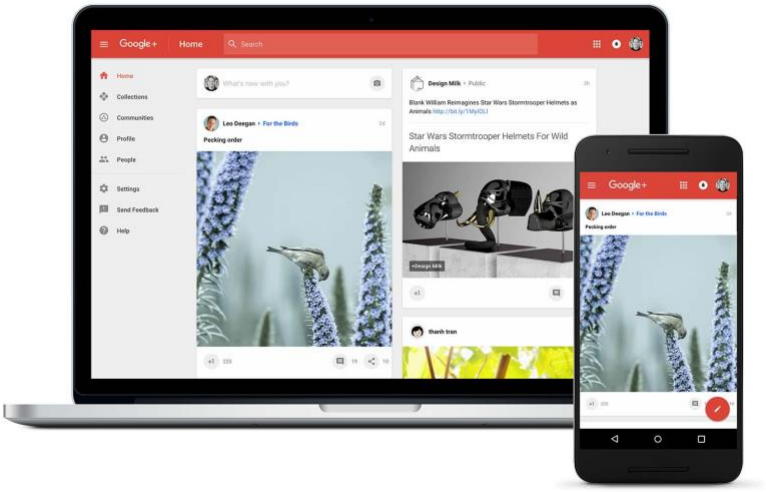
ISyE 6740 – Fall 2022
Final Report

Team Member Names: Syefira Shofa

Project Title: Google Plus User Analysis

Problem Statement

Social network platforms have risen in popularity since the very first networking site launched in 1997 and is now a part of everyday life for most internet users. They are commonly used as tools for socializing, business and dating. While online relationships are often seen as less valuable than offline ones, social platforms can serve to complement offline relationships.



Major companies have tried entering the social networking market. While some have become successful, others have failed. Google+ was a failed venture that was live from 2011-2019. During its lifetime, Google+ provided an extensive list of features and a redesign in 2015. The primary purpose of Google+ was to increase the quality of its search engine results by enabling Google to factor in "social cues" to deliver search results. Thus, Google+ activity was a major factor in SEO. The goal of this project is to identify how a user operated within the Google+ ecosystem and see how Google+ fulfilled the needs of the user.

.

Data Source

The data used for this analysis was the Google Plus dataset provided by the Stanford Large Network Dataset Collection https://snap.stanford.edu/data/. The Google Plus dataset includes node features (profiles), circles, and ego networks.

The data were separated into 132 separate sets of data, one for each ego. Each set of data has the following:

- nodeId.edges : The edges in the ego network for the node 'nodeId'. Edges are directed (a follows b). The 'ego' node does not appear, but it is assumed that they follow every node id that appears in this file.

- nodeId.circles : The set of circles for the ego node. Each line contains one circle, consisting of a series of node ids. The first entry in each line is the name of the circle.

- nodeId.feat : The features for each of the nodes that appears in the edge file.

- nodeId.egofeat : The features for the ego user.

- nodeId.featnames : The names of each of the feature dimensions. Features are '1' if the user has this property in their profile, and '0' otherwise.

For this project, we focus on user_id 100129275726588145876. For this user's social connections, the following attributes were available:

- Gender
- Institution
- Job Title
- Last Name
- Place
- University

Methodology

Clustering

Clustering will be used in order to find the types of users our user has an interest in following and placing into a circle. KModes clustering is an unsupervised machine learning algorithm used to cluster categorical variables. It uses the mode in order to evaluate the similarity and dissimilarity between data points and defines clusters based on the number of matching categories between data points.

The silhouette method is used to find the optimal number of clusters. It computes silhouette coefficients of each point that measures how much a point is similar to its own cluster compared to other clusters. This measure ranges from -1 to 1. A silhouette coefficient near +1 indicates that the sample is far away from neighboring clusters, 0 indicates that the sample is on or very close to the decision boundary between two neighbors and negative values indicate that samples might have been assigned to the wrong cluster.

Feature Importance

Classification models are used in order to see what features leads our user to place someone that they follow into a circle. For this, we compare and contrast the results of bagging and boosting. The main difference between the 2 is that bagging is a method of merging the same type of predictions whereas boosting is a method of merging different types of predictions.

Random forest is a type of bagging approach. Bagging is a weak learners model that learns from each other independently in parallel and combines them for determining the model average. Bagging runs weak learners on bootstrap replicates of the training set. We then average weak learners and reduce the variances. In bagging, each model receives an equal weight, models are

built independently and training data subsets are drawn randomly with a replacement for the training dataset.

