Thanks for joining we will start at 4pm 🕓

Developer Student Clubs

Today's Agenda

Understanding the Big Mart Sales Prediction Dataset

- Understanding the dataset and problem statement.
- Data Cleaning and formatting.
- Feature Encoding. (Ordinal/Nominal)
- Feature Scaling. (MinMax)
- Train Test Split.
- Modeling and predictions.
- Q/A Session.

Developer Student Clubs

Google Developer Students Club
Institute of Business Administration

Data Science for Everyone

NABEEL AHMED

ZOHAIR ABBAS

MUHAMMAD ARHAM

ABDUL REHMAN ABBASI

ZOHAIB AZAM

ABDUL AHAD IMTIAZ

Developer Student Clubs

WHO AM I?

- Technical Lead - GDSC IBA
- Google certified Data Analyst
- SAP Functional Analyst - 3STEM
- Deputy Team Lead - TCF ADP
- ML/DL Enthusiast
- Loves to bring useful insights out of Data.

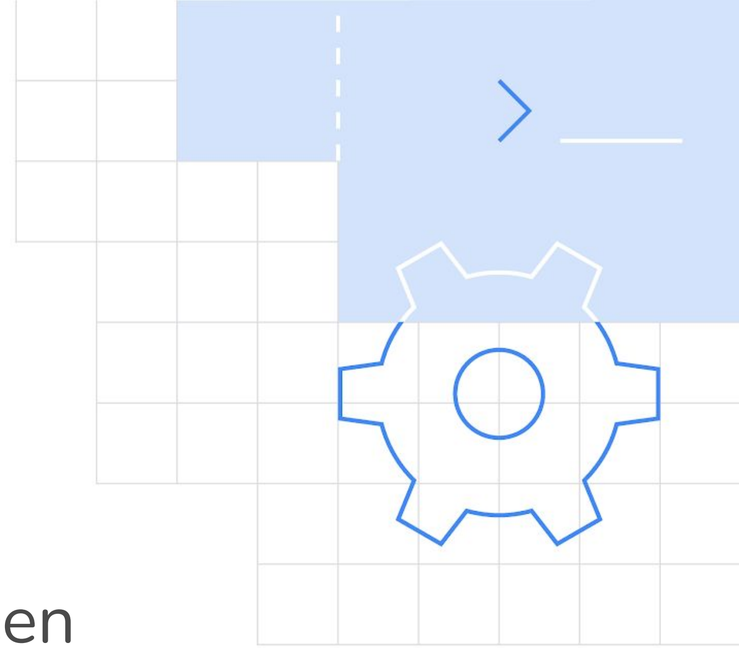
Google Developers

Developer Student Clubs

Big Mart Sales Prediction

2013 sales data for 1559 products across 10 stores in different cities

Challenge ?

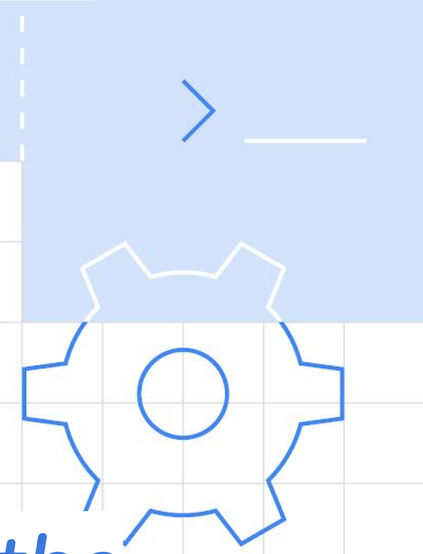


-The data scientists at BigMart have collected **2013** sales data for **1559** products across **10 stores** in different cities.

-Also, certain attributes of each product and store have been defined.

-The aim is to build a **Predictive Model** and predict the sales of each product at a particular outlet.

