“Classification Model Evaluation” Module Transcript

Chapter 1

Intro, Topics Covered, & Learning Outcomes

Hey. My name is Hayley, and I'm on the One AI team here at One Model. In the "Classifications" module, you learned all about classifications and what problems they help solve. Now, we will move into evaluating the performance of these models to determine their effectiveness, helping you answer the million dollar question - Is this model any good or does it need further refinement before trusting its predictions?

We will cover a classification model evaluation overview, an introduction to some common evaluation metrics, feature importance, and the importance of interpretation plus iteration.

Once you've completed this module, you will grasp the importance of evaluating classifications to determine their accuracy and effectiveness in making predictions. Understand how each evaluation metric provides unique insights, so using them in conjunction is key. See the importance of considering the context in which the model will be applied during evaluation. And recognize that model building as an iterative process involving refinement based on evaluation, domain knowledge, and stakeholder feedback.

Chapter 2

Classification Model Evaluation Overview

Section 2 - Classification Model Evaluation Overview

Once a classification model run is completed, it's important to evaluate its performance to determine if the model is producing accurate predictions. This involves comparing the model's prediction with actual outcomes and using various performance metrics based on the specifics of the model and dataset. Evaluating performance allows you to compare different versions of the model and select the best one for your needs. Well performing models enable stakeholders to make informed decisions based on its predictions.

If the model performed poorly, you should refine it before deploying it for practical applications. Providing strong models and transparent performance metrics builds trust

and confidence among both model builders and stakeholders. When evaluating a model's performance, we should consider the broader context of its application to ensure our evaluation aligns with the practical problem it aims to address.

The primary question to ask is what are we using this model for? Are we exploring and gaining insights into organizational dynamics and employee behaviors or relying on it for critical decisions, such as promotions or medical diagnoses? Or is it somewhere in between critical decisions and data exploration? For models impacting critical decisions, achieving nearly perfect performance may be very important. However, for exploratory purposes, which is what we encourage with One AI models, absolute perfection is less important than digging into the predictions and data exploration.

Furthermore, considering the characteristics of the dataset is important. For example, if the dataset contains imbalanced classes or if the model was trained on skewed data, we will want to keep that in mind during model evaluation. Moreover, we must evaluate if the model was trained on a representative dataset to ensure it will generalize well to new unseen data. A highly performant model trained on data vastly different from its future application dataset is actually not a very good model.

Chapter 3

Introduction to Classification Evaluation Metrics

Section 3 - Introduction to Common Classification Evaluation Metrics

We can accurately measure how well the model performed by using cross validation. As you learned in the cross validation module, before running a model, we split the dataset into two subsets, training folds and a validation fold. Once the model is trained, we evaluate its performance on the validation fold to see how well it will perform on new unseen data.

There are many excellent tools available to evaluate the performance of a classification model. Each evaluation metric offers unique insights into various aspects of the model's behavior, strengths, and weaknesses. Using them together provides a more comprehensive understanding of the model's performance and is the best strategy when possible. Let's go over some of the most common and relevant evaluation metrics.

We'll begin with accuracy. Accuracy is one of the most straightforward and commonly used evaluation metrics for classifications. It measures the proportion of correctly predicted instances out of the total number of predictions in the dataset. In other words,

it indicates how often the model correctly predicts the actual outcome or even more simply, how often the model was right. Accuracy's main strengths are that it's intuitive to understand and very easy to calculate. Its primary weakness is that it can be very misleading with imbalanced classes where the majority class dominates the accuracy metric, potentially masking poor performance in minority classes. For example, if you are predicting voluntary terminations and only 20 employees out of 500 voluntarily terminated, a model predicting that everyone stayed would still be 96% accurate. While this accuracy score seems impressive, the model fails to provide any insightful information and gets every voluntary termination prediction wrong. Most of the models we work with in One AI generally have imbalanced datasets, so we offer other more suitable metrics to measure performance.

Precision quantifies the accuracy of positive predictions made by the model indicating the proportion of correctly predicted positive instances out of all instances predicted as positive. It is calculated by dividing the number of true positives by the total number of positive predictions, which includes both true positives and false positives. Precision's main strength is its usefulness when minimizing false positives is critical, such as in medical diagnostics. However, precision does not consider false negatives, potentially overlooking missed positive instances. It may not be suitable as a standalone metric in cases where both false positives and false negatives are important. That's where recall comes into play.

