# Phase 5 - Submission

## TEAM NAME: Data Digits

Sravani Ganipisetti - sgani2@unh.newhaven.edu

Divya Astalapuram – dasta1@unh.newhaven.edu

Bindu Priya Basa- bbasa4@unh.newhaven.edu

## Please introduce your selected data set and research question.

The dataset we are working with is "H&M sales 2018 data". The dataset is taken from Kaggle website (https://www.kaggle.com/datasets/tulasiram574/hm-sales-data ). This dataset contains 15 attributes and 100 rows. This dataset is about different products purchased from H&M company in United States and about the profits.

- This dataset has information on products sold by H&M in the year of 2018.
- Each product is represented by a unique ID(Product ID) and Order ID number and includes information such as the Customer Id, Order Date, Sales, Quantity, Discount and Profit.
- And this dataset includes information about the Category and Sub-Category of each product and ordered from which City, State and Region.

**RESEARCH QUESTION:**

To what extent does the geographic location, encompassing city, state and regional factors, influence the underlying dynamics of product co-purchasing behaviours, as revealed through market basket analysis? Furthermore, how can these discerned geographic variations inform the strategic customization of marketing and sales approaches for diverse locations, ensuring optimal alignment with local consumer preferences and demands?

**HARDWARE USED:**

- Processor Used: i5
- RAM: 8GB
- Operating System: Windows 11 64-bit
- Tools Used: Jupyter Notebook
- Language: Python

**List Of Data Mining Techniques Used:**

- Random Forest
- Regression
- Time Series

**Random Forest:**

- Random Forest is a machine learning algorithm used for classification and regression tasks due to its high accuracy, robustness, feature importance, versatility and scalability. Random Forest reduces overfitting by averaging multiple decision trees and is less sensitive to noise and outliers in the data.

```
model = RandomForestRegressor(random_state=42)


grid_search = GridSearchCV(model, param_grid, cv=5, scoring='neg_mean_squared_error')
grid_search.fit(X_train, y_train)

best_model = grid_search.best_estimator_

best_model.fit(X_train, y_train)

y_pred = best_model.predict(X_test)

mse = mean_squared_error(y_test, y_pred)
mae = mean_absolute_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
```

```
Mean Squared Error: 930.9124861898647
Mean Absolute Error: 18.672686361978556
R-squared (R2) Score: -0.229782114090763
```

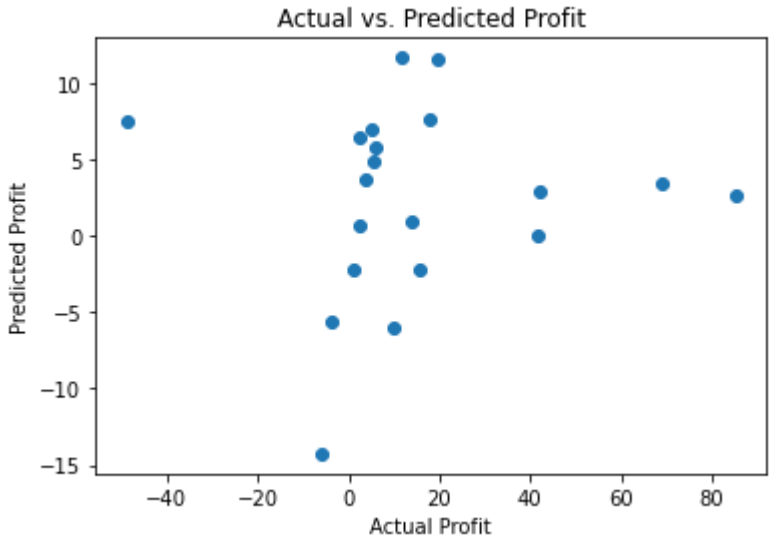
- The R2 score suggests that random forest model is okay for the dataset we have used.

**Parameters Used:**

'n_estimators' :[50, 100, 150]

**Hyper Parameters Used:**

grid_search = GridSearchCV(model, param_grid, cv=5, scoring='neg_mean_squared_error')




**Regression:**

- Regression is a machine learning technique used for investigating the relationship between independent variables or features and a dependent variable or outcome. It is used as a method for predictive modelling in machine learning, in which an algorithm is used to predict continuous outcomes.

