A MACHINE LEARNING MODEL OF PREDICTING THE VIEWERSHIP OF MISINFORMATIONAL VIDEOS ON COVID - 19

Course Name – Business Analytics
Course Code – K301

Submitted To –

Khan Muhammad Saqiful Alam
Adjunct Faculty
Institute of Business Administration,
University of Dhaka

Submitted By –

Vaskor Roy
Roll – 24
Section – B
BBA 27th Batch

Date of Submission – December 11, 2021

Table of Contents

Dataset Details

Dataset Description

This dataset contains metadata from all Covid-related YouTube videos that were removed by YouTube due to misinformation. Between November 2019 and June 2020, 8,122 videos were shared. There are unique IDs for the videos and social media accounts that posted them, as well as statistics on social media interaction.

Dataset Link

A dataset of Covid-related misinformation videos and their spread on social media | Zenodo

Drive Folder Link for Related Analysis and Visualization Files

https://drive.google.com/drive/folders/1B7XlAlyjlW6d3JCaeqj2JmeTXwezSm26?usp=sharing

Reasons Behind Choosing This Specific Dataset

Misinformation can induce people to make choices that are not only wrong, but also potentially life-threatening in some cases. This has become even more evident during the COVID – 19 pandemics, where misinformation about the disease has claimed and continues to claim lives. Therefore, working on this dataset can provide insights on how social media influences the spread of misinformation.

Expectations From This Dataset

1. Determining the factors that attracts people to these kinds of fake videos carrying misinformation and disinformation

2. Determining the significance of each factor over people's tendency of watching these videos

Notes

1. In order to create proper visualizations and machine learning models, rows carrying empty values for specific variables such as "view_count" and "subscriber_count" have been removed, leaving only 758 rows of data representing 758 videos.

2. The list of keywords used in these visualizations are not yet complete and may need to be refined even more to get a more comprehensive picture of the situation.

Exploratory Visualizations

Before conducting analysis on the dataset, let us explore the various variables in the dataset itself.

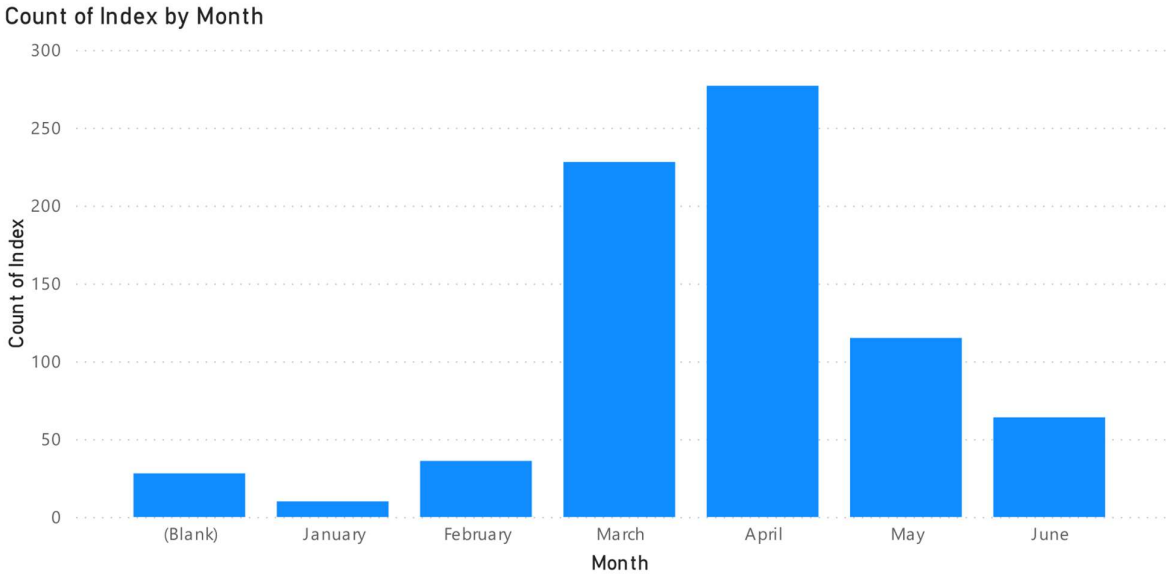


Figure 1 - Column Chart of Number of Videos by Month

This column chart indicates the number of videos that were uploaded across the months of 2020, and as seen in the chart, April 2020 experienced the highest number of video uploads. The (blank) column here indicates those videos for which the upload date was not found in the dataset.

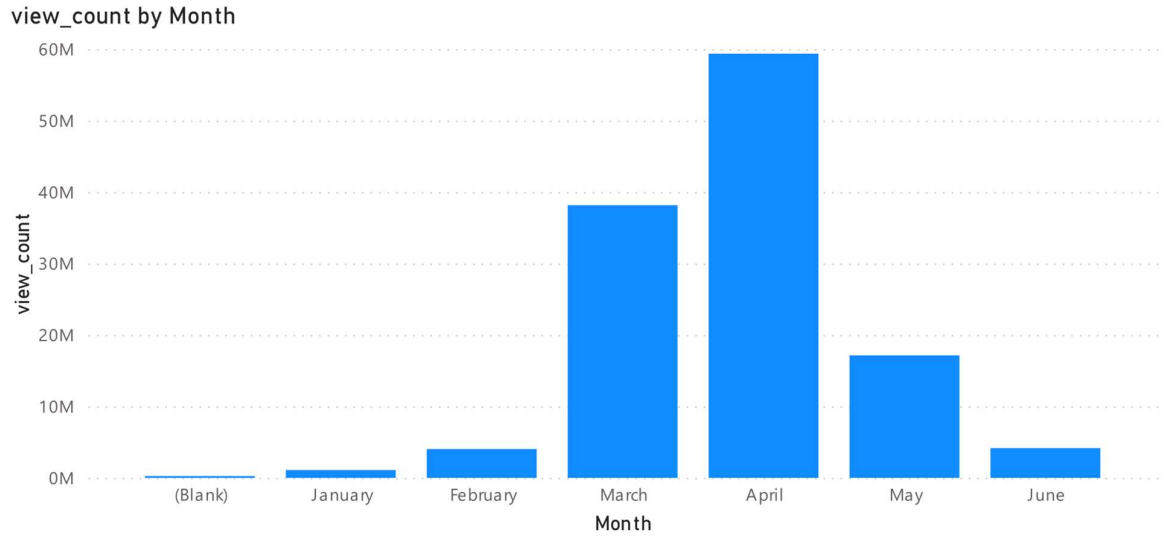


Figure 2 - Column Chart of Sum of view_count by Months

This second column chart, on the other hand, represents the total number of views garnered by these videos during the first 6 months of 2020. Again, as seen from this chart, April 2020 is the month when these videos were able to get the greatest number of views from their audience

Figure 3 - Word Cloud Chart of video_title

This word cloud represents the most found words in the titles of all the 758 videos in the dataset. The most used word is obviously COVID and coronavirus. But surprisingly enough, there are also names of individuals like Bill Gates, the founder of Microsoft, Dr. Anthony Fauci, the director of the National Institute of Allergy and Infectious Diseases, Dr. Judy Mikovits, a former scientist who has discredited due to her anti-vaccine claims and conspiracy theories. At the same time, there are also words like "Hoax, Fake" etc.

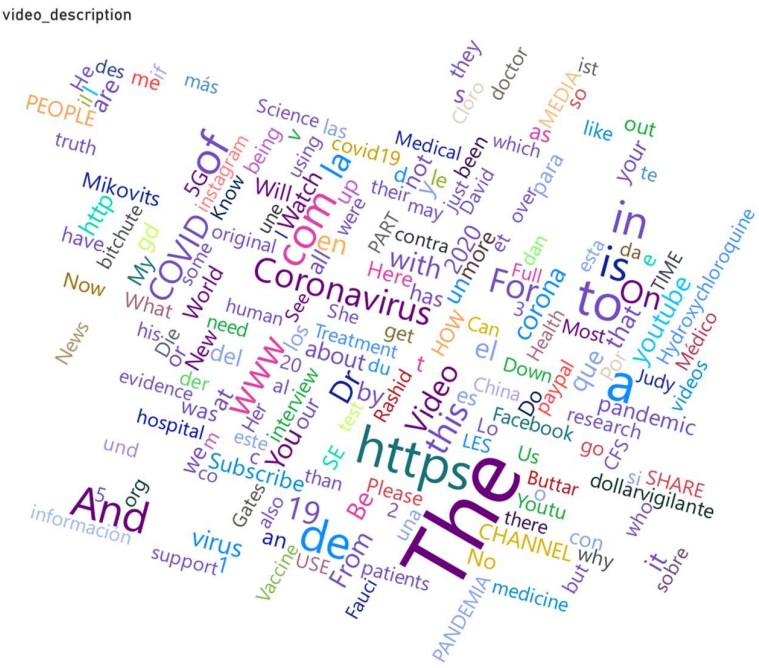


Figure 4 - Word Cloud Chart of video_description

This second word cloud represents the words most used in the description of the 758 videos of the dataset. Setting aside words like "The", "Subscribe"' the most used word related to the context is again "Coronavirus". At the same time, similar to the word cloud for the video titles, words like "Mikovits", "Hydroxychloroquine", "Pandemic" are prevalent here as well.

