Identify Data-level Issues in Machine Learning Models

Ravikiran Srinivasulu
SOFTWARE CONSULTANT

ravikirans.com | ravikirans.com/YouTube

Agenda

Imbalanced data in Classification problems

Data scale issues in distance-based problems

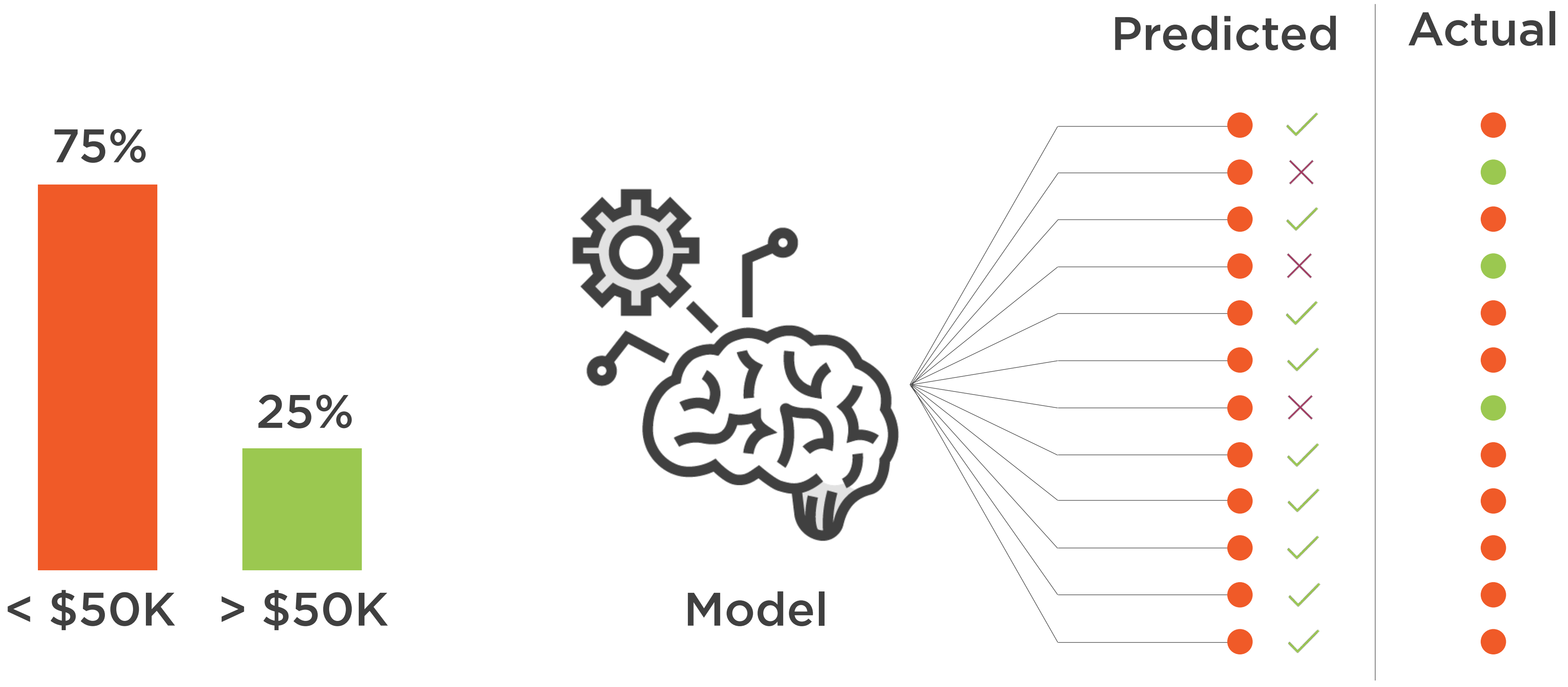
Multicollinearity issues in Regression

Outliers in Regression

Problem with high-dimensional datasets

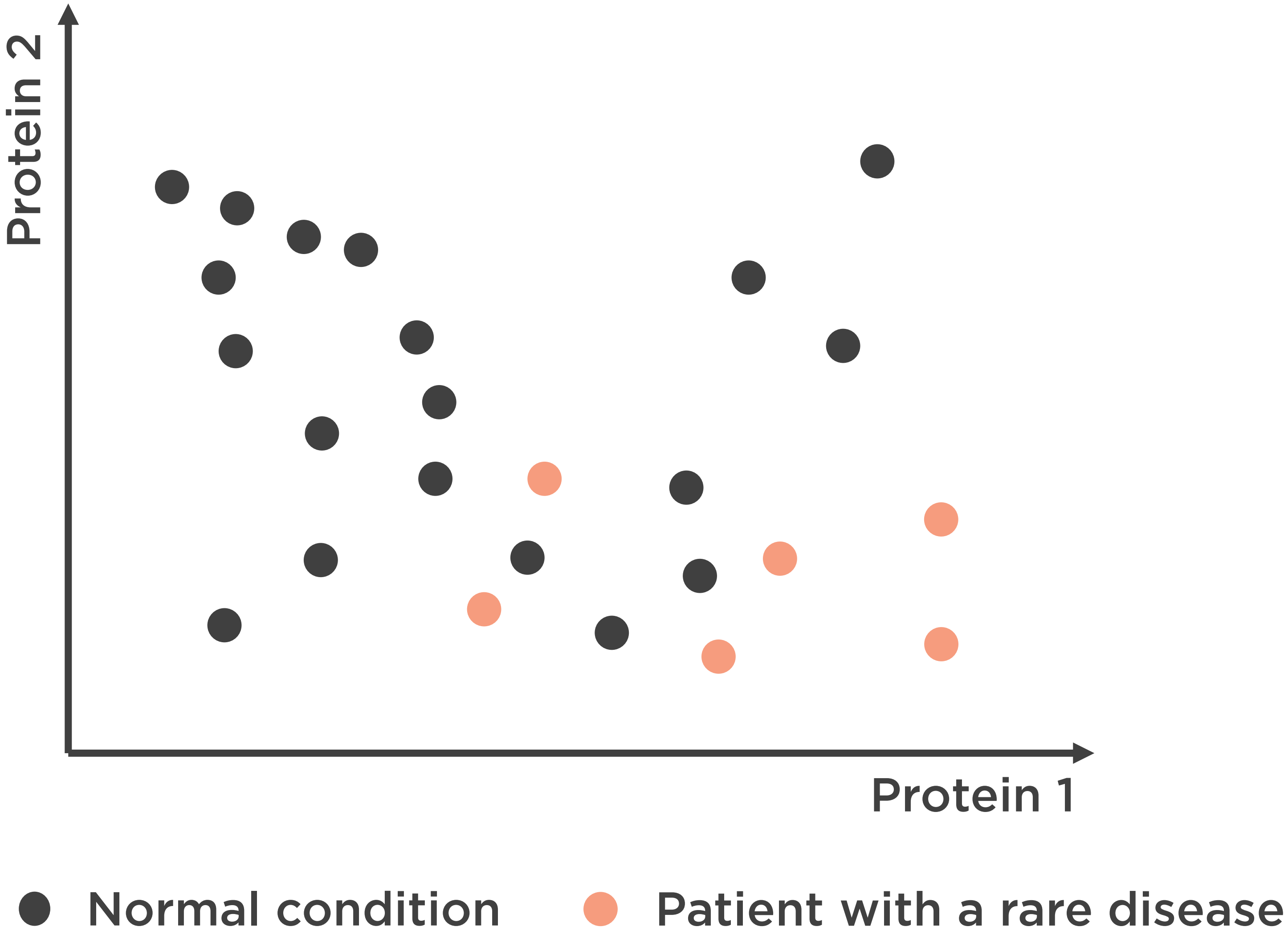
Imbalanced Dataset for Classification Problems