Gradient boosting is a type of boosting technique. Boosting is a weak learners model where the learners learn sequentially and adaptively to improve model predictions of a learning algorithm. Boosting runs weak learners on a weighted set. The weak learners are combined linearly. This typically requires knowledge on the performance of weak learners. In boosting, models are weighed based on their performance, new models are affected by a previously built model's performance and every new subset comprises the elements that were misclassified by previous models

Data Preparation

For the analysis, last name was dropped as a variable as it did not provide useful, generalized insights for the user. The dataset had to be cleaned manually for the attribute's institution, job title, place and university. The values seemed to have been inputted manually and, thus, there were misspellings and variations of the same value. For example, Google was also inputted as Google (Android), Google Inc., Google+, and Google, Inc.

Evaluation and Final Results

User

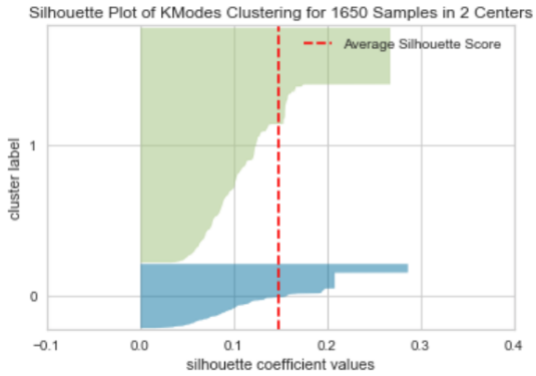
This analysis focused on user id 100129275726588145876. This particular user had the following attributes:

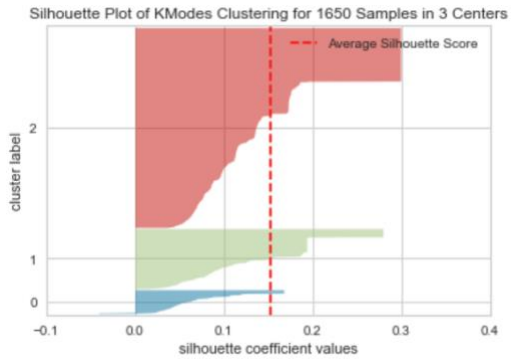
- Gender: 1
- Job Title: Researcher, undergraduate, university
- Place: Fajardo, San Juan
- University: Polytechnic University of Puerto Rico

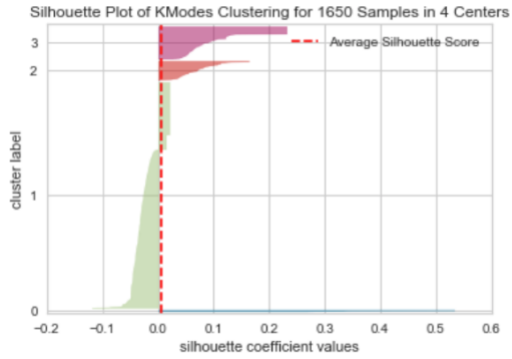
Following

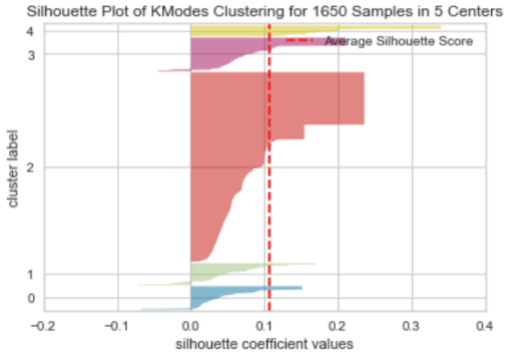
A user had the ability to follow other users in the Google+ ecosystem and did not require that the other user followed them back. Our dataset contains 1650 users that our user followed. Clustering was used in order to find natural groups the user had an interest in following.

