

Lesson 4.4: Linear Regression

# DISTRIBUTED COMPUTING WITH SPARK SQL

Linear Regression



**Brooke Wenig/Conor Murphy**  
Machine Learning Practice Lead/  
Data Scientist, Databricks

**UC DAVIS**  
Continuing and Professional Education

Slide 2: Welcome Back!



# Welcome Back!

Examine linear regression algorithm

Slide 3: Learning Objectives



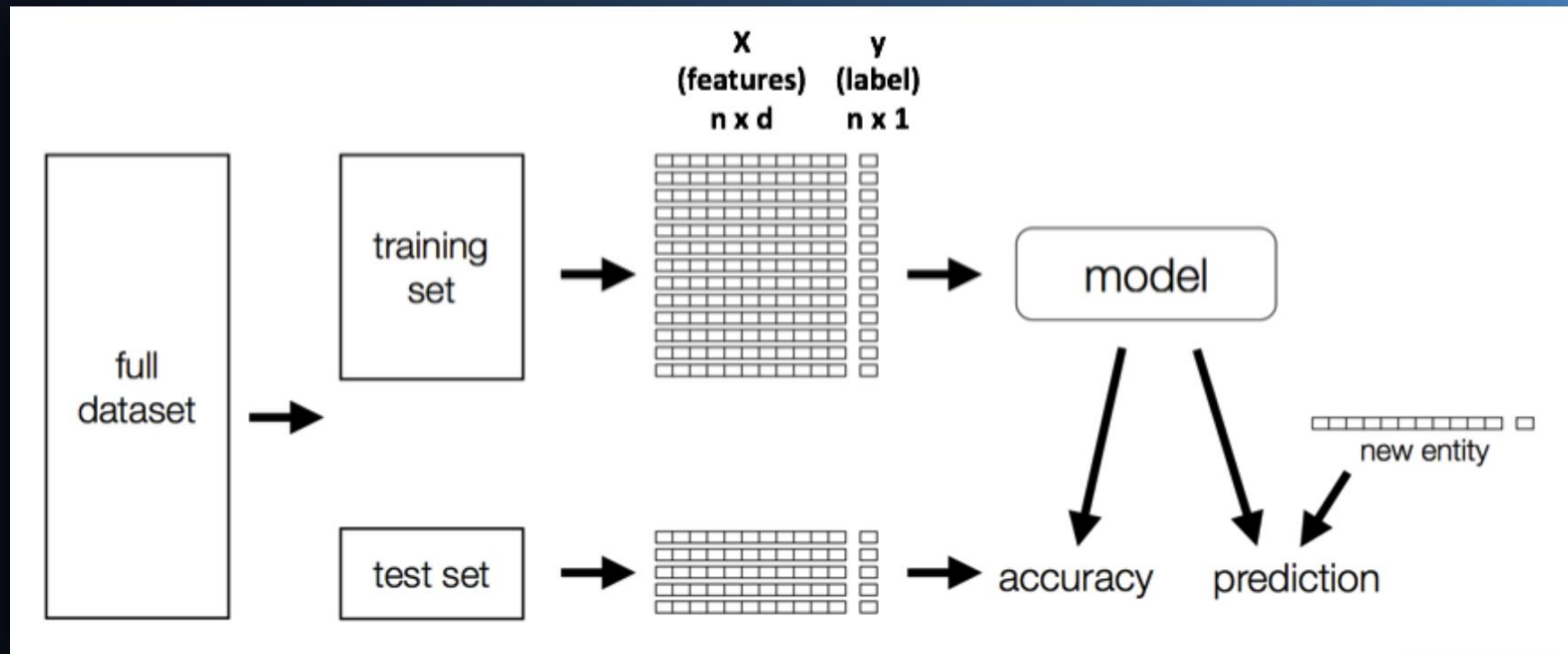
## Learning Objectives

Interpret machine learning models

Identify which use cases you can apply  
linear regression to

## Slide 4: Model Training Lifecycle

## Model Training Lifecycle



Slide 5: How Do We Determine Which Model to Use?

**How Do We  
Determine  
Which Model  
to Use?**



Slide 6: Ask Key Stakeholders

## Ask Key Stakeholders

Do they need an interpretable model?

Explain what independent variables contributed to end prediction



Slide 7: Opting for Interpretable Models Over Accurate Models

## Opting for Interpretable Models Over Accurate Models

Algorithm predicts 80% success

Procedure fails, don't blame the algorithm

You must understand why the algorithm made the prediction





Slide 8: Opting for Interpretable Models Over Accurate Models



## **Interpretable**

Linear regression

Decision trees



## **Accurate**

Neural networks



Slide 9: Accounting for Assumptions

## Accounting for Assumptions

Algorithms have different assumptions about underlying distribution of data

Algorithms are like recipes – a set of step-by-step instructions



## Slide 10: Linear Regression

## Linear Regression

Goal: Find line of best fit

$$y \approx \hat{y} = w_0 + w_1x + \epsilon$$

X: feature

y: label



Slide 11: Assumptions of Linear Regression



## Assumptions of Linear Regression

There is a linear relationship between input features and the output

## Slide 12: Multivariate Regression

## Multivariate Regression

Use of linear regression with multiple variables

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \dots + \beta_p X_p$$

p = total number of features in the dataset

Slide 13: Interpreting Linear Regression

## Interpreting Linear Regression

A highly interpretable model

Examine coefficients to see how a prediction is made

E.g.: If the coefficient for number of calls is -0.5 – then the response time decreases by half a minute for every additional call received

Slide 14: Coming Up



## Coming Up

Train a linear regression model

Apply the linear regression model in SQL  
using a custom function