Predicting the Popularity of Online News Articles

Authors:

Vinaya Chinti

Madhuri Ghadiyaram

Course Project Professor's Name:

Dr. Liao

Course Name:

Analytics: Big Data to Information

Course Number and Section:

AIT 580

Date:

7/23/2020

Title: Predicting the popularity of online news articles.

Abstract

Predictive analytics uses statistical and machine learning methods to predict future events based on historical events with a certain degree of precision. Almost all the things are offered as online services nowadays, people tend to share and read the new articles online. Advertising agencies and the new article publishers are assessing the popularity of a particular news article by considering the number of shares the article gets. The dataset consists of articles from Mashable which is a famous online news website. This project focuses on predicting the popularity of news articles by considering various features from the dataset. Feature selection and dimensionality reduction are some of the major steps in model building to achieve significant accuracies. This project implements 3 algorithms and they are trained using datasets that are preprocessed using different techniques. The models are evaluated using various accuracy metrics along with the confusion matrix and ROC curve.

Keywords: Predictive Analytics, Online Articles, K-Nearest Neighbour, Random Forest, Regression, Principal Component Analysis.

1. Introduction

The easy availability of the internet has increased the use of online resources. As all the data is available at the fingertip so are the news articles. The news article should have an impact on the large population and for this, the trends and human sentiments need to be analyzed. Data science, machine learning, and data analytics have made it easy to pre-process the raw data and to gain insights into the data. Machine learning algorithms are used for prediction, classification, sentiment analysis, and anomaly detection. Features of particular data play a very important role in building machine learning models. The news articles can be analyzed by using various features such as the length of the article, the day on which it was published, keywords, the media through which it was published which are amongst the few. The advertising agencies can use these factors to target a particular audience and analyze how people react to particular articles or what are their likes and dislikes. These insights can be used to boost the business of the advertising agency as well.

The online news popularity dataset has several features with the number of shares as the target variable. The dataset can be used for prediction as well as classification. The dataset will be used for exploratory analysis which answers many questions related to how various features affect the popularity of news articles. Pre-processing techniques will be applied to the data, the data will be divided into training, testing, and validation datasets. Logistic Regression, Random Forest, and SVM classifiers will be trained and tested on the testing and the validation datasets. The models will be evaluated using various attributes along with the confusion matrix and the ROC-AUC curve.

2. Related Work

Referring to papers related to online news popularity prediction helped us in gaining some information about the features and the algorithms that can be used for prediction. We analyzed various papers to understand the preprocessing techniques that can be applied as the dataset has many attributes. The paper Ren, H., & Yang, Q. [1] explains various attributes selection methods and they have applied 10 different learning algorithms which include various regression algorithms, random forest, and SVM. They have used fisher's criterion and mutual information to rank the important attributes. Amongst all the algorithms, the random forest algorithm performed better. Uddin, T., Hossain, M., Patwary, M., & Ahsan, T. [2] have considered 2 scenarios: prediction of popularity after the article is published and prediction of popularity before the article is published. They have used a Gradient Boosting Algorithm with 5-fold cross-validation and have executed the algorithm in 20 runs with random seed to get the best accuracy.

3. Objectives

Following are the objectives of the project:

1) Analyze the data using various visualization techniques.
2) Pre-process the data to remove the null values and identify the outliers in the data.
3) Identify the target variable and the predictors.
4) Find the correlation between the predictors and the target variable for feature selection.
5) Use Principal Component Analysis for dimensionality reduction.
6) Divide the dataset into training, testing, and validation datasets.
7) Build prediction models to predict the number of shares
8) Perform comparative study amongst all the models built using the accuracy, precision, and recall.

The dataset will be used to answer the following questions by using various visualization techniques:

1) Does the day of the week affect the shares or popularity of a news article? Does this mean people read articles only on particular days like on weekends?
2) Do people tend to read lengthy articles?
3) Does the data channel used for publishing the article affect the number of shares?
4) What is the effect of the inclusion of images, videos, and references to other content on the number of shares?

