**AN ANALYSIS OF SPAM SMS FEATURES**

IFN 701 Project 1- Data Analysis and Research Project

Project Student Name: Harshita Jain

Student ID: n9539361

Academic Supervisor: Dr. Guido Zuccon

Project Coordinator: Dr. Charles Wang

# **ABSTRACT**

The increasing dependence on short text messages for communicating with other people has led to an increase in submission of Spam Short Message Service(SMS) which intends to extract confidential, personal and valuable information from the recipients. This growth is difficult to curb because of the availability of affordable and unlimited prepaid SMS packages and high response rates as customers are more comfortable with sharing their confidential and personal information via SMS. The previous solutions built in this context are simple, straightforward and do not consider all the core characteristics of Spam SMS. One of the major problems that has limited the research on this topic is scarce data available for research on Spam SMS.

This document details a data analysis and research project involving the investigation of SMS Spam. This project aims to create a data analysis to analyze the features that differentiate a Spam SMS from a Legitimate SMS and a predictive model that is enabled to understand and automatically identify whether an SMS is a Spam message or a Legitimate message. This project will be performed using a publicly available dataset acquired from Kaggle and by creating a supporting codebase in R Markdown.

The outcome of analysis is the most frequent words used in Spam SMS that differentiate it from a Legitimate SMS. Moreover, the result of building predictive models using 4 classifiers: *Naïve Bayes, Support Vector Machine, Logistic Regression and Decision Tree* indicated that Support Vector Machine performed better than the rest of the three classifiers and can be used in mobile devices to prevent arrival of Spam SMS to the subscriber.

The outcome of this project would benefit the mobile networks operators as, now, they would not have to spend more on maintaining the networks and operations which get hampered by Spam SMS. Customers would be benefitted as their confidential, personal and valuable information would remain protected.

# **1. INTRODUCTION**

This document aims at elaborating a data analysis and a research project. The project used a publicly available dataset from Kaggle to analyze the features that differentiate a Spam SMS from a Legitimate SMS and to build a predictive model capable of accurate prediction of a Spam SMS. The project was performed by creating a codebase in R Markdown.

## **1.1 Context of the Project**

SMS is an integral feature of mobile devices that facilitates communication through exchange of short text messages. (Ahonen, Tomi T., 2011) Informa Telecoms and Media has found that there has been an increase in this exchange from 5 trillion messages being exchanged in 2010 to 10.7 trillion SMS being exchanged in 2015. (Global SMS traffic to reach 8.7 trillion by 2015: study)Also, the revenue generated by SMS has increased from US$132.5 billion in 2010 to US$136.9 billion in 2015.(Global SMS traffic to reach 8.7 trillion by 2015: study)

However, this exchange has been hindered due to submission of unsolicited and unwanted SMS in bulk to recipients without their authorization and consent. Such SMS intends to extract confidential, valuable and personal information out of the recipients. According to Cloudmark Analysis, 92% of such SMS are fraud.(Whitepapers)

The most forms of Spam SMS are (Whitepapers)

* Have won a Gift Card Message,
* Account Phishing Spam Message,
* SMS Service Message,
* Accident Compensation Spam Message, and
* Payment Protection Insurance (PPI)Compensation Spam Message.

Cloudmark Analysis revealed that there has been a growth in reception of Spam SMS of 300% from 2011 to 2012. (Whitepapers) This growth can be accredited to the following two reasons:

1. The availability of affordable and unlimited prepaid SMS packages due to which extraction of confidential, valuable and personal information of the recipients becomes a cost effective opportunity for the spammers. (Khemapatapan, C, 2011)
2. Customers feel more comfortable sharing their confidential, valuable and personal information vis SMS. (Khemapatapan, C, 2011) Therefore, it was revealed in June 2013 statistics that 43% of such messages are everted within 15 minutes of reception. (SMS Marketing Statistics, 2017)

Spam SMS adversely affects consumers and mobile network operators. The reasons that make this problem a significant one are:

1. Mobile network operators bear a heavy loss in order to maintain networks, operations and increased customer care services to the customers. Spamming ruins their reputation making them lose on many valuable customers(Khemapatapan, C, 2010).
2. Customers are also left annoyed and worried as their confidential, personal and valuable information is at stake(Khemapatapan, C, 2010).
3. Many network operators have provided means to their customers to block Spam SMS, which sometimes leads to filtration of legitimate message as a spam due to its characteristics matching to those of a spam message(Khemapatapan, C, 2010).

## **1.2 Related Work and Research Gap**

There have been a number of anti-spam techniques built in this context to solve this problem like (Khemapatapan, C, 2010)–

* Blacklisting - This technique forbids access to a service if the name is written on the list.
* Simple Filtering - This technique analyses the traffic data and identifies the individual subscriber causing huge volumes of it.
* Spoofing/Faking Detection Techniques

These techniques are brittle, simple and straightforward in nature. They do not consider the core characteristics of Spam SMS. Also, they perform in an ad-hoc an d post-hoc manner. That is, a consumer can only blacklist a number as and when it sends him a Spam SMS. Blacklisting one number does not guarantee him complete prevention from receiving Spam SMS. Moreover, spamming methods have advanced in a way that they make a Spam SMS appear as a Legitimate one. Therefore, there is an urgent need to build a more sophisticated and appropriate model to eradicate this issue. (Khemapatapan, C, 2010)

Moreover, not much research could be done in this context due to scarce data available on Spam SMS.(Khemapatapan, C, 2010)

### **1.2.1 How this Project Addresses the Problem**

I have worked on bridging the gap by building a predictive model to accurately predict whether an SMS is a Spam SMS or a Legitimate one by working on the publicly available dataset available at Kaggle. It comprises of 5,574 English, real and non-encoded text messages, submitted to Grumble text website. All claims made on this site about the text message being spam are identified and investigated through carefully scrutinizing over a hundreds of webpages. (SMS Spam Collection) All messages have accurately been tagged as legitimate and spam. In total, the dataset consists of 747 Spam messages.

## **1.3 Aims and Objectives of the Project**

The objective of this project would be to answer the following two questions:

1. What are the characteristics that distinguish Spam messages from Legitimate messages?
2. What is the effectiveness of the classification methods – Support Vector Machine, Decision Trees, Logistic Regression or Bayesian Classifiers in identifying SMS Spam?

Therefore, the purpose of this project is:

1. To analyze the data to understand the differentiating features of SMS Spam.
2. To build a predictive model which can accurately predict whether the SMS is a Spam SMS or a Legitimate SMS.

Particularly, the main aim of this project is

* Carrying out an exploratory analysis on the dataset to explore and learn about the data.
* Statistically predicting and modelling the data
* Result Interpretation

## **1.4 Brief Overview of Methods used in the Project**

In this project, I will develop a data analysis including the investigation of a number of predictive models. This analysis is structured in 4 phases (Guo, P., 2013):

1. **Preparation Phase -**

This phase includes acquiring data from Kaggle, formulating key questions for analysis and cleaning the data.

1. **Exploration Phase -**

This phase includes carrying out exploratory analysis on the dataset to analyze the features that differentiate a Spam SMS from a Legitimate SMS. This analysis starts with analyzing "How Length of Messages and Number of Messages relate to each other for each label", followed by finding words that appear most frequently in Spam SMS.

1. **Data Preparation Phase -**

This phase includes preparing data to be used to build predictive models in Classification phase (Next Phase). In this phase, a corpus is created and cleaned by transforming all the text to lower case, removing punctuations, numbers, stop words and white space. The, the data was split into 70% training set and 30% test set.

1. **Classification Phase -**

This phase includes building 4 different classifiers in 2 settings. The settings are:

* 1. ***Setting 1:*** Considering all features of the data
  2. ***Setting 2:*** Considering manually engineered features

The 4 classifiers used are:

* 1. Naïve Bayes
  2. Decision Tree
  3. Support Vector Machine
  4. Logistic Regression

After having modelled all classifiers for each setting, a comparison would be made to determine the best model and the corresponding setting.

1. **Final Delivery Phase -**

This phase includes submitting outputs from Exploration Phase and Classification Phase in the form of R Markdown and an Analysis Report.

Therefore, the two **target deliverables** of the project would be -

1. R Markdown
2. Analysis Report

## **1.5 Outcome of the Project**

The outcome of this project will facilitate:

* An effective and a deeper knowledge of characteristics and features that make a Spam SMS different from a legitimate SMS.
* A predictive model that can accurately predict whether an SMS is spam or legitimate.

These outcomes would indirectly affect the society in a better way. The learnings and research could be converted into operational products in future that would aid accurate identification and filtration of spam SMS.

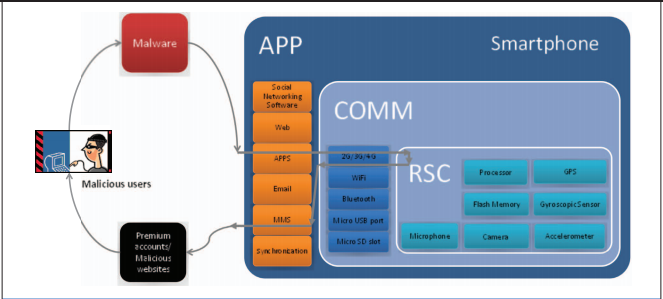
# **2.Literature Review of Previous Work**

In this section, I ,would discuss the possible threats a mobile phone user is vulnerable to, due to the increase in submission of Spam SMS. I would then discuss the various solutions built in this context, their drawbacks, and how this project aims at solving them.

## **2.1 Possible Threats from a Spam SMS**

In a mobile phone, a user receives a malware in the form of an SMS, from a malicious attacker. The Spam SMS appears to be sent from a legitimate sender, making the receiver believe that it is a Legitimate SMS.

Figure 1 shows that User A was sent a Spam SMS infected with a malware. As she opened the SMS, the malware ran through her mobile phone and extracted all the information including his contact book. The malware sends all the sensitive information sored in the mobile phone to the spammer.



*Figure 1: Spam SMS Attack*

Such a malicious Spam SMS was first reported in 2004. The incidences of such Spam SMS increased by 3,325% by 2011. The affects of such Spam SMS vary from being minor issues, such as, degradation of performance of mobile devices and slow operations of mobile devices, to being major issues, such as, financial loss and being unable to receive or make a call.

Therefore, a number of solutions were proposed in this context to prevent Spam SMS. They have been discussed in the following section.

## **2.2 Previous Work Done in this Context**

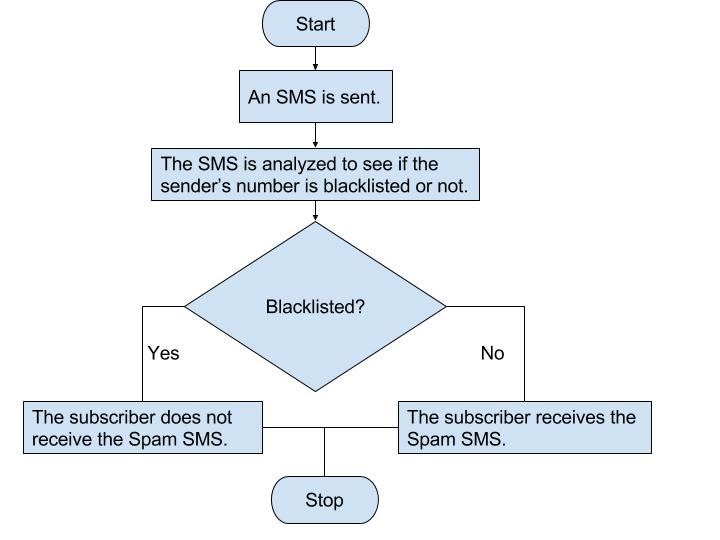
The solutions built in this context are:

1. **Blacklisting:**

Blacklisting is a process of forbidding a person or a group from using certain resources and services, by adding them to a list called blacklist. (Blacklist Doc)

***How does it work?***

Researchers suggested that SMS spamming could be avoided and, eventually stopped, by blacklisting the spammers' number. They suggested that mobile network operators should authenticate each subscriber and create a blacklist for SMS Spammers. Therefore, this would facilitate prevention of submission of Spam SMS to other recipients and reception of SMS only from legitimate and trusted users. (Khemapatapan, C, 2010)



*Flowchart 1: The Process of Blacklisting*

Figure 2 depicts the process of SMS Blacklisting. Firstly, the spammer sends a Spam SMS to a recipient. This SMS would be analyzed by the intended recipient's network to verify if it a Spam SMS or a Legitimate SMS. It is verified by analyzing the sender's number. If the spammer's number has been blacklisted by the recipient in past, the recipient will not be able to receive the message and hence, Spam SMS Reception has been prevented. On the other hand, if the recipient has not blacklisted the spammer's number, he will receive the Spam SMS. (blacklist Doc)

***Drawbacks of Blacklisting***

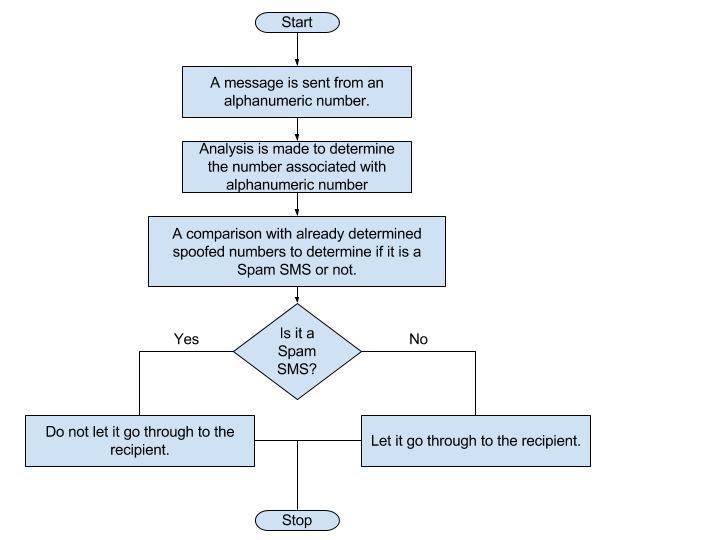
It was experienced that blacklisting was not an efficient way of solving the problem of SMS Spamming as it did not guarantee complete prevention from receiving a Spam SMS.