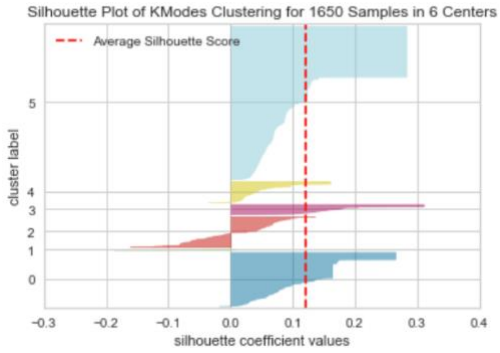
Silhouette Plot Results







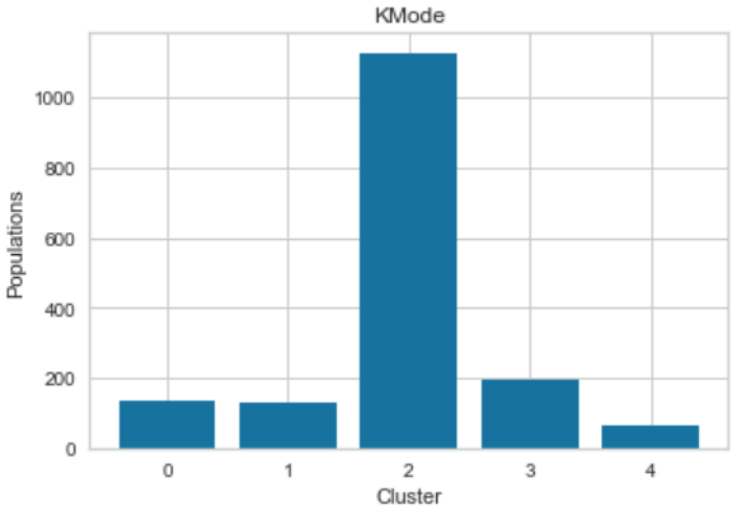




Cluster size of 5 was chosen as the clusters had above average silhouette scores and had decent group sizes. Cluster sizes of 2 and 3 were also under consideration.

KMode Results

	gender:1	gender:2	institution:Google	job_title:developer	job_title:engineer	job_title:school	job_title:software
0	1	0	0	1	0	0	0
1	1	0	1	0	1	0	1
2	1	0	0	0	0	0	0
3	0	1	0	0	0	0	0
4	1	0	0	0	0	1	0

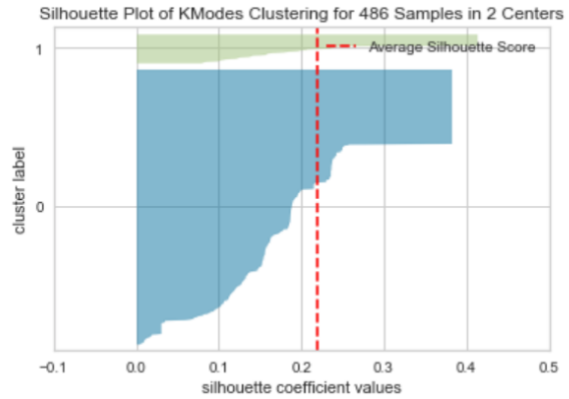


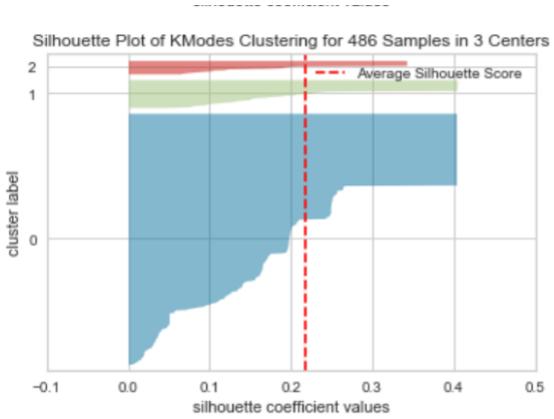
Clusters 2 & 3 of the clusters are simply defined by gender. The other 3 clusters give us insight into the type of users our user likes to follow: developer, software engineer at Google and school. Here, school most likely refers to someone attending a school and not yet working.

