“Model Deployment” Module Transcript

Chapter 1

Intro, Topics Covered, & Learning Outcomes

Hey. My name is Josh, and I'm on the One AI team here at One Model. In past modules, we discussed how to build machine learning models and improve model performance.

In this module, we're going to talk about what to do once we have a model we're satisfied with. This is the point at which we are able to achieve the full value of the model by sharing the insights it generates with stakeholders. One AI facilitates this sharing by allowing deployment of results into the One Model data model and, ultimately, storyboards. In this module, we will cover important considerations prior to model deployment, what deploying a model entails, the difference between deploy, deploy and persist, and ignore, and how to deploy a model and or results in One AI.

After completing this module, you'll be comfortable evaluating important considerations before deploying and or persisting, including model performance, interpretability, and alignment with business objectives.You will understand what deploying a machine learning model generally means and how that differs from deploying in One AI. You will understand model and results deployment and be able to differentiate deployment options and how they impact accessibility of information. And, you will learn how to deploy machine learning results in One AI, moving them from the One AI tab into the One Model data model and, ultimately, Explore and Storyboards.

Chapter 2

Considerations Prior to Deployment

Before deploying a machine learning model or results, several important considerations should be taken into account to ensure its effectiveness, reliability, and suitability for real-world use.

Here are some key factors to consider. First, we have model performance and fit. Evaluate the model's performance to ensure it meets the desired F1, precision, recall, or other relevant metrics for the specific problem. Additionally, check for biases in the model predictions and ensure fairness across different demographic groups or data segments.

It's important to ensure that the model is interpretable prior to deployment, specifically for the problem domain you're working in. Evaluate interpretability by ensuring appropriate features were selected and that preprocessing and configuration happened as expected. Consider the specific business requirements that the model is intended to address. Ensure that the model aligns with the business objectives and that it provides meaningful insights and predictions that add value to the intended use case. Assess whether the model is ready to be integrated into the production environment. This includes considerations such as scalability, compatibility with existing systems or frameworks, and deployment infrastructure.

In One Model, integration involves working with your customer success team to build the necessary tables and dimensions so you can build machine learning metrics and explore and include them in storyboards. You should also enable SHAP prior to running and deploying your model for better interpretation of features.

Deployed models should be continuously monitored to maintain performance and relevance over time. This involves updating and rerunning when new input data becomes available, such as when employees join or leave the organization, new data sources are added to One Model, or simply as time passes.

Establish procedures for model retraining, updates, and version control to incorporate new data and improve model performance.

It's a best practice to develop a monitoring and maintenance plan prior to deployment. Depending on the application domain, consider regulatory, compliance, and ethical considerations related to deploying the model. Ensure that the model's use adheres to relevant laws, regulations, and ethical guidelines in accordance with your organization.

Chapter 3

Deployment in Machine Learning

In past modules, we've talked a lot about how to create a machine learning model with strong performance, good fit, well selected features, and the like. Once you have a great model that you're ready to share, deployment comes into play. In machine learning in general, deploying a model refers to the process of making the trained model available for use in a production environment where it can receive new input data, make predictions, and provide outputs or actions based on those predictions. You're essentially taking your model from the developmental or experimental stage to one where it can be integrated into applications, systems, or processes to serve a practical purpose.

Chapter 4

Deployment in One Model

Deployment is a little bit different in One Model since all of the machine learning models you work with are technically in a production environment and can be viewed by other users possessing the necessary application access. In One Model, deploying refers to moving the predictions and other results of running a model from within the One AI tab to the data model and, ultimately, Explore and Storyboards.

Deploying also changes the status of the model run, indicating to users that it has been deployed. Even once results are deployed, models should be continuously monitored to maintain performance and relevance over time. This involves updating and rerunning when new input data becomes available. If the performance, fit, and selected features still make sense, you should deploy the new model results.

