

An Analysis of Spam SMS Features

Data Analysis and Research Project

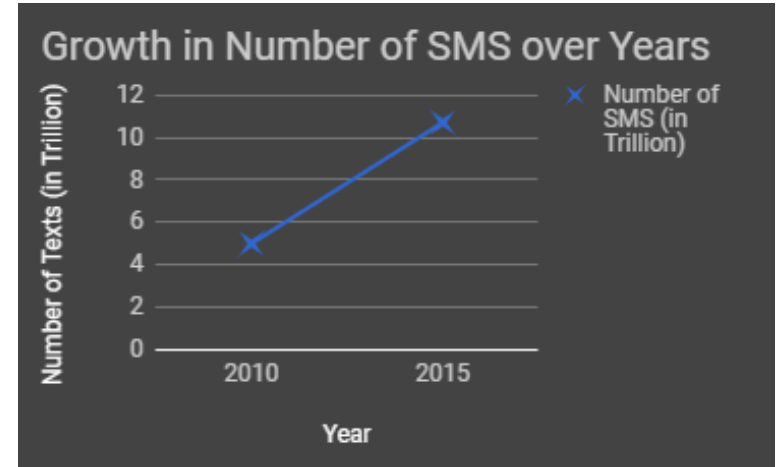
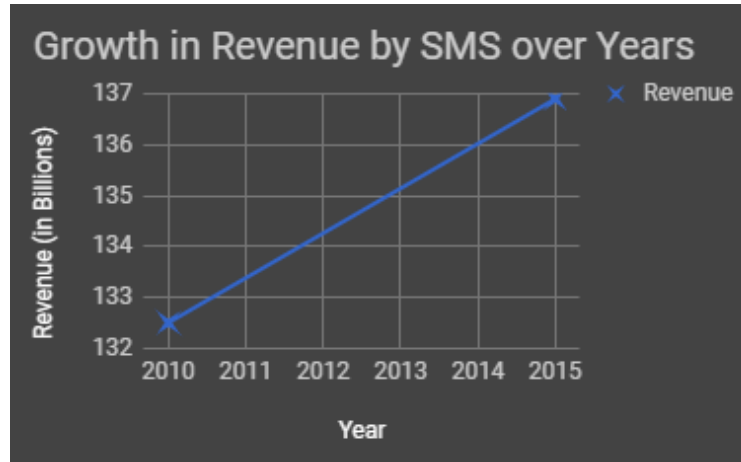
Harshita Jain

n9539361

Supervisor: Dr. Guido Zuccon



Context



- 92% of spam SMS are fraud
- Overall rate of receipt grew by 300% from 2011 to 2012.

Reason and Effects

Reason

- The availability of affordable unlimited pre-pay SMS packages
- Customers being more comfortable in sharing their confidential information via SMS
 - 43% of messages were responded in the first 15 minutes of receiving them

Effects

- Mobile network operators suffer a loss
 - Higher network costs
 - Higher operating costs
 - Increased customer care costs
 - Tarnished reputation
- Annoying for customers
 - Loss of confidential and valuable personal information

Gap in Previous Solutions

- Simple solutions are used - Blacklisting and Spoofing/Faking Detection
 - Brittle by nature
 - Do not take content of messages into account
 - Require on-going management
- Not much data available for research studies
- Available datasets are different

Data Set

v1	v2
ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
ham	Ok lar... Joking wif u oni...
spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
ham	U dun say so early hor... U c already then say...
ham	Nah I don't think he goes to usf, he lives around here though
spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, 1.50 to rcv
ham	Even my brother is not like to speak with me. They treat me like aids patent.
ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune
spam	WINNER!! As a valued network customer you have been selected to receive a 900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.

Source: Kaggle

Project Purpose and Deliverables

- **Purpose of the Project:**
 - Analyze the data to understand features that make Spam SMS different from a Legitimate SMS.
 - Build a predictive model which can accurately predict if an SMS is a Legitimate SMS or a Spam SMS
- **Project Deliverables:**
 - R Markdown
 - Analysis Report

APPROACH

Data Analysis and Research

Preparation Phase

Input	Output
Acquired data from Kaggle	Key question for Analysis. Clean data.

Exploration Phase

- Analyzed Length of Messages v/s Number of Texts for each Label
- Manually Selected Differentiating Features of Spam SMS
 - Verified by Producing Word Cloud for Spam SMS
 - Visualized Uni-Grams using Bar-Plots
 - Visualized Bi-Grams and Tri-Grams using Venn Diagram

Input	Output
Clean data	Analysis of features that make a Spam SMS different from a Legitimate SMS

Data Preparation Phase

- Prepared Data to be used to Build Predictive Models
 - Created a Clean Corpus by Transforming Text to Lower-Case, Removing Numbers, Stop Words, Punctuation and White Space.
 - Split the data into a ratio of 7:3: 70% Training Set and 30% Test Set

Input	Output
Clean data	Data ready to be used to build predictive models.

