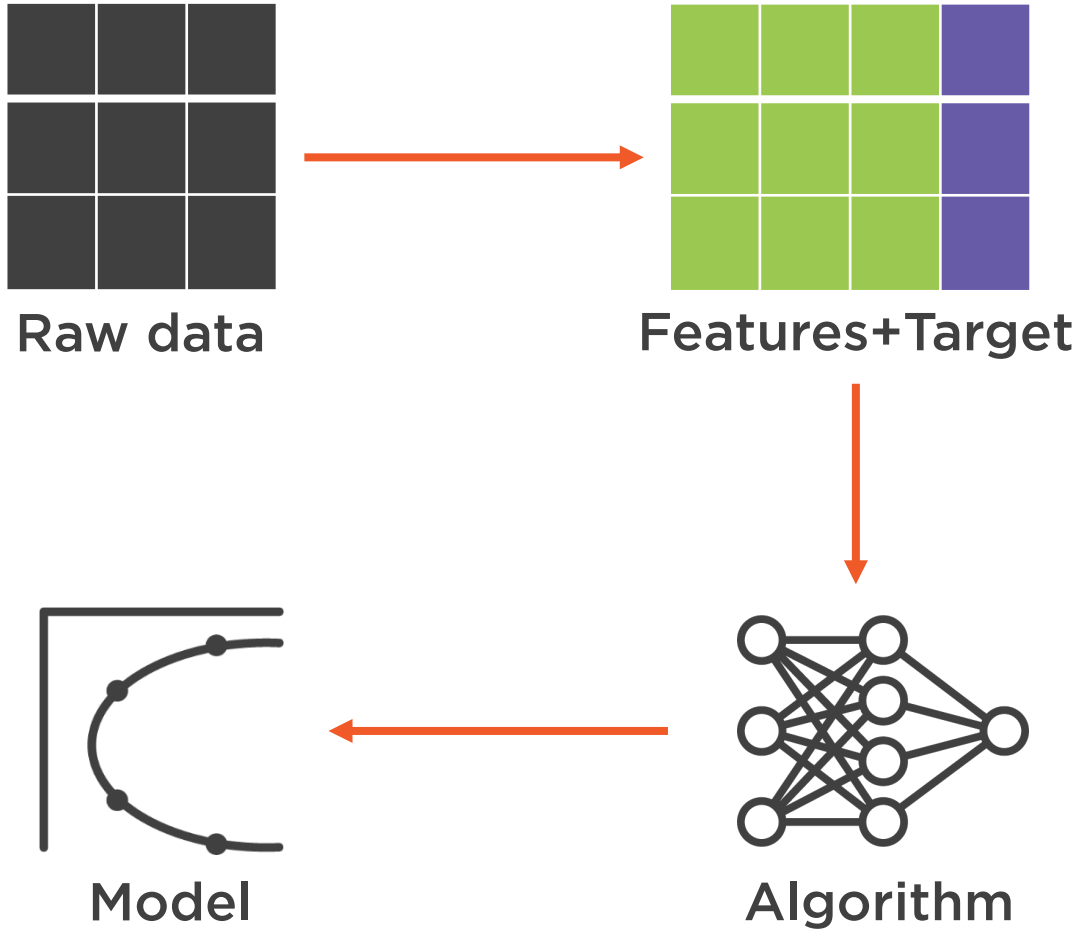
Differentiating Data, Features, Targets & Models

Ravikiran Srinivasulu
SOFTWARE CONSULTANT

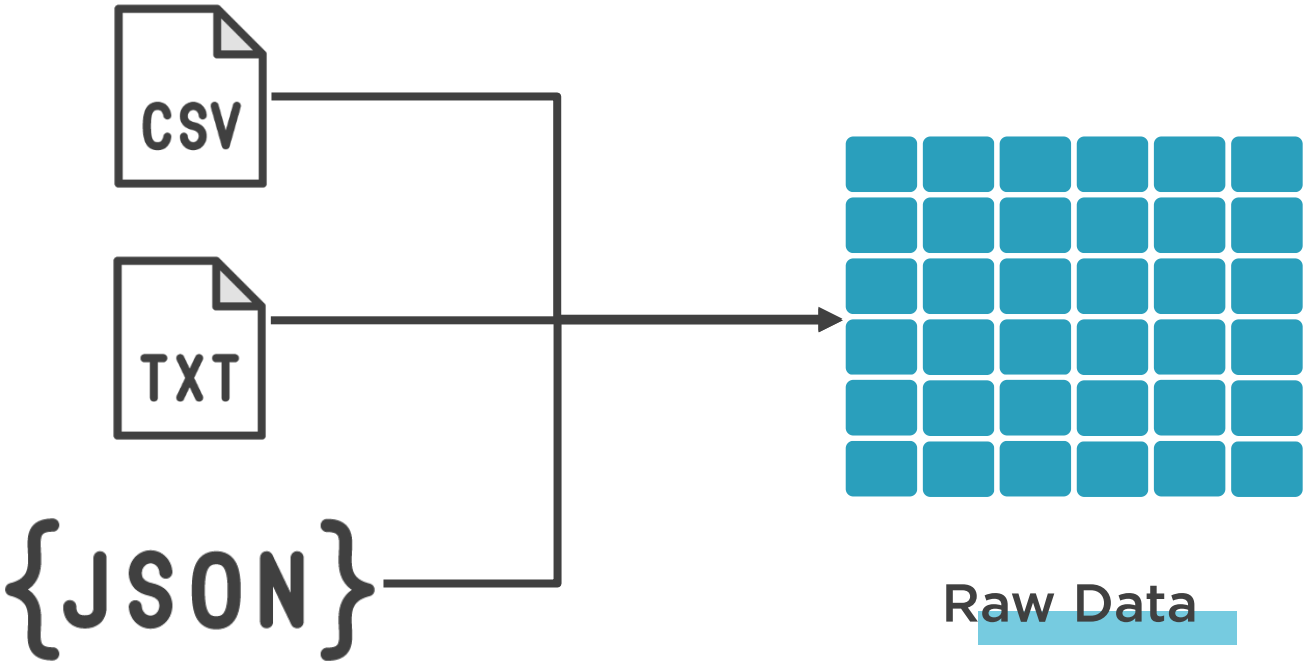
ravikirans.com | go.ravikirans.com/YouTube