Imbalanced Dataset for Classification Problems

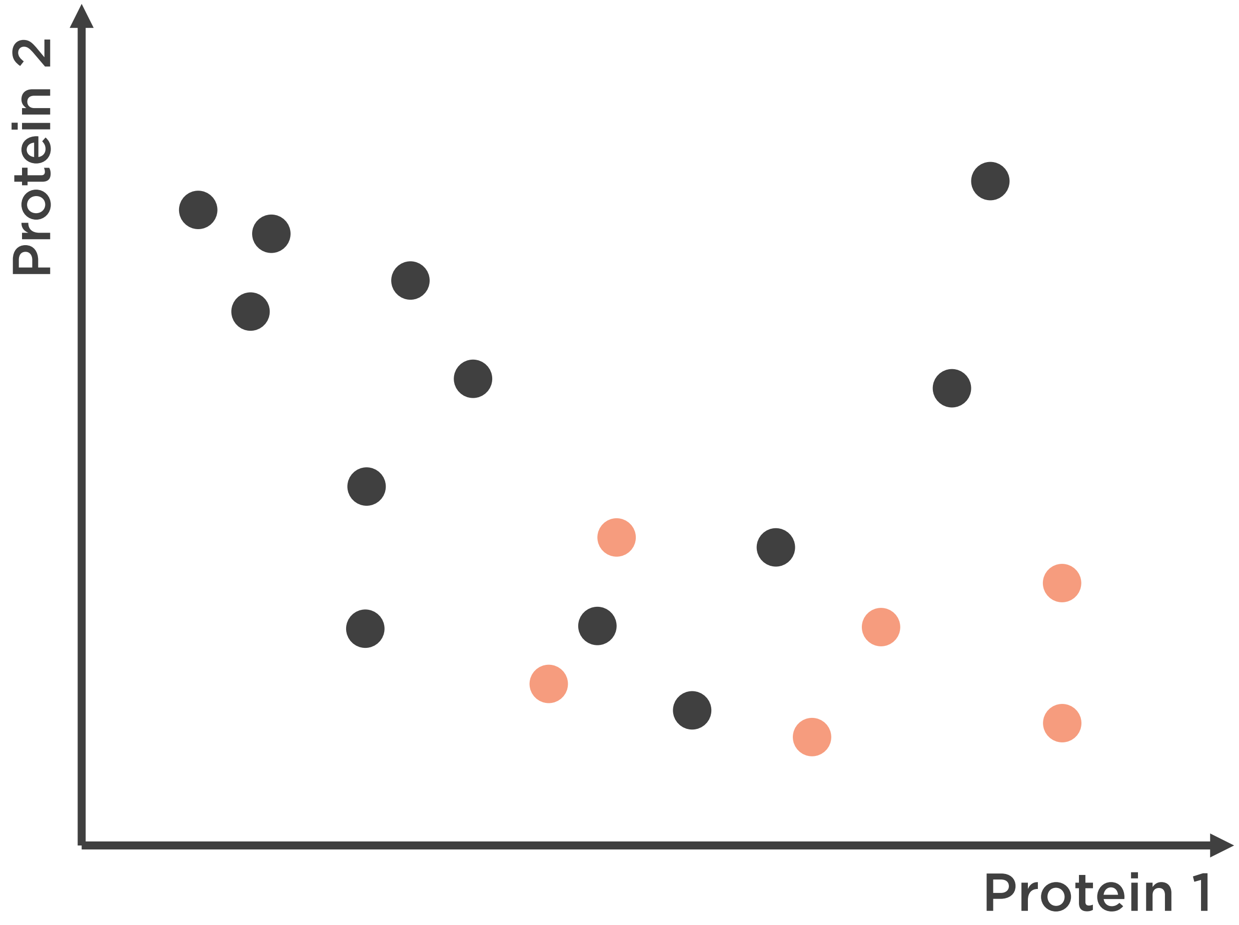


75% accuracy

Imbalanced Dataset for Classification Problems

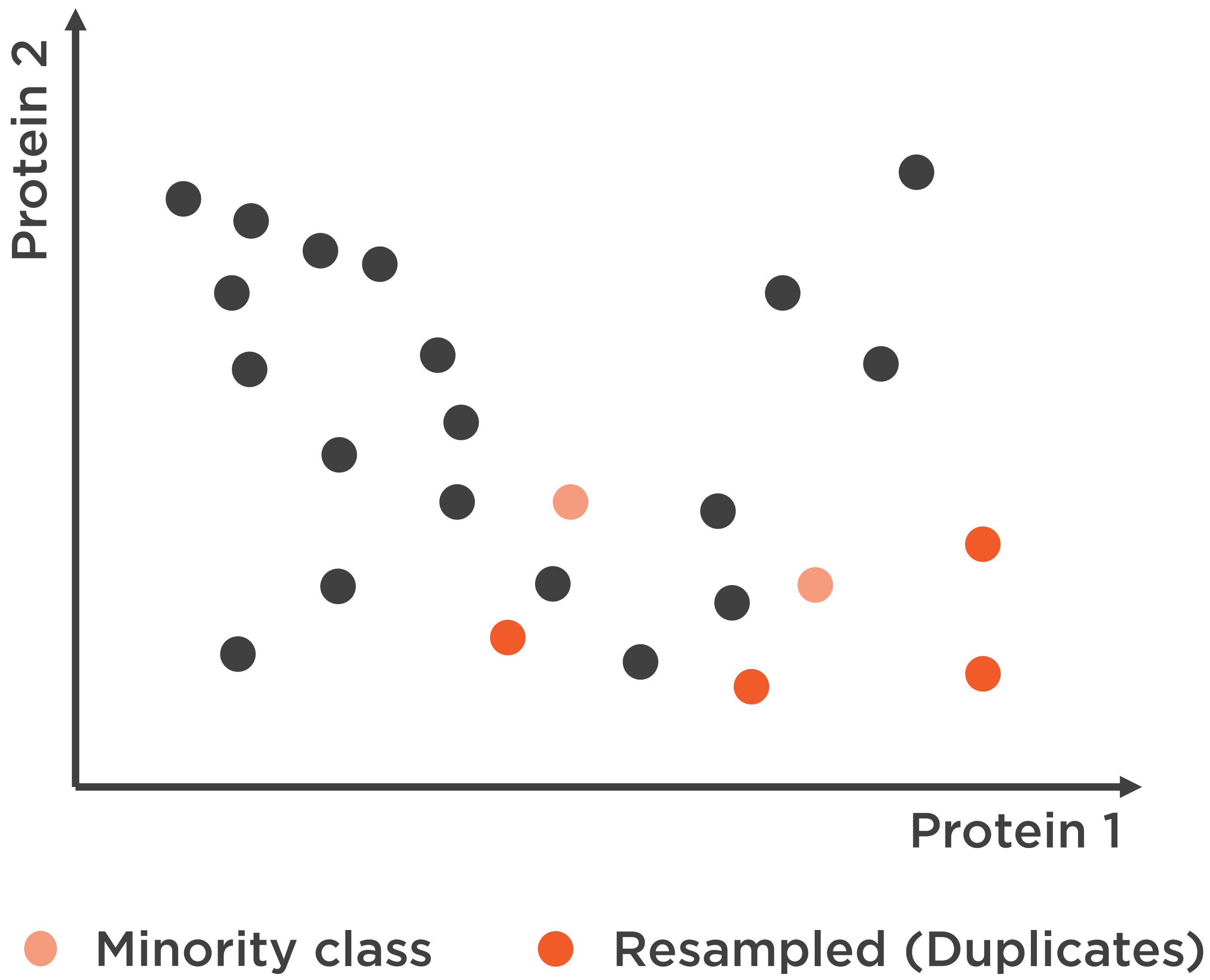


Undersampling

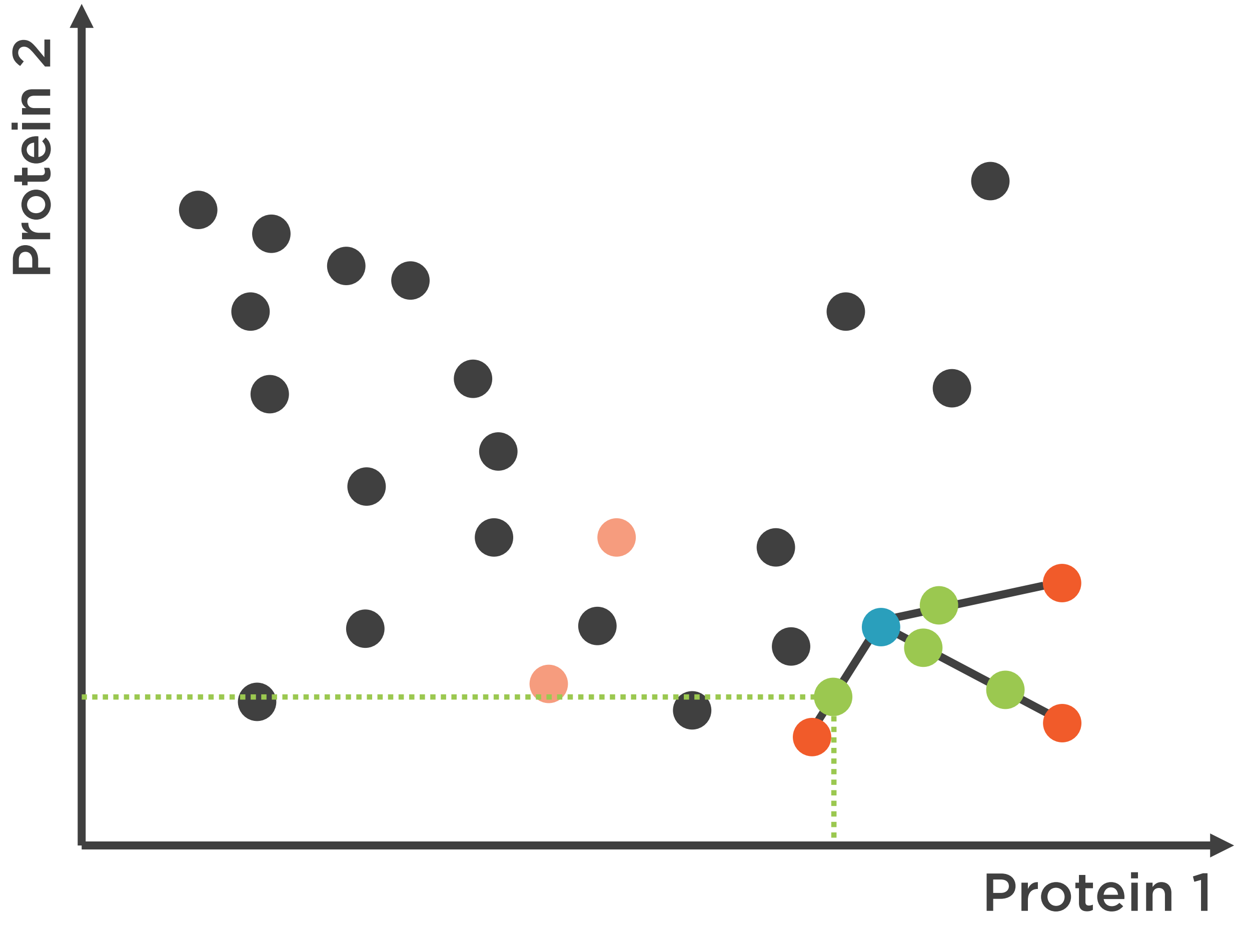


● Normal condition ● Patient with a rare disease

Random Oversampling



Synthetic Minority Oversampling Technique

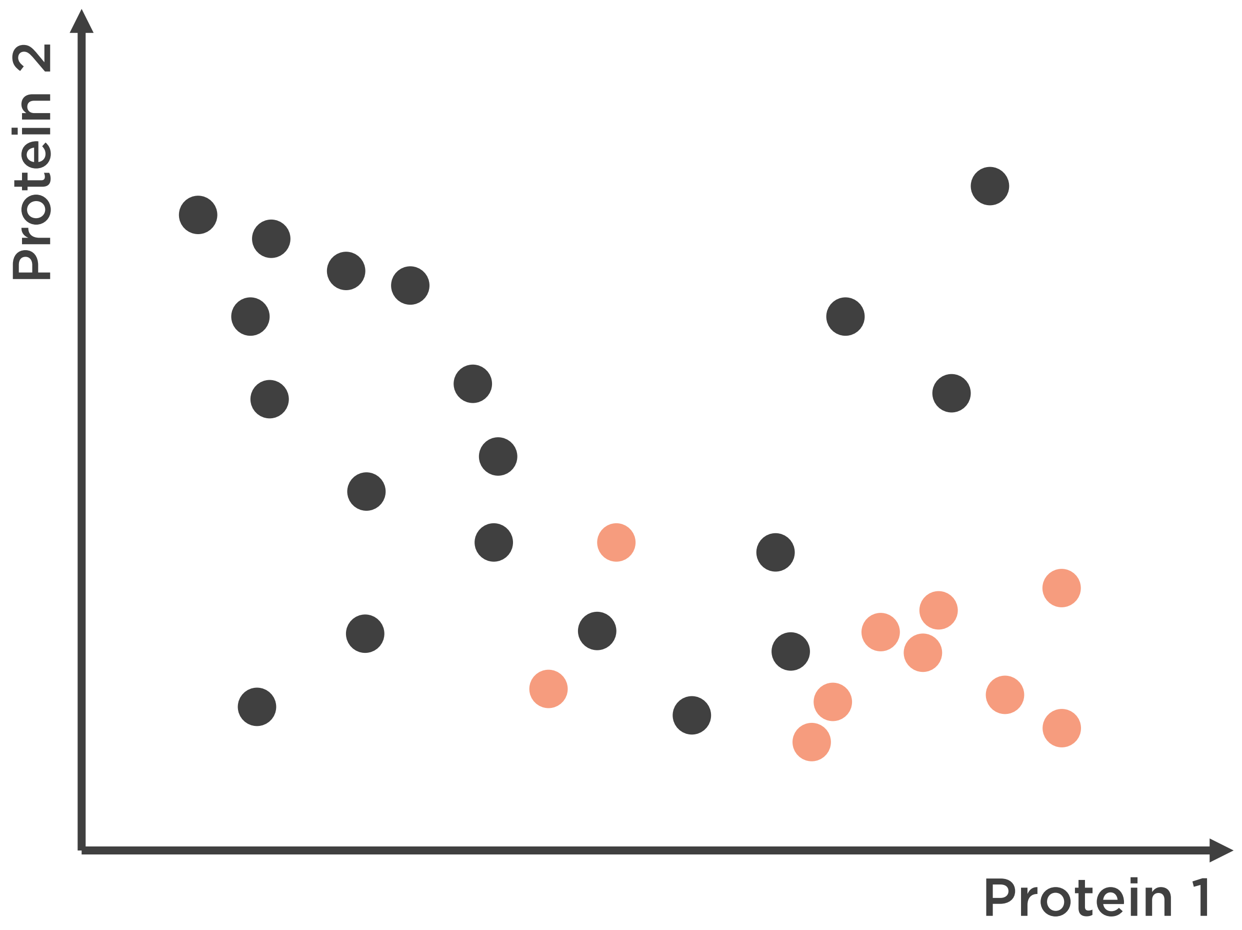


● Observation of interest

● Nearest neighbor

● Synthetic data points

Synthetic Minority Oversampling Technique



● Normal condition ● Patient with a rare disease

Demo

Use SMOTE to increase minority samples in Census dataset

