“SHAP” Module Transcript

Chapter 1

Intro, Topics Covered, & Learning Outcomes

Hey there. My name is Hayley, and I'm on the One AI team here at One Model. In previous modules, you've heard how One AI makes interpreting models easy with robust reporting and the ability to model key findings and storyboards. This is made possible in part by the topic of this module - SHAP - which is a framework for interpreting the output of machine learning models that clearly visualizes the importance of different features in model predictions.

We will cover a background and overview of Shapley values, an introduction to SHAP in machine learning, strengths and weaknesses of SHAP, and how to interpret SHAP in One AI.

After completing this module, you will have a clear understanding of Shapley values and how they're adapted to provide interpretability in models through SHAP, understand how SHAP values are used to explain individual model predictions that can be aggregated to larger groupings, identify how One AI leverages the strengths and mitigates the weaknesses of SHAP, and gain practical insights into interpreting SHAP visualizations available in the One AI Results Summary and model storyboards in One Model.

Chapter 2

Background & Overview of Shapley Values

Section 2 - Background and Overview of Shapley Values

First, I will discuss Shapley values because SHAP builds upon this concept to provide a simple framework for explaining machine learning model predictions. Shapley values are a concept from cooperative game theory that have been adapted and applied in various fields, including machine learning for model interpretability.

They provide a clear, numerical way to assign a value or importance to each player, or in our case each feature, within a cooperative game, or in our case within a predictive model. For example, if you've ever worked in a group project, you know that the work is not always divided equally with each person contributing different amounts to the project's overall completion.

Similarly, features in predictive models contribute differently to the final predictions. Each feature's Shapley value represents its average contribution to model predictions across all possible feature combinations.

Positive SHAP values indicate features that tend to increase predictions, while negative values indicate features that tend to decrease predictions. For example, in a model that predicts voluntary terminations, a positive Shapley value for a lower than average salary means that lower than average salaries contribute positively to predicting voluntary attrition, making it more likely.

A negative Shapley value for higher than average salary means higher than average salaries contribute negatively to predicting voluntary attrition, making it less likely.

Understanding the distinction between positive and negative Shapley values, as opposed to considering absolute values alone, is important because it provides insight into the direction and impact of each feature on model predictions.

This allows us to understand both sides of the coin: what drives employees to voluntarily terminate and what motivates them to remain employed.

Shapley values also provide insights into importance and help explain why certain features are more influential in the model's decision making process. They represent feature importance in numerical terms, making it easier for individuals to grasp and interpret as people often understand numbers better. These numerical values serve as a concrete anchor for comparing different features, enhancing understanding and model analysis.

Chapter 3

Intro to SHAP in Machine Learning

Section 3 - Intro to SHAP in Machine Learning

SHAP, which stands for SHapley Additive exPlanations, is an extension of the concept of Shapley values. It is a method used in machine learning to explain individual model predictions by highlighting the importance of different features contributing to the prediction of a particular instance.

For example, SHAP can reveal how much a lower than average salary influenced employees to voluntarily terminate or how working at the Orlando office contributed to new hire failure. Let's examine how SHAP does this.

As you have previously learned, machine learning models make predictions based on input features like age, salary, and country, each contributing differently to the final prediction.

SHAP explains why a specific prediction was made for a particular instance. For example, it can reveal why an employee is predicted to voluntarily terminate in the next year.

In One AI, SHAP values can and should be aggregated to provide group insights such as why the engineering org unit has more employees with a high risk of voluntary attrition than sales.

SHAP works by systematically excluding different features from the model and observing the impact on predictions. This process measures each feature's importance by quantifying the change in predictions when that feature is included or excluded.

SHAP considers all possible combinations of features and their contributions to predictions. For each feature, it calculates how much adding that features information changes the prediction compared to the predictions without that feature. After evaluating the impact of each feature across multiple combinations, SHAP assigns a Shapley value to each feature, providing a fair way to distribute importance among contributing features. This value represents the average marginal contribution of the feature across all possible combinations of features.

