Megastore Profit Prediction

Team ID	**SC_31**

Name	ID
محمد محمود فوزي خليل	20201700736
فتحي ناصر فتحي	20201700575
مازن عبداللطيف لطفي عبد اللطيف	20201700642
احمد ياسر محسن صالح	20201700095
محمد ابراهيم عبدالحميد محمد	20201700651
محمد عبد الله عبد الحكيم سالم	20191700835

The Report

1) Preprocessing Techniques

1: we explore the data and explore its columns and rows.

```
Index(['Row ID', 'Order ID', 'Order Date', 'Ship Date', 'Ship Mode',
       'Customer ID', 'Customer Name', 'Segment', 'Country', 'City', 'State',
       'Postal Code', 'Region', 'Product ID', 'CategoryTree', 'Product Name',
       'Sales', 'Quantity', 'Discount', 'Profit'],
      dtype='object')
```

2: check that the data has a null value or not and if a null is found we delete the rows that have null values.

```
RangeIndex: 7995 entries, 0 to 7994
Data columns (total 20 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Row ID         7995 non-null   int64
 1   Order ID       7995 non-null   object
 2   Order Date     7995 non-null   object
 3   Ship Date      7995 non-null   object
 4   Ship Mode      7995 non-null   object
 5   Customer ID    7995 non-null   object
 6   Customer Name  7995 non-null   object
 7   Segment        7995 non-null   object
 8   Country        7995 non-null   object
 9   City           7995 non-null   object
 10  State          7995 non-null   object
 11  Postal Code    7995 non-null   int64
 12  Region         7995 non-null   object
 13  Product ID     7995 non-null   object
 14  CategoryTree   7995 non-null   object
 15  Product Name   7995 non-null   object
 16  Sales          7995 non-null   float64
 17  Quantity       7995 non-null   int64
 18  Discount       7995 non-null   float64
 19  Profit         7995 non-null   float64
dtypes: float64(3), int64(3), object(14)
```

3: splitting the data to train and test (70 % for train & 30 % for test).

4: we apply feature engineering:

We calculate a new column (days to deliver) from the columns (Ship Date, Order Date).

Days to deliver = Ship Date – Order Date.

5: the data have 2 columns containing date. We select the month from each date in each column and replace the 2 date columns with the new month date columns.

6: the data have a dictionary that have a 2 key (“Main Category”, “Sub Category”).

To handle this column we extract the “Main Category” and the “Sub Category” with its values and creates a new 2 columns for them.

7: the data have 13 columns that values are Categorical.

We handle that by applying the feature encoding on those columns.

8: The ranges of data are very different we handle it by applying data scaling. We use standard scaler to scale the data.

9: other techniques that improve the results:

- Applying outlier detection .improves the model by handle outlier data that decreases the error of the models