Recall, often measured alongside precision and also known as sensitivity, measures how often the model correctly identifies true positives from all the actual positive samples in the dataset. It is calculated by dividing the number of true positives by the total number of actual positives, which includes both true positives and false negatives. Recall is particularly useful in scenarios where detecting all positive instances is important, such as anomaly detection. However, high recall may come at the expense of precision. These two metrics are naturally in tension, so improving precision typically reduces recall and vice-versa. Due to this trade-off, it can be challenging to determine if one model is better than another based solely on precision or recall. This is where the F1 score can be useful.

The F1 score is the harmonic mean of precision and recall providing a balanced measure of the model's performance that considers both false positives and false negatives. It helps quantify the value of the trade-off between precision and recall. For instance, determining if giving up 5 points of precision was worth gaining 10 points of recall. The F1 score's main strengths are that it's suitable for imbalanced datasets and provides a single metric that balances precision and recall. However, F1 score does not provide insight into the relative importance of precision and recall.

A class balance chart offers a visual representation of the distribution of predicted labels offering insights into the model's ability to correctly classify instances across different categories. It also helps users identify potential imbalances in the dataset. For example, this chart indicates a dataset imbalance with 'No Termination' representing the majority class and 'Termination' representing the minority class. In a balanced dataset, these bars would be more similar in length.

A confusion matrix is a table that summarizes the model's predictions compared to the actual labels showing counts of true positives, true negatives, false positives, and false negatives. It's a powerful visual because it provides a detailed breakdown of the model's performance, allowing for the identification of air types and areas for improvement. However, this visual can be tricky to interpret for models with more than two classes.

The Receiver Operating Characteristics (ROC) curve is a graphical representation that shows the performance of a binary classification model at various classification thresholds. It plots the true positive rate against the false positive rate at different threshold settings. In simpler terms, it helps us understand how well a model can distinguish between positive and negative instances. A curve that hugs the top left corner of the plot indicates a better performing model, while a curve closer to the diagonal line suggests weaker performance.

The ROC curve is commonly used alongside the Area Under the Curve (AUC) metric. The AUC represents the area under the ROC curve and provides a single value for the overall performance of the model. Higher AUC values closer to 1 indicate better model performance and discrimination ability, meaning the model can better separate positive and negative instances regardless of the threshold chosen. An AUC of 0.5, on the other hand, represents a model that performs no better than random guessing.

Chapter 4

Feature Importance

Section 4 - Feature Importance

When evaluating a model, we can't only consider evaluation metrics, but should also look at feature importance. Feature importance assesses the contribution of individual features towards making accurate predictions. It helps us understand which features have the most significant impact on the model's performance and how they influence the classification outcomes. In simpler terms, it ranks and scores the features selected

by the model, giving users an idea of how important each feature is for making accurate predictions.

Model performance metrics like F1 score provide an overall measure of how well the model performs in terms of predictive accuracy. However, they do not tell you which features are driving that performance.

Feature importance complements these metrics by identifying the most influential features, providing insights into the underlying relationships in the data. Additionally, understanding feature importance allows us to interpret the model's decision making process. It helps identify which features have the most influence on the predicted outcomes, making the model more interpretable and transparent.

Finally, we want to ensure that the model isn't making predictions based on features that should not be included, such as data that has not been validated, is completely random, or is a cheat column.

Chapter 5

Importance of Interpretation + Iteration

Section 5 - The Importance of Interpretation + Iteration

When interpreting your models with the tools we've covered, it's important to understand the model's strengths and weaknesses based on the results. This helps inform the edits or improvements you might want to make for future runs before deployment. Model building is an iterative process that involves refining the model based on insights from evaluation results, domain knowledge, and feedback from stakeholders.

A model is never truly "done" while it's in use. It's more of a feedback loop. You should reevaluate the model's performance after implementing improvements to ensure the changes have the desired effect and do not introduce unintended consequences. This approach ensures the model continuously evolves to meet your changing needs.

Chapter 6

Conclusion & Thanks

Understanding the performance of classifications is crucial for making informed decisions in organizational settings. Through this module, we've explored various

evaluation metrics, feature importance, and stress the importance of an iterative approach to model building. With these tools, you're well equipped to navigate the complexities of model evaluation and drive success in your organization. Happy modeling!