Blacklisting a number facilitates prevention from receiving Spam SMS from that particular number only. A spammer can resort to identity changes by buying a new SIM Card, or by sending Spam SMS through emails, which is the cheapest option due to availability of free email services like Yahoo, Hotmail, Gmail, etc.(Blacklist Doc) Therefore, spammer can make new email accounts periodically and send Spam SMS to the subscribers.

1. **Spoofing/Faking Detection Techniques**

SMS Spoofing is a process which makes the sender and contents of a Spam SMS appear like a Legitimate one.

***How does it Work?***



*Flowchart 2: The Process of SMS Spoof Detection*

Figure 3 depicts the process of SMS Spoofing Detection Technique. First, the spammer converts his phone number to a combination of alphanumeric text, which he uses to send the SMS to conceal his real identity. This SMS is analyzed by the network of intended recipient to determine the actual phone number associated with the alphanumeric text. After the actual number has been extracted, a comparison is made with already determined spoofed numbers to determine that whether this SMS is a Spam SMS or a Legitimate SMS. If the received SMS is a Legitimate SMS, it is sent to the intended recipient. Whereas, if it is a Spam SMS, it is not sent to the intended recipient.

***Drawbacks of Spoofing Detection Technique***

However, Spoofing Detection Technique could not completely prevent reception of Spam SMS as a spammer can resort to identity changes and continue with his activities of sending Spam SMS to the subscribers. Also, it was observed that sometimes implementation of this technique incorrectly detected a Legitimate SMS as a Spam SMS on account of the sender's alphanumeric number.

1. **Implementation of Email Spam Filters to Filer Spam SMS**

Many studies indicate that the proposal of implementing content-based email spam filters, to prevent SMS Spam from being sent to the subscribers, was made with a perception that a Spam Filter would be able to perform its job with all types of Spams. *Content-based Spam filters vary form filtering a simple keyword to automatic classification of complex text. These approaches use supervised machine learning algorithm to train a model how to accurately predict whether a message is a Spam SMS or a Legitimate SMS.* However, the most common length of a Spam SMS is 160 characters and makes it inappropriate to be used by content-based filter systems. (Khemapatapan. C, 2010) Also, the type of language used in the two services is quite different; from abbreviated language, emoticons, bad punctuations, etc. being used in SMS to a formal language being used in emails. (Khemapatapan. C, 2010)

Therefore, it was concluded that email spam filters would not be an appropriate way to prevent Spam SMS.

## **2.3 Supporting Decisions**

Therefore, my project aims at combatting the above discussed drawbacks by working on the following decisions:

1. An **exploratory analysis** should be carried out to analyze the features that differentiate a Spam SMS from a Legitimate SMS. This would help us combat the drawback discussed in section 2.2*: "Also, it was observed that sometimes implementation of this technique incorrectly detected a Legitimate SMS as a Spam SMS on account of the sender's alphanumeric number."*
2. A **predictive model** should be built that is capable of accurate prediction of whether an SMS is a Spam SMS or a Legitimate SMS. This would help us combat all the drawbacks discussed for Blacklisting and Spoofing Detection Technique.

# **3.Project Methodology**

## **3.1 Project Management Approach**

The project management methodology I chose for this project was Dynamic Systems Development Method (DSDM). This approach encourages iterative and incremental delivery, essentially keeping parameters: time, cost and quality fixed (The DSDM Agile Project Framework, 2014). It focusses on solution optimization and control risk by permitting change of requirements throughout the development period and active involvement of stakeholders through continuous communication, review and feedback. I have chosen to work in this agile framework because of two main reasons:

* *Time Constraint* – The project needs had to be delivered in 8 weeks time. Therefore, working in timeboxes ensured On-Time Delivery.
* *Quality Control* – Meeting the laid quality standards is as important as delivering the product on time. Therefore, frequent review and feedback, of work done, by the Supervisor helped produce deliverables of the expected quality.

### **3.1.1 MoSCoW Prioritization**

|  |  |
| --- | --- |
| **Prioritization** | **Deliverables** |
| Must Haves (60%) | Data Analysis of features that differentiate Spam SMS from Legitimate SMS.    Investigation to understand the most effective classification method to predict whether an SMS is a Spam SMS or a Legitimate SMS. |
| Should Haves (20%) | Recommending development of a model of the most effective classification method. |
| Could Haves (20%) | To evaluate if development of this model is feasible or not. |
| Won't Haves | Building an operational product accurately identifying a spam SMS. |

*Table 1: MoSCoW Prioritization for Scope*

**3.1.2 Detailed Weekly Plan**

In this project, activities like – setting up project team of Project Student and Supervisor and describing the overall project and its high level requirements was completed in the first four weeks (Week 1 – Week 4), i.e., the foundation and the feasibility phases.

The plan for the remaining 8 weeks (Week 5 – Week 13) was to divide them into increments, which further were divided into timeboxes. The breakup of these weeks into increments and timeboxes is as follows:

|  |  |
| --- | --- |
| **Item** | **Details** |
| Number of Increments | 2 |
| Duration of each Increment | 4 weeks |
| Timeboxes in each Increment | 3 |
| Duration of each Timebox | 1 week |
| Deployment Period in each Increment | 1 week |

*Table 2: Weeks breakup into Increments and Timeboxes*

Each timebox and increment had a review and feedback session at the end. This was conducted in weekly meetings and also focused on the creation of new tasks list for next timebox.

Detailed plan for each increment is as follows:

**Increment 1 -** Analyze Differentiating Features of Spam SMS

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Phases** | **Feasibility** | **Foundation** | **Timebox 1** | **Timebox 2** | **Timebox 3** | **Deploy** |
| **Goal** | Initial project setup | Project planning |  |  |  |  |
| **Items** | Finalize project team    Define project objective and scope.    Ethics Clearance | Investigation of the project's background    Document project proposal | Find research papers related to the topic.    Explore the dataset and read it.    Share the research papers with the Supervisor and determine the relevant ones. | Read at least half of the articles.    Clean the data to remove missing and noisy data.    Share the work done with the Supervisor and seek feedback. | Read the remaining papers.    Write the code in R Markdown to Analyze Features that Differentiate Spam SMS from Legitimate.    Share the work done with the Supervisor and seek feedback. | Finalize the code in R Markdown and Analysis for Differentiating Features of Spam SMS. |
| Deliverables | Project Unit Study Agreement    Ethics exemption | Project plan/proposal document | 25 – 30 research papers    Dataset knowledge | Clean data | Knowledge on characteristics of spam SMS. | Code and Analysis Report. |
| Duration | 2 weeks (24/07 - 06/08) | 2 weeks (07/08 - 20/08) | 1 week (21/08 - 27/08) | 1 week (28/08 03/09) | 1 week (04/09 - 10/09) | 1 week (11/09 - 17/09) |

*Table 3: Detailed Weekly Plan for Increment 1*

**Increment 2 -** Build Predictive Models

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Phases** | **Timebox 1** | **Timebox 2** | **Timebox 3** | **Deploy** | **Concluding the Project** |
| **Items** | Write the code in R Markdown using Naïve Bayes Classification Method and Logistic Regression.    Share the work done with the Supervisor and seek feedback. | Write the code in R Markdown using Support Vector Machine.    Share the work done with the Supervisor and seek feedback.. | Write the code in R Markdown using Decision Tree.    Share the work done with the Supervisor and seek feedback. | Finalize the code in R Markdown and Analysis Report for all Classification Methods.    Investigation of effectiveness of each of the used Classification methods in determining if the SMS is spam or legitimate. | Project Presentation.    Project Report.    Submit report. |
| Deliverables | Knowledge on effectiveness of Naïve Bayes and Logistic regression Classifiers. | Knowledge on effectiveness of Decision Tree. | Knowledge on effectiveness of Support Vector Machine. | Code and Analysis Report for all Classification Methods. | Project presentation    Final report |
| Duration | 1 week (18/09 - 24/09) | 1 week (25/09 01/10) | 1 week (02/10 - 08/10) | 1 week (09/10 - 15/10) | 2 weeks (16/10 - 29/10) |

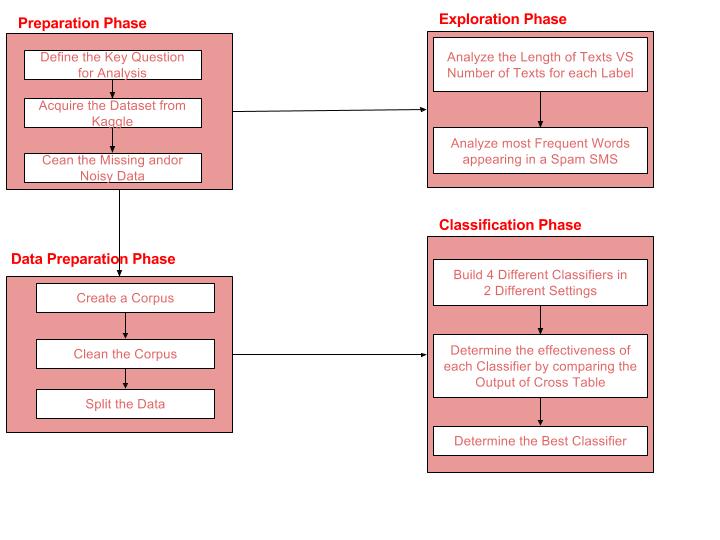
*Table 4: Detailed Weekly Plan for Increment 2*

## **3.2 Project Methodology**

This section would give a clear idea of all the methods that were used to make this project a success, keeping in mind the time constraint and the target deliverables.

The project methodology that best suited this project is ***Data Analysis***. This methodology would be a conjuncture of 4 phases(Guo, P., 2013):

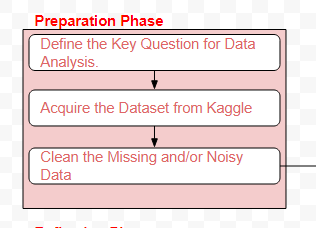
1. Preparation Phase
2. Exploration Phase
3. Data Preparation Phase
4. Classification Phase



*Flowchart 3:Project Methodology Workflow(Guo, P., 2013)*

All of these 4 phases are a further conjuncture of various activities. These will be explained in detail in the following sections.

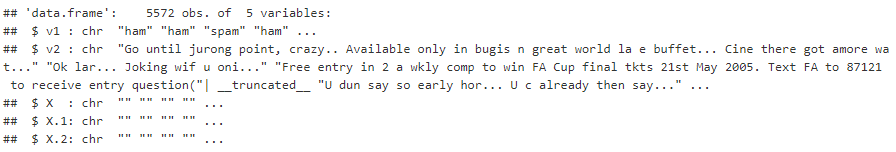
### **3.2.1 Preparation Phase**



*Figure 2: Preparation Phase*

Preparation phase allows us to lay the foundation for the analysis. This phase started with defining the objective of the project – what the project is about and what do we aim to do in it.

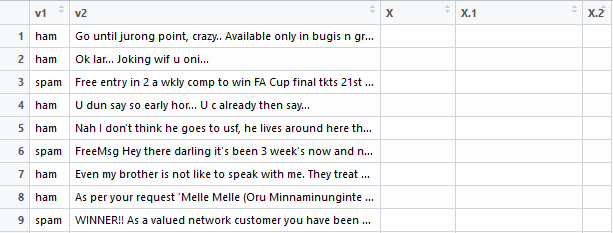
Once the objective has been clearly laid and understood, I acquired the data that would be helpful in achieving the objective. As mentioned earlier, I have acquired the dataset from Kaggle.



