Microsoft

# Introduction to regression models using R and Tidymodels

Follow along with these modules at
https://aka.ms/learnlive-20220909A

Join the chat at
https://aka.ms/LearnLiveTV


**Eric Wanjau**

Data Scientist and Researcher
Leeds Institute for Data Analytics
University of Leeds
@ericntay



**Karani Wachira**

Undergrad student
DeKUT R Community Lead
Dedan Kimathi University of Technology
@mrAfricanSeed

# Say hi to our moderator 👋!

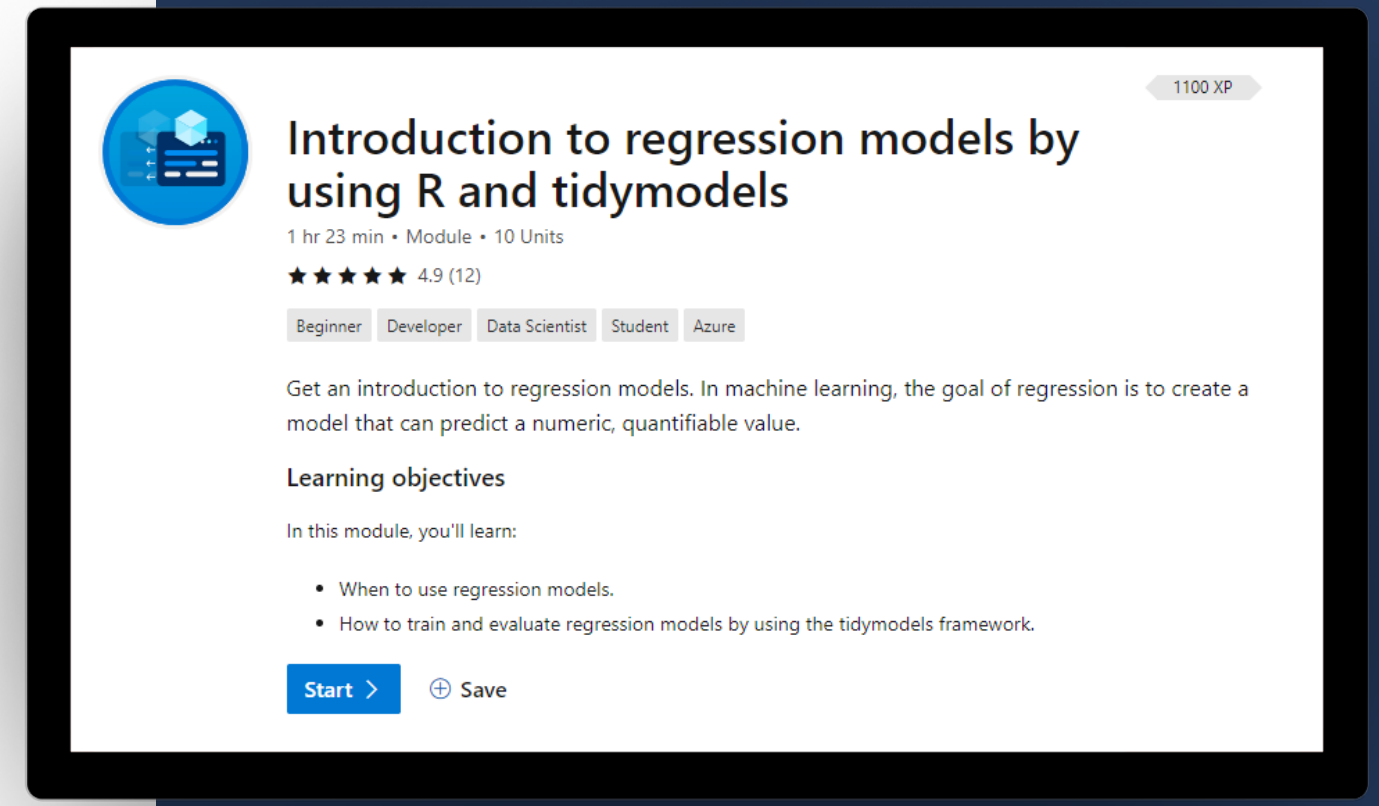
**Pascal Burume**

Network and telecommunications engineer, Microsoft Student Ambassador
Youth Africa Technologies
LinkedIn: /in/pascal-burume-909520218/

# /learn

https://aka.ms/learnlive-20220909A

Complete interactive learning exercises, watch videos, and practice and apply your new skills.



