**BUAN 6356: Business Analytics with R**

**Group Project**

**Classification Model for Bifurcating Employees to Suitable Department**

**Group 1:**

**Arsh Syed**

**Rohan Pon Surya Murugan**

**Shobhit Pachauri**

**Contents:**

1. **Background/Context**
2. Domain
3. Brief Description of the Scenario
4. Decisions of Interest
5. Decision Makers
6. **Business Understanding**
7. Business Objectives
8. Situation Assessment
9. Data Mining Goals
10. **Data Understanding**
11. Data Requirements
12. Describe Data
13. Sources
14. Quality

**Project Proposal**

1. **Background/Context**
2. **Domain**

Classification model that companies/employers can use to bifurcate employees into suitable departments based on their Sales profile and personality.

1. **Brief Description of the Scenario**

* In the current economy, when a team/department in any company is unable to achieve the targets and outperform the competition, the employees are assessed based on their scores. And due to this, we can observe mass layoffs.
* This affects the well-being of the employees and affects them financially.
* However, this process does not result in any improvement in the performance of the team and companies/employers must restart the process of hiring new employees. This wastes the time and resources of the company.
* One of the major reasons for this is because the model used for assessing the performance of the employees and making decisions based on this metric is outdated. And the only result of this prescriptive model results in terminating the employment or assigning the training to the employees all over again.
* A reason for this is that majority of the times, employees/any individual is unable to utilize their skills their personality traits in the job role that they are working on, and if the performance in one aspect of performance tracking metrics is low for that individual, it’s difficult to change roles within the company or the department.
* Due to this, employees feel unsatisfied with the work and are not able to grow at it. This affects the performance of the team.
* A classification model like the one we are proposing will help understand the psychological traits of the employees. Until now, only the “Numbers” are used as a performance tracking metrics.
* Introducing a psychological component in this process will be a great way to enable the employees to work to their maximum potential.
* Such a process will help avoid the unnecessary sudden termination of employment.

1. **Decisions of interest:**

Based on the classification model proposal, the decisions of interest can be explained as follows:

* Instead of following the same outdated performance assessing model, and either terminate the employment or reassigning the trainings, we focus on assessing the overall personality and performance metrics of the employees and associating them into a team/department or roles, that are more suitable for their personality type, Feedback, experience, age and NPS rating.
* On the deepest level, the process is introducing the psychological personality types and traits to the decision making associated with performance tracking done by the employer.
* It’s in human nature to feel completely satisfied when the mind is simulated intellectually. Employees often tend to resort to “Silent Quitting” and “Minimum efforts”, just to survive at a job, that doesn’t help them grow. It’s also the case when the employees feel stuck or when the job gets mundane.
* In simple words, classification is just to assign suitable roles to employees. For example: Assigning data analysis related work to the employees who have analyst traits. Or assigning the work that requires a lot of communication and people pleasing skills, to the employee who has more outgoing and extroverted skills based on his personality type.
* It is important to understand, we are not suggesting omitting the basic performance tracking metrics like feedback and NPS score. All these variables are still being used.
* We’re simply modifying the process, adding a psychological touch to the method.

1. **Decision makers:**

This classification technique can be used by several positions that oversee assessing the performance or improving the output of the team. These decisions makers are as follows:

* VPs and Skip managers:

On the highest level, it is observed in any corporate setting, that when it comes to team-based decisions, VPs and Skip managers oversee tracking the performance of a wide range of teams/markets/departments.

* Managers:

Next, we have the managers of each team. They are responsible for carefully assessing the performance of each employee individually. They are the ones responsible for understanding the needs of employees and helping them find a more suitable work/project.

They are also responsible for deciding whether an employee needs to repeat the training or be assigned a new training module.

* Human Resources Department:

The Human Resources team is generally responsible for hiring new employees and attending to any change within the team or department.

One of the key roles of HR department is what we call Organizational Psychology. The classification model that we are proposing can directly be implemented by the team responsible for understanding the psychological needs of the organization and suggesting plans to make the workplace a safer and more efficient place for the employees as well as the organization.

1. **Business Understanding**
2. **Business Objectives**

* The objective is to increase the team’s performance by utilizing the personality traits of the employees, by assigning them into suitable roles. The purpose of designing such a classification model is to avoid a mismatch between employees and their roles. If the employees can completely utilize their skills in their job, we can impact the customer satisfaction ratings and improve the business through Business Development department.
* The next major objective is to save the company’s resources that are used for hiring new employees or training the current employees.