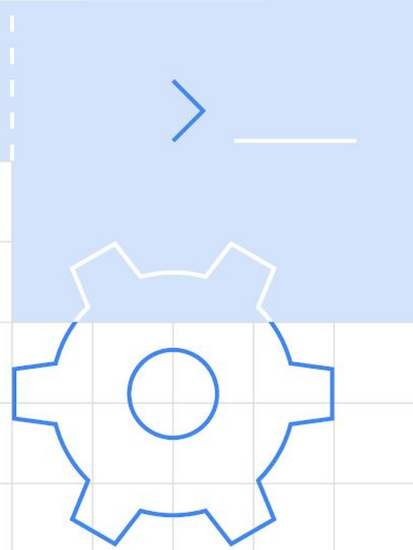
Developer Student Clubs



In this challenge, we ask you to build a predictive model that answers the question: "What would be the sale of each product at a particular outlet given its properties?"

Developer Student Clubs

Understanding of Data



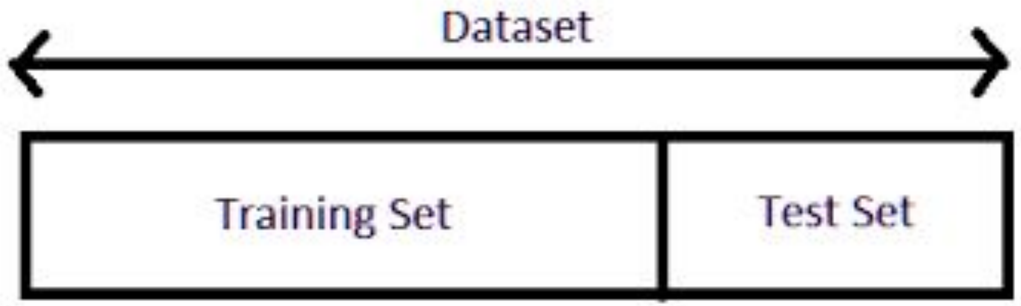
Developer Student Clubs

Features:

Variable	Description
Item_Identifier	Unique product ID
Item_Weight	Weight of product
Item_Fat_Content	Whether the product is low fat or not
Item_Visibility	The % of total display area of all products in a store allocated to the particular product
Item_Type	The category to which the product belongs
Item_MRP	Maximum Retail Price (list price) of the product
Outlet_Identifier	Unique store ID
Outlet_Establishment_Year	The year in which store was established
Outlet_Size	The size of the store in terms of ground area covered
Outlet_Location_Type	The type of city in which the store is located
Outlet_Type	Whether the outlet is just a grocery store or some sort of supermarket
Item_Outlet_Sales	Sales of the product in the particular store. This is the outcome variable to be predicted.

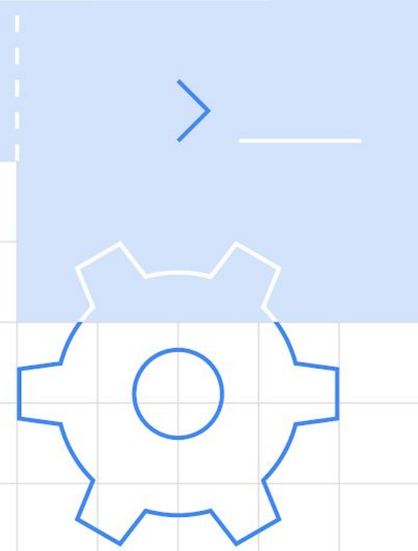
What is Train/Test data

- Train/Test is a method to measure the accuracy of your model.
- It is called **Train/Test split** because you split the the data set into two sets: a training set and a testing set.
- You ***train*** the model using the training set.
- You ***test*** the model using the testing set.
- ***Train*** the model means *create* the model.
- ***Test*** the model means test the accuracy of the model.



Developer Student Clubs

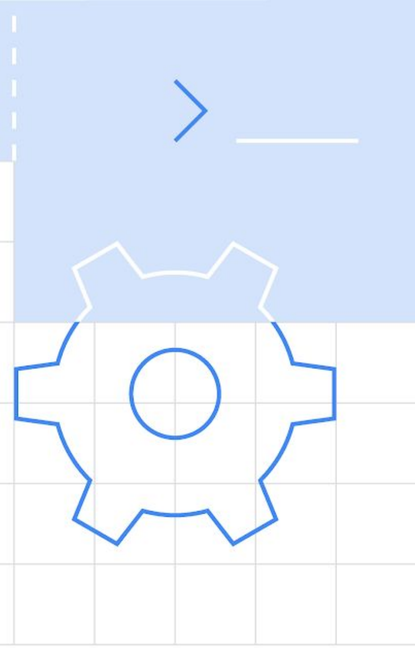
Machine Learning Techniques



1) **Feature Selection**: is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data.
2) **Feature Encoding**: Machine learning models can only work with numerical values. For this reason, it is necessary to transform the categorical values of the relevant features into numerical ones.
3) **Feature Scaling**: Feature Scaling is a technique to standardize the independent features present in the data in a fixed range.

