(11:15)

Next, Retsef discusses some of the pros and cons of the two algorithms, along with methods you can use to determine when these algorithms will work well and when issues may arise so that you can prepare.

Video 4.11: Module Summary (3:07)

In this video, Retsef walks you through the key takeaways and lessons from this module

Discussion 4.2: Identifying Considerations for Applying Clustering to Datasets [20 Minutes]

Learning Outcome Addressed:

- Identify important considerations for applying clustering to datasets.

This is a required discussion and will count toward course completion.

Clustering is a method for grouping data points based on their similarities. In business, you can use clustering to manage, explore, and derive insights from data. How you use clustering depends on your business needs. For example, if you are working with another machine learning (ML) algorithm that does not work well with large datasets, you may use clustering to separate the large dataset into smaller sizes that are more manageable and amenable to the other ML methods. You may also use clustering to explore patterns and features of your data. For example, if your organization entered into a new market and does not know much about your customers, you may use clustering to identify market segments. Clustering can also be used to identify outliers, such as detecting abnormal transactions.

Like all ML algorithms, clustering algorithms have their advantages and limitations. Some clustering algorithms can generate accurate predictions while continuing to evolve as more data is being fed through the algorithms. The trade-off with accuracy is speed and cost—since clustering has to account for all data points, the execution can get slow, expensive, and even computationally infeasible when working with big data.

Take the role of an organization's executive attempting to use a clustering model to extract new information from your organization, or an organization of your choice's data. This model could seek to detect abnormal transactions, gain insight on customer demographics, or any other use case you have for clustering algorithms. First, you must come up with a business use case so that you can determine what data is needed. Complete your own research and develop an outline of what data your model will be analyzing. Then, answer the following questions:

- What is your input data, and how did you decide to use the data?
- What are your expected clusters or output?
- How many clusters or groups should your model create, and why? How would this help the organization as a whole? What can you infer from the results?

Submit your response as a discussion, written out in paragraph form. Finally, take a few minutes to reply to a peer and either explain why you agree with their ideas or whether you have found some areas for improvement.

Be sure to read the statements posted by your peers. Engage with them by responding with thoughtful comments and questions to deepen the discussion.

Suggested Time: 20 minutes

Rubric: Discussion 4.2

Criteria	Exceeds expectations	Meets expectations	Below expectations
Thoughtful and complete	**4 pts**	**3 pts**	**0 pts**

response to the question(s)	Fully responds to the question(s), post is supported by connections to the reading and real-life examples, and post makes additional connections to the field of data engineering with novel ideas, critical thinking, or extensive application of how to use the topic in future work.	Fully responds to the question(s), and post is supported by connections to the content or real-life examples.	Partially responds to the question(s), or connections to the content are missing or vague.
Engagement with the learning community	**2 pts** Posts thoughtful questions or novel ideas to multiple peers that generate new ideas and group discussion.	**1.5 pts** Asks questions or posts thoughtful responses to generate a single peer's response.	**0 pts** No responses to peers or posts minimal or vague responses to peers that do not motivate a response (e.g., “I agree.”).

Search entries or author   Unread   ⇧ ⇩   ✓ Subscribed

↰ Reply

○

https:/

Manjari Vellanki (https://classroom.emeritus.org/courses/9054/users/231480)

Apr 18, 2024

For my current job role, I never got chance to work on performing analysis by clustering the data other than categorical variables that are derives as part of analysis. After learning through this module, I created a use case where we can implement the Clustering algorithms which aids in further understanding of identical behavior of subgroups.

Use case: Identify the similarities and dis-similarities in subgroups of subjects/patients that are participating in Oncology study vs results/scores (Eg: Tumor measurements).

Data Collected: Patient demographics like Patient_id, age, age_at_diagnosis, weight, BMI , gender, Date of enrollment, Study participation duration, Treatment related information (like Treatment codes, Start dates and end dates), sample_id, test result values/scores that are collected as part of analysis, cancer severity grades.

Step1: Import required libraries and read in the dataset by importing the file.

Step2: Generate Distribution plots for all the columns to see the distribution across the columns.

Step3: Create a new range variable for BMI and Study participation duration (in days).

Step4: Using the elbow method to find out the optimal number of clusters.

Step5: Plot the elbow graph and identified the ideal number of clusters (in this case, #5 clusters).

Step6: Kmeans algorithm fits to the dataset

Step7: Use fit_predict method that returns for each observation which cluster it belong to.

Step8: Plot the scatter plot for cluster representation.

In this case, I have given a try using different variable combinations to see the impact age vs score, BMI vs score, days of enrollment vs severity grades…

Edited by **Manjari Vellanki (https://classroom.emeritus.org/courses/9054/users/231480)** on Apr 25 at 3:13pm

Reply (1 like)

○

(http

Javier Di (https://classroom.emeritus.org/courses/9054/users/226884)

Apr 21, 2024

Great example Manjari. Interested in how you would determine the optimal number of clusers with the elbow method?

And which variables you think are likely to be the biggest drivers here? If age, BMI or other? Thank you

Reply

○

(http

Manjari Vellanki (https://classroom.emeritus.org/courses/9054/users/231480)

Apr 22, 2024

Hi Javier-

Thanks for your response. Though I'm aware of number of possible clusters, for practice purpose I have used "Elbow method" to derive by choosing values ranges from 1:10. Age, BMI are definitely biggest drivers, but surprisingly study participation duration also has some impact on forming the clusters.

Reply (1 like)

○

(http [Ricardo Anaya (https://classroom.emeritus.org/courses/9054/users/228915)](https://classroom.emeritus.org/courses/9054/users/228915)

Apr 24, 2024

I had the same question, bu I saw you already answered it.

I thinkg it will be a interesting relation with cancer severity grades and Treatment codes on which the values is when the Doctors need to provide sense to the data.

Reply

○

(http [Manjari Vellanki (https://classroom.emeritus.org/courses/9054/users/231480)](https://classroom.emeritus.org/courses/9054/users/231480)

Apr 25, 2024

Thanks Ricardo :)

Reply

○

(http [Mariana Flores (https://classroom.emeritus.org/courses/9054/users/237198)](https://classroom.emeritus.org/courses/9054/users/237198)

Apr 24, 2024

Hi Manjari, so nice to connect on the discussion board again. Great post, cluster analysis across the medical field can have such a significant positive impact on so many lives. I agree with your with methodology and data collection.

The real-world application of cluster analysis is remarkable - thank you for sharing.

Reply

○

(http [Manjari Vellanki (https://classroom.emeritus.org/courses/9054/users/231480)](https://classroom.emeritus.org/courses/9054/users/231480)

Apr 25, 2024

Thanks Mariana :)

Reply

○

(http

Haitham Farag (https://classroom.emeritus.org/courses/9054/users/233864)

Apr 24, 2024

Thanks for sharing the steps, I found your approach ideal to be emulated as a checklist (framework) for analysis, especially step 7, which I overlooked in my response.

Could you please elaborate on the distinction between steps 4 and 6.

I would also appreciate some detail on step 5, KMeans *fitting* the dataset.

Reply

○

(http

Manjari Vellanki (https://classroom.emeritus.org/courses/9054/users/231480)

Apr 25, 2024

Hi Haitham-

Thank you so much. There is a typo btw step5 and step6(corrected).

Step4: Using the elbow method to identify the number of clusters by generating a graph.

Step5: Involves in examine the generated graph to figure out the ideal number of clusters.

Step6: updating K-means algorithm with the "No of clusters" and "Random state" options to actually fits into dataset.

Reply (1 like)

○

(http

Mhelissa Yayalar (https://classroom.emeritus.org/courses/9054/users/233590)

May 1, 2024

Nice work, Manjari! May I suggest creating derived variables to simplify complex features, such as BMI alone might not provide clear insights. Categorizing patients into BMI ranges (e.g., underweight, normal weight, overweight) simplifies the data to make informed decisions based on the data. In this case, we might create a new variable called "BMI Range" based on the BMI values.

Example Categories:

- Underweight: BMI < 18.5
- Normal Weight: 18.5 ≤ BMI < 24.9
- Overweight: 25 ≤ BMI < 29.9
- Obese: BMI ≥ 30

Reply

○

https:/

Roy Nunez (https://classroom.emeritus.org/courses/9054/users/229552)

Apr 20, 2024

Hi Everyone,

As an executive in this bank, I propose we begin using clustering models to enhance customer relationship management by segmenting customers based on their transaction behaviors and financial profiles. This segmentation will inform targeted marketing, product offerings, risk assessment, credit jumps next course of action and customer service improvements.

