CAPSTONE PROJECT – FOUNDATIONS OF DATA SCIENCE

**Increasing Engagement Efficiency – An investigation into the optimal type and post time for Facebook Live Sellers in Thailand**

**William Belcher**

Many small businesses who wish to grow are instructed to use social media as it is a ‘powerful tool to promote your business to potential customers and gain valuable” [1] Ensuring the content these businesses post engages with the target audience effectively will reduce time and money spent on social media management. This report aims to uncover the most advantageous combination of post type and time posted that will yield greater engagements by analysing the post data of 10 Thai fashion and cosmetic retailers. The methodological approach is to utilise the data science practices of data pre-processing and exploratory visual techniques to draw conclusions from the data set. This investigation exposed, that after applying the processes outlined above, the Analytics data set contains a wealth of information that can be extracted and further analysed. Overall, the inquiry concluded that Videos engaged more the followers with the best time to post an advert being on either a Monday or Tuesday between the hours of 6 to 10AM. A larger data set containing more information on the engagement patterns of the pages followers could provide more insight into what trends will lead to a higher engagement rate.

**Introduction**

In an ever-changing digital world, online presence has become key. Facebook, one of the world’s largest social media platforms enables businesses to share insight and promotions with those who wish to follow the said business’s page. The business can then create a post; either a link, photo, video, or status, for the pages followers to see. A follower can then engage with a post by carrying out “actions such as reacting to, commenting on or sharing the ad, claiming an offer, viewing a photo or video, or clicking on a link”[2]. However, for the duration of this report, an engagement will be limited to either a reaction, comment, or share. Ensuring the content a business creates is efficient at engaging users is imperative as any time spent on creating non-engaging posts is money squandered. Understanding what type of content is most engaging as well as precisely what time to post the content will enable a business to grow exponentially without incurring a significant financial burden. The objectives for this investigation will be to determine which factors have the most significant effect on the engagement efficiency of a post enabling a business to maintain a large return on investment in social media marketing.

**Data**

The Facebook Sellers data set contains the post information of “10 Thai fashion and cosmetics retail sellers, posts of a different nature (video, photos, statuses, and links) and engagement metrics consisting of comments, shares, and reactions.” [3] The data set is 6 years’ worth of post data from each business collated into one file and downloaded from the UCI Machine Learning Repository [3]. The following table details all 12 variables

|  |  |
| --- | --- |
| Variable | Type |
| status\_id | qualitative, nominal |
| status\_type | qualitative, nominal |
| status\_published | quantitative, continuous |
| num\_reactions\_ | quantitative, continuous |
| num\_comments | quantitative, continuous |
| num\_shares | quantitative, continuous |
| num\_likes | quantitative, continuous |
| num\_loves | quantitative, continuous |
| num\_wows | quantitative, continuous |
| num\_hahas | quantitative, continuous |
| num\_sads | quantitative, continuous |
| num\_angrys | quantitative, continuous |

Table 1: List of variables and their associated types

In total, there were 7051 observations. It is worth noting that each post was advertising a product and the status\_type describes the content of each ad.

**Methods**

A range of data science methods were implemented to pre-process the data set before it could be explored further. The software R-Studio was utilised along with its associated language, R [4]. A select few third-party libraries were also installed to enabling the data wrangling process to be made more seamless, namely, **dplyr**, **ggplot2**, **tidyr.**

* Data Representation

The original data set is a ‘comma-separated values’ (.csv) file. By making use of the **read.csv** function and its associated arguments arguments **header** = TRUE, **sep** = “,”, **dec** = “.” and **fileEncoding** = “UTF-8-BOM”, the data set can be imported into the R workspace. The use of the fileEncoding argument was used to remove trailing characters in the column names of the data set.

* Data Cleaning

The data set was first cleaned by removing 4 columns which were not part of the original data. These columns are likely to have appeared from format conversion. The data set was then filtered through a dplyr function to remove NA values. This step proved to be unnecessary as no rows were removed from filtering. Duplicate entries of the status\_id was then checked, revealing that 53 observations had been duplicated. The **distinct** function from dplyr removed the duplicate entries with the remaining data set contain 6997 entries.

* Type Conversion

The read.csv function handled many of the data types correctly, leaving only the status\_published as an unknown data type. This issue was rectified by stripping the time components from the column through the use of the **strptime** function and passing in the associated format **"%m/%d/%Y %H:%M**", representing “month/day/year hour:minute”. From there, the time components are then converted to a POSIXct data type which enables the date to be sorted in either an ascending or descending fashion.

* Missing Value Imputation

Noting that the num\_reactions column seemed to be a row wise sum of each reaction type, the assumption had to be validated. By checking the num\_reactions value against the sum of each reaction type, 9 columns were found to contain mismatching data. This issue was rectified by overwriting the num\_reactions value with the correct calculated value as opposed to the other option of removing the row completely.

