Evaluating the Impact of Unsupervised Learning on Credit Risk Classification

Yunzhen Wu

Feb. 23, 2025

Abstract

The purpose of this work is to study the effects of unsupervised learning approaches on improving the efficiency of credit risk classification models. Traditional logistic regression is often used to predict creditworthiness, but many studies have speculated that unsupervised pre-training may improve prediction accuracy and reduce misclassification costs. This study uses Gaussian mixture models (GMM), K-means, and hierarchical clustering methods on the German credit dataset to form new features and evaluate their impact on logistic regression performance. The result indicates that no one of the unsupervised learning methods could significantly raise the classification accuracy and reduce the costs as compared to the existing supervised approaches. This fact implies that the dataset does not contain distinguishable clusters. Our research sheds light on the inadequacies of clustering techniques in such exporting-related contexts and recommends employing alternative approaches such as feature engineering, as well as, ensemble learning.

Introduction

Credit risk assessment is an important task for financial institutions, whose goal is to expand their customer base while minimizing potential financial risk. The cost of granting credit to a bad customer is much higher than the cost of not granting credit to a good customer, so accurate classification is critical. Traditional techniques, such as logistic regression, that are used for credit assessment depend on the data being labeled for dividing applicants into categories of good or bad credit risks. Nevertheless, much recent research has demonstrated that unsupervised learning may be productive in enhancing the performance of traditional models via pre-training.

This study investigates whether combining unsupervised learning methods can improve the performance of logistic regression in credit risk classification. Specifically, Gaussian mixture models (GMM), K-means, and hierarchical clustering were applied to identify hidden structures in the dataset and evaluate their performance in logistic regression.

Literature Review

Several studies have explored the integration of unsupervised learning into credit risk assessment. Zweig and Campbell (1993) and Gallop et al. (2003) discuss effective cutoffs in classification problems, with the focus on the fact that precision-recall trade-off is particularly important in cost-sensitive situations. More recently introduced work by Lucas and Jurgovsky (2020) deals with the role of anomaly detection and clustering techniques realizing the financial risks, while the article by Pang et al. (2020) offers to provide a thorough review of unsupervised pretraining methods for the classification tasks.

Unsupervised clustering methods have been widely applied in financial area such as fraud detection and customers' groups design. Furthermore, the question whether they are able to increase the accuracy in traditional credit classification remains still open, especially in scenarios where the inherent structure does not conform to the naturally formed clusters.

Methods

1. Data Preparation

The dataset used in this study is the German credit dataset from OpenML, which contains 1,000 instances with 20 explanatory variables and a binary class variable representing credit risk. The dataset consists of both numerical and categorical features. The key steps in data preparation included:

- **Detecting Missing Data**: There were no missing values in the dataset.
- **Feature Encoding**: Categorical variables were encoded using one-hot encoding to facilitate compatibility with machine learning models.
- **Feature Scaling**: Min-Max normalization was applied to ensure that all variables were on a comparable scale before applying clustering algorithms.

2. Exploratory Data Analysis (EDA)

EDA helps understand the distribution of characteristics and potential correlations. The main findings are summarized as follows:

- **Credit Amount Distribution** (Figure 1): Credit limit shows a right-skewed distribution, indicating that most applicants request a lower credit limit.

- **Personal Status Distribution** (Figure 2): Most applicants are single men; this is the reason why credit allocation can possibly lead to demographic bias.

3. Baseline Model: Logistic Regression

As the baseline classifier, a traditional logistic regression model was trained on the basis of K-Fold cross-validation. The model was evaluated by the precision, recall, and F1-score, with the decision threshold (0.086) optimally determined by the work done by Zweig and Campbell (1993).

4. **Unsupervised Learning Approaches**

GMM, K-Means, and hierarchical clustering were applied to generate new features for the data. These cluster labels were then used as additional features in logistic regression to evaluate their impact on classification performance. The new models were also evaluated using precision, recall, and F1 score.

