“Regression Model Evaluation” Module Transcript

Chapter 1

Intro, Topics Covered, & Learning Outcomes

Hi, my name is Hayley and I'm on the One AI team here at One Model. In the "Regressions" module, you explored the fundamentals of regression and the types of problems they can help solve. Now we will focus on evaluating the performance and feature selection of these models to determine their effectiveness, helping you answer that million dollar question, is this model reliable or does it need further refinement before we can trust its predictions?

During this module, we will cover a regression evaluation overview, an introduction to common evaluation metrics, feature importance, and the importance of interpretation and iteration.

Once you've completed this module, you will understand and apply key regression evaluation metrics to determine model accuracy and reliability, analyze feature importance to identify which variables most significantly impact your model's predictions and make informed decisions about model refinement. And you will recognize that model building is an iterative process, involving refinement based on evaluation, domain knowledge, & stakeholder feedback.

Chapter 2

Regression Evaluation Overview

Section 2 - Regression Evaluation Overview

Once a regression model run is complete, it is important to evaluate its performance to ensure the model produces accurate and reliable predictions. This involves comparing the model's predictions with actual outcomes and using various performance metrics tailored to the specifics of the model and the model dataset.

Cross validation enables us to accurately measure how well the regression model performs. As you learned in the "Cross Validation" module, One AI splits the data into two subsets before running a model: training folds and a validation fold.

After the model is trained on the data from the training folds, we evaluate its performance on the validation fold to assess how well it performs on new, unseen data. This process helps us evaluate the model’s generalization ability, ensuring it will work effectively with new, real-world data.

It's important to identify whether the model is underfitting or overfitting the training data. If the model performs well on the training data but poorly on the validation data, it suggests the model has learned the noise in the training data rather than the underlying pattern, which indicates overfitting.

Conversely, if the model performs poorly on both the training and validation data, it indicates the model is too simple to capture the underlying patterns in the data, which is known as underfitting.

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.

It's really important to make sure everyone that will be using the model results is comfortable with and understands the model's configuration, feature selection, and performance. 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 also 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 are we relying on it for critical decisions like determining salaries or medical diagnoses? Or is it somewhere in between the critical decisions and data exploration?

Different domains have unique data properties, goals, and constraints which must be considered for the model to be genuinely useful and effective. 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 much less important.

Furthermore, considering the characteristics of the dataset is essential. A representative dataset helps the model generalize well to real world scenarios. Addressing issues like distribution, skewness, and outliers improves model performance and managing multicollinearity ensures stability. Tailoring the evaluation to domain specific needs ensures the model meets practical requirements.

Finally, we should 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 not actually a very good model.

Chapter 3

An Introduction to Evaluation Metrics

Section 3 - An Introduction to Evaluation Metrics

There are many excellent tools available to evaluate the performance of regression models. 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 whenever possible.

Let's go over some of the most common and relevant evaluation metrics.

Chapter 4

Explained Variance Score

We'll start with the explained variance score, which measures the proportion of the variance in the target variable that is captured by the model. Here is the actual equation so you know how it's calculated. Variability in the target variable refers to how much the values of the target variable, which is the outcome you're trying to predict, differ from each other. For example, if we are predicting employee salaries, the target variable is salary, and the variability in salaries means how much salaries differ across employees. Some employees might earn $30,000 a year while others earn $90,000.

The explained variance score tells us how well our model captures this variability among employees versus just outputting the average value. The explained variance score ranges from zero to one, with one meaning the model performs perfectly, and zero meaning the model performs no better than guessing the average value.

So, for our salary example, if the score is close to one, it means the model is good at capturing most of the differences in salaries. It can predict high and low salaries accurately based on the input features. But if the score is close to zero, it means that the model doesn't capture the difference as well, and its predictions are often far from the actual salaries and closer to the average of all of the salaries.

Let's head over to One Model to look at a salary prediction regression model in One AI so we can practice analyzing explained variance scores.

This example model here is ultimately trying to predict employees' salaries, but in doing so, also explain why different employees earn different salaries. The explained variance score can be found in the Results Summary toward the bottom in the Regression Report section.

0.9037 means that this model explains about 90.4% of the variation in employee salaries. The remaining 9.6% is what the model model doesn't explain or what the model cannot predict. This unexplained part could be due to factors we didn't measure. For example, if you don't have performance review data or something else important that the model needs to learn from, could be due to natural randomness, or errors in the data. Overall, 0.9037 is a great explained variance score for the purposes of this exploratory model where 100% perfection is not necessary.

