TRINITY COLLEGE DUBLIN

REPORT

ST3002 Final Project 2019/2020

BY

BABATUNJI OMONIWA

18343546

LECTURER

PROF. SUSAN CONNOLY

Purpose of Project:

As a consultant to the online news portal, I am mandated to solve two problems:

P1. To come up with a way to predict the number of shares an article will receive using some subset of the attributes provided in the dataset provided.

P2. To investigate whether the articles that are published in the World section on Thursdays fall into defined groups.

Method

To address the problems P1 and P2 above, we carry out two methods M1 and M2, respectively.

M1: We provide a summary of our methodology for predicting the number of shares in Fig. 1. We carried out data cleaning on the raw data. We categorize our To predict an outcome (target) variable based on the values of other "predictor" variables, we split the cleaned data into training and test (80%, 20%). We applied three models; a baseline, a linear regression model; and a decision tree model.

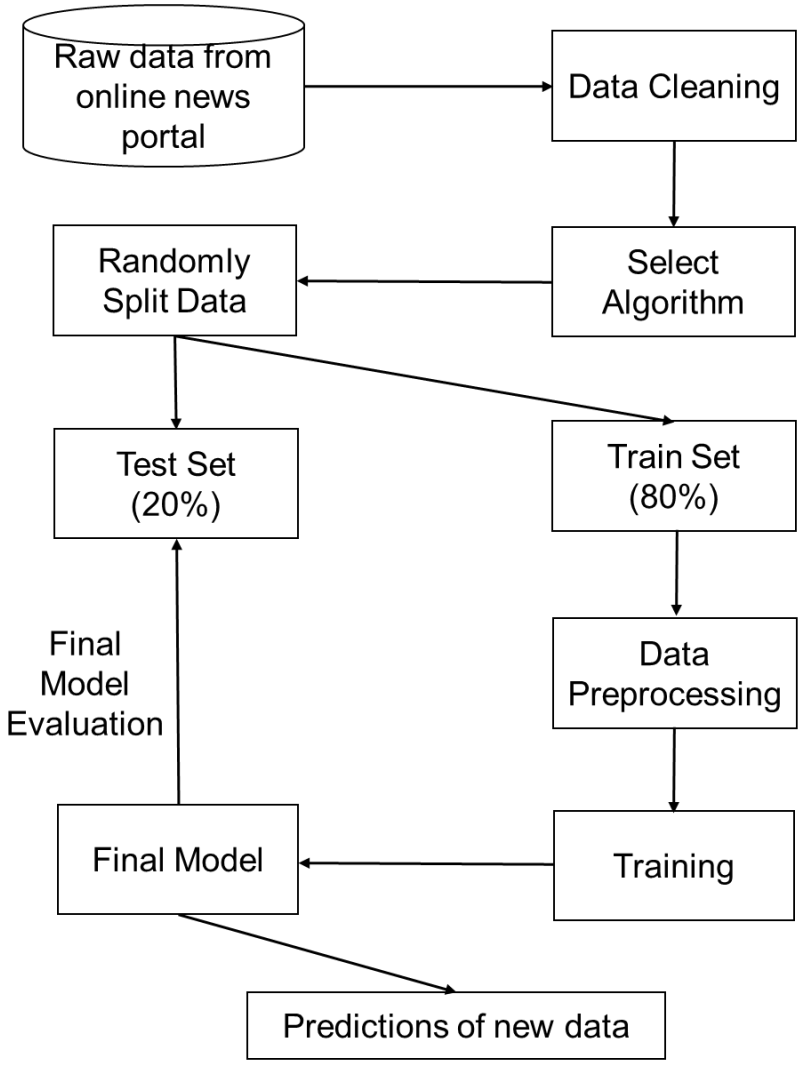

Figure 1: *Method for predicting the number of shares.*

M2: We provide a summary of our methodology for knowing whether the articles that are published in the World section on Thursdays fall into defined groups in Fig. 2. The raw data was cleaned. We then transformed the binary data to categorical data of one variable. On this note, we then applied a univariate clustering algorithm.

