ONLINE NEWS POPULARITY USING MACHINE LEARNING

Nalabolu Sriram, Srikanth Kolluri, Kalyan Badeti Parameswara, Samanth Satvai

*Abstract*—In our project, news article statistics of a company mashable.com are used to evaluate the popularity and noteworthy information of the target variable using various Machine Learning techniques. These days, news is only one of the many kinds of information that can be located on social media platforms like Facebook and Twitter. A deluge of online news items engages readers in a few different ways on a daily basis. The purpose of this study is to investigate how the distinct qualities of each post, as evaluated by the number of shares, impact the posts' overall levels of popularity. An indication of how far the news has been disseminated online may be gleaned from the number of shares and likes that a post receives. In addition, data from social media platforms may be helpful when attempting to forecast how many people will like a post and the reasons why. Therefore, we have a great deal of interest in social media data analytics in order to gain a better understanding of and improve our ability to forecast the level of interest indicated by the number of shares. Specifically, we want to know how the level of interest in specific websites and individual articles is affected by the various types of data, variables, and presentational approaches. Working with data mining algorithms in huge datasets is frequent these days, especially with the growth of online news, which has made it highly valuable. A well-known internet news website called Mashable served as the source of the data. The findings were evaluated using precision, recall, and the F-measure, and then those evaluations will compare against one another to see which one was superior. In addition to this, we compare our findings to those of previously published publications using the same dataset. First, we prepared and cleaned our data. Then we altered our target variable “shares” to be predicted. We created a binary class where the output based on the number of shares is either 0 or 1. We did hyperparameter tuning to increase the accuracy of the models. We also visualized our results, compared our accuracies, and decided on the best models for our prediction.

Keywords—News, popularity, regression, classification, xgboost

# Introduction

In the modern world, many different social media platforms have rapidly evolved into one of the most essential channels of communication for the spread of a wide variety of information, including the news. There are a significant number of daily posts of internet news, and each of these articles connects with people in a diverse range of ways or tactics that are chosen by them on their own. The many different characteristics that go into making up a post have the potential to influence its level of popularity, which is the primary subject of this research and is determined by the number of shares. [1] Our idea will be investigated as part of this research. The amount of support a specific post receives from users on social media (measured in the form of likes) can be used as an indication of how widely the information it contains is being shared. In addition, the information that can be gleaned from social media platforms is really important when it comes to estimating the level of adoration that visitors would have for a variety of content. This is due to the fact that the information may be gleaned from the likes, shares, and comments provided by users. Because of this, we have a significant interest in analytics of the data that is obtained from social media platforms in order to investigate and forecast the popularity that is shown by the number of shares. Specifically, we want to examine and forecast the popularity that is shown by the number of likes. For instance, one area of research that piques our attention is determining the kinds of data, characteristics, and writing styles that have an impact on the volume of traffic that is received by websites or publications. The dataset has a large number of different aspects, each of which possesses the potential to be of value in the inquiry. Some examples of these characteristics are the total number of words, the total number of links, the total number of photos and videos, the date the material was initially made available, and a great deal of additional information.

It is impossible to emphasize the significance of popularity prediction models for online news, which were built with the assistance of machine learning algorithms, in the news industry of today. These models were designed for online news. These algorithms are able to provide accurate predictions regarding the level of reader interest in online news stories because they take into account a wide range of criteria. These criteria include the subject matter of the article, the time of release, and the level of activity on social media sites surrounding the post.

As a direct consequence of the technological developments that have taken place in this day and age, it has become an essential component of our day-to-day lives to keep up with the latest news by reading it online. Because there are now an ever-increasing number of news stories that can be accessed over the internet, it has become increasingly difficult for publishers and editors to forecast which pieces will be the most widely read and will pique the interest of their audience. This is because there are now an ever-increasing number of news stories that can be accessed over the internet. This is because there are now a growing amount of news items that can be obtained on the internet. This trend will continue as long as the internet remains popular. On the other hand, with the assistance of algorithms that are developed for machine learning, it is now feasible to make predictions about the amount of people who will be interested in online news that are both more accurate and effective. This is the case because it is now possible to make predictions about the amount of people who will be interested in online news. The purpose of the machine learning algorithms is to make an educated guess as to how popular an article will become in the future by taking into account a range of criteria. The subject matter of the piece, the date it was published, and the quantity of engagement it received on social media are some examples of these characteristics. This technology has enormous ramifications for the media sector as a whole as a result of the fact that it may provide publishers and editors with assistance in the production of content that will connect with their respective audiences. This, in turn, has the potential to contribute to an increase in income, a decrease in the propagation of disinformation, the detection of new trends, and an enhancement in the quality of the user experience. Throughout the course of this essay, we are going to investigate the importance of popularity prediction models for online news, as well as their possible applications and the difficulties that are connected with putting these models into practice.

