**Data Mining II Performance Assessment Task #2**

**D212**

**By Josue Gonzalez**

Contents

[Part I: Research Question 3](#_Toc180522048)

[Part II: Technique Justification 3](#_Toc180522049)

[Part III: Data Preparation 5](#_Toc180522050)

[Part V: Data Summary and Implications 9](#_Toc180522051)

[1. Explain the Quality of the Clusters: 9](#_Toc180522052)

[2. Discuss the Results and Implications: 9](#_Toc180522053)

[3. Discuss One Limitation of the Analysis: 10](#_Toc180522054)

[4. Recommend a Course of Action: 10](#_Toc180522055)

# Part I: Research Question

**A. Purpose of the Data Mining Report**

**A1.Proposed Research Question:**

*How can principal component analysis (PCA) be applied to identify key customer characteristics within a telecommunications company’s customer base, in order to enhance targeted product offerings and improve long-term profitability?*

**A2.Goal of the Analysis:**

To use principal component analysis (PCA) to identify the key characteristics of telecommunications customers from the available data, focusing on variables like monthly charges, bandwidth usage, and customer satisfaction metrics (represented by items such as 'Item1' through 'Item8') which will be renamed accordingly based on their data dictionary names. This analysis aims to reduce dimensionality while preserving critical customer insights, thereby enabling the company to better target products and improve strategic decision-making.

# Part II: Method Justification

**B1.Explaination of PCA**

The goal of PCA is to simplify complex datasets by reducing their dimensionality while retaining the most important information. This technique works by converting a set of correlated variables into fewer uncorrelated components, each capturing a significant portion of the total variance. These components are ranked based on how much variance they explain. In this analysis, PCA helps identify the most influential factors related to customer churn, offering clearer insights into which demographic, service usage, and satisfaction variables have the greatest impact, thereby enabling more targeted strategic decisions. (Jolliffe & Cadima, 2016).

The expected outcome of applying PCA to the telecommunications customer dataset is a streamlined set of principal components that represent the most critical variations in customer characteristics. These components will reduce the complexity of the data, helping to capture the essential traits of customers without needing every variable. For instance, rather than analyzing each individual variable, such as monthly charges, bandwidth usage, and satisfaction scores separately, PCA will transform them into fewer, comprehensive factors. This dimensionality reduction will make it much easier to interpret and analyze patterns, as we will be able to focus on a handful of key components instead of a large number of disparate variables.

Through these principal components, the company can expect to uncover distinct customer characteristics that provide actionable insights. For example, one of the principal components may emerge as a measure of "cost sensitivity," indicating how customer monthly charges and other pricing-related factors contribute to their overall behavior. Another component may capture "satisfaction-driven" tendencies, based on aggregated customer ratings across several satisfaction indicators. Identifying these primary components will allow the company to segment its customers more effectively, targeting each group with tailored products, services, or marketing campaigns. This approach will ultimately lead to more strategic, data-driven decisions that can improve customer retention and maximize profitability by aligning company offerings with the actual needs and preferences of different customer groups.

**B2. PCA Assumptions:**

PCA relies on several key assumptions to effectively extract meaningful patterns from the data. First, it assumes that relationships among the variables are linear, making it best suited for linear combinations that explain maximum variance. Second, PCA performs optimally when variables follow a normal distribution, improving the clarity of the components. Additionally, PCA accounts for multicollinearity, where variables are highly interrelated, by converting them into uncorrelated components, maximizing the variance explained and minimizing redundancy. In this analysis, PCA aims to simplify the dataset, focusing on the most influential factors to enhance understanding of customer behavior related to churn.

# Part III: Data Preparation

**C1.Continuous Data set variables:**

The analysis focuses on continuous variables representing key customer characteristics, including age, income, outage seconds per week, tenure, monthly charges, bandwidth usage, and standardized which aligns them with the continuous variables in the analysis. This approach ensures that each variable, including satisfaction scores, contributes proportionately to the PCA. By using these standardized continuous variables, the analysis seeks to pinpoint the most influential factors in customer retention, effectively addressing the research question regarding churn drivers in the telecommunications sector.

**C2. Standardization of data set**

Before applying PCA, it is essential to standardize the continuous variables, given PCA's sensitivity to scale differences. Variables like age, income, outage seconds per week, tenure, monthly charges, and bandwidth usage. This ensure that each variable contributes equally to the analysis while also generating more reliable components. By aligning the scales, this process eliminates biases arising from different units across variables, making the data more suitable for PCA and enhancing its ability to uncover meaningful patterns.

A screenshot of a computer

Description automatically generated

# Part IV: Data Analysis

**D1.Principal Components**

Once the dataset was standardized, PCA was applied to reduce dimensionality and identify the primary drivers of churn. By transforming the data, PCA generates a series of principal components, each representing a combination of the original variables that captures the highest variance. These components act as condensed variables, summarizing the key patterns within the dataset. This analysis used all available continuous variables to generate the components, which were then assessed to determine their role in explaining churn. The insights provided by PCA help reveal which factors have the strongest impact, guiding the creation of targeted retention strategies to boost customer satisfaction. The PCA process produces a loadings matrix, which shows how much each original variable contributes to each principal component. This matrix helps pinpoint the variables most relevant to customer churn, such as age, monthly charges, and satisfaction levels, which had strong loadings on the initial components. The PCA transformation generates a matrix of loadings, which shows how much each original variable contributes to each principal component. This matrix helps to understand which variables have the greatest influence on the components. For example, in our analysis, variables such as **age**, **monthly charges**, and **satisfaction levels** showed strong loadings on the first few principal components, indicating their significance in explaining customer churn.

