PROJECT TITLE

MARKET BASKET INSIGHTS

Problem statement

The problem of market basket insights aims to uncover patterns and relationships within consumer shopping data to better understand how products are purchased together. This analysis is crucial for various stakeholders, including retailers, e-commerce platforms, and marketing teams, as it provides valuable information for optimizing sales, marketing strategies, and overall business operations.

Key Elements of the Problem:

1. Transaction Data: The core of this problem is transactional data, which includes records of customer purchases over a specific period. Each transaction lists the products bought and their quantities.

2. Association Patterns: The primary objective is to identify association patterns among products. This involves finding out which products tend to be bought together frequently and quantifying the strength of these associations.

3. Business Goals: -Product Recommendations: Retailers want to recommend complementary products to customers, increasing sales and enhancing user experience.

-Inventory Management: Efficiently manage inventory by understanding which products have correlated demand, reducing wastage, and optimizing stock levels.

- Marketing and Promotion: Targeted marketing campaigns and cross-selling strategies based on market basket insights can boost revenue.

- Store Layout and Product Placement: For physical stores, insights can help arrange products in-store to encourage additional purchases.

4. Data Challenges: Challenges include handling large volumes of transaction data efficiently, dealing with noise and outliers, and ensuring data privacy and security compliance

5. Algorithmic Approaches: Market basket insights are typically obtained through techniques like association rule mining, with algorithms such as Apriori and FP-Growth. Machine learning models may also be employed for more advanced analysis.

6. Visualization and Interpretation: Beyond discovering associations, presenting the results in an understandable and actionable format is crucial. Data visualization tools and dashboards play a significant role in this aspect.

7. Continuous Improvement: Shopping patterns evolve, so the solution should be dynamic and adaptable, with the ability to continuously monitor and update insights

Design thinking:

1. Empathize: Understand the User and the Problem - Begin by empathizing with the users (e.g., retailers, e-commerce platforms) to understand their goals and challenges. - Conduct interviews, surveys, or observational research to gather insights about their needs and pain points related to market basket analysis.

2. Define: Frame the Problem - Clearly define the problem by creating a problem statement or user story that encapsulates the challenges and goals of the users. - For example, "How might we improve product recommendations based on market basket insights to increase sales?"

3. Ideate: Generate Ideas - Brainstorm creative solutions for collecting and analyzing market basket data. - Encourage cross-functional teams to collaborate and come up with innovative approaches. - Consider using techniques like ideation workshops and mind mapping.

4. Prototype: Build a Solution - Create a prototype or proof of concept for your market basket analysis system. - This could involve designing a data pipeline for collecting transaction data, implementing algorithms for association rule mining, and developing a user interface for visualizing insights.

5. Test: Gather Feedback - Test your prototype with real users to gather feedback and iterate on your solution. - Analyze the effectiveness of your market basket insights and whether they meet user needs.

6. Implement: Deploy the Solution - Once you've refined your solution based on user feedback, implement it in a real-world setting. - Monitor its performance and make necessary adjustments.

7. Iterate: Continuously Improve: - Market basket insights are not a one-time task; they require ongoing analysis and improvement. - Continuously collect and analyze data to identify changing consumer behaviors and preferences.

8. Scale: Expand and Optimize: - As your market basket insights solution proves its value, consider scaling it to serve a larger audience or expanding its capabilities. - Optimize your solution to handle larger datasets and provide more advanced insights.

In the Market Basket Insights Project, the combination of ensemble methods and deep learning can significantly enhance the analysis of transaction data, customer behavior, and market trends. Here's a detailed explanation of how these two approaches can be used in synergy:

1. Ensemble Methods:

Ensemble methods, such as Random Forest, Gradient Boosting, or AdaBoost, can be employed for various tasks within the project:

a. Market Basket Analysis: Ensemble methods can be used to identify frequent itemsets and association rules. For instance, using the Apriori algorithm in combination with ensemble methods can help discover interesting item combinations and associations in customer transactions.

b. Customer Segmentation: By applying ensemble clustering techniques like K-Means or Hierarchical Clustering, the customer base can be segmented into different groups based on purchase patterns, enabling targeted marketing strategies.

c. Recommendation Systems: Ensemble methods can be used to create robust recommendation engines that suggest complementary or frequently co-purchased products to customers.

