MLUL2

Group-12 Assignment Report

Submitted By:

Student Name	PG ID
Ojasvi Ashish Chauhan	12320060
Mahesh Chandankar	12320037
Kashika Sharda	12320047
Kriti Joshi	12320021
Maulin Shah	12320030

Q1. Recommender System:

Exploratory Data Analysis:

On Performing basic Statistical Analysis, we found that we have 638 unique customers who bought 115 products labelled with 632 unique SKUs.

Insights from Distribution of orders over time Graph:

1. **Seasonal Peaks**: Noticeable spikes in orders around May 2012 and mid-2013, suggesting seasonal demand or promotional effects.
2. **Growth Trend**: Increasing order volume from early 2012 to mid-2013, indicating growth in customer base or successful marketing.
3. **Volatility**: High fluctuation in daily orders, showing variability in customer purchasing behavior.
4. **Recent Decline**: Decrease in order consistency towards the end of 2014 and early 2015, possibly due to market saturation or competition.

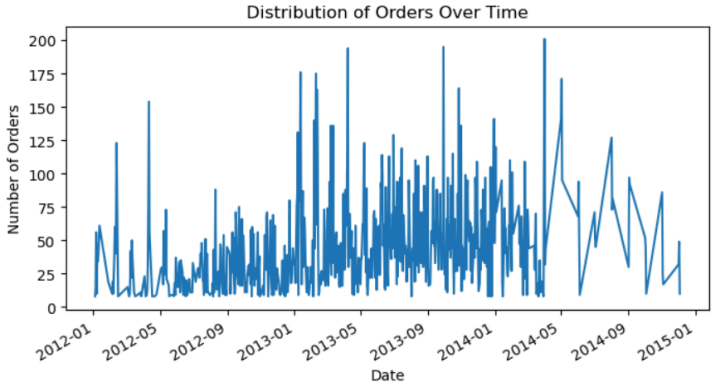

Figure 1

Insights from Top 10 frequent ordered products Graph:

1. **Top Products**: "Other Dals" and "Whole Spices" are the most frequently ordered products, each with over 2000 orders.
2. **Popular Categories**: Pulses (dals) and spices dominate the top product categories.
3. **Diverse Preferences**: Besides dals and spices, beans, other vegetables, root vegetables, organic fruits & vegetables, and specific items like "Moong Dal" and "Toor Dal" are also popular.
4. **Demand Concentration**: The top 10 products have a high concentration of orders, indicating these are staple or frequently used items for customers.

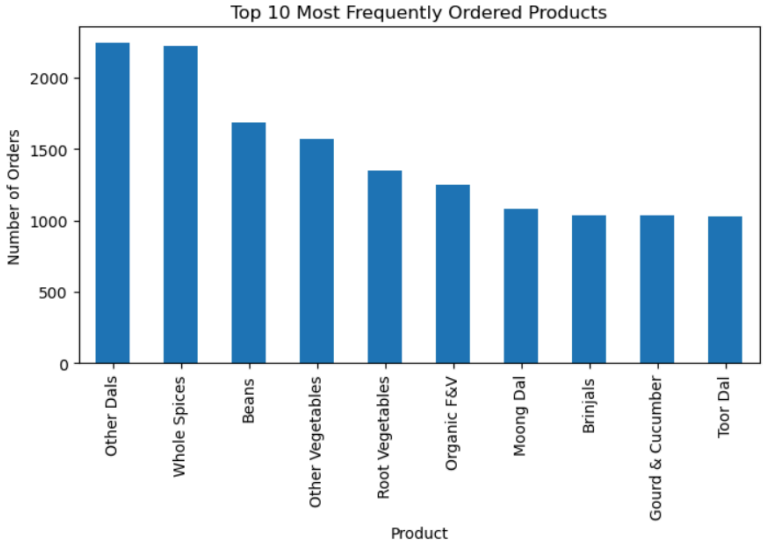

Figure 2

Insights from the Distribution of Number of Orders for Top 10 Customers:

1. **Top Customer**: The customer "SWOZECO" is the most active, with over 350 orders, significantly higher than the rest.
2. **High Engagement**: The second and third most active customers, "SWLLREW" and "SWOEOHC," have around 250 orders each.
3. **Order Distribution**: The top 10 customers have a relatively high number of orders, ranging from approximately 200 to 350, indicating a strong engagement with the platform.
4. **Loyal Customer Base**: The consistent order volume among the top 10 customers suggests a loyal customer base that regularly places orders.

