Lab ML for Data Science: Part 1

Getting Insight into Unsupervised Dataset

Outline

THE PROJECT OBJECTIVE

METHOD AND ALGORITHM

EXPERIMENTAL RESULT

DISCUSSION AND INSIGHT

The Project Objective

The Project Objective

Dataset description: the annual spending of wholesale customers for six different product categories (Fresh, Milk, Grocery, Frozen, Detergents_Paper, and Delicassen) at wholesale stores in Portugal. The data contains information about channel (1-Horeca and 2-Retail) and region (1-Lisbon, 2-Oporto, and 3-other) where the purchase happened.

- **Anomaly Detection:** detecting anomalous spending behaviour using unsupervised Machine Learning technique
- **Additional Explanation:** exploring the reason of the anomaly (e.g. why specific instance is predicted to be anomalous)
- **Reproducibility:** introducing a mechanism that favour robust resampling

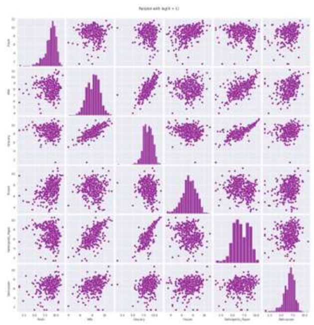
Method and Algorithm

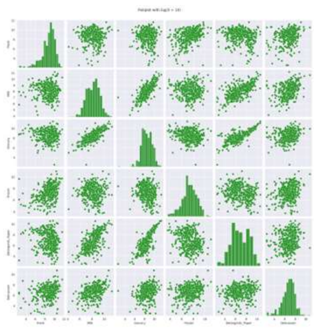
Initial Data Analysis

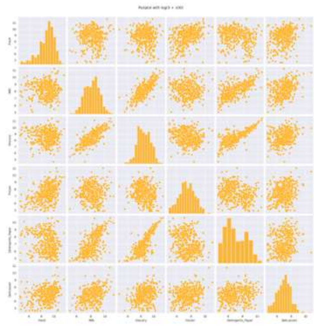
Non-linear transformation:

$$x \to \log(x + \theta)$$

1. Create a scatter plot each for $\theta = 1$, $\theta = 10$, $\theta = 100$
2. Choose θ that shows a result that is more normally distributed



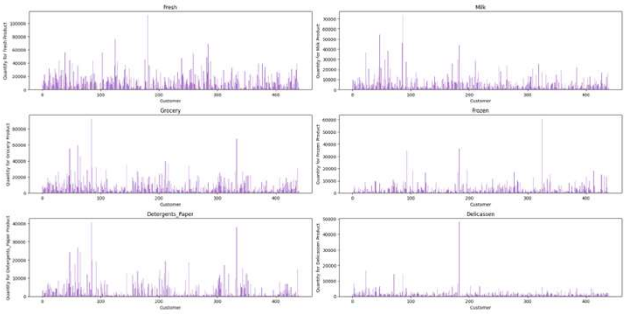




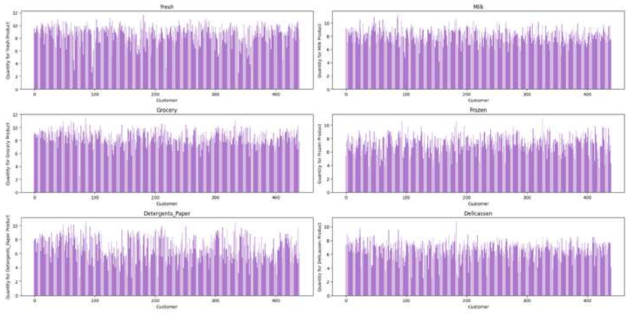
Initial Data Analysis

Non-linear transformation

Before transformation



After transformation



Detecting Anomalies

Calculating hard minimum (squared distance):