4. Dataset

4.1 Dataset Selection

The data set is from the UCI repository and is in the form of a CSV file. The dataset is downloaded using the following link

Data source: https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity

4.2 Dataset Description

The dataset consists of various attributes related to the articles that are published by Mashable from 7th January 2013 to 7th January 2015. The data has 61attributes out of which 2 attributes are non-predictive (URL,timedelta), remaining 58 are predictive attributes, and 1 target field(Shares) and has 39644 instances.

4.3 Dataset Schema

The dataset has the following columns:

Aspects	Attributes
Number of Words	n_tokens_title : Number of words in the title n_tokens_content: Number of words in the content n_unique_tokens: Rate of unique words in the content n_non_stop_words: Rate of non-stop words in the content n_non_stop_unique_tokens: Rate of unique non-stop words in the content average_token_length: Average length of the words in the content
Links	num_hrefs: Number of links num_self_hrefs: Number of links to other articles published by Mashable
Media	num_imgs : Number of images num_videos: Number of videos
Published Day	weekday_is_monday: Was the article published on a Monday? weekday_is_tuesday: Was the article published on a Tuesday? weekday_is_wednesday: Was the article published on a Wednesday? weekday_is_thursday: Was the article published on a Thursday? weekday_is_friday: Was the article published on a Friday? weekday_is_saturday: Was the article published on a Saturday? weekday_is_sunday: Was the article published on a Sunday? is_weekend: Was the article published on the weekend?
Number of Keywords	Number of keywords in the metadata, Worst/Best/Average keyword (MIN,MAX,AVG shares),
Article Category	data_channel_is_lifestyle: Is data channel 'Lifestyle'? data_channel_is_entertainment: Is data channel 'Entertainment'? data_channel_is_bus: Is data channel 'Business'? data_channel_is_socmed: Is data channel 'Social Media'? data_channel_is_tech: Is data channel 'Tech'? data_channel_is_world: Is data channel 'World'?

NLP	LDA_00: Closeness to LDA topic 0 LDA_01: Closeness to LDA topic 1 LDA_02: Closeness to LDA topic 2 LDA_03: Closeness to LDA topic 3 LDA_04: Closeness to LDA topic 4 global_subjectivity: Text subjectivity global_sentiment_polarity: Text sentiment polarity global_rate_positive_words: Rate of positive words in the content global_rate_negative_words: Rate of negative words in the content rate_positive_words: Rate of positive words among non-neutral tokens rate_negative_words: Rate of negative words among non-neutral tokens avg_positive_polarity: Avg. polarity of positive words min_positive_polarity: Min. polarity of positive words max_positive_polarity: Max. polarity of positive words avg_negative_polarity: Avg. polarity of negative words min_negative_polarity: Min. polarity of negative words max_negative_polarity: Max. polarity of negative words title_subjectivity: Title subjectivity title_sentiment_polarity: Title polarity abs_title_subjectivity: Absolute subjectivity level abs_title_sentiment_polarity: Absolute polarity level
Target	shares: Number of shares

4.4 Data Preprocessing

```
# check duplicates
dups = data.duplicated()
print ('Number of duplicate rows = %d' %(dups.sum()))
```

```
Number of duplicate rows = 0
```

Fig 1: Checking for duplicate values

```
#Before removing
print('Number of instances = %d \n Number of attributes =%d' %(data.shape[0],data.shape[1]))
print('----------------------')
data.isnull().sum()
```

```
Number of instances = 39644
 Number of attributes =61
----------------------

url                            0
timedelta                      0
n_tokens_title                 0
n_tokens_content               0
n_unique_tokens                0
                              ..
title_subjectivity             0
title_sentiment_polarity       0
abs_title_subjectivity         0
abs_title_sentiment_polarity   0
shares                         0
Length: 61, dtype: int64
```

Fig 2: Checking for null values

The data does not have any duplicate or missing values. The ‘URL’ and 'timedelta' columns are non-predictors so they are deleted from the data frame. The shares column is used to form a new column called Popularity which will be the binary target variable. The weekdays and channels used to publish are merged into a single column.

