International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-4, April 2019

Bias Detection in Predictive Models Using Fairml

Vijay Kumawat, Vaibhav Bangwal, Lavanya K

Abstract. In Machine Learning, predictive models are used in decision making processes. Policy- makers, auditors and end users have concern regarding the prediction model whether these predictive models are making bias/unfair decision or not. There is the possibility that model can generate wrong decision due to bias. This bias can be intentional or unintentional discrimination due to some of the features present in the dataset. Bias arises in many industries like Banking, Housing, Education, Finance, Insurance, etc. which uses AI model for prediction. If the significance of the feature is high and also the feature is considered as protected attribute, namely race, religion, gender, then the feature can possibly contribute to bias in the prediction. To deal with this problem FairML model could help us . FairML is a framework that is put to use to discover bias in the predictive ML models. Basically it consists of four ranking algorithms (Iterative orthogonal feature projection (IOFP), Minimum Redundancy, Maximum Relevance (mRMR), Lasso Regression, Random forest) which helps in finding the significance of the features. FairML ranking algorithms handles both linear and non-linear dependencies.

In this paper we have studied different feature algorithm for different prediction models in order to get the significant features as prediction models are used in every field.

Keywords: IOFP, mRMR, LASSO, FairML, Bias ,Variable Ranking, Feature Significance.

I. INTRODUCTION

In this paper, we are comparing the different predictive models on the basis of Bias and detect which attribute is helping in to increase the bias. The main reason for the bias is the presence of bias in the training data set as the dataset may contain missing values or the dataset is oversampled. Due to these bias model gets fail to capture the required regularities.

Bias can be understood in terms of the following:

- Lack of suitable set of features- In this case mode is said to be underfitted.
- Lack of suitable data set- In this case it does not care about whether appropriate feature is present or not, bias arises due to lack of the appropriate dataset.

Attributes in the dataset like Race, Gender, Religion, Color, Age, Martial status, etc. are considered as protected attributes which may result in bias. One need to pay attention when these attributes are present in the dataset. Bias could be intentional or unintentional discrimination in many industries like Educational, Banking, Fraud, Insurance, etc.

Manuscript published on 30 April 2019.
***** Correspondence Author (s)
Vijay Kumawat B-Tech in Computer Science Engineering(3 rd Year) in VIT University, Vellore, (Tamil Nadu) India
Vaibhav Bangwal, B-Tech in Computer Science Engineering(3 rd Year) in VIT University, Vellore, (Tamil Nadu) India
Dr. K.Lavanya Associate Professor in the School of Computer Science and Engineering(SCOPE) in VIT, Vellore , (Tamil Nadu) India

FairML is a framework which examines the predictive models in the machine learning. We use FairML in order to detect which features contributes more towards the bias. FairML consists of the ranking algorithm which are used to calculate the significance of the features. Ranking algorithms are discussed in the methodology.

II. THEORY

1.1 FairML

FairML is a framework which helps in finding the relative significance to detect the bias in the ML models. It consists of the four ranking algorithm which helps to calculate the relative significance of the linear and non-linear model. Ranking algorithms are discussed in further sections.

