Project Proposal

The BEST Fit - Olivia Encarcion, Leo Yang, Philip Lin, Allison Yang

```r
library(tidyverse)
library(tidymodels)
library(ggplot2)
library(dplyr)

newspopularity <- read_csv("data/OnlineNewsPopularity.csv")
```

Introduction

Introduction To Subject Matter: We are investigating the online news popularity, and the different features of articles that affect that popularity. In this case, the popularity is measured by the number of shares that the online article got. Due to the rise of the internet in recent years, many people around the world are interested in knowing what features of an article makes it trendy. Some of these key features of an article include keywords, digital media content, and earlier popularity of the news mentioned. This research topic can benefit researchers, content providers, advertisers, large companies, and politicians because online news has become so impactful, influencing the knowledge that the general public receives, and knowing what makes online news popular can help people know what types of articles reach the public.

Tsai, M.-J., & Wu, Y.-Q. (2022). Predicting online news popularity based on machine learning. Computers and Electrical Engineering, 102, 108198. https://doi.org/10.1016/j.compeleceng.2022.108198

Research Question: How do different article attributes (ex. Polarity, Positive/Negative Sentiments, Number of Images, etc.) relate to its virality on social media?

Motivation/ Importance: In this project, we hope to analyze what factors dictate the "virality" of news headlines on social media, to better understand how it might influence the spread of news. Gaining a more clear understanding of how news spreads online might help inform policies to avoid undue suppression or accidental spread of misinformation.

Hypothesis: We hypothesize that more polarized, negative, and concise articles have a higher rate of virality on social media.

…

Data description

Source: The University of California Irvine Machine Learning Repository released this data set in 2015, known as "Online News Popularity". It includes features about articles published by Mashable in two years (from 2013 to 2015). Mashable Inc. is a digital media website founded in 2005 and as of November 2015, it has over 6,000,000 Twitter followers and over 3,200,000 fans on Facebook.

Data Collection: Alongside predictors such as number of words in title, number of images, etc., the data set also tracks the number of shares in social networks to track how popular any given article is. Estimated relative performance values (such as polarity of positive/negative words, titles, etc.) were tracked by the authors using a Random Forest classifier and a rolling windows as assessment method.

General Characteristics: The data set includes a total of 39644 observations, each of which is an individual article. Each observation includes characteristics such as Number of Words in Title/Content, Rate of Unique Words, Number of Images, Data Channel, Day Published, Rate of Positive/Negative Words, Polarity, etc. Generally, the intention of the data set is to predict the number of shares/popularity of an article based on different factors.

…

Data Processing

To prepare the OnlineNewsPopularity dataset for analysis, we need to perform several processing steps. The first step is go through the dataset and handle and missing data. This includes checking for missing values and remove them from the dataset. If a feature has too many missing values, we may decide to drop it completely. We also will chose to merge different data channels into one categorical variable. The dataset includes multiple variables representing different data channels. Instead of having separate columns for each channel, we can create a single categorical variable with different levels. Additionally, since the dataset has multiple variables for whether the article was uploaded on a weekday or weekend we are going to merge these variables in a meaningful way. This transformation simplifies modeling and visualization. Lastly, we can convert highly skewed variabls using log transformations to improve model performances.