2. Deep Learning:

Deep learning models, particularly Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks, can be used to uncover intricate insights within the project:

a.Sequence Analysis: LSTMs are well-suited for modeling sequential data, which is inherent in customer purchase histories. By training LSTM models on this data, you can predict the next item in a customer's basket or capture temporal dependencies within purchase sequences.

b.Predictive Modeling: Deep learning models can be employed to predict customer behavior, such as the likelihood of churning or the probability of making specific purchases.

c. Anomaly Detection: LSTM networks can be used to detect unusual purchase patterns or anomalies in customer transactions, which can be valuable for fraud detection or identifying unique customer behavior.

3.Integration:

Ensemble methods and deep learning can be integrated for even more robust insights:

a.Feature Engineering: Features extracted from ensemble methods, such as item association rules, can be used as input features for deep learning models to improve their accuracy and predictive power.

b. Hybrid Models: Hybrid models can be created, where ensemble methods and deep learning models work together to provide comprehensive insights. For instance, ensemble methods might identify interesting item associations, and deep learning models can then predict which of these associations are most likely to be adopted by individual customers.

By combining the strengths of ensemble methods for rule-based analysis and deep learning for sequence and pattern recognition, the Market Basket Insights Project can unlock valuable insights that aid in optimizing inventory management, personalizing marketing campaigns, and improving the overall retail experience.

Processing the dataset in the Market Basket Insights Project involves several essential steps:

1.Data Collection:

- Gather transactional data from various retail sources, including purchase histories, customer profiles, and item information. This data may contain details like transaction timestamps, customer IDs, and purchased items.

2. Data Cleaning:

- Handle missing values, duplicates, or erroneous entries in the dataset to ensure data quality. Remove or impute missing data appropriately.

3. Data Transformation:

- Transform the raw data into a format suitable for analysis. This might involve converting timestamps to a standardized format, normalizing numerical data, and encoding categorical variables.

4.Feature Engineering:

- Create new features or modify existing ones to enhance the dataset's informativeness. Features could include purchase frequency, basket size, popular item categories, and customer purchase history.

5.Sequence Generation:

- Organize the data into sequences representing individual customer purchase histories. Each sequence comprises a chronological list of items purchased by a customer.

6. Padding Sequences:

- Standardize the length of sequences by padding them with a special token or zero-padding. This ensures uniform input dimensions for model training.

7. Train -Test Split:

- Divide the dataset into training and testing sets to evaluate model performance. The training set is used to train the model, while the testing set is reserved to assess its predictive accuracy.

8. One-Hot Encoding:

- Encode categorical variables, such as item IDs, using one-hot encoding to represent them as binary vectors. Each item is represented as a unique binary pattern.

9. Model Training Data Preparation:

- Structure the data to feed into the chosen deep learning model (e.g., LSTM). Prepare input-output pairs where the input is a sequence of purchases, and the output is the next item in the sequence.

These processed and prepared datasets are then fed into the chosen deep learning architecture for training and subsequent analysis to derive valuable insights about customer behavior, purchase patterns, and market trends.

