gTech <professional services/>

WiDS Datathon

Introduction to Machine Learning

gTech gPS Data Science
Christiane Ahlheim, Yan Sun

Feb 2022

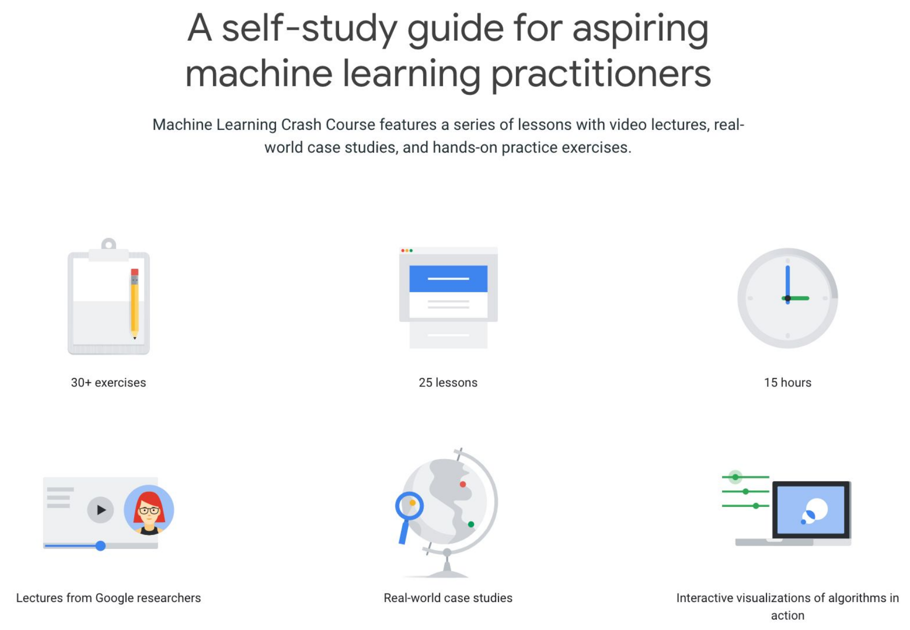
What we'll cover in the next 45 minutes

- What is Machine Learning?
- Common distinctions: Supervised vs Unsupervised
- Model Generalization
- Supervised Learning
    - Classification
    - Regression

For more details...

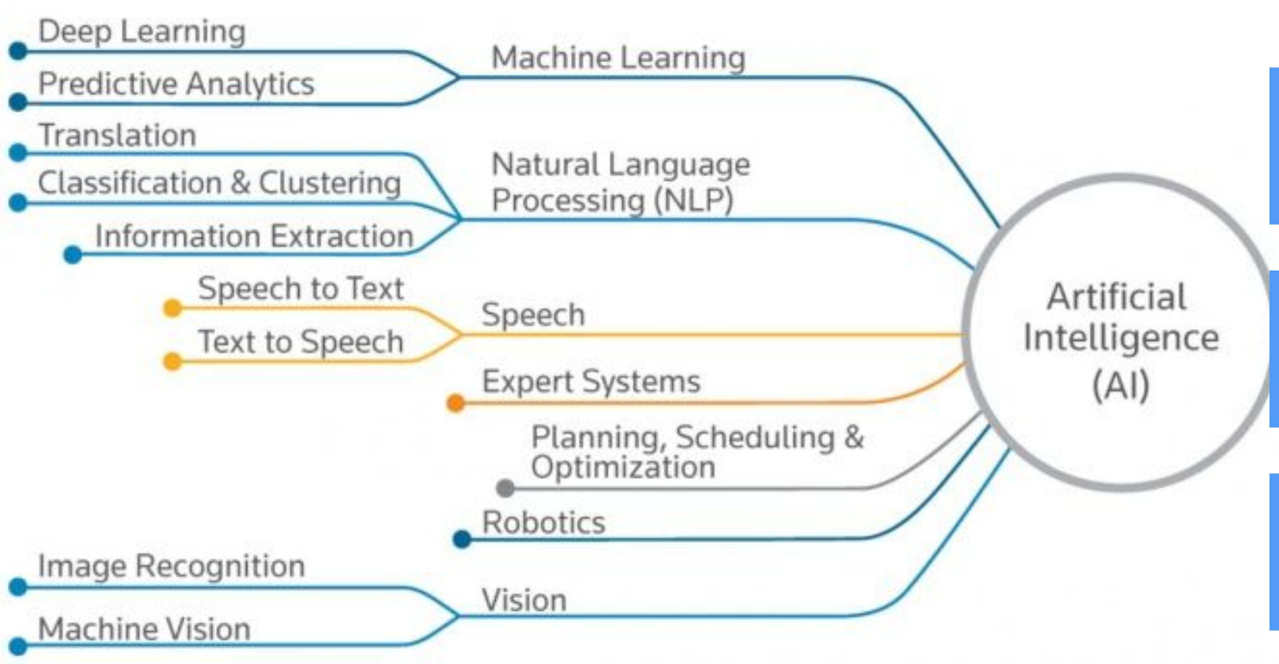
[Machine Learning Crash Course | Google Developers]

Source of most of the content shared here.



