INTERVIEW QUESTIONS
For
PREDICTIVE MODELING

Website: www.analytixlabs.co.in
Email: info@analytixlabs.co.in

ANALYTICS TOOLS FOR DATA SCIENCE

Q: What are some of the tools used for statistical analysis?
Ans: Some popular tools for statistical analysis include

- **SAS:** A suite of analytics software developed by SAS
- **R:** An open source language and environment for statistical computing
- **Python:** An open source language and environment for statistical computing
- **Julia:** An open source language and environment for statistical computing
- **Spark:** An open source language and environment for statistical computing
- **WEKA:** A suite of machine-learning free software written in Java
- **SPSS:** A software for statistical analysis, currently owned by IBM
- **EViews:** Mostly used for econometric analysis, a software developed by Quantitative Micro Software
- **Minitab**: A statistical tool developed at Pennsylvania State University

Q: What are some of the visualization tools available in the market today?
Ans: Visualization tools can be divided into three broad categories:
Graphical tools:

MS Excel: Microsoft Excel is the standard offering in the Microsoft Office bundle. It is used mostly by analysts for all lightweight analysis, as well as a visualization tool.
D3.js: A JavaScript library to create graphs in HTML and related web technologies
FusionCharts: A JavaScript library for graphs on the Web
Google Charts: Interactive charts for web and mobile devices
Power BI: Microsoft product

Dashboard tools:

Tableau: A US-based software company with a flagship product that helps create dashboards on raw data
Qlikview: A dashboard software product by the US-based company Qlik
Spotfire: Dash boarding software by TIBCO
OBIEE: By Oracle
Business Objects: By SAP
Cognos: By IBM
MSBI: By Microsoft
Pantaho:
JasperSoft:
Palantir:
Rshiny: As part of R
Bokeh/Dash: As part of Python

Infographic tools:

Infogram
plotly
Picktochart

Data Audit & Data Sanitization

Data validation

- Total number of observations.
- Total number of fields.
- Each field name, Field type, Length of field.
- Format of field, Label.

Basic Checks

- Are all variables as expected (variables names & variable types).
- Are there some variables which are unexpected?
- Are the data types and length a cross variables correct?
- For known variables, is the data type as expected (For example if age is in date format something is suspicious)
- Have labels been provided and are sensible?
- If anything suspicious we can further investigate it and correct accordingly

Data Validation – snapshot of data

Printing the first few observations all fields in the dataset. It helps in better understanding of the Variable by looking at its assigned values.

Check points for data snapshot output:

1. Do we have any unique identifier? Is the unique identifier getting repeated in different records?
2. Do the text variables have meaningful data? (If text variables have absurd data as '&^%*HF' then either the variable is meaningless or the variable has become corrupt or wasn't properly created.)
3. Are there some coded values in the data? (if for a known variable say State we have category codes like 1-52 then we need definition of how they are coded.)
4. Do all the variables appear to have data? (Incase variables are not populated with non missing meaningful value it would show in print. We can further investigate using means statistics.)

Categorical fields and Frequencies

- Calculate frequency counts cross-tabulation frequencies. Especially for categorical, discrete & class fields.
- Frequencies
  - Help us understanding the variable by looking at the values it's taking and data count a teach value.
  - They also helps us in analyzing the relationships between variables by looking at the cross tab frequencies or by looking at association.

Check points for looking frequency table

1. Are values as expected?
2. **Variable understanding:**
   a. Distinct values of a particular variable, missing percentages.
   b. Are there any extreme values or outliers?
   c. Any possibility of creating a new variable having small number of distinct category by clubbing certain categories with others.