# Motivation

Quite a few different lines of reasoning lend credence to the relevance of models that attempt to anticipate the level of interest in internet news. To begin, they provide assistance to publishers and editors in the process of optimizing the information they produce in order to achieve the maximum degree of engagement that is humanly feasible. If publishers are able to make an accurate prediction as to which articles are most likely to be read by their audience, they are in a better position to allocate their resources toward the development of high-quality content that resonates with that audience.

Second, the use of algorithms that project the amount of interest that people will have in online news may be of assistance to publishers in obtaining better levels of financial success. Finding out which of their articles are the most popular may help publishers raise the number of page views and clicks on their website, which in turn can lead to an increase in the amount of money they earn from advertising.

Thirdly, these models have the ability to help limit the spread of misleading information and fake news, which is an important goal of mine. Models that predict the popularity of online news articles by analyzing the content of those articles are able to identify items that are likely to include inaccurate or misleading information. These models are able to do this because the models analyze and predict the popularity of online news articles. This may assist to avoid the spread of deceptive material and ensure that readers are provided with news that is based on facts and can be relied upon.

The fourth argument is that models that predict the popularity of online news may be used to detect emerging issues and trends. These models can be used to identify developing subjects and trends. These computers are able to predict topics that are likely to become popular in the not-too-distant future by analyzing the activity that occurs on social media platforms and the patterns that occur in people's search searches. Because of this, it can be simpler for publishers to stay one step ahead of their rivals and develop content that is both current and relevant at the same time.

The user experience on news websites may be improved by using these models, which have the potential to be used for this purpose. By undertaking an examination of user behavior and preferences, algorithms that forecast the popularity of online news are able to identify the kind of articles that consumers are most interested in reading. It is possible for this to be of aid to publishers in the process of customizing their content and offering users with an experience that is more engaging and gratifying.

# Main Contributions and objectives

* Data exploration: review the type and distribution of data, the relationship between variables and how they relate to the predicting target.
* Data cleansing: Identify and rectify inconsistencies, anomalies, missing data, outliers.
* Visualization of the data: shows the data’s statistics and basic analysis to see the nature of data.
* Data structuring: consider the machine learning algorithms that will be used in the study.
* Feature engineering and selection: create new variables to improve model’s output and choose relevant and eliminate non relevant variables.
* Running machine learning models
* Comparing the models, selecting the best model, visualizing the results.

# Related work

Over the course of the previous several years, predicting trends in popularity has been the subject of a substantial amount of research. This is resulting in the emergence of new challenges, as well as new areas of scientific interest. A substantial number of scholars are dedicated to determining the location of the spot where the most information is concentrated. The ability to anticipate what will be widely used is a crucial step that will result in significant advantages in practice.

According to Ren et al. [2,] the Online News Popularity Dataset was analyzed using a broad variety of machine learning techniques, all of which could be found in the "UCI machine learning library." They sorted the features by using the Fisher criteria and the Mutual information, and based on the rankings they obtained, they selected the top 20 characteristics. In order to make their prediction, Ren et al. made use of a variety of machine learning strategies, some of which are listed below: "Linear Regression, Logistic Regression, Support Vector Machine, Random Forests, K-Nearest Neighbors, REPTree, Kernel Partial Least Square, Kernel Perceptron, and C4.5 Algorithm." They used a 5-fold criteria of validation cross to test their hypothesis. Logistic regression performed better than most of the other models when it was utilized as a classification model, producing a fair degree of accuracy as a result. The Random Forest method was the one that proved to be the most effective in overcoming this classification obstacle. They were able to employ a variable number of decision trees as well as a variable quantity of features for each decision point. This gave them a great deal of flexibility. As a consequence of this, they were able to attain an accuracy of roughly 66 percent in linear regression and approximately 55 percent in support vector machine analysis by using a polynomial kernel with a value of nine. They began with just five trees and worked their way up to a total of five hundred throughout the course of the project. This is the most accurate algorithm that has ever been constructed, with a degree of accuracy that reached a high of 69 percent.

