By: Jaslyn Toh Lixuan

**Introduction**

As the world become more technology driven, the impacts are seen when leading E-commerce companies such as Alibaba and Amazon emerged and won the hearts of online shoppers (Singapore Economic Development board, 2017). With added convenience via online transactions, customers can shop anywhere, at any time. This leads to higher competition for in-stores retailers. The successes of leading retail companies are no doubt, contributed by effective marketing of products on social media. Digital marketing is one of E-commerce’s best strategy. With targeted marketing and extensive promotions on social media sites, sales can be generated in a seamless and effective way (Spring, 2019). Therefore, Thai Association of Cosmetics and Fashion Companies should leverage on such marketing strategies to boost their sales.

**Project Objectives**

The objective of this project is to help Thai Association of Cosmetics and Fashion Companies maximise their social media strategy, by identifying the most effective engagement metrics to leverage on to sell products via the social media platform Facebook. Facebook is a platform where sellers can leverage to conduct live streaming, postings of products and customer engagement in the most seamless and convenient manner. Positive and negative sentiments can be detected just by looking at the reactions of customers.

**Expected results & Hypothesis**

With time metrics given, there is potential to look at trends and prediction of engagement with time data. By looking at the relationship between given variables, the expected results of the prediction should show the best metrics contributing to the most to customer engagement on Facebook, and at what time of the year.

Hypothesis: By identifying the right type of posts, seasonality of engagements and relationship of engagements, the right type of posts can be posted at the right time, generating more engagements with customers as they are more likely to view the post.

**Description of dataset**

A picture containing background pattern

Description automatically generated

*(Fig 1. Glimpse of Dataset)*

The given Facebook dataset consists of 7,050 rows and 16 columns. There are 4 columns (“Column1, Column2, Column3, Column4”) with missing values. The original data set has a combination of 3 characters and 9 integer values. The num\_reactions is the sum of num\_likes, num\_loves, num\_wows, num\_hahas, num\_sads and num\_angrys.

As observed, the mean values of the numeric columns are not evenly distributed (See Appendix A). There is a high disparity between mean values for **A: num\_reactions, num\_comments and num\_likes** and **B: num\_loves, num\_shares, num\_wows, num\_hahas, num\_sads, num\_angrys.** Group A columns have more than 200 counts per observation, while group B columns have less than 40 counts per observation. The number of reactions per post is the highest compared to the rest of the metrics. A particular post has garnered a total of 20,990 comments. However, this data point is an outlier because it exceeds the outlier score (> 67.5), marking it as an extreme outlier.

**Data cleaning and processing**

1. Drop missing values

A total of 4 columns (“Column1, Column2, Column3, Column4”) have been dropped from the data set to remove missing values contributing to noise in data.

1. Partitioning

The dataset is then partitioned into Group A and Group B, where group A is used for the modelling while group B will be used for exploratory data analysis.

A: num\_reactions, num\_comments and num\_likes, status\_type, status\_published

B: status\_type, num\_loves, num\_shares, num\_wows, num\_hahas, num\_sads, num\_angrys

1. Splitting and Encoding Time Columns

The status\_published column denotes a time data. The status\_published column in group A is first split into new columns: day, day of week, month and hour. Thereafter, one hot encoding is applied to convert the categorical variables into columns of dummy variables.

1. New column – Engagements

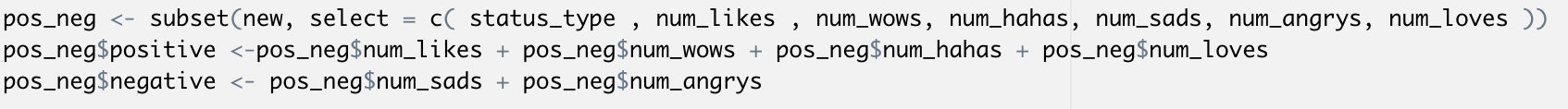
The number of engagements is a sum of the num\_likes, num\_comments and num\_shares. As these three metrics are the default buttons on a post, engagement levels can be measured by combining the effects of all three metrics.

**Overview of analysis and methods used**

1. **Time Series plot: Trends in number of comments, shares and reactions before and after live videos**

After the introduction of Facebook live, there is no immediate increase in engagement in all three metrics (See Appendix B). However, we can observe an increase in all three types of engagement from the late 2017 to 2018 onwards, which could be contributed by the increasing awareness of Facebook live videos. One notable observation is that the number of shares and comments after Facebook live has significantly increased over time as compared to before live. The number of reactions remain consistent, albeit slight increase.

