IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Issue-6, June 2019 E-ISSN: 2347-2693

Predicting Voting Outcomes Using Data Analytics and Machine Learning Algorithms

Urjit Desai[1*], Ameya Dalvi[2], Atharva Dhuri[3]

[1,2,3] Computer Engineering, St. Francis Institute of Technology, Mumbai University, Mumbai, India

[*]Corresponding Author: urjitdesai1998@gmail.com. Tel.: +91 9820731508

DOI: https://doi.org/10.26438/ijcse/v7i6.742745 | Available online at: www.ijcseonline.org

Accepted: 11/Jun/2019, Published: 30/Jun/2019

***Abstract*—**Voting is the right of every eligible citizen. It is the power vested upon the people which allows them to choose a party or a person who will represent them as a part of the government. On one side of the coin, are the people whereas on the other side, are the parties. Every election, a stupendous sum of money is spent by the parties in doing social work, promotion of candidates and many more such fields. Thus, it would be of strategic importance to a party, if they are able to predict the voting outcomes in an area in advance, as it can help them to carry forward their campaign judiciously. In this proposed work, a dataset from Show of Hands is used which contains multiple features, several of them hidden, which were discovered after data analytics. The aim is to correctly predict the party a person is most likely to vote for, in the USA presidential election. For this purpose, first after collecting the data, we perform data cleaning and feature extraction. Next, the data is given as input to our model. The model is trained using multiple machine learning algorithms like Logistic Regression, Support Vector Machine (SVM), Naïve Bayes Classifier and Random Forest. The accuracy of these models is compared and the prediction report is generated.

***Keywords*—**Voting, Data Analytics, Data Cleaning, Machine Learning.

I. INTRODUCTION

Voting is of utmost importance for any state and its' respective citizens. The day to day functioning, and the development of a state is greatly influenced by the ruling party. A ruling party's aim is to get the majority seats in its state as well as try to win seats in other states. For this purpose, ample amount of money is spent behind campaigning, promotions and other social events. A majority proportion of the money can be strategically used in certain other aspects instead of its haphazard usage, in order to improve the desired output of a party's actions.

The most popular methodology used for predicting voting outcomes is the exit poll- an opinion poll of people leaving a polling station, asking how they voted. But it is too late for a party to take the required measures once the voting has already taken place. Other voting prediction systems predict the voter turnout i.e. whether or not a person will cast a vote. Some systems use sentiment analysis from different social media platforms. This proposed work differs in the aspect that it aims to predict the party for which a person is most inclined to vote, thus providing the party an opportunity to channelize its efforts.

The dataset used contains data from Show of Hands, an informal polling platform used on mobile devices and the web, to see what aspects and characteristics of people's lives predict how they will be voting for the presidential election. The dataset contains 107 columns (features) which will be used to train the model and predict the stated goal. The proposed system will use data analytics for completely exploring the data, identifying relationships among those columns, finding some patterns of work, cleaning and refining the dataset to reduce complexity of data. Then, the system will extract useful features from the bulky dataset and will also derive new features for making system processing easier. Since it is a classification problem; Logistic Regression, SVM and Random Forest algorithms will be used.

Section II describes the related works done in this field and it also explains the limitations, Section III briefs about the procedures and methods followed to achieve the desired goal with block diagram, Section IV describes results and discussion, while section V concludes the research work with future scope.

II. Related Work

A project on similar lines was solved using a different approach [1]. The algorithms used in the work were Artificial Neural Networks (ANN) and Support Vector Regression (SVR). Six independent variables such as GDP, unemployment rate, the president's approval rate and others were considered in a stepwise regression to identify significant variables. Support Vector Regression proved to be superior as it successfully predicted the outcome of previous three elections.

Ashenafi predicted the party affiliation of the Congressmen as Democrats or Republican in the work [2]. KNN and Naïve Bayes algorithm were used in the prediction process. This proposed work differs from [2] in the dimension that the later predicts the vote of a congressmen whereas this work aims to predict the vote of a citizen.

Social media platforms like twitter plays a huge role in molding people's opinions. A large dataset of 370,000 tweets gathered in 2016 was carefully validated against Google Trends to create a legitimate dataset [3]. A Gaussian process regression model was used to predict the election outcome.

Voter turnout prediction is the category that has drawn most of the attention for research. Voter turnout prediction is the likelihood of an individual turning out to cast his/her vote. Research conducted in this area generally make use of Logistic regression, SVM and Naïve Bayes algorithm for the purpose of prediction.

