“Regressions” 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 "Supervised Learning" module, you learned that supervised learning can be categorized into classifications and regressions.

We dove into classifications in the previous module and will now begin to explore regressions to give you some insight into another major type of model you can build in One AI with the group attrition recipe being a regression model client favorite.

This module will be structured similarly to the classification models module. We will cover an overview of regressions, common regression algorithms, strengths and weaknesses of regressions, and some people analytics use cases.

After completing this module, you will understand regressions as a supervised machine learning technique for predicting continuous outputs.

You will be introduced to common regression algorithms, such as linear regression, ridge, lasso, elastic net, and random forest.

You will identify how One AI takes advantage of the strengths and mitigates the weaknesses of regressions, and you will explore regression model use cases and how these models enable organizations to extract valuable insights from employee data and inform decision-making.

Chapter 2

Regression Overview

Section 2 - Regression Overview

Regression is a common supervised machine learning technique for predicting continuous values.

The goal is to build a model that can accurately predict a target variable, which is also called a dependent variable, based on one or more input features, which are also called Independent Variables.

The input features are part of the labeled historical data that the model learns from. It models the relationship between input features and the target variable by plotting a line or curve of best fit to the data.

In simple linear regression, a line of best fit is plotted to the data. For polynomial or other nonlinear regressions, a curve is plotted. This is based on the number of input features and output labels. A regression is classified as linear when there is one input and one output. It's classified as multiple when there are many inputs and one output, and multivariate when there are many inputs and many outputs. The line of best fit or curve serves as a predictive model to estimate the output variable based on the input variables.

Once trained, the model can make predictions on new or unseen data and fill gaps in missing data.

Training a regression model requires labeled data, which is a necessity for all supervised machine learning models. The model learns from datasets containing examples with known input features and the corresponding target values. For example, if using regression to predict employee performance ratings, the target value is the rating and the data points like job role, education, skills, and location are the input features. The model uses this labeled data to learn relationships between these features and the target value. Once trained, it can predict performance ratings based on input features.

Typically, the dataset is split into a training set for the model training and a testing set for performance evaluation on unseen data. As with all supervised machine learning, special care should be taken to ensure the labeled training data is representative of the overall population.

If not, the model may overfit to unrepresentative data, leading to inaccurate predictions on new, unseen data once the model is deployed. Additionally, regression analysis requires careful selection of features to accurately capture relationships between features and outcomes.

Chapter 3

Regularization

Regularization is a technique used to prevent overfitting and improve the generalization of regression models. Overfitting occurs when a regression learns not only the underlying patterns in the training data, but also the noise and random fluctuations leading to poor performance on new data.

Regularization introduces a penalty term to the model's loss function, discouraging it from fitting the training data too closely and controlling the model's complexity.

This encourages the model to prefer simpler patterns that generalize better to new data. The goal is to balance fitting the training data well and avoiding excessive complexity. We'll discuss the two main regularization methods in the algorithm section of this module.

Chapter 4

Regression Algorithms

Section 3 - Common Regression Algorithms

Machine learning models use algorithms to learn from data, identify patterns, make predictions, or perform tasks without explicit programming.

An algorithm is the mathematical procedure, technique, or set of rules that the model follows to do so.

Regression algorithms fall into the supervised learning algorithms category, in which algorithms learn from labeled training data, where each example is associated with a known output or target label.

Some common examples are linear regression, which is one of the simplest and most widely used regression algorithms.

It estimates the linear relationship between input features and a continuous target variable by minimizing prediction errors through parameter optimization.

The ridge regression is a variant of linear regression that uses L2 regularization to prevent overfitting.

It adds a penalty term to the loss function, which penalizes large coefficients, making it effective for dealing with multicollinearity.

Lasso is similar to Ridge in that it also uses regularization. However, it uses L1 regularization, which tends to produce sparse models by setting some coefficients to zero, making it useful for feature selection and building simpler models.

And then we have elastic net, which combines the penalties of ridge and lasso regression. It includes both L1 and L2 regularization terms in the loss function, balancing between feature selection and regularization.

It's useful in tasks with datasets that have high dimensional features and potential collinearity.

Next is the decision tree regression, which builds a decision tree to model the relationship between the input variables and the target variable. It splits the data into smaller subsets based on feature values and predicts the target variable for each subset.