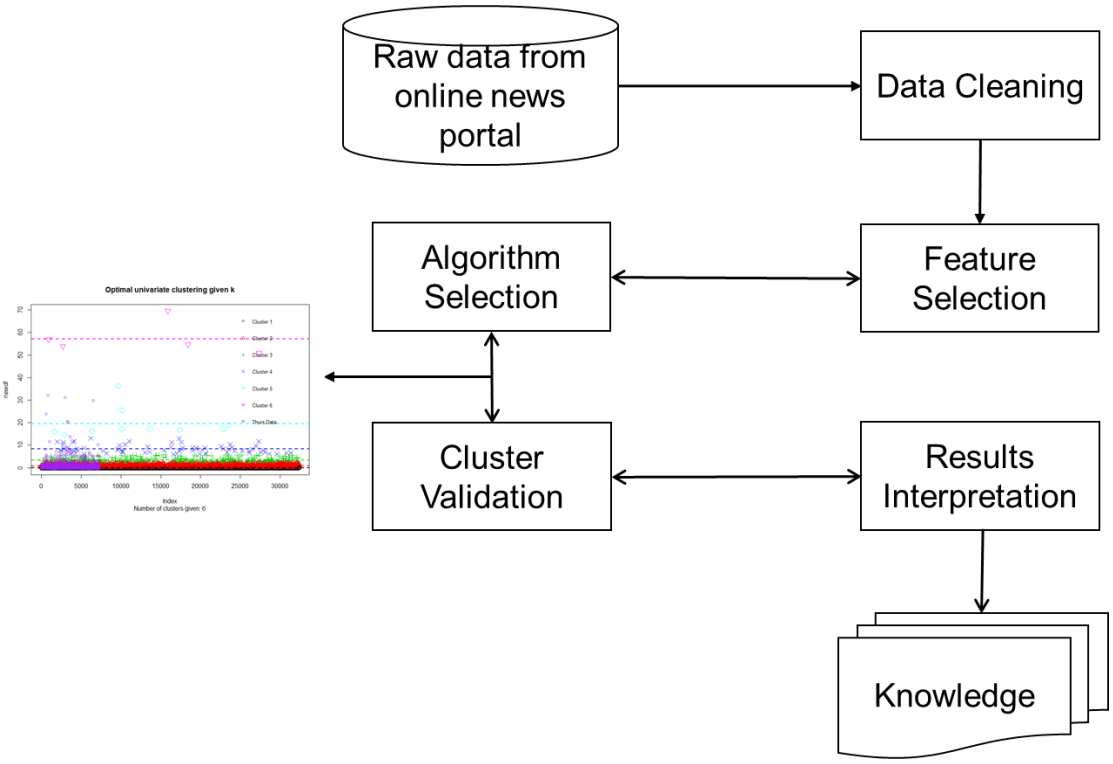

Figure 2: *Method for clustering.*

Evaluation

In this section, we present the evaluation metrics, and the baseline used to evaluate our models.

Evaluation Metrics: For the continuous shares outcome, the main error metrics we will use to evaluate our models are the mean absolute error (MAE) and root mean squared error (RMSE), which are two of the most commonly used metrics for measuring the accuracy for continuous variables. In this report, we evaluate the accuracy of our model using both the MAE and RMSE. We compare the performance the model against a baseline.

Mean Absolute Error (MAE): MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight [1].

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

Root Mean Squared Error (RMSE): RMSE is a quadratic scoring rule that also measures the average magnitude of the error. It's the square root of the average of squared differences between prediction and actual observation [1].

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Models used: We consider comparing the linear regression and decision tree models against a baseline. In the absence of any predictor, all we have is the dependent variable (shares). What would be our best guess if we had to predict the shares? I would say that the mean of the shares values, in the training data, is the best value we can come up with. Recall we can only use the training data to build models; the testing data is only there to evaluate them, after making the predictions. Since we are dealing with continuous outcome data, we employ the use of linear (multiple) regression and decision tree models since they are known to offer high interpretability and decent accuracy.

The linear (multiple) regression is probably the best known statistical model. In specific terms, the dependent variable is assumed to be a linear function of several independent variables (predictors), where each of them has a weight (regression coefficient) that is expected to be statistically significant in the final model. On the other hand, the decision tree (also known as regression tree for continuous outcome variables) is a simple and popular machine learning algorithm, with a few interesting advantages over linear models: they make no assumptions about the relation between the outcome and predictors (i.e., they allow for linear and non-linear relations) [2].