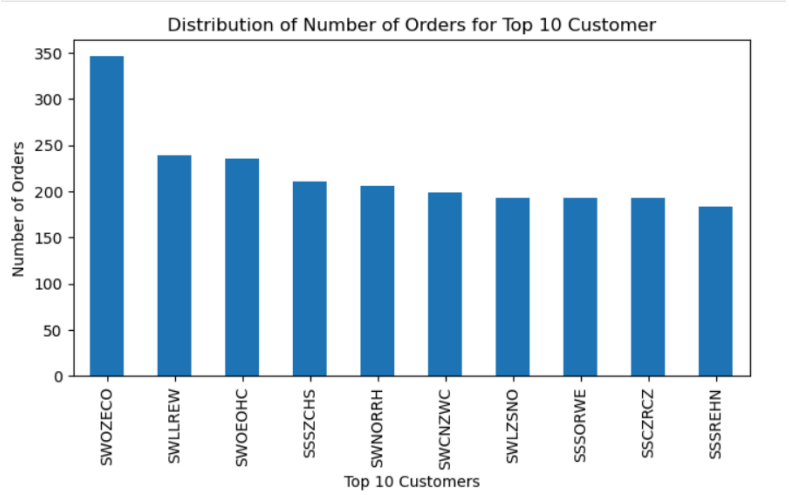

Figure 3

Top 10 SKU Combinations bought together most frequently:

- The most frequently bought together item pair is SKU 15668379 and SKU 15668460, which were purchased together 156,683 times.
- Other frequently bought together items include SKU 15668379 and SKU 15668688 (purchased together 125,47 times), and SKU 15668688 and SKU 15668460 (purchased together 119,32 times).
- There is no information about the specific items that these SKUs represent.

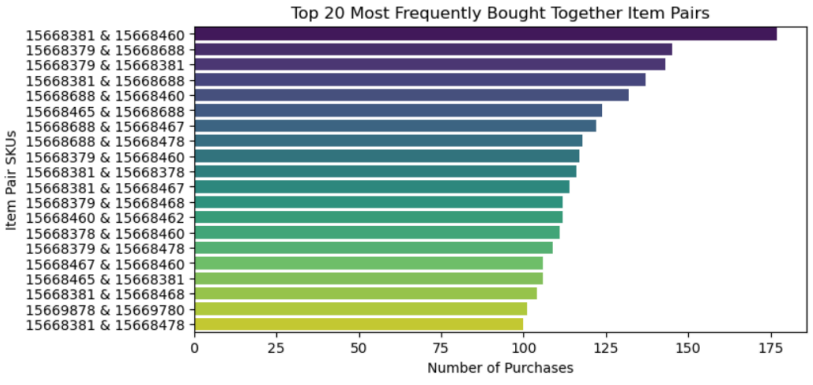

Figure 4

Customer Ordering Patterns Over Time:

This graph shows customer ordering patterns over time:

- Customer SWCRLZN (Blue): Orders primarily in mid-2013, peaking at around 5.

- Customer SWOZECO (Orange): Orders concentrated from early 2012 to mid-2013, with peaks reaching around 20-25.

- Customer SSNCWZO (Green): Orders from early 2013 to mid-2013, peaking at around 5.

- Customer SWOWELS (Red): Orders from early 2013 to mid-2013, peaking at around 5.

- Customer SWNECLZ (Purple): Orders from early 2012 to mid-2012, peaking at around 10.

General Observations:

- Order Concentration: Most ordering activity is from early 2012 to mid-2013.

- Peak Activity: Customer SWOZECO has the highest peak orders, reaching up to 25.

- Variability: Significant variability in the frequency and timing of orders among customers.

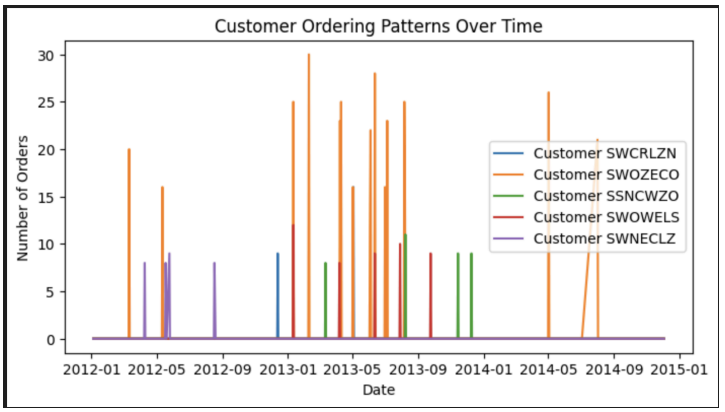

Figure 5

<u>3 Key Insights from above EDA charts</u>:

1. Customer Activity Over Time:

- Order Concentration: Most ordering activity is concentrated from early 2012 to mid-2013.