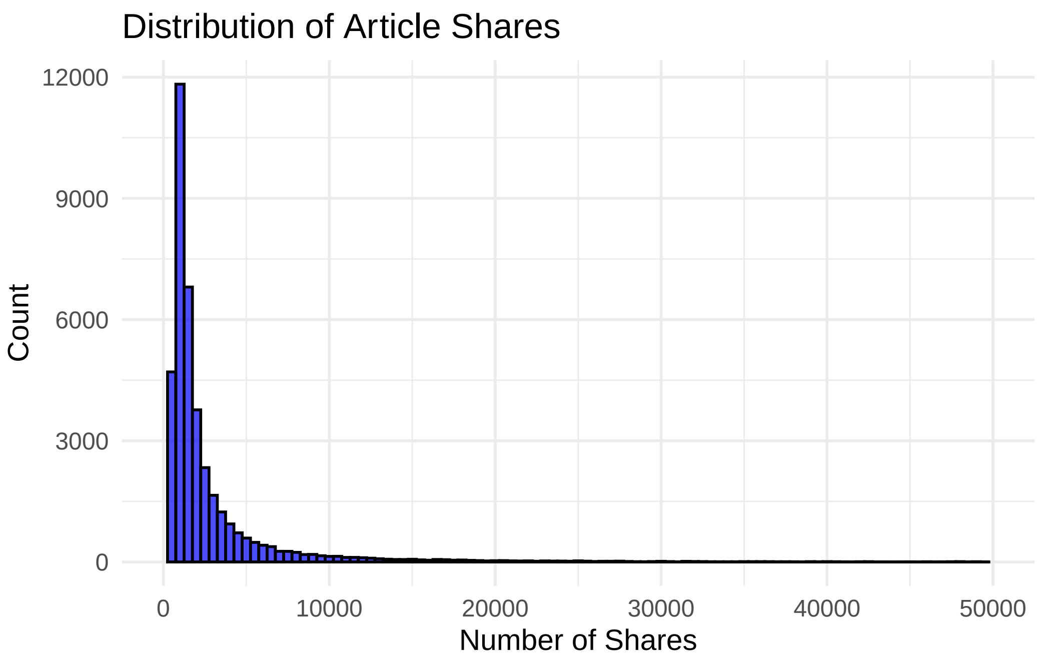
```r
summary(newspopularity$shares)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      1     946    1400    3395    2800  843300 
```

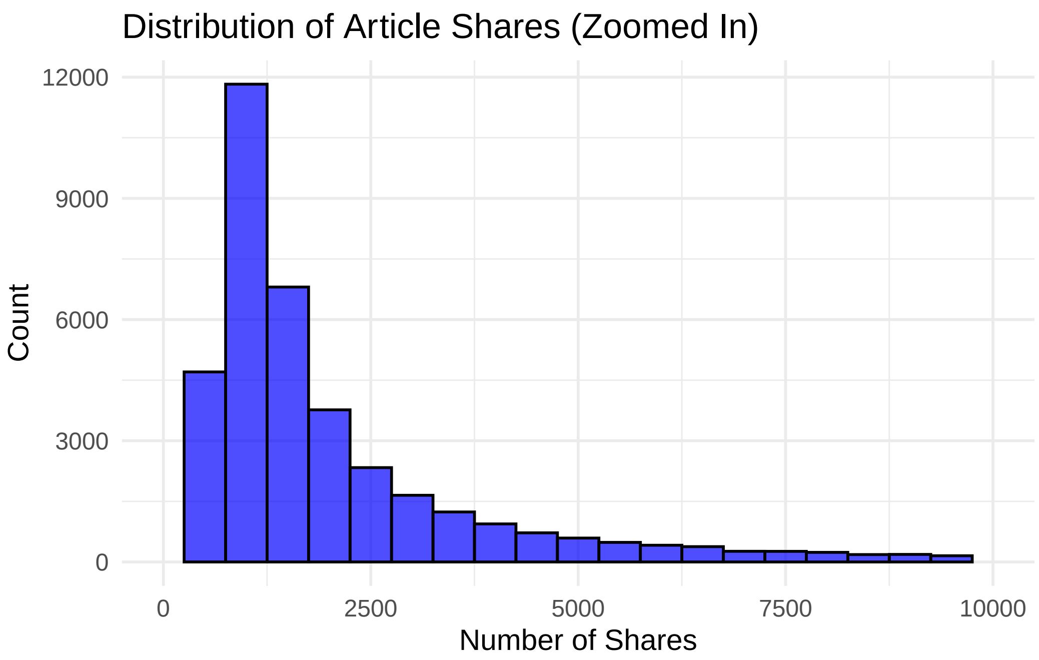
```
newspopularity |>
  summarize(
    mean_shares = mean(shares),
    median_shares = median(shares),
    sd_shares = sd(shares),
    min_shares = min(shares),
    max_shares = max(shares),
    q1 = quantile(shares, 0.25),
    q3 = quantile(shares, 0.75)
  )
```

```
# A tibble: 1 x 7
  mean_shares median_shares sd_shares min_shares max_shares    q1    q3
        <dbl>         <dbl>     <dbl>      <dbl>      <dbl> <dbl> <dbl>
1       3395.          1400    11627.          1     843300   946  2800
```