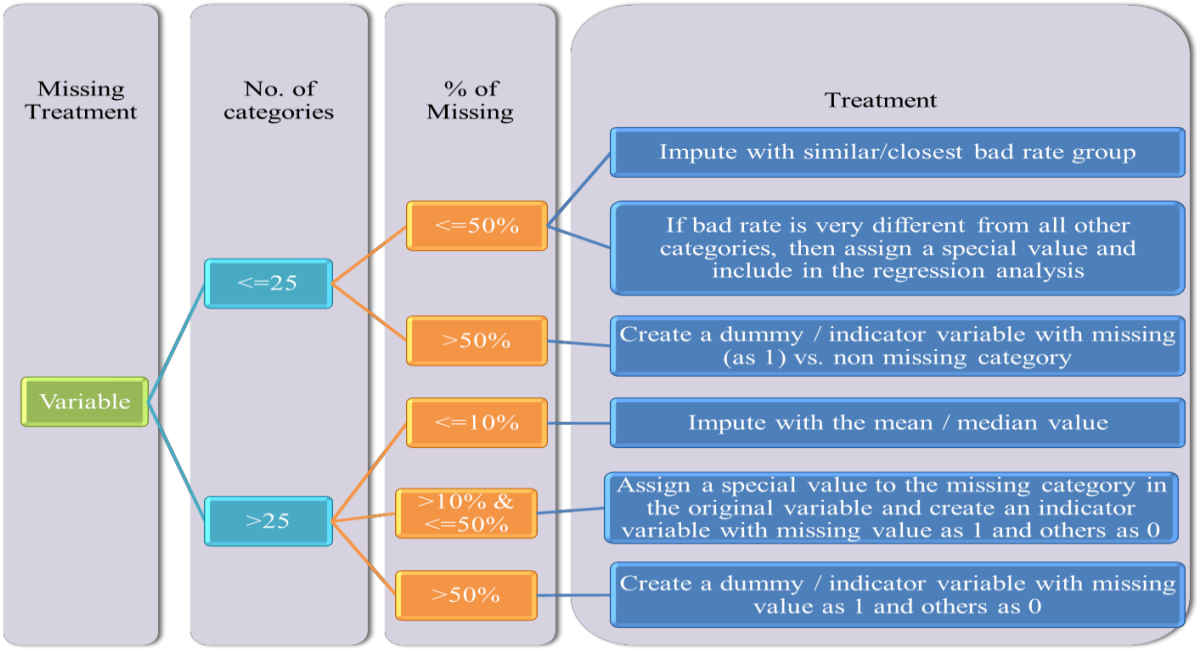
Descriptive statistics for continuous fields

- Distribution of numeric variables by calculating.
  - N–Count of non-missing observations.

- N miss–Count of Missing observations.
- Min, Max, Median,Mean.
- Quartile numbers & percentiles–P1,p5,p10,q1(p25),q3(p75),p90,p99
- Stddev
- Var
- Skewness
- Kurtosis

- **CheckList**
  - Are variable distribution as expected?
  - What is the central tendency of the variable? Mean, Median and Mode across eachvariable
  - Is the concentration of variables as expected? What are quartiles?
  - Indicates variableswhich are variables are with std dev=0; the variableswhich are useless for the current objective.
  - Are there any outliers/extreme values for the variable?
  - Are outlier values as expected or they have abnormally high values-for ex for Age if max and p99 values are10000.Then should investigate if its the default or there is some error in data
  - What is the % of missing value associated with the variable?

Missing values and outlier treatment



REGRESSION PROBLEMS: LINEAR REGRESSION

In this topic, we cover different types of regression models, assumptions and questions related to them, and the estimation method used commonly with regression models.

Q. What is regression analysis?

Ans: Regression analysis is a statistical technique for estimating the relationships among variables. It basically try to measure and identify the cause and effect relationship among the variables. Regression analysis comes indifferent flavors: Logistic, Multiple choice logistic, multinomial, multiple etc.

Q: What is meant by the term "linear regression"?

Ans: Linear regression is a statistical modeling technique that attempts to model the relationship between an explanatory variable and a dependent variable, by fitting the observed data points on a linear equation, e.g., modeling the body mass index (BMI) of individuals by weight.

A linear regression is used if there is a relationship or significant association between the variables. This can be checked by scatter plots. If no association appears between the variables, fitting a linear regression model to the data will not provide a useful model.

A linear regression line takes equations in the following form: **Y = a + bX,**
Where, X = explanatory variable and
Y = dependent variable.
b = slope of the line
a = intercept (the value of y when x = 0).

Q: What are the various assumptions that an analyst takes into account while running a regression analysis?

Ans: Regression analysis depends on the following assumptions:

- The relationship between the variables should be linear (or approximately linear) over the range of population being studied.
- Y variable in the regression analysis should be normal, i.e., should follow the normal curve (exactly or approximately).
- There should be no multicollinearity, i.e., the independent variables should not show correlation among themselves.
- There should be no autocorrelation in the data, i.e., the residuals should be independent of each other.
- The condition of homoscedasticity, i.e., the error terms or residuals along the regression, should be equal.
- Errors normal, Errors iid.