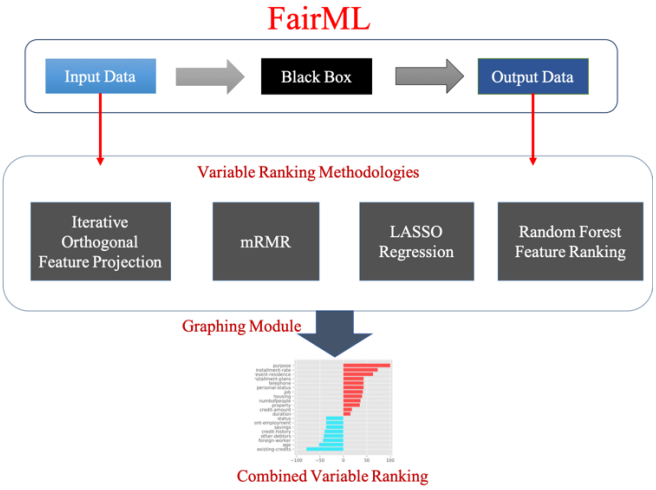

Fig. 1. FairML Schematic Diagram

1.2 Data and Input Processing Module

First of all, we need to perform different types of test such as detecting and handling of the missing values, checking for the duplicate values and converting the data into the categorical value. The aim of data and input processing module is to validate both input and output data and make it in the form of matrix as this matrix is input for the various ranking algorithms.

1.3 Methodology

The output of the data and input processing module acts as the input to the different ranking algorithms. The ranking algorithms are Iterative orthogonal feature projection (IOFP), Minimum Redundancy, Maximum Relevance (mRMR), Lasso Regression and Random Forest.

These algorithms are appropriate for the ranking the inputs to predictive models of machine learning. LASSO and IOFP handles the linear dependencies between the input variables while mRMR and Random Forest handles with the non-linear input variables. Every ranking algorithm, out of the four mentioned ranking algorithms, generates a set of score. The four sets of scores which are obtained from the ranking algorithms are clustered into a combined score for every attribute. This combined score is used for final ranking of an input attribute.Ranking algorithms are discussed in this section.

1.3.1Iterative Orthogonal Feature Projection(IOFP)

This algorithm is used to evaluate the relative significance of the inputs to the prediction models. Protected feature attributes are selected. One protected attribute is removed from protected attributes and the other features are transformed orthogonal to the removed feature. Prediction is calculated of transformed data and compared to the initial prediction of the original data. For all protected attributes this step is repeated. If the difference between the significance of the both orthogonal and initial data is high then it can lead to have bias. That feature requires more attention while prediction as this could bring some discrimination while prediction.

1.3.2 Minimum Redundancy, Maximum Relevance.

This methodology deals with the mutual information between target variable and input feature. Higher the mutual information with the target variable maximum the relevance in the prediction model. It helps to remove the insignificant feature for detecting bias in the prediction model. This technique can deal with the high dimension dataset.

1.3.3 Least Absolute Shrinkage Selection Operator.

This technique deals with the shrinking of the coefficient of the features. The coefficient of the insignificant features is set to be zero and helps to confirm the significant features. This correlation signifies the strength of the relationship between explanatory variable and the outcome of the black box. Prediction accuracy may be improved due to shrinking of the insignificant features.

1.3.4 Random Forest.

This technique used for determining the feature importance. It combines different decision tree models like ID3, MARS, CART, etc. and deploys two method to calculate importance. First method uses depth of the and attribute in decision tree as a measure of importance. Second method uses permutation to quantify variable significance.

1.3.5 Dataset Description

The dataset is taken from the census bureau database. Score generated from different ranking algorithm is normalized

Preprocessing of the dataset.

The adult dataset contains many of the missing values. So, the missing values are replaced by the mode of the column. As the FairML need all the values numerical, so the string type data was converted to the numerical type. The attributes which were converted form the string type to the integer type are work class, marital status, occupation, relationship, race, sex, native country and the label. The values assigned to the different types are given below.

Workclass

Private = 1, Self-emp-not-inc =2, Self-emp-inc=3, Federal-gov=4, Local-gov=5, State-gov=6, Without-pay=7, Never-worked=8.

Marital-status

Married-civ-spouse = 1, Divorced=2, Never-married=3, Separated=4, Widowed=5, Married-spouse-absent =6, Married-AF-spouse=7.

Occupation

Tech-support =1, Craft-repair =2, Other-service=3, Sales=4, Exec-managerial=5, Prof-specialty=6, Handlers-cleaners=7, Machine-op-inspct=8, Adm-clerical=9, Farming-fishing=10, Transport-moving=11, Priv-house-serv=12, Protective-serv=13, Armed-Forces=14.

