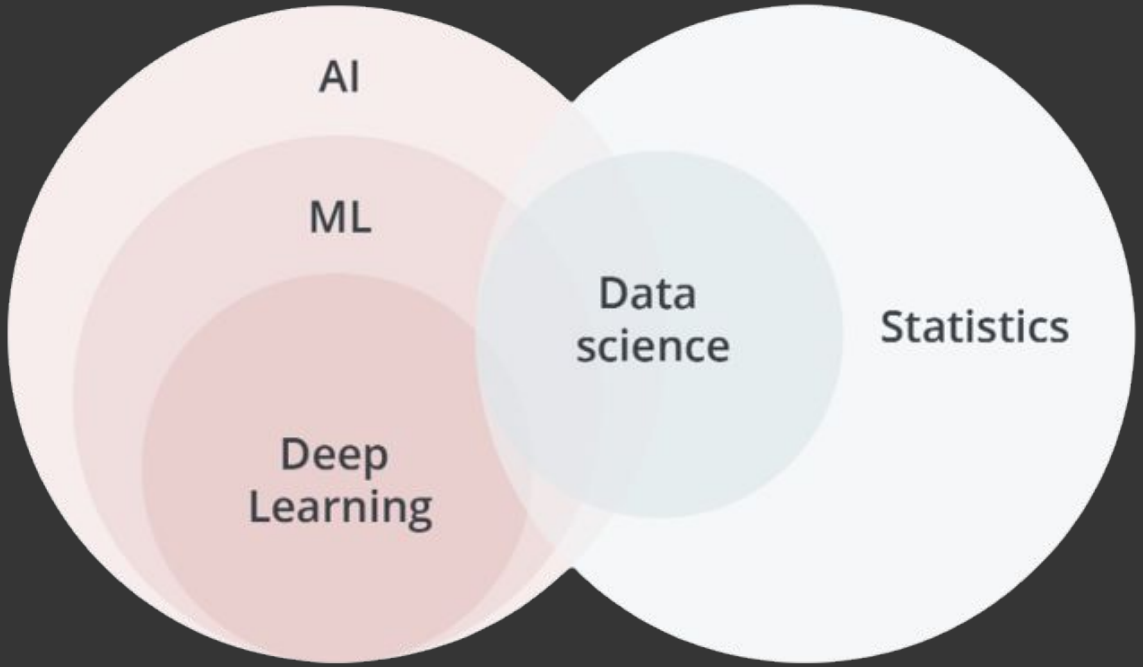
Machine Learning Demo: Analyzing Airbnb Data

By: Anna Hauk

What is Machine Learning?



- Deep learning is a subset of machine learning, and machine learning is a subset of artificial intelligence.
- Data science utilizes ML and statistics to uncover patterns and insights in data.
- ML-based AI learns its behavior from data.

Formal Definitions

- **Artificial Intelligence:** all-encompassing term that captures the research and implementation of systems that are capable of performing tasks intelligently in a given environment
  - AI is different because it uses its environment to shape its behavior, whether through experience or on the fly.
  - It can deal with the unknown and provide a generalized response for previously unseen situations.
- **Machine Learning:** subset, or implementation, of AI that deals with the research and implementation of systems that shape its behavior based on experience
  - Solves problems by “learning” from past data in order to make decision
  - needs historical data in order to perform well. It always goes through a training phase, where it consumes data (often in high volume) in order to update its inner data structure
  - Not all AI systems employ ML

Formal Definitions

- **Deep Learning (DNN) :** subset of machine learning that uses neural networks as its underlying algorithm
  - often used in the fields of image recognition, language processing, and speech recognition

- **Data Science:** discipline concerned with finding patterns and providing insights from data
  - Machine learning automates these processes and can pick up subtle relationships within data that could otherwise be missed using a traditional statistics approach.

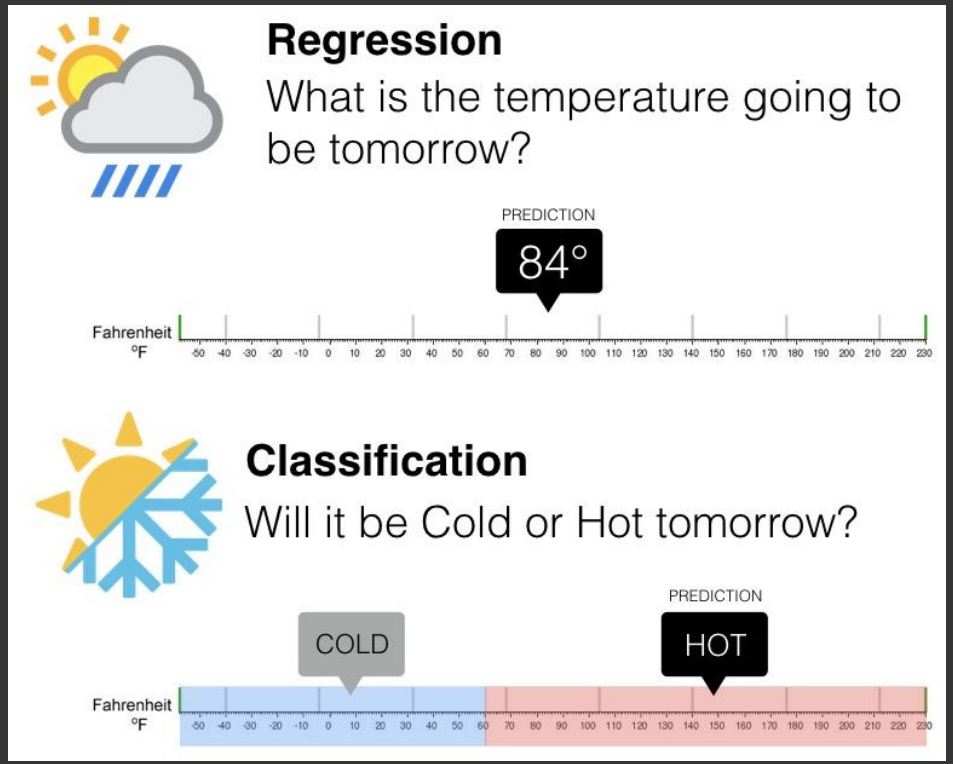
Supervised vs Unsupervised

Label is given

- Binary
- Multiclass
- Regression

Raw data without any labels

Regression vs Classification



Labels

Binary Classification

- Spam
- Not spam

Multiclass Classification

- Dog
- Cat
- Horse
- Fish
- Bird
- ...

Multi-label Classification

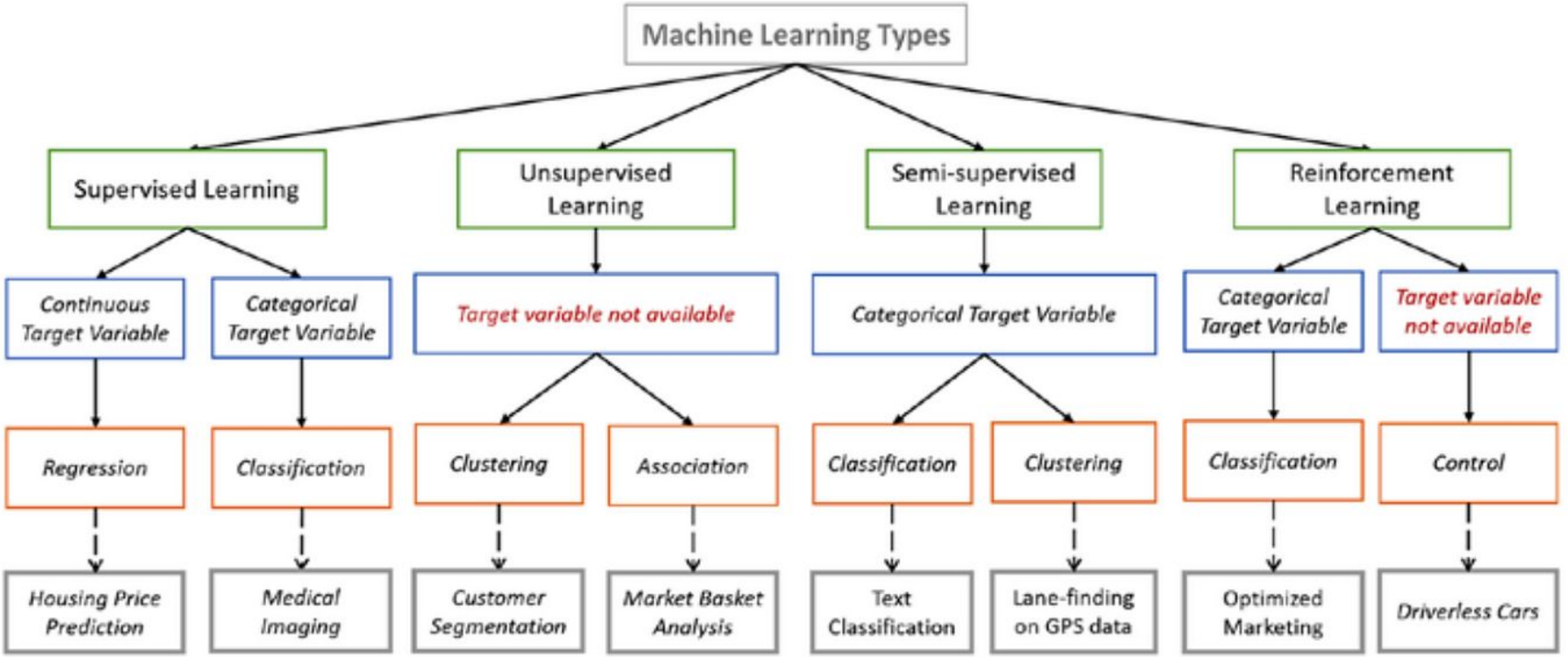
- Dog
- Cat
- Horse
- Fish
- Bird
- ...

Regression

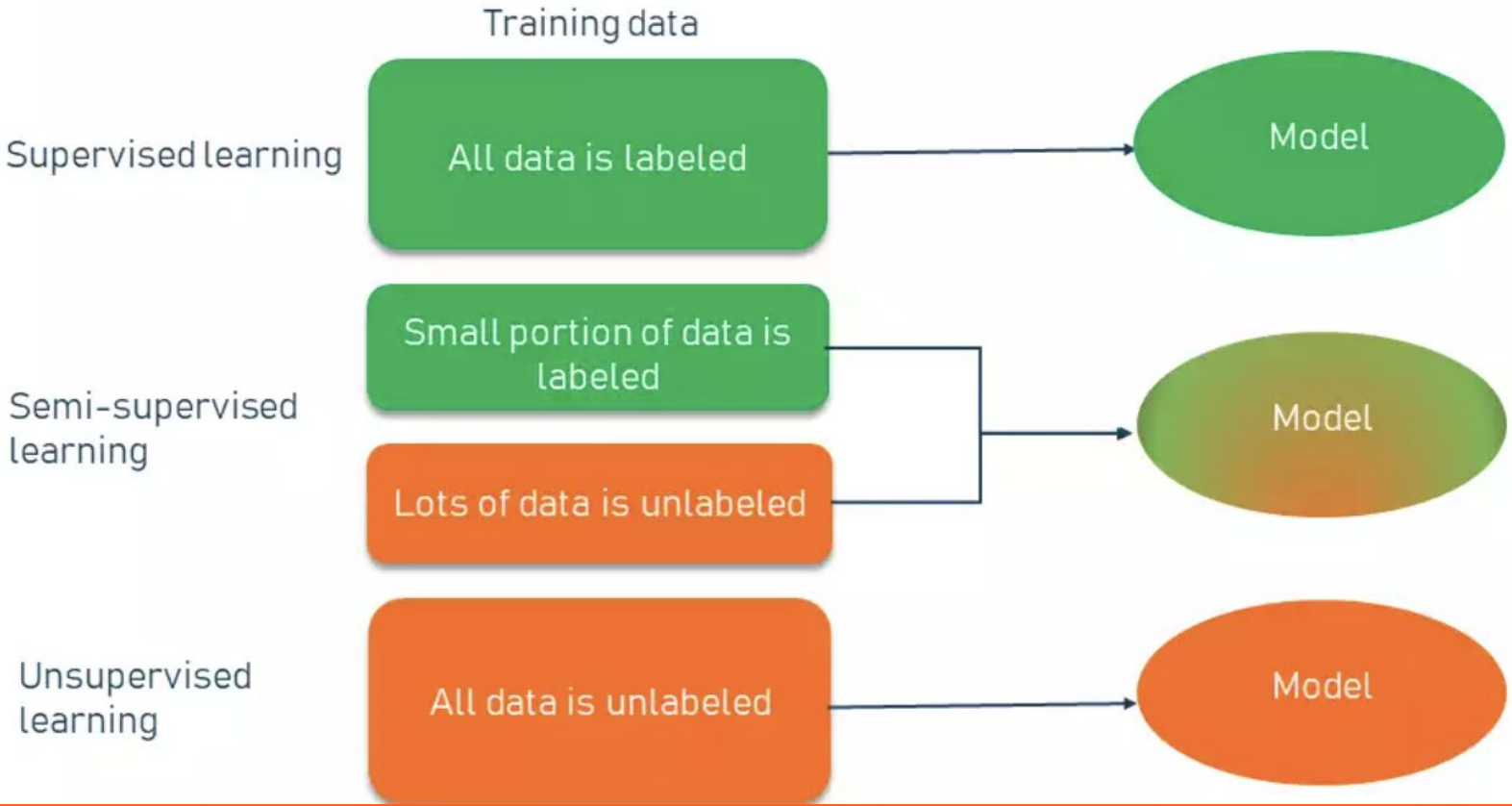
height of the animal?

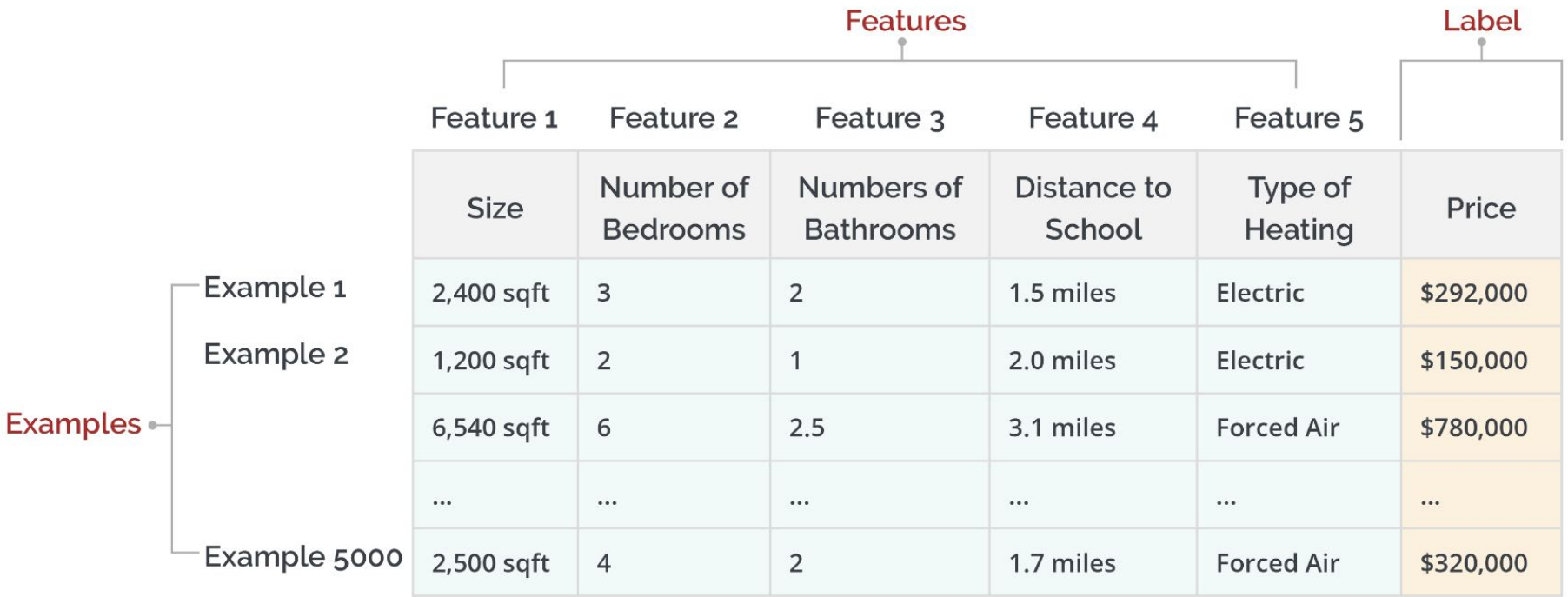
temperature for tomorrow?

credit score of a person with a specific financial profile?


Machine Learning Types
Supervised Learning
Unsupervised Learning
Semi-supervised Learning
Reinforcement Learning
Continuous Target Variable
Categorical Target Variable
Target variable not available
Categorical Target Variable
Categorical Target Variable
Target variable not available
Regression
Classification
Clustering
Association
Classification
Clustering
Classification
Control
Housing Price Prediction
Medical Imaging
Customer Segmentation
Market Basket Analysis
Text Classification
Lane-finding on GPS data
Optimized Marketing
Driverless Cars

SUPERVISED LEARNING vs SEMI-SUPERVISED LEARNING vs UNSUPERVISED LEARNING




Features
Label
Feature 1
Feature 2
Feature 3
Feature 4
Feature 5
Size
Number of Bedrooms
Numbers of Bathrooms
Distance to School
Type of Heating
Price
Examples
Example 1
Example 2
Example 5000
2,400 sqft
3
2
1.5 miles
Electric
$292,000
1,200 sqft
2
1
2.0 miles
Electric
$150,000
6,540 sqft
6
2.5
3.1 miles
Forced Air
$780,000
...
...
...
...
...
...
2,500 sqft
4
2
1.7 miles
Forced Air
$320,000

Data Preparation

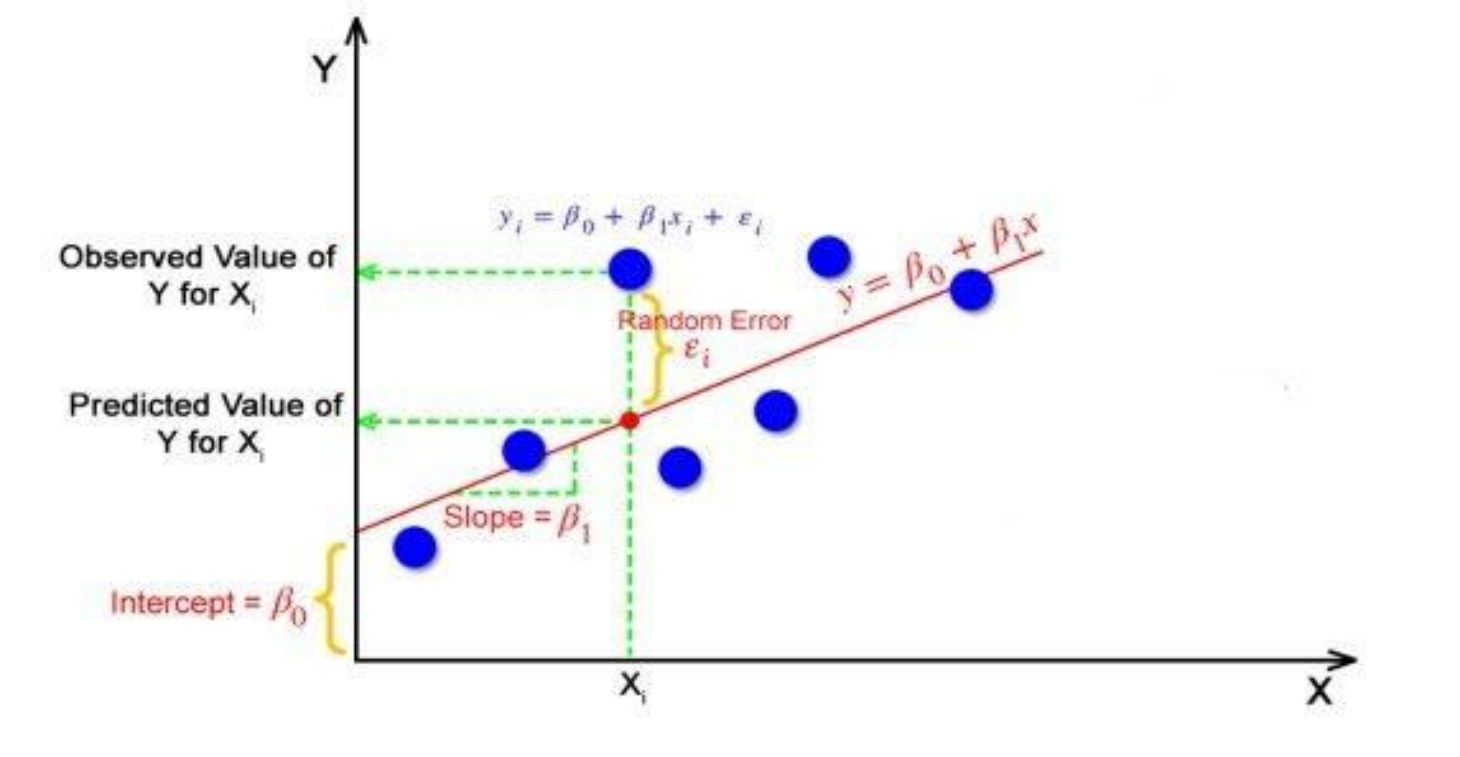
Sample dataset

Name	Income	Credit Score	Occupation	Job Sector	Loan Status
John Doe	$76,000	650	Engineer	Engineering	Good
Gill Bates	$85,000	760	Nurse	Healthcare	Defaulted
Jane Doe	“95000.00”	0	Banker	Financial	Good
John Doe	$76,000	650	Engineer	Engineering	Good
Melon Usk		810	Flight Attendant	Transportation	Excellent
Barren Wuffet	5000/mo	35000	Contractor	Construction	Defaulted

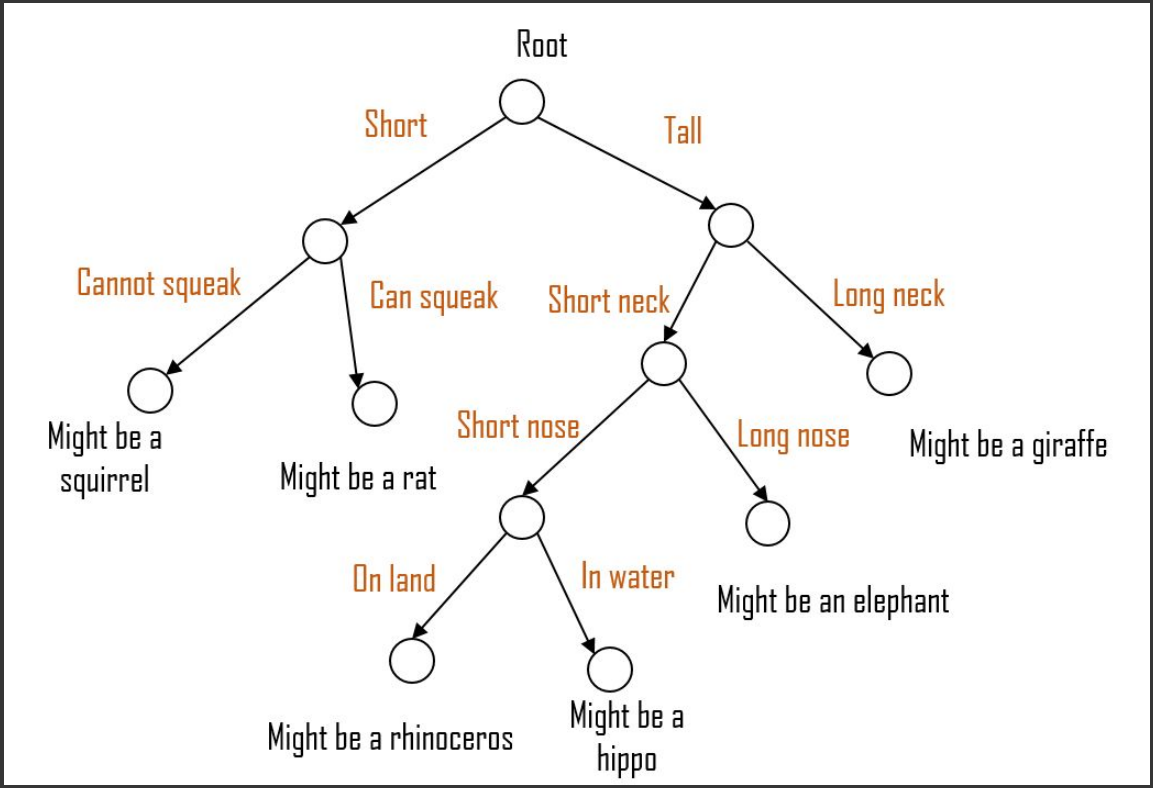
Problems in the data:

- The entry for John Doe is repeated
- Income format is not standardized
- The entry for Melon Usk contains a missing value
- Credit score contains two outliers — 0 and 35000, likely to be errors as typical range is between 350 to 850
- Occupation and Job Sector are somewhat redundant as they tell similar stories

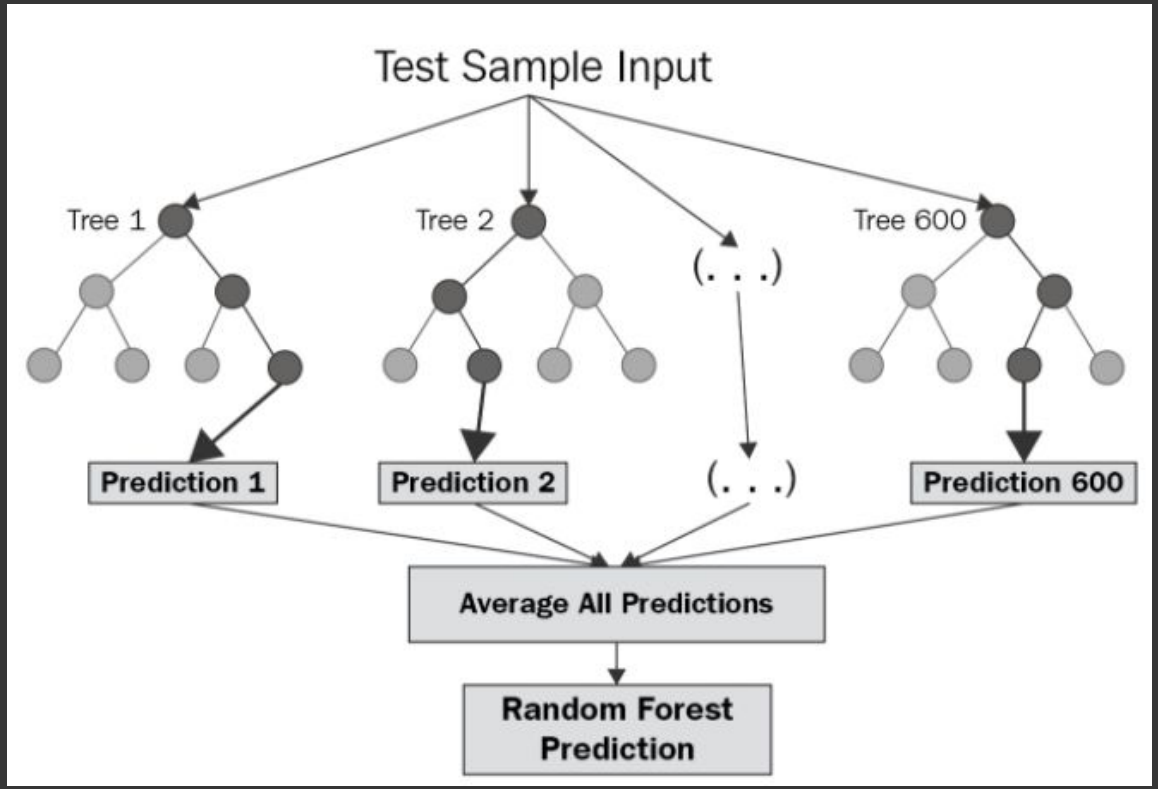
Linear Regression



Decision Trees

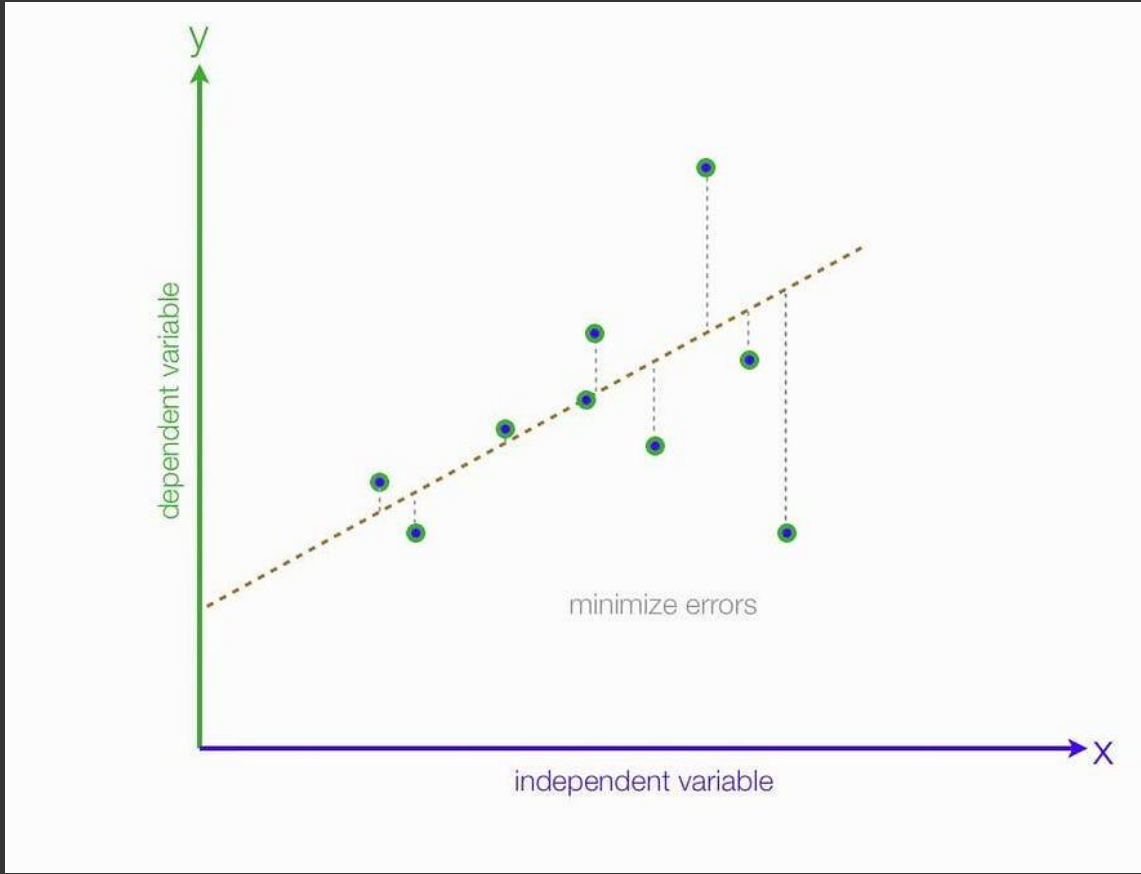


Random Forest



RMSE & R Squared

Root Mean Squared Error



$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

shows how well a regression model (independent variable) predicts the outcome of observed data (dependent variable)

Where,

$\hat{y}$ – *predicted value of y*

$\bar{y}$ – *mean value of y*

R-Squared

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$