Demystifying Black Box Models with Neural Networks for Accuracy and Interpretability of Supervised Learning

Jagadeesh Prabhakaran
Department of Information Systems and Decision Sciences,
California State University, Fullerton, CA 92831, USA
jagadesh2205@gmail.com

***Abstract*—Intensive data modelling on large datasets that were once limited to supercomputers and workstations can now be performed on desktop computers with scripting languages such as R and Python. Analytics, a field that is popular over this aspect of access to high computational capability enables people to try out different mathematical algorithms to derive the most precise values with just calling pre-written libraries. This precision in the case of black box models such as Neural Networks and Support Vector Machine comes at the cost of interpretability. Specifically, the importance of interpretability is realized while building classification models where understanding how a Neural Network functions in solving a problem is as important as deriving precision in values. The Path Break Down Approach proposed in this paper assists in demystifying the functioning of a Neural Network model in solving a classification and prediction problem based on the San Francisco crime dataset.**

***Index Terms*—Neural Networks, Machine Learning, Black box Models, Supervised Learning**

I. Introduction

Analytics and statistical modelling are age old approaches to building predictive models. Its applications and usability have tremendously increased with the rise of computational capabilities where just anyone can utilize them without deep knowledge in statistics or even the domain of application. Data science which was once programmed only on supercomputers and mainframe computers for very advanced calculations such as aircraft scheduling, space explorations and military research in being applied to every field of business owing to the power of data driven decision making. For any decision to be made in a wide variety of businesses, the accuracy of predictive models is as important as its interpretability since the whole purpose of having an analytics function is to have actionable insights that suggest strong recommendations based on the data.

There are a wide variety of prediction and classification techniques that have been used on datasets. The two-main categorization of data modelling are white box models and black box models. White box models are the ones that have a straight forward mathematical formula to interpret how the model has evaluated the training data to come up with the data model. Examples of such techniques of classification are decision trees and logistic regression. On the other hand, black box models are the ones where there is no straight formula to understand the functioning of how the model was built but the logic on how the accuracy was derived seems substantial to depend on it for further predictions. Some examples of black-box classification modelling techniques are Support Vector Machine and Neural Networks. These are techniques where the conceptual functioning seems clear but the derivation of results is complex to interpret. The complexity in interpretation is because there are multiple iterations of values worked on a trial and error basis to identify the best fit model across the entire training set.

II. Problem definition

The advantage of the white box model is that they are easy to interpret but their accuracy measures are not as good as black box models. The black box models are more often the better performing models but their incapability to be interpreted makes it difficult to identify the key influencers or independent attributes that drive the output variable. This is very essential for them to be widely adapted in a business setting because a management that has for years depended on logical intuitions for decision making when opening to data driven decision making would look for substantial evidence that the model suggests something based on the deeper logics than something haywire. This brings up the question on how a black box model with neural networks be interpreted when achieving the best accuracy measures for classification problems.

III. Classifier

A classifier is a functional element that is trained to observe the features of a dataset and make inferences in a manner that can help predict outcomes of variables in the dataset when a different set of parameters are given as input. Decision making based on classifiers are primarily to classify elements based on a given set of observed categories. For any classifier to be built, it is essential for splitting a given dataset into training and testing data generally on a 60-40 proportion that must be picked randomly for each set to avoid any selection bias. A classifier can be built using a dataset that could have any set of features ranging from numbers, textual patters, voice or image. To experiment on this neural network classifier, the San Francisco Crime dataset from Kaggle is utilized along with its categorical variables and textual descriptions of crimes reported between 2009 and 2015 which consisted of hundred thousand records. The purpose of this classifier on this dataset is to analyze the independent variables in the dataset and come up with a model to accurately predict whether the San Francisco Police Department (SFPD) would take action or take no action on a given case based on the learnings of the past using the dependent variable called Resolution which is converted to categorical binary dummies.

