Mid Term Report

Project Title: US Regional Sales Channel Prediction

Group Details:

- Fayaz Shaik
- Mahesh Maddineni
- Nathaniel Yee
- Vivekananda Reddy Pyda

Data:

Type of Data: The data is presented in a comma-separated values format, featuring a combination of text, dates, and numerical information.

Dataset Information:

There are 7992 attributes and 16 attributes.

- **OrderNumber:** A unique identifier for each order.
- **Sales Channel:** The channel through which the sale was made (In-Store, Online, Distributor, Wholesale).
- **WarehouseCode:** Code representing the warehouse involved in the order.
- **ProcuredDate:** Date when the products were procured.
- **OrderDate:** Date when the order was placed.
- **ShipDate:** Date when the order was shipped.
- **DeliveryDate:** Date when the order was delivered.
- **SalesTeamID:** Identifier for the sales team involved.
- **CustomerID:** Identifier for the customer.
- **StoreID:** Identifier for the store.
- **ProductID:** Identifier for the product.
- **Order Quantity:** Quantity of products ordered.
- **Discount Applied:** Applied discount for the order.
- **Unit Cost:** Cost of a single unit of the product.
- **Unit Price:** Price at which the product was sold.

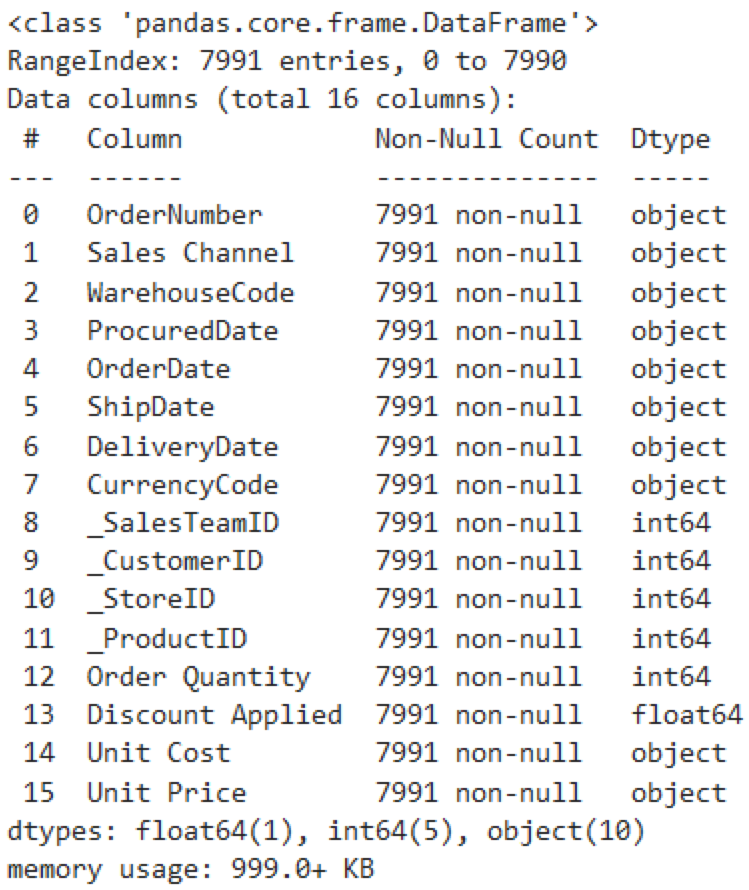
Below is the information of column names and their data types.

Column Name	Data Type
OrderNumber	object
Sales Channel	object
WarehouseCode	object
ProcuredDate	object
OrderDate	object
ShipDate	object
DeliveryDate	object
CurrencyCode	object
_SalesTeamID	int64
_CustomerID	int64
_StoreID	int64
_ProductID	int64
Order Quantity	int64
Discount Applied	float64
Unit Cost	object
Unit Price	object

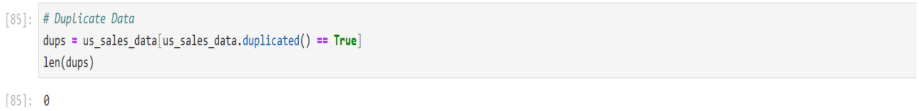
Data Processing:

Below are steps processing steps:

1. Checked for missing data, duplicate data. There are no such kind of data.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7991 entries, 0 to 7990
Data columns (total 16 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   OrderNumber       7991 non-null   object 
 1   Sales Channel     7991 non-null   object 
 2   WarehouseCode     7991 non-null   object 
 3   ProcuredDate      7991 non-null   object 
 4   OrderDate         7991 non-null   object 
 5   ShipDate          7991 non-null   object 
 6   DeliveryDate      7991 non-null   object 
 7   CurrencyCode      7991 non-null   object 
 8   _SalesTeamID      7991 non-null   int64  
 9   _CustomerID       7991 non-null   int64  
 10  _StoreID          7991 non-null   int64  
 11  _ProductID        7991 non-null   int64  
 12  Order Quantity    7991 non-null   int64  
 13  Discount Applied  7991 non-null   float64
 14  Unit Cost         7991 non-null   object 
 15  Unit Price        7991 non-null   object 
dtypes: float64(1), int64(5), object(10)
memory usage: 999.0+ KB
```