Data Scale Issues in Distance-based Models

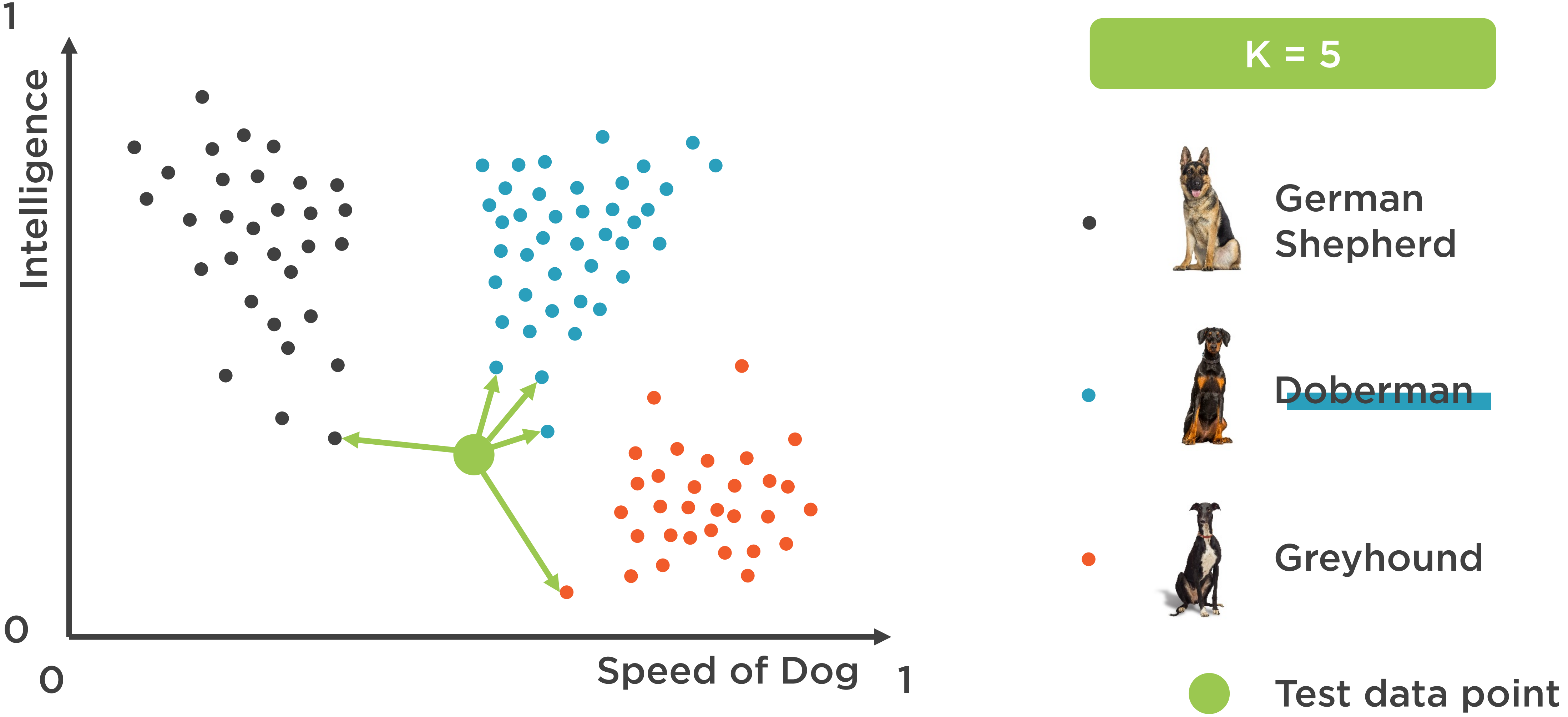
Data Scale Issues in Distance-based Models

Greyhound

German Shepherd

Doberman

Data Scale Issues in Distance-based Models



K = 5

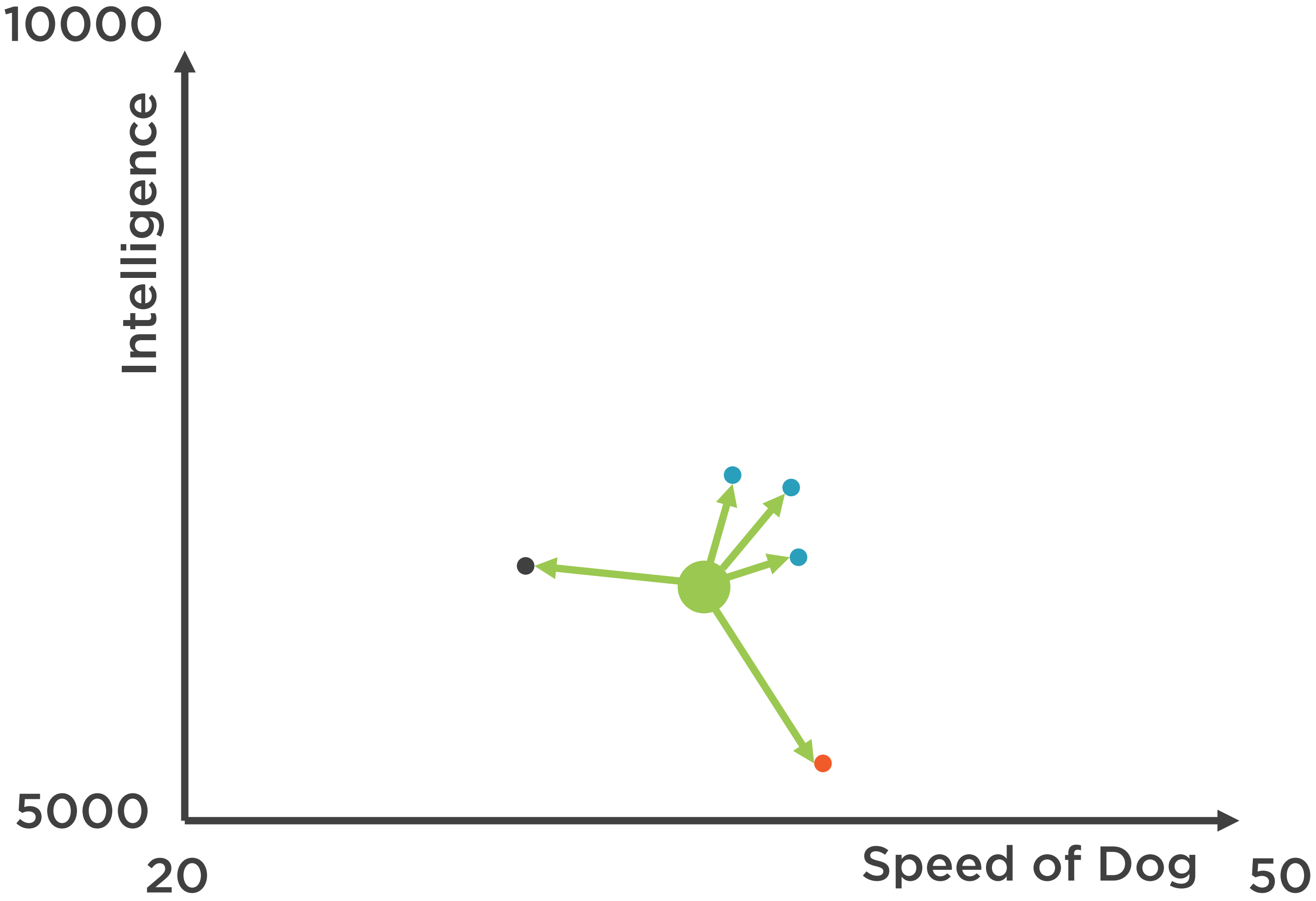
● German Shepherd

● Doberman

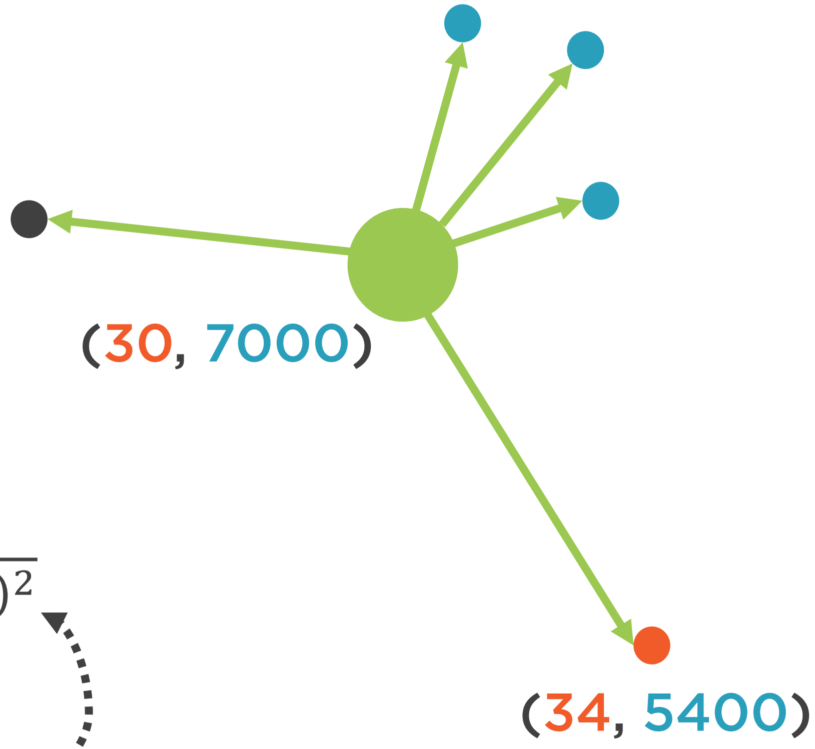
● Greyhound

● Test data point

Data Scale Issues in Distance-based Models



Data Scale Issues in Distance-based Models



$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

$$\sqrt{(30 - 34)^2 + (y_1 - y_2)^2}$$

$$\sqrt{(30 - 34)^2 + (7000 - 5400)^2}$$

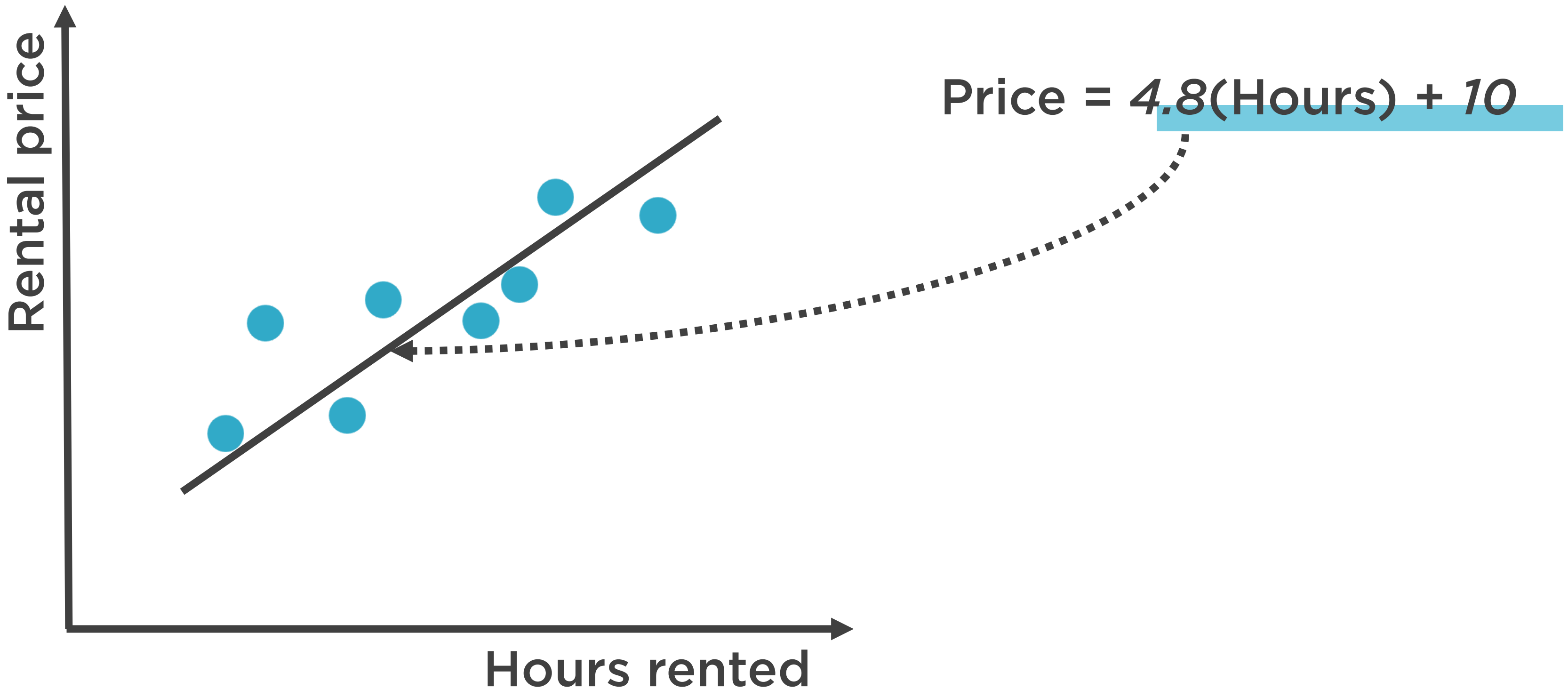
Intelligence dominates distance calculation

