International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-6, April 2020

Important Feature Selection for Predicting Human Freedom Index Score using Machine Learning Algorithms

Poonam Kumari, S Prasad Babu Vagolu, Sunil Chandolu

Abstract*— Human freedom index refers to the state of human freedom in various countries based their personal and economic attributes. Human freedom can help us identify nobility of citizens in a country. For an individual of a country freedom is of great value and hence it is worthy to measure. Though there are many attributes to measure the human freedom index both in personal as well as in economic factors, here we are interested to find only those features which contribute the most and are relevant to predict the outcome i.e. human freedom index score. We will go through various features engineering process like removing strongly correlated attributes, filtering method using Mutual Information (Entropy) and then use Select KBest algorithm to select top features filtered through Mutual information. These steps will help reduce the training time, increase accuracy and reduce overfitting when model is created to predict the human freedom index score which is a Machine Learning Regression problem.

Index Terms— Corelated Feature, Decision Tree Regression, Feature Engineering, Human Freedom Index, Linear Regression, Machine Learning, Mutual Information, Random Forest Regression, Regression, SelectKBest.

I. INTRODUCTION

Human Freedom is the term that refers to the degree to which the citizens of a country are truly free to enjoy their civil liberties such as freedom of speech, freedom of movement, safety and security, etc.[1,2,3]. These liberties are divided into personal and economic freedom. Personal freedom includes procedural justice, civil justice, criminal justice, women's security, etc. Economic freedom includes government consumption, judicial independence, impartial courts, etc. There are so many features that influences the prediction of human freedom index score for a country. But not all features are of high importance. If we use all the feature that can influence the prediction, it will increase the training time of the model and will make the model more complex.

Revised Manuscript Received on April 30, 2020.
***** Correspondence Author
Poonam Kumari*, is currently pursuing Master of Computer Applications in Department of Computer Science,GIS, GITAM (Deemed to be University), Viskahapatnam, India, E-mail: poonamkgupta71@gmail.com
S Prasad Babu Vagolu is currently an Assistant Professor in Department of Computer Science, GIS, GITAM (Deemed to be University), Viskahapatnam, India, E-mail: prasadray@live.com
Sunil Chandolu is currently an Assistant Professor in Department of Computer Science,GIS, GITAM (Deemed to be University), Viskahapatnam, India, E-mail: kunny0306@gmail.com

Although using many features can help us get more accurate data, but still it will take a lot of time to predict the outcome. So, the goal is to remove highly correlated attributes and to find only those features that will influence the predicting model the most. So that minimum number of features we can get the most accurate prediction using any machine learning regression algorithm.

Less the number of features to consider for the prediction model less will be the training time and more will be the accuracy as all unrelated feature are removed and not considered. This will in turn also reduce the overfitting of the model.

II. PROCEDURE FOR PAPER SUBMISSION

The human freedom index dataset, collected from kaggle.com has many economic freedom and personal freedom features which basically revolves around rules of law, security and safety, movement and religion factor of the country. There are total 79 different indicators that contribute towards predicting human freedom index score, but not all of these features are of great importance for the prediction. So the goal is to reduce the number of correlated data and filter to select only those feature which highly influence the prediction of the outcome. This will help us increase the accuracy with less number of features and will reduce training time so that model can be trained with less attributes and hence reduce the complexity involved due to high degree of features involved.

Once only the highly important feature is filtered out of the lot. We can use various machine learning regression algorithms like linear regression, decision tree regression and random forest regression to find which algorithm give the maximum accuracy and then parameter tuning can be done to increase the accuracy further.

This will give us a human index score prediction model which consist of only important feature, will take less predicting time and will provide the maximum accuracy for the predicted data.

III. PRE-PROCESSING THE DATASET

Firstly, we need to prepare the dataset for applying various algorithms.

A. *Cleaning the Data*

Dataset can have many missing or noisy data which need to be handled before going further. Missing values can be handled using Imputer package of sklearn .preprocessing [7],

This package will help us fill the missing values using one of the following methods:

- Using mean along the axis.
- Using median along the axis.
- Using most frequent along the axis.

For this problem statement mean value along the axis is used.

B. *Data Reduction*

The data reduction involves attributes subset selection, Dimensionality reduction. This can also be referred as feature engineering.

C. *Feature Engineering*

Let's discuss various methodology using which we can reduce the number of features.

Find correlated attributes and remove: It is important to determine and compute the degree to which features in dataset are reliant on each other. This understanding can help enhance the data to meet the potential of any machine learning algorithms, such as linear regression, whose performance will reduce with the existence of these interdependencies [4]. The statistical relationship between two features is referred to as correlation.
Basically, Correlation are of three types:

- Positive Correlation: both attributes changes in same direction
- Neutral Correlation: No relation in the change of attributes values.
- Negative correlation: both attributes changes in opposite direction.