IV. Modelling Approach

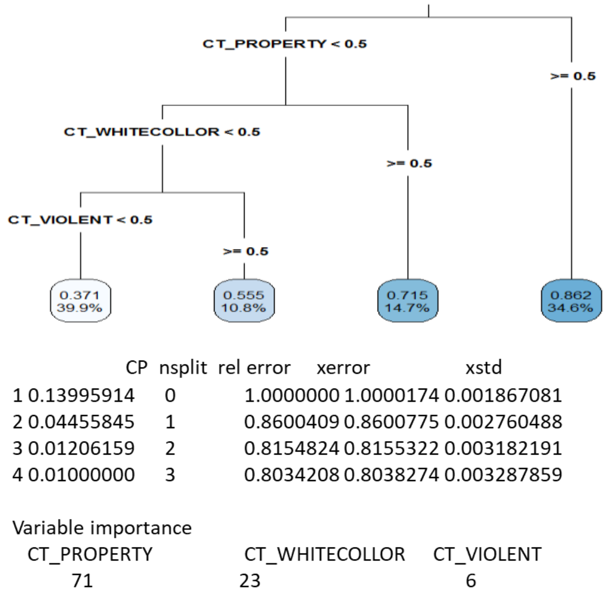
Broadly there are two kinds of data modelling approaches, namely white box modelling and black box modelling. As the names suggest white box modelling refers to those techniques that are easily understandable and can be explained using mathematical equations or operators. Examples of such techniques are Decision Trees and Logistic Regression. On the other hand, black box modelling refers to the computationally intensive techniques that are difficult to comprehend. These models also follow a mathematical structure but go through multiple iterations trying out different combinations of values until it derives the best accuracy. Linear programming is a typical example of how this will function. With respect to classification problems, the modelling techniques that belong to this category black box are Neural Networks and Support Vector Machine.

V. White Box Model Interpretation

In the white box modelling method of decision trees, the full tree structure built can be visualized and interpreted based on the variables assigned to the values for a classification problem. The figure 1 shows the decision tree representation of a value 1 for the binary attribute no-action. The independent variables in the tree (crime type – white collar, crime type – violent, crime type – property) are the top influencers of the output variable of whether action would be taken or not where the flag value 1 in the output suggests no action. In the decision tree, the intensity of the blue color in the leaf node suggests the values being closer to one or zero and the percentage values in the leaf node shows the distribution of values under that bucket.

Using the values of the nsplit and relative error shown in Fig. 1 it can be inferred that the relative error keeps reducing as the nodes are split in order to predict the outcome variable with better precision. The variables that are of importance are listed with their scores for the decision trees model. This model was formulated using the rpart and rpart.plot functions in R.

The interpretation of this model is quiet simple owing to the use of greater than less than symbols and the tree structure explaining the decision-making process clearly. The other white box technique employed to solve problem is Logistic Regression which utilizes the logit function on the linear equation to derive binary output variables based on the independent attributes set for a given row in the dataset.



```
          CP  nsplit  rel error    xerror         xstd
1 0.13995914     0    1.0000000 1.0000174 0.001867081
2 0.04455845     1    0.8600409 0.8600775 0.002760488
3 0.01206159     2    0.8154824 0.8155322 0.003182191
4 0.01000000     3    0.8034208 0.8038274 0.003287859

Variable importance
  CT_PROPERTY         CT_WHITECOLLOR    CT_VIOLENT
     71                 23                  6
```

Fig.1. Decision Trees Model Representation and Evaluation

The mathematical equation for the logistic regression function can be given as:

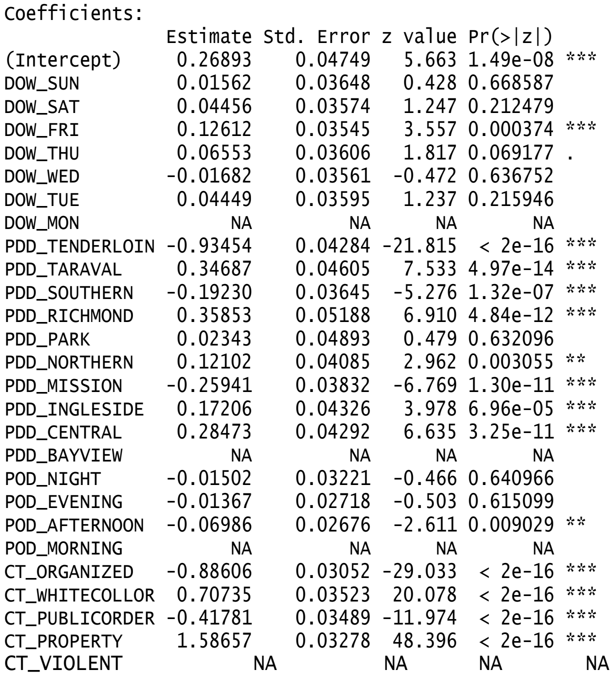
$$Ln\left(\frac{P}{1-P}\right)=\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_kX_k$$

where p - refers to the probability of the event – no action would be taken on the crime represented by values between 0 and 1

X1, X2 and Xk - represents the independent variable values beta0, beta1, beta2 and betak - represents the coefficient values computed to adjust the weights of importance to be given to each variable in the list.