I have the below suggestions for some of the features we can focus on and am open to further discussions:

- Transaction Frequency: per day/month/quarter
- Transaction Amounts: how much customers spend per transaction with and without fees
- Net worth: checking and savings balances, holdings i.e. equities, investment accounts, real estate, - offer snap shots of financial health
- Categories of Transactions: Understanding where customers are spending their money (e.g., groceries, travel, dining) can help provide insight into lifestyle choices and preferences and helps tailor product offers and rewards in personalized marketing.
- Engagement Metrics: interaction data with bank channels i.e. online banking logins, app usage, pages entered

We expect to kick off with the following clusters:

- High-Value Customers: High net worth, frequent large transactions.
- Tech-Savvy, Young Customers: High engagement with digital tools, lower balances but frequent transactions.
- Traditional, Conservative Customers: Prefer in-branch banking, higher average age, stable but larger account balances.
- Bargain Seekers and Reward Maximizers: Engage primarily during promotional periods or for specific rewards.

Note that I believe there may be some challenges in segmentation may appear when comparing High-Value and Traditional Customers based on transaction frequency versus net worth. We should expect that this will be an iterative process and will go through much refinement throughout this initiative's early stages and quarterly when we reach maturity as customer behavior change over time and our models should evolve with the dynamics . I think we should start with 4-6 clusters, providing a manageable yet distinct grouping that aligns with known customer behavior patterns and business needs. These are the customer segments our marketing strategies are poised to personalize to. The elbow method and silhouette scores can help to determine if we selected a statistically appropriate number of clusters and we can adjust our cluster size accordingly and discuss the overall business marketing stagey to ensure alignment.

Given these expectations we believe this insights will help our organization with the following:

- Enhanced personalization, with tailored marketing communications and product offers that can increase conversion rates and customer satisfaction.
- Better understanding of customer segments can lead to improved risk assessment, mitigation strategies and credit jumps
- It will help with improved efficient allocation of resources across marketing, sales, and customer service based on segmented needs and behaviors.

Using the results from our analysis we foresee a few things we can infer such as:

- Personalization potential increase in customer loyalty and retention as products and services better meet the specific needs of each segment.
- Risk Management By understanding the transaction behaviors and financial profiles of each cluster, we can improve our risk assessments associated with different segments. This can improve our tailored risk management strategies, such as adjusting credit limits or loan offers based on the cluster's typical behavior.

- Resource Allocation these insights prefer digital interaction may not need as much in-branch support, allowing the bank to allocate resources towards enhancing digital platforms and uncover pain points unique between web and mobile.
- Predictive Insights: Beyond immediate marketing and service improvements, these clusters can help predict future trends in customer behavior. For example, identifying a growing trend in the Tech-Savvy, Young Customers cluster could inform the bank’s strategy in digital tool development.

The insights into gather from activity levels and financial habits and derived from this data can directly inform product development, marketing, and customer service strategies. I suggest we host a weekly reoccurring meeting on this initiative and keep track of progress.

Reply

○

(http

Javier Di (https://classroom.emeritus.org/courses/9054/users/226884)

Apr 21, 2024

Great business example Roy. So in terms of the clusters to make sure I understand, you would plot different variables (transaction frequency amounts, etc) Vs Net Worth and see the clusters to then determine business actions to make marketing and spending more efficient?

Thank you, Javier

Reply

○

(http

Roy Nunez (https://classroom.emeritus.org/courses/9054/users/229552)

Apr 21, 2024

Hi Javier,

Thank you for your question. Indeed, plotting a couple of variables like transaction frequency and net worth against each other could provide some initial insights, and a pairwise correlation matrix can be applied to analyze how these variables relate across our dataset

However, in this example I was considering a multivariate clustering approach with at least the five features listed net worth, engagement, Transaction Frequency, amounts, categories. But considering your question, the correlation matrix will help in determine the correlated features that can be teased out, since these are not ideal for clustering methods and skewed results as Retsef mentioned.

Thanks so much for question! It’s an important aspect of preparing our data for effective clustering.

Reply 👍

○

(http

Turki Alghusoon (https://classroom.emeritus.org/courses/9054/users/229165)

Apr 21, 2024

Hi Roy,

Great example. I like how you incorporate the elbow method and silhouette scores help you determine the number of clusters. I am wondering how the clusters could help you identify the bargain seekers: what combination of the datapoints you listed is usually the tell-tale sign of that cluster?

Best,

Turki

Reply 👍

○

(http

Roy Nunez (https://classroom.emeritus.org/courses/9054/users/229552)

Apr 21, 2024

Hi Turki,

This is a great question! Let me elaborate on this. When considering bargain seekers we will be looking at the engagement for promotional products, like when a customer clicks on a nudge our models saw as a good fit to display at a particular screen. Interactions with our promotional nudges I think will be a strong indicator of a bargain seeker.

Also transaction amounts when there are reward points or discounts offered. I think the transactions categories will also help here, especially if there is increased spending in promotional periods. The transactions frequencies of course play a role in most in not all these analyses.

Reply 👍

○

(http

Ricardo Anaya (https://classroom.emeritus.org/courses/9054/users/228915)

Apr 24, 2024

I assume the net worth calculated in anual Income + other assets, etc. this

this info should had from all banks to have a full assesment, in case of the transactions are no only on this specific bank, there should be something to link other banks to have a full assement on transaction and net worth right?

Reply

○

(http

Roy Nunez (https://classroom.emeritus.org/courses/9054/users/229552)

Apr 25, 2024

HI Ricardo,

Net worth will include assets, not annual income. We have information from aggregated accounts from other banks and their transactions. So whatever assets are aggregated become part of the portfolio as well as debts. This could included loans, equities, etc.

Reply

○

(http

Haitham Farag (https://classroom.emeritus.org/courses/9054/users/233864)

Apr 24, 2024

I echo colleagues' comments, a very good example indeed. And thank you for pointing out *silhouette scores*.

In practice, would segmentation (clustering) be done first to inform of opportunities/risks or would we expect clustering to be designed to address the business objective? or maybe it's a hybrid?

would we expect different clustering approaches (of the same client dataset) contingent on the business objective?

Your point on customer behaviour patterns is very interesting and raises a question of what would be a practical approach to identify these patterns (e.g. movement of certain observations between clusters over time).

Thanks for sharing Roy

Reply

○

(http

Roy Nunez (https://classroom.emeritus.org/courses/9054/users/229552)

Apr 25, 2024

Hi Haitham,

Thank you for your questions!

Yes the clustering can serve both purposes informing about opportunities that were not initially observed and guide and pursue business objectives.

Yes, we expect an iterative process and expect to use different clustering approaches that will change dynamically with changing trends, customer behavior/preferences and feed into other models. I am thinking we will refine cluster definitions over time and as new data comes.

Interesting question to raise!! In terms of practical approaches, there are more advanced data science techniques that we can use to observe the temporal patterns, and movement of data points/customers from cluster to cluster to help understanding the shifts. I will surely be considering this. Thanks for bringing this point up!

Reply (1 like)

○

(http

Priscilla Annor-Gyamfi (https://classroom.emeritus.org/courses/9054/users/226376)

Apr 25, 2024

Great post Roy. I think in addition to the benefits of this clustering, you could also add the essence of being able to help you target marketing strategies peculiar to each segment which is more effective.

Reply

○

(http

Roy Nunez (https://classroom.emeritus.org/courses/9054/users/229552)

Apr 25, 2024

Hi Priscilla,

Thank you. Yes, agreed. It should increase the effectiveness of our marketing efforts and also optimize resource allocation.

Reply

○

(https:/

Javier Di (https://classroom.emeritus.org/courses/9054/users/226884)

Apr 21, 2024

- The business case applies to an investment company and it's research idea process on what companies to focus on to drive better returns for the investment company.
  Should we focus on and put our research efforts on:

  1) High return on capital businesses (ROIC), irrespective of the valuation at which we buy them? Is it better for future returns to just buy the most profitable companies?

2) Low purchase valuation, irrespective of the quality and return of capital on the business? If we just buy a business very cheaply, that by itself could ensure an attractive return going forward

This will help focus and direct the research efforts of the investment company and drive conclusions.

- What is your input data, and how did you decide to use the data?
  The data will be for the Sample we use as broad as possible:

  [Companies, Valuation P/E Multiple (t0), ROIC (t0), Returns (t+1)]
  There will be 4 columns

- What are your expected clusters or output?
  I would expect the data to form the clusters charted below. Basically for the 2 different hypothesis we are testing:

High return on capital businesses: would expect the data to group around High ROIC/ High Expected Returns, Low ROIC/ Low Expected Returns

2) Low purchase valuation: would expect the data to group around the opposite quadrants of the previous example 1), meaning High Entry Valuation/ Low Expected Returns, Low Entry Valuation/ High Expected Returns

- How many clusters or groups should your model create, and why? How would this help the organization as a whole? What can you infer from the results?

  I would expect each model/premise to have 2 clear clusters on the opposite ends as outlined above because higher or lower ROICs of a business should drive higher or lower returns.
  On the opposite side lower entry valuation drives higher returns and higher entry valuations drive lower returns as you get either the benefit of the P/E expanding from a low valuation or contracting from a high valuation.
  This would help make better conclusions through the use of data and direct research efforts more effectively for the organization.