*Figure 5:Output of Data Acquisition*

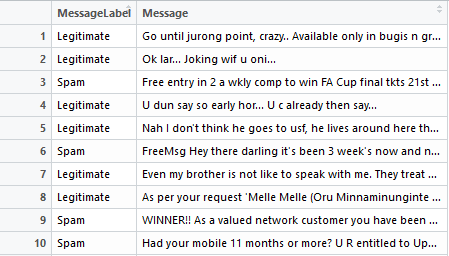
Post acquiring the data, I scanned through the data to see if there are any missing, noisy or semantically erroneous data in the dataset as removing these entries from the dataset would help me perform the analysis in an appropriate manner. (Guo, P., 2013)

I found that, there were 3 null columns (*columns X, X.1 and X.2 in Figure 6*) present in the data and the remaining two columns *(columns v1 and v2 in Figure 6)*, for Message and Label, were not named in a comprehensible manner.



*Figure 3:Raw Data*

Therefore, I removed the 3 null columns and renamed the first column as ***MessageLabel*** and the second column as ***Message***.

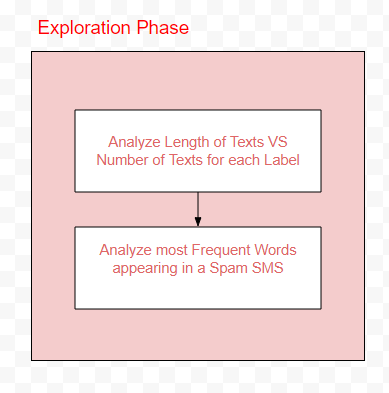


*Figure 4:Clean Data*

|  |  |
| --- | --- |
| **Input** | **Output** |
| Data acquired from Kaggle | Key question for data analysis.    Clean data |

*Table 5: Input-Output Table for Preparation Phase*

**3.2.2 Exploration Phase**



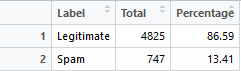
*Figure 5:Exploration Phase*

In this phase, I carried out an exploratory analysis to analyze the features that differentiate a Spam SMS from a Legitimate SMS.

***Analyzing Distribution***

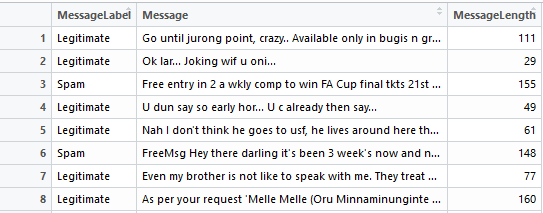
I started the analysis with calculating the percentage of Spam SMS in the dataset.

1. I counted the total number of Legitimate SMS and Spam SMS in the dataset and put that in a data frame.
2. This was followed by calculating percentage values for each Label and renaming the column names to make them more comprehensible.
3. After having known the percentage values for each, I plotted a pie chart to visualize the distribution.



*Figure 6:Total Percentage of Legitimate SMS and Spam SMS in the dataset*

***Analyzing Length of Texts VS Numbers of Texts***  
This analysis was made to determine how the length of texts and number of texts are related to each other for each label. I calculated the number of characters for each text and plotted them against the number of texts for each label.



*Figure 7:Column Message Length added to the Dataset*

***Analysis of the Most Frequent Words Occurring in a Spam SMS***

1. In this analysis, I explored the data and manually engineered words that appeared most frequently in Spam SMS. I observed that words like *Call, Mobile, Prize, Text, etc*. appear most frequently in Spam SMS.
2. To verify my manual findings, I produced a word cloud for Spam SMS. ***A word cloud is an image consisting of words belonging to a text, document or a dataset, in which the size of the words depict their importance and frequency in that text.*** *(Thesaurus)*
3. Now, after having a visualized interpretation of all the words that appear most often in a Spam SMS from a word cloud, I made 6 categories. That is,
   1. *winner, win, won, award, selected, prize and claim* were put into one category '*Winner*',
   2. *congratulations and congrats* were put into one category 'Congratulation',
   3. *xxx, babe, naked, dirty, flirty* were put into one category '*Adult*',
   4. *urgent, attention, bonus, immediately, now, stop* were put into one category '*Attention*',
   5. *free* was put into one category '*Free*', and
   6. *ringtone, call, mobile, text, txt* were put into one category '*Ringtone*'.
4. After having made these 6 categories, I ran for loop for each to assign a 'y' or an 'n' to messages depending on occurrence of words from these categories in the messages.
5. Now, I plotted these 6 categories on a bar plot to determine the most important category of all.
6. Following this, I verified the output of bar plot by using Importance function of Random Forest which extracts importance of all tokens as assigned by randomForest() and plots it. (https://www.rdocumentation.org/packages/randomForest/versions/4.6-12/topics/importance).

## 

## *Figure 8:A Sample of how each Category has been Created using a For Loop*

|  |  |
| --- | --- |
| Input | Output |
| Clean data from Preparation Phase | Features that make a Spam SMS different from a Legitimate SMS |

*Table 6: Input-Output Table for Exploration Phase*

## **3.2.3 Data Preparation Phase**

## 

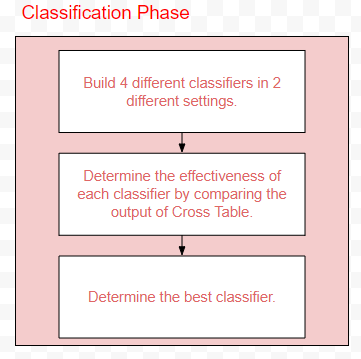
*Figure 9:Data Preparation Phase*

This is the second data preparation phase wherein I prepared the data to be used for building predictive models. Therefore, I created a corpus and cleaned it by transforming all the text to lower case, removing punctuation, numbers, white space and stop words. Then, I split the data into 70% training set and 30% test set.

|  |  |
| --- | --- |
| Input | Output |
| Clean Data from Preparation Phase | Data ready to be used to build predictive models. |

*Table 7: Input-Output Table for Data Preparation Phase*

### **3.2.4 Classification Phase**



*Figure 10:Classification Phase*

In this phase I had built 4 different classifiers in 2 different settings. The 2 settings are:

1. Setting 1: Considering all the features of data
2. Setting 2: Considering manually engineered features, discussed in section 3.2.2.

The 4 different classifiers used are:

1. Naïve Bayes
2. Logistic Regression
3. Decision Tree
4. Support Vector Machine

***Steps involved in building each classifier***

1. First, for each classifier, I defined the model to train the data. Model definition depended on the setting the model was being built in.



*Figure 11:Model Definition for Support Vector Machine being built in Setting 1*



*Figure 12:Model Definition for Support Vector Machine being built in Setting 2*

1. Now, the trained model will be used to test unseen data, that is, the test data. The outcome of this test would decide the effectiveness of the classifier.



*Figure 13:Test the Model Support Vector Machine being built in Setting 1*



*Figure 14:Test the Model Support Vector Machine being built in Setting 2*

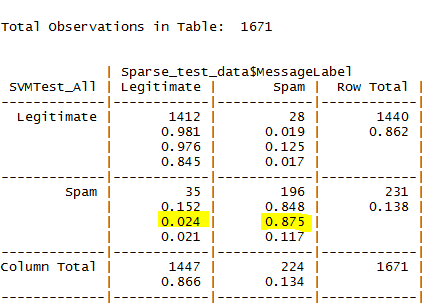
1. I compared the output of Cross Table in order to evaluate the effectiveness of each classifier in each setting. I specifically focused on evaluating the effectiveness on terms of ***recall*** for Spam class and ***incorrect prediction*** for Legitimate class. (the highlighted values)



*Figure 15:Cross Table evaluation of Support Vector Machine being built in Setting 1*

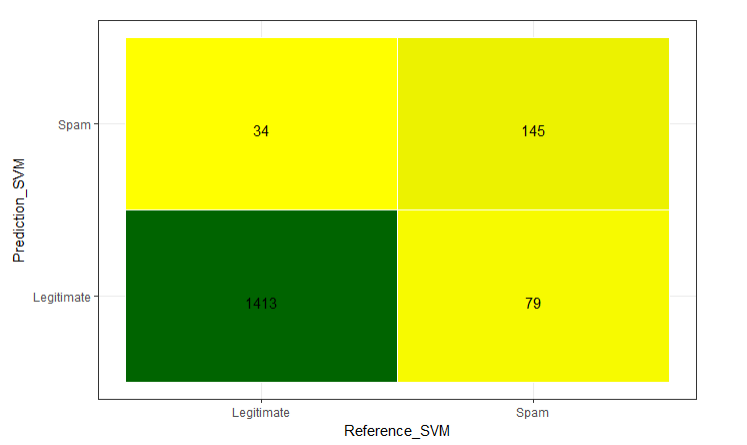


*Figure 16:Cross Table evaluation of Support Vector Machine being built in Setting 1*



*Figure 17:Cross Table Output Format*

1. Then, I plotted the True Positive False Positive Matrix to visualize the prediction.



*Figure 18:Prediction Visualization*

|  |  |
| --- | --- |
| **Input** | **Output** |
| Data from Data Preparation Phase | Precision, Recall and Accuracy Measures for each classifier in each setting.  Most effective classifier and the corresponding setting. |

*Table 8: Input-Output Table for Classification Phase*

# **4.Outcomes of the Project**

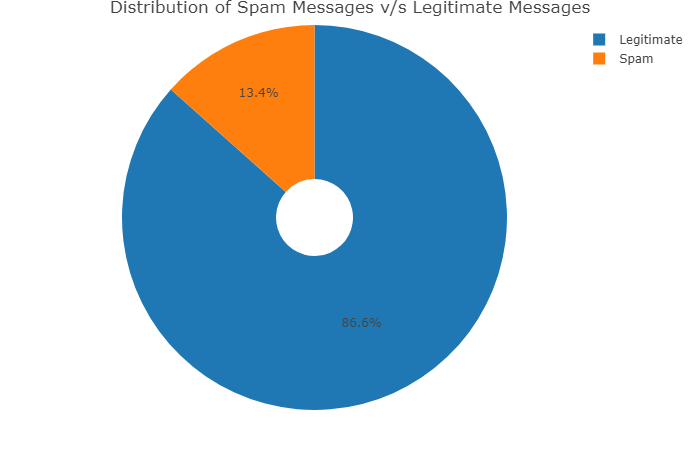
This section details the outcomes of the project achieved in response to the objectives of the project, as mentioned in section 1.3. The outcomes of this project have been attained in Exploration Phase and Classification Phase, as discussed in sections 3.2.2 and 3.2.4 respectively.

## **4.1 Outcomes of Exploration Phase**

In exploration phase, I carried out an exploratory analysis to analyze the features that differentiate a Spam SMS from a Legitimate SMS. This analysis is a conjuncture of 3 analyses, out of which two main analyses (Pts. 2 and 3) help in achieving the objective of the project:

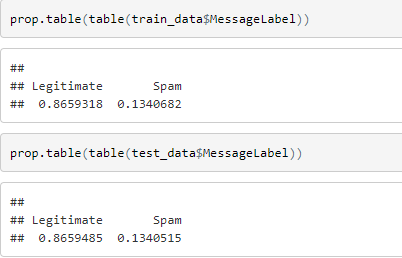
1. Analyzing the distribution of Spam SMS and Legitimate SMS in the Dataset.
2. Analyzing how Length of Texts and Number of Texts are related to each for each Label.
3. Analyzing the most frequent words occurring in a Spam SMS.

