Microsoft Security: User Segmentation through PCA

Sponsor: Microsoft

Hillary Zhang, Brevin Simon

Washington State University
School of Electrical Engineering and Computer Science

CySER

Project Statement of Purpose:

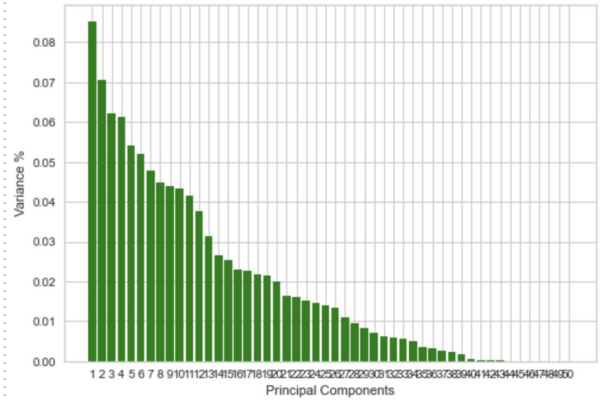
This project aims to create accurate user profiles from large data to better detect usage anomalies within Microsoft Security.

Microsoft

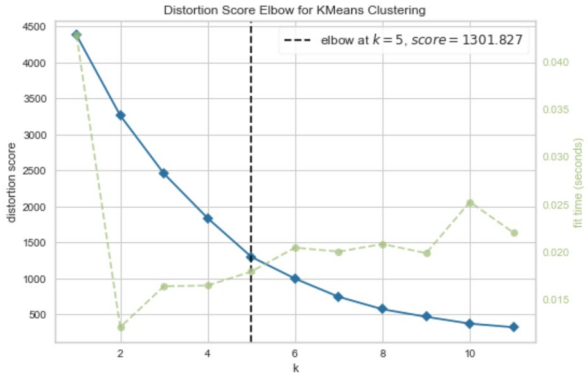
Telemetry Data (Ordinal)

UserID	IntelligencePage	IndicatorPage
L82SA	2	2
8QPZ2	6.3	2.3

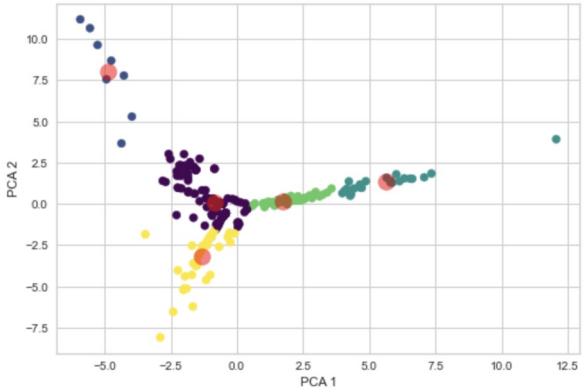
Finding The Principle Component



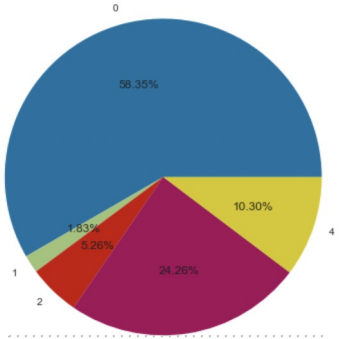
Number of Clusters: Elbow Method



KMeans Clustering with Centroids



After segmenting the users we were able to have our KMeans model assign each user into their clusters that most matched their behavior.



Final Cluster Distribution

Then we still had to go through each cluster and figure out what each cluster meant from a persona standpoint. For example, if the users who were in Cluster 2 were the only users to work in IntelligencePage we can denote them “Security”.

Glossary

CSV: Comma Separated Values
Ordinal: Data built sequentially, for example First, Second, Third.
Principal Component Analysis (PCA): Algorithm to reduce dimensionality of data and remove data features that matter less than others.
Azure: Microsoft’s cloud server application
KMeans: Unsupervised segmentation machine learning technique
Key Performance Indicators (KPI): Critical indicators of progress toward a resulting goal.

Special thanks to Dr. Assefaw, who helped guide this project to completion