**Cyclistic Bike-share Analysis for Targeted Casual-users marketing**

**Tamojit Roy1**

*1Vellore Institute of Technology, Chennai, India*

*1E-mail:* [*tamojit.roy2019@vitstudent.ac.in*](mailto:tamojit.roy2019@vitstudent.ac.in)

**Abstract**

Although the age calls for motor-vehicles as the major shareholders of the commute industries, cyclists still continue to have impressive shares in some of the developing heavyweights and developed nations around the world. Some of the countries like Sweden, Denmark, Germany, UK, Japan and even China have a wonderful landscape for cycling. Although some of them are casual riders, quite a many are annual members for the major cycle manufacturing companies all over the world. Data analytics plays an important role in analyzing and boosting the sales of any company. It has an upper hand when it comes to implementing market plans to target the set of customers who are the most vulnerable by suggesting them various specialized schemes and membership benefits. Hence, it is no doubt one of the most potent tools helpful in boosting the sales of a product. Hence, it can play a pivotal role in increasing the annual membership of cycles of any company for good profits. This research work intensely focusses on analyzing all the major aspects and most of the if not all of the attributes of bike-share sales of a prominent bike-share company in Chicago named Divvy. This research work mainly revolves around understanding how subscribers and customers of Divvy bike-share service use bikes differently. The comparison along with other tasks have been used to design marketing strategies aimed at converting customers of the company to the subscribers of its services.

**Keywords**

Support Vector Machine (SVM), Logistic Regression, Naïve Bayes, Random Forest, Voting Classifier, F1 Score, Passive Aggressive Classifier, Count Vectorizer, Term Frequency-Inverse Document Frequency (TF-IDF)

**1. Introduction**

In today's Internet world, people rely on different online services / platforms for news. As the use of social media platforms such as Facebook, Twitter, and news websites increases, news spreads rapidly to millions of people in a short period of time. News websites publish news and provide authentication sources. The problem is how to authenticate messages and articles distributed on social media such as WhatsApp groups, Facebook pages, Twitter, other microblogging and social networking sites. Believing in rumors of pretending to be news is bad for society. The spread of fake news has widespread consequences, including the generation of biased opinions that affect the outcome of elections in favor of a particular candidate. In addition, spammers use compelling headlines to monetize their ads through Clickbait. Especially in developing countries like India, it takes time to put an end to rumors and focus on the right certified news articles. This white paper presents models and methods for detecting fake news. Social media platforms such as Facebook, Instagram, and Twitter provide a cheaper way to deliver news online much faster by spreading the network more easily. It's better than traditional news media, but it's a lot of fake news. News articles contain intentionally false information created online for a variety of purposes for economic and political gain. Therefore, there is an urgent need for a fake news detection system that not only distinguishes between fake news and real news, but also displays the relevant real news articles that are closest to the original. Since dealing with fake news requires precision and with current advances in machine learning algorithms to detect such fake news, it would be foolish to depend on a single algorithm/method to perform such complex classification. Therefore, it makes much more sense to divide the work into phases to completely separate the work by data mining operations such as data collection, data pre-processing, feature extraction, feature selection and implementation of machine learning models to perform predictions to classify the news as True or False and also predict the probability that the news belongs to the predicted label. Several machine learning models will be used to classify news as true or false. Each will be evaluated and compared against each other based on metrics like accuracy, f1 score, precision and recall. After the following machine learning models - SVM, Logistic Regression, Naïve Bayes and Random Forest will be trained and adjusted. A voting classifier will be implemented that will combine all the models mentioned above and form a composite classifier that uses all these classifiers to predict class labels and probabilities, and use soft voting to make the final prediction. A suitable dataset that has been divided into training set, validation set and test set will be used and will be preprocessed to apply feature extraction techniques in the next steps. Models like Random Forest Classifier, Naïve Bayes, Logistic Regression and SVM Classifier will be used to train the dataset. Finally, the different scores for each category will be calculated and compared. Below is a detailed flowchart that summarizes the entire process.