  Conclusions:
  From the results you can infer that the research efforts should focus on companies with a low P/E multiple and ideally with a high ROIC as well. This is the most fertile area to direct research efforts in order to obtain high future returns on our stock investments for our investment management company.

Clusters_.pptx (https://classroom.emeritus.org/files/2496378/download?download_frd=1&verifier=P5uGuEQvwCMmCudg5ZCpVxG47pdPzJiQcd9thY8a)

Reply

○

(http

Turki Alghusoon (https://classroom.emeritus.org/courses/9054/users/229165)

Apr 21, 2024

Hi Javier,

Cool idea for clustering and thank you for including the ppt slide. it helped me visualize the clusters. I have a question on the conclusion: does the model suggest that companies with high Valuation P/E Multiple are potentially bad investments in general? or this conclusion related to a certain time frame (i.e. long-term vs. short-term investments)? and if time is a factor here, would you split the analysis by time frame?

Thank you.

Turki

Reply

○

(http

Javier Di (https://classroom.emeritus.org/courses/9054/users/226884)

Apr 23, 2024

Thank you Turki and that's a great question. Given high PE and lower future expected lower return will form a cluster as indicated by the PPT slides, majority of high P/E companies will not be great investments. However, it's just a cluster and there will be outliers at high P/E multiples that will have a high return.
In the business world, this is because few high P/E companies can grow at high enough rates to justify that high PE and will de-rate quickly when they eventually don't grow as fast.

And yes, this exercise is done as T0 and T+1 but in investing if you have 10years or more for example the entry multiple will matter less and the dominant component will be the earnings growth over that longer period.

In shorter periods it's all about the multiple P/E expansion or contraction. Hope this helps

Reply

○

(https:/

Haitham Farag (https://classroom.emeritus.org/courses/9054/users/233864)

Apr 21, 2024

Business use Case

Turn ad hoc/ occasional donors of earmarked funds (i.e. for a specific response) to a monthly donation programme that is not earmarked (i.e. can be used flexibly to respond to any emergency.

Customer Input Data

- Age
- Gender
- Employment Status
- Donation history within the past 12 months (amounts, frequency, periodicity)
- Donation type (tax-deductible)
- Postal Code

Input Data Rational

The above features were identified based on their relevance to the business use case and accessibility. While other features like annual income, job title and employer would be good alternatives for feature; however, this info is privileged and protected by law.

Expected Clusters

Given the limitation of both K-Means and AHC clustering approaches in dealing with multidimensional, a decision is required on which of the above customer features to cluster donation against. This decision can be supported by a regression model or by observing the generated set of clusters of donations against each of the above features. The outcome of either process should assist in deciding on the 2 most likely correlated features.

Number of Clusters

Execute a Script Plot (K vs Within Cluster Sum of Squares) and arbitrarily select one of the K values at the curve's elbow. The available (market campaign) budget and business objectives could eventually influence this decision.

Cluster Shapes

The organization has better potential of benefiting from the clustering exercise when:

- clusters are convex-shaped
- points in each cluster are closer to each other and the cluster centroid
- many points in each cluster
- clusters are spread (dispersed) horizontally with little to no overlap (no cluster falls in another shadow)
- limited number to no outliers

The benefit of having clusters with the above attributes would reflect in the impact of the marketing campaign to reach a wider audience with greater precision (i.e. higher return on investment). The result will inform the design of the market campaign and the target group (cluster). Nevertheless, the cluster model remains descriptive and may not shed enough light on how to shift donors (points) from one cluster to the desired cluster (e.g. monthly donations, higher donations).

Edited by **Haitham Farag (https://classroom.emeritus.org/courses/9054/users/233864)** on Apr 21 at 8:19pm

Reply

○

(http

Diego Milanes (*He/Him*) (https://classroom.emeritus.org/courses/9054/users/228518)

Apr 22, 2024

Hi Haitham

Thanks a lot for this nice example. I wonder if convex-shaped clusters (for better potential on benefit as you mentioned) can be easily created using k-means or AHC, or if other type of metric/method is needed to obtain these cluster shapes.

cheers!

Reply

○

(http

[Haitham Farag (https://classroom.emeritus.org/courses/9054/users/233864)](https://classroom.emeritus.org/courses/9054/users/233864)

Apr 23, 2024

Good day Diego

Thanks for the feedback. The Python clustering subject matter is new to me and because I have a somewhat a better grasp of KMeans over AHC, the framework was based on using KMeans.

To my understanding, KMeans can produce convex shape clusters (contingent on the data), [resource: www.cs.ubc.ca (course material), (https://www.cs.ubc.ca/~schmidtm/Courses/340-F19/L9.pdf)](https://www.cs.ubc.ca/~schmidtm/Courses/340-F19/L9.pdf) while AHC does not.

Please feel free to rectify the above point.

Sharing any additional material on handling Python AHC and its interpretation would be much appreciated.

Regards

Reply (1 like)

○

(http

[Priscilla Annor-Gyamfi (https://classroom.emeritus.org/courses/9054/users/226376)](https://classroom.emeritus.org/courses/9054/users/226376)

Apr 25, 2024

Great post Haitham,

I like the kind of data you are able to collect on your customers for this analysis. However, I was looking forward to seeing a clear indication on the number of clusters used and which part of the data collected were used in the clustering. For instance, you would like to analyze the frequency of donations by customers and get to know the age group that donates more. You will do this in conjunction with other features like employment status and donation history.

Reply

○

(http

Haitham Farag (https://classroom.emeritus.org/courses/9054/users/233864)

Apr 25, 2024

Thanks, Priscilla for the feedback.

Your suggestions are well noted. I am looking into how PCA can assist in identifying the key features.

Reply

○

(https:/

Turki Alghusoon (https://classroom.emeritus.org/courses/9054/users/229165)

Apr 21, 2024

- **Business Case:** As the organization's executive, I would like to analyze the compensation distribution for employees in the organization to gain assurance on pay equity across the organization. By creating cluster for staff within each job family, I can identify staff that over-underpaid relative to their peers and take corrective actions to ensure that people are equally compensated for equal work and the value they bring to the organization.

- **Input data:** For this analysis I will use the following input data:
  - Input Data
    - Employee number
    - Employee Job Family
    - Employee Salary
    - Employee Years of Experience
    - Employees education and certification data
  - How to Use the Data:
    - Separate employee data by job family
    - For each job family, I will create clusters based on salary and years of experience.

- **Expected Cluster Output:**
  - I expect employees within each job family to be generally within clusters based on years of experience and salary. More junior employees will be clustered at the lower pay range, where more senior employees will be clustered around the higher pay range.
  - I also expect there to be anomalies:

        - Some employees with rare credentials and higher education levels might be higher on the pay range than their age cluster and visa versa.
        - There could also be rare cases where somebody is severely underpaid or overpaid with no valid reason.
        - The root cause for all outliers will be determined by isolating the outliers and taking a deeper look into their credentials to understand the context.

- **How Many Clusters:**
    - I want the model to create 4 clusters per job family.
    - **Rationale:** Since employees in general start their careers in their twenties and retire in their early sixties, that would create a natural grouping per decade of experience which could serve as the starting point.

- **How Would this help the organization as a whole:**
    - This will help the organization in many ways:
        - It will send a strong message to staff that organization cares about their well-being
        - It will motivate staff to be more productive and aspire to stay in an organization where they have confidence they will be treated fairly. This could also help the organization attract more talent
        - It could also unearth any issues of staff being undercommented or overcompensated in an unfair manner, which is not only wrong, but also detrimental to the reputation of the organization

- **What I can infer from the results:**
    - What job families have the highest pay range.
    - What job families have the highest variance in pay (widest pay range).
    - How consistent is organization in paying employees relative to their experience and job families.
    - Are there cases where employees could potentially be unfairly overpaid or underpaid.

Reply

○

(http

Roy Nunez (https://classroom.emeritus.org/courses/9054/users/229552)

Apr 21, 2024

Hi Turki,

Great example! I just wonder why performance/productivity is not being considered in the input data. I think this could help explain some of the rare cases where "somebody is severely underpaid or overpaid with no valid reason." as many raises are performance based, and some of the value they bring to the organization is commonly measured by some metric.

Reply (1 like)

○

(http

Ricardo Anaya (https://classroom.emeritus.org/courses/9054/users/228915)

Apr 24, 2024

+1 on the productivity question, how is is mesured?

more on the efficient time at work (some can do full day job in hours while some need more time, more than a day)

both efficient but more Key KPIs would be needed to asses a compensation.

thought task though.

Reply

○

(http

Yossr Hammad (https://classroom.emeritus.org/courses/9054/users/229118)

Apr 22, 2024

Hello turki,

Great example. i was wondering why did you decide creating 4 clusters. i read the rationale but couldnot get exactly why 4. in my example i couldnot decide on how many clusters since i think there are specific techniques , thats why i was wondering did you use any technique to come up with that specific number.

thank you

Edited by Yossr Hammad (https://classroom.emeritus.org/courses/9054/users/229118) on Apr 23 at 6:46pm

Reply

○

(http

Chris Cosmas (*He/Him*) (https://classroom.emeritus.org/courses/9054/users/226607)

Apr 25, 2024

Hello Turki,

Very nicely presented, it is a very important topic and often creates insecurities and unpleasant feelings with employees if they learn they aren't compensated correctly.

