Data Science with Python Career Program - Internship Project on Amazing Mart Dataset

- By Mohit Bohra

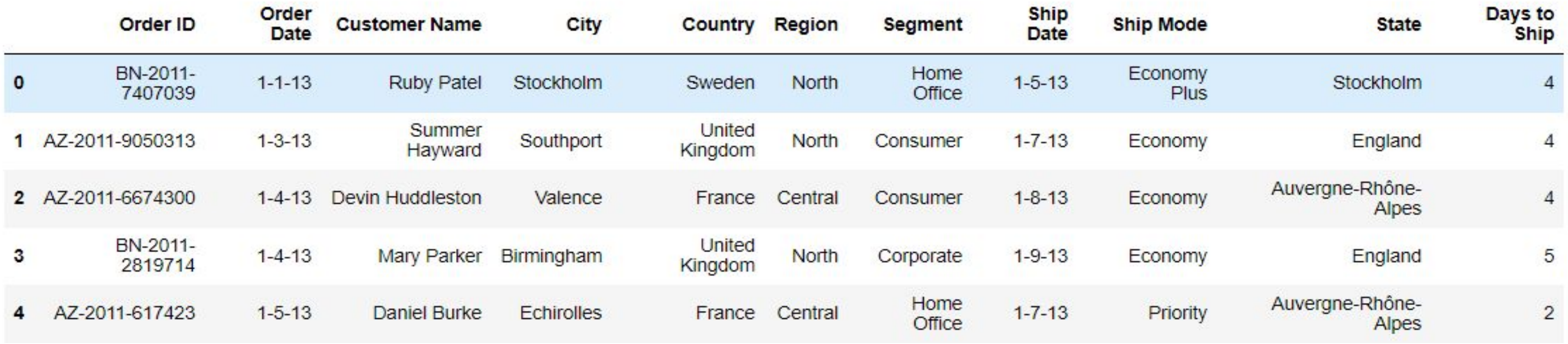
Agenda

- Data Exploration
- Data insights
- EDA Graphs.
- Graphical Analysis and conclusion on Data
- Data Cleaning & Pre-Processing Steps.
- ML Modeling
- Deployment of ML Models using Streamlit.

Data Exploration

- Data file one Order Data which is a TSV file
  (This file is converted to csv file to read data)

The dataset provided is a summary of orders made by customers in various countries and regions. It contains 11 columns providing different details of each order such as Order ID, Order Date, Customer Name, City, Country, Region, Segment, Ship Date, Ship Mode, State, and Days to Ship. Each row represents a single order made by a customer and provides information on when the order was made and shipped, the mode of shipping, the customer's location and segment, and how long it took to ship the order. This dataset could be used for various analytical purposes, such as identifying trends, analyzing shipping times, and understanding customer behavior.

