Facebook User
Engagement Data Analysis

Craig Lynch
Goalcast Take Home Test

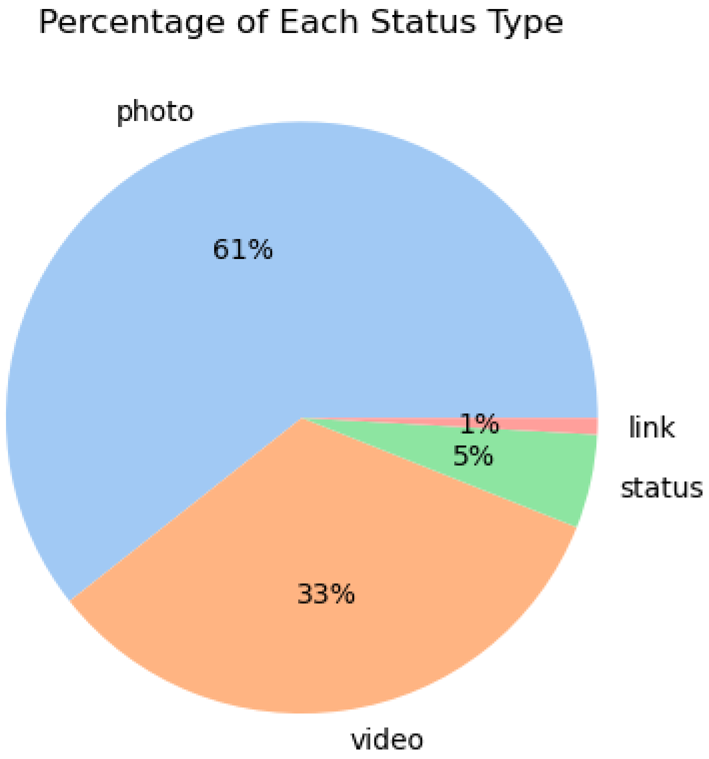
November 13, 2021

Introduction

As part of the Data Analyst application at Goalcast, Goalcast provided a csv file containing Facebook pages posts of different retail sellers and asked for an analysis of the data with visualizations and to develop a predictive model that will help forecast engagement metrics of future posts. The posts were photos, videos, links, or statuses and engagement metrics such as comments, shares, and reactions were collected. “Reactions” comprises a number of possible different responses, such as “like”, “love”, ”wow”, ”haha”, ”sad”, and “angry”.