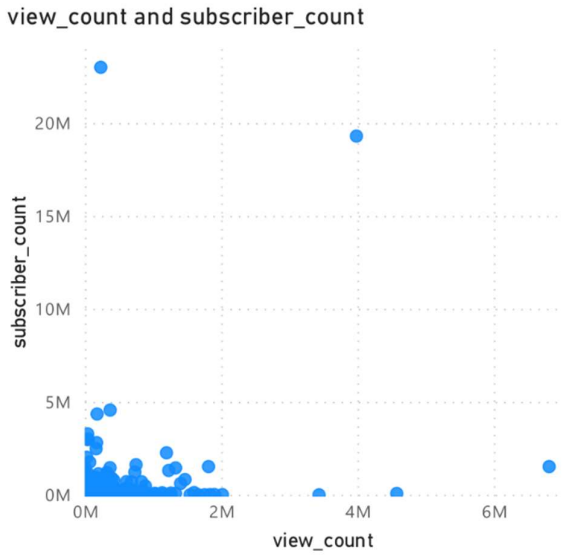


Figure 5 - Scatter Plot of subscriber_count vs view_count

This visualization is a scatter plot with the view count of the video in the horizontal axis and the subscriber count of the channel uploading the video in the vertical axis. This scatter plot looks mostly empty due to the presence of extreme outliers as seen in the plot.

Therefore, this second scatter plot given below was made where the both the variables in concern were filtered to only show values with 3000000 or less, in order to get a clearer picture about the distribution of these videos.

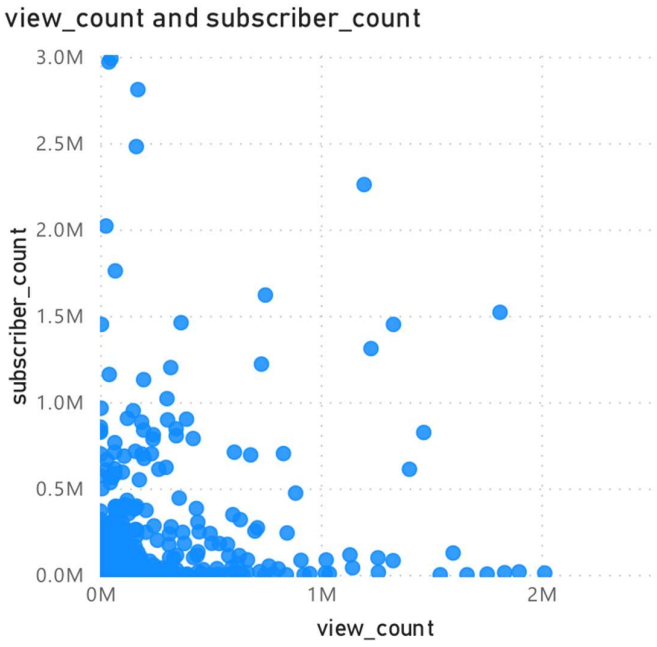


Figure 6 - Modified Scatter Plot of subscriber_count vs view_count

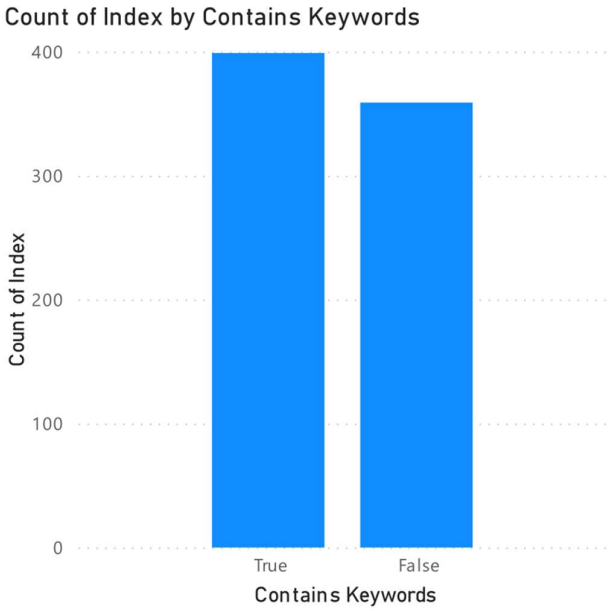


Figure 7 - Column Chart of Number of Videos by the variable - "Contains Keywords"

This column chart is a simple representation of the titles or the descriptions of the videos having certain keywords which is being assumed as one of the possible factors behind the audience watching these fake videos. As seen on the chart, of the 758 videos considered, 399 of them had these keywords, while the rest did not contain those keywords. It is important to note, however that the list of these keywords is not yet complete and thus this number can vary in the future.

The donut chart given below represents the number of videos that contain specific keywords in their titles or descriptions. Suffice to say that, this chart can change depending on the introduction of other keywords or other variables that may or may not have been considered.

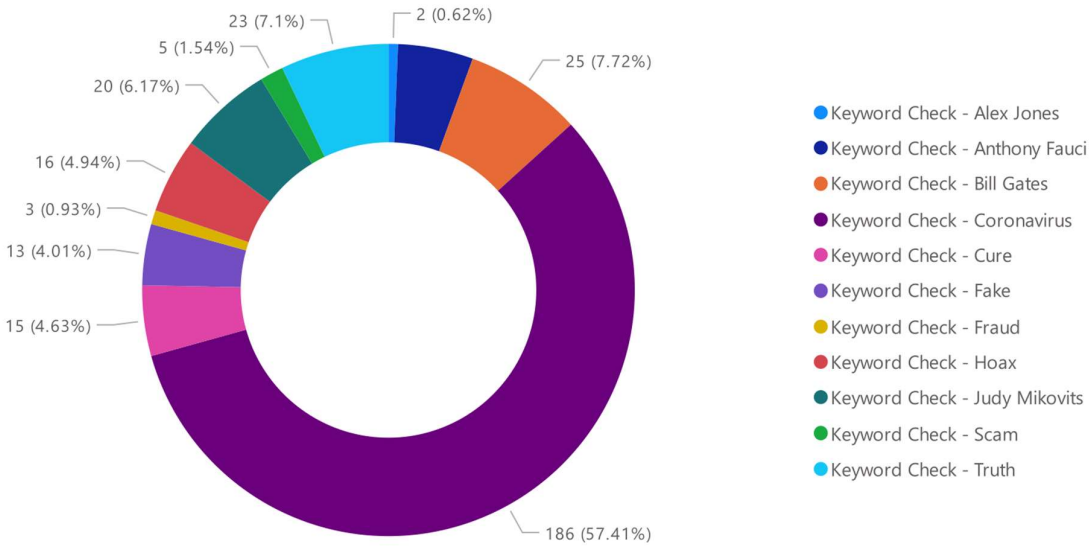


Figure 8 - Donut Chart of Representation of Various Keywords

Exploratory Analysis

Introduction of Variables

There are 5 numerical variables in the dataset in concern –

a) **view_count** – the number of views each video received
b) **subscriber_count** – the number of subscribers each channel publishing the video maintained
c) **facebook_graph_reactions** – the number of reacts each video got when shared on Facebook
d) **facebook_graph_comments** – the number of comments each video got when shared on Facebook
e) **facebook_graph_shares** – the number of times each video was shared on Facebook

On the other hand, there is only one categorical variable in the dataset –

a) **ContainsKeywords** – Whether the title or description of the video contains specific keywords that are assumed to be related to the current pandemic of COVID 19. For this variable, the value 1 indicates that the video title or description contains any one of those keywords, whereas the value 0 indicates that there are none.

Summary Statistics of Variables

The following table provides summary statistics (i.e., maximum or minimum value of the variable, mean of the variable, standard deviation of the variable etc.) for the numeric variables – "view_count, subscriber_count, facebook_graph_comments, facebook_graph_shares, facebook_graph_reactions" of the dataset.