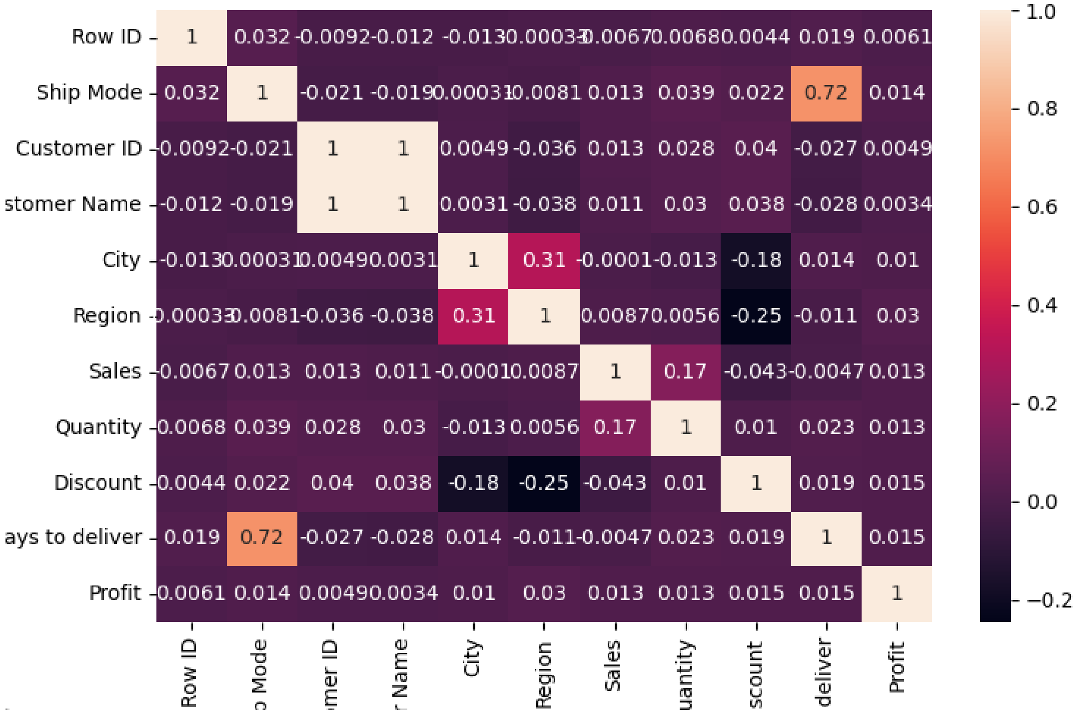
2) Feature Selection (Perform Analysis)

- We apply Correlation on dataset after the preprocessing techniques.

    1- Apply concatenation on training data and testing data.

    1- Apply correlation that the correlation of target column > 0.001.

    2- We got on the top features from the correlation to train data and test data.



3) Regression Models

1) Apply Linear regression.
    a. Fit train data and target train data.
    b. Predict test data.
    c. Calculate the mean square error on this data.
    d. MSE = 3749.089509321926
    e. Calculate the r2_score = 29.25445261684535

```
mean squared error for linear regression model test :  3749.089509321926
r2_score for linear regression model:
29.25445261684535
```

2)Apply Polynomial regression.

a. Apply polynomial regression.
b. Fit and transform on the train data.
c. Fit the transformed features to linear regression.
d. Fit the polynomial data and target train data.
e. Predict on test data.
f. Calculate the mean square error = 1745.9811297553447
g. Calculate the r2_score = 67.05322974069483

```
Mean Square Error for polynomial regression model 1745.9811297553447
r2_score polynomial regression model:
67.05322974069483
```

3)Apply Ridge regression.

a. Fit train data and target train data.
b. Predict test data.
c. Calculate the mean square error on this data.
d. MSE = 3749.1377888548927
e. Calculate the r2_score = 29.253541579117748

```
Mean Squared Error: 3749.1377888548927
R-squared Score for ridge regression: 29.253541579117748
```

4)Apply Lasso regression.

f. Fit train data and target train data.
g. Predict test data.
h. Calculate the mean square error on this data.
i. MSE = 3753.26
j. Calculate the r2_score = 29.253541579117748

```
Mean squared error: 3753.26
R-squared Score for lasso regression: 29.253541579117748
```

4) Differences Between Each Model

1 When we applied polynomial regression the mean square error and the accuracy became better from linear regression and other models.

2 When we applied ridge regression the mean square error and the accuracy became almost like linear regression.

5) Conclusion

In this report, we applied various preprocessing techniques such as handling missing values, scaling, and encoding categorical variables. We also performed feature selection and correlation analysis to identify the most important features for the models. After preprocessing, we applied four regression models: linear regression, polynomial regression, ridge regression, and lasso regression to predict the target variable.

We evaluated the models' performance using metrics such as mean squared error (MSE), and R-squared score. Our results showed that polynomial regression model outperformed other models. We also found that feature selection improved the model's performance by reducing overfitting.

Overall, our findings suggest that using polynomial regression model along with proper preprocessing techniques and feature selection can lead to accurate predictions of the target variable.