School of Management

Coursework Submission Sheet

Student name: Suphakorn Homnan

Department: Management

Programme and Year of Study: Business Analytics 2023-2024

Name of lecturer: Brittany Davidson

Unit title and code: MN50752 Data Mining & Machine Learning

Number of pages in assignment: 15 pages **Word count: 2669**

Declaration

I/we certify that I/we have read and understood the entry in the relevant Student Handbook for the School of Management on Cheating and Plagiarism and that all material in this assignment is my/our own work, except where I/we have indicated with appropriate references. I/we agree that, in line with Regulation 15.3(e), if requested I/we will submit an electronic copy of this work for submission to a Plagiarism Detection Service for quality assurance purposes. I/we also confirm that the percentage allocation of work is as shown above.

Upload your coursework via the Unit's Moodle Page by the specified submission date and time. For **Moodle submissions,** this form should be available on Moodle site for the relevant unit and can be pasted into the front page of your assignment if requested by your tutor. Copies are also available from Management Reception.

You should aim to hand your work in before the deadline given by your lecturer/ tutor. The University guidelines on penalties for late submission are as follows:
Any assessment submitted late without an agreed extension, will receive a maximum mark of 40% or the relevant pass mark. Any assessment submitted more than 5 working days without an agreed extension will receive a mark of zero.

For all **Moodle Submissions**, your marks and feedback should be returned to you via Moodle.

All work is internally moderated. For all work that contributes towards a final degree classification, a sample of work with supporting documentation for each unit is sent to an External Examiner for review and comment.

Please note that any mark given is provisional and is subject to confirmation by the relevant Boards of Examiners for Units, Programmes and Boards of Studies. These normally take place at the end of each Semester.

Contents

Introduction

This study will explore Z social media data to finding the insights, which we will focus on applying the machine learning knowledge to divide the kind of user behaviours to be the clusters and define the meaning of each group of users to bring these data to make the strategies to expand user-based or retain the prior members. First section, we will consider the raw data to understand the information before doing the advanced data-mining approaches. Following that we will analyse the unsupervised model to classify the user roles and analyse these results to find a strategy to develop this business. Next section, we will do the prediction model to anticipate the clusters of each member and apply it to find the roles transition of this platform between prior and new users. Lastly, revealing the constraints of this research and conclude the important points and give some suggestions to improve this study in the future.

Wrapping Up Exploration

Data Exploration

In this part, it will present the ways in which explore the social media data by doing correlation matrix, finding the most appropriate features of this information [4], and then we will apply these features to separately plot to identify the distribution of these data for analysing that are they appropriate to select doing the cluster algorithm as feature selection in the Clustering analysis section.

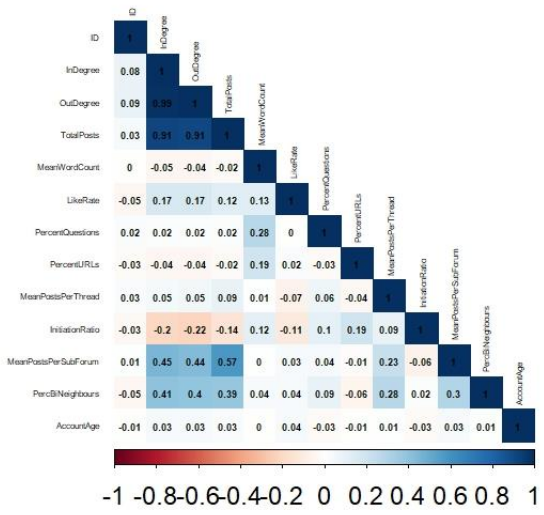

Fig 1: plots the correlation analysis of social media data, showing their relationship.

Fig 1 depicts it has five significant features which has the high correlation rate > 40%. They have eight pairs of features, which we will plot these data to observe the distribution of each pair feature to find the appropriate pair features for doing the clustering as feature selection in the next section continuously.

Fig 2 shows these data have the distribution appropriately even though they have some a bit outlier, but it does not affect much to build the model; then we decided to choose these features except the PercBiNeighbours feature because it has more vary distributions when compared to other pair features.

