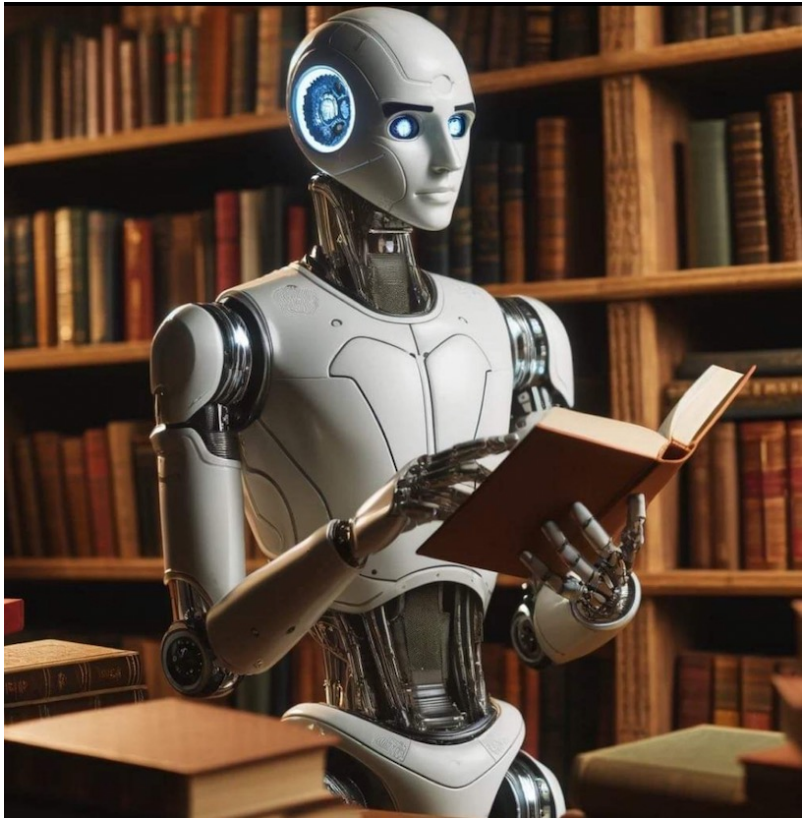
Introduction to Machine Learning (Part III)