Agenda



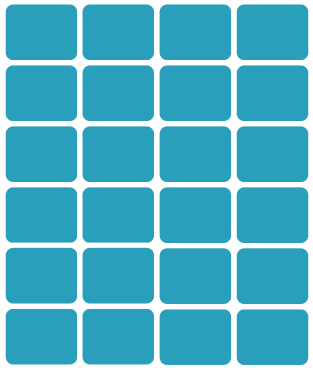
Moving from Raw Data to Features

Moving from Raw Data to Features



Customer Id	Customer Name	Last Order Date
14097	Anna	12-10-2019

Moving from Raw Data to Features

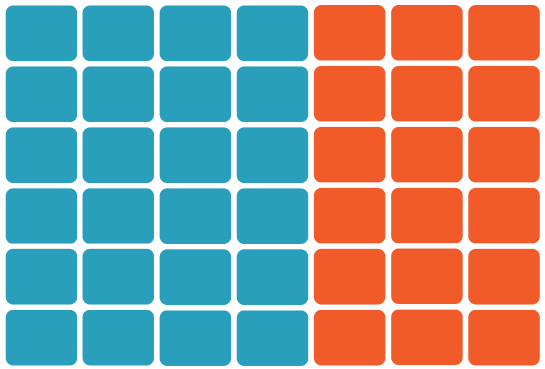


Raw Data

Customer Id	Customer Name	Last Order Date
~~14097~~	~~Anna~~	12-10-2019

Moving from Raw Data to Features

- Learning with Counts
- Binning

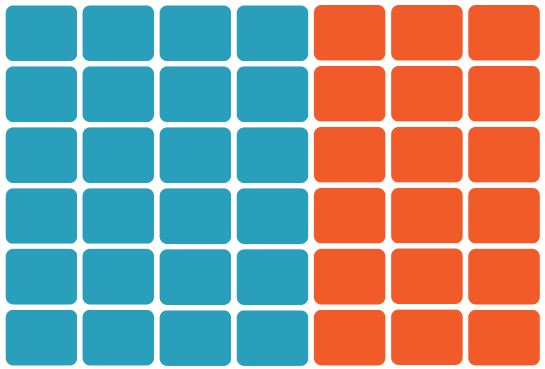


Feature Engineering

Customer Id	Customer Name	Last Order Date
~~14097~~	~~Anna~~	12-10-2019

Moving from Raw Data to Features

- Learning with Counts
- Binning



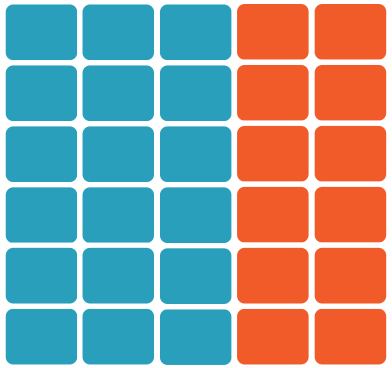
Feature Engineering

Customer Id	Customer Name	Last Order Date	Days Since the Last Order
~~14097~~	~~Anna~~	~~12-10-2019~~	5

Moving from Raw Data to Features

- **Filter Based Feature Selection**
- **Fisher LDA**
- **...**

Moving from Raw Data to Features



Feature Matrix

$$\begin{bmatrix} X_1, X_3, X_6, X_7, X_9 \end{bmatrix}$$

6 Characteristics of a Good Feature

6 Characteristics of a Good Feature

Features Must Be Related to the Problem

Features Must Be Related to the Problem

Age	Cholesterol	Sugar	Family History	Marital Status	Heart Disease?
33	200	125	0	0	1
54	199	115	1	1	0
45	162	127	1	1	0
60	198	129	0	1	1
38	212	132	0	0	0
44	198	130	1	1	1
72	240	140	0	0	0

6 Characteristics of a Good Feature

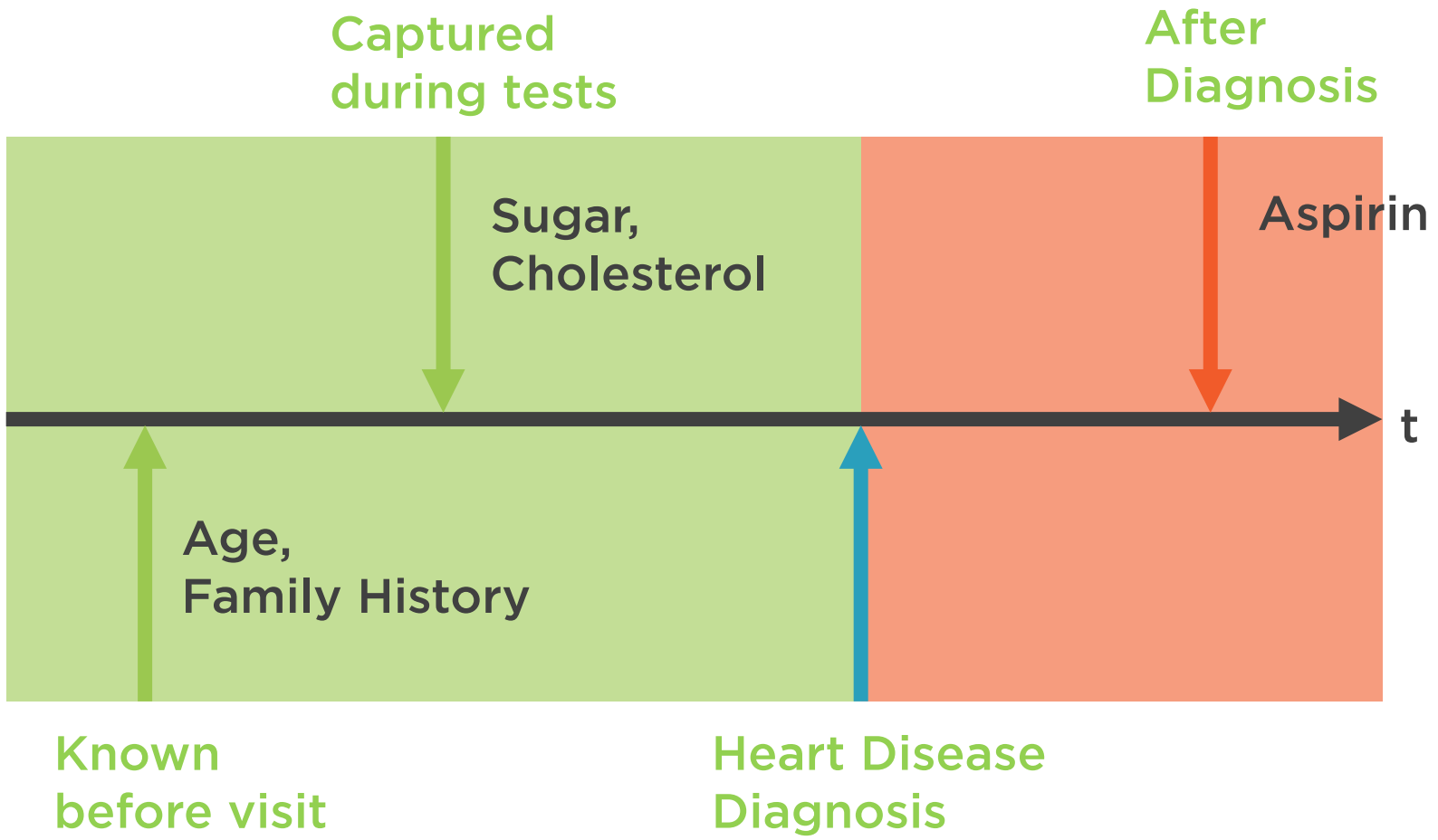
Features Must Be Related to the Problem

Features Values Must Be Known At Prediction Time

Features Must Be Known at Prediction Time

Age	Cholesterol	Sugar	Family History	Aspirin Consumption	Heart Disease?
33	200	125	0	1	1
54	199	115	1	0	0
45	162	127	1	0	0
60	198	129	0	1	1
38	212	132	0	0	0
44	198	130	1	1	1
72	240	140	0	0	0

Features Must Be Known at Prediction Time



Features Must Be Known at Prediction Time

Age	Cholesterol	Sugar	Family History	Aspirin Consumption	Heart Disease?
33	200	125	0	1	1
54	199	115	1	0	0
45	162	127	1	0	0
60	198	129	0	1	1
38	212	132	0	0	0
44	198	130	1	1	1
72	240	140	0	0	0