# Learning objectives

- When to use regression models.
- How to train and evaluate regression models by using the tidymodels framework.

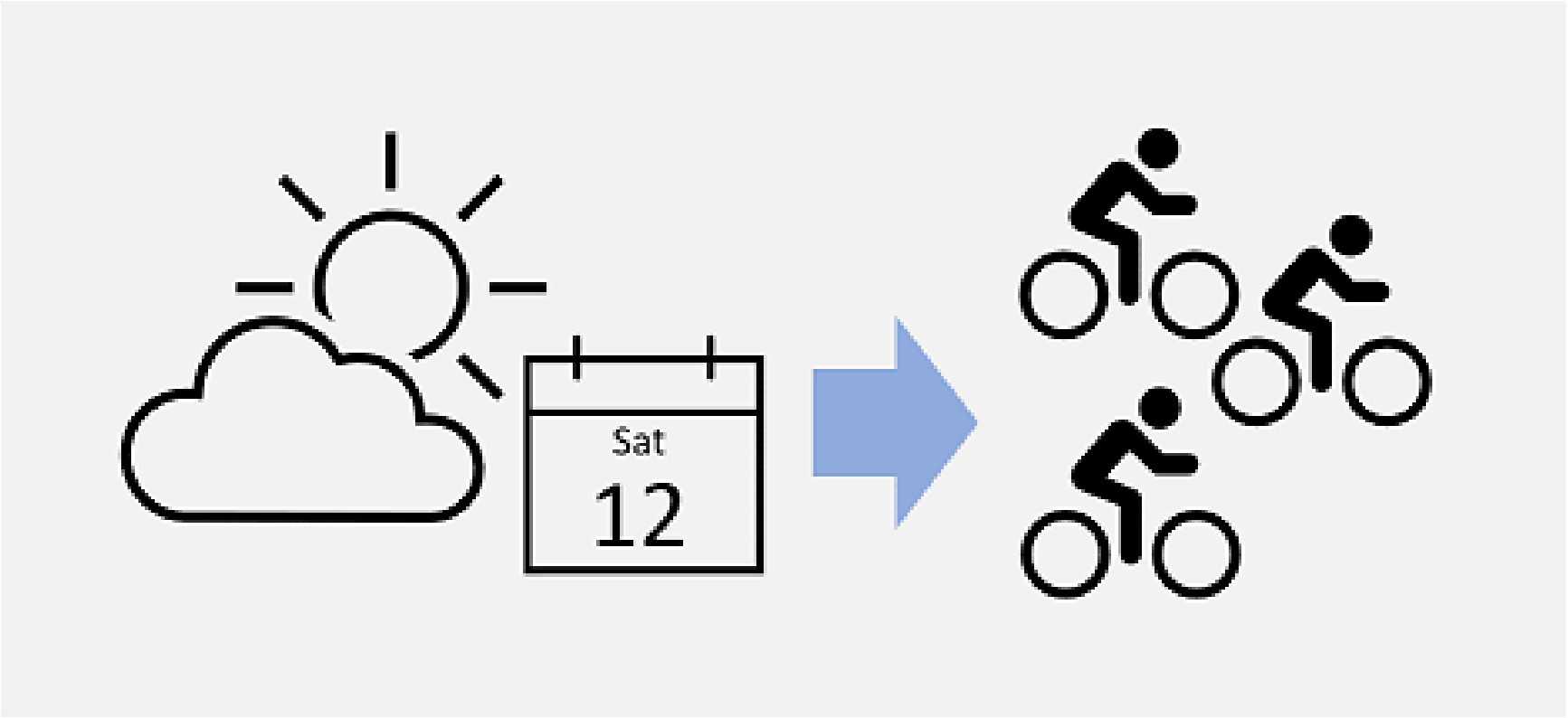
# What is regression?