```
[85]: # Duplicate Data
      dups = us_sales_data[us_sales_data.duplicated() == True]
      len(dups)

[85]: 0
```

2. Converted object data type in OrderDate, ShipDate, DeliveryDate to datetime data type and also converted object data type of Unit Cost, Unit Price to float data type.

3. Derived new features “DaysToShip”, “DaysToDeliver” using OrderDate, ShipDate, Delivery date and also derived Profit using Unit Cost, Unit Price.

4. A total of 19 plots were created for the purpose of data visualization, all of which are available for reference in the attached notebook.

5. Employed encoding techniques to translate categorical values into their corresponding numeric representations.

6. Employed box plots to detect outliers within the data and subsequently removed these outliers using the Interquartile Range (IQR) method.

7. Computed a correlation matrix and created visualizations to identify the features that exhibit significant correlations with the target variable.

8. Split the dataset into training and testing sets, adhering to an 80:20 ratio.

Which attributes you use and which one you don't use? Why?

- Dropped OrderNumber as it is unique for every column.
- Dropped ProcuredDate is manufacturing date of the product which doesnot require to predict sales channel
- Dropped CurrencyCode is USD for all orders and it is also not affecting our target data.
- Dropped OrderDate, ShipDate, DeliveryDate as we derived new features DaysToShip and DaysToDeliver.
- Using all other columns in the dataset including the derived features.