Relationship

Wife =1, Own-child=2, Husband=3, Not-in-family=4, Other-relative=5, Unmarried=6.

Race

White =1, Asian-Pac-Islander=2, Amer-Indian-Eskimo=3, Other=4, Black=5.

Sex

Female = 0, Male =1.

Fnlwgt (salary)

>50K = 0 ,<=50K=1

Native-country

United-States =1, Cambodia =2, England=3, Puerto-Rico=4, Canada =5, Germany=6, Outlying-US(Guam-USVI-etc)=7, India=8, Japan =9, Greece =10, South =11, China=12, Cuba=13, Iran=14, Honduras=15, Philippines=16, Italy=17, Poland=18, Jamaica=19, Vietnam = 20, Mexico=21, Portugal=22, Ireland=23, France=24, Dominican-Republic=25, Laos=26, Ecuador=27, Taiwan=28, Haiti=29, Columbia=30, Hungary=31, Guatemala=32, Nicaragua=33, Scotland=34, Thailand=35, Yugoslavia=36, El-Salvador=37, Trinadad&Tobago=38, Peru=39, Hong=40, Holand-Netherlands=41.

Other attributes have the numerical values so we do need to alter those.

1) Prediction Models.

***a)* SVM.**

Support Vector Machine (SVM) is supervised learning machine, which classifies the labeled data into categories. SVM works well when there are only two categories. To classify the data, the separation line is calculated. For the separation line few parameters are set i.e. kernel, regularization, margin, and gamma. Kernel is the part of the SVM where there are some linear algebra functions which transforms the non-linear input data into the required linear form for the classification.

***b)* Random Forest.**

Random forest is the supervised learning algorithm which uses weak models such as decision trees as the learning algorithm. Random forest is the example of bagging. Different decision trees are formed from the random selected training data. The accuracy of all the decision trees are calculated and the highest weight is given to the decision tree with maximum accuracy.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

c) **Logistic Regression.**

Logistic Regression Model (LRM), is used when the output is categorical e.g. the news is fake or real. There are different types of LRM i.e. binary, multinomial and ordinal. LRM predicts the probabilities using the logistic function. With the small training data, the model can over fits. In LRM the features are correlated.

d) **Naïve Bayes.**

Naive Bayes is the simplest supervised learning algorithm. Naive Bayes calculate the probabilities for each feature. The features are assumed independent

1.4 Graphical Representation of the feature dependence

All these four predictive models are used for the prediction and significant features are obtained for all the predictive models.

Importance is calculated using the feature selection algorithm and score is normalized so that it can take values between -100 to 100. A bar plot has been made for each prediction model for the features' score which is generated by the IOFP ranking algorithm.

a) Logistic Regression Model.

Table 1. Feature Importance in Logistic Regression

Feature	Importance
Age	0.9192465901710327
Workclass	0.9177620387839049
fnlwgt	0.8547304611387746
Education	0.9191538057093372
Marital status	0.9151640738564315
Occupation	0.9144217981628676
Relationship	0.9184733863235703
Race	0.9170506912442397
Sex	0.6068413076423468
Capital-gain	0.005010360931555996
Capital-loss	0.0244332415798101
Hours	0.9178238950917019
Native Country	0.87310178455448

Fig. 2. Feature dependence in Logistic Regression model

b) Naïve Bayes

Table 2. Feature Importance in Naïve Bayes

Feature	Importance
Age	0.8700089691646306
Workclass	0.870039897318529
fnlwgt	0.8090186496768008
Education	0.8700089691646306
Marital status	0.8700089691646306
Occupation	0.8700089691646306
Relationship	0.8700089691646306
Race	0.8700089691646306
Sex	0.5701295889648347
Capital-gain	0.0832895184486438
Capital-loss	0.04633037453994371
Hours	0.8700089691646306
Native Country	0.8700089691646306