Measurements	view_count	subscriber_count	facebook_graph_comments	Facebook_graph_reactions	Facebook_graph_shares
Mean	163900.0343	204562.3113	17433.76781	21175.6504	11060.39578
Standard Error	16060.6794	42091.33607	4204.091977	3906.865355	1624.22763
Median	32927.5	11500	725.5	1763.5	1696.5
Mode	4	1	2	264	321
Standard Deviation	442179.4098	1158850.238	115746.2186	107563.0348	44717.90993
Sample Variance	1.95523E+11	1.34293E+12	13397187128	11569806452	1999691468
Kurtosis	92.20504199	295.2457862	200.2971278	138.1375794	97.7999833
Skewness	8.039936477	16.25848274	12.97154116	10.83216103	9.265455604
Range	6810192	22999999	2191347	1731306	580076
Minimum	0	1	0	0	2
Maximum	6810192	23000000	2191347	1731306	580078
Sum	124236226	155058232	13214796	16051143	8383780
Count	758	758	758	758	758
Confidence Level (95.0%)	31528.76302	82629.61529	8253.064292	7669.57791	3188.525638

Figure 9 - Summary Statistics for the Numeric Variables

Outlier Detection

There were 67 instances of outlier data in the 758 instances of the entire dataset, however since there are 5 numeric variables in dataset, scatter plots have been utilized to showcase the outlier data –

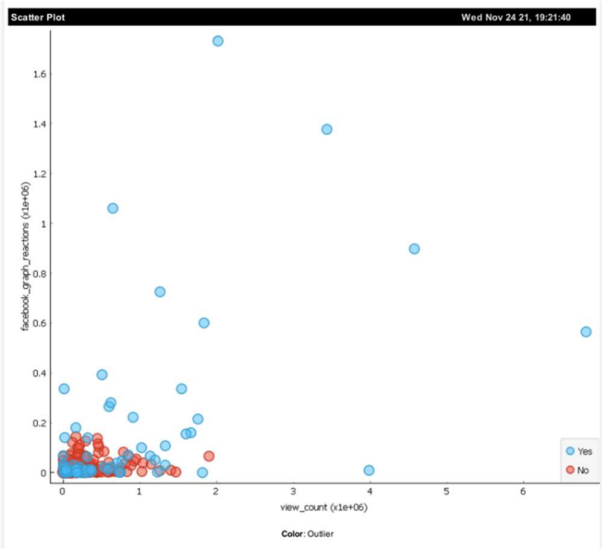


Figure 10 - Scatter Plot showcasing Outliers for variables facebook_graph_reactions vs view_count

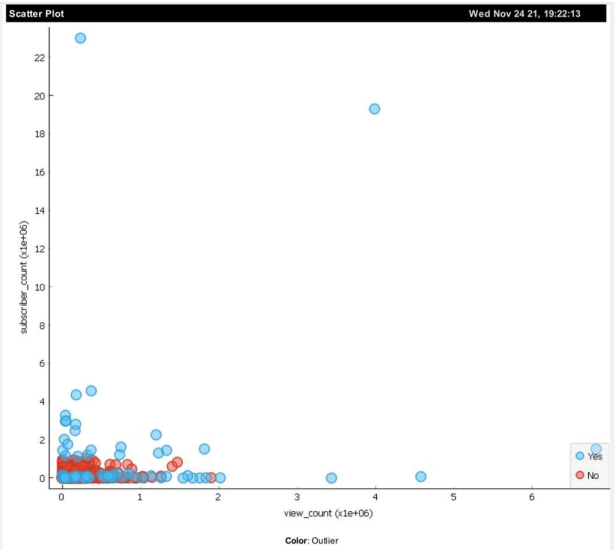


Figure 11 - Scatter Plot showcasing Outliers for variables subscriber_count vs view_count

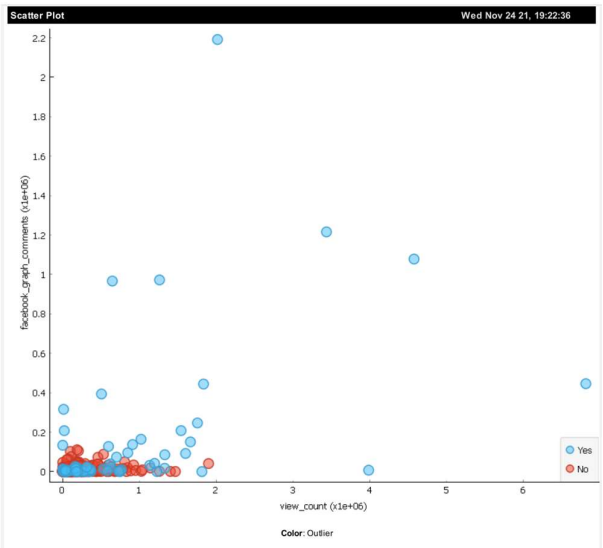


Figure 12 - Scatter Plot showcasing Outliers for variables facebook_graph_comments vs view_count

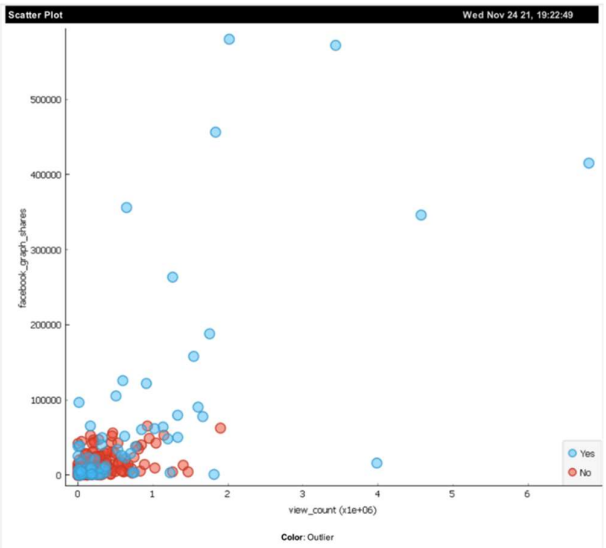


Figure 13 - Scatter Plot showcasing Outliers for variables facebook_graph_shares vs view_count

Statistical Analysis

Since the end goal of this project is to determine the variables that influence "view_count", statistical analysis has been conducted to determine –

1. whether any correlation exists between the target variable "view_count" and the other numeric and categorical variables, and if so, whether they are positive or negative and whether the correlation is strong or weak
2. whether the categorical variable "ContainsKeywords" influence the target variable "view_count"
3. whether there are differences between the population segments based on the categorical variable "ContainsKeywords"

For this purpose, the following statistical techniques have been utilized –

1. Pearson's Correlation Test (Both 1 tailed and 2 tailed)
2. Independent Sample T-Test or Student's T-Test
3. One Way ANOVA

Pearson's Correlation Test

Pearson's bivariate correlation coefficient, r, is used to determine the strength and direction of linear correlations between pairs of continuous variables. By extension, the Pearson Correlation test determines if there is statistical evidence for a linear relationship between identical pairs of variables in a population, as measured by a population correlation coefficient ("rho"). Pearson Correlation is a parametric correlation coefficient.

For this correlation test, both one tailed significance and 2 tailed significance was tested for the variables. In the one tailed significance testing, the null (H_0) and alternate (H_1) hypothesis were the following for each pair of variables –

H_0: $\rho = 0$ ("the population correlation coefficient is 0; there is no association")
H_1: $\rho > 0$ ("the population correlation coefficient is greater than 0; a positive correlation could exist")
OR
H_1: $\rho < 0$ ("the population correlation coefficient is less than 0; a negative correlation could exist")

Correlations

		view_count	subscriber_count	facebook_graph_reactions	facebook_graph_comments	facebook_graph_shares
view_count	Pearson Correlation	1	.253**	.573**	.517**	.697**
	Sig. (1-tailed)		.000	.000	.000	.000
	N	758	758	758	758	758

Figure 14 - Pearson's Correlation Test (One-Tailed)

Through the one tailed significant testing, it can be seen that the variable view_count has significant correlation at the 0.01 significance level with the variables "subscriber_count, facebook_graph_comments, facebook_graph_shares, facebook_graph_reactions"; with the correlation being weak with the variable "subscriber_count" and the correlation being strong for the rest of the numeric variables.

In the two tailed significance testing, the null (H_0) and alternate (H_1) hypothesis were the following for each pair of variables –

H_0: $\rho = 0$ ("the population correlation coefficient is 0; there is no association")
H_1: $\rho \neq 0$ ("the population correlation coefficient is not 0; a nonzero correlation could exist")

Correlations

		view_count	subscriber_count	facebook_graph_reactions	facebook_graph_comments	facebook_graph_shares
view_count	Pearson Correlation	1	.253**	.573**	.517**	.697**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	758	758	758	758	758

Figure 15 - Pearson's Correlation Test (Two-Tailed)

In the two tailed significance testing, the results were the same as that of the one tailed significance testing, indicating that the correlation between the target variable "view_count" and the other numeric variables were significant at the 0.01 significance level, with the correlation being weak between "view_count" and "subscriber_count"; and the correlation being strong between "view_count" and other numeric variables.