I had a thought while reading your example. Wouldn't tenure and age be more correlated creating elliptical shapes rather than circular clusters?

The datapoints will be mostly shaped in straight lines which might affect the cluster outcomes.

Reply

○

(https:/

Roman Jazmin (https://classroom.emeritus.org/courses/9054/users/225803)

Apr 22, 2024

As a company executive using a cluster model to extract valuable information from the company's recent global survey of untouched oil reserves, I would cluster my data based on (1) longitudinal data sets, (2) latitude data sets, (3) measured depths of all the untapped oil reserves, (5) regions or areas with highest concentrations and number of oils reserves, and (6) filter out areas that are already occupied.

How I would envisioned analyzing these data sets are to have them overlap with each other and after careful analysis of data points with the most color over laps, we can predict or at least understand the conditions that we need to look for when searching for new oil reserves that would put us ahead of our competitors at the same time reduce the number of costly oil reserve explorations in the future.

Reply

○

(http

Jignesh Dalal (https://classroom.emeritus.org/courses/9054/users/229173)

Apr 24, 2024

It's fascinating how you're planning to use clustering to analyze the spatial distribution of untapped oil reserves. This approach not only seems efficient for identifying high-potential areas but also strategic in terms of reducing unnecessary exploratory costs and staying ahead of competition.

Given the complexity of the data sets you mentioned—spanning longitudinal and latitude data, depths, and regional concentrations—I'm curious about the specific clustering algorithms you find most effective for handling such multidimensional data. Are there particular techniques or tools that help in managing the vast amount of data and ensuring the accuracy of your predictive analysis?

Reply

○

(https:/

Diego Milanes (*He/Him*) (https://classroom.emeritus.org/courses/9054/users/228518)

Apr 22, 2024

My hypothetical case is the identification of student performance groups for an educational virtual platform (something like canvas)
The input data will be the student's interaction with the platform, such as engagement with course materials, time spent on the platform, courses taken, assessments completed, etc.; student-related data like age, location, gender, and academic background; and finally, some performance data such as grades and scores.
I expect to categorise students into different academic groups. The number of clusters is then determined by the academic performance levels (high-, average-, low-performance), engagement level (high-, moderated-, low-engagement), and student-related characteristics.
For instance, if a cluster of highly engaged corresponds to high-performing students, then the online platform must offer advanced courses. Conversely, some personalised tutoring or other learning strategies must be applied for low-engagement and low-performance students.
This strategy (clustering) will help identify students' needs and preferences and prioritise activities to improve their academic performance. We can also identify outliers and abnormal patterns and take appropriate actions to resolve the issues.

Reply (1 like)

○

(http

Manjari Vellanki (https://classroom.emeritus.org/courses/9054/users/231480)

Apr 22, 2024

Hi Diego-

Interesting example. I'm just eager to know, what would be the other possible variable against Score/grades that feeds in to the requirement to assess performance levels. Also, what would be the ideal number of clusters and any further classification of variables is needed?

Reply

○

(http

Mariana Flores (https://classroom.emeritus.org/courses/9054/users/237198)

Apr 24, 2024

Hi Diego, so nice to connect on the discussion board again. Great post, cluster analysis across the education field is fascinating. I agree with you in that anomaly detection is an important area where cluster analysis can help to resolve issues.

The real-world application of cluster analysis is remarkable - thank you for sharing.

Reply

○

(http

Koffi Henri Charles Koffi (https://classroom.emeritus.org/courses/9054/users/208039)

May 1, 2024

hey Diogo , great post . what will be the optimal number of cluster here ? is the number of cluster beeing determine upfront ? that will be great if the online platform like emeritus can implement this model it will really improve student experiment and facilitate learning

Reply

○

(https:/

Yossr Hammad (https://classroom.emeritus.org/courses/9054/users/229118)

Apr 22, 2024

As an executive to our hospital i've come to realize how data can improve how we care for our patients as well as running the operations. i believe that using clustering models to analyze patients information will benefit us.

The input data includes patients attributes such as : medical history, diagnoses, treatment received, length of stay and costs. these data will allow us see similarities and abnormalities that might not be obvious.

Expected clusters would be grouped for example as follows: maternity patients, pediatric patients, elderly patients, and chronic disease patients. Each cluster would have unique needs, different resource utilization and different treatment patterns.

The output or the goal is to group patients into clusters based on their health history and how they utilize our care.

Getting the right number of clusters is important and critical. if few we would miss the key difference between groups which is the main goal to use clusters, too many it will get overly fragmented that wont help and non actionable. we will need to go through analytical exercises to get the optimal number.

The benefits:
1- personalizing healthcare for each cluster differently based on their needs and health profile.

2- best utilization of our resources, knowing which cluster of patients need more care, longer hours.. will definitely help allocate our resources in the best way possible.

3- identify the high risk clusters like those patients who frequently are admitted to the hospital, and those who have complicated medical history so we can proactively mitigate health risks.

4- we can develop cost effective strategies by analyzing resource allocations and find opportunities for cost saving.

5- indicate the best practice that lead to best outcomes for different clusters to improve quality.

At the end of the day , the clusters approach will allow us to improve and provide each cluster with their needs and run our hospital in the best way possible financially.

Edited by **Yossr Hammad (https://classroom.emeritus.org/courses/9054/users/229118)** on Apr 22 at 7:12pm

Reply

○

(http

STEPHEN HUTSON (https://classroom.emeritus.org/courses/9054/users/233645)

Apr 24, 2024

Hi Yossr! I think this was a good example of a clustering use case in that you were able to infer really useful potential insights into managing patients within a hospital. I particularly liked the concept of identifying patients more frequently admitted to the hospital with more complex medical histories to know who the higher risk patients are, which I think could be expanded as potential populations to reach out to for medical trials.

Reply (1 like)

○

(https:/

Jignesh Dalal (https://classroom.emeritus.org/courses/9054/users/229173)

Apr 23, 2024

Business Use Case: Enhancing Fraud Detection and Understanding Customer Transaction Patterns:

Objective:

1. Detect unusual transaction patterns that might indicate fraudulent activities.
1. Segment customers based on their transaction behaviour to tailor financial products and advice.

Data Collection and Analysis Outcomes:

Data Input and Reasons for Data selection: The effectiveness of our banking fraud detection system relies on the strategic selection and use of several key data types. The primary dataset includes detailed transaction logs, such as the amount, date/time, location, transaction types (e.g., online, ATM, app, POS), and any suspicious activity flags. This data is vital for spotting fraud patterns and understanding customer behavior trends.

We also use customer demographic data, such as age, occupation, income level, account type, and the customer's relationship duration with the bank. This information helps profile customers, allowing us to potentially link specific demographic features with certain transaction types or identify potential fraud suspects.

Finally, we use data on past transactions confirmed as fraudulent. This historical data is crucial for training our models to recognize and predict patterns associated with confirmed fraud.

This comprehensive approach to data collection and analysis aims to improve the accuracy and effectiveness of our fraud detection.

Expected Cluster and Outcomes

Expected Cluster:

The expected clusters for our analysis of bank transactions aim to identify and categorize distinct types of financial behaviors and potential fraud scenarios. One cluster is designed to capture fraudulent transactions, which are those that deviate significantly from typical customer patterns. Such anomalies are crucial for detecting potentially illicit activities. Another cluster focuses on high-value transactions, characterized by unusually large transaction amounts compared to typical customer behavior, which might signify unique purchasing needs or, conversely, red flags for further scrutiny. Additionally, there's a cluster for frequent small transactions; this pattern may suggest a different kind of financial behavior, such as regular low-value payments, or it could indicate test transactions typically carried out by fraudsters probing the system. Lastly, demographic-based transaction patterns form another cluster, grouping customers who exhibit similar transaction behaviors based on demographic factors. This helps in understanding customer segments more profoundly and tailoring fraud detection mechanisms and financial services to fit distinct demographic profiles.

Model Specifications and Organizational Benefits

Number of Clusters and why?

For our clustering model, we are targeting the formation of approximately 4-6 clusters. This range is chosen to effectively capture distinct patterns in the data without rendering the model too complex for interpretation. Opting for 4-6 clusters strikes a balance between granularity and manageability, ensuring that each cluster is statistically significant and actionable without being overwhelming. This approach allows for a detailed analysis that remains practical and interpretable, providing clear insights that can be directly applied to improve decision-making and operational strategies.