Common Terminology

Machine Learning is...



Source: Neota Logic

One branch of the field of Artificial Intelligence

A way of solving problems without explicitly codifying the solution

A way of building systems that improve themselves over time

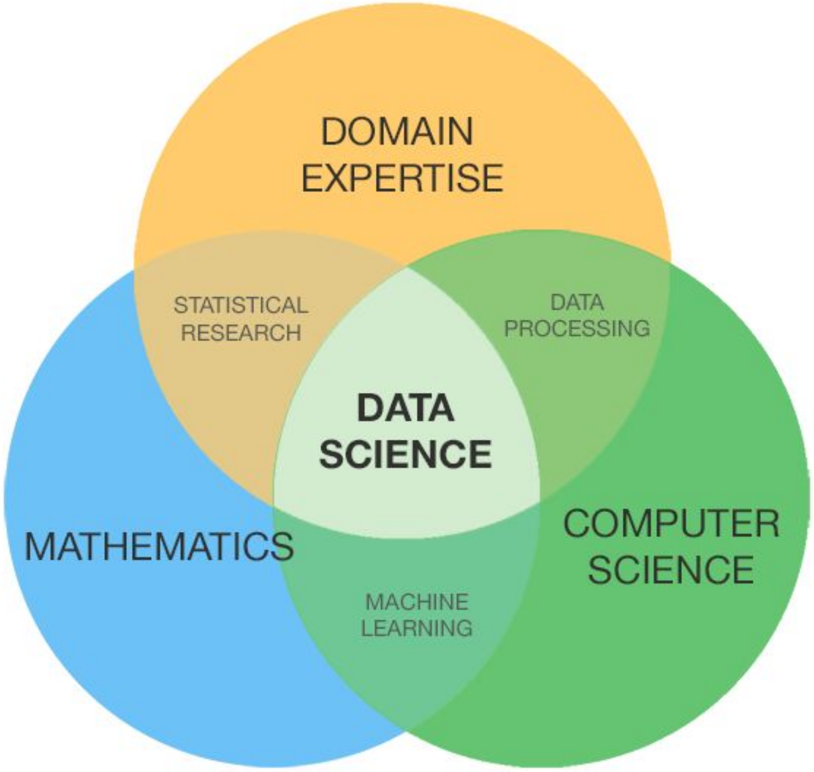
Machine Learning ≠ Data Science

Data Science:

- Solving business problems in a data-driven way
- Include define problem statement, data processing and model building

Machine Learning:

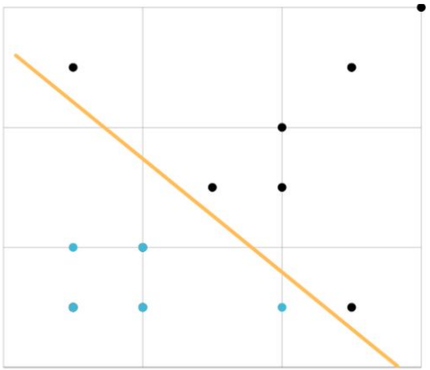
- A practice of using algorithms to capture the insights from big data
- One of the tools that Data Scientist uses



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

Supervised Learning

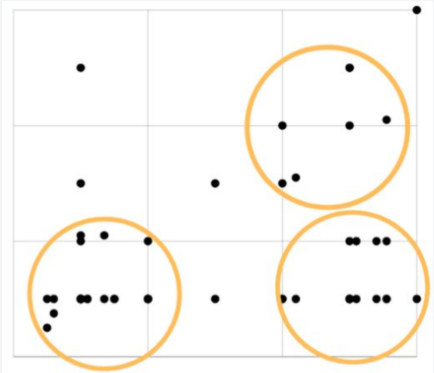
- Supervised learning is the machine learning task that **use labeled datasets** to train algorithms which will classify data or predict outcome



- **Classical examples:**
  - Time Series Forecasting: Stock price, Sales forecast
  - Classification: Handwriting Recognition, Tumor Detection
  - Regression: House rent, Car price prediction

Unsupervised Learning

- Unsupervised learning is the type of algorithm that learn pattern from **untagged** data



- **Classical examples:**
  - Customer segmentation
  - Feature reductions

gTech
<professional services/>

This year's WiDS datathon

> "[...] predict the energy consumption using building characteristics and climate and weather variables . "