$$z_{jk} = \left\| x_j - x_k \right\|^2$$

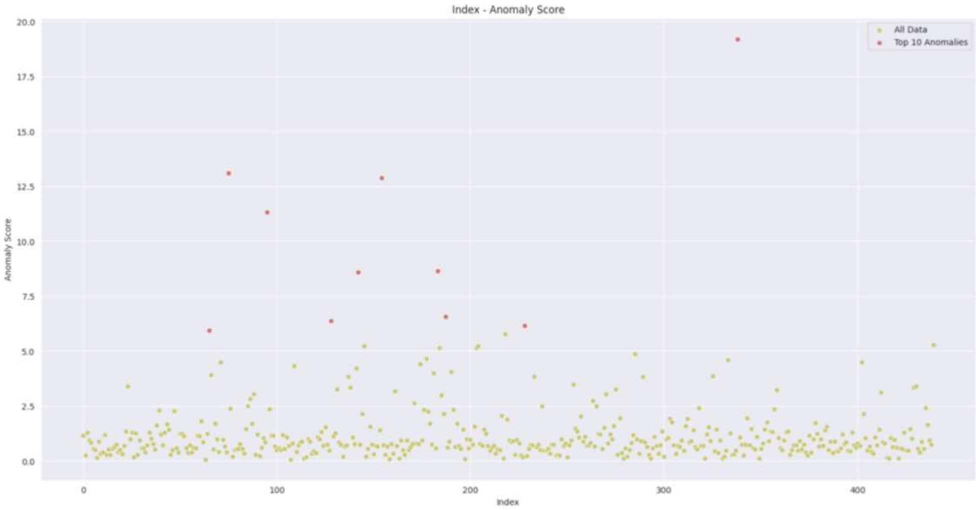
$$y_j = \min_{k \neq j} z_{jk}$$

Implementation:

```
def hard_min(data):
    squared_distances = scipy.spatial.distance.cdist(data, data) ** 2 # $1 Hard min should use the square euclidean distance
    np.fill_diagonal(squared_distances, np.nan)
    hardmins = np.nanmin(squared_distances, axis=1)
    return hardmins
```

Detecting Anomalies

Calculating hard minimum (squared distance):



Detecting Anomalies

Calculating soft minimum (considering multiple neighbour):

$$y_j = soft\min_{k \neq j}\{\overrightarrow{z_{jk}}\}$$

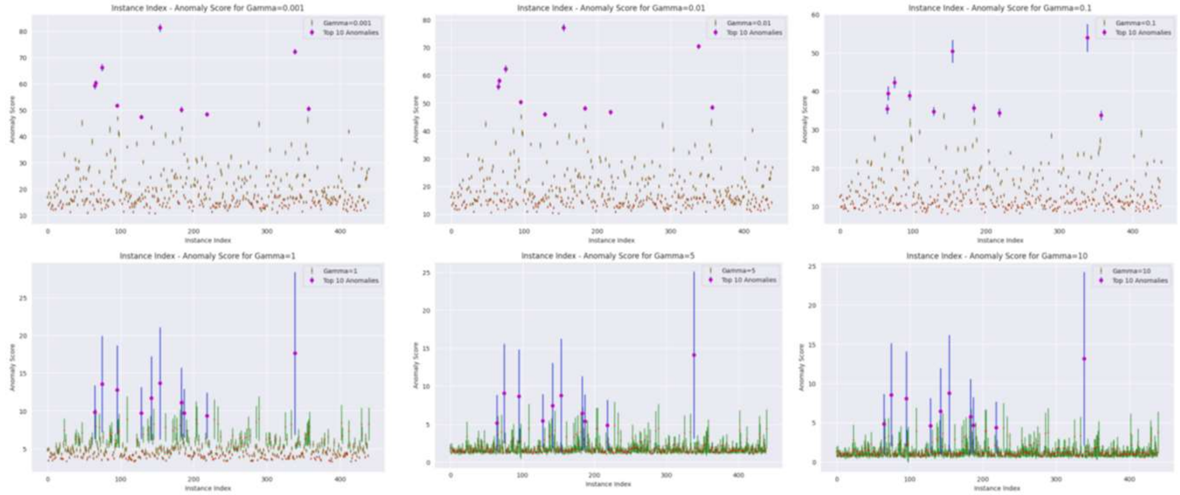
$$soft\min_{k \neq j}\{z_{jk}\} = -\frac{1}{\gamma}\log\left(\frac{1}{N-1}\sum \exp(-\gamma z_{jk})\right)$$