(1) Gaussian Mixture Model (GMM)

- GMM was chosen because it is able to model complex nonlinear data distributions and assign soft probabilities to clusters, which can provide more nuanced grouping of applicants.
- Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) were used to determine the optimal number of clusters (Figure 3).

(2) K-Means Clustering

- K-Means was chosen because it is able to efficiently partition data into different groups by minimizing the intra-cluster variance, making it a widely used technique in customer segmentation.

- The Elbow Method was applied to determine the optimal number of clusters (Figure 4).

(3) Hierarchical Clustering

- Hierarchical Clustering allows for a hierarchical representation of customer relationships without requiring a predefined number of clusters, making it suitable for exploring natural structures in the dataset.
- A dendrogram (Figure 5) was generated to visualize the hierarchical structure.
- The silhouette score was computed to assess cluster quality.

Each clustering technique was applied to the scaled dataset, and the resulting cluster labels were incorporated as additional features in logistic regression. The updated models were re-evaluated to determine if these new features improved classification accuracy and reduced misclassification costs.

Results

1. Baseline Logistic Regression Performance

The initial logistic regression model, trained using customer financial and demographic data, served as the baseline for comparison. Cross-validation results indicate an **F1 score of 0.825**, an **average cost per fold of 60.20**, and a **precision-recall balance optimized using cost-sensitive thresholding** (Table 1).

2. Gaussian Mixture Model (GMM) Results

GMM clustering was applied to the dataset to extract new features for subsequent logistic regression analysis. The optimal number of clusters was determined based on **Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC)** scores (Figure 3). The best clustering configuration resulted in **three clusters**. When these cluster assignments were incorporated into the logistic regression model, the performance showed no improvement: the **F1 score remained 0.825**, the **average cost per fold increased slightly to 296.20** and Precision & Recall **remain the same level (0.703 and 0.999)** (Table 2).

3. K-Means Clustering Results

K-Means clustering was tested with an optimal **K value selected using the Elbow Method** (Figure 4). From the graph, there is no obvious elbow point, which to some extent indicates that this dataset may not be suitable for K-means. Despite optimizing for different cluster counts, the derived clusters did not improve model performance. The **final F1 score remained at 0.825, and the average cost per fold was 293.60** (Table 3). The Precision **slightly increased by 0.1%** while the Recall slightly decreased by 0.3%.

4. Hierarchical Clustering Results

Number of clusters was evaluated using a **dendrogram for optimal cluster selection** (Figure 5), and **15** clusters were chosen. The silhouette score for hierarchical clustering was **0.08**, indicating weak cluster cohesion (Figure 6). The subsequent logistic regression model using hierarchical clusters as features yielded an **F1 score of 0.826** with an **average cost per fold of 293.20**, reflecting no substantial improvement over the baseline (Table 4). The Precision **slightly increased by 0.2%,** and the **recall remained at 99.9%.**

Across all three unsupervised learning methods, the resulting logistic regression models failed to outperform the baseline classifier. The clustering structures detected did not provide additional discriminative power for the classification task, indicating that unsupervised feature extraction did not meaningfully enhance predictive accuracy in this case.

Conclusion

This study examined the potential of unsupervised learning to enhance credit risk classification. The results show that all clustering methods tested (GMM, K-Means, hierarchical clustering) did not provide meaningful improvements over the baseline logistic regression model. The low silhouette scores indicate that the dataset does not exhibit a strong clustering structure, so these methods are not effective for feature engineering in this case. Thus, no corresponding recommendations and guidance can be provided to companies.

While unsupervised learning remains a powerful tool in a variety of applications, its effectiveness is highly dependent on dataset characteristics. In this case, clustering methods did not provide meaningful segmentation, which emphasizes the importance of data-specific method selection.

Given these findings, other machine learning methods not yet mentioned in this course would be used in the future to test and improve this scenario. This includes alternative feature engineering techniques (e.g., polynomial features, interaction terms) and ensemble learning methods (e.g., random forests, XGBoost).