Model Generalization

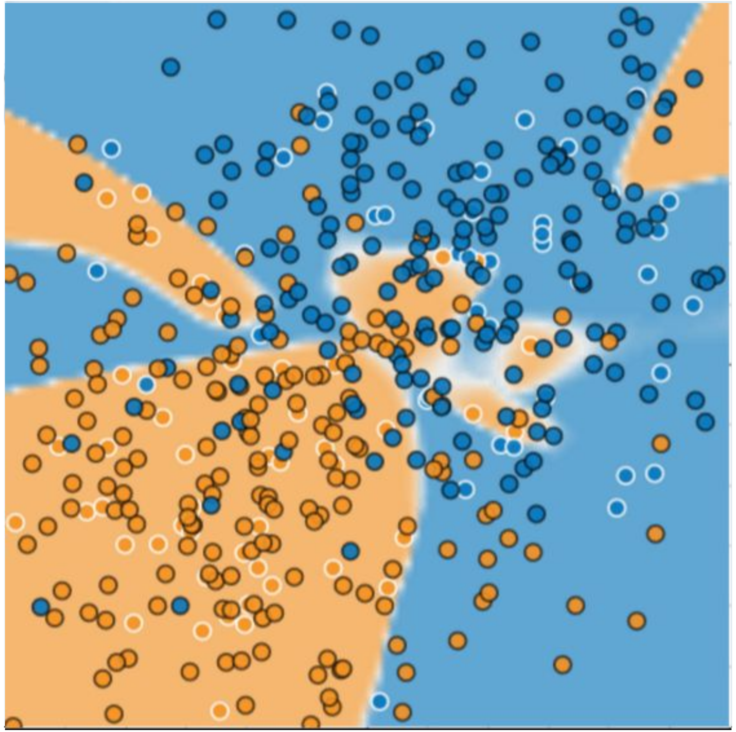
Generalization: Over- and Underfitting

The goal for each ML algorithm: predict well on **new data.**

Risk: (Complex) models can **overfit** peculiarities in your data, instead of learning the true signals.

This results in **poor performance** on new data points.

Source: [Generalization: Peril of Overfitting | Machine Learning Crash Course | Google Developers]

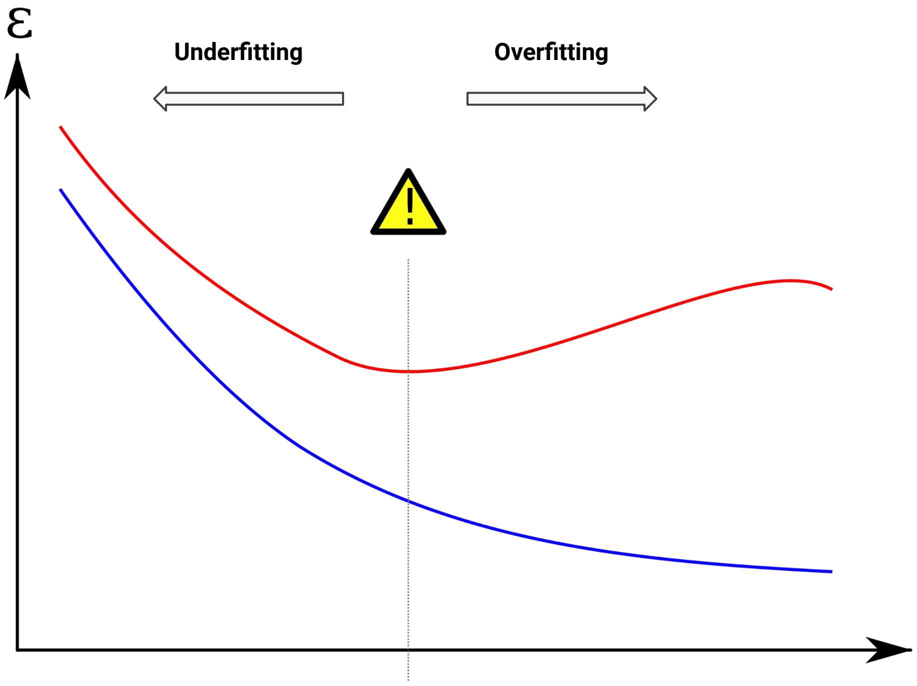


Generalization: Over- and Underfitting

We can diagnose over- and underfitting by inspecting the model performance on our training data (blue) and new data (red).

Overfitting: The error on the training data decreases, but *increases* on the new data

Underfitting: The error on the training data is still too high and could go down further.



Generalization: Training- and Test-Set

How can we know how our model will perform on new data points?

