D214: DATA ANALYTICS GRADUATE CAPSTONE

Executive Summary and Implications

An Analysis of Online News Popularity using Multiple Linear Regression

Darian Gurrola

August 6, 2024

Problem Statement and Hypothesis

As the digital media landscape grows more saturated, media companies such as Mashable must use any available resource to gain a competitive advantage. These organizations rely heavily on advertising revenue from brands that are willing to pay large sums of money to attract new customers (Miles, 2020). The problem that digital media companies face is identifying which factors are most predictive of higher viewership, and by extension, revenue.

The question being investigated in this analysis is "To what extent do the number of words in an article title, the number of words in an article, the number of links in an article, the number of keywords in the metadata, and the text sentiment polarity of an article influence the number of shares an online news article receives?" This research question is being investigated because of its business applications in the digital media space. By understanding what factors drive reader engagement, digital media companies will be able to maximize their advertising revenue.

The null hypothesis of this analysis is that the variables "n_tokens_title", "n_tokens_content", "num_hrefs", "num_keywords", and "global_sentiment_polarity" do not have a statistically significant effect on "shares". The alternative hypothesis is that the variables "n_tokens_title", "n_tokens_content", "num_hrefs", "num_keywords", and "global_sentiment_polarity" have a statistically significant effect on "shares".

Summary of Data Analysis Process

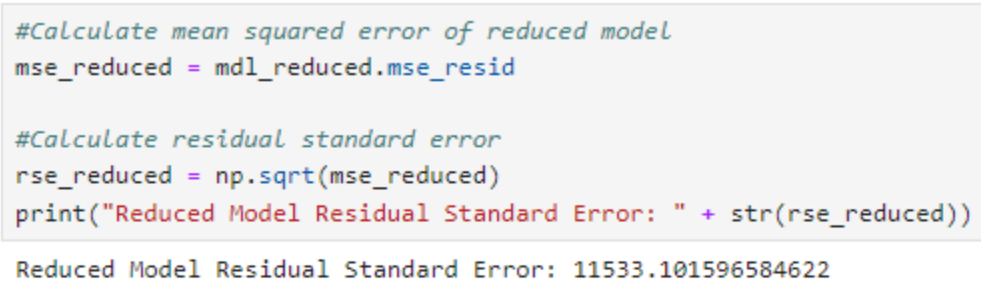
To collect data for the analysis, I used the online news popularity dataset from the UC Irvine Machine Learning repository. This dataset summarizes a set of features about articles published by Mashable in a period of two years (Fernandes et al., 2015). It contains 58 features and 39,727 observations. The amount of data was more than sufficient for performing multiple linear regression.

The next step was to explore the data using Python and several libraries such as pandas, numpy, and matplotlib. Upon review of the data, I discovered that there were several variables that were redundant or unnecessary for the analysis. These were removed to prevent multicollinearity and a negative impact on model performance. I also generated univariate and bivariate visualizations to gain an understanding of each feature and how they relate to the dependent variable, "shares". None of the selected features had a linear relationship with the dependent variable.

To clean the data, I took steps to identify duplicate rows, null or missing values, and outliers. Using pandas, I discovered that there were no duplicates or null or missing values in the dataset. However, using numpy and matplotlib, I found several thousand outliers across my selected features. I ultimately retained these outliers to preserve the sample size of the data and avoid introducing bias.

With the data now adequately prepared, the next step was to build a linear regression model and reduce it using backwards stepwise elimination. The model was first reduced by iteratively removing variables with a variance inflation factor greater than 10. This was done to treat multicollinearity and ultimately produce a more reliable model. After that, I continued reducing the model by iteratively removing variables with a p-value greater than 0.05. The reduced model was left with twelve statistically significant variables. These were "num_hrefs", "num_imgs", "num_videos", "data_channel_is_entertainment", "kw_avg_avg", "self_reference_avg_sharess", "weekday_is_tuesday", "weekday_is_wednesday", "weekday_is_thursday", "weekday_is_friday", "avg_negative_polarity" and "abs_title_sentiment_polarity".

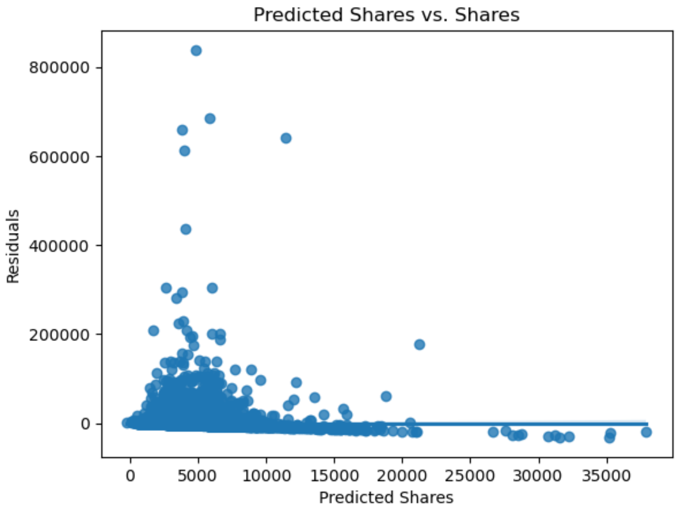
To evaluate the fit and performance of the model, I calculated the residual standard error of the model and plotted the residuals against the predicted values. I also generated a histogram of the residuals.