References

Gallop, Robert J., Paul Crits-Christoph, Larry R. Muenz, and Xin M. Tu. 2003. "Determination and Interpretation of Optimal Operating Point ROC Curves Derived Through Generalized Linear Models." *Understanding Statistics* 2, no. 4: 219-242.

Lucas, Yvan, and Johannes Jorgovsky. 2020. "Credit Card Fraud Detection Using Machine Learning: A Survey."

Pang, Guansong, Chunhua Shen, Longbing Cao, and Anton van den Hangel. 2020. "Deep Learning for Anomaly Detection: A Review."*ACM Computing Surveys*.

Zweig, Mark H., and Gregory Campbell. 1993. "Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine." *Clinical Chemistry* 39, no. 4: 561-577. See page 572 for a discussion of cost-based cutoff determination.

Appendix

Figure 1: Credit Amount Distribution

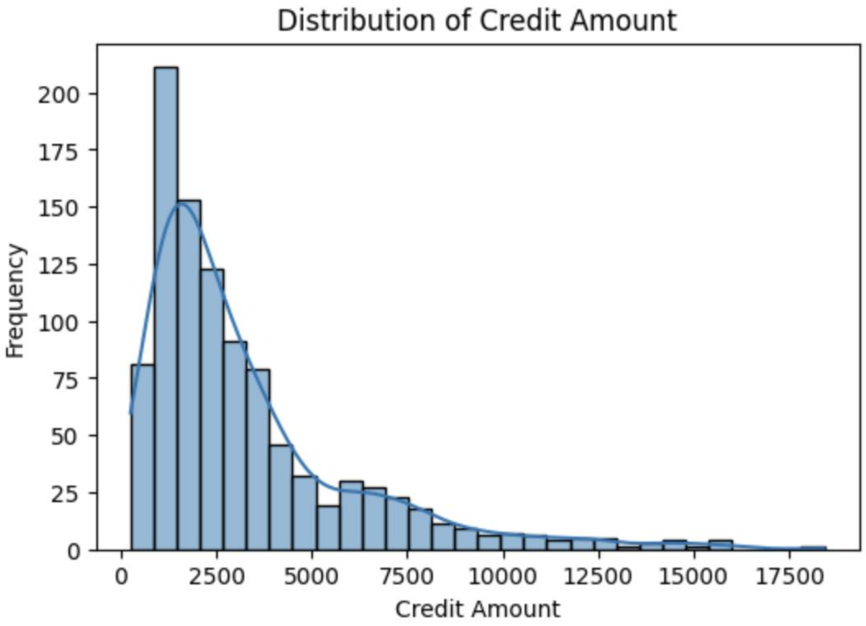


Figure 2: Personal Status Distribution

