Interpretable Machine Learning Models

Summary

Machine Learning (ML) models make predictions based on large input data. These models can deal with both linear and non-linear data relationships. Most of them demonstrate good accuracy and confidence rates. Although, when the input data is large, interpreting even the best of the models like decision trees becomes a challenge due to the depth of the tree split. When it comes to interpreting black-box models like neural nets and deep leaning models, the difficulty in explaining the model outputs is two fold. Oftentimes, they are not capable of providing a rationale for their predictions. Neural networks tend to have a very high accuracy amongst all the ML models, however operationalising them is a challenge if the rationale behind the decision is not fully understood by the business or in some cases, even the modellers themselves.

To tackle this issue and to shine light on the black-box algorithms, there is growing interest in a new field of study called “Interpretable Machine Learning”. It aims to clarify the classification rules behind the final prediction or output. These models are built after the black box model has been developed, and they aim at decoding the decision making process of the black-box models.

This study explores and compares the most advanced interpretable Machine Learning techniques available today.We can used LIME and SHAP to derive a logical explanation of the model outputs, and also use the results to evaluate if the decisions are valid. This will help the business to better understand why or why not a particular prediction was made by the model, and will also help the modellers in confidently operationalizing neural nets without having to trade accuracy for interpretability.

We are trying to answer the most important question in the world of Machine Learning and Artificial Intelligence.

“ Can you explain how your model works?”

Objective

Numerous machine learning models use statistical and mathematical techniques to classify data and predict outcomes. Deep learning/neural networks, go further into mimicking how human brain functions. Neural networks can get so complicated that they become entirely unexplainable. As a modeller we only know the inputs to the neural networks and can only read the outputs. But what happens inside the algorithm is unknown, and that’ why they are known as black-box algorithms. Due to the mystery of the decision making process, the usage of such models has been limited in highly regulated environments like CRA and other government departments which require interpretability for decision making.

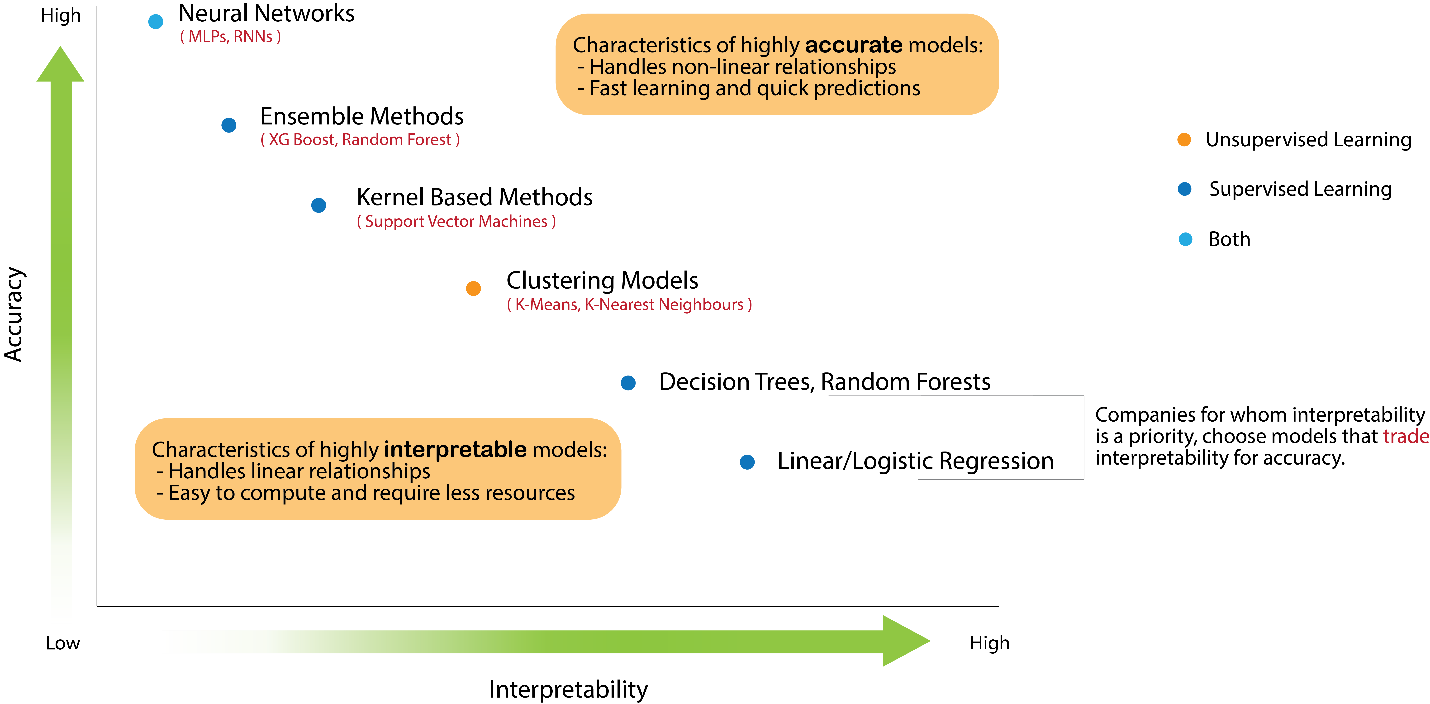
There are three main components of neural networks: an input layer, the hidden layer or the model, and the output layer. Simply put, some input data is fed to the model and a series of hidden layers (manually chosen) group the input data based on patterns discovered by the model. The final layer is the output or the prediction that the model makes.