An Online News Popularity dataset was used as the training environment for the many distinct machine-learning approaches that Frenandes et al. [3] applied. These machine-learning techniques include "Random Forests, AdaBoost, Support Vector Machine, K-Nearest Neighbor, and Naive Bayes." The combined use of AdaBoost and SVM was successful in achieving an accuracy of around 66 percent. The best result was obtained by a Random Forest (RF), which "had an overall area under the Receiver Operating Characteristic (ROC) curve of 73 percent, which corresponded to an acceptable discrimination and an accuracy of approximately 67 percent." However, KNN and Naive Bayes both obtained an accuracy of approximately 62 percent.

The previous work done by Junli Kong and his colleagues [4] was improved, and a deep learning model called Gated Recurrent Unit was successfully coupled with the multivariate features taken from metadata (GRU). Because it is able to overcome the long-distance dependencies that are present in "Recurrent Neural Networks (RNN)", "The Gated Recurrent Unit (GRU)" has the benefit of maintaining the relevant characteristics via gate. This is made possible as a result of the fact that it is able to maintain the relevant characteristics. In comparison to the "Long-term short-term memory" (LSTM), the "Gated Recurrent Unit" (GRU) has an architecture that is easier to understand and needs less time to train. The Gated Recurrent Unit, often known as the GRU, has shown to be very effective in the field of natural language processing (NLP), which encompasses a wide variety of applications, such as voice recognition and machine translation.

With the help of Random Forest regression, Shreya and her team [5] were able to make their forecast about the amount of shares obtained by news articles. "The parameters of the RFR model were altered in such a way that the number of decision trees, also known as n estimators, would be set to 10, and the quality of the split measure, also known as criterion, would be set to be the mean square error (mse)." "The coefficient of determination score was determined to be -0.45 on the test set after training the RFR on sixty percent of the data, which was around 23,500 articles. This was after the test set was run. The R2 score is low because the RFR model does not include any parameters that have been tuned, and the share count distribution is extremely skewed in the wrong direction. This is the explanation for the poor result on the R2 test. The R2 parameter on the RFR model may therefore be optimized by first identifying the best feasible values for each parameter, and then using those values for the parameter. The RFR model is optimized via parameter selection by altering the values of two parameters: the number of decision trees, as well as the appropriate amount of features to split the tree at the nodes. This is done in order to get the best possible results from the model (max features). n estimators and max features were chosen because, respectively, they are the most significant elements in defining the error that is associated with each tree in the forest and the correlation that exists between the trees. This led to the selection of n estimators and max features. Following the optimization of these parameters, an improved version of the RFR model was constructed in contrast to the one that had been developed before.

To determine how well the RFR model performed when it was trained on the two distinct kinds of data, an out-of-bag estimate, also known as an oob score, was applied to the score that was produced from the training data set. This was done in order to quantify how well the model performed overall. Although the number of estimators was first set at 100 and then increased by increments of 50, the maximum number of features was fixed at a value equal to the square root of the number of features data. This was done while the number of estimators was raised by increments of 50. As can be seen in the figure that follows, the oob score continues to progressively fall until the number of estimators reaches 400, at which point it starts to gradually rise. This pattern may be seen until the end of the section. As a result, the number 400 is the optimal choice for n estimators to consider.

[6] makes the suggestion of using a regression model for load predictions. In order to create an accurate estimate of the load, the GPR model was used. The GPR models, which are nonparametric probabilistic models based on kernels, are the models that are recommended to be utilized as the final models. These models are probabilistic. During the course of their investigation, they came to the conclusion that the GPR is a method that is useful for accurately predicting load. It is nonparametric, which means that it is not limited by a functional form. As a result, rather than calculating the probability distribution of the parameters of a specific function, GPR calculates the probability distribution over all admissible functions that fit the data. This is in contrast to the traditional approach, which calculates the probability distribution of the parameters of a single function. As a result of the fact that it is nonparametric, it is not constrained by a particular functional form. GPR is able to discover meaningful patterns by leveraging the already-existing training datasets and by conducting data extrapolation on the patterns it has learned. This allows it to expand on the information it has gained from its previous learning. A GPR is able to make predictions despite having access to only a limited number of training datasets and implementing a prediction model that is both computationally and financially efficient. This is because of the GPR's ability to implement a prediction model that is both computationally and financially efficient. In addition to this, it provides a predicted distribution, which is described by the mean value in combination with the variance that is linked with each value. As a consequence of this, the algorithms known as Rational Quadratic GPR and Exponential GPR are the ones that have to be employed while carrying out load forecasting.