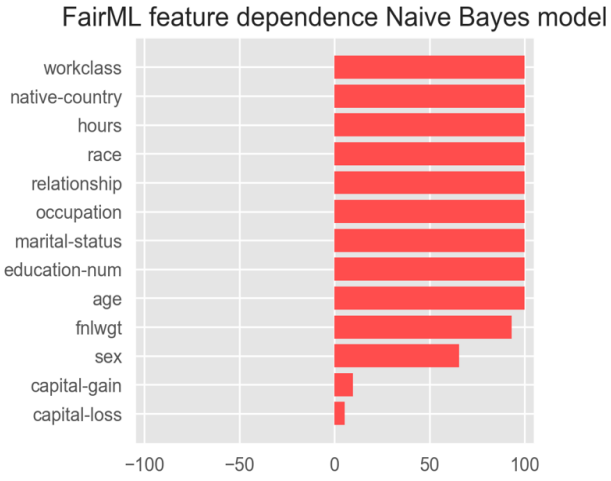

Fig.3. Feature dependence in Naïve Bayes Model

c) Support Vector Machine

Table 3. Feature importance in SVM

Feature	Importance
Age	0.9192465901710327
Workclass	0.9177929669378034
fnlwgt	0.8547304611387746
Education	0.9191538057093372
Marital status	0.9176692543222095
Occupation	0.9147001515479541
Relationship	0.9186280270930628
Race	0.9170506912442397
Sex	0.6068413076423468
Capital-gain	0.005010360931555996
Capital-loss	0.025051804657779977
Hours	0.9178548232456004
Native Country	0.8732254971700739

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

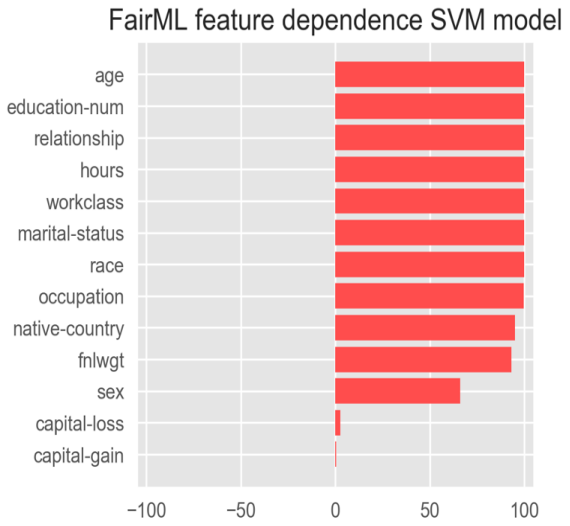

Fig. 4. Feature dependence in SVM Model

d) Random Forest

e) Table 4. Feature importance in Random Forest Model

Feature	Importance
Age	0.13979525562119197
Workclass	-0.1839297312343426
fnlwgt	0.10042371570840937
Education	0.04444375715213559
Marital status	-0.12092908174311075
Occupation	0.38162249095351497
Relationship	0.12949618037299354
Race	-0.1556304704172208
Sex	0.05597995855627377
Capital-gain	0.07936164290353509
Capital-loss	0.043701481458571736
Hours	0.04172207960906813
Native Country	-0.04997989669996598