# What is regression?

Supervised ML technique that works by establishing a relationship between variables in the data to predict a numeric, quantifiable value.

- **A training dataset.** You'll apply an algorithm that determines a function that encapsulates the relationship between the feature values and the known label values.
- **A validation or test dataset.** You can use it to evaluate the model by using it to generate predictions for the label and comparing them to the actual known label values.

# What is regression: Predicting bike rentals



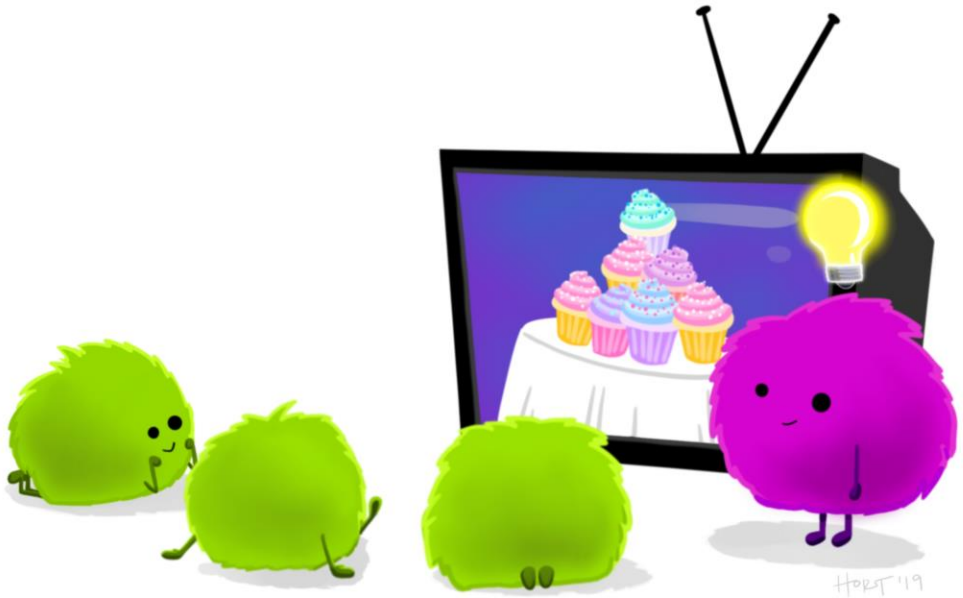
EDA

Data Budgeting

Fit and evaluate a linear regression model

Can we do better?