LOADING AND PREPROCESSING THE DATASET

```
import pandas as pd
data=pd.read_csv('assignment1_data.csv',on_bad_lines='skip')
print(data)
```

```
       BillNo                         Itemname  Quantity  \
0      536365   WHITE HANGING HEART T-LIGHT HOLDER         6
1      536365                WHITE METAL LANTERN         6
2      536365     CREAM CUPID HEARTS COAT HANGER         8
3      536365  KNITTED UNION FLAG HOT WATER BOTTLE         6
4      536365       RED WOOLLY HOTTIE WHITE HEART.         6
...       ...                                ...       ...
99994  545059          RIBBON REEL HEARTS DESIGN         1
99995  545059         RIBBON REEL STRIPES DESIGN         1
99996  545059           BROWN CHECK CAT DOORSTOP         1
99997  545059                    CARD PARTY GAMES        12
99998  545059                         WICKER STAR         4

                  Date  Price  CustomerID         Country
0      01-12-2010 08:26   2.55     17850.0  United Kingdom
1      01-12-2010 08:26   3.39     17850.0  United Kingdom
2      01-12-2010 08:26   2.75     17850.0  United Kingdom
3      01-12-2010 08:26   3.39     17850.0  United Kingdom
4      01-12-2010 08:26   3.39     17850.0  United Kingdom
...                 ...    ...         ...             ...
99994  27-02-2011 13:04   1.65     14157.0  United Kingdom
99995  27-02-2011 13:04   1.65     14157.0  United Kingdom
99996  27-02-2011 13:04   4.25     14157.0  United Kingdom
99997  27-02-2011 13:04   0.42     14157.0  United Kingdom
99998  27-02-2011 13:04   2.10     14157.0  United Kingdom

[99999 rows x 7 columns]
```

```
data.describe()
```

Out[4]:

	BillNo	Quantity	Price	CustomerID
count	99999.000000	99999.000000	99999.000000	64769.000000
mean	540650.859369	9.886779	4.220347	15384.034229
std	2501.404483	239.273822	45.728639	1768.882343
min	536365.000000	-2600.000000	0.000000	12346.000000
25%	538379.000000	1.000000	1.250000	13875.000000
50%	540646.000000	3.000000	2.460000	15356.000000
75%	542711.000000	9.000000	4.210000	17059.000000
max	545059.000000	74215.000000	13541.330000	18283.000000

```
pd.set_option('display.max_columns', None)
pd.set_option('display.expand_frame_repr', False)
print(data)
```

```
      BillNo                         Itemname  Quantity            Date  Price  CustomerID         Country
0     536365   WHITE HANGING HEART T-LIGHT HOLDER         6  01-12-2010 08:26   2.55     17850.0  United Kingdom
1     536365                WHITE METAL LANTERN         6  01-12-2010 08:26   3.39     17850.0  United Kingdom
2     536365     CREAM CUPID HEARTS COAT HANGER         8  01-12-2010 08:26   2.75     17850.0  United Kingdom
3     536365  KNITTED UNION FLAG HOT WATER BOTTLE         6  01-12-2010 08:26   3.39     17850.0  United Kingdom
4     536365     RED WOOLLY HOTTIE WHITE HEART.         6  01-12-2010 08:26   3.39     17850.0  United Kingdom
...      ...                                ...       ...              ...    ...         ...             ...
99994  545059          RIBBON REEL HEARTS DESIGN         1  27-02-2011 13:04   1.65     14157.0  United Kingdom
99995  545059         RIBBON REEL STRIPES DESIGN         1  27-02-2011 13:04   1.65     14157.0  United Kingdom
99996  545059           BROWN CHECK CAT DOORSTOP         1  27-02-2011 13:04   4.25     14157.0  United Kingdom
99997  545059                  CARD PARTY GAMES        12  27-02-2011 13:04   0.42     14157.0  United Kingdom
99998  545059                       WICKER STAR         4  27-02-2011 13:04   2.10     14157.0  United Kingdom

[99999 rows x 7 columns]
```

```
data = pd.read_csv('assignment1_data.csv')
null_values = data.isnull().any().any()
if null_values:
    print("There are null values in the dataset.")
else:
    print("There are no null values in the dataset.")
```

```
There are null values in the dataset.
```