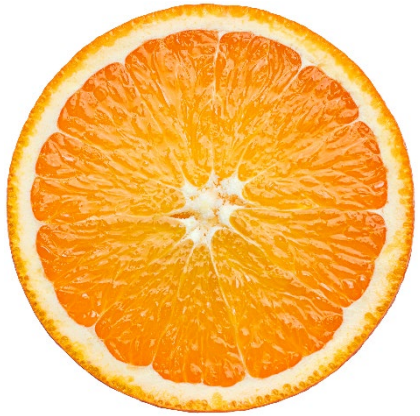
6 Characteristics of a Good Feature

Features Must Be Related to the Problem

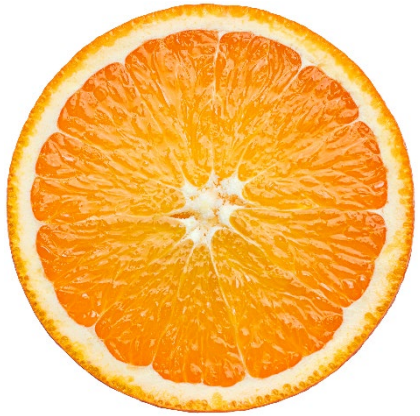
Features Must Be Known At Prediction Time

Feature Values Should Have Enough Variation

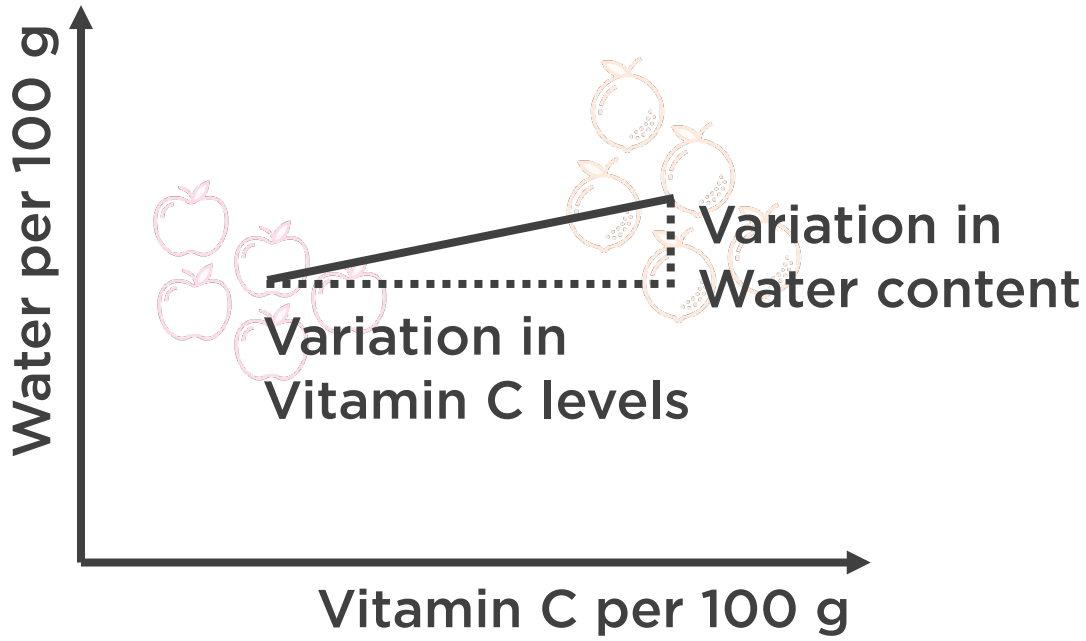
Features Values Should Have Enough Variation



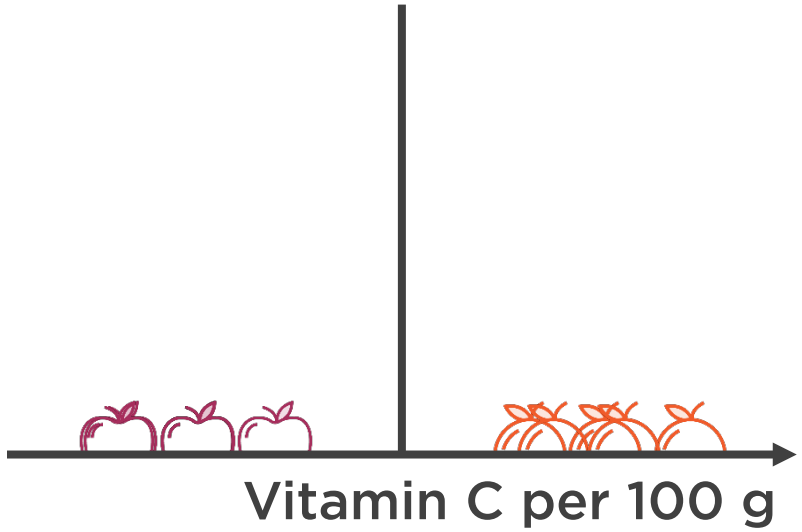
Features Values Should Have Enough Variation



Features Values Should Have Enough Variation



Features Values Should Have Enough Variation



6 Characteristics of a Good Feature

Features Must Be Related to the Problem

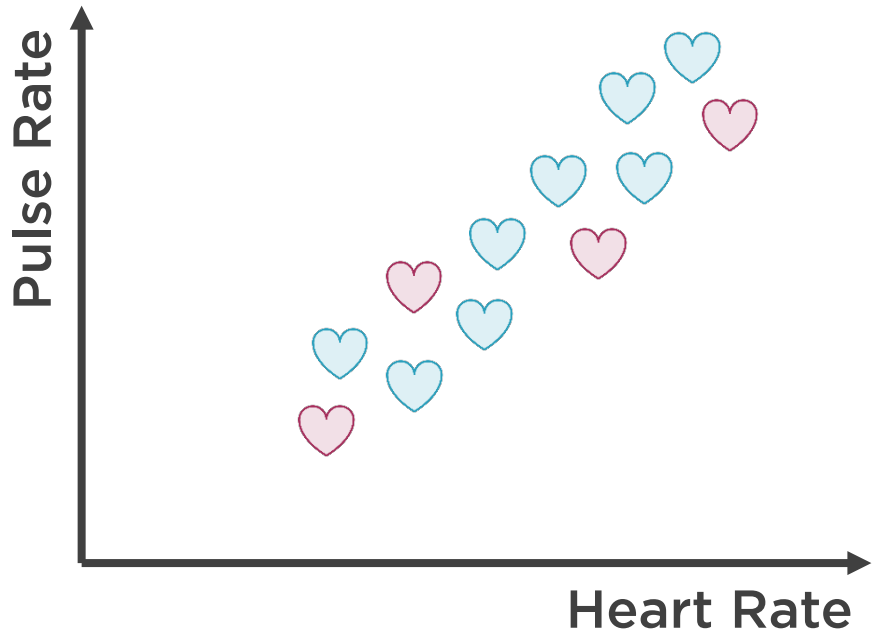
Features Must Be Known At Prediction Time