[7] They build a regression model that is able to provide an accurate prediction of the price of electricity for the next day. In this study, an innovative hybrid machine learning technique referred to as ARD-ETR is offered as a means to forecast day-ahead electricity prices. ARD-ETR was developed as a result of the research presented here. The methodology utilizes ensemble tree bagging in addition to linear regression as its foundation. A time series representation of the dataset that contains the prices of energy was gathered from the Nord Pool spot market. This market provides the prices. After that, this dataset is analyzed in order to determine whether or not the proposed technique is successful. In addition, the approach that was recommended is able to handle the difficulties caused by the features of the time series thanks to the combination of the ensemble tree-based bagging model ETR and the linear regression model ARD. Both of these models are linear regression models. After that, the historical dataset is converted into a supervised learning method by taking the value from one week in the past and dividing it into two phases. The training phase is used to develop the model for forecasting, and the test phase is used to establish the model for evaluating. Both phases use the value from one week in the past. According to the empirical findings, the ARD-ETR hybrid strategy that was developed has attained the greatest performance in terms of MAE, RMSE, and MSE when compared to the individual and other hybrid techniques that were employed on their work. To be more precise, the findings indicate that the ARD-ETR technique achieved the lowest MAE, RMSE, and MSE values (£/MWh) with corresponding values of 2.03, 3.09, and 16.7. In addition, when contrasted with the other methods that were used in the prior studies for the day-ahead EPF, the hybrid ARD-ETR technique produced the lowest feasible MAE scores and RMSE values. This was the case when compared to the other approaches. Therefore, a recently developed hybrid approach demonstrates greater improvement when evaluated with the method of the benchmark, as well as a significant reduction in testing MAE value (£/MWh) with 32.1 and RMSE value (£/MWh) with 29.12 in comparison to earlier hybrid models. These results can be attributed to the hybrid approach's ability to better account for the effects of the technique of the benchmark.

[8] They constructed a regression model to predict the flow of traffic, which is a situation that is similar to our project and may be compared to the same. They have set as their objective the creation of a strategy that is capable of producing more accurate estimates of the volume of traffic. This not only makes a significant contribution to the planning and resolution of transportation issues, but it also provides a more effective methodology for the development of navigation software in circumstances in which real-time data on traffic flow may not be available everywhere or at all times. This is especially useful in situations in which real-time data on traffic flow is required for planning and resolution of transportation issues. Utilizing three distinct types of traffic data, their study proposes a method that is both practicable and methodical for estimating the flow of urban traffic. It is generally accepted that measurements of topology and geometry belong to the category of independent variables, but traffic flow is generally accepted as belonging to the category of dependent variables. When the results of the estimates are compared, it is shown that a combination of topological and geometrical metrics provides a larger R value. This was observed when the R values were compared.

# Proposed framework

It is intriguing to make predictions about how popular internet news will be in the future for a variety of reasons, one of which is that the outcomes might have significant implications for news organizations, journalists, and advertising. When news organizations are able to make educated decisions about which stories to promote and how to devote resources based on a prediction of the popularity of an online news piece before it is released, this is beneficial to both parties. This is because both parties can make more informed decisions about which stories to promote.

In addition, marketers may use these projections to identify which articles have the potential to generate the greatest traffic and, as a consequence, the most cash from advertising revenue. This information may also help journalists better customize the material they generate to the unique interests and needs of their audience members, which is another way in which this information can be of use to them.

The popularity prediction of online news is intriguing for the same reason that it highlights the potential of machine learning and data analytics in the field of journalism. When seen from a more macro perspective, the popularity forecast of online news is noteworthy. As the amount of data that is made available to journalists continues to grow, there is a growing demand for tools and methods that can assist them in making sense of the information that is available to them and in determining the stories that are most likely to resonate with their readers. This demand is driven by the fact that the amount of data that is made available to them is continuing to expand. The prediction of the popularity of online news is only one example of how these tools and tactics may be applied to enhance the quality and relevance of the content that can be viewed on online news sites.