Multicollinearity in Multiple Regression

Multicollinearity

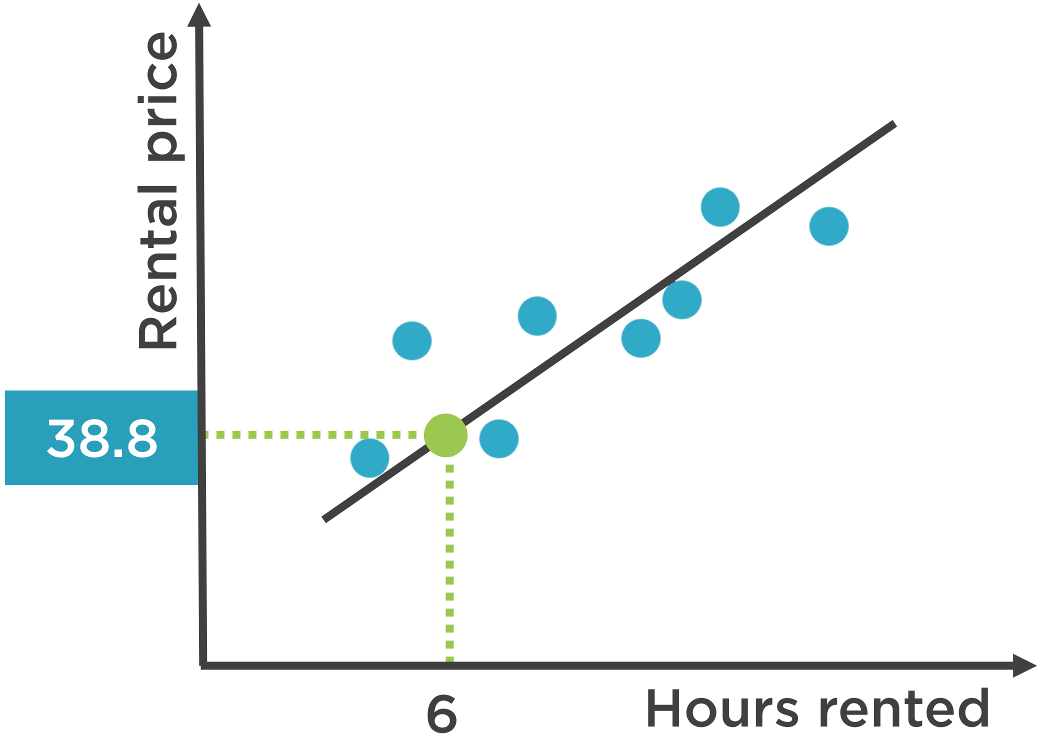
When one predictor variable in multiple regression can be linearly predicted from the others with a substantial degree of accuracy