Independent Samples T-Test or Student's T-Test

The Independent Samples T-Test examines the means of two independent groups to ascertain whether statistical evidence exists that the associated population means are statistically substantially different. The t Test for Independent Samples is a parametric test.

For the purpose of this project, the independent sample t-test was conducted with the target variable "view_count" as the test variable and the categorical variable "ContainsKeywords" as the grouping variable. Thus, the null (H_0) and alternate (H_1) hypothesis were the following for this pair of variables –

H_0: $\mu_1 = \mu_2$ ("the two population means are equal")
H_1: $\mu_1 \neq \mu_2$ ("the two population means are not equal")

where μ_1 and μ_2 are the population means for group 1 representing the videos with the value of the categorical variable "ContainsKeywords" = 0, and group 2 representing the videos with the value of the categorical variable "ContainsKeywords" = 1, respectively.

At the same time, since the independent sample t-test requires the assumption that both the groups being compared have the same variance, we also ran Levene's Test for Equality of Variance, where the null (H_0) and alternate (H_1) hypothesis were the following –

H_0: $\sigma_1^2 - \sigma_2^2 = 0$ ("the population variances of group 1 and 2 are equal")
H_1: $\sigma_1^2 - \sigma_2^2 \neq 0$ ("the population variances of group 1 and 2 are not equal")

Group Statistics

	Contains Keywords - 1 = True, 0 = False	N	Mean	Std. Deviation	Std. Error Mean
view_count	0	359	168202.16	387655.696	20459.685
	1	399	160029.20	486498.267	24355.377

Figure 16 - Group Statistics for the Independent Sample T-Test

From Levene's Test for Equality of Variance, it can be seen that, since the significance value of Levene's Test is extremely close to 1, we can conclude that we cannot reject the null hypothesis and thus use the values in the row "Equal variances assumed" to assess the results of the independent sample t-test.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
view_count	Equal variances assumed	.018	.893	.254	756	.800	8172.969	32186.071	-55011.728	71357.666
	Equal variances not assumed			.257	745.302	.797	8172.969	31808.538	-54272.027	70617.965

Figure 17 - Independent Sample T-Test between the test variable "view_count" and the grouping variable "ContainsKeywords"

In the independent sample t-test, it can be seen that, since the significance value of the t-test (0.800) is larger than the chosen confidence value (0.05), we can say that the null hypothesis cannot be rejected and thus the two categorical groups of videos doesn't have any significant difference in their means.

One-Way ANOVA (Analysis of Variance)

While, both the independent sample t-test and one-way ANOVA can both the compare of means between two categorical groups of population, one way ANOVA is typically used to conduct the comparison for three or more groups from the categorical variable in concern. Therefore, the purpose of conducting one-way ANOVA in this context is mainly to validate the results of the independent samples t-test. As a result, the one-way ANOVA has been conducted with the target variable "view_count" as the dependent variable and the categorical variable "ContainsKeywords" as the factor.

For the one-way ANOVA, the null (H_0) and alternate (H_1) hypothesis were the following –

H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$ ("all k population means are equal")
H_1: At least one μ_i different ("at least one of the k population means is not equal to the others")

Here μ_i is the population mean of the i^{th} group (i = 1, 2, ..., k)

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
view_count	Between Groups	1.262E+10	1	1.262E+10	.064	.800
	Within Groups	1.480E+14	756	1.958E+11		
	Total	1.480E+14	757			

Figure 18 - One-Way ANOVA

From the results of the one-way ANOVA test, it can be seen that since the significance value for the ANOVA is greater than the chosen confidence level of 0.05, we cannot reject the null hypothesis and thus there are no statistically significant difference between the two categorical groups of videos in terms of means.

From the independent sample t-test and the one-way ANOVA test, it can be seen that the categorical variable "ContainsKeywords" play no significant role in influencing the values of the variable "view_count".

Supervised Machine Learning

After conducting exploratory analysis to determine the variables that may influence the value of the target variable "view_count", supervised machine learning has been conducted to create a machine learning model that can predict the viewer count of any video given that there is sufficient information regarding the other variables of the video, such as the subscriber count of the channel publishing the video.

Predicted Variable – "view_count" (numeric) - the number of views each video received

Predictor Variable –

1. facebook_graph_reactions (numeric) - the number of reacts each video got when shared on Facebook
2. facebook_graph_comments (numeric) - the number of comments each video got when shared on Facebook
3. facebook_graph_shares (numeric) - the number of times each video was shared on Facebook
4. subscriber_count (numeric) - the number of subscribers each channel publishing the video maintained
5. Contains Keywords (categorical) - Whether the title or description of the video contains specific keywords that are assumed to be related to the current pandemic of COVID 19. For this variable, the value 1 indicates that the video title or description contains any one of those keywords, whereas the value 0 indicates that there are none.

Ignored Variables –

1. published_timestamp
2. removal_timestamp
3. Keyword Check – Hoax
4. Keyword Check – Truth
5. Keyword Check – Fraud
6. Keyword Check – Fake
7. Keyword Check – Scam
8. Keyword Check – Coronavirus
9. Keyword Check – COVID
10. Keyword Check – Cure
11. Keyword Check – Anthony Fauci
12. Keyword Check – Bill Gates
13. Keyword Check – Alex Jones
14. Keyword Check – Judy Mikovits

Using Machine Learning Models to Predict "view_count"

For supervised machine learning, my primary target for prediction is a numeric variable, namely "view_count". Therefore, in my supervised machine learning endeavor, I have mainly used 6 models –

1. Linear Regression
2. Tree Model
3. Random Forest Model
4. kNN Model
5. Neural Network Model
6. Stacking Model

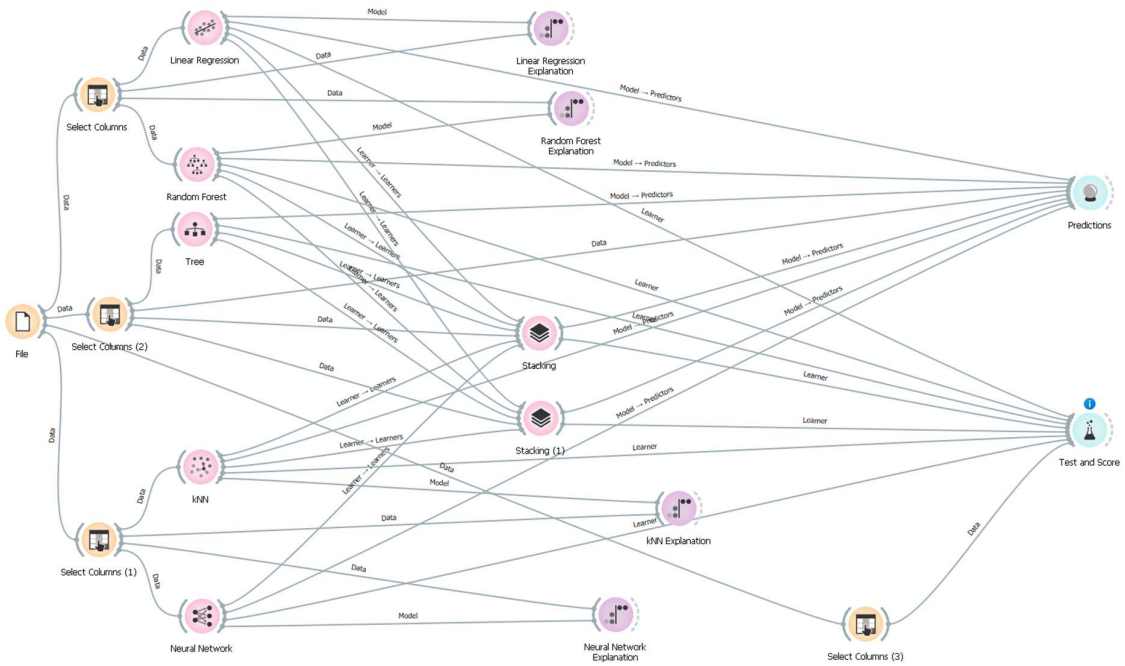


Figure 19 - Machine Learning Model Building Workflow

1. kNN Model

- For 5 neighbors, RMSE = 380175.913 and R^2 = 0.260
- For 10 neighbors, RMSE = 386515.420 and R^2 = 0.235
- For 20 neighbors, RMSE = 384124.578 and R^2 = 0.244
- For 40 neighbors, RMSE = 392089.216 and R^2 = 0.213
- For 50 neighbors, RMSE = 396099.667 and R^2 = 0.197
- Since, the kNN model with 5 neighbors has the least value of RMSE and the highest value of R^2, therefore, this specific kNN model will be used to compare with other machine learning models.

