Report

Capstone Project

Title: Predicting Social Media Shares

Authored by

Shiva Chaitanya Goud Gadila

Reg No:811252042

Problem Statement:

The main aim of this project is to develop a predictive model that can estimate the number of shares a social media post or article will receive based on various factors provided in the dataset.

Objectives of the Project:

This project aims to achieve the following objectives to solve the problem statement:

- *Finding Insights from the Data:* The initial step of the project will involve performing exploratory data analysis (EDA). By visualizing the distribution of various features and analyzing correlations among variables, the project aims to identify factors that significantly impact the number of social media shares. EDA can reveal trends and hidden patterns that can guide content creators and influencers in tailoring their content strategy for higher engagement.

- *Developing Predictive Model: Given* that the goal is to predict the number of social media shares, this problem can be formulated as a Regression task. The project will involve the implementation of Regression algorithms such as Regression, Decision trees and Clustering algorithms. These algorithms will be trained on the dataset, which will be divided into training and test sets. The predictive model's performance will be evaluated using appropriate metrics like accuracy, precision, recall, and F1-score.

- *Recommend Strategies to Increase Social Media Shares:* The insights gained from the exploratory data analysis and the predictive model will provide valuable information on which features have the most influence on social media shares. Based on these findings, the project will suggest actionable strategies for content creators and influencers to increase their social media shares and improve their engagement rates. These strategies could include content optimization based on popular topics, formats, or timing, leveraging specific media types (images, videos), and understanding the audience's preferences.

Significance of the data:

The dataset from the UC Irvine Online News Popularity Data Set is highly relevant and significant in the context of understanding and predicting the engagement of social media posts and articles. Here are some key reasons why this dataset is valuable:

Real-world Data: The dataset is sourced from actual social media posts and articles, making it representative of real-world scenarios. This enhances the practicality and applicability of any model developed using this data.

Rich and Diverse Information: With 61 columns and 39,644 observations, the dataset provides a wide range of features that can potentially influence the number of shares. Analyzing this data can reveal valuable insights into what factors drive higher engagement and popularity on social media platforms.

Real world application of this Project:

The demand for social media influencers and content providers is enormous since social media has integrated itself into daily life. Based on engagement, likes, comments, and shares, their posts can generate thousands of dollars for them. Predicting engagement is crucial for optimizing their content strategy and partnering with brands effectively. Data analysis and predictive modeling using the provided dataset can reveal insights into factors influencing engagement, leading to more effective content creation, and marketing strategies.

Dataset and Variables description:

This data set is taken from the Machine Learning Repository of the University of California, Irvine Online News Popularity Data Set (UC Irvine). It contains 61 columns and 39,644 observations including different factors like URL, number of tokens in the title and content, number of images, videos, audience reactions, comments and many more, which could significantly affect the engagement rate of content creators on social media.

The studied dataset contains a wide variety of characteristics that define the content of articles that the renowned media source Mashable produced over a two-year period. The dataset, which is a comprehensive and rich source of information, is described as heterogeneous since it includes a number of data types, including binary, category, and numerical data.

Each column in the dataset contains a different trait or characteristic related to the articles, and each entry in the dataset is a specific article published by Mashable.

The variable which must be predicted is shared, based on the values of this variable, it can be concluded that this problem is a Regression type problem.

Data Cleaning:

- **Check for percentage of null values in each column:**

  During this step, using dplyr library, few transformations were performed on the dataset in such a way that it results in the percentage of null values/missing values in each column. The rows with missing values will be handled in the best way, based on the % of missing values.

url	timedelta	n_tokens_title
0	0	0
n_tokens_content	n_unique_tokens	n_non_stop_words
0	0	0
n_non_stop_unique_tokens	num_hrefs	num_self_hrefs
0	0	0
num_imgs	num_videos	average_token_length
0	0	0
num_keywords	data_channel_is_lifestyle	data_channel_is_entertainment
0	0	0
data_channel_is_bus	data_channel_is_socmed	data_channel_is_tech
0	0	0
data_channel_is_world	kw_min_min	kw_max_min
0	0	0
kw_avg_min	kw_min_max	kw_max_max
0	0	0
kw_avg_max	kw_min_avg	kw_max_avg
0	0	0
kw_avg_avg	self_reference_min_shares	self_reference_max_shares
0	0	0
self_reference_avg_sharess	weekday_is_monday	weekday_is_tuesday
0	0	0
weekday_is_wednesday	weekday_is_thursday	weekday_is_friday
0	0	0
weekday_is_saturday	weekday_is_sunday	is_weekend
0	0	0

As a result of this step, no column in the dataset has missing values. Since there are no null values, no additional steps were to be taken to ensure the consistency of the data.

The following are the few more steps that were implemented as part of Data Cleaning.

- **Removing irrelevant variables based on Domain knowledge:**
  Removing irrelevant variables from the dataset which wouldn't add much value to the prediction of social media shares. Following variables have been removed from the dataset:
  Url
  TimeDelta

  Number of rows and variables in the dataset:

```
dim(Data1)
```

```
[1] 39644    59
```

There are 39,644 rows and 59 variables in the dataset.

