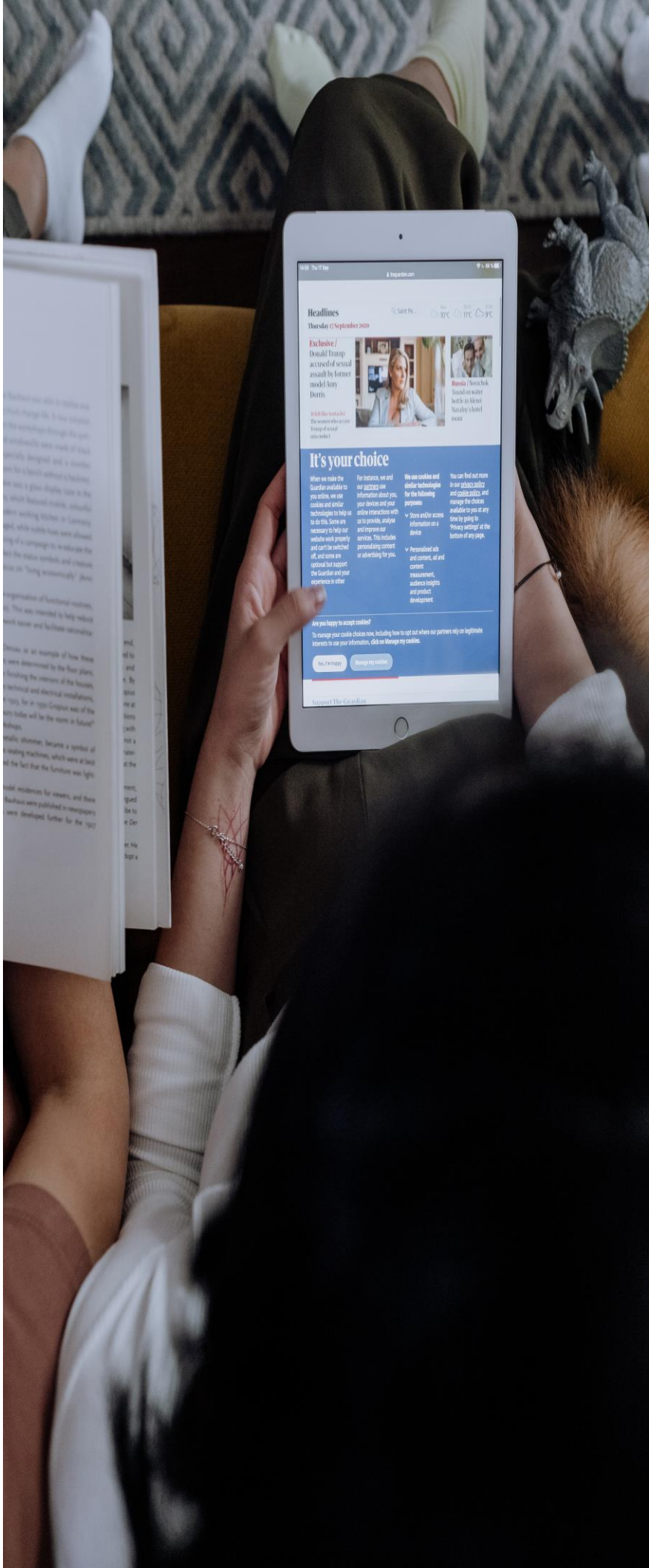
STATISTICAL TECHNIQUES

Online News Popularity

PREPARED BY:

Bobbi McDermott



CONTENTS

INTRODUCTION

This is an examination of an Online News Popularity Dataset. News consumption has transformed dramatically from paper format to online. This has led to changes in how the news is delivered. However the growing interconnection and complexity of the news ecosystem, as well as worries about the overreporting of news consumption (Atske, 2020) has led to questions on how to measure it and does reporting affect the content. With social media, being the main source of news online, traffic diverted over shares lends weight to how online resources analyze their data to increase this traffic their way (Atske, 2020). This obviously contributes to the content delivery overall.

THE DATASET

This dataset has 61 features as per the data dictionary. It is information that will provide valuable insight into the parameters of 'shares', therefore that is the target variable. The idea is to predict the number of shares in social networks (popularity).

