**Project**

Organization campaign via Facebook in Thailand

**Introduction**

Due to the rapid growth of technology, there has been a huge development and growth in the industry. In addition, technologies are used more efficiently. Social media is an integral part of today's technology and becoming part of our daily lives. The use of social media is growing exponentially. One of them is Facebook, which is one of the social media that have the highest number of users in the world. There are many different uses for Facebook. such as sharing a moment, communicating, searching for information, and including news updates as well. In addition, organizations have adapted Facebook to communicate between organizations and customers. in terms of promotion, advertisement and selling products and services of the organization.

Therefore, to use Facebook in the most efficient way, it is necessary to provide information about the needs of consumers on Facebook, how customers use Facebook, what kind of Facebook posts are there to respond to, and how to respond. if these data were analyzed we can offer services and promotions and corporate information that genuinely meet the needs of consumers.

**Problem**

Organizations want to advertise their campaigns to reach as many customers as possible. and want this campaign to be shared including comments about this campaign to reach the target audience and users as much as possible. The main channel for advertising this campaign is Facebook.

Therefore, if the organization wants a campaign to be successful by commenting, reacting, and sharing with others, you need to understand the user's behaviour and know how to get the most users to share this campaign. What are some factors that will help users share this campaign with friends, family, or other close people?

### **Dataset**

We found a suitable dataset online, which contains a dataset of 7,050 records that’s data cover the number of reactions, shares, and comments. Including the type of Facebook post and the type of reactions.

We are confident that this dataset can be used for analysis. And can get the results for organizations to advertise campaigns to reach customers effectively. Moreover, this data set can be analyzed for other perspectives in the future as well.

### **Hypotheses**

According to studying the usage of Facebook, we understand that if the organization wants the campaign to continue to be shared with a large number of users. This allows us to make assumptions for successful campaigns as follows:

1. The campaign should have a lot of comments from the participants. will affect the campaign will continue to be shared with others.
2. Campaigns where users have a lot of reactions. This will increase the opportunity of the campaign being shared as well.
3. Campaigns with video-type can attract more users than other campaigns type.

### **Data Preparation**

The dataset in this project comes from https://archive.ics.uci.edu. In this dataset, there are 15 features in total. All of which contain information related to user engagement. including the number of shares and comments as well. here are the details for each column.

|  |  |
| --- | --- |
| **Feature** | **Meaning** |
| status\_id | Status id of Facebook's post |
| status\_type | Type of post (e.g., video, photo, link, and status) |
| status\_published | Date and Time |
| num\_reactions | Total number of reactions |
| num\_comments | Total number of comments |
| num\_shares | Total number of shares |
| num\_likes | Total number of like reaction type |
| num\_loves | Total number of love reaction type |
| num\_wows | Total number of wow reaction type |
| num\_hahas | Total number of haha reaction type |
| num\_sads | Total number of sad reaction type |
| num\_angrys | Total number of angry reaction type |
| Column1 | Null |
| Column2 | Null |
| Column3 | Null |

Due to this dataset having three empty columns, Therefore, it needs to be dropped before proceeding to the next step. we will get the original dataframe as below image.

Application

Description automatically generated with medium confidence

After we get the original dataframe, we still need to transform some data. Below is a step of data transformation.

1. Convert data type from object to datetime for status\_published column.
2. Create a new column for the hour that can get data from status\_published column.
3. Create a dummy value for the status\_type and join it with the original dataframe
4. Drop unneeded column (status\_id, status\_type and status\_published)

# Now, we will get the summary statistics of this dataframe.

# Table Description automatically generated

### We decided to keep the data point whose reactions, comments and shares value is between 5% and 95%. we use univariate analysis to see the distribution of some interesting features (histogram, distplot, boxplot).

# Displot

Chart, line chart

Description automatically generated

* Histogram and Boxplot

## And we use Multivariate Analysis to find the relationship and correlation between two or more features.

* Correlation between num\_shares and other interesting features.

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

### The heatmap of the correlation

We can find the correlation between each feature from this heatmap. It shows that "like" is the most Facebook users react to the post. the number of shares (we are interesting) is correlated with the number of comments and number of love reactions.

And this heatmap shows that the user commented on a Facebook post. Users will share with their friends and comment on that post to participate with each other in the post as well. Also, if users respond with a love reaction will make users want to share and comment on that post even more. Therefore, the Facebook post should focus on the posts that users will respond to with a love reaction in order for users to share and comment on that post in the most efficient way.

However, the response of users to video posts is higher than images. This shows that most of the users like to comment and share more video-type posts.

### **Model Planning**

Regarding a past similar project, we had the experience with a linear regression model and we can get satisfactory results from that project. we think we can produce that model and technique to get the best result for this project as well.