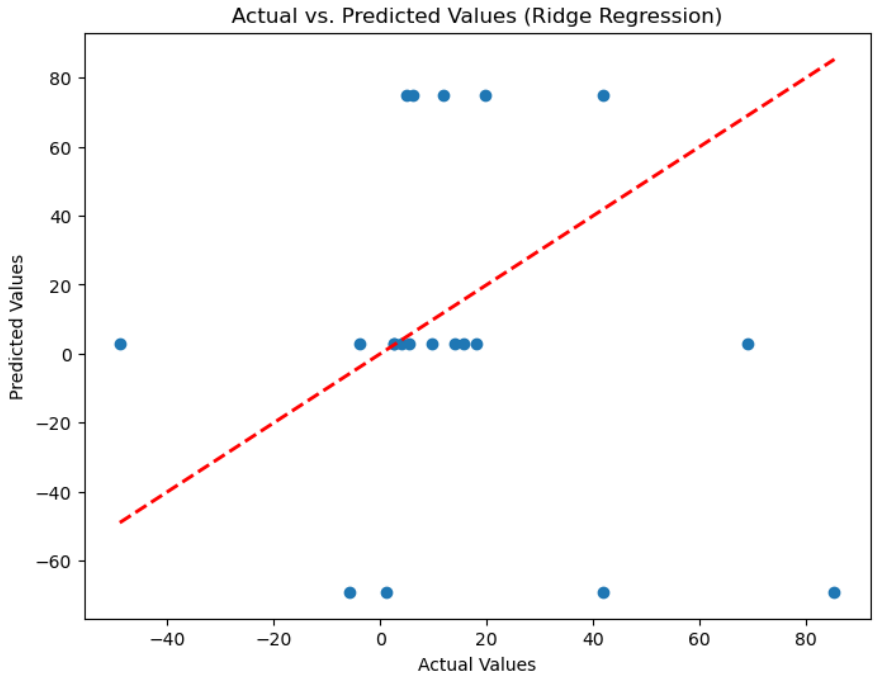
```
print(f"Ridge Regression Coefficients: {best_ridge.coef_}")
print(f"Ridge Regression MSE: {ridge_mse}")
print(f"Ridge Regression R2 Score: {ridge_r2}")
plt.figure(figsize=(8, 6))
plt.scatter(y_test, ridge_predictions)
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], linestyle='--', color='red', linewidth=2)
plt.title("Actual vs. Predicted Values (Ridge Regression)")
plt.xlabel("Actual Values")
plt.ylabel("Predicted Values")
plt.show()
```

```
Ridge Regression Coefficients: [-44.27896295   0.        ]
Ridge Regression MSE: 3518.2462079360403
Ridge Regression R2 Score: -3.6477798113935034
```

- A Mean Squared Error (MSE) of 3518.24 indicates that, on average the model's predictions deviate from the actual values.
- The R-Squared (R2) has value of -3.64 indicates that model is not performing good and not performing well.

**Parameters Used:**

'alpha': [0.001, 0.01, 0.1, 1, 10]




**Time Series Analysis:**

- Time series analysis is a statistical technique used to analyze and make predictions about data that is collected over a period of time.

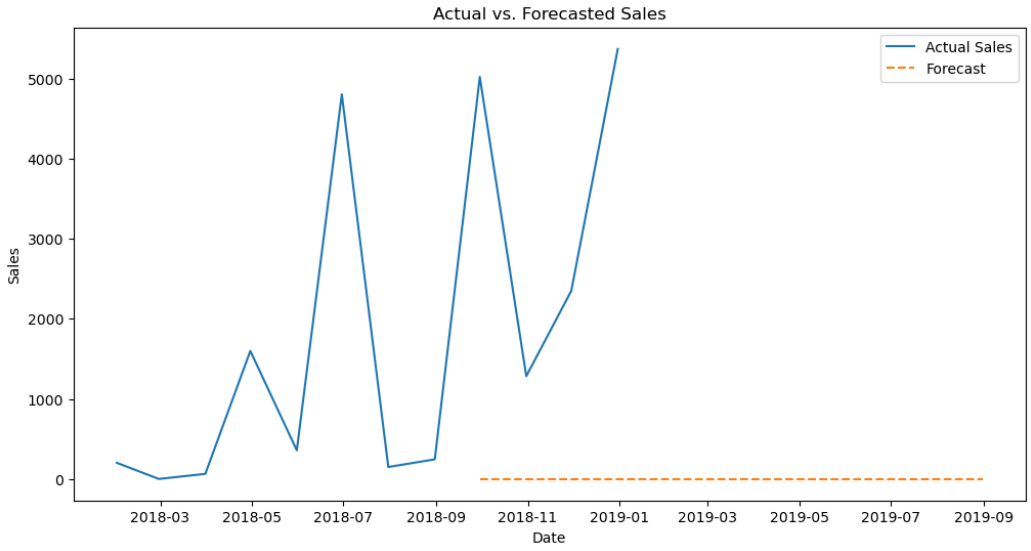
```
model = ExponentialSmoothing(time_series_data['Sales'], trend='add', seasonal='add', seasonal_periods=12)
model_fit = model.fit()


forecast = model_fit.forecast(steps=12)
forecast = pd.Series(forecast, index=pd.date_range(start=time_series_data.index[-1], periods=12, freq='M'))


time_series_data = time_series_data.resample('M').sum()
forecast = forecast.resample('M').sum()


mse = mean_squared_error(time_series_data['Sales'], forecast)
r2 = r2_score(time_series_data['Sales'], forecast)
print("Time Series MSE:", mse)
print("Time Series R2 Score:", r2)
```

```
Time Series MSE: 7256081.388057724
Time Series R2 Score: -0.7854833702247861
```

- The R-Squared (R2) has value of -0.785 indicates that it is capturing the variance in the data.

