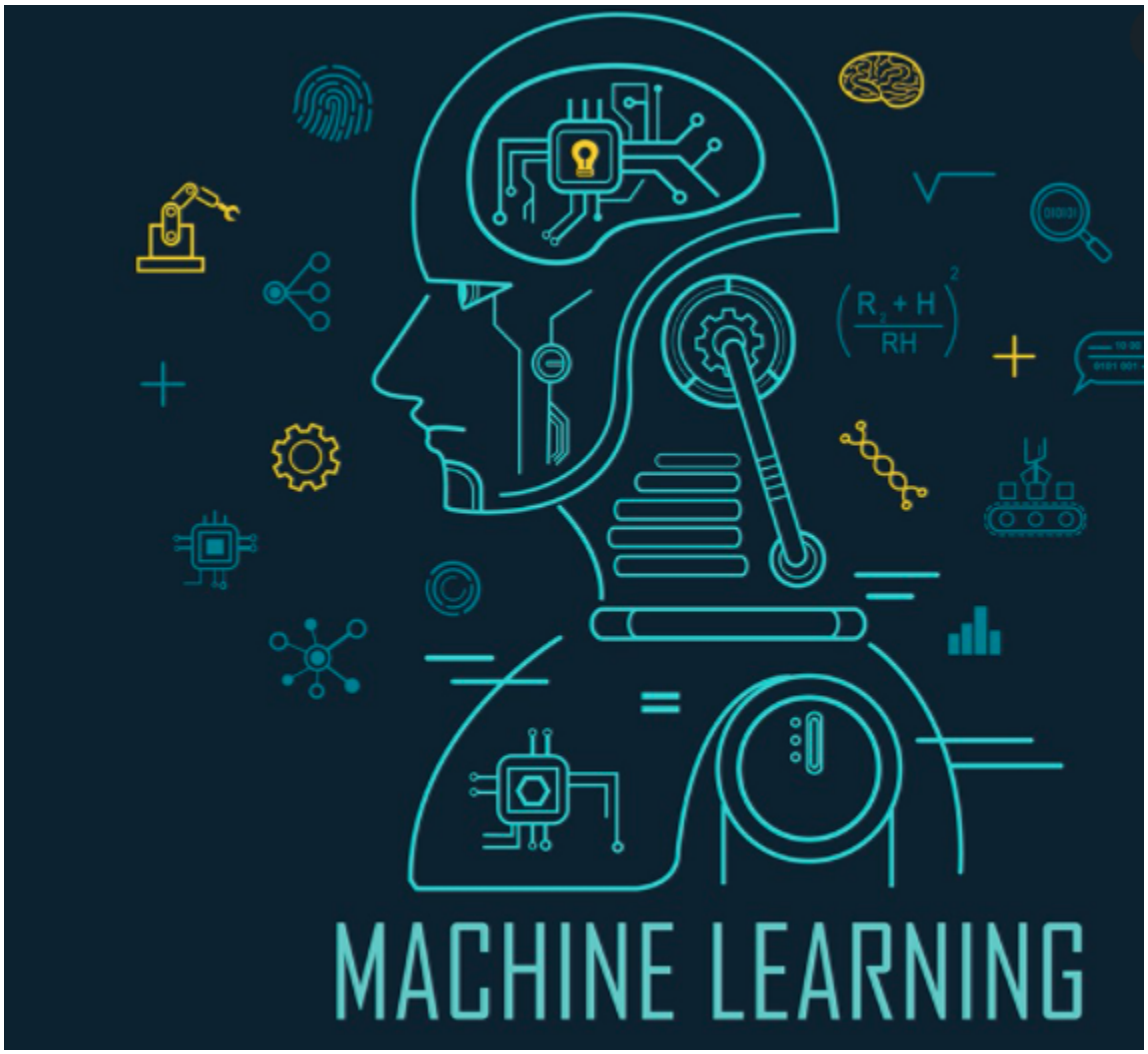


MARKET SEGMENTATION ANALYSIS OF MACHINE LEARNING JOB MARKET IN INDIA

Feynn Labs - Machine Learning Internship Batch - 12 : Team-B Task -3

Contributors: Aman Sai Krishna Bukkapatnam,Tanmay Ture,Aman Sharma,Fazlullah Bokhari .

Problem Statement:

Machine learning (ML) is the study of computer algorithms that can improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so.

Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks. A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory, and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. Some implementations of machine learning use data and neural networks in a way that mimics the working of a biological brain. In its application across business problems, machine learning is also referred to as predictive analytics.

Finding Companies most probable to hire an ML Engineer/Data Analyst Applicant in respect to his/her skillset.

