPENDIX

Slide Header

- Please mention any other pointers (if needed)

Great Learning

Happy Learning !