# Introduction

This document details a plan for a data analysis and research project involving the investigation of Short Message Service(SMS) Spam. This project aims to create a data analysis and 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.

## 1.1 Context of the Project

SMS is an integral feature of common mobile devices that allows users to exchange communication as short text messages(1). According to Informa Telecoms and Media, several trillion SMSs are exchanged yearly and there has been a growth in this exchange from 5 trillion worldwide messages in 2010 to 8.7 trillion messages in 2015(2). SMS messages are an important source of revenue for mobile network operators with this component of revenue being estimated to $136.9 billion in 2015(2).

The exchange of SMS messages has however been hampered by the submission of unsolicited messages that are sent in bulk to subscribers, without their consent and authorization, with the intent to acquire their confidential, personal and valuable information and possibly to misuse it(3). According to a Cloudmark analysis, 92% of spam messages are fraud(3).

The most common types of spam messages sent to the recipients are(3):

* 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.

There has been a growth in reception of spam messages of 300% from 2011 to 2012(3). This growth attributes to two main reasons:(4)

1. The availability of affordable unlimited pre-pay SMS packages which has made SMS spamming a cost-effective opportunity for spammers to extract valuable information out of the recipients(4).
2. Messaging is regarded as a trusted service among the subscribers which makes them more comfortable sharing their confidential information. As a result, messages have a higher response rate as compared to any other service. According to June 2013 statistics, 43% of messages were responded in the first minutes of receiving them.(5)

Spam messages not only affect the consumers but also the Mobile Network Operators. The reasons that make this problem a significant one are:

Mobile Network Operators(MNOs) suffer a huge loss on account of maintaining their network, operations and providing increased customer care services to the customers. Spamming also tarnishes their reputation making them lose many valuable customers(4).

Customers are also left annoyed and worried as their confidential, personal and valuable information is at stake(4).

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(4).

## 1.2 Related Work and Research Gap

There have been a number of anti-spam measures built to solve this problem like –

* 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(4)

These techniques are brittle, simple and straightforward in nature as they do not really take the special and core characteristics of spam messages into account. Moreover, with advances in spamming methods and careful fabrication of spam messages to make them appear as legitimate, there is an urgent need to build a more sophisticated and appropriate model to eradicate this issue. (4)

Also, because of the private nature of SMS exchanges, data to study this problem is scarce.(4)

1.2.1 How tis Project Addresses the Problem

I would work on bridging up the gap by working on the publicly available dataset available at Kaggle. It comprises of 5,574 English, real and non-encoded text messages. All of the messages have accurately been tagged as legitimate and spam. The dataset consists of a total of 425 Spam messages manually selected from Grumbletext website. All of the claims made on this site about the text message being spam are identified and investigated through carefully scrutinizing over a hundreds of webpages. (6)

I would address the problem of SMS Spam by carrying out an exploratory analysis on this dataset and building a data analysis and a predictive model that would be able to accurately identify whether the SMS is Spam or Legitimate.

## Aims and Objectives of the Project

**Objective of the Project** –

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.

The project objective can be achieved by answering the below 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?*

Particularly, the main **aim of this project** is to carry out data analysis on the dataset acquired from Kaggle in the following order:

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

## 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:

1. **Preparation Phase -**

This phase includes data acquisition and manipulation.

1. **Analysis Phase -**

This phase includes carrying out exploratory analysis on the dataset followed by writing the code to build the predictive model. The code would then be executed and refined in case of any issues or bugs.

1. **Reflection Phase -**

This phase would go parallel with the Analysis Phase. The core activity that will be carried out in this phase is to think and communicate the ideas and outputs of the code written in the Analysis Phase.

1. **Final Delivery Phase -**

In this phase, I would finalize the code in R markdown and write down all the observations in the Analysis Report.

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

1. R Markdown
2. Analysis Report

## Outcome of the Project

The result of this project will produce:

* A better understanding of characteristics and features that make a spam SMS different from a legitimate SMS.
* A predictive model that that can accurately identify 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 of spam SMS.

# 2. Project Methodology

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

The project methodology that would best suit this project is ***Data Analysis***. This methodology would be a conjuncture of 4 phases(7):