Data Collection/Scraping based on

1. Geography
2. Company's field of work
3. Company size

4. Upcoming vacancies in respect to the company's growth (IPO/Funding etc.)

5. Machine Learning/Data Analysis Skills currently most demanded in the market in respect to i) Experience required, ii) Time required to acquire the skill, iii) Vacancies open iv) Salary, etc.

Analyzing Machine Learning Job Market in India with respect to the given problem statement using Segmentation analysis and outlining the segments most optimal to apply or prepare for Machine Learning Jobs.

Data Collection:

Data collection is defined as the procedure of collecting, measuring, and analyzing accurate insights for research using standard validated techniques. A researcher can evaluate their hypothesis on the basis of collected data. In most cases, data collection is the primary and most important step for research, irrespective of the field of research. The approach of data collection is different for different fields of study, depending on the required information.

Regardless of the field of study or preference for defining data (quantitative or qualitative), accurate data collection is essential to maintain research integrity. The selection of appropriate data collection instruments (existing, modified, or newly developed) and delineated instructions for their correct use reduce the likelihood of errors. A formal data collection process is necessary as it ensures that the data gathered are both defined and accurate. This way, subsequent decisions based on arguments embodied in the findings are made using valid data. The process provides both a baseline from which to measure and in certain cases an indication of what to improve.

According to the problem statement, the dataset should mainly contain Job title, Experience, Skills, Location, etc..

Web scrapping Naukri website.
https://www.kaggle.com/vikasbhadoria/exploring-data-science-jobopportunities-in-india/data

https://www.kaggle.com/lekuid/data-scientist-job-listings-in-india
https://www.kaggle.com/ankitkalauni/predict-the-data-scientists-salary-in-india
https://www.kaggle.com/andrewmvd/data-analyst-jobs
https://www.kaggle.com/halhuynh/it-jobs-dataset

Exploring Data:

Data exploration is the first step of data analysis used to explore and visualize data to uncover insights from the start or identify areas or patterns to dig into more. Using interactive dashboards and point-and-click data exploration, users can better understand the bigger picture and get to insights faster.

This process makes deeper analysis easier because it can help target future searches and begin the process of excluding irrelevant data points and search paths that may turn up no results. More importantly, it helps build a familiarity with the existing information that makes finding better answers much simpler.

Many times, data exploration uses visualization because it creates a more straightforward view of data sets than simply examining thousands of individual numbers or names.

In any data exploration, the manual and automated aspects also look at different sides of the same coin. Manual analysis helps users familiarize themselves with information and can point to broad trends.

Exploring Machine Learning Jobs Dataset,

	Job Title	Company Name	Exp	Location	Skills
0	Data Engineer: Machine Learning	IBM	4-8 Yrs	Bangalore/Bengaluru	deep learning|Interpersonal skills|Time manage...
1	Data Engineer: Machine Learning	IBM	4-6 Yrs	Bengaluru/Bangalore	deep learning|Interpersonal skills|Time manage...
2	Manager - Machine Learning Engineer (Data Sci...	Pylon Management Consulting Pvt Ltd	7-9 Yrs	Remote	Data Science|Machine Learning|Deep Learning|IT...
3	Data Scientist-Python Machine Learning	Jubna	3-5 Yrs	Noida, NCR	mapping|ML algorithms|analyses data|machine le...
4	Senior/Lead Data Scientist - Machine Learning/...	Squareroot Consulting Pvt Ltd.	1-6 Yrs	Bangalore/Bengaluru	Visualization|Exploratory Testing|Machine Lear...