Implementation:

```
def soft_min(data, gamma):
    squared_distances = scipy.spatial.distance.cdist(data, data) ** 2
    np.fill_diagonal(squared_distances, np.nan)
    N = data.shape[0]
    softmins = np.zeros(N)
    for j in range(N):
        softmins[j] = (-1 / gamma) * np.log(np.sum(np.exp(-gamma * squared_distances[j, np.arange(N) != j])) / (N-1))
    return softmins
```

Detecting Anomalies

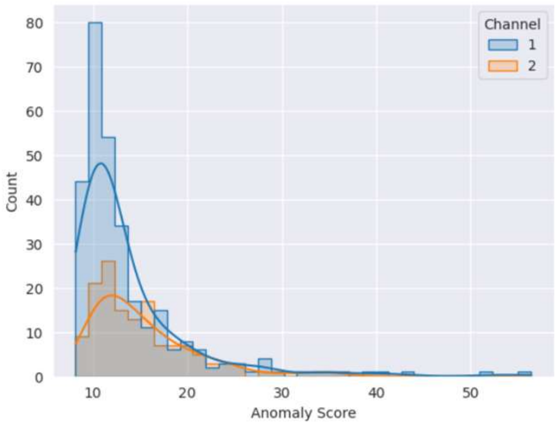
Selecting suitable parameter γ using bootstrap with sample size = 440 and the number of sampling = 1000



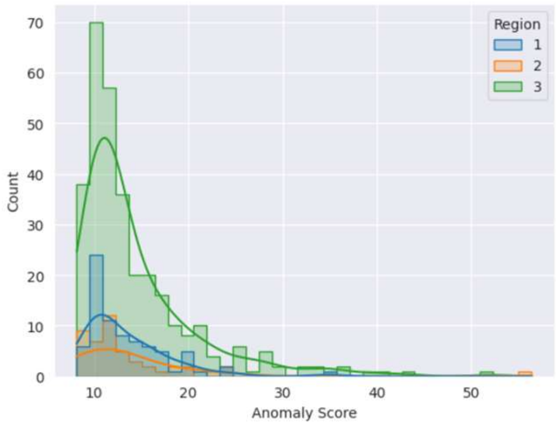
Bootstrap: a larger sampling leads to more accurate estimates of the sampling distribution. The sampling that we choose is sampling with replacement. We test the gamma parameter with value 0.001, 0.01, 0.1, 1, 5, 10, and visualize the result on this graph. We decide gamma 0.1 is the best, because we can see clearly the separability, while still having a decent spread.

Getting Insights into Anomalies

Anomaly Score Based on Channel



Anomaly Score Based on Region



This histogram is created by using the soft-minimum anomaly score.

Getting Insights into Anomalies

Identifying input features that drive anomaly

Contribution of data point k to the anomaly score of instance j

$$R_k^{(j)} = \frac{\exp\left(-\gamma z_{jk}\right)}{\sum_{k \neq j} \exp\left(-\gamma z_{jk}\right)} \cdot y_j$$

```
def compute_contribution_to_anomaly_score(data, gamma):
    num_instances = len(data) # 440

    squared_distances = scipy.spatial.distance.cdist(data, data) ** 2 # $
    anomaly_scores = soft_min(data, gamma)
    contributions = np.zeros((num_instances, num_instances))
    for j in range(num_instances):
        # Exclude the self-distance
        distance_j = np.delete(squared_distances[j], j)
        total_exp_distances = np.sum(np.exp(-gamma * distance_j))
        for i in range(num_instances):
            if i != j:
                exp_distances =np.exp(-gamma * squared_distances[j][i])
                contributions[j][i] = exp_distances / total_exp_distances * anomaly_scores[j]

    return contributions
```

Contribution of input feature i to the anomaly score of instance j.