The strategy that we are going to use to [9] deal with the prevalence of internet news is as follows:

Gathering of Information: The very first thing that has to be done is to collect data on a variety of news items and the qualities that go along with them (e.g. author, headline, publication time, category, sentiment, etc.). These information are obtainable by way of application programming interfaces (APIs), news websites, and third-party data sources.

The first processing of data: It is possible that the data will need to be cleaned and preprocessed before it can be used for analysis. This is done in order to guarantee that the data acquired is of a high quality and can be utilized. At this point, you can be requested to perform things like normalize the data, delete duplicates, or fill in missing numbers.

Feature Extraction: The second phase, which comes after the data has been cleaned and preprocessed, is to extract essential qualities that may be used to create predictions about how popular certain news items will be. This step occurs after the data has been preprocessed and cleaned.

Training the Model The last step that has to be taken is to train a machine learning model on the characteristics that were selected. The decision tree, the random forest, and the neural network are a few examples of well-known models for estimating the degree to which people will be interested in a certain piece of news.

Model Evaluation Once the model has been trained, it is essential to undertake a performance analysis using measurements such as accuracy, precision, recall, and F1 score. This should be done as soon as possible after the training has been completed. Holdout testing and cross-validation are two examples of methods that might be used at this stage of the process.

Deployment of the Model: The last phase is the preparation of the model for deployment in a production environment, where it will be used to make real-time predictions on the popularity of a variety of different news stories.

The use of machine learning to the challenge of forecasting the degree of interest in online news is notable for a variety of reasons, including the following:

The use of popularity prediction models for online news can be beneficial to journalists and editors because it enables them to gain a deeper understanding of the types of news articles that are most likely to be well received by their audience. This, in turn, enables them to improve the quality of the content they produce. By doing a study of the patterns and trends that are included within the data, they are able to discover the topics, headlines, and formats that are most likely to generate clicks, shares, and interaction with the audience.

Better Marketing: These models may also be used by news organizations to better focus their marketing efforts to the particular audiences for whom they are writing, making for more effective marketing. For instance, by doing a study of the patterns of interaction for certain age groups, geographic locations, or interest categories, they are able to construct advertising that are more precisely targeted and are thus more likely to reach the proper individuals.

When media companies are aware of the articles that are most likely to be seen by a big audience, they are better equipped to optimize the effectiveness of their advertising strategies and placements, which ultimately results in an increase in the potential money that they may generate. This is of the highest relevance in an industry that has been seeing a fall in advertising income, and in which organizations in the media business are exploring for new methods to monetize their content.

The capacity to predict the degree of interest in online news may also assist to enhance the user experience for those who read it, which is great news for everyone in the publishing industry. By recommending articles that are most likely to be of interest to particular users, news providers are able to provide a more personalized and engaging experience for their audience, which encourages readers to continue returning for more content. This in turn encourages readers to continue returning for more content.

In general, machine learning has the potential to be of assistance to media companies in improving the efficiency of their content, marketing, and revenue strategies, in addition to the quality of the user experience that is offered to their readers.

# Data Description

We created prediction models by employing a variety of categorization machine learning approaches, such as decision trees, random forests, logistic regression, SVM, KNN, and Naive Bayes. The dataset used to generate these models was obtained from the UCI repository. The next step is to evaluate each approach and decide which one is the most effective.

There were neither missing nor null data in the dataset, which had a total of 39644 samples and 61 characteristics.

The total number of occurrences is 39797

61 is the total number of qualities (58 predictive attributes, 2 non-predictive, 1 goal field)

The Breakdown of the Classes: The worth of the class is subjected to continuous evaluation at all times (shares). Before moving further with the project, we first converted the work into a binary format using a decision threshold of 1400.

Variation in Share Prices: The number of occurrences inside the range is as follows:

1400 below 18490

1400 above 21154

There were neither missing nor null data in the dataset, which had a total of 39644 samples and 61 characteristics.

The 'URL' and 'timedelta' attributes of the dataset were removed since it was determined that they did not contribute anything useful to the analysis.

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Graphical user interface, text

Description automatically generated

# Results

Decision Tree Classifier:

This is an example of a supervised machine learning algorithm known as a decision tree classifier. This type of algorithm comes to conclusions in a manner that is comparable to that of humans.