* Mutating Columns

The cleaned and corrected data was first passed through a custom function used to extract the month that the status had been posted from its status\_published. This function can be found in appendix X. The base R function weekdays was able to extract the day of the week the status was published with the status\_type being converted to title case through the use of **tools::toTitleCase**. A final mutation was completed to create a column containing the total engagements (combination of reactions, comments and shares) of each status. A final mutation was utilised to extract the hour a status was posted, giving each post an hour, day and season, which could be utilised later. This cleaned, corrected and mutated data became the basis for further mutations and summarisations which were completed as required on subsets of importance.

* Group-based Data Summarisation

Several groups were selected from the data, including the season posted, weekday posted, hour posted and status type. A summarisation was then completed on each group to find the mean engagements received.

* Exploratory Visualisation using ggplot2:

The basic commands **ggplot**, **geom\_boxplot, geom\_bar, geom\_point, geom\_line, geom\_histogram, geom\_density, geom\_tile** and **geom\_step,** along with the applicable arguments were used to create visualisations of the chosen subsets.

**Results and Discussion**

The original data was a comma separated file with 7,050 observations of 12 variables. After the pre-processing of the data was completed, the remaining data set contained 6,997 rows of 17 variables. Once the initial data had been processed, the initial visual exploration of the data could be completed. An initial probing into the number of each post type was completed, revealing that the majority of sellers are relying on photo and video posts to interact with the followers with these post types combining for almost 88% of posts. Figure 1 illustrates this disparity.

Chart, bar chart

Description automatically generated

Figure 1: A tally count of each status type

A brief look into the mean engagements received on each status type revealed that Videos were able to receive on average over double the amount of engagements as statuses and over quintuple times that of a Photo. Figure 2 displays this fact.

Chart, bar chart

Description automatically generated

Figure 2: Mean engagements per status type

An investigation into the mean number of engagements received on posts per season was conducted next. Figure 3 shows the relationship between the mean engagements and the season, noting the significant increase in reactions during spring as opposed to those during the remaining season.

Chart, bar chart

Description automatically generated

Figure 3: Average engagements received on a post per season

Further breaking down the mean engagements on each post by looking at the weekday each status was posted reveals a decrease in engagements over the mid to late working week followed by a resurgence over the weekend. The data was then grouped by status type to add a further dimension to the plot. Figure 4 illustrates this trend.

Chart, line chart

Description automatically generated

Figure 4: Mean engagements received each day, separate by status type

Investigating the hour at which each status was posted reveals an interesting trend. Figure 5 shows the maximum engagements received on a post per hour it was posted with the mean values overlayed in a line graph. Note the differing scales of the y axis.

Chart, histogram

Description automatically generated

Figure 5: Maximum engagements per hour the status was posted, line indicates mean

A statistical analysis into the correlation between when a Video was posted, and the engagement received was then conducted in order to find what variables impacted the number of engagements the most. As such, a correlation heatmap was constructed and can be found below. A value of one indicates a strong linear correlation with a value of negative one indicating a strong negative correlation, a value of 0 indicates there is no correlation between the two variables.

Chart

Description automatically generated

Figure 6: Correlation Heatmap for the data set

When looking at the engagements row, it can be noted that modifying the status type alters the engagements with a negative correlation of -0.24. The remaining adjustable variables provide no correlation with the number of engagements. The decimal value is due to the correlation being calculated from the factor level of the status type.

These levels are:

|  |  |
| --- | --- |
| Decimal | Level |
| 1 | Video |
| 2 | Status |
| 3 | Link |
| 4 | Photo |

Table 2: Levels of the status type factor

A decision importance tree was then graphed to provide more insight into which variables provided the most impact on the engagements. From the plot, it can be seen that the status type provides the most change in engagements. Followed by season, weekday and then hour.

Chart, bar chart

Description automatically generated

Figure 7: Decision importance tree

A final engagement heatmap can then be constructed to provide a quick and visual way of finding when the best time to post a video is. The resulting heatmap can be seen below.

A screenshot of a cell phone

Description automatically generated

Figure 8: Engagement heatmap

**Conclusion**

When looking at the decision importance tree, as well as figure 2, figure 3, figure 4 and figure 5, it can be seen that posting a video will drastically improve the engagement efficiency of the post. Posting in spring will, on average, bring a 20% increase to the engagements received. Altering the weekday and hour a status is posted provide minimal benefits when compared to the status type and season. If one were so inclined, posting on either a Monday or Tuesday between the hours of 6 to 10am will provide an approximate 45% increase in engagements over a post during any other time period.

In conclusion, the best combination of post type and time posted that will maximise engagements is a Video posted between 6 to 10AM on either a Monday or Tuesday.

**References**

* [1] Western Australia Small Business Development Corporation, retrieved October 2020, <https://www.smallbusiness.wa.gov.au/business-advice/marketing/8-ways-to-grow-your-business>
* [2] Post Engagement | Facebook Business Help Centre, retrieved October 2020,  
  <https://www.facebook.com/business/help/735720159834389>
* [3] Facebook Live Sellers in Thailand Data Set, retrieved October 2020, <https://archive.ics.uci.edu/ml/datasets/Facebook+Live+Sellers+in+Thailand>
* [4] RStudio Team (2020). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA <http://www.rstudio.com/>

Mode: Desktop. Version 1.25042