### **4.1.1 Distribution of Spam and Legitimate SMS in the Dataset**

The pie chart visualization of distribution of Spam SMS and Legitimate SMS in the dataset (*refer* ***Analyzing Distribution*** *of section 3.2.2 for its methodology*) revealed that 13.42% of all the SMS were Spam SMS *(refer Figure 19).* 

*Figure 19:Distribution of Spam SMS and Legitimate SMS in the Dataset*

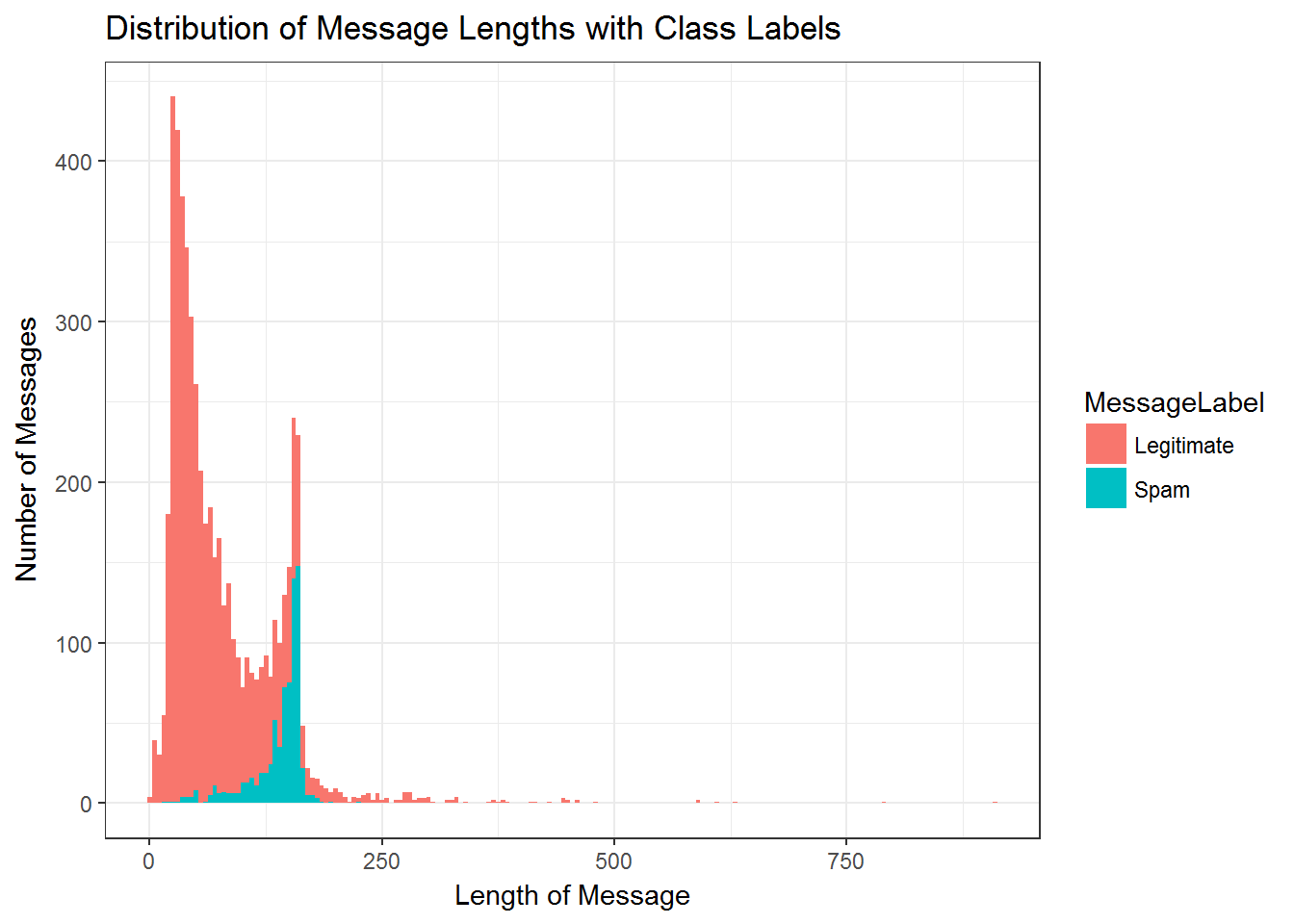
The distribution was analyzed to confirm if the same proportion of Spam and Legitimate SMS have been distributed in the training and the test sets, created in Data Preparation Phase.



*Figure 20: Validation of Distribution of Spam and Legitimate SMS in Training and Test sets*

### **4.1.2 Analysis of Length of Texts VS Number of Texts for each Label**

Figure 20 reveals that the most common length of a Spam SMS is 160 characters, while the most common length of a Legitimate SMS is 20 characters. But, the length of a Legitimate SMS can vary from being as small as just on "OK" in a text, that is, only 2 character SMS, to being as long as containing 900-1000 characters in an SMS. Moreover, Legitimate and Spam SMS overlapped each other for the entire range of length of Spam SMS.

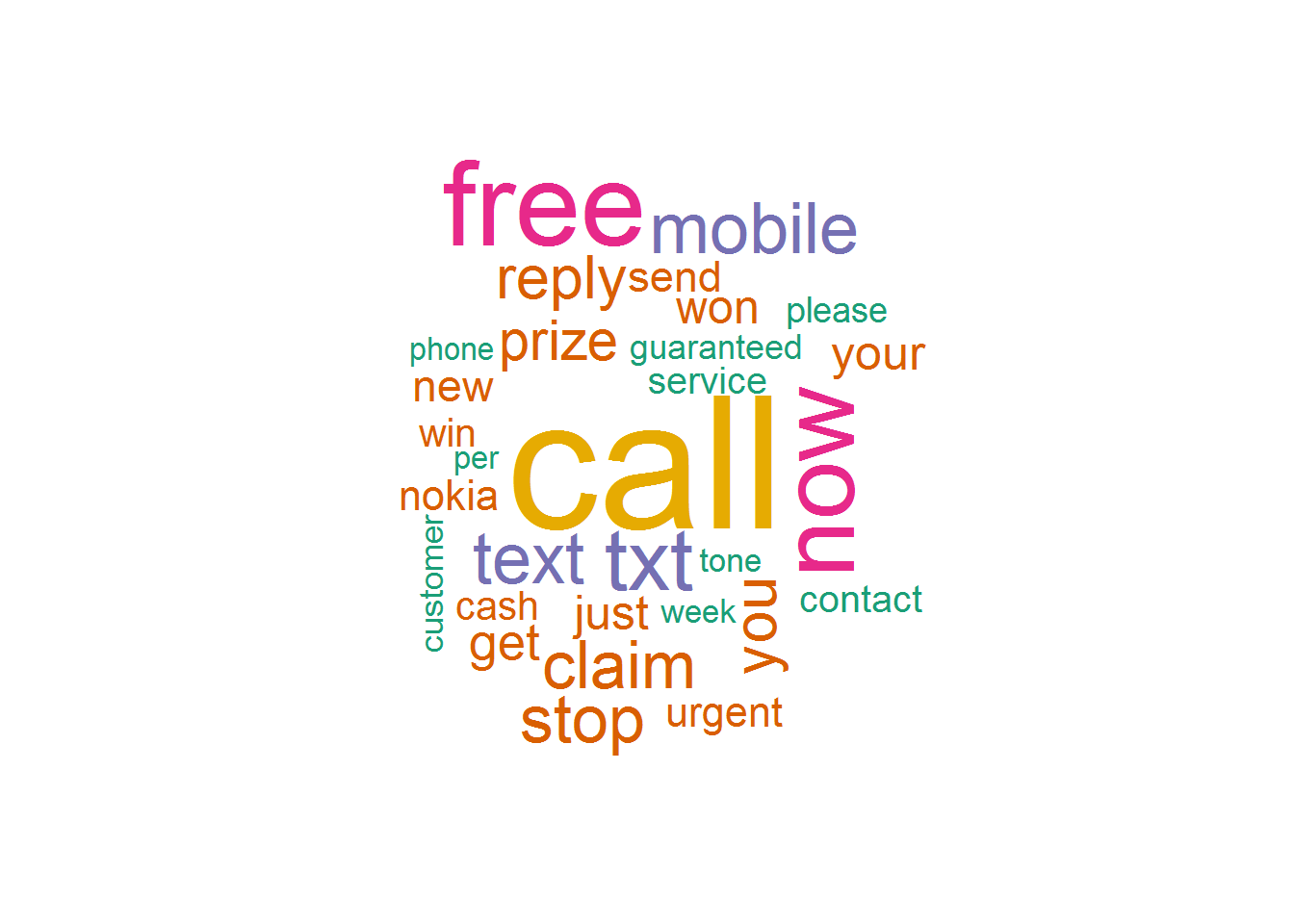


*Figure 20: Length of Texts VS Number of Texts for each Label*

Therefore, the analysis of Length of Texts VS Number of Texts for each label did not reveal much about the nature and differentiating features of Spam SMS. Thus, I resorted to analysis of words that appear more frequently in a Spam SMS, which is discussed in the following section.

### **4.1.3 Analysis of the Most Frequently Occurring Words in a Spam SMS**

I explored the data and manually engineered words to determine the most important words used in a Spam SMS. As a result of the manual exploration of dataset, I concluded that the most important words in a Spam SMS are: Free, Call, Text, Prize, Winner, etc. To verify manual exploration, I produced a word cloud, for Spam SMS, to find the most frequent words occurring in a Spam SMS. It revealed that the most important words in a Spam SMS are: Call, Free, Now, Mobile, Text, Txt, etc. *(refer Figure 21)*



*Figure 21: Word cloud for Spam SMS*

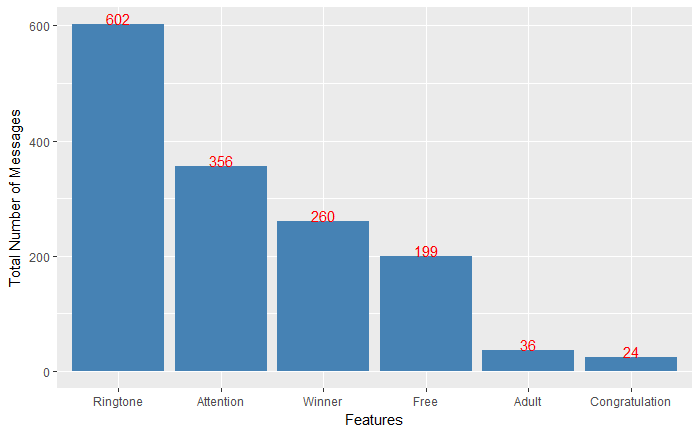
After having confirmed the most frequent words in Spam SMS, I categorized them as discussed in ***Analysis of the Most Frequent Words Occurring in a Spam SMS*** of section 3.2.2. This added 6 more columns in the dataset. *(refer Figure 22)*



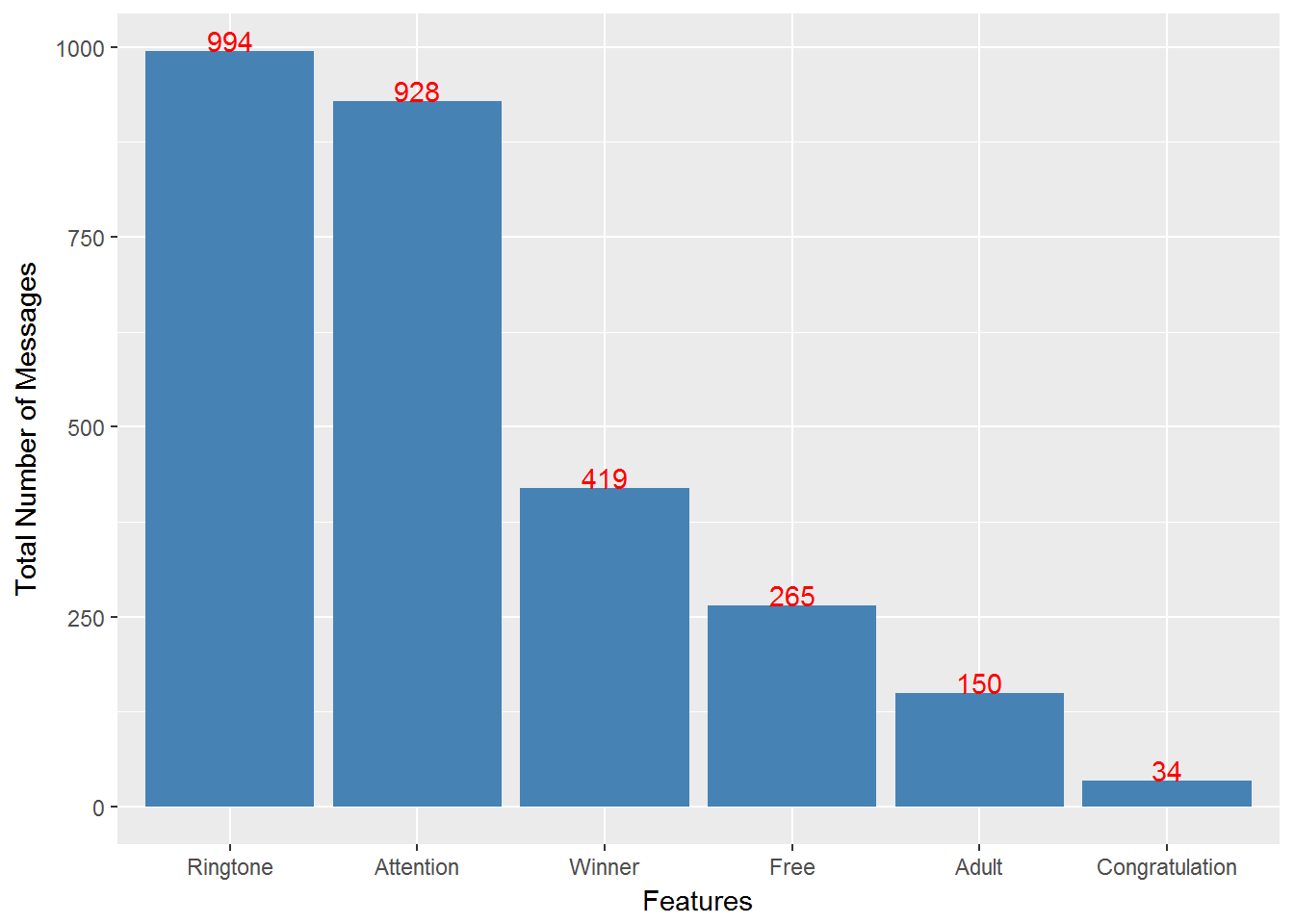
*Figure 22:Dataset containing Values for the 6 Categories*

To determine the importance of each category, I plotted them on a bar plot. The bar plot, in Figure 23, revealed that the most important category of all is Ringtone, as 602 out of 747 Spam SMS contain words belonging to this category. Also, when I plotted the total number of texts in the dataset containing words belonging to category Ringtone, Figure 24 revealed that, in total, 994 SMS contained those words. With this, I inferred, that only 392 Legitimate SMS contained those words, that is, the number of Spam SMS containing those words exceed the number of Legitimate SMS containing those words by 54%%.