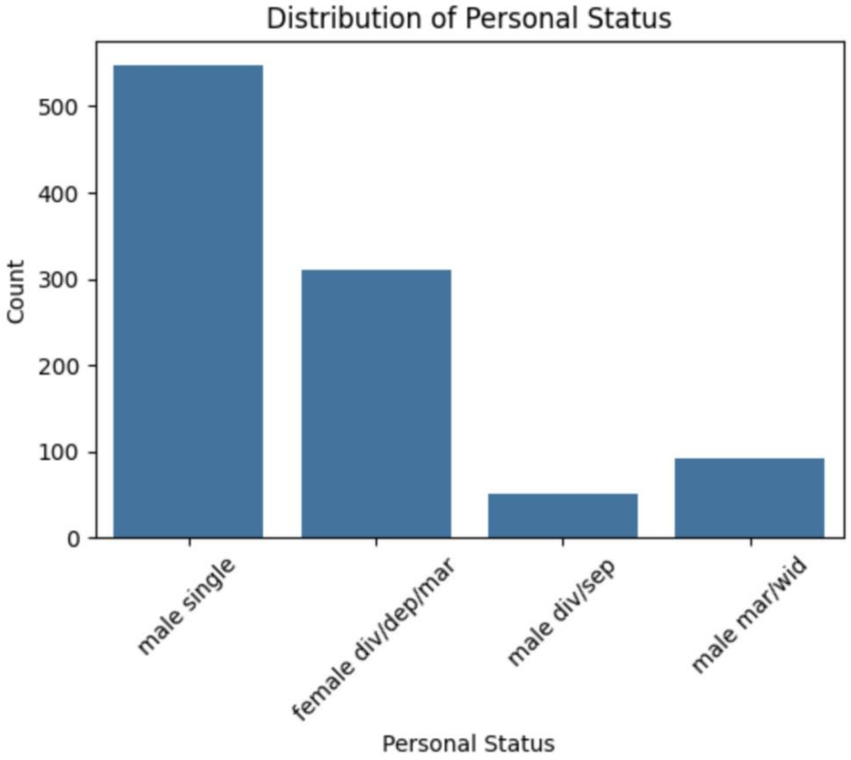


Figure 3: GMM Model Selection using BIC/AIC

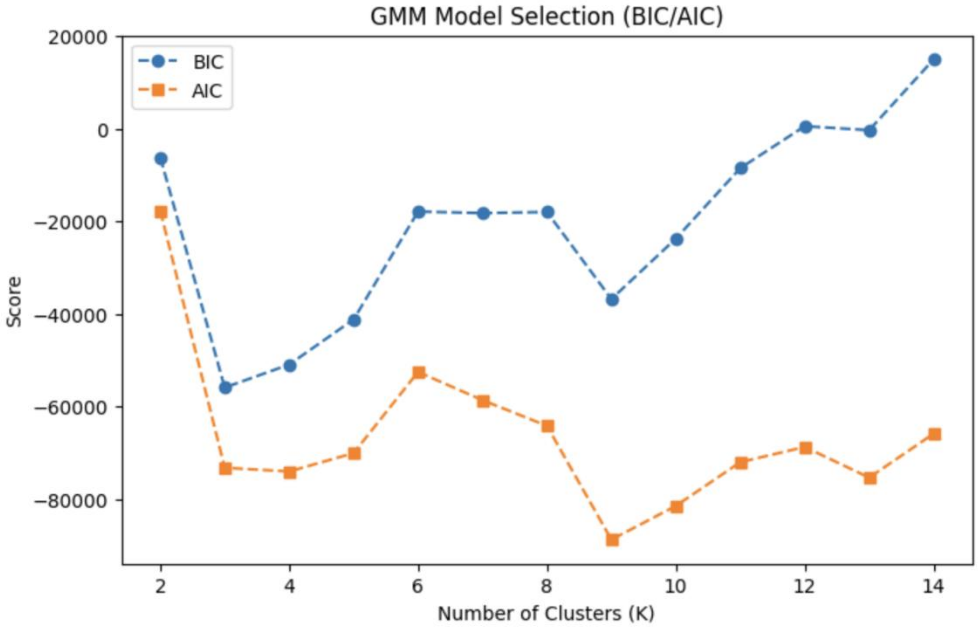


Figure 4: K-Means Elbow Method for Optimal K

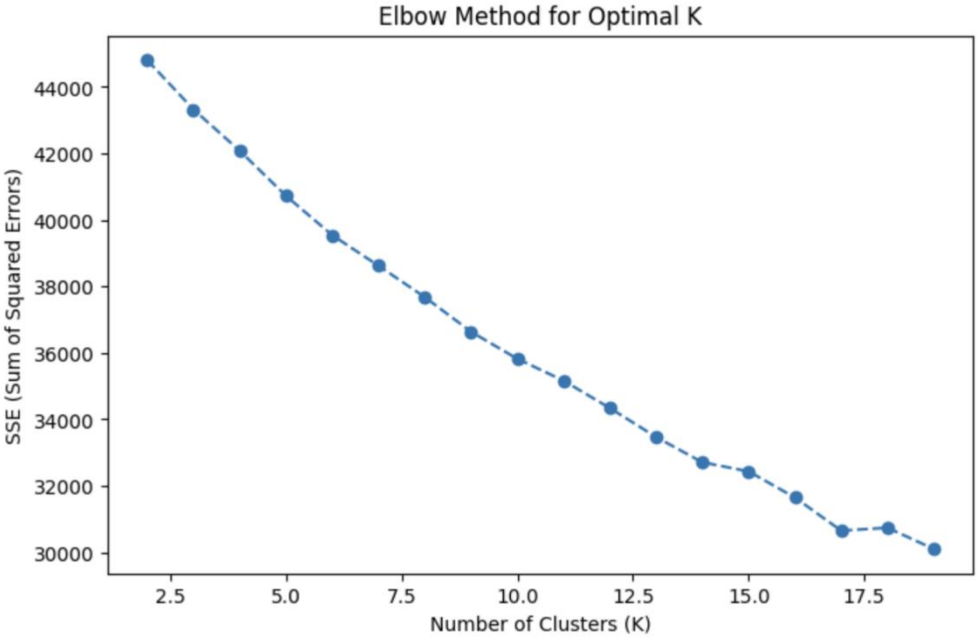


Figure 5: Hierarchical Clustering Dendrogram

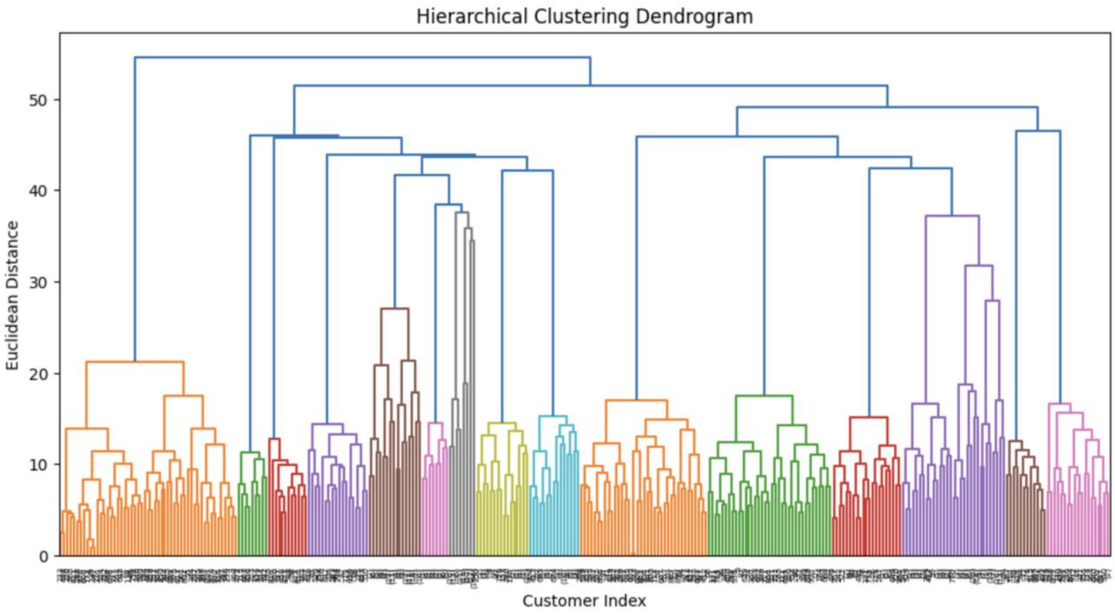


Figure 6: Hierarchical Clustering Silhouette Analysis

```
Hierarchical Clustering Silhouette Score: 0.080
```