Q: How would you execute regression on Excel?

Ans: Regression on Excel can be performed by using three built-in functions to calculate slope, intercept, and R^2 values or by using the Regression function provided in the Data Analysis toolbar (after installing Analysis ToolPak add-ins). The built-in functions are SLOPE() , INTERCEPT() , and RSQ() .

Q: What is the multiple coefficient of determination or R-squared?

Ans: The multiple coefficient of determination, R^2, is a method by which to calculate the overall effectiveness (in terms of percentage similar to linear regression) of all the independent variables in explaining the dependent variable.
For example, if $R^2 = 0.8$, this means that the independent variables have 80% of the variation in the value of dependent variables.

Unfortunately, R^2 alone may not be a reliable measure of the accuracy of the multiple regression model, as R^2 increases every time a new variable is added in the model, even though the variable might not be statistically significant. If there is a large number of independent variables, the value of R^2 may be high, even though the variables do not explain the dependent variable that well. This problem is called overestimating the regression.

By adjusting the R^2 value for the number of independent variables, the problem of overestimating the regression can be overcome.

Q. What is the difference between R^2and Adjusted R^2?
Ans:

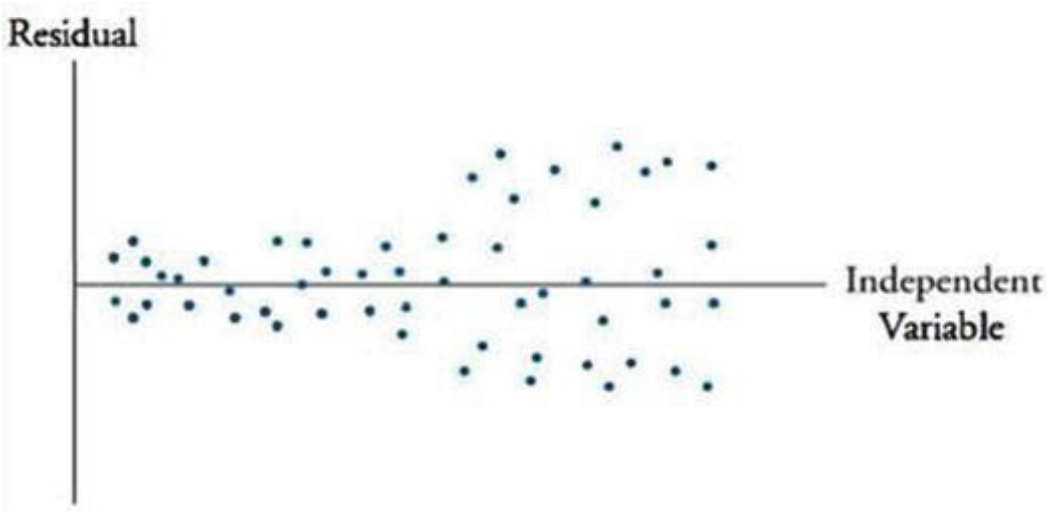
a) R2 is a statistic that will give some information about the goodness of fit of a model. In regression, the R2 coefficient of determination is a statistical measure of how well the regression line approximates all data points. An R2 of 1.0 indicates that the regression line perfectly fits the data.
b) R2 increases whenever we add a new independent variable, Adjusted R2 might increase or decrees based on the variable explanatory power. Too many independent variables can decrease the value of Adj R2
c) Adjusted R2 is a modification of R2 that adjusts for the number of explanatory terms in a model. Unlike R2, the adjusted R2 increases only if the new term improves the model more than would be expected by chance. The adjusted R2 can be negative, and will always be less than or equal to R2. Adjusted R2 is not always better than R2: adjusted R2 will be more useful only if the R2 is calculated based on a sample, not the entire population. For example, if our unit of analysis is a state, and we have data for all counties, then adjusted R2 will not yield any more useful information than R2

Q. Why do we minimize squares of deviations (OLSE – Ordinary lease square estimator) why cannot we use absolute differences?
Ans: It is hard to deal with absolute differences when you are differentiating and integrating.

Q: What is meant by "heteroscedasticity"?
Ans: When the variance of the residuals differs across observations in the sample, this is called heteroscedasticity. It is one of the errors in regression analysis that analysts have to test before running the regression analysis. One of the assumptions of multiple regression is that the variance of the residuals is constant across observations.