Chapter 5

Mean Squared Error & Root Mean Squared Error

Next, let's talk about mean squared error or MSE. MSE is the average of the squared differences between predicted and actual values. It's useful when you want to emphasize larger errors as squaring the difference penalizes larger errors more than smaller ones. This can help identify models that have significant outliers or large prediction errors. The value of MSE is in squared units of the target variable, which can sometimes make it harder to interpret directly.

Sometimes looking at the root mean squared error or RMSE can provide more context and make MSE more interpretable. RMSE is the square root of the MSE. It converts the MSE back into the original units of the target variable. It's more intuitive to understand because it gives an error metric in the same units as the target variable, making it easier to relate to real world scenarios.

Using both MSE and RMSE provides a more comprehensive evaluation of the model. While RMSE gives a clear picture of the average error, MSE can highlight the presence of large errors that might need to be addressed.

Let's go back into One Model to take a look at that salary prediction regression and practice analyzing MSE and RMSE.

One hundred and 101,674,610 seems really large, but that's partially because salaries are quite large and it's squared. This large value indicates that there are errors in this prediction. Taking the square root of this MSE gives an RMSE of about 10,000, which means the average prediction error is about $10,000 per employee in the model dataset. This tells us that there is room for improvement with this regression model.

Chapter 6

Residuals Plot

The first visual I will cover is the residuals plot, which is a graph that shows the residuals on the vertical axis and the predicted values on the horizontal axis. Residuals are the differences between the actual values and the predicted values by the model. The residuals plot helps visualize how well the regression model fits the data. It can highlight any patterns in the residuals, which indicate issues with the model.

In a good model, the residuals should be randomly scattered around the horizontal axis, the zero line, showing no clear pattern. This suggests that the model's predictions are unbiased and errors are random.

We will use the image here from the salary prediction regression model in One AI to practice interpreting a residuals plot. First, let's talk about the r squared in the upper left corner. The Train R^2 equals 0.893, which indicates the model explains 89.3% of the variance in the training data. The Test R^2 equals 0.903, which indicates the model explains 90.3% of the variance in the test data. This is a good indicator that the model generalizes well to new and unseen data.

Next, we want to look at the distribution of residuals. As you can see, the residuals are mostly scattered around the zero line, which is good. However, there are some noticeable patterns and clusters, especially as the predicted values increase. There are some points with very high residuals, which are large errors, both above and below the zero line. These could be outliers or cases where the model significantly under or over predicted the salary.

Then, we want to look at the spread of the residuals. Here we see they increase as the predicted values increase, suggesting that the model may not perform as well for higher salary predictions.

Finally, look at the histogram on the right side of the graph. The distribution of residuals on the right shows most residuals are clustered around zero, but with some extreme values. This suggests the model performs well for many predictions, but has significant errors for some.

My overall takeaway from this residuals plot is that the model performs well and generalizes effectively, but there are some areas, particularly for higher salary predictions, where the model could be improved.

Chapter 7

Prediction Error Plot

Next is the prediction error plot, which shows the predicted values against the actual values to help visualize how close the model's predictions are to the actual values. Ideally, if predictions were perfect, all points would lie on the 45 degree line, which is called the identity line, where the predicted value equals the actual value.

We will use the image here from the salary prediction regression model in One AI to practice interpreting a prediction error plot. The R^2 equals 0.903 indicates that the model explains 90.3% of the variance in actual salary data. This is a strong indication of good model performance.

Additionally, most of the points are clustered close to the identity line, indicating that the model's predictions are generally close to the actual salaries.

Some points are further away from the identity line, especially at higher salary values. This deviation indicates that the model's predictions are less accurate for these cases, which is similar to what we observed in the residuals plot.

Chapter 8

Prediction Distribution Plot

Next is the prediction distribution plot, which shows the distribution of predicted values from the regression. It is often compared to a normal distribution to see how the predicted values are spread out. It helps visualize the range and frequency of predicted values and shows whether the predictions are concentrated around certain values or spread out and if there are outliers. We will use the image here from the salary prediction regression to practice interpreting a prediction distribution plot.

First, we want to look at the shape of the distribution. The predicted values form a distribution that appears somewhat skewed to the right, indicating that most salary predictions are concentrated at the lower end.

Next, we can see that the peak of the distribution is around $50,000 to $100,000, suggesting that the majority of salary predictions fall within this range. The spread of the predicted values shows a long tail towards higher salaries, indicating some high salary predictions, but in much smaller numbers.

