Analytics in Retail Sales: A Machine Learning Approach to Enhancing Customer Experience

Team Members :

Nithya Madhulapally

Bhushan Karande

Dakshayeni Bujunuru

AGENDA

- Introduction
- Data Overview
- Data Details
- Data Pre-Processing - Bhushan Karande/Dakshayeni Bujunuru
- Exploratory Data Analysis - Dakshayeni Bujunuru/Bhushan Karande
- Solutions
  - Linear Regression - Dakshayeni Bujunuru
  - Artificial Neural Network - Bhushan Karande
  - K-means Clustering - Nithya Madhulapally
- Conclusion
- Future Directions

Introduction

Problem Statement

Leverage transactional data from the retail store to derive insights and actionable strategies for store performance optimization, product positioning, and customer engagement for a retail chain.

Objective

- To predict sales of a future item in a particular store using linear regression
- Product segmentation that would give products often bought together
- To predict which categories or items a customer will likely purchase.

DATA OVERVIEW

- Source: Kaggle
- https://www.kaggle.com/datasets/nishchay331/retail-store/data
- Represents the data on sales from a Retail clothes store from the year 2018 to 2022.
- 10 Features & 8454383 data points.

DATA DETAILS

About Dataset columns -

- Transaction data - Date on which the transaction made
- Bill ID- ID number of transaction
- User ID- ID number of user who made the transaction
- Line item amount - amount of item
- Item description - Description of item
- Inventory category - category of item
- Color- colour of item
- Size- size of item
- Zone_name- name of zones where stores are located
- Store name - name of stores according to the zones

DATA PRE-PROCESSING / ETL

- Duplicate Removal
  - Ensured uniqueness of data by removing duplicate entries.
- Feature Completeness
  - Excluded datapoints with insufficient feature data to maintain data quality.
- Enhancement of Categorical Data
  - Augmented categorical information by extracting product details from descriptions.
- Consistency Check
  - Eliminated data that did not align with the intended feature representations.
- Null Value Management
  - Addressed missing values by assigning 'Unknown' where applicable.
- Focused Data Selection
  - Dropped rows lacking essential information on inventory category/product for targeted analysis.

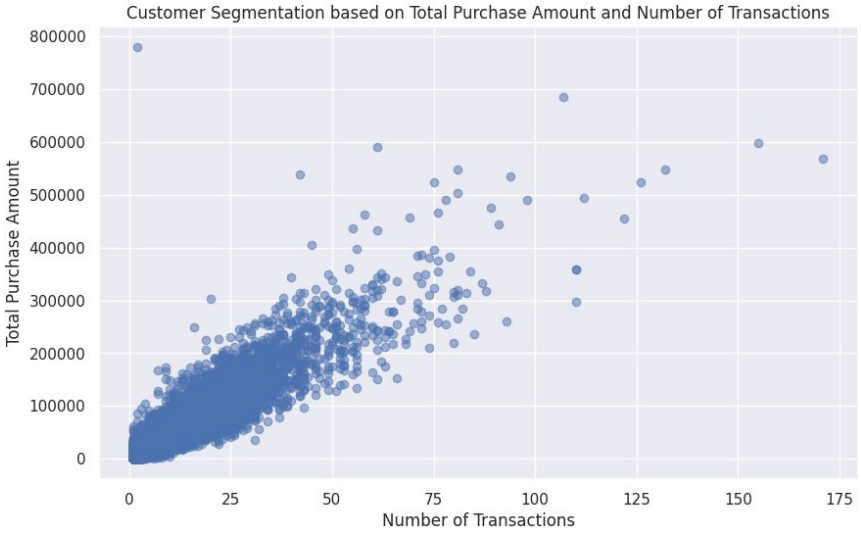
DATA PRE-PROCESSING / ETL

Handling Duplicates :