1. **Situation Assessment**

Until before the proposal of this model, we have seen employees being laid off on a large scale. Many of these layoffs could not be explained. However, we feel that there could be a way in which such layoffs could be greatly reduced and the jobs of many could be saved. Our objective is to assign an employee to a suitable role so that his/her’s skillset is utilized to the utmost potential. In our dataset, we have multiple variables that could be used to assess an employee’s caliber and the chances of being promoted or assigned to a different department.

* We have the general classification of domain of work i.e., business. It could take two values, namely hardware and software.
* Secondly, we have the age category. It is a qualitative variable that describes the demographic of the work force. This could be helpful in giving promotions to the managerial roles based on seniority.
* Then we have the female variable. This gives us the employees population by gender i.e., Male or Female.
* The years of experience variable gives us an understanding of how much technical knowledge the employees have gained from working in the company. It could also be a driving factor in giving promotions based on experience and loyalty to the company.
* The personality of the employee is probably one of the most important attributes in our data analysis. It has four categories, namely Analyst, Diplomat, Explorer, and Sentinel. An analyst is known for being rational, impartial and intelligent. One expects an analyst to be objective in decision-making; to solely believe in facts and figures. A diplomat is known for being calm and composed in any situation and is expected to make the right choices. An explorer is known for being spontaneous and flexible. One can expect an explorer to be creative; to defy norms to put forward a solution that could increase profits. A sentinel is known for being practical and straightforward and is responsible in maintaining order in a workplace.
* The number of certificates tells us the diverse skills that employees possess.
* The feedback column is given by employers on evaluating the employees based on performance.
* We have a salary column and a NPS (Net Promoter Score). The NPS score is an indication of the approachability and conversational ability of an employee. A high NPS score indicates that the employee is good at communication and could oversee Sales.

1. **Data Understanding**

**a**. **Data Requirements**

To begin, it’s vital we establish a clear and comprehensive understanding of what information is crucial for our exploration. We’re not just gathering numbers and facts; we are piecing together a coherent and multifaceted view of our sales representatives' professional lives. It’s about having varied and diverse information that enables us to delve deep and draw substantive, informed conclusions about our business questions, that will help the organisation to assess their existing sales representatives.

1. **Describe Data**

The dataset contains **21,993** **rows** and **11 columns**, detailing various attributes of sales representatives, such as their age, business type, gender, years of experience, education, personality type, number of certificates, feedback score, salary, and Net Promoter Score (NPS), with a few missing values in each column. Speaking in more detail, each row in the dataset represents an individual sales representative with the following attributes:

* **Sales\_Rep**: A unique identifier for the sales representative (Numeric).
* **Business**: The type of business they are involved in, e.g., Hardware, Software (Categorical).
* **Age**: The age of the sales representative (Numeric).
* **Female**: A binary variable indicating the gender of the sales representative, where 1 represents female and 0 represents male (Binary).
* **Years**: The number of years of experience the sales representative has (Numeric).
* **College**: Indicates whether the sales representative has attended college, with values 'Yes' or 'No' (Binary).
* **Personality**: The personality type of the sales representative, e.g., Diplomat, Explorer (Categorical).
* **Certificates**: The number of certificates the sales representative has earned (Numeric).
* **Feedback**: A score representing the feedback received by the sales representative (Numeric).
* **Salary**: The salary of the sales representative (Numeric).
* **NPS**: The Net Promoter Score of the sales representative, which is a measure of customer satisfaction and loyalty (Numeric).

**c. Sources**

Our information comes from several places. We don't have the specifics on where each piece of data comes from, but usually, such valuable insights are gathered from our own organizational databases and HR records, or they are put together from different surveys and reports specific to our industry. Each piece of information has been meticulously gathered to ensure a rich and varied understanding of our subject.

Dataset: Tech Sales Reps Data

Source: Book: Business Analytics by Sanjiv Jaggia, 2e

Appendix A. Big Data Sets: Variable Description and Data Dictionary. Data 7.

1. **Quality**

To make sure our data is reliable, we typically check for any missing pieces of information, any entries that are repeated or duplicated, any values that don’t seem right, and whether the information in each part is consistent and accurate. We began by taking a closer look at these areas in the provided dataset and could find a few issues with the quality of the data. We noticed several areas that could impact the reliability of the insights drawn from it. There were missing values and the varying number of unique values across columns were identified. Below, we delve deeper into each aspect to offer a clearer understanding of the dataset's specifics.

1. Missing Values: Each column in the dataset has 3 missing values.

3. Unique Values:

* Sales\_Rep: 21,990 unique values.
* Business: 2 unique values, indicating it is a categorical variable.
* Age: 45 unique values.
* Female: 2 unique values, likely a binary variable.
* Years: 13 unique values.
* College: 2 unique values, likely a binary variable.
* Personality: 4 unique values, indicating it is a categorical variable.
* Certificates: 7 unique values.
* Feedback: 292 unique values.
* Salary: 596 unique values.
* NPS: 10 unique values.