The attributes which are highly correlated can be removed as it provides almost same information which other attributes provide.
Using heat map, we can find the features which are highly correlated. Here we are interested to remove those features which are negatively or positively correlating to 85 % or more. Once the correlated features are found then it can be removed from the dataset. In this observation we found total 17 correlated feature which is removed from dataset.

Feature selection based on Mutual Information (Entropy): Mutual information is a filtering method which use an entropy measure called mutual information to find which features should be included in the reduced data set XS[3].
Mutual information is a degree concerning two random variables X and Y, that measures the quantity of information gained about one random variable, through the other random variable[8]. The mutual information is calculated using (1):

$$I(X;Y)=\int_X \int_Y p(x,y) \log (p(x,y)/p(x)p(y)) dxdy \quad (1)$$

Where p(x,y) refers to joint probability density function of X and Y, p(x) and p(y) refers to marginal density functions. The mutual information defines how similar the joint distribution p(x,y) is to the products of the factored marginal distributions. If X and Y are wholly unrelated i.e. independent,
then p(x,y) would equal p(x)p(y), and this integral would be zero.
When it comes to feature selection, the goal is to maximize the mutual information between the subset of selected feature XS and the target variable y.
Formula (2) can be use:

$$\tilde{s}=\arg\max_{|S|=k} [\![I(x\_s;y), \quad \text{s.t. } |S|=k]\!] \quad (2)$$

Here k refers to the number of features we want to select out of the lot. This quantifier is referred to as a joint mutual information and maximizing the quantifier can be done using NP hard Optimization problem approach as the set of possible combinations of features grows exponentially. In this observation, using mutual information the features can be arranged with its importance factor as:

pf_score	1.326986
pf_rol	0.836149
ef_trade	0.704244
ef_trade_tariffs_mean	0.688765
pf_ss	0.674518
pf_expression_influence	0.664561
pf_rol_procedural	0.657653
ef_trade_movement_visit	0.609997
ef_trade_regulatory_compliance	0.574722
ef_legal_military	0.552688
pf_rol_civil	0.545516
ef_trade_tariffs	0.545424
ef_legal_integrity	0.522093
ef_legal_enforcement	0.515596
ef_legal_gender	0.504320
ef_regulation_business_compliance	0.485062
ef_trade_tariffs_revenue	0.481474
ef_regulation_business	0.473839
pf_movement	0.469811
ef_regulation_business_bureaucracy	0.467554

These are the top filtered features which are filtered using mutual information.
SelectKBest method to select top feature using Mutual information: Now using SelectKBest algorithm we can select the top most important feature which we found using the mutual information filtering method. If all the feature is considered and linear regression algorithm is used, it gives 100% accuracy but, it consumes a lot of time that is 0.029918909072875977s as so many feature are used it also take time to train the data set. So even though it is giving 100% accuracy, still we need to reduce the number of feature it uses in order to reduce the training time and still have a good accuracy.

Now by using the mutual information we have found the important feature still we need to select the most important one of all. So using SelectKBest algorithm we can select the top k important feature of the mutual information filtered algorithm

Suppose we give k=50 then top 50 most important feature is selected and then we can give those features to the algorithm to predict the accuracy and the time taken to train the model. This help us find the best accuracy and also minimum number of feature. Once the number of feature is reduced we can apply different regression algorithm to predict the most accurate score. Here, 22 most important feature is selected by giving k=22, so the total number of features which was 114, is now reduced to 22.

IV. REGRESSION ALGORITHMS

Now since we have reduced the features, now we can apply different regression algorithms. Here three type of regression algorithms are used.

A. Linear Regression

Linear regression model the connection between two features by fitting a linear equation to given data [2].

One of the features is measured as explanatory variable, and the other as dependent feature. A scatterplot can be used to determining the strength of the connection between two features. If no association is observed between explanatory and dependent features, then fitting a linear regression model will not give a productive model. A valued numerical measure of association between two features is referred as correlation coefficient, which has value between -1 and 1 which indicate the forte of the connection between two features. A linear regression line is calculated using Y = a + bX, where X is the independent feature and Y is the dependent feature, b is the slope of the line, and a is the intercept.

Using 22 most important feature gives 0.99 i.e. 99% accuracy which is a good accuracy.

B. Decision Tree Regression

Decision Tree is a decision-making tool which uses a flowchart-like tree structure. We can also say that it is a model of decisions and uses all the possible results, which includes results, input costs and effectiveness. Decision-tree algorithm falls under the class of supervised learning algorithms. It is uses with both continuous and categorical output feature. Decision tree make regression model in the arrangement of a tree structure. It breaks down a dataset into smaller subsets and at the same time associated decision tree is incrementally build. The outcome is a tree with decision nodes and leaf nodes. A decision node has two or more branches individually representing values for the attribute verified. Leaf node signifies a result on the numerical target.

