**Modeling:**

The goal of the analysis is to determine whether the news article is classified as good or bad related to its popularity. Here, we used the median value of the shares to determine whether the shares are classified as good or bad. The reason that we haven’t used the mean value for the classification is it produced the class imbalance. So, we have used classification algorithms like KNN, Logistic Regression, Decision Trees, Random Forest, and SVC. While cleaning the data we removed the columns which has high multi-collinearity and the columns which are unnecessary for the modeling.

**KNN ALGORITHM:**

The k-nearest neighbor’s method, generally known as KNN or k-NN, is a non-parametric, supervised learning classifier that utilizes proximity to classify or predict the grouping of a single data point.

For the KNN algorithm, we have plotted the error rate vs accuracy plot, based on the plot we can identify that the error rate was low when the k value is at 17. So, by looking at that plot we found that the optimum value of the k is 17.

**Error rate vs k value**

Chart, line chart

Description automatically generated

So, for the KNN algorithm, we choose the number of neighbors as 17 and the accuracy of the model with this K value was 63%. The confusion matrix and the classification report for the KNN model are shown below:

From the confusion matrix, we can interpret that 2557 are classified as false negative and 1662 are classified as false positive.

Chart, treemap chart

Description automatically generated

Table

Description automatically generated

**Logistic Regression:**

When the dependent variable is dichotomous, logistic regression is the proper regression strategy to use (binary). While implementing the logistic regression model, the accuracy of the model is 65%. The confusion matrix and the classification report are shown below.

From the confusion matrix, we get 2316 as a false negative and 2105 as a false positive.

Chart, treemap chart

Description automatically generated

Table

Description automatically generated

The AUC for the logistic model is 0.7, from the AUC we can say that 70% chance that the model will be able to distinguish between positive class and negative class, and the logistic model is said to be a considerable model.

Chart, line chart

Description automatically generated

**Decision Tree:**

We have implemented the decision tree for the given data set with a maximum depth of 5 first, we considered the decision tree with the default parameters. The accuracy of the model is 62%. So, then we adjusted some of the parameters in the decision tree function and tried different depths. We considered the decision tree with maximum depth at 5 is considered as optimum with an accuracy of 64%. The classification and confusion matrix for the decision tree algorithm

is shown below

Table

Description automatically generated

From the confusion matrix, we can analyze that 2107 rows are classified as false negative and 2159 are classified as false positive.

Chart, treemap chart

Description automatically generated

**Random Forest:**

During training, random forests (also known as random choice forests) generate a huge number of decision trees to use as an ensemble learning approach for classification, regression, and other problems. The output of a random forest is the class selected by the vast majority of trees, which is useful for solving classification issues. When a regression task is given, the average prediction of the individual trees is given back. Decision trees may overfit their training data, although random decision forests mitigate this problem. Random forests are more effective than decision trees in most cases. So, for the random forest, we just used the default parameters, and the accuracy of the model is 66%. Moreover, we tried random forest with different parameters, but the accuracy of the model is not increased. The confusion matrix and classification report for the random forest is shown below.

From the confusion matrix, we can identify that 2166 are classified as false negative and 1793 are classified as false positive.

Table

Description automatically generated

Chart, treemap chart

Description automatically generated

**Support Vector Machines (SVM):**

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.

An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification, implicitly mapping their inputs into high-dimensional feature spaces. From this model, accuracy is 65%.

**Table

Description automatically generated**

From the confusion matrix, we can identify that 2058 are classified as false negative and 1897 are classified as false positive.

**Chart, treemap chart

Description automatically generated**