Q: How do you differentiate between conditional and unconditional heteroscedasticity?
Ans: Unconditional heteroscedasticity occurs in cases in which the level of independent variables does not affect heteroscedasticity, i.e., it doesn't change systematically with changes in the value of independent variables. Although this is a defilement of the equal variance assumption, it frequently causes no serious problems with the regression.
Conditional heteroscedasticity is heteroscedasticity that is related to the level of (i.e., conditional upon) the independent variables.

Q: What are the different methods of detecting heteroscedasticity?
Ans: There are two methods of detecting heteroscedasticity: examining scatter plots of the residuals, and using the Breusch-Pagan chi-square test. Plotting the residuals against one or more of the independent variables can help us spot trends among the observations.



The residual plot in the figure indicates the presence of conditional heteroscedasticity. Notice how the variation in the regression residuals increases as the independent variable increases. This indicates that the variance of the dependent variable about the mean is related to the level of the independent variable.

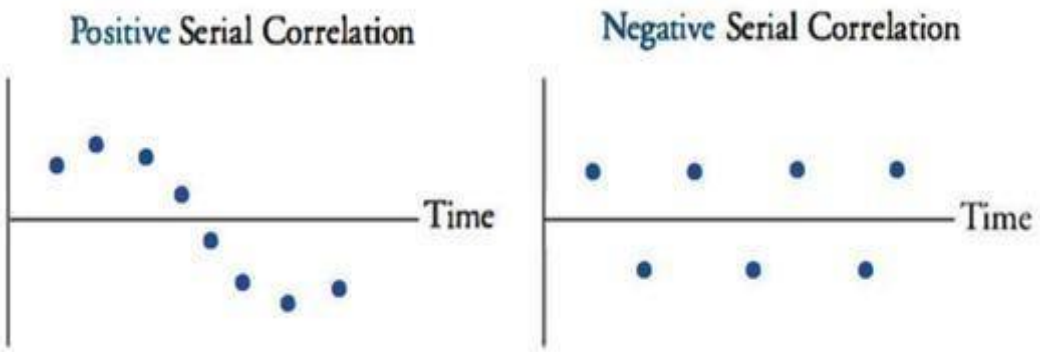
The more common way to detect conditional heteroscedasticity is the Breusch-Pagan test, which calls for the regression of the squared residuals on the independent variables. Independent variables contribute significantly in explaining squared residuals in case of conditional heteroscedasticity.

Q: What are the different methods to correct heteroscedasticity?
Ans: The most common remedy is to calculate robust standard errors. The t-statistics is recalculated using the original regression coefficients and the robust standard errors. A second method to correct for heteroscedasticity is to use generalized least squares, by modifying the original equation.

Q: What is meant by the term "serial correlation"?
Ans: Serial correlation, or autocorrelation, is the phenomenon commonly observed in time series data, in which there is a correlation between the residual terms. It is of two types: positive and negative.
When a positive regression error in one time period increases the probability of observing a positive regression for the next time period, this is a positive serial correlation. In a negative serial correlation, the positive regression error causes the probability of observing a negative error to increase.

Q: What are the different methods to detect serial correlation?
Ans: There are two methods that are commonly used to detect the presence of serial correlation: residual plots and the Durbin-Watson statistic.

A scatter plot of residuals vs. time, can reveal the presence of serial correlation.



Scatter plot of residuals vs. time indicating positive and negative serial correlations
The more common method is to use the Durbin-Watson statistic (DW) to detect the presence of serial correlation.

Q: What are the different methods to correct multicollinearity?

Ans: The most common method to remove multicollinearity is to omit independent variables having a high correlation with the variable set. Unfortunately, it is not always an easy task to identify the variable(s) that are the source of the multicollinearity. There are statistical procedures that may help in this effort, such as stepwise regression, which systematically removes variables until multicollinearity is reduced.

A summary of violations of the assumptions of multiple regression is offered in Table.

	Conditional Heteroscedasticity	**Serial Correlation**	**Multicollinearity**
What is it?	Residual variance related to level of independent variables	Residuals are correlated	High correlation among two or more independent variables
Effect?	Coefficients are consistent. Standard errors are underestimated. Too many Type I errors	Coefficients are consistent. Standard errors are underestimated. Too many Type I errors (positive correlation)	Coefficients are consistent (but unreliable). Standard errors are overestimated. Too many Type II errors
Detection?	Breusch-Pagan chi-square test	Durbin-Watson test	Conflicting t and F statistics; correlations among independent variables if k = 2
Correction?	Use White-corrected standard errors	Use the Hansen method to adjust standard errors.	Drop one of the correlated variables.