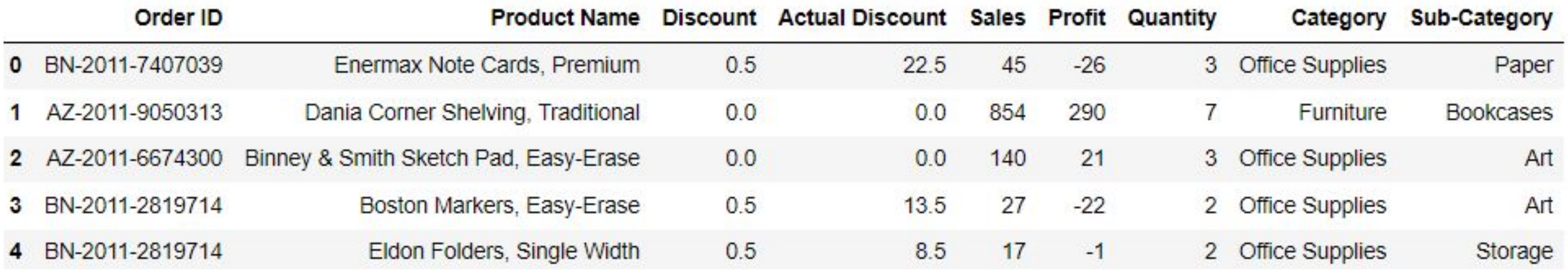


	Order ID	Order Date	Customer Name	City	Country	Region	Segment	Ship Date	Ship Mode	State	Days to Ship
0	BN-2011-7407039	1-1-13	Ruby Patel	Stockholm	Sweden	North	Home Office	1-5-13	Economy Plus	Stockholm	4
1	AZ-2011-9050313	1-3-13	Summer Hayward	Southport	United Kingdom	North	Consumer	1-7-13	Economy	England	4
2	AZ-2011-6674300	1-4-13	Devin Huddleston	Valence	France	Central	Consumer	1-8-13	Economy	Auvergne-Rhône-Alpes	4
3	BN-2011-2819714	1-4-13	Mary Parker	Birmingham	United Kingdom	North	Corporate	1-9-13	Economy	England	5
4	AZ-2011-617423	1-5-13	Daniel Burke	Echirolles	France	Central	Home Office	1-7-13	Priority	Auvergne-Rhône-Alpes	2

Data Exploration

- Data file Two Order Breakdown Data which is in JSON format
  (This file is converted to csv file to read data)

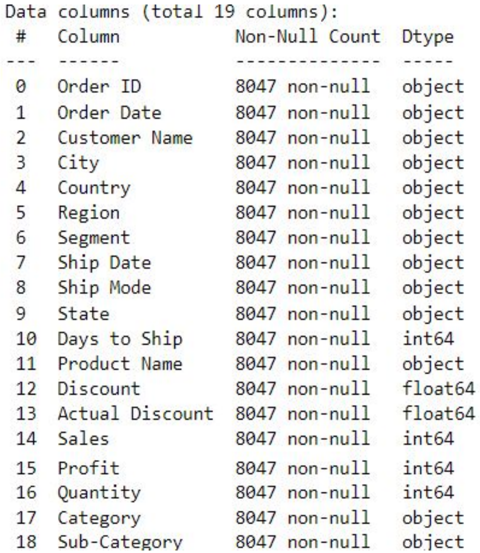
The dataset provided is a summary of sales made for various products in different categories and subcategories. It contains 9 columns providing details such as Order ID, Product Name, Discount, Actual Discount, Sales, Profit, Quantity, Category, and Sub-Category. Each row represents a single product sold, providing information on the product name, discount, actual discount, sales amount, profit made, quantity sold, and category and subcategory of the product. This dataset could be used for various analytical purposes, such as identifying popular products or categories, analyzing discounts and profitability, and understanding sales trends over time.



	Order ID	Product Name	Discount	Actual Discount	Sales	Profit	Quantity	Category	Sub-Category
0	BN-2011-7407039	Enermax Note Cards, Premium	0.5	22.5	45	-26	3	Office Supplies	Paper
1	AZ-2011-9050313	Dania Corner Shelving, Traditional	0.0	0.0	854	290	7	Furniture	Bookcases
2	AZ-2011-6674300	Binney & Smith Sketch Pad, Easy-Erase	0.0	0.0	140	21	3	Office Supplies	Art
3	BN-2011-2819714	Boston Markers, Easy-Erase	0.5	13.5	27	-22	2	Office Supplies	Art
4	BN-2011-2819714	Eldon Folders, Single Width	0.5	8.5	17	-1	2	Office Supplies	Storage