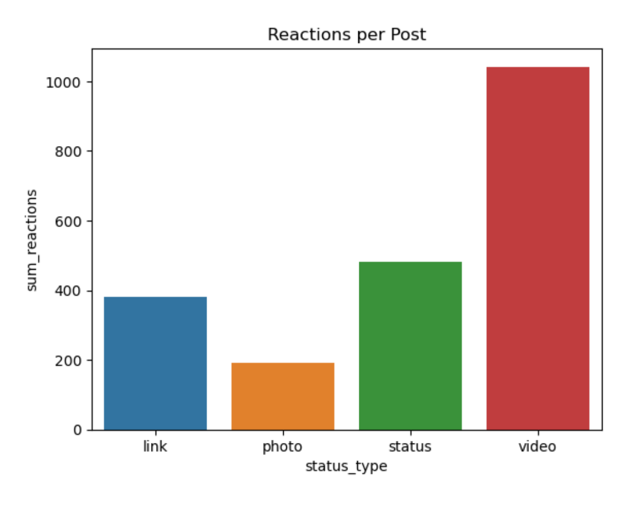
Exploratory Data Analysis

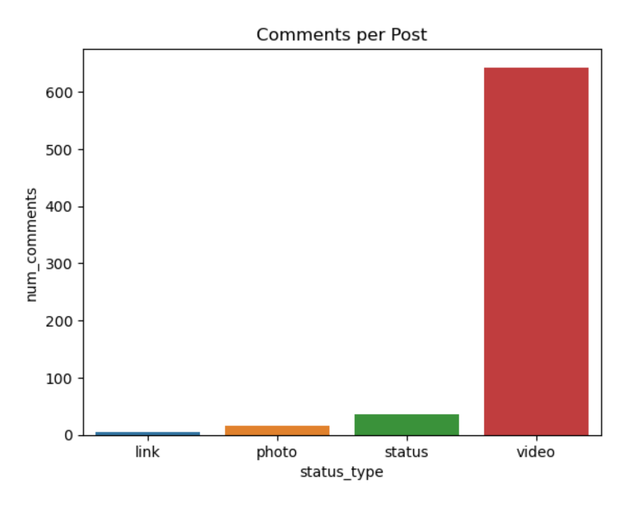
The original dataset contained 7050 observations containing facebook pages posts of different retail sellers and the posts were either photos, videos, links, or statuses. The data ranges from July, 2012 to June, 2018. Photos have been around since the beginning, however Videos first launched in September, 2013, Links launched in March, 2014, and Statuses launched in January, 2013.

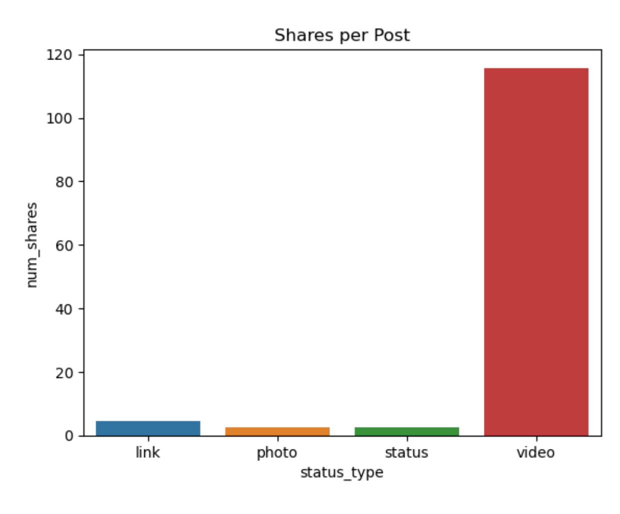


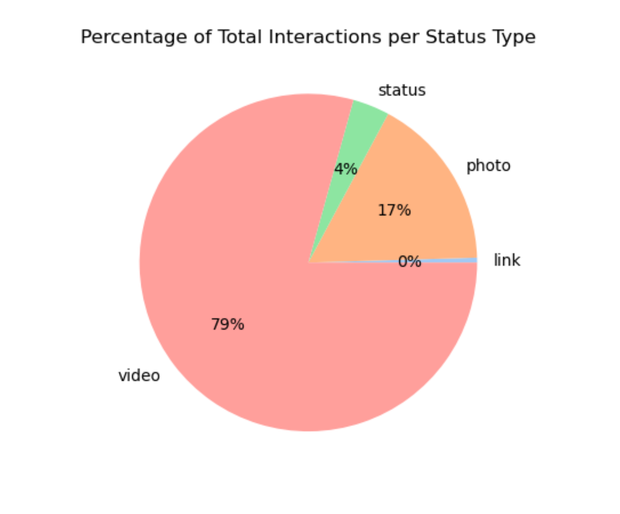
As shown above, photos make up the majority of the posts.

Because photos have been used throughout the entire dataset, and other post types introduced later, I wanted to find out the average user engagement per post. As shown below, videos generate the highest amount of reactions, comments, and shares per post, consisting of 79% of total interactions.

