MN50752: Data Mining & Machine Learning

Abhratanu Majumder

Word Count - 2705

CONTENTS:

Introduction

In an era dominated by social media platforms, understanding user behaviour is important for companies striving to provide tailored experiences and enhance user engagement. Social media company Z recognizes the significance of leveraging behavioural data to gain insights into their user base and optimise their platform's performance. In order to identify user categories and provide strategic recommendations, behavioural data is explored, analysed, and interpreted. To find patterns and insights, our method combines data exploration, cluster analysis and supervised learning. The dependability of our results is ensured by exacting data preparation and rigorous statistical techniques. Important behavioural metrics that provide light on user preferences and interactions include post frequency, word count, question ratio, and frequency of sharing URLs. These metrics may be used to improve user engagement and retention techniques.

Exploratory Data Analysis

Understanding the structure and properties of the dataset through exploratory data analysis (EDA) is crucial because it offers insightful information that directs further analysis and modelling endeavours.

Data Overview: The dataset contains 2307 rows and 13 columns, representing various behavioural metrics of users on social media platform Z. A glance into each variable's distribution and core patterns may be obtained using summary statistics. For instance, the mean number of posts per user is approximately 79.7, with a wide range of values observed across different metrics.

Correlation Matrix: The correlation matrix reveals the relationships between different variables. For example, there is a strong positive correlation between InDegree and OutDegree (0.989), indicating that users who receive many posts also tend to make a significant number of posts themselves.

Duplicate Data: In order to maintain data accuracy and integrity for further research, the analysis found and eliminated duplicate rows from the dataset. Fortunately, the dataset had no duplicate rows.

Box-plot: For an individual variable, each box shows the dataset's interquartile range (IQR), and the line inside the box shows the median. Most variables have a very limited range of values with a few outliers, whereas certain variables such as "TotalPosts" and "meanWordCount" have wider ranges and more outliers, implying increased variability.
In comparison to the other variables, the "TotalPosts" variable has a notably wide range of values and a large number of outliers, indicating a more diverse distribution. Data outliers can skew mean values and inflate measures of variability, such as standard deviation, which can have a major impact on statistical analysis. This may result in incorrect interpretations of the dispersion and central tendency of the data.

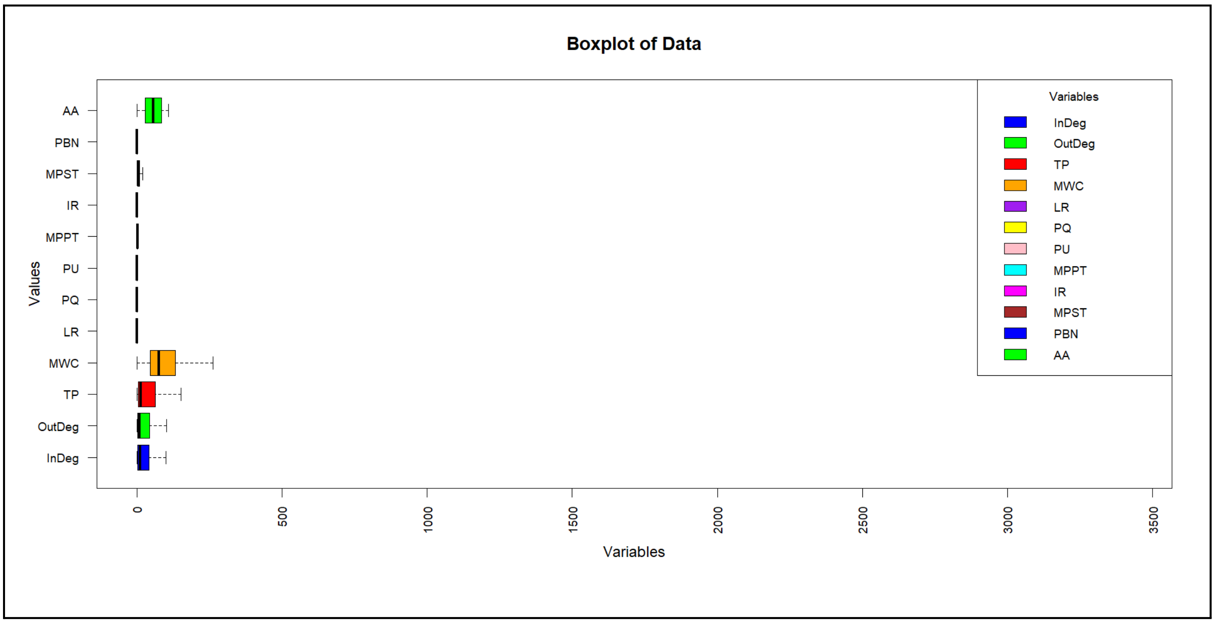


Figure 1: Boxplot of the data

The noticeable diversity observed among users in variables like "TotalPosts" and "meanWordCount" highlights a significant difference in the level of activity and content length within the dataset. The wide range of values in "TotalPosts" indicates that some users are highly active, posting numerous times, while others are less engaged. Similarly, the extensive variation in "meanWordCount" suggests that users' contributions vary greatly in length, with some opting for detailed and elaborate posts, while others prefer shorter, more concise submissions.

Pair-plot:

Upon reviewing the pair plot, it's evident that certain variables, like 'InDegree' and 'OutDegree', display a positive relationship, showing an upward trend in their scatterplot. The histogram shows that the distribution of 'AccountAge' seems to be reasonably constant along its range. Nevertheless, scatterplots for other variables, such "PercentQuestions" and "PercentURLs," show fewer distinct patterns, suggesting weaker relationships with other factors. Relationships can be classified as either exponential or sigmoidal, while some appear to be linear and others lack a clear trend. All things considered, the pair plot analysis offers insightful information about the connections between various variables, pointing out both regions in need of more research as well as significant correlations.