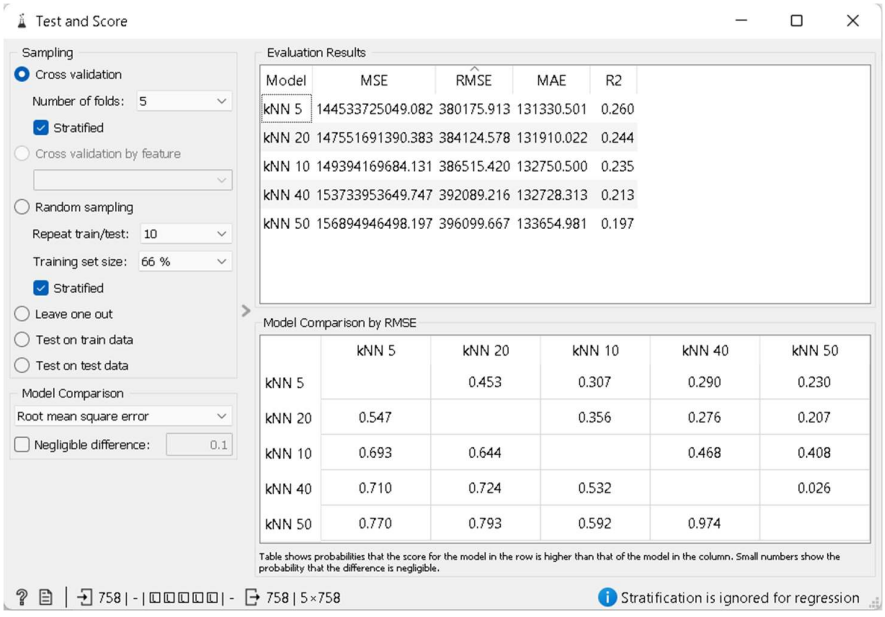


Figure 20 - kNN Model Comparison for Different number of Neighbors

2. Linear Regression

- In this model, with no regularization, RMSE = 356699.245 and $R^2 = 0.348$
- Since the value of R^2 is still low for this model, therefore there is a need to check whether there are any unnecessary columns in the predictor variables
- With Elastic Net Regularization, the value of R^2 is virtually unchanged and even the value of RMSE doesn't experience any significant changes, thus indicating that there are no unnecessary columns.

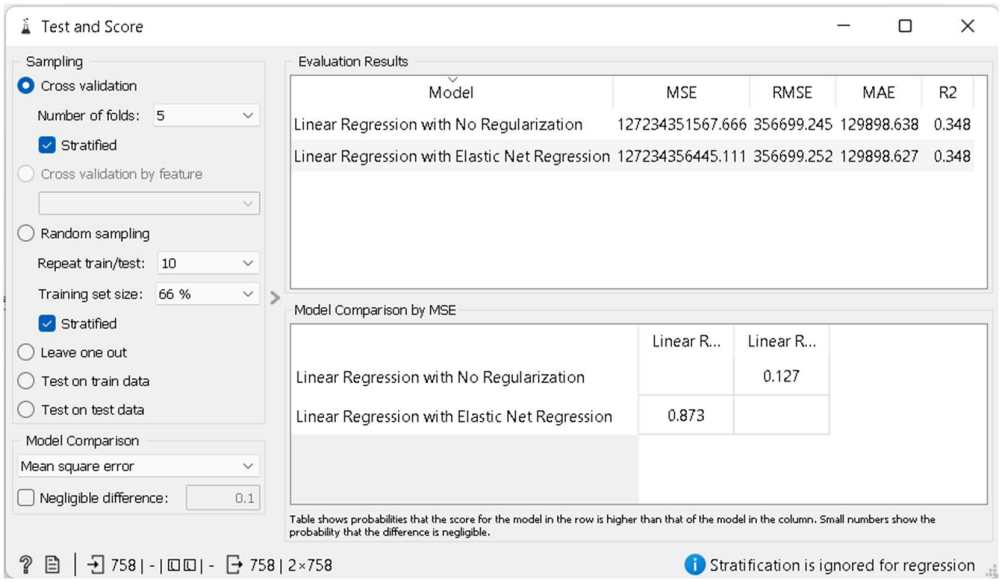


Figure 21 - Comparison Between the Linear Regression Model with No Regularization and Linear Regression Model with Elastic Net Regression

- The probability of the value of RMSE for kNN being greater than the value of RMSE for linear regression is 59.4% with both no regularization and elastic net regression.
- As expected, the linear regression model is statistically better for predicting "view_count" than the kNN model.

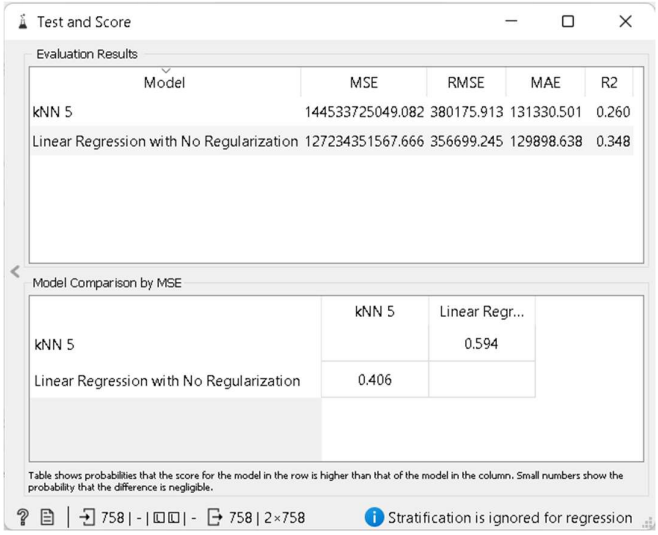


Figure 22 - Comparison Between kNN and Linear Regression with No Regularization

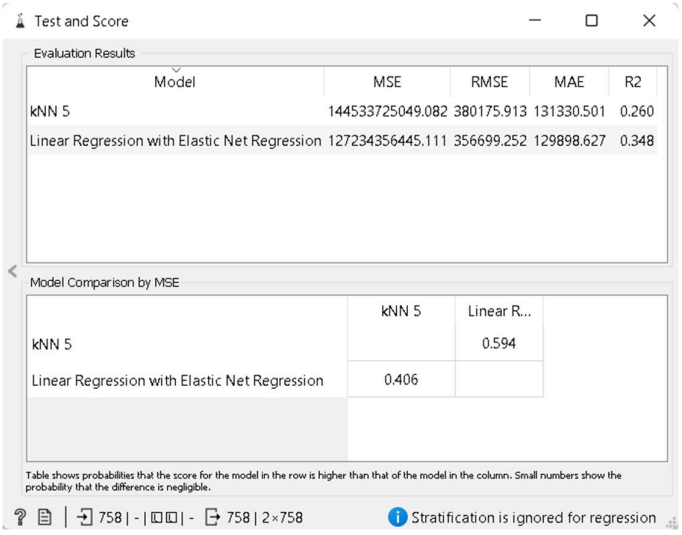


Figure 23 - Comparison Between kNN and Linear Regression with Elastic Net Regression

3. Tree Model

- For this model, RMSE = 413068.573 and R^2 = 0.126
- Since the value of RMSE is the largest in this model and at the same time the value of R^2 is the smallest in this model as well, thus this model is the worst among the three models discussed so far in predicting the variable – "view_count".

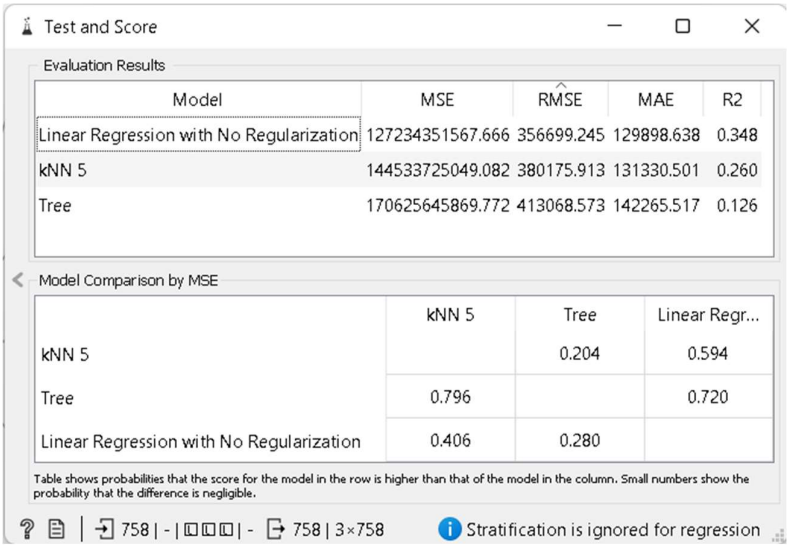


Figure 24 - Comparison Between the Tree, kNN and Linear Regression Model

4. Random Forest Model

- For 10 trees, RMSE = 369770.034 and R^2 = 0.300
- For 50 trees, RMSE = 367346.965 and R^2 = 0.309
- For 100 trees, RMSE = 353353.085 and R^2 = 0.361
- For 200 trees, RMSE = 363675.934 and R^2 = 0.323
- For 500 trees, RMSE = 362464.725 and R^2 = 0.327
- Since among the variants tested here, the model with 100 trees has both the highest value of R^2 and the lowest value of RMSE, this variant of the model will be used to compare with other models utilized so far.