Q. What is MLE (maximum likelihood estimator?)
Ans: Used in estimating statistical parameters, It assumes a (NO) distribution of the parameter and maximize sits joint probability distribution, estimate is obtained at the point where probability distribution of parameter is maximum.

Starts with arbitrary values of the regression coefficients and constructs an initial model for predicting the observed data. Then evaluates errors in such prediction and changes the regression coefficients so as make the likelihood of the observed data greater under the new model. Repeats until the model converges, meaning the differences between the newest model and the previous model are trivial.

It assumes the distribution of the variable under consideration and finds out the parameters of that distribution by maximizing the likelihood function (JDF).

Q. When do you go for generalize d linear models (GLM)?
Ans: When above assumptions fail……. eg: errors are not normal; OR When you have discrete independent variable eg: yes/no; 1 or 0

Q. How do you find parameters of GLM (OLSE or MLE)?
Ans: MLE

Q. What is Multicollinearity?
Ans:

1) Multicollinearity refers to a situation in which two or more explanatory variables in a multiple regression model are highly correlated. We have perfect multicollinearity if the correlation between two independent variables is equal to 1 or-1. In practice, we rarely face perfect multicollinearity in a dataset. More commonly, the issue of multicollinearity arises when there is a high degree of correlation (either positive or negative) between two or more in depend invariables.
2) Multicollinearity is a statistical phenomenon in which two or more predictor variables in a multiple regression model are highly correlated. In this situation the coefficient estimates may change erratically in response to small changes in the model or the data. Multicollinearity does not reduce the predictive power or reliability of the model as a whole; it only affects calculations regarding individual predictors. That is, a multiple regression model with correlated predictors can indicate how well the entire bundle of predictors predicts the outcome variable, but it may not give valid results about any individual predictor, or about which predictors are redundant with others.

Q. How do you detect Multicollinearity and how will you remove it?
Ans: Multicollinearity is Interdependency of independent (predictor variables).

a) High Variance inflation factor (VIF)
b) High F but low t values
c) Conditional index (CI)

Use PCA or FA, Drop problematic variables, Ridge regression

Q. What is variance inflation factor (VIF)?
Ans: Each of the predictor variables is regressed upon other predict or vars. If that R-squared is high then this variable has co linearity with others.

Q. Why do we check for multi co-linearity in nonlinear regression models?
Ans: Model is nonlinear i.e: the regression coefficients are nonlinear not predictor variables, so there is a chance of co linearity relation between the predictor vars.

Q. What procedure do you use to fit regression model in SAS?
Ans: Procreg, proc glm

Q. What is the procedure for non-linear regression in SAS?
Ans: Proc logistic, Proc genmod.

Q. What is log-linear model?
Ans: Use log as link function instead of logit, which is used for poisson response variable (y=0,1,2,3,…)

Q. How do you find good ness fit of your model in GLM?
Ans:

a) It's not R-square, here it is Chi-square.
b) Percent Correct Predictions
c) Hosmer and Lemeshow Goodness-of-Fit Test
d) ROC curves
e) Somers'D
f) Gamma
g) Tau-a
h) C
i) More than a dozen "R2"-type summaries.

Q. Why linear regression is called linear?
Ans: Because the output variable is modeled as a **linear** function of the input variables. The case of one explanatory variable is **called** simple **linear regression**. For more than one explanatory variable, the process is **called** multiple **linear regression**.

Q. What is Ks test?
Ans: In statistics, the Kolmogorov–Smirnov **test** (**KS test**) is a non parametric **test** of the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample KS **test**), or to compare two samples (two-sample KS **test**).

LINEAR REGRESSION (Questions without answers)

1. Can you define regression in layman terms?
2. Why do you require regression?
3. Can you explain what different situations you have been used linear regression are?
4. Explain the major steps in linear regression model building?
5. What are the basic assumptions of linear regression?
6. What are the consequences when assumptions failed?
7. How do you test assumptions?
8. How do you define dependent variables & Independent variables?
9. How do you sample the data for validation?
10. What is the global hypothesis in linear regression? How do you interpret it from the output?
11. How do you interpret your linear regression SAS output? What are the different tables you look into the output to finalize the model?
12. What is importance of constant in the equation?
13. What is VIF? And Condition Index? How do you calculate VIF/Condition Index?
14. What are standardized Betas? How these are different from normal betas?
15. How do you test goodness of Fit?
16. What is R-square/Adjusted Square? How do you calculate R-Square & Adjusted R-Square?
17. What is t-value and p-value and their significance with respect to linear regression?
18. How do you validate linear models? What are the statistics you look for validation of models?
19. What is decile analysis? What is the importance of decile analysis in linear regression models?