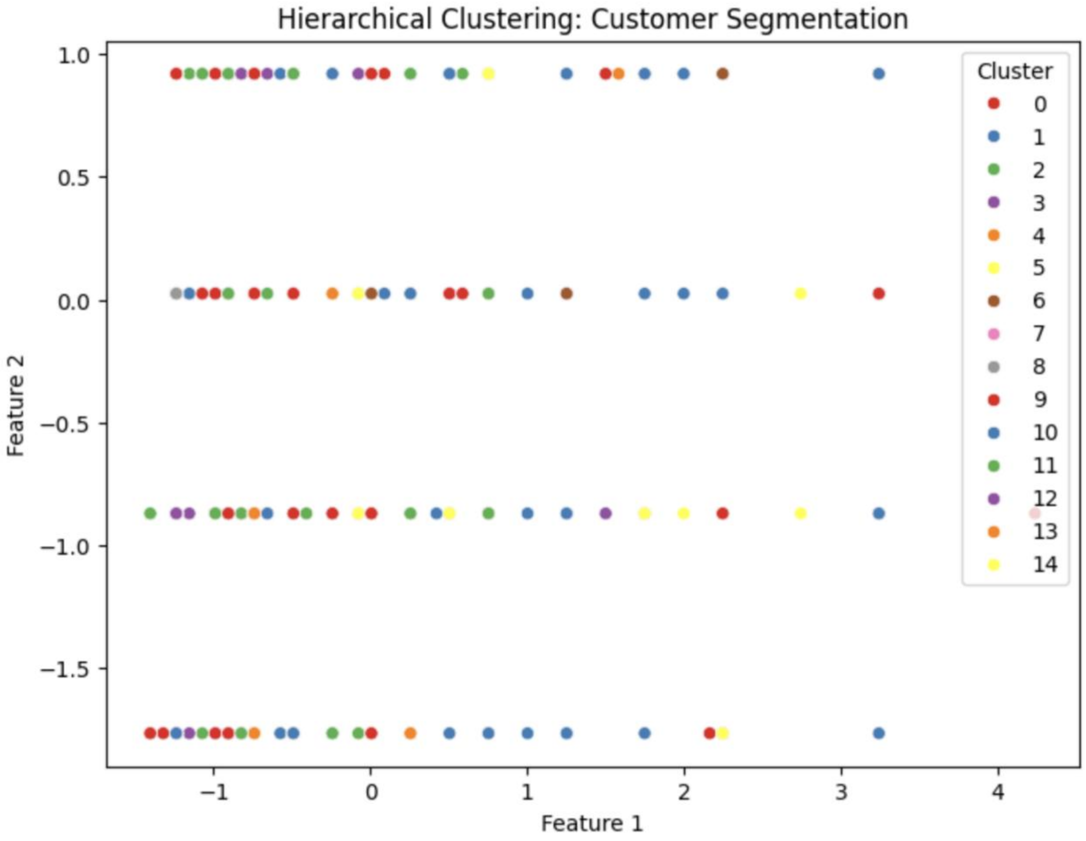
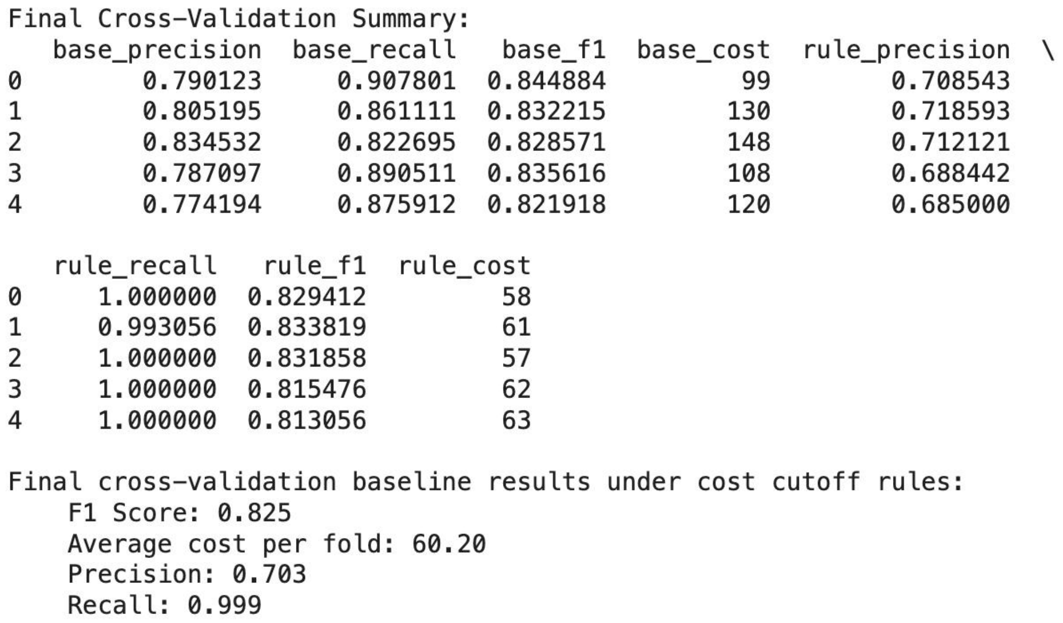