Developer Student Clubs

Feature Selection



- When building a machine learning model in real-life, it’s almost rare that all the variables in the dataset are **useful** to build a model. Adding redundant variables **reduces** the generalization capability of the model and may also reduce the overall **accuracy** of a classifier.
- The goal of feature selection in machine learning is to find the **best set of features** that allows one to build useful models of studied phenomena.

  A. Filter methods (Chi-Square Test, Fisher’s Test, Pearson’ Correlation..)

  B. Wrapper methods

  C. Embedded methods

  D. Hybrid methods

- Developer Student Clubs

Feature Encoding

- **Ordinal Data:** The categories have an inherent order
- **Nominal Data:** The categories do not have an inherent order

	Degree
0	High school
1	Masters
2	Diploma
3	Bachelors
4	Bachelors
5	Masters
6	Phd
7	High school
8	High school

	Degree
0	1
1	4
2	2
3	3
4	3
5	4
6	5
7	1
8	1

Index	Animal
0	Dog
1	Cat
2	Sheep
3	Horse
4	Lion

One-Hot code

Index	Dog	Cat	Sheep	Lion	Horse
0	1	0	0	0	0
1	0	1	0	0	0
2	0	0	1	0	0
3	0	0	0	0	1
4	0	0	0	1	0

Developer Student Clubs

Feature Scaling

I’m sure most of you must have faced this issue in your projects or your learning journey. For example, one feature is entirely in **kilograms** while the other is in **grams**, another one is **liters**, and so on. How can we use these features when they vary so vastly in terms of what they’re presenting?

	Student	CGPA	Salary '000
0	1	3.0	60
1	2	3.0	40
2	3	4.0	40
3	4	4.5	50
4	5	4.2	52

	Student	CGPA	Salary '000
0	1	-1.184341	1.520013
1	2	-1.184341	-1.100699
2	3	0.416120	-1.100699
3	4	1.216350	0.209657
4	5	0.736212	0.471728

Cont.

- **Normalization**: is a scaling technique in which values are shifted and rescaled so that they end up ranging between **0** and **1.** It is also known as **Min-Max** scaling.