There are two ways to look at this dataset, we could do a **regression model** by looking at the values of the target variable or we could do a classification model by assessing the values to see whether something is popular or not.

As this is a statistics assignment and the remit involves binomial distribution and the calculations of two probabilities, I have decided to approach this with a **classification method**

SECTION 1: EXPLORATORY DATA ANALYSIS

I employed EDA to understand the data, find patterns and relationships within it to develop theories. I need to establish any flaws to be resolved and organize the data for visualization and summary statistics.

Part 1: Cleaning and Observation

After uploading the csv I observed through df.info() that there were delimiters that would need to be cleaned. From there, I identified and dealt with whitespace but there were no null or duplicate values. There were however zero values, that in context with the data dictionary would need to be removed. The 'n_tokens_content' was the feature I observed that this value affected as it indicated no content. These were removed losing an acceptable 2.98% of the dataset (Figure 1). The URL was also dropped as it was the only categorical feature and provided no information that was beneficial to the remit of the assignment.

Understanding the data dictionary is a crucial as it provides context, what type of attribute you're dealing with and having one negates the need to expanded research (Nicholas, 2022). However there were features that I didn't understand that I needed to research further, namely LDA which is used to classify text in a document to a particular topic (Li, 2018), and measures how closely relation they are.

I then visualized the distribution of the target variable. There were duplicates which led me to measure the unique values in context, which resulted in a 3.73% (Figure 2), (Figure 3). This low cardinality tells me that transforming this variable to binary will have very little affect. (usɛr11852, 2019)

Outliers

I used two methods to identify the outliers. **1) Box Plots** and **2) Interquartile Range.** IQR was used essentially removing the central tendency that boarders the median displaying the upper and lower ranges by adding or subtracting 1.5 IQR (Bhandari, 2021) **(**Figure 4). There are significant outliers among numerous features but non that appear contextually incorrect as per the data dictionary. It is observed that if we do have this distribution, there may be something interesting in the data that warrant further investigation (Kumar, 2020)

Part 2: Correlation

I created a correlation matrix to get a visual as strength of the relationships between the variables, (Figure 5)

As there are numerous features, it is difficult to pinpoint so I set the 'threshold' to get a list and I can compare them to the visualization. There features most strongly correlated can be removed to help the learning process and produce results comparable to the whole model (Lannge and Fowler, 2021) Removing strongly correlate features that do not pose a significant alteration to the prediction, are beneficial in that it prevents overfitting and improves generalization(Mubasir, 2020)

Part 3: Descriptive Statistics

I have outlined certain aspects to the dataset in previous sections that are descriptive statistics in the distribution of the target variable and outliers by way of approaching cleaning the dataset. Understanding the central tendency of 'shares' help comprehend the underlying patterns and trends in the data and can influence judgments based on that data, knowing the central tendency of the target variable can be valuable in a decision-making setting (M, 2016)

I overview the central tendency in the dataset again using the 'describe()' function. I also look at the central tendency of the target variable, I established that the mean was separate to the median and mode, but still within the centralized distribution: **Mean: 3355, Median: 1400 and Mode: 1100** (Figure 6) There are numerous ways we can interpret the data through statistical techniques and understanding media-audience relationship through consumer information diets (Makhortykh et al., 2021), helps content creators use these reader habits to drive more traffic and shares.

There are two elements I will look at: **1. Time based consumption habits**: This become all the more relevant with online content as it's immediately accessible through multiple devices(Makhortykh et al., 2021). While devices are relevant, dependency is more relevant for a news distribution. **2. Content based consumption habits:** This is usually looked at in reverse, as traffic for business, entertainment and lifestyle are filtered through social media usage but it is an important factor for content creators. Given this, it would be prudent to look at central tendency of the days of the week content is consumed and the channels that are consumed.

From these I observed that the mu values of the days of the week were highest on Tuesday, Wednesday and Thursday (Figure 7). The data channels centered around World, Technology, Business and Entertainment (Figure 8). I then went to look at the variance of the two highest mu values in each category, which returned 0.1522 (weekday_is_wednesday) and 0.1673 (data_channel_is_world).