One could consider an algorithm for classifying data that makes use of machine learning to be something that is intended to make decisions.

In most cases, you will state that the model can correctly forecast the category of the brand-new, never-before-seen input; but, behind the scenes, the algorithm must select which category to assign.

You could come across references to decision trees referred to as the CART algorithm due to the fact that they are capable of performing classification and regression tasks: A tree depicting both classification and regression. This is an umbrella phrase covering tree-based algorithms in general, not just decision trees in particular. Specifically, this term refers to decision trees.

The purpose of using the characteristics of the dataset to generate yes/no questions and then repeatedly segmenting the dataset until it is possible to single out all of the data points that belong to each class is what Decision Trees are attempting to accomplish.

You are constructing a hierarchical organization of the data as you implement this strategy.

Bear in mind that the addition of a new question results in the creation of a new node in the tree. The first node that one encounters when navigating a tree is referred to as the root node.

This method works at totally segmenting the dataset in such a way that every leaf node, or a node that is no longer responsible for splitting the data, is assigned to a single category. The term "pure leaf nodes" can also refer to these.

Baseline model: Logistic regression

In logistic regression analysis, the primary emphasis is placed on grouping individuals into the many categories that can be distinguished. The aim that is supposed to be served by logistic regression analysis is the same purpose that is intended to be served by other methods in the field of statistics for the construction of models. When conducting an investigation of this nature, the primary goal is to develop a model that is both plausible and capable of describing the correlation between dependent (predicted) and independent (predictive) variables while accounting for the smallest amount of variance possible. This should be accomplished with as little variation as is humanly possible.

The investigation yields data that can be put to use in comparatively assessing the several application steps of the logistic regression analysis. When the coefficient predictions of the intended model variables at the end of the study are taken into consideration, it is shown that an increase of one unit in the predictive variable of scientific thinking skills results in an increase of 14.4 percent in the likelihood of having good critical thinking. This is demonstrated by the fact that an increase of one unit in the predictive variable of scientific thinking skills also results in an increase in the likelihood of having good scientific thinking skills. Additionally, it has been discovered that the likelihood of having strong critical thinking increases by 4.9 percent for every one unit that is added to the predictive variable of epistemological belief [12].

After adjusting, the accuracy went up to 0.64 from a previous value of 0.60.

The precision is 0.65, which indicates that out of all the samples that are projected to be positive, there are 3672 samples that accurately predict, whereas 1993 samples predict incorrectly. The recall is 0.62, which indicates that out of all the actual positive samples, there are 3672 samples that make accurate predictions and 2235 samples that make inaccurate predictions.

Xgboost:

Even though decision trees are one of the models that can be comprehended with the least bit of effort, their behavior can be quite unpredictable at times. Take, for example, a single training dataset that has been arbitrarily divided in half in order to produce two distinct datasets. Now that we have all the components, let's put each one through their own decision tree training so that we can have two distinct models.

Chart, treemap chart

Description automatically generated

The results that we would receive from putting these two models through their paces would be substantially different from one another. One school of thought contends that decision trees are associated with a significant amount of volatility on account of this tendency. Methods of aggregation such as bagging and boosting are two examples of strategies that can help decrease the amount of variety that exists in each student. Bagging is a method of education that involves the production of numerous decision trees all at once in order to construct the fundamental students. The instruction that these students receive is based on data that has been sampled multiple times via replacement. The outcome that is obtained by averaging the output of all of the learners is the greatest forecast that can be made.

When boosting is used, the trees are formed in a sequential way, and the objective of each consecutive tree is to reduce the errors that were caused by the tree that came before it in the construction process. Each tree learns new information from its predecessors and corrects any mistakes that are still present in the database. As a direct result of this, the following tree in the sequence will gain knowledge by employing a refined version of the residuals.

The students with the lowest scores are referred to as the base learners, and they are used in the process of boosting. These students show a significant amount of bias, and their ability to anticipate outcomes is only marginally superior to that of guessing at random. The fact that each of these weak learners provides some crucial information for prediction makes it possible for the boosting strategy to effectively combine these weak learners and construct a robust learner as a result of the combination of these weak learners. The final strong learner brings both the bias and the variance down to values that are acceptable.