Feature Values Should Have Enough Variation

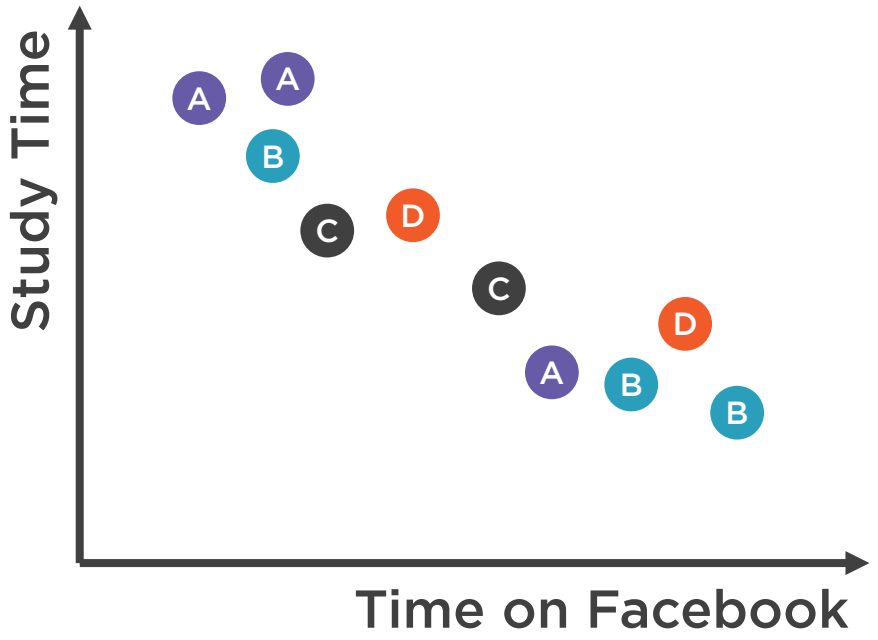
Features Should Not Be Highly Correlated

Features Should Not Be Highly Correlated

Predict Heart Disease



Predict Student Grades



6 Characteristics of a Good Feature

Features Must Be Related to the Problem

Features Must Be Known At Prediction Time

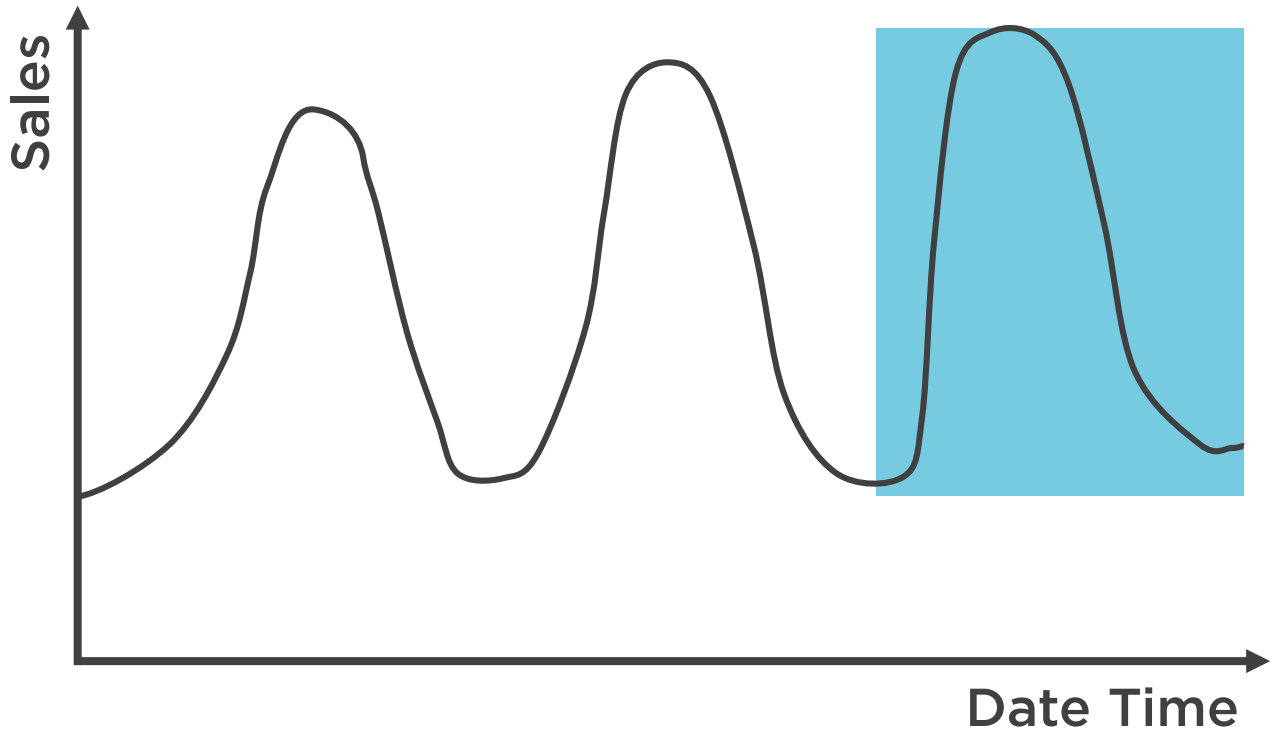
Feature Values Should Have Enough Variation