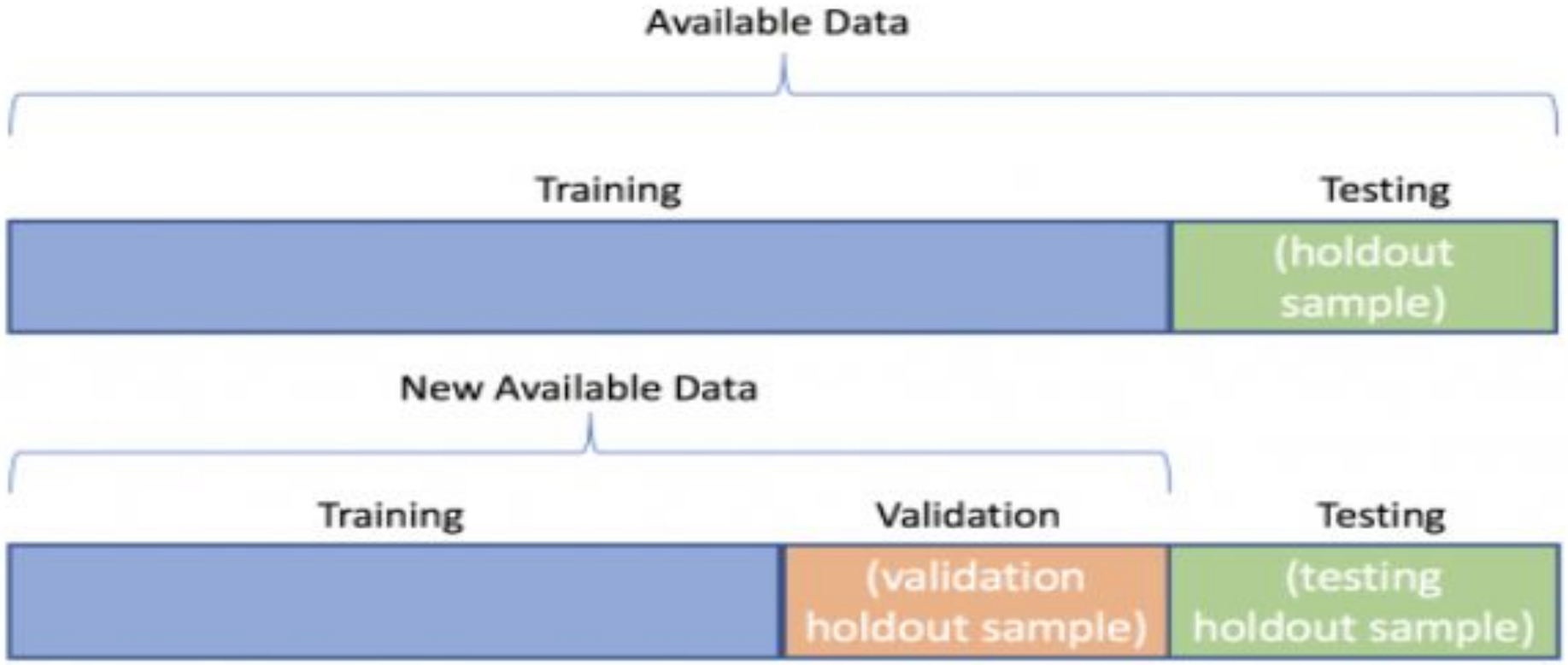
$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- **Standardization:** is another scaling technique where the values are centered around the mean with a unit standard deviation.

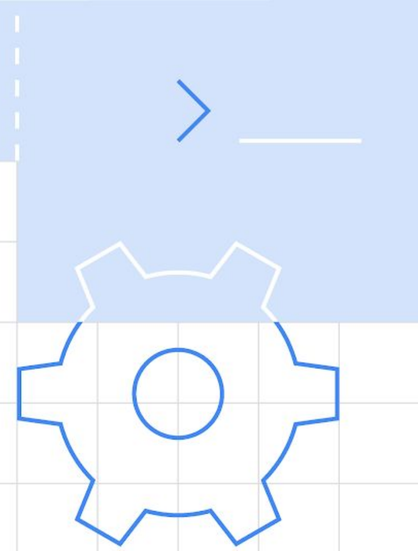
$$X' = \frac{X - \mu}{\sigma}$$

How Kaggle Works?

- We have given Train and Test data.



Supervised Learning Algorithms

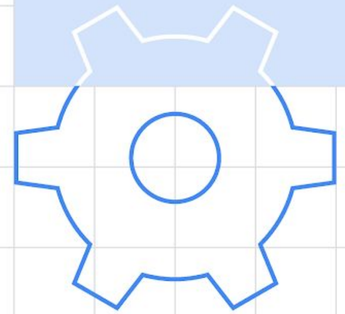


- Linear Regression
- Logistic Regression
- K Nearest Neighbor - KNN
- Decision Trees
- Random Forest
- Naive Bayes’
- Support Vector Machine - SVM