**Parameters Used:**

(time_series_data['Sales'], trend='add', seasonal='add', seasonal_periods=12)




**Association Rules:**

Association Rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures.

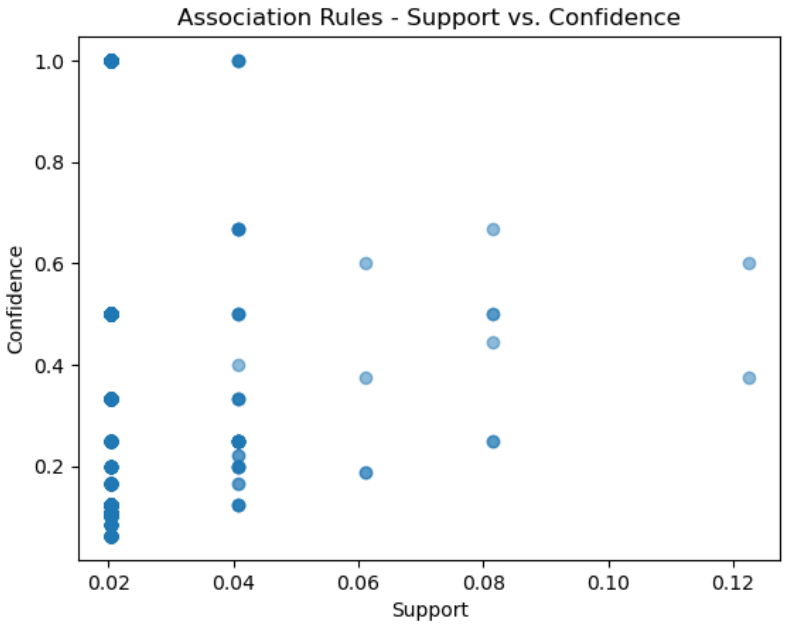
```
df = new_data_cleaned[['Order ID', 'Sub-Category']]

basket = df.groupby('Order ID')['Sub-Category'].apply(list).reset_index()

basket = basket.set_index('Order ID')['Sub-Category'].str.join('|').str.get_dummies().reset_index()

frequent_itemsets = apriori(basket.drop('Order ID', axis=1), min_support=0.005, use_colnames=True)

rules = association_rules(frequent_itemsets, metric="lift", min_threshold=1.0)

print(rules)
```

```
        antecedents                       consequents  \
0       (Flip flops)                           (Bags)
1            (Bags)                     (Flip flops)
2     (Heels & Flats)                           (Bags)
3            (Bags)                  (Heels & Flats)
4          (Jackets)                           (Bags)
...              ...                             ...
1495       (Belts)  (Tops, Dresses, Formals, Jackets, Sport shoes)
1496   (Sport shoes)       (Tops, Belts, Formals, Jackets, Dresses)
1497      (Formals)    (Tops, Belts, Dresses, Jackets, Sport shoes)
1498      (Jackets)   (Tops, Belts, Dresses, Formals, Sport shoes)
1499      (Dresses)    (Tops, Belts, Formals, Jackets, Sport shoes)

      antecedent support  consequent support   support  confidence      lift  \
0               0.061224            0.183673  0.020408    0.333333  1.814815
1               0.183673            0.061224  0.020408    0.111111  1.814815
2               0.061224            0.183673  0.020408    0.333333  1.814815
3               0.183673            0.061224  0.020408    0.111111  1.814815
```

```
4          0.163265          0.183673  0.081633    0.500000  2.722222
...             ...       ...       ...       ...       ...
1495       0.122449          0.020408  0.020408    0.166667  8.166667
1496       0.163265          0.020408  0.020408    0.125000  6.125000
1497       0.163265          0.020408  0.020408    0.125000  6.125000
1498       0.163265          0.020408  0.020408    0.125000  6.125000
1499       0.326531          0.020408  0.020408    0.062500  3.062500
...
1498  0.017076   1.119534        1.000000
1499  0.013744   1.044898        1.000000
```

[1500 rows x 10 columns]

- The association rules is used for analyzing customer order data to discover and print the sub-categories of items are frequently purchased.



**Conclusion:**

To check which model suits best, In our analysis we used several modelling techniques like Random Forest, Time Series, Regression Analysis. Among these models Random Forest is the best performer on our dataset achieving R2 score of -0.22.

We've mostly used regression data mining technique as because some of the variables are categorical and some are not. We've calculated Mean square error, Rsquare score , Mean absolute error as in regression, these are the testing parameters. We've tried Ridge regression, Random forest and Time series analysis , Associate methods (Apriori) as our data mining techniques for this data set. We found Random forest to be the best fit model for this based on the R square score we calculated. Even apriori results are moderate and better. We split the dataset into 80% training set and 20% testing set to assess the performance.

**Github Link:**

https://github.com/Sravani1698/Data_Digits_Phase5.git