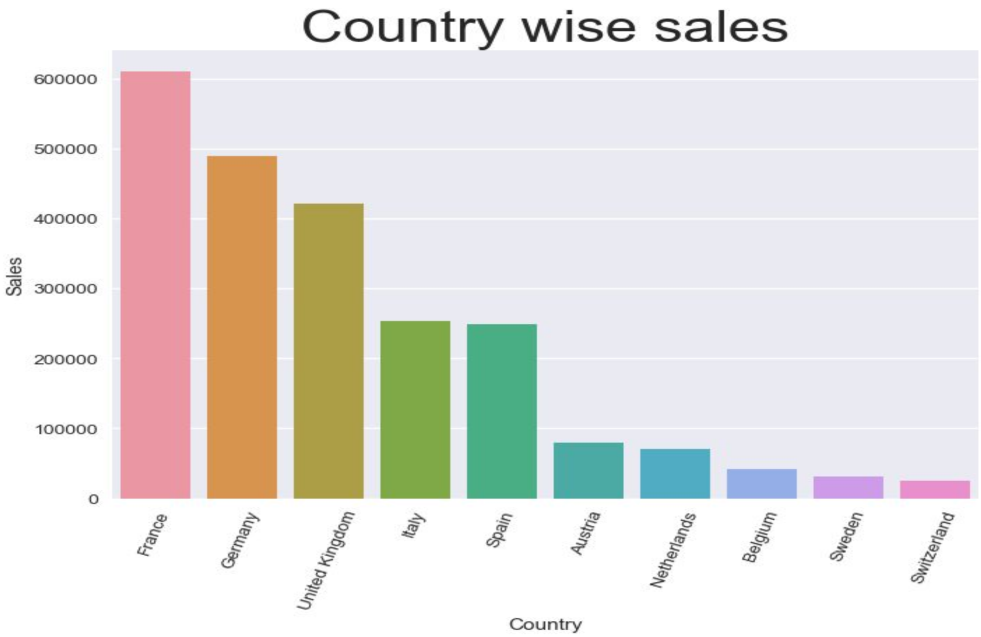
Data insights

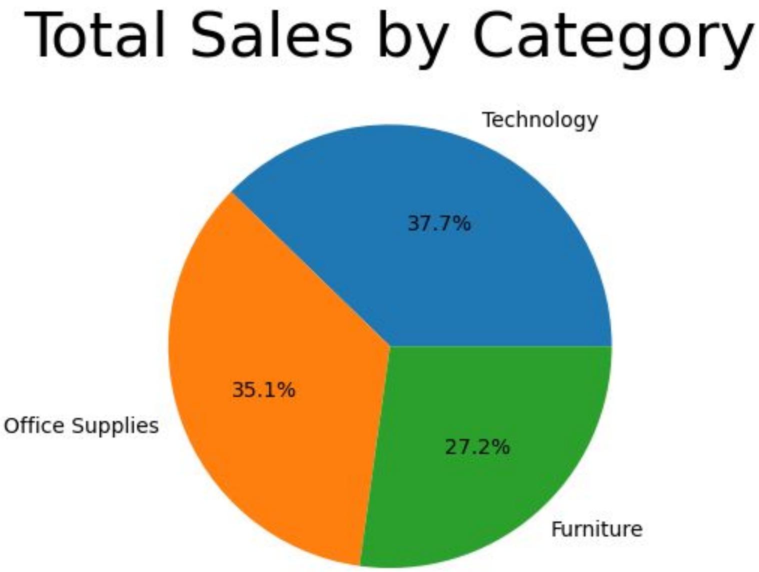
- DataFrame has 8047 rows and 19 columns. This information can be useful for understanding the size and structure of the dataset, and for performing various data manipulation and analysis tasks.
- There are no null values
- There are only two duplicates
- The consists of both categorical and numerical data



```
Data columns (total 19 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   Order ID         8047 non-null   object 
 1   Order Date       8047 non-null   object 
 2   Customer Name    8047 non-null   object 
 3   City             8047 non-null   object 
 4   Country          8047 non-null   object 
 5   Region           8047 non-null   object 
 6   Segment          8047 non-null   object 
 7   Ship Date        8047 non-null   object 
 8   Ship Mode        8047 non-null   object 
 9   State            8047 non-null   object 
 10  Days to Ship     8047 non-null   int64  
 11  Product Name     8047 non-null   object 
 12  Discount         8047 non-null   float64
 13  Actual Discount  8047 non-null   float64
 14  Sales            8047 non-null   int64  
 15  Profit           8047 non-null   int64  
 16  Quantity         8047 non-null   int64  
 17  Category         8047 non-null   object 
 18  Sub-Category     8047 non-null   object 
```