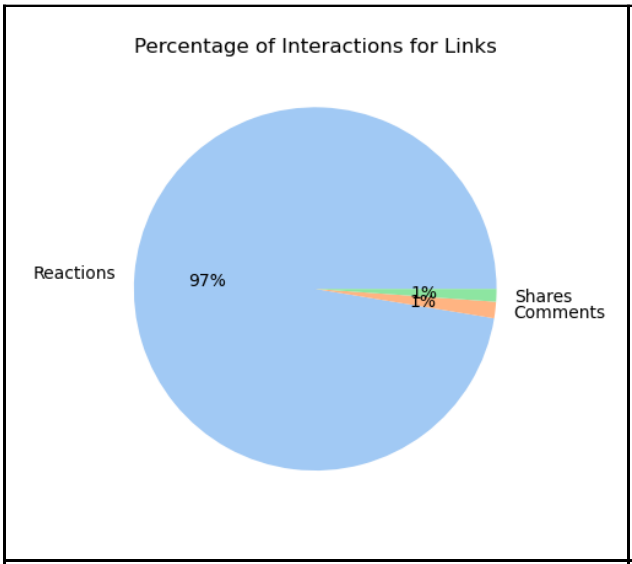


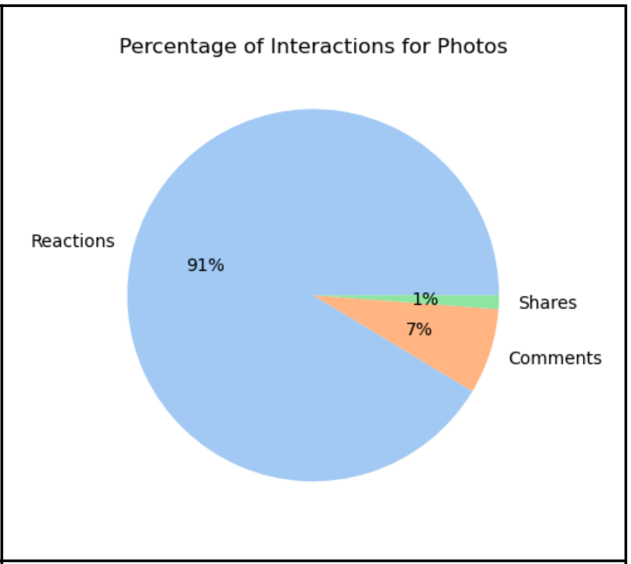


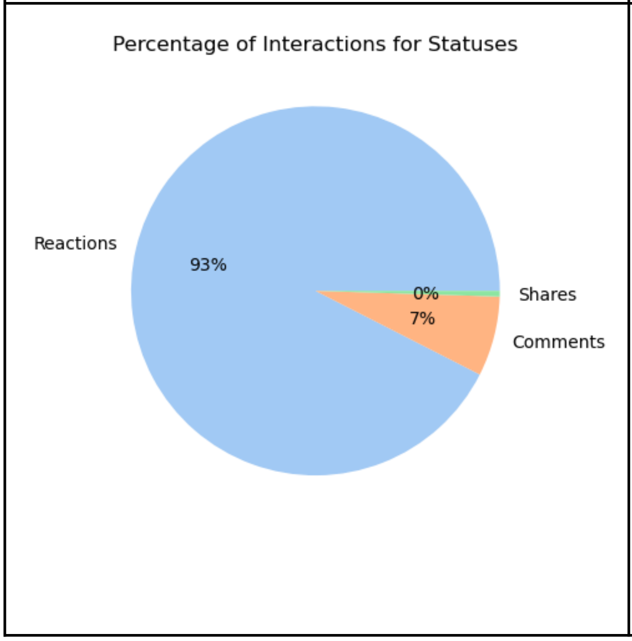


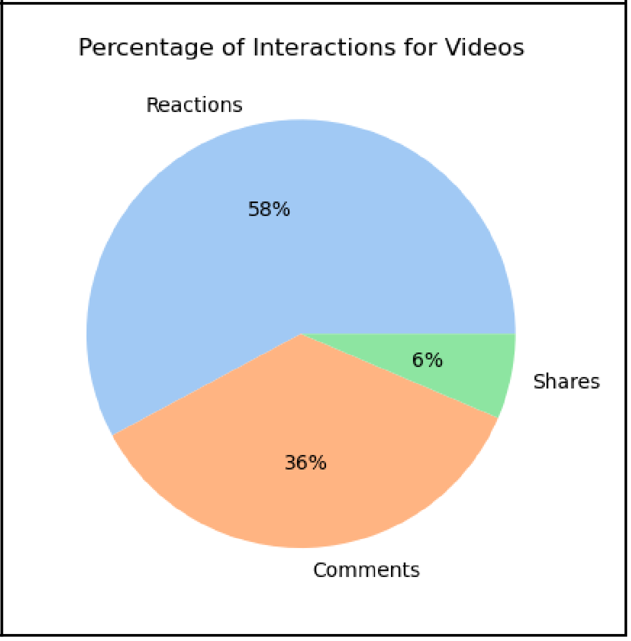


Looking at each post type individually, we see that reactions make up the majority of the user engagement. However, a large portion of video engagement is comments.