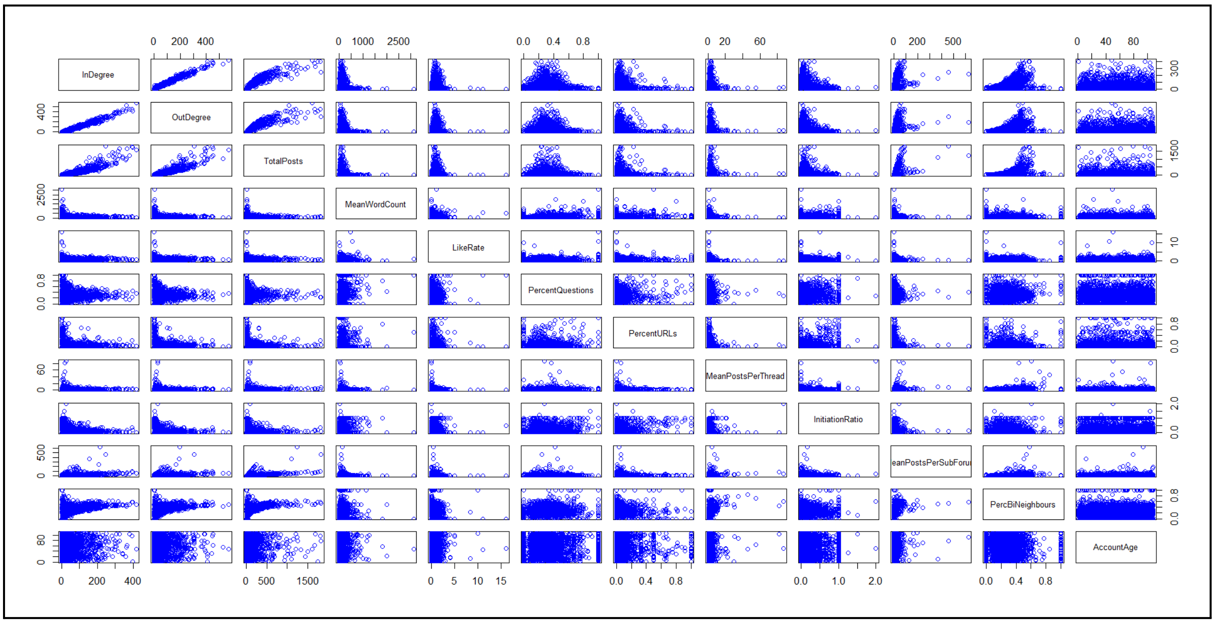


Figure 2: Pairplot of the data

Visualising Total Posts Distribution: This type of graph is used to show the frequency distribution of a quantitative variable, in this case, "Total Posts." The distribution of the data is skewed to the right, suggesting that a few users are quite active, while the majority have a comparatively low total number of postings. Few people publish regularly, and most users rarely participate. There may be outliers or very active users, as indicated by the distribution's extended tail at the upper end.

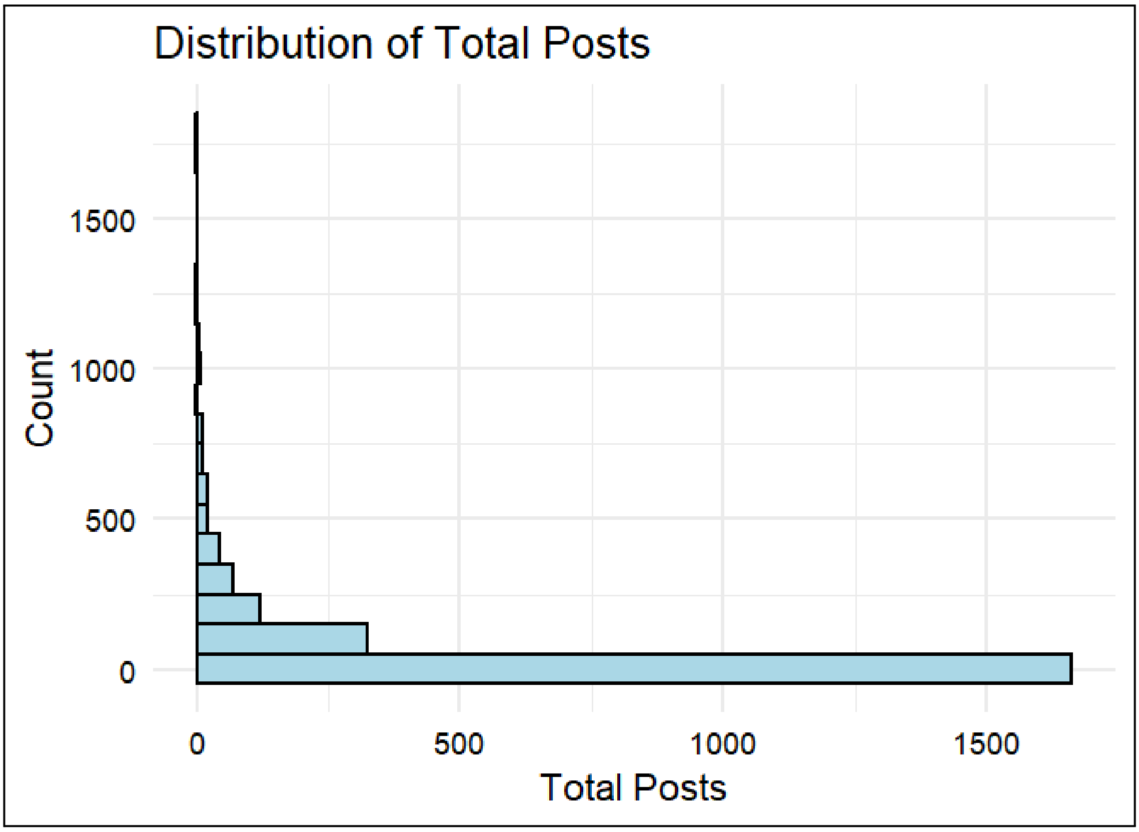


Figure 3: Distribution of Total Posts

Distribution of Account Age: The account creation process appears consistent over time, with no notable spikes or dips observed. Variations in account creation show random fluctuations rather than distinct trends, indicating a steady pace of new user registrations without any unusual activity periods. This stable pattern suggests a consistent user base with minimal attrition and no sudden surges in new sign-ups. The absence of recent spikes suggests no significant events or marketing initiatives impacting user growth. Overall, the distribution of account ages reflects a stable and steady user base on the platform