5. System

5.1 System Architecture

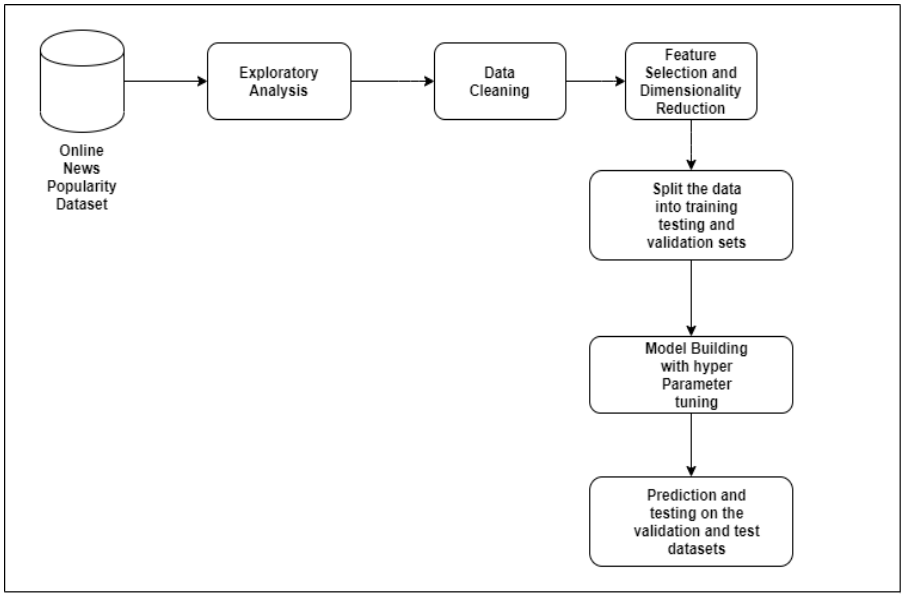

Fig 3: System Architecture

The system consists of the following components:

1) **Data Source:** The data is taken from the UCI data repository and consists of attributes for the articles published by Mashable online new website.
2) **Exploratory Analysis:** The dataset was analyzed to form different questions that are addressed in the objectives part of the paper. The data was analyzed to see on which day is most of the articles published, do people tend to read lengthy articles, does the number of images, videos, and reference to external media affect the popularity of the article.
3) **Data Cleaning:** The non-predictor columns were removed from the dataset. The data did not have any null or duplicate values so there was no need of cleaning the data.

4) **Feature Selection and Dimensionality Reduction:** For dimensionality reduction, we have used principal component analysis: Following diagram represents the scatter plot for the principal component one and principal component two.

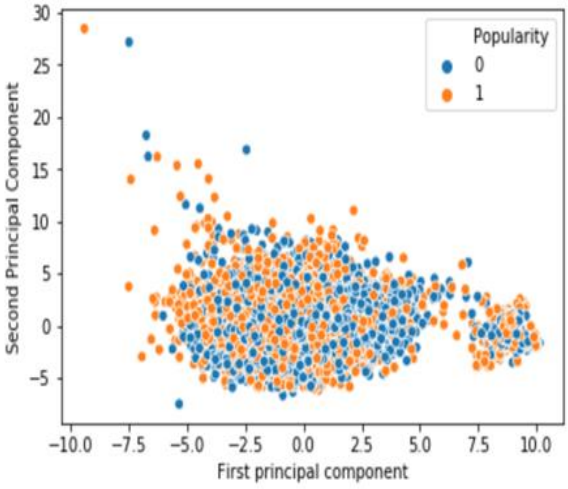

Fig 4: Principal Component Analysis

From the above diagram, it can be observed that PCA is not able to cluster the data into two separate groups because of which each component will have very less explained variance.

5) **Split the data into training testing and validation datasets:** 80% of the data was used for training the models and 10% of the training data is used for validation testing. 20 % of the entire dataset is used as the test data.
6) **Model building:** The models used for training are logistic regression, random forest, support vector machine, and K-Nearest Neighbor. For K-Nearest Neighbor we checked the accuracy for K ranging from 1 to 10 and chose the K which gives the highest accuracy is chosen for model building.
7) **Prediction on the test and validation datasets:** The models are tested by using the validation and testing datasets. The models are evaluated using the accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC score.