- **Removing rows where number of words in the article are zero:**

  I noticed that a few rows in the dataset have number of words in content to be 0. For any social media post, if there are articles with no content in it, those articles will not be shared and having such articles in the dataset might affect the results of our analysis as those articles can be considered as outliers.
  Hence, identifying and removing such articles is crucial before performing analysis.

```
mydata = Data1 %>% filter(n_tokens_content!=0)
dim(mydata)
```

```
[1] 38463    59
```

After removing rows in which the count of words in an article is zero, there are 38463 rows and 59 variables in the new dataset.

- **Eliminating variables by developing a Preprocess model to remove variables with near zero variance and high correlation:**

  During this step, created a preprocess model which removes variables with near-zero variance and applies correlation-based filtering to remove highly correlated columns.

  Near-zero variance is a method used to remove columns with very low variance, which typically means that these columns have almost constant values and therefore contribute little information. *Correlation* based filtering removes columns which are highly correlated to reduce multicollinearity and redundancy in the data.

```
preProcessModel <- preProcess(mydata[,-c(59)], method = c("nzv", "corr"))
Preprocessed_data <- predict(preProcessModel, mydata)

preProcessModel
```

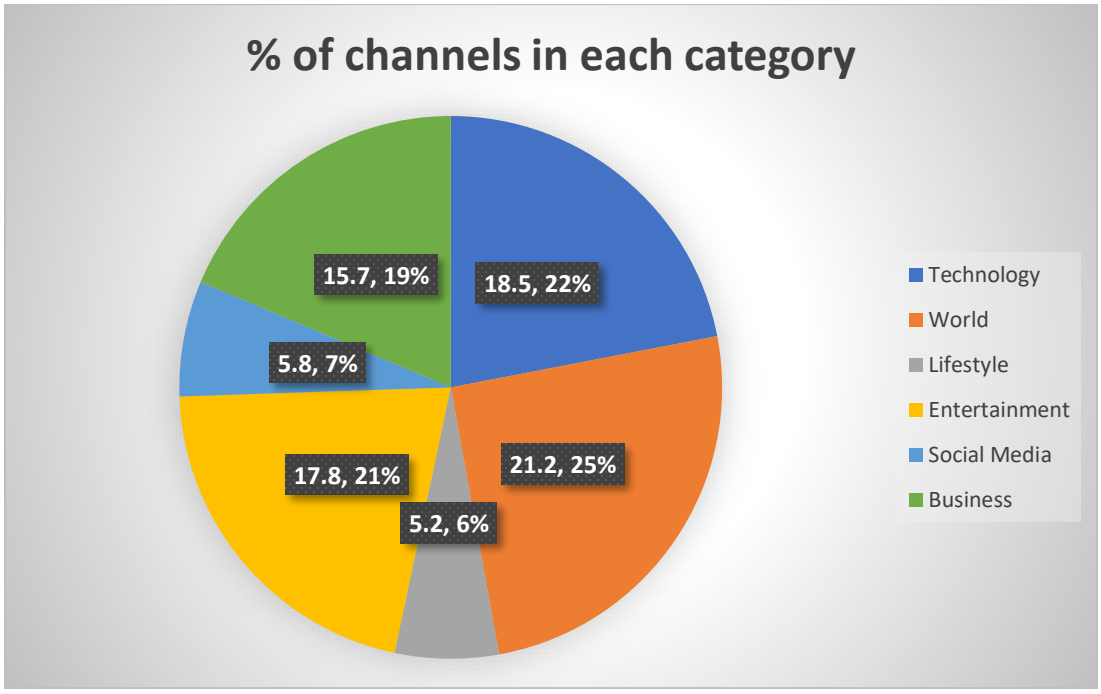
```
Created from 38463 samples and 5 variables

Pre-processing:
  - ignored (0)
  - removed (5)
```

5 variables have been removed after predicting this preprocessing model on the dataset.

Towards the end of the Data Cleaning Process, 54 variables were considered for further analysis.

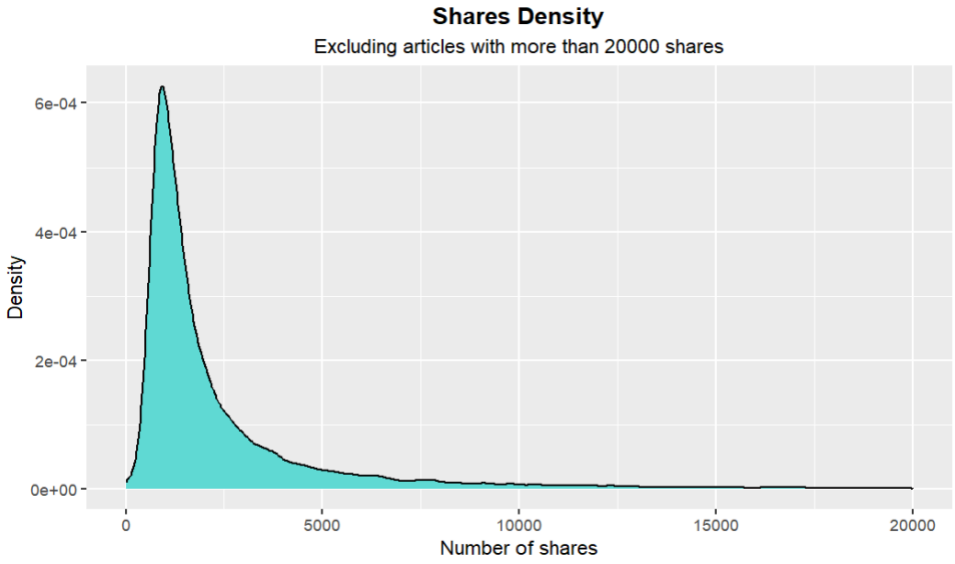
Data Exploration:

- **Checking the percentage of data channels in Lifestyle, Entertainment, Technology, Business, social media, and World:**



  - 25% of the channels post content related to World topic, whereas 22% are of technology and 21% are of Entertainment category.
  - Only 7% of social media coounts post content relevant to social media topic.

- **Checking the density of Number of shares variable:**

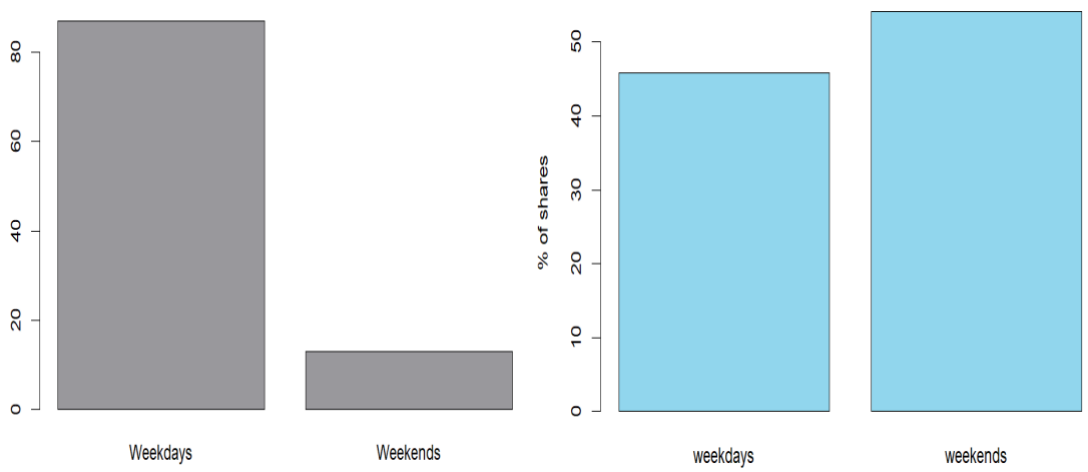


- **Checking the percentage of articles published on weekend:**

  Checking if posting articles on weekends has any effect on number of shares:

is_weekend <int>	count <int>	percentage <dbl>
0	34454	86.90849
1	5190	13.09151

is_weekend <int>	avg_shares <dbl>	percentage <dbl>
0	3277.805	45.84901
1	3871.324	54.15099

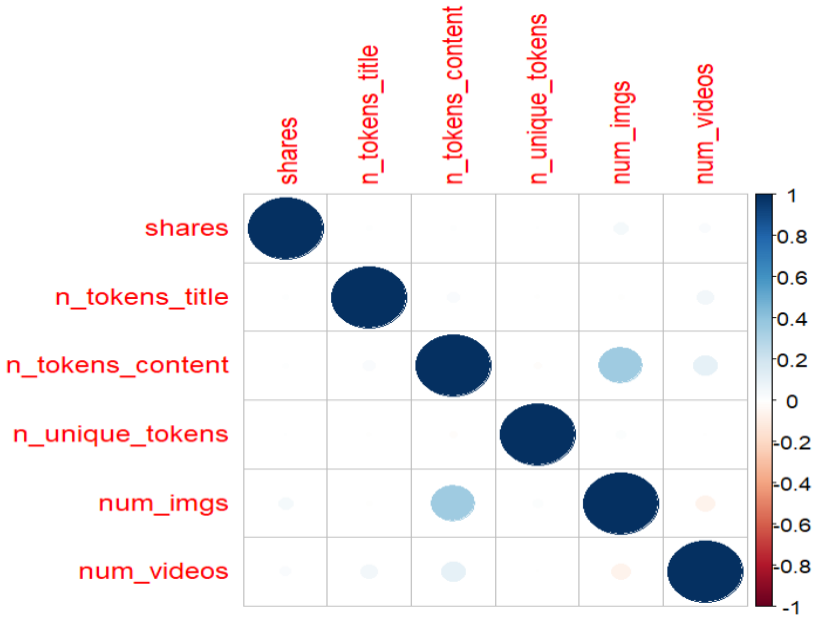


  - Only 13% of articles were published on weekends. However, articles posted on weekends are attracting 8% more number of shares when compared to articles posted on Weekdays.

- **Checking if there is any correlation between number of shares and number of words in an article:**

  Based on the below correlation plot:
  - Having a greater number of words in the title or in the article does not help in gaining more shares.

- A little correlation can be observed between the number of words in the article and number of images/videos.