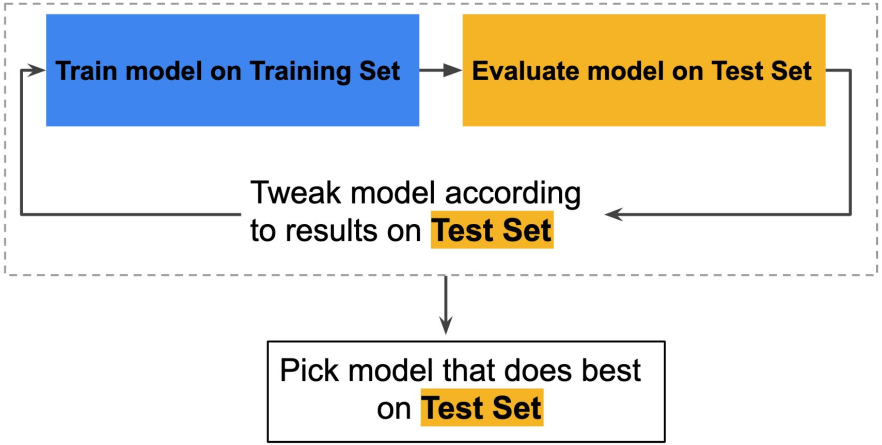
We split the data!

The test set needs to:

- Be large enough to yield statistically meaningful results
- Be representative of the whole dataset
- Be independent of the training data

Never train on test data! If your model performance is too good, check that the training data has not leaked into the test data.

Rule of thumb: 80/20 split

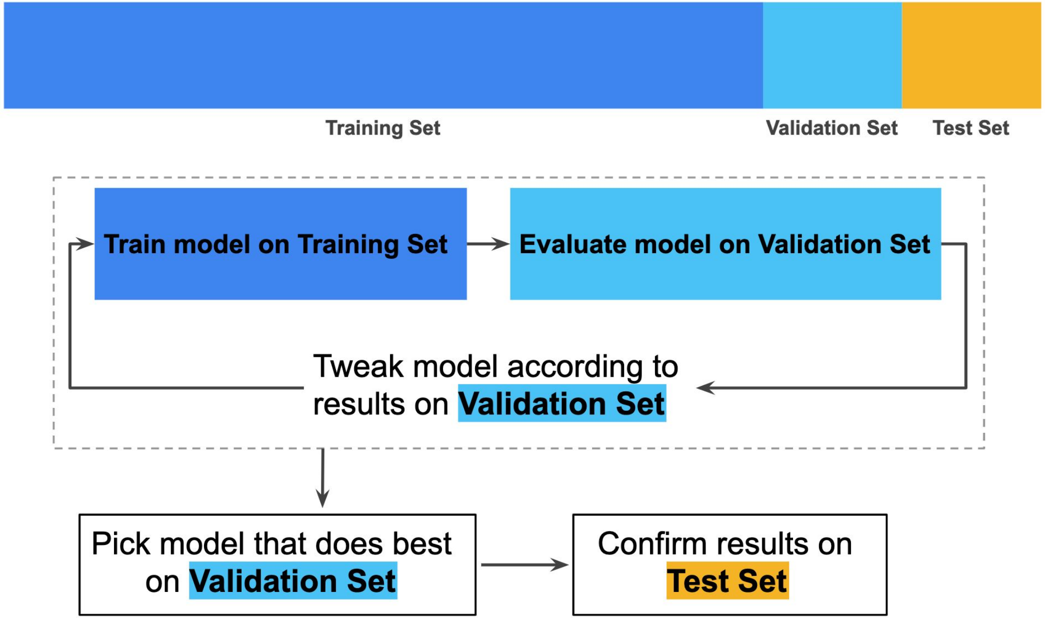


Generalization: Validation Set

Introducing a test-set already reduces the risk of overfitting greatly, but we still risk overfitting to the *test set*.

This is why general best practice is to have three splits: training, validation, and test set.

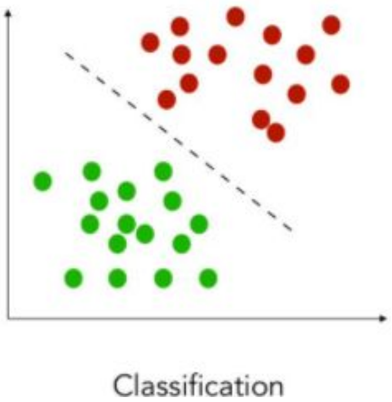
In this workflow, only the final model is checked against the test set, and risks of overfitting are thus reduced further.