5.2 Data Processing

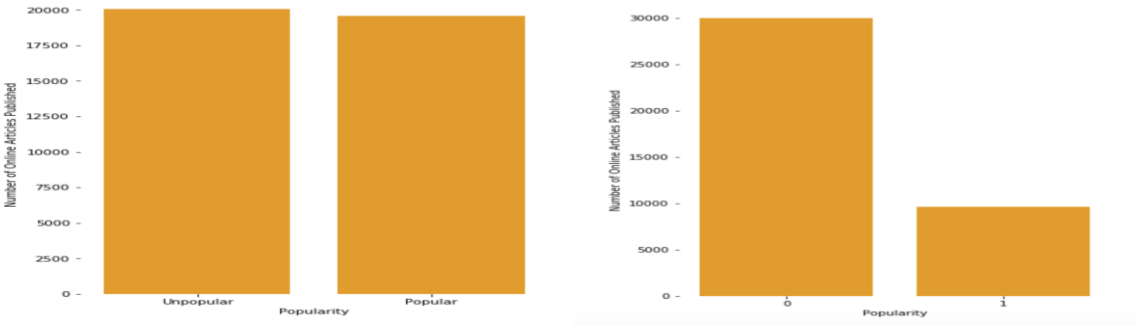

Fig 5: Data divided to have balanced target labels and unbalanced target labels

The target column is formed by using the shares column. The target column has two labels namely Popularity and Unpopularity where 1 stands for popularity and 0 for unpopularity. Two data frames are used for training the algorithms, the first data frame has an almost equal number of popular and unpopular class labels which is balanced data whereas the second data frame has more unpopular labels as compared to the popular labels which are unbalanced data. For balanced data if the number of shares is more than 1400 then it is classified as popular if it is less than 1400 then it is classified as unpopular. 1400 is chosen as the median for the share column is 1400. For unbalanced data articles having shares less than 2800 is classified as unpopular and having shares more than 2800 is classified as popular. 2800 is the third quartile of the shares column. The weekdays' columns are merged into one column called ‘Day Article Published’ where 1 is for Monday, 2 is for Tuesday and so on until Sunday. Similarly, the channels column is merged into a single column called ‘Data Channel’ where 1 is for lifestyle, 2 for entertainment, 3 is for bus, 4 for socmed, 5 for technology, and 6 for the world. The image above shows how the number of target labels is balanced for the balanced data and the unbalanced data there are more records with target labels as unpopular as compared to popular.

6. Data Analytics Algorithms

Following existing algorithms were used for model building and predicting the popularity of online news articles as the algorithms are supervised learning methods and are suitable for binary data where the target variable is binary.

1) **Logistic Regression:** Logistic Regression is used to define the relationship between the target variable and the predicted variables. It is suitable for problems where the target variable is categorical. It does not perform well when the independent variables are not correlated with the dependent variable. It requires fewer computations and does not require the data to be scaled. There are the following three types of logistic regression models:

a) **Binary Logistic Regression:** The target variable is binary. In this project, the target variable is binary as the article is either popular or unpopular.
b) **Multinomial Logistic Regression:** The target variable can have multiple values. For example, for sentimental analysis, the categories of target variables can be positive, negative, and neutral.
c) **Ordinal Logistic Regression:** The target variable can have multiple categories but the order of the data matters as well. For example, the ratings of a movie.

2) **Random Forest:** Random forest is a type of ensemble learning. It uses many decision trees and each tree predicts the class label and the label with most votes will be the predicted label of the instance. The trees will have very low correlation amongst themselves so that they are not similar to one another. The random forest always performs better as compared to the decision trees because as a majority the trees will give a better prediction.