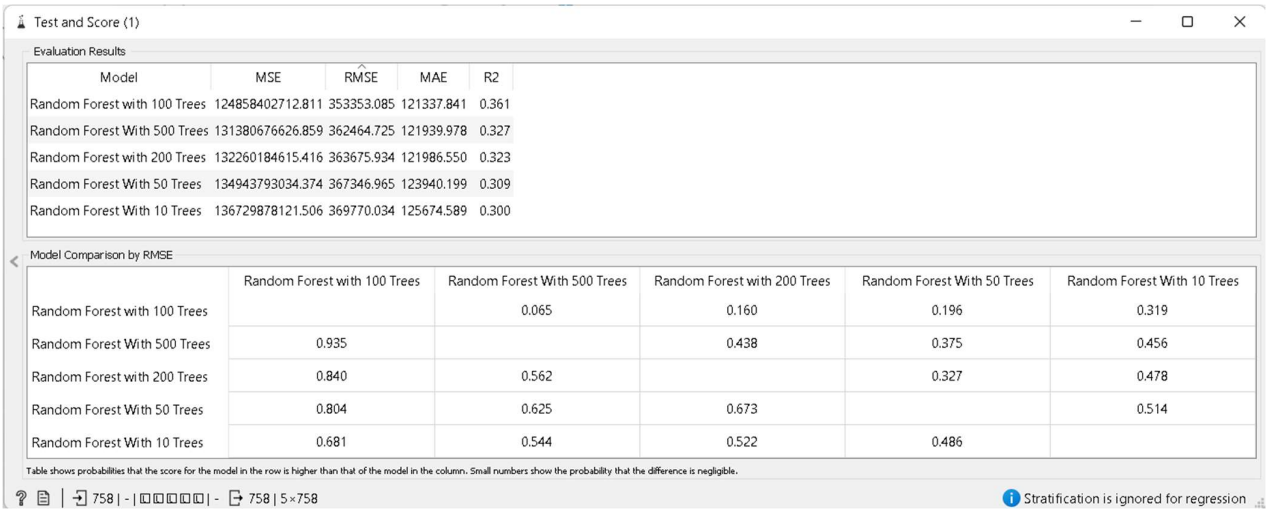


Figure 25 - Comparison between different variants of the Random Forest Model

- Among the four models discussed so far, the random forest model with 100 trees is actually better than both kNN with 5 neighbors and the Tree model in predicting "view_count"

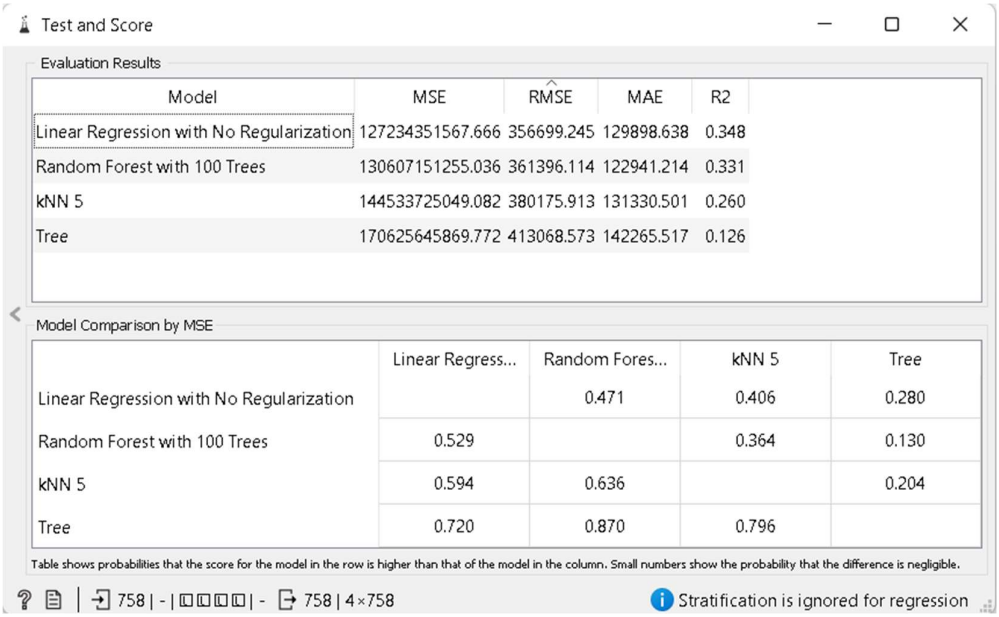


Figure 26 - Comparison Between the kNN, Tree, Random Forest & Linear Regression Model

5. Neural Network Model

- For this model, RMSE = 471222.127 and R^2 = -0.137
- Considering the fact that not only this model has the highest value of RMSE among the models considered so far, but also carries a negative value of R^2, this model is completely unsuitable for predicting "view_count".

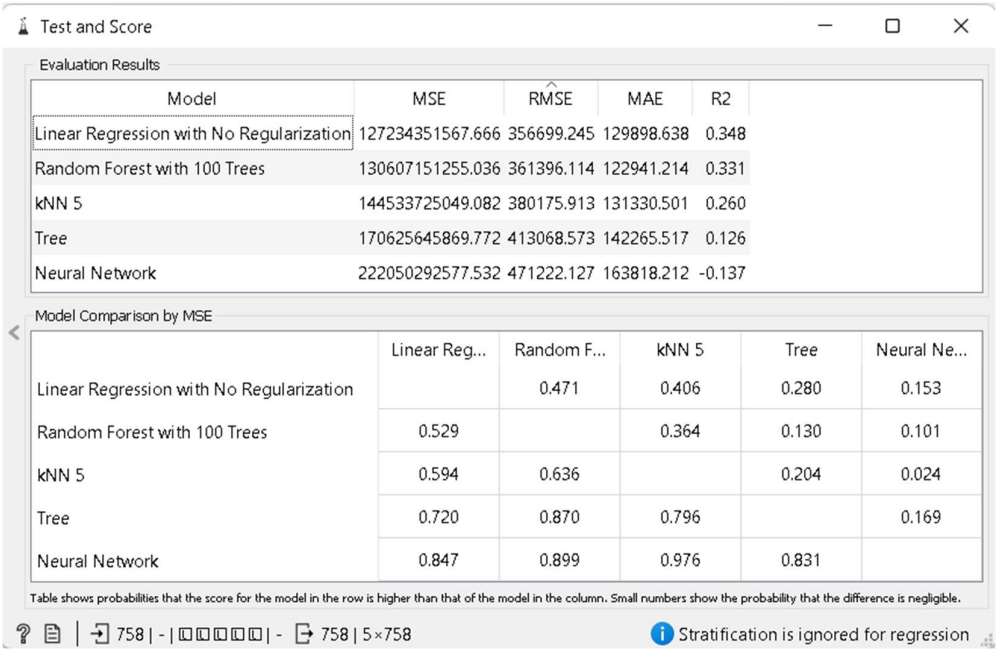
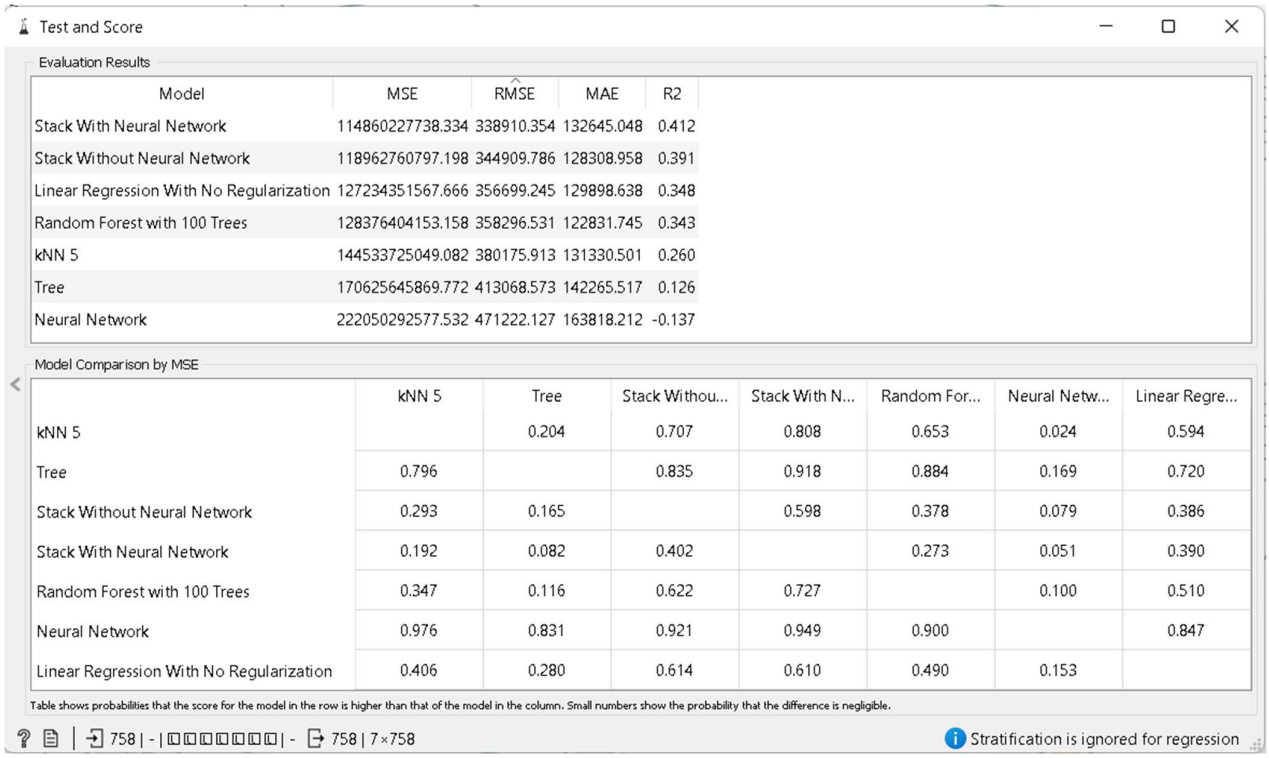