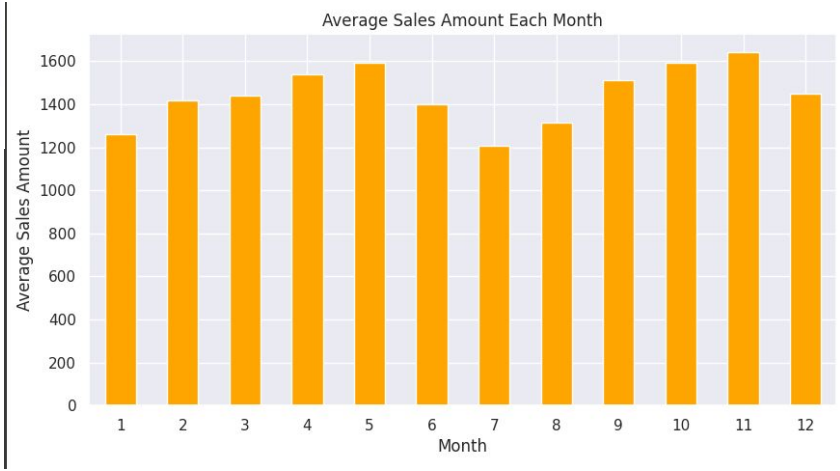
We have 1898076 duplicate rows . Dropped the duplicate records

```
Number of rows before removing duplicates: 8454383

Number of duplicate rows: 1898076

Preview of duplicate rows:
     user_id    bill_id  line_item_amount  bill_discount transaction_date  \
1  519644808  741961800             559.6            0.0       2022-04-20
4   12626591  768369011             519.6            0.0       2022-10-09
5   12626591  768369011             519.6            0.0       2022-10-09
6   12626591  768369011             519.6            0.0       2022-10-09
7   12626591  768369011             519.6            0.0       2022-10-09

                                 description inventory_category  \
1         MBL ITA16BLT004 Regular Casual Tan 32          MENS BELT
4  MBL ITA16BLT001 Regular Casual Dark Brown 32          MENS BELT
5  MBL ITA16BLT001 Regular Casual Dark Brown 32          MENS BELT
6  MBL ITA16BLT001 Regular Casual Dark Brown 32          MENS BELT
7  MBL ITA16BLT001 Regular Casual Dark Brown 32          MENS BELT

       colour size zone_name  store_name  year
1         Tan   32     North  North_6023  2022
4  Dark Brown   32     South  South_6017  2022
5  Dark Brown   32     South  South_6017  2022
6  Dark Brown   32     South  South_6017  2022
7  Dark Brown   32     South  South_6017  2022

Number of rows after removing duplicates: 6556307
```

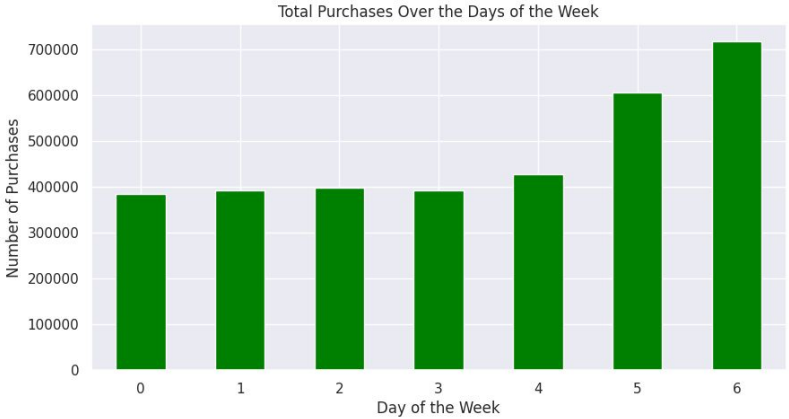
Exploratory Data Analysis

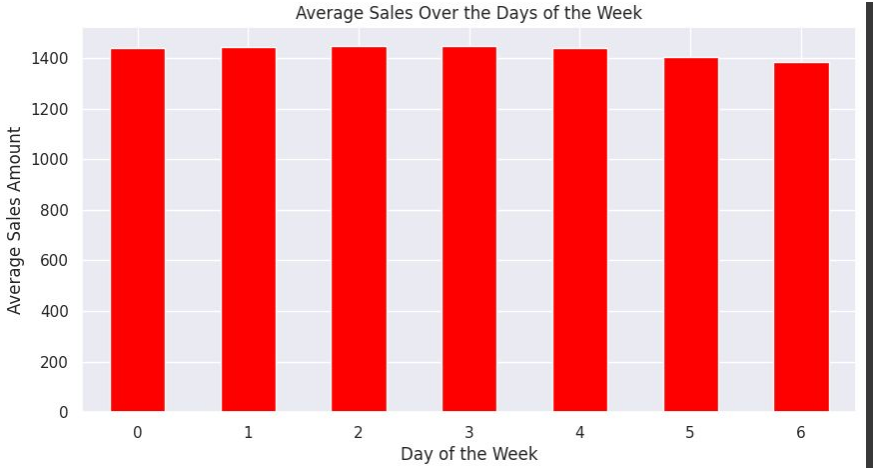


Exploratory Data Analysis



Exploratory Data Analysis





Linear Regression

- Used Linear Regression for predicting future sales.
- Divided our grouped data into 80% for training and 20% for testing and evaluating the model.
- Encoded the categorical features, as linear regression can’t handle categorical variables.
- Scaled the numerical features for better performance of model.
- Used PolynomialFeatures for best performance.

Linear Regression

```
[86]: (2.3448192519191067, -0.056848387699349745)
```

```python
# Split data into training and test sets
# Assuming 80-20 split for training and test
train_size = int(0.8 * len(X))
X_train, X_test = X[:train_size], X[train_size:]
y_train, y_test = y[:train_size], y[train_size:]

# Initialize the Linear Regression model
model = LinearRegression()

# Train the model
model.fit(X_train, y_train)

# Make predictions
y_pred = model.predict(X_test)

# Calculate metrics
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

mse, r2
```

ARTIFICIAL NEURAL NETWORK

- Leverage ANNs to gain insights into the purchasing behaviors of frequent customers.
- Employed bill data as input for the neural network, incorporating product details such as color, size, and total amount.
- Recognized the complexity of customer purchase patterns, often influenced by multiple interrelated factors.
- Utilized ANNs for their proficiency in capturing these nonlinear relationships.