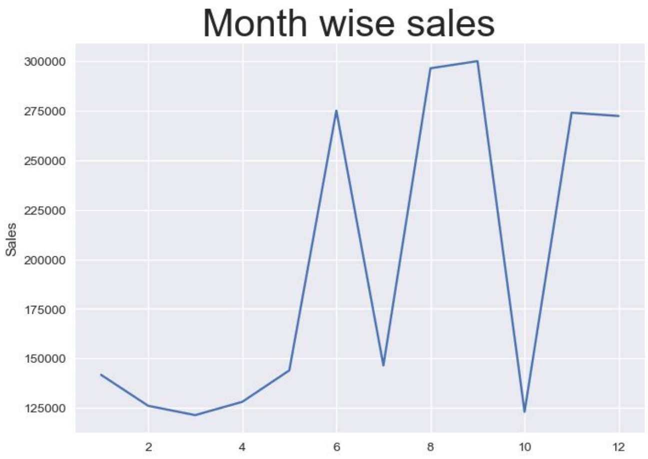
EDA Graphs.


Country wise sales
600000
500000
400000
300000
200000
100000
0
Sales
France
Germany
United Kingdom
Italy
Spain
Austria
Netherlands
Belgium
Sweden
Switzerland
Country

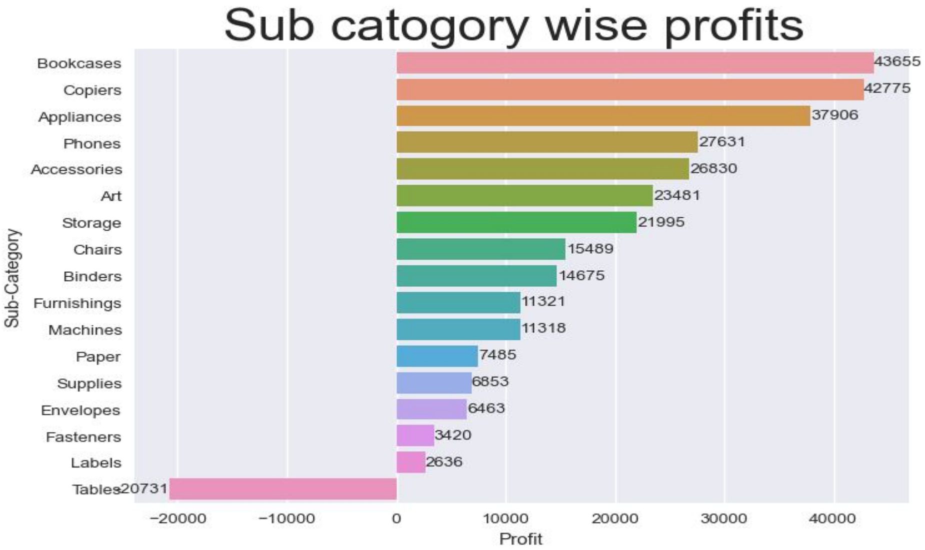

Total Sales by Category
Technology
37.7%
Office Supplies
35.1%
27.2%
Furniture

Graphical Analysis and conclusion on Data

skill academy

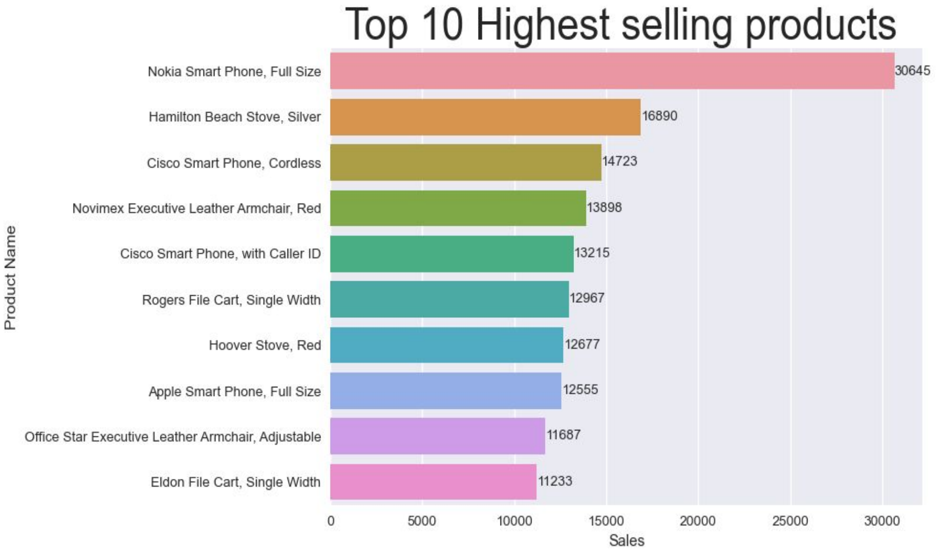


This Graph shows monthly sales data, with the highest sales amounts in August and September and the lowest in February, March, and October. The graph provides insights into sales patterns and trends over time, which