- Peak Activity: Customer SWOZECO has the highest peak orders, reaching up to 25 orders.

- Variability: There is significant variability in the frequency and timing of orders among customers, indicating different purchasing behaviors and cycles.

2. Top 10 Most Frequently Ordered Products:

- Top Products: "Other Dals" and "Whole Spices" remain the most frequently ordered products, each with over 2000 orders.

- Popular Categories: Pulses (dals) and spices dominate the top product categories.

- Diverse Preferences: Besides dals and spices, beans, other vegetables, root vegetables, organic fruits & vegetables, and specific items like "Moong Dal" and "Toor Dal" are also popular.

- Demand Concentration: The top 10 products have a high concentration of orders, indicating these are staple or frequently used items for customers.

3. Distribution of Orders for Top 10 Customers:

- Top Customer: The customer "SWOZECO" is the most active, with over 350 orders, significantly higher than the rest.

- High Engagement: The second and third most active customers, "SWLLREW" and "SWOEOHC," have around 250 orders each.

- Order Distribution: The top 10 customers have a relatively high number of orders, ranging from approximately 200 to 350, indicating a strong engagement with the platform.

- Loyal Customer Base: The consistent order volume among the top 10 customers suggests a loyal customer base that regularly places orders.

General Observations:

- Order Concentration: Most ordering activity is from early 2012 to mid-2013.

- Peak Activity: Customer SWOZECO has the highest peak orders, reaching up to 25.

- Variability: Significant variability in the frequency and timing of orders among customers.

Clustering:

The dataset does not have enough information to cluster the users in categories based on their past purchases as initial analysis to determine the number of clusters returns a number of clusters close to 600 which means almost every user is its own cluster.

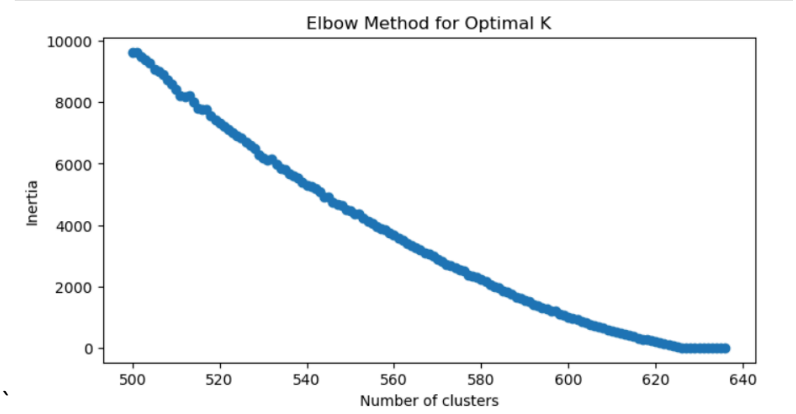

Figure 6

Report Summary Algo 1:

Objective:

The goal is to develop a recommendation system using collaborative filtering to suggest top products (SKUs) for customers based on their historical purchase data.

Methodology:

1. Data Preparation:

- Created a user-item matrix from the training data to represent purchase history.

- Computed cosine similarity for users and items to understand similarities based on their purchase patterns.

2. Recommendation Algorithms:

- Item-Based Collaborative Filtering: Calculates similarity between items and recommends items similar to those the user has previously purchased.

- User-Based Collaborative Filtering: Identifies similar users and recommends items that these users have purchased.

- Hybrid Method: Combines scores from both user-based and item-based approaches to provide recommendations.

3. Evaluation Metrics:

- Evaluated the recommendation algorithms using precision, recall, and F1-score on a test dataset to measure their effectiveness.