Also, the least important category for Spam SMS is Congratulation, as 34 out of 747 Spam SMS contain words belonging to this category *(refer Figure 23).* And, none of the Legitimate SMS containing words belonging to this category *(refer Figure 24).*

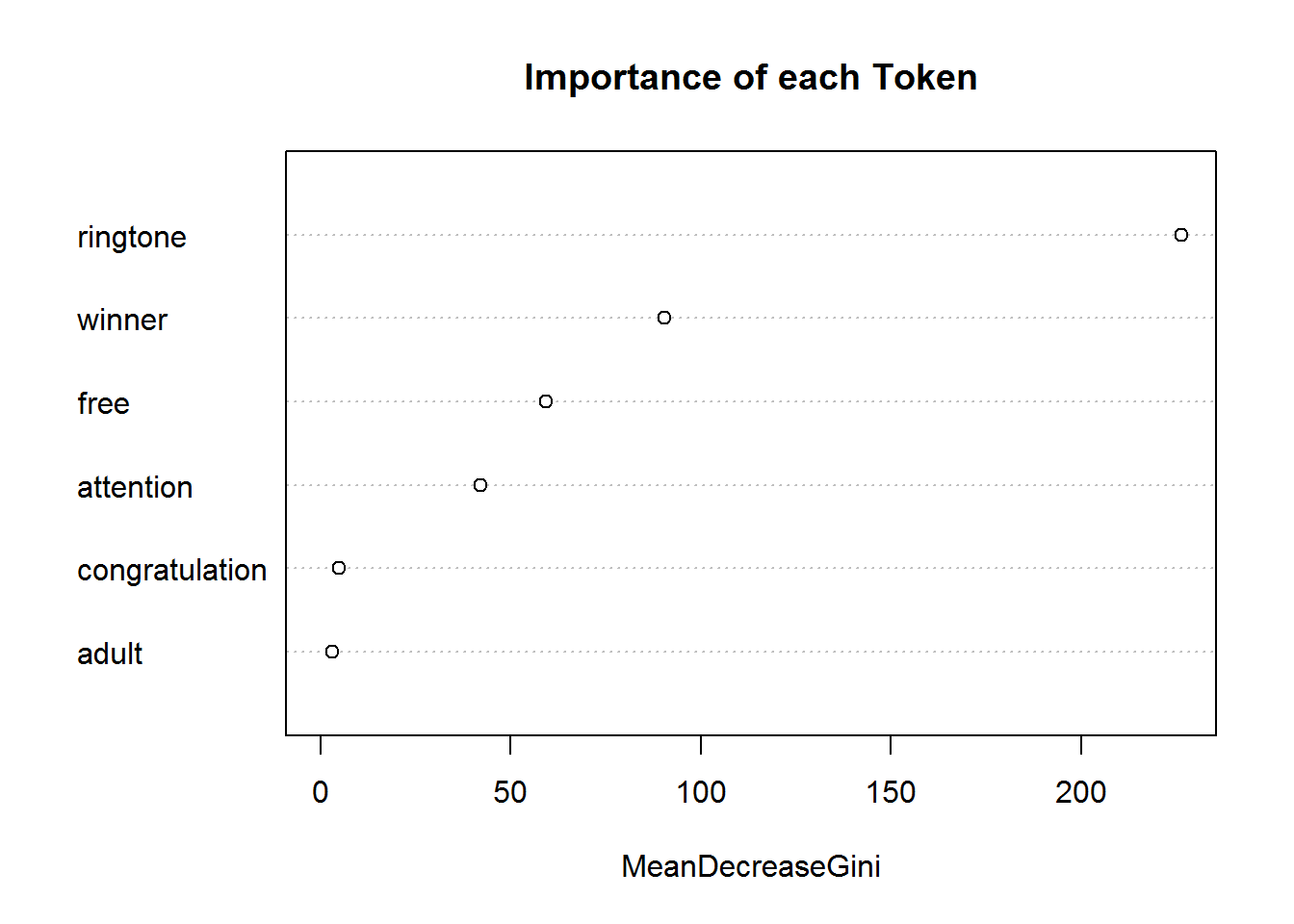


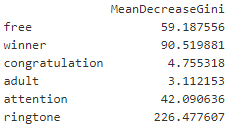
*Figure 23: Bar Plot Depicting Importance of Category for all Spam SMS in the Dataset*



*Figure 24: Bar Plot Depicting Importance of Category for all SMS in the Dataset*

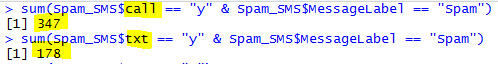
The interpretation of bar plot was verified using Importance function of Random Forest. It substantiated the output of the bar plot and revealed that the most important category for Spam SMS is Ringtone. *(refer Figure 25)*





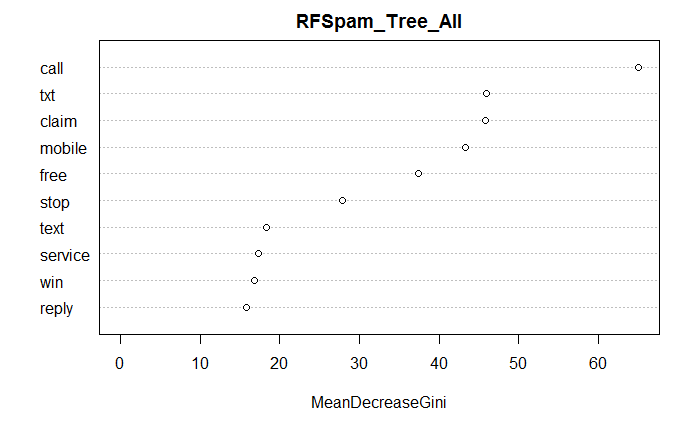
*Figure 25: Importance of Categories in Spam SMS in Plot-form and Tabular-form*

Category ***Ringtone*** contained the word ***call and txt,*** which were depicted as the most frequent words in the word cloud. 46% of Spam SMS contained the word call in them and 24% of them contained the word txt in them (refer Figure 26). Therefore, I concluded that the combined outcomes of word cloud, bar plot and importance plot revealed that ***the most important category for Spam SMS is Ringtone and the most important word for Spam SMS is call***.



*Figure 26: Reveals the Number of Spam SMS containing the Word "Call" and "Txt"*

Also, the importance function of Random Forest built for all the features of the data (Setting 1) substantiated the findings and confirmed that the most important words for Spam SMS are **Call and Txt.**



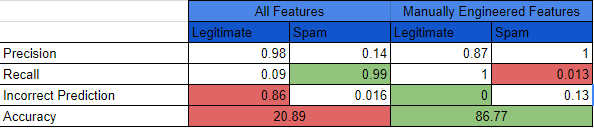
*Figure 27: Importance of Words in Spam SMS*

## **4.2 Outcomes of Classification Phase**

In classification phase, I had built 4 different classifiers using 2 different settings, as discussed in section 3.2.4. I evaluated the effectiveness of each classifier for each setting, specifically using recall for spam class and incorrect prediction for legitimate class.

### **4.2.1 Effectiveness of each Classifier**

1. **Naïve Bayes**



*Figure 28: Comparative study of the Output of Cross Table for Naïve Bayes in both the Settings*

Figure 27 depicts that this model accurately predicted 222 out of 224 Spam SMS, while it incorrectly predicted 1320 out of 1447 Legitimate SMS as Spam SMS, for Setting 1.

On the other hand, this model predicted only 3 out of 224 Spam SMS, while it did not predict any Legitimate SMS as Spam SMS, for Setting 2.

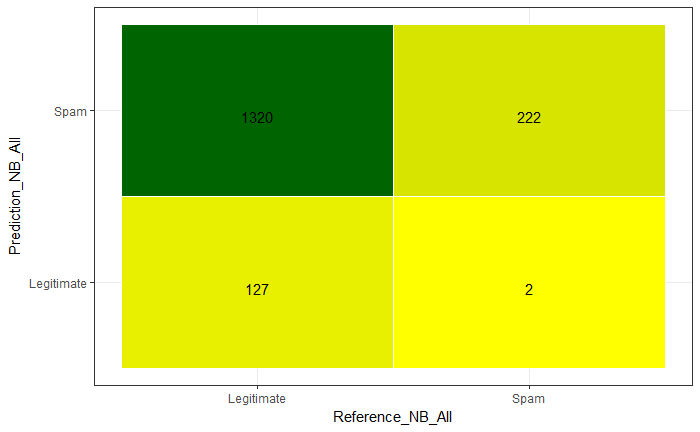
The accuracy for this model in Setting 1 is only 20.89%, while for setting 2 is 86.77%. Even though, model built in Setting 2 has higher accuracy but it cannot be claimed as the best classifier as it predicted only 3 out of 224 Spam SMS. Therefore, I conclude that Naïve Bayes classifier could not be used in mobile devices to accurately predict Spam SMS.

The visualization of true positive and false positive matrix is depicted in Figure 28.

Below figure reveals that 127 out of 1447 Legitimate SMS are predicted accurately. This cell is known as the True Positives and has the darkest shade cell depicting that high percentage of accurate prediction (8.78%) has been made for True Positives.

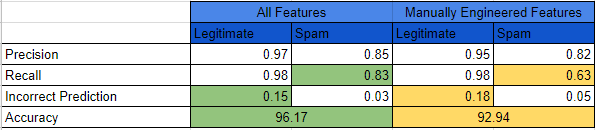
222 out of 224 Spam SMS are predicted accurately. This cell is known as True Negatives and has the second darkest shade cell depicting the high percentage of accurate prediction (99.10%) has been made for True Negatives.

False Positive is the cell depicting that 38 out of 2 Spam SMS have been incorrectly predicted as Legitimate SMS. False Negative is the cell depicting that 1320 out of 1447 Legitimate SMS have been incorrectly depicted as Spam SMS.



*Figure 28: True Positive and False Positive Matrix Visualization for Setting 1*

1. **Logistic Regression**



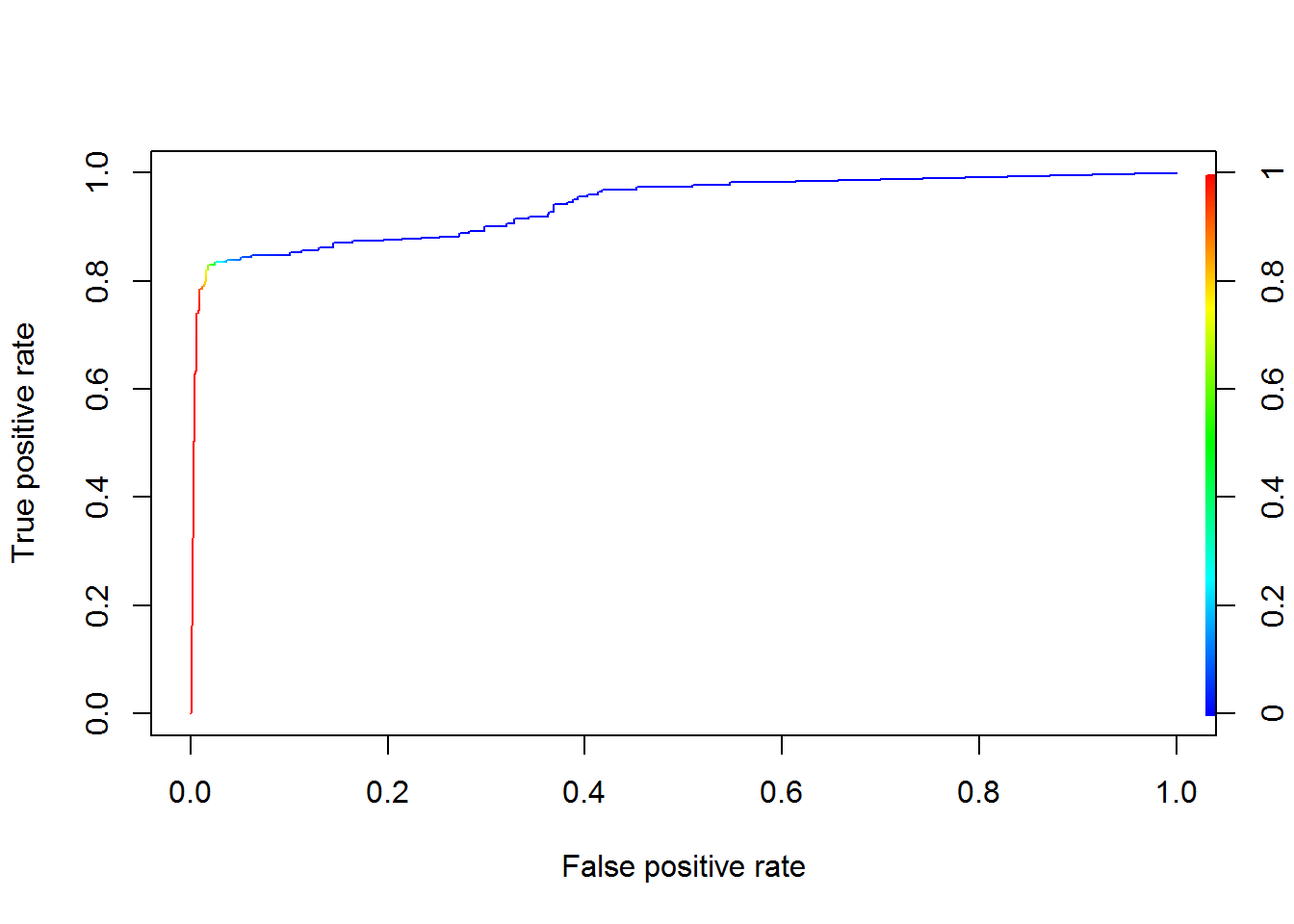
*Figure 29: Comparative study of the Output of Cross Table for Logistic Regression in both the Settings*

Figure 29 depicts that this model accurately predicted 186 out of 224 Spam SMS, while it incorrectly predicted 32 out of 1447 Legitimate SMS as Spam SMS, for Setting 1.