SHAP values can be visualized in different ways, such as bar charts and beeswarm plots, to help interpret and understand which features drive predictions up or down. Positive SHAP values indicate features pushing predictions higher, while negative values indicate features pulling predictions lower. In this image, the green bars represent negative SHAP, which pulls the prediction of voluntary termination lower. Features with large green bars, such as "Annual Salary Higher than Average" or "Team Average Tenure Higher than Average" are drivers of retention.

Chapter 4

Strengths & Weaknesses of SHAP

Section 4 - Strengths and Weaknesses of SHAP

Like any machine learning interpretation method, SHAP has its strengths and weaknesses. Understanding these helps leverage its advantages and mitigate its drawbacks.

Starting with the strengths, the primary reason we use SHAP to visualize models is to explain complex machine learning data in a transparent and interpretable way, supporting ethical AI. SHAP enhances model transparency and clarity, allowing for scrutiny of potential biases and unfairness, enabling practitioners to intervene where necessary. It also helps evaluate whether models treat individuals from different demographic groups fairly and highlights areas needing intervention to address fairness.

Finally, SHAP empowers decision makers to make ethical decisions based on model interpretations because they can actually understand what is driving their predictions.

SHAP offers a clear intuitive way to understand how individual features contribute to model predictions.

Visualizing these values on a storyboard makes it easier to grasp and communicate these insights. By visualizing SHAP values, you can easily identify which features have the most significant impact on model predictions. This aids feature selection, debugging models, and improving model performance.

Onto the weaknesses:

Visualizations using SHAP must be designed carefully to ensure clarity and avoid misinterpretation, especially for non-technical users. We offer a variety of storyboard templates using SHAP that clearly convey what the model is showing and can be customized to fit organizational needs.

Creating informative and interactive storyboards with SHAP visualizations can be time consuming and resource intensive. It requires knowledge in both machine learning interpretation and data visualization. Our templates take care of this as well.

Finally, visualizing SHAP values for models with many features can be challenging. Simplification techniques and dimensionality reduction in One AI helps address this concern.

Chapter 5

Interpreting SHAP in One AI

Section 5 - Interpreting SHAP in One AI

SHAP is used in various contexts throughout the One Model platform. SHAP values, however, are not generated with each model run by default because it is resource

intensive and increases the time required for models to run. However, this trade off is often worthwhile due to the powerful insights SHAP provides. To enable SHAP, you must check the 'Override' box and the 'Generate SHAP Values' checkbox in the Global Settings before running the model.

If SHAP is generated, one of the places SHAP values will appear is in the Results Summary report in the Feature Importance section. It appears first in the SHAP Beeswarm Chart, which is a feature impact visualization. Since SHAP provides a numerical importance for each feature for every individual prediction, this visualization plots each value as a dot. The horizontal axis indicates how predictive that feature is for that instance, either in a positive or negative direction. Additionally, this chart uses color coding to show how each feature value for an instance compares to the average for the entire model population.

For example, the "is_future_manager" feature in this chart clearly shows that being a future manager is a strong indicator of voluntarily terminating, which is what the model that this chart accompanies predicts. For scaled features, the coloring is a gradient. In this example, a more recent date of birth is an indicator of voluntary termination. So for date of birth, March 1, 2000 might be red; March 1, 1985 purple; and March 1, 1970 blue.

Next, SHAP appears in the SHAP average bar chart, which shows the average absolute value of the SHAP values for each feature. It is a great indicator of how important each feature was to the set of predictions, but does not show whether the feature made a positive or negative classification more likely, as it uses absolute values. In this chart, we can see that the cost center had the highest impact on this set of predictions, while tenure had a more modest impact.

Additionally, SHAP values can appear on storyboards for deployed models, provided a data engineer has added the necessary data tables to your One Model. In this example, the top bar shows that having an annual salary higher than average almost always drives retention as indicated by the longer green bar compared to the shorter red bar. Conversely, the second bar shows that having an annual salary lower than average almost always increases the prediction of attrition. As you can see, SHAP is an effective way to simply explain machine learning model prediction data.

Chapter 6

Conclusion & Thanks

This module covered SHAP and its role in interpreting machine learning models within One AI. By understanding Shapley values and SHAP, you can now interpret model

predictions, enhancing transparency and trust. You are equipped to identify the impact of features on model outcomes, aiding in debugging, feature selection, and improving model performance. Happy modeling!