Boosting is a form of bagging that makes use of trees that have a fewer number of splits than other bagging methods, such as Random Forest, which allow trees to develop to their fullest potential. These relatively small trees, the roots of which do not extend very far, lend themselves quite well to interpretation. An optimal decision regarding parameters such as the number of trees or iterations, the rate at which the gradient boosting algorithm learns, and the depth of the tree could be made through the utilization of validation strategies such as k-fold cross validation. This would allow for an optimal decision to be made. If there are a significant number of trees in the region, there is a possibility of overfitting taking place. As a result, it is very necessary to exercise utmost caution while selecting the stopping conditions for the boost.

After tuning, there was neither an increase nor a decrease in the accuracy, and the process of tuning itself takes a very long time to complete. The primary model has consequently been selected.

The precision is 0.67, which indicates that out of all the anticipated positive samples, there are 3886 samples that correctly forecast, while there are 1939 samples that incorrectly guess. The recall is 0.66, which indicates that there are 3886 samples that properly predict actual positive samples, while there are 2021 samples that incorrectly forecast actual positive tests.

For this particular model, the data channel is entertaining characteristic is of the utmost significance. To put it another way, the entertainment pieces might have a higher likelihood than other sorts of content do of becoming popular or unpopular.

Naive Bayes:

The precision is 0.66, which indicates that of the samples that are anticipated to be positive, there are 760 samples that correctly forecast, while there are 400 samples that incorrectly predict. Because the recall is only 0.13, this indicates that out of the real positive samples, there are only 760 samples that make accurate predictions, while the remaining 5100 samples make incorrect predictions.

The issue with this model is that it forecasts a relatively small number of positive outcomes (1,160 samples total), while forecasting primarily unfavorable outcomes (10,700 samples). Because of this, its precision is very good, but its recall is only moderate.

KNN:

The precision is 0.56, which indicates that among the samples that are projected to be positive, there are 3200 samples that correctly forecast, while there are 2500 samples that incorrectly predict. The recall is 0.54, which indicates that out of all the actual positive samples, there are 3200 samples that make accurate predictions and 2700 samples that make inaccurate predictions.

Because the accuracy, precision, and recall are all just a hair above the random model's threshold of 0.5, we can conclude that this model is not adequate and that it requires additional tuning or additional data wrangling.

Table

Description automatically generated

After performing experiments and performing analyses on a wide variety of classification and regression models in order to select the best models for the project, the logistic regression model is considered to be the base model, and the XG model is considered to be the major model. Both of these models were chosen after we conducted these experiments and performed these analyses. According to the study that classifies things, numerous different models came up with differing degrees of accurate results and displayed varying levels of consistency across the board. As a result, the following is a thorough picture of the findings that were reported in the classification report for the numerous models that were run. This allows one to choose which model is the most accurate based on the proportion of true predictions it made. The accuracy rates of each model appear to be roughly equal to one another when compared to one another; however, the accuracy rate for the principal model, which is the XGBoost model, is much greater than that of the other models, coming in at 0.67.

In the end, despite the fact that the accuracy of each model is not all that great, we came to the conclusion that XG Boost was the best model that was readily available. Its recall is at a score of 0.66, its accuracy and precision are both at a score of 0.67, and its F-1 score is also at 0.66. On the other hand, if there was additional tweaking done to the models in the future, the state would be in a far better form. Regression modeling, which is one of the various forms of modeling that we do, and classification modeling, which is another sort of modeling, are the two types of modeling that we do the most frequently. It is vital for us to have an awareness of the processes that can enhance the level of precision of the model because there is always a desire for extremely exact models. You can try retrieving the estimates of the model's performance by using score metrics, finding and diagnosing the best parameter values by using validation curves, holding out the cross validation such as K fold cross validation, using the grid search to optimize the hyperparameter combination, and continuously and constantly fine tuning the parameters of machine learning models. All of these factors have the potential to contribute to an increase in the accuracy of the models.