Results and Discussion

In this section, we present results from the experiments carried out on the given dataset. We observe that the accuracy improved when compared to the decision tree model, and is just about close to the performance of the linear regression model. We have succeeded in evaluating the accuracy of three different models: baseline, linear regression, and decision tree. From Table I, we show that all methods beat the baseline, regardless of the error metric, with the linear regression offering the best performance.

Table I: *Results showing the accuracy of our model.*

Method	*Mean absolute error (MAE)*	*Root Mean Square Error (RMSE)*
Baseline – Mean guess	3238.6928	11687.3448
Linear regression model	3125.4790	11614.6538
Decision tree model	3195.5141	11771.1820

Fig. 3 show the number of shares according to the days of the week. From the figure, it is difficult to know whether the articles that are published in the World section on Thursdays fall into defined groups. As such, we carry out a univariate k-means clustering algorithm as shown in Fig. 4.

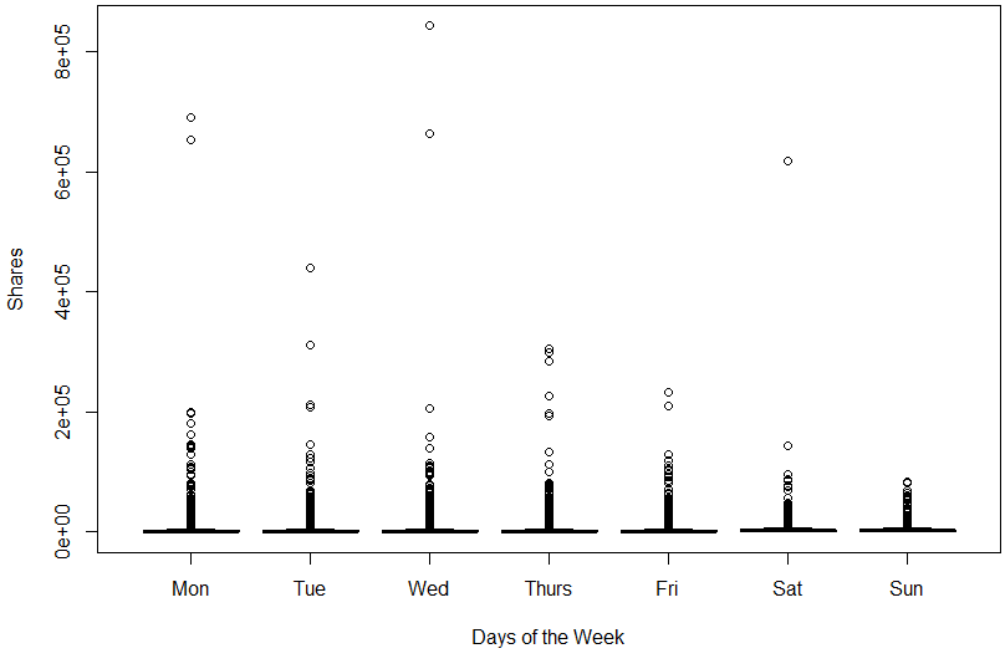

Figure 3: *Number of shares according to days of the week.*

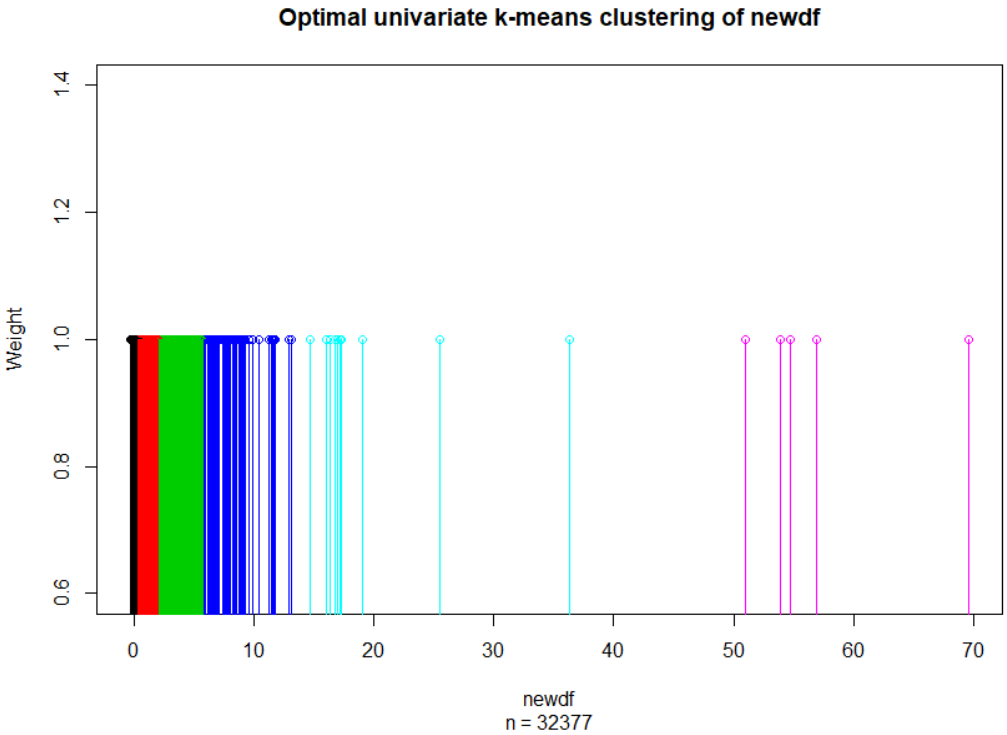