**Conclusion:**

This dataset has some missing values in each column that need attention, but there are no duplicate rows when we don’t count the rows with missing values. The number of unique values in each column provides insights into the nature of the variable, whether it is categorical, binary, or continuous.

1. **Data Preparation**
2. **Data Selection**

This specific dataset is taken from the book “Business Analytics – Communicating with Numbers Edition 2”, from the Appendix section (Data 7: Tech Sales Rep Data). This dataset contains 21,993 rows and 11 columns. As there are many records, we’ve decided to subset 1000 records from the actual dataset so that we get well-defined graphs and plots that are non-overlapping.

To ensure quality of data, we’ve taken a subset of our dataset to contain 1000 rows to ensure well-defined plots. In our dataset, there were **no null values** in it. We’ve taken 1000 random observations off from the dataset. The rest of the observations were **omitted.**

Our final dataset that we worked upon contained 1000 rows and 11 columns. Some of the columns were Business, Age, Years, College, Personality, NPS, etc.

Data types:

|  |  |
| --- | --- |
| **Column Name** | **Data type** |
| Business | Categorical – Nominal |
| Age | Numerical – Ratio |
| Female | Categorical – Binary |
| Years | Numerical – Ratio |
| College | Categorical – Binary |
| Personality | Categorical - Nominal |
| Certificates | Numerical – Ratio |
| Feedback | Numerical – Ratio |
| Salary | Numerical – Ratio |
| NPS | Numerical - Ratio |

1. **Data Cleaning**

Our initial dataset consisted of 21,993 rows and 11 columns. We’ve identified 1000 random observations to ensure that there is no bias and to also have well-defined graphs and plots. Therefore, the remaining 20,993 records were omitted.

There were no missing or duplicate values in the dataset. The dataset’s outliers were omitted.

We had to perform One-Hot encoding on the Personality column. It could take 4 unique values, namely Analyst, Diplomat, Explorer, and Sentinel. Hence, we had to create (4-1) = 3 dummy variables to represent these values. The other variables in consideration were scaled as well.

1. **Preparing Data**

We initially gathered a dataset containing 21,993 rows. We realized that the dataset was too big and had to be reduced to make any significant conclusions on the data. So, we trimmed the dataset to end up with 1000 entries. We performed data cleaning, where we had to identify for outliers, missing values, and other inconsistencies. We concluded that the dataset did not require much cleaning, and it showed data integrity.

However, we were not able to perform analysis with ~22,000 records. Hence, we picked 1,000 records at random (no bias) to implement the model. We used all the 11 columns for the analysis. The categorical columns were all converted to numeric for implementing the model. Our objective is to classify the employees into four clusters or positions based on various factors.

1. **Modeling – Building Models**
2. **Agglomerative Clustering**
3. **Describing data in detail**

We have taken 1,000 rows and 11 columns for the clustering analysis. However, we’ve also realized that not all columns are significant in determining whether a particular employee needs to be assigned a new, high-paying role. Therefore, after understanding the data and the problem, we have chosen Personality, Feedback and NPS to be the predictor variables.

‘Personality’ is a categorical variable that takes one out of the four values: Analyst, Diplomat, Explorer, and Sentinel. These values are derived from <https://www.16personalities.com/> , a personality-identifying website that evaluates a person based on a set of behavioral questions and gives a personality out of the above-mentioned four.

‘Feedback’ is a numerical variable which is continuous in nature. It means that it can take infinitely many values between two numbers (decimal values). It is a review system created by the employers, who evaluate the employees on a scale of 1 to 5, 1 being a poor-performing employee, and 5 being an excellent one.

NPS (Net Promoter Score) is an indication of how well an employee can promote the company’s products to the customers. A high NPS score indicates that the employee can communicate well and can work in Sales.

After studying the personality types, the proper metrics of Feedback and NPS, we have created a table of conditions that we will apply in our cluster. Based on our understanding, for those particular roles, these personality types show more suitable traits.

A screenshot of a computer

Description automatically generated

1. **Decision-making model chosen**

The model we have chosen for analysis and evaluation is Agglomerative Clustering, also known as AGNES. In AGNES, each observation is considered as an individual cluster. Further, these ‘clusters’ are combined based on similarity to form larger clusters, until there’s only one large cluster that exists.

For our project, however, we precisely need **four** clusters. In any form of clustering analysis, there is no prior information on what kind of records each cluster contains. It is a form of **unsupervised learning**.

1. **Rationale for the choice of model**

The idea that we have chosen requires an unsupervised learning model to be implemented. As the job roles are to be assigned to the employees and there is no prior information on what employees were assigned to which roles, we must perform unsupervised learning.