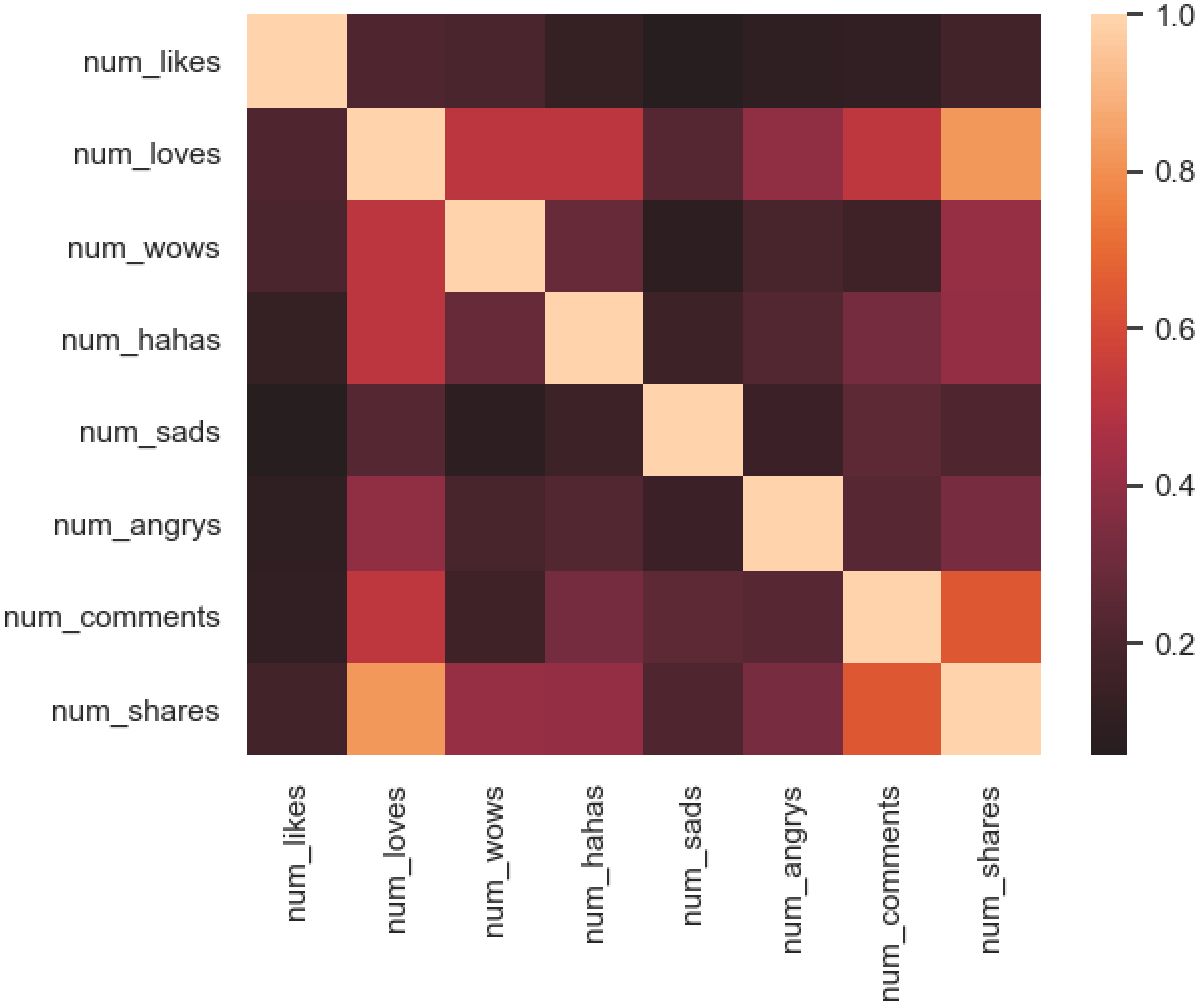






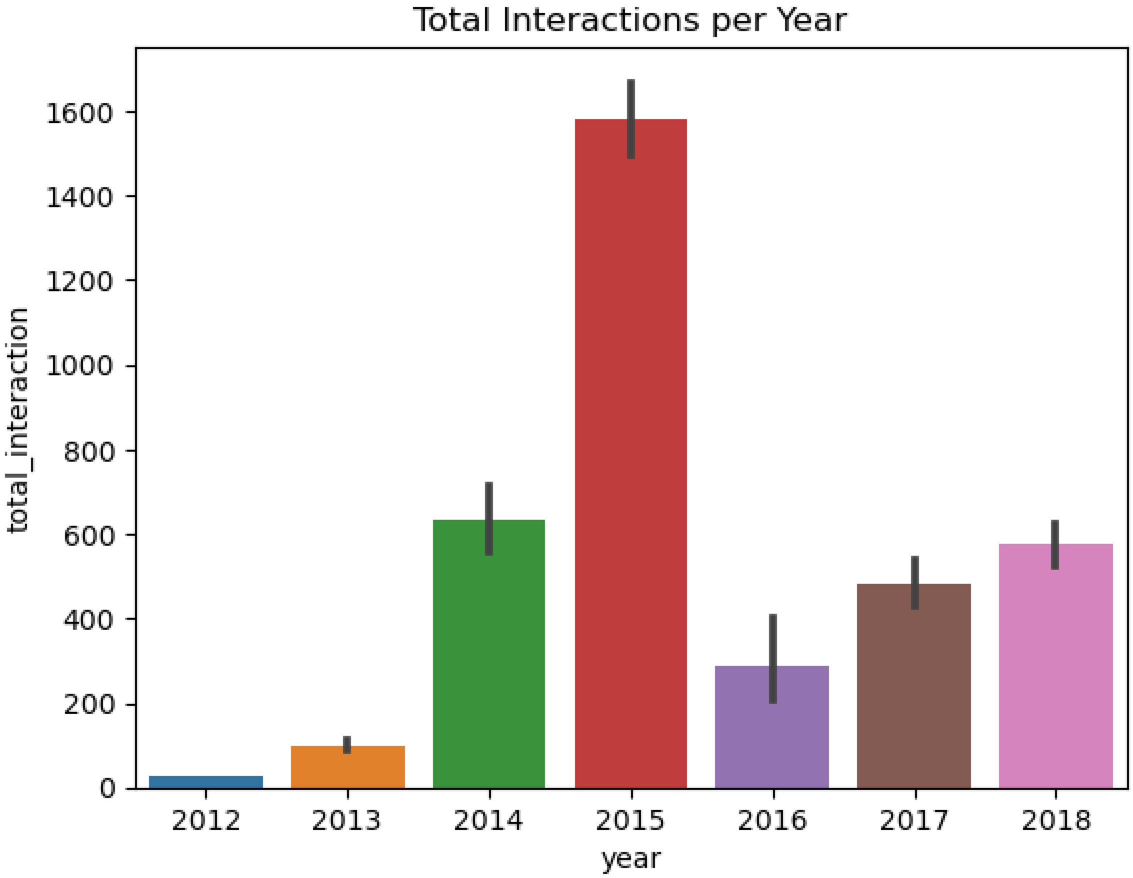


There is correlation between the number of comments on a post and the amount of shares. Also, users are more likely to comment or share if they “love” the status type.

