Interpretable Machine Learning – Increase Trust and Eliminate Bias

While today's models have unprecedented levels of accuracy, this has come at the cost of greater complexity. Improved model accuracy often comes with a trade-off – oftentimes we don't understand the inner workings of modern models and struggle to explain the calculation process.

In a perfect world, our models would provide both accurate decisioning and an understanding of the underlying reasoning for those decisions. This would offer not only insights into the problem at hand but also a better understanding of why a model might fail. Interpretability and transparency also help institutions meet regulatory standards and allow model decisions to be challenged or changed when they produce an adverse outcome.

To achieve the twin goals of accuracy and interpretability, organizations are relying on a class of models known as interpretable machine learning (IML) models. IML models aim to match the performance of "black-box" models but with the desired properties of interpretability and transparency.

Why is model interpretability important?

The ability to understand and articulate the inner workings of a complex model can be particularly beneficial in the following contexts:

- **Stakeholder Buy-In** – In high-risk domains, such as credit decisioning, health, and safety, stakeholder buy-in is critical to the adoption of sophisticated modeling solutions. Before complex model processes can be implemented, all stakeholders must fully understand what the model does and how the model arrives at decisions. Stakeholders are more confident in models when explanations are easily understandable and intuitive.

- **Trust/Reliability** – When using advanced models, we want to be sure that the logic is sound, and that spurious patterns and noise were not used to train the model. We also want to ensure that models are stable and not susceptible to minor perturbations. If a slight variation in the data produces an unexplainable change in the model prediction, stakeholders will lose confidence in the model's reliability.Higher customer acquisition costs to replace lost balances

- **Compliance/Control** – Model interpretability is critical to ensure compliance with internal and external auditors and government regulations. Regulators want assurance that model processes and outcomes are reasonably understood, and they closely monitor the impact of model decisions on the public.

In modeling, there may be an accuracy-interpretability tradeoff, meaning that as models become more accurate, they lose interpretability, and vice-versa. This stems from the difference between parametric and non-parametric modeling methods. Parametric methods make assumptions about the underlying data distribution, and are generally inflexible. For example, linear regression assumes that a linear function relates the input to the output, and these models are only capable of estimating linear shapes. Conversely, non-parametric methods can estimate many more functional shapes. However, while this offer flexibility, these shapes can be quite complex and are usually less interpretable than a simple linear function.

Consider credit risk decisioning models. The stakes are extremely high in credit risk modeling, and small improvements in accuracy can enhance financial performance by large amounts. However, making these small improvements is challenging. Models must consider thousands of factors, such as income history, past borrowing behavior, debt levels, and future potential earnings.

The importance of using interpretable machine learning (IML) for credit decisioning quickly becomes apparent for several reasons. First, simple linear functions, while being desirable due to their transparency and interpretability, suffer from low accuracy and generate numerous false positives (false alarms). Second, although traditional black-box machine learning models like neural networks can provide high predictive accuracy, they lack transparency. Regulators mandate transparency in credit decisioning models to combat bias and lending discrimination.

An Explainable Boosting Machine (EBM), an IML model developed by Microsoft Research, is often as accurate as a black-box model, while remaining completely interpretable. EBMs belong to a class of models called general additive models (GAMs), which are an extension of linear models. GAMs relax the restriction that the relationship must be a simple weighted sum, and instead assume that the outcome can be modelled by a sum of arbitrary functions of each feature.

EBMs are tree-based, Generalized Additive Models, offering a "best of both worlds" approach with both strong accuracy and an interpretable, additive structure. The final result of an EBM is a GAM of the form:

$$y = a_0 + a_1 \cdot f_1(x_1) + \ldots + a_k \cdot f_k(x_k)$$

where y is the prediction and x_1, …, xk are the input features.

