IS482 ANALYTICS PRACTICUM

ANALYTICS PRACTICUM PROPOSAL

Prepared by: Team 3GAG (Ang Jia En Clara, Chiang Ling Yi, Sennett Khong Wei Quan

1-1-2016

Table of Contents

INTRODUCTION & BACKGROUND

SGAG is a digital content firm which based their content on Singaporean's daily life. It started off among a group of friends in Singapore Management University where their first meme fuelled the attention of many people during the first McDonald's shortage of curry sauce. The founders, Xiao Ming and Karl Mak, thought that it would be interesting to have a localised 9GAG, focusing on all Singapore related issues. As more incidents in Singapore surfaced, SGAG feeds on them to create more relatable memes, which is their main selling point - to showcase what Singaporeans are really feeling. In 2013, the team developed a structured content strategy and started working with various clients to create hype around issues Singaporeans could all relate to.

SGAG has since continued to grow strong, has over 500,000 likes on their Facebook page and has also expanded to creating videos on Facebook as well as YouTube. With its listicles being very successful, SGAG wishes to place more focus on their videos, where competition is of abundance. Therefore, by studying historical data, SGAG would like to find out which kinds of videos are more popular and more views, in hopes of knowing the consumer's preferences to in turn increase viewership.

BUSINESS PROBLEMS & MOTIVATION

As social media and the internet become increasingly widespread in Singapore, users are seen to be trending towards video related content from text based and pictured contents. For SGAG's videos to remain popular amidst current competitive forces, they seek to place their focus on understanding their audience in order to bring curated contents which are relevant to their target audience. These objectives raised specific business questions like:

1. Which type of videos does my audience like most? (Based on Content type, Video Quality, Character preferences)
2. What have we done well/wrong (Evaluate) and what improvements should we make (Improve)?
3. Which are the popular and trending videos on the internet space to emulate (Future plans)?

PROJECT OBJECTIVES

Through this project, our sponsor wishes to improve their content strategy with an emphasis on the company's video content channels (Facebook and YouTube). Based on the audience interaction data provided through these media channels, the sponsor seeks to understand their target audience's preferences in their video content (e.g. the virtual characters used in the videos, the different type of videos being posted.)

This project seeks to provide insights to answer questions aforementioned and allow our sponsor's media unit to accurately curate content which are tailored towards their audiences' preferences.

SCOPE OF PROJECT

With the objectives mentioned above, our team aims to seek solution for our sponsor's business problems through adopting the analytical methodology with the following steps which will be mentioned in details below:

1. Discovery
    a. Leading business domain
    b. Business problem discovery
    c. Business research
2. Data Preparation
    a. Data Collection
    b. Data Preparation
    c. Initial Exploratory data analysis
    d. Data cleaning
    e. Data integration
3. Model Planning
    a. Interactive exploratory data analysis
4. Model Building
    a. Analysis and modelling
    b. Model validation
    c. Insights discovery and actionable recommendations
5. Communication of result
    a. Presentation of result to sponsor
    b. Analytics Practicum Conference
    c. Final Paper

Through analysis of the data from Facebook/YouTube, our group aim to at the basic level provide summary statistics followed by highlighting trends which could be hidden among the data. This include but not limited to comparisons between derived columns comparing percentages between unique users, total impressions and comparison of video views between the different view categories.

Video specifications, like production complexity, would also assist to answer business questions like what have been done well and what could have been done better. By analysing the production complexity on the use of professional format videos or instant iPhone exported videos, the sponsor would be able to know the type of video formats preferred by their audiences in the past, and whether it is it starting to change in recent days.

Lastly, to identify the popular video constituents (which includes genre, video type, SGAG characters) and compare the performances of videos of based on its constituents that are pre-defined, including views, likes, as well as comments, to see if the audiences have a dominant preference towards certain constituent or not. Some examples of video types include public submission, republished videos and SGAG-produced video.