May 23, 2024

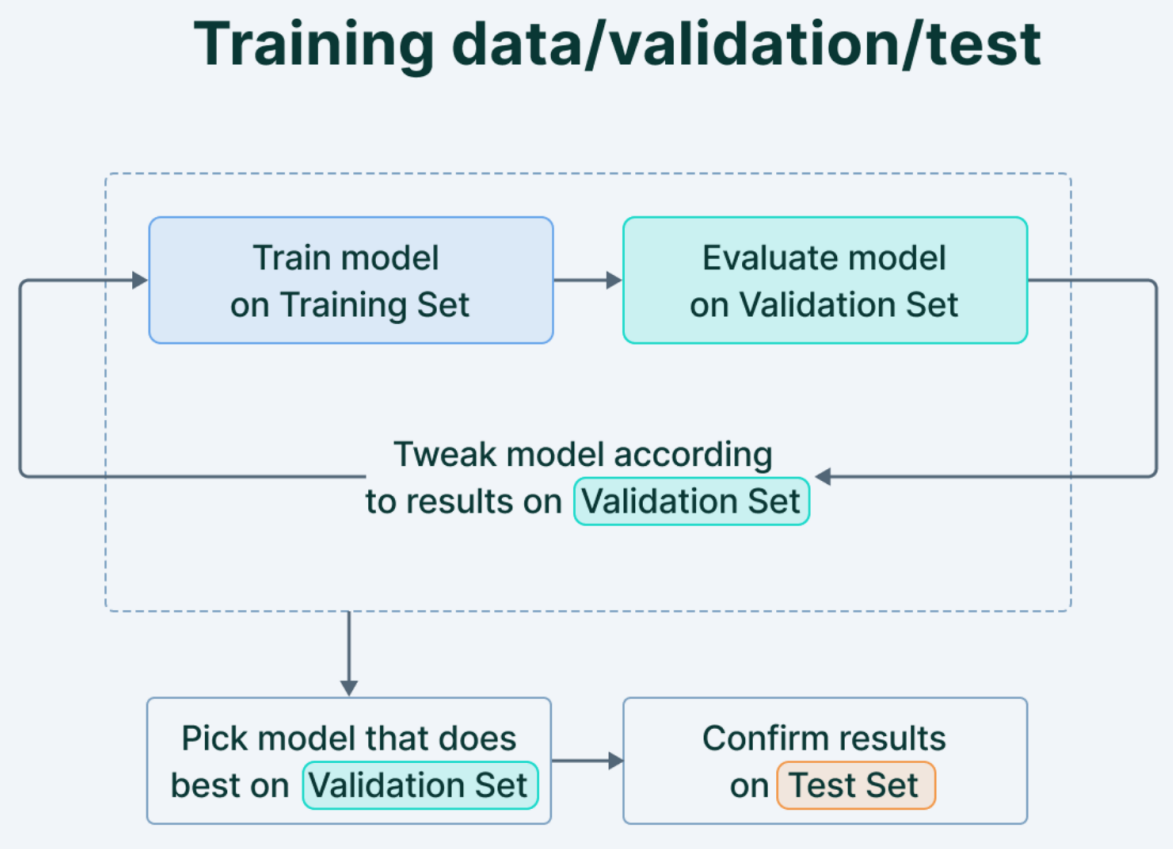
Overview of tonight's lesson: ML 3

1. ML regression project!
2. Exploratory data analysis
3. Preprocessing pipeline
4. Random forest model



ACENET
accelerate discovery

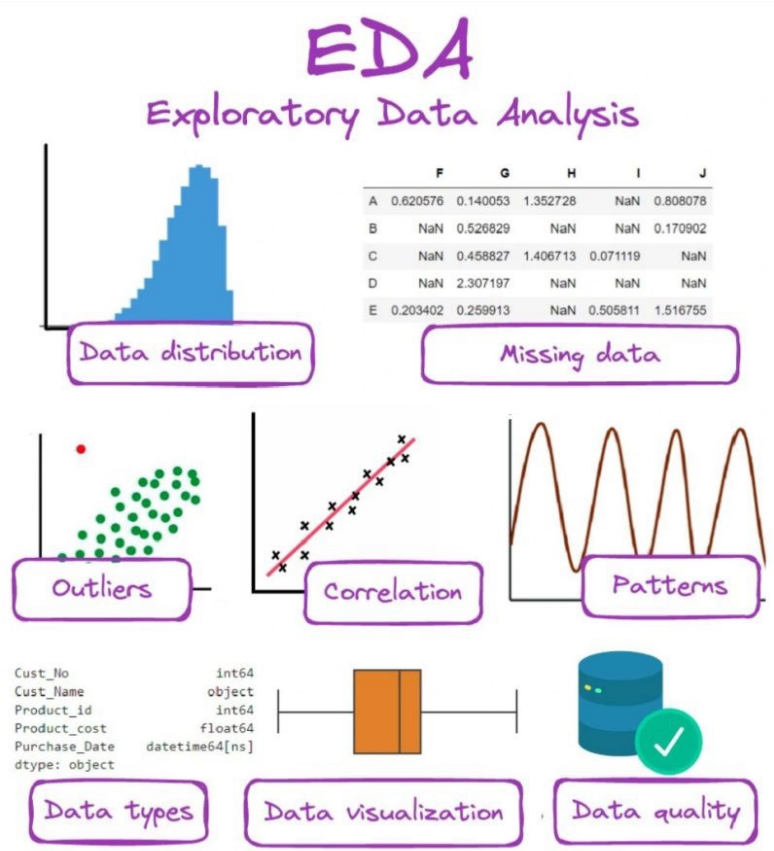
Splitting data:



ACENET
accelerate discovery

Exploratory data analysis:

- Understand the structure and content of your data
- Discover potential outliers, missing values, inconsistencies
- Prepare data for ML analysis



ACENET
accelerate discovery

Preprocessing data:

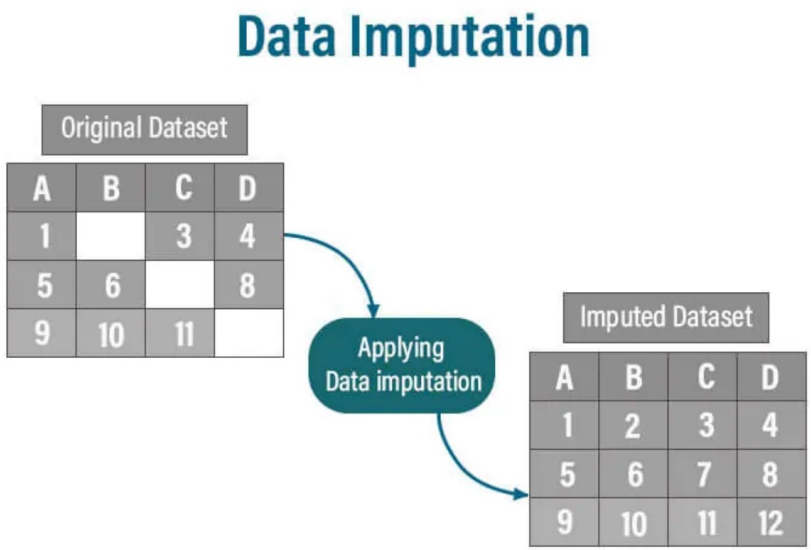
- Transform raw data into a clean and usable format suitable for analysis or machine learning models

Preprocessing data:

- Imputing the data (filling in the NaN values)
- Scaling the numerical data so that all of the features are in the same range
- Encoding the categorical variables

Imputing the data:

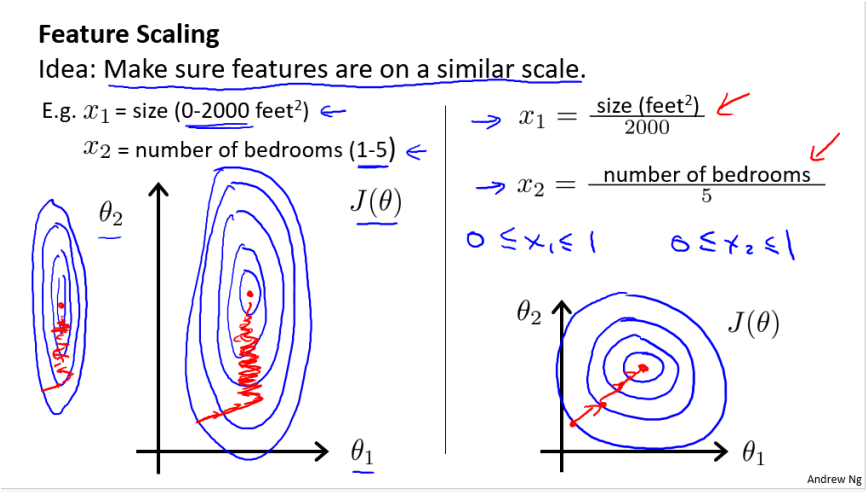
- This replaces the missing values based on some method
- Methods can include using the mean/median/mode, KNN, or some other model
- Can also do this with categorical variables
- Imputation introduces some assumptions about the data!



ACENET
accelerate discovery

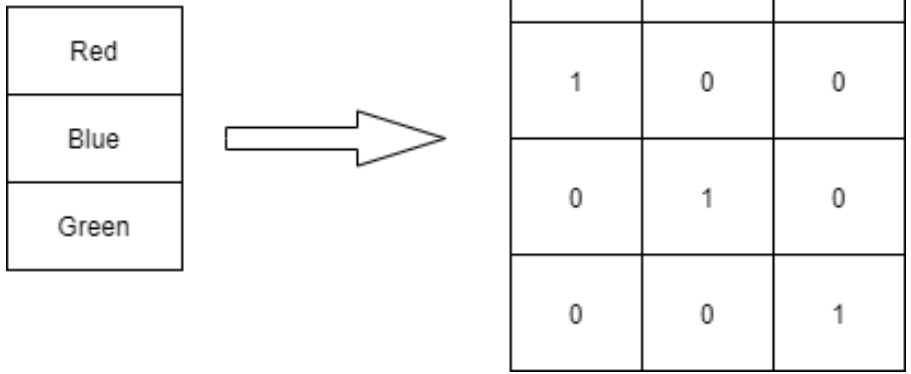
Scaling the numerical data :

- Scale features to be in a similar range
- For example, can set the mean and standard deviation to zero



