Principal Component Analysis of ChatGPT vs. DeepSeek Dataset

Introduction

In this analysis, we examined a dataset comparing two prominent large language models (LLMs): ChatGPT and DeepSeek. The dataset included multiple numerical metrics, such as Active Users, Response Tokens, Session Duration, Response Accuracy, and more. Given the complexity and high dimensionality of this data, interpreting it directly was challenging. Therefore, Principal Component Analysis (PCA) was chosen as an effective method to simplify the data while preserving its most critical information.

Problem and Objective

The main challenge was the high dimensionality of the dataset, making it difficult to visualize or interpret meaningful insights directly. PCA was applied to reduce these multiple dimensions into fewer, interpretable components. The goal was to uncover underlying patterns, identify primary metrics driving differences between the platforms, and simplify future analyses.

PCA Results

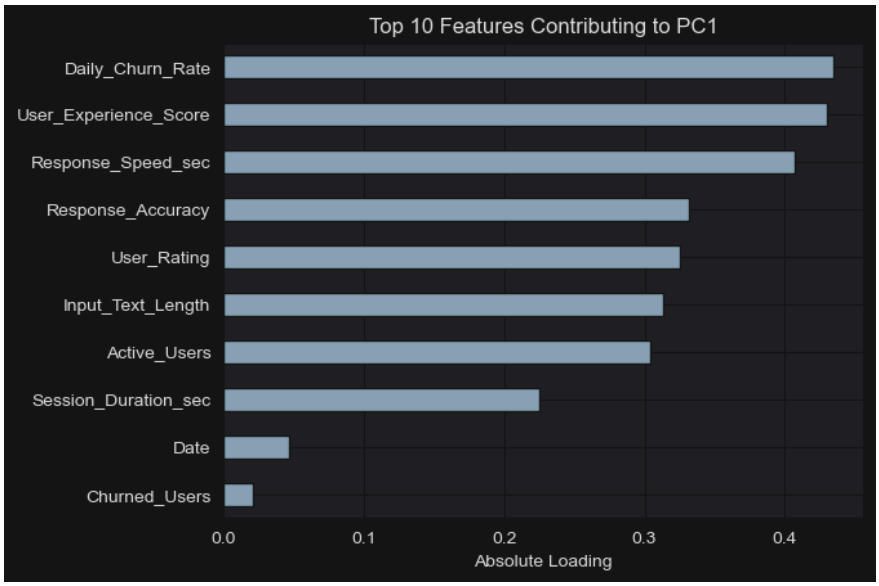
The PCA analysis provided significant insights:

- The first Principal Component (PC1) alone explained approximately **30.9%** of the dataset's total variance, indicating that a substantial portion of the differences across sessions could be captured by this single composite feature.
- Adding the second Principal Component (PC2) explained another **11.04%**, bringing the cumulative variance captured by the first two components to approximately **42%**.
- Including the third component (PC3), the cumulative variance reached around **52%**.
- Approximately **80%** of the variance could be explained by including seven to eight components, effectively simplifying the dataset significantly.

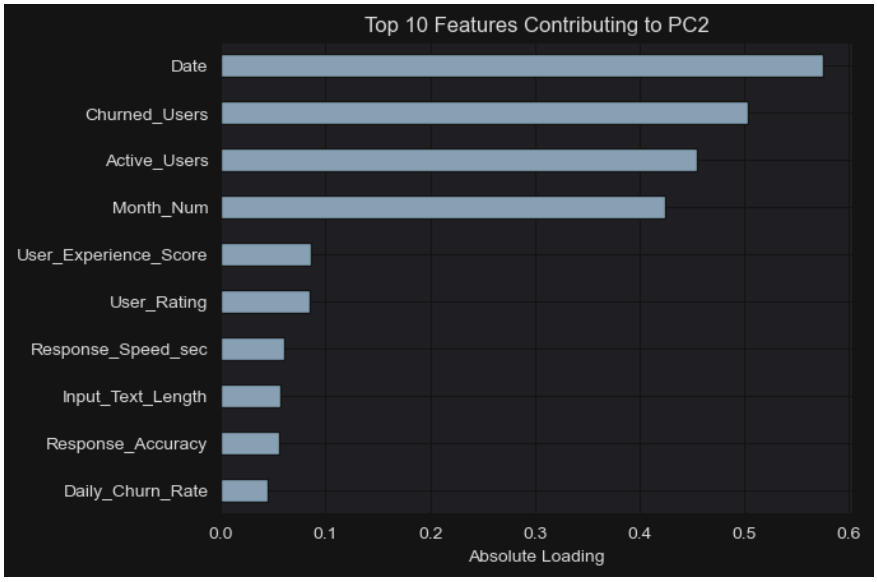
Interpreting PCA Loadings

To better understand what each principal component represents, we examined the loadings, the contributions of each original metric:

- **PC1** was primarily influenced by features like **Daily Churn Rate**, **User Experience Score**, and **User Experience Score**. This indicates that PC1 predominantly captures overall engagement and response complexity.



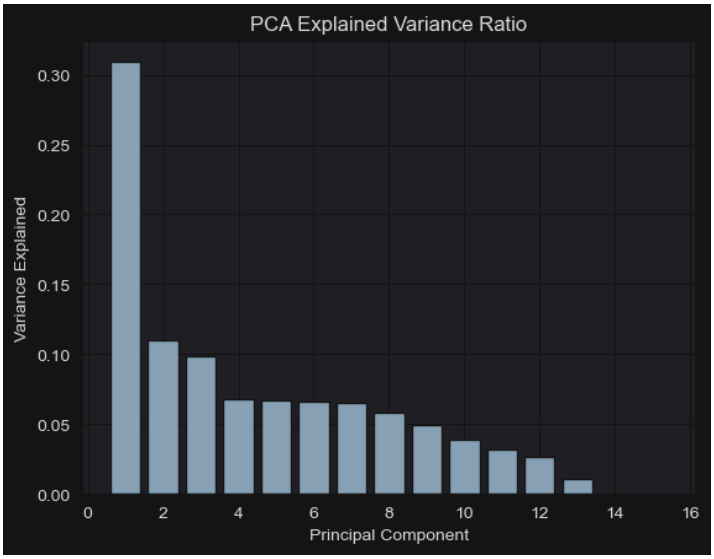
- **PC2** was influenced most strongly by features such as **Date**, **Churned Users**, and **Active Users**, suggesting this component mainly focused on the user interaction rate.



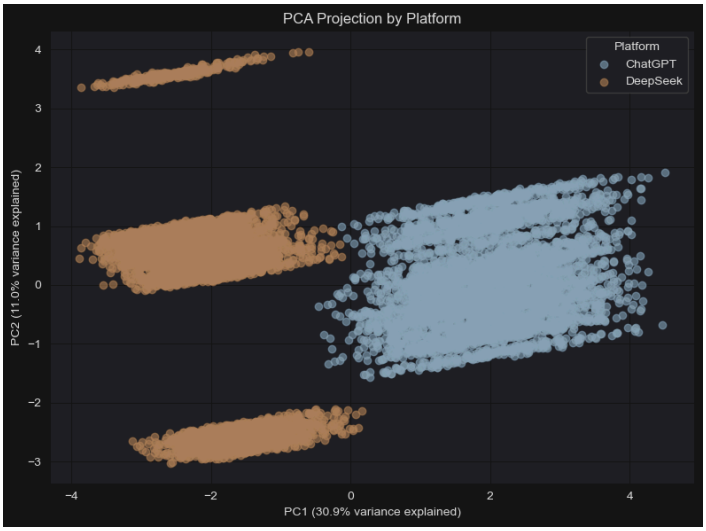
Visualization Insights

- **Explained Variance Bar Plot:**
    - This plot clearly demonstrates diminishing returns from each additional component. The steep drop-off after the first few components provided a visual

rationale for focusing analysis on fewer dimensions.



- **Scatter Plot of PC1 vs. PC2:**
  - The scatter plot clearly illustrates distinct clustering patterns between ChatGPT and DeepSeek, showing how effectively these principal components differentiate between the platforms. Each data point represented a session projected onto these two primary dimensions, revealing insights into similarities or differences in platform behaviors and user experiences.



Conclusion

Applying PCA successfully simplified a complex dataset, highlighting key dimensions that differentiate ChatGPT and DeepSeek platforms. These principal components offer a concise

yet comprehensive view into user interaction patterns and platform performance characteristics, significantly easing subsequent data-driven decisions or further analytical investigations.