# The plan: Predicting bike rentals

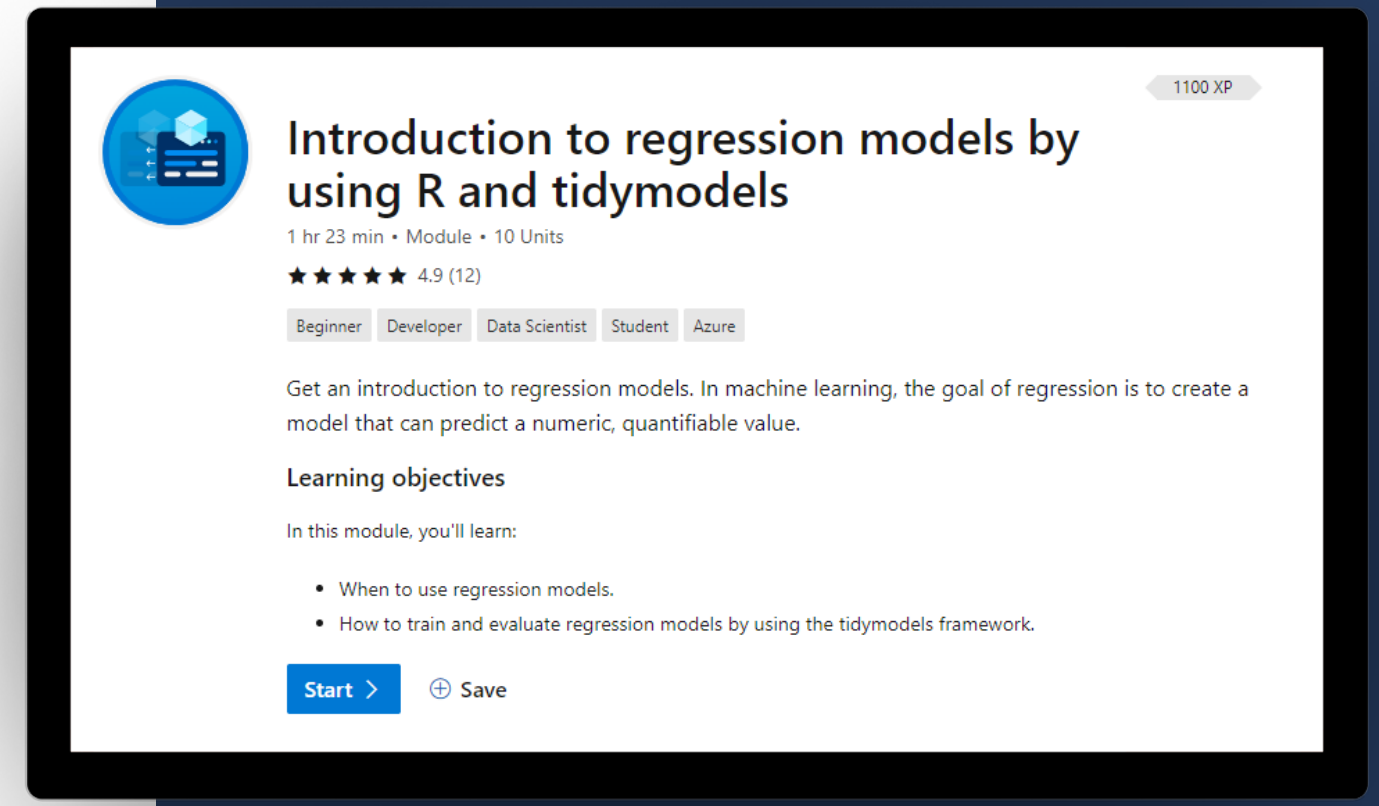


artwork by @allison_horst

# /learn

https://aka.ms/learnlive-20220909A

Complete interactive learning exercises, watch videos, and practice and apply your new skills.



# Knowledge Check

Test your knowledge in the chat

# Question 1

Vote at

**The tidymodels framework was used in R to train a regression model from a dataset of sales data. To evaluate the model to ensure it will predict accurately with new data, what should be done?**