Moreover, in unsupervised learning, we can perform clustering to assign records into various groups. Our first model was K-Means clustering, where *k* centers were initially chosen, records assigned to the closest centers based on Euclidean distance, centers recalculated, and process iterated until we get the least distance error.

However, we have chosen AGNES (Agglomerative Clustering) as the second model for comparison. In this model, each observation is taken as the cluster itself. They are later joined to form larger clusters based on the distance between each cluster. This process continues until we have achieved the desired number of clusters. We can also extend this process to one big cluster containing all the observations. But for this project we only need **four** clusters.

The predictor variables are a mix of numerical and categorical types. Therefore, we must use the AGNES clustering using Gower’s Method. Gower’s Method inherently handles variables of different scales and types without performing normalization. It considers both numerical and categorical variables in such a way that it provides meaningful distances between observations.

1. **Detailed Model development and Output**

The predictor variables are **Personality, Feedback, and NPS**. The Personality variable is categorical. So, it must be converted into a numerical variable with the help of One-Hot encoding, that allows us to convert a categorical variable into a numerical variable with the help of dummy variables. For four categories, we need three dummy variables. The feedback and NPS variables are numerical.

Then, we calculate the distances between observations using the daisy() function. It takes **two** parameters, namely, data frame and metric. The metric that we will be using is “gower”. This metric ensures that Gower’s Method is implemented. After this step, we perform the AGNES clustering using the agnes() function. It takes **three** arguments: ‘d’ (distance vector from the Gower’s Method), ‘diss = True’ indicates the presence of a dissimilarity matrix or distance vector, ‘method’ indicates the type of linkage that is implemented in forming the clusters. There are **four** linkage methods, namely Single, Complete, Average, and Ward’s method. The Single linkage method calculates the distance between the two closest points between different clusters. It is prone to be affected by large values or outliers, which results in long, undefined boundaries for the clusters. The Complete linkage method calculates the distance between two farthest points across different clusters. It results in the formation of small and compact clusters that are not affected by outliers. The Average (Centroid) linkage method calculates the average pairwise distance between the points belonging in either of the clusters. It strikes a fine balance between Single and Complete linkage methods as it produces well-balanced clusters in terms of the definition of boundaries and the number of observations in each of the clusters. Finally, the Ward’s linkage method minimizes the variance within each cluster with the help of **Error Sum of Squares (SSE)**. It is the sum of square differences between each observation and the cluster center. This results in the formation of clusters with similar shapes (boundaries) and sizes (number of observations).

We made the dendogram plots for each of the linkage methods to illustrate how the clusters join to get the desired result. We have queried **over 1,000** records for the analysis. Therefore, the dendogram looked overlapping and we were not able to uniquely identify each observation or individual clusters.

The **silhouette width** is calculated for each linkage method. Silhouette method is an indication of the separating distance across various clusters. It is a measure of how close each point in a cluster is to its neighboring points across different clusters. The range of values is **[-1, 1]. ‘1’** indicates a perfect cluster formation, whereas **‘-1’** indicates the worst formation. **‘0’** indicates overlapping clusters. So, we can say that a clustering method with a higher silhouette width (coefficient) forms better, well-defined clusters.

For Single Linkage Method:

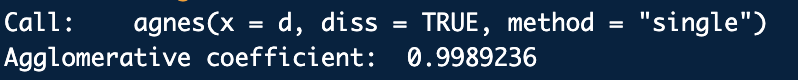


Fig: Agglomerative coefficient of Single Linkage Method

A screenshot of a graph

Description automatically generated

Fig: Dendogram of Single Linkage Method

A blue background with yellow text

Description automatically generated

Fig: Summary of clusters (number of observations in each cluster) using Single Linkage Method

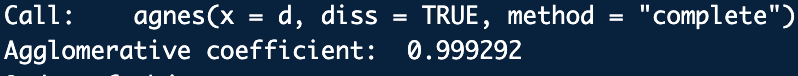
 For Complete Linkage method:

Fig: Agglomerative coefficient of Complete Linkage Method

A screenshot of a computer

Description automatically generated

Fig: Dendogram of Complete Linkage Method

A blue background with yellow text

Description automatically generated

Fig: Summary of clusters (number of observations in each cluster) using Complete Linkage Method

For Average Linkage Method:

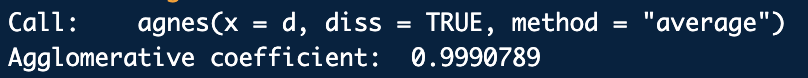


Fig: Agglomerative coefficient of Average Linkage Method

A screenshot of a computer

Description automatically generated

Fig: Dendogram of Average Linkage Method

A blue background with yellow text

Description automatically generated

Fig: Summary of clusters (number of observations in each cluster) using Average Linkage Method