Table 1: Baseline Logistic Regression Performance

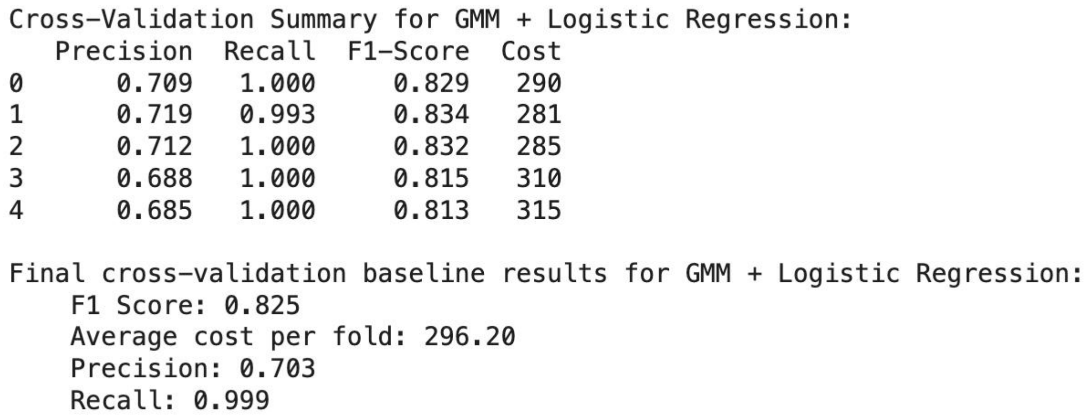


```
Final Cross-Validation Summary:
   base_precision  base_recall   base_f1  base_cost  rule_precision  \
0        0.790123     0.907801  0.844884         99        0.708543
1        0.805195     0.861111  0.832215        130        0.718593
2        0.834532     0.822695  0.828571        148        0.712121
3        0.787097     0.890511  0.835616        108        0.688442
4        0.774194     0.875912  0.821918        120        0.685000

   rule_recall   rule_f1  rule_cost
0     1.000000  0.829412         58
1     0.993056  0.833819         61
2     1.000000  0.831858         57
3     1.000000  0.815476         62
4     1.000000  0.813056         63

Final cross-validation baseline results under cost cutoff rules:
    F1 Score: 0.825
    Average cost per fold: 60.20
    Precision: 0.703
    Recall: 0.999
```