Features Should Not Be Highly Correlated

Features Should Be Simple

Features Should Be Simple

Date Time: 29-05-2019 11:44:12 -> Day of the Week



6 Characteristics of a Good Feature

Features Must Be Related to the Problem

Features Must Be Known At Prediction Time

Feature Values Should Have Enough Variation

Features Should Not Be Highly Correlated

Features Should Be Simple

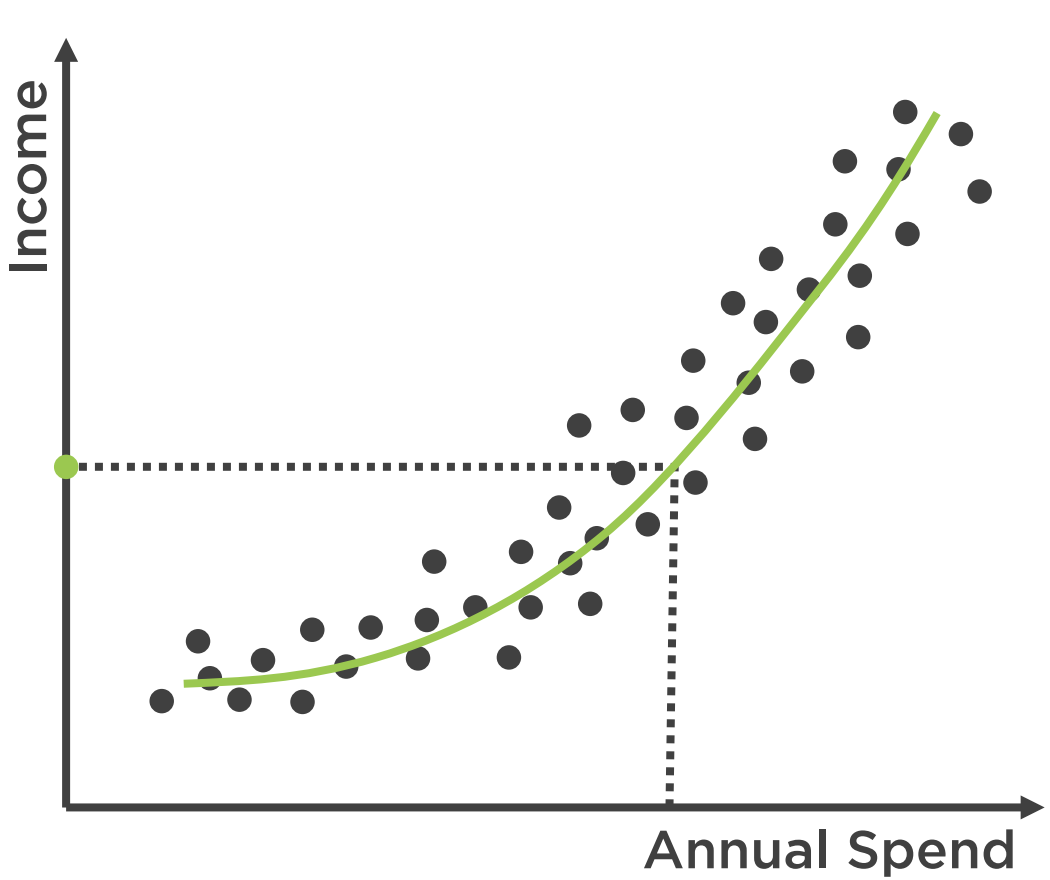
Features Should Have Enough Examples

Features Should Have Enough Examples

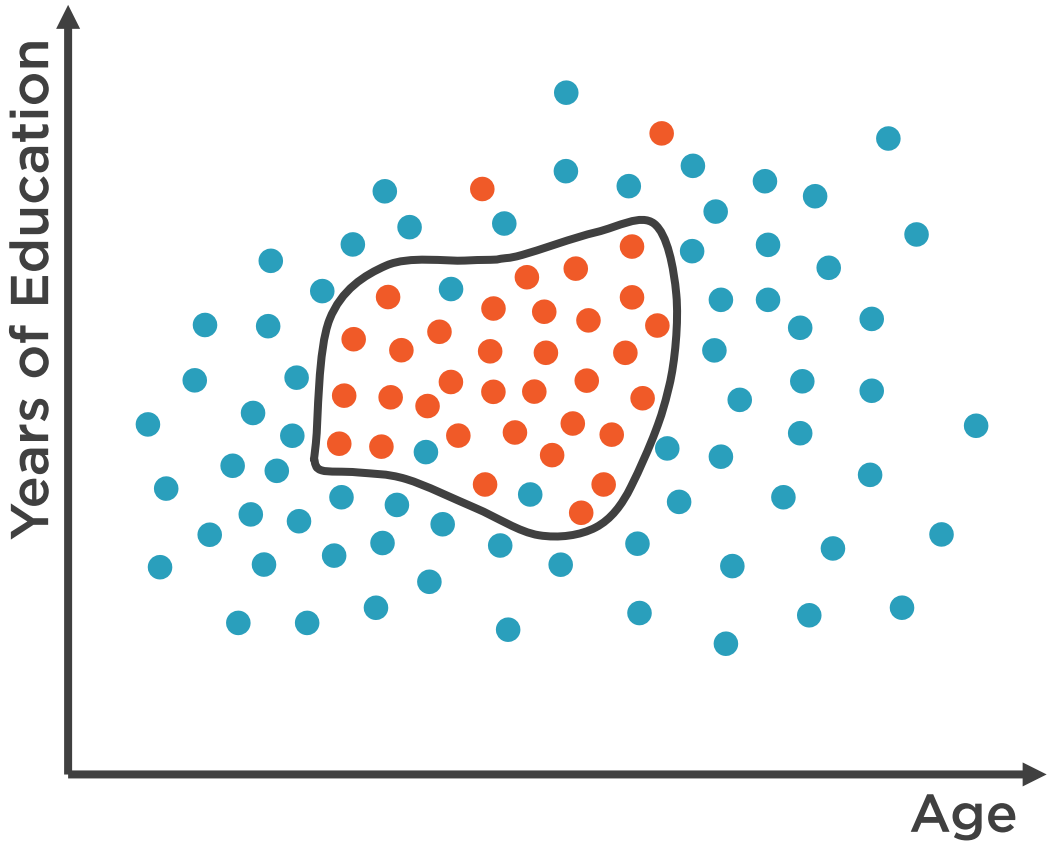
Amount (in $)	Zip Code	Target
330	80201	1
54	32501	0
670	60602	0
1200	52808	0
5600	52804	0
207	50321	1
700	83254	0