For Ward’s Linkage Method:

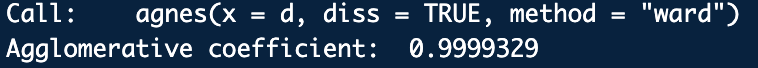


Fig: Agglomerative coefficient of Ward’s Linkage Method

A screenshot of a computer

Description automatically generated

Fig: Dendogram of Ward’s Linkage Method

A blue background with yellow text

Description automatically generated

Fig: Summary of clusters (number of observations in each cluster) using Ward’s Linkage Method

A screenshot of a computer

Description automatically generated

Fig: Result of clusters using the **four** mentioned cluster linkage methods

**K-Means Clustering**

Rationales for Using K-Means Clsutering:

1. Known Value of k:

* K-Means clustering is chosen when the number of clusters (k) is known in advance. Here, we have defined 4 clusters based on the input variables. The clusters are the output variables “Suitable Roles”
* This is advantageous in situations where there is prior knowledge or a specific requirement for a predetermined number of clusters.
* The algorithm partitions the data into k clusters based on similarities, making it suitable for cases where the desired number of groups is pre-defined.

1. Large Dataset:

* K-Means is well-suited for large datasets due to its efficiency in terms of computational resources. The original dataset contains 25000+ records.
* The algorithm's time complexity is relatively low, making it scalable for datasets with a substantial number of observations. After many experiments, we conducted the final test on 1000 random records so we can compare the results and generate a more understandable observation.
* This characteristic makes K-Means an attractive choice when dealing with extensive datasets, providing a balance between accuracy and computational efficiency.

1. Prior Knowledge About Classification Basis:

* K-Means clustering assumes that clusters are spherical and equally sized, with a centroid representing each cluster.
* The algorithm assigns data points to the cluster whose centroid is closest to them in Euclidean space.
* If there is prior knowledge about the characteristics or basis of classification, such as specific features that define clusters, K-Means can be effective.
* For this model, we have researched the personality types and performance measuring parameters (feedback and NPS), to determine appropriate bucket/limits for a particular cluster.
* We applied these conditions using a nested IF-Else statement to implement the clusters.

Detail model development and output

1. Data Import and Preprocessing

Properly importing and preparing the data set for the model is a crucial step. We start with importing the required packages and libraries that are used for plots, creating clusters, reading and writing excel files, etc.

Next, we imported the data set and stored the data under “project\_data” for further processing.

Next crucial step is to clean the data, we did that by omitting the null values using “na.omit(project\_data.xlsx)”.

Now, as we have 25000+ rows in our dataset, its inefficient to create a cluster with that many entries. The clusters that are plotted are difficult to understand. The main purpose of the project is to create a easily understandable report and elaborate the functioning of Cluster analysis models.   
To achieve this, we reduce the number of input rows to 1000 randomly selected rows. The rows are selected randomly to avoid any sequential bias that might occur due to any pre-sorting of the data.

The observations that are made on these 1000 entries can be approximated for the entire data.

Another benefit of reducing the dataset is to optimize the program and reduce the time it takes to compile and generate the outputs.

A screen shot of a computer program

Description automatically generated

1. Data Preprocessing - Dummy Variables:

A black screen with white text

Description automatically generated

Rationale:

1. Handling Categorical Data:

"Personality Type" is categorical, necessitating a transformation for numeric algorithm compatibility.

Dummy variables convert categories into binary indicators (0 or 1).

1. Numeric Representation:

Dummy variables assign numeric values to each personality type, essential for algorithms like K-Means.

Maintains categorical distinctions without assuming ordinal relationships.

1. Compatibility with K-Means:

K-Means relies on numeric input; dummy variables facilitate the inclusion of categorical information.

Enables distance metric computation crucial for clustering.

1. Implementation:

mutate and as.numeric functions in R create binary indicators for each personality type.

Benefits:

1. Compatibility: Numeric representation aligns with K-Means requirements.
2. Information Preservation: Dummy variables retain categorical distinctions.
3. No Assumption of Order: Binary indicators avoid implying order among personality types.
4. K-Means Clustering

A computer screen shot of a black and white screen

Description automatically generated

Rationale:

1. Cluster Creation:

The kmeans function initiates the K-Means clustering algorithm.

The specified value of centers (k=4) determines the number of clusters to form.

1. Random Seed for Reproducibility:

set.seed(123) ensures reproducibility by starting the random number generator from a defined state.

Facilitates consistent results across runs.