CLASSIFICATION TECHNIQUES

A classification technique plays an important role in the whole of analytics-based decision making. This is probably the most important group of techniques to be learned by an analytics professional. Also, given the breadth and depth of these techniques and their vast usage, you are bound to get many questions on this subject in a job interview. So, arm yourself with some basic concepts.

Further, I will delve into some analytics tools, such as R, SAS, and Tableau, which will surely come up in an interview. Besides, having some basic knowledge of databases, SQL, and big data will help you showcase an all-around knowledge of this subject.

I will also briefly touch upon big data, although it is not within the scope of this book.

Q: What is understood by "classification"?
Ans: Classification is the grouping of a data set, based on some predefined criteria. The criteria are usually based on some historic information, and classification tries to classify the data set, based on information received from that historic criteria.

An example: A company wants to have a database of 1 million customers in the United States, including their demographic information. It wants to identify the top 50,000 customers who have highest propensity to respond to an offer campaign.

The company's analyst retrieves past data on response rates for a similar campaign on 200,000 customers. Their response rate is trained on a classification technique that tries to separate respondents with non respondents and also create a scorecard for the customers. The model is then executed on a 1-million-customer base, to classify respondents from non respondents and pick the top 50,000 respondents who should be sent the new campaign.

Other examples include

- Google identifying whether a mail is spam, based on its content and other information
- Assessing whether an employee would attrite, based on his/her past information

Q: Can you name some popular classification methodologies?
Ans: There are numerous classification techniques today. This is probably the most widely studied area and encompasses techniques that are so vast and differentiated from one another that the topic itself is mammoth in proportion.
Some of the more widely known techniques are

- Logistic regression
- Neural network
- Decision tree
- Random forest
- Discriminant analysis

Q: Briefly, what is understood by "logistic regression"?
Ans: Logistic regression is the technique of finding relationships between a set of input variables and an output variable (just like any regression), but the output variable, in this case, would be a binary outcome (think of 0/1 or yes/no).
For example: Will there be a traffic jam in a certain location in Bangalore? is a binary variable. The output is a categorical yes or no.

The probability of occurrence of a traffic jam can be dependent on such attributes as weather conditions, day of the week and month, time of day, number of vehicles, etc. Using logistic regression, we can find the best-fitting model that explains the relationship between independent attributes and traffic jam occurrence rates and predict the probability of jam occurrence.

Q: What is an odds ratio?

Ans: Odds is the relative occurrence of different outcomes, expressed as a ratio of the form a:b. For example, if the odds of an event are said to be 5:2 in favor of the first outcome, this means that the first outcome occurs five times for the second outcome to occur twice. Odds are related to probability and can be shown mathematically as follows:

Odds = a:b

Probability = a/(a+b)

Probability = Odds/(1 + Odds)

Odds = Probability/(1 - Probability)

Q: How is linear regression different from logistic regression?

Ans: Linear regression is applicable on numerical or continuous variables, but logistic regression is applicable when the dependent variable is categorical (a commonly dichotomous variable). The output of logistic regression is between 0 and 1, where 1 denotes “success” and 0 denotes “failure.” But in linear regression, the output is continuous, which can assume any range of value.

Linear regression predicts numerical outputs, such as sales or profit, whereas logistic regression predicts dichotomous output, such as yes and no or living and dead.

Q. What is logistic regression, when do u use it?

Ans:

a) When basic assumption of regression fail,
b) When there is binary (categorical) response.

Q. Give an example where you can use logistic regression?

Ans:

a). Response/Non response
b). Good customer, Bad customer……all binary cases

Q: Can you give a brief overview of decision trees?