data.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99999 entries, 0 to 99998
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   BillNo      99999 non-null  int64  
 1   Itemname    99698 non-null  object 
 2   Quantity    99999 non-null  int64  
 3   Date        99999 non-null  object 
 4   Price       99999 non-null  float64
 5   CustomerID  64769 non-null  float64
 6   Country     99999 non-null  object 
dtypes: float64(2), int64(2), object(3)
memory usage: 5.3+ MB
```

print(data.dropna())

	BillNo	Itemname	Quantity	Date	Price	CustomerID	Country
0	536365	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	2.55	17850.0	United Kingdom
1	536365	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
2	536365	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.75	17850.0	United Kingdom
3	536365	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
4	536365	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
...	...	...	...	...	...	...	...
99994	545059	RIBBON REEL HEARTS DESIGN	1	27-02-2011 13:04	1.65	14157.0	United Kingdom
99995	545059	RIBBON REEL STRIPES DESIGN	1	27-02-2011 13:04	1.65	14157.0	United Kingdom
99996	545059	BROWN CHECK CAT DOORSTOP	1	27-02-2011 13:04	4.25	14157.0	United Kingdom
99997	545059	CARD PARTY GAMES	12	27-02-2011 13:04	0.42	14157.0	United Kingdom
99998	545059	WICKER STAR	4	27-02-2011 13:04	2.10	14157.0	United Kingdom

[64769 rows x 7 columns]

print(data.notnull().sum())

```
BillNo        99999
Itemname      99698
Quantity      99999
Date          99999
Price         99999
CustomerID    64769
Country       99999
dtype: int64
```

data.isnull().sum()

```
Out[13]: BillNo            0
         Itemname         301
         Quantity           0
         Date               0
         Price              0
         CustomerID     35230
         Country            0
         dtype: int64
```

data.dropna(subset=['Itemname', 'CustomerID'], inplace=True)

data.isnull().sum()

```
Out[16]: BillNo         0
         Itemname       0
         Quantity       0
         Date           0
         Price          0
         CustomerID     0
         Country        0
         dtype: int64
```

PERFORMING ASSOCIATION ANALYSIS AND GENERATING INSIGHTS

```
# Import necessary libraries
import pandas as pd
from mlxtend.frequent_patterns import apriori
from mlxtend.frequent_patterns import association_rules

# Load the dataset (replace 'dataset.csv' with your dataset)
data = pd.read_csv('assignment1_data.csv')

# Data preprocessing
basket = (data.groupby(['CustomerID', 'Itemname'])['Itemname']
        .count().unstack().reset_index().fillna(0)
        .set_index('CustomerID'))
```

```
# Convert to a binary format (0/1 encoding)
def encode_units(x):
    if x <= 0:
        return 0
    if x >= 1:
        return 1

basket_sets = basket.applymap(encode_units)

# Generate frequent itemsets using Apriori
frequent_itemsets = apriori(basket_sets, min_support=0.05, use_colnames=True)

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

# Display the association rules
print("Association Rules:")
print(association_rules)

# Generating Insights
# You can analyze the association rules and provide insights based on high-confidence rules and lift values
confidence_threshold = 0.7
lift_threshold = 1.0

meaningful_rules = association_rules[
    (association_rules['confidence'] > confidence_threshold) &
    (association_rules['lift'] > lift_threshold)
]