So, we decided and plan to implement the machine learning pipeline with the linear regression model by applying a different combination of feature selection, feature transformation and feature scaling methods.

As mentioned above, we use Feature Selection using different techniques, Variance Threshold Selection and the K-Best method, both of which are widely accepted and effective. we also can get good and effective results.

After that, we will produce the Linear Regression model with the features that we provided and use a Linear Regression model along with a Polynomial Features transformation and a Linear Regression model and a feature scaling method (MinMax). Then we will also use the Linear Regression model again but will change the dataframe obtained from the Variance Threshold Selection and K-Best methods.

Therefore, we will get a total of 12 linear regression model scores, and we will compare the results and select the best feature section with the best result for the conclusion. we try to use the techniques that we have studied and experienced in the past projects. It is also an efficient and standardized way to achieve the most accurate results and clearest

### **Model Implementation**

We have learned more machine learning models and there are several models that can be used in the project. therefore, originally intended to implement only the linear regression model. we decided to implement several models and apply them with a different combination of feature selection, feature transformation and feature scaling methods. To find the model that yields the highest and most suitable score for our dataset to predict the results according to the objectives of the organization. All models that we choose to implement are as follows:

* LinearRegression
* SGDRegressor
* RandomForestRegressor
* DecisionTreeRegressor
* GradientBoostingRegressor
* AdaBoostRegressor
* XGBRegressor
* CatBoostRegressor
* MLPClassifier

We also use a variety of feature selection (SelectFromModel, VarianceThreshold, SelectKBest), feature transformation (PolynomialFeatures) and feature scaling (RobustScaler, StandardScaler, and MinMaxScaler) methods to work with the models listed above and helped to find out the best model that can provide the best score for this project.

In this project, we implement Pipeline for convenience and accuracy and to prevent confusion because we use a variety of models and A variety of feature selection, feature transformation and feature scaling methods as well, pipelines can help manage and process results from different models. then we can compare all the scores from those models to find the best score for this dataset and use it to predict the most effective results.

It is based on the hypothesis that we need to know how much a new campaign has been shared in order to estimate the campaign's results. By the relationship between the number of shares and other information as previously project description.

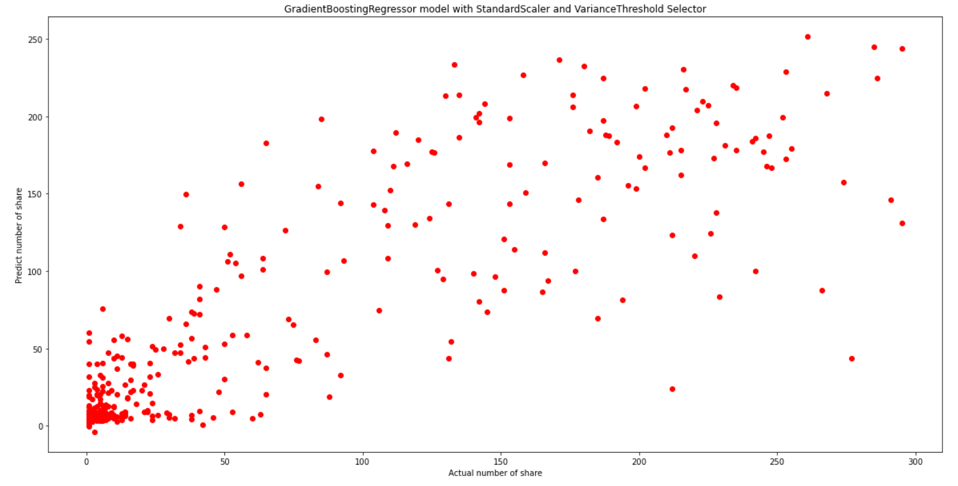
This requires us to use a variety of machine learning techniques to test this hypothesis. We started implementing each technique with the same goal from the same dataset. In addition, each model is applied to different techniques such as feature selection, feature transformation and feature scaling methods, and this allows us to know the different results of each model. We try to make this project covers as many testing model techniques as possible for the best results from each machine learning model.

The results of the machine learning models and techniques are numerous. So, we filtered the scores of all the different models to find the best results. After that, we can get the best model and techniques for this project to use for predicting the project objectives.

**Results Interpretation and Implications**

The results of the selected model can be plotted on a graph for the correlation between the actual data and the predicted data (shown below). The trend of the chart is a position relation. which is quite accurate and consistent with the hypothesis and the accuracy score was 78%, Although, we still can see some errors as you can see from the graph. However, this model is effective and suitable to be used for project objective predictions.