In summary, we would be analysing the videos provided by SGAG based on these main categories which is aimed to provide actionable insights for our sponsor:

1) Sponsored and Non-sponsored content video (Genre,video type, characters involved, video resolution, video dimensions)
2) Video Comments Text-mining
    a) To identify sentiment e.g positive, negative, neutral for videos
    b) To identify viewer's perception e.g degree of hard-sell perception]

Through cross referencing video post with Likes, Reach, Shares, New Page likes and Page unlikes with the video impression, unique views, percentages views we will be able to dive deeper within into finding out the impact of how different categories and subcategories has effect on SGAGs videos.

DATA COLLECTION AND DESCRIPTION

The data will be provided by our sponsor and will be split into two sets upon receipt of the data. This is to facilitate the future process of training of an analytical model and secondly for testing the model.

FACEBOOK

The data provided from our sponsor has a limit of one year from current day. As the focus of the project is on improving their video content strategy, the data provided will be extracted from video segments under Facebook Insights. The following tables below are some of the sample fields which are provided from our sponsor for the analysis.

Video Metrics Total vs. Unique

Column Name	**Description**
Post ID	Unique identifier given to each post entry
Permalink	Static link given for each post entry
Post Message	Message written along with each post entry
Type	Type of content posted
Countries	No Information available in this column
Language	No Information available in this column
Posted	Date and Time post entry was uploaded online
Lifetime Post Total Impressions	Lifetime: The number of impressions of your Page post. (Total Count). This includes the number of times it was shown on the Facebook Homepage and repeated views
Lifetime Post Total Reach	Lifetime: The total number of people your Page post was served to. (Unique Users)
Lifetime Total Video Views	Lifetime: Total number of times your video was viewed for more than 3 seconds. (Total Count)
Lifetime Total 30-Second Views	Lifetime: Total number of times your video was viewed for 30 seconds or viewed to the end, whichever came first. (Total Count)

Lifetime Unique 30-Second Views	Lifetime: Number of unique people who viewed your video for 30 seconds or to the end, whichever came first. (Unique Users)
Lifetime Total Views to 95%	Lifetime: Total number of times your video was viewed to 95% of its length. (Total Count)
Lifetime Unique Views to 95%	Lifetime: Number of unique people who viewed your video to 95% of its length. (Unique Users)

Video Metrics Auto play (AP) vs. Click to play (CTP)

Column Name	**Description**
Post ID	Unique identifier given to each post entry
Permalink	Static link given for each post entry
Post Message	Message written along with each post entry
Type	Type of content posted
Countries	No Information available in this column
Language	No Information available in this column
Posted	Date and Time post entry was uploaded online
Lifetime Total Video Views	Lifetime: Total number of times your video was viewed for more than 3 seconds. (Total Count)
Lifetime Auto-Played Video Views	Lifetime: Number of times your video started automatically playing and people viewed it for more than 3 seconds. (Total Count)
Lifetime Clicked-to-Play Video Views	Lifetime: Number of times people clicked to play your video and viewed it more than 3 seconds. (Total Count)
Lifetime Total 30-Second Views	Lifetime: Total number of times your video was viewed for 30 seconds or viewed to the end, whichever came first. (Total Count)
Lifetime Auto-Played 30-Second Views	Lifetime: Number of times your video started automatically playing and people viewed it for 30 seconds or to the end, whichever came first. (Total Count)

Lifetime Clicked-to-Play 30-Second Views	Lifetime: Number of times people clicked to play your video and viewed it for 30 seconds or to the end, whichever came first. (Total Count)
Lifetime Total Views to 95%	Lifetime: Total number of times your video was viewed to 95% of its length. (Total Count)
Lifetime Auto-Played views to 95%	Lifetime: Number of times your video started automatically playing and people viewed it to 95% of its length. (Total Count)
Lifetime Clicked-to-Play views to 95%	Lifetime: Number of times people clicked to play your video and viewed Fit to 95% of its length. (Total Count)

