DATA PRE-PROCESSING AND DATA WRANGLING

TODAY'S TOPICS

- Data Pre-Processing
- Data Wrangling

WHAT IS DATA PRE-PROCESSING ?

- Our dataset usually contains different types of data in different formats. But a machine can only understand mathematical data.
- Sometimes data values are in different ranges which can cause some problems in learning. So they need to be converted into some fixed ranges.
- Data preprocessing is nothing but converting these types of data into some more valid numerical form for a Machine Learning Algorithm is known as Data Preprocessing.
- Data Preprocessing is one of the most important step Machine Learning. This is the first step in a Machine Learning pipeline. Data preprocessing consists of different parts. In each part, we apply some modifications to our data so that we can use the data.

WHAT IS DATA WRANGLING ?

- Data wrangling is the process of cleaning, structuring and enriching raw data into a desired format for better decision making in less time.
- Data has become more diverse and unstructured, demanding increased time spent culling, cleaning, and organizing data ahead of broader analysis.
- At the same time, with data informing just about every business decision, business users have less time to wait on technical resources for prepared data.

STEPS IN DATA WRANGLING

- There are typically six iterative steps that make up the data wrangling process as:
  - Discovering
  - Structuring
  - Cleaning
  - Enriching
  - Validating
  - Publishing

DISCOVERING

- Before we can dive deeply, we must better understand what is in our data, which will inform how we want to analyze it.
- How we wrangle customer data, for example, may be informed by where they are located, what they bought, or what promotions they received.

STRUCTURING

- This data wrangling step means organizing the data, which is necessary because raw data comes in many different shapes and sizes.
- A single column may turn into several rows for easier analysis. One column may become two.
- Movement of data is made for easier computation and analysis.

CLEANING

- What happens when errors and outliers skew your data? You clean the data.
- What happens when state data is entered as CA or California or Calif.? You clean the data.
- Null values are changed and standard formatting implemented, ultimately increasing data quality, which is the goal of data wrangling.

ENRICHING

- Here we take stock in our data and strategize about how other additional data might augment it.
- Questions asked during this data wrangling step might be: what new types of data can I derive from what I already have or what other information would better inform my decision making about this current data?

VALIDATING

- Validation rules are repetitive programming sequences that verify data consistency, quality, and security.
- Examples of validation include ensuring uniform distribution of attributes that should be distributed normally (e.g. birth dates) or confirming accuracy of fields through a check across data.

PUBLISHING

- Analysts prepare the wrangled data for use downstream – whether by a particular user or software – and document any particular steps taken or logic used to wrangle said data.
- Data wrangling gurus understand that implementation of insights relies upon the ease with which it can be accessed and utilized by others.
- The data is now ready for analytics.

USE OF PANDAS

- Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data -
  - Load
  - Prepare
  - Manipulate
  - Model
  - Analyze.

HANDLING MISSING DATA

- There are some specific rows in our dataset which doesn't contain any information for some specific columns.
- As we know Machine Learning algorithms are just mathematical equations. So we cannot send these missing values in Machine Learning algorithm.
- Missing data can be handled commonly in two ways-
  - Removing data
  - Imputation

REMOVING NULL DATA

- In many cases, the solution is just removing the specific row. If we use pandas, then it is very simple. We just need to use one pandas methods called dropna().
- Suppose we have a panda dataframe df, which contains some missing values. So a simple code to delete those specific rows is given below-
  - df = df.dropna(axis = 1)
- But this is not the best solution. Sometimes delete a row can delete some important information of other columns. So there is a better way called 'Imputation'.

IMPUTATION

- In imputation, instead of deleting a row, we fill the missing values by some other values.
- The imputed values might be median, mean,0 etc.
- The imputed values might not be exactly the same but it is very accurate to the right value.
- Syntax : data['Age'] = data['Age'].fillna(data['Age'].mode()[0])

Z-SCORE, STANDARD SCALAR & MIN-MAX SCALAR

- Z-scores are expressed in terms of standard deviations from their means. Resultantly, these z-scores have a distribution with a mean of 0 and a standard deviation of 1.
- The idea behind Standard Scaler is that it will transform your data, such that the distribution will have a mean value of 0 and a standard deviation of 1. The mathematical formula for Standard Scaler is given below: ***(xi–mean(x))/stdev(x)***
- The Min-Max Scalar uses the following formula for calculating each feature: **(xi–min(x))/(max(x)–min(x))**
- It uses the min and max values, so it's very sensitive to outliers.
- It can be used as an alternative to Standard Scaler when data is not normally distributed.

DATA ENCODING

- We know that Machine Learning algorithms require data in numerical form.
- But many times, datasets contain some features in some other form. So we need to convert these value into numerical form.
- Scikit-Learn provides different encoding methods for Data Encoding.

LABEL ENCODING

- To understand Label Encoding, first, let's assume a dataset contains three columns weight, height, and gender.
- Now in this dataset, the gender column is not in numerical form. That means we need to convert it into some type of numerical form. To achieve that, we can use Label Encoding.
- We know that the gender columns contains three unique values, Male, Female and Third Gender. If we apply Label Encoding algorithm on this column, then it replaces the values by 0 , 1 and 2 respectively. (Male = 0 , Female = 1 and Third Gender=2).
- But a disadvantage with Label Encoding is that the machine will think that 0>1>2 (i.e. male>female>third gender). But in actual this is not so. So we went to One Hot Encoding.

ONE HOT ENCODING

- One Hot Encoding takes a column which has categorical data, which has been label encoded and then splits the column into multiple columns.
- The numbers are replaced by 1s and 0s, depending on which column has what value.

ALGORITHMS SENSITIVE TO OUTLIERS

- Linear Regression - Yes
- Logistic Regression - Yes
- SVM - Somewhat
- Decision Tree - No
- Random Forest - No
- k-Means Clustering – Yes
- kNN – Somewhat

FEATURE ENGINEERING

- *The features you use influence more than everything else the result. No algorithm alone, to my knowledge, can supplement the information gain given by correct* ***feature engineering*** *— Luca Massaron*
- Feature Engineering efforts mainly have two goals:
  - Preparing the proper input dataset, compatible with the machine learning algorithm requirements.
  - Improving the performance of machine learning models.

THANK YOU