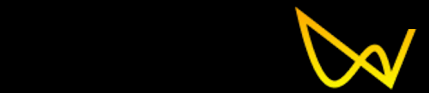
DATA PREPROCESSING

CxC Workshop – February 10, 2024

TREVOR YU & CARTER DEMARS



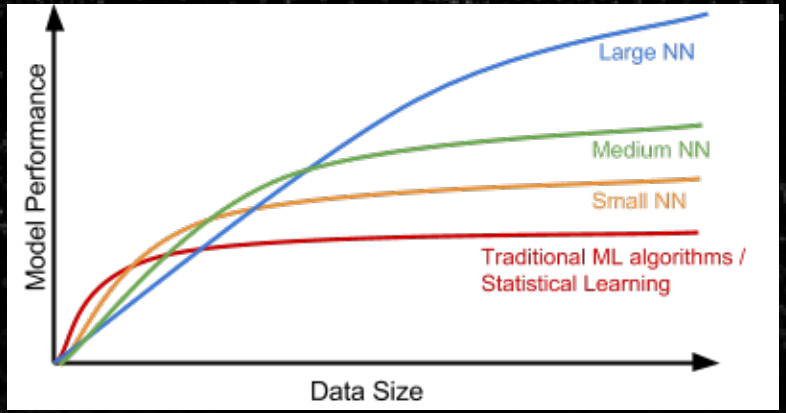
TODAY'S AGENDA

- Learn to perform basic exploratory data analysis (EDA) and data visualization
- Identify outliers, handle missing values, and perform other common data operations such as normalization
- Understand the intuition behind various preprocessing techniques for both categorical and continuous features to prepare for classification tasks

ML PRACTITIONERS NEED GOOD DATA

- Most machine learning applications require clean data in the form of vectors
- Most models expect data inputs to be passed in a consistent way
- Different modalities of data (tabular data, images, text) require different techniques

- For some learning algorithms, such as neural networks, increasing the size of the training dataset can have a huge impact on the effectiveness of the algorithm.
- On the other hand, insufficient data or poor data quality will often result in an underperforming model



USEFUL PYTHON LIBRARIES

NUMPY (NUMERICAL PYTHON)

- Built on top of C
- Library for working with arrays and matrices and python, with the associated high-level functions to operate on these arrays

MATPLOTLIB

- Open-source plotting library that closely resembles plotting in MATLAB

USEFUL PYTHON LIBRARIES

PANDAS

- Open-source data analysis & manipulation tool
- Reads data into a series/DataFrame
- Features for dealing with missing data, changing data format, aggregating data, slicing, sorting and applying transformations

SCIKIT-LEARN

- Tools for data analysis and machine learning
- Highly popular library with classification, regression, and clustering algorithms
- Built on top of NumPy and SciPy

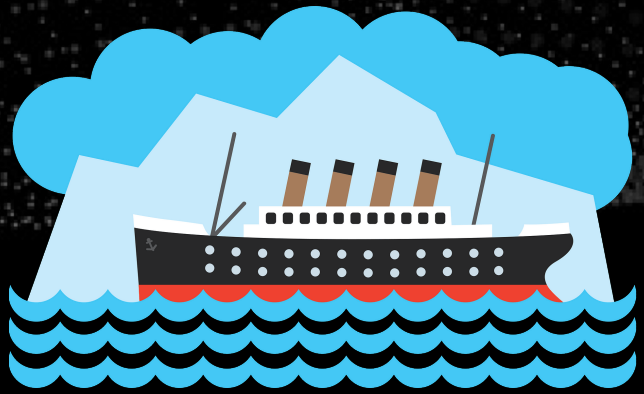
WAT.ai

CODE-ALONG ACTIVITY

Link to Colab Notebook

THE CHALLENGE: MACHINE LEARNING FROM DISASTER

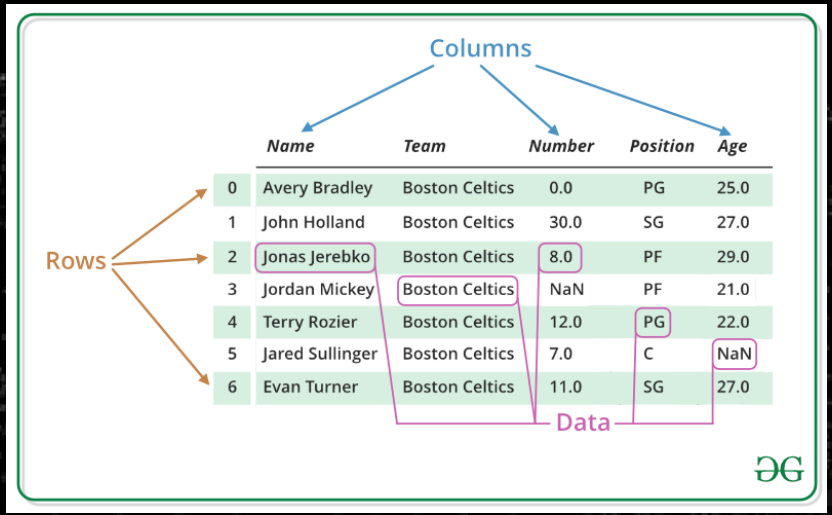
On April 15, 1912, during her maiden voyage, the widely considered "unsinkable" RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren't enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.