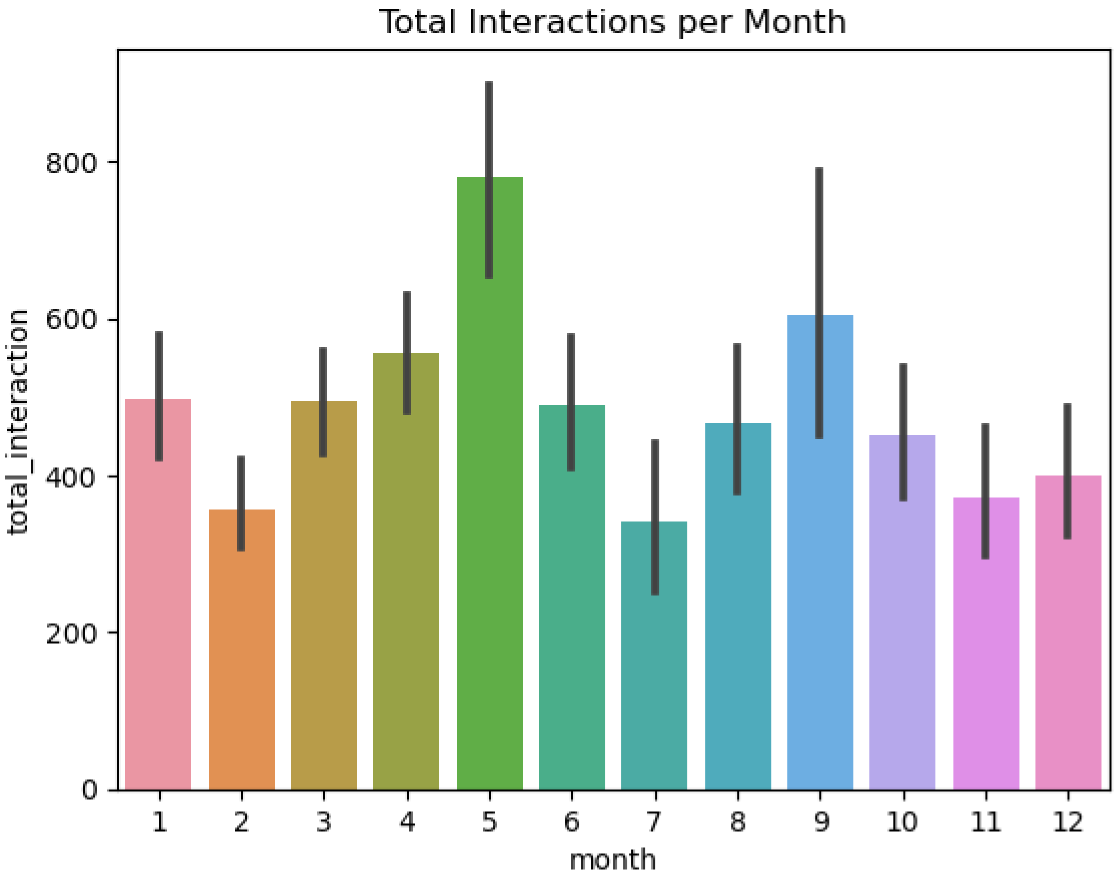


Looking into the timeline of the posts, there is a large drop in user interaction in 2016. This is most likely due to Facebook launching "reactions" early that year. This introduction may have affected how their algorithms calculate interactions.

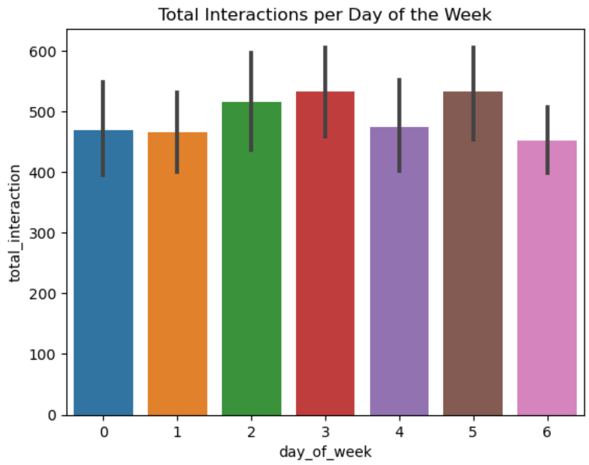
Please note that the dataset begins in July, 2012 and ends in June, 2018, so the data for these years are incomplete.

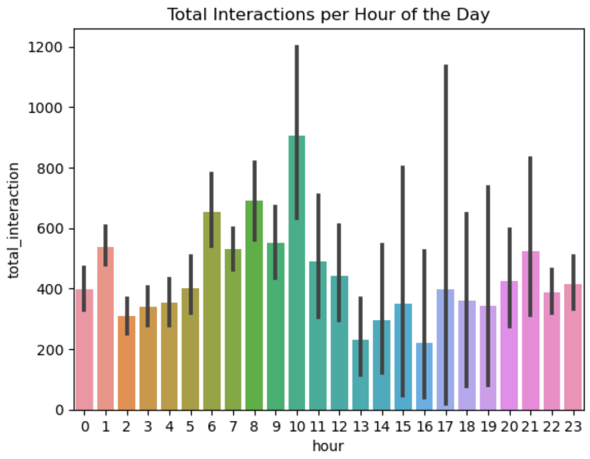


Looking into monthly data, Spring and Fall appear to have more user interactions. This is possibly because buyers are getting ready for the upcoming seasons.