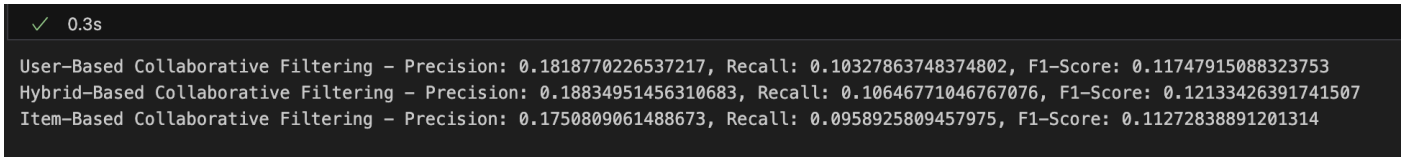

Figure 7

4. Final Recommendations:

- Applied the hybrid method to generate top 5 SKU recommendations for each customer in the test set.
- Generated final recommendations for the last orders subset and saved the results to a CSV file.
- Final recommendations have been saved in `GR12_rec_5_sets.csv.csv`.

Report Summary Algo 2:

Recommendation System Using Collaborative Filtering and SVD

Process Overview:

1. Data Loading and Preprocessing:

- Load datasets into pandas DataFrames.
- Display initial rows to understand data structure.

2. Encoding Categorical Variables:

- Combine datasets for consistent encoding.
- Use `LabelEncoder` to convert `SKU` and `Member` columns to numerical values.

3. Aggregating Data:

- Group historical orders by `Member_encoded` and `SKU_encoded`, counting occurrences.
- Create a user-item interaction matrix with order counts.

4. Applying SVD:

- Convert the matrix to a numpy array.
- Perform Singular Value Decomposition (SVD) with 90 latent factors.
- Reconstruct the matrix to predict ratings.

5. Generating Recommendations:

- For each order in the last order subset, predict ratings.
- Exclude already purchased items.
- Generate top-5 recommendations for each member.

6. Saving Recommendations:

- Create a DataFrame for recommendations.
- Save to `GR12_rec_5_sets2.csv`.

- Recommendations saved in `GR12_rec_5_sets2.csv`.

Evaluation

- Based on testing on Kaggle, the recommendation algorithm using SVD (Collaborative Filtering) demonstrated better accuracy compared to other methods.

Q2. Outlier Detection:

We have used two data sources to get the financial data for 54 Companies related to the IT Industry.

- https://www.moneycontrol.com/
- https://www.screener.in/

The dataset has 7 large Cap, 9 mid cap and 38 small cap companies.

The ratios used for Outlier Detection as per problem statement are:

1 Current ratio: Current Assets / Current Liabilities
2 Debt to equity ratio: Total Debt / Total Equity
3 Inventory turnover: Cost of goods sold / Average value of inventory.
4 Return on equity (ROE): Net income / Total equity
5 Operating margin: (Gross profit - selling and admin exp - other exp - depreciation/ Net sales)
6 Price earnings ratio (P/E): Share price / Earnings per share
7 Book value per share ratio: (Total Equity – Preferred Equity) / Total shares outstanding

Additional ratios used:

8 ROCE: EBT/ (Equity + reserve + borrowings
9 ROA: Net Profit / Total Assets
10 Quick Ratio: (Current assets- inventory)/Current Liabilities
11 Asset Turnover Ratio: Sales/Fixed assets
12 Interest Coverage ratio: EBIT/Interest

Ratio Explanation:

- Current ratio is denoted by CR. This measures a company's ability to pay its short-term obligations with its short-term assets. A higher ratio means the company is in a better position to cover its debts.
- Debt to equity ratio denoted by DE. This ratio shows how much debt a company is using to finance its assets compared to the amount of equity. A lower ratio indicates a more financially stable company.
- Inventory turnover denoted by IT. This measures how often a company sells and replaces its inventory over a period. A higher ratio indicates efficient inventory management and strong sales.
- Return on equity denoted by ROE. This measures how effectively a company uses shareholders' equity to generate profit. A higher ROE indicates better financial performance and efficiency in using investments.
- Operating margin denoted by OPR. This ratio shows the percentage of revenue that remains after covering operating expenses. A higher operating margin indicates a more profitable and efficient company.
- Price earnings ratio denoted by PE. The P/E ratio shows how much investors are willing to pay for each dollar of a company’s earnings.
- Book value per share ratio denoted by BVPS. This shows the value of a company's assets that each share would receive if the company were liquidated. It's a measure of the company's worth per share.
- Return on Capital Employed denoted by ROCE. This shows how well a company is generating profits from its capital, which includes both equity and debt.
- Return on Assets denoted by ROA. It tells you how good a company is at turning its assets (like buildings, equipment, and cash) into profits. It’s like seeing how much money the company makes for every dollar it owns in assets.
- Quick Ratio denoted by QR. This measures a company's ability to pay its short-term obligations with its liquid assets (cash, accounts receivable). It indicates whether a company can quickly pay its bills without needing to sell inventory.
- Asset Turnover Ratio denoted by ATR. This measures how efficiently a company uses its assets to generate sales. It shows how many dollars of sales are produced for every dollar invested in assets.
- Interest Coverage ratio denoted by ICR. This measures a company's ability to pay interest on its debt with its earnings. It shows how many times a company can cover its interest expenses with its operating income.