In this challenge, we ask you to build a predictive model that answers the question: what sorts of people were more likely to survive? You'll be using passenger data, such as name, age, gender, and socio-economic class.

DATAFRAME BASICS

- A DataFrame is a 2-dimensional tabular data representation
- Uniquely labeled axes, called rows and columns
- Each row is typically one collected data point
- Each column contains values of a "feature" across many examples

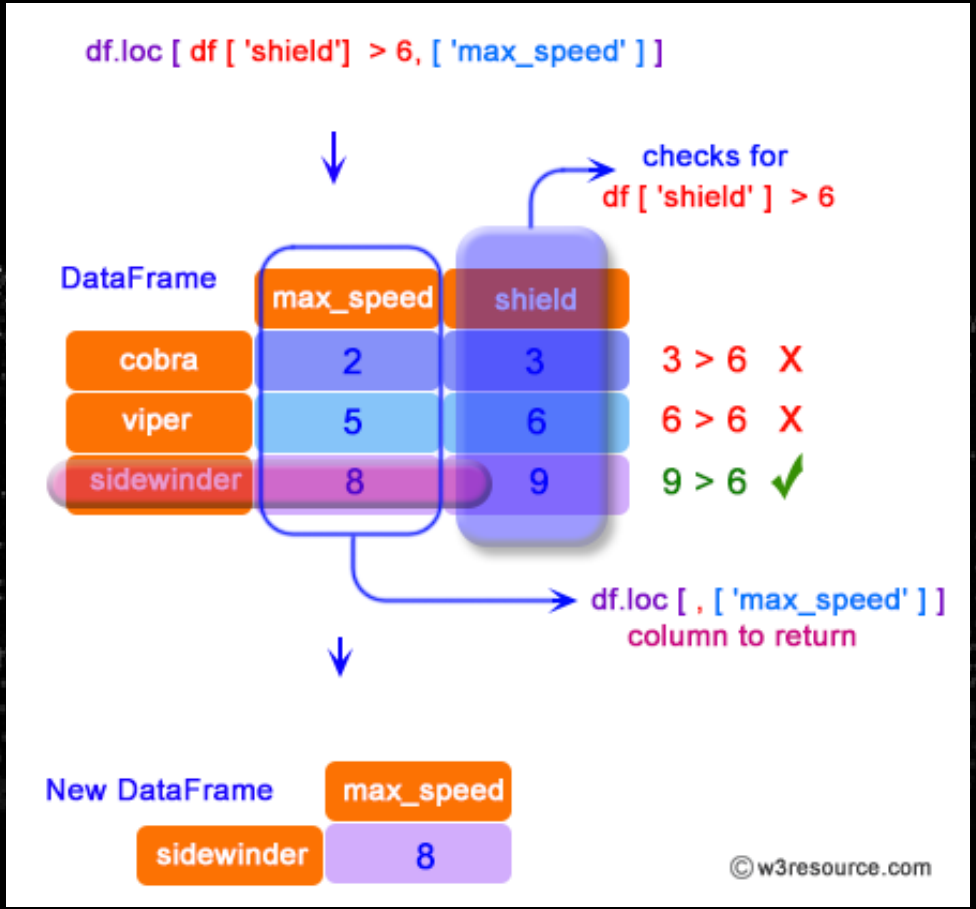


DATAFRAME BASICS

- By convention, refer to dataframes as `df` in code
- `df.index, df.columns, df.values` can access these aspects of the dataframe
- `df.shape` gives the number of (rows, columns)

VIEWING SUBSETS

- Common operations:
  - Selecting columns
  - Selecting rows by index
  - Selecting rows by condition
  - Sorting results by value
- Selections do not happen in-place, results must be assigned to variable to persist



HANDLING MISSING VALUES

TYPES OF MISSING VALUES

Missing Not At Random - when a value is missing for a reason related to the true value. (Ex: if a survey respondent chooses not to disclose their income, this could be because they have an abnormally high or low income)

Missing at Random - when a value is missing for a reason related to another observed variable. (Ex: many age values are missing for survey respondents of a particular gender)

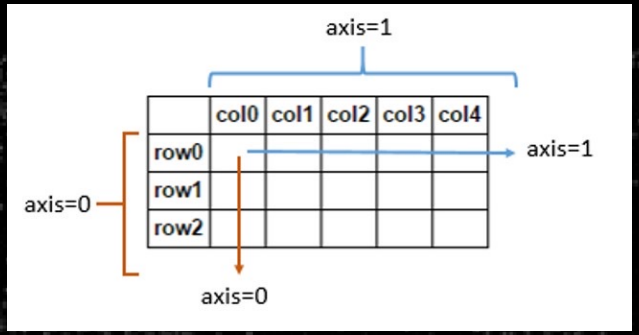
Missing Completely at Random - when there's no patterns in the missing values.