Multicollinearity in Multiple Regression



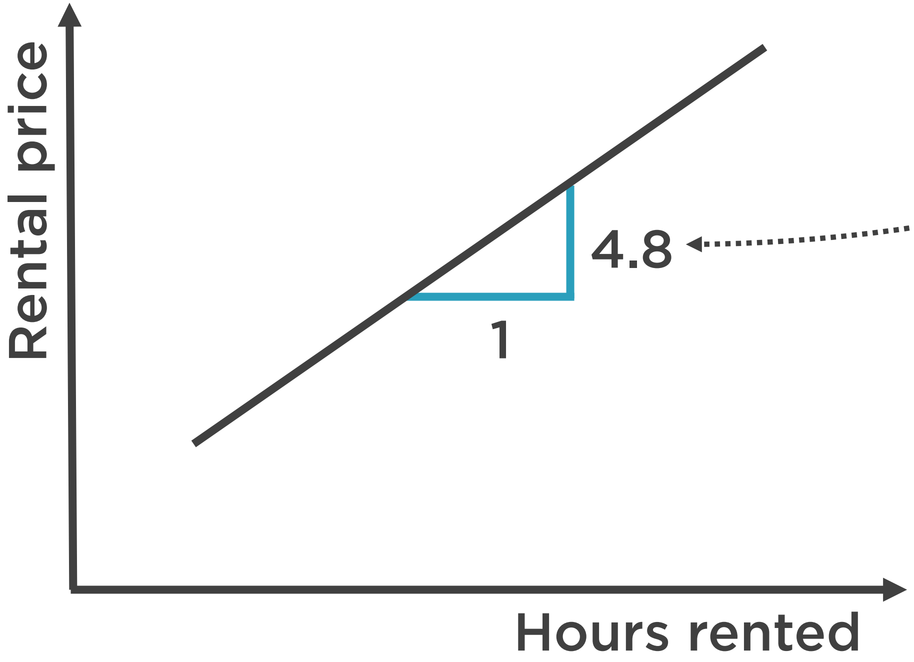
Price = *4.8*(Hours) + *10*

Multicollinearity in Multiple Regression



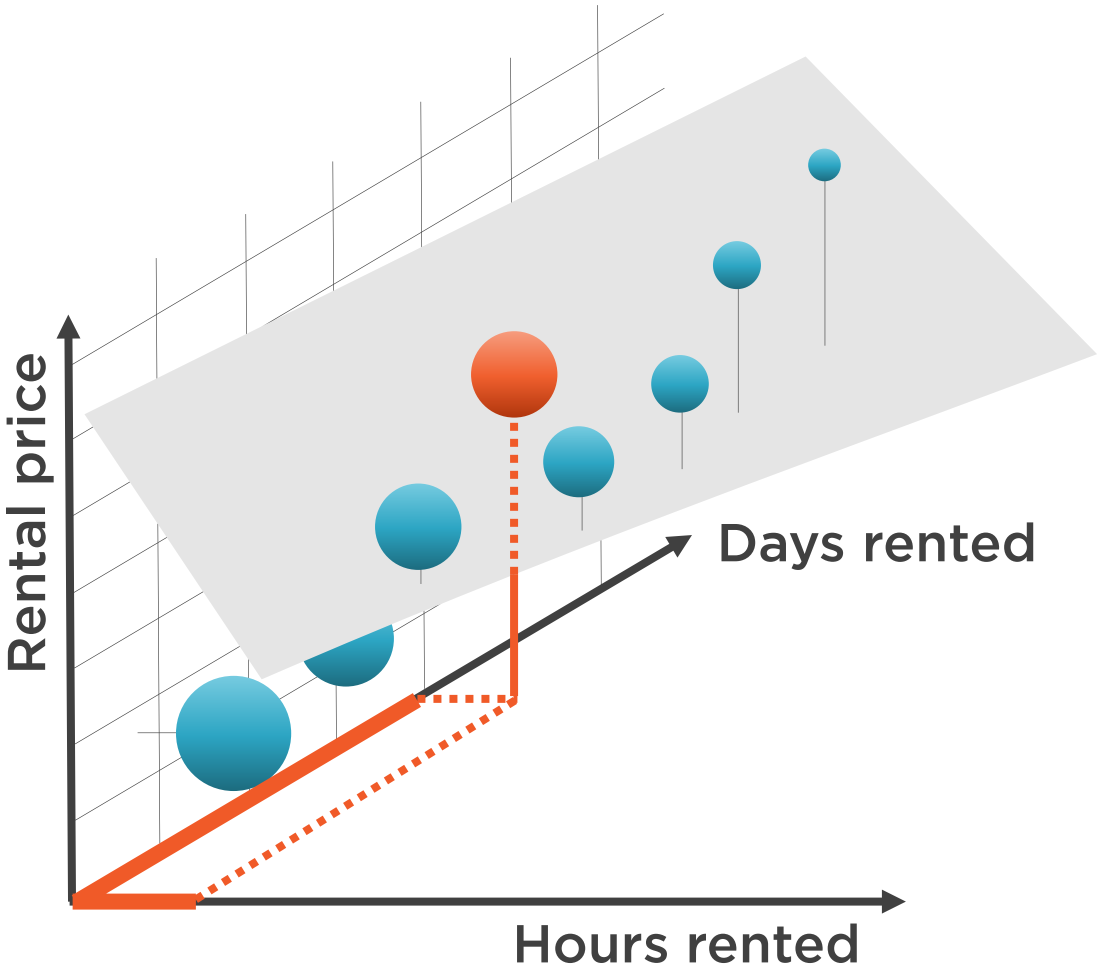
$\boxed{38.8} = 4.8(6) + 10$

Multicollinearity in Multiple Regression



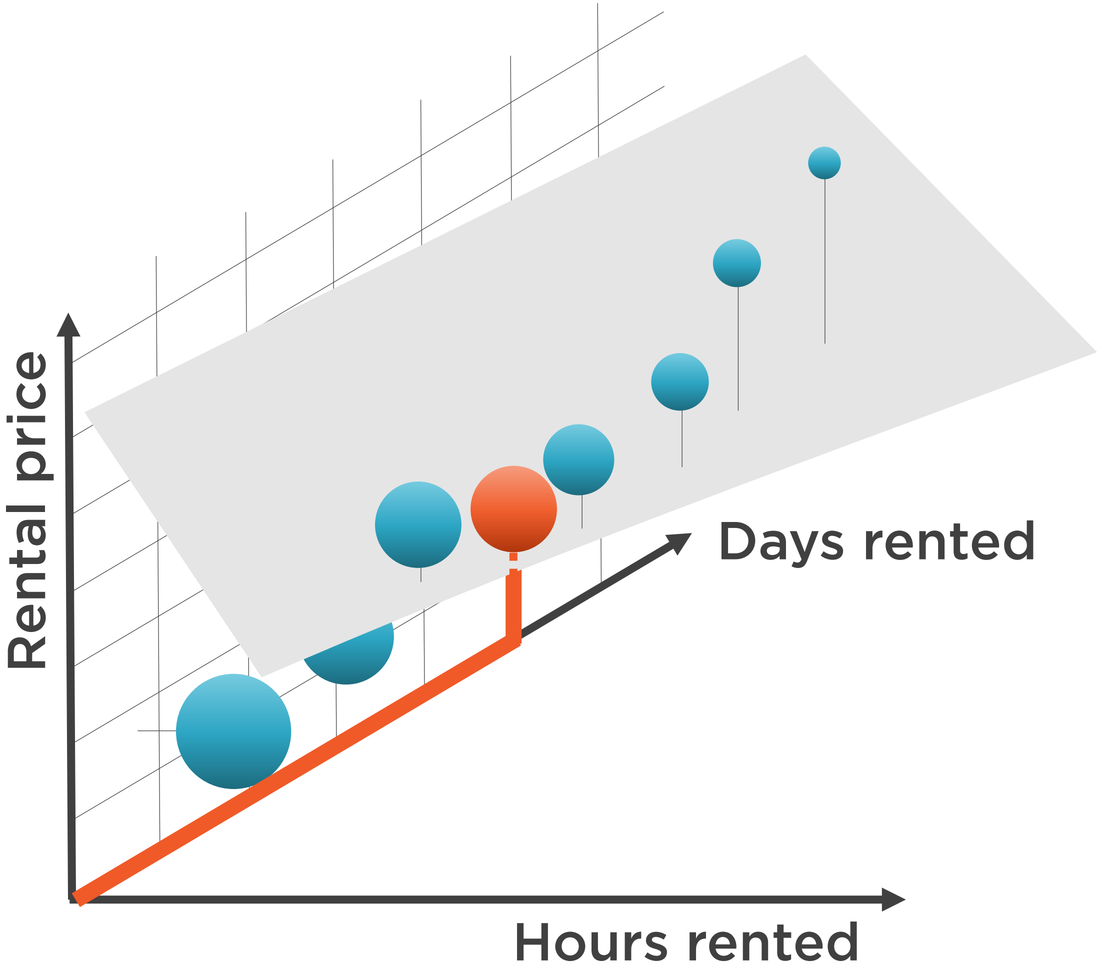
Price = *4.8*(Hours) + *10*

Multicollinearity in Multiple Regression



Price = *.12*(Hours) + *100*(days) + C

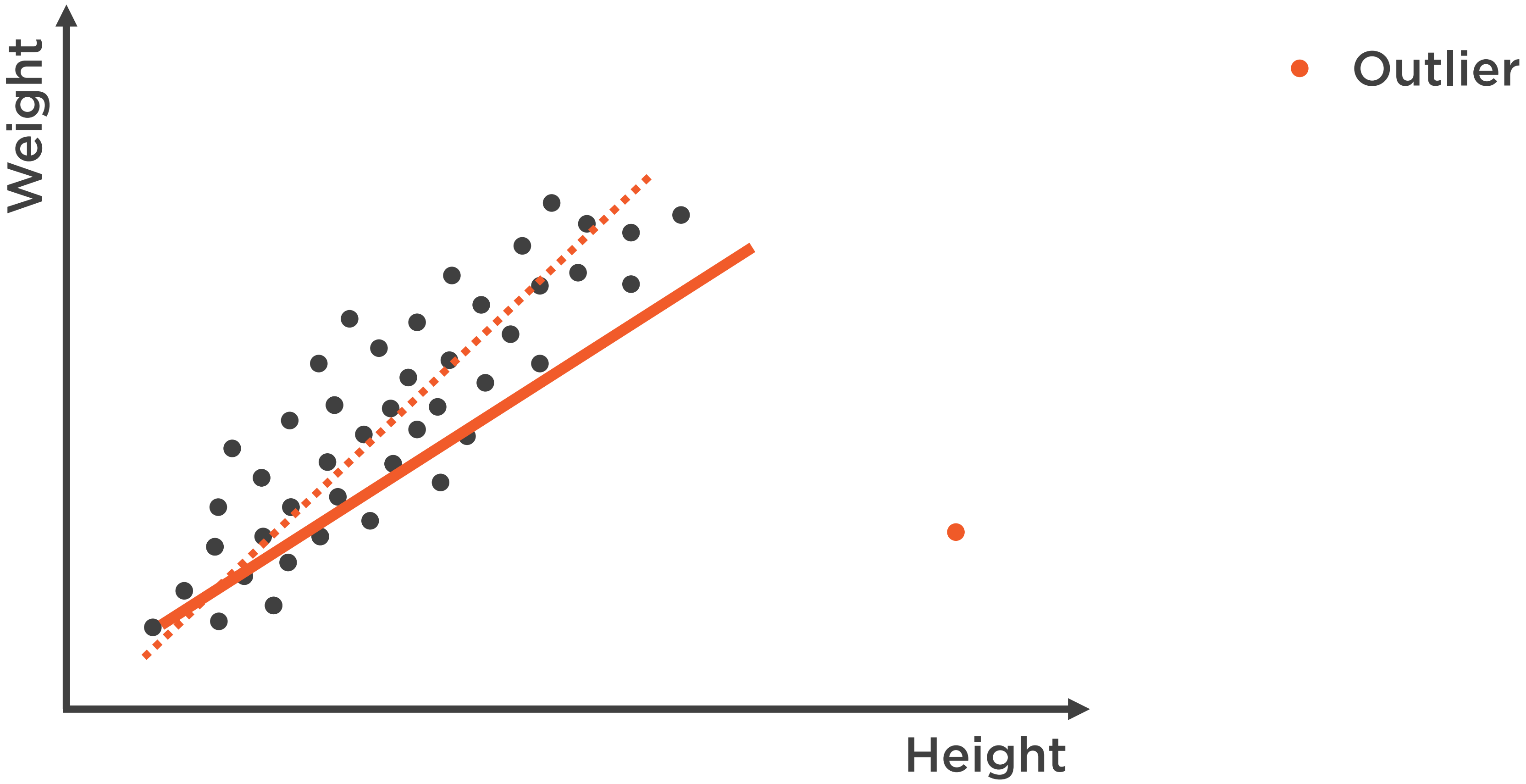
Multicollinearity in Multiple Regression