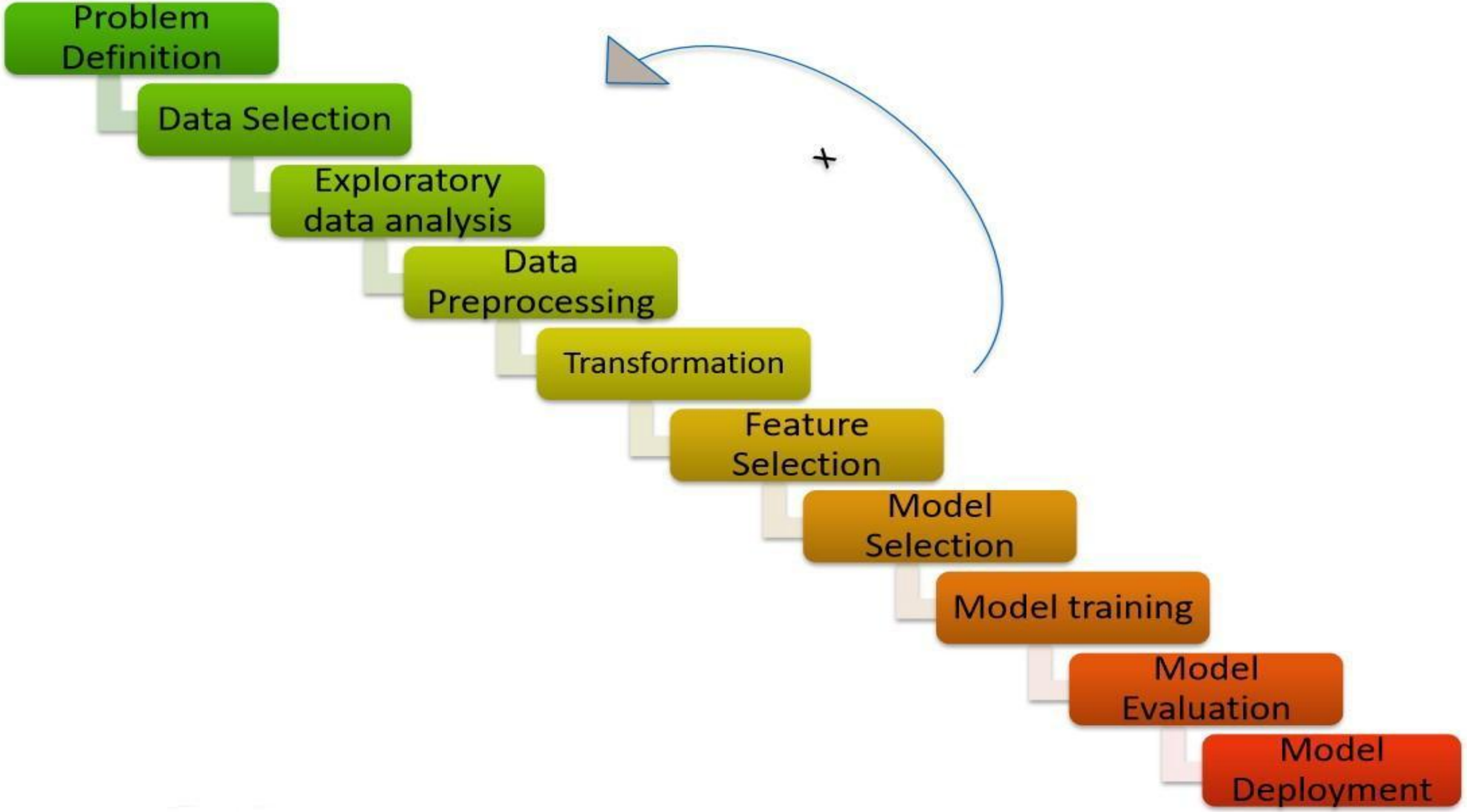
and many more...

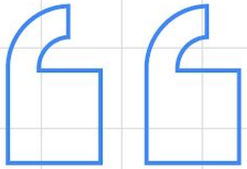
Developer Student Clubs

Life Cycle of Data Science Project



Developer Student Clubs


Problem Definition
Data Selection
Exploratory data analysis
Data Preprocessing
Transformation
Feature Selection
Model Selection
Model training
Model Evaluation
Model Deployment



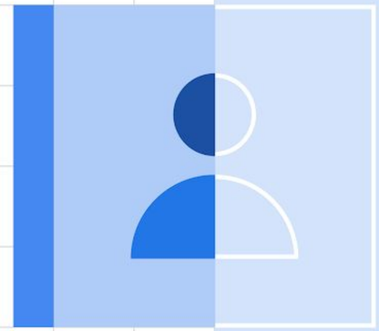
Are you all ready Now to kick-start your journey as **Data Scientists**?

Let's get started!

Developer Student Clubs

Thank you so much!!!

Hope you all are pumped up and excited for upcoming DSC Sessions!



Google Developers

Developer Student Clubs

Q/A Session