##### References

1. C. Newberry, "16 Key Social Media Metrics to Track in 2023," HootSuite, 2022. [Online]. Available: https://blog.hootsuite.com/social-media-metrics/.
2. H. Ren and Q. Yang, "Predicting and Evaluating the Popularity of Online News.," 2012.
3. K. Fernandes, P. Vinagre and P. and Cortez, "A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News.," Progress in Artificial Intelligence, pp. 535-546, 2015.
4. K. Junli, W. Baocang, L. Caiyun and W. Gaofei, "An Approach for Predicting the Popularity of Online Security News Articles," IEEE CNS 2018 - 1st International Workshop on System Security and Vulnerability (SSV), 2018.
5. S. R, A. D.M, M. B.S, S. B and A. C.M, "Predicting Popularity of Online Articles using Random Forest Regression," 2016 Second International Conference on Cognitive Computing and Information Processing (CCIP), 2016.
6. M. Madhukumar, A. Sebastian, X. Liang, M. Jamil and M. N. S. K. Shabbir, "Regression Model-Based Short-Term Load Forecasting for University Campus Load," IEEE Access, vol. 10, pp. 8891-8905, 2022.
7. A. N. Alkawaz, A. Abdellatif, J. Kanesan, A. S. M. Khairuddin and H. M. Gheni, "Day-Ahead Electricity Price Forecasting Based on Hybrid Regression Model," IEEE Access, vol. 10, pp. 108021-108033, 2022.
8. L. Pun, P. Zhao and X. Liu, "A Multiple Regression Approach for Traffic Flow Estimation," IEEE Access, 2020.
9. "Data Preparation in Machine Learning: 6 Key Steps," SearchBusinessAnalytics, [Online]. Available: https://www.techtarget.com/searchbusinessanalytics/feature/Data-preparation-in-machine-learning-6-key-steps.
10. Jim, "Choosing correct type of regression analysis," [Online]. Available: https://statisticsbyjim.com/regression/choosing-regression-analysis/.
11. B. Carolina, "Medium," Medium, 28 June 2021. [Online]. Available: https://towardsdatascience.com/decision-tree-classifier-explained-in-real-life-picking-a-vacation-destination-6226b2b60575. [Accessed 03 March 2023].
12. Ö. ÇOKLUK\*, "Logistic Regression," Educational Sciences: Theory and Practice, vol. 10, no. 3, pp. 1397-1407, 2010.
13. guest\_blog, "An End-to-End Guide to Understand the Math behind XGBoost," Analytics Vidya, 23 September 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/.

Brownlee and J, "XGBoost for regression. Machine Learning Mastery.," 6 Novemeber 2022. [Online]. Available: https://machinelearningmastery.com/xgboost-for-regression/ .

1. A. Bandari, S. Asur, and B. A. Huberman, "The Pulse of News in Social Media: Forecasting Popularity," Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, Dublin, Ireland, 2012, pp. 10-17.
2. J. M. DiMicco, D. R. Millen, R. G. Wong, and J. A. Payne, "People Sensemaking and Relationship Building on an Online News Site," Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Boston, MA, USA, 2009, pp. 205-214.
3. A. B. Kahsay and R. Graefe, "The Impact of User Comments on the Popularity of Online News," Journalism & Mass Communication Quarterly, vol. 95, no. 1, pp. 28-49, 2018.
4. L. Liu and Y. Li, "Predicting Online News Popularity with Recurrent Neural Networks," Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, Shanghai, China, 2014, pp. 797-806.
5. S. Nagarajan, A. B. Whinston, and Y. Zhang, "Building Trustworthiness for Online News through Social Network Analysis," Decision Support Systems, vol. 77, pp. 44-54, 2015.
6. E. Panciera, A. Diakopoulos, and D. R. Potts, "Curation in Context: A Comparative Analysis of News Aggregator Use," Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vancouver, BC, Canada, 2011, pp. 1457-1466.
7. M. Scharkow, "The Accuracy of Self-reported Internet Use—A Validation Study Using Client Log Data," Communication Methods and Measures, vol. 6, no. 4, pp. 228-243, 2012.
8. J. Vosoughi, D. Roy, and S. Aral, "The Spread of True and False News Online," Science, vol. 359, no. 6380, pp. 1146-1151, 2018.
9. Y. Wang, C. Li, S. Li, Y. Li, and Q. Li, "Predicting Popularity of News Articles Using Deep Learning," Proceedings of the IEEE International Conference on Data Mining, Atlantic City, NJ, USA, 2015, pp. 1103-1108.

Y. Zhang, Y. Liu, and L. Wang, "Social Network Analysis of Online News Sources," Journal of the Association for Information Science and Technology, vol. 68, no. 9, pp. 2223-2236, 2017.