# BigMart sale prediction

For machine learning

# Our team member

## Truong

21110809

## Duy

21110759

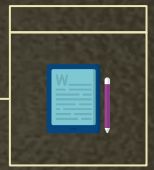
## Phuong

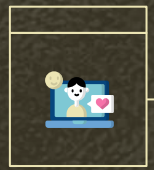
21110792

# Purpose

## Role

Help the mart to understand the properties of products and stores





## Predictions

impact business outcomes

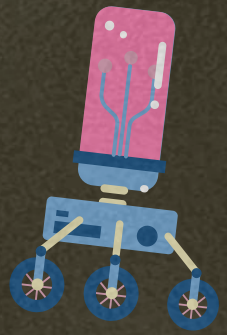
## Data analysis skills

Enhance your machine learning and data analysis skills.

# Input and output

## Input

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



## Output

BigMart will try to understand the properties of products and outlets which play a key role in increasing sales.

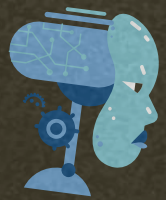
# Dataset

Collected from Kaggle, there is 11 classes with different train and test size

| Variables | Descripions |
|---|---|
| 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 (MRP) of the product |
| Outlet_Identifier | Unique store ID |
| Outlet_Establishment_Year | The year in which store was establiished |
| 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 |

# Model

## Linear Regression



## Pros

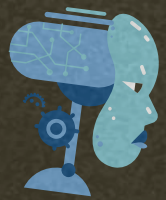
Fast to train and works well with large datasets.

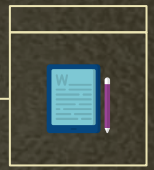
## Cons

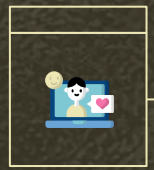
A linear relationship between features and target variable might not always hold true. It's also sensitive to outliers.

# Model

## Decision Tree







## Pros

Can handle both numerical and categorical data, and doesn't require much data preprocessing.

## Cons

Prone to overfitting if not properly tuned, and can create biased trees if some classes dominate.

# Model

## Random Forest

## Pros

Reduces overfitting problem in decision trees and handles large datasets with higher dimensionality. It can estimate what variables are important in the classification.

## Cons

More complex and computationally intensive than decision trees.

# Project Plan

| Task | Assignment | Day |
|---|---|---|
| Choose Topic | All members | 10-11/10 |
| Determine the required variables | All members | 12-13/10 |
| Collect data | Duy, Truong | 16-21/10 |
| Filter the data | Truong, Phuong | 22/10 |
| Determine the algorithm to use | All members | 23-24/10 |
| Conduct assessment test | Duy, Phuong | 25/10 |
| Algorithm construction | Truong, Phuong | 27/10-5/11 |
| Code design | Duy, Truong | 5-23/11 |
| Test and fix bug 1 | All members | 24-25/11 |
| Report | Duy, Phuong | 25-30/11 |
| General test | All members | 6-7/12 |
| Test and fix bug 2 (if have) | Individual | 8-9/12 |

# THANKS FOR WATCHING

GROUP 5