Word Count: 1945

Plagiarism Percentage 3%

Matches

1 **World Wide Web Match**
View Link

2 **World Wide Web Match**
View Link

3 **World Wide Web Match**
View Link

4 **World Wide Web Match**
View Link

Suspected Content

Online News Popularity Ananya Joshi Student of

> PES University Bengaluru, India ananyajoshigoa @gmail.com Shruthi Shankaran Student of PES University Bengaluru, India shruthishankaran10 @gmail.com 2

Shree Akshaya A T Student of

> PES University Bengaluru,India shreeakshaya2000 @gmail.com Abstract— The objective of 2

the project

> is to predict the popularity of the news based on the number of shares 1

done in an online platform. As there is drastic development and expansion of the internet, more number of people find it convenient in reading and sharing the news articles online. The dataset is from the UCI ML repository[1]. For the prediction of popularity, we performed a transformation for the target variable, a logarithmic conversion ,and then divided the popularity into two categories, whether the news is popular or

not, based on ‘shares’, the target variable.Then seven models were implemented i.e., Logistic Regression, DecisionTree, KNN,Adaboost,Gradient Boosting, Random forest and Integrating different models (LDA,Random forest, Adaboost, Gradient boosting) and using soft max voting. Finally the achieved accuracy was 67%. Keywords: Classification, random forest, prediction, popularity prediction, feature selection ,regression I. Introduction and background News is the information about all the current events that are happening all across the world. As the internet has become more and more popular, so has digital news which reaches the world very fast and is very convenient.Apps which can summarise the entire day’s news headlines are becoming increasingly popular due to less time being wasted in advertisements and more in actual information being gained.The browsers we use on our smartphones are able to predict which news articles we would like to click on, and are quite relevant, after a while of us using them. These Recommendation systems keep track of your interests based on whether you click on the article or not and dynamically update their databases. The dataset used has 61 columns (a target variable, namely ‘shares’ , 2 non predictive features - ‘url’ and ‘timedelta’) and 39644 rows. The reason the dataset interested us,topic wise, was because news is extremely relevant and the problem statement itself - ‘what would people like, what would people choose to watch’ is something which applies to most walks of life- shopping,social media, etc. Therefore by choosing this problem statement we wished to learn how to analyse real-life data to find out which news actually appeals to people and to be able to predict its popularity, accurately. Predicting the popularity of online news also helps in knowing what kind of news will be popular, so the news portal can arrange news according to their homepages. This can be influenced by the number of shares that are done through social media,as well as the title and the content of the news. It can also be helpful for the advertiser to know popular news, so it will be easy to customize eye-catching ads in order to get the reader's attention. The number of features would allow us to explore various feature selection techniques, and to be able to produce a better predictive model. II. Past Work In the paper,Online News Popularity Prediction[2] published in 2018, they used top 20 features according to fisher scores. They used various models including Multilayer Perception (MLP) with 4 layers, Badding Bagging , AdaBoost , Random Forests, a Naive Bayes classifier ,

K nearest neighbors , Logistic Regression, Support Vector Machine. Result :Random forest 3

model resulted in greatest precision,accuracy and recall and F measure for them. In the paper

,Predicting and evaluating the popularity of online news –published by He Ren, Quan Yang 1

[3].They used theFisher method for selection of one third of the attributes in the total dataset. The paper includes a lot of machine learning approaches i.e., Linear and Logistic regression, Support vector machine with linear, polynomial and gaussian kernels, and Random forest.Result: Comparing all the methods in terms of accuracy and recall value, the Random forest approach gave the best result of accuracy 69%. In the paper, Online News Popularity Prediction by Shuo

Zhang Research School of Computer Science, Australian National University 4

[4], Feature selection was performed by the implementation of evolutional algorithm.The threshold for