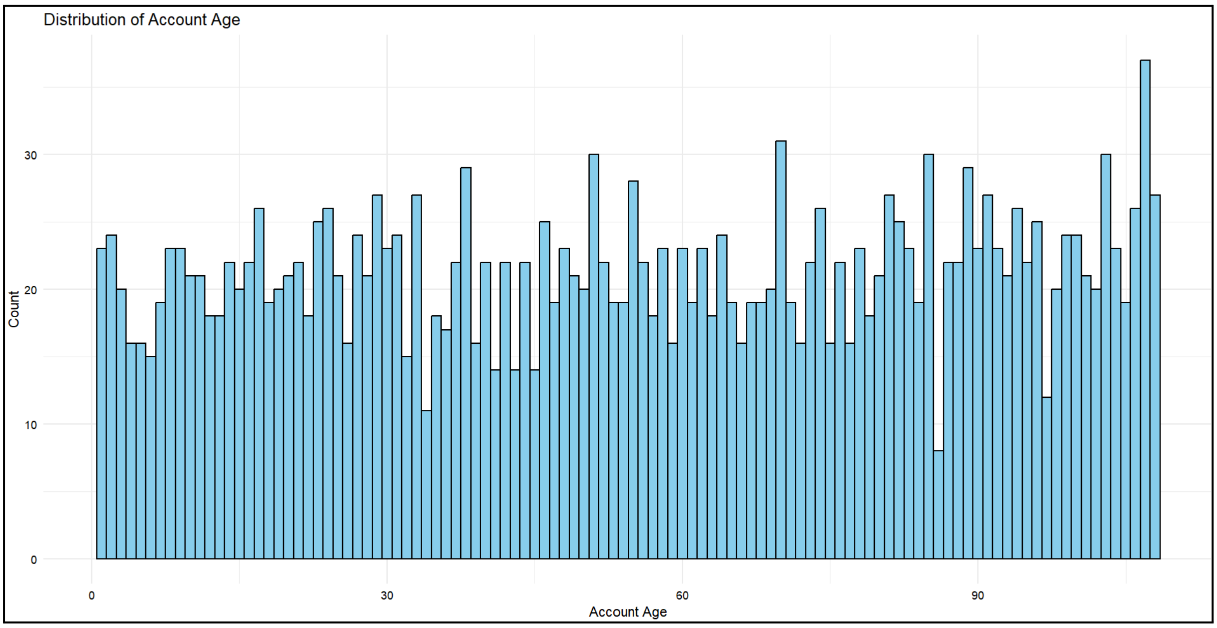


Figure 4: Distribution of Account Age

Scatter plot for InDegree vs OutDegree: A positive correlation between InDegree and OutDegree is shown by the scatter plot analysis, indicating that nodes with a high number of incoming connections also have a high number of outgoing connections. As is common with network data, the majority of nodes have low to moderate connection counts. On the other hand, a small number of outliers with abnormally high connection scores indicate nodes that may be important or crucial to the network. A reinforcing process wherein active nodes attract and form connections is implied by this correlation. The density of points in the vicinity of the origin implies a scale-free characteristic of the network. The possible effects of these outliers on the dynamics and structure of networks call for more research. Understanding this link is essential to recognizing key nodes and understanding network expansion, which may guide information-dissemination and interventions.

