Proposal for Lab 2

Arisa Nguyen, Ayman Bari, Emanuel Mejía, Jorge Bonilla

Introduction

Businesses and organizations often create Facebook pages as a way to build community, advertise, and keep their followers in the loop. Posts made by these organizations are the primary way of communicating to their audience. Group 3's Lab 2 will investigate how specific attributes of the post, Facebook page, and organization affect engagement on each post. The data set to be analyzed comes from UC Irvine's Machine Learning Repository, and has scraped information from 2,770 pages, 57,000 posts, and 4,120,532 comments. Since the data set contains 5 variants (each variant of the data set scraped Facebook at different times), the variant containing data for the most posts will be used for this analysis.

Research Question

For this project we'll refer to **engagement** as the number of comments made on a post in a 24 hour period. Our research question is as follows: *how is user engagement on the Facebook platform affected by post length while controlling for other variables such as post date, post share counts, page category, and page popularity?*

Data Set

The data set to be analyzed is the "Facebook Comment Volume Dataset Data Set". Our variables of interest are:

- CC2: Number of comments in the last 24 hours before base time
- Base time: Time data was scraped
- Post share count: Number of shares of the post
- Post length: Character count of the post
- Page popularity: Number of likes of the page
- Page talking about: Number of people interacting with the page

Plan of Action

We first conceptualize a causal theory for the number of comments a post will receive during the last 24 hours relative to when the post was initially published. For our research project we consider the number of comments that a post receives as our metric for product success. Our causal theory states that the length of a Facebook post affects the number of comments within the next 24 hours while controlling for other variables such as post share count, page popularity and the base time at which the post was published relative to when the data was collected. Considering the variables that we believe to impact engagement as represented by the number of comments that a post receives, we form the following structural model:

To evaluate our hypothesis we established the following regression model:

$$\text{CC2} = \beta_0 + \beta_1 \text{PostPopularity} + \beta_2 \text{PostShares} + \beta_3 \text{PostLength} + \beta_4 \text{PageTalkingAbout} + \beta_5 \text{BaseTime} + \epsilon$$

We will start with an exploration of the data, checking the joint distributions of each causal variable with our outcome variable. Following an initial assessment to check our assumptions, we will evaluate possible transformations for our variables that may better model the impact on engagement. The source of the data set stems from a program written in Java and Facebook Query Language to scrape the pages, posts, and comments[1]. As the dataset includes five variants, each collected at a different time, it is unclear as to whether

[1] Singh, Kamaljot. "Comment Volume Prediction using Neural Networks and Decision Trees." International Conference on Modelling and Simulation, 2015, https://ijssst.info/Vol-16/No-5/paper16.pdf

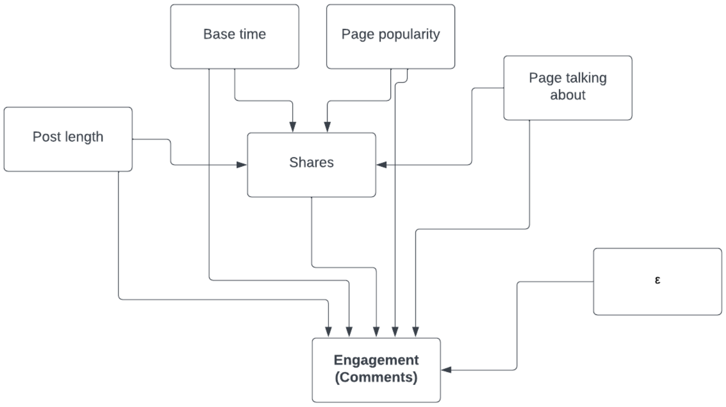

Figure 1: Structural Model for Product Engagement.

the same post may have been captured in multiple variants. We will therefore assess our model across each variant independently, and look for consistency in the results. We expect to find consistency in the strength of the causal relationship between the explanatory factors we model and engagement across each variant. As part of the exploratory data analysis, we will produce the joint distribution graphs between number of comments against the number of page shares, popularity, and “talked about” instances. Moreover, we will ensure that the assumptions of the Classical Linear Model are met in order to conduct an Ordinary Least Squares regression. We plan to include in our model a number of control features, which we do not believe to have a direct causal effect on our outcome variable, but may be associated with higher or lower levels of engagement. During the modeling stage of the project, ANOVA F-Tests will be conducted to assess the efficacy of adding additional variables to the model to determine how successful the Facebook platform is in terms of number of comments which is the engagement metric we wish to analyze. We will do this process to develop at least three models:

1. The base model where we regress the number of comments on the day on which the post was published
2. A model where we regress the number of comments on all of the desired covariates
3. A third model where we add interaction terms for our covariates