The coefficient values of the logistic regression model assist in interpreting the input variables. The signs in front of the coefficients of the variables represent whether the variables are directly or inversely influencing the output decision.

The Fig. 2 represents the coefficients, standard errors and z scores of each variable included in the model to predict the no-action attribute. Interpretation of these results suggest that the variables CT-Property, CT-WhiteCollar, CT-Organized, PDD-Tenderloin are the top influencers of the output variable (CT – Crime Type, PDD – Police Department District, POD – Part of Day, the variables have been changed from categorical variables to binary dummies representing categories). The standard errors represent the extent to which these values of independent variables can vary. The NA values in the coefficients list display the cases of singularities where the values directly correlate with other variables in the list. This could have been eliminated by removing the attributes from the list, but to maintain consistency of variables across the different modelling techniques the same set of variables were used without any elimination. While listing the variables for the model building function the tilde followed by (.,) was used to input all available independent variables in the data frame.



```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)    
(Intercept)     0.26893    0.04749   5.663 1.49e-08 ***
DOW_SUN         0.01562    0.03648   0.428 0.668587    
DOW_SAT         0.04456    0.03574   1.247 0.212479    
DOW_FRI         0.12612    0.03545   3.557 0.000374 ***
DOW_THU         0.06553    0.03606   1.817 0.069177 .  
DOW_WED        -0.01682    0.03561  -0.472 0.636752    
DOW_TUE         0.04449    0.03595   1.237 0.215946    
DOW_MON              NA         NA      NA       NA    
PDD_TENDERLOIN -0.93454    0.04284 -21.815  < 2e-16 ***
PDD_TARAVAL     0.34687    0.04605   7.533 4.97e-14 ***
PDD_SOUTHERN   -0.19230    0.03645  -5.276 1.32e-07 ***
PDD_RICHMOND    0.35853    0.05188   6.910 4.84e-12 ***
PDD_PARK        0.02343    0.04893   0.479 0.632096    
PDD_NORTHERN    0.12102    0.04085   2.962 0.003055 ** 
PDD_MISSION    -0.25941    0.03832  -6.769 1.30e-11 ***
PDD_INGLESIDE   0.17206    0.04326   3.978 6.96e-05 ***
PDD_CENTRAL     0.28473    0.04292   6.635 3.25e-11 ***
PDD_BAYVIEW          NA         NA      NA       NA    
POD_NIGHT      -0.01502    0.03221  -0.466 0.640966    
POD_EVENING    -0.01367    0.02718  -0.503 0.615099    
POD_AFTERNOON  -0.06986    0.02676  -2.611 0.009029 ** 
POD_MORNING          NA         NA      NA       NA    
CT_ORGANIZED   -0.88606    0.03052 -29.033  < 2e-16 ***
CT_WHITECOLLOR  0.70735    0.03523  20.078  < 2e-16 ***
CT_PUBLICORDER -0.41781    0.03489 -11.974  < 2e-16 ***
CT_PROPERTY     1.58657    0.03278  48.396  < 2e-16 ***
CT_VIOLENT           NA          NA       NA        NA
```

Fig.2. Logistic Regresssion Model Interpretation

VI. NECESSITY FOR INTERPRETATION

In each classification based prediction problem there are two important aspects: predicting the category or class with precision and identifying the variables, factors or influencers that dictate which class would be the outcome. Both the white box models: decision trees and logistic regression give a clear picture of which variables act as key influencers of the output variable and the coefficients suggest the extent to which each variable contributes towards accurately predicting the output variable. The black box models have long known to be uninterpretable. Due to their complex nature of computation anybody looking to solve simple problems of neural networks by hand might have to work for weeks to get all the combinations right. Interpretation is important for business decisions to be made and suggest actionable recommendations to be made. Managers who have long known to follow their intuitions to make critical decisions have found it difficult to rely on black box models which do not clearly mention the variables that influence the output variable.