Benefits to the Organization

Our clustering model provides several strategic benefits to the organization. First, it enhances fraud detection capabilities by enabling the prompt identification and response to potential fraud. This not only helps in reducing financial losses but also increases customer trust by demonstrating a proactive approach to safeguarding their transactions. Secondly, the model offers deeper customer insights by analyzing transaction behaviors, which aids in providing personalized financial advice and tailoring product offerings to meet individual customer needs more effectively. This customization enhances customer satisfaction and loyalty by aligning services with their specific financial habits and preferences. Lastly, the model contributes to operational efficiency by streamlining monitoring processes. It focuses efforts on transactions that are most likely to be fraudulent, thereby optimizing the use of resources and allowing for more efficient management of risk and security measures within the organization. These improvements collectively enhance the overall functionality and responsiveness of our financial services, leading to better customer service and operational success.

Inference from the Results

Our clustering model has demonstrated several key outcomes that contribute to improving both security measures and customer engagement. Firstly, by identifying vulnerable groups that are more susceptible to fraud, we are able to implement additional security measures or targeted advisories to protect these specific segments, enhancing their safety and trust in our services. Secondly, there is an ongoing refinement of our fraud detection algorithms. As the model is continuously refined with additional data, its accuracy and effectiveness in detecting fraudulent activities are enhanced, ensuring our systems remain robust against evolving threats. Lastly, the insights gained from analyzing customer spending behaviors are utilized to tailor marketing strategies and product offerings more effectively. This customization ensures that our marketing efforts are more aligned with customer needs and preferences, thereby increasing the relevance and attractiveness of our products. Collectively, these strategic outcomes not only boost our operational capabilities but also significantly improve customer satisfaction and security.

In the above scenario, Clustering model acts a powerful tool to decode complex dataset into actionable insights. This significantly contributing to strategic decision-making and operational improvements in the banking sector.

Below we will see an step by step example with synthetic dataset:

Scenario Setup:

Step 1: Create a synthetic dataset

- Transaction Amount: Values are all over the place, but the bigger ones don't show up as much.
- Transaction Type: Grouped into ATM withdrawals, POS transactions, and online purchases.
- Customer Age: The age range spans from 18 to 70 years.
- Number of Transactions per day: Certain clients may have numerous small transactions. These could be either standard or potentially alarming, depending on established patterns.

Step 2: Apply Clustering

- We're going to apply K-means clustering to spot clusters in transaction patterns. We've picked K-means due to its proven track record in efficiently grouping data into distinct clusters, which is influenced by variations in transaction behaviours.

Step 3: Analyze Clusters

We'll have a look at the clusters to spot any that could indicate fraudulent activities—like uncommonly high transaction volumes or extremely large transaction amounts.

Step 4: Visualization

Let's take a look at the clustering results to get a better understanding of transaction patterns and possible suspicious activities.

Let's explore each of these steps in detail, using Python code and visualizations.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

np.random.seed(42)
data = {
```

```
'Transaction Amount': np.concatenate([np.random.exponential(scale=150, size=950),
np.random.exponential(scale=1000, size=50)]),
'Transaction Type': np.random.choice(['ATM', 'POS', 'Online'], size=1000),
'Customer Age': np.random.randint(18, 70, size=1000),
'Transactions per Day': np.concatenate([np.random.poisson(3, size=900),
np.random.poisson(10, size=100)])
}

df = pd.DataFrame(data)

df_encoded = pd.get_dummies(df, columns=['Transaction Type'])

scaler = StandardScaler()
scaled_features = scaler.fit_transform(df_encoded)

#k-means
kmeans = KMeans(n_clusters=4, random_state=42)
kmeans.fit(scaled_features)
df['Cluster'] = kmeans.labels_