On the other hand, this model predicted 142 out of 224 Spam SMS, while it incorrectly predicted 32 out of 1447 Legitimate SMS as Spam SMS, for Setting 2.

The accuracy for this model in Setting 1 is only 96.17%, while for setting 2 is 92.94%. Therefore, the outcome of the model built in 2 different settings reveals that it works best when modelled in Setting 1.

The Receiver Operating Characteristic (ROCR) curve, which is a plot of true positives VS False Positives for the different cut-points of a test, substantiated the high accuracy of the model achieved in Setting 1. Area under the curve depicts accuracy of the test. Therefore, more the curve runs closer to the left hand and the top border of the ROC space, more accurate the test is. (ROCR Curve Bookmarked Doc for reference) (refer Graph 1)



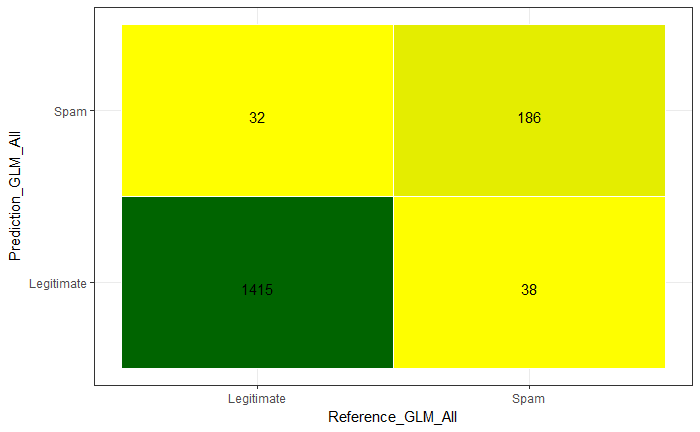
*Graph 1: ROCR Curve in Setting 1*

The visualization of true positive and false positive matrix for Setting 1 is depicted in Figure 30.

Below figure reveals 1415 out of 1447 Legitimate SMS are predicted accurately. This cell is known as the True Positives and has the darkest shade cell depicting that high percentage of accurate prediction (97.78%) has been made for True Positives.

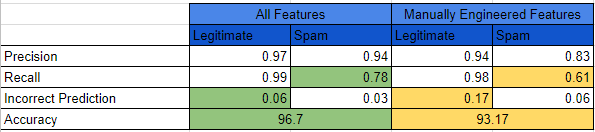
186 out of 224 Spam SMS are predicted accurately. This cell is known as True Negatives and has the second darkest shade cell depicting the high percentage of accurate prediction (83.03%) has been made for True Negatives.

False Positive is the cell depicting that 38 out of 224 Spam SMS have been incorrectly predicted as Legitimate SMS. False Negative is the cell depicting that 32 out of 1447 Legitimate SMS have been incorrectly depicted as Spam SMS.



*Figure 30: True Positive and False Positive Matrix Visualization for Setting 1*

1. **Decision Tree**



*Figure 31: Comparative study of the Output of Cross Table for Decision Tree in both the Settings*

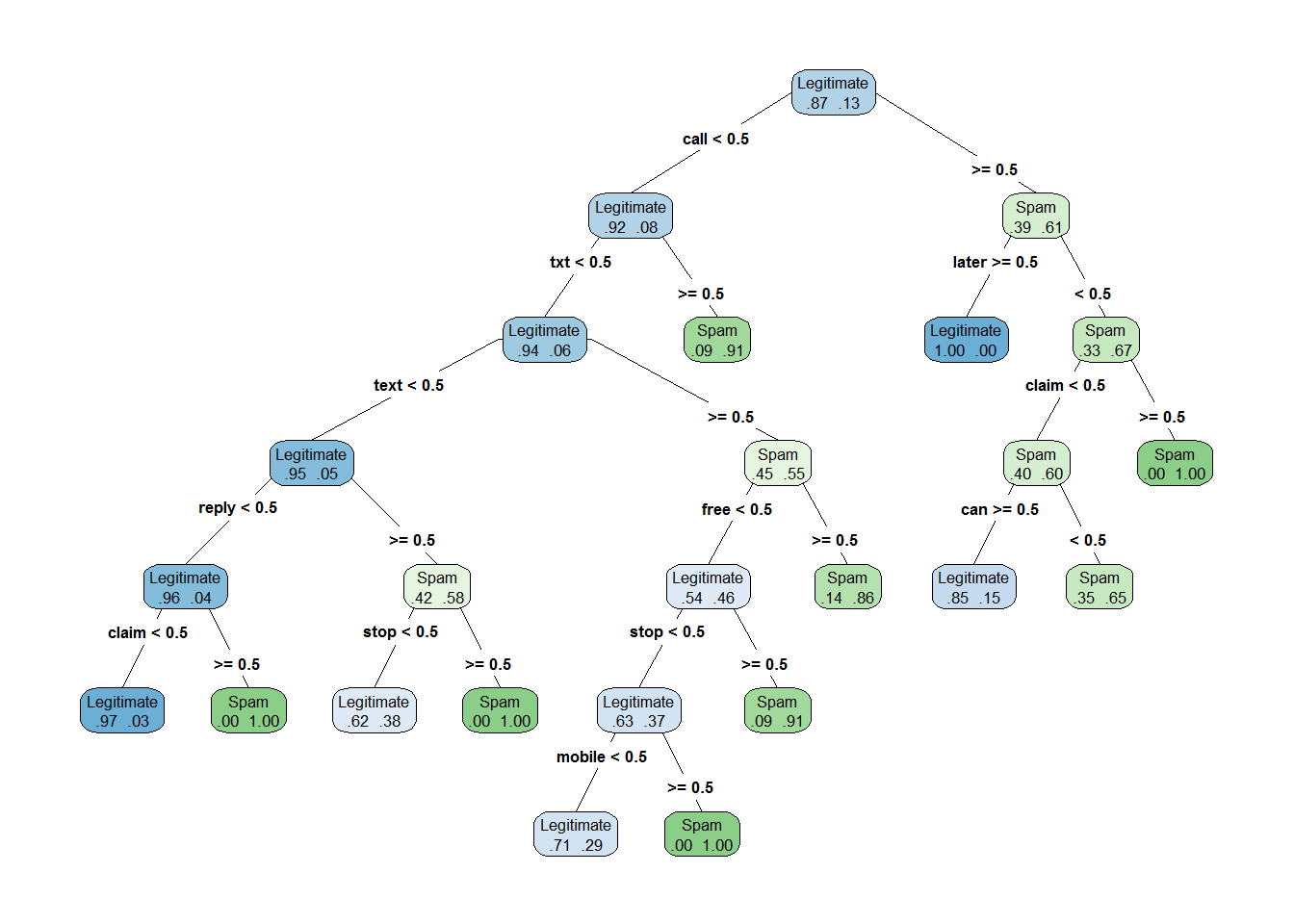
Figure 30 depicts that this model accurately predicted 175 out of 224 Spam SMS, while it incorrectly predicted only 7 out of 1447 Legitimate SMS as Spam SMS, for Setting 1.

On the other hand, this model predicted 137 out of 224 Spam SMS, while it incorrectly predicted 28 out of 1447 Legitimate SMS as Spam SMS, for Setting 2.

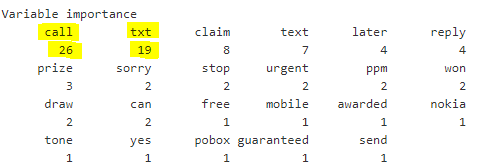
The accuracy for this model in Setting 1 is only 96.7%, while for setting 2 is 93.17%. Therefore, the outcome of the model built in 2 different settings reveals that it works best when modelled in Setting 1.

Building a Recursive Partitioning Decision Tree for both the settings substantiated the outcome of ***Analysis of the Most Frequently Occurring Words in a Spam SMS*** discussed in section 4.1.3.

The decision tree built in Setting 1 revealed that the most important words / tokens in Spam SMS are call and txt, with variable importance for both being 26 and 19 respectively. The least important words, like mobile, nokia, send, were assigned variable importance of only 1.

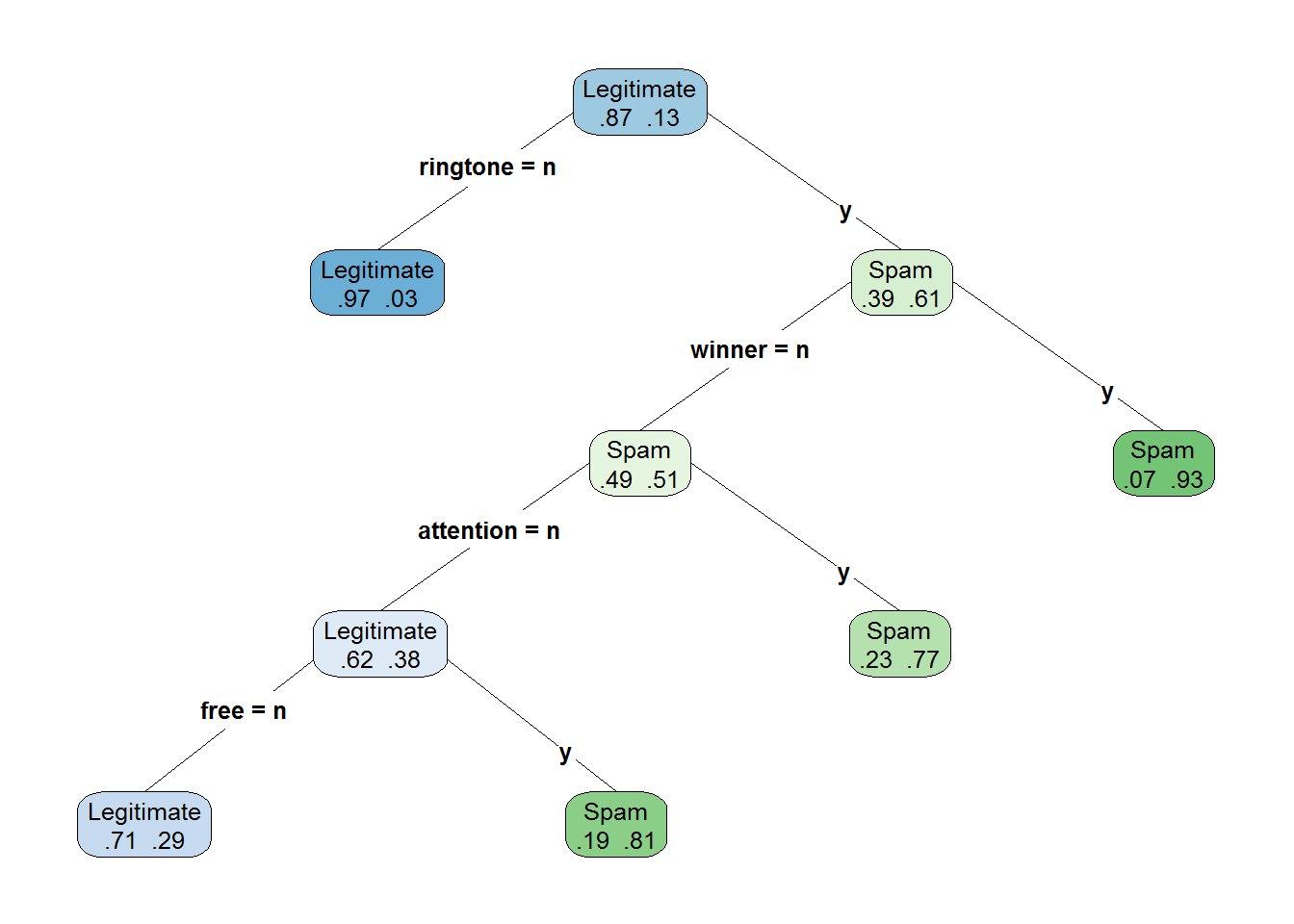


*Figure 31: Recursive Partitioning Decision Tree for Setting 1*



*Figure 32: Variable Importance of Words*

The decision tree built in Setting 2revealed that the most important category in Spam SMS Ringtone, with variable importance for it being 68. The least important category Congratulation was assigned variable importance of only 1.