Both these levels of variance are low, which shows us the data is closely aligned with the mean, and for context, would indicate that there is consistency in the data itself (Warton and Hui, 2017). These results would be exactly what a digital marketer would be looking for. Statistics provide a support for strategies such as SEO and Advertisements to optimize consistency for gaining traffic, return on ad revenue and click through rates, to name but a few. (Martin, n.d.)

SECTION 2: FEATURE SELECTION AND TRANSFORMATION/ SCALING AND FEATURE IMPORTANCE

Part 1: Feature Selection and Transformation

We have two types of data in this dataframe, one is binary to reference a categorical value and the other is numerical, which would indicate a split would be easier to handle the data. I identify and split the dataframe for these two categories for means accuracy, df_num for numerical and df_cat for categorical. As previously stated, there are significant outliers, however I do not want to eliminate them altogether as I've determined they provide relevant data. However, I will use winsorization to tidy up the features, clipping the lower quartile at 5% and upper at 95% to avoid any loss of pertinent information (Horsch, 2021).

The numerical data was then passed through f_regression to see the important features by their relationship with the target variable which is statistically important (Frost, 2017). The data was split into train and test, 80% to 20%, however I determined all the features in the dataframe are relevant. Reverting to the target variable by way of measuring the data against it, I determined that creating a threshold value to make the feature binary was the most affective. I used central tendency, namely median as it's robust to outliers, to determine the minority and majority classes for a framework for analytics of the dataframe overall.

Part 3: Scaling and Feature Importance

Min / Max Scaling was used to scale so that the input space is not dominated by any one feature(Kumar, 2022). I ran Random Forest Classifier is employed to display the feature importance in order. This will give features to concentrate on as it works on majority ranking using hyperparameters (R, 2021). I got 66% accuracy on this; however, I ran it also without scaling and got 7% accuracy which indicates correct data processing. The two most important features resulted as 'kw_max_avg' and 'kw_avg_avg'. I will use these when performing real average range due to their importance's and provide insight to the differences.

SECTION 3: HYPOTHESIS TEST

Part 1: Introduction

We use the hypothesis test to compare the mean of the variables that are similar, using t-test calculating p-value, which is a measurement of the likelihood that the difference between the two groups is the result of chance, and t-value, which is a measure of the difference between the two groups.

In hypothesis testing, the null hypothesis represents the current understanding or default position, while the alternative hypothesis represents the new or alternative position that is being tested. The goal of the hypothesis test is to determine whether the data provides sufficient evidence to reject the null hypothesis in favor of the alternative hypothesis. When choosing a significance level (also known as the alpha level) for the hypothesis test, it is common to choose a value that corresponds to a low probability of rejecting the null hypothesis when it is actually true (Frost, 2020).

It is generally desirable to have a low standard deviation because it indicates that the data is more precise, and the results of the experiment are more reliable. It means the data points are relatively close to the mean, whereas a high standard deviation means that the data points are more spread out and less reliable (Minitab, 2021) In a hypothesis test, having a low standard

deviation can help to increase the power of the test, which means that it is more likely to detect a significant difference between the means of the two groups if one exists.

Part 2: Content: Test

The feature with the lowest standard deviation is 'global_rate_negative_words' which is the rate of negative words in the content. I performed a hypothesis test against 'max_negative_polarity' which is the rate of negative words in an article which would allude to the overall negative sentiment of it

Part 3: Results:

The Test ran as Ttest_indResult(statistic=382.2157549518096, pvalue=0.0). The larger the t-value, the more evidence there is against the null hypothesis. In this case, the t-value is very large, which suggests there is a significa nt difference between the means of the two samples.
A p-value of 0.0 indicates the extreme probability of obtaining this t-value is extremely low, given that the null hypothesis is false and the alternative is true. This indicates that the difference between the means of the two samples is highly significant and provides strong evidence against the null hypothesis (Dementyev, 2022) I therefore reject the null hypothesis in favour of the alternative hypothesis, and that the relationship between the global rate of negative words and negative polarity is not due to chance (Figure 9)

