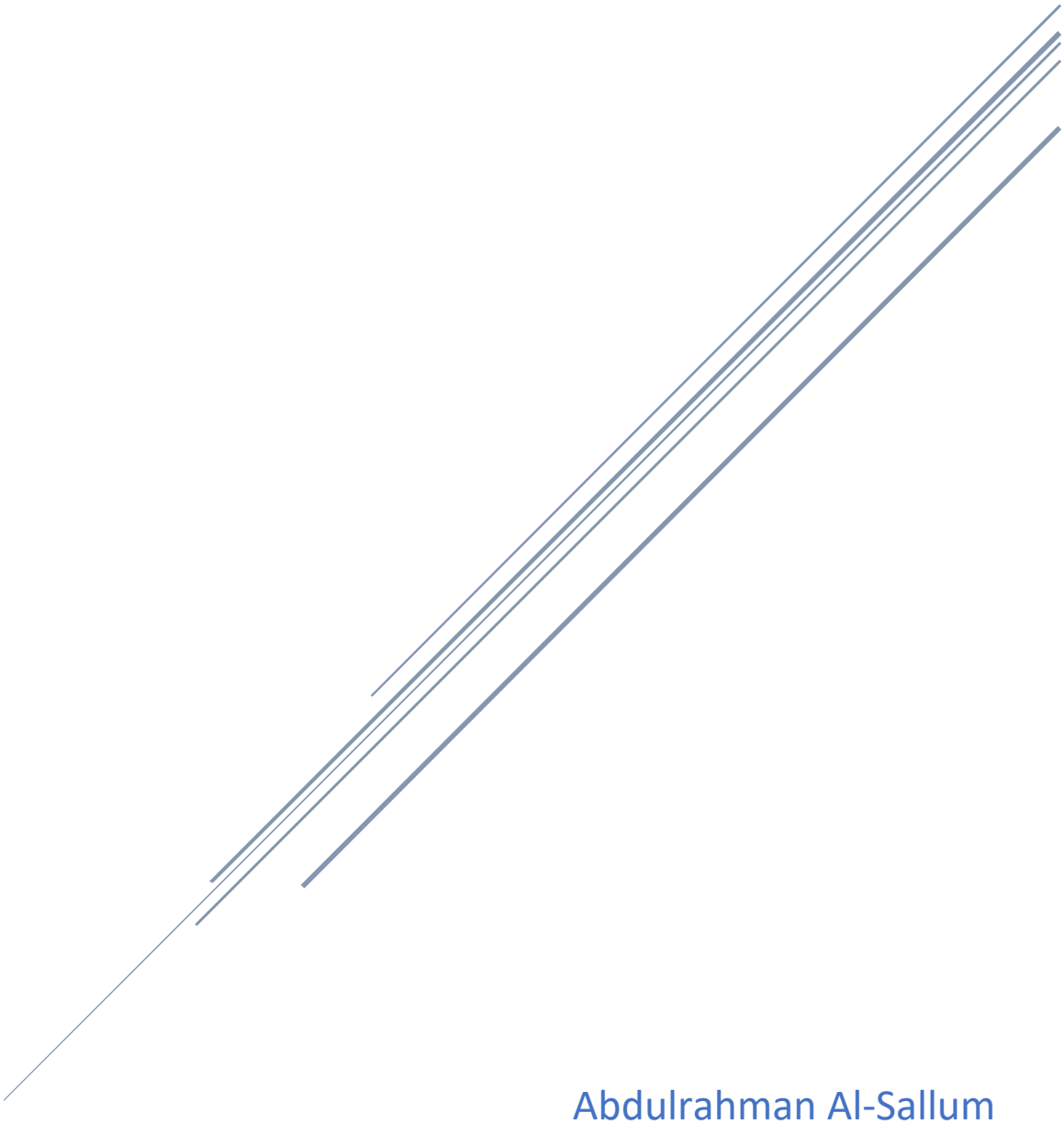
FINAL SUBMISSION

Big Mart Sales Data



Abdulrahman Al-Sallum
T5-Bootcamp

- **Abstract:**

The Goal of this project is to understand and predict the sales of customers in Big Mart stores.
The data worked on was collected in 2013 and uploaded to Kaggle. Using this data, we can get insights to the different outlets of Big mart and products.

- **Design**

The purpose is to understand customer behavior and predict the sales of Big Mart outlets. By Viewing Vital correlation and visualization that allow us to answer important questions like: Which outlet sold the most? What factors to consider that effect the sales?

- **Data**

The data contain around 8500 records with 12 features categorical and numerical. A few feature highlights include product type, MPR(PRICE), Outlet Type. With the targeted is the sales.

- **Algorithm**

Preprocessing:

1- The data columns were in a good consistent shape but there were missing values in two columns.
2- The two columns are: Product Weight which is numerical. Outlet size which is categorical (small – medium – large)
3- For the numerical data (product weight) we linked all the products with their ID using a python dictionary, so any missing product weight we can find its weight from the dictionary.
4- For the categorical data (Outlet size) we used the mode (most occurred value) to fill the missing values.
5- After dealing with the missing values the categorical were converted to dummy variable.

Modeling:

Since this is a regression problem, the following algorithms were used:

Lasso, Ridge, XGBoost and Random Forest.
The metric used is R^2 which indicate how well your model perform (1 being the highest).

Cross validation was used to see the average performance of the models:

Model	R^2
Lasso	0.55
Ridge	0.52
XG Regressor	0.558
Random Forest	0.54

- **Communication**

A dedicated presentation was made to show some key visualization and insights (GitHub).