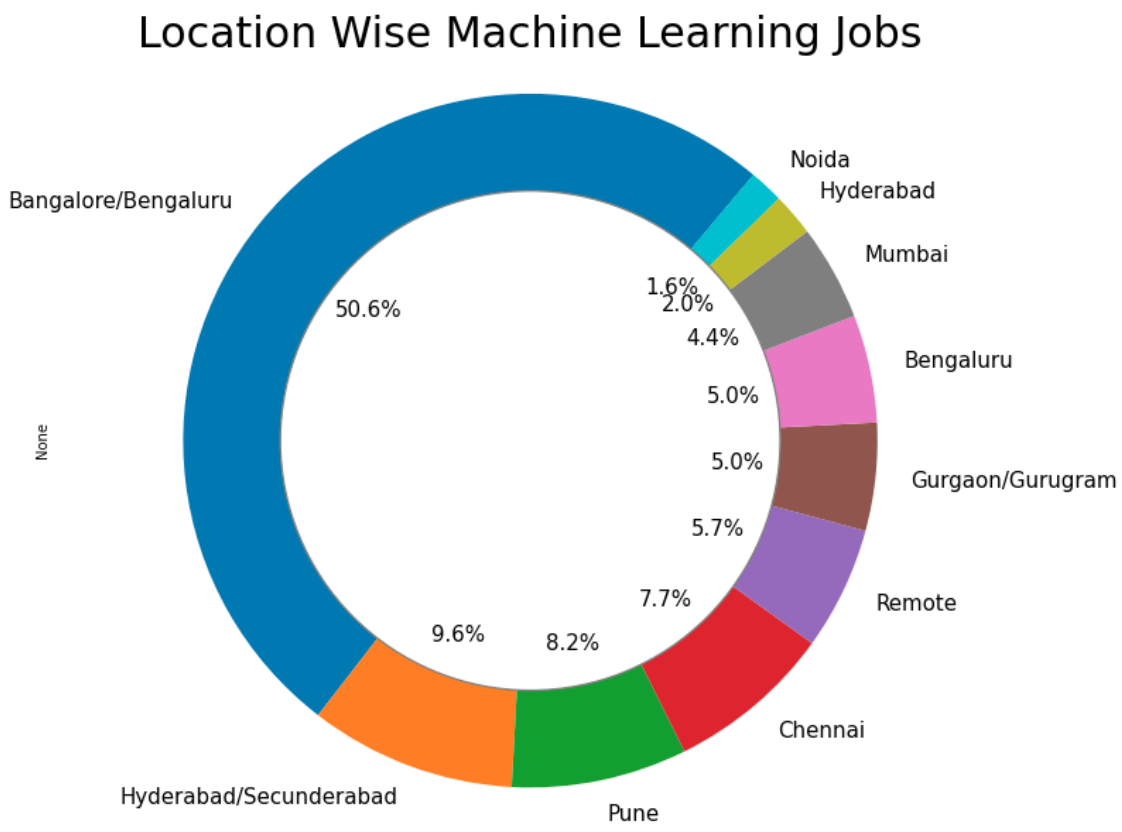
Checking for null values in the dataset

```
: Job Title         0
  Job URL           0
  Company Name      0
  Company URL       0
  Exp               0
  Salary            0
  Location          0
  Skills            0
  Posted          368
  dtype: int64
```

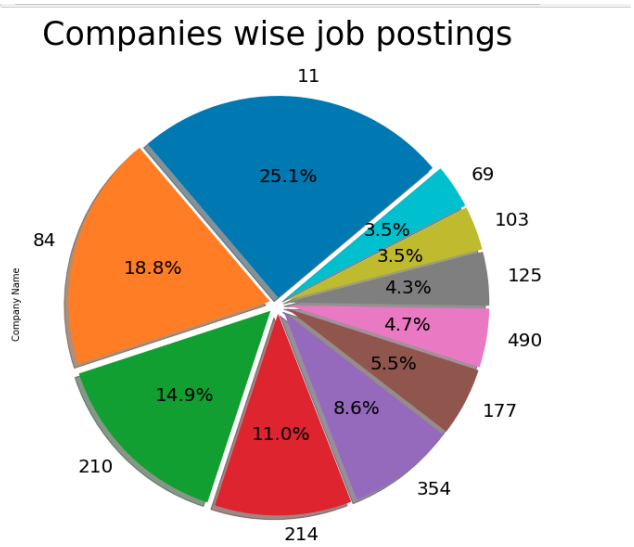
Columns of the dataset

```
Index(['Job Title', 'Job URL', 'Company Name', 'Company URL', 'Exp', 'Salary',
       'Location', 'Skills', 'Posted'],
      dtype='object')
```

Location Wise Machine Learning Jobs:

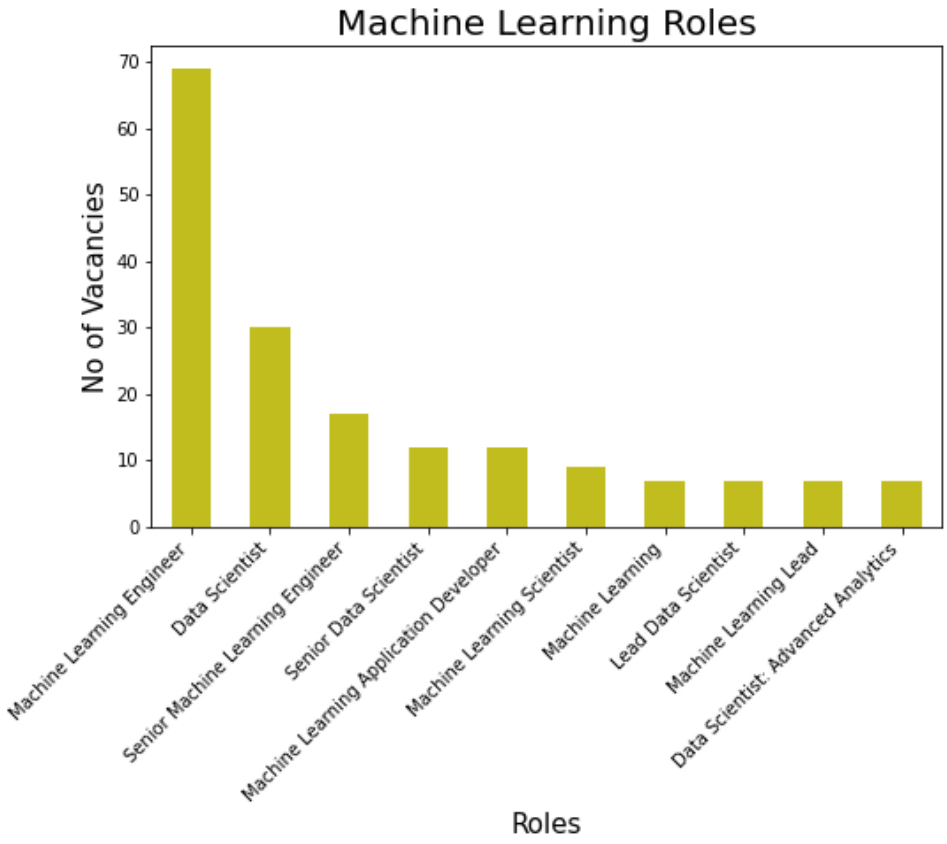


Company-wise Job Postings:

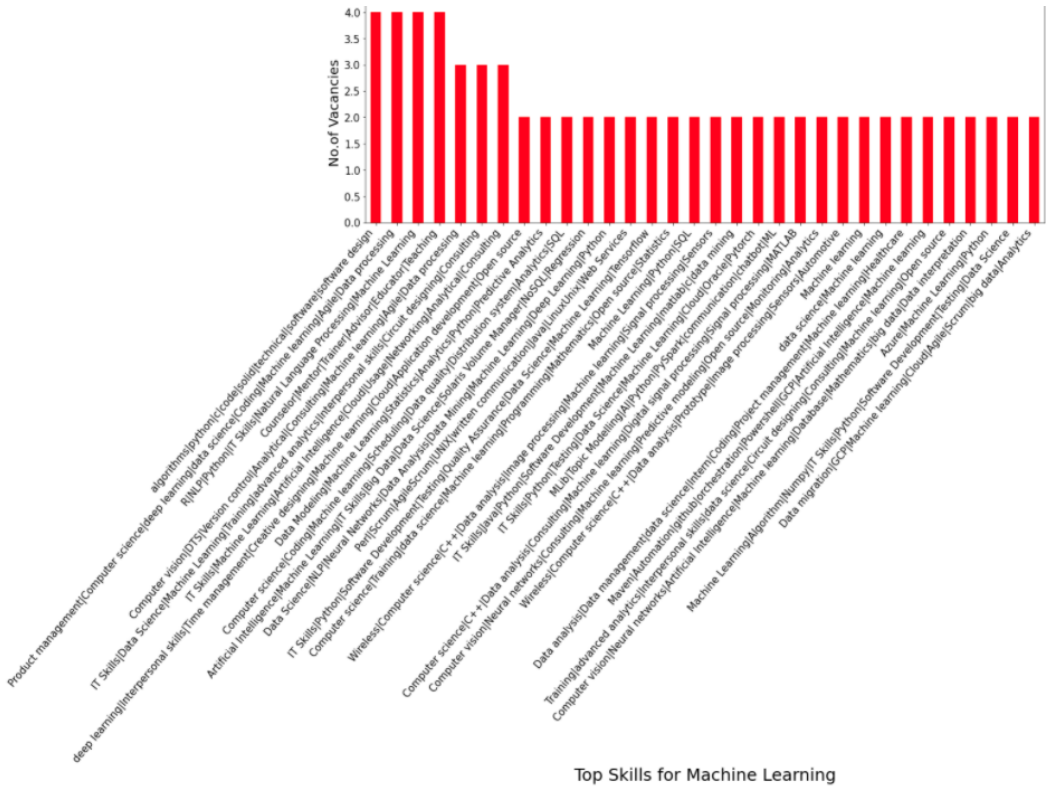


Experience wise No. of Vacancies:

Machine Learning Roles in the Market :



Skills wise plot of Machine Learning jobs in the Market:



Info of the dataset:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 994 entries, 0 to 993
Data columns (total 9 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Job Title     994 non-null    object
 1   Job URL       994 non-null    object
 2   Company Name  994 non-null    object
 3   Company URL   994 non-null    object
 4   Exp           994 non-null    object
 5   Salary        994 non-null    object
 6   Location      994 non-null    object
 7   Skills        994 non-null    object
 8   Posted        626 non-null    object
dtypes: object(9)
memory usage: 70.0+ KB
```

Extracting Segments

Methodologies that can be used:

1.Distance-Based Method:

Most of the time, data available for market segmentation is in the form of tables, where each column represents features about customers and rows represents customers. Numerous approaches to measuring the distance between two vectors exist, several are used routinely in cluster analysis and market segmentation.

1.1 Distance Measures:

The most common distance measures used in market segmentation analysis are a] Euclidean distance: b] Manhattan or absolute distance: Both Euclidean and Manhattan distance treat all dimensions of the data equally. The way this kind of algorithms works is, they find the distance of each vector with all other vectors and then group those which are closer to each other. Usually, these algorithms are used when the data available is small.

1.2 Hierarchical Methods:

Hierarchical clustering methods are the most intuitive way of grouping data because they mimic how a human would approach the task of dividing a set of n observations (consumers) into k groups (segments). Market segmentation analysis occurs between those two extremes. First method is Divisive hierarchical clustering methods start with the complete data set X and splits it into two market segments in a first step. Then, each of the segments is again split into two segments. This process continues until each consumer has their own market segment, Second method is Agglomerative hierarchical clustering approaches the task from the other end. The starting point is each consumer representing their own market segment (n singleton clusters). Step-by-step, the two market segments closest to one another are merged until the complete data set forms one large market segment. Underlying both divisive and agglomerative clustering is a measure of distance between groups of observations (segments). This measure is determined by specifying (1) a distance measure d(x, y) between observations (consumers) x and y, and (2) a linkage method. The linkage method generalises how, given a distance between pairs of observations, distances between groups of observations are obtained. Single linkage: distance between the two closest observations of the two sets.

Complete linkage: distance between the two observations of the two sets that are farthest away from each other.

Average linkage: mean distance between observations of the two sets. Clustering in general, and hierarchical clustering in specific, are exploratory techniques. Different combinations can reveal different features of the data. The result of hierarchical clustering is typically presented as a

dendrogram. A dendrogram is a tree diagram. The root of the tree represents the one-cluster solution where one market segment contains all consumers. The leaves of the tree are the single observations (consumers), and branches inbetween correspond to the hierarchy of market segments formed at each step of the procedure. The height o the branches corresponds to the distance between the clusters. Higher branches point to more distinct market segments. Dendrograms are often recommended as a guide to select the number of market segments. however, dendrograms rarely provide guidance of this nature because the data sets underlying the analysis are not well structured enough.

1.3 Partitioning Methods:

Hierarchical clustering methods are particularly well suited for the analysis of small data sets with up to a few hundred observations. For larger data sets, dendrograms are hard to read, and the matrix of pairwise distances usually does not fit into computer memory. This means that – instead of computing all distances between all pairs of observations in the data set at the beginning of a hierarchical partitioning, cluster analysis using a standard implementation – only distances between each consumer in the data set and the centre of the segments are computed.

A partitioning clustering algorithm aiming to extract five market segments, in contrast, would only have to calculate between 5 and 5000 distances at each step of the iterative or stepwise process (the exact number depends on the algorithm used).

a] k-Means and k-Centroid Clustering:

The most popular partitioning method is k-means clustering.

Let X = {x1,..., xn} be a set of observations (consumers) in a data set.

Partitioning clustering methods divide these consumers into subsets (market segments) such that consumers assigned to the same market segment are as similar to one another as possible, while consumers belonging to different market segments are as dissimilar as possible In addition, the algorithm requires the specification of the number of segments In fact, the choice of the distance measure typically has a bigger impact on the nature of the resulting market segmentation solution than the choice of algorithm .

1.4 Hybrid Approaches:

Several approaches combine hierarchical and partitioning algorithms in an attempt to compensate the weaknesses of one method with the strengths of the other. The strengths of hierarchical cluster algorithms are that the number of market segments to be extracted does not have to be specified in advance, and that similarities of market segments can be visualised using a dendrogram. The biggest disadvantage of hierarchical clustering algorithms is that standard implementations require substantial memory capacity, thus restricting the possible sample size of the data for applying these methods. Also, dendrograms become very difficult to interpret when the sample size is large. The strength of partitioning clustering algorithms is that they have minimal memory requirements during calculation, and are therefore suitable for segmenting large data sets.

The disadvantage of partitioning clustering algorithms is that the number of market segments to be extracted needs to be specified in advance. Partitioning algorithms also do not enable the data