$$R_i^{(j)} = \sum_{k \neq j} \frac{\left[x_k - x_j\right]_i^2}{\left\| x_k - x_j \right\|^2} \cdot R_k^{(j)}$$

```
def propagate_contributions_to_input_features(data, contributions):
    num_instances, num_features = data.shape
    feature_contributions = np.zeros((num_instances, num_features))

    for j in range(num_instances):
        for i in range(num_features):
            feature_sum = 0
            for k in range(num_instances):
                features_distance = data[k] - data[j]
                if k != j and np.linalg.norm(features_distance) !=0:
                    feature_sum += (np.abs(features_distance)[i]) ** 2 / np.linalg.norm(features_distance) ** 2 * contributions[j][k]
            feature_contributions[j, i] = feature_sum
    return feature_contributions

contributions = compute_contribution_to_anomaly_score(X_arr, 0.1)
feature_contributions = propagate_contributions_to_input_features(X_arr, contributions)
```

Experimental Result

What makes the top 10 anomalies?

Feature contributions of the top 10 anomalous instance

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
339	$\boxed{29.485070}$	4.529298	2.130403	10.513941	8.336905	1.416954
155	5.396845	9.010229	7.602488	10.874156	7.685597	$\boxed{11.600233}$
76	1.457854	1.880539	$\boxed{29.035958}$	1.691701	7.810054	1.224231
67	$\boxed{18.047837}$	3.503522	2.141368	2.129333	3.326700	10.889680
96	$\boxed{24.699626}$	1.402416	1.364338	2.039177	7.271388	2.609820
184	1.876089	6.547149	2.469367	6.747506	4.913448	$\boxed{13.491530}$
66	$\boxed{12.297407}$	2.762617	3.180834	10.643841	4.395246	2.467247
129	9.469591	3.571826	1.930515	3.789911	3.459209	$\boxed{12.881751}$
219	$\boxed{21.756676}$	0.983913	0.946765	2.394896	2.036644	6.622685
358	$\boxed{13.352435}$	5.538051	3.031242	3.414835	2.867980	5.813387

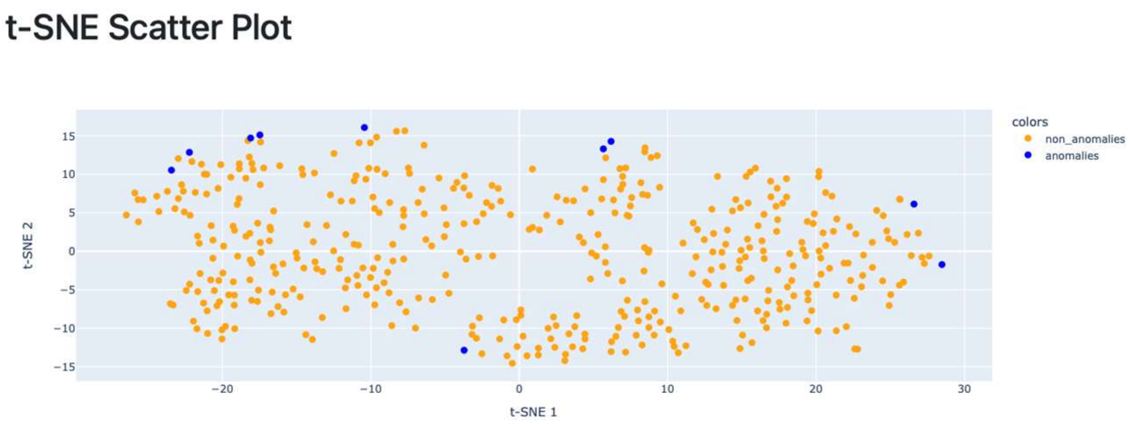
The soft-minimum score of the top 10 anomalous instance