1. Cluster Label Assignment:

Cluster labels generated by K-Means are assigned back to the original dataset.

as.factor converts cluster labels to a factor for categorical representation.

1. Implementation:

The algorithm operates on the cluster\_data subset, including relevant features for clustering.

1. Benefits:
2. Output:

kmeans\_model contains cluster information, and data$Cluster represents the assigned clusters for each data point.

1. Mapping Clusters to Suitable Roles:

Based on the table that was previously attached, an algorithm consisting of if-else statement is designed to implement the conditions.

A screenshot of a computer

Description automatically generated

Role Assignment Criteria:

* Criteria based on cluster, personality type, feedback, and NPS scores determine suitable roles.
* A nested ifelse structure ensures role assignment based on specified conditions.

Cluster-Specific Criteria:

* Each cluster is associated with distinct personality type and performance criteria.
* Roles are assigned based on these criteria for effective cluster segmentation.

Handling Unspecified Cases:

* The final NA statement handles cases not covered by the defined criteria.
* Ensures completeness in the role assignment process.

Implementation:

* Utilizes a nested ifelse structure to check multiple conditions for each cluster.
* Conditions include personality type, feedback, and NPS score ranges.

Output:

data$Suitable\_Role\_KMeans column stores the assigned roles for each data point based on cluster and criteria.

A computer screen shot of text

Description automatically generated

Visualizing K-Means Clustering:

fviz\_cluster generates a visual plot illustrating the distribution of data points across clusters.

The use of stand = FALSE maintains the original feature scale, and frame.type = "norm" adds a normalizing frame for enhanced clarity.



Output:

A graph of different colored circles

Description automatically generated

K-Means model summary

**A black background with yellow and blue text

Description automatically generated**

Output:

Here is the summary for the model. Clustering vector shows the number of cluster that each record belongs too.

A screen shot of a computer

Description automatically generated

Cluster Size Visualization:

1. Quantitative Overview:

* cat and table functions provide a numerical summary of the count of items within each cluster.
* The resulting cluster\_table is a tabulation of the cluster-wise item distribution.

1. Visual Representation:

* barplot transforms the cluster\_table into a bar plot, visually displaying the distribution of items among clusters.
* Colors (col) differentiate clusters, and the y-axis limit (ylim) ensures an appropriate scale.

1. Role Labeling:

* Text annotations (text) atop bars incorporate role labels in a specified sequence.
* Enhances the understanding of the roles associated with each cluster.
* This code segment combines numerical and visual techniques to comprehensively convey the distribution of items across clusters, aiding in the interpretation of cluster composition and size.

A computer screen with text on it

Description automatically generated

Output:

A computer screen shot of a number of items

Description automatically generated

A chart with red green and blue squares

Description automatically generated

Summary for Adding Cluster Labels and Writing to Excel:

1. Cluster Label Integration:

The cbind function combines the original data with additional columns for cluster labels (Cluster) and assigned roles.

1. Numeric Conversion:

Ensures consistency by converting the Cluster column to numeric format.

1. Data Export:

The write\_xlsx function exports the enriched dataset with cluster information to a new Excel file named "Project\_With\_Clusters.xlsx."

This process enhances the dataset with cluster details and roles, facilitating further analysis and external sharing of the enriched data.

A screen shot of a computer code

Description automatically generated

Output:

Here is a screenshot of the attached Final Dataset to show the output clusters created by K-Means and multiple types of AGNES clustering.

A screenshot of a computer

Description automatically generated

1. **Model Evaluation**

For Model Evaluation, we are comparing the performance of all the models, i.e.

1. Hierarchical Agglomerative Clustering (AGNES)
2. Single Linkage
3. Complete Linkage
4. Average Linkage
5. Ward’s Linkage
6. K-Means

By comparing 2 major metrics:

1. Silhouette Score
2. Comparing the cluster structures based on number of items in each cluster.

Silhouette Score:

AGNES methods:

A computer code with yellow text

Description automatically generated

Fig: Silhouette width for Single linkage

A computer screen with yellow text

Description automatically generated

Fig: Silhouette width for Complete linkage

A screenshot of a computer code

Description automatically generated

Fig: Silhouette width for Average linkage

A screenshot of a computer

Description automatically generated

Fig: Silhouette width for Ward’s Method Linkage

K-Means method:

A black screen with white text

Description automatically generated

Fig: Silhouette width for K-Means method

Cluster quality is indicated by the silhouette score, which has a range of **-1 to 1**. Values nearer 1 indicate non-overlapping clusters. Surprisingly, in our situation, the **Single Linkage** approach produces the greatest Silhouette score. On closer inspection, however, the predetermined ranges for all three of the input variables naturally produce overlapping clusters, which is an intrinsic feature of the pre-established cluster criteria. As a result, it may be deduced that the clusters are naturally inclined to display overlap. Given this insight, it would be prudent to choose the **K-means model** over the others because it considers the data clusters' intrinsic overlapping nature.