1. **Bar plot: Investigating the positive and negative sentiments towards different status type**



Chart, bar chart

Description automatically generated

*(Fig 2. Positive and Negative sentiments grouped by status\_type)*

The positive sentiments (blue plot) are categorised by the number of likes, wows, hahas and loves, while negative sentiments (red plot) are categorised by the number of sads and angrys. In the purple bar plot, the number of likes have been removed as it has a huge disparity of counts with the rest of the metrics. Videos generally received a higher average of positive and negative reviews. This can be due to the high potential for engagement (See Appendix C), resulting in more actions taking by users to take action and review the lives videos.

1. **Scatter plot: Relationships between engagement metrics**

A picture containing diagram

Description automatically generated

*(Fig 3. Relationship between num\_reactions, num\_comments and num\_shares)*

It is observed that the number of comments and number of shares have a positive correlation (Fig. 3). This means that as the number of comments increase, the number of shares will likely increase as well. The number of comments and number of shares is moderately positively correlated with the number of reactions. This shows that there is potential for predicting num\_comments with the num\_shares, vice versa.

1. **Time Series: Average engagements by hour**

According to the trend as seen in the figure, the best time to post is at 10am which garners the most engagements (See Appendix D). The worst time to post is at 1pm as the engagements is at the lowest. With this observation, we can examine whether we can predict engagements with date and hour.

**Summary of predictive models**

The predictive models used are Linear Regression and Multi-linear regression, Random Forest, Support Vector Machine. A combination of different models are used to conduct trial and error to find the best Facebook metrics which can be used to predict engagement levels. Principal Component Analysis is then used to reduce the dimensionality of the dataset.

Linear Regression is utilised to find relationship between two or more metrics. With linear regression, we can observe how one metric can complement another to increase engagement rates. Random Forest is an ensemble method, where many decision trees are used to derive at a prediction (Swaminathan, 2019). Random Forest is utilised to find out the importance of features to predicting engagements on Facebook. It is utilised since ensemble learning methods aggregates multiple learning models to produce better outcomes and accuracy (Khandelwal, 2020). Support Vector Machine is a supervised learning model, used for classification and regression tasks. The algorithm creates a hyperplane which separates the data into classes (Pupale, 2019).

**Model, Evaluation and Insights**

**Principal Component Analysis**

**Chart, line chart

Description automatically generated**

*(Fig. 4 Scree plot)*

The PCA method is used to reduce the number of components in the data and extract the more important ones. As the cut-off is 1.0, this scree plot indicates that 2 components should be retained. The principal components are then applied to the training (70%) and testing (30%) dataset (See Appendix E).

**Linear Regression:** Predicting number of engagements with the number of reactions.

Y = Engagements, X = num\_reactions

Diagram

Description automatically generatedChart, scatter chart

Description automatically generated

The results shown in the residual plots shows a positive result as the data points are evenly spread along the red dotted lines. The linear model obtained an R2 score of 0.4639, RMSE score of 382.4785 and MAE of 290.5487. This shows that the model is moderately good at predicting the observations.

**Multi-linear Regression: Predicting the number of comments with num\_shares, num\_reactions and status\_type**

X = status\_type, num\_shares and num\_reactions, Y = num comments

**Diagram, schematic

Description automatically generatedChart, scatter chart

Description automatically generated**

From the results of the residual plots, we observe that the data points are well fitted along the red linear lines, showing positive results. The R2 score is at 0.2403, RMSE is at 247.468 and MAE score is 106.6441. The model has a low but adequate R2 score, showing that the model fits the observations and num\_shares, num\_reactions and status\_type can be used to predict engagements.

**Random Forest**

**Chart

Description automatically generated**

The Mean decrease Accuracy %IncMSE and the Mean Decrease Gini (IncNodePurity) is plotted and we observe that the RC1-3, num\_reactions and status\_type\_video are the top 5 attributes. The higher the %IncMSE, the more important the attributes. The higher the IncNodePurity, the more useful the variable. This shows that the number of reactions and video type best predict engagements.

The R2 score is at 0.9870, RMSE is at 59.63 and MAE score is 16.5449. The model has a high R2 score, showing that the model fits the observations almost perfectly for predicting engagements levels.

**Support Vector Machine**

The model is ran on the same training and testing data set as the previous models. The R2 score is at 0.9913, RMSE is at 48.7014 and MAE score is 38.0316. The model has a very high and almost perfect R2 score, indicating that the model fits the observations very well and can be used to predict engagement levels with the given inputs.

**Overall Results for predicting Engagements**

*Red – Linear Regression, Blue – Random Forest, Green – SVM*

**Chart, scatter chart

Description automatically generated**

Data points are close to 0, showing that the predicted results are similar to the target results. Random Forest and SVM performed the best as seen in the diagram.

**Issues faced**

The main issue faced is to fine-tune the models with the input attributes, to get the best model score. Initially, the linear model was applied to predict engagement levels with time columns (day, month, date). However, the results generated negative R2 scores which indicates that the models are underperforming with these inputs.