**2. Literature Survey**

This paper deals with analyzing and examining the various factors, which portray the bicycle-industry as a potent industry which can lead to the green recovery and sustainable development of economy and environment in Bangladesh. The authors performed a SWOT analysis after collecting and analyzing information with regards to the bicycle industry in the south-asian developing economy from many research publications, and by interviewing industry experts, goverment officials and university professors and finally a joint discussion by all the experts. Based on the findings in which both internal factors (strengths and weaknesses) as well as external factors (opportunities and threats) were taken into consideration, they came up with an internal factor evaluation matrix (IFEM) to find out that the bicycling industry in Bangladesh has enormous potential, given that some reforms be conducted upon the same with a view towards reducing carbon-emissions and decarbonizing the commute sector by facilitating green investment in bicycling as a non-motorized transport (NMT) option to reduce GHG emissions in the post-pandemic settings. Suggested potential strategies to facilitate the adoption of bicycles as a resilient and sustainable transport option such as by developing local manufacturing capability, dedicated infrastructure, reducing import duties, attracting FDIs, eliminating gender differences, etc., were also mentioned. [1]

The analysis of the pattern of bike-sharing in a region plays the most significant role in predicting the bike usage in any particular region. The authors in this paper, have provided a comprehensive review of bike-sharing usage prediction with many invaluable approaches to predict the bike-sharing usage pattern with deep learning. Following a set of procedures of the following modules: data aggregation to build the prediction input-features and targets, defining 3 data formats (time-series format, grid format and graph format), addressing 3 types of prediction problems (time series-input prediction, graph-input prediction and grid-input prediction), quantifying the prediction error using different evaluation metrics, and finally some prediction challenges (complex spatial dependencies and complex temporal dependencies) and prediction models like FFNN, LSTM, RNN, GNN, GRU, MLP, SVR, etc., were illustrated with a different section for each type of model was mentioned for both docked and dock-less bike-sharing systems. Finally, application scenarios within the bike-sharing systems and beyond were brought up followed by the challenges and the development directions in the research paper. More open datasets, various applications based on bike usage prediction and potential research directions were summarized to encourage future research. [2]

A vision of developing a sustainable bike-sharing system as viewed a PSS (product-service system) was expressed by the authors in this research paper. Although, it was initially developed considering a single focal-company, restricted to a particular region (southern region of developing country, Brazil) only, the authors emphasized the significance of their research-work claimimg that it was the first one to be done keeping in mind the scenario of a developing country. The design of the system-model organized in 4 stages: (i) value proposition, (ii) value configuration, (iii) value delivery and (iv) value capture was introduced. Although, the strategy doesn't hold the providers of the service as the manufacturers themselves, the business-model analyzed was a use-oriented in the context of shared mobility. By conducting face-to-face interviews with those involved in developing the business model, a research protocol was developed. The authors expressed that the PSS business model analyzed by them could represent significant contributions to improve micro-mobility. [3]

This paper entirely deals with understanding, analyzing and illustrating the multiple modes and forms of relationships between build environment (i.e., land-use, transportation system and urban design) and bike-share usage. Quite a many variances between the build environment and bike-usage were stated and described with some outliers in notable cases. Variance in relationship in the build environment across different mobility patterns, docked and dockless bike-share patterns, w.r.t. trip purpose, between arrival and departure patterns, based upon the day of week, etc. and the bike-share usage were elaborated. The paper concluded with a brief summary of the major findings of the authors and them encouraging the recommendations for the future research works. [4]

This paper deals with the application of Natural Language Processing techniques to identify when a news source may be producing fake news. Collection of labeled real and fake news articles to build a classifier that can make decisions about information based on the content from the collection has been implemented. A text classification approach, using four different classification models to analyze the results has been implemented. LSTM implementation was the best performing model. The model focused on identifying fake news sources, based on multiple articles originating from a source. High predictions were made when the source was flagged as a producer of fake news that future articles from this source will also be fake news. As multiple data points come from each source, it broadens the misclassification tolerance of the article. This is discussed in a precise and better way in this paper [5].

In this paper, authors looked at how Natural Language Processing techniques may be used to detect when a news source is spreading bogus news. To develop a classifier that can make judgments about information based on the content in the corpus, the authors employed a corpus of labeled actual and false new articles. A text classification approach was applied and assessed the outcomes using four distinct classification models. The LSTM implementation produced the best results. Based on many articles originating from a source, the algorithm focuses on identifying fake news sources. It is anticipated with high confidence that any future publications from a source classified as a producer of fake news will also be fake news. Focusing on sources broadens the scope of our article misclassification tolerance, because we then have multiple data points coming from each source [6].

