Final Project: Data-Driven Content Strategy for Streaming Service

Pelin Nisa Top

Introduction

This report synthesizes a comprehensive analysis of a significant dataset from a premier streaming service, encompassing 8,804 individual titles. The data embodies a wide array of content characteristics including genre classifications, audience ratings, release years, and more. The intent behind this analysis is to distill strategic insights that can directly influence content acquisition, customer engagement, and competitive positioning.

Clustering with K-Means

Methodological Justification for K=5 Clustering

In deploying the K-Means clustering algorithm, we sought to discern a structure within our content repository that could be segmented into meaningful groups. The determination to segment the content into five distinct clusters was guided by the analysis of within-cluster sum of squares, a statistical measure indicating the homogeneity of items within each cluster. The decision for five clusters represents an equilibrium—sufficiently detailed to provide nuanced insights while avoiding an overly fragmented view that could obfuscate strategic action.

Business Insights from Clustering

Strategic content clustering affords us the opportunity to unveil patterns that can inform targeted marketing campaigns, content recommendations, and inventory management. By grouping titles into five clusters, we can assess content performance, identify potential investment opportunities in specific genres, and enhance user engagement through personalized content discovery paths.

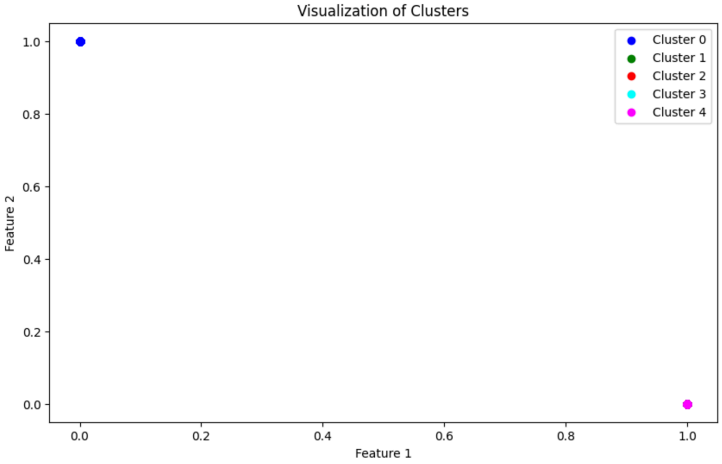


Figure 1: Visualization of Clusters

Analytical Summary of Clustering

The following table encapsulates the key performance indicators of our clustering model:

Metric	Value
Mean Absolute Error (MAE)	11.21
R^2 Score	0.888
Optimal Number of Clusters (k)	5
Variance Explained by Clusters	34.18%

The MAE and R^2 Score serve as quantifiable measures of the clustering model's precision and robustness. An MAE of 11.21, while moderately low, suggests there is room for model refinement, potentially through the integration of additional content attributes. The R^2 score, approaching the upper threshold of 1, reflects a strong model fit, corroborating the clusters as reflective of inherent data patterns.

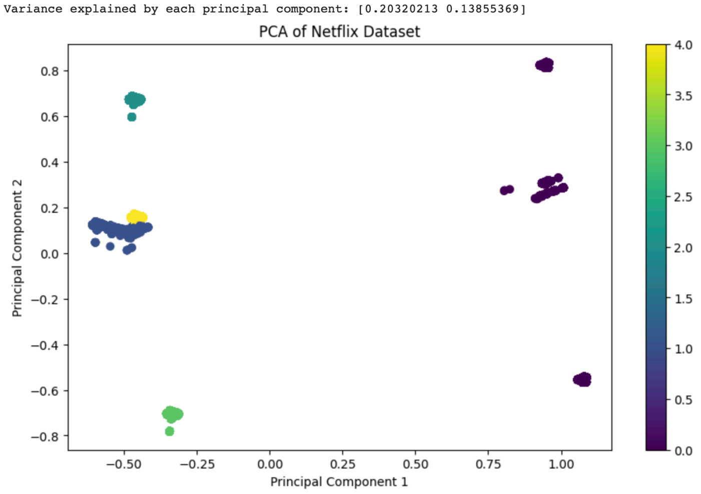
Dimensionality Reduction with PCA

PCA's Role in Data Interpretation

Principal Component Analysis (PCA) has been applied to distill the high-dimensional dataset into a two-dimensional graphical representation, which facilitates the understanding of the content's distribution and diversity. The PCA model's first two components capture 34.18% of the dataset's variance, an indicator of the complexity and richness of the data. While not exhaustive, this level of variance captured is significant in discerning broad patterns and relationships.

Strategic Application of PCA Insights

The visual distribution gleaned from PCA offers a strategic vantage point from which to comprehend the breadth and depth of the content library. These insights are invaluable for content positioning, identifying gaps in the content catalog, and directing investment to fulfill the uncovered market demands.



```
Variance explained by each principal component: [0.20320213 0.13855369]
```

Figure 2: PCA of Netflix Dataset

Conclusion

Through the methodologies of K-Means clustering and Principal Component Analysis, this report has illuminated the underlying content structures and relationships within the streaming service's extensive library. The actionable intelligence derived from this analysis equips content strategists and decision-makers with a data-driven foundation to optimize content offerings, enhance user satisfaction, and solidify market position in an increasingly competitive landscape.