HANDLING MISSING VALUES

DELETION

Column deletion: removing a column that has too many missing values and is non-essential for your model

Row deletion: removing rows with missing values, ideally if the missing values are Missing At Random, to avoid biasing your model



IMPUTATION

- Fill missing values with their defaults (empty string, zero, etc...)
- Fill missing values with the mean, median, or mode
- Backward or forward fill

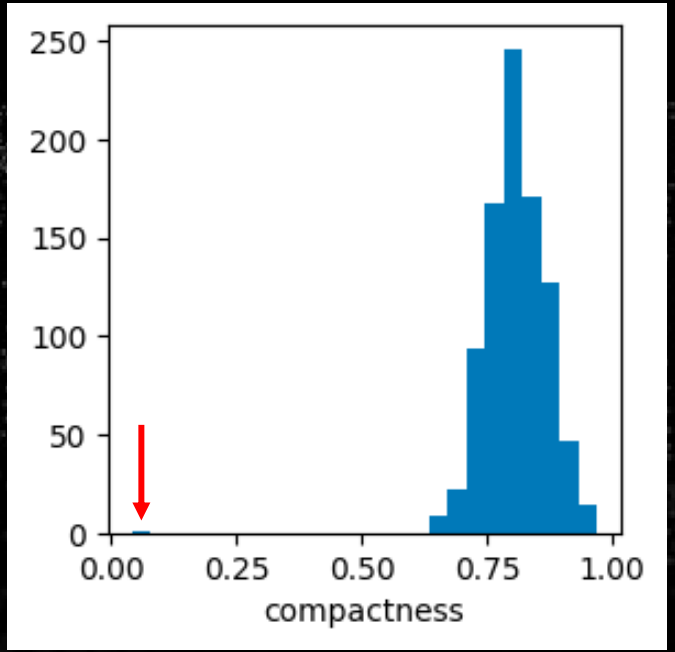
FEATURE ENGINEERING

- Feature engineering is the practice of processing raw data to extract more informative characteristics for our models
  - Sometimes, raw features are uninformative
  - Domain knowledge on useful features

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

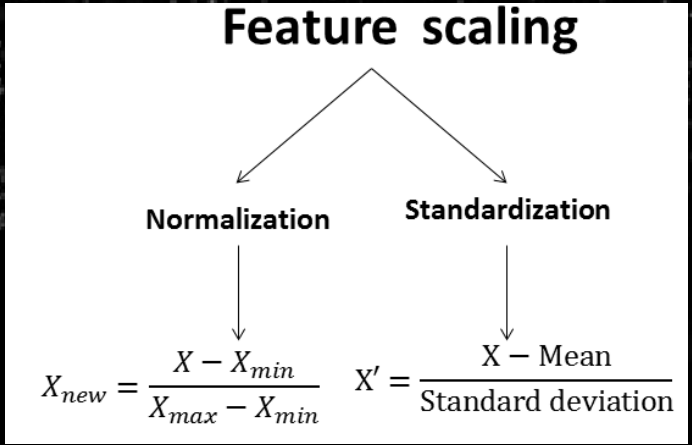
OUTLIER REMOVAL

- Outliers are data points that deviate significantly from the mean distribution of the data
  - Often represent incorrect measurements
  - Domain knowledge is often required to interpret the meaning of outliers
- Heuristic methods like thresholds or statistical methods like Z-scores
- Removing too many data points could adversely affect a model's ability to generalize



FEATURE NORMALIZATION

- Many ML models assume that the input data (as vectors) are roughly normally distributed with 0 mean and unit standard deviation
  - Sometimes, scaling data between [-1, 1] is also used
- Model performance generally improves when features are normalized
- Typically, this is the last step before passing the data to a model



HANDLING CATEGORICAL DATA TYPES

- **Ordinal categorical variables** have discrete categories whose order matters (ex: small, medium, and large)
- **Nominal categorical variables** have discrete categories without order, so concepts such as the mean have no interpretation (ex: gender)

id	color
1	red
2	blue
3	green
4	blue

One Hot Encoding

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0