In this study we will be demonstrating various interpretable systems that explain exactly how and why the black-box model made the decisions or predictions. This will eventually help the officers in reviewing their decision making process confidently. This will also help in eliminating biases and result in better results. Interpretable ML is important for the agency because our core values reflect accountability and trust.

Accuracy vs Interpretability

Model Accuracy and Interpretability have trade-offs. As you can see in the graph below which shows different ML algorithms starting from Linear/Logistic Regression to Neural Networks. On the far left are algorithms like Neural Networks, Ensemble models that are highly accurate but not very interpretable. And on the far right, we have models like Linear Regression and Logistic Regression that are highly interpretable but not as accurate.

This study aims at investigating ML techniques that can optimize for both simultaneous, accuracy as well as interpretability.



Exploring model interpretability with demo project

For the purpose of this study, we will be using open data from a bank to determine the loan eligibility. We will follow the CRISP-DM framework for model building to identify and understanding the business problem, removing outliers, and engineer new features engineering. With some limited domain knowledge we will be creating a model that makes recommendations based on the input data. We will try different models and confirm whether Neural Networks provide better accuracy over other models.

After the models are built, we will them use LIME and SHAP to derive logical explanations of the model outputs for each record in the data.

Why is this approach better?

Without this technique, the second best-option is to rely on a model that is generally explainable, which is in this example, logistic regression. Deep Learning models are able to extrapolate more information from the same dataset by incorporating complex equations into their algorithms. Unfortunately, by introducing this complexity in the model, the model becomes less interpretable. However, this novel approach allows the added benefit of keeping complexity without compromising the interpretability of the model.