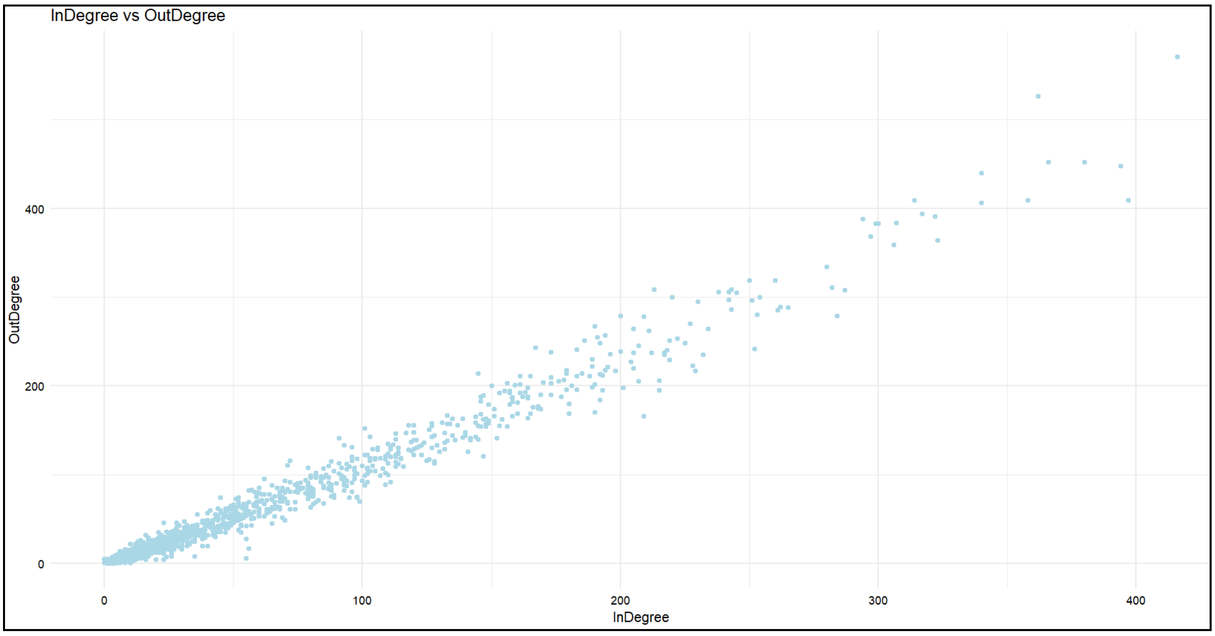


Figure 5: InDegree Vs OutDegree Scatterplot

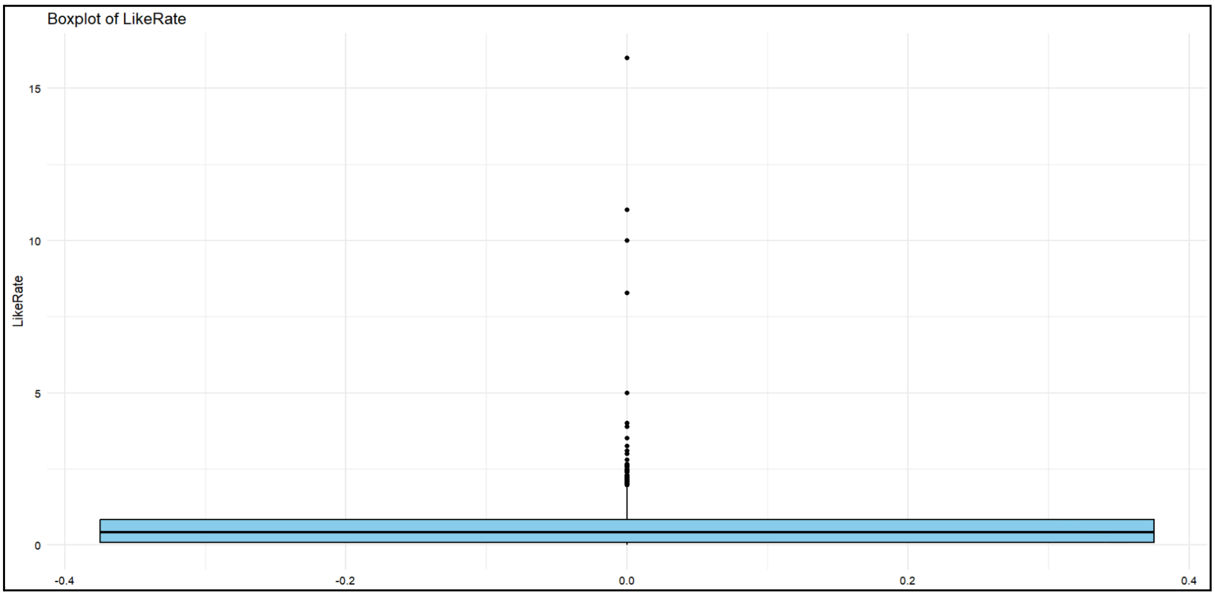
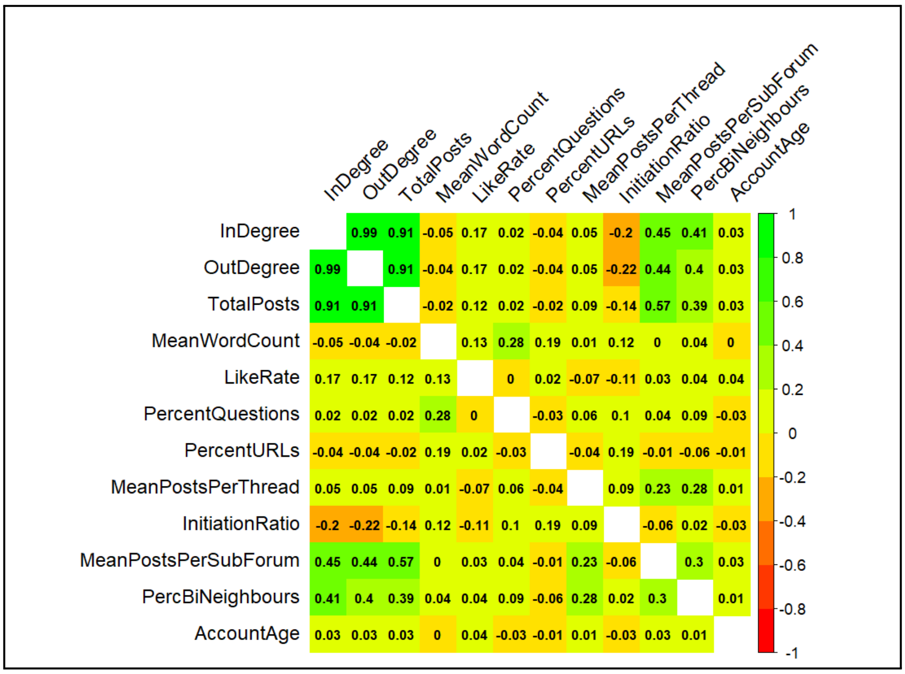


Figure 6: Distribution of LikeRate

The above plot provides a visual summary of the variable's central tendency and dispersion. According to the boxplot analysis, the median LikeRate for the dataset is primarily low or neutral, with the LikeRate distribution centred around zero. The minimal variability among the middle 50% of LikeRate values is indicated by the small interquartile range. Nonetheless, there are prominent outliers on both ends of the distribution, especially those that are skewed higher, indicating rare instances of LikeRates that are noticeably higher or lower than average. There are notable increases in LikeRates, suggesting material that connects more strongly with the users, even if most LikeRates often linger around zero.

Interpreting the Correlation Heatmap: The below heatmap visually displays the correlations between different variables in the dataset. The analysis reveals several noteworthy correlations among variables. InDegree and OutDegree have a significant positive correlation (0.99), suggesting that entities with a high number of incoming connections also typically have a high number of outward connections. InDegree and OutDegree have a significant positive correlation (0.99), suggesting that entities with a high number of incoming connections also typically have a high number of outward connections. Furthermore, there are substantial positive correlations (0.91) between InDegree and OutDegree and TotalPosts, indicating that entities with greater degrees of connectedness tend to be more active in terms of posting. LikeRate and MeanWordCount have a somewhat positive association (0.28), suggesting that lengthier postings often get more likes. A somewhat positive association (0.3) is also seen between MeanPostsPerSubForum and PercentBiNeighbours, suggesting that entities that post in more subforums often have a larger percentage of simultaneous neighbour connections. Nonetheless, AccountAge shows very little association with other variables, indicating that account age has little effect on the other activities that are examined.