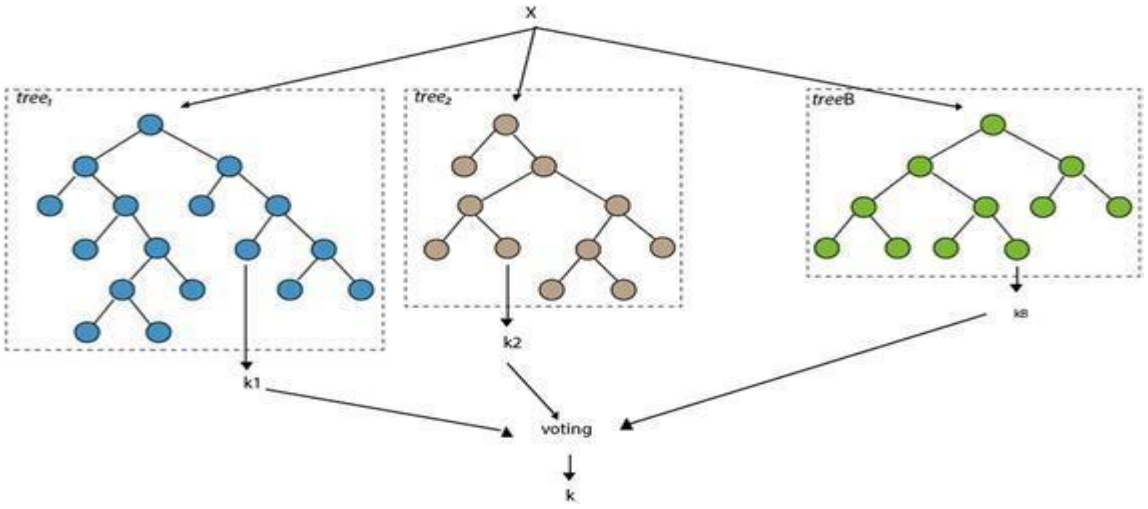
Ans: Decision trees, as the name suggests, are tree-shaped visual representations by which one can reach a particular decision, by laying down all options and their probability of occurrence. Decision trees are extremely easy to understand and interpret. At each node of the tree, one can interpret what the consequence of selecting that node or option will be.

Q: Can you explain briefly the random forest method of classification?

Ans: Random forest is currently the most accurate of all classification techniques available. Random forest is an ensemble method that works on the principle that many weak learners can come together to make a strong prediction. In this case, the weak learner is a simple decision tree, and random forest is strong learner.

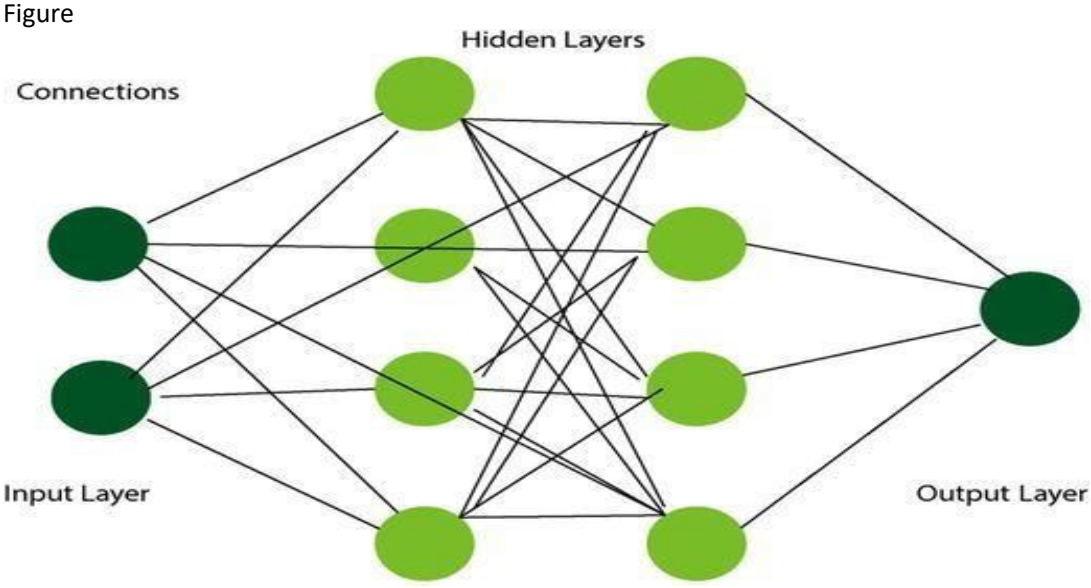
Random forest optimizes the output from many decision trees formed from samples of the same data set. In general, the higher the number of trees, the better the accuracy of the resulting random forest ensemble will be. Yet, at higher numbers, the gain in accuracy decreases. So, the analyst has to decide on the number of trees, based on the cost of implementation that he/she will face with higher numbers of trees.