Chapter 5

Deploy, Deploy & Persist, & Ignore

Once a machine learning model has run successfully and is in a pending status, you have a few options for what action to take next depending on your objectives. You can deploy your model, which as we discussed in the last section, is how you move your data from within the One AI tab to other parts of One Model where it can be more widely shared. Deploying loads the results of that model run into the data model, feeding any metrics, dimensions, columns, and storyboards created for reporting on this data. Storyboards created from machine learning data are generally configured to display the most recent deployed run and should automatically refresh with the newly deployed results once your site has subsequently been processed.

This is an easy way to share the data and insights from the mall with stakeholders in a clear interpretable way. By enabling SHAP, you are also able to create visualizations that showcase the predictive drivers and their directionality, which allows users to understand what is driving the model's predictions and how. What deploying model results does not do is save the model. Since One AI is an AutoML platform, by default, it selects an algorithm and settings based on the data it's trained on.

This means that rerunning a machine learning model can result in changes to that model if the training data has changed. This can happen even if you change nothing in the configuration.

This is where deploy and persist comes into play. Deploy and persist not only deploys the results of the model run, but it also saves the model in a frozen state, meaning it retains its characteristics from run to run, regardless of whether the training data changes. After persisting a model, a new model will not be trained on subsequent runs. Instead, the existing model, including the estimator, configuration settings, and features will be used. The only thing that may change is the predictions since new data can be added to the predict set. So, if your model is making predictions on active employees and new employees join your organization, they will also receive predictions.

Model creators and practitioners also have the option to ignore model runs. This action prevents it from ever being deployed or persisted in the future. This may be done if you were doing something experimental that you do not want to share. It also indicates to others who might view the model runs that this one should be ignored.

You are still able to view the model configuration and accompanying reports, so it's a nice alternative to deleting a model once your experimentation is complete. Note that ignoring only applies to that particular run of the model, and subsequent runs can still be deployed if desired.

Finally, you are also able to leave your model in a pending state for as long as needed until you wish to take another action. In the next section, we will move over to the demo site so I can show you how to deploy, deploy and persist, and ignore models in One AI from the Results Summary report.

Chapter 6

Deploying a Model in One Model

The capability to deploy, persist, or ignore a machine learning model run can be found on the Results Summary report for completed runs of a machine learning model. These actions are taken from the bottom of the Results Summary because the information from the EDA report and the Results Summary report, such as selected features and model performance, should be taken into consideration prior to taking these actions. Click on the One AI tab in the main ribbon menu and scroll down to the model you wish to deploy, persist, or ignore. Click on the 'Runs' button, and then select a run of the machine learning model that's in a 'Pending' status.

Remember, only models in a pending status can be deployed, persisted, or ignored.

This will automatically open the EDA report. Review the report to ensure that it's the correct run and you're happy with the selected features. Then click the 'View Results Summary' button. Again, review the Results Summary report and make sure you're happy with how the model performed. Once you've done this, you will find the 'Ignore' button, 'Deploy' button, and 'Deploy and Persist Model' button located at the bottom right.

Ignoring a model takes effect immediately.

If you deploy or deploy and persist a model, the output from the model run is fed into a data source on your One Model site. Your site's data must refresh before the model results will be available for use in Explore and Storyboards.

Plan accordingly as this usually takes place overnight. If it's urgent, you can work with your customer success team to manually refresh your site's data off cycle. Be mindful that this generally takes anywhere from 1-6 hours. It's worth noting that the first time each model is deployed on your site, there may be data engineering work necessary.

This is work that your data engineer performs in the processing script to build the necessary tables and dimensions to enable building machine learning metrics and storyboards. The bulk of this work only needs to be performed once on your site, but changes may be necessary to accommodate output from new recipe types. As I've mentioned, if you have multiple runs of the same model deployed or persisted, the iteration deployed or persisted most recently will usually automatically be the version available in Explore and Storyboards.

Chapter 7

Conclusion & Thanks

This module has equipped you with the knowledge and skills to deploy models and results within One Model. You can now differentiate between deployment actions and share model insights with stakeholders. Remember to consider model performance, interpretability, scalability, and compliance before deploying models. By following best practices and leveraging One AI's features, you can confidently manage machine learning models to drive impactful business decisions and optimize your data-driven workflows. Happy modeling!