**Data Science Assignment Report**

* Created a tool that estimates the selling prices of items

**Data exploration**

* I did some basic exploration of the dataset with the help of the methods like info, describe etc.

**Data cleaning**

# Dropped irrelevant columns

# Found the number of missing values in each column

# quantity tons 0

# customer 1

# country 28

# status 2

# item type 0

# application 24

# thickness 1

# width 0

# delivery date 1

# selling\_price 1

# Replaced the missing values with mode and median wherever applicable.

**Handling Outliers**

* Found the outliers in the data using zscore measure and 2 records were found with zscore>3, so I removed them.

**Data visualisation**

* Plotted pie charts, bar graphs, line charts to gain the following insights about the data:
  + 1. Majority of the items have ‘Won’, ‘Lost’ and ‘Not lost for AM’ as respective status.
    2. Total Selling price of Items with item type S and W is exponentially greater than that of other item types.
    3. Maximum number of orders were delivered on dates 22021-01-01 and 2021-04-01.
    4. Item types W and S are more prominent in the given dataset.

**Preparing Data for model training (feature engineering)**

* First, I found out number of different categories in categorical features.
* Then I performed One Hot Encoding for columns with many categorical features.
* For this, I took top 10 features that were most prominent in the dataset (took reference from a Kaggle competition).

**Model Building**

* I split the data into train and test sets with a test size of 10%.
* I also standardized the dataset so that converging of the gradient descent becomes faster.
* I tried 4 different models and evaluated them using mean squared error as the MSE is great for ensuring that our trained model has no outlier predictions with huge errors.

**Model Performance**

The random forest model far outperformed the other approaches.

* **Random forest: MSE= 21077.12**
* **Linear Regression: MSE=32212.77**
* **Lasso Regression: MSE=32110.28**
* **Ridge Regression: MSE=32110.28**

After that, Hyperparameter tuned the best regression model (random forest) using randomised search cv with parameters such as n estimators, max\_features, max\_depth, min\_samples\_split, min\_samples\_leaf.

**R2 score calculated value: 0.52**