	Num. of the customer	Score	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	339	56.412571	$\boxed{2.564949}$	5.837730	8.858084	9.655731	3.218876	6.327937
1	155	52.169549	6.448889	4.174387	4.990433	4.442651	2.833213	$\boxed{2.890372}$
2	76	43.100336	9.923682	7.044905	$\boxed{2.564949}$	8.393216	2.564949	6.892642
3	67	40.038439	$\boxed{2.944439}$	7.342132	8.912877	5.220356	8.154213	3.610918
4	96	39.386764	$\boxed{2.564949}$	7.982758	8.742255	6.109248	5.451038	6.577861
5	184	36.045089	10.514801	10.691035	9.912447	10.506272	5.517453	$\boxed{10.777977}$
6	66	35.747192	$\boxed{4.553877}$	9.950800	10.732869	3.828641	10.095801	7.267525
7	129	35.102803	5.010635	9.088963	8.251403	5.023881	6.977281	$\boxed{2.564949}$
8	219	34.741580	$\boxed{3.332205}$	8.924523	9.630037	7.166266	8.477828	8.761237
9	358	34.017930	$\boxed{3.850148}$	7.158514	10.011534	4.990433	8.818334	4.787492

Table on the left: Feature contributions
Table on the right: Score – soft minimum score, Column Products – number after log transformation
We are trying to look on which feature that contributes the most for these points identified as the anomaly. We are also trying to see from the data after log transformation, and it is pretty evident that majority it has shown as the lower number of purchase of those products.

Here is the soft-minimum score of the top 10 anomaly purchase in comparison with other data.

Top 10 Anomalies Purchase



Discussion and Insight

Insight for The Top 10 Anomalies

The most contributing features for *the top-10 anomalies* are generally Fresh and Delicatessen product.

In the Fresh product category, most anomalies are associated with low purchase volumes, while anomalies in the Delicatessen and Frozen product categories tend to occur with middle to low purchase volumes. Conversely, the anomalies in the Grocery product category are primarily due to high purchase volumes.

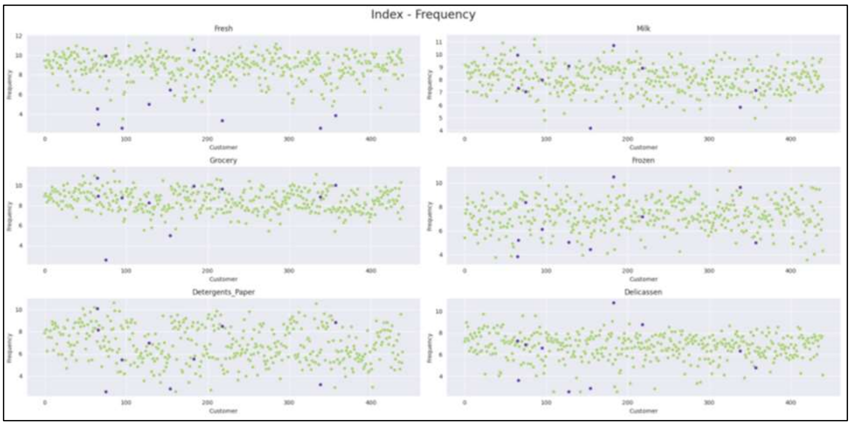


Table of Feature Contributions.

Insight for The Top 10 Anomalies

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
339	$\boxed{29.485070}$	4.529298	2.130403	10.513941	8.336905	1.416954
155	5.396845	9.010229	7.602488	10.874156	7.685597	$\boxed{11.600233}$
76	1.457854	1.880539	$\boxed{29.035958}$	1.691701	7.810054	1.224231
67	$\boxed{18.047837}$	3.503522	2.141368	2.129333	3.326700	10.889680
96	$\boxed{24.699626}$	1.402416	1.364338	2.039177	7.271388	2.609820
184	1.876089	6.547149	2.469367	6.747506	4.913448	$\boxed{13.491530}$
66	$\boxed{12.297407}$	2.762617	3.180834	10.643841	4.395246	2.467247
129	9.469591	3.571826	1.930515	3.789911	3.459209	$\boxed{12.881751}$
219	$\boxed{21.756676}$	0.983913	0.946765	2.394896	2.036644	6.622685
358	$\boxed{13.352435}$	5.538051	3.031242	3.414835	2.867980	5.813387

Channel	Anomaly Score	Region
1	56.412571	2
1	52.169549	3
1	43.100336	3
1	40.038439	3
1	39.386764	3
1	36.045089	3
2	35.747192	3
1	35.102803	3
2	34.741580	1
2	34.017930	3