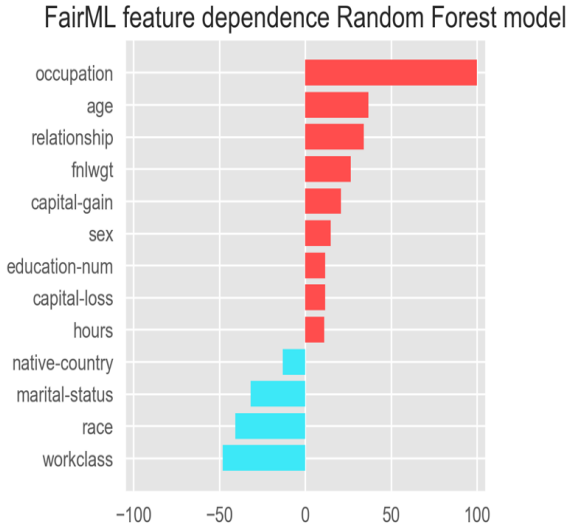

Fig. 5. Feature dependence in Random Forest Model

III. DISCUSSION

For Logistic Regression Model , SVM Model and Naïve Bayes model, most of the features given in the dataset have high dependence, i.e. they contribute more in predicting whether the income of an individual exceeds $50k/year or not. It can also be inferred that if the variable has higher significance it can possibly contribute more to the bias in the results of the predictive models. In the logistic regression, Naïve Bayes and SVM models, most of the variables of the data set have high feature dependence, so there can be a greater chance of bias in the prediction as the features like race ,marital-status and sex should not be given higher significance while predicting the income of an individual.

In random forest model occupation is more significant as compared to other features. Other features in this model are less significant so these features may contribute less to the prediction results of the models hence less contribution to the bias. Attributes native-country, marital status, race and workclass have the negative significance which means they do not contribute to the prediction model.

So manually we need to pay more attention to those attributes which are more significant to the prediction model as these contribute more in prediction model. There could be possibly bias occur because of high variance in the data which can lead to intentionally or unintentionally discrimination while the decision making.

From the above result we can say that Random Forest model's prediction will have less unfair decision in comparison to the other three prediction models. Random Forest prediction model will require less time for bias checking and will give high prediction accuracy in compare to other models as discussed in this paper.

IV. CONCLUSION

In this paper we focused on the dependence of the attribute on the different prediction model. Significant features are calculated which contribute more to the result of the predictive model. Bias can occur in any prediction model. Mostly bias occurs due to the large variance in the data and over- sampled of the dataset. Bias can be reduced to a certain limit. After this limit if we try to reduce bias any further then there is the possibility of increasing bias in the prediction model which will lead in unfair decision-making process. FairML helps in assessing the prediction model for variable significance and hence bias Policy makers and end user will get benefit in decision making process.

Bias can be mitigated using some machine learning algorithm in order to improve the prediction accuracy of the predictive models. IBM is working on AI 360 which will help in mitigating the bias to a certain limit using ML algorithms.

REFERENCES

1. Adebayo, J. A. (2016). FairML : ToolBox for Diagnosing Bias in Predictive Modeling by, (2012).
2. Uri, D., & Bellows, M. (2018). Exploration Of Classifying Sentence Bias In News Articles With Machine Learning Models
3. Glauner, P., Valtchev, P., & State, R. (2018). Impact of Biases in Big Data, (July).
4. Courtland, R. (2018). Bias detectives : the researchers striving to make algorithms fair, (June).

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

5. Potash, P., Romanov, A., & Rumshisky, A. (2017). Tracking Bias in News Sources Using Social Media : the Russia-Ukraine Maidan Crisis of 2013 – 2014, 13–18.
6. Tan, S. (2018). Interpretable Approaches to Detect Bias in Black-Box Models, 1–2. debayo, J., & Kagal, L. (2015). Models, *37*.
7. Chouldechova, A. (2017). Fair Prediction with Disparate Impact : A Study of Bias in Recidivism Prediction Instruments, *5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

AUTHORS PROFILE

Vijay Kumawat is currently pursuing B-Tech in Computer Science Engineering(3 rd Year) in VIT University, Vellore.
Email id: vijaykumawat256@gmail.com

Vaibhav Bangwal is currently pursuing B-Tech in Computer Science Engineering(3 rd Year) in VIT University, Vellore.
Email id: bangwal.vaibhav04@gmail.com

Dr. K.Lavanya is currently working as an Associate Professor in the School of Computer Science and Engineering(SCOPE) in VIT, Vellore.
Email id: lavanya.k@vit.ac.in