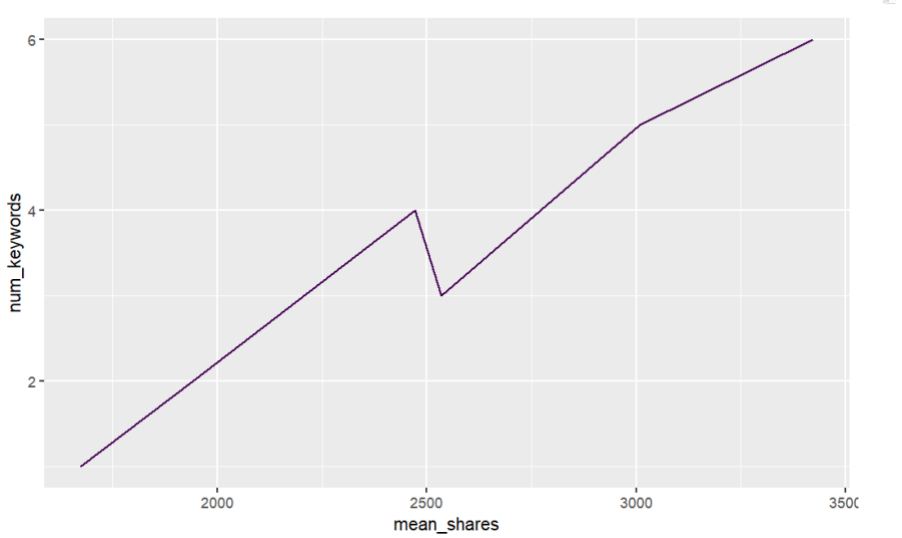
- **Checking if adding more images and Videos in the article result in greater number of shares or not:**

Graphics <int>	mean_shares <dbl>
57	629.00
72	796.00
69	888.00
67	1025.75
85	1100.00
77	1228.50

- Having a greater number of images and Videos in an article does result in better number of shares.

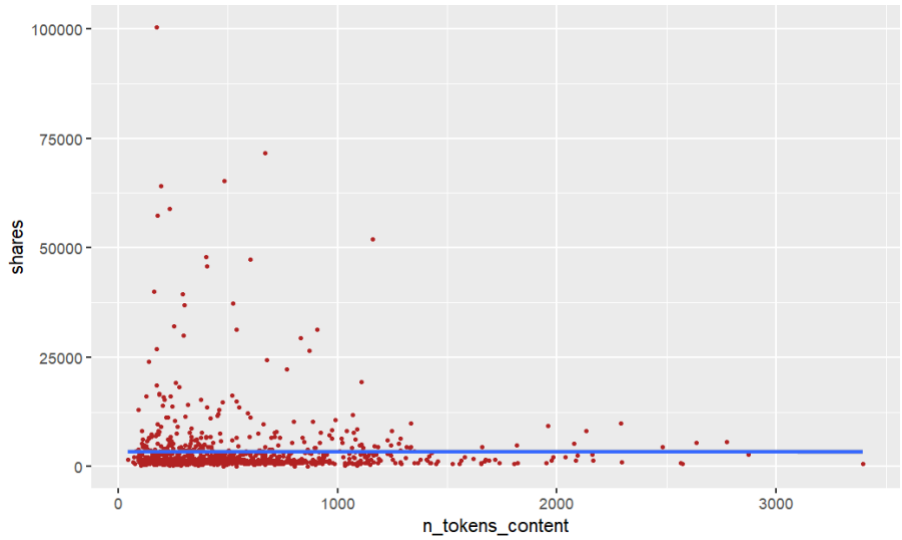
- **Checking the distribution of Number of Keywords and if having more keywords result in more shares or not:**

num_keywords <int>	mean_shares <dbl>
1	1673.600
2	1941.422
4	2471.083
3	2534.948
5	3008.349
6	3421.353



- Clearly, having a greater number of keywords in the article results in more shares.

- **Checking if having a greater number of words will lead to more shares or not:**



- We could see that having a greater number of words does not help in getting more shares. In fact, having more words resulted in lesser shares.

Data Preparation:

- *Variable Selection:* Removing a few variables from the dataset which would not add much contribution in predicting sales variable. All these variables have binary values 1s and 0s. After removing all such variables, 40 variables were selected to be used for model building.

- *Data Partition:* The dataset has been divided into a Training Data subset, containing 70% of the observations, and a Test Data subset, containing the remaining 30%. The Training Data is used to train the machine learning model, allowing it to learn patterns and relationships present in the data. The Test Data is then used to evaluate the model's performance on unseen data. This evaluation is crucial to understand how well the model generalizes to new, previously unseen instances.

```
set.seed(123)
samples=createDataPartition(Model_variables$shares,p=0.7,list=FALSE,times=1)  #Training And Testing
training = Model_variables[samples, ]
test = Model_variables[-samples,]
```

```
dim(training)
```

```
[1] 26925    40
```

```
dim(test)
```

```
[1] 11538    40
```

- *Data Normalization: In* this dataset, all the variables are numeric. Normalizing the information guarantees that every variable is on a comparable scale, which can work on the exhibition of AI calculations and keep specific highlights from ruling others because of their scale.