CHANNEL	Frequency
Horeca	298
Retail	142
Total	440

REGION	Frequency
Lisbon	77
Oporto	47
Other Region	316
Total	440

- **The top 10 anomalies are dominated from Channel 1 (Horeca) and Region 3 (Other).** This could be because data imbalance since the data are coming from this particular channel and region has significantly higher than the others. So, the possibility of anomaly is also higher.
- **Despite Channel 2 (Retail) exhibiting fewer anomalies compared to Channel 1 (Horeca), analysis of the top 10 data points consistently indicates that Fresh products are the primary contributors to these anomalies.** These anomalies are primarily due to low purchase quantities. It is plausible that customers visiting retail stores tend to purchase a wider variety of goods, given the broader range of products.
- **Although only one data point from the Oporto region appears in the top 10 anomaly list, this region exhibits the highest anomaly score compared to all other regions.**

Table at the left: Feature Contributions

Here, we would like to see the top 10 anomalies and its relation with the Channel and Region where the transaction happened.

- The top 10 anomalies are dominated with Channel 1- Hotel restaurant and café and Region 3 – Other. (Because data imbalance)
- For channel 2, all the anomaly are consistently showing that Fresh products are the most contributed factors. (possible because retailer stores has broader range of products)
- Region 2 (Oporto) is only 1 data that is inside the top 10 list, but it has the highest anomaly score.

Insight for Overall Purchase Data

The most contributing features for *overall purchase* are Fresh product, followed by Detergents and Frozen product.

Fresh products are challenging to handle and must be consumed and sold quickly leading to a clear distinction between Fresh products and the others.

Retail shops have demonstrated a significant influence in selling Fresh products, despite the lower frequency of purchases compared to Horeca.

The Oporto region appears to be a promising location with high potential for purchasing Fresh products.

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
total	2.943971	1.694657	1.494041	2.740328	2.751812	2.547973
channel 1	2.635883	1.739868	1.506352	2.631433	2.765948	2.415437
channel 2	3.590522	1.599779	1.468205	2.968855	2.722148	2.826111
region 1	2.648085	1.698856	1.442940	2.470154	2.794168	2.281499
region 2	3.276541	1.611121	1.315416	2.696144	2.677278	2.267846
region 3	2.966606	1.706059	1.533060	2.812733	2.752577	2.654569

REGION	Frequency
Lisbon	77
Oporto	47
Other Region	316
Total	440

CHANNEL	Frequency
Horeca	298
Retail	142
Total	440

To remember:

- Overall purchase feature contribution: Fresh, followed by Detergents. The reason might be: Fresh products need to be consumed quickly
- Retail shops has much more contribution than Horeca channel, even though the frequency of purchase less than Horeca
- The feature contribution for Oporto is higher compared to other region for Fresh product, even though the frequency is also less than other region

Discussion

These findings can inform the distribution and marketing strategies for the products.

- The most effective channel and location for selling Fresh product is through retail shop in Oporto.
- For detergents, the optimal channel and location is through hotels, restaurants, and cafés in Lisbon.
- Additionally, Fresh products have demonstrated strong purchase potential across all other regions in Portugal.

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
total	2.943971	1.694657	1.494041	2.740328	2.751812	2.547973
channel 1	2.635883	1.739868	1.506352	2.631433	2.765948	2.415437
channel 2	3.590522	1.599779	1.468205	2.968855	2.722148	2.826111
region 1	2.648085	1.698856	1.442940	2.470154	2.794168	2.281499
region 2	3.276541	1.611121	1.315416	2.696144	2.677278	2.267846
region 3	2.966606	1.706059	1.533060	2.812733	2.752577	2.654569

Comparing the top-10 list feature contribution and the overall feature contribution:

- The top 10 list contribution doesn't really align with the overall feature contribution. In the top 10 list, the feature that contributes the most are Fresh products and followed by Delicatssen. While in the overall feature contribution, it is Fresh products and followed by Detergents.
- The channel and region for top 10 list are mainly channel 1 and region 3. While in overall, it is channel 2 and region 2.

- Fresh is better in retail in Oporto
- Detergents in horeca and Lisbon