analyst to track changes in segment membership across segmentation solutions with different number of segments because these segmentation solutions are not necessarily nested. The basic idea behind hybrid segmentation approaches is to first run a partitioning algorithm because it can handle data sets of any size. But the partitioning algorithm used initially does not generate the number of segments sought. Rather, a much larger number of segments is extracted. Then, the original data is discarded and only the centres of the resulting segments (centroids, representatives of each. The advantage of using hierarchical clustering in the second step is that the resulting dendrogram may provide clues about the best number of market segments to extract.

Bagged clustering is suitable in the following circumstances (Dolnicar and Leisch 2004; Leisch 1998):
• If we suspect the existence of niche markets.

• If we fear that standard algorithms might get stuck in bad local solutions.

• If we prefer hierarchical clustering, but the data set is too large.

Bagged clustering can identify niche segments because hierarchical clustering captures market niches as small distinct branches in the dendrogram. The increased chance of arriving at a good segmentation solution result from:

(1) drawing many bootstrap samples from the original data set,

(2) repeating the k-means analysis – or any other partitioning algorithm – many times to avoid a suboptimal initialisation (the random choice of initial segment representatives),

(3) using only the centroids resulting from the k-means studies in the second (hierarchical) step of the analysis, and

(4) using the deterministic hierarchical analysis in the final step.

Bagged clustering is an example of a so-called ensemble clustering method (Hornik 2005). These methods are called ensemble methods because they combine several segmentation solutions into one. Ensembles are also referred to as committees.

An additional advantage of bagged clustering – compared to standard partitioning algorithms is that the two-step process effectively has a built-in variable uncertainty analysis.

2.Model-Based Methods:

Distance-based methods have a long history of being used in market segmentation analysis. More recently, model-based methods have been proposed as an alternative.

According to Wedel and Kamakura (2000, p. XIX) – the pioneers of model based methods in market segmentation analysis – mixture methodologies have attracted great interest from applied marketing researchers and consultants. Model-based methods are viewed as one additional segment extraction method available to data analysts. Given that extracting market segments is an exploratory exercise, it is helpful to use a range of extraction methods to determine the most suitable approach for the data at hand. Having model-based methods available is particularly useful

because these methods extract market segments in a very different way, thus genuinely offering an alternative extraction technique.

As opposed to distance-based clustering methods, model-based segment extraction methods do not use similarities or distances to assess which consumers should be assigned to the same market segment. Instead, they are based on the assumption that the true market segmentation solution – which is unknown – has the following two general properties:

(1) each market segment has a certain size, and

(2) if a consumer belongs to market segment A, that consumer will have characteristics which are specific to members of market segment A. Modelbased methods use the empirical data to find those values for segment sizes and segment-specific characteristics that best reflect the data.

Model-based methods can be seen as selecting a general structure, and then finetuning the structure based on the consumer data. the segment sizes π (positive values summing to one), and the segment- specific characteristics θ.

The values that need to be estimated are called parameters. Different statistical frameworks are available for estimating the parameters of the finite mixture model. Maximum likelihood estimation (see for example Casella and Berger 2010) is commonly used. Maximum likelihood estimation aims at determining the parameter values for which the observed data is most likely to occur. i.e., An alternative statistical inference approach is to use the Bayesian framework for estimation. If a Bayesian approach is pursued, mixture models are usually fitted using Markov chain Monte Carlo methods. consumers in the empirical data set can be assigned to segments using the following approach. First, the probability of each consumer to be a member of each segment is determined. This is based on the information available for the consumer, which consists of y, the potentially available x, and the estimated parameter values of the finite mixture model:

The consumers are then assigned to segments using these probabilities by selecting the segment with the highest probability. A standard strategy to select a good number of market segments is to extract finite mixture models with a varying number of segments and compare them. Selecting the correct number of segments is as problematic in modelbased methods as it is to select the correct number of clusters when using partitioning methods.

At first glance, finite mixture models may appear unnecessarily complicated. The advantage of using such models is that they can capture very complex segment characteristics, and can be extended in many different way.

2.1 Finite Mixtures of Distributions:

The simplest case of model-based clustering has no independent variables x, and simply fits a distribution to y. To compare this with distance-based methods, finite mixtures of distributions basically use the same segmentation variables: a number of pieces of information about consumers, such as the activities they engage in when on vacation.

The statistical distribution function f () depends on the measurement level or scale of the segmentation variables y.

2.2 Finite Mixtures of Regressions:

Finite mixtures of distributions are similar to distance-based clustering methods and – in many cases – result in similar solutions. Compared to hierarchical or partitioning clustering methods, mixture models sometimes produce more useful, and sometimes less useful solutions. Finite mixtures of regression models. Finite mixture of regression models assume the existence of a dependent target variable y that can be explained by a set of independent variables x. The functional relationship between the dependent and independent variables is considered different for different market segments.