INITIAL DATA PREPARATION

As part of preparation work to prepare for this analytics project, we came up with some derived columns to allow us to see the differences between the numbers and to understand their relationships. Here is a list of the derive columns which have been added

Video Metrics Total vs. Unique

Column Name	**Description**
Video Type	Simple categories of the videos on our sponsor's Facebook Page
% Reached	Percentage of unique users compared to the total impressions
% videos more than 3 seconds (Total)	Percentage of the number impression video was viewed more than 3 seconds compared to total impressions
% videos more than 3 seconds (Unique)	Percentage of unique people who viewed more than 3 seconds compared to total unique views
% videos more than 30 seconds (Total)	Percentage of the number impression video was viewed for 30 seconds or to the end compared to total users
% videos more than 95% (Total)	Percentage of the number impression video viewed to 95% compared to total impression
% videos more than 95% (Unique)	Percentage of unique people who viewed the video to 95% compared to total unique views
Difference in percentage (3 Seconds vs 30 Seconds) for all impressions	Difference in percentage of impression who viewed more than 3 seconds compared to those who viewed at least 30 seconds or completed the video (which ever comes first).
Difference in percentage (3 Seconds vs 95%) for all unique views	Difference in percentage of unique views more than 3 seconds compared to those unique views up to 95%.

Video Metrics Auto play (AP) vs. Click to play (CTP)

Column Name	**Description**
% of Autoplay	Percentage of autoplay views
% of click-to-play	Percentage of click to play views
% differences (AP vs CTP)	Difference in percentage between autoplay and click to play
% of Autoplay with 30 seconds view (Total)	Percentage of the number of views which automatically started playing and people viewed it for 30 seconds or to the end, whichever came first
% of Click to play with 30 seconds view (Total)	Percentage of the number of views where people clicked to play the video and viewed it for 30 seconds or to the end, whichever came first
% difference between AP and CTP for 30 seconds view	Difference in percentage between % of Autoplay with 30 seconds view (Total) and % of Click to play with 30 seconds view (Total)
% of Autoplay with 95% view (Total)	Percentage of the number of views which automatically started playing and people viewed it for 95%
% of Click to play with 95% view (Total)	Percentage of the number of views where people clicked to play the video and viewed it for 95%
% difference between AP and CTP for 95% view	Difference in percentage between % of Autoplay with 95% view (Total) and % of Click to play with 95% view (Total)

METHODOLOGY

Our team plans to adopt the data analytics lifecycle through a continuous iterative process with the following processes repeated over the span of the semester.

Leading Business Domain

Our team aims to understand SGAG through exploring the content published by SGAG through multiple channels, namely SGAG's social network services, website and media contents.

Business Problem Discovery

Through interviews and speaking with the founder of the company, our team aims to understand the business problems of SGAG to assist us in identifying problems and translating them into analytics objective.

Business Research

Through online case studies on web content analytics, Facebook and YouTube analysis, we aim to gain a deeper understanding of the various techniques applied for similar analysis.

Data Collection

SGAG will be able to provide us with data of the uploaded video posts, generated from their Facebook and YouTube pages with variables listed above. Apart from what is given, we would also like to gather additional information of the videos, including the comments on the videos, characters involved in the video, video resolution, as well as the video type, whether it is an original or reposted video. As SGAG uploads videos of various categories and resolutions (Filmed using a phone versus video camera), we would like to accurately categorise them to further increase the dimension of the data set.

Data Preparation

Some of the datasets provided are not structured in the format suitable for analysis to be done (for example: Additional derived columns need to be added, tables need to be traversed etc.).

Data required for analysis on the video specifications in relation to audience receptivity will require additional use of open source tools like EXIFTool to extract the EXIF (Camera data) data like video width and video height.