The following table presents the PCA component matrix, demonstrating the weight that a given feature contributes to the identified principal components. This matrix aids in identifying which variables have the most impact, guiding the interpretation of the PCA results.

**D2.total number of components**

To determine the optimal number of principal components, a scree plot was generated. This plot illustrates the explained variance for each component and helps identify the 'elbow point,' where the curve starts to flatten, indicating diminishing returns in variance capture. This point suggests that additional components add little to the total variance explained and may not significantly enhance the model's interpretability.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A graph with lines and dots

Description automatically generated

In this analysis, the scree plot revealed that the first **4 components** capture over 80% of the total variance, making them the most significant contributors to explaining customer churn. The elbow point was identified at the 8th component, after which the additional variance explained by subsequent components decreased sharply.

By focusing on these top 8 components, we retain the majority of the meaningful variance while reducing dimensionality, making the model both efficient and interpretable. This approach allows for a more streamlined analysis that captures the key patterns without unnecessary complexity.

**D3.variance of each component**

After determining that four principal components capture over 80% of the total variance, the PCA process was refined to focus on these components. The adjusted PCA results provide a clearer understanding of the variance explained by each of the top four components. By limiting the analysis to continuous variables, such as age, income, outage seconds per week, tenure, monthly charges, and bandwidth usage, this PCA analysis retains the most meaningful variance while minimizing the overall complexity of the model. Each component now reflects essential continuous characteristics, capturing key dimensions of customer demographics, service usage patterns, and financial factors without including satisfaction scores or other ordinal data.

A screenshot of a computer

Description automatically generated

The table above shows the **explained variance ratio** for each of the selected 8 components, along with their **cumulative variance ratio**. As observed, the first principal component (PC1) accounts for approximately 33.2% of the total variance, while the cumulative variance for the top 4 components exceeds 80%. This distribution demonstrates the efficiency of the PCA in capturing the essential patterns of the dataset.

With the decision made to retain only 5 components, the PCA loadings were also adjusted accordingly. The final component matrix shows the weight each variable contributes to these selected components. This revised PCA provides a more focused and interpretable model, supporting strategic decisions for customer churn analysis.

**D4.Total variance by components**

A screenshot of a computer code

Description automatically generated

The total variance captured by the components was calculated to evaluate the overall effectiveness of the PCA. The results show that the first four components collectively capture over 80% of the variance, with all components together capturing 100%. This indicates that these top four components are sufficient to summarize the key patterns in the data, making it an optimal choice for analysis while minimizing complexity. These components represent essential continuous dimensions, such as age, income, tenure, and bandwidth usage, which provide valuable insights into customer characteristics and potential churn predictors.

**D5. Summary of data analysis**

The analysis aimed to identify the key factors influencing customer churn in the telecommunications sector using Principal Component Analysis (PCA). The goal was to simplify the dataset and focus on the most significant continuous variables that drive churn, enabling more effective strategic decision-making. Six continuous variables, covering demographics and service usage patterns, were standardized to ensure equal contribution to the PCA, allowing for a fair comparison of variables regardless of their original scales. By excluding ordinal satisfaction scores and focusing solely on continuous factors, the analysis centers on quantifiable characteristics like monthly charges, bandwidth usage, and tenure to identify influential drivers of churn.

A screen shot of a graph

Description automatically generated

The scree plot with cumulative variance indicated that the initial components captured a significant share of the total variance. Specifically, the first principal component (PC1) accounted for around 33.2%, with the second component (PC2) contributing an additional 17%. Combined, the first five components represented over 60% of the total variance, effectively simplifying the dataset while preserving essential information. This analysis supports keeping the top eight components, as they collectively account for more than 80% of the total variance.

A screen shot of a graph

Description automatically generated

The biplot provides a detailed view of how each original continuous variable contributes to the first two principal components. In this adjusted PCA, PC1 reflects a combination of financial and usage-related characteristics, with variables like "MonthlyCharge" and "Bandwidth\_GB\_Year" contributing strongly. PC2, on the other hand, is more aligned with "Tenure" and "Outage\_sec\_perweek," highlighting the importance of service reliability and customer duration. This view helps pinpoint the most influential continuous factors in the dataset, aiding in understanding customer retention and churn drivers.

These findings suggest that a combination of tenure, monthly charges, and bandwidth usage are the most influential continuous factors for predicting churn. The insights emphasize the need for retention strategies that focus on improving service reliability, fostering long-term relationships, and addressing bandwidth needs. By reducing the dataset to a smaller set of principal components based solely on continuous variables, the analysis provides a clearer understanding of churn drivers, allowing for more targeted interventions that enhance customer satisfaction and profitability.

**G Code sources**

[*https://www.statology.org/scree-plot-python/*](https://www.statology.org/scree-plot-python/)

<https://scikit-learn.org/dev/modules/generated/sklearn.decomposition.PCA.html>

H.References

* Cattell, R. B. (1966). "The scree test for the number of factors." Multivariate Behavioral Research, 1(2), 245-276.
* Jolliffe, I. T., & Cadima, J. (2016). "Principal component analysis: A review and recent developments." Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.