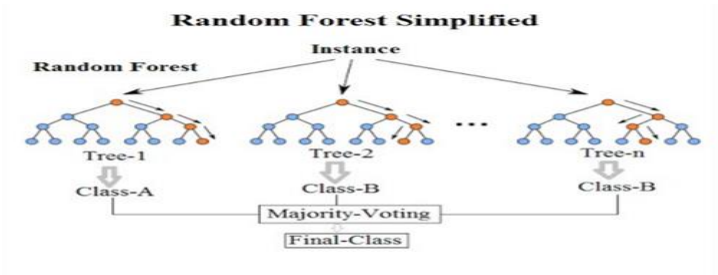

Fig 6: Random Forest

3) **Support Vector Machine:** SVM is a classification method used for two-class classification. It uses a hyperplane to divide the data points into two classes. The distance of the data point from the hyperplane matters for correct classification. The best hyperplane is the one in which all the data points are away from the hyperplane. If the datapoint point is close to the hyperplane then it may be the case that it is wrongly classified.

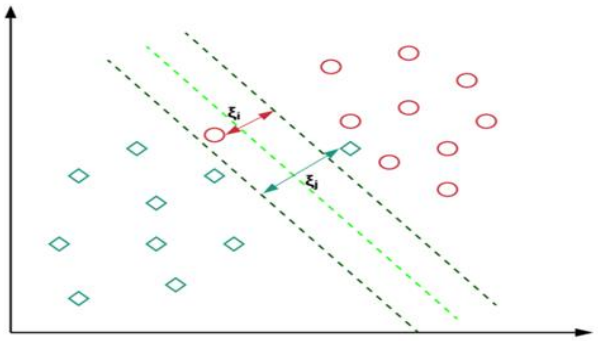

Fig 4: Support Vector Machine Algorithm

4) **K-Nearest Neighbor:** K- Nearest Neighbor is a supervised learning algorithm. KNN is an instance-based learning method. Distance is calculated between the data point to be

classified and the points near to it. The point is classified based on which points it is close to. The value of K is chosen by running the KNN algorithm multiple times. K that minimizes the error is chosen. Euclidean distance is the most common distance measure used for calculating the distance between the data point and its neighbors.

7. Software and Hardware Development Platforms

- **Hardware:** Laptops with configuration Intel I7 8th gen processor, 512 SSD, and 16 GB RAM.
- **Software:** Jupyter Notebook
- **Libraries:** Numpy, Seaborn, sklearn, Matplotlib

8. Experimental Results and Analysis

8.1 Exploratory Analysis:

8.1.1 Exploratory Analysis for the days of the week and the articles published.

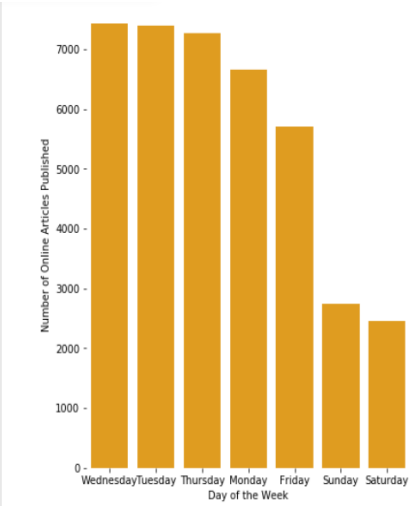

Fig 7: Popularity of Articles according to the day of week

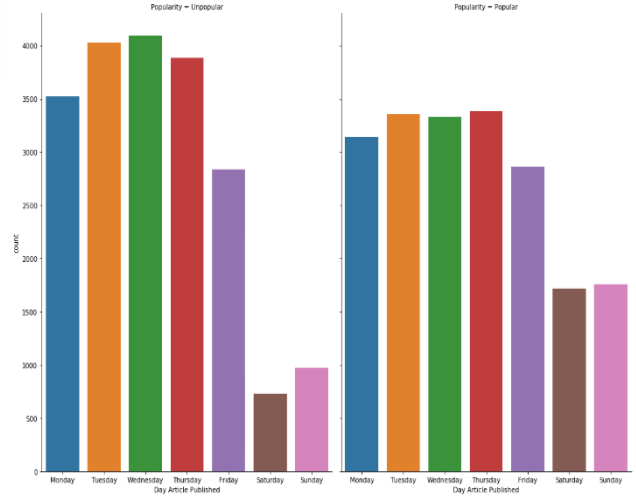

Fig 8: Number of articles published on each day of the week

From the above images, it can be observed that most of the articles are published on Wednesday and Tuesday. The articles published on Thursday become popular whereas the articles published on Wednesday are unpopular. In comparison, the articles published on weekends become more popular.