Define Target for ML Problems

Regression



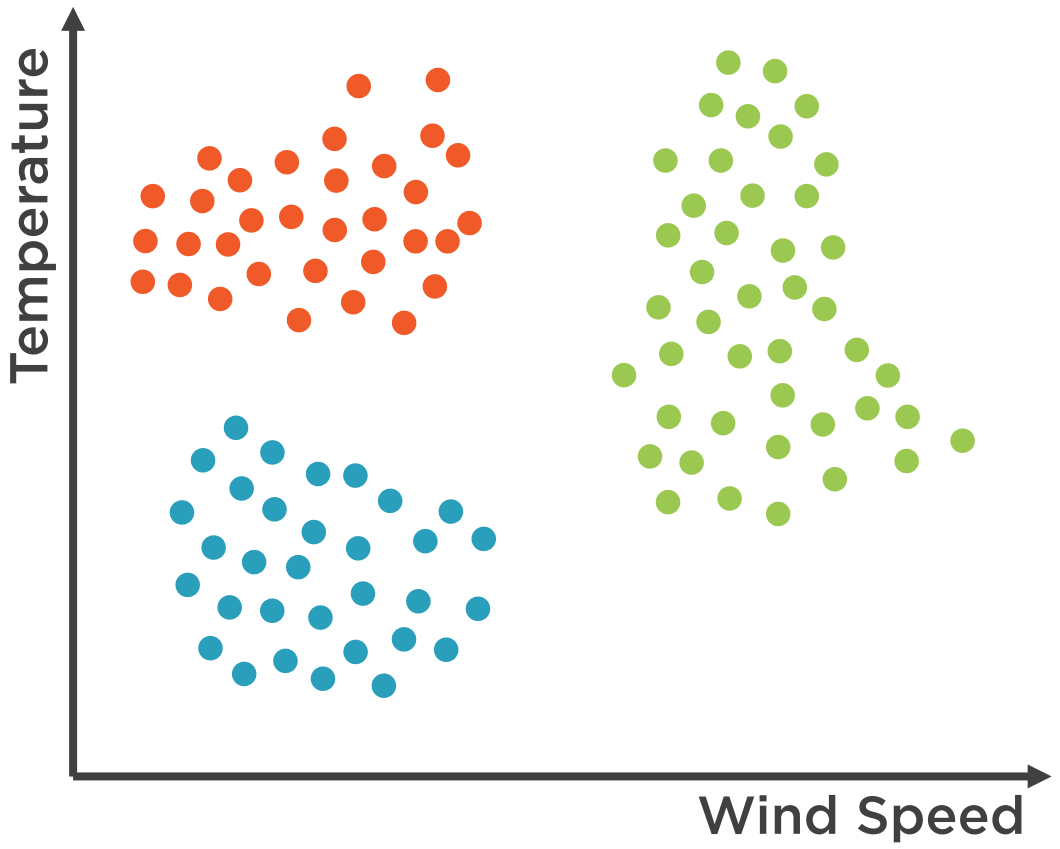
Classification



> 50K $

< 50K $

Multi-Class Classifier

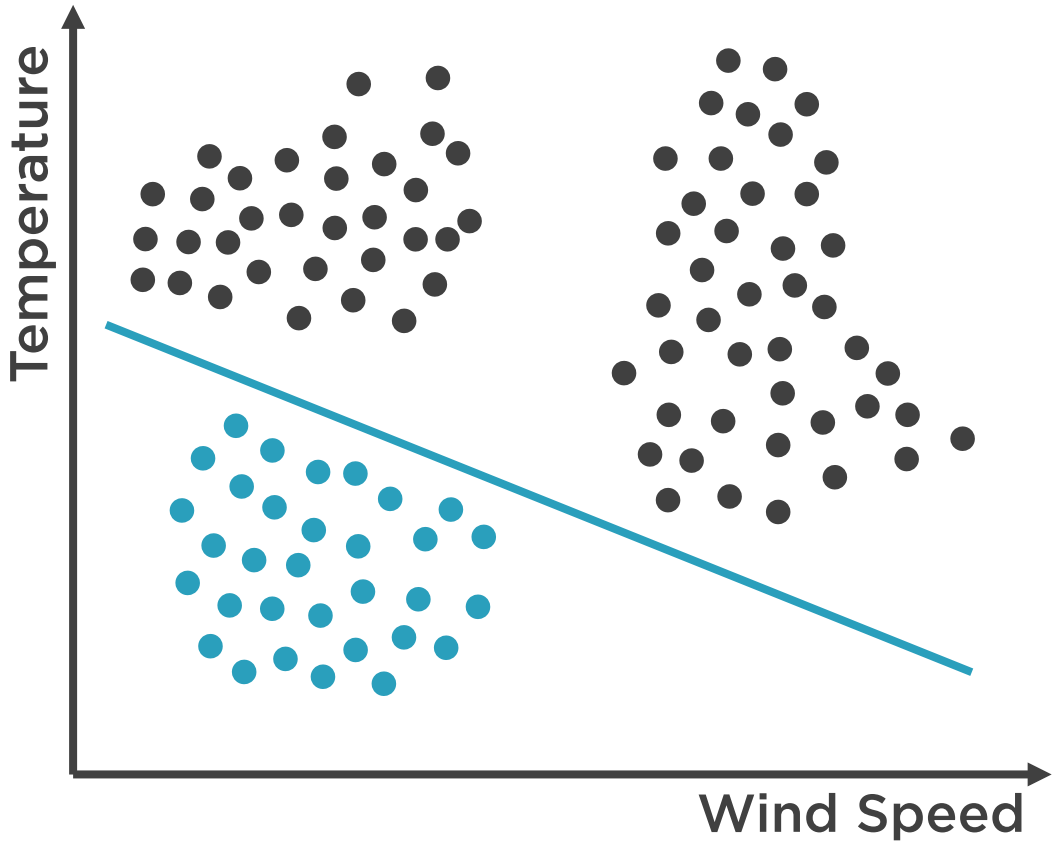


Snowy

Windy

Thunderstorm

Multi-Class Classifier

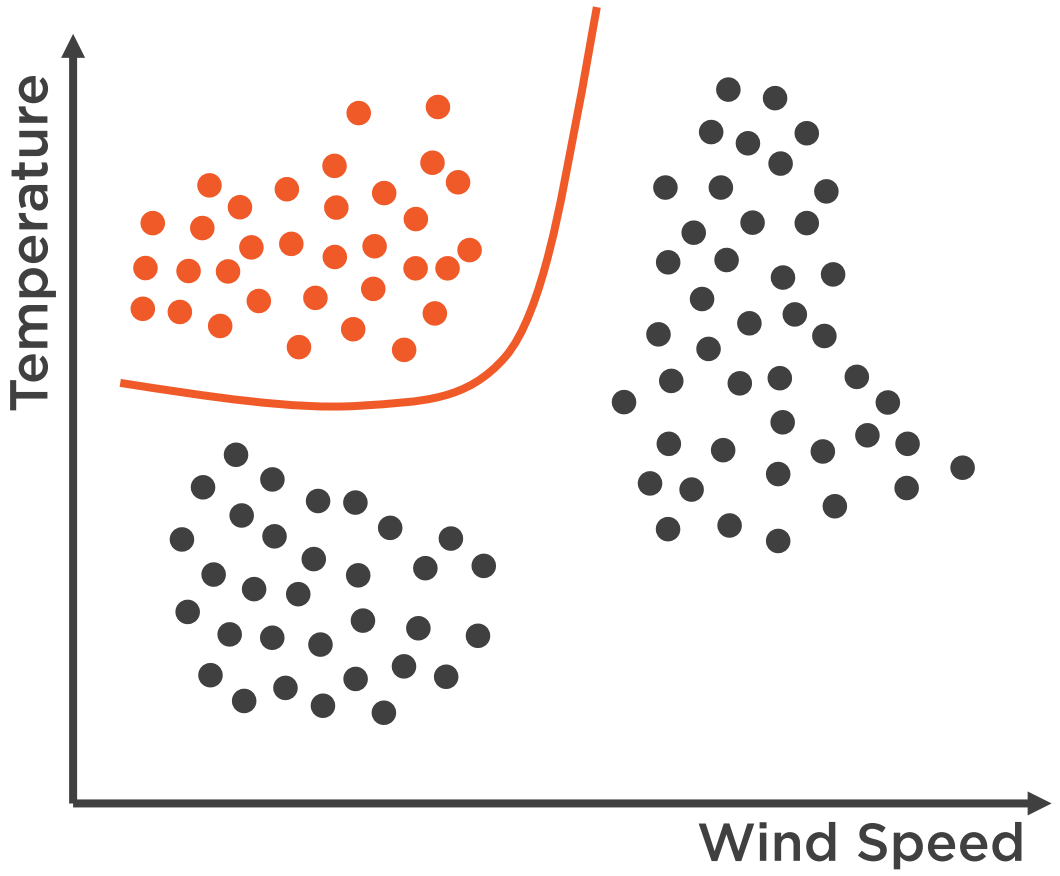


Snowy

Windy

Thunderstorm

Multi-Class Classifier

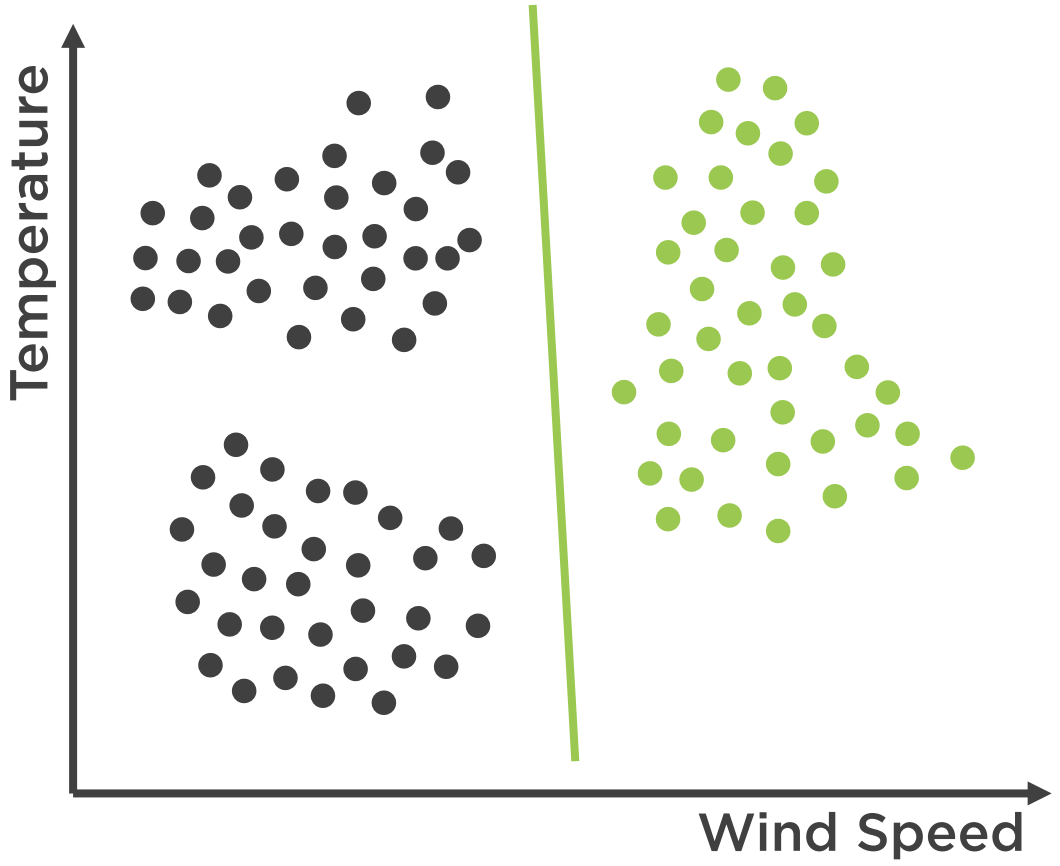


Snowy

Windy

Thunderstorm

Multi-Class Classifier

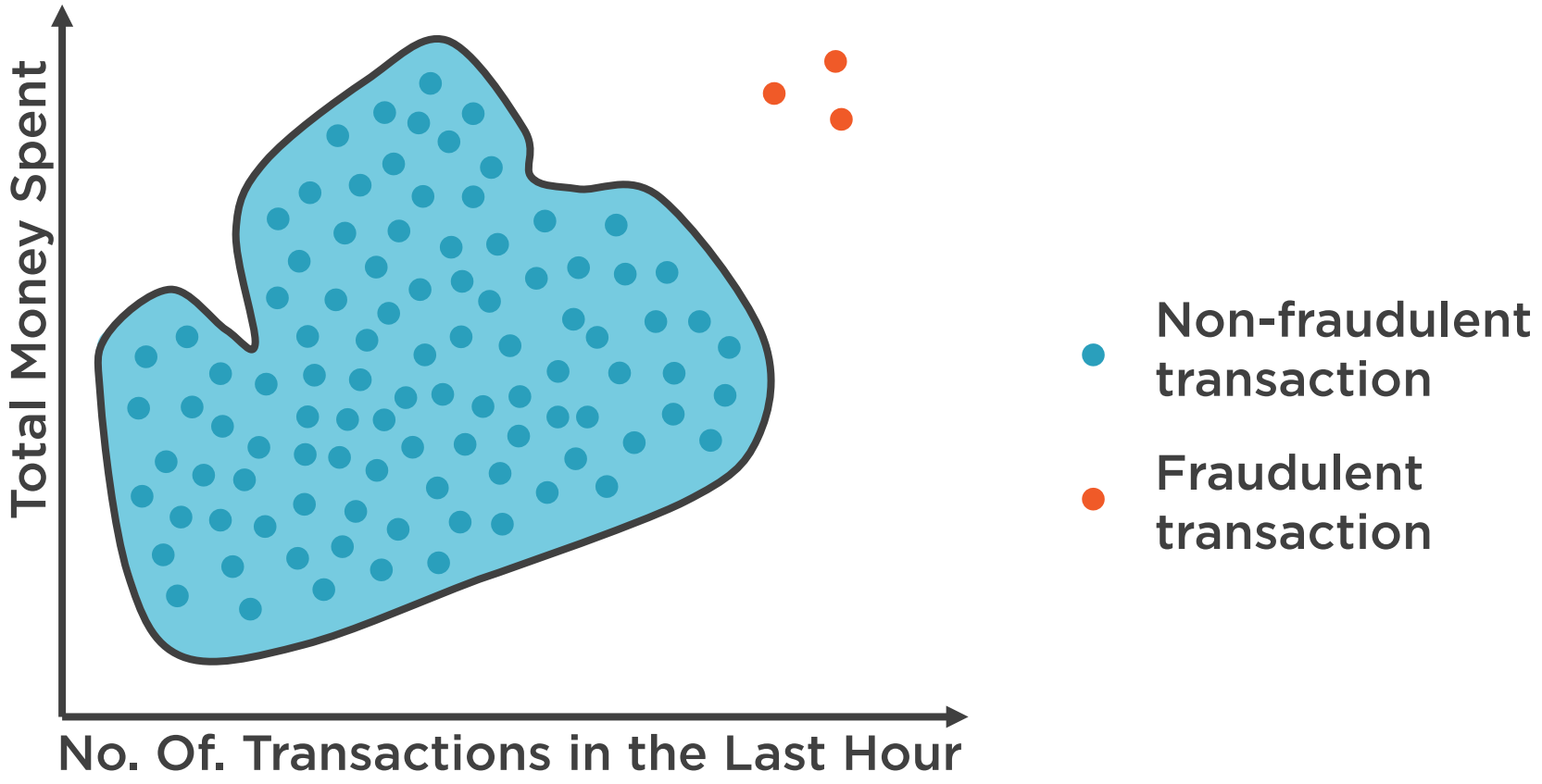


Snowy

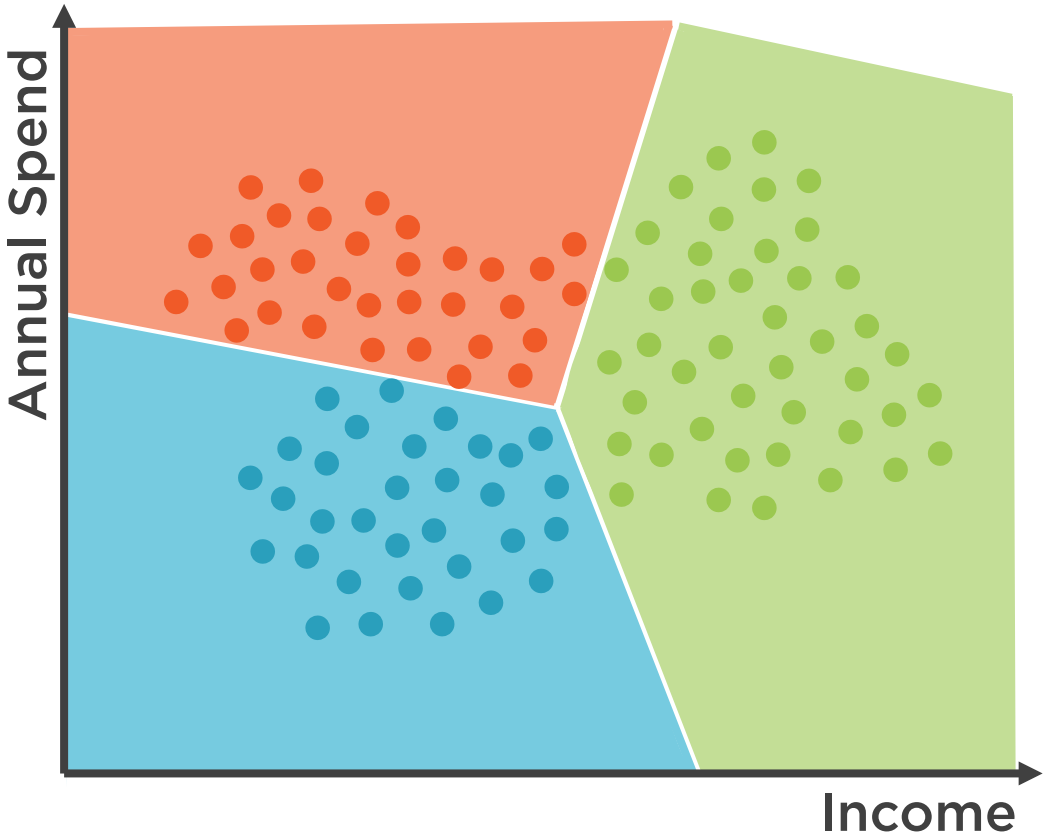
Windy

Thunderstorm

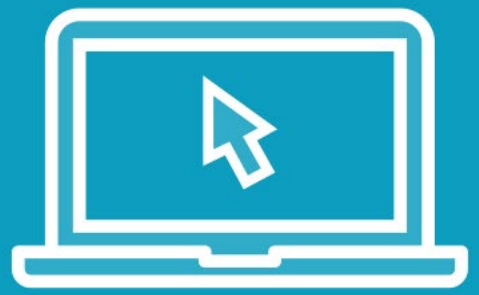
Anomaly Detection – Credit Card Fraud



Clustering



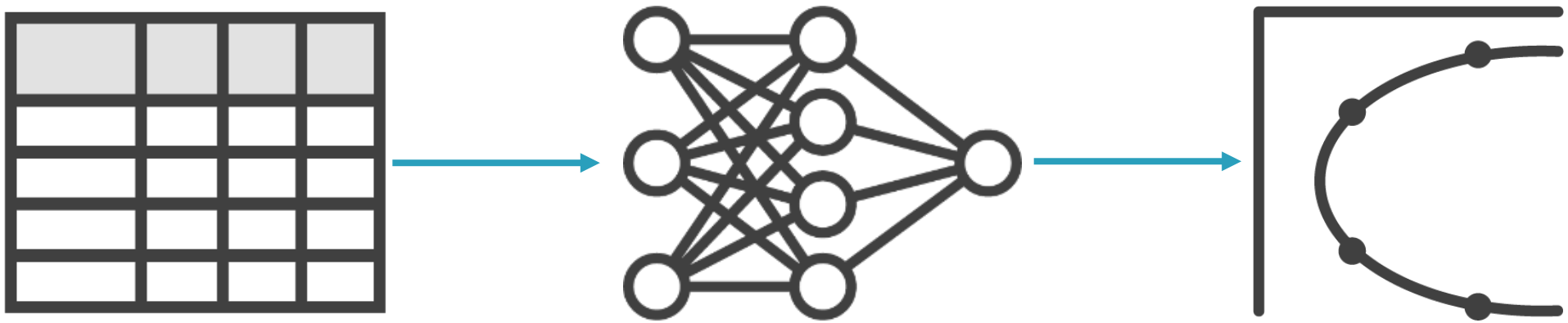
Demo



Explore datasets for different ML problems