Clusters Structure

AGNES methods:

A blue background with yellow text

Description automatically generated

Fig: Record count for each cluster – Single Linkage method

A blue background with yellow text

Description automatically generated

Fig: Record count for each cluster – Complete Linkage method

A blue background with yellow text

Description automatically generated

Fig: Record count for each cluster – Average Linkage method

A blue background with yellow text

Description automatically generated

Fig: Record count for each cluster – Ward’s Linkage method

K-Means method:

A computer screen shot of a number of items

Description automatically generated  
Fig: Record count for each cluster – K Means method

Observation:

By observing the summary table/cluster structure for all the models, we can observe:

1. For the Single Linkage method, the data distribution is highly skewed. Cluster 1 has most of the data while the other clusters have a very low number of items. This can be a possible consequence because in the Single Linkage method, the clusters are sensitive to chaining effects. And as the conditions of defining the clusters is overlapping in all the cases, this can result in improper distribution of the data.
2. The methods Complete Linkage and Average Linkage, have a substantially better outcome. However, these methods just improve the effect of chaining and outliers. As our data is mostly categorical, outliers have a negligible effect on the clusters. Therefore, we can improve the results by deploying a different method that focusses less on the effects of chaining and outliers.
3. Among the AGNES methods, Ward’s Linkage provides the most suitable results. Ward’s method tends to minimize the variance within clusters and create well defined, compact clusters.

However, even Ward’s Linkage has its limitations.

* Hierarchical clustering methods, including Ward's, can be computationally demanding, particularly for large datasets, due to the need to compute a distance matrix and build a hierarchical structure. And we have a large dataset.
* May face scalability issues for very large datasets due to its quadratic time complexity. Therefore, this method takes a higher amount of time to compute and compile.
* Generates a hierarchical structure represented by a dendrogram. Interpretation may involve choosing a specific level in the hierarchy, which can be more nuanced but may also require more subjective judgment. Dendrograms are not the most appealing plots to represent the clusters when we have high number of items in the dataset. Therefore, it is difficult to interpret the results of the models.

To overcome these shortcomings, we implement K-means clustering method.

1. From the table above, we can observe that the cluster structure for K-Means is much more evenly distributed compared to AGNES methods.
2. Also, we need to optimize the code and reduce the computation time for our model. We performed the experiment on random 1000 data from the dataset, but the compilation load increased excessively for the complete dataset of 25000 values.
3. K-means is often preferred for its computational efficiency, scalability, and simplicity, especially when the number of clusters is known in advance. The ward’s method may be preferred when interpretability, cluster shape, and sensitivity to outliers are more critical, and when a hierarchical structure is desired. Based on our goals, K-Means methods seemed more appropriate.

Discussion (10)

a. Based on the Model, what would your decision/recommendation be? Why?

* Based on the model developed for the purpose of reducing the number of layoffs within the company, the decision and recommendation would be to leverage the insights gained from the K-Means clustering method. The model's utilization involved predicting and identifying more suitable roles for employees based on their Personality Type, Feedback, and NPS (Net Promoter Score) to create meaningful clusters.
* After a comprehensive analysis of the outcomes and the clusters generated by the K-Means model, it became evident that these clusters could serve as a foundation for determining Suitable Roles for each employee. This classification, driven by the employees' inherent characteristics and performance metrics, offers a strategic approach to optimizing the workforce structure.
* The recommendation, therefore, is for the employers and HR department to actively incorporate this model into their decision-making processes. By leveraging the insights provided by the K-Means clustering, companies can strategically transfer employees to roles that align more closely with their inherent strengths, personality traits, and past performance indicators. This strategic realignment is anticipated to have a positive impact on employee engagement, satisfaction, and overall performance.
* The envisioned outcome of implementing this recommendation is a tangible improvement in individual and team performance, thereby fostering overall growth within the department. By aligning employees with roles that better suit their skill sets and characteristics, the company stands to benefit from a more efficient and harmonious workforce.
* One of the key advantages of this approach is its potential to reduce the incidence of layoffs and terminations resulting from skill mismatches. By proactively addressing the issue of job fit through the insights provided by the clustering model, the company can mitigate the risk of involuntary separations. This not only contributes to a more stable and motivated workforce but also aligns with the goal of retaining valuable talent within the organization.

In summary, the decision based on the K-Means clustering model's outcomes is to recommend the strategic reallocation of employees to roles that better align with their individual characteristics. This approach, driven by data-driven insights, has the potential to enhance overall workforce efficiency, reduce layoffs, and foster a positive working environment conducive to sustained growth and success.