A. Split the data randomly into two subsets. Use one subset to train the model and the other subset to evaluate it.

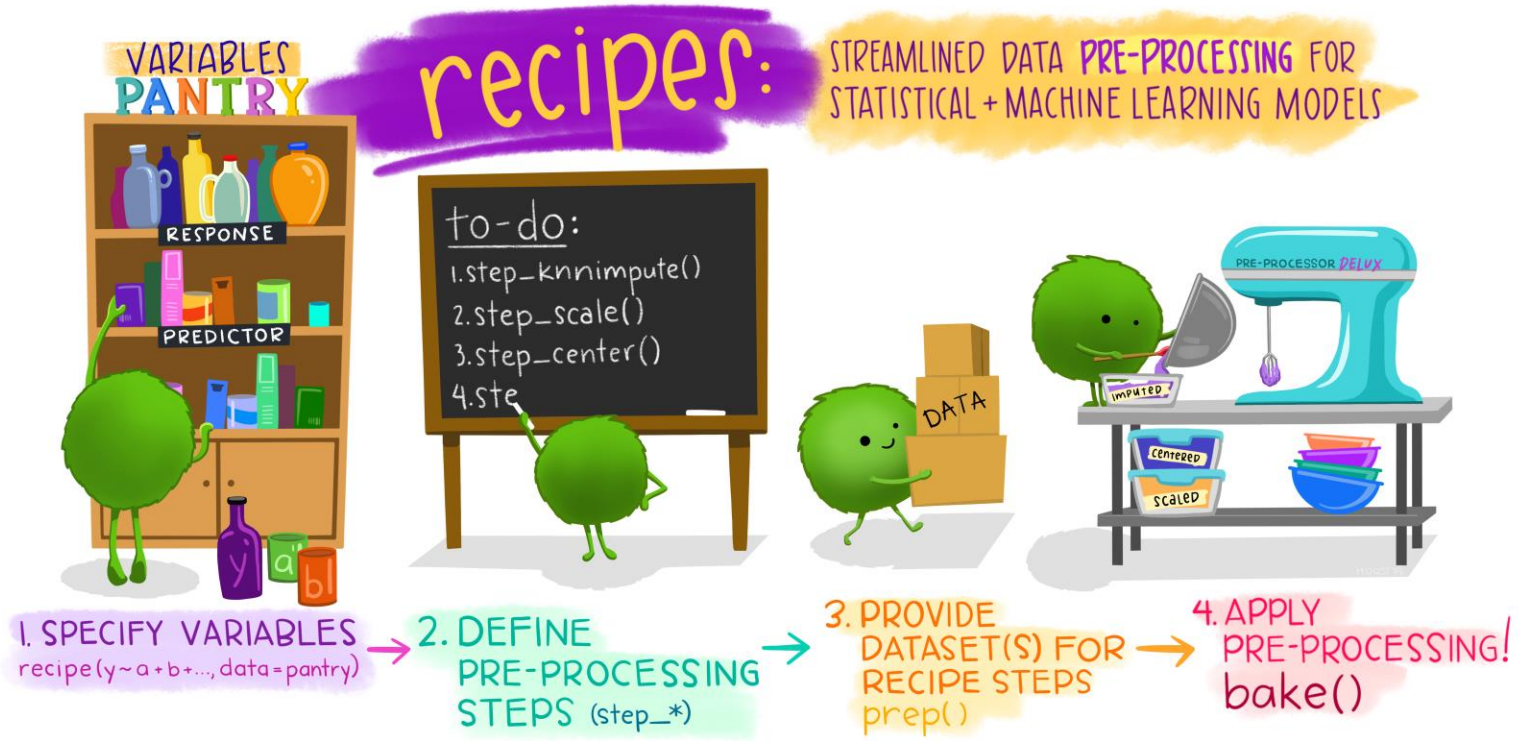
B. Use all the data to train the model. Then use all the data to evaluate it.

C. Train the model by using only the feature columns. Then evaluate it by using only the label column.

Vote at https://aka.ms/polls

# Can we do better?

# Feature Engineering with recipes




## Normalize data

- step_normalize()

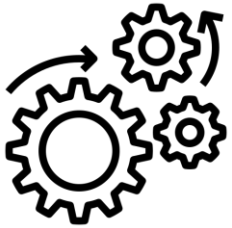
## Impute missing values

- step_impute_mean()

# Try out a different model


# Model Tuning

Some model parameters cannot be estimated directly from the training data.



Instead of learning these kinds of hyperparameters during model training, we tune them

- Try different values
- Measure performance
- Pick the best

# Let's pause for a sec...

- Typically we can't decide on which final model to use with the test set before first assessing model performance.