	InDegree	OutDegree	TotalPosts	MeanWordCount	LikeRate	PercentQuestions	PercentURLs	MeanPostsPerThread	InitiationRatio	MeanPostsPerSubForum	PercBiNeighbours	AccountAge
InDegree		0.99	0.91	-0.05	0.17	0.02	-0.04	0.05	-0.2	0.45	0.41	0.03
OutDegree	0.99		0.91	-0.04	0.17	0.02	-0.04	0.05	-0.22	0.44	0.4	0.03
TotalPosts	0.91	0.91		-0.02	0.12	0.02	-0.02	0.09	-0.14	0.57	0.39	0.03
MeanWordCount	-0.05	-0.04	-0.02		0.13	0.28	0.19	0.01	0.12	0	0.04	0
LikeRate	0.17	0.17	0.12	0.13		0	0.02	-0.07	-0.11	0.03	0.04	0.04
PercentQuestions	0.02	0.02	0.02	0.28	0		-0.03	0.06	0.1	0.04	0.09	-0.03
PercentURLs	-0.04	-0.04	-0.02	0.19	0.02	-0.03		-0.04	0.19	-0.01	-0.06	-0.01
MeanPostsPerThread	0.05	0.05	0.09	0.01	-0.07	0.06	-0.04		0.09	0.23	0.28	0.01
InitiationRatio	-0.2	-0.22	-0.14	0.12	-0.11	0.1	0.19	0.09		-0.06	0.02	-0.03
MeanPostsPerSubForum	0.45	0.44	0.57	0	0.03	0.04	-0.01	0.23	-0.06		0.3	0.03
PercBiNeighbours	0.41	0.4	0.39	0.04	0.04	0.09	-0.06	0.28	0.02	0.3		0.01
AccountAge	0.03	0.03	0.03	0	0.04	-0.03	-0.01	0.01	-0.03	0.03	0.01	

Figure 7: Correlation Heatmap

Unsupervised Model:

Clustering analysis is a vital technique in machine learning used to discover patterns and structures within unlabeled data. The objective of this modelling is to provide a concise overview of clustering algorithms and their applications. Applications for clustering may be found in many different disciplines, such as document clustering, anomaly detection, image and customer segmentation. It allows deeper insights into complicated datasets and helps in decision-making.

The data has been standardised using z-score standardisation, where each feature's mean is subtracted from the data and then divided by the standard deviation. This transformation makes the features comparable by giving them a mean of 0 and a standard deviation of 1.

Optimal Cluster Selection Using the Elbow Method for K-means Clustering: This graph is used to determine the optimal number of clusters (k) for k-means clustering by plotting the within-cluster sum of squares (WCSS) against the number of clusters. The within-cluster sum of squares (WCSS) is plotted versus the number of clusters, and at k=3, a clear elbow is seen.

The elbow denotes a substantial shift in the curve's slope, meaning that adding more clusters after this point will only slightly enhance cluster compactness. A balance between WCSS and model complexity is achieved by using k=3, which makes it easier to create internally homogenous yet well-separated clusters.

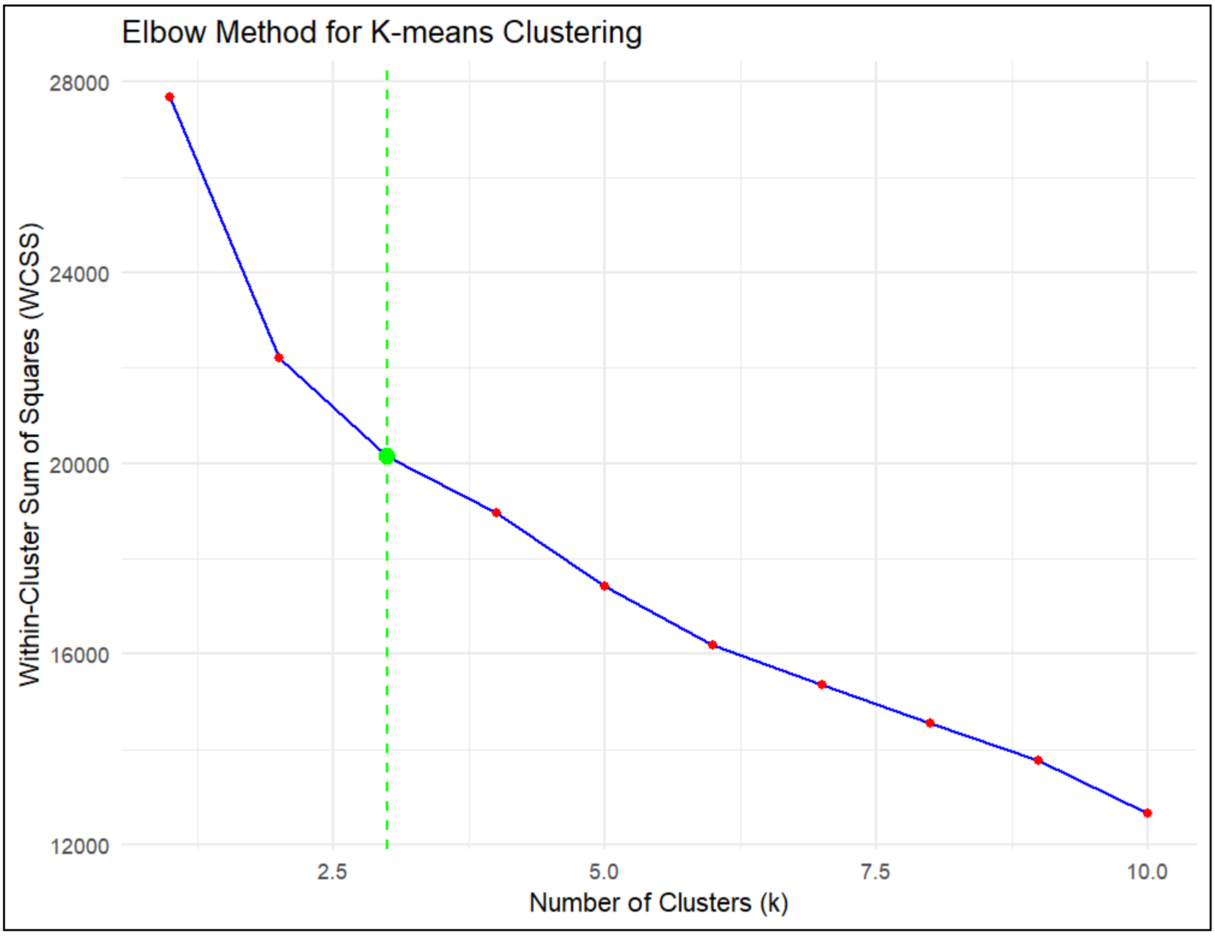


Figure 8: Elbow Method for K-means Clustering

The scaled dataset with three clusters was subjected to k-means clustering after the elbow approach. Three unique clusters with cluster sizes of 475, 1579, and 253 were produced as a consequence of the clustering. Different mean values for different behavioural variables are displayed by each cluster, suggesting unique patterns of user behaviour within each group.

Insights from K-means Clustering: The "K-means Clustering" scatter plot illustrates three distinct user clusters based on their network connectivity behaviour. Cluster 1 (green) consists of less active users, while Cluster 2 (orange) represents moderately engaged users, and Cluster 3 (red) includes highly active and influential users. Because of this segmentation, communication, marketing, and resource allocation methods may be customised to target higher levels of engagement within Cluster 1, promote growth in Cluster 2, and capitalise on the influence of users in Cluster 3. Decisions on how to improve the platform's community management techniques, engagement tactics, and user experience are informed by an understanding of these clusters.

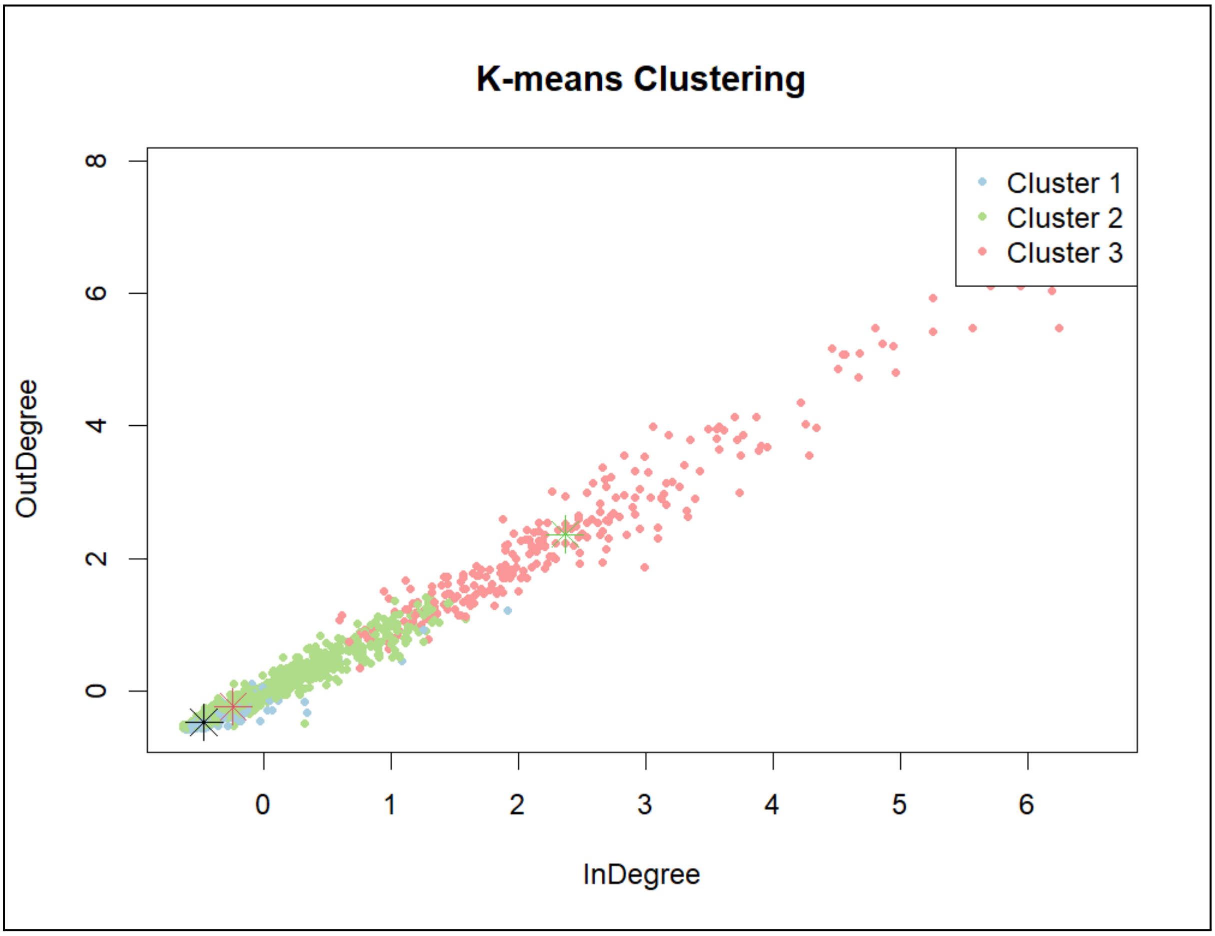


Figure 9: K-means Clustering (InDegree vs OutDegree)

The K-means clustering model's silhouette analysis shows how well the grouping is done. Higher silhouette scores correspond to better-defined clusters; the range of silhouette values is -1 to 1. The average silhouette width in our analysis was 0.2711827, indicating a moderate to high degree of clustering. Every data point's individual silhouette score indicates how well it fits into the cluster to which it has been given; positive scores denote proper grouping, while negative scores suggest possible misclassification. These ratings can help determine the ideal number of clusters and are useful in evaluating the quality of the clustering technique.

Center of variables for three clusters: Different behavioural patterns were identified by the clustering model for each of the three user categories. The users in Cluster 3 exhibit the highest levels of involvement, as seen by their leading mean values in MeanPostsPerSubForum, OutDegree, and TotalPosts. Cluster 2 distinguishes itself with users who contribute longer posts and frequently include URLs, as well as a higher tendency to initiate interactions. In the meantime, users in Cluster 1 have the greatest like ratings on their contributions even if they have the fewest posts overall. Interestingly, there is not a significant disparity in account age between clusters, indicating that user duration is not a determining factor in the grouping of users. These insights help to customise community management and communication tactics to improve user experience and engagement for various user categories.

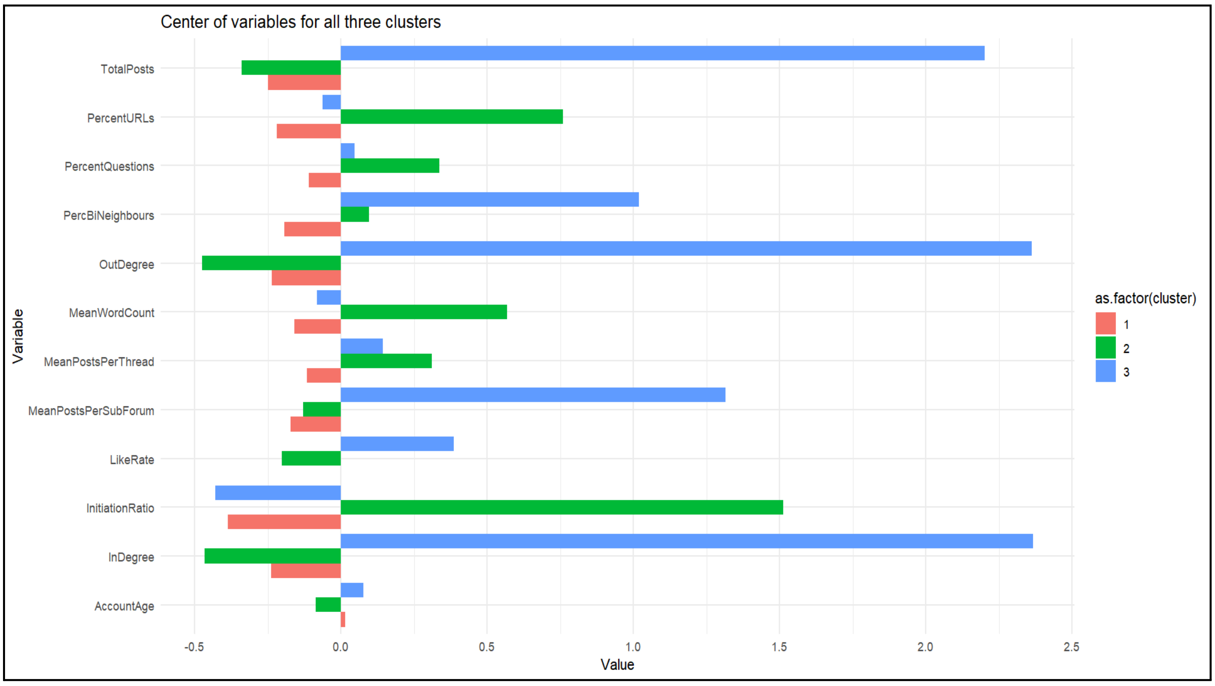


Figure 10: Center of variables for three clusters

Silhouette Plot: To test our clustering model, silhouette analysis has been used, a technique that measures how closely clustered and well-separated our clusters are from one another. An average silhouette width of 0.27 is obtained from the silhouette plot, showing respectable degree of cluster clarity. On closer examination, Cluster 1 revealed negative silhouette values, indicating that not all of the data points in this cluster would fit neatly together. While not quite reaching our ideal criterion, Cluster 2 showed a decent match, with most locations having positive silhouette values. However, Cluster 3 was distinguished by its strong silhouette values, which suggest that its members have great internal similarity and that the clusters are clearly defined.

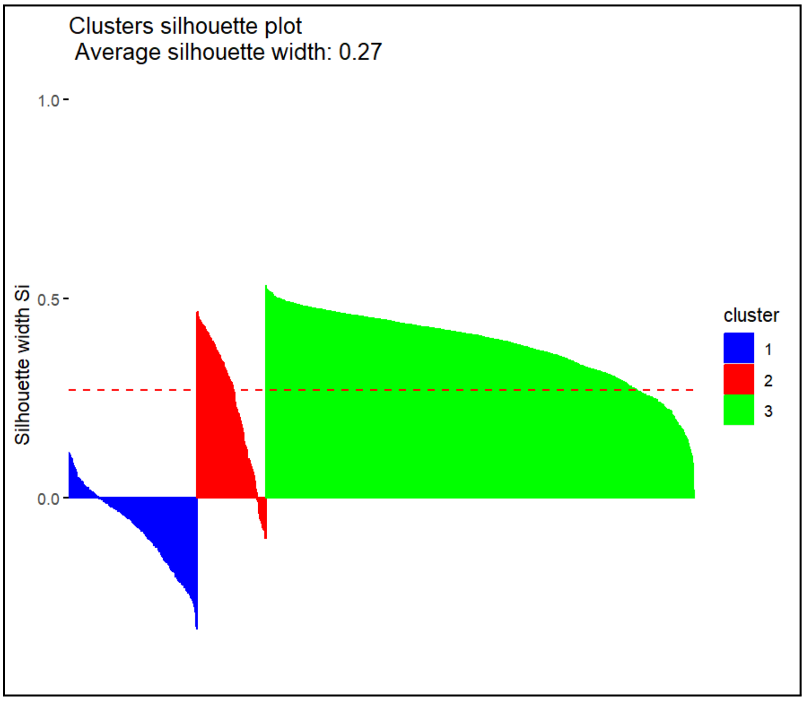


Figure 10: Silhouette Plot

Correlation with Clusters: The correlation analysis uncovered nuanced relationships between cluster assignments and various features. InDegree and InitiationRatio showed slight positive correlations with the clusters, suggesting instances with higher values tend to group together. Conversely, OutDegree and MeanWordCount exhibited negative correlations, with MeanWordCount displaying the strongest negative relationship. This implies instances with higher MeanWordCount are likely to be in different clusters from those with higher InDegree or InitiationRatio values. These results provide insightful information about the distinctive qualities of each cluster, which may help direct focused strategies based on clustering results.

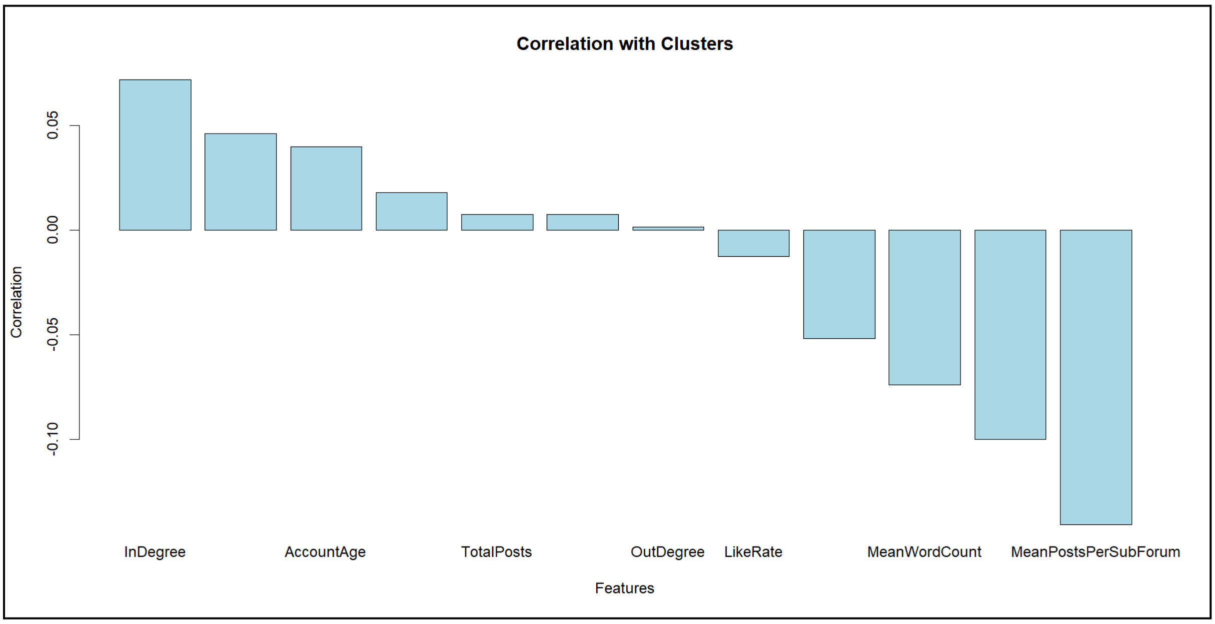


Figure 10: Correlation with Clusters

Supervised Model - SVM

The SVM model's confusion matrix shows excellent accuracy for all predicted classes. With 100% accuracy in class 1 classification, both sensitivity and specificity were achieved. With only one misclassification, Class 2 also demonstrated excellent specificity of 99.86% and flawless sensitivity. In the same way, class 3's sensitivity of 99.60% showed precise predictions. The model performed well in categorising occurrences into their appropriate clusters, as demonstrated by its overall accuracy of 99.96%.

High values for sensitivity, specificity, positive predictive value, and negative predictive value are also shown in the statistics by class for all classes, indicating the model's ability to successfully differentiate between various clusters. The model's overall performance is further supported by the balanced accuracy, which takes into account the sensitivity and specificity for each class. Values close to 1 indicate outstanding classification competencies.

Confusion Matrix:

Prediction	Reference 1	Reference 2	Reference 3
1	475	0	0

2	0	1579	1
3	0	0	252

Although it appears remarkable, the Random Forest model's 100% accuracy on the training set of data raises questions about overfitting. When a model overfits, it fails to understand underlying patterns from the training set and performs poorly when applied to new data. The out-of-bag error rate of 3.64% for the model implies misclassification even in the training set, suggesting possible limits in fully capturing the intricacies of the data.

As a result, the Random Forest model could not generalise well to new data while having a high training accuracy. The SVM model, on the other hand, shows good performance metrics and accuracy in all classes, which makes it a more dependable option for identifying the clusters.

The Support Vector Machine (SVM) model displays exceptional accuracy, achieving 100% precision in classifying users across all predicted clusters. Notably, class 1 achieved perfect precision, while class 2 exhibited only a minor misclassification. Sensitivity in class 3 is also high, at 99.60%. These results underscore SVM's robustness in understanding user behaviours and accurately categorising them. Given SVM's superior performance, it is advisable to periodically retrain the model with updated data to ensure sustained accuracy and relevance in capturing evolving user behaviours. This approach will enable Social Media Company Z to effectively leverage SVM in enhancing user engagement and tailoring strategic initiatives to meet diverse user needs.

Conclusion:

Our analysis uncovered valuable insights into user behaviour on social media platform Z, facilitating informed decision-making and targeted interventions to enhance engagement. We were able to identify three separate user groupings with different levels of impact and activity by combining supervised learning, clustering, and exploratory data analysis. Despite Random Forest's flawless accuracy on the training set, overfitting was a problem. Support Vector Machine (SVM), on the other hand, showed strong performance, which makes it a dependable choice for cluster detection. By utilising this information, Z will be able to fortify its place in the cutthroat social media market and cultivate closer ties with its user base.

Recommendations:

It is clear from examining user activity on social media platform Z that many user groups exist, each with differing degrees of contact and participation. It is advised to customise communication and community management tactics to the preferences of each group in order to maximise user experience and retention. Giving special benefits and recognition to highly involved users might help them feel like they belong and motivate them to keep participating. It would be possible to provide incentives for users who share links frequently and write lengthier articles to start conversations and provide insightful content. Users that obtain a lot of likes but post less frequently should also be recognized for their contributions and given tailored engagement opportunities to encourage them to stay involved.

In order to further boost user happiness and cultivate a feeling of community among users, community-building activities and ongoing platform enhancement based on user input are recommended. Platform Z can successfully foster user engagement and retention across various user categories by putting these methods into practice.