Glossary:

Current Assets:	Current assets are all the resources a company owns that can be easily converted into cash within one year, like cash, inventory, and accounts receivable (money owed by customers).
Current Liablities:	Current liabilities are all the debts and obligations a company must pay off within one year, like short-term loans, accounts payable (money owed to suppliers), and wages.
Total Debt:	Total debt is the sum of all the money a company owes to lenders, including both short-term(<12 months) and long-term borrowings(>12 months), such as loans and bonds.
Total Assets:	Total assets are everything a company owns that has value, including cash, buildings, equipment, inventory, and investments.
COGS:	Cost of Good Sold(COGS(refers to the direct cost of producing the goods a company sells. It includes the cost of materials, labor, and overhead directly involved in making the product.
Average Inventory:	The average inventory is like an estimate of the typical amount of money a company has tied up in its stock on hand. It's calculated by averaging the value of inventory at the beginning and end of a specific period.
Net Income:	It's calculated by subtracting all expenses, including operating costs and taxes including depriciation and interest paid on loans from total revenue during a specific period.
Total Equity:	It's calculated by subtracting total liabilities from total assets.
Gross Profit:	This is the initial profit a company makes after accounting for the direct costs of producing and selling its goods or services. It's calculated by subtracting the cost of goods sold (COGS) from total revenue.
EPS:	This indicates a company's profitability per share of common stock. It's calculated by dividing the company's net income (profit after all expenses) by the number of outstanding common shares.
Preferred Equity:	Preferred shareholders typically receive a fixed dividend payout before common shareholders and have priority in claiming assets during liquidation (after debt is settled).
Total Share Outstanding:	This represents the total number of a company's common shares currently held by all its shareholders. It encompasses shares available for trading on the open market, restricted shares owned by insiders, and institutional holdings.
EBT:	Earning Before Tax (EBT) reflects a company's profit before income taxes are deducted. It's essentially the last subtotal on the income statement before net income. EBT is calculated by subtracting all business expenses (excluding taxes) from revenue.
Reserves:	Reserve refers to funds set aside from a company's profits for specific future purposes. These purposes can vary, such as covering unexpected expenses, funding future investments, or meeting potential legal liabilities.
EBIT:	It reflects the profit generated by the core business activities before considering interest expenses and income taxes.

Figure 8

Algorithm 1: Local Outlier Factor:

We have used different contamination thresholds to identify outliers:

- At contamination level at 5%, 3 companies are outliers.
- At contamination level at 10%, 6 companies are outliers.
- At contamination level at 15%, 8 companies are outliers.
- At contamination level at 20%, 11 companies are outliers.

Logic to identify Outliers based on feature value:

- Step1 - First Mean, Standard Deviation for each feature has been calculated.
- Step2 - Range of Mean has been calculated, i.e., min value and max value of mean. This has been identified with the help of Mean and Standard Deviation. (Mean-/+standard deviation) will give mean value and max value .
- Step3 - If the feature value of the identified company is outside the calculated range values of the mean of the dataset, then the said company is outlier due to a particular feature.