Initial Exploratory Data Analysis

Before we do a more concrete and complete exploratory data analysis (EDA), we first would like to do a basic EDA to identify outliers that might, in future, affect our analysis results.

Data Cleaning

For data cleaning, we would be going through the data sets to identify missing data, inconsistent values, inconsistency between fields and duplicated data. For each row of data with missing value(s), we would then check the number of missing fields it has to decide whether we should omit the row of data completely, or should we predict a value for the missing field(s) using suitable techniques such as association rule mining and decision trees.

Extracted data from EXIFTools will require the removal of irrelevant fields to remove less important data. The required fields from the EXIF data of videos are:

1) Video height
2) Video width
3) Video length

The video width and height will allow us to quickly classify the type of production be it professional film crew or hand phone quality.

Data Integration

For the videos uploaded on Facebook, we intend to collect 3 sets of data - Video data on Facebook, data of the post on Facebook that was used to share the video and the data on video specifications (resolution, dimensions etc.). Facebook data of the video and post can be integrated using the full ID (xxxxx_xxxxxx). The dataset on the video specifications can be integrated by using the last 16 digits (digits after the underscore) of the ID.

Interactive Exploratory Data Analysis

Post integration, further analysis will be conducted on the data to uncover basic trends. This is to bring out insights which might not be clear on Facebook and YouTube individually, but as a whole. It also includes summary statistics to highlight to our sponsor the key high performers based on the available metrics.

Outliers for the data may also include viral videos which the sponsor have curated. For these videos, they may be put together for further analysis. This may also include videos with interesting video titles. An example of this identifying this could be a trend with videos post which may have higher video plays but lower “at least 30 seconds view” or “view to 95% length”.

Results can then be presented in an interactive dashboard using softwares such as SAS Visual Analytics, Tableau or QlikView.

Analysis and Modelling

We will then proceed to do our analysis of the various components and create analytical models using related variables. For our prediction models, suitable techniques, such as stepwise regression, can be carried out. We will then determine which variables are the best predictors using a threshold, and remove other variable which are not producing the best results.

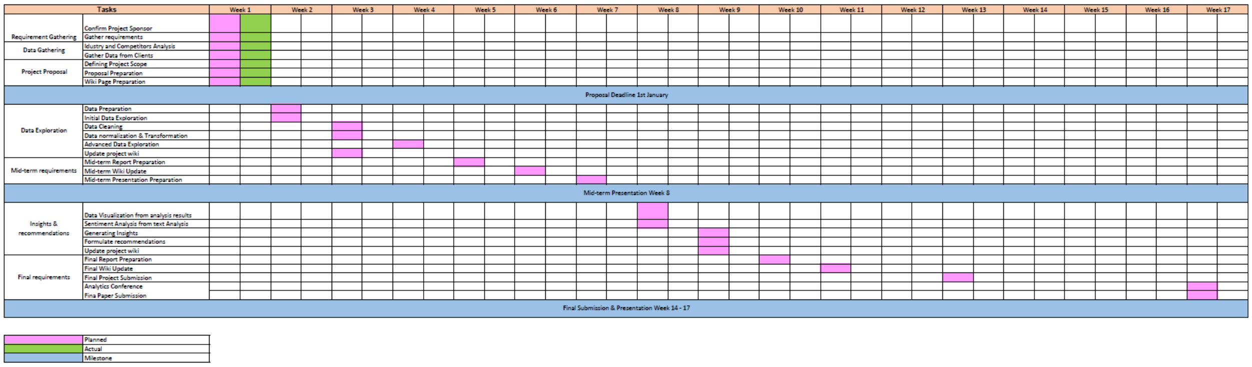
Model Validation

Model validation ensures that the models meet the intended requirements with regards to the methods used and the results obtained. Ultimately, we aim to have models that addresses the business problem and provides results with fairly high accuracy. We will split our data sets into 2 parts - training data and testing data. After we have built our models using the training data (Decision tree, linear/multilinear regression etc.), we need to check if the model is over or under fitting by running the test dataset through the models and compare both the results of the test and training data. For regression models, we also need to check the R-square values to ensure the model's accuracy.