To understand EBMs, it is important to understand tree-based models in general. Tree-based models build classification or regression models by breaking down a data set into smaller and smaller subsets and associating a decision with each subset of data (see Figure 1). However, as a decision tree becomes larger, it quickly loses interpretability. State-of-the-art decision trees use ensemble modeling techniques, such as boosting and bagging, which achieve strong levels of accuracy. However, as we alluded to previously, these complex boosting and bagging methods lead to a decrease in interpretability. While simple decision trees can be converted into a set of understandable "if-then" rules, boosting and bagging create numerous decision trees to arrive at a result, becoming less and less interpretable.

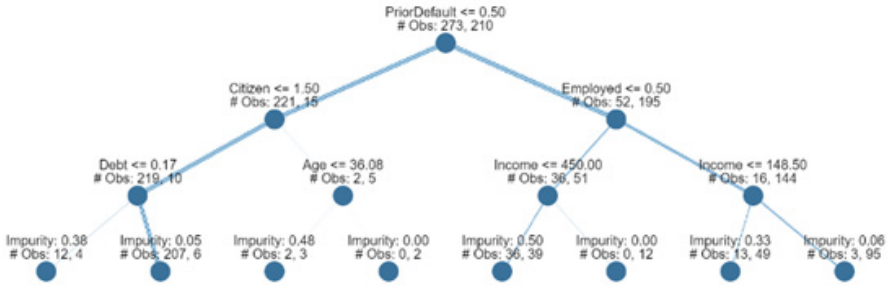

Figure 1: Example of a Decision Tree for Credit Modeling

The EBM training procedure is quite similar to a tree-based method called "gradient boosting". However, EBM training differs in that each tree can only be built with a single feature, which provides EBMs with their attractive interpretable properties. To prevent bias in the algorithm towards particular features, a very small learning rate is used, which ensures that the order that features are introduced to the model does not matter.

Figure 2. Explainable Boosting Machine Tree-Based Training Procedure

Once we have all the trees (Figure 2), we aggregate them feature-wise to produce a contribution graph for each feature. Essentially, the sum of all trees for each feature is the aforementioned f. The contribution graphs can be thought of as dictionaries or lookup tables. For each feature value, they hold the value's contribution towards the final prediction. The contribution graph for each feature shows the shape of the function f that relates the feature values to the output variable.

$$
\begin{aligned}
y \quad = \quad & a_0 \\
& + a_1 \cdot \underbrace{\left(T_1^{(1)}(x_1) + \ldots + T_r^{(1)}(x_1) \right)}_{= f_1(x_1)} \\
& \vdots \\
& + a_k \cdot \underbrace{\left(T_1^{(k)}(x_k) + \ldots + T_r^{(k)}(x_k) \right)}_{= f_k(x_k)}
\end{aligned}
$$

As an example, we developed and estimated a credit decisioning EBM for consumer loans. Figure 3 shows the estimated contribution of debt-to-income (DTI) ratio to the model prediction. A feature risk score (see y-axis) above zero represents a contribution to the model towards "Approval," while a score below zero denotes a contribution towards "Denial." The shaded grey area represents the confidence region.

The plot shows that DTI has a nonlinear, U-shaped relationship with credit approval, as DTI ratios less than 35% and greater than 50% contribute to a predicted approval, while DTI ratios between 35% and 50% contribute to a predicted denial. This implies that higher DTI increases the chance of credit denial, but only up to a certain point. The U-shape of the DTI feature contribution graph surprisingly shows that DTI greater than 50% is associated with credit approval. While this may seem unintuitive, the data may contain individuals with a high net worth, who are strong credits but may carry large amounts of debt relative to income.

