UNIVERSITA
DEGLI STUDI
DI TORINO

Data Mining Project 2025

Student: Varaga Haghoubians

Professors: Mirko Polato, Robert Birke

University of Turin

1 Dataset: Iranian Churn

Dataset Reference:

- Iranian Churn [Dataset]. (2020). UCI Machine Learning Repository. `https://doi.org/10.24432/C5JW3Z`

This dataset comprises 3,150 rows from an Iranian telecom company, spanning 12 months. Each row represents a customer and provides 13 features:

- Call failures
- Frequency of SMS
- Number of complaints
- Number of distinct calls
- Subscription length
- Age group
- Charge amount
- Type of service
- Seconds of use
- Status

- Frequency of use
- Customer Value
- Churn (0 or 1)

Columns other than `Churn` are aggregated data from the first 9 months; the `Churn` label reflects each customer's state after 12 months (with a 3-month "planning gap").

2 Project Description

The aim of this project is to showcase three distinct data mining tasks on the *Customer Churn.csv* dataset:

- **Clustering** using K-Means,
- **Classification** using a Decision Tree,
- **Anomaly Detection** via Local Outlier Factor (LOF).

We have conducted each method separately and visualized the outputs.

3 Goal of the Project

- **Clustering (K-Means):** Uncover latent customer segments or usage patterns (high usage vs. low usage, etc.).
- **Classification (Decision Tree):** Predict churn (class 0 or class 1) given a customer's feature profile.
- **Anomaly Detection (LOF):** Identify atypical customers whose usage or behavior diverges significantly from their local peer group.

4 Libraries Used

The Python ecosystem was employed, specifically:

- `pandas` & `numpy` for data manipulation,
- `scikit-learn` (`sklearn`) for K-Means, Decision Tree, LOF, scaling, metrics,
- `matplotlib` & `seaborn` for visualizing results,
- `PCA` from `sklearn.decomposition` to reduce dimensionality for 2D plotting.

5 Interpretation of the Outputs

5.1 Anomaly Detection (LOF)

After dropping the `Churn` label (since anomaly detection is unsupervised), we scaled the numeric attributes and applied **Local Outlier Factor (LOF)**. LOF outputs:

- A label (+1 for inlier, −1 for outlier),
- A continuous `LOF_Score` measuring local density deviance.

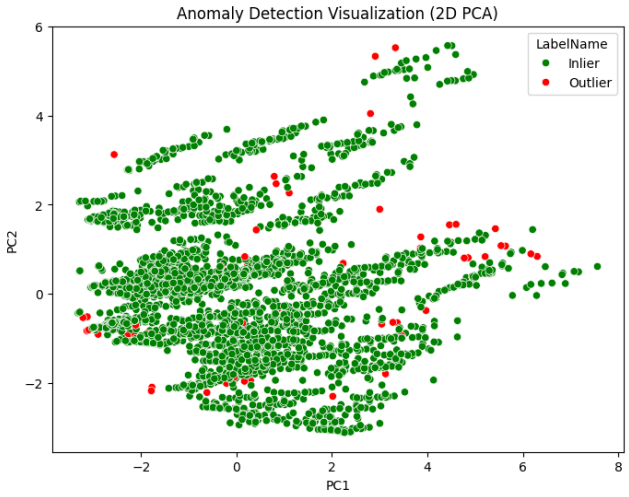

Figure 1: Anomaly Detection Visualization (2D PCA). Green dots = Inliers; Red dots = Outliers.

Figure 1: Shows the dataset in 2D (via PCA). The majority of points in green are deemed normal. The red points, scattered in various pockets, are flagged as anomalies, possibly due to extreme usage or complaint patterns.

Interpretation:

- If around 2% are outliers, it matches typical contamination assumptions.
- Investigate outliers to see if they reflect suspicious/fraudulent behaviors or simply rare but valid user profiles.