Figure 27 - Comparison between the kNN, Tree, Random Forest, Linear Regression and Neural Network Models

6. Stacking Model

- By stacking different models, both the value of RMSE and the value of R^2 improved with RMSE being 338910.354 and R^2 being 0.412.
- However, since the neural network was determined to be completely unreliable in predicting “view_count”, another stacking without the neural network was conducted and the result was that both the value of RMSE and the value of R^2 degraded by a relatively significant margin. This indicates that, while neural network by itself is invalid in this scenario; using it while stacking can lead to increased accuracy in predicting “view_count”.


Test and Score

Evaluation Results

Model	MSE	RMSE	MAE	R2
Stack With Neural Network	114860227738.334	338910.354	132645.048	0.412
Stack Without Neural Network	118962760797.198	344909.786	128308.958	0.391
Linear Regression With No Regularization	127234351567.666	356699.245	129898.638	0.348
Random Forest with 100 Trees	128376404153.158	358296.531	122831.745	0.343
kNN 5	144533725049.082	380175.913	131330.501	0.260
Tree	170625645869.772	413068.573	142265.517	0.126
Neural Network	222050292577.532	471222.127	163818.212	-0.137

Model Comparison by MSE

	kNN 5	Tree	Stack Withou...	Stack With N...	Random For...	Neural Netw...	Linear Regre...
kNN 5		0.204	0.707	0.808	0.653	0.024	0.594
Tree	0.796		0.835	0.918	0.884	0.169	0.720
Stack Without Neural Network	0.293	0.165		0.598	0.378	0.079	0.386
Stack With Neural Network	0.192	0.082	0.402		0.273	0.051	0.390
Random Forest with 100 Trees	0.347	0.116	0.622	0.727		0.100	0.510
Neural Network	0.976	0.831	0.921	0.949	0.900		0.847
Linear Regression With No Regularization	0.406	0.280	0.614	0.610	0.490	0.153	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

758 | - | - 758 | 7×758

Stratification is ignored for regression

Figure 28 - Comparison between all the models utilized in the prediction of "view_count"

Conclusion

Among the models utilized, the stacking model is the most accurate in terms of predicting “view_count”. However, none of the models have neither a high enough value of R^2, nor a relatively low enough value of RMSE. Therefore, the models used are not very good predictors of the variable – “view_count”.

Determining variables that impact the prediction of "view_count" in the models

Across all the models utilized in the prediction of "view_count", the common variable that had the most significant impact in predicting "view_count" was "facebook_graph_shares". However, the variable "facebook_graph_reactions" had the most negative impact in predicting "view_count" in the Linear Regression model. Other than that, all the predictor variable had positive impact in predicting "view_count". The level of impact, however, varies across the model among the predictors. Figures representing the level of impact of variable in the prediction for each model except the stacking model are given below –

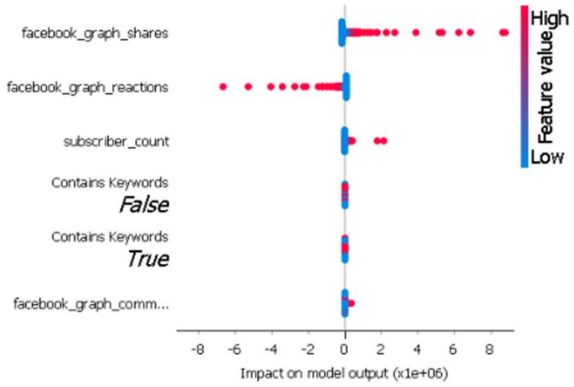


Figure 29 - Variable impact in Prediction of view_count for the linear regression model

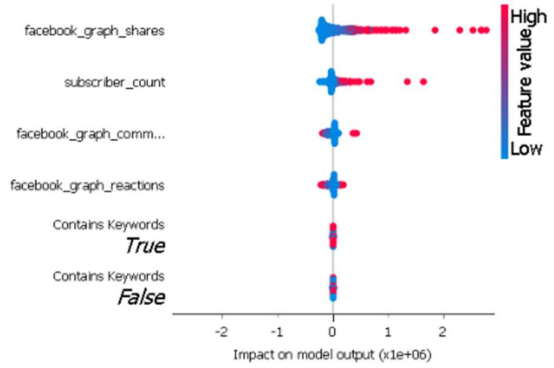


Figure 30 - Variable impact in prediction of view_count for the Random Forest Model

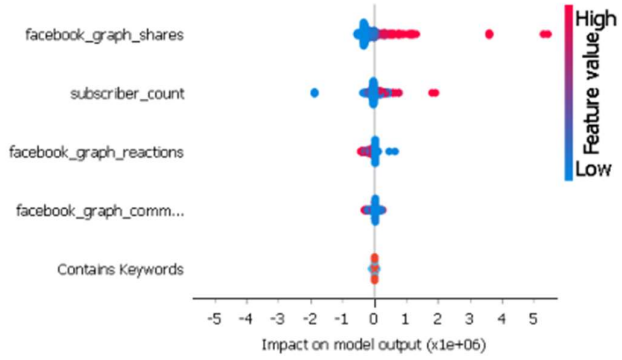


Figure 31 - Variable impact in Prediction of view_count for the Tree Model

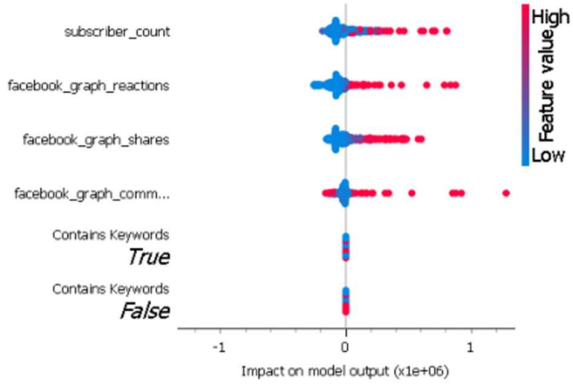


Figure 32 - Variable impact in Prediction of view_count for the kNN Model

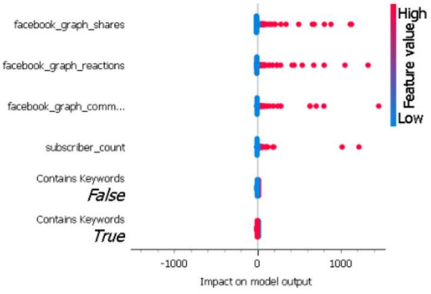


Figure 33 - Variable Impact in Prediction of view_count for the Neural Network Model

Constraints & Model Fallbacks

Trying to Predict Whether a Video Was Fake or Not

One initial intention I had while working with this dataset was to predict whether a video was fake or not based on other predictor variables present in the dataset. But the prime obstacle to fulfilling that intention was a simple fact that all the videos represented in the dataset was fake and focused on presenting misinformation. This meant that trying to predict whether a video was fake or not using this dataset through machine learning would be impossible; since a model trained by this dataset would always predict every video tested to be fake. As a result, using this specific dataset to predict whether a video was fake or real would be quite impossible.