The goal of this paper is to investigate four well-known machine learning algorithms, namely the random forest, the Naïve Bayes, the neural network, and the decision trees, to see how effective they are at detecting fake news. On a widely used public dataset, LIAR, an experiment was performed, and the findings demonstrate that the Naïve Bayes classifier outperforms the other algorithms significantly. This result is tested and verified in this paper and the authors have mentioned the metrics very justly and the work has been done optimally and utmost time has been devoted to increase the accuracy of the results in this paper [7, 12].

This paper deals with the collection of dataset from Kaggle. The information is divided into two categories: true and false news, and then integrated into a single dataset. The machine learning model is trained using this dataset. The authors attempted to develop a machine learning model utilizing four different classifiers and the Tf-idf vectorizer in this project. The goal is to foresee news that would mislead the user and cause pandemonium [8].

The authors extracted and analyzed real-time data from multiple domains across Twitter. The dataset was preprocessed and the user\_verified column played an important role. Next, some machine algorithms were run on the features extracted from the preprocessed dataset. Logistic regression and support vector machines each yielded promising results with an accuracy of over 92%. Naive Bayes and long / short term memory did not achieve the desired accuracy. The model can also be applied to images and videos for better detection of fake news. This was mentioned by the authors in their future work which could be done in a proper and better way that would lead to not only textual but also media news to be judged for correctness and false detection [9].

To distinguish bogus news from genuine news, a deep learning-based technique was applied in this study. The proposed model was created using an LSTM neural network. A gloVe word embedding was employed for vector representation of textual words in addition to the neural network. Tokenization has also been used for feature extraction and vectorization. The concept of N-grams is applied to improve the suggested model. The results of a comparison of several false news detection systems were examined. The suggested model's outcomes were assessed using accuracy measures. The model outperformed with a precision of 99.88 percent [10].

Machine learning is offered as a method for detecting bogus news. Using vectorization of the news title and our dataset to analyze the tokens of words. The dataset that the authors’ have worked with is a pre-curated collection of news items that have the property of being false or not. Their goal was to create a model that can determine whether an article is real or false. This was done using the sentiment analysis procedure and later the same was embedded into the base code and a hybrid model was designed by the authors of this paper [11].

This research looks into advanced and cutting-edge false news detection systems in depth. The authors started with the negative repercussions of bogus news. Then they talked about the dataset that was used in earlier research and the NLP approaches that were used. To categorize representative methods into several categories, a complete overview of deep learning-based techniques has been presented. The most often used evaluation measures in the detection of false news are also reviewed. Nonetheless, in future research paths, they proposed additional recommendations to improve fake news detection techniques. This could be subtly implemented inside the base code of the existing model. Then later the model can either be constructed into a hybrid system or may be divided into different modules for better modularity, so that the same can be later implemented inside an application and used in real time and be available to use within the reach of the local public [13].

**3. Problem Statement**

There are many ways to detect fake news and articles using machine learning algorithms. Given today's advances in machine learning and the level of knowledge reached, it is not appropriate to blindly rely on one or two such algorithms to classify fake news. Therefore, instead of collecting more data and extensive training to improve accuracy and comparing accuracy scores for better understanding, a holistic approach for detecting fake news is the need of this crucial moment. Therefore, we decided to take a step-by-step systematic approach to classify fake news using data mining technology. The implementation of data mining operations such as data collection, data preprocessing, feature extraction, feature selection, and machine learning models can be used to make predictions to classify messages as true or false and predict the probability of the news belonging to the predicted level. A set of machine learning models will be implemented to compare the performance of machine learning models based on metrics such as accuracy, f1 value, fit rate, and recall. The primary determinant for assessing model performance can be selected as the f1 score, which takes into account the trade-off between fit and recall. After the following machine learning models (naive Bayes, SVM, logistic regression, random forest) have been trained and tuned, all of the above models are combined to form an ensemble classifier that predicts using all of these classifiers. A voting classifier will be implemented. Label and class probabilities that use the soft voting method to make the final prediction. There are two approaches for machine learning models to gain insights from headings: CountVectorizer and TfidfVectorizer. These models are trained using features extracted from both CountVectorizer and TfidfVectorizer. All models are then hyperparameter adjusted in GridSearchCV with 5 holdout cross-validation sets for all different possible parameters. This hyperparameter adjustment is intended to improve the model's f1 score. After the models have been tuned, they are tested in the test set and the model's evaluation metrics are calculated. The trained and voted models are combined into a voting classifier that uses them all as the basic estimator. When new test data is passed to the voting classifier, all underlying models are built to predict sample specifications. After receiving the labels from all models, the final label of the test sample is predicted using the soft voting mechanism.