Using 22 most important feature gives 0.96 i.e. 96% accuracy which is not better that linear regression. Training time of the algorithm is 0.002957582473754883s.

C. Random Forest Regression

A Random Forest is technique which is capable of executing both regression and classification activity with by using multiple decision trees and using method called Bootstrap Aggregation also called bagging. The elementary idea is to combine various decision trees to determine the concluding output rather than depending on single decision tree. The random forest model is a sort of additive model that predict the outcome by merging decisions of various decision tree. More correctly we can write this class of models using equation (3).

$$g(x)=f0(x)+f1(x)+f2(x)+... \quad (3)$$

Where the model g is the sum of various base models fi. This technique to use multiple models for better prediction is known as model ensembling. In random forest regression each base models are created separately using distinct subsample of the data.

Using 22 most important feature gives 0.98 i.e. 98% accuracy which is not better that linear regression, but better than decision tree regression.

V. RESULTS

When all the features are considered in linear regression algorithm it gives 100% accuracy but it takes training time of 0.029918909072875977s which is a lot of time. Then feature engineering is used to remove features which are 85% or more correlated to each other. Then using mutual information filtering method features are arranged according to importance in prediction the human freedom index score.

Then using SelectKBest algorithm top 22 feature are found and then given to different regression algorithm. Which gave the following accuracy.

1. **Linear Regression: 99%**
2. **Decision Tree Regression: 96%**
3. **Random Forest Regression: 98%**

Therefore, Linear Regression gives the maximum accuracy. Using feature engineering we have reduced the 114 feature to 22 most important features.

Comparing Training time gives the following observation:

1. **Linear Regression before feature engineering: 0.029918909072875977s**
2. **Linear Regression after feature engineering: 0.002957582473754883s.**

After feature engineering the training of the model take much less time than before. Hence it gives good accuracy and reduces the training time

VI. CONCLUSION

It is concluded that, when feature engineering is done on human freedom index dataset, it helps to reduce the number of features to only features which are of high importance. Once the feature engineering is done it helps the machine learning regression algorithms to train the data in a lesser time than before. When model was trained before feature engineering it took more time than after feature engineering was done.

For future scope different feature engineering algorithms can be used to reduce the features even more and give high accuracy. So that, the training time of the model can be reduced further and still giving accurate results.

ACKNOWLEDGMENT

The authors wish to thank Dept of Computer Science, GIS, GITAM, Visakhapatnam for their support.

REFERENCES

1. Porčnik, Tanja & Vásquez, Ian. (2015). The Human Freedom Index: A Global Measurement of Personal, Civil, and Economic Freedom.
2. Porčnik, Tanja & Vásquez, Ian. (2018). The Human Freedom Index 2018: A Global Measurement of Personal, Civil, and Economic Freedom.
3. Porčnik, Tanja & Vásquez, Ian. (2019). The Human Freedom Index 2019: A Global Measurement of Personal, Civil, and Economic Freedom.
4. Rong Rong, Shen & Bao-wen, Zhang. (2018). The research of regression model in machine learning field. MATEC Web of Conferences. 176. 01033. 10.1051/matecconf/201817601033.
5. Vergara, Jorge & Estevez, Pablo. (2014). A Review of Feature Selection Methods Based on Mutual Information. Neural Computing and Applications. 24. 10.1007/s00521-013-1368-0.
6. Hall, Mark. (2000). Correlation-Based Feature Selection for Machine Learning. Department of Computer Science. 19.
7. Grzymala-Busse, Jerzy & Grzymala-Busse, Witold. (2005). Handling Missing Attribute Values. 10.1007/0-387-25465-X_3.
8. Mordohai, Philippos & Medioni, Gérard & Fua, Pascal & Ross, Arun & Soh, Jung & Deravi, Farzin & Triglia, Alessandro & Bazin, Alex & Roli, Fabio & Bigun, Josef & Roui-Abidi, Besma & Abidi, Mongi. (2009). Mutual information. 10.1007/978-0-387-73003-5_971.

AUTHORS PROFILE

Poonam Kumari is pursuing her Master's degree in Computer Application's from Department of Computer Science, GIS, GITAM. She is passionate about research wotrk in Machine Learning and its Applications, Artificial Intelligence, and Cloud Computing.

S.Prasad Babu Vagolu is an Assistant Professor in Department of Computer Science, GIS, GITAM. He has 8 years of Software Development Experience and 10 years of Teaching Experience. He is CSI lifetime member and passionate about research in Wireless Mesh Networks, Machine Learning and Artificial Intelligence.

Snuil Chandolu is an Assistant Professor in Department of Computer Science, GIS, GITAM. He has 8 years of Teaching Experience. He is passionate about research in Wireless Networks, Machine Learning and Artificial Intelligence.