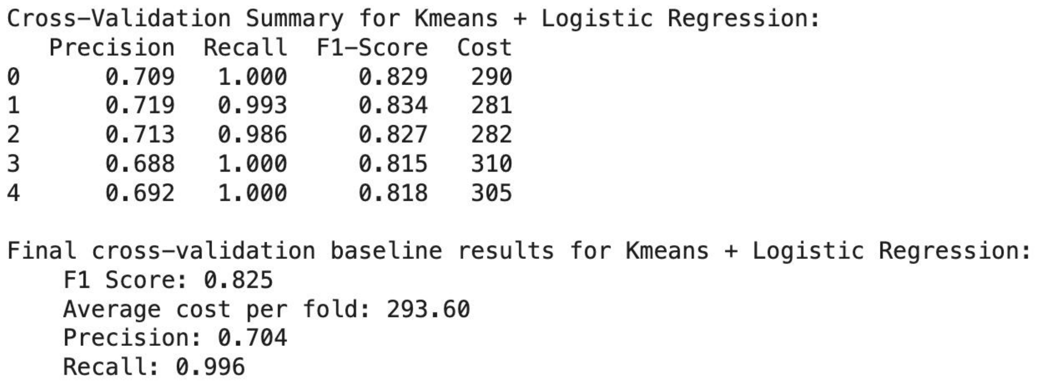
Table 2: Cross-Validation Results for GMM + Logistic Regression Performance



```
Cross-Validation Summary for GMM + Logistic Regression:
   Precision  Recall  F1-Score  Cost
0      0.709   1.000     0.829   290
1      0.719   0.993     0.834   281
2      0.712   1.000     0.832   285
3      0.688   1.000     0.815   310
4      0.685   1.000     0.813   315

Final cross-validation baseline results for GMM + Logistic Regression:
    F1 Score: 0.825
    Average cost per fold: 296.20
    Precision: 0.703
    Recall: 0.999
```

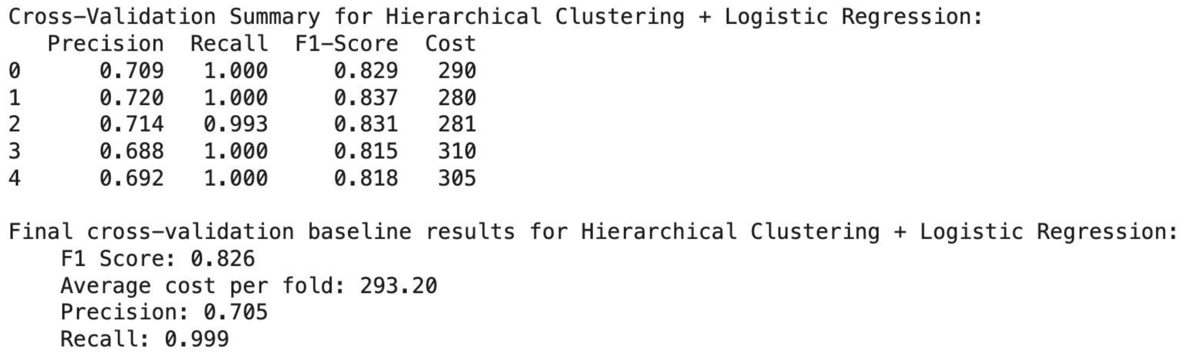
Table 3: Cross-Validation Results for K-Means + Logistic Regression Performance



```
Cross-Validation Summary for Kmeans + Logistic Regression:
   Precision  Recall  F1-Score  Cost
0      0.709   1.000     0.829   290
1      0.719   0.993     0.834   281
2      0.713   0.986     0.827   282
3      0.688   1.000     0.815   310
4      0.692   1.000     0.818   305

Final cross-validation baseline results for Kmeans + Logistic Regression:
    F1 Score: 0.825
    Average cost per fold: 293.60
    Precision: 0.704
    Recall: 0.996
```

Table 4: Cross-Validation Results for Hierarchical Clustering + Logistic Regression Performance



```
Cross-Validation Summary for Hierarchical Clustering + Logistic Regression:
   Precision  Recall  F1-Score  Cost
0      0.709   1.000     0.829   290
1      0.720   1.000     0.837   280
2      0.714   0.993     0.831   281
3      0.688   1.000     0.815   310
4      0.692   1.000     0.818   305

Final cross-validation baseline results for Hierarchical Clustering + Logistic Regression:
    F1 Score: 0.826
    Average cost per fold: 293.20
    Precision: 0.705
    Recall: 0.999
```