How Algorithms Learn Models?

Algorithm vs. Model

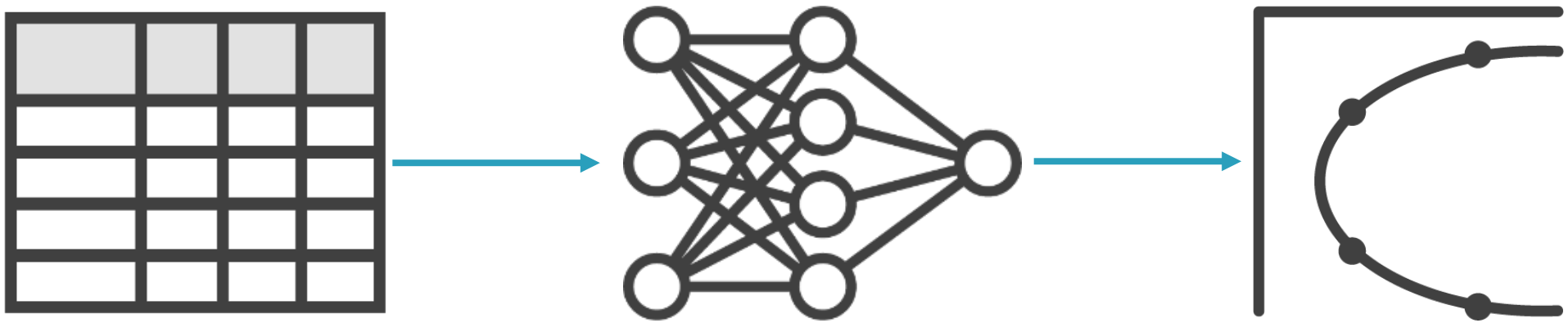


Dataset

Machine Learning Algorithm

Model

Algorithm vs. Model



Dataset

Unlearned Model

Learned Model

How Algorithms Learn Models?

$$a_1X_1 + a_2X_2 + a_3X_3 = y$$

Feature Vector

6 observations

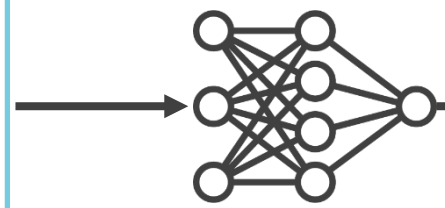
$$\begin{bmatrix} X_{11} & X_{21} & X_{31} \\ X_{12} & X_{22} & X_{32} \\ X_{13} & X_{23} & X_{33} \\ X_{14} & X_{24} & X_{34} \\ X_{15} & X_{25} & X_{35} \\ X_{16} & X_{26} & X_{36} \end{bmatrix}$$

How Algorithms Learn Models?

$$a_1X_1 + a_2X_2 + a_3X_3 = y$$

Feature Vector

$$\begin{bmatrix} X_{11} & X_{21} & X_{31} \\ X_{12} & X_{22} & X_{32} \\ X_{13} & X_{23} & X_{33} \\ X_{14} & X_{24} & X_{34} \\ X_{15} & X_{25} & X_{35} \\ X_{16} & X_{26} & X_{36} \end{bmatrix}$$



Machine Learning Algorithm

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

Model Parameters

How Algorithms Learn Models?

$$a_1X_1 + a_2X_2 + a_3X_3 = y$$

Feature Vector

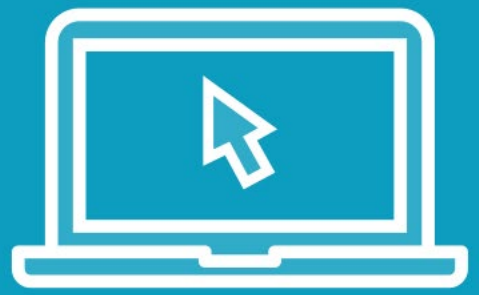
$$\begin{bmatrix} X_{17} & X_{27} & X_{37} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = y$$

New Observation

Model Parameters

Predicted Value

Demo



Modify the metadata of dataset

Summary

Data Quality is fundamental to ML

Dataset => Set of features + Target

Features => Input predictors

ML models predict different Target values

Raw Data can be directly used as features

Use Feature Selection to select relevant features