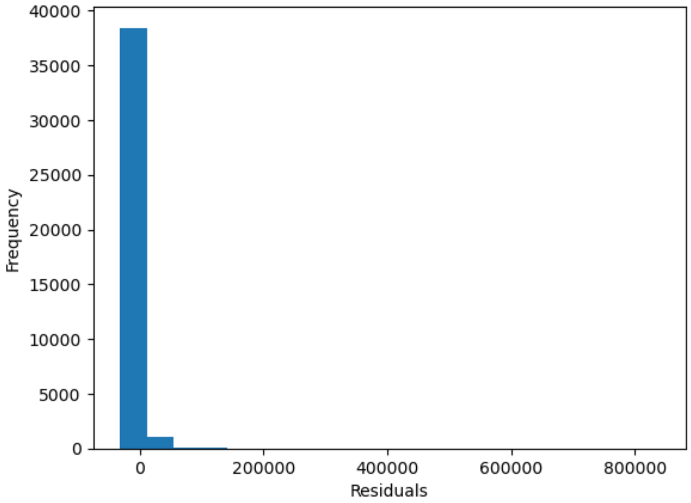


```
#Calculate mean squared error of reduced model
mse_reduced = mdl_reduced.mse_resid

#Calculate residual standard error
rse_reduced = np.sqrt(mse_reduced)
print("Reduced Model Residual Standard Error: " + str(rse_reduced))
```

```
Reduced Model Residual Standard Error: 11533.101596584622
```





Outline of Findings

Based on the results of the analysis, I was unable to reject the null hypothesis. The variables "n_tokens_title", "n_tokens_content", "num_keywords", and "global_sentiment_polarity" did not have a statistically significant effect on "shares". The only hypothesized feature with statistical significance was “num_hrefs”.

Based on the p-value of the F-statistic of the final model, it appears that the model itself is statistically significant. This indicates that model fit the data better than a model with no independent variables (Frost). Despite the statistical significance of the twelve remaining variables and the model itself, the reduced model did not perform as expected.

The adjusted R-squared value of 0.016 indicated that only 1.6% of the variance in the dependent variable could be explained by the independent variables (IBM). Additionally, the residual standard error of 11533.10 was rather alarming considering that median value of "shares" was 1400. Overall, these two metrics demonstrate that the model has poor predictive value and simply does not fit the data well.

Limitations of Tools and Techniques

A limitation of the analysis is that linear regression models are prone to outliers. They appeared in every variable and comprised a sizable portion of the observations. The decision to retain them likely caused the data to violate several assumptions of linear regression.

First, it was apparent during the data exploration phase that none of the features had a linear relationship with the dependent variable. This violated the primary assumption of linear regression. Secondly, the residual plot of the reduced model indicated that the residuals were concentrated around certain value ranges in the dependent variable. This violated the assumption regarding the independence of residuals. Lastly, the histogram of the residuals resulted in a left-skewed distribution. This meant that the normality assumption was violated. The failure of data to meet these criteria almost certainly resulted in a poorly fit model.

Summary of Proposed Actions

Because of the model's poor performance and the failure of the data to meet the assumptions of linear regression, the model does not have much practical use. The recommended course of action should be to perform another analysis using a different approach. The first approach would be to attempt linear regression again after imputing or removing outliers in the dataset. Although this could introduce bias in the dataset, it may help the data meet the assumptions required for a meaningful model. The second approach would be to use the same data with a different type of model such as lasso or ridge regression. These models do not have the same limitations as linear regression and may prove more suitable for the dataset used. Either of these proposed actions might allow a digital media company to obtain more meaningful results and understand what truly drives article shares.

Expected Benefits of the Study

One expected benefit of the study was the reduction of the number of features in the linear regression model. Reducing the model would allow the media organization to filter out the least important features in predicting online news article popularity. I expected to iterate on the initial model reduce the number of variables by first removing those with a VIF greater than 10, and then those with p-values greater than 0.05.

The second expected benefit of the analysis was to create of a statistically significant model. The p-value of the F-statistic of the final model was less than 0.05, indicating that this was also achieved.

The last expected benefit of the analysis was to produce a model with an adjusted R-squared value of at least 0.5. It seemed reasonable for the features in the final model to explain at least 50% of the variation in the dependent variable. Ultimately, however, the actual model did not perform nearly as well as expected.

References

Miles, S. (2020, June 17). How Do Digital Media Companies Make Money?. Webpublisher Pro. https://webpublisherpro.com/how-do-digital-media-companies-make-money/

Fernandes,Kelwin, Vinagre,Pedro, Cortez,Paulo, and Sernadela,Pedro. (2015). Online News Popularity. UCI Machine Learning Repository. https://doi.org/10.24432/C5NS3V.

Frost, J. (n.d.). How to Interpret the F-test of Overall Significance in Regression Analysis. Statistics by Jim. https://statisticsbyjim.com/regression/interpret-f-test-overall-significance-regression/

Adjusted R squared. (n.d.). IBM. https://www.ibm.com/docs/en/cognos-analytics/12.0.0?topic=terms-adjusted-r-squared