The overlaid normal distribution curve shows that the predicted values do not perfectly follow a normal distribution, which is common in real world data where salaries are often right-skewed.

Chapter 9

Prediction Comparison Plot

Next, we have the prediction comparison plot, which shows the distribution of the predicted values compared to the distribution of the actual values, which are the previous period's target values. It provides an easy way to compare the distributions and see if the model captures the overall pattern and spread of the target variable.

We'll use the image here from the Salary Prediction regression to practice interpreting this type of plot. So, the plot shows two distributions: the previous period's actual salary distribution in blue and the predicted salary distribution in green. And the extent of overlap between these distributions indicates how well the model's predictions match the actual values.

Both distributions have similar shapes with peaks around the same range, $50,000 to $100,000. This suggests that the model is reasonably good at predicting the overall distribution of salaries.

When looking at the differences in spreads, we can see that the predicted distribution has a slightly narrower spread compared to the actual distribution. This suggests that the model may be underestimating the variability in salaries.

Finally, the long tail in the actual salary distribution, again, the blue, shows that there are some very high salaries while the predicted distribution in green does not capture these well, indicating that the model struggles with predicting higher salaries as we've seen in other visuals.

Chapter 10

Outlier Dataframe

And finally, we have the outlier dataframe, which lists records that have unusually high or low residuals (or errors) compared to the majority of data. These records are considered outliers because their predictions significantly deviate from their actual values. For our salary example, this would mean their predicted salaries are far from their actual current salaries.

It utilizes Z-scores, which are a statistical measure that describes a value's relationship to the mean of a group of values. In this context, it shows how many standard deviations a residual is from the mean residual, and a high absolute value indicates an outlier.

For example, Person ID 7285 with a Z Score of 13.6 is a significant outlier. Key features include not being a manager, high annual salary range, and high future manager potential. The model significantly underestimates or overestimates this employee's salary, indicating potential missing features or nonlinear effects not captured by the model.

It's helpful to look for patterns and common features among the outliers to gain insights into potential weaknesses in the model and areas for improvement. For example, many outliers in this particular model happen to be managers or future managers, indicating the model may not fully account for the salary dynamics of managerial roles.

Chapter 11

Feature Importance

Section 4 - Feature Importance

When evaluating a model, we can't only consider evaluation metrics, but should also look at feature importance. It's beneficial to use both of these tools because performance metrics give us insights into the accuracy and reliability of the predictions, but the feature importance analysis explains why the model makes certain predictions, revealing the underlying drivers of the outcomes.

Feature importance analysis helps identify which input variables have the most significant impact on the target variable in a regression model. Here's the general process. First is training the model using input features and the target variable. Next is calculating feature importance, and this all depends on the type of model. So, different methods are used to calculate feature importance depending on what type of regressor is used.

For example, with a linear regression, feature importance is directly related to the model coefficients. Each feature's coefficient represents the change in the target variable for a one unit change in the feature, assuming all other features remain constant. So positive coefficients indicate that as the feature value increases, the predicted value of the target variable increases, while negative coefficients indicate that as feature value increases, the predicted value of the target variable decreases. And then for the magnitude of coefficients, the larger the absolute value, the greater the importance.

However, with tree-based models, feature importance measures how often and how significantly a feature is used to split data across all trees. So, again, it just all depends on the regressor.

In general, this process ranks and scores the selected features to convey how important including each feature in the model is to making accurate predictions.

This analysis helps reveal if the model is relying too heavily on certain features, which might indicate a bias or an overfitting issue. For example, in this model, there is a positive coefficient for gender male, which could indicate potential gender bias and likely warrants further investigation to ensure fair salary predictions.

Additionally, feature analysis can help with identifying important features leading to a more concise model by eliminating irrelevant or less significant features. This can reduce complexity and improve generalization. You can remove features with less importance or reduced dimensionality.

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 12

Interpretation & Iteration

Section 5 -Interpretation and Iteration

When interpreting your model with the tools we've covered, it's helpful to understand the model 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 that models continuously evolve to meet your changing needs.

Chapter 13

Conclusion & Thanks

We covered the key aspects of evaluating regression models, including performance metrics and feature importance analysis. Using both tools together provides a complete picture of your model's accuracy and the factors influencing its predictions. Remember, model building is an iterative process that involves continuous refinement based on evaluation results and feedback to ensure the model remains effective and reliable. Happy modeling!