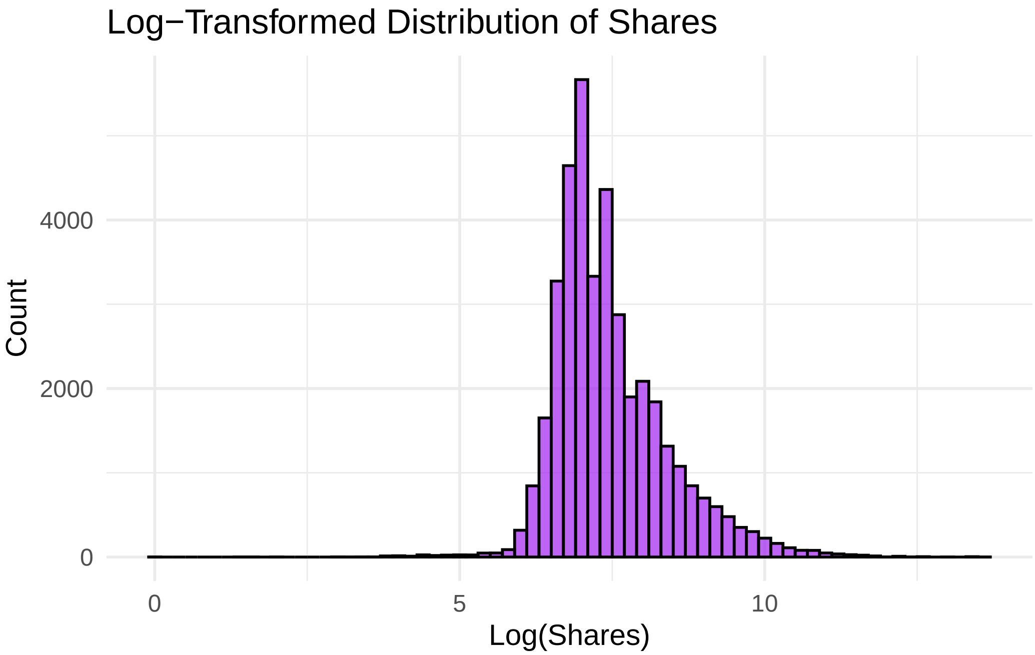
```
ggplot(newspopularity, aes(x = shares)) +
  geom_histogram(binwidth = 500, fill = "blue", color = "black", alpha = 0.7) +
  scale_x_continuous(limits = c(0, 50000)) +  # Limit for better visualization
  labs(title = "Distribution of Article Shares",
       x = "Number of Shares",
       y = "Count") +
  theme_minimal()
```



```r
ggplot(newspopularity, aes(x = shares)) +
  geom_histogram(binwidth = 500, fill = "blue", color = "black", alpha = 0.7) +
  scale_x_continuous(limits = c(0, 10000)) +  # Zoom in on the first 10,000 shares
  labs(title = "Distribution of Article Shares (Zoomed In)",
       x = "Number of Shares",
       y = "Count") +
  theme_minimal()
```



```
ggplot(newspopularity, aes(x = log(shares))) +
  geom_histogram(binwidth = 0.2, fill = "purple", color = "black", alpha = 0.7) +
  labs(title = "Log-Transformed Distribution of Shares",
       x = "Log(Shares)",
       y = "Count") +
  theme_minimal()
```



...

Analysis approach

This dataset features 58 predictive attributes, and one response variable– the number of shares. Some predictors of particular interest to us are:

rate__positive__words - The rate of positive words among non-neutral tokens, which captures how emotionally charged the language is.

Rate__negative__words is likewise, but negative words

Title__sentiment__polarity - A measure of how polarizing the title is

N__tokens__content - A measure of how long the article's content is

N__tokens__title - A measure of how long the article title is

Data__channel__is__entertainment, data__channel__is__bus, data__channel__is__socmed, data__channel__is__tech, and data__channel__is__world – Indicator variables for the article belonging to different categories or "data Channels", we can merge into one categorical variable.

Weekday__is__monday, weekday__is__tuesday, weekday__is__wednesday, weekday__is__thursday, weekday__is__friday,weekday__is__saturday, weekday__is__sunday

– These are indicator variables for what day the article was published, which we can merge into a categorical variable of Day_published

In terms of modeling techniques, a multiple linear regression would make the most sense, since we have a quantitative "shares" response var.

…

Data dictionary

The data dictionary can be found here.