III. Methodology

The proposed system is composed of four phases:

1. Data Input
2. Exploratory Data Analysis (EDA)
3. Machine Learning Model Selection
4. Prediction

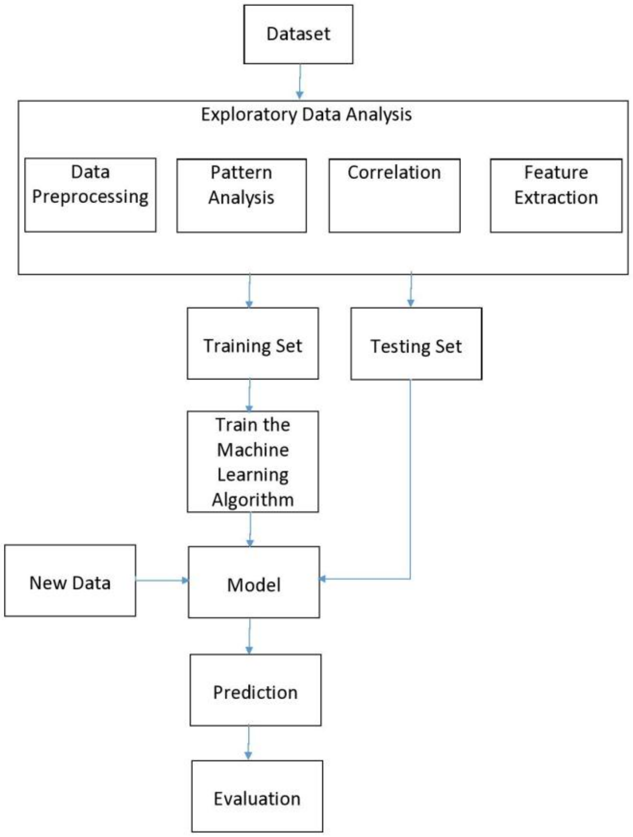


Figure 1: Block Diagram

Phase 1: Data Input
The input data file is in a .csv file which will be provided to the system, containing 5569 rows (records). This data contains 107 columns from which important features are extracted after performing data analytics. The data is not in a suitable form for directly using it, and thus needs to be cleaned before training the machine learning algorithm. Thus, exploratory data analysis has been performed in order to convert data into a suitable format.

Phase 2: Exploratory Data Analysis
The initial data available contains numerous null values, continuous and textual data which need to be handled before using it. There are different ways of handling null values; like discarding the rows that contain a lot of null values, performing mathematical imputations, randomly assigning values to the null fields, etc. The dataset used in this proposed work contains several null values and thus discarding them will do more harm than good. The following steps were implemented to get the data in a desired format:

1. **Feature selection**: Out of the 107 columns, not all were relevant in determining the aim of the work. Features for the model were selected on the basis of knowledge about the USA presidential election, self-understanding and correlation matrix obtained from data analytics. For example, questions like- 'Does the weather have a large effect on your mood?', 'Do you wear glasses or contact lenses?', 'Do you take a daily multivitamin?' have nothing to do with which party a person might vote for. On the other hand, a person's income, age, gender, education and questions like- 'Do/did you get most of your K-12 education in public school, or private school?', 'Do you currently have a job that pays minimum wage?', 'Are you currently employed in a full-time job?' have a significant impact on a person's vote. In the end, 17 features were shortlisted for training the model.
2. **Data imputation**: The final features taken into consideration still had missing values, which needed to be filled. For demographics related data such as- the records having their year of birth as null values, the average of years of births' of all records was imputed for it. For the records having null values for their range of income, their income was imputed based on their educational qualifications. For the questions that had not been answered, the response was considered to be as negative i.e. zero.
3. **Data conversion**: In this phase, continuous and textual values were converted into categorical values like age being divided into 5 age groups, the income into 5 classes, etc. For columns having binary values, dummies were created. Thus, finally the data was converted into a format suitable for training the model.

Phase 3: Machine Learning Model Selection

Machine learning is a science that uses statistical techniques to give computer systems an ability to learn from the given dataset without being explicitly programmed. Since the data available was well-labelled, supervised learning algorithms were used for prediction. As the task in hand was a classification problem, the following machine learning algorithms were implemented:

1. Logistic Regression
2. Support Vector Machine
3. Random Forest
4. Naïve Bayes Classification

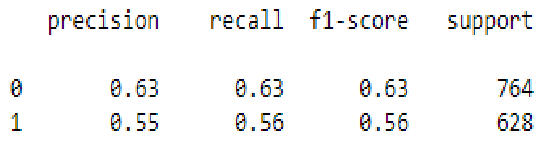
Phase 4: Prediction and Evaluation

The data was divided into training and testing data. 80% of the data was used for training and the rest, for testing. After the model was trained with the training set, the test set was given as input to the model. The model generated the predicted output, and this output was compared with the actual output. Classification reports were generated for each of the algorithms, used to evaluate their performances.

IV. Results and Discussion

In this proposed work, there was an attempt to predict the party a person is likely to vote based on Show of Hands data. On giving the testing data as input to the model, the following results as shown in the figure were obtained. Below are the classification reports for the different algorithms used where 0 represents the Republican while 1 represents the Democrat party. The columns of the classification report answer the following:

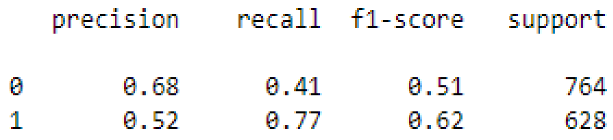
1. Precision: What proportion of positive identifications was actually correct?
2. Recall: What proportion of actual positives was identified correctly?
3. F1-score: It is a measure of the test's accuracy. It is calculated as the harmonic mean of precision and recall.
4. Support: It is the number of occurrences of each class.



```
          precision    recall  f1-score   support

0              0.63      0.63      0.63       764
1              0.55      0.56      0.56       628
```

