Microsoft Security: User Segmentation through PCA

Sponsor: Microsoft

Hillary Zhang, Brevin Simon

Project Statement of Purpose:
This project aims to create accurate user profiles from large data to better detect usage anomalies within Microsoft 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.

User ID | IntelligencePage | IndicatorPage
--- | --- | ---
L825A | 2 | 2
B292 | 6.3 | 2.3

Finding The Principle Component

<table>
<thead>
<tr>
<th>Principal Components</th>
<th>Variance %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.08</td>
</tr>
<tr>
<td>2</td>
<td>0.06</td>
</tr>
<tr>
<td>3</td>
<td>0.04</td>
</tr>
<tr>
<td>4</td>
<td>0.03</td>
</tr>
<tr>
<td>5</td>
<td>0.02</td>
</tr>
<tr>
<td>6</td>
<td>0.01</td>
</tr>
<tr>
<td>7</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Telemetry Data (Ordinal)

Number of Clusters: Elbow Method

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

KMeans Clustering with Centroids

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”.

Final Cluster Distribution

Glossary

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