VII. NEURAL NETWORKS

Artificial neural networks as it is commonly called is an approach devised in 1958 to model how the brain processed visual data and recognize objects. The core principle of this mechanism remains the same even today. This method is inspired by the neural network in the brain and human body where the neurons (nerve cells) are responsible for transmitting the input details to the next layer to make sure the information is carried forward in the right direction.

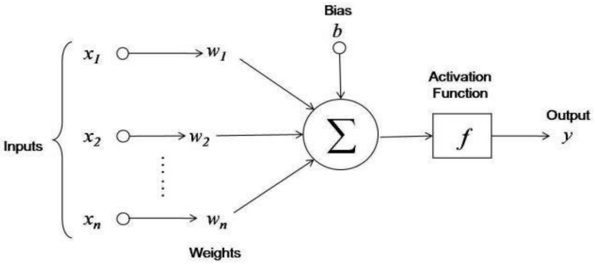


Fig. 3. Neural Networks Function

In a neural network based computation model the inputs are received from the n nodes where n represents the number of input variables. Specialized weights are assigned to each variable based on the ideology to reduce the sum of squared errors in predicting the correct outcome. As inferred from Figure 3 the inputs (x1,x2) are multiplied with the weights generated (w1,w2) to serve as inputs for the next layer of computation. These are often known as the hidden layers. These layers in turn incorporate an approach such as feed forward or back propagation as suggested in the model parameter to try a combination of mathematical functions to derive at the best possible accuracy. The sigma denoted by b is the bias which is the error term that is incorporated to adjust for all the biases that exist in computation. The final activation function formats the output to a desired range, in our case to binary or probabilistic values between 0 and 1 to make a prediction of the crime case solving motivation which will be assigned to y.

VIII. PATH BREAKDOWN APPROACH

The step by step approach for path break down was constituted to replicate the computational method of neural network by hand and this has guided towards providing a better understandability of how the weights are generated for a neural network. The weights here can be compared to the coefficients in logistic regression but are not mere correlation coefficients. These weights are modified as each input row is fed into the neural model for it to learn on deriving at the best possible value. By the path break down approach we can assume the inputs to be a matrix of dimension (100000, 26) where it represents all the rows and columns of the Independent variable set. The number of weights that will exist in the model will depend on the number of nodes setup for the hidden layers. In this case this has been set to 2 nodes since 1 node did not converge and 3 had lower accuracy measures. The weights can be represented by a matrix od dimension (26,2) where 26 represents the weights marked near the input nodes and 2 represents the two nodes in the hidden layer which mean 2 sets of 26 weights will have to computed. These weights are values that were computed once all the Input combinations were read by the model and all the extreme values were considered to assign the constant as a weight value that when multiplied with the Input would send the right signal to the next node in the hidden layer. What happens when the inputs and weights are combined here is matrix multiplication of (100000,26) and (26,2) matrices that result in the values for the hidden layer. This can be given by this equation.

$$Z^{(2)} = X * W^{(1)} \qquad (1)$$

where X= Input values, W = weights computed (totally 52 for this dataset), Z = activity for the hidden layer

In Fig. 4 the blue lines representing the constant values are the biases that are adjusted when the inputs converge in each node. Once the entire dataset has been converged onto the 2 nodes in the first hidden layer, the sigmoid function is used as the activation function and applied to every individual element in the matrix $Z^{(2)}$. This activation function after applied to each element will give an output In the same dimension as the input dimension, which In this case is a (100000,2) matrix obtained by combining (100000,26) and (26,2). This can be given by this equation.

$$a^{(2)} = f(Z^{(2)}) \qquad (2)$$

where a = Activity Value, f(Z) = Sigmoid function on the Z matrix elements.