*Figure 33: Recursive Partitioning Decision Tree for Setting 2*



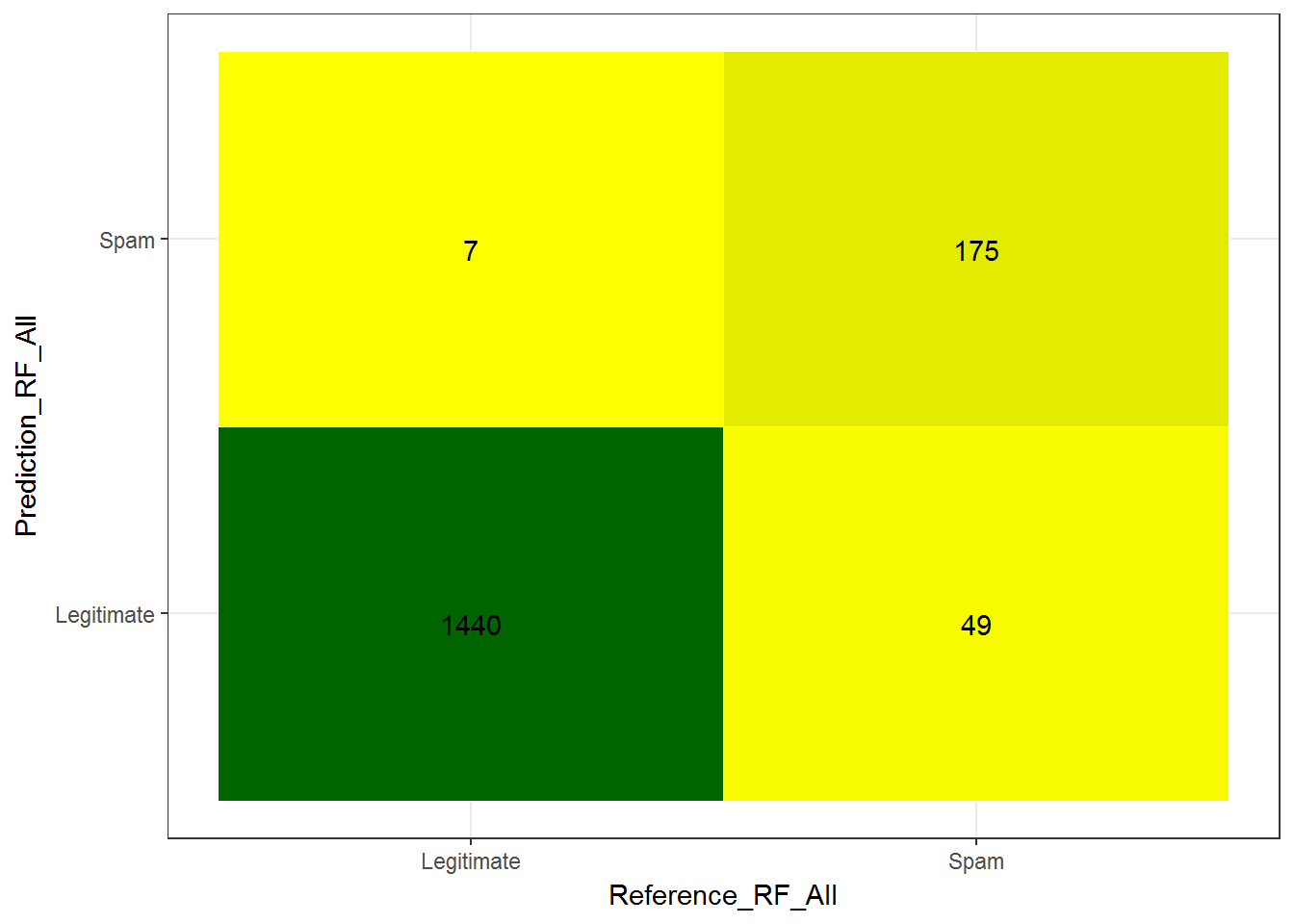
*Figure 34: Variable Importance of Categories*

The visualization of true positive and false positive matrix for Setting 1 is depicted in Figure 35. For this matrix, the positive class is Legitimate Class and the Negative Class is Spam. It reveals the value of correct and incorrect predictions. These predictions are used to calculate the precision, recall, accuracy and incorrect prediction for a test.

Below figure reveals that as the percentage of accurate predictions decrease, the color of the cell gets lighter. 1415 out of 1447 Legitimate SMS are predicted accurately. This cell is known as the True Positives and has the darkest shade cell depicting that high percentage of accurate prediction (97.78%) has been made for True Positives.

175 out of 224 Spam SMS are predicted accurately. This cell is known as True Negatives and has the second darkest shade cell depicting the high percentage of accurate prediction (78.125%) has been made for True Negatives.

False Positive is the cell depicting that 49 out of 224 Spam SMS have been incorrectly predicted as Legitimate SMS. False Negative is the cell depicting that 7 out of 1447 Legitimate SMS have been incorrectly depicted as Spam SMS.



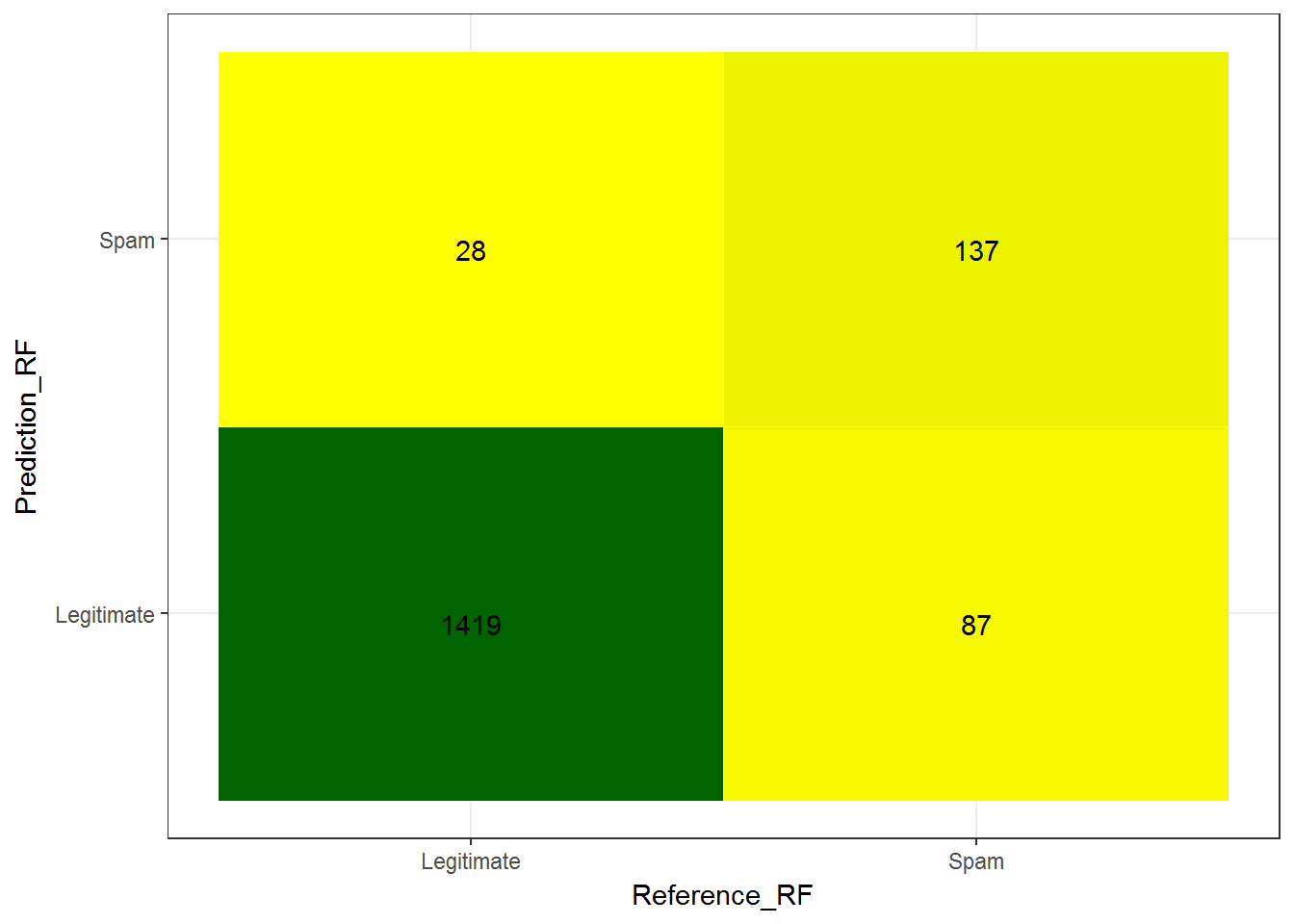
*Figure 35: True Positive and False Positive Matrix Visualization for Setting 1*

Figure 36 depicts true positive and false positive visualization for Setting 2.

Below figure reveals that 1419 out of 1447 Legitimate SMS are predicted accurately. This cell is known as the True Positives and has the darkest shade cell depicting that high percentage of accurate prediction (98.06%) has been made for True Positives.

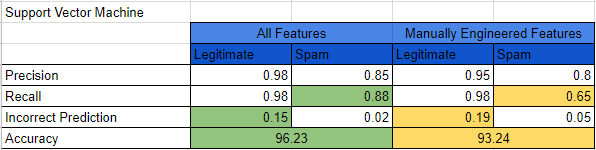
137 out of 224 Spam SMS are predicted accurately. This cell is known as True Negatives and has the second darkest shade cell depicting the high percentage of accurate prediction (61.16%) has been made for True Negatives.

False Positive is the cell depicting that 87 out of 224 Spam SMS have been incorrectly predicted as Legitimate SMS. False Negative is the cell depicting that 28 out of 1447 Legitimate SMS have been incorrectly depicted as Spam SMS.



*Figure 36: True Positive and False Positive Matrix Visualization for Setting 2*

1. **Support Vector Machine**



*Figure 37: Comparative study of the Output of Cross Table for Support Vector Machine in both the Settings*

Figure 37 depicts that this model accurately predicted 196 out of 224 Spam SMS, while it incorrectly predicted 35 out of 1447 Legitimate SMS as Spam SMS, for Setting 1.

On the other hand, this model predicted 145 out of 224 Spam SMS, while it incorrectly predicted 34 out of 1447 Legitimate SMS as Spam SMS, for Setting 2.

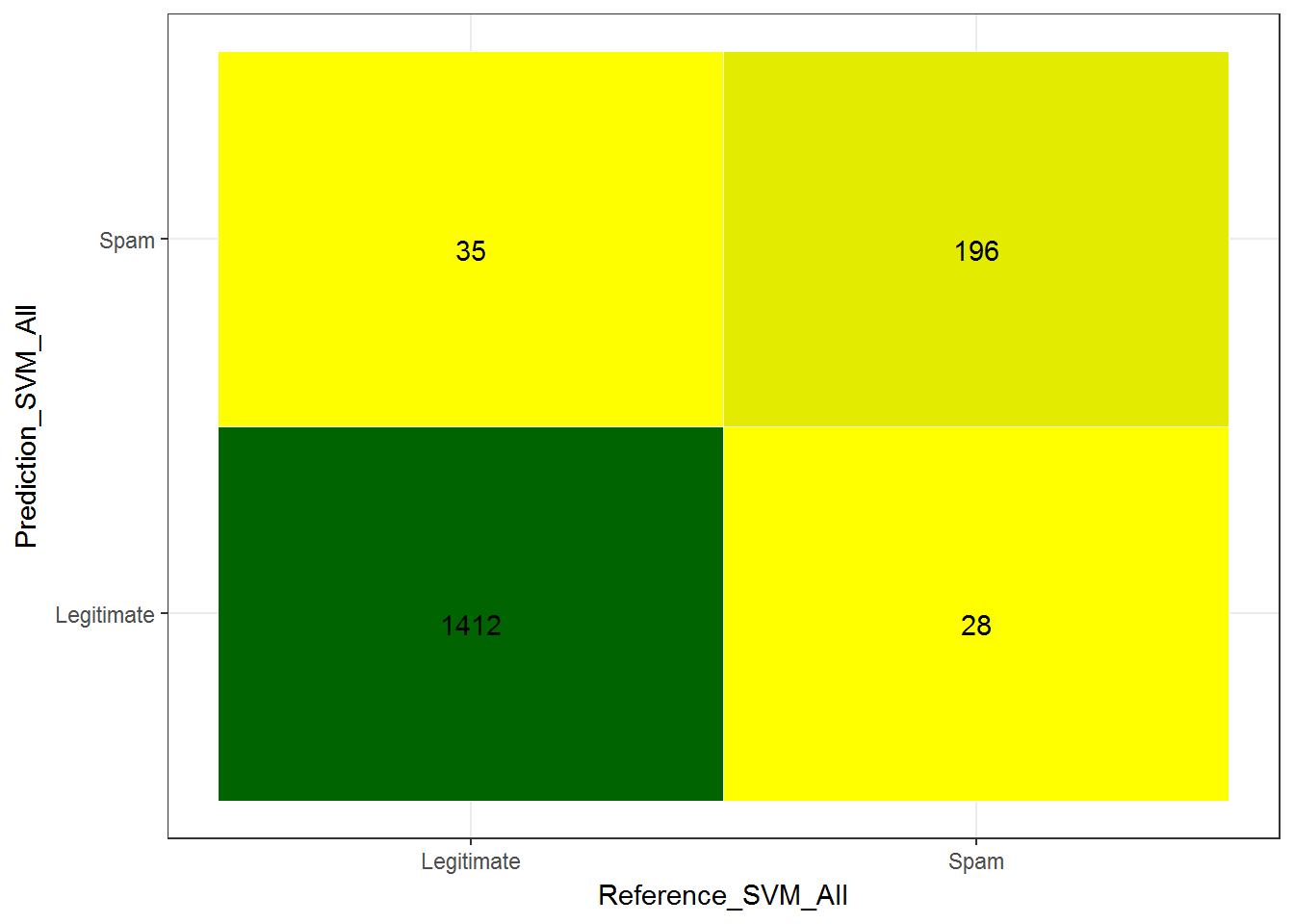
The accuracy for this model in Setting 1 is only 96.23%, while for setting 2 is 93.24%. Therefore, the outcome of the model built in 2 different settings reveals that it works best when modelled in Setting 1.