# Graphs
plt.figure(figsize=(10, 6))
scatter = plt.scatter(df['Customer Age'], df['Transaction Amount'], c=df['Cluster'], cmap='viridis',
alpha=0.6)
plt.title('Clustering of Bank Transactions')
plt.xlabel('Customer Age')
plt.ylabel('Transaction Amount')
plt.colorbar(scatter)
plt.show()
```

Insights from the Graph:

- **Clusters primarily differ based on the transaction amount**, with distinct groups for regular and high-value transactions.
- **Age distribution** across clusters appears to be relatively uniform, indicating that age might not be the main distinguishing factor in this clustering model.
- Clusters with higher transaction volumes may suggest areas to inspect for unusual or fraudulent activities, especially if these transactions significantly deviate from a customer's

usual behaviour.

CustomerAge_Transaction.png (https://classroom.emeritus.org/files/2505558/download?download_frd=1&verifier=ceUXg8GtKPwgm1LSvkdL3j2T0sQxwMfzUlU6YGNE)

Reply

○

(http

Chris Cosmas (*He/Him*) (https://classroom.emeritus.org/courses/9054/users/226607)

Apr 25, 2024

Hello Jignesh,

Very well put, this is an excellent scenario.

I was having a hard time understanding why cluster algorithms are good for fraud detection, I hadn't understood that these would create a cluster of their own indicating deviations from the norm. I thought it would create outliers which would still be input in one of our defined clusters. Thank you for clarifying.

Reply

○

(https:/

Mariana Flores (https://classroom.emeritus.org/courses/9054/users/237198)

Apr 23, 2024

As an executive of a direct-to-consumer product company attempting to increase the value of each customer purchase through upsell and cross-sell strategies, I would tailor product recommendations so that recommendations are relevant and complementary to each customer segment. Leveraging a clustering model to extract data from a centralized repository for the purpose of understanding customer characteristics and preferences and using input data like purchase history and customer demographics to match to each customer's product needs and tailor recommendations accordingly. The model should create the most personalized clusters or customer groups while balancing speed and cost. I would then use these clusters as inputs for my recommendation engine to generate product recommendations based on customer segments. I would also implement hypothesis testing through an experimental design and analysis methodology as well as measurement strategy to be able to evaluate the performance of these personalized product recommendations and offerings to both evaluate past performance and improve future investment accordingly. This would help the organization by intelligently guiding strategic decision-making and maximizing return on resource allocation.

Reply (1 like)

○

(http

Shahrod Hemassi (*He/Him*) (https://classroom.emeritus.org/courses/9054/users/224267)

Apr 23, 2024

Hi Mariana. Great post. I agree with how you have identified clustering as a way to analyze your customer's shopping behavior. It would be interesting to see which data elements contributed to patterns of behavior that may not have been initially apparent to the company. Gaining insight from this analysis could provide significant value to your company's marketing investment decisions and could lead to significant profit gains over time. Thanks for posting this topic.

Reply

○

(http

Lee Lanzafame (https://classroom.emeritus.org/courses/9054/users/231975)

Apr 23, 2024

I work at a telecommunications company and that's how we do it! great work

Reply (1 like)

○

(http

Mariana Flores (https://classroom.emeritus.org/courses/9054/users/237198)

Apr 24, 2024

Genuinely, it means a lot :)

Reply

○

(https:/

Shahrod Hemassi (*He/Him*) (https://classroom.emeritus.org/courses/9054/users/224267)

Apr 23, 2024

In this module, we learned about the value of using Clustering algorithms to find insights in large volumes of data that may have similarities in different groupings of the data. I feel like this would be advantageous for professional sports and sports betting businesses in monitoring fair play among the professional athletes. With the increase in sports betting, there have been instances where some professional sports athletes have been approach by large betters who offer them money in order to swing a betting line in their favor. The player may accept the offer (which is illegal) and adjust their play slightly to affect game results and/or individual statistical results. This illegal activity can be difficult to detect, but I see this as an area where data scientists can use clustering algorithms to identify normal patterns of behavior and in turn detect abnormal patterns of behavior.

First, we would need to process data about the performance of individual athletes across many games and identify similarities in the performance data. Of course, some athletes are better than others and the typical performance relative to each athlete can vary significantly. Some athletes perform at a high level, others and medium levels, and others at low levels. It may initially appears as if we are unable to compare the statistical data for a high performance athlete against a low performing athlete, but I suspect that if we cluster the athletes together, we will see similarities in their performance over time based on the cluster that they are in and relative to the performance of other players in the same cluster.

Similarly, we can analyze betting patterns on games. Typically, betting platforms set betting odds on games and on performance of individual athletes in games to a level where they will expect roughly the same amount bet in each direction. For example, the betting odds (the betting line) for a basketball player's performance in a particular game may be set at 10.5 points and the expectation is that roughly an equal amount will be bet that the player will score more than 10.5 points in the game as less than 10.5 points. Betting platforms will adjust these lines as bets are placed in order to try to ensure that they have roughly an equal amount bet in either direction. The reason for this is that the betting platform gets a small amount of each bet placed so as long as they have about the same amount bet on either side of a betting line, the betting platform is sure to gain some profit from all the bets collectively.

Data scientists can again help the betting platforms with analyzing their data. They can use clustering to detect similarities in betting patterns. They may have some customers who bet large amounts and others who bet small amounts and it may initially seem like they cannot compare the betting patterns of these different types of customers but using clustering, we may find similarities and insights into the betting patterns. In particular, we would look to identify common betting patters so we can identify abnormal betting patterns.

At times, a player has a particularly good game or a particularly bad game. That happens and it may not be an indication of anything other than the player having a better-than-normal or worse-than-normal game. Also, there might be times where the betting community bets heavily in one direction on a betting line and this might just be a case where the betting line was not set at the right point or that the betting community had a different feeling about a particular betting line. But if we have situations where abnormal player performance aligns with abnormal betting patterns, we may have a situation where an athlete is illegally altering the results of a game or individual statistics illegally.

In fact, something like this recently occurred in the NBA and a player named Jontay Porter received a lifetime ban from playing in the NBA. The NBA apparently is already working with betting platforms to identify these outliers in player performance and betting patterns. Maybe they have hired data scientists who are using clustering algorithms to provide this insight. There are many professional sports consisting each of many sports leagues who each have many athletes on the teams in the leagues. Betting is available all over the world on all sports, sports leagues, and sports athletes. The importance of data science and methods such as clustering algorithms to these businesses in immense.

This is one area where it is evident that the techniques we are learning will be of great value.

Reply

○

https:/

Lee Lanzafame (https://classroom.emeritus.org/courses/9054/users/231975)

Apr 23, 2024

I'm taking the position of an executive at a telecommunications company. We currently have a rewards store where customers can accumulate points over time based on; the products they subscribe to, if they pay their bills on time and a few other measures.

Customers can use these points to purchase things like Fitbits, speakers and mobile phones. (Over 1000 items and 20 categories). The current system hasn't been working well with people not claiming their points or even logging into the system.

Our current website doesn't tailor specific things to people its just static pages. We have a machine learning algorithm that tries to cluster customers into categories but we have over 20 parents categories and 100's of sub categories and the business refuses to adopt the current model because of its inability to correctly cluster customers.

We have done feature engineering on a variety of customer attributes (Demographic information, interaction data and account information).

I think the current approach isn't working due to the number of categories, subcategories, number of products and we need to change the way we think about clustering.

I first propose that we reduce the number of subcategories and products.

We can still use K-means and the elbow method, we just need to change the way we think about clustering. The new idea is to expect 5 clusters ie (High value customers, low value customers, early adopters, loyal customers and infrequent customers)

We can monitor how this new approach goes and we can see if website use increases, redemption increases, customer retention improves and we have the right product mix on the site.

Reply

○

(http

Dawn Prewett (https://classroom.emeritus.org/courses/9054/users/233112)

Apr 24, 2024

You present the vision of a few of the companies I've worked for the past, so I felt kind of pulled into the narrative. Which made me wonder whether the perceived value of the points matches the perceived value of the reward. There is this tendency when working on the company side of this equation to see "anything that is free is good", but from the customer's standpoint they earned those points, so they weren't free - they had to do something to get them. The perceived value of a single point will vary by person as will the perceived effort to utilize the points. If the value of the points is too high, it may cause what the customer to feel that it's seemingly impossible to earn enough points to get the prize they perceive as desirable. If the points are right sized, but the process of using them is too intrusive, suddenly the value of the prize decreases drastically. It would be interesting to see if the mix is right. Additionally, based on your description of the website, I wonder if users are experiencing fatigue even trying to find a prize they want, putting the whole process into the category of "why bother". But as you noted, with too many categories and buckets, the picture is so obscure it's hard to tell what is going on.

Reply

○

(http

Ahmad Abu Baker (https://classroom.emeritus.org/courses/9054/users/234460)

Apr 24, 2024

Hi Lee Lanzafame,

Your insights into reimagining customer clustering at your telecommunications company are both insightful and practical. It's evident that the current rewards system and website setup face significant challenges, particularly with low customer engagement and the inability to effectively categorize customers for tailored experiences.

Reducing the number of subcategories and products seems like a logical first step in simplifying the clustering process and making it more manageable. By focusing on broader customer segments like high value, low value, early adopters, loyal customers, and infrequent customers, you can create more meaningful and actionable clusters that align with your business goals.

Your proposal to continue using K-means and the elbow method while adjusting the clustering approach demonstrates a pragmatic approach to leveraging existing tools and techniques. It's essential to adapt and evolve your methods to better meet the needs of your business and customers.

Monitoring the outcomes of this new clustering approach, including website usage, redemption rates, customer retention, and product mix optimization, will provide valuable insights into its effectiveness. This data-driven approach will enable you to iterate and refine your strategies over time, ensuring continuous improvement and better alignment with your business objectives.

Overall, your proposal showcases a thoughtful and strategic approach to addressing the challenges faced by your telecommunications company. By embracing innovation and leveraging data-driven insights, you're well-positioned to drive meaningful improvements in customer engagement, retention, and overall business performance.

Best regards,

Ahmad Baker

Reply

○

(https:/

Dawn Prewett (https://classroom.emeritus.org/courses/9054/users/233112)

Apr 24, 2024

In my previous role, we had a great deal of analysis to calculate how many people we actually needed to hire for lower-wage, warehouse roles in order to actually fill the positions since not everyone who accepts a role, starts work. Despite crunching data and having a mountain of it at our disposal, we never could quite determine what factors influenced a candidate to not start their new position, despite accepting it. We hypothesized that increased pay and schedule flexibility would mitigate this.

If I were to return to the same job and dive in to test this hypothesis, I would conduct a targeted analysis by examining data on actual job starts versus non-starts, and correlating this with varying pay scales (adjusted for regional cost of living), alongside the flexibility of working hours. This analysis would involve measuring deviations from median work hours, aiming to understand the impact of flexibility.