# Display meaningful association rules
print("Meaningful Association Rules:")
print(meaningful_rules)
```

Association Rules:

	antecedents	consequents	antecedent support
0	(SET OF 6 SPICE TINS PANTRY DESIGN)	(SET OF 3 CAKE TINS PANTRY DESIGN)	0.092129
1	(SET OF 3 CAKE TINS PANTRY DESIGN)	(SET OF 6 SPICE TINS PANTRY DESIGN)	0.143990
2	(JAM MAKING SET PRINTED)	(JAM MAKING SET WITH JARS)	0.127517
3	(JAM MAKING SET WITH JARS)	(JAM MAKING SET PRINTED)	0.109823
4	(SET OF 3 CAKE TINS PANTRY DESIGN)	(JAM MAKING SET WITH JARS)	0.143990
5	(JAM MAKING SET WITH JARS)	(SET OF 3 CAKE TINS PANTRY DESIGN)	0.109823
6	(SET OF 3 CAKE TINS PANTRY DESIGN)	(RECIPE BOX PANTRY YELLOW DESIGN)	0.143990
7	(RECIPE BOX PANTRY YELLOW DESIGN)	(SET OF 3 CAKE TINS PANTRY DESIGN)	0.087858
8	(WHITE HANGING HEART T-LIGHT HOLDER)	(RED HANGING HEART T-LIGHT HOLDER)	0.198292
9	(RED HANGING HEART T-LIGHT HOLDER)	(WHITE HANGING HEART T-LIGHT HOLDER)	0.104942
10	(WHITE HANGING HEART T-LIGHT HOLDER)	(CANDLEHOLDER PINK HANGING HEART)	0.198292
11	(CANDLEHOLDER PINK HANGING HEART)	(WHITE HANGING HEART T-LIGHT HOLDER)	0.056132
12	(WHITE HANGING HEART T-LIGHT HOLDER)	(HEART OF WICKER SMALL)	0.198292
13	(WHITE HANGING HEART T-LIGHT HOLDER)	(HEART OF WICKER LARGE)	0.148871
14	(WHITE HANGING HEART T-LIGHT HOLDER)	(HEART OF WICKER SMALL)	0.128127
15	(HEART OF WICKER LARGE)	(HEART OF WICKER SMALL)	0.198292
16	(HEART OF WICKER LARGE)	(HEART OF WICKER LARGE)	0.128127
17	(HEART OF WICKER SMALL)	(HEART OF WICKER LARGE)	0.148871

	consequent support	support	confidence	lift	leverage	conviction	zhangs metric
0	0.143990	0.070775	0.768212	5.335167	0.057509	3.693071	
1	0.092129	0.070775	0.491525	5.335167	0.057509	1.785479	
2	0.109823	0.056132	0.440191	4.008187	0.042128	1.590145	
3	0.127517	0.056132	0.511111	4.008187	0.042128	1.784625	
4	0.109823	0.051861	0.360169	3.279543	0.036047	1.391270	
5	0.143990	0.051861	0.472222	3.279543	0.036047	1.621913	
6	0.087858	0.051861	0.360169	4.099429	0.039210	1.425599	
7	0.143990	0.051861	0.590278	4.099429	0.039210	2.089244	
8	0.104942	0.072605	0.366154	3.489106	0.051796	1.412106	
9	0.198292	0.072605	0.691860	3.489106	0.051796	2.601771	
10	0.056132	0.050031	0.252308	4.494916	0.038900	1.262375	
11	0.198292	0.050031	0.891304	4.494916	0.038900	7.375717	
12	0.148871	0.067114	0.338462	2.273518	0.037594	1.286590	
13	0.198292	0.067114	0.450820	2.273518	0.037594	1.459827	
14	0.198292	0.064674	0.504762	2.545553	0.039267	1.618834	
15	0.128127	0.064674	0.326154	2.545553	0.039267	1.293876	
16	0.148871	0.089689	0.700000	4.702049	0.070614	2.837096	
17	0.128127	0.089689	0.602459	4.702049	0.070614	2.193165	

```
17          0.128127  0.089689    0.002459  4.702049  0.070614    2.195165

    zhangs_metric
0        0.895022
1        0.949247
2        0.860201
3        0.843103
4        0.811999
5        0.780833
6        0.883242
7        0.828888
8        0.889842
9        0.797036
10       0.969837
11       0.823766
12       0.698699
13       0.658130
14       0.696384
15       0.757330
16       0.903029
17       0.925038
Meaningful Association Rules:
                          antecedents                          consequents  \
0   (SET OF 6 SPICE TINS PANTRY DESIGN)    (SET OF 3 CAKE TINS PANTRY DESIGN)
11    (CANDLEHOLDER PINK HANGING HEART)  (WHITE HANGING HEART T-LIGHT HOLDER)

    antecedent support  consequent support   support  confidence      lift  \
0             0.092129            0.143990  0.070775    0.768212  5.335167
11            0.056132            0.198292  0.050031    0.891304  4.494916

    leverage  conviction  zhangs_metric
0   0.057509    3.693071       0.895022
11  0.038900    7.375717       0.823766
```