The trees are combined according to a voting mechanism. The voting is based on the success criteria of each tree. The best results are gained using a weighted approach, wherein the votes are weighted, based on the accuracy of individual trees.



Q: What is understood by "neural network"?
Ans: Neural network (also known as artificial neural network) is inspired by the human nervous system: how complex information is absorbed and processed by the system. Just as with humans, neural networks learn by example and are configured to a specific application.

Neural networks are used to find patterns in complex data, and thus can forecast and classify data points. Neural networks are normally organized in layers. Layers are made up of a number of interconnected nodes. Patterns are presented to the network via the input layer, which communicates to one or more hidden layers, in which the actual processing is done. The hidden layers then link to an output layer, where the answer is output, as shown in Figure



Q: How is neural network different from conventional computing?
Ans: Conventional computing comprises predefined instructions that form the building blocks of its processing system. Neural networks, on the other hand, do not have predefined steps to processing a system. Rather, they learn from past experiences to chart their own steps in processing.

Q: Can you explain discriminant analysis, in brief?
Ans: Discriminant analysis –based classification works according to the concept of analysis of variance (ANOVA), which is to test whether there is a significant difference between the mean of two or more groups with respect to a particular variable. If the mean of a variable is significantly different in different groups, it can safely be said that this variable classifies the data set into groups.
To extend this concept, MANOVA, or multivariate analysis of variance, can be executed to classify a data set based on multiple variables.

Q. What is purpose of discriminant analysis?
Ans: The main purpose of a discriminate function analysis is to predict group membership based on a linear combination of the interval variables. The procedure begins with a set of observations where both group membership and the values of the interval variables are known. The end result of the procedure is a model that allows prediction of group membership when only the interval variables are known. A second purpose of discriminate function analysis is an understanding of the dataset, as a careful examination of the prediction model that results from the procedure can give insight into the relationship between group membership and the variables used to predict group membership.

Q: How would you assess the performance of a classification model?
Ans: The performance of a classification model is assessed by a table called a confusion matrix. It is based on the count of records that are accurately predicted vs. counts of records incorrectly predicted.
Following table is a confusion matrix for a two-class problem.

		Predicted values	
		Class 1	**Class 2**
Actual Values	Class 1	A	B
	Class 2	C	D

Here, A is the number of records of Class 1 that are correctly predicted to be Class 1.B is the count of records of Class 1 that are incorrectly predicted to be Class 2. So, total correct predictions are A+D. Total incorrect predictions are B+C.
Accuracy of a model = Total correct predictions/Total records = A+D/A+B+C+D.
Error Rate of a model = Total incorrect predictions/Total records = B+C/A+B+C+D.
A robust classification model aims to increase the accuracy rate or decrease the error rate of a prediction.

Q. What is: lift, KPI, robustness, model fitting, design of experiments, 80/20 rule?
Ans:

Lift:
It's measure of performance of a targeting model (or a rule) at predicting or classifying cases as having an enhanced response (with respect to the population as a whole), measured against a random choice targeting model. Lift is simply: target response/average response.
Suppose a population has an average response rate of 5% (mailing for instance). A certain model (or rule) has identified a segment with a response rate of 20%, then lift=20/5=4
Typically, the modeler seeks to divide the population into quantiles, and rank the quantiles by lift. He can then consider each quantile, and by weighing the predicted response rate against the cost, he can decide to market that quantile or not.
"if we use the probability scores on customers, we can get 60% of the total responders we'd get mailing randomly by only mailing the top 30% of the scored customers".

KPI:
- Key performance indicator
- A type of performance measurement
- Examples: 0 defects, 10/10 customer satisfaction
- Relies upon a good understanding of what is important to the organization

More examples:
Marketing & Sales:
- New customers acquisition
- Customer attrition
- Revenue (turnover) generated by segments of the customer population
- Often done with a data management platform

IT operations:
- Mean time between failure
- Mean time to repair

Robustness:
- Statistics with good performance even if the underlying distribution is not normal
- Statistics that are not affected by outliers
- A learning algorithm that can reduce the chance of fitting noise is called robust
- Median is a robust measure of central tