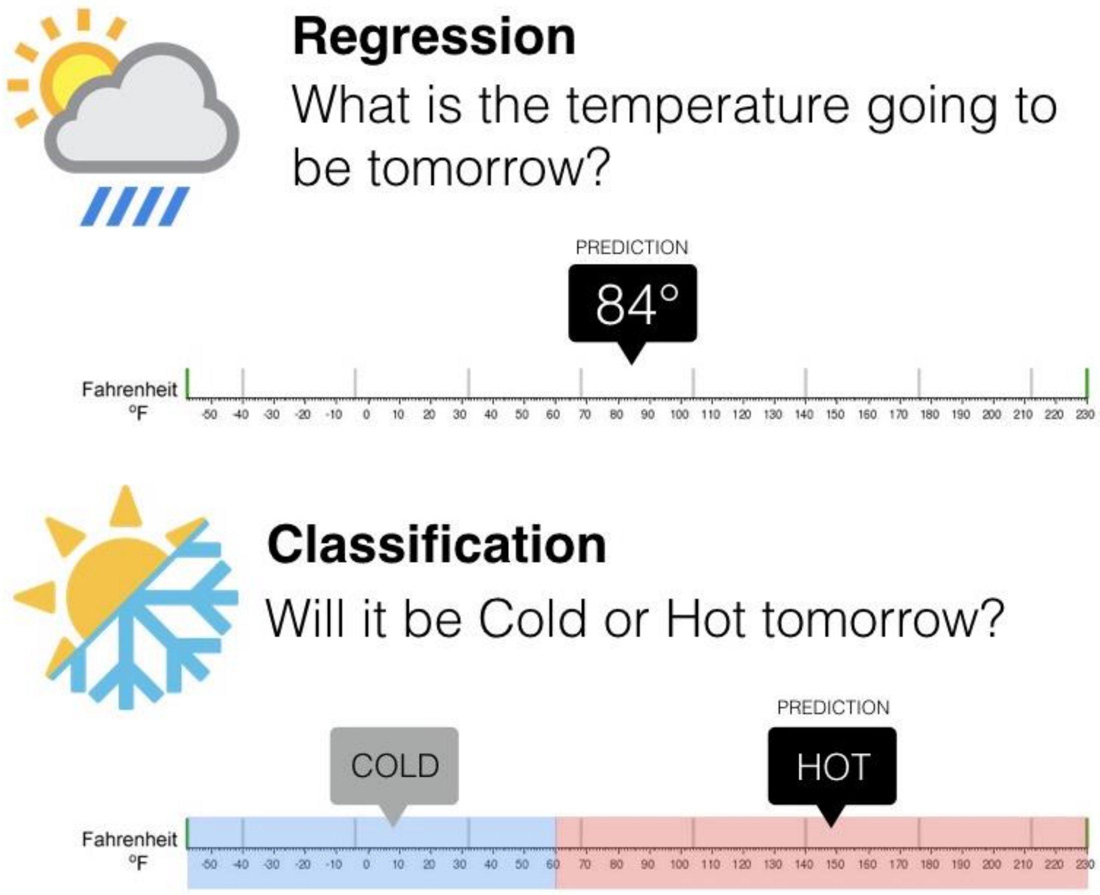
Data Driven Modeling 2: Classification

Himaghna Bhattacharjee

“To be or not to be...”

Rapid Advancement in Process Intensification Deployment

Classification



Regression

What is the temperature going to be tomorrow?

Classification

Will it be Cold or Hot tomorrow?

- Classification is the machine learning task where you want to predict a categorical output, called a 'class'

Examples

Is a catalyst good, bad or average?