While there does not appear to be higher engagement on any particular day of the week, higher activity can be seen during the morning hours between 7am - 11am.

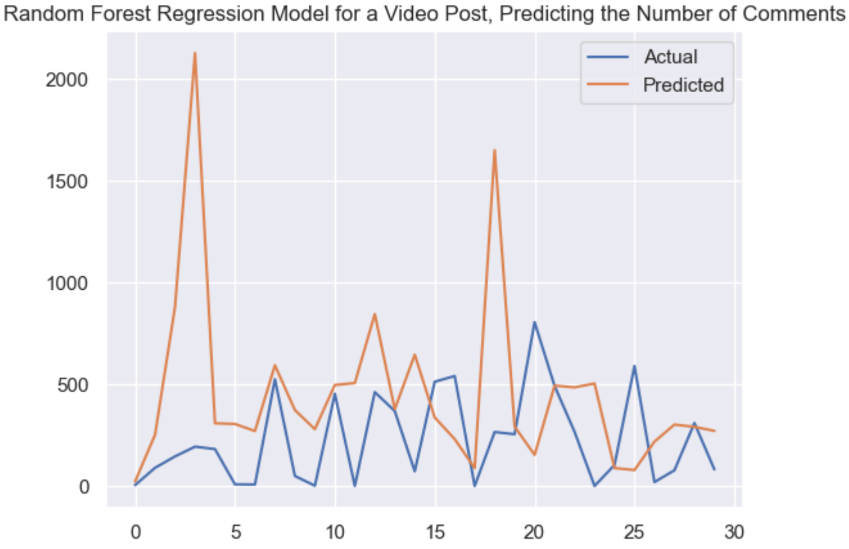


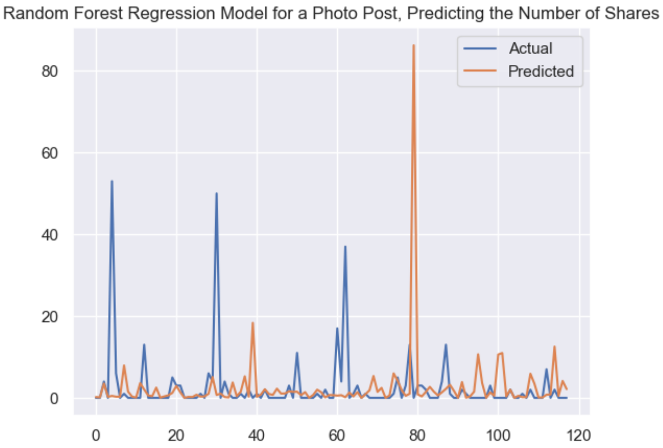


Predictive Modeling

As mentioned above, there is a large drop in user interaction in 2016 most likely due to Facebook launching "reactions". As a result, I have used only data collected in 2016 and onwards for the predictive modeling.

Additionally, outliers for the sum of reactions, number of comments, and number of shares have been removed to improve model performance. I have classified outliers and data points that are greater than three times the standard deviation of each.


Random Forest Regression Model for a Video Post, Predicting the Number of Comments


Random Forest Regression Model for a Photo Post, Predicting the Number of Shares

I chose to proceed with the RandomForest model, however it could be improved by further outlier removal and tuning hyperparameters. Also,there is a large data imbalance; "status" and "link" do not have enough data to provide accurate predictions.

Uses

With further improvements to the model, sellers can optimize when and how they advertise their products. Knowing that videos generate the highest number of comments and shares, and get the most overall engagement per post, sellers can focus on making high quality videos to increase their user engagement and sales.

Additionally, knowing that higher levels of engagement occur during the hours of 7am - 11am, and during Spring and Fall months, sellers can optimize inventory and supply chain planning.

Moving Forward

In order to improve my model, more data would be needed, especially for statuses and links. Also, having access to users' comments would allow for Natural Language Processing and can determine specific words that generate higher user engagement. Having individual sellers’ follower count would allow the calculation of average engagement and applause (positive reaction) rate and would allow the calculation of the amplification rate (number of shares/number of followers) and the virality rate (number of unique views/number of shares).