**Conclusion**

From the analysis and prediction results, we can conclude that video status types is the best strategy and indeed can increase engagement levels as it proves to engage audiences the most. From our analysis, a high number of reactions, whether positive or negative were contributed by videos. This shows that videos can attract the attention of people to take an action. The Thai Association of Cosmetics and Fashion Companies should seriously consider leveraging Facebook live to promote their products to maximise engagement levels with their customers.

**Usefulness of results**

The results achieved a high positive score, which shows that the models are effective to use to predict engagement levels. The models utilised, such as SVM and Random Forest avoids overfitting of data. This helps the model achieve a more accurate outcome when tested on new datasets.

**Limitations of results**

One limitation of the models used is that random forest does not work well with large number of data as a large number of trees will result in a slow and ineffective real-time predictions. If the Thai Association wants to adopt this method of prediction to improve real time sales during Facebook live, it will be an added cost because to generate the best possible outcomes, more iterations of trees have to be ran and this will increase their costs.

**Further analysis**

A potential Facebook metric which can be proposed for the retailers to track is Reach. Reach is the number of people your post or content is seen by on Facebook (Jackson, 2020). Reach is beneficial to track the views of advertisements or page posts. This gives opportunity to see what your audiences like, why certain posts are more seen and preferred by others. This metric will definitely improve the social media marketing strategies of the Thai Association of Cosmetics and Fashion Companies, helping them effectively promote their products to targeted audiences.

**Appendix A**

**Summary Statistics**

**A screenshot of a cell phone

Description automatically generated**

A picture containing graphical user interface, table

Description automatically generated

Graphical user interface, application

Description automatically generated

|  |  |
| --- | --- |
| **Before live** | **After live** |
| **Chart, histogram  Description automatically generated** | A picture containing table, sitting, light, room  Description automatically generated |
| Chart  Description automatically generated | Chart, histogram  Description automatically generated |
| A picture containing chart  Description automatically generated | **Chart, histogram  Description automatically generated** |

**Appendix B**

**Comparison of metric counts before live and after live**

**Appendix C**

**Average engagements by status type**

Video has the highest average engagement (comments, share, reactions), indicating that such posts will more likely garner more interaction with the content posted by sellers.

**Chart, bar chart

Description automatically generated**

**Appendix D**

**Average engagements by hour**

Chart, bar chart

Description automatically generated

**Appendix E**

**Principal Component Analysis Parameters**

**Graphical user interface, text, application, email

Description automatically generated**

**Appendix F**

**Linear Regression**

**Parameters:**

**A picture containing text

Description automatically generatedGraphical user interface, text, application

Description automatically generated**

**Graphical user interface, text, application

Description automatically generated**

**Score for Linear regression:**

**Graphical user interface, text, application, email

Description automatically generated**

**Score for Multi-linear regression:**

**Graphical user interface, text, application, email

Description automatically generated**

**Appendix G**

**Random Forest**

**Random Forest parameters:**

A picture containing graphical user interface

Description automatically generated

**Random Forest results:**

**Graphical user interface, text, application

Description automatically generated**

**Appendix H**

**Support Vector Machine**

**Parameters:**

**Graphical user interface, text, application

Description automatically generated**

**Support Vector Machine results:Graphical user interface, table

Description automatically generated**

**Graphical user interface, application

Description automatically generated**

**References**

Singapore Economic Development Board (2017, January 25). How will Alibaba shake up the Southeast Asian e-commerce landscape? Retrieved from <https://www.edb.gov.sg/en/news-and-events/insights/innovation/how-will-alibaba-shake-up-the-southeast-asian-e-commerce-landsca.html>

Spring, T. (2019, June 10). Council Post: E-Commerce Strategy: Keys To Mastering Multi-Channel Distribution. Retrieved October 02, 2020, from <https://www.forbes.com/sites/forbestechcouncil/2019/06/10/e-commerce-strategy-keys-to-mastering-multi-channel-distribution/>

Swaminathan, S. (2019, January 18). Linear Regression - Detailed View. Retrieved October 02, 2020, from <https://towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86>

Khandelwal, R. (2018, November 14). Decision Tree and Random Forest. Retrieved October 02, 2020, from <https://medium.com/datadriveninvestor/decision-tree-and-random-forest-e174686dd9eb>

Pupale, R. (2019, February 11). Support Vector Machines(SVM) - An Overview. Retrieved October 02, 2020, from <https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989>

Jackson, D. (2020, August 31). 11 Facebook Metrics Every Brand Needs to Track. Retrieved October 02, 2020, from <https://sproutsocial.com/insights/facebook-metrics/>