1. Preparation Phase
2. Analysis Phase
3. Reflection Phase
4. Final Delivery Phase

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

## 2.1 Pictorial Representation

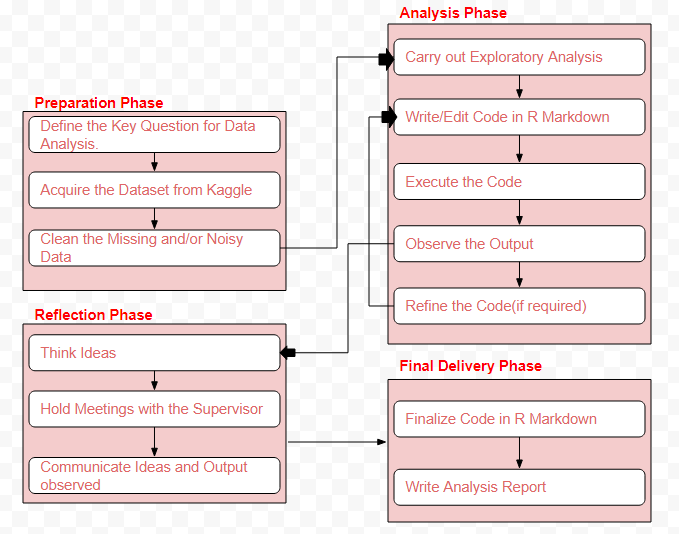
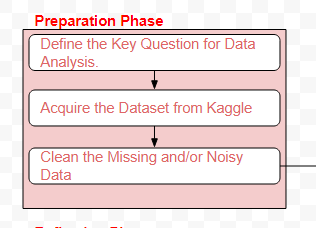


Figure 1: Project Methodology Workflow(7)

### 2.1.1 Preparation Phase



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

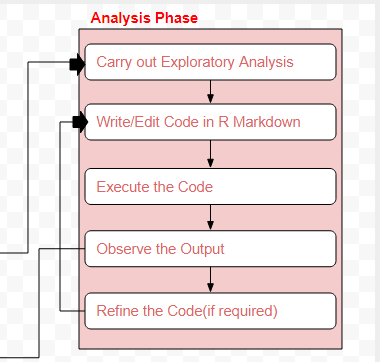
Once the objective has been clearly laid and understood, we acquire the data that would be helpful in achieving the objective. As mentioned earlier, I have acquired the dataset from Kaggle.

Post acquiring the data, I would want to scan through the data to see if there are any missing, noisy or semantically erroneous data in the dataset(7). Removing these entries from the dataset would help me perform the analysis in an appropriate manner.

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

### Table 1: Input/Output Table for Preparation Phase

### 2.1.2 Analysis Phase



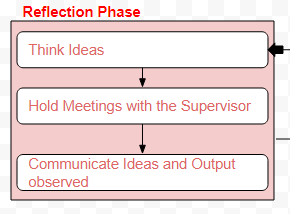
Analysis phase is the most important phase of this project where I would be working towards the predictive model. This phase would start with an exploratory analysis of the dataset followed by continuous iterative cycle of writing the code in R Markdown, executing it an defining it till we get to the desired outcome(7).

I would be continuously consulting my Supervisor, online forums like – Stack Overflow, etc., and relevant research papers while iterating the code.

|  |  |
| --- | --- |
| Input | Output |
| Clean data from Preparation Phase. | Features and characteristics of spam SMS.  Predictive model |

Table 2: Input/Output Table for Analysis Phase

### 2.1.3 Reflection Phase



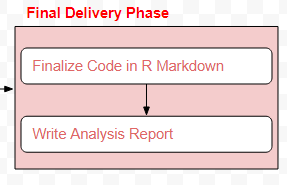
Reflection phase goes parallel with analysis phase. This phase would involve meetings with the Supervisor wherein the input would be the observed output in the analysis phase(7).

Discussions in the meeting would lead to new and better ideas of how to make the predictive model better. Therefore, the output of the meeting would be a new to-do list for me(7).

|  |  |
| --- | --- |
| Input | Output |
| Output from the Analysis Phase | Suggestions to make the project better.  New To-Do List. |

Table 3: Input/Output Table for Reflection Phase

### 2.1.4 Final Delivery Phase



The concluding phase for this project would be the Final Delivery Phase in which I would be disseminating all the observations and learnings in the form of a Presentation and a Report for the final submission.

|  |  |
| --- | --- |
| Input | Output |
| Outputs and observations from Analysis and Reflection Phase | Code in R Markdown  Analysis Report |

Table 4: Input/Output Table for Final Delivery Phase

## 2.2 Breakdown of Tasks

|  |  |  |  |
| --- | --- | --- | --- |
| Period | Period Activities | Deliverables | Duration |
| Period 1: Initial Project Setup | Finalize project team.  Ethics Clearance  Finalize project scope and objective  Make a repository in GitHub  Produce Project Proposal  Download R  Download dataset | Project Unit Study Agreement  Ethics exemption  Repository in GitHub  Project Plan | 4 weeks (Week 1 – Week 4) |
| Period 2: Data Analysis | Clean the data  Explore the data  Statistical Prediction and Modelling of the data  Interpret the results  Write the Analysis Report | Code in R Markdown  Analysis Report | 6 weeks (Week 5 – Week 10) |
| Period 3: Recommendations | Provide recommendations to the Supervisor  Seek recommendations from Supervisor | Recommendations and Suggestions | 1 week (Week 11) |
| Period 4: Wrap Up | Project conclusion activities | Prepare Project Presentation  Prepare Project Report | 2 weeks (Week 12 – Week 13) |

Table 5: Breakdown of Tasks

# Project Management Approach

The project will be managed by *Dynamic Systems Development Method (DSDM)*. This approach encourages iterative and incremental delivery, essentially keeping parameters: time, cost and quality fixed (Agile project management, 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 to be delivered in the next 8 weeks. Therefore, working in timeboxes would ensure that the product is delivered on time.
* *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 would help producing deliverables of the expected quality.