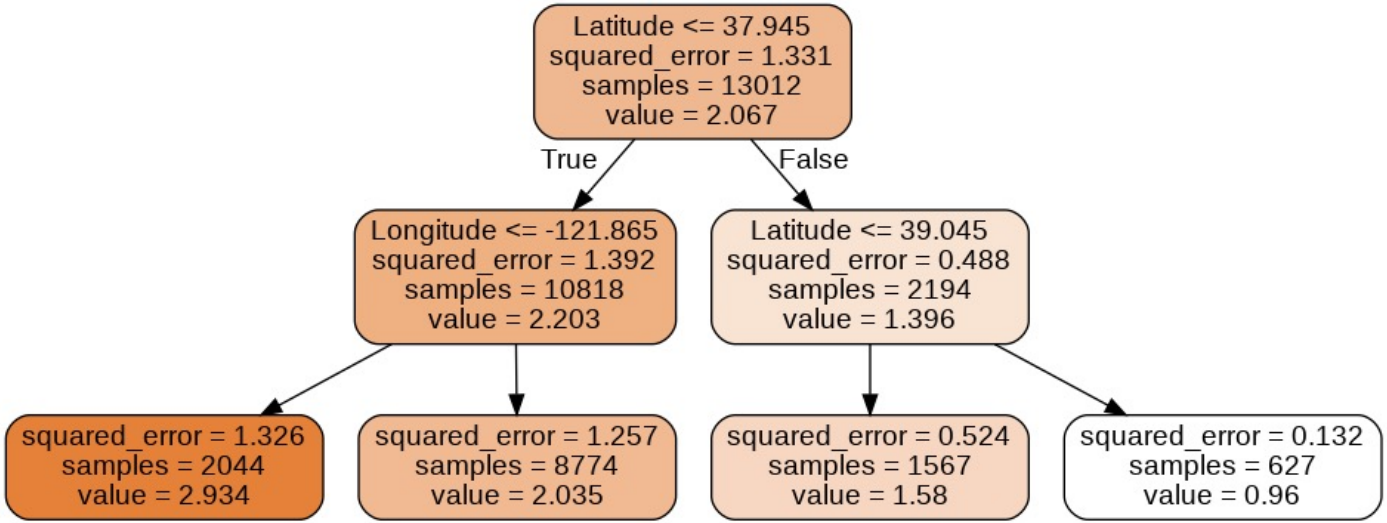
Encoding:

- Encoding is the process of transforming categorical data (data with labels) into numerical representations.
- Need to think about: are the categories ordered or non-ordered?



Trees and forests

- A decision tree in ML is essentially a set of if and else statements to subset data
- A random forest is a collection of decision trees where the if else statements are slightly different to get a more robust estimate



Two important metrics for regression

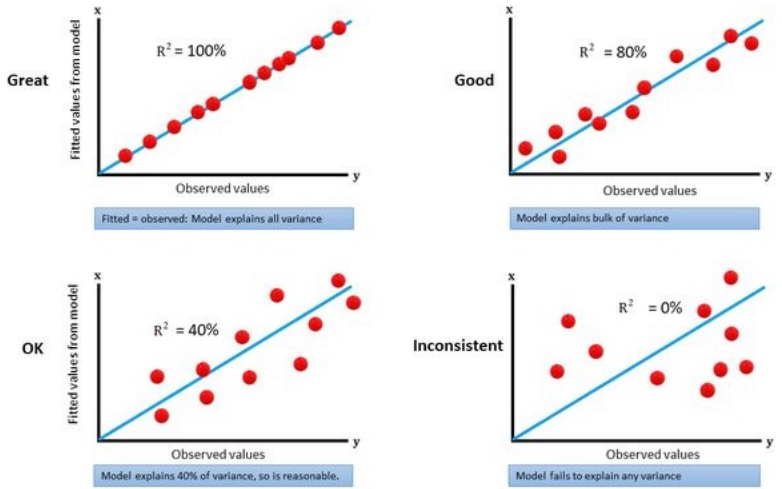
- Root mean squared error
- R square

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

$\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are predicted values

$y_1, y_2, \ldots, y_n$ are observed values

n is the number of observations



ACENET
accelerate discovery

NEXT STEPS FOR ML:

- ML&HPC!

ACENET
accelerate discovery

ace-net.ca

info@ace-net.ca

certificate@ace-net.ca