Data Mining Task

Task: We explored 3 different multi-class classification algorithms to address the problem, given the limited unique values in the target variable. The prominent algorithms include K-Nearest Neighbors, Decision Tree, Random Forest.

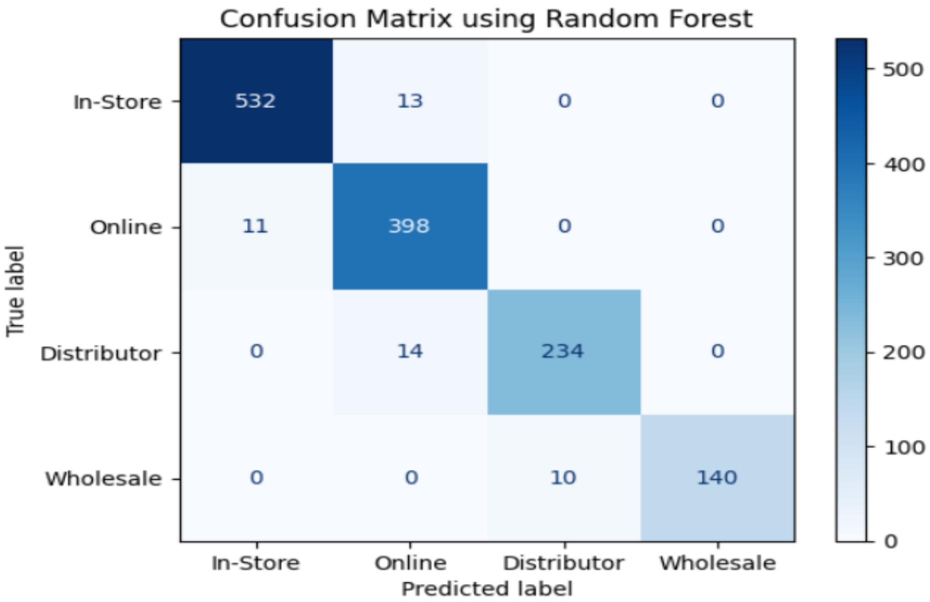
Progress:

We have implemented 3 algorithms on our data after pre-processing:

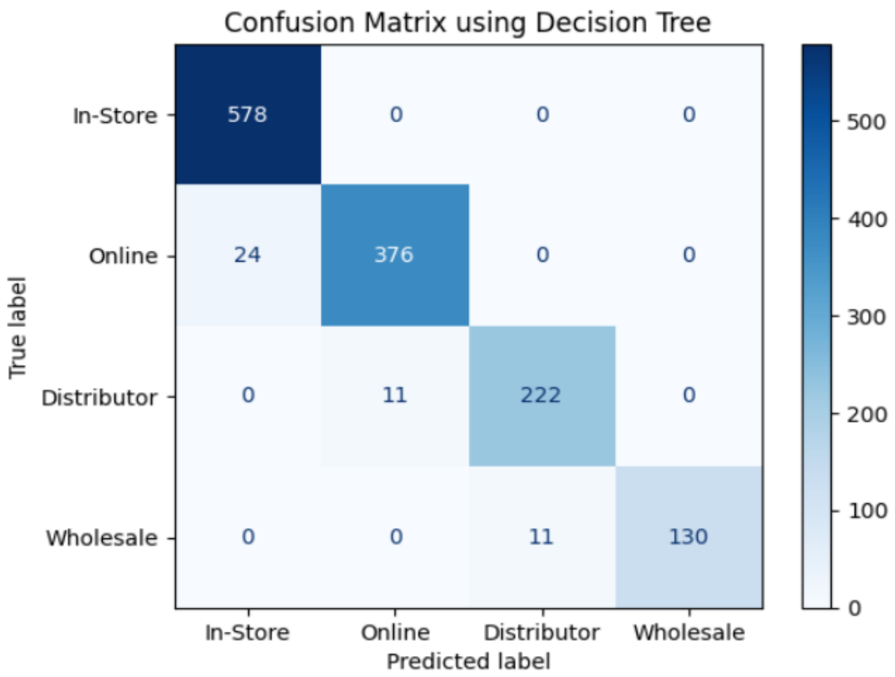
- Random Forest Classification
- Decision Tree Classification
- K Nearest Neighbors Classification

Preliminary Results:

1. Random Forest Classification:
    - Accuracy: 0.96
    - Precision: 0.96
    - Recall: 0.95
    - F1 Score: 0.95
    - Confusion Matrix:

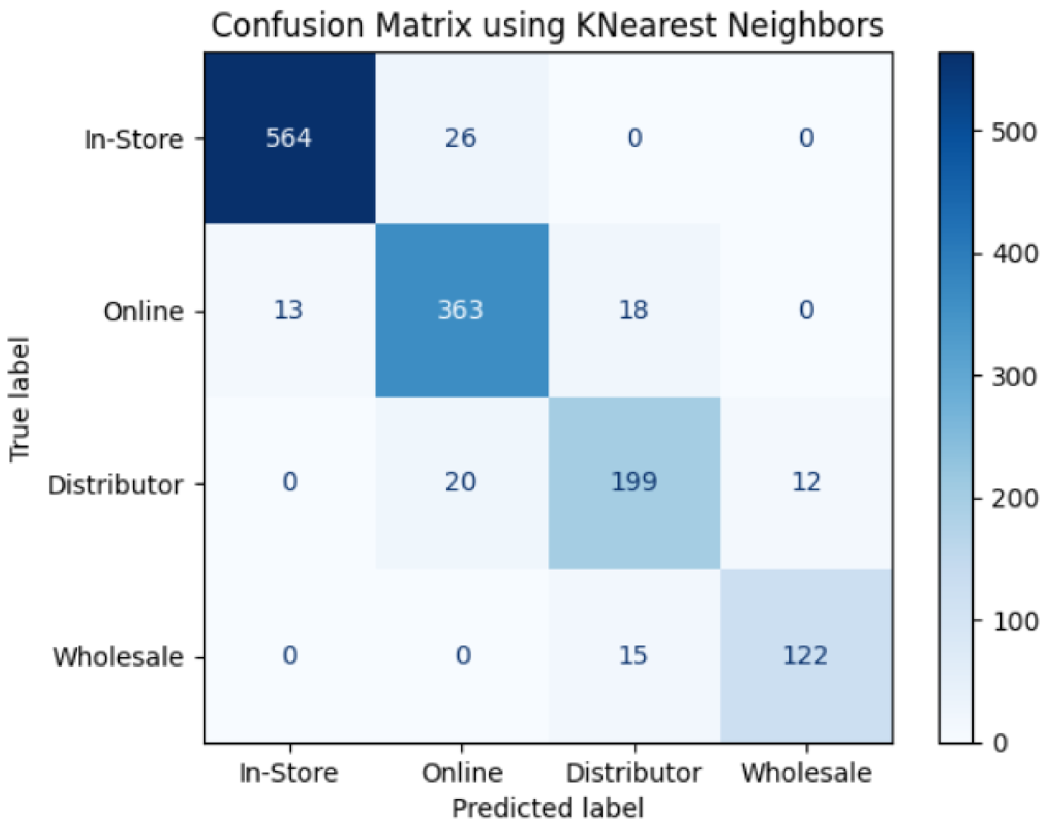


2. Decision Tree Classification:
    - Accuracy: 97
    - Precision: 97
    - Recall: 95
    - F1 Score: 96
    - Confusion Matrix:



	In-Store	Online	Distributor	Wholesale
In-Store	578	0	0	0
Online	24	376	0	0
Distributor	0	11	222	0
Wholesale	0	0	11	130

3. K Nearest Neighbors Classification:
   - Accuracy: 0.92
   - Precision: 0.91
   - Recall: 0.91
   - F1 Score: 0.91
   - Confusion Matrix:



Confusion Matrix using KNearest Neighbors

True label \ Predicted label	In-Store	Online	Distributor	Wholesale
In-Store	564	26	0	0
Online	13	363	18	0
Distributor	0	20	199	12
Wholesale	0	0	15	122

Schedule:

- Use other classification Models (23rd Oct – 29th Oct)
- Tuning Classification Models (30th Oct – 5th Nov)
- Final Testing and Validation (6th Nov – 12th Nov)
- Presentation Preparation (13th Nov – 26th Nov)
- Project Presentation (27th Nov)