Because we are implementing several models, we have achieved 162 results from those models and techniques. Based on predictive accuracy. the Gradient Boosting Regressor model is the best model. Therefore, we select the Gradient Boosting Regressor with 'StandardScaler' and the 'VarianceThreshold' selecter for out-of-sample predictions.



Below are the top 20 results.  
  
Graphical user interface, application, table, Excel

Description automatically generated

**Does the model appear valid and accurate on the test data?**

According to the test results, the model has an accuracy of 78%, which is considered high accuracy. And when compared to the test data, it is quite accurate and close to the actual data.

**Does the model output/behavior make sense to the domain experts?**

The selected model shows whether the model is suitable for a domain or dataset.

**Do the parameter values make sense in the context of the domain?**

The parameters are considered comprehensive and appropriate for this domain. because there are various parameters that can be used effectively There are related parameters that affect others and the results. Therefore, the parameters used are considered appropriate and efficient.

**Is the model sufficiently accurate to meet the goal?**

According to the results of the model, it was up to 78% accuracy, which is the high accuracy. and after using the model to predict the results, the results are very satisfactory.

**Are more data or inputs needed?**

The dataset is complete for the purpose of the project, which is to achieve the project objective completely. So, no additional input is required.

**Is a different form of the model required to address the problem?**

This project uses a pipeline with a variety of models and methods to produce. and we selected the model with the best results. however, the selected model is the best and most suitable model and can be used for the purpose of this project. Therefore, there is not necessary for other models to solve the problem.

**Communicate and document the key findings and major insights derived from the analysis**

Machine learning models help the process and planning more efficiently. For example, this project can use a machine learning model to help predict the outcome of a new company's campaign, which allows the company to predict the results in advance. In addition, knowing the various correlations that affect the desired outcomes which can help to plan and optimization of the campaign to achieve the campaign results and the goals of this campaign.

**Out-of-sample Predictions**

We have created 3 sample data. to be used for predicting results with the selected model. The sample data has 3 different values:

1. Sample post with a normal number of reactions and comments

2. Sample post with a high number of reactions and comments.

3. Sample post with extreme high number of reactions and comments.

We can predict and get different results from all 3 samples of data. which makes the company can know the prediction of result of this campaign.

In addition, the sample data was also adjusted to test the results of the relationship between the number of love reactions and the number of shares. which is a greater number of reactions can get more shares as well. So, we adjusted the sample data to use the same total number of reactions, but with fewer likes reactions and more love reactions. Which made it more satisfying that the model was able to predict that the number of shares would increase significantly if the company can get more love reactions from the customers.

This is a choice for the campaign content that the campaign should be presented in the form of love, warmth, and softness, to increase the customer's feelings to make it easy to understand what the campaign wants to convey. Therefore, the number of shares will increase as the prediction results. Below is the 2-sample data and prediction result for the original sample data and adjusted data.

* Sample 1, The total number of reactions is 168 (166 likes, 3 loves) and 127 comments. Facebook post's share prediction result is 10 shares.

Text

Description automatically generated

After increasing the number of love reactions to 30 and decreasing the number of like reactions to 136 with the same other condition. *the Facebook post's share prediction result is* ***66*** *shares instead of* ***10*** *shares.*

Text

Description automatically generated

* Sample 2, The total number of reactions is 240 (220 likes, 18 loves) and 197 comments. Facebook post's share prediction result is 50 shares.Text

  Description automatically generated

After increasing the number of love reactions to 60 and decreasing the number of like reactions to 178 with the same other condition. *the Facebook post's share prediction result is* ***133*** *shares instead of* ***50*** *shares*.Text

Description automatically generated

* Sample 3, The total number of reactions is 1300 (1150 likes, 140 loves) and 1290 comments. Facebook post's share prediction result is 211 shares.Text

  Description automatically generated

After increasing the number of love reactions to 300 and decreasing the number of like reactions to 990 with the same other condition. *the Facebook post's share prediction result is* ***216*** *shares instead of* ***211*** *shares*.Text

Description automatically generated

**Concluding Remarks**

Machine learning models can be used to effectively predict the result of campaigns based on company objectives and purpose. It helps the company to plan and execute campaigns that meet the target audience efficiently. The most important thing is to check the relationship between the dataset, and parameters Including managing the data process before using that dataset with the model. In terms of model selection, It is necessary to find a model suitable for the purpose and type of the dataset and need to compare models to find the most efficient model for the company's objectives and dataset.

However, the machine learning model is simply a tool used to analyze, process, and predict the result of a company's objectives. The different models have different advantages and disadvantages. Therefore, we need to understand the objectives, dataset, and limitations of the project. to select the best and most efficient model for our project.