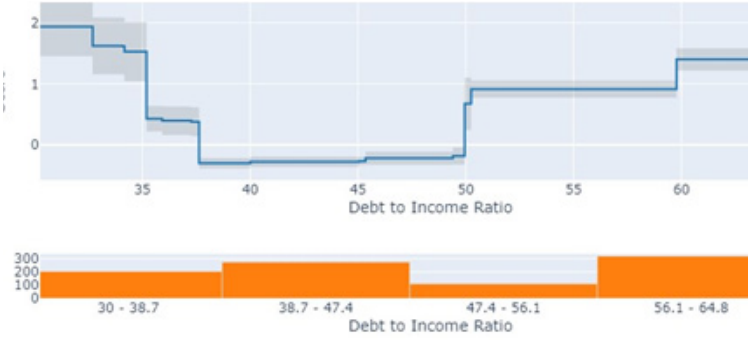

Figure 3. Contribution of Debt-to-Income Ratio on Model Predictions

Figure 4 shows the feature contribution graph for monthly residual income (MRI). The figure shows that, intuitively, as MRI increases, the likelihood of credit approval rises. However, as with DTI, the relationship is nonlinear. While higher MRI contributes to credit approval, the shape of the function is piece-wise linear and lacks smoothness. The contribution graph suggests that MRI below $2,000 contributes to a predicted denial, MRI between $2,000 and $3,200 makes no significant contribution, and MRI above $3,200 contributes to a predicted approval.

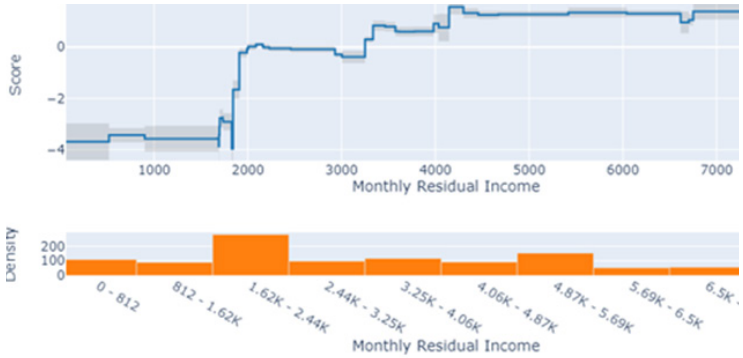

Figure 4. Contribution of Monthly Residual Income on Model Predictions

Explainable Boosting Machines can also show two-way variable interactions. We can visually examine two-dimensional heatmaps of our most predictive interactions. The heatmap for the interaction between DTI and credit score (Figure 5) shows that, in general, having a high DTI lowers the chances of approval unless the applicant also has a high credit score. Similarly, a low DTI can compensate for a low credit score and raise the likelihood of approval.

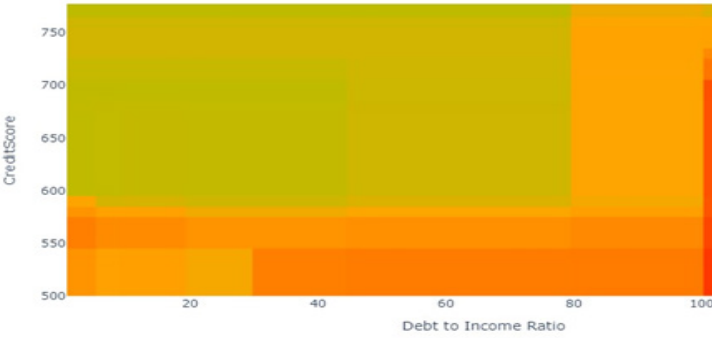

Figure 5. Two-way Interaction of Debt-to-Income Ratio and Credit Score

The heatmap for the interaction between MRI and credit score (Figure 6) shows a similar pattern. As discussed above, in general, an MRI below $3,200 lowers the chances of approval, but the heatmap shows that a high credit score makes approval more likely. Likewise, high monthly income can compensate for a low credit score and raise the chances of approval.

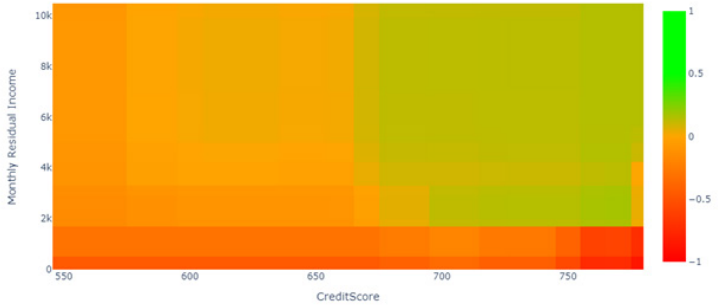

Figure 6. Two-way Interaction of Monthly Residual Income and Credit Score

EBMs can also rank the overall importance of each variable in our model. Figure 7 shows that MRI is the most prominent feature in our credit decisioning model, followed by DTI and credit score. Figure 7 is an example of "global" explainability, attempting to understand how a model works holistically. Global interpretability seeks to understand model behavior across all predictions.