8.1.2 Exploratory Analysis for different types of Data Channels

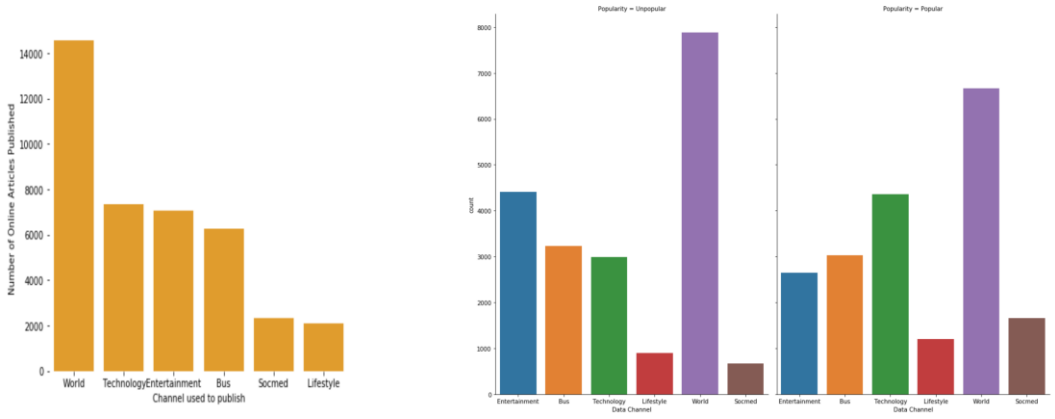

Fig 9: Number of articles published on different types of data channels in a week and according to their popularity(popular/unpopular)

From the above images, it can be seen that the articles related to the world become more popular followed by articles related to technology and entertainment.

8.1.3 Exploratory Analysis for the length of the Articles:

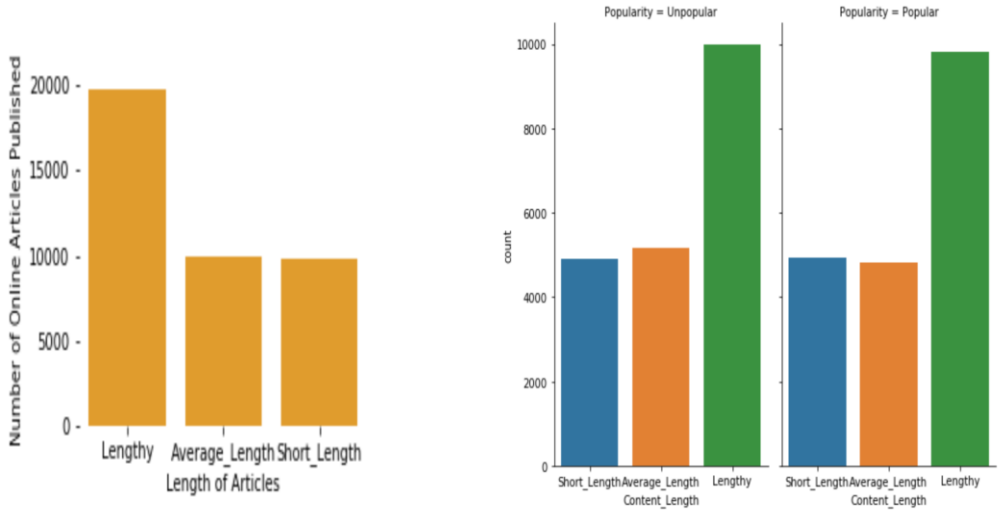

Fig 10: Number of articles that are present according to their length in a week and according to their popularity.