Next, we have random forest regressor, which is an ensemble learning method that builds multiple decision trees and then averages and combines their predictions to improve accuracy and reduce overfitting.

And finally, we have the LightGBM regressor, which is a gradient boosting framework that uses tree based learning algorithms.

It grows trees vertically - leaf wise - rather than horizontally - level wise - to achieve faster training speed and lower memory usage. It's great for large datasets with high dimensional features and or missing values.

Chapter 5

Strengths & Weaknesses of Regressions

Section 4 - Strengths and Weaknesses

Regressions have their own unique set of advantages and disadvantages.

It's important to understand these to determine if a regression model is a smart tool for your organization's problem domain. Let's start with the strengths. A major strength of the regression model lies in its interpretability.

It's easy for those outside of the field to understand how changes in input features affect the predicted output as coefficients represent the feature's contribution to the prediction.

Additionally, regressions are relatively simple and fast in that algorithms are computationally efficient and can be trained quickly on large datasets compared to more complex models.

Regression models are also versatile. They can handle various types of data and are suitable for both simple and complex tasks. This allows users to gain understanding on simpler versions and then scale up to more complex tasks if needed.

And finally, some algorithms, such as lasso and ridge regression, provide feature selection capabilities by shrinking coefficients towards 0, effectively highlighting important features. Others like decision trees and ensemble methods can automatically assess and rank the input feature importance, providing insights into which variables are most influential in predicting the target.

Moving on to the weaknesses.

Most of the weaknesses that general regression models have are mitigated by how One AI was built, but they're they are still important to understand.

More complex regressions are prone to overfitting. Regularization techniques can help prevent this and improve model generalization.

Linear regression and other parametric models have limited capacity to capture complex nonlinear relationships in the data. They may underperform when the relationship is highly nonlinear.

Additionally, regressions may not handle missing data well without preprocessing techniques, which can introduce bias or reduce model performance. One AI offers null filling strategies to mitigate this risk. And finally, regression models require labeled training data, which isn't always available and can be complicated to obtain. One AI uses the data that you already have loaded into One Model, which takes the pain out of this weakness.

Chapter 6

People Analytics Use Cases

Section 5 - People Analytics Use Cases Where Regressions are Applicable

Regressions are invaluable tools for people analysts and their organizations offering deep insights into employee and recruiting data. By analyzing continuous variables like performance ratings, salary progression, and job satisfaction scores, these models reveal critical relationships and factors influencing workforce dynamics. They help understand the impact of various factors on employee outcomes, such as productivity, compensation satisfaction, and job engagement.

Identifying key predictors and quantifying their influence enables data driven decisions on talent development, compensation strategies, and organizational policies. The insights from regression analysis optimize human resources practices, enhance employee satisfaction, and improve overall organizational performance.

Some common regression models that you can build in One AI include group attrition regression, which is a recipe that helps you predict the amount of attrition for defined groups in your organization by leveraging generative group features. This model can help identify features that impact termination and attrition amounts by group. In a similar vein, we offer employee attrition regressions, which is a custom model that helps you predict the likelihood of an employee turning over based on factors, such as job satisfaction, salary, tenure, performance ratings, and demographic variables.

Next, we have a salary and compensation regression, which is a custom model that can help you predict salaries for your employees. This is useful in determining salary factors, compensation planning, equity analysis, and more.

Next, we have employee engagement regressions, which are custom models that help you analyze survey responses, sentiment analysis, and feedback data to predict employee engagement levels and satisfaction. By understanding the factors influencing engagement, companies can implement targeted interventions to improve employee morale and retention.

And, of course, if you have any other ideas for regression models, custom models can be built as long as you have the appropriate data and a defined metric in One Model. The sky's the limit.

Chapter 7

Conclusion & Thanks

This module has equipped you with essential knowledge for leveraging regressions and people analytics.

Understanding regression fundamentals, including model types and their strengths and weaknesses, prepares you to predict and analyze key employee data aspects like salary factors, attrition risks, and engagement levels. With the versatility and applicability of regression models demonstrated through specific use cases within One AI, you are empowered to drive organizational performance improvements through data-driven approaches. Happy modeling!