Section 4: Real Average Range of Two Variables

It is calculated **as** real average range = (sum of all ranges) / (number of data points)**.** I found it difficult to get information on this from an abstract point of view. It seems to be industry focused and used for process productivity in different working fields. It measures the dispersion or spread of the data around the mean. A larger RAR value indicates a greater spread of values and a smaller RAR value indicates a more concentrated set of values (Hayes, 2022)

I used kw_max_avg(keyword max shares) and kw_avg_avg (keyword avg shares) based on feature importance already performed. The code returned 0.187 and 1.098 respectively (Figure 10) These are far from the mean and would indicate a low level of variability. This could indicate that the process producing the data is stable and consistent(Bhandari, 2020)

Section 5: Binomial Distribution and two probabilities of this distribution

Part 1. Introduction

The binomial distribution is a discrete probability distribution that models the number of successes in a given number of independent trials, where each trial has two possible outcomes: success or failure(Vaidya, 2022)

When you choose a feature for a binomial distribution, you should consider the following factors: The feature should have two possible outcomes. These outcomes should be mutually exclusive and exhaustive, and there should be no overlap between them. The feature should be independent of other features, and the probability of success should be constant for each trial (Glen, 2018)

Part 2: Approach and Implementation

As per the process of binomial distribution, I visualized correlations amongst the binary features (Vaidya, 2022) (Figure 11) and ran a sorted view to decide the optimum feature to represent this distribution. I want to choose the feature with the lowest correlation as it reduces complexity and increases accuracy.

I will therefore take the feature 'data_channel_is_tech' as it's the lowest correlation value and it has the structure appropriate for a binomial distribution. As per the remit, I calculated two probabilities using this distribution. Firstly, 6 shuffles, with a probability of 23.5% (the probability of getting one of two values and, 2, being the number of times a positive value will occur (Figure 12). This resulted in a 28.4% probability. My second example will do a shuffle of 10 with the number of times of success being 4 (Figure 13), which resulted in a probability of 12.9%.

SECTION 6: CONCLUSION

The data was capable of being interpreted by either regression or classification; however I chose to interpret the data as a classification problem. The unique 'share' values were low in comparison to the count, which would make a machine learning algorithm easier to predict feature values that influence the target variable. Low cardinality of the target variable will make the algorithm less complex (Sangani, 2021). The descriptive statistics demonstrated a wide range in data values, which pointed towards the necessity of scaling. Further indication of this was looking at related features, such as days of the week, they were closely aligned so that it affirmed the integrity of the dataset.

Variance was able to tell me that target variable was not changing drastically in relation to the changing of the features. I performed feature selection, transformation, scaling and importance, in order to prepare the data for statistical interpretations for the most accurate readings. It was of interest to perform a hypothesis test between features to affirm a null hypothesis and therefore a relationship, which is vital towards machine learning. The low level in Real Average Range shows low variance which can be helpful in identifying trends in the data.

Lastly, binomial distribution provided a concise output of feature probability of successful outcomes. Overall showing that the data is prepared and of suitable balance to perform machine learning for prediction and provide online new providers, vital information to deliver content that is absorbed by the consumer.

SECTION 6: BIBLIOGRAPHY

Atske, S., 2020. Measuring News Consumption in a Digital Era. Pew Research Center's Journalism Project. URL https://www.pewresearch.org/journalism/2020/12/08/measuring-news-consumption-in-a-digital-era/ (accessed 1.6.23).

Bhandari, P., 2021. How to Find Outliers | 4 Ways with Examples & Explanation [WWW Document]. Scribbr. URL https://www.scribbr.com/statistics/outliers/ (accessed 1.7.23).

Bhandari, P., 2020. How to Find the Range of a Data Set | Formula & Examples [WWW Document]. Scribbr. URL https://www.scribbr.com/statistics/range/ (accessed 1.9.23).

Dementyev, A., 2022. T-test and Hypothesis Testing (Explained Simply) [WWW Document]. Medium. URL https://towardsdatascience.com/t-test-and-hypothesis-testing-explained-simply-1cff6358633e (accessed 1.9.23).

Frost, J., 2020. Failing to Reject the Null Hypothesis. Statistics By Jim. URL https://statisticsbyjim.com/hypothesis-testing/failing-reject-null-hypothesis/ (accessed 1.9.23).