From the above images we can see that the length of the articles does not affect the popularity

8.1.4 Exploratory analysis for images, references, and videos.

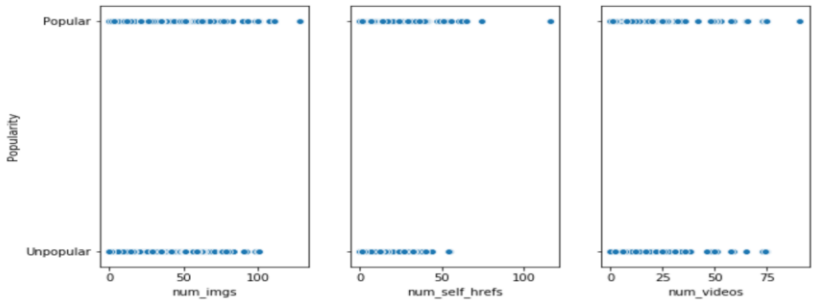

Fig 11: Total number of images, videos,hrefs according to popularity.

From the above images, it can be observed that the articles with a greater number of images, references and videos become popular.

8.2 Experimental Results

	Logistic Regression	Random Forest	K-NN	SVM
Validation set	71.53%	76.10%	73.89%	74.65%
Test set	71.63%	76.22%	74.14%	74.78%

Table 1: Accuracy of models for unbalanced data

	Logistic Regression	Random Forest	KNN	SVM
Validation set	59.55%	68.19%	58.00%	58.7%
Test set	59.78%	66.40%	56.91%	57.7%

Table 2: Accuracy of models for balanced data

	Logistic Regression	Random Forest	KNN	SVM

Validation set	56.52%	53.78%	52.80%	53.72%
Test set	56.07%	52.36%	52.66%	53.02%

Table 3: Accuracy of models after applying PCA to balance data

The above table shows the accuracies that were obtained on the testing and the validation datasets. After comparing the accuracies, it can be observed that the random forest algorithm performed well in all three scenarios. This is the case because the random forest uses several decision trees that vote to assign a label to an instance. PCA is used for dimensionality reduction but it did not perform well on this dataset and the accuracies are lowest for the models where data with PCA was used for training and testing. The explained variance ratio for the first principal component is 0.09 and for the second component it is 0.07 which is very less variance and both the components represent very less information because of which models with PCA did not perform well. The models trained using the unbalanced data have high accuracy as the data was unbalanced and has more data with unpopular labels, so it was able to perform well on the test data as well. The models trained using the balanced data gave real accuracies as the data was balanced.