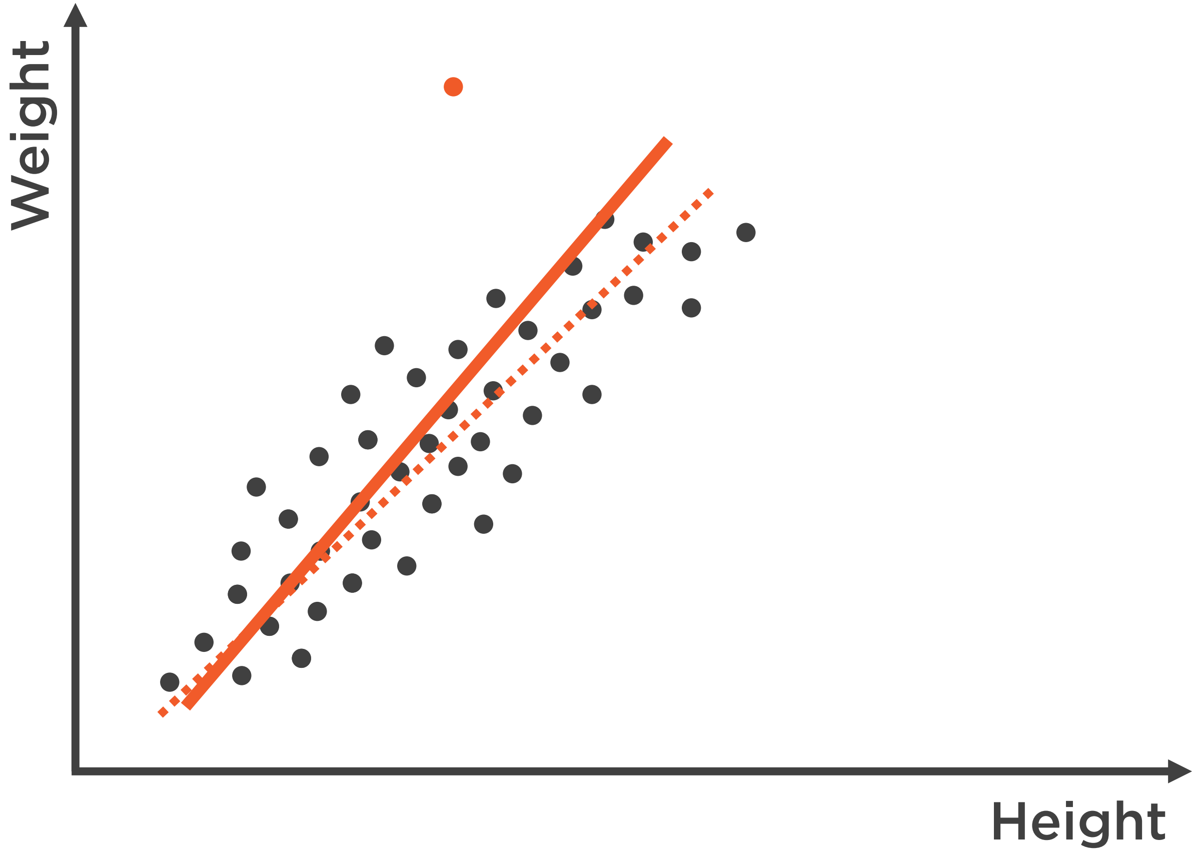
Price = *.12*(Hours) + *100*(days) + C

Outliers in Regression Models

Outliers in Regression Models

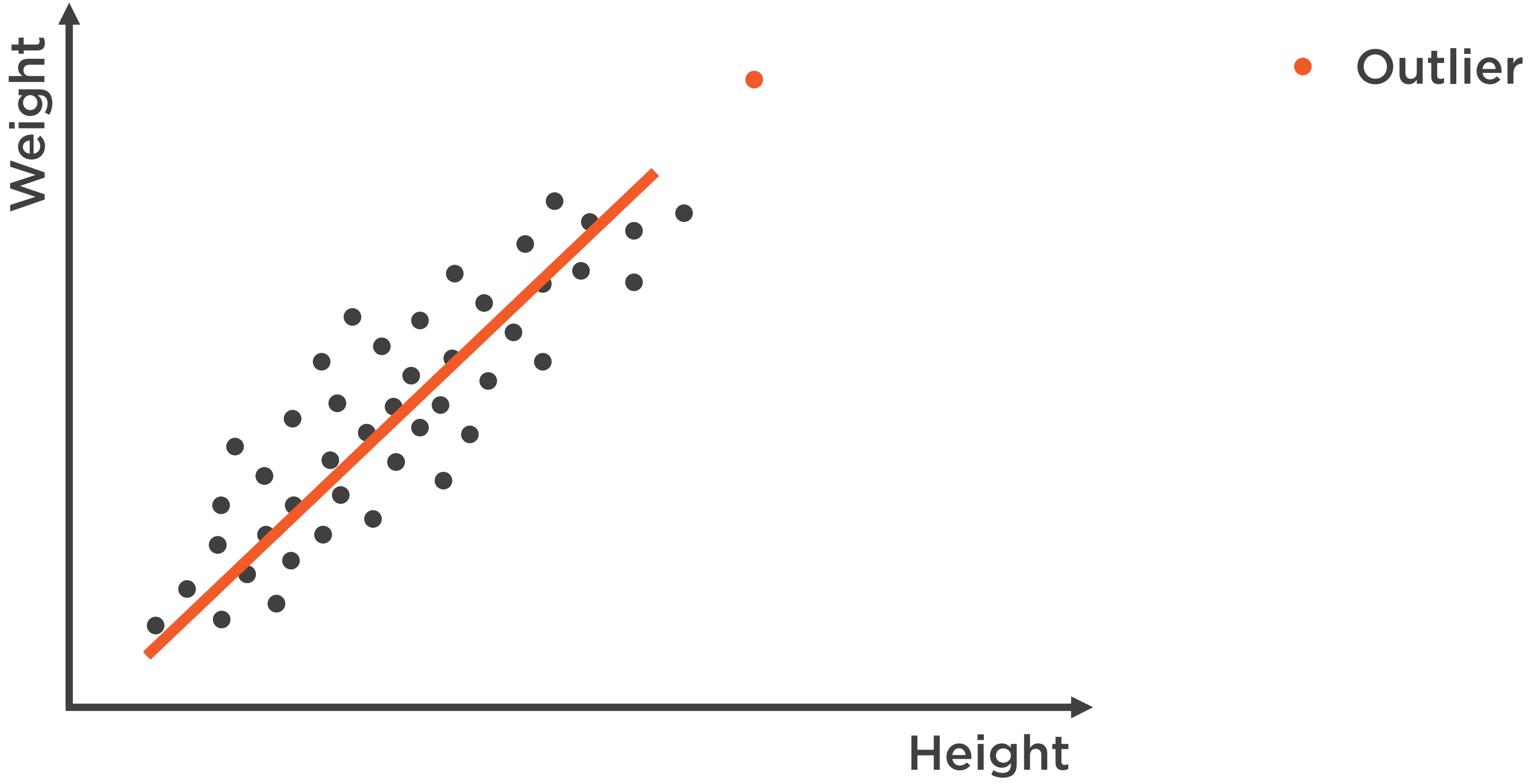


Outliers in Regression Models

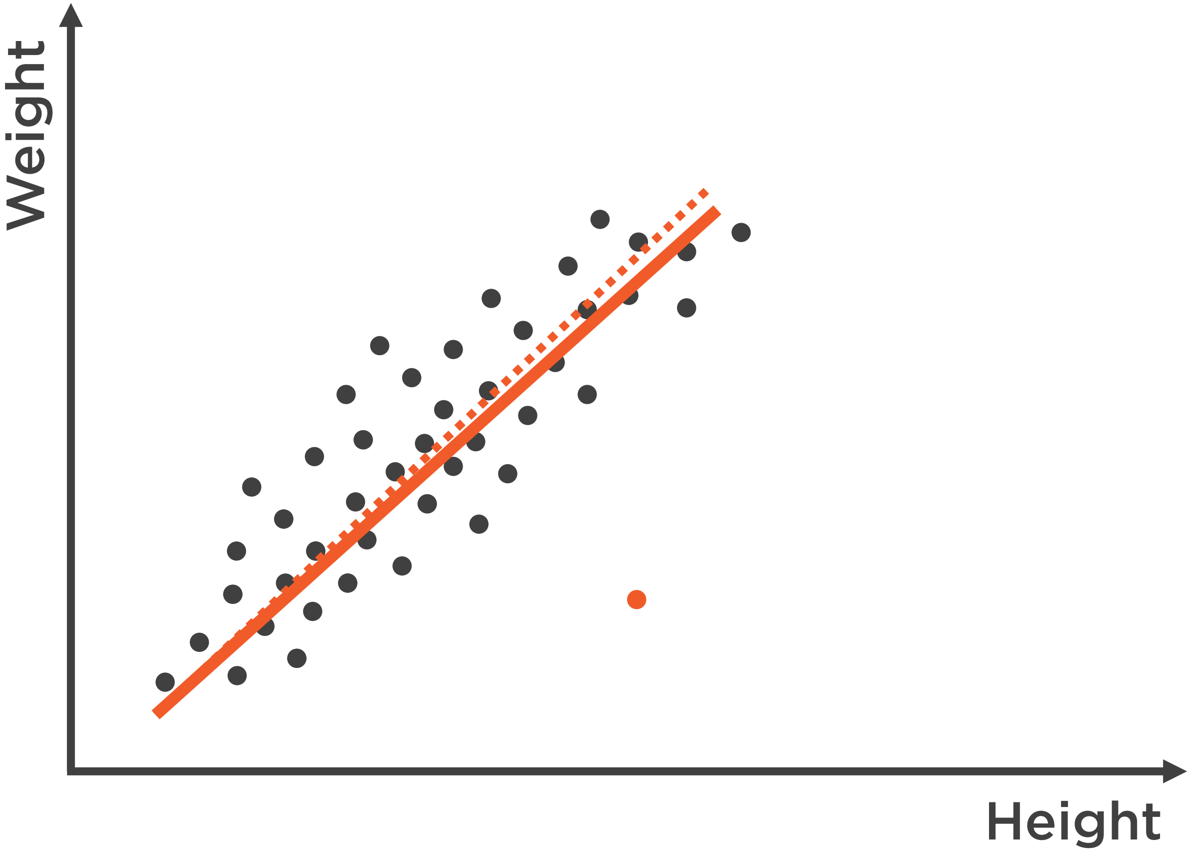


● Outlier

Outliers in Regression Models

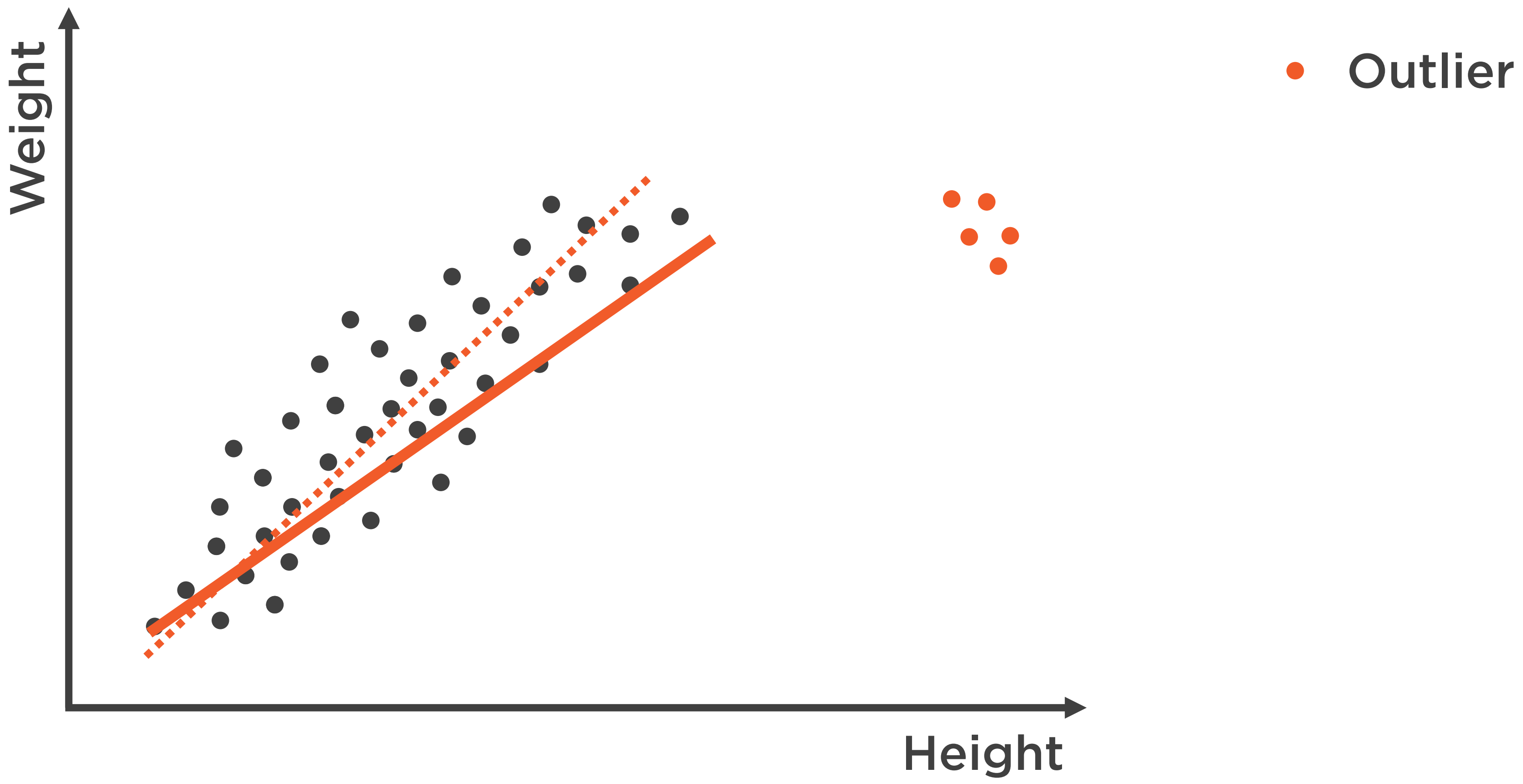


Outliers in Regression Models

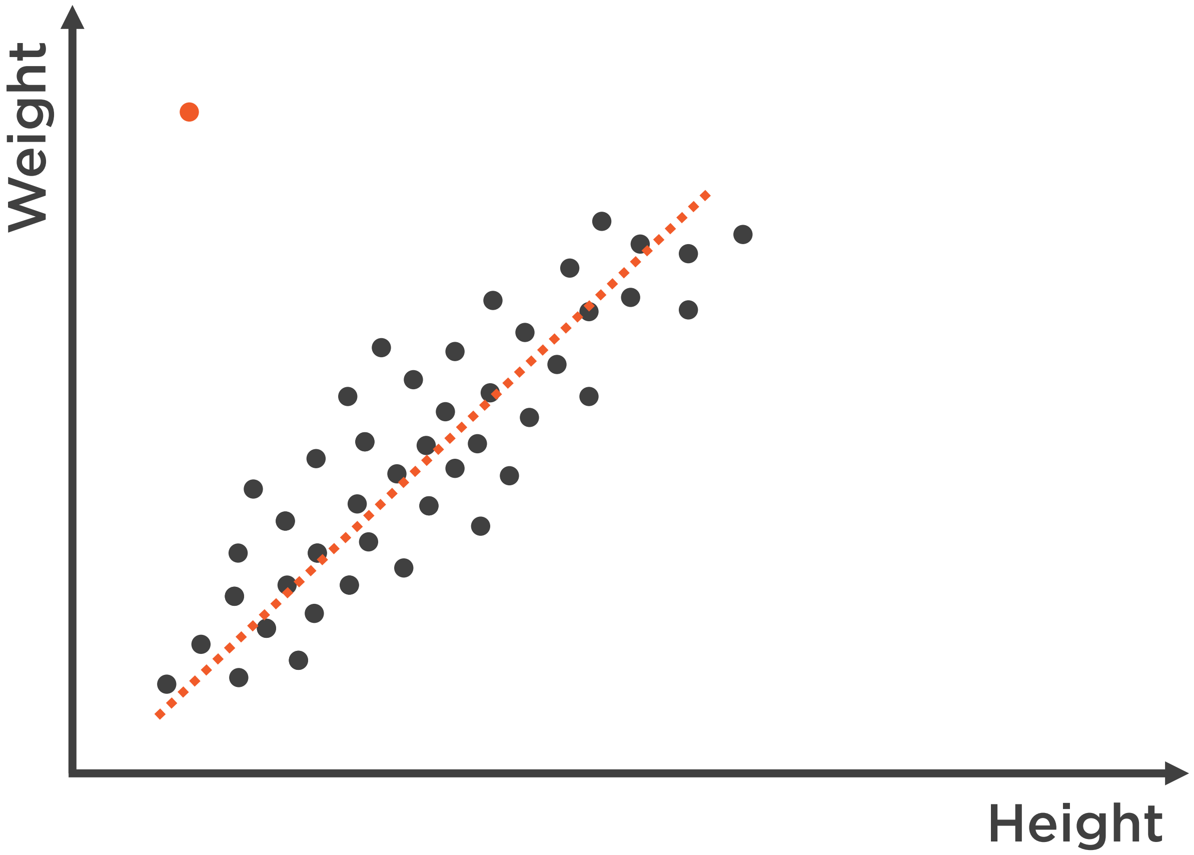


● Outlier

Outliers in Regression Models

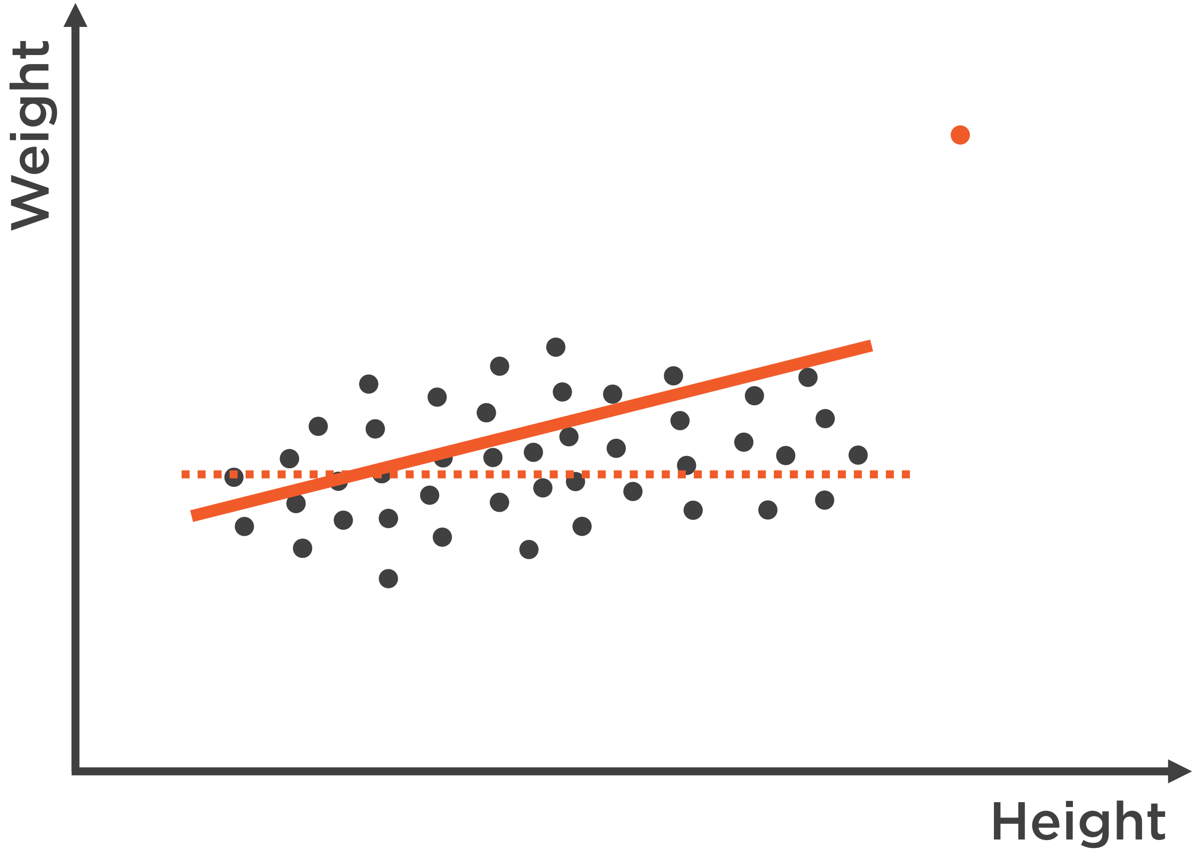


Outliers in Regression Models



● **Outlier**