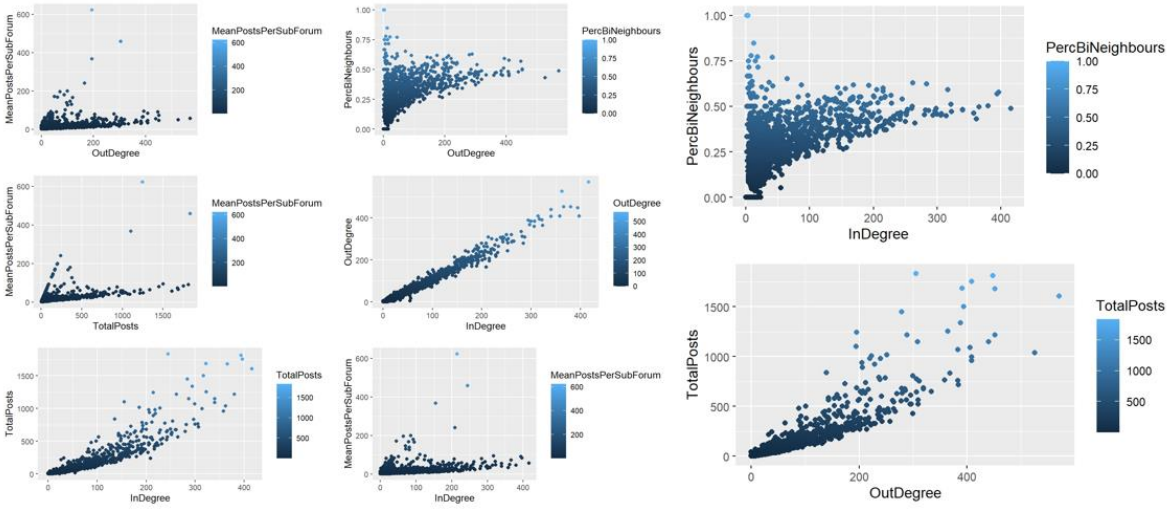

Figure 2: The distribution of significant features pairs.

Data Metric Definition

Following correlation analysis, this subpart explores every feature in more detail. We will provide a clear definition for each metric, outlining its significance in understanding user behaviour on this social media platform; it is split into six categories by Chan, Hayes, and Daly (Davidson et al., 2019). These categories are used to consider separating types of user behaviours on this platform; we apply the Reader-to-Leader Framework (RtLF) [2] approach to categorizing users; we will present this detail in the clustering interpretation.

Category	Feature	Meaning
Structural Features	InDegree	total number of unique network neighbours replying to (or quoting) a user
	OutDegree	total number of unique network neighbours receiving posts from (or being quoted by) a user
Popularity Features	LikeRate	mean average number of likes per post.
Content Features	PercentQuestions	percentage of a user's posts that contain question marks (excluding within URLs)
	PercentURLs	percentage of a user's posts that contain URLs
	TotalPosts	total number of posts of a user
	MeanWordCount	mean average word sound for all user's posts
Initiation Features	InitiationRatio	number of threads initiated / number of threads participated i
Persistence Features	MeanPostsPerThread	total number of posts / number of threads participated in
	MeanPostsPerSubForum	total number of posts / number of sub forums participated in
Cooperation Features	PerBuNeighbours	number of neighbours a user has both received posts from and posted replies to / total number of unique network neighbours

Table 1: The data Metric of the social media platform data shows their meanings.

The identified metrics are important for understanding the user behaviours more and we can split it into groups reasonably. This information will be considered by doing the feature selection approach before running the clustering algorithm. It can aid developing identified segments accurately. In the next section, explaining step-by-step to find the user behaviour groups to explore the insights for development to this business.

Clustering Analysis

This section will show the methods that produce the clusters of user behaviors by considering the relevant features by doing feature selection and find the optimal numbers of cluster before executing the clustering algorithm to separate the user groups, identifying the empirical results, and doing evaluation. Lastly, providing the definition of each cluster and highlighting the potential opportunities to client's firm.

Data Preprocessing

Feature Selection

Regarding the correlation matrix in the prior work, we will select four relevant features for significantly building the clusters. By choosing the right features, we can develop the quality and efficiency of our clustering models. This approach aids us in understanding the relationship between variables and the impacts on their clustering outcomes. In our experiments, we revisit feature selection to ensure we are using the most appropriate features to remain optimized for delivering meaningful and actionable insights.

Finding the Optimal Number of Clusters

Determining the ideal number of clusters is vital for effective clustering analysis. It involves identifying the optimal number of groups that best represent patterns in the data. Through our experiments, we strive to find this optimal number by evaluating the boundary of clustering numbers using the elbow method [3]. This process ensures that we strike the right balance between capturing meaningful patterns and avoiding overfitting cases. By selecting the optimal number of clusters, we can enhance the interpretability and usefulness of our clustering results, enabling better decision-making and insights for our business objectives. As a result, Fig 3 shows the elbow begins to plateau from k=4, which means it indicates this would be a reasonable number of clusters.

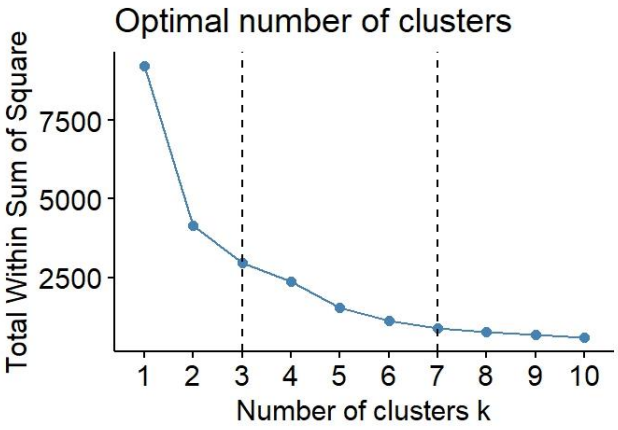

Fig 3: Elbow plot for the community Z.

Clustering Findings and Evaluation

After we receive the optimal number of clusters, we will bring the boundaries of these number values to produce the consequences to prove these results that k=4 can provide more reasonable outcomes than the other results, so it can give the appropriate findings, splitting the user behaviour groups. We decided to select the k-means method because our exploration in the previous study shows that every feature of this data is numeric, and it does not have the missing value. Thus, it is appropriate to choose this approach, which is simple and efficient, to do the data segment by dividing the whole data into small sections called “clusters” [3] to define each centroid and identify these groups of samples.

The flowchart [Fig 4] presents the steps to produce the results precisely. Firstly, we will do the feature selection (Correlation matrix) to find the relevant features and then transforming the data by doing normalization data for reducing outlier and adjust scale. Following that we will run the k-mean and silhouette algorithms by the k boundary numbers (k= 3, 4…,7), observing the distribution, separation of the findings and analyses the metric scores which value is the most appropriate score.

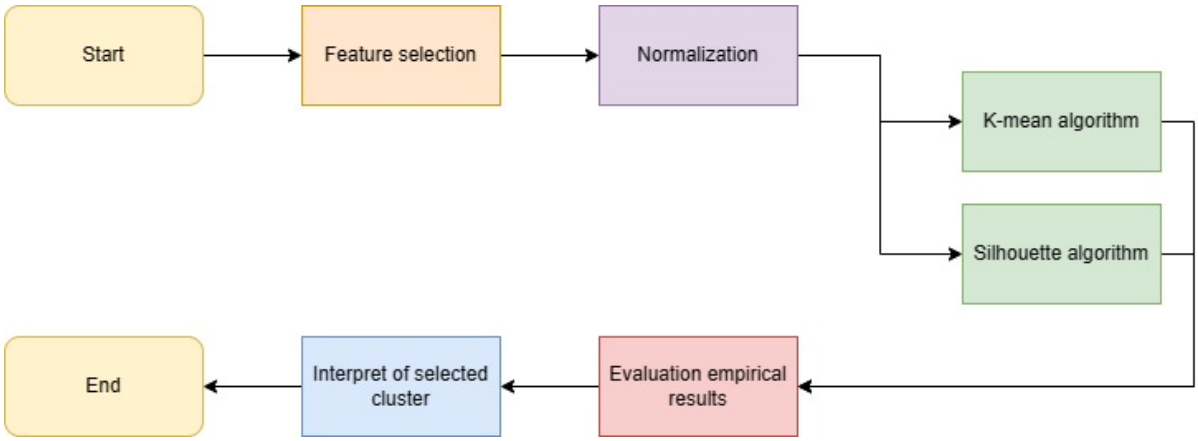

Fig 4: Steps to produce the cluster by using k-mean approach.

After finishing executing the building of the clusters, we received the set of outcomes of boundary ideal numbers of clusters of the prior work [Fig 5], and then we considered the

silhouette score as a metric to evaluate these findings. Overall, it reveals remarkable outcomes between 64-73%; these values present the selected features significant with the community data because it displays the suitable distribution of each group of user behaviour with a satisfactory metric score. Although each result provides good results, we decided to choose the consequence of k=4 because this value matches two criteria: rule of thumb [Silhouette score table] and elbow method, which shows that it should be the most appropriate cluster number of this data.

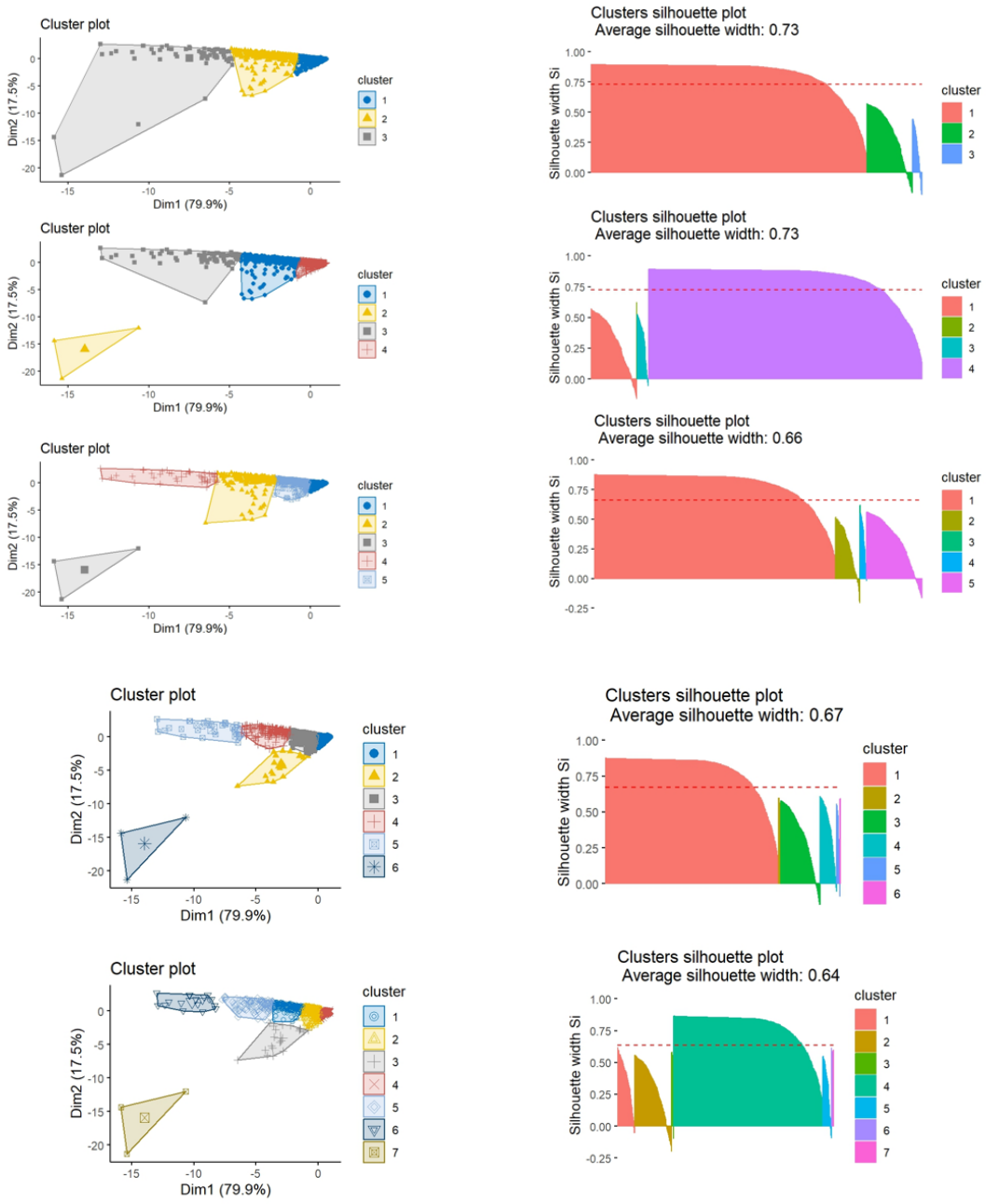

Fig 5: The empirical outcomes of running k-means algorithm and silhouette diagnosis.

Interpretation of Clusters

We can divide the roles of the community into four, which can be mapped based on RtLF theory [2], as shown in Tables 2 and 3. It presents new users as people who have the highest initiation rate, but they have the lowest in- and out-degree features and total posts when compared to others. In the influencer role, they are users who have the highest like rates and a lot of connections. Engager and Blogger, as collaborators in this platform, have average values in all metrics, but Blogger has the highest interaction rate with the community.

Role name	Description
Novice	Lowest in and out degree and total posts, but generally type long word in their posts and have the highest initiation rate.
Engager	Users who have average in all metrics.
Blogger	High interaction rate per both thead and forum, and high upload the posts.
Influencer	Users who have the highest engagement (LikeRate) and have a lot of friends (Highest in and out degree)

Table 2: the role interpretation mapping with Reader-to-Leader Framework.

Category	Input variable	Overall	Engager	Blogger	Influencer	Novice
Structural Features	InDegree	145.40	113.8082	205	248.443038	14.34452
	OutDegree	165.89	125.3553	231.3333	292.632911	14.22234
Content Features	TotalPosts	615.32	240.8365	1395	804.734177	20.7247
	MeanWordCount	107.77	103.5308	103.4525	107.222666	116.8801
	PercentQuestions	0.32	0.304018	0.370914	0.30992109	0.293855
	PercentURLs	0.06	0.055782	0.051	0.06375497	0.070855
Popularity Features	LikeRate	0.69	0.826129	0.432288	0.95846349	0.532287
Initiation Features	InitiationRatio	0.15	0.111574	0.132016	0.08803446	0.255408
Persistence Features	MeanPostsPerThread	3.73	2.576467	7.594816	2.63730947	2.11212
	MeanPostsPerSubForum	139.66	25.67058	483.25	44.553697	5.149901
Cooperation Features	PercBiNeighbours	0.40	0.385706	0.542646	0.48017823	0.204019

Table 3: the average value of each metric, using colours to compare within the feature: red (low), yellow (medium), green(high).

Fig 6 depicts that the newbie is much of the population, at 82.7%. It is the fourth per fifth of all members. Noticeably, it has very few bloggers, which may reflect that this platform does not have enough professionals to aid the new members when they ask questions. It may also have less diverse content types because it has just three people who can do that if we do not ask for help from engagers. However, this information shows that the community can amend any sections to develop their social media in the future.

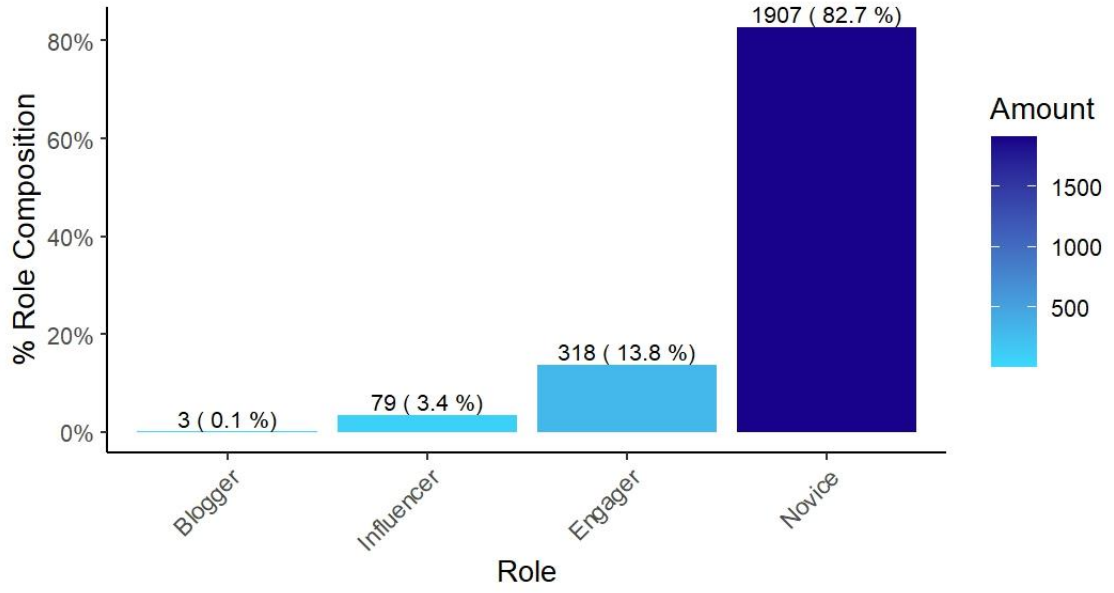

Fig 6: The proportion of roles composition of the Z social media.

Discussion

As mentioned in the previous subsection, this community has many newbie members and needs more collaborator roles to support those users. Fig 7 shows that they have collaborators, around 14% of the members. It is one-fifth of the number of contributors. It may affect the lack of exchange of information in some cases. We have a recommendation: their high-experience users should conduct conferences in their community to provide the essential information to develop their members to step up to be collaborators or leaders in the future. This can aid them in enhancing their platform efficiency and attracting potential new visitors through their exciting actions.

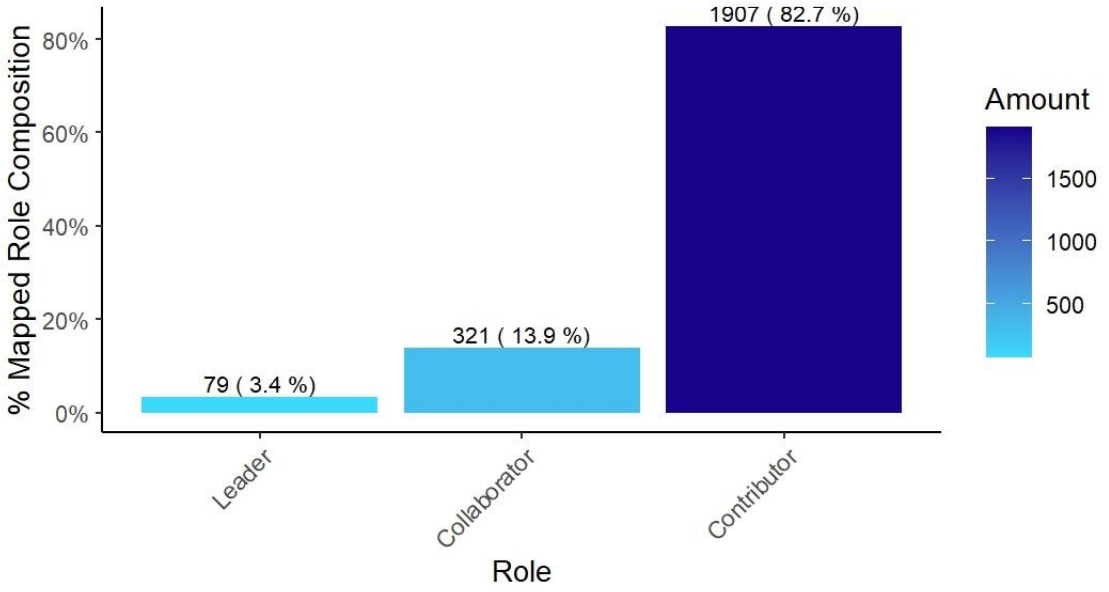

Fig 7: the proportion of users in each category of the reader-to-Leader Framework (RtLF).

Predict Clustering Groups

The aim of this section is to understand the features that have affected the community's position so that we can support clients in developing their community by using a prediction model.

As the initial step, we will outline the process of constructing the decision model. This model is designed to anticipate the data cluster, and it does so by dividing the data into two parts: training and test. The test data, which consists of new users with less than two years of experience, is used to validate the model's accuracy. The remaining data is then used to train the prediction model.

After we get the model, it is used to produce forecasts and analyse the comparison of new members with the prior members' data, identifying position transitions for finding strategies to help the clients.

Pipeline Explanation

First, we will split the data to train and test the dataset; we will apply the data of users with more than two years of experience on this platform, training the prediction model. In this case, we decided to select the k-nearest neighbour because it is KNN and doesn't make any assumptions about the distribution of data [5], making it robust for varied feature types and suitable for clustering-derived targets. It applies to doing experiments of the set of k-values with cross-validation, avoiding the overfitting model by trying various datasets. Finally, we will choose the best performance model to predict the consequences of new member clusters.

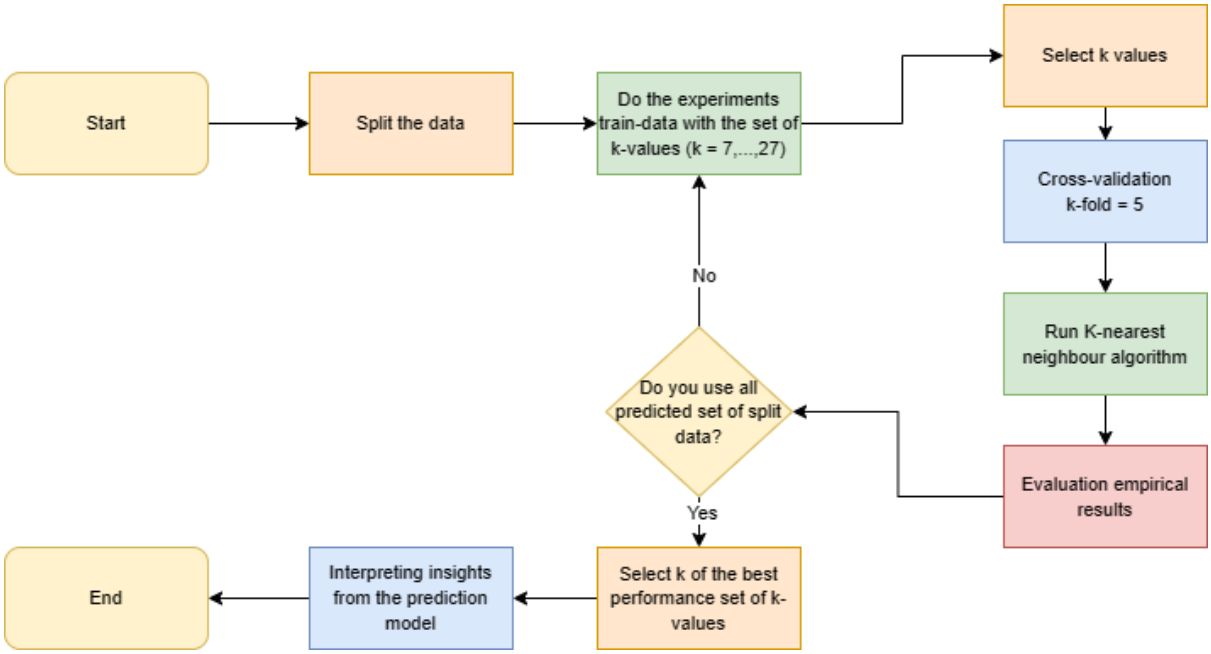

Fig 8: The workflow to do the prediction clustering analysis.

Model Training and Evaluation

Regarding the training model, we decided to select 21 values to experiment with the k-nearest neighbour method to find the most outstanding model and apply the cross-validation concept to examine the different datasets to reduce the bias and make the mode more diverse in the appropriate decision model. We defined the k-fold = 5, which means we split the test data as 20% and 80% data to train the model, doing these five different training and test datasets and calculating the average accuracy of the model before changing the k-value. Table 4 shows that k=12, 13 is the best performance model at around 97.9%. In this case, we decided to select k=12 to build the prediction model and continuously forecast the cluster of the test datasets.

k	7	8	9	10	11	12	13
Accuracy	0.975669	0.975117	0.977327	0.978432	0.978432	0.978984	0.978984
k	14	15	16	17	18	19	20
Accuracy	0.977876	0.977876	0.977324	0.976771	0.976217	0.976217	0.976217
k	21	22	23	24	25	26	27
Accuracy	0.976216	0.976216	0.975111	0.974558	0.974558	0.974558	0.974006

Table 4: The empirical consequence of the k-nearest neighbour (k=7...,27)

The prediction model produces the forecasting outcomes and applies the confusion matrix [Table 5]. It shows remarkable consequences with few error findings. Although this model does not provide the blogger position, it got the results because of the limitations of the data,

which has only one blogger as a testing data set. However, if we want more diverse outcomes, the dataset should gather more records or data to examine reasonably.

In terms of evaluation metrics, the precision and specificity rates are 100%, which means the prediction model can perform the satisfied results because it does not have false positive and negative outcomes. Additionally, providing high sensitivity and accuracy rates reflects the excellent performance of this model, which can produce accurate results, and the F-score shows that the model can give well-classified results of almost 100% (99.76%). Although it can display outstanding outcomes, it should train with a lot of data to learn the unseen pattern to avoid the over-biased model in the future.

Prediction	Actual				Total = 498
	Engager	Blogger	Influencer	Novice	
Engager	59	0	1	0	60
Blogger	0	0	0	0	0
Influencer	0	1	11	0	12
Novice	2	0	0	424	426

Table 5: The confusion matrix of test dataset.

Precision	100%
Sensitivity	99.53%
Specificity	100%
Accuracy	98.33%
F-measure	99.76%

Table 6: the values of each evaluation metrics.

Strategic Insights: Analysis Outcomes and Recommendation

This part will consider comparing prior and new members of the Z community, new members' data from the prediction model, and previous members' data from the clustering model. We will observe the change in both data sets through the bar chart visualization [Fig 9]. As you have seen, the prior users have more segments of collaborator and leader positions; these data came from mapping the role of the four user behaviours based on the Reader-to-Leader Framework approach [2]. The population of this community is in the novice role; even people who have spent more than two years on this platform cannot show a significant increase in the number of other positions.

Fig 9: The proportion of role transition of the Z community

Regarding the role transition, we experimented with the Z community data, which has 2307 members. It revealed the tendency to step up from the contributor to the higher roles. As we see in Table 7, the percentage of contributors decreased when they spent more than two years on this platform, at 3.5%. In contrast, 1.3% and 2.3% become the leader and collaborator, respectively. However, these numbers still need to improve to make this community the top social media platform because the quality of members needs to improve; it still has many newbie users and few professional people to drive this community up.

Our exploratory data suggests a promising avenue for the community's growth. We recommend the manager to invest in training their employees or admins to control content quality. By enhancing user interactions and improving the user experience, we can create a more inviting platform. This can attract more visitors, increase membership, and open new opportunities for the community in the future.

Position	Prior members	% Prior member	New members	% New members
Contributor	259	14.3%	60	12.0%
Leader	67	3.7%	12	2.4%
Contributor	1483	82.0%	426	85.5%

Table 7: The values of the proportion of role transition of the Z social media.

Limitation

For the prior experiment, we found the two main constraints of this project. Firstly, the small size of the datasets can cause us to miss some critical points. If this platform has more users visiting, the prediction model may forecast unreasonable outcomes even if diagnosed with a

very high correctness rate from evaluation metrics; it shows this model may be overfitting, leading to unreliable findings and potentially biased predictions. Secondly, it is computationally expensive, especially when analysing many features, as distance calculations become more computationally intensive by exploring the optimal k values for these models. We can even use the elbow method to reduce the exploratory data to find the appropriate k, but K-nearest is different; we must examine the various k-values to identify the best one that can produce the proper consequences; it will take a lot of time and computational cost.

Conclusion

This study investigates user behaviour within the Z social media community by applying machine learning techniques. Through cluster analysis, we aim to identify distinct user segments and understand their characteristics. This will provide valuable insights for developing strategies to expand the user base and retain existing members.

We received four user characteristics by examining the k-means algorithm to categorize the customer segments. It can apply these numbers to find insights to enhance the quality of the community and use the k-nearest neighbour model to produce the forecasts to compare the prior members and new members to study the changing position of this social media to find a strategy to develop the user skills to improve the community to support the new visitors.

In terms of further development, we propose to collect more data and features to find various potential user characteristics. This can aid the people who recently visited this website in getting a satisfied experience and expanding their user-based in the future.

References

1. Davidson, B.I., Jones, S.L., Joinson, A.N. and Hinds, J. (2019c). The evolution of online ideological communities. *PLOS ONE*, 14(5). Available from: https://doi.org/10.1371/journal.pone.0216932 [Accessed 4 April 2024]
2. Preece J, Schneiderman B. The Reader-to-Leader Framework: Motivating Technology-Mediated Social Participation. Trans Human-Computer Interact. 2009; 1(1):1–25.
3. Bholowalia, P. and Kumar, A. (2014). EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. *International Journal of Computer Applications*, [online] 105(9), pp.975–8887. Available at: https://research.ijcaonline.org/volume105/number9/pxc3899674.pdf.
4. Luna, Z. (2021). *Feature Selection in Machine Learning: Correlation Matrix | Univariate Testing | RFECV*. [online] Geek Culture. Available at: https://medium.com/geekculture/feature-selection-in-machine-learning-correlation-matrix-univariate-testing-rfecv-1186168fac12.
5. Anon, (n.d.). *K-Nearest Neighbors (KNN) – Theory*. [online] Available at: http://www.datasciencelovers.com/machine-learning/k-nearest-neighbors-knn-theory/.

Appendix

Appendix 1: Rule of Thumb for the values (Silhouette score)

Range of SC	Interpretaion
0.71-1.0	A strong structure has been found
0.51-0.70	A reasonable structure has been found
0.26-0.50	The structure is weak and could be artificial
<0.25	No substantial structure has benn found

Appendix 2: The formula of each evaluation metrics

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

$$Precision = \frac{True\ Positive}{True\ positive + False\ Positive}$$

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

$$F - measure = \frac{2\ \times precision\ \times recall}{recall\ + precision} = \frac{2\ \times TP}{2\ \times TP\ + FP + FN}$$