can inform strategic decisions related to inventory, pricing, and marketing. This information is valuable for businesses looking to optimize their sales performance.

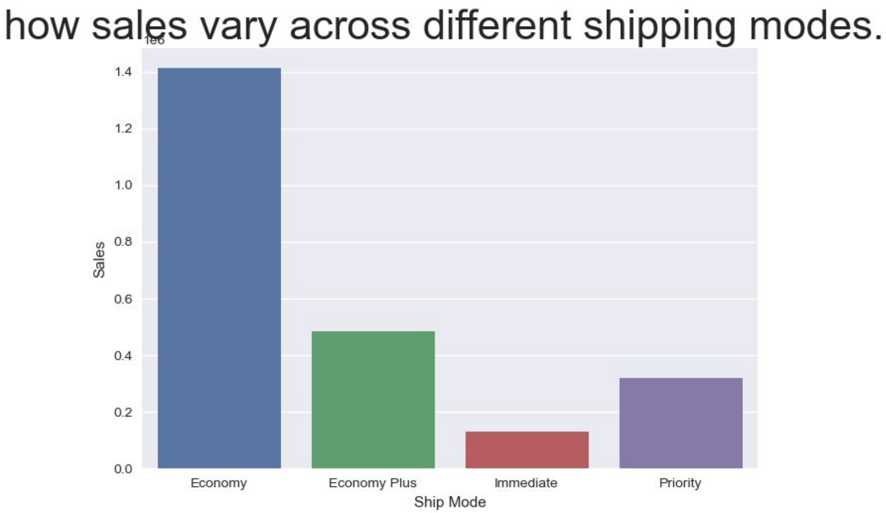


This graph is a summary of profit data for different sub-categories of products. The graph includes the top ten sub-categories by profit, with the highest profits in Bookcases, Copiers, and Appliances. The data provides insights into the performance of various sub-categories, which can inform decisions related to product development, marketing, and pricing. This information is valuable for businesses looking to optimize their profitability.