## 3.1 MoSCoW Prioritization for Scope

|  |  |
| --- | --- |
| Prioritization | Deliverables |
| Must Haves (60%) | * Data Analysis of different features of SMS Spam. * Investigation to understand the most effective classification method. |
| Should Haves (20%) |  |
| Could Haves (20%) |  |
| Won't Haves | Building an operational product accurately identifying a spam SMS. |

Table 6: MoSCoW Prioritization for Scope

## 3.2 Detailed Weekly Plan/ Task Breakdown Structure

In this project, activities like – setting up project team of Project Student and Supervisor and describing the overall project and its high level requirements have been 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) is to divide them into increments, which further would be 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 7: Weeks breakup into Increments and Timeboxes

Each timebox and increment would end with a review and feedback session. This will be conducted in weekly meetings and would also focus on the creation of new tasks list for next timebox.

Detailed plan for each increment is as follows:

**Increment 1 -** Analyze 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 using Decision Tree Classification method.  Share the work done with the Supervisor and seek feedback. | Finalize the code in R Markdown and Analysis Report for this Classification Method. |
| 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 for Decision Tree Classification Method. |
| 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 8: Detailed Weekly Plan for Increment 1

**Increment 2 -** Analyze Features of Spam SMS

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Phases | Timebox 1 | Timebox 2 | Timebox 3 | Deploy | Concluding the Project |
| Items | Write the code in R Markdown using Support Vector Machine Classification method.  Share the work done with the Supervisor and seek feedback. | Write the code in R Markdown using Logistic Regression.  Share the work done with the Supervisor and seek feedback.. | Write the code in R Markdown using Bayesian Classifiers.  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. | Prepare Project Presentation.  Prepare Project Report.  Submit report. |
| Deliverables | Knowledge on different characteristics of spam SMS. | Knowledge on different characteristics of spam SMS. | Knowledge on different characteristics of spam SMS. | 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 9: Detailed Weekly Plan for Increment 2

3.4 Communication Plan

In order to ensure success of the project, on-time delivery of the deliverables and meeting the laid quality standards for the outcome, the project team (Project Student and the Supervisor) would follow the Agile Manifesto - "We value Individuals and Interactions over Processes and Tools" very closely(8).

Therefore, we would follow the below laid down communication plan very closely to ensure frequent and active engagement of all the stakeholders.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item | Purpose of Communication | Communication Strategy Opted | Frequency | Participating Stakeholders |
| Status Meeting | To discuss the progress and status of the project. | Face to Face Communication | Weekly | * Project Student * Supervisor |
| Collaborative Working | To work closely with the Supervisor and take feedback. | Working in the assigned Lab at University | Weekly | * Project Student * Supervisor |
| Status Report/Document | To keep a track of the tasks completed/in progress since last week. | * Prepare the report. * Email to the Supervisor. * Put it on GitHub. * Show it in the weekly meeting. | Weekly | * Project Student |
| Quick Communication | To consult the Supervisor in case of any issues. | * E-mail * Slack | As and when required. | * Project Student * Supervisor |
| Review and Feedback Sessions | To get the work done reviewed by the Supervisor and get constructive feedback on it. | * Face to Face Communication * E-mail * Slack | At the each of each increment. | * Project Student * Supervisor |

Table 10: Communication Plan

3.5 Potential Project Risks and Risk Mitigation Strategies

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Potential Risk | Consequences | Risk Level | Risk Severity | Overall Risk | Risk Mitigation Plan |
| Lack of active involvement of the Supervisor | The supervisor might be busy and not have enough time to attend the scheduled meeting. This could result in a solution deviated from expectations, or a solution not meeting the laid quality standards. | High | Medium | High | Weekly meeting with the supervisor.  Quick communication over Slack. |
| Solution Deviation from Expectations | If the solution does not meet the requirements, it might consume a lot of time to re-work on mending it to make it aligned with the requirements. | Medium | High | Medium | Frequent and Incremental Delivery of Solution. |
| Loss of Key Resources | Loss of code and data stored on a hard-drive or computer might pose a risk in case the physical storage device is disrupted. | Medium | High | High | Store the code online – GitHub Repository |
| Project Delivery Exceeding the Promised Time | It will lead to breach of one of the core principles of DSDM: On-Time Delivery. | Medium | High | Medium | Follow the Detailed Weekly Plan closely.  Inform the Supervisor of all the impediments faced. |

Table 11: Risk Assessment

# Ethics

There was no ethics clearance required for the project.

# References

# Corrections made in Response to Feedback for Week 3 Presentation