- There is a gap between our need to measure performance reliably and the data splits (training and testing) we have available

# Resampling for evaluating model performance

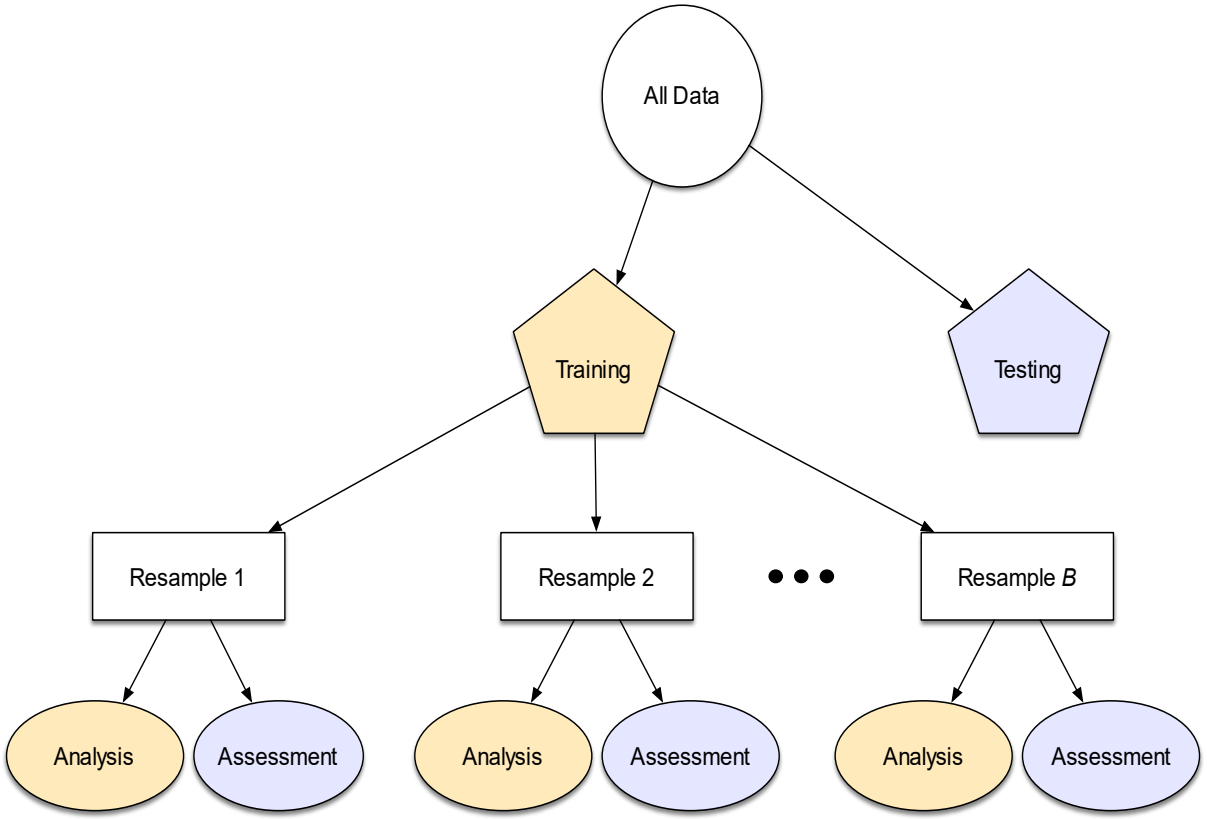



Image from Tidy Modeling with R

## Cross Validation

Image from Julia Silge's ML course

# Knowledge Check

Test your knowledge in the chat

# Question 2

Vote at

**A regression model specification was created by using the linear_reg() function in the tidymodels parsnip package. What should be done to train the model?**

A. Call the predict() function and specify the model specification, formula, and data.

B. Call the recipe() function and specify the model specification, formula, and data.

C. Call the fit() function and specify the model specification, formula, and data.

Vote at https://aka.ms/polls

# Question 3

Vote at

**A regression model was trained by using the tidymodels framework. When it's evaluated with test data, the model achieves an R-squared metric of 0.95. What does this metric say about the model?**