Classes

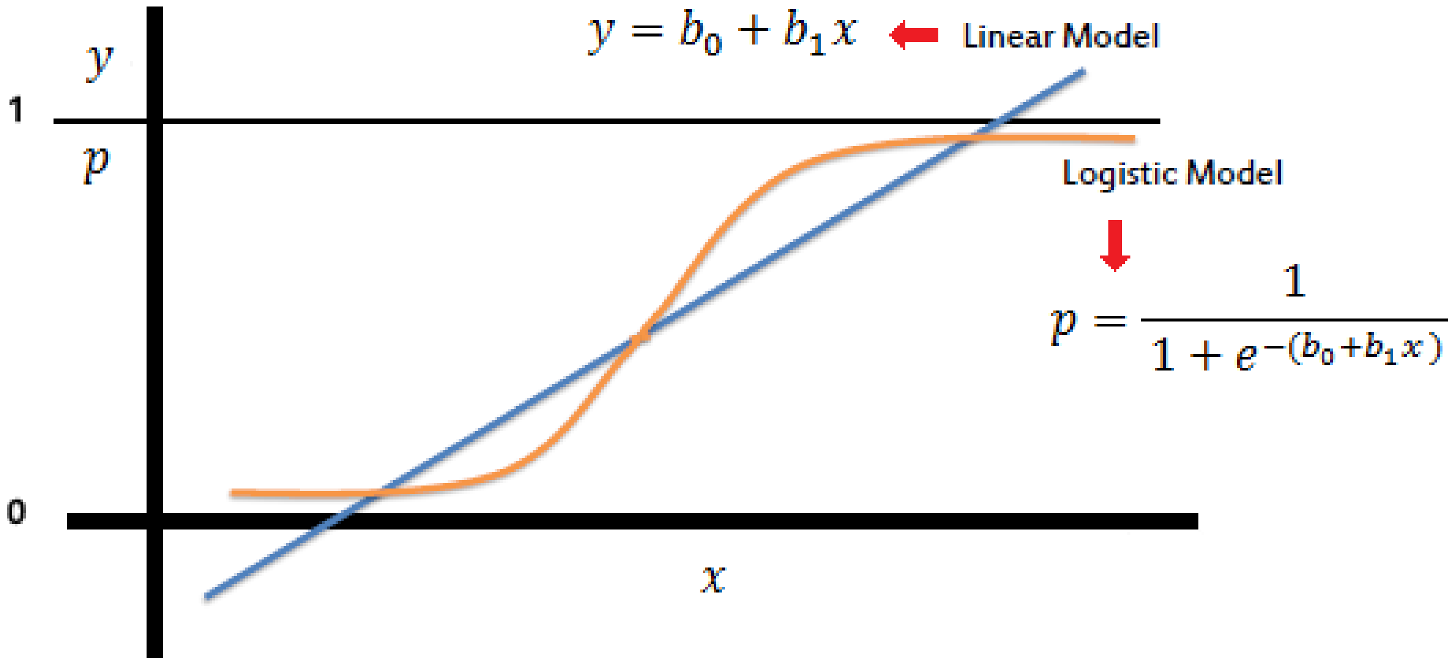
Good Average Bad

Is the flow laminar or turbulent?

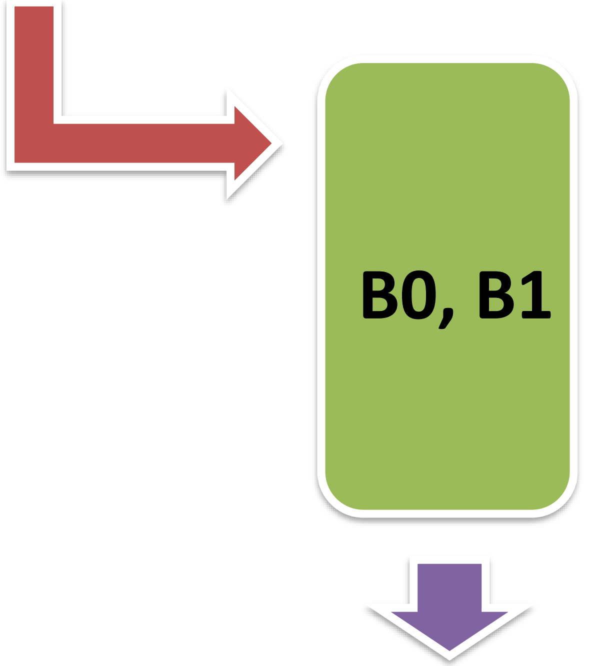
Classes

Laminar Turbulent

Logistic Regression



Descriptors

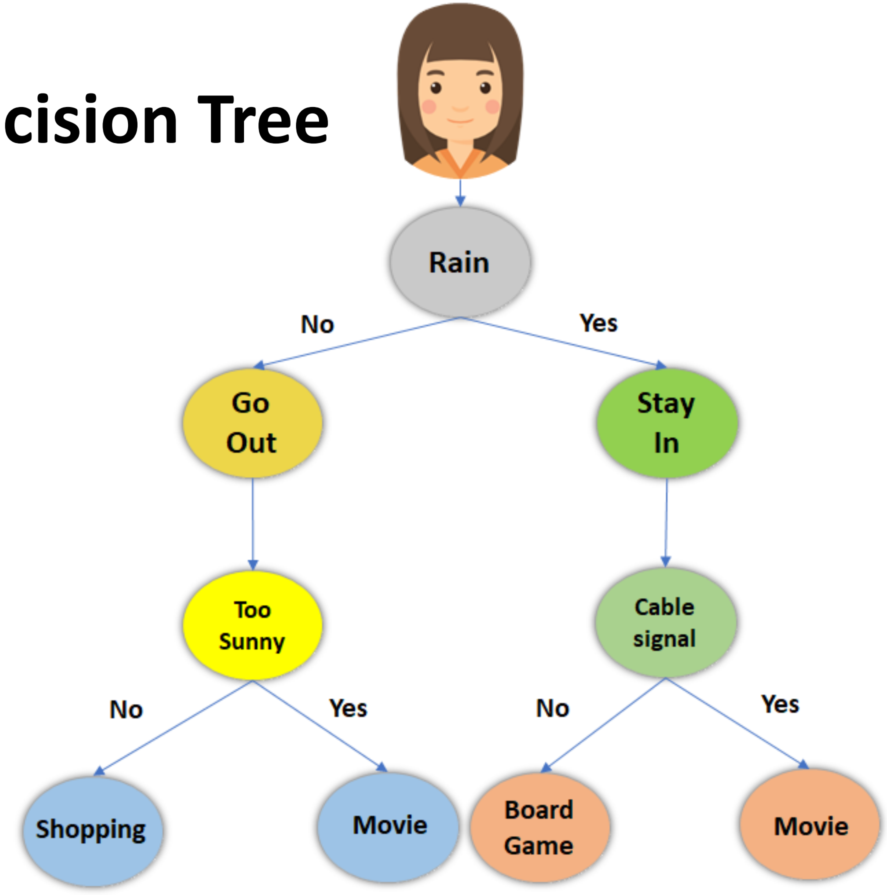

B0, B1

Class Membership

- Logistic regression is a linear model for **binary classification**

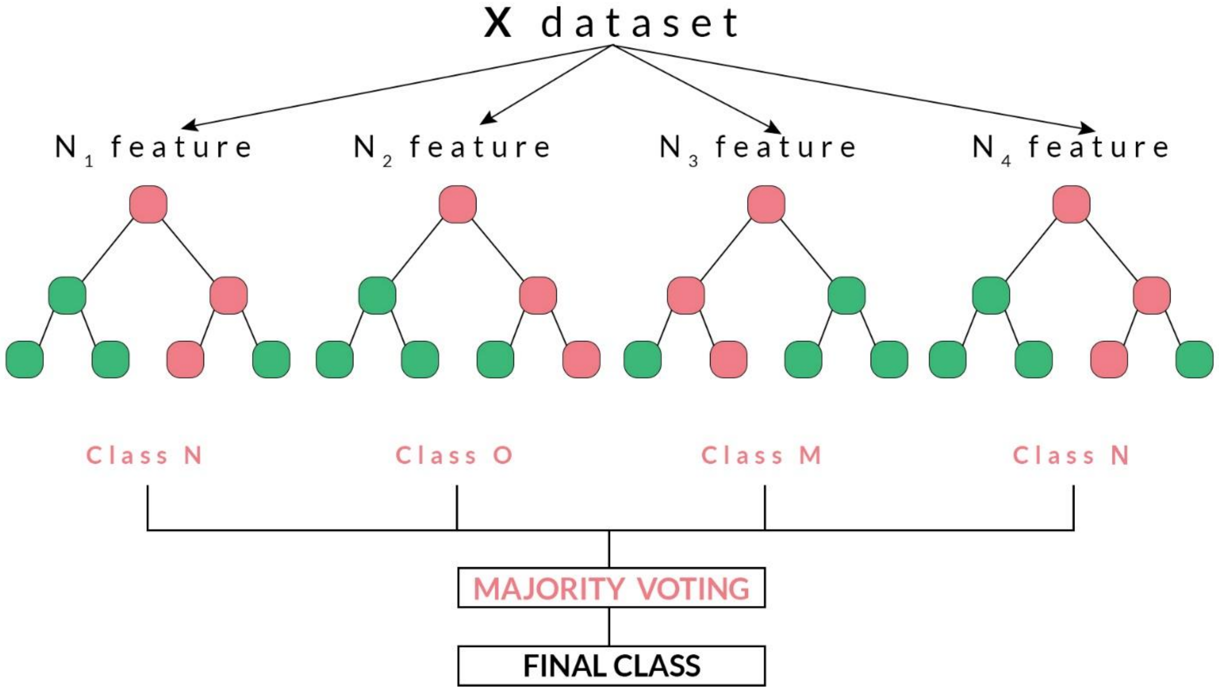
Random Forest

Decision Tree



Jane's Weekend Plan

A Random Forest



- Random forest uses a collection of decision trees to classify a sample
- **Can be used for multiclass classification**

Confusion Matrix

Actual Class	+	True Positive	False Negative
	-	False Positive	True Negative
		+	-
		Predicted Class	

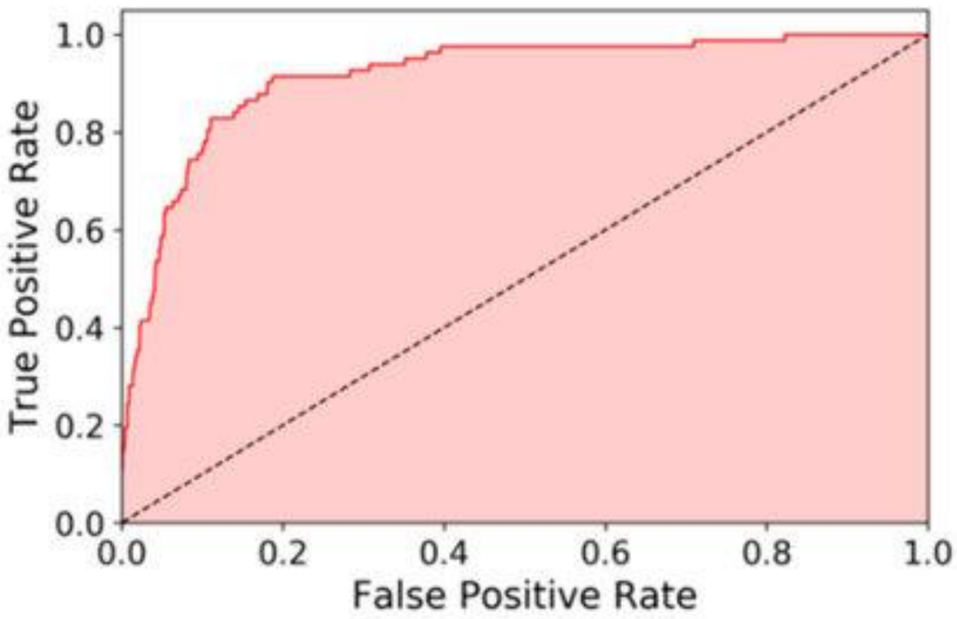
Measures and Metrics

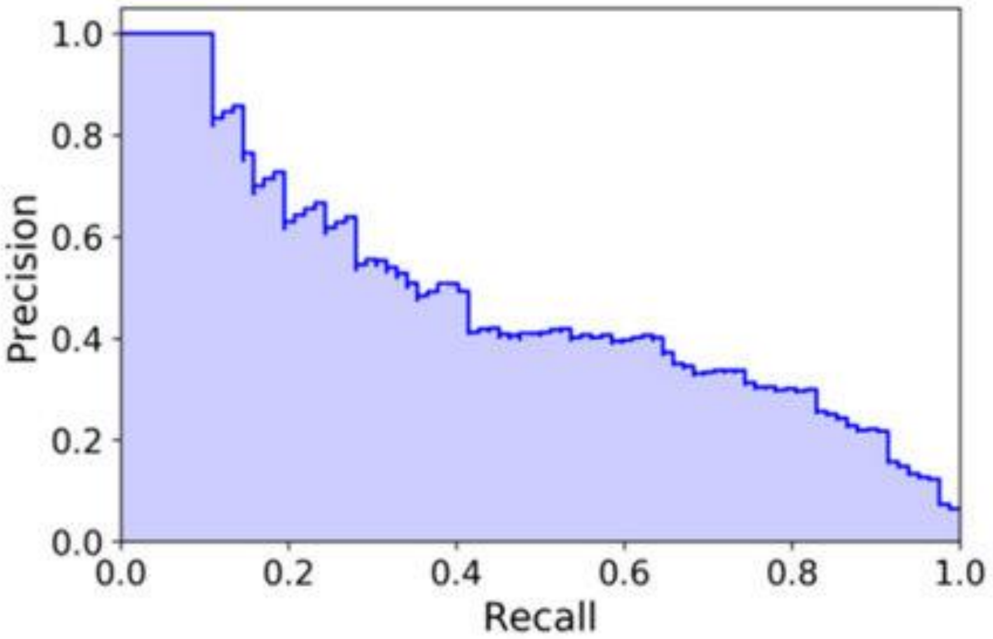
Accuracy: $\dfrac{Correct\ Predictions}{Total\ Predictions}$

Precision and Recall curve are good measures for imbalanced datasets with lot of negatives

Recall: $\dfrac{True\ Positive}{True\ Positive + False\ Negative}$

Precision: $\dfrac{True\ Positive}{True\ Positive + False\ Positive}$





Problem Set-up

- We will look at the California dataset and instead of predicting exact house price (regression), we will predict if a house has a high or low price (classification) based on some cutoff value.