**What are the limitations of the Model you have used?**

**1. Data-Based Limitations:**

* **Personality Types:** The model relies on personality types such as Diplomat, Analyst, Explorer, and Sentinel for clustering. However, it's essential to acknowledge that as data analysts, not psychologists, our understanding of these personality traits might be oversimplified. More nuanced and accurate cluster conditions could be established with the input of experts in psychology who possess a deeper understanding of the intricacies of personality.
* **Limited Suitable Roles:** The model currently identifies employees' suitability for only four roles. A sales department, for instance, encompasses a broader range of roles with diverse skill requirements. Expanding the model to incorporate more specific and varied roles would demand additional data and more complex conditions, making it more representative of the actual workforce dynamics.
* Numeric Variable Interpretation: The model conditions are based on our interpretation of numeric variables like Feedback and NPS (Net Promoter Score). It's crucial to recognize that our understanding, as data analysts, might not capture the nuanced corporate dynamics that influence these metrics. Seeking input from individuals with more experience in the corporate setting could refine the conditions and enhance the model's accuracy.

**2. K-Means Limitations:**

* **Assumption of Spherical Clusters:** K-Means assumes that clusters are spherical and equally sized. In real-world scenarios, where clusters may have irregular shapes or varying sizes, this assumption might not hold true. The algorithm may struggle to accurately identify non-spherical or unevenly distributed clusters.
* **Sensitivity to Initial Centroids:** K-Means results can be sensitive to the initial placement of cluster centroids. Different initializations may lead to different final clusters. Employing multiple initializations and choosing the most representative result can mitigate this sensitivity, but it doesn't eliminate the issue.
* **Requirement for Pre-specified Number of Clusters:** K-Means requires the user to specify the number of clusters (K) in advance. Choosing an inappropriate K value can result in suboptimal clustering. While methods like the elbow method or silhouette analysis can aid in determining K, there's often an element of subjectivity involved. We have picked 4 clusters, but as explained above, if a greater number of roles exist, our algorithm will get more complicated and will require more time for computation and compilation.

1. **What cognitive biases would you expect (most likely) to influence the decision‐making process? How does decision support mitigate some/all of these?**

The dataset that we have used for this project involves a **Personality** column, that judges an employee on various factors to assign one out of the four types, namely Analyst, Diplomat, Explorer, and Sentinel. These personality evaluations are done using 16personalities, a website where anyone can get a personality on answering a couple of behavioral questions. Therefore, this evaluation seems objective, and it requires a psychologist’s personal evaluation to achieve better results.

We could have used the help of a **Human Resource (HR)** person for proper assignment of roles to employees. That person would have had better decision-making capabilities and could have understood better on how this process works.

We also could have included more variables in our analysis. Even though the dataset looked intensive, the number of variables were not sufficient to make highly distinctive clusters.

These are some of the biases that we have encountered in this project. We have observed them keenly and did appropriate changes in our methodologies to get valid results for the same.

1. **What enhancements would you aim for to enable better decision support for this task?**

**Psychologists for Better Limits Based on Personality Types:**

* Collaborating with psychologists or experts in personality assessment can significantly enhance the model's accuracy. Psychologists can provide deeper insights into the nuances of personality types, ensuring that cluster conditions align more closely with the intricacies of human behavior. Their expertise can refine the interpretation of personality traits and contribute to more accurate and meaningful clustering.

**HR Involvement for More Understanding on the Roles:**

* Involving HR professionals in the decision-making process ensures a holistic understanding of the roles within the organization. HR experts possess valuable insights into the specific skill sets, competencies, and responsibilities associated with different positions. Their involvement can lead to more accurate role definitions and better alignment between the model's clusters and the actual job requirements, contributing to the overall effectiveness of the decision support system.

**More Performance Measuring Metrics in the Dataset:**

* Including a broader set of performance metrics in the dataset, especially numeric values, enhances the model's ability to generate objective and precise data. Metrics such as sales targets achieved, customer satisfaction scores, or project completion rates provide quantifiable indicators of employee performance. This richer dataset enables more comprehensive and fine-grained analyses, allowing decision-makers to make more precise decisions based on a broader range of performance dimensions.

**Running the Model on a Larger Dataset for Better Train and Test Split:**

* Expanding the dataset to a larger scale provides the model with a more diverse set of examples, improving its generalization to new, unseen data. A larger dataset allows for a more robust training of the model, reducing the risk of overfitting to the specific characteristics of the initial dataset. A better train and test split ensures that the model's performance is evaluated on a representative sample, enhancing its reliability and applicability to real-world scenarios.