Walking into one of these facilities, you quickly notice that the population varies greatly. While younger individuals, usually fresh out of high school or still in college, make up a massive employee base for these positions, there are also a number of individuals that range the gambit of ages – including more elderly individuals. It would make sense that younger workers, possibly less encumbered by financial responsibilities, were more willing to accept lower wages, but what about the others? Looking at hiring trend for older workers, there are two distinct reasons they are taking lower wage jobs – the need for supplemental income and something to do with all of their free time. Those who work more than one job, are also a common thread in these jobs.

From the extensive dataset available, spanning several years, it would also be crucial to consider external factors such as significant economic shifts, natural disasters and global pandemics, as well as normal cyclic hiring shifts—that could influence hiring patterns and cluster formations, then normalize the data to allow for accurate and meaningful analysis.

Based on my knowledge of the data and what we've learned about clustering algorithms, I predict there will be five distinct clusters:

- The Necessity Cluster: These individuals will accept any available job due to urgent financial needs, regardless of pay or flexibility. They may be in sensitive groups that find it harder to secure work due to a having a record, battling addiction or homelessness.
- The Supplemental Income Cluster: These individuals prioritize flexible jobs that can accommodate their primary, and more financially fruitful, employment. These individuals may be looking for additional short time work to manage unexpected bills, pay down a mountain of debt, or get "back on their feet" after a major life change.
- The Dual Job Cluster: For these individuals, the job serves as a primary source of income, not the only source of income, so flexibility is needed. These individuals are likely in the middle of the payscale for these low-wage jobs, but still aren't making quite enough to meet all of their financial needs.
- The Comfortable Cluster: These employees hold higher-paying roles where the pay meets their living requirements, but the positions lack flexibility. This group will be denser than the Ideal Job Cluster, but much smaller than the previous groups.
- The Ideal Job Cluster: A very small, sparse group that holds one of the higher paying and flexible positions available.

These clusters provide a group that potentially represents the average population that works in these positions. Seeing how pay and flexibility impact willingness to start could aid in the understand why people show up when others do not. Doing a similar exercise with non-starts could create a better rounded view of the issue as well.

Reply

○

https:/

Ricardo Anaya (https://classroom.emeritus.org/courses/9054/users/228915)

Apr 24, 2024

- What is your input data, and how did you decide to use the data?
  - In my case:
    - Wireless Bands for Wireles Carriers
    - Bandwithds
    - Combinations of Bands 4G + 4G
    - Combination of bands 4G +5G
    - Products supporting bands and Combination 4G+4G and 4G+ 5G
- What are your expected clusters or output?
  - Clusters with higher numbers with Hihger specs with Certain Premiun Products
  - Going below in tiers, the suppor dimishes, I want to minimize the risk of having a Critical combination not supported by a lower tier modem
  -
- How many clusters or groups should your model create, and why?
  - 1 in the low band, 2 in the middle band 1 in the Higher band and one in the mmwave
  -
- How would this help the organization as a whole?
  - have better products supporting all needed bands and combinations
  -
- What can you infer from the results?
  - some products will need update, future products in the lowe tier would need to adress higher capacity
  -

Reply

○

https:/

STEPHEN HUTSON (https://classroom.emeritus.org/courses/9054/users/233645)

Apr 24, 2024

For my use case I'll approach clustering as an executive in a manufacturing organization looking to optimize production and operational efficiencies within my organization. The types of input data I would use would include production data, like production dates/times, information around machine operations, and product quality metrics including results around yields and defective product rates. Other valuable input data would include information around our suppliers and materials, inventory levels, and the length of time it takes for each manufacturing step to take place. This information would be valuable to analyze because it would inform our organization around our manufacturing process as well as insights into what may impact our production efficiency in order to analyze areas where we may need to target improvements to reduce the number of defects and see what factors are slowing down our production processes. For our cluster outputs, we would look at clusters based around metrics like the number of defects by product type to see if certain products tend to come out defective more than others. I would determine the number of clusters by looking at the elbow method to ensure we're factoring in the optimal number of clusters for our analysis. The overall benefits to this type of analysis in this example would help with our quality control efforts to target areas where additional investment may be needed to reduce defect rates, as well as looking at how we're utilizing our inventories and suppliers to identify areas for increased manufacturing efficiencies.

Reply

(http

Timothy Andrew Ramkissoon (https://classroom.emeritus.org/courses/9054/users/226697)

May 1, 2024

Your choice of input data is comprehensive and relevant. By including production data, machine operations, quality metrics, supplier information, inventory levels, and process durations, you cover critical aspects of the manufacturing process. Your approach is robust, and by fine-tuning specific aspects, you can drive substantial improvements in manufacturing efficiency. Keep iterating and collaborating with stakeholders to maximize impact.

Reply

(https:/

Ahmad Abu Baker (https://classroom.emeritus.org/courses/9054/users/234460)

Apr 24, 2024

Hello everyone,

As an executive at EY, a future leader in the business consulting sector, leveraging clustering models presents an exciting opportunity to extract valuable insights from our vast repository of data. One compelling business use case for clustering algorithms within our organization is to enhance client segmentation and better understand their diverse needs and preferences.

The input data for our clustering model would encompass a wide range of client-related variables, including industry vertical, revenue size, geographical location, past engagement history, and specific consulting needs. We decided to utilize this data because it offers a comprehensive view of our client base and enables us to identify distinct patterns and segments within it.

The expected output of our clustering model would be well-defined client segments or clusters, each representing a unique profile or segment of clients with similar characteristics and consulting requirements. These clusters would empower us to tailor our consulting services and solutions more effectively to meet the specific needs and preferences of each client group.

In determining the number of clusters, we would leverage advanced clustering techniques and methodologies to identify the optimal number of segments that best capture the underlying structure of our client data. This would enable us to avoid under- or over-segmentation and ensure that the resulting clusters are actionable and meaningful.

The insights derived from our clustering model would provide significant benefits to our organization as a whole. By gaining a deeper understanding of our client segments, we can personalize our service offerings, develop targeted marketing strategies, allocate resources more efficiently, and drive client engagement and satisfaction to new heights. Additionally, it would enable us to identify emerging trends and opportunities within specific industries or regions, positioning us as trusted advisors and thought leaders in the business consulting space.

In summary, leveraging clustering algorithms to enhance client segmentation and insights holds tremendous potential for EY's business consulting sector. By harnessing the power of data-driven analytics, we can unlock new opportunities for growth, innovation, and client value creation, reaffirming our position as a global leader in the consulting industry.

I'd love to hear your thoughts on how we could further refine our approach or any additional insights you may have on leveraging clustering algorithms within EY's business consulting sector.

Best regards.

Ahmad Baker

Reply

○

(http

Swati Sharma (https://classroom.emeritus.org/courses/9054/users/236938)

Apr 30, 2024

Hello Ahmed :Your insights on leveraging clustering models for client segmentation at EY are impressive! Your approach to utilizing diverse client variables and advanced techniques shows a deep understanding of the potential impact on consulting services. I agree that incorporating real-time data streams and exploring internal applications could further enhance your strategy. Keep up the great work!

Reply

○

(https:/

Chris Cosmas (*He/Him*) (https://classroom.emeritus.org/courses/9054/users/226607)

Apr 24, 2024

One of the most important activities performed in a previous organization I worked with was putting different investment products into categories. This would allow asset managers to gauge their success compared to other managers, it would also allow the company to compare its returns compared to a peer group. This would make their investment vehicles more visible and can make them more attractive to investors looking to invest in a particular category. These categories take into account many different financial parameters such as asset allocation, portfolio weights, investment style, and risk of the investments. There are many different broad categories which are each divided further into more specific categories.

Let us look at a set of categories for the sake of simplicity:

- U.S. Equity
- Europe Equity
- China Equity
- Japan Equity
- MENA Equity

For a fund to identify as an equity fund it must have more than 70% of its portfolio invested in stocks.

The most fitting of the two models presented would be the Agglomerative Hierarchical Clustering. The data to be fed into the model would be the following two variables: The % of the portfolio invested in stocks, and the geographical location in which the fund invests the most.

The different regions will have to be translated into numerical values:

- U.S. stocks = 100
- European stocks = 200
- China stocks = 300
- Japan stocks = 400
- MENA stocks = 500

The model might create more clusters than mentioned as there might be other regions in which the funds invest in and the equity % might be lower than the required 70%.

This will ensure that funds (observations) are grouped with other funds that share a high level of similarity, making the categories more reliable and homogenous so that comparisons are made with similar funds to not mislead investors or managers to compare performance with funds that might be operating in a very different manner.

Reply

○

(https:/

Priscilla Annor-Gyamfi (https://classroom.emeritus.org/courses/9054/users/226376)

Apr 25, 2024

Business Use Case:

I am a marketing manager for a retail chain with stores across different regions in the country. I want to have a better understanding of our customer base and tailor marketing strategies to specific demographics and trends in shopping behaviors. I decided to use the clustering algorithm on our customer data.

Our input data includes customer attributes such as age, gender, location, shopping frequency, purchase history, preferred product categories, average transaction value, and engagement with marketing channels (e.g. Our social media pages). I included additional data on seasonal trends (sales per month or year), promotional campaigns, and store locations across the regions.

Using this dataset enables a thorough analysis of customer segments based on demographics, shopping behaviors, and preferences. Through clustering customers who share common characteristics, the model uncovers unique segments within the customer base, facilitating precise marketing opportunities/strategies.

The clustering model aims to generate clusters representing different customer segments such as "Young Urban Professionals", "Family Shoppers" and "Budget conscious Seniors " considering factors like their "spending score or average spending amount, shopping frequency, purchase history and engagement in promotional activities. Each cluster would help identify the unique shopping behaviors and preferences of these groups of customers.

There would be three clusters created for this model namely the "Young Urban Professionals", "Family Shoppers" and "Budget conscious Seniors ". This is because the aim for this analysis is to better understand customer demographics and their respective shopping behaviors as far as their age group and other characteristics are concerned.

Clustering customers will help my organization to personalize marketing messages, promotions, and product assortments to resonate with each segment's preferences. By targeting specific customer segments more effectively, the company can enhance customer engagement, loyalty, and ultimately, sales revenue. Additionally, insights from clustering analysis can inform decisions regarding store layouts, inventory management, and expansion strategies.

Inference from Results:

Upon analyzing clustering results, below are some insights we derived:

1. **Targeted Marketing Strategies:** Tailoring promotional offers, advertisements, and loyalty programs to cater to the unique needs and preferences of each customer segment.