Graphical Analysis and conclusion on Data


Top 10 Highest selling products

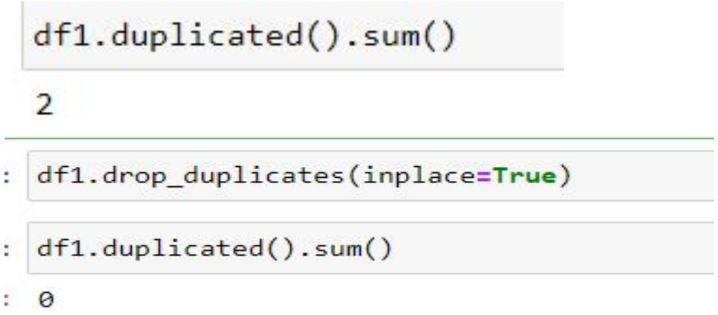
The given graph provides a summary of sales data for different products, including the name of the product and the total sales amount. The graph highlights the top ten products ranked by sales, with the highest sales recorded for Nokia Smart Phone, Full Size, Hamilton Beach Stove, Silver, and Cisco Smart Phone, Cordless. The graph can provide valuable insights into the popularity of different products, which can be utilized to make informed decisions about inventory management, marketing, and pricing, ultimately helping businesses optimize their sales performance.


how sales vary across different shipping modes.

This is a summary of sales information for different shipping modes, including the name of the mode and the total sales amount. The graph displays four different shipping modes, ranked by sales, with the highest sales recorded for Economy mode. The graph provides insights into the performance of different shipping modes, which can inform decisions related to shipping strategy, pricing, and customer satisfaction. This information is valuable for businesses looking to optimize their shipping and sales performance.

Data Cleaning & Pre-Processing Steps.

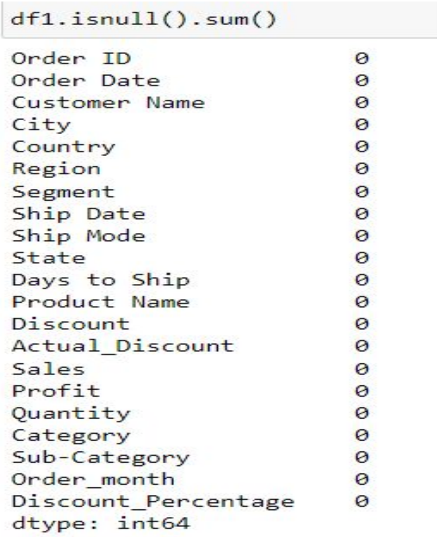
- Checked duplicate values
  Findings 2 duplicate values
- Removed duplicate values



```
df1.duplicated().sum()

2
```

```
: df1.drop_duplicates(inplace=True)

: df1.duplicated().sum()

: 0
```

- Checked null values
  Findings there are no null values



```
df1.isnull().sum()

Order ID                0
Order Date              0
Customer Name           0
City                    0
Country                 0
Region                  0
Segment                 0
Ship Date               0
Ship Mode               0
State                   0
Days to Ship            0
Product Name            0
Discount                0
Actual_Discount         0
Sales                   0
Profit                  0
Quantity                0
Category                0
Sub-Category            0
Order_month             0
Discount_Percentage     0
dtype: int64
```

ML Modeling

- The dataset used in this analysis is a sales dataset that contains information on various products sold by a company. The dataset includes information on the order ID, order date, customer name, city, country, region, segment, ship date, ship mode, state, days to ship, product name, discount, actual discount, sales, profit, quantity, category, sub-category, order month, and discount percentage.
- The first step in the analysis was to select the dependent and independent features for the dataset. The column names in the DataFrame were modified, a new column for discount percentage was created, and a subset of columns was selected as independent features including information on the country, region, segment, category, sub-category, discount percentage, actual discount, quantity, and order month. The dependent feature was set to the sales column of the same DataFrame.
- The next step was to evaluate the performance of various regression models on the dataset. Linear Regression, Ridge Regression, Lasso Regression, KNN Regression, Decision Tree Regression, and Random Forest Regression models were applied to the dataset, and the evaluation metrics of MAE, MSE, RMSE, and R2 score were used to evaluate the performance of each model. Based on the evaluation metrics, the Random Forest Regression model was selected as the best model for predicting "Sales" on the given dataset.

ML Modeling

1. Linear Regression: This model is a basic regression model that assumes a linear relationship between the independent and dependent variables. It generates a straight line that best fits the data. The evaluation metrics of this model are MAE of 172.77, MSE of 83229.65, RMSE of 288.50, and an R2 score of 0.53.