2.3 Extensions and Variations:

Finite mixture models are more complicated than distance-based methods. The additional complexity makes finite mixture models very flexible. It allows using any statistical model to describe a market segment. As a consequence, finite mixture models can accommodate a wide range of different data characteristics: for metric data we can use mixtures of normal distributions, for binary data we can use mixtures of binary distributions. For nominal variables, we can use mixtures of multinomial distributions or multinomial logit models.

Ordinal variables are tricky because they are susceptible to containing response styles. To address this problem, we can use mixture models disentangling response style effects from content-specific responses while extracting market segments. If the data set contains repeated observations over time, mixture models can cluster the time series, and extract groups of similar consumers. Alternatively, segments can be extracted on the basis of switching behaviour of consumers between groups over time using Markov chains. This family of models is also referred to as dynamic latent change models, and can be used to track changes in brand choice and buying decisions over time. Mixture models also allow to simultaneously include segmentation and descriptor variables. Segmentation variables are used for grouping, and are included in the segment-specific model as usual.

3.Algorithms with Integrated Variable Selection:

Most algorithms focus only on extracting segments from data. These algorithms assume that each of the segmentation variables makes a contribution to determining the segmentation solution. But this is not always the case. Sometimes, segmentation variables were not carefully selected, and contain redundant or noisy variables.

Variable selection for binary data is more challenging because single variables are not informative for clustering, making it impossible to pre-screen or prefilter variables one by one.

A number of algorithms extract segments while – simultaneously – selecting suitable segmentation variables.

3.1 Biclustering Algorithms:

Biclustering simultaneously clusters both consumers and variables. Biclustering algorithms exist for any kind of data, including metric and binary. Several popular biclustering algorithms exist; in particular they differ in how a bicluster is defined. In the simplest case, a bicluster is defined for

binary data as a set of observations with values of 1 for a subset of variables. Each row corresponds to a consumer, each column to a segmentation variable. Biclustering is particularly useful in market segmentation applications with many segmentation variables. Standard market segmentation techniques risk arriving at suboptimal groupings of consumers in such situations. 1. No data transformation. 2. Ability to capture niche markets. Biclustering methods, however, do not group all consumers. Rather, they select groups of similar consumers, and leave ungrouped consumers who do not fit into any of the groups.)

3.2 Variable Selection Procedure for Clustering Binary Data (VSBD):

VSBD method is based on the k-means algorithm as clustering method, and assumes that not all variables available are relevant to obtain a good clustering solution. In particular, the method assumes the presence of masking variables. They need to be identified and removed from the set of segmentation variables.

Removing irrelevant variables helps to identify the correct segment structure, and eases interpretation. The procedure first identifies the best small subset of variables to extract segments. Because the procedure is based on the k-means algorithm, the performance criterion used to assess a specific subset of variables is the withincluster sum-of squares (the sum of squared Euclidean distances between each observation and their segment representative). After having identified this subset, the procedure adds additional variables one by one. The variable added is the one leading to the smallest increase in the within-cluster sum-of-squares criterion. The procedure stops when the increase in within-cluster sum-ofsquares reaches a threshold. The number of segments k has to be specified in advance.

3.3 Variable Reduction: Factor-Cluster Analysis:

The term factor-cluster analysis refers to a two-step procedure of data-driven market segmentation analysis. In the first step, segmentation variables are factor analysed. The raw data, the original segmentation variables, are then discarded. In the second step, the factor scores resulting from the factor analysis are used to extract market segments.

Running factor-cluster analysis to deal with the problem of having too many segmentation variables in view of their sample size lacks conceptual legitimisation and comes at a substantial cost:

1. Factor analysing data leads to a substantial loss of information.

2. Factors-cluster results are more difficult to interpret.

4 Data Structure Analysis:

Extracting market segments is inherently exploratory, irrespective of the extraction algorithm used. Validation in the traditional sense, where a clear optimality criterion is targeted, is therefore not possible. Ideally, validation would mean calculating different segmentation solutions, choosing different segments, targeting them, and then comparing which leads to the most profit, or most success in mission achievement. This is clearly not possible in reality because one organisation cannot run multiple segmentation strategies simultaneously just for the sake of determining which

performs best. The term validation in the context of market segmentation is typically used in the sense of assessing reliability or stability of solutions across repeated calculations Data structure analysis provides valuable insights into the properties of the data. These insights guide subsequent methodological decisions. Most importantly, stability-based data structure analysis provides an indication of whether natural, distinct, and well-separated market segments exist in the data or not. If they do, they can be revealed easily. If they do not, users and data analysts need to explore a large number of alternative solutions to identify the most useful segments for the organisation.

4.1 Cluster Indices:

Because market segmentation analysis is exploratory, data analysts need guidance to make some of the most critical decisions, such as selecting the number of market segments to extract. So-called cluster indices represent the most common approach to obtaining such guidance.

Cluster indices provide insight into particular aspects of the market segmentation solution. Which kind of insight, depends on the nature of the cluster index used. Generally, two groups of cluster indices are distinguished: internal cluster indices and external cluster indices. Internal cluster indices are calculated on the basis of one single market segmentation solution, and use information contained in this segmentation solution to offer guidance.

External cluster indices cannot be computed on the basis of one single market segmentation solution only. Rather, they require another segmentation as additional input.

4.2 Gorge Plots:

A simple method to assess how well segments are separated, is to look at the distances of each consumer to all segment representatives. Similarity values can be visualised using gorge plots.

Each gorge plot contains histograms of the similarity values separately for each segment. The x-axis plots similarity values. The y-axis plots the frequency with which each similarity value occurs high similarity values indicate that a consumer is very close to the centroid (the segment representative) of the market segment. Low similarity values indicate that the consumer is far away from the centroid.

If natural, well-separated market segments are present in the data, we expect the gorge plot to contain many very low and many very high values. This is why this plot is referred to as gorge plot Producing and inspecting a large number of gorge plots is a tedious process, and has the disadvantage of not accounting for randomness in the sample used. These disadvantages are overcome by stability analysis, which can be conducted at the global or segment level.

4.3 Global Stability Analysis:

An alternative approach to data structure analysis that can be used for both distance and model-based segment extraction techniques is based on resampling methods To assess the global stability of any given segmentation solution, several new data sets are generated using resampling methods, and a number of segmentation solutions are extracted.

Resampling methods offer insight into the stability of a market segmentation solution across repeated calculations. Resampling methods – combined with many repeated calculations using the same or different algorithms – provide critical insight into the structure of the data.

Global stability analysis acknowledges that both the sample of consumers, and the algorithm used in data-driven segmentation introduce randomness into the analysis.

Bootstrapping generates a number of new data sets by drawing observations with replacement from the original data. These new data sets can then be used to compute replicate segmentation solutions for different numbers of segments.

4.4 Segment Level Stability Analysis:

The globally best segmentation solution does not necessarily mean that this particular segmentation solution contains the single best market segment. Relying on global stability analysis could lead to selecting a segmentation solution with suitable global stability, but without a single highly stable segment. It is recommendable, therefore, to assess not only global stability of alternative market segmentation solutions, but also segment level stability of market segments contained in those solutions to protect against discarding solutions containing interesting individual segments from being prematurely discarded. After all, most organisations only need one single target segment.

Selecting the target segments:

Target market represents a group of individuals who have similar needs, perceptions and interests. They show inclination towards similar brands and respond equally to market fluctuations. Individuals who think on the same lines and have similar preferences form the target audience.

Target market includes individuals who have almost similar expectations from the organizations or marketers.

To select a target market, it is essential for the organizations to study the following factors:

- Understand the lifestyle of the consumers
- Age group of the individuals
- Income of the consumers
- Spending capacity of the consumers
- Education and Profession of the people
- Gender
- Mentality and thought process of the consumers
- Social Status
- Kind of environment individuals are exposed to

Customising the Marketing Mix

9.1 Implications for Marketing Mix Decisions

Marketing was originally seen as a toolbox to assist in selling products, with marketers mixing the ingredients of the toolbox to achieve the best possible sales results (Dolnicar and Ring 2014). In the early days of marketing, Borden (1964) postulated that marketers have at their disposal 12 ingredients: product planning, packaging, physical handling, distribution channels, pricing, personal selling, branding, display, advertising, promotions, servicing, fact finding and analysis. Many versions of this marketing mix have since been proposed, but most commonly the marketing mix is understood as consisting of the 4Ps: Product, Price, Promotion and Place (McCarthy 1960).

9.2 Product

One of the key decisions an organisation needs to make when developing the product dimension of the marketing mix, is to specify the product in view of customer needs. Often this does not imply designing an entirely new product, but rather modifying an existing one. Other marketing mix decisions that fall under the product dimension are: naming the product, packaging it, offering or not offering warranties, and after sales support services.

9.3 Price

Typical decisions an organisation needs to make when developing the price dimension of the marketing mix include setting the price for a product, and deciding on discounts to be offered

9.4 Place

The key decision relating to the place dimension of the marketing mix is how to distribute the product to the customers. This includes answering questions such as: should the product be made available for purchase online or offline only or both; should the manufacturer sell directly to customers; or should a wholesaler or a retailer or both be used.

9.5 Promotion

Typical promotion decisions that need to be made when designing a marketing mix include: developing an advertising message that will resonate with the target market, and identifying the most effective way of communicating this message. Other tools in the promotion category of the marketing mix include public relations, personal selling, and sponsorship.

Code_Implementation :

Link_1

Implementation on other dataset:

Link_2