2. **Product Assortment:** Optimizing product selection and placement within our stores across the regions to align with the preferences of different customer segments.
3. **Customer Lifetime Value:** Identifying high-value customer segments for prioritized retention efforts and personalized service.
4. **Market Expansion:** Identifying untapped market segments or geographic areas for potential expansion or targeted marketing campaigns.

Reply

(https:/

Todd Engle (https://classroom.emeritus.org/courses/9054/users/228910)

Apr 29, 2024

At the Credit Union, I was contracting with, I was helping them build the Commercial side of the Credit Union, they currently only offered retail side products. At first, the executives planned the commercial side to be an online service only. They just recently hired an executive to head up deposits, and he has been pushing to open physical branches, which took other executives back. Eventually, he justified his idea and gained support. Now, the question is, where are the best places to locate these commercial branches?

To determine optimal branch locations for the credit union, I would need to analyze various data points. Below are the three data points I would start with:

- **Small Business Demographics**: Industry type, number of employees, location, years in operation, annual revenue. This helps identify potential small business clients and their financial needs.
- **Financial Data**: Existing loan information (if any), business bank statements, credit scores, and tax returns. Analyzing financial health allows for responsible lending decisions.
- **Credit Union Offerings**: Existing loan products, interest rates, loan terms, eligibility criteria. This ensures the model considers the credit union's capacity to serve different business needs.

I would aim to identify distinct small business segments based on their industry, size, financial health, and location. The expected output would be:

- **Clusters representing small business segments:** Restaurants, retail shops, professional services etc. Each segment might have typical loan requirements and growth trajectories.
- **Targeted loan products:** Developing loan products with specific features (interest rates, terms, collateral requirements) tailored to the needs of each cluster.

As I learned in this module, the optimal number of clusters would depend on the data and the level of granularity desired. The optimal number of clusters again depends on the data's granularity. Here's a possible approach: Start with broad industry clusters (e.g., restaurants, retail, professional services). Analyze each cluster's financial data and location to identify sub-clusters based on factors such as size (revenue, employees) and creditworthiness. This would help the credit union Develop targeted loan products**.** They can now create loan packages with features (e.g., lower interest rates, shorter terms) that cater to the specific financial needs of each segment (e.g., short-term working capital loans for restaurants, and equipment financing for professional services). This would also help the Credit Union streamline its loan application process, helping the Loan Officer pre-qualify loan applications based on cluster membership, and potentially offering faster approvals for businesses with strong financial profiles within specific segments.

By analyzing the clusters, the credit union can extrapolate:

- **Market potential for small business loans:** Identify the size and location of high-potential small business segments in the credit union's service area.
- **Loan product development:** Tailor loan products based on the dominant cluster in each market segment, increasing the credit union's competitiveness in attracting new business clients.
- **Targeted marketing:** Focus marketing efforts on reaching businesses within specific high-potential clusters, potentially through industry-specific events or partnerships.

I would hope this approach would allow the credit union to leverage data-driven insights to become a more attractive option for small businesses seeking loans and credit services.

Reply

○

(https:/

Mhelissa Yayalar (https://classroom.emeritus.org/courses/9054/users/233590)

Apr 30, 2024

An example that I think about related to my company's Services business and their Maintenance, Repair, and Overhaul (MRO) planning services - where my organization helps a large global airline transform its maintenance and engineering division into a profit-seeking MRO business.

In the MRO industry, our organization helps by resolving the following challenges for customers:

- Lack of formal technical services agreement: Prices, responsibilities, and processes were not clearly defined.

- Industry-standard organization structure: They needed guidance on setting up an airline MRO structure.

- Culture transformation: Shifting from an internal division mindset to that of a third-party business.

The input data for this MRO business case would include various factors related to the airline's maintenance and engineering division that include some of the following relevant data sources:

- Historical Maintenance Costs: Information on past maintenance expenses for different aircraft types.

- Performance Metrics: Metrics related to aircraft availability, turnaround time, and reliability.

- Market Data: Information about competitors, market demand, and industry trends.

For clusters or groups that make sense for the MRO business, I would consider the following:

- Aircraft Types: Grouping similar aircraft models together.

- Maintenance Complexity: Clusters based on the complexity of maintenance tasks.

- Geographical Regions: Clusters based on the airline's operational regions (e.g., Europe, Africa, Asia).

- Service Offerings: Clusters related to specific services offered (e.g., engine maintenance, avionics, airframe).

- Cost Categories: Clusters based on cost components (e.g., labor, spare parts, facilities).

Given the diverse business of MRO, the specific objectives will be focus on the optimizing its service offerings, and therefore, we can develop clusters based on different types of maintenance services. For instance, one cluster for heavy maintenance and another for line maintenance. The benefits of these different clustering allows targeted process improvements. For instance, optimizing turnaround time for specific aircraft types. Also, addresses the challenges, such as with high maintenance costs that may need specialized skills or additional investments.

Reference: MRO Business Planning Services Case Study | Boeing Services. https://services.boeing.com/resources/case-studies/mro-business-planning.

Reply

○

(https:/

Swati Sharma (https://classroom.emeritus.org/courses/9054/users/236938)

Apr 30, 2024

Hello Team : In my past i was working in a e-commerce company, and we were always struggling with our marketing efforts(how much is too much) after learning clustering, in a e-commerce company i would understand the customers and there shopping behaviors. To do this, I'would deep dive into our customer data. We will take everything from who's buying what to where they're from and how often they shop. It's like peering into the minds of our customers, getting to know them completely.

I want to group our customers into different "tribes" based on similarities. By doing this, we can tailor our messages and offerings to each tribe, making sure we're speaking their language and offering what they really want. while doing this there can be a-lot of tribes in the digital marketplace. to group them properly I would play around with different groupings, testing the waters to see which ones make the most sense.

Once we've nailed down our tribes, the possibilities are endless. We can create personalized marketing campaigns that resonate with each group, making them feel like we really "get" them. Plus, we can identify our VIP customers—the ones who shop regularly from our business and make sure they feel extra special.

In the end, it's all about building stronger connections with our customers and making their shopping experience feel like a breeze. And with the power of data clustering, we're one step closer making this a reality.

Edited by Swati Sharma (https://classroom.emeritus.org/courses/9054/users/236938) on Apr 30 at 10:03pm

Reply

(http

Isabella Tockman (https://classroom.emeritus.org/courses/9054/users/207395)

May 13, 2024

Your approach of categorizing customers into tribes and communicating with them in a way they understand is spot on! This strategy will help strengthen your relationship with customers, making them feel valued and increasing their loyalty.

Reply

(https:/

Koffi Henri Charles Koffi (https://classroom.emeritus.org/courses/9054/users/208039)

May 1, 2024

Team Fit

In my organization to assigned a newly join employee to a team

The employee meet with around 3 to 8 team where the employee ask different questions to the project manager and product owner ,

Different set of question is asked during the interview with project manager such as

1. The technologies stack ?
2. The programming language used to build the application?
3. The team size
4. How the manager evaluate the employee performance
5. The challenging aspect of the project etc…

After the response provided the employee select the team that best fit him

Here the organization can design a model by creating different clusters

- Each cluster correspond to a team
- A cluster will have all what it needs by the employee to succeed in the project .

Input Data

- Employee Name
- Employee specialization

- Employee skill set
- Employee future goal
- Employee education

Cluster Output

- Each cluster will be associated with a team , the employee will answer a couple questions and the model will assign an employee to a team (cluster) .

Base of the feature(input data)

Number of cluster

Here since we are trying to assign employee to a team the number of cluster K is the number of team available in the company

Reply

○

(https://

Timothy Andrew Ramkissoon (https://classroom.emeritus.org/courses/9054/users/226697)

May 1, 2024

I'm an Asset Integrity Manager for an Oil & Gas Company.

Input Data:
The input data for clustering typically consists of features (attributes) that describe the entities you want to group. In our context, this could include data related to offshore assets, equipment, maintenance records, inspection results, and operational parameters. I would decide on the data based on its relevance to our business objectives. For example, if we're interested in optimizing maintenance schedules, relevant features might include asset age, corrosion rates, and historical failure data.

Expected Clusters or Output:
Clustering aims to group similar data points together. The output would be clusters of assets or equipment that share common characteristics.
For instance, we might discover clusters of aging assets with similar corrosion patterns, which could guide targeted inspection and maintenance strategies.

Number of Clusters:
Determining the optimal number of clusters (K) is crucial. We can use techniques like the "elbow method" or silhouette score to find the right K. The choice of K depends on business context. Too few clusters may oversimplify, while too many may lead to redundancy.

For example, if we’re segmenting assets for maintenance planning, we might choose K based on resource allocation and operational efficiency.

Organizational Benefits:
Clustering can provide several benefits including maintenance optimization by grouping similar assets, we can tailor maintenance schedules, prioritize inspections, and allocate resources efficiently. Risk Assessments (e.g., high risk assets prone to corrosion or fatigue); cost reduction for targeted interventions based on clusters can reduce unnecessary maintenance costs; performance benchmarking allow us to compare asset performance within and across groups; root cause analysis can reveal patterns related to failures or anomalies; predictive maintenances can inform predictive models for early fault detection.

Inferences from Results:
Asset Similarities for assets within the same cluster share common characteristics and maintenance priorities for high-risk clusters need more attention. Operational insights may highlight operational inefficiencies or opportunities resulting in the use of segmentation strategies to guide customized strategies for different asset groups.

Reply

○

(https:/

Isabella Tockman (https://classroom.emeritus.org/courses/9054/users/207395)

May 13, 2024

We can think about an auditing firm that wants to implement a clustering model to detect anomalies in financial transactions and identify potential instances of fraud or errors in their clients' financial records.

1. What is your input data, and how did you decide to use the data

The input data includes various financial transaction features such as transaction amount, date, type of transaction, account involved, and other relevant financial metrics. The decision to use this data was based on its relevance to detecting anomalies in financial transactions. These data are collected from the client's financial records, including accounting databases, transaction logs, and financial statements.

2. What are your expected clusters or output?

The expected output of the clustering model is to identify clusters representing different patterns of financial transactions.

Specifically, the model aims to distinguish between normal or typical financial transactions and abnormal or suspicious transactions. This helps in identifying potential instances of fraud, errors, or financial misstatements.

3. How many clusters or groups should your model create, and why? How would this help the organization as a whole? What can you infer from the results?

The model should create clusters based on the client's specific financial data and auditing objectives. The optimal number of clusters can be determined through techniques like the elbow method.

Detecting anomalies in financial transactions helps auditing firms in identifying potential fraud, errors, or financial misstatements, enhancing financial reporting integrity. Detecting problems early helps auditors look deeper into the issue, suggest ways to fix it, and make plans to avoid similar issues in the future.

Reply