2. Ridge Regression: Ridge regression is a type of linear regression that uses L2 regularization to prevent overfitting. It adds a penalty term to the cost function that controls the size of the coefficients. The evaluation metrics of this model are MAE of 172.31, MSE of 82980.75, RMSE of 288.06, and an R2 score of 0.54.

3. Lasso Regression: Lasso regression is another type of linear regression that uses L1 regularization to prevent overfitting. It adds a penalty term to the cost function that shrinks the coefficients towards zero. The evaluation metrics of this model are MAE of 170.04, MSE of 86403.60, RMSE of 293.94, and an R2 score of 0.52.

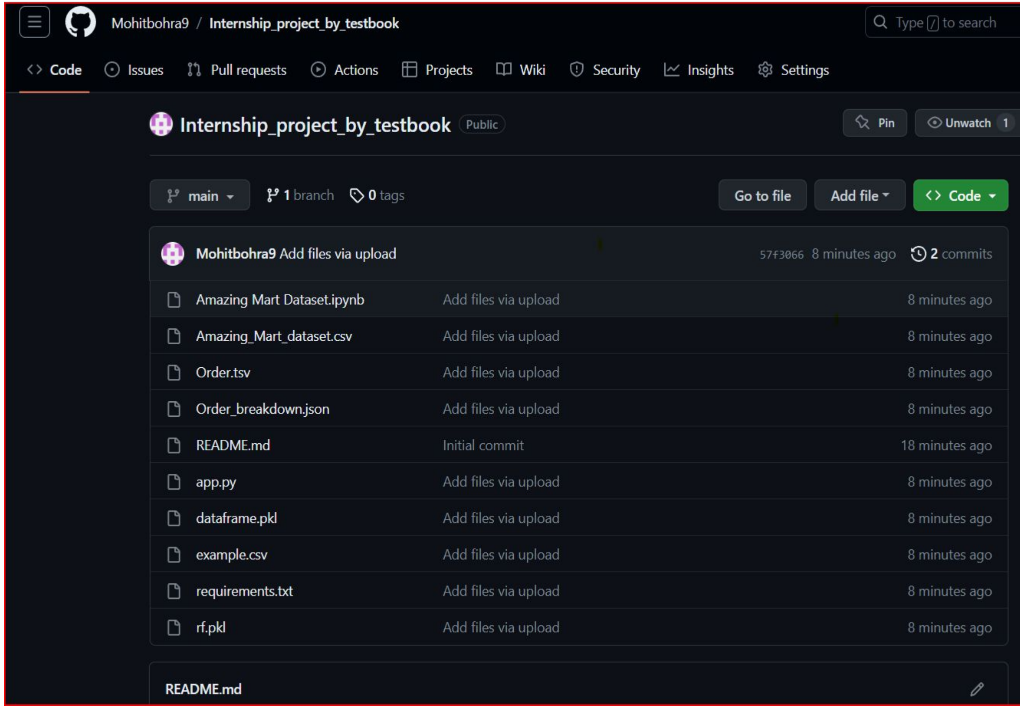
4. KNN Regression: KNN regression is a non-parametric model that uses the k-nearest neighbors to predict the target variable. It finds the k closest neighbors to a data point and takes the average of their target values as the predicted value. The evaluation metrics of this model are MAE of 102.82, MSE of 67548.35, RMSE of 259.90, and an R2 score of 0.62.

5. Decision Tree Regression: Decision tree regression is a model that uses a tree-like structure to predict the target variable. It splits the data into smaller subsets based on the values of the independent variables and generates a tree that predicts the target variable at the end of the branches. The evaluation metrics of this model are MAE of 103.33, MSE of 56561.01, RMSE of 237.83, and an R2 score of 0.68.

6. Random Forest Regression: Random forest regression is an ensemble model that uses multiple decision tree models to predict the target variable. It generates multiple decision trees on random subsets of the data and takes the average of their predictions. The evaluation metrics of this model are MAE of 95.90, MSE of 45316.31, RMSE of 212.88, and an R2 score of 0.75.

