Assignment 4 (Week 9): Unsupervised Learning for Policyholder Segmentation

Course: Data Analytics for Actuarial Science

Overview

Topics: Principal Components Analysis (PCA), K-Means Clustering, Hierarchical Clustering.
Dataset: `policyholder_churn.csv`.

Deliverables:

- One Jupyter Notebook: `Assignment4_unsupervised.ipynb`.
- One PDF report (2–3 pages) summarizing findings and actuarial interpretation.

Assume the dataset contains the following columns:

- `age`: Policyholder age.
- `tenure_years`: Years with the insurer.
- `num_claims`: Number of claims filed.
- `policy_type`: Categorical variable (`Life`, `Health`, `Motor`).
- `annual_premium`: Annual premium charged.
- `churn`: Binary indicator (1 = lapsed, 0 = retained). This is used only for interpretation, not for training.

Unless explicitly stated, use only the explanatory variables for PCA and clustering. You may use `churn` later for interpretation.

Part A – Data Preparation & Exploration (15 pts)

A1. Load the dataset `policyholder_churn.csv` into a pandas DataFrame.

A2. Check for missing values and basic descriptive statistics for the variables `age`, `tenure_years`, `num_claims`, `annual_premium`, and `policy_type`.

A3. Identify any obvious outliers or strong skewness in the numerical variables and briefly comment on them.

A4. Encode `policy_type` into numerical form using one-hot encoding (e.g., `policy_type_Health`, `policy_type_Motor`).

A5. Standardize all numerical features (zero mean and unit variance) before applying PCA or clustering.

What to show in the notebook for Part A:

- A table of missing values for each column and your chosen handling method (e.g., removal or imputation).
- `describe()` output or an equivalent summary table for the main variables.
- A short (3–4 line) comment on any outliers or skewness and how this may affect your analysis.

Part B – Principal Components Analysis (PCA) (25 pts)

B1. Apply PCA to the standardized feature matrix (excluding `churn`) and extract at least the first 5 principal components.

B2. Report the proportion of variance explained by each of the first 5 components and the cumulative variance explained.

B3. Produce a scree plot (component number vs. explained variance ratio).

B4. Choose a suitable number of principal components k (e.g., 2 or 3) and justify your choice based on the scree plot and cumulative variance.

B5. Produce a scatter plot of the first two principal components, coloring points by `policy_type`.

Questions to answer in your notebook/report for Part B:

- How much total variance is explained by the first k components?
- Which original variables contribute most to PC1 and PC2 (based on PCA loadings)?
- What latent dimensions might these components represent in an actuarial context (e.g., overall risk level, size of premium, tenure)?

Part C – K-Means Clustering (30 pts)

C1. Choose an input space for clustering: either

- the full standardized feature matrix, or
- the first k principal components from Part B.

State clearly which option you use.

C2. For $K = 2, 3, 4, 5, 6$, fit K-Means clustering models and for each K compute:

- The inertia (within-cluster sum of squares),
- The average silhouette score.

C3. Create two plots:

- Inertia vs. K (Elbow plot),
- Silhouette score vs. K.

C4. Based on these plots, choose a final number of clusters K and justify your choice.

C5. Fit a final K-Means model with K clusters and assign each observation to a cluster.

C6. For each cluster $(0, 1, \ldots, K^{-1})$, compute the mean values of `age`, `tenure_years`, `num_claims`, and `annual_premium`.

C7. Provide a verbal profile or label for each segment (e.g., "young low-premium low-claims customers").

C8. Optional (for interpretation only): compute the churn rate in each cluster by calculating the mean of `churn` for each cluster.

Questions to answer in your notebook/report for Part C:

- Which K did you choose, and what evidence supports this choice (elbow, silhouette)?
- How do the cluster profiles differ in terms of age, tenure, claims, and premiums?
- Are there clusters with notably higher or lower churn rates (if calculated)?

Part D – Hierarchical Clustering (15 pts)

D1. Draw a random subsample of 500–800 policyholders from the same feature space used for K-Means (either standardized features or k PCs).

D2. On this subsample, perform hierarchical clustering using Ward's method.

D3. Plot the dendrogram for the subsample.

D4. Cut the dendrogram to obtain K clusters (same number as in Part C).

D5. Compare the hierarchical clustering solution with the K-Means solution qualitatively (for example, by looking at cluster sizes or profiles).

Questions to answer for Part D:

- Do the hierarchical clusters appear broadly similar to the K-Means clusters?
- Give one advantage and one disadvantage of hierarchical clustering compared to K-Means in actuarial applications.

Part E – Report & Actuarial Interpretation (15 pts)

Prepare a short report (2–3 pages) in PDF format. The report should include:

E1. A brief description of the dataset and main variables used in the analysis.

E2. A summary of the PCA results (variance explained, interpretation of principal components).

E3. A summary of the clustering results (number of clusters, key characteristics of each cluster).

E4. Interpretation of the segments from an actuarial/business perspective, including at least two concrete recommendations. Examples:

- Which clusters might be targeted for retention campaigns?
- Which clusters might be underpriced or overpriced given their claim behavior?
- How could underwriting or marketing strategies differ across segments?

The report should focus on interpretation, not on detailed code, and should not exceed three pages.