Classification and Regression

Classification

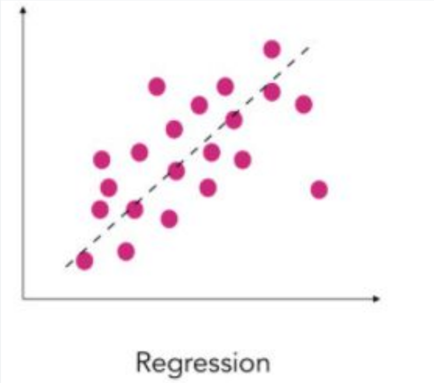
- Labels are **categorical**, which can be two (binary) or more (multiclass)



- Classification model predict each observation's category
  - Output the probability for each category
- Classical examples:
  - Tumor detection
  - Handwriting recognition
  - …

Regression

- Labels are (usually) **continuous**, but could, e.g., only be integers



- Regression model predict each observation's value
  - Output the actual value as prediction
- Classical examples:
  - Stock market
  - Sales
  - …

By Maël Fabien,
https://maelfabien.github.io/machinelearning/ml_base/#supervised-vs-unsupervised

This year's WiDS datathon

> “[...] predict the energy consumption using building characteristics and climate and weather variables . ”

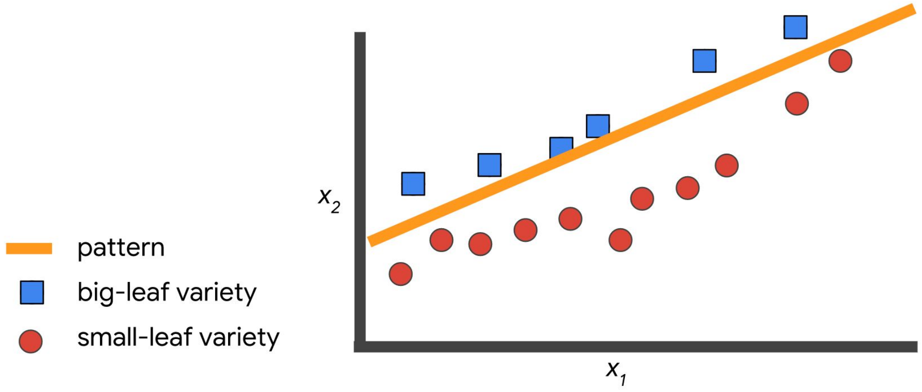
Classification Deep-Dive

Classification Problems

Classification: predicting categorical labels (e.g. plant type, hair color, image category)

Easiest case: binary classification, with only two labels (e.g., cat vs dog)

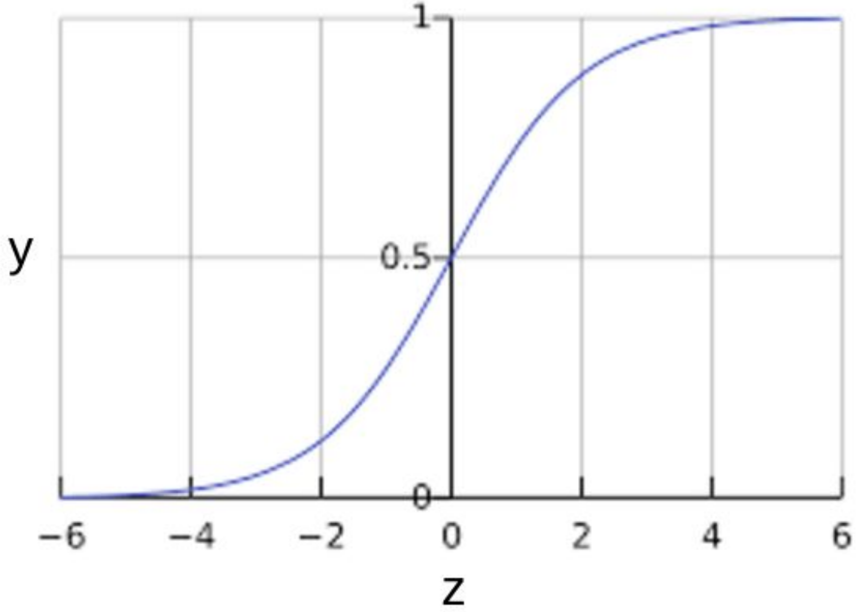
Output: predicted (probability of) label → probabilities are turned into label-predictions via **thresholding**



Example Algorithms

Logistic Regression:
Supports binary and multiclass classification

Tree-based models:
Also support regression (see next section), range from
Decision Trees to
Random Forests
and gradient-boosted Trees like **LightGBM.**



Model performance: Confusion Matrix

Ideally, we want high values in the green cells and low values in the red cells.

But: often, we have consider trade-offs between those four outcomes.