```
import pandas as pd
import plotly.express as px
import plotly.io as pio
import plotly.graph_objects as go
pio.templates.default = "plotly_white"

data = pd.read_csv("assignment1_data.csv")
print(data.head())
```

```
   BillNo                             Itemname  Quantity         Date  \
0  536365   WHITE HANGING HEART T-LIGHT HOLDER         6  01-12-2010 08:26
1  536365                  WHITE METAL LANTERN         6  01-12-2010 08:26
2  536365       CREAM CUPID HEARTS COAT HANGER         8  01-12-2010 08:26
3  536365  KNITTED UNION FLAG HOT WATER BOTTLE         6  01-12-2010 08:26
4  536365       RED WOOLLY HOTTIE WHITE HEART.         6  01-12-2010 08:26

   Price  CustomerID         Country
0   2.55     17850.0  United Kingdom
1   3.39     17850.0  United Kingdom
2   2.75     17850.0  United Kingdom
3   3.39     17850.0  United Kingdom
4   3.39     17850.0  United Kingdom
```

```
fig = px.histogram(data, x='Itemname',
                   title='Item description')
fig.show()
```

Item description
count
400
200
0
HEART BUTTONS JEWELLERY BOX
GOLD/M.O.P PENDANT ORBIT NECKLACE
ASSORTED SHAPES PHOTO CLIP SILVER
found
BOX/12 CHICK & EGG IN BASKET
BLACK CRYSTAL DROP EARRINGS
NEW BAROQUE JEWELLERY BOX
PINK FLOCK GLASS CANDLEHOLDER
GREY ACRYLIC FACETED BANGLE
ECONOMY HOLIDAY PURSE
SMALL KITCHEN FLOWER POTS PLAQUE
DROP EARRINGS W FLOWER & LEAF
GREEN ENAMEL FLOWER RING
FILIGREE DIAMANTE EARRINGS
HOUSE SHAPE PENCIL SHARPENER
PINK CHICK EGG WARMER + EGG CUP
WRAP BLUE RUSSIAN FOLKART
FRENCH BLUE METAL DOOR SIGN 7
PAPERWEIGHT SAVE THE PLANET
CARD SUKI BIRTHDAY
BLUE CIRCLES DESIGN TEDDY
3 WHITE CHOC MORRIS BOXED CANDLES
VINTAGE PHOTO ALBUM PARIS DAYS
SET/4 SKULL BADGES
PAPERWEIGHT VINTAGE PAISLEY
LUSH GREENS RIBBONS
VINTAGE BLUE TINSEL REEL
DIAMANTE BOW BROOCH BLACK COLOUR
GROW YOUR OWN FLOWERS SET OF 3
EMBOSSED GLASS TEALIGHT HOLDER
CHRYSANTHEMUM SKETCHBOOK
3D SHEET OF DOG STICKERS
HOME GARLAND PAINTED ZINC
VINTAGE RED TEATIME MUG
GREEN CHRISTMAS TREE STRING 20LIGHT
FRENCH BLUE METAL DOOR SIGN No
ANTIQUE ALL GLASS CANDLESTICK
LETTER "D" BLING KEY RING
SCOTTIES CHILDRENS APRON
PAPERWEIGHT KINGS CHOICE
GLITTER STAR GARLAND WITH BELLS
NO SINGING METAL SIGN
SET/12 TAPER CANDLES
SWEETHEART CAKESTAND 3 TIER
SET OF 2 TEA TOWELS APPLE AND PEARS
VINTAGE PAISLEY STATIONERY SET
KINGS CHOICE BISCUIT TIN
SET 10 LIGHTS NIGHT OWL
UNION JACK FLAG LUGGAGE TAG
ROMANTIC PINKS RIBBONS
PICTURE DOMINOES
ROSE COTTAGE KEEPSAKE BOX
DOORMAT FANCY FONT HOME SWEET HOME
WOODEN FRAME ANTIQUE WHITE
WHITE HANGING HEART T-LIGHT HOLDER
Itemname