A. The model is 95% accurate.

B. The model explains most of the variance between predicted and actual values.

C. On average, predictions are 0.95 higher than actual values.

Vote at https://aka.ms/polls

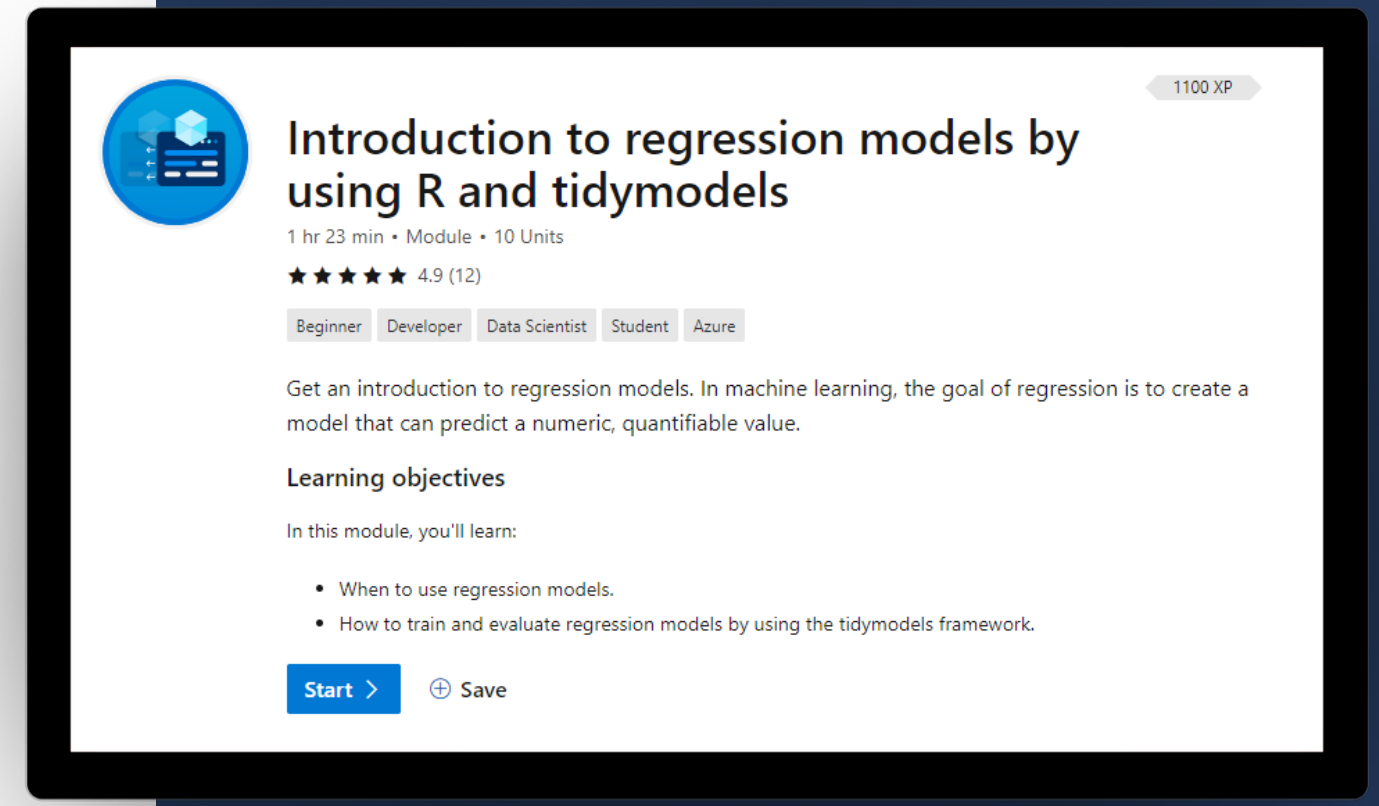
# Summary

- When to use regression models.
- How to train and evaluate regression models by using the tidymodels framework.

# /learn

https://aka.ms/learnlive-20220909A

Complete interactive learning exercises, watch videos, and practice and apply your new skills.



# Don’t miss!

## Introduction to classification models by using R and tidymodels

**September 16, 2022 4:00PM - 5:30PM (Central Europe)**

Continue your learning with the next module in this series!!

https://aka.ms/learnlive-machine-learning-r-tidymodels-Ep3