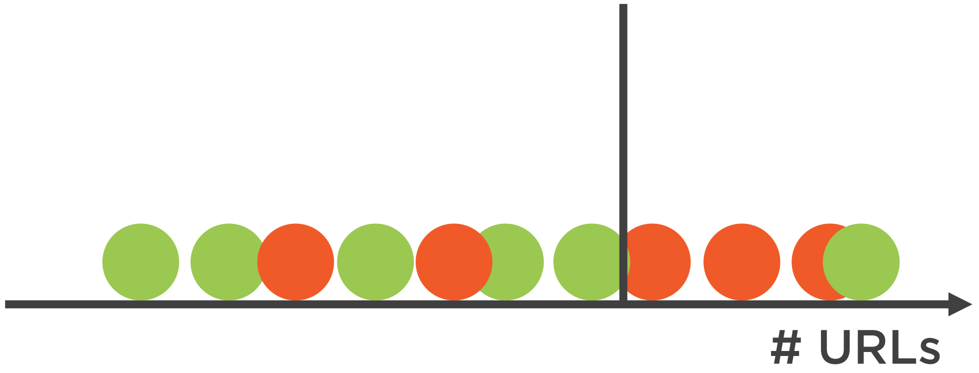
Outliers in Regression Models

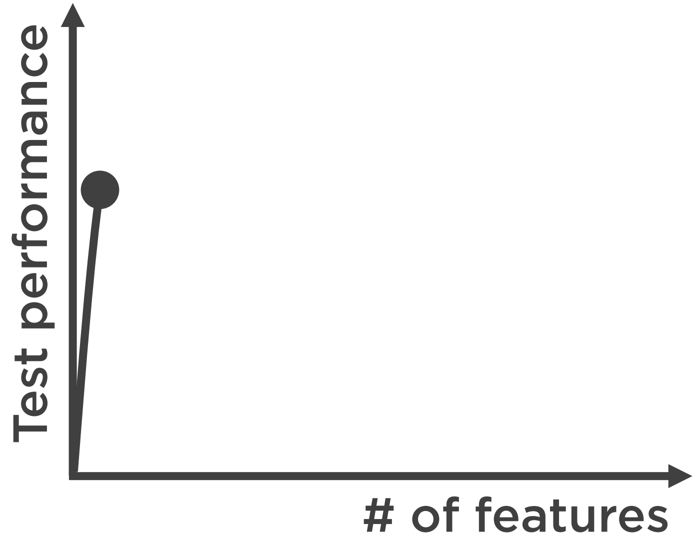


● Outlier

Problem with High-dimensional Datasets

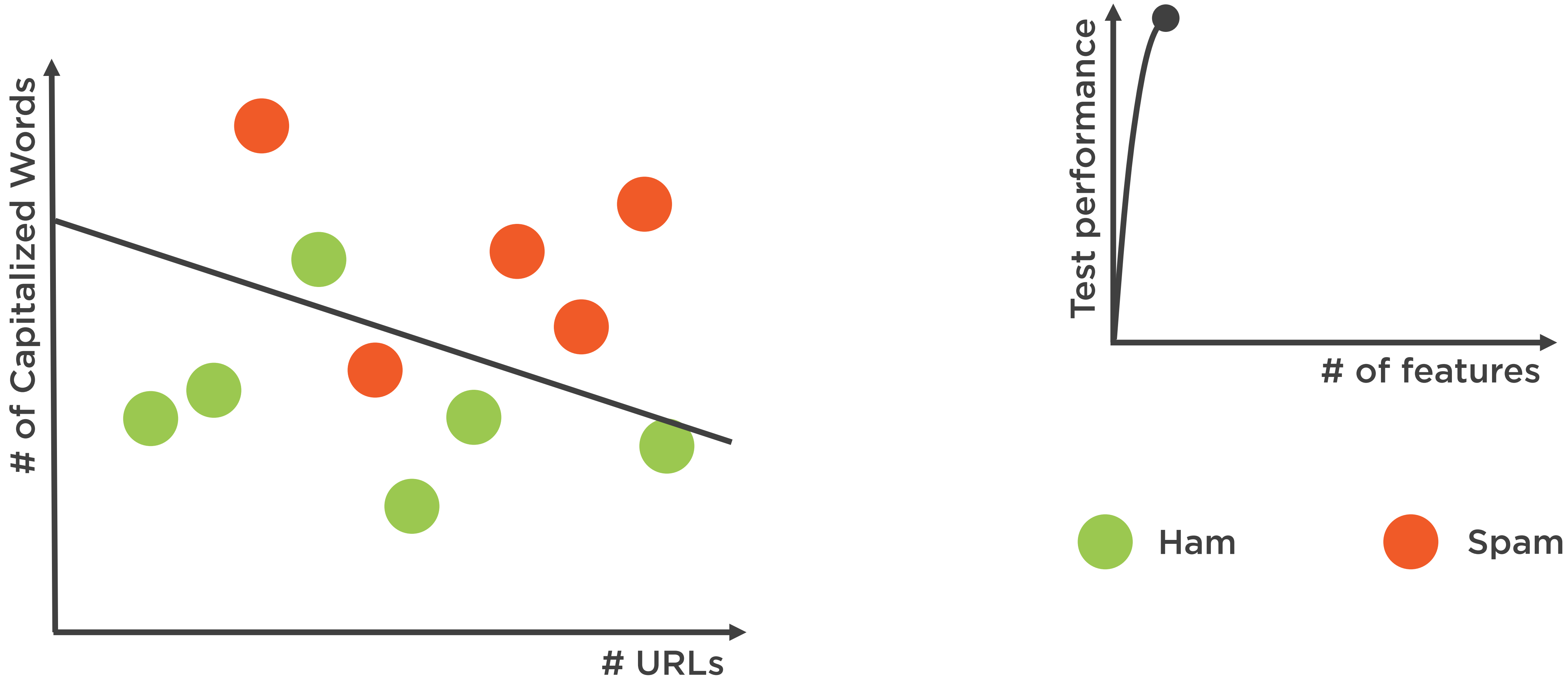
Problem with High-dimensional Datasets



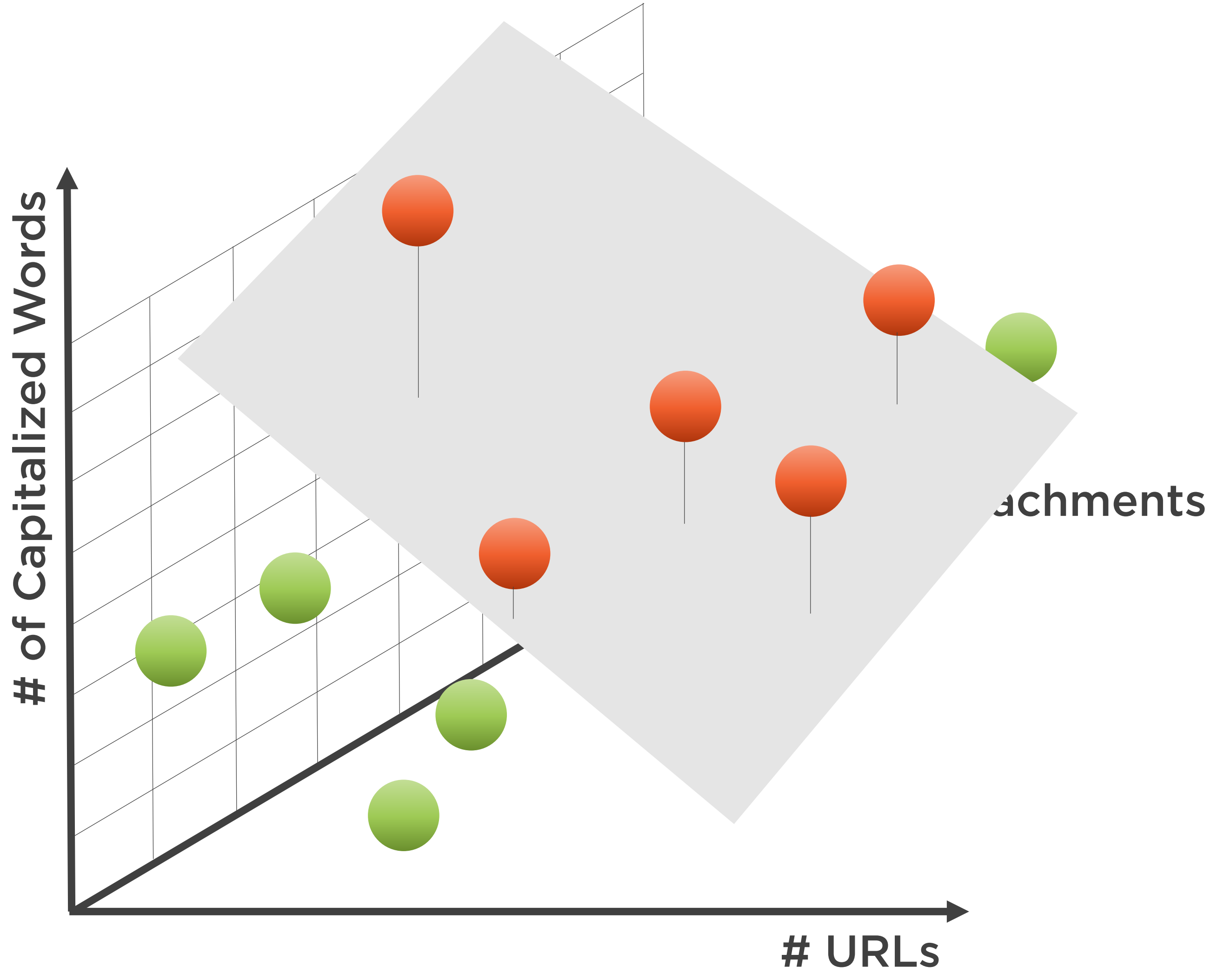


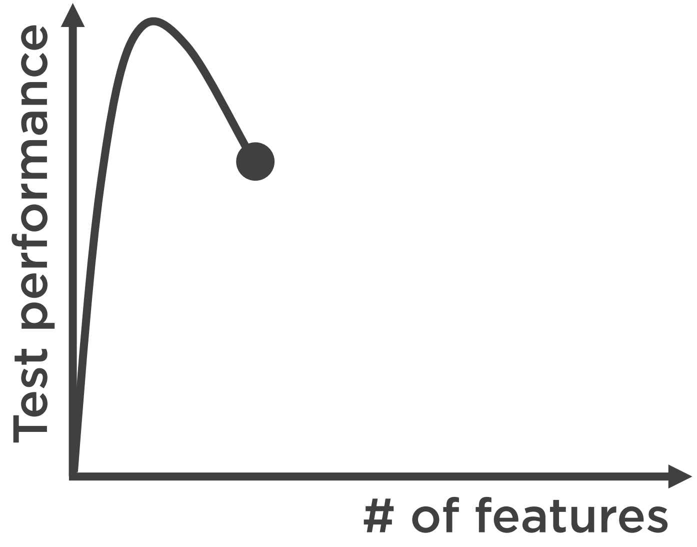
🟢 Ham 🔴 Spam

Problem with High-dimensional Datasets



Problem with High-dimensional Datasets

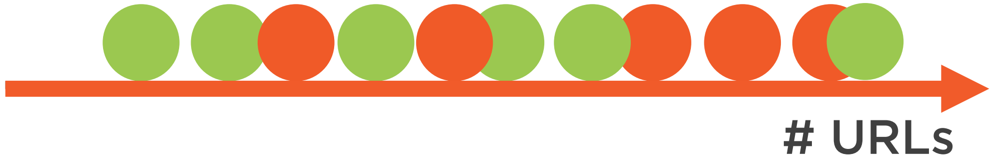




● Ham

● Spam

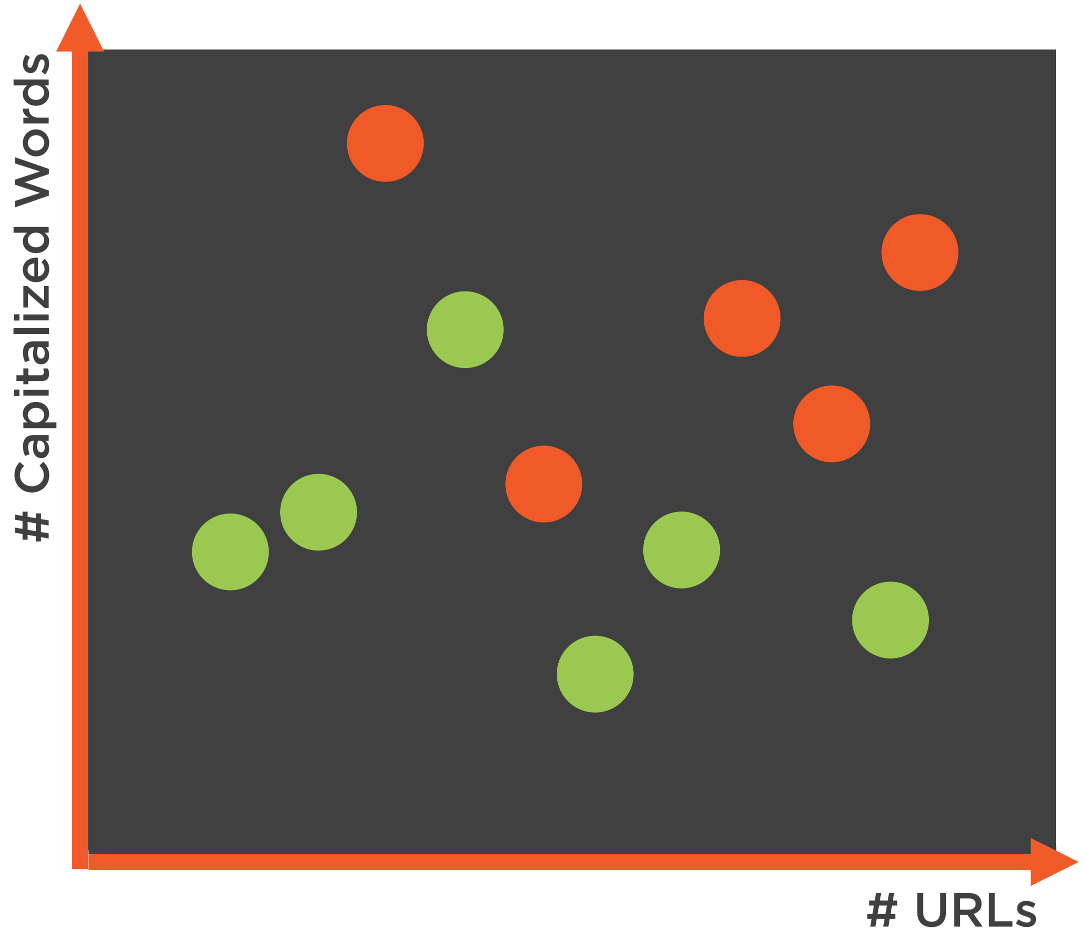
Problem with High-dimensional Datasets