Frost, J., 2017. How to Interpret the F-test of Overall Significance in Regression Analysis. Statistics By Jim. URL http://statisticsbyjim.com/regression/interpret-f-test-overall-significance-regression/ (accessed 1.9.23).

Glen, S., 2018. Binomial Distribution: Formula, What it is, How to use it [WWW Document]. Statistics How To. URL https://www.statisticshowto.com/probability-and-statistics/binomial-theorem/binomial-distribution-formula/ (accessed 1.9.23).

Hayes, A., 2022. Average True Range (ATR) Formula, What It Means, and How to Use It [WWW Document]. Investopedia. URL https://www.investopedia.com/terms/a/atr.asp (accessed 1.9.23).

Horsch, A., 2021. Winsorizing – Towards Data Science [WWW Document]. Winsorizing – Towards Data Science. URL https://towardsdatascience.com (accessed 1.8.23).

Kumar, A., 2022. Feature Scaling in Machine Learning: Python Examples. Data Analytics. URL https://vitalflux.com/python-improve-model-performance-using-feature-scaling/ (accessed 1.21.23).

Kumar, P., 2020. How to Handle Outliers | LinkedIn [WWW Document]. URL https://www.linkedin.com/pulse/how-handle-outliers-piyush-kumar/ (accessed 1.7.23).

Lannge, E.J., Fowler, P., 2021. Does removal of correlated variables affect the classification accuracy of machine learning algorithms?

Li, S., 2018. Topic Modeling and Latent Dirichlet Allocation (LDA) in Python [WWW Document]. Medium. URL https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24 (accessed 1.7.23).

M, M., 2016. What are the Measures of Central tendency? Business Jargons. URL https://businessjargons.com/measures-of-central-tendency.html (accessed 1.7.23).

Makhortykh, M., de Vreese, C., Helberger, N., Harambam, J., Bountouridis, D., 2021. We are what we click: Understanding time and content-based habits of online news readers. New Media & Society 23, 2773–2800. https://doi.org/10.1177/1461444820933221

Martin, N., n.d. How Social Media Has Changed How We Consume News [WWW Document]. Forbes. URL https://www.forbes.com/sites/nicolemartin1/2018/11/30/how-social-media-has-changed-how-we-consume-news/ (accessed 1.6.23).

Minitab, 2021. Increase power [WWW Document]. URL https://support.minitab.com/en-us/minitab/20/help-and-how-to/statistics/power-and-sample-size/supporting-topics/increase-power/ (accessed 1.9.23).

Mubasir, M., 2020. Don't Overfit! II — How to avoid Overfitting in your Machine Learning and Deep Learning Models [WWW Document]. Medium. URL https://towardsdatascience.com/dont-overfit-ii-how-to-avoid-overfitting-in-your-machine-learning-and-deep-learning-models-2ff903f4b36a (accessed 1.20.23).

Nicholas, J., 2022. What Is A Data Dictionary | What Is Its Use? | Introduction And Guidance In 2023 | BusinessAnalystMentor.com [WWW Document]. URL https://businessanalystmentor.com/data-dictionary/ (accessed 1.7.23).

R, S.E., 2021. Random Forest | Introduction to Random Forest Algorithm. Analytics Vidhya. URL https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/ (accessed 1.9.23).

Sangani, R., 2021. Dealing with features that have high cardinality [WWW Document]. Medium. URL https://towardsdatascience.com/dealing-with-features-that-have-high-cardinality-1c9212d7ff1b (accessed 1.21.23).

usɛr11852, (https://stats.stackexchange.com/users/11852/us%ce%b5r11852, 2019. Answer to "(Low cardinality) categorical features handling in gradient boosting libraries." Cross Validated.

Vaidya, D., 2022. Binomial Distribution. WallStreetMojo. URL https://www.wallstreetmojo.com/binomial-distribution/ (accessed 1.9.23).

Warton, D.I., Hui, F.K.C., 2017. The central role of mean-variance relationships in the analysis of multivariate abundance data: a response to Roberts (2017). Methods in Ecology and Evolution 8, 1408–1414. https://doi.org/10.1111/2041-210X.12843