	CR	DE	IT	ROE	OPR	PE	BVPS	ROCE	ROA	QR	ATR	ICR	Outlier	LOF Score
mean_value	3.860370	0.222222	575.897407	0.158704	0.142963	33.269444	278.096296	0.244630	0.130000	3.739444	7.284444	43.750185	0.592593	1.588470
stand_dev	6.572549	0.366146	1504.810917	0.205252	0.094062	164.900179	530.606241	0.164118	0.086853	6.612478	12.502754	48.515411	0.813066	0.965424
min_value	-2.712178	-0.143924	-928.913509	-0.046549	0.048901	-131.630734	-252.509945	0.080511	0.043147	-2.873034	-5.218309	-4.765226	-0.220473	0.623046
max_value	10.432919	0.588368	2080.708324	0.363956	0.237025	198.169623	808.702538	0.408748	0.216853	10.351922	19.787198	92.265596	1.405659	2.553894

Figure 9

Figure 9 consists of Mean Value, Standard Deviation, Min & Max value(i.e., Range of Mean) of each feature for the set of 54 companies.

Figure 10 displays values of 11 companies which are outliers at contamination 20%.

	Name	CR	DE	IT	ROE	OPR	PE	BVPS	ROCE	ROA	QR	ATR	ICR
10	Honeywell Automation India Ltd	3.34	0.01	15.28	0.13	0.13	99.76	3606.29	0.19	0.10	3.20	22.85	86.75
18	Redington Ltd	1.31	0.48	11.44	-0.50	-0.04	7.03	88.63	-0.28	0.11	0.93	90.47	-7.57
23	Bls International Services Ltd	2.94	0.01	2106.29	0.25	0.13	44.68	19.54	0.30	0.43	2.94	6.49	86.79
28	Netweb Technologies India Ltd	0.00	0.38	6.60	0.48	0.15	353.46	18.39	0.50	0.00	0.00	19.05	16.04
30	CE Info Systems Ltd	4.53	0.05	9.36	0.20	0.38	96.81	101.03	0.26	0.17	4.51	5.02	50.91
31	Latent View Analytics Ltd	46.53	0.02	0.00	0.13	0.25	66.89	58.93	0.16	0.12	46.53	23.23	69.86
36	Rategain Travel Technologies Ltd	19.55	0.02	0.00	0.10	0.09	60.95	65.52	0.14	0.01	19.55	1.43	20.33
37	Zen Technologies Ltd	2.69	0.02	3.12	0.03	0.12	451.09	39.45	0.12	0.08	2.36	2.91	8.82
42	PG Electroplast Ltd	1.36	1.46	5.66	0.05	0.04	-1016.12	174.09	0.08	0.06	0.87	3.74	1.86
46	Black Box Ltd	1.30	2.12	16.31	-0.84	-0.02	12.15	17.63	-0.15	0.01	1.23	7.92	-1.18
48	Hinduja Global Solutions Ltd	3.44	0.12	186.49	0.04	-0.05	36.97	1627.28	0.07	0.06	3.42	1.60	2.58

Figure 10

<u>Interpretation @ Contamination 20%</u>:

At contamination level 20%, we have found 11 companies in outliers.

- Honeywell Automation India Ltd is an outlier for feature BVPS & ATR, since values of BVPS & ATR are lying outside the range (Min & Max value) as observed above.
  Here BVPS value is 3606.29 while BVPS range(for dataset of 54 companies) is Min Value:-252.50 & Max Value:808.70.
  Similarly ATR value 22.85 while ATR range is Min Value:-5.21 & Max Value:19.78.

Other companies are outliers due to various factors as mentioned below:

- Redington Ltd is an outlier due to ROE, OPR, ROCE, ATR, ICR.
- Bls International Services Ltd is an outlier due to IT, ROA.
- Netweb Technologies India Ltd is an outlier due to ROE, PE, ROCE, ROA.
- CE Info Systems Ltd is an outlier due to OPR.
- Latent View Analytics Ltd is an outlier due to CR, OPR, QR, ATR.
- Rategain Travel Technologies Ltd is outlier due to CR, ROA, QR.
- Zen Technologies Ltd is outlier due to PE.
- PG Electroplast Ltd is outlier due to DE, OPR, PE, ROCE.
- Black Box Ltd is outlier due to DE, ROE, OPR, ROCE, ROA.
- Hinduja Global Solutions Ltd is an outlier due to OPR, BVPS, ROCE.