Figure 4: *Optimal univariate k-means clustering of the data (k=6).*

From Fig. 5, we can see that the articles that are published in the World section on Thursdays overlap more with that in cluster 1 and 2, thereby implying they fall into defined groups.

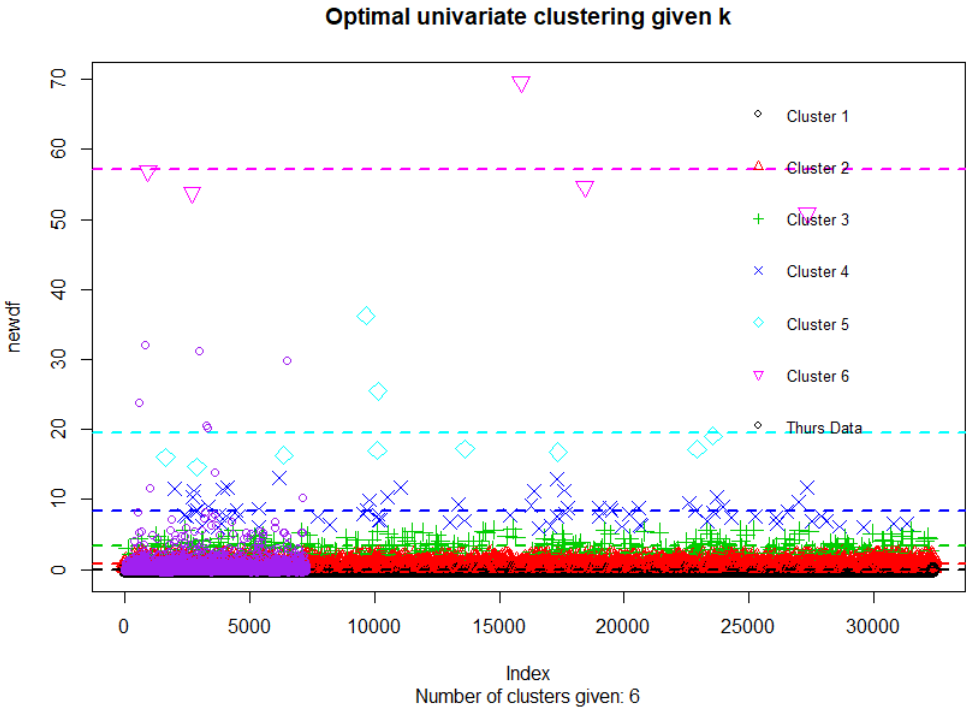

Figure 5: *Optimal univariate clustering given K.*

Conclusion

In this report, we provide accurate prediction of the number of shares an article will receive using some selected attributes provided in the dataset. We perform data cleaning, split the data into training and test set. We then performed training using different models and evaluated the accuracy of the models using the MAE and RMSE. The models were compared against a baseline to validate the accuracy. Furthermore, to investigate whether the articles that are published in the World section on Thursdays fall into defined groups, we adopted the use of a univariate clustering algorithm. We observe that the articles published on Thursdays relate more to clusters 1 and 2 of the data, implying that they belong to defined groups.

References

[1] https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d

[2] https://www.r-bloggers.com/part-4a-modelling-predicting-the-amount-of-rain/