

Facebook Metrics

By: Aurelio Barrios

Introduction

In the modern day data is an extremely powerful tool that is being used to the advantage of successful businesses. Social media paired with data driven decisions can be crucial to the success of a modern business. This is why evaluating and predicting a successful Facebook post can be an incredible insight into social media accounts. How can accounts maximize the impact of their social media posts in order to achieve their goals? What can metrics like engagement, reach and impressions tell us about the success of a post? This project aims to create a machine learning model that analyzes the metrics of a post in order to get a better understanding of what metrics have a greater impact on the interactions of a post.

Data Wrangling

It is important to note that the data for this project was gathered from the UC Irvine machine learning repository. This means that our dataset was fairly clean to begin with, but there were still some extra steps needed to prepare the data for the purpose of this project.

After some initial peeks at the data it was noticed that there were still some missing values in the dataset. The rows with the missing values were dropped since there was not a significant benefit to keeping them and dropping these rows wouldn't affect the dataset too much since it was only 4 rows. It was also necessary to make sure that all our values made sense in the context of their column. For example, columns like "Post Month", "Post Weekday" and "Post Hour" all had a specific range in the context of the data that they were trying to keep. For example it would not make sense for there to be the value "20" in the "Post Month" column since there is not a 20th month in our calendar. This same logic was applied to the other

columns. Lastly this step was finished by tidying up the dataset in a way that made it easier for it to be used in the future steps like improving the captions for our columns.

Exploratory Data Analysis

In our Exploratory Data Analysis stage we found information about our data that can be useful for the next step of modeling. Through this analysis it was found that there were multiple outliers in our dataset, but two of these were significantly larger than the other outliers. It was here that the question had to be asked about how to handle the outliers in the dataset. After careful exploration of these outliers it was found that it would be best to keep the outliers in the dataset since these instances were not clearly as a result of error. It was most likely that these outliers were unique events that could be useful in our analysis.

Modeling

In the modeling phase we used our data to create different models with different approaches. The initial approach was a baseline model using all available features. This was used as a baseline for our future models to beat.

The first additional model that was compared to the baseline model was a linear regression model using principal component analysis. The idea being that since our features were all heavily correlated to each other, with the principal component analysis model we could remove some of the multicollinearity and get a better understanding of our features. This model proved to be a failure when compared to the baseline model.

The second additional model that was compared to the baseline model was a Lasso regression model. This model proved

to be better than our baseline model and was also useful in performing a form of feature selection. This was due to the fact that the Lasso regression model uses a regularization technique that adds a penalty term which forces some coefficients to be exactly zero, essentially performing feature selection. This was a great model that performed but also gave us a better insight into what features were key in our analysis.

Conclusions & Future Endeavors

Through our analysis and modeling it was found that features like "lifetime_engaged_users" and "lifetime_consumers" were two of the features that were most significant to our target variable. This can provide some insight as to what social media accounts should focus on when maximizing the impact of their posts. Although we gained a better insight as to what Facebook metrics are essential to total interactions on a posts, there is still some additional work to be done. Firstly, there can be more analysis into what features drive total interactions. This project scratches the surface in this aspect. This project can also go deeper in the machine learning aspect and optimize a machine learning model that converges on all aspects of data like the training, testing and validation sets. It could prove useful to build a model that has more precision on our metric of Mean Absolute Error.