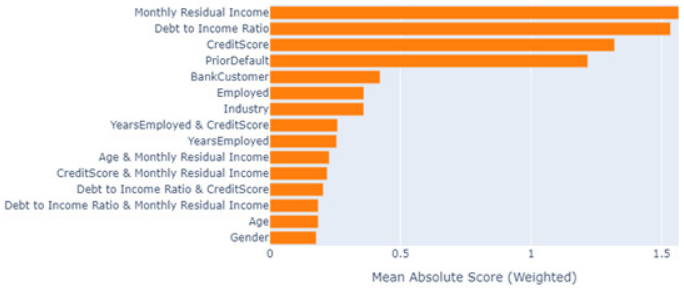

Figure 7. Overall "Global" Feature Importance

While the previous visualization provided "global" explainability for the machine learning model, we can also provide "local" explainability, which are explanations for each individual prediction. Figure 8 shows a single record from our data. We see for this individual record, credit score, DTI, and MRI had the strongest influence over the model's approval of the credit application. Local explainability is a valuable tool for financial institutions to prevent lending discrimination. For example, local explainability can help lenders generate adverse action notices to explain why a particular applicant was denied credit.

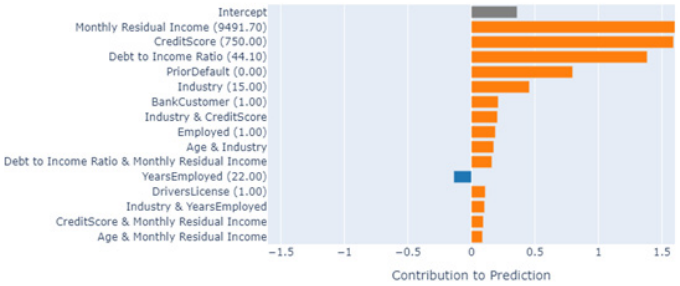

Figure 8. "Local" Explanation of an Individual Prediction

While black box models can deliver exceptional accuracy, organizations need assurance that their processes and outcomes are well-understood. Interpretable Machine Learning models can enable model developers to ensure that their systems work as expected, help institutions meet regulatory standards, and allow those affected by a machine learning decision to challenge or change that outcome.

Author: Robert Chang

Rob is FI Consulting's Modeling & Analytics Domain Leader, managing Data Science teams at clients including the U.S. Department of Health & Human Services, the U.S. Small Business Administration (SBA), Capital One Bank, and the Federal National Mortgage Association (Fannie Mae). Rob leads FI's Graph Practice, implementing graph-based modeling, analytics, and knowledge management solutions for FI's Federal and Commercial clients. Prior to joining FI, Rob was a Financial Engineer and Derivatives Trader for banks and hedge funds in New York, London, Singapore, and Hong Kong. Rob holds a Master's degree in Mathematics of Finance from Columbia University, and a Master's degree in Information Management Systems from the Harvard Extension School.

Author: Fiona Meng

Fiona Meng is a data science consultant specializing in machine learning, neural networks and advanced graph data analytics. She designs and deploys scalable models using Python, SQL, and PySpark. Fiona has also authored machine learning academic publications.

If you are interested in learning more about how FI Consulting can support your organization, please email Rob Chang at rchang@ficonsulting.com or call us at 571.255.6889.

FI·CONSULTING

2107 Wilson Boulevard | Suite 400 | Arlington, VA 22201 | 571.255.6900 | ficonsulting.com | info@ficonsulting.com