		Predicted: True	Predicted: False
Actual	True	True Positives True label = 1, predicted label = 1	False Positives True label = 0, predicted label = 1
	False	False Negatives True label = 1, predicted label = 0	True Negatives True label = 0, predicted label = 0

Model Performance: Accuracy

Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right. Formally, accuracy has the following definition:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Example Data

		Predicted: True	Predicted: False
Actual	True	True Positives True label = 1, predicted label = 1 1	False Positives True label = 0, predicted label = 1 1
	False	False Negatives True label = 1, predicted label = 0 8	True Negatives True label = 0, predicted label = 0 90

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{1+90}{1+90+1+8} = 0.91$$

*Which problem could we have with **Accuracy** as a metric?*

Model Performance: Precision and Recall

Precision

What proportion of positive identifications was actually correct?

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall

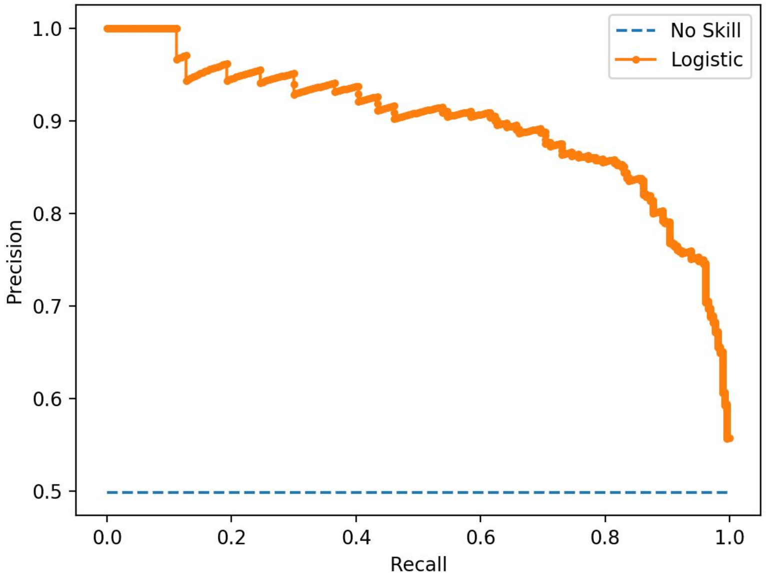
What proportion of actual positives was identified correctly?

$$\text{Recall} = \frac{TP}{TP + FN}$$

Model Performance: Precision and Recall

Both metrics need to be examined to fully evaluate the effectiveness of a model.

Usually, they are in tension: improving precision reduces recall and vice versa.



https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/

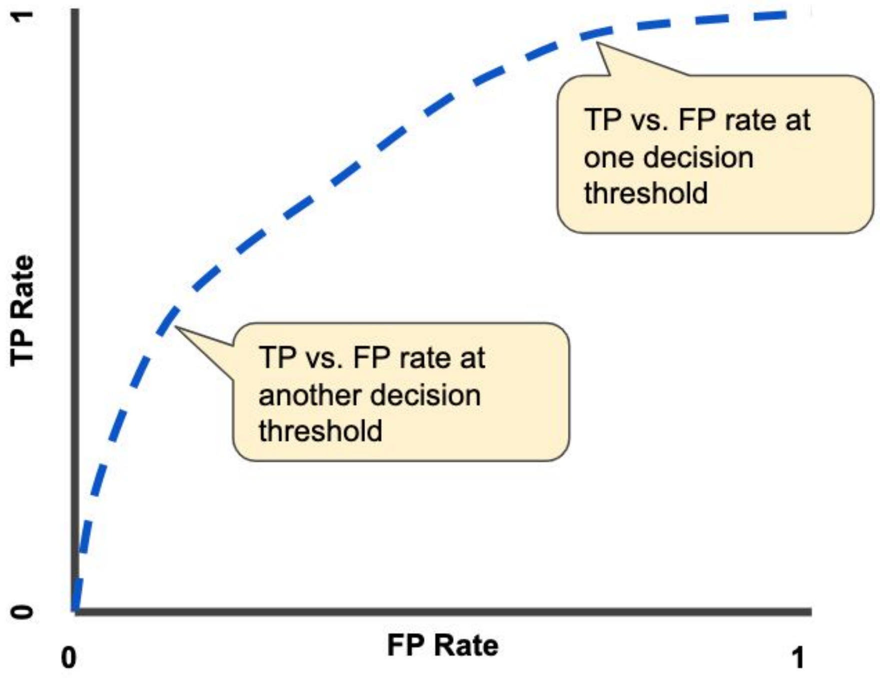
Model Performance: ROC curve and AUC

Receiver operator characteristic (ROC) curve:
Performance of a classification model at all classification thresholds, by plotting **True Positive Rate** (TPR) and **False Positive Rate** (FPR)