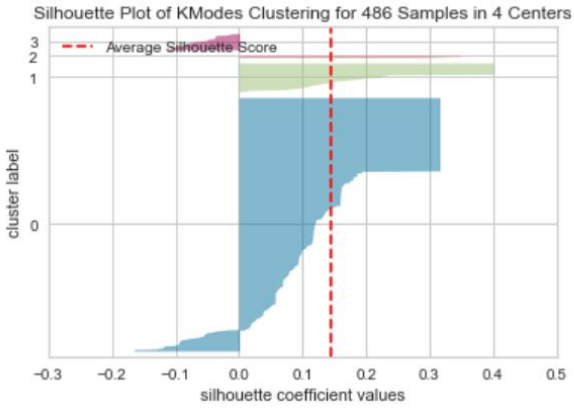
Circles

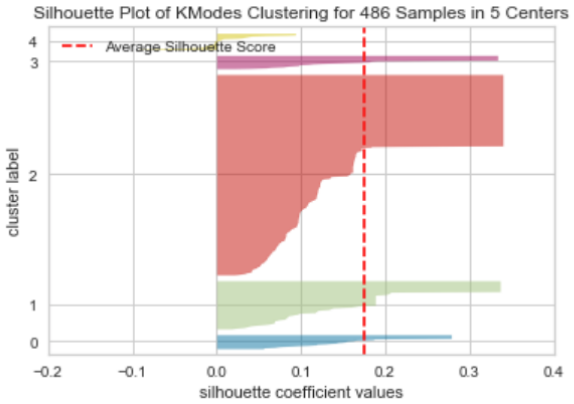
Circles was a feature that lets users group contacts. Users had either the option of using the default circles provided by Google+ or could create their own. Whenever a user posted a stream update, they had the option of choosing which circles could see that update. A single contact could be in multiple circles at once and users that our user were following can also be placed in a circle. This user had 486 contacts placed in circles and the data shows that in the platform the user had 2 circles. Clustering was used in order to find natural groups the user had an interest in placing into a circle.

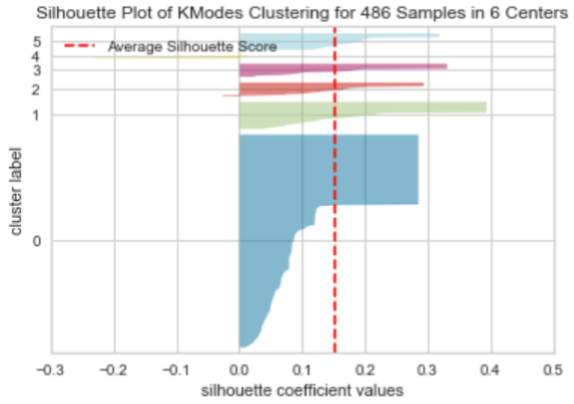
Silhouette Plot Results







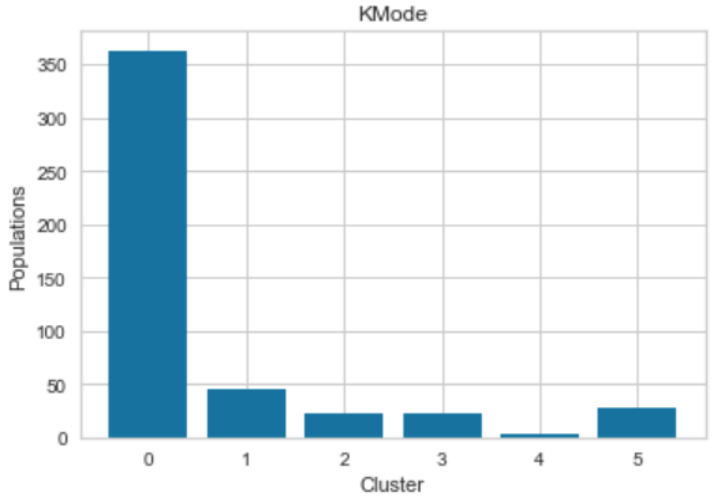




Cluster size of 6 was chosen as the clusters had above average silhouette scores and had decent group sizes. Cluster sizes of 3 and 5 were also under consideration.

KMode Results

	gender:1	gender:2	job_title:developer	job_title:network	job_title:programmer	job_title:school	job_title:software
0	1	0	0	0	0	0	0
1	0	1	0	0	0	0	0
2	1	0	0	0	1	0	0
3	1	0	1	0	0	0	1
4	1	0	0	1	0	0	0
5	1	0	0	0	0	1	0

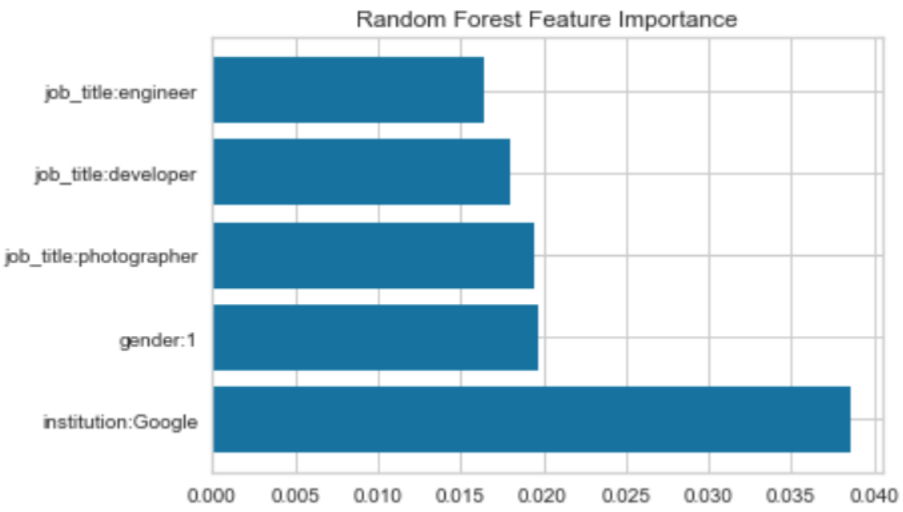


2 of the clusters classified are defined solely by gender. Interestingly, 4 of the clusters were primarily defined by job titles: programmer, software developer, networker and school. Here, the job title school most likely means that the user is still attending school as our user is currently an undergraduate student.

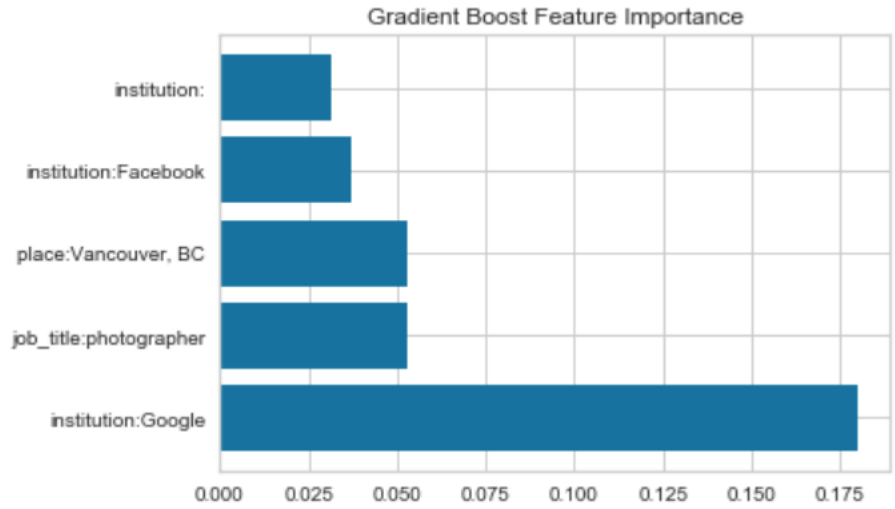
Feature Importance

Our user only put 29.45% of the users they followed into a circle. Random forest and gradient boosting models are used in order to see what features leads our user to place someone that they follow into a circle.

Random Forest Results



Gradient Boosting Results



Feature Importance Conclusion

Based on the results of the two models, our user was more likely to place someone into a circle if they worked at Google or Facebook. Moreover, their job titles most likely had to have been either an engineer or developer. Our model also indicates that our user might have had an interest in photography since they were more likely to follow photographers.

Conclusion

We set out in order to understand how a user who was an undergraduate student at a polytechnic university used the Google+ platform. Google+ had a system where users could follow another user and then place them in a circle if they wanted to. We learned that natural groups that our user follower were those that were developers, software engineer at Google and students. Also, our user liked to place programmers, software developers, networkers and students in circles. But, not everyone that our user followed was placed into a circle. Our user

had a higher chance of placing into a social circle someone that was an engineer/developer at Facebook or Google as well as someone who was a photographer. From this, it’s possible to hypothesize that our user was using Google+ in order connect with those whose careers they aspired to. Google+, when it was alive, did not find success as a social media platform. But, given that our random user turned out to use it in for personal interest and career aspirations, we can hypothesize that perhaps if Google had used SEO to power Google+ as well or found better ways for followers to interact with those they followed, it may have been more successful.