Algorithm 2: Isolation Forest:

Following Charts display the outliers detected by Isolation Forest Algorithm at different contamination levels:

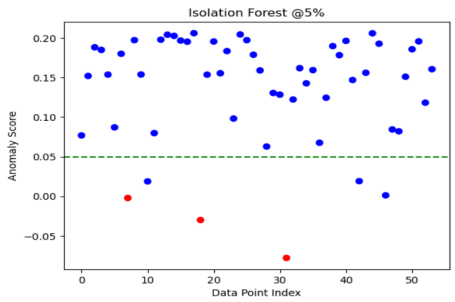

Figure 11

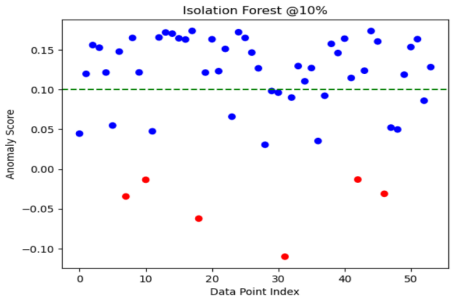

Figure 12

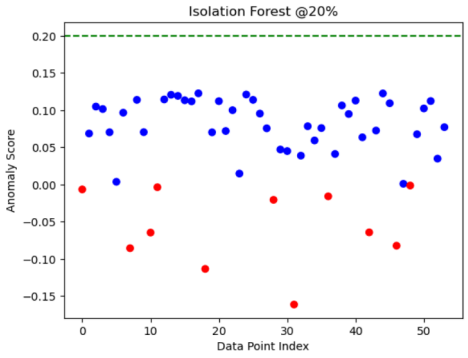

Figure 13

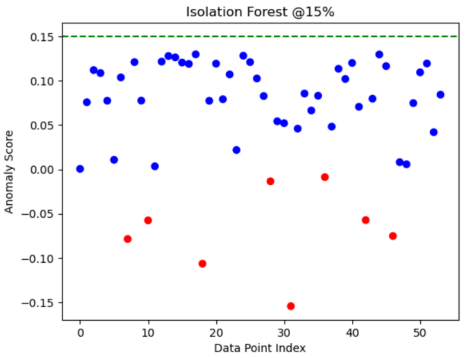

Figure 14

Interpretation @ Contamination 20%:

	Name	CR	DE	IT	ROE	OPR	PE	BVPS	ROCE	ROA	QR	ATR	ICR	Outlier	LOF Score	anomaly	anomaly_score
0	Tata Consultancy Services Ltd	2.53	0.09	4621.54	0.47	0.24	30.06	247.12	0.63	0.29	2.53	10.99	74.05	1	1.142292	-1	-0.006624
7	Oracle Financial Services Software Ltd	10.10	0.01	0.00	0.24	0.42	38.15	863.33	0.35	0.27	10.10	6.73	203.99	1	1.566285	-1	-0.085702
10	Honeywell Automation India Ltd	3.34	0.01	15.28	0.13	0.13	99.76	3606.29	0.19	0.10	3.20	22.85	86.75	-1	4.701914	-1	-0.064712
11	Tata Elxsi Ltd	4.83	0.09	4809.62	0.36	0.28	57.15	334.92	0.45	0.27	4.83	9.43	55.24	1	1.077150	-1	-0.003678
18	Redington Ltd	1.31	0.48	11.44	-0.50	-0.04	7.03	88.63	-0.28	0.11	0.93	90.47	-7.57	-1	4.293931	-1	-0.113536
28	Netweb Technologies India Ltd	0.00	0.38	6.60	0.48	0.15	353.46	18.39	0.50	0.00	0.00	19.05	16.04	-1	2.699845	-1	-0.020638
31	Latent View Analytics Ltd	46.53	0.02	0.00	0.13	0.25	66.89	58.93	0.16	0.12	46.53	23.23	69.86	-1	4.966854	-1	-0.161452
36	Rategain Travel Technologies Ltd	19.55	0.02	0.00	0.10	0.09	60.95	65.52	0.14	0.01	19.55	1.43	20.33	-1	2.771178	-1	-0.015903
42	PG Electroplast Ltd	1.36	1.46	5.66	0.05	0.04	-1016.12	174.09	0.08	0.06	0.87	3.74	1.86	-1	3.846810	-1	-0.064387
46	Black Box Ltd	1.30	2.12	16.31	-0.84	-0.02	12.15	17.63	-0.15	0.01	1.23	7.92	-1.18	-1	3.592921	-1	-0.082346
48	Hinduja Global Solutions Ltd	3.44	0.12	186.49	0.04	-0.05	36.97	1627.28	0.07	0.06	3.42	1.60	2.58	-1	2.007984	-1	-0.001417

Figure 15

At contamination level 20%, we have found 11 companies in outliers(refer to Figure 15):

- For Oracle Financial Services Software Ltd, the company shows strong financial health with a very high BVPS (863.33) and a significant PE ratio (38.15). This suggests it is potentially overvalued or has strong market expectations.
- Honeywell Automation India Ltd, is an outlier mainly due to its exceptionally high BVPS (3606.29). This could indicate a high valuation of its assets.
- Redington Ltd has negative ROE (-0.50) and ROCE (-0.28), which are strong indicators of financial distress or operational inefficiency. These metrics are likely driving its classification as an anomaly.