Classification Phase

- Built 4 Different Models for 2 Different Scenarios
- Compared each Model for each Scenario

Classifiers used:

- Decision Tree
- Support Vector Machine
- Logistic Regression
- Naïve Bayes

Scenario 1: Modelled for the 'manually selected differentiating features' of Spam SMS

Scenario 2: Modelled for 'all the features' of the data

Input	Output
Data from Data Preparation Phase	Precision, Recall, F1 and Accuracy Measures for each model in each scenario
	Best Scenario for each Model

Reflection Phase

- Think Ideas
- Meetings with the Supervisor
- Communicate Ideas and Observed Output

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

Final Delivery Phase

Input	Output
Output from Exploration Phase	Code in R Markdown
Output from Classification Phase	Analysis Report
Output from Reflection Phase	

Contributions(1)

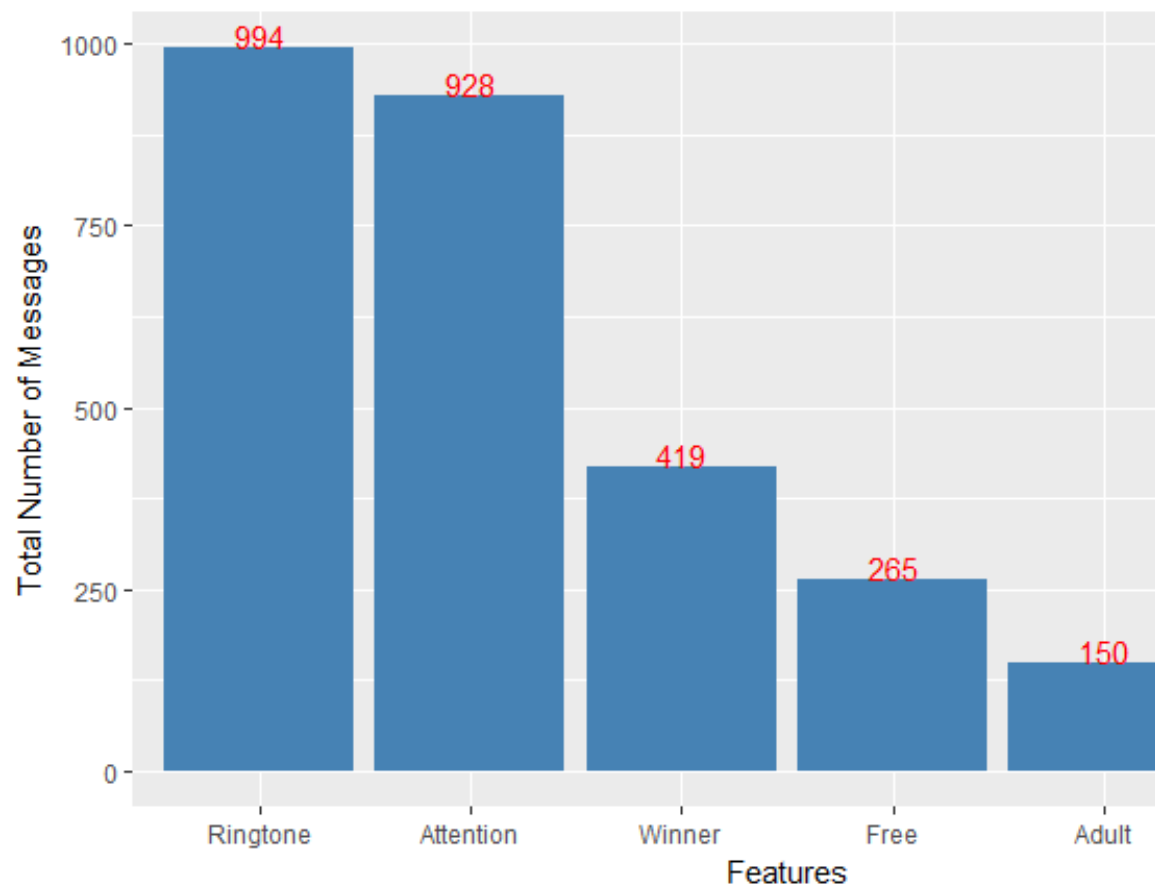
- *Analyze the data to understand the differentiating features of Spam SMS.*
 - Explored Length of Messages
 - Explored words that occur most frequently in Spam SMS.

A histogram showing the distribution of message lengths for two groups. The x-axis is labeled 'Length of Message' and ranges from 0 to 1000 with major ticks at 0, 250, 500, and 750. The y-axis is labeled 'Number of Messages' and ranges from 0 to 400 with major ticks at 0, 100, 200, 300, and 400. The red series has a primary peak of approximately 450 messages at a length of about 50, and a secondary peak of about 250 messages at a length of about 180. The teal series has a single prominent peak of about 150 messages at a length of about 180. Both distributions are right-skewed, with most messages being shorter than 250 characters.