The visualization of true positive and false positive matrix for Setting 1 is depicted in Figure 38.

Below figure reveals that 1412 out of 1447 Legitimate SMS are predicted accurately. This cell is known as the True Positives and has the darkest shade cell depicting that high percentage of accurate prediction (97.58%) has been made for True Positives.

196 out of 224 Spam SMS are predicted accurately. This cell is known as True Negatives and has the second darkest shade cell depicting the high percentage of accurate prediction (87.5%) has been made for True Negatives.

False Positive is the cell depicting that 28 out of 224 Spam SMS have been incorrectly predicted as Legitimate SMS. False Negative is the cell depicting that 35 out of 1447 Legitimate SMS have been incorrectly depicted as Spam SMS.



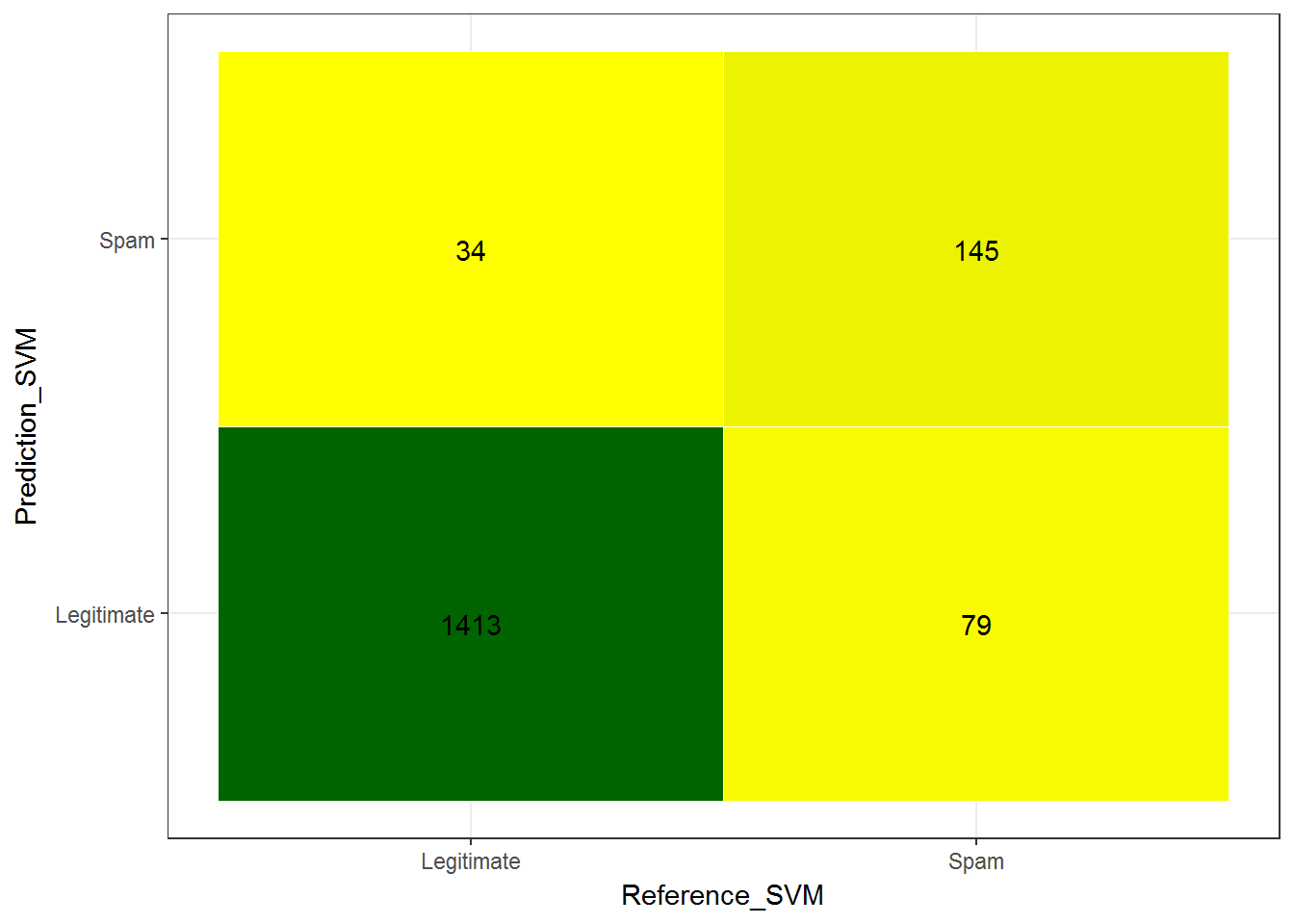
*Figure 38: True Positive and False Positive Matrix Visualization for Setting 1*

The visualization of true positive and false positive matrix for Setting 1 is depicted in Figure 39.

Below figure reveals that 1413 out of 1447 Legitimate SMS are predicted accurately. This cell is known as the True Positives and has the darkest shade cell depicting that high percentage of accurate prediction (97.65%) has been made for True Positives.

145 out of 224 Spam SMS are predicted accurately. This cell is known as True Negatives and has the second darkest shade cell depicting the high percentage of accurate prediction (64.73%) has been made for True Negatives.

False Positive is the cell depicting that 79 out of 224 Spam SMS have been incorrectly predicted as Legitimate SMS. False Negative is the cell depicting that 34 out of 1447 Legitimate SMS have been incorrectly depicted as Spam SMS.



*Figure 39: True Positive and False Positive Matrix Visualization for Setting 2*

We have witnessed that Setting 1 is the best setting, a classifier could be built in. But, the model built would be used in mobile devices, in which memory consumption and CPU storage are one of the factors that determine the performance and efficiency of the device.(Methods and Data Doc) Implementing classifier built in setting 1 would mean an accuracy of approximately 96% but, lower efficiency and performance of the device as more the features a model includes, more would be the memory consumption and CPU storage of the mobile devices. On the other hand, implementing classifier built in setting 2 would mean an accuracy of approximately 93%, that is decrease in accuracy by only 3%, but higher efficiency and performance of the device as only 6 features are used to build this model.

Nowadays, when processing time of a device a crucial requirement amongst users (Methods and Data Doc), therefore, it is important that classifiers built in setting 2 be implemented. Therefore, I suggest that Support Vector Machine built in setting 2 be implemented to predict whether an SMS is a Spam SMS or a Legitimate SMS, in mobile devices. On the other hand, I would suggest Support Vector Machine built in setting 1 if the model is to be implemented in networks.

# **5.Discussion**

## **5.1 Critical Analysis of Findings**

The report details a number of different approaches that have been tested to predict Spam SMS and achieve the objectives of the projects, as discussed in section 1.3.

1. What are the characteristics that distinguish Spam messages from Legitimate messages?

As discussed in section 4.1.3, data analysis and research has revealed the features that differentiate a Spam SMS from a Legitimate SMS. Analysis of features revealed that the most frequently occurring and the most important word in a Spam SMS is **Call**. 46% of all Spam SMS contained the word Call in them. This analysis helps in understanding the core characteristics of Spam SMS, which lacked in the implementation of previous solutions (blacklisting and spoofing detection technique).

1. What is the effectiveness of the classification methods – Support Vector Machine, Decision Trees, Logistic Regression or Bayesian Classifiers in identifying SMS Spam?

As discussed in section 4.2, research has suggested that the classification of Spam SMS is very effective with its accuracy being **96%.** Effectiveness of classifiers has been reviewed using evaluation metrics: Recall for Spam Class, Incorrect Prediction for Legitimate Class and Accuracy of the classifier. The classifiers can be implemented in mobile devices, in networks, or in both. Opposed to previous solutions, implementation of classifiers to predict Spam SMS would reduce the events where a Legitimate SMS would be detected as a Spam SMS and is filtered out.

## **5.2 Significance of Outcomes**

The solutions that have been built in this context like, blacklisting and spoofing/faking detection techniques, are centric in nature and are fully dependent on network and spam policy of the operators, as discussed in detail in section 2.2. They are brittle, straightforward and do not consider the core characteristics of Spam SMS. Therefore, implementing the proposed outcomes of this project would eliminate dependency on networks and operators. They would also eliminate post-hoc and ad-hoc activities, discussed in section 1.2, as these outcomes would filter Spam SMS before it could reach the intended recipient.

I motivated the concern that how Spam SMS has a damaging impact on consumers and mobile network operators. Implementation of proposed solutions would benefit mobile network operators as the quality of messages and services to the customers would improve and the cost of maintaining the networks and operations would reduce. Customers would be benefitted as their confidential, valuable and personal information would remain safe and secure.

## **5.3 Limitations of the Project**

Research would haven been benefitted more if I had more related datasets to work on. More data available on Spam SMS would facilitate production of a more generalized solution. I would have been able to investigate more features that differentiate a Spam from a Legitimate SMS. I could have compared the outcomes of this dataset with the outcomes of another dataset to confirm if the classifiers would produce an accuracy of 96%, in predicting whether an SMS is a Spam SMS or a Legitimate SMS, for the Spam SMS of all the different datasets. Therefore, having restricted to one dataset gave a solution appropriate for that dataset only.

This dataset consisted of only single-language text messages (briefly described in section 1.2.1). Therefore, all of the research including process of tokenization (Data Preparation Phase discussed in section 3.2.4) and feature analysis (Exploration Phase discussed in section 3.2.3) was done only for English language. Therefore, this solution is not robust for a multilingual region, where text messages are exchanged in more than one languages.

## **5.4 Future Work**

A more sophisticated and a generalizable solution can be produced if more data on Spam SMS could be collected and made available. It is difficult to detect particular patterns in Spam SMS due to the content of text messages being less. Increased availability of datasets would mean more text messages to analyze. As submission of Spam SMS increases, it becomes imperative that more data on Spam SMS be collected to facilitate production of an advanced solution.

Datasets should be collected with different languages so that researches and analyses can be set up to build a solution which is appropriate for a multi-lingual region, where text messages are exchanged in more than one languages.

As the number of Spam SMS increases, it becomes important that the proposed solution is collaborated with the industry to validate its performance in real-world.

# **6.Conclusion**

In this project, a data analysis and predictive model is built to accurately predict whether an SMS is a Spam SMS or a Legitimate SMS. Analysis helped in identifying features that differentiated Spam SMS from Legitimate SMS. It showed that the most frequently occurring and the most important words for Spam SMS are Call and Txt, which constitute almost 46% and 24% respectively, of all Spam SMS. Also, effectiveness of 4 classifiers, Naïve Bayes, Logistic Regression, Decision Tree and Support Vector Machine, was evaluated in 2 different settings. It revealed that the most effective classifier for mobile devices is Support Vector Machine built in setting 2, with an accuracy of 93%, and the most effective classifier for mobile networks is Support Vector Machine built in setting 1, with an accuracy of 96%.

Increase in submission of Spam SMS needs an urgent and advanced solution. Therefore, more datasets, of different languages, should be collected and made available in order to achieve it. The proposed solution should be integrated into and put into practice in industry in order to determine its efficiency and performance in real-world.