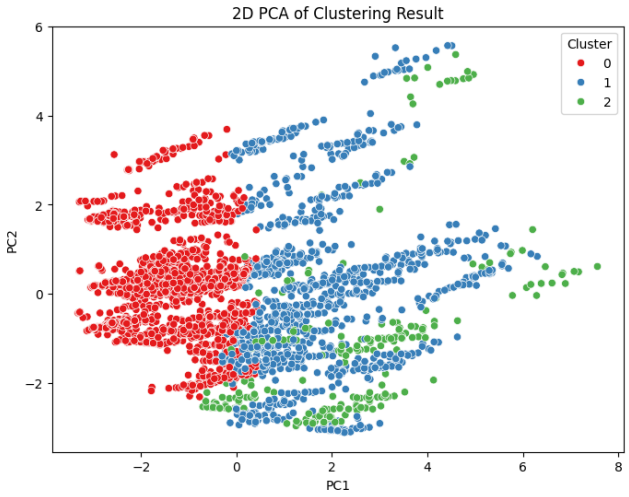

Figure 2: 2D PCA of K-Means Clustering Result.

5.2 Clustering (K-Means)

We performed **K-Means** with $k = 3$ after scaling numeric features and excluding `Churn`. Each row was assigned a cluster: 0, 1, or 2.
Figure 2:

- *Cluster 0* (e.g. red): Possibly high usage or distinct patterns.
- *Cluster 1* (e.g. blue): Another major group, maybe moderate usage.
- *Cluster 2* (e.g. green): A smaller group with different usage or subscription length.

Interpretation:

- Each cluster corresponds to different usage or complaint profiles.
- Summaries of these clusters can help identify potential marketing or retention strategies per cluster.

5.3 Classification (Decision Tree)

For classification, we used the `Churn` column as the label (0 or 1). We split the data (train/test), trained a **Decision Tree**, then evaluated via confusion matrix.

Figure 3:

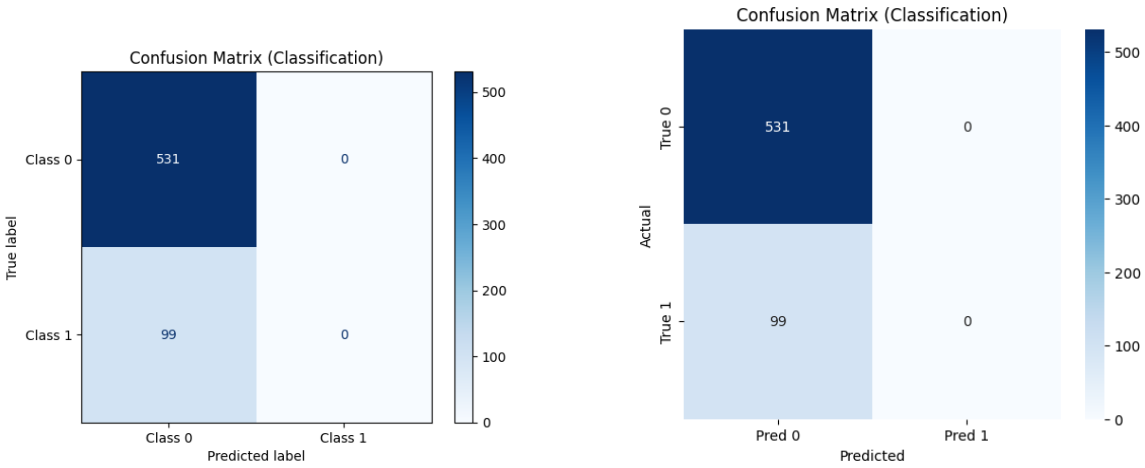

Figure 3: Two versions of the Confusion Matrix labeling. Left: True vs. Pred. Right: Class 0 vs. Class 1.

- Example matrix:

```
[[531   0]
 [ 99   0]]
```

- This means the model predicted everything as Class 0 (non-churn).

Interpretation:

- This yields no false positives but many false negatives (99 churners missed).
- The high imbalance in churners can lead to a trivial predictor that always picks Class 0.
- Addressing imbalance (or adjusting thresholds) is crucial if we want to catch more churners.