This neural network model build uses the feed-forward technique where the values are propagated only in the forward direction and there Is no back propagation taking place. Back propagation method can also be used be the previous layers to accept reverse values and send recomputed results forward, but for this model since only a single hidden layer is utilized feed-forward method is appropriate. The heavy task of computation is complete for this model while applying the activation function. The next step on this path breakdown approach will be to multiply the activation values with the respective weights as a (100000,2) matrix by the two weights values represented on the two lines converging from the hidden layer given as a (2,1) matrix. The resultant of multiplying these two matrices (100000,2) and (2,1) will result in the activity of the third layer as a (100000,1) matrix. This represented by $Z^{(3)}$ that can be given by this equation

$$Z^{(3)} = a^{(2)} * W^{(2)} \qquad (3)$$

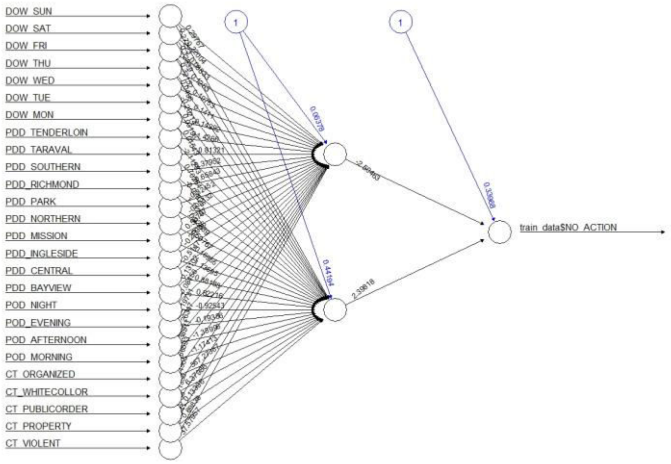


Fig.4. Neural Network Visualization for this Supervised Learning Problem

The final step in this approach will be the application of the activation function to $Z^{(3)}$ to derive the Y^ (y-hat) value that represents the output variable. The final output can be given by the equation:

$$Y\hat{} = f(Z^{(3)}) \qquad (4)$$

The final application of the activation function on the $Z^{(3)}$ matrix will give the desired binary outputs of 0 or 1 which represent action would be taken or not respectively for a case filed at the San Francisco Police Department. Briefly the purpose of this Neural Network model would be to predict this aspect of the scenario with precision.

IX. MODEL PERFORMANCE COMPARISON

The accuracy measures of the different classification models were noted in a confusion matrix where the actual values and predicted values of the validation dataset were compared. Since there are 3 classes in this binary classification of whether action will be taken on the crime or no action will be taken, a 2 by 2 matrix was formed for all the three models. Based on the count values in the confusion matrix, the misclassification error rates were calculated.

Misclassification Error Rate = (Count of Incorrect Predictions / Count of rows in the Validation set) * 100

TABLE 1

MISCLASSIFICATION ERROR RATES FOR THE MODELS

Models	Misclassification Error Rate
Decision Trees	28.10%
Logistic Regression	29.02%
Neural Networks	27.27%

Through Table 1 it can be inferred that the misclassification error rate has been the lowest for Neural Networks which is a black box model and has been the highest for Logistic Regression which is a white box model. Even through the margin of difference is very minute for the dataset of 100,000 records that we have taken, when inferences made from this sample dataset is applied to a population dataset of over a million records in the San Francisco crime dataset between 2009 and 2015 even the 2 % better accuracy would help predict 20 more cases with accuracy using a Neural Networks model than the logistic regression model.

X. CONCLUSION

The path breaks down approach proposed in the paper will demystify the core functioning of the neural nets libraries utilized to solve several classification problems. The success of the neural networks model lies in the understanding of which technique to utilize (feed forward, radial basis or back propagation) and the number of hidden layers for nodes. Finalizing on the best approach will be initially through trial and error and later through experience where the best accuracy achieved by models involving complex techniques and multiple hidden layers could also take a lot of computation time and hence requires a trade-off to be taken when deployed in production. The interpretability element of white box methods for business recommendations along with the functional understandability and precision of black box models such as neural networks will help eliminate complexities in supervised learning.

REFERENCES

[1] G.P. Zhang, "Neural networks for classification: a survey" IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) Vol. 30 Issue: 4, Nov 2000

[2] Stephen Dreiseitl, Lucila, "Logistic regression and artificial neural network classification models: a methodology review) Elsevier Journal of Biomedical Informatics Volume 35, Issues 5–6, October 2002, Pages 352-359

[3] San Francisco Crime Dataset, SFOpenData platform, 2016 [Online] Avaiable: https://datasf.org/opendata/

[4] Fabio Fabris, *A review of supervised machine learning applied to ageing research*: Springer-Biogerontology, 2017; 18(2): 171–188.