Interpreting machine learning and neural network models: Understanding the Why behind the decision?

Yogesh Dhingra

yogesh.dhingra@okstate.edu

Oklahoma State University, Oklahoma, Stillwater, United States

Machine Learning algorithms are primarily known for making highly accurate predictions but at the cost of interpretability. It is often difficult to comprehend the reasons behind the output of a model. Such "black-box" nature of machine learning models creates a barrier in the adoption of machine learning models in the business world especially in domain of finance and pharmaceuticals which are highly regulated domains of the present day. For instance predicting whether a drug is harmful or vice versa is not suffice in a real world, knowing what factors make it harmful is also equally important. A highly interpretable model ensures no bias is involved in model output and builds the trust of the end user in the complex machine learning algorithms. This paper aims to demonstrate techniques that can be used to make machine learning and deep learning models more transparent by extracting human understandable insights. A financial dataset has been used as an example to demonstrate the techniques to encourage the adoption of complex machine learning algorithms in the highly regulated sectors.

OBJECTIVE: To illustrate techniques that can explain the reason behind the model output i.e. understand underlying factors that drove the model behavior. The dataset used can be found at Kaggle : Dataset-URL . The data contain complete loan data for all loans issued through the 2007-2018, including the current loan status (Fully Paid, Default.) and latest payment information. The data consist of 955,000 observations and 75 variables.

METHODS: The target here is a binary indicator indicating if the loan defaulted or not. The data was transformed, normalized and partitioned into 80% training and 20% validation datasets. Variables that were highly correlated i.e. VIF >10 (variance inflation factor) were removed and variables with more than 50% of the value as missing were dropped. Random forest and deep neural network models were used to predict the target variable. The models were evaluated based on the area under ROC curve. To understand model behavior and explain the relationship between features following methods were used : Partial Dependence plot ,Shap Values and LIME(Local Interpretable Model-agnostic Explanations)

RESULTS: Random Forest Interpretation using Partial dependence plot and Shap values: An optimized random forest using GridSearchCV was built .The model ROC was 0.96. Let's consider an example from the data to understand the model behavior

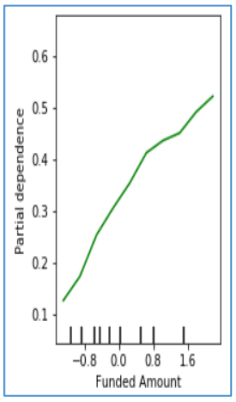
Do people with high amount funded for loan are likely to default more?

- Figure 1 shows Partial dependence plot between variable "Funded Amount" and binary target "Loan Status" set as "Default" i.e. Loan status=1 .It can be inferred from the plot that the "Funded Amount" has a linear relationship with "Loan status" and the probability that a loan will default increases with increase in the "Funded Amount".

How is the impact of total payment received against a loan till date?

- Figure 2 shows a Shap summary plot, each point on the plot represent a datapoint, the color shows whether feature has a high or low value. The Shap value on the x-axis

indicates whether the impact on the prediction value is positive or negative. For variable "Total Payment" it can be inferred that a low total payment received against a loan increases the chances that loan will likely default in the future.



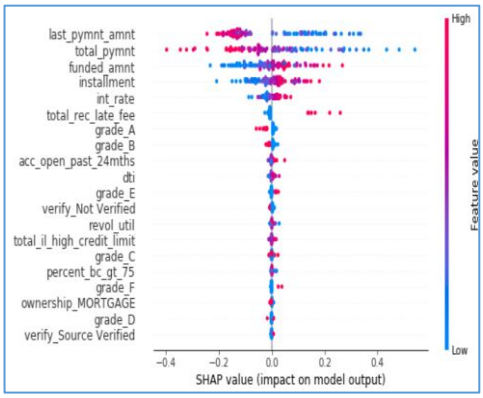


Figure 1:Partial Dependence plot

Figure 2: SHAP Summary plot

Deep learning network interpretation using LIME: A deep learning model was built using TensorFlow framework and LIME interpretability library with 2 hidden layers. The first layer with 4 hidden units and second with 4 units, each with relu nonlinearity. The model ROC was 0.97. LIME library explains how model behave locally for a given data point by identifying the reasons for a given prediction. For example:

What are the factors used by model to identify a loan as "Default" which has these characteristics: Funded Amount: $24,700, Revolving Balance:$28,861,Reason for Loan: Debt Consolidation, Home Ownership: Mortgage, Revolving line utilization rate: 65%

- Intuitively a loan account with such characteristics is likely to be labeled as a "Default" account because of high revolving utilization rate and high revolving balance by the end user. The model on the other hand although predicts the output correctly but gives more weightage to different set of variables(shown in Figure 3)

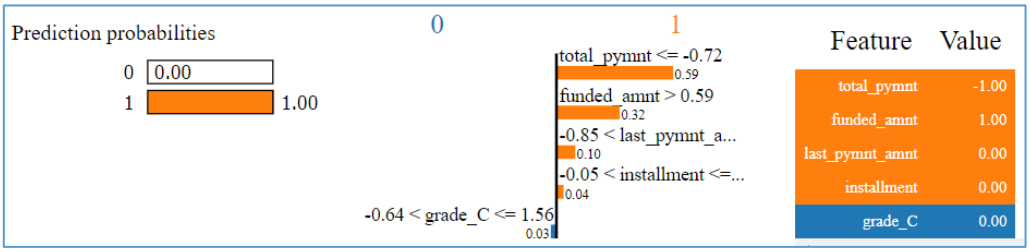


Figure 3: Explanation for the prediction using LIME

CONCLUSION: This paper highlighted some of the techniques that can be used to explain the reasons behind the prediction of machine learning models. Such techniques enables model output to be trusted and encourage their usage in finance and other regulated fields. However, these techniques currently works well when the size of the dataset is small, for future research computational optimization methods such as parallelizing can be incorporated to make these techniques more efficient.