ARTIFICIAL NEURAL NETWORK

- **Model Framework:** Built using the Sequential API.
- **Input Layer**: Comprises 128 neurons, directly receiving input features.
- **Hidden Layer**: A dense layer with 64 neurons for deeper data representation.
- **Activation Function:** ReLU used in the first two layers to address non-linearity.
- **Optimization and Loss Function:** Adam optimizer paired with binary cross-entropy, suitable for multilabel classification issues.
- **Training Parameters**: Set to train for 20 epochs with a batch size of 128.
- Achieved a Hamming loss of 2 * 10^-6, indicating high accuracy in predicting product purchase likelihood.

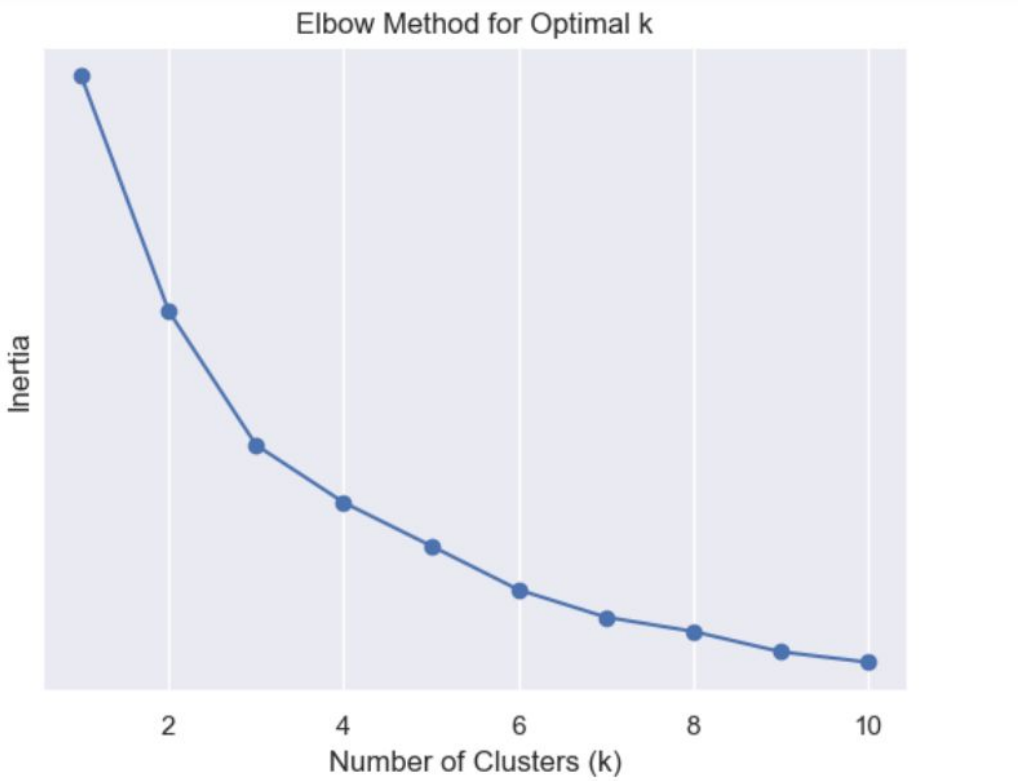
```
# Split the data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = Sequential()
model.add(Dense(128, activation='relu', input_dim=X_train.shape[1]))
model.add(Dense(64, activation='relu'))
model.add(Dense(y_train.shape[1], activation='sigmoid'))
```

```
# Compile the model
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['mean_absolute_error'])

# Train the model
model.fit(X_train, y_train, epochs=20, batch_size=128)

# Evaluate the model
y_pred = model.predict(X_test)
y_pred_binarized = (y_pred > 0.5).astype(int)  # Threshold to determine if a label should be assigned
```

K-MEANS

- Used K-means algorithm to categorize data into k groups by minimizing the variance within each cluster.
- K-means for product segmentation
- We have experimented with multiple number of clusters and concluded that the number of clusters increases as the inertia decreases.

K-MEANS


Elbow Method for Optimal k
Inertia
2
4
6
8
10
Number of Clusters (k)

FUTURE DIRECTIONS

Dynamic Pricing Models: Develop pricing strategies based on customer behavior, inventory levels, and market trends using machine learning models. This could involve reinforcement learning to dynamically adjust prices in real-time.

Anomaly Detection: Implement anomaly detection to identify unusual transactions that could indicate fraud, data entry errors, or operational inefficiencies.

Lifetime Value Prediction: Utilize historical transaction data to model and predict the lifetime value of customers, aiding in focusing retention efforts on high-value customers.

CONCLUSION

- Our machine learning models, encompassing regression, clustering, and artificial neural networks, have demonstrated the ability to predict future sales, segment products, and anticipate customer purchasing patterns.
- The insights derived from our analysis empower the retail chain to optimize store performance, refine product positioning, and enhance customer engagement, leading to an improved shopping experience and increased profitability.

THANK YOU

Zoom link - to the video recording