- Netweb Technologies India Ltd stands out due to a very low debt-to-equity ratio (0.00) and high BVPS (353.46). The absence of debt could indicate a conservative financing strategy or limited access to debt markets.
- Latent View Analytics Ltd has an extremely high current ratio (46.53) and zero debt, suggesting the company has a large amount of liquid assets relative to its liabilities and no leverage. This might indicate inefficiency in using its assets.
- Rategain Travel Technologies Ltd has a high current ratio (19.55) and significant PE ratio (60.95) indicates strong liquidity and high market valuation. The company might be seen as having strong future growth potential but could also be overvalued.
- PG Electroplast Ltd has a negative BVPS (-1016.12) points to potential financial instability or significant liabilities exceeding its assets. This is a critical indicator of financial distress.
- Black Box Ltd has a high debt-to-equity ratio (2.12) and negative ROE (-0.84) signals high leverage and poor profitability. These factors contribute to its classification as an outlier and suggest potential financial troubles.
- Tata Elxsi is marked as an anomaly due to its high inventory turnover ratio, indicating efficient management of inventory but also potential overvaluation.
- Latent View Analytics Ltd has The extremely high current ratio implies that the company holds a lot of liquid assets compared to its liabilities, which could indicate inefficiency in utilizing its assets. The lack of debt suggests a very conservative financing strategy or potential issues with accessing debt markets.
- Hinduja Global Solutions Ltd is flagged as an anomaly primarily due to its high book value per share, indicating a high valuation of assets. The LOF score suggests it is somewhat isolated compared to its neighbors in the dataset, potentially due to the high asset valuation.

LOF vs. Isolation Forest Comparison as 20% Contamination:

Both Isolation Forest and LOF methods are effective in identifying outliers, with LOF being more sensitive to local deviations. Companies flagged by both methods generally exhibit extreme financial metrics, such as high or negative asset valuations, high liquidity ratios, and poor profitability metrics.

Both algorithms consistently flag companies with extreme values in key financial metrics, such as high or negative BVPS, high CR, high DE, and poor profitability metrics.

LOF scores tend to be higher for more pronounced outliers, indicating greater sensitivity to local data density compared to Isolation Forest, which provides a more global perspective.

Both the methods mostly agree on which companies are outliers.

Common Companies detected as outliers:

- Black Box Ltd
- Hinduja Global Solutions Ltd
- Honeywell Automation India Ltd
- Latent View Analytics Ltd
- Netweb Technologies India Ltd
- PG Electroplast Ltd
- Rategain Travel Technologies Ltd
- Redington Ltd

Document 3 keys Lessons learnt in the assignment:

1. One key learning is effective use of Singular Value Decomposition (SVD) for collaborative filtering in recommendation systems. SVD helps in reducing the dimensionality of the user-item interaction matrix, capturing latent factors that represent the underlying relationships between users and items. By applying SVD, the code efficiently predicts user preferences for items they haven't interacted with, leading to accurate and personalized recommendations. This approach is crucial for handling large-scale datasets with sparse interactions, enhancing the scalability and performance of the recommendation system.
2. Another key learning is effectiveness of hybrid recommendation algorithms that combine user-based and item-based collaborative filtering. By leveraging both user similarity and item similarity, the hybrid approach enhances the recommendation quality by capturing diverse aspects of user preferences and item characteristics. This method mitigates the limitations of relying on a single type of filtering, resulting in more accurate and robust recommendations. Specifically, user-based scores are combined with item-based scores, leading to a comprehensive recommendation system that can adapt to various user behaviors and item interactions.
3. Both Isolation Forest and LOF methods are effective in identifying outliers, with LOF being more sensitive to local deviations. LOF scores tend to be higher for more pronounced outliers, indicating greater sensitivity to local data density compared to Isolation Forest, which provides a more global perspective.