In conclusion, based on the evaluation metrics, the Random Forest Regression model is the best model for predicting "Sales" on the given dataset. However, it is important to note that the performance of each model may vary depending on the dataset and the problem at hand.

Deployment of ML Models using Streamlit.

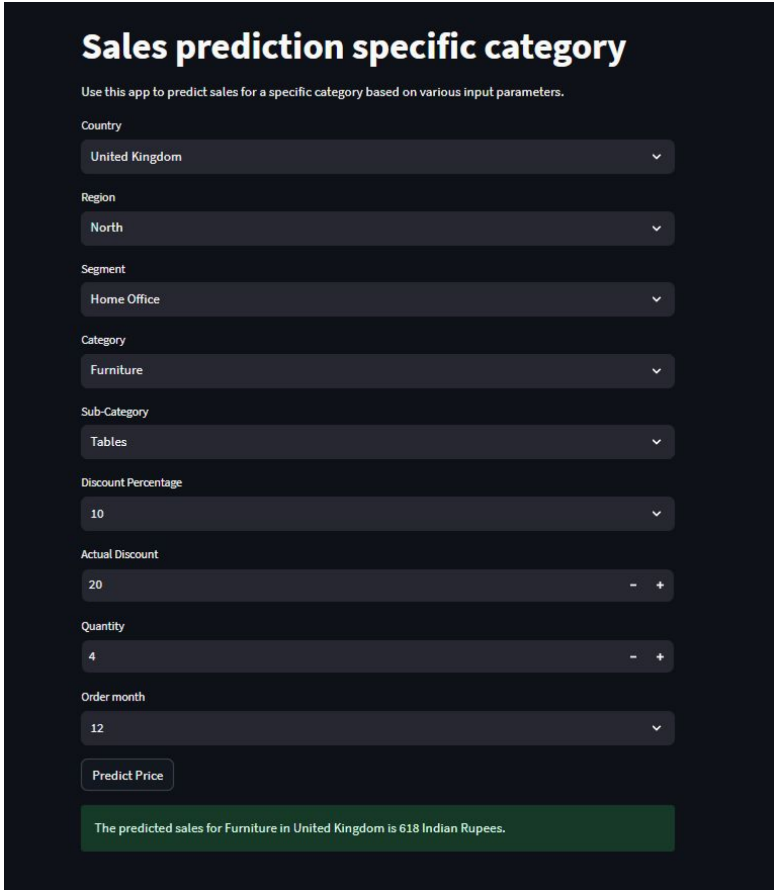


Internship_project_by_testbook Public

Mohitbohra9 Add files via upload 57f3066 8 minutes ago 2 commits

Amazing Mart Dataset.ipynb	Add files via upload	8 minutes ago
Amazing_Mart_dataset.csv	Add files via upload	8 minutes ago
Order.tsv	Add files via upload	8 minutes ago
Order_breakdown.json	Add files via upload	8 minutes ago
README.md	Initial commit	18 minutes ago
app.py	Add files via upload	8 minutes ago
dataframe.pkl	Add files via upload	8 minutes ago
example.csv	Add files via upload	8 minutes ago
requirements.txt	Add files via upload	8 minutes ago
rf.pkl	Add files via upload	8 minutes ago

README.md



Sales prediction specific category

Use this app to predict sales for a specific category based on various input parameters.

Country: United Kingdom

Region: North

Segment: Home Office

Category: Furniture

Sub-Category: Tables

Discount Percentage: 10

Actual Discount: 20

Quantity: 4

Order month: 12

The predicted sales for Furniture in United Kingdom is 618 Indian Rupees.

Endnotes

Reference Links:-

Github : https://github.com/Mohitbohra9/Internship_project_by_testbook

Streamlit website : https://internshipprojectbytestbook-ero5rgx3hoh.streamlit.app/