However, as long as this dataset is supplemented by data on videos presenting real information about COVID – 19 that have been factchecked by relevant experts and doctors in the field, building a model to predict whether a video about a specific topic was fake or not may be actually possible.

Utilizing Neural Network for Prediction

While trying to utilize supervised machine learning models to predict the viewership of a video based on other specific metrics, one model – specifically the neural network model – had given out a result of an R-squared value of – 0.137, which was quite abnormal when it comes to prediction. This is because of the fact that, when Orange ML, the software utilized for supervised machine learning, implements neural network model on a dataset for prediction, it preprocesses the data in the following manner –

1. *"removes instances with unknown target values"*
2. *"continuizes categorical variables (with one-hot-encoding)"*
3. *"removes empty columns"*
4. *"imputes missing values with mean values"*
5. *"normalizes the data by centering to mean and scaling to standard deviation of 1"*

Therefore, that abnormal result was concerning to a certain extent. Therefore, in order to find a workaround for this issue, a manual preprocessing was carried out on the data by normalizing the features and continuizing the sole discrete variable in the dataset.

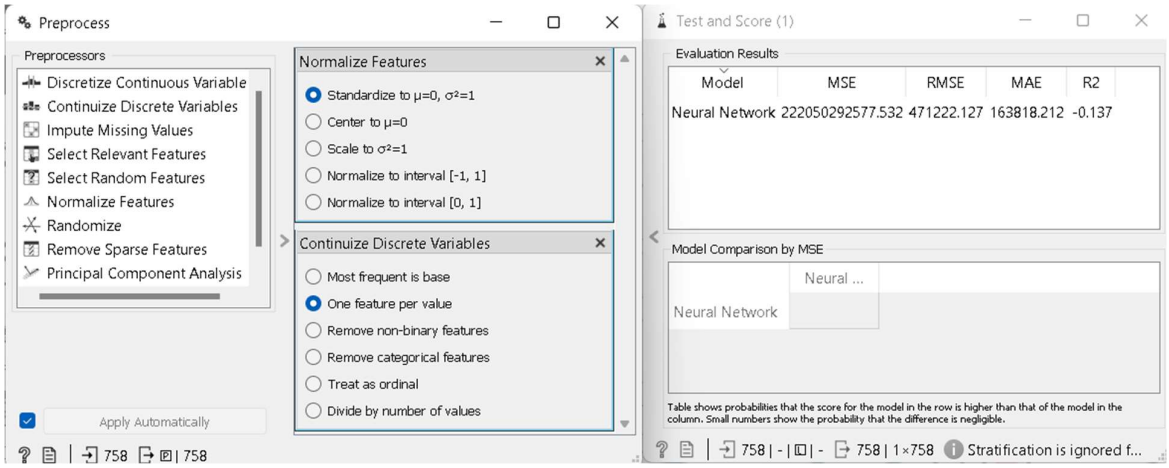


Figure 34 - Attempt 1 of Workaround

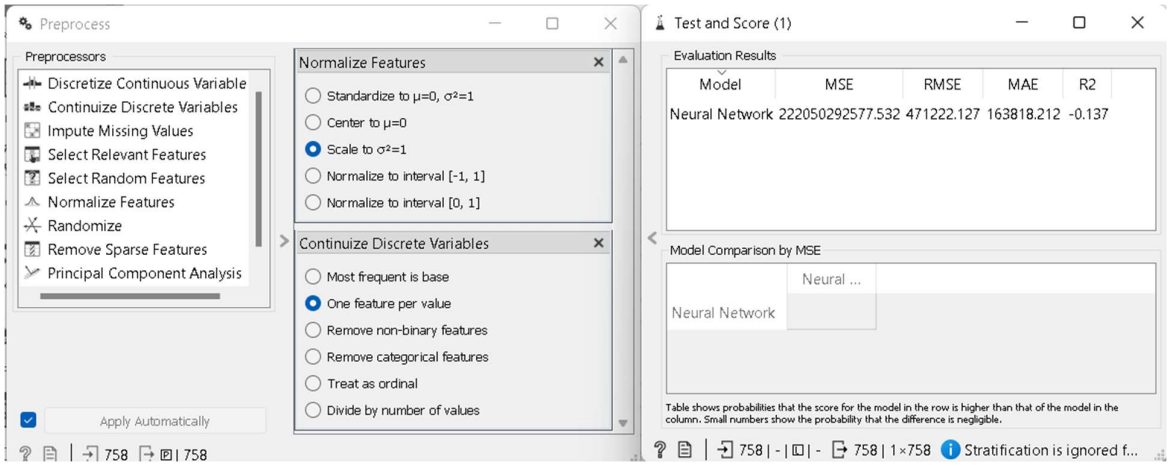


Figure 35 - Attempt 2 of Workaround

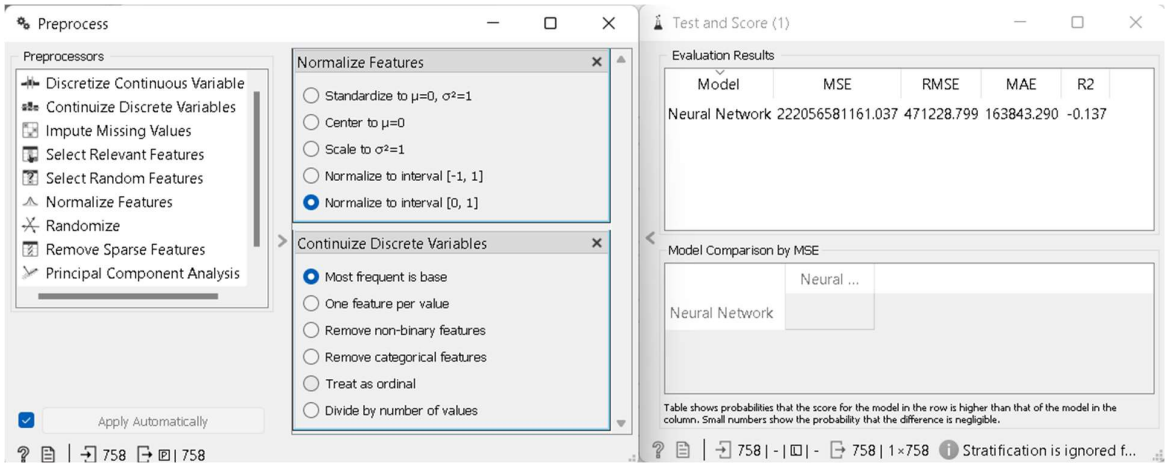


Figure 36 - Attempt 3 of Workaround

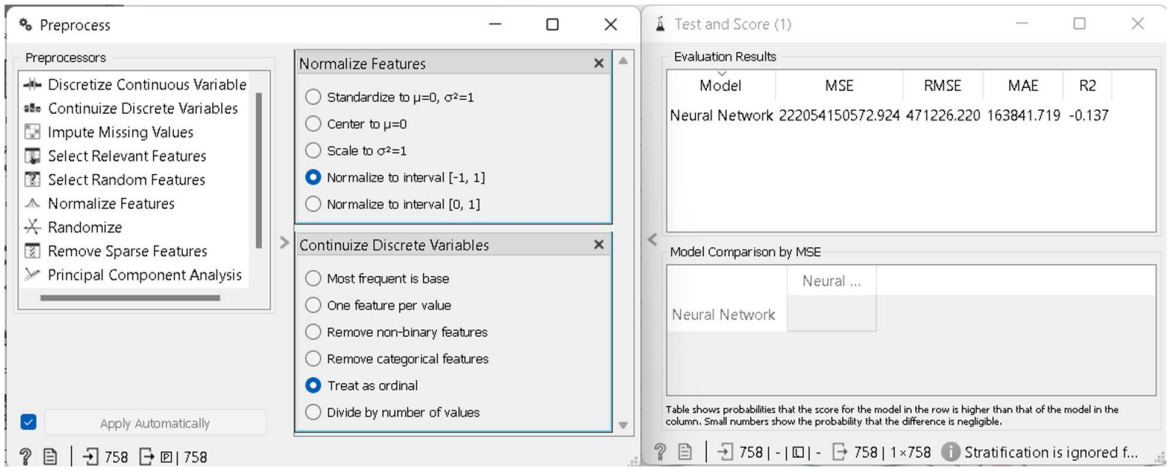


Figure 37 - Attempt 4 of Workaround

No matter the choices made, there was no improvement in the value of R-squared of the model, thus leading me to the conclusion that neural network was quite unsuitable for making predictions on this dataset.