Word Cloud for Messages tagged as Spam



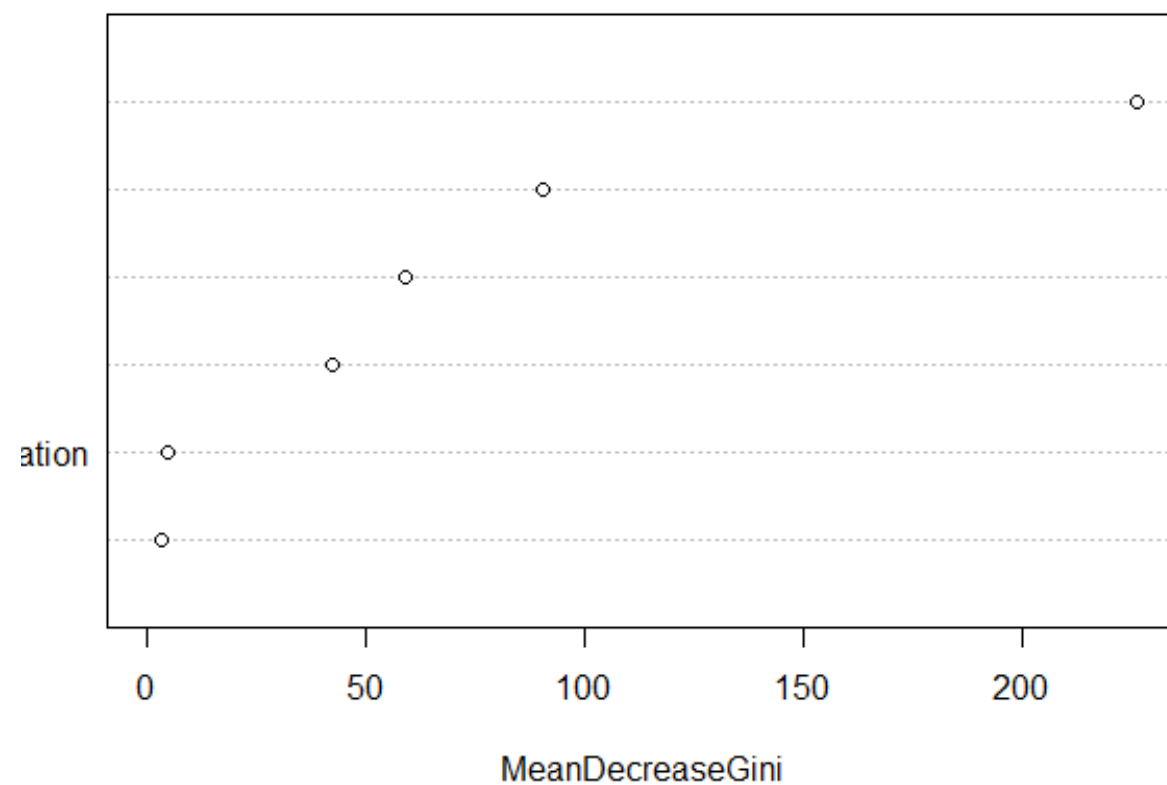
Distribution of Message Lengths v/s Number of Texts



Features v/s Number of Messages

Importance of each Feature

Importance of each Token



Contributions(2)

- *Build a predictive model which can accurately predict if an SMS is a Legitimate SMS or a Spam SMS*
 - *Built 4 models using different classifiers:*
 - *Decision Tree with Random Forest*
 - *Support Vector Machine*
 - *Logistic Regression*
 - *Naïve Bayes*
 - *Built on two types of data:*
 - *Manually explored and selected 6 most important features of the Spam SMS*
 - *All features of the data*



Thank You!!

Any questions?

References

- Team, A. V., Shaikh, F., Jain, K., Gupta, A., & Gupta, D. (2016, October 11). A Complete Tutorial to learn Data Science in R from Scratch. Retrieved August 09, 2017, from <https://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scratch/>
- Guo, P. (2013, October 30). Data Science Workflow: Overview and Challenges. Retrieved August 09, 2017, from <https://cacm.acm.org/blogs/blog-cacm/169199-data-science-workflow-overview-and-challenges/fulltext>
- Delany, S. J., Buckley, M., & Greene, D. (2012). SMS Spam Filtering: Methods and Data. Retrieved from <http://arrow.dit.ie/cgi/viewcontent.cgi?article=1022&context=scschcomart>
- Whitepapers. (n.d.). Retrieved August 09, 2017, from <https://www.cloudmark.com/en/s/resources/whitepapers/sms-spam-overview>
- (n.d.). Retrieved August 09, 2017, from [https://archive.ics.uci.edu/ml/datasets/SMS Spam Collection](https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection)
- The DSDM Agile Project Framework (2014 Onwards). (2017, April 18). Retrieved August 09, 2017, from <https://www.agilebusiness.org/resources/dsdm-handbooks/the-dsdm-agile-project-framework-2014-onwards>