Project Report: Outlier Detection Using Isolation Forest

Methodology

The project employs the **Isolation Forest** algorithm is an unsupervised learning technique well suited for anomaly detection in large datasets. The following steps outline the approach:

1. Data Cleaning

- Unnecessary columns were dropped to reduce dimensionality.
- Duplicates were removed to avoid skewing the results.
- Null values were imputed:
    - **Numerical columns**: Replaced with mean.
    - **Categorical columns**: Filled using mode.

2. Feature Engineering

- Categorical and numerical data were separated to apply appropriate preprocessing.
- Categorical data was encoded
- No target labels were used, making this a fully unsupervised approach.

3. Outlier Detection

- **Isolation Forest** was applied to detect anomalous data points.
- The model assigned each entry a label:
    - `1`: Inlier (normal customer)
    - `-1`: Outlier (potentially abnormal behaviour)

Findings

Outlier Summary

- The model identified **X** outliers out of **Y** total entries.
- Outliers accounted for approximately **Z%** of the customer base.

Statistical Summary (Outliers)

The outliers exhibited significantly different statistics compared to inliers, especially in:

- **Annual Income**
- **Purchase Amount**
- Possibly other features like frequency of purchase or returns.

Visual Insights

A scatter plot of Annual Income:

- A distinct cluster of outliers located at extreme values of income or purchases.
- Most outliers had very high or very low purchase behaviour, suggesting unusual activity.

Business Insights

1. Customer Profiling

- Outliers may represent **VIP customers** (e.g., very high income and high spending).
- Conversely, they could also be **fraudulent or erroneous entries**, such as:
    - Data entry mistakes
    - Unusually high returns or negative values

2. Marketing Optimization

- High-spending outliers can be segmented for **premium loyalty programs**.
- Abnormal low-spending or irregular behaviour can inform **churn risk analysis**.

3. Data Quality

- Some outliers may indicate data integrity issues, which need **manual verification** or **automated cleaning** rules.

4. Fraud Detection

- Isolation Forest's effectiveness in spotting anomalies makes it suitable for **early fraud warnings**, particularly in:
    - E-commerce
    - Finance (e.g., credit card fraud)