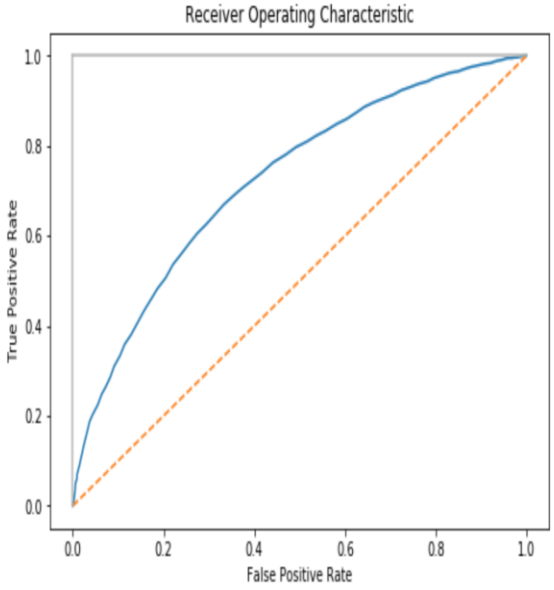

Fig 12: Random Forest model for balanced data

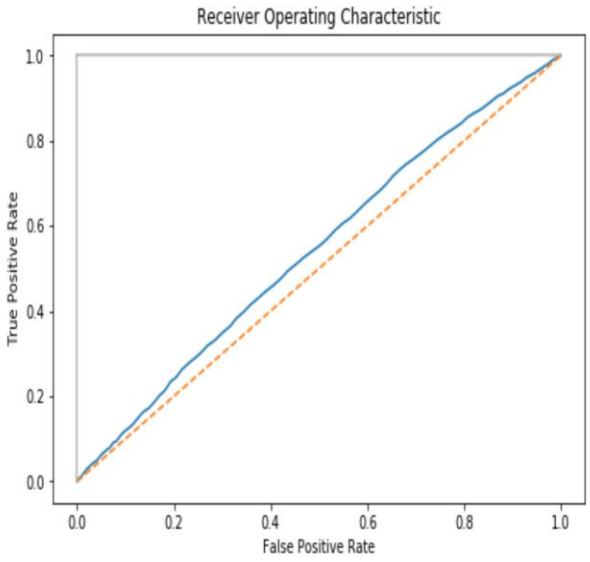

Fig 13: Random Forest model and with PCA components

The ROC curve is used to assess the accuracy of the model and it is a plot of false-positive rate against the true positive rate. The ideal situation is the top left corner of the graph where the true positive rate is 1 and the false positive rate is 0. In the diagram for the random forest model for balanced data, it can be seen that the area under ROC is less as compared to the other image where the model is trained after applying PCA. Lesser the areas under ROC better is the model. From the

images as well it can be observed that random forest performed well on balanced data as compared to the model trained using PCA applied data.

9. Conclusion

To conclude, the random forest model performed well on the data as compared to the other models. PCA is used for dimensionality reduction when the dataset has many attributes form this implementation it was observed that models trained using PCA have the least accuracies. So PCA does not work well in all the cases. The balance of the target attributes also plays a crucial role in model accuracies.

10. Definitions

Predictors: Predictors are independent variables that are used in the regression analysis to predict the dependent variable.

Target variable: Target variable is a dependent variable that is used in the regression analysis to be modeled and predicted by other variables

PCA: Principal Component Analysis (PCA) is a statistical technique used for dimensionality reduction which is also called feature extraction method.

Logistic Regression: Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more independent variables.

Random Forest: Random forests create decision trees on randomly selected data samples, get a prediction from each tree, and select the best solution using voting.

Support Vector Machine: SVM utilizes a hyperplane to imperfectly separate two classes, is known to be a very reliable algorithm with high accuracy.

KNN: K- Nearest Neighbour is a supervised learning algorithm. KNN is an instance-based learning method.

Precision: Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives.

Recall: Recall is defined as the number of true positives divided by the number of true positives plus the number of false negatives.

F1-Score or F-Measure: F1 score calculates the harmonic mean of precision and recall.

ROC: Receiver Operating Characteristic (ROC) curve is a measure of accuracy and a plot for the true positive rate(sensitivity) against the false positive rate(specificity).

11. References

[1] Ren, H., & Yang, Q. Predicting and Evaluating the Popularity of Online News. Department of Electrical Engineering, Stanford University.

[2] Uddin, T., Hossain, M., Patwary, M., & Ahsan, T. (2020). Predicting the Popularity of Online News using Gradient Boosting Machine [E-book]. International Islamic University Chittagong, Bangladesh.

[3] Random forest. (2020, July 11). Retrieved July 20, 2020, from https://en.wikipedia.org/wiki/Random_forest

[4]Misra, R. (2020, June 07). Support Vector Machines - Soft Margin Formulation and Kernel Trick. Retrieved July 20, 2020, from https://towardsdatascience.com/support-vector-machines-soft-margin-formulation-and-kernel-trick-4c9729dc8efe

[4] Chauhan, N. S. (2020, April 18). Real world implementation of Logistic Regression. Retrieved May 4, 2020, from https://towardsdatascience.com/real-world-implementation-of-logistic-regression-5136cefb8125

[5] Yiu, T. (2019, August 14). Understanding Random Forest. Retrieved July 23, 2020, from https://towardsdatascience.com/understanding-random-forest-58381e0602d2

https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[6] Srivastava, T. (2020, April 01). K Nearest Neighbor: KNN Algorithm: KNN in Python & R. Retrieved July 23, 2020, from https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

[7] Official seaborn tutorial¶. (n.d.). Retrieved July 23, 2020, from https://seaborn.pydata.org/tutorial.html