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

Area under the ROC Curve (AUC)
measures the **entire two-dimensional area** underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1)



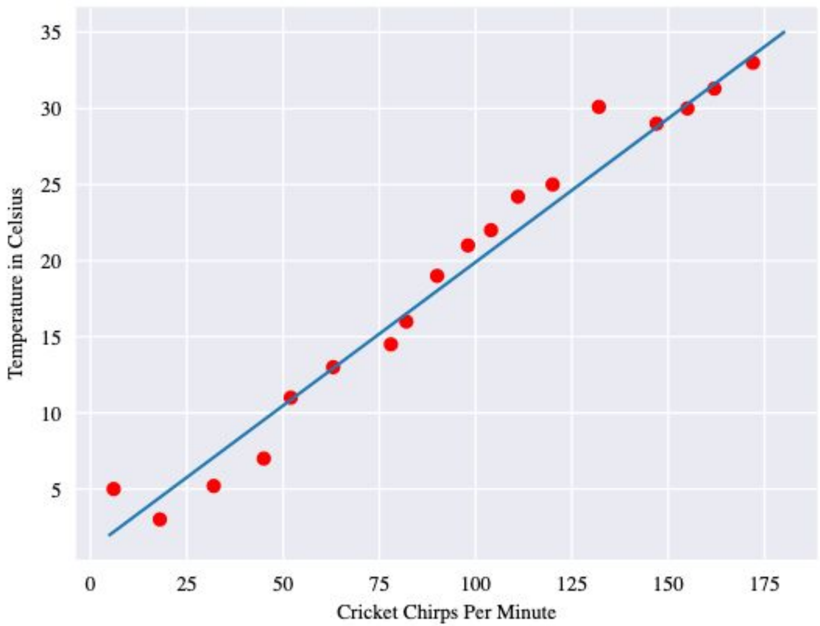
Regression Deep-Dive

Regression Problems

Regression: predicting continuous target values (e.g. temperature, costs, height)

Can be formulated as **linear** or **non-linear** models

Output: (usually) predicted **target values**



Common Algorithms

Linear Regression

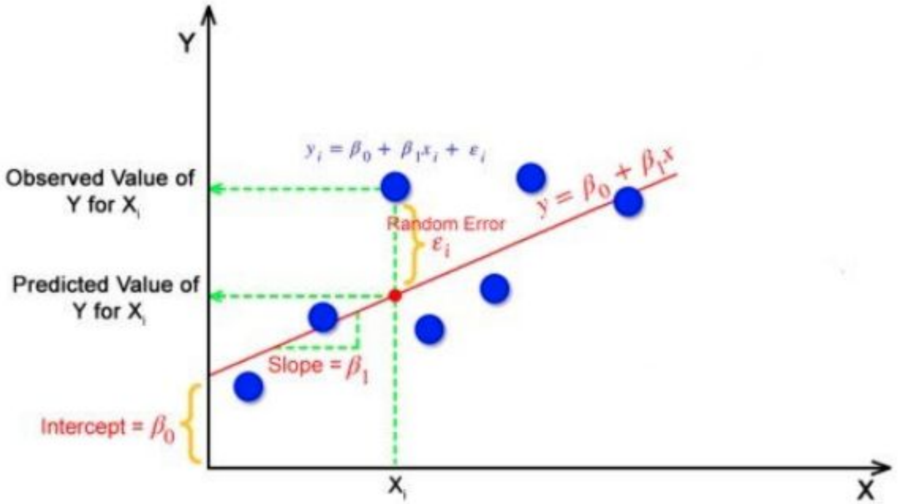
Estimate target value with a linear function of intercept and other predictors

Tree based models

Random forest regression, Gradient boosted regression

Neural Networks

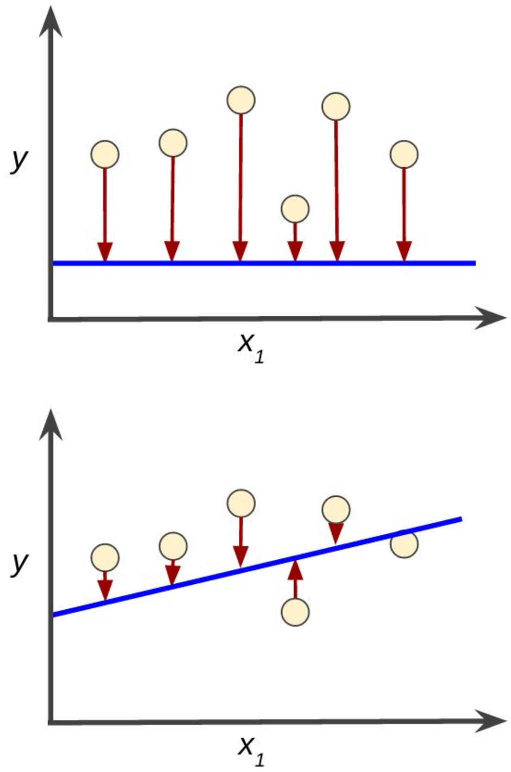
Deep Neural Network: Train a network with multiple hidden (transformation) layers to predict target value