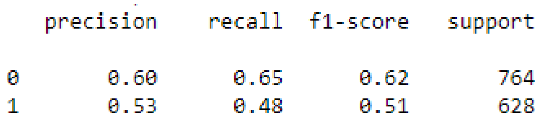
Figure 2: Logistic Regression



```
        precision    recall  f1-score   support

0            0.68      0.41      0.51       764
1            0.52      0.77      0.62       628
```

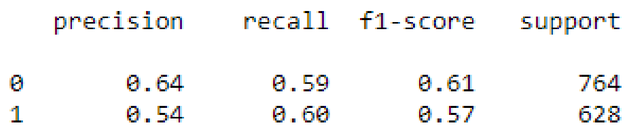
Figure 3: Naïve Bayes Classifier



```
          precision    recall  f1-score   support

0              0.60      0.65      0.62       764
1              0.53      0.48      0.51       628
```

Figure 4: Random Forest



```
     precision    recall  f1-score   support

0         0.64      0.59      0.61       764
1         0.54      0.60      0.57       628
```

Figure 5: Support Vector Machine

It can be seen from the classification reports that Logistic Regression and SVM have similar results.

V. Conclusion and Future Scope

From the results obtained, it can be concluded that although none of the algorithms serve as a satisfactory classifier, Logistic Regression and SVM have comparatively better performances among the four. It can also be concluded that the data obtained by Show of hands method is not sufficiently reliable. One of the shortcomings of the work is that several null instances are considered to be 0 (zero) which might not be the ideal case. Also, if more data is available and more detailed imputation of values is performed,then it can improve the accuracy furthermore.

This proposed work is a generic approach and the idea behind it can be used to predict voting outcomes not only in USA but also in India, given that an appropriate dataset is available. The approach can be used for prediction purposes in fields other than voting as well. In future, such surveys can be taken at regular intervals on a larger scale, so that a party can monitor the outcome of its efforts and take necessary actions to channelize its money and manpower in required areas.

References

[1] Zolghadr, M. Niaki, S.A. Niaki, "*Modelling and Forecasting US Presidential Election using learning algorithms*", International Journal of Industrial Engineering, Vol.14, Issue.3, pp.491-500, 2018.

[2] A. Wakjira, "Predicting voting Affiliation Using Machine Learning Algorithms ", Metropolia Ammattikorkeakoulu Publisher,2014

[3] P. Kassraie, A. Modirshanechi and H. Aghajan, "*Election Vote Share Prediction using a Sentiment-based Fusion of Twitter Data with Google Trends and Online Polls.",* In the proceedings of the 6th International Conference on Data Science, Technology and Applications (DATA 2017),pp.363-370.

[4] P. Salunkhe, S. Deshmukh, "*Twitter Based Election Prediction and Analysis* ", International Research Journal of Engineering and Technology (IRJET), Vol.04, Issue.10, pp.539-544, 2017.

[5] Marie Fernandes , "*Data Mining: A Comparative Study of its Various Techniques and its Process*", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.1, pp.19-23, 2017.

[6] A. Jenita Jebamalar, "*Efficiency of Data Mining Algorithms Used In Agnostic Data Analytics Insight Tools*", International Journal of Scientific Research in Network Security and Communication, Vol.6, Issue.6, pp.14-18, 2018.

[7] K. Sree Divya, P. Bhargavi, S. Jyothi, "Machine Learning Algorithms in Big data Analytics", International Journal of Computer Sciences and Engineering, Vol.6, Issue.1, pp.63-70, 2018.

[8] WEKA Manual for Version 3-6-8, The University of Waikato, 2012.

Authors Profile

Mr. Urjit Desai is currently pursuing a Bachelor of Engineering from St. Francis Institute of Technology, India; since 2016. He is an active member of IEEE and CSI since 2017. He has worked on several projects in the field of Data Analytics, Machine Learning, DBMS, Data Structures and Algorithms. He has completed two internships and aspires to pursue Ph.D. in the field of Data Science. His research work focuses on Data Mining, Artificial Intelligence, Big Data Analysis and Machine Learning.

Mr. Ameya Dalvi is currently pursuing a Bachelor of Engineering from St. Francis Institute of Technology, India since 2016. He is an active member of CSI since 2018. He has worked on projects in the field of Machine Learning, DBMS, Data Structures and Web Frame Development . He aspires to pursue Masters. in Computer Science with a specialization in the field of Data Science. His research work focuses on Data Mining, Big Data Analysis and Machine Learning.

Mr. Atharva Dhuri is currently pursuing a Bachelor of Engineering from St. Francis Institute of Technology, India since 2016. He is an active member of CSI since 2017. He has worked on projects in the field of Machine Learning, DBMS, Data Structures and Data Analytics. He aspires to pursue MSc. in the field of Data Science. His research work focuses on Data Mining, Artificial Intelligence, Deep Learning, Big Data Analysis and Machine Learning.