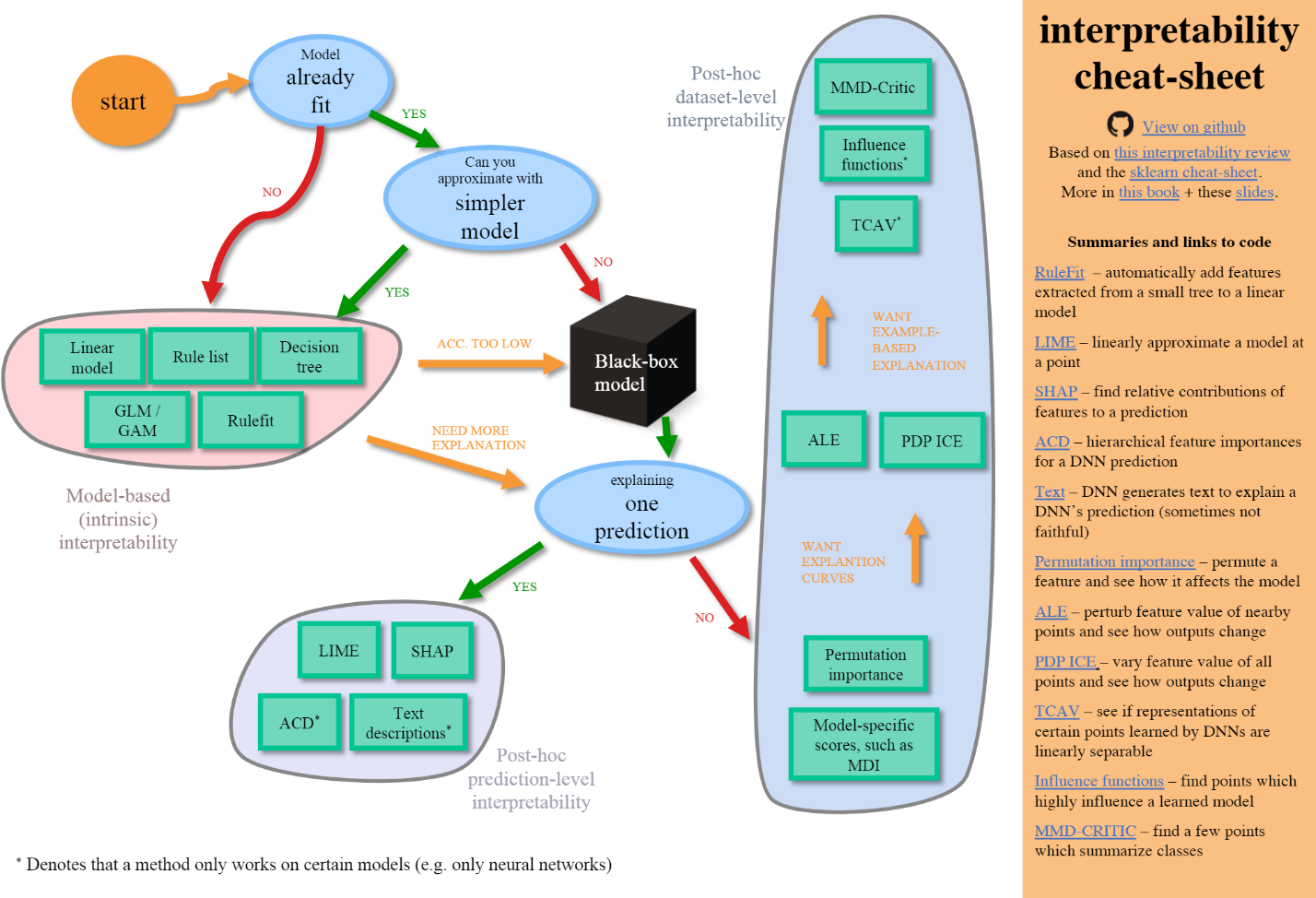
Available Solutions:

One way to explain deep learning models is to examine the feature importance: for a given prediction, how important is each input feature. For simpler models, interpretability is often provided by the model itself. This is especially true for tree-based models, where the algorithms can assess the importance of each feature, given a set of input features.

Model Agnostic Interpretability

Interpretable AI or ML is a new and rapidly evolving field. As the demand and need for explainable AI grows, any new solutions are being developed. Currently there some libraries that are available:

1. LIME – **L**ocal **I**nterpretable **M**odel-Agnostic **E**xplanations
2. SHAP – **SH**apely **A**dditive Ex**P**lanations
3. ElI5 – **E**xplain **L**ike **I** am **5**



Regardless of the type, our aim is to design and build models with more transparency, accuracy and reliability. As a result, better decision making will result in improved services, and people satisfaction. Our goal is to get the buy-in from the business and build their trust in machine learning models.

There are two kinds of Interpretation:

**Local**: Which explains how and why a specific prediction was made and what features contributed to that prediction.

* Why did the model predict that taxpayer X will not pay voluntarily.

**Global**: Which explains how the model works overall and overall what features were termed important by the model in making predictions. For e.g. **Feature importance** is a global interpretation of the model.

CRISP – DM

Business

Understanding

Data

Understanding

Data

Preparation

Modeling

Evaluation

Deployment

Data

Interpretability

Some concepts are extremely difficult to explain even when communicating in the same language. For example, kindergarteners are unlikely to grasp astrophysics even if they are taught in their native language. Because some topics are inherently difficult or impossible to explain.

The cons of using them are important to keep in mind as we evaluate the final performance of our models.