Insights Discovery and Actionable Recommendation

Insights can then be gathered from the various results, which we can leverage on to provide recommendations to SGAG in hopes of improving viewership of their videos.

PROJECT SCHEDULE



Tasks		Week 1		Week 2		Week 3		Week 4		Week 5		Week 6		Week 7		Week 8		Week 9		Week 10		Week 11		Week 12		Week 13		Week 14		Week 15		Week 16		Week 17	
Requirement Gathering	Confirm Project Sponsor	🟪	🟩																																
	Gather requirements	🟪	🟩																																
Data Gathering	Idustry and Competitors Analysis	🟪	🟩																																
	Gather Data from Clients	🟪	🟩																																
Project Proposal	Defining Project Scope	🟪	🟩																																
	Proposal Preparation	🟪	🟩																																
	Wiki Page Preparation	🟪	🟩																																
Proposal Deadline 1st January																																			
Data Exploration	Data Preparation			🟪																															
	Initial Data Exploration			🟪																															
	Data Cleaning					🟪																													
	Data normalization & Transformation					🟪																													
	Advanced Data Exploration							🟪																											
	Update project wiki					🟪																													
Mid-term requirements	Mid-term Report Preparation									🟪																									
	Mid-term Wiki Update											🟪																							
	Mid-term Presentation Preparation													🟪																					
Mid-term Presentation Week 8																																			
Insights & recommendations	Data Visualization from analysis results															🟪																			
	Sentiment Analysis from text Analysis															🟪																			
	Generating Insights																	🟪																	
	Formulate recommendations																	🟪																	
	Update project wiki																	🟪																	
Final requirements	Final Report Preparation																			🟪															
	Final Wiki Update																					🟪													
	Final Project Submission																									🟪									
	Analytics Conference																																	🟪	
	Fina Paper Submission																																	🟪	
Final Submission & Presentation Week 14 - 17																																			

🟪	Planned
🟩	Actual
🟦	Milestone

DELIVERABLES

The final deliverables of this project are:

- Project Proposal
- Midterm Report
- Midterm Presentation
- Final Report
- Final Presentation
- Project Wiki Page
- Project Poster

STAKEHOLDERS

Main Stakeholders of this project includes:

- Project Sponsor: Karl Mak, Co-Founder of SGAG and MGAG
- Project Supervisor: Prof. Kam Tin Seong, Associate Professor of Information Systems

REFERENCES

Analytics Kick Start Workshop by Prof. Kam Tin Seong, Associate Professor of Information Systems

Bynne, T. (2007, October 30). Web Analytics and Web Content Management. Retrieved February 5, 2017.

Jacky Yap (December 2014). The Story Of SGAG Singapore Revealed: How It All Started & Tips To Virality. Retrieved from https://vulcanpost.com/266631/story-sgag-sg/

Jarvinen, J., & Karjaluoto, H. (2015, April 21). The use of Web Analytics for Digital marketing performance measurement.

Social Media Masters - Facebook Case Study: Taco Bell pt 1. (2016, November 30). Retrieved January 05, 2017, from http://www.nextanalytics.com/facebook-case-study-taco-bell-pt-1/

Social Media Masters - Facebook Case Study: Taco Bell pt 2. (2016, November 30). Retrieved January 05, 2017, from http://www.nextanalytics.com/facebook-case-study-taco-bell-pt-2/

We’re Karl Mak & Xiao Ming, co-founders of SGAG. Ask us anything! (2016, April 21). Retrieved January 05, 2017 from https://www.techinasia.com/talk/sgag-ama

Zhao, S. (2014). Content Analysis of Facebook pages: Decoding Expressions Given Off. Retrieved January 5, 2017.