Model Performance | Minimizing Loss

Goal: find model parameters so that predicted values are most similar to actual values, i.e. that **minimize the loss**.

$$MSE = \frac{1}{N} \sum_{(x,y)\in D} (y - prediction(x))^2$$

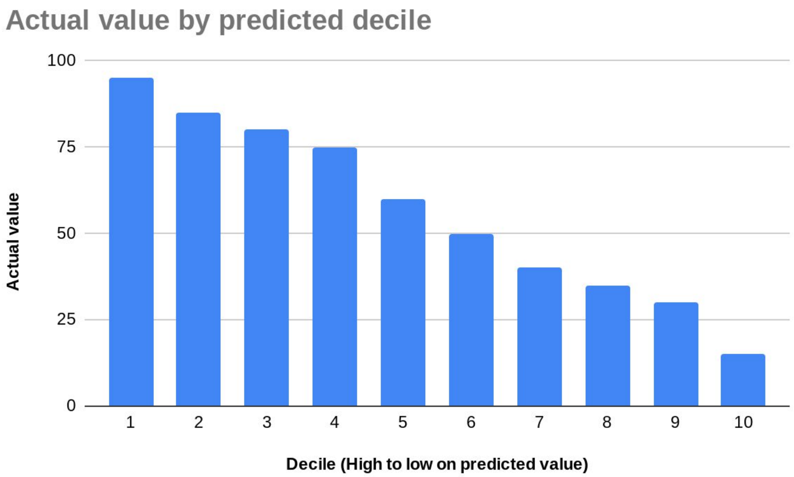


The arrows represent loss.
The blue lines represent predictions.

Model Performance | Other Metrics

Decile Lift Chart:

Average of actual value within each predicted decile



Mean Average Percentage of Error:

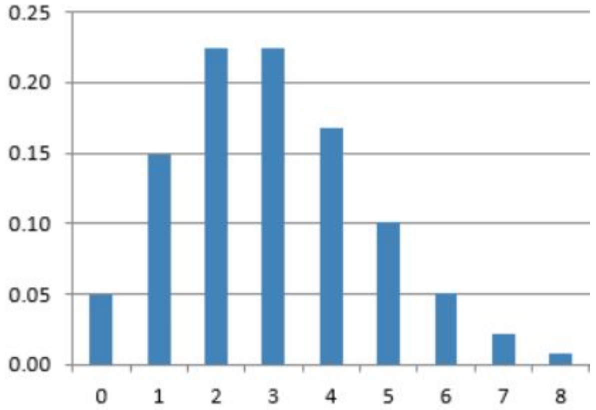
$$\text{MAPE} = \frac{100\,\%}{n} \sum_{t=1}^{n} \left| \frac{Actual\ value - Predicted\ value}{Actual\ value} \right|$$

Measure of prediction accuracy in forecasting model

Special Regression Cases

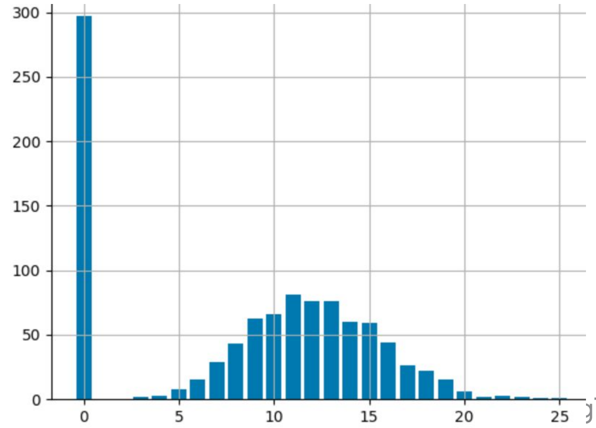
Poisson regression:

Poisson regression is applied when response variable are count data
Example: # of ER visits, # of car accident each year



Tweedie Loss/Zero-inflation regression:

Zero-inflated model is applied when you data contain excess zero-count data



Thank you!

Questions?

Further Resources

Good Resources for Data Science and ML

Courses:

[Machine Learning Crash Course | Google Developers]

Code, Models, Frameworks (usually with examples):

Scikit-learn

https://keras.io/

Blogs:

https://towardsdatascience.com/

Books:

1 An Introduction to Statistical Learning

The Elements of Statistical Learning