  By performing Data Normalization, every one of the factors in the dataset have been rescaled to a typical reach, making it more straightforward for the AI calculations to gain from the information and make expectations.

	n_tokens_title <dbl>	n_tokens_content <dbl>	n_unique_tokens <dbl>	num_hrefs <dbl>	num_self_hrefs <dbl>
1	0.7636987	-0.73152575	0.0212510472	-0.6344722	-0.3582373
2	-0.6544434	-0.65486869	0.0074684620	-0.7224200	-0.6178331
3	-0.6544434	-0.74856065	0.0005331746	-0.7224200	-0.6178331
4	-0.6544434	-0.06716457	-0.0161745442	-0.1947330	-0.8774288
5	1.2364127	1.08482068	-0.0368168486	0.6847453	4.0548905
6	-0.1817293	-0.40999197	-0.0030361505	-0.8103678	-0.3582373

Model Selection:

During this step, below is the strategy I used before selecting a model:

1. Select models which would suit the best for our problem type.
2. Carefully weigh the pros and cons of each model.
3. Assuming that there are any drawbacks of the model which could influence the exhibition of our model while taking care of the issue, then, at that point, carry out another model which could beat the impediments of the past model.
4. Finally, evaluate the performance of each model using different performance metrics and choose a model with the least errors and best performance.

As the *Shares* variable which we are trying to predict is a continuous numerical value, I started with a basic Multiple Linear Regression Model.

1. **Linear Regression:** A statistical method called regression analysis is used to examine the connection between two or more variables in a way that allows the dependent variable to be predicted using one or more independent variables.

Advantages of Linear Regression:

- The model's independent variable coefficients shed light on the magnitude and direction of their influence on the dependent variable. Understanding the relative significance of various predictors is made easier by this.
- The relevance of each predictor's coefficient is statistically tested and determined by p-value in multiple linear regression. This makes it easier in determining whether a predictor significantly affects the dependent variable's variation.
- You can eliminate potential confounding variables—variables that could affect both the dependent and independent variables—by including appropriate predictors in the model. More precise inferences can be drawn as a result.

Disadvantage of Linear Regression:

- Overfitting, when the model performs well on the training data but badly on new data, can result from adding too many predictors to the model without sufficient validation.
- For multiple linear regression to yield accurate results, the sample size must be large enough in relation to the number of predictors. Unstable estimations and faulty inferences might result from a lack of data.
- The coefficient estimations can be strongly influenced by outliers and significant points, which could result in false findings. Regression techniques that are robust can help to solve this problem.
- The complexity of the model rises along with the number of predictors. As a result, the model could be more challenging to understand and may need bigger sample sizes to produce valid results.

- Regression can reveal connections between variables, but it cannot establish causality. The associations seen in the model may be caused by other confounding factors since correlation does not always imply causation.

After studying the advantages and disadvantages of Regression model, I believe that this wouldn't be a perfect model to estimate the social media shares as our objective is just not to predict the shares but also to be able to tell what factors influence these shares and what strategies/recommendations we could provide in order to increase the shares.

But, before exploring a different algorithm, I have implemented Regression model to see if there are any other conclusions that can be drawn through this model.

Implementation of Linear Regression:

```
Model<-lm(log(shares)~.,data=Normalized_data)
summary(Model)
```

```
Call:
lm(formula = log(shares) ~ ., data = Normalized_data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6788 -0.9320  0.1085  1.0311  5.7722

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               -1.63258    0.02309 -70.706  < 2e-16 ***
n_tokens_title             0.06327    0.02166   2.921  0.00350 **
n_tokens_content           0.02747    0.03202   0.858  0.39098
n_unique_tokens            2.62596    1.35485   1.938  0.05265 .
num_hrefs                  0.05877    0.02163   2.718  0.00660 **
num_self_hrefs            -0.06891    0.02196  -3.138  0.00171 **
num_imgs                   0.02573    0.02257   1.140  0.25435
num_videos                 0.01687    0.02046   0.824  0.40975
average_token_length      -0.02976    0.02247  -1.324  0.18546
num_keywords              -0.05845    0.02516  -2.323  0.02024 *
kw_min_min                 0.06228    0.03804   1.637  0.10166
kw_avg_min                -0.01575    0.02200  -0.716  0.47403
kw_max_max                 0.08894    0.04288   2.074  0.03809 *
kw_avg_max                -0.05741    0.03278  -1.751  0.07993 .
kw_min_avg                -0.10335    0.02605  -3.967 7.36e-05 ***
```

```
Residual standard error: 1.527 on 5404 degrees of freedom
  (21483 observations deleted due to missingness)
Multiple R-squared:  0.05248,   Adjusted R-squared:  0.046
F-statistic:  8.09 on 37 and 5404 DF,  p-value: < 2.2e-16
```

- R square of this model is just 5%. R square indicates the proportion of variability explained by the independent variable in predicting the dependent variable.

  I'm thinking that the model is underfitting because the R-square is too low, which suggests that it isn't strong enough to depict the link between the dependent and independent variables.

- The P-value is used to understand the significant relationship between each independent variable and the dependent variable.

  As may be obvious, a couple of factors have a critical relationship with the dependent variable.

  I have made the decision to investigate many models before assessing the performance of this one.

2. **Decision Trees:** Decision trees operate through the recursive division of the input data into subsets according to the values of the input features. Each split is chosen by choosing the feature and threshold value that, in accordance with a given criterion, produces the best separation of the target variable.

Regression decision trees are built similarly to classification decision trees, but they forecast numerical values rather than class labels.

Advantages of Decision Trees:

- Decision trees are simple to comprehend and visualize. It is possible to explain how a choice is made based on input features because of the intuitive tree structure.
  This was one of the disadvantages of the Regression model which can be overcome by implementing Decision Trees.
- Decision trees offer a ranking of the relevance of each feature, which can be useful in determining which features have the greatest predictive power.
- Since they base their conclusions on the majority of votes in each leaf node, decision trees are less vulnerable to outliers than linear regression models.

Disadvantages of Decision Trees:

- Decision When allowed to grow deeply, trees are particularly prone to overfitting. They can build intricate trees that precisely suit the training set of data, but they struggle to generalize those results to brand-new, untested data. Finding the ideal balance between complexity and generalization can be difficult, however regularization techniques like pruning and establishing maximum depth can assist to alleviate this.
- When trained on various subsets of the data, decision trees can yield wildly divergent outcomes due to their high variance. Pruning and ensemble approaches are effective in reducing variation, but they may also create bias.
- Decision trees may have trouble capturing more intricate interactions between characteristics, even while they can collect some sorts of correlations in the data. They lack the innate ability to represent interactions, such as those that polynomial regression, for instance, captures.

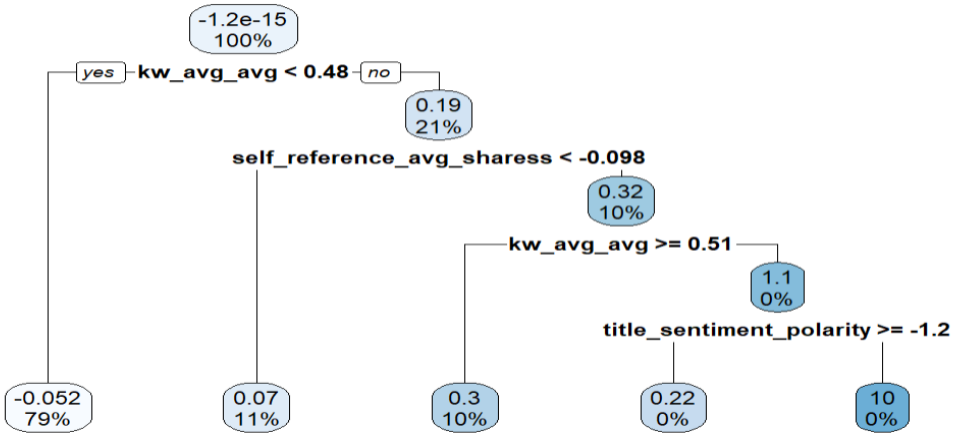
Most of the limitations of Regression model like lack of interpretability and inability to capture complex relationships can be overcome by using Decision Trees.

Hence, implementing Decision Trees.

Implementation of Decision Trees:

```
library(rpart)
library(rpart.plot)

DT_Model<- rpart(Normalized_data$shares~.,data=Normalized_data,method = "anova")

summary(DT_Model)

rpart.plot(DT_Model)
```

Method *Anova* is the method used for Regression problems in Decision Tree, whereas *class* is used for Classification.



Based on the above visual of decision trees, it can be interpreted that the variable kw_avg is given the first preference to split, while title_sentiment_polarity has been given the least preference. Decision trees offer a ranking of the relevance of each feature, which can be useful in determining which features have the greatest predictive power.

After studying Decision Trees, it looks like Decision Trees could be the best model for our problem. However, to overcome the limitations of Decision Trees, I was curious to explore Random Forest algorithm as well.

3. **Random Forest**: In order to increase prediction accuracy and generalization, the ensemble learning technique Random Forest combines several decision trees.

By training a group of decision trees on multiple subsamples of the training dataset, a Random Forest for regression is created for dealing with regression problems. The forecasts of individual trees are combined to get the final prediction after each tree is independently constructed.

Bagging is a method that Random Forest utilizes to provide varied training subsets. The original training dataset is used to draw several random samples (with replacement). A different decision tree is trained using each of these subsets.

A random subset of features is taken into account for the optimum split for each node of a decision tree. This randomization aids in lowering the correlation between the trees, strengthening the ensemble.

Advantages of Random Forest:

- Overfitting is reduced by combining the predictions from various trees, which improves generalization to fresh data.
- The ensemble is less sensitive to data noise because of the randomization that is included during training.
- The value of each feature in creating predictions can be revealed through Random Forest, assisting in feature selection.

Implementation of Random Forest:

```
library(randomForest)
rf_model <- randomForest(shares ~ ., data = Normalized_data,ntree=150, mtry=3,maxnodes=60 )

rf_model
```

```
Call:
 randomForest(formula = shares ~ ., data = Normalized_data, ntree = 150,      mtry = 3, maxnodes
60)
               Type of random forest: regression
                     Number of trees: 150
No. of variables tried at each split: 3

          Mean of squared residuals: 0.9890115
                    % Var explained: 1.1
```

Hyperparameters in the above model are ntree, mtry and maxnodes.

Ntree- Number of trees

Mtry- Number of variables randomly sampled for splitting at each node.

Maxnodes- Maximum number of terminal nodes in each tree.

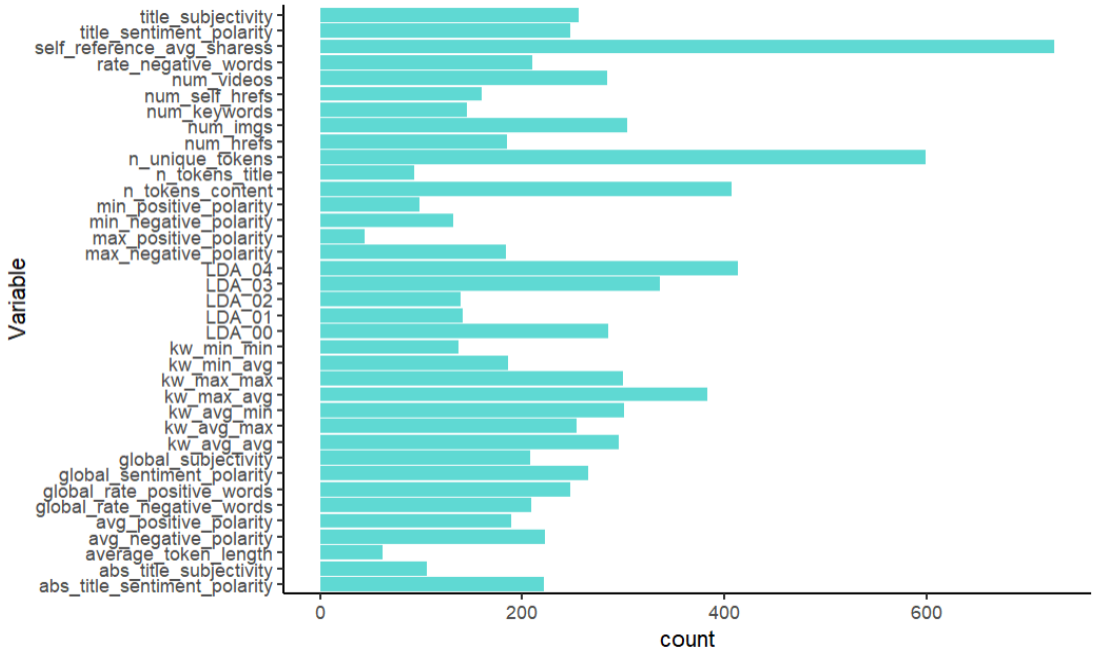
All of the above hyperparameters have been used to tune the model's performance. First, we started with 200 number of trees, 5 mtry value and 60 to be the maximum number of nodes.

This model was developed and tested on unseen data. When the performance of this model was giving higher RMSE and MAE, to reduce the error, different combinations of these hyperparameter values were tried until the error had decreased.

Variable importance from Random Forest algorithm:

As Decision Trees and Random Forest helps in understanding the importance of variables in predicting the target variable, below code was implemented to know variable importance.

```
#Visualizing the variable importance:
var_imp <- as.data.frame(importance(rf_model))
var_imp$Variable <- rownames(var_imp)
```



- From the above plot, it can be concluded that the highest importance was given to the self_reference_avg_shares variable, as the algorithm considered it to be the one which has the highest predictive power in predicting the shares variable.

Evaluating the performance of the models:

There are several metrics like Sum Squares of Errors (SSE), Root Meant Square Error (RMSE), Mean Absolute Error (MAE), R square etc. to evaluate the performance of a model for Regression type problems.

Each of these performance metrics have both advantages and disadvantages. Few of the **disadvantages** are listed below:

- SSE increases there are a greater number of data points; actual value of SSE does not say much about model goodness.
- Moreover, SSE is expressed in squared units of the dependent variables.
- On the other hand, RMSE foes consider the number of data points and residual is divided by the number of points.
- Minimizing RMSE is same as minimizing SSE.

I have specifically chosen RMSE and MAE both to evaluate the performance.

1. **Linear Regression:**

```
Testing<- predict(Model, Normalized_test)
View(Testing)
```

```
RMSE(Normalized_test$shares,Testing)
```

```
1.919897
```

```
MAE(Normalized_test$shares,Testing)
```

```
1.636684
```

2. **Decision Trees:**

```
MAE(Normalized_test$shares,predictions)
```

```
0.263379
```

```
RMSE(Normalized_test$shares,predictions)
```

```
0.9805839
```

3. **Random Forest:**

```
Predictions_rf<-predict(rf_model,Normalized_test)
```

```
RMSE(Normalized_test$shares,Predictions_rf)
```

```
0.9770674
```

```
MAE(Normalized_test$shares,Predictions_rf)
```

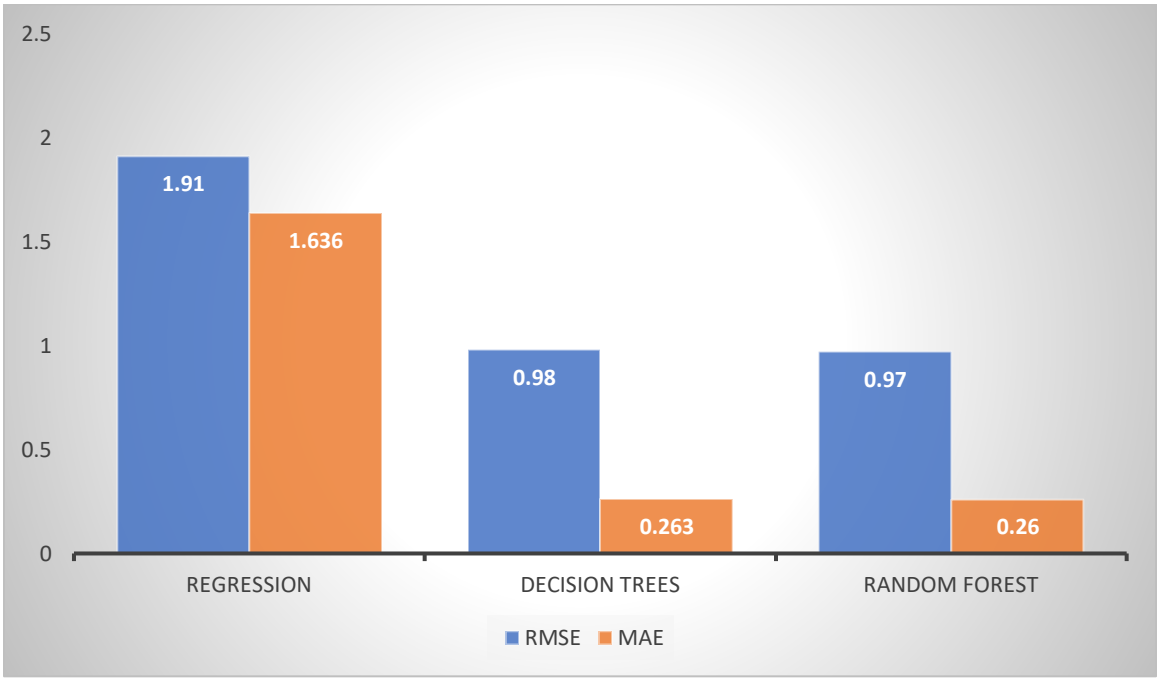
```
0.2606968
```

Interpreting the performance of all models:

After evaluating the performance of all models on unseen data, I have compared and interpreted their loss functions, in order to check which models have performed better than the other models.

Model	RMSE	MAE
Linear Regression	1.91	1.636
Decision Trees	0.98	0.263
Random Forest	0.97	0.260

Visualizing the performance of each model:



- Based on the above results, it can be concluded that Decision Trees improved the performance of the model significantly compared to Regression model.
- A drastic decrease in the MAE was observed when Decision Trees are used over Regression.
- Even though Random Forest model is expected to reduce overfitting, in comparison to Decision Trees, they do not show much of an improvement.
- Hence, we conclude that Decision Trees algorithm works best in predicting the number of shares variable in our dataset.

Insights:

- 25% of the channels post content related to World topic, whereas 22% are of technology and 21% are of Entertainment category.
- Only 7% of social media coounts post content relevant to social media topic.
- Only 13% of articles were published on weekends. However, articles posted on weekends are attracting 8% more number of shares when compared to articles posted on Weekdays.
- Having a greater number of words in the title or in the article does not help in gaining more shares.
- A little correlation can be observed between the number of words in the article and number of images/videos.
- Having a greater number of images and Videos in an article does result in better number of shares.
- Having a greater number of keywords in the article results in more shares.

Recommendations:

To improve the engagement rate on social media posts, we suggest the following strategies based on our analysis:

- As incorporating images and videos proved to be increasing the number of shares, we recommend inserting relevant images and videos in the articles along with content.
  In general, visuals attract the attention of audiences more than just content. Hence, it is always preferred to add relevant images and videos only.
- There is no significant amount of increase in the number of shares by increasing the number of words in the content. In fact, the shorter the length of the content, the better the reach.
- Sharing the article links to other profiles helps in keeping up the engagement rate and inviting other creators in your profession to like and comment on posts along with sharing will help your profile's overall visibility.
- Posting articles on weekends will reach a wider audience and improve number of shares when compared to posting on weekdays.

Conclusion:

In conclusion, this project focuses on using data analysis and predictive modeling to uncover insights and strategies for optimizing social media engagement. Through exploratory data analysis, we identify influential factors in content sharing. By developing predictive models, we offer accurate tools for estimating engagement rates. The project's significance lies in its potential to empower content creators and influencers to tailor their strategies based on data-driven recommendations, ultimately enhancing engagement and impact in the competitive landscape of social media.