We will now look at the 3 most important model interpretability options available:

1. ELI5 – **E**xplain **L**ike **I** am **5**
2. LIME – **L**ocal **I**nterpretable **M**odel-Agnostic **E**xplanations
3. SHAP – **SH**apely **A**dditive Ex**P**lanations

ELI5

Can be used to explain and create visualizations for interpretable models like Decision trees, random forests, Logistic and Linear regression, etc. It can be used for both local and global interpretations for white-box models. However, when it comes to black-box models like Neural nets, it can only provide global interpretations. It can only generate feature importance charts when it comes to interpreting neural nets. Since most of the white-box models already provide the feature importance calculations built within SPSS, we will not be doing demo of this package.

LIME and SHAP

Data

Black-box model

Predictions

Explainer model

Explanations

Charts/Dashboards

 LIME/SHAP

LIME

The basic idea behind LIME is to create a surrogate model to explain a given record. For e.g. if we want to understand why a person was denied loan, we should understand the most important features behind this prediction. LIME is such a technique that can explain why a single data point was classified as a specific class. It treats every model as a black-box and uncovers the mystery. LIME can explain tabular data, image data and text data.

How it works:

The algorithm first chooses a record for observation. It then creates a new dataset around the record by sampling form the distributions learnt from the training dataset. Next, it calculstes distances between the new points and the record, which is termed as the measure of similarity.

Now it uses the specific model we used to predict the class of the new points and finds the features (n features) that have the strongest relationship with the target class. Then it fits a linear model on the fake data in n dimensions weighted by similarity. These weights of the new linear model are used to explain the predictions.

SHAP

SHAP approximates the effect of removing a feature under various combinations of presence or absence of the other features. The effects can be negative or positive since a feature can affect the prediction in either way. SHAP gives both; global and local interpretability. This helps to study both the global importance of features as well as the impact of a particular feature impact for a given record.

For example, when we take a look at a single loan application and study with SHAP, we can see that the person’s profile being unskilled and non-resident, not having a lot EXPLAINABLE MACHINE LEARNING FOR CREDIT LENDING 27 of money in the bank account all put the profile in a risky zone thus increasing the likelihood of default.

How it works:

SHAP works on the Shapley values method from game theory. In SHAP, the weight of each feature is computed first. For e.g. Suppose there are n number of features. To get the importance of a feature1, SHAP first calculates all the subsets of features n that do not contain feature1. Then it computes the effect on the predictions after adding feature 1 to those subsets.

Approach

We will first build a model

then use SHAP to explain the model both globally as well as locally.

We will also train a neural network and we will use an advanced interpretability technique like SHAP and LIME to explain the predictions. Thus combining the best of the accuracy and interpretability techniques, which would definitely yield better results.

Algorithms like logistic regression and decision trees are already an explainable machine learning algorithms. However, with increasing input features interpreting a tree is a difficult task. Therefore, we will deploy the explainable machine learning algorithms on top of the logistic regression as well to test local interpretability.

This study mainly focuses on the model interpretation part and not the data quality. Therefore, we assume the data quality is good and proceed with our analysis.

LIME machine learning algorithm generates data around the desired points then uses those dumpy data to formulate a linear regression line in order to create an explanation. Thus, we assume that the dataset could be explained by a linear regression line even though there might be some features that would be eliminated in the process.

Methodology:

Problem Solution

In order to solve this problem we will follow the 5 step process:

1. Clean the Kaggle dataset

2. Build features that are important

3. Train models from Logistic Regression to GBDT

4. Measure the accuracy of the models

5. Use SHAP to explain each of these models

To train the models we will use the Scikit-Learn open-source ML package to train Logistic Regression and Gradient Boosted Tree models. For the Boosted tree models, we will use LightGBM library. And we will use SHAP to help us with feature selection so we create the best possible set of features to train. After that, we will train a few different ML models and measure the AUC(Area Under the Curve), Precision and Recall scores. All the experiments will be conducted in Python and Jupyter Notebook will be used for visualization.

Similarly, SHAP provides the ability to look into individual prediction explanations like below. For example, when we take a look at a single loan application and study with SHAP, we can see that the person’s profile being unskilled and non-resident, not having a lot of money in the bank account all put the profile in a risky zone thus increasing the likelihood of default.

LIME provides . . .

Conclusions:

1. Model Agnostics methods for interpretability help in decoding the machine learning models and spot biases in the data.
2. Can be used to explain to non-technical audience why or why-not a prediction was made.
3. Dashboard feature helps creating a what-if scenario for analysing the impact of different values on the predictions.
4. Build confidence in the predictions and helps make informed decisions.