popular vs unpopular number of shares was taken as 1400,thus making it a binary classification problem. The final model achieved an accuracy of 70% . III. Assumptions We assumed a threshold value, namely median, for the number of shares to make it easier to classify.We compared various feature selection algorithms, and based on accuracy of the final model used selectkbest technique for the final hybrid integrated model, thus rejecting the other selection algorithms.We also assumed that the features which our feature selection method did not select did not drastically affect the model. We assumed that the outliers of certain columns could be kept as removing them would lead to loss of information which could be helpful in building a better model,in fact,a comparison was done between building the models with and without outliers as well. IV. Proposed solution We sought to implement methods of feature selection like kbest features, extra trees classifier, RFE. and build models using an integration of linear discriminant analysis, Adaboost, Gradientboost and Random Forest. Various combinations were tested and the model having best accuracy was implemented. The features which contribute most to an article's popularity had to be determined .38 features are positively correlated, 21 are negatively correlated, however the correlation values are not very significant (none of the features are strongly or moderately correlated >0.3) . There were no missing values.Various methods for detecting outliers were implemented including IQR,95 percentile method,z-score,DBSCAN, to check which would remove least rows.It was decided to prevent loss of information by not executing the removal of too many rows.Initially, IQR method was used for those rows based on the visualisation of the features which seemed to affect the target variable the most. The features selected for outlier detection were n_non_stop_unique,n_unique_tokens and the target variable 'shares'. Scaling: As different features have differing ranges of values, some features may influence the prediction more than the rest. Hence, scaling is necessary. We have used min-max scaling such that every feature value falls between minimum (0) and maximum (1). Since there were 60 features we attempted to use feature extraction techniques based on which were most relevant to the problem statement. Several algorithms were compared , some of which contradicted each other- eg one which gave a feature a higher ranking , and another would reject the very same feature.We referred to past literature to identify which techniques had worked better in history and built a better model. We did the feature selection by finalizing and comparing 2 techniques, namely the Recursive Features Elimination(RFE) which is a wrapper method, which uses Linear Regression model, and selectKbest features:attributes were scored using f_regression function. The latter gave an improvement of 2% in accuracy in the final testing dataset, so the model was implemented based on it. On visualization, it was observed that 'shares' has skewed distribution. Since 'shares' doesn't have a normal distribution, a log transformation was performed to result in a normal distribution data as some machine learning models rely on this. The transformation did help in improving the models. Also, the problem statement was converted to a binary classification one, where a threshold based on the median value was decided,and binary values of 1 and 2 were given to separate it before building the model.(It was checked with ternary classification based on IQR, but binary classification was better suited for the models.) For model building several approaches were used - logistic regression, random forest, adaboost, decision tree, gradient boost, k nearest neighbours and an integration of 4 models with softmax voting (linear discriminant analysis, Adaboost, Gradientboost and Random Forest). On evaluation, we found the top 35 features selected using k best features and the model built using the integrated model to have the highest accuracy of 67.12% for a data split of 70%-30% for training and testing set respectively. V. Experimental results We find that gradient booting, decision tree, adaboost and integrated model with Linear Discriminant Analysis, random forest, adaboost and gradient boost using soft max voting perform better for our dataset than the rest, KNN performs the worst. We found that removal of outliers led to lesser accuracy by 1%. Table-I(comparison of model with outliers vs without outlier) Model Accuracy(with outliers) Accuracy(without outliers) Logistic Regression 65.33 65.22 Random Forest 67.02 66.77 Gradient boost 66.35 66.01 Adaboost 66.41 65.66 K-nearest neighbours 61.21 61.72 Decision Tree 64.02 64.19 Integrated

model(LDA,Adab oost,gradient boost, random forest) 67.12 66.49 Table-II (Comparisons of models considering precision, recall ,f1-score) Model Precision Recall F1-Score Logistic Regress ion 0.65 0.65 0.65 Random Forest 0.67 0.66 0.66 Gradien t boost 0.66 0.66 0.66 Adaboo st 0.66 0.66 0.66 K-neare st neighbo urs 0.61 0.61 0.61 Decisio n Tree 0.64 0.64 0.64 Integrat 0.67 0.67 0.67 ed model(LDA,A daboost, gradient boost, random forest) From the table we can see that the integrated model with Linear Discriminant Analysis random forest, adaboost and gradient boost using soft max voting gives better accuracy of 67.12% . The AUC score of this model = 0.72762 And the ROC- curve for the model: Conclusions Some insights gained from data visualization were: Tech news has most popular news, World news has the most no. of unpopular news. Also, most popular news are the ones published on the weekend, unpopular on weekdays.In general, good articles have upto 20 visuals (images/ videos) and lesser words in their content, with title length 5-15 words, avg word length 4-6, higher no. of keywords. The integrated model performs significantly better than other models by ~2%, and thus is selected. final accuracy achieved is 67.12%. Gradient boost and random forest perform quite well too. One shortcoming of the hybrid/ integrated model is its execution time, which is longer than the other models used.There is scope of improvement of accuracy by selecting a different number of features and changing the weights. Contributions PES1201801677:Cleaning,Visualization:z-score and scatter plots, PCA,selectkbest for feature selection, checking model accuracy with outliers PES1201801466: Visualization:histogram, Box plot(outlier removal), Implementation of RFE for feature selection , logarithmic transformation of target variable,checking model accuracy without outliers PES1201800170: Visualization: line plots, IQR method(outliers),Lasso regression, DBscan for feature selection, Comparison of model accuracies of classifying into 2 vs 3 (based on quantiles) All members tried various combinations of models, checking accuracies with and without removal of outliers. References [1].http://archive.ics.uci.edu/ml/datasets/Onl ine+News+Popularity [2]Online News Popularity Prediction Feras Namous ,Ali Rodan (Nov 2018) https://www.researchgate.net/publication/33 1347814_Online_News_Popularity_Predicti on [3]Predicting and evaluating the popularity of online news – published by He Ren, Quan Yanghttp://www.academia.edu/download/53 460177/popularityOnlineNews.pdf [4]Online News Popularity Prediction by Shuo Zhang Research School of Computer Science, Australian National University, http://users.cecs.anu.edu.au/~Tom.Gedeon/c onf/ABCs2018/paper/ABCs2018_paper_167 .pdf Appendix Box plot for shares PCA on 2 dimensions Line plots PCA on 3 dimensions Correlation matrix (below, left) Histograms 1) Count of popular, unpopular news over genres, mean 2) Count of popular, unpopular news over genres, using median 3) Count of popular, unpopular news over genres, distributed over days in a week Scatter plots Histograms for all features