Data density:

10/10 = 1

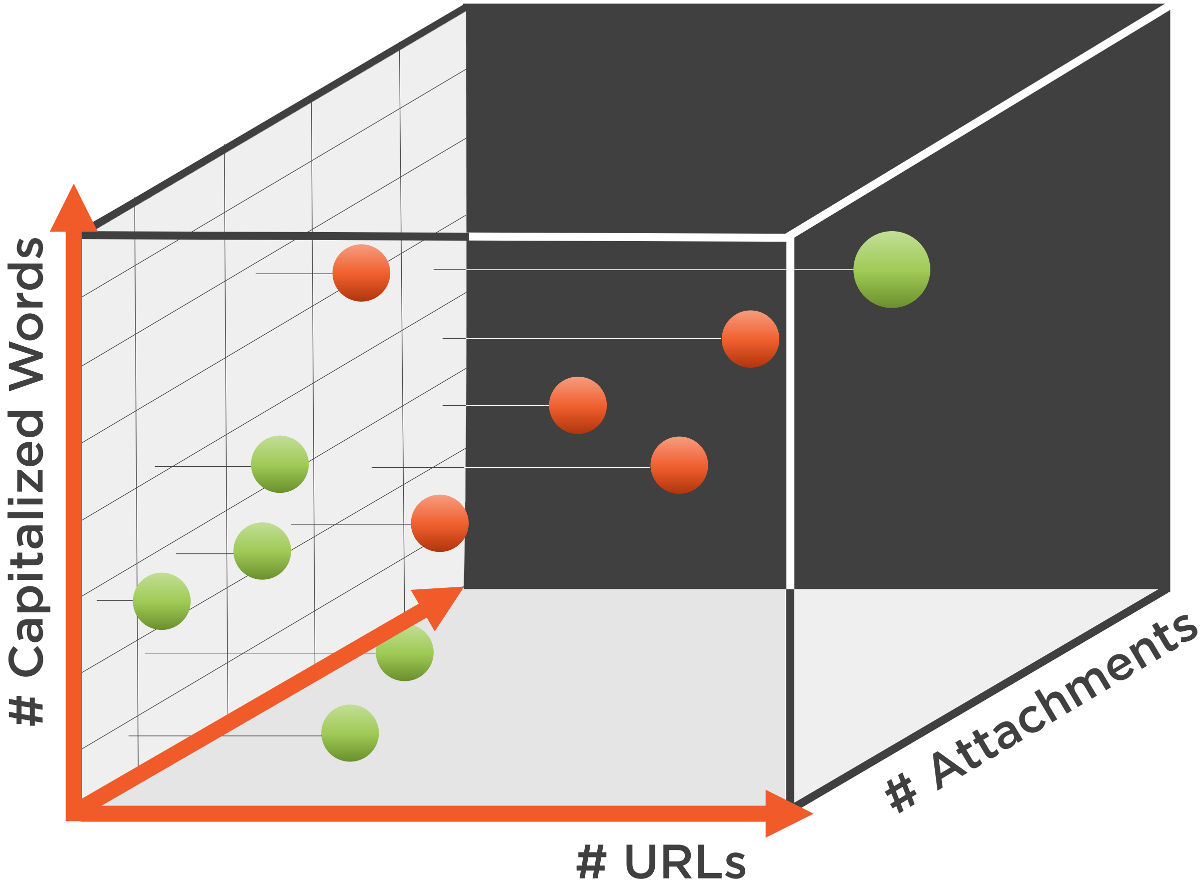
Problem with High-dimensional Datasets



Data density:

10/100 = 0.1

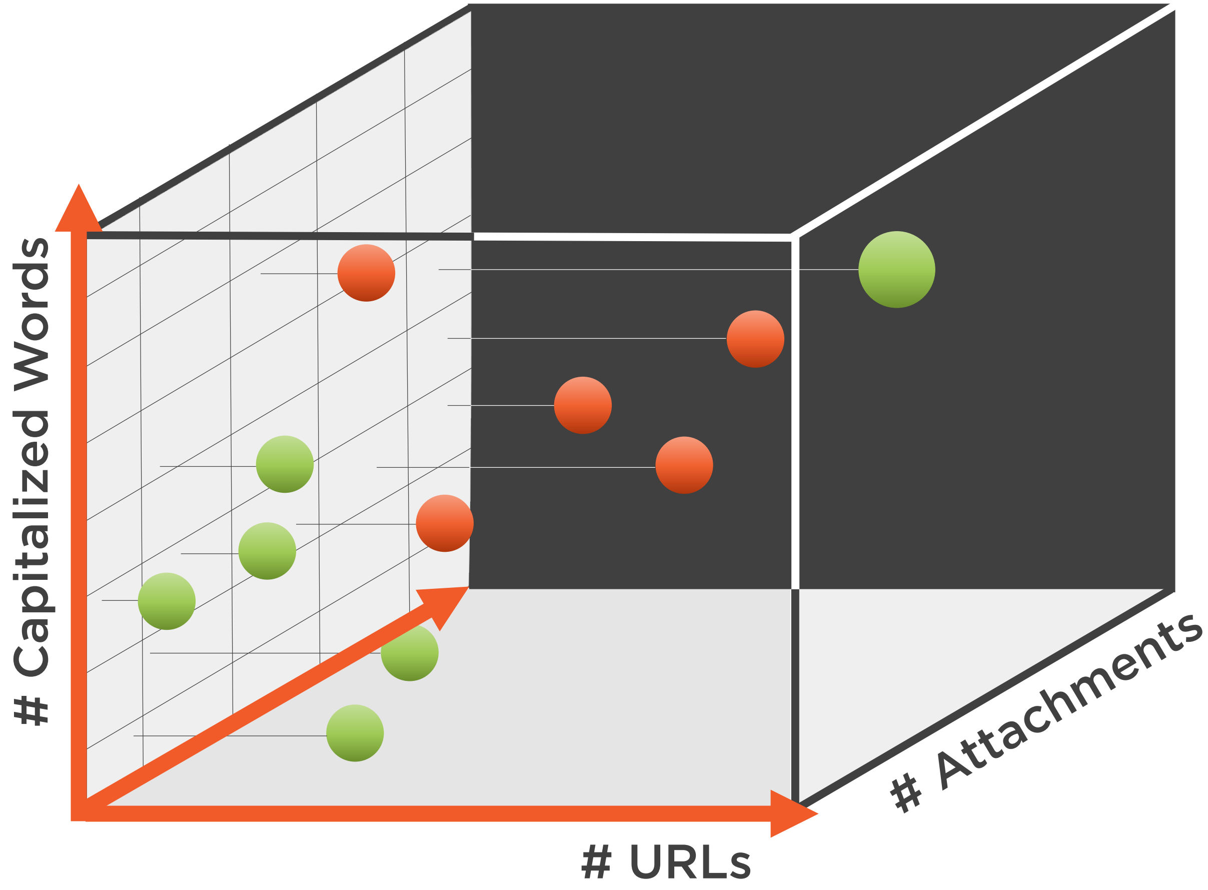
Problem with High-dimensional Datasets



Data density:

10/1000 = 0.01

Problem with High-dimensional Datasets



- Increase the number of observations
- Remove unnecessary features
- Use PCA

Summary

Data-level issues indicate the importance of data transformation

Play with SMOTE to improve model performance

PCA can help solve a variety of issues