# Introduction

This document proposes a plan for a development project, called Short Message Service(SMS) Spam Prediction Model. This project aims at building a predictive model in R which can accurately predict whether an SMS is a spam message or a legitimate message. In this project, I would perform data analysis on the dataset acquired from Kaggle followed by writing a code in R markdown to build the predictive model.

## 1.1 Context of the Project

### 1.1.1 Brief Introduction

Short Message Service (SMS) is that integral feature of most mobile devices that allows them to exchange communication as text messages(1). According to Informa Telecoms and Media, exchange of messages has seen an enormous growth worldwide from 5 trillion messages in 2010 to 8.7 trillion messages in 2015(2). Undoubtedly, the growth in this component is a big opportunity for mobile network operators as it attracts a huge amount of revenue for them ($136.9 billion in 2015)(2), but it has a downside to it as well.

The downside is that unsolicited messages are sent in bulk to the subscribers, without their consent and authorization, with an intent to procure their confidential, personal and valuable information and misuse it(3). According to 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,
* PPI Compensation Spam Message, etc.

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

1. 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)

### 1.1.2 Consequences of SMS Spam

Spam messages not only affect the customers but also affect the Mobile Network Operators.

* 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).

As a result, it is the need of the hour to build a predictive model which can accurately predict if the SMS is a spam SMS or a legitimate SMS.

## 1.2 Problem to be Addressed

There have been some anti-spam measures built in this context like – Blacklisting, Simple Filtering, 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. Also, 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, since ubiquitousness of SMS and drop in prices of the unlimited SMS packages has been recent, not much data is available to be researched upon. Even if the data is available, it is radically very different making it difficult for the researchers to reach to a consensus.(4)

1.2.1 Significance of the Project

Therefore, I would work on bridging up the gap by working on the 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 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 the predictive model which would be able to accurately distinguish whether the SMS is Spam or Legitimate.

## Aims and Objectives of the Project

**Objective of the Project** – 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 question:

*Which filtering technique would help make the predictive model better – Support Vector Machine, Decision Trees, Logistic Regression or Bayesian Classification Filter?*

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

The methodology I would be applying to this project would be ***Data Analysis.*** The process of Data Analysis would be carried out in four phases:

1. **Preparation Phase -**

The two activities that will be carried out in this phase would be be b – Acquiring the data and Reformatting the data.

1. **Analysis Phase -**

In this phase, I would be 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 goes 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 success of this project would benefit all the subscribers and the service providers (mobile network operators) in the following ways -

The customers would no longer receive spam messages and nor would their legitimate messages be filtered out as spam message. This would keep their personal, confidential and valuable information secured and safe.

Likewise, mobile network operators would be saved from spending huge amounts for maintaining networks and operations.

Therefore, the outcome of this project would result in accurate prediction of whether a message is a spam message or a legitimate message. Therefore, this would help in reducing the magnitude of frauds made using spam messages by big folds.

# 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

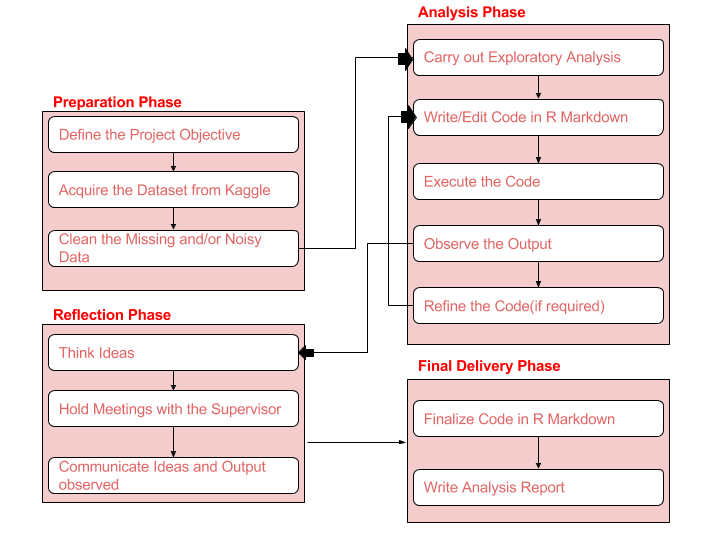
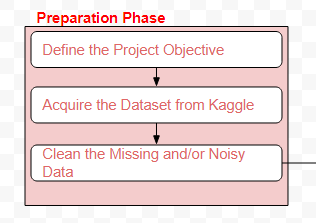


Figure 1: Project Methodology Workflow(7)

### 2.1.1 Preparation Phase

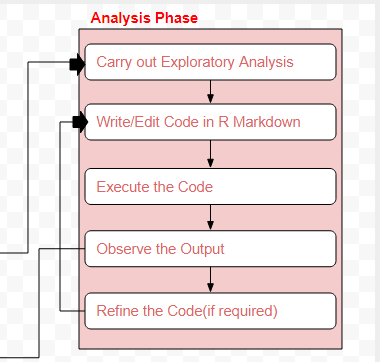


Preparation phase allows us to lay the foundation for the analysis. This phase starts with defining the objective of the project – what the project is about and what do we aim to do in it.

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.

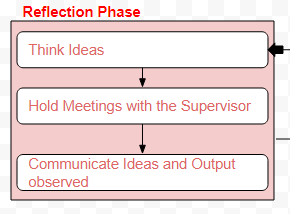
### 2.1.2 Analysis Phase



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.

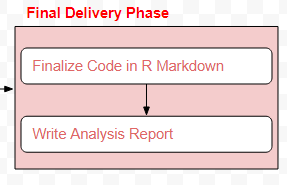
### 2.1.3 Reflection Phase



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).

### 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.

# Project Management Approach

## 3.1 MoSCoW Prioritization for Scope

## 3.2 Task Breakdown Structure

## 3.3 Detailed Weekly Plan 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. * 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 1: 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 Related and Important Resources | Loss of code 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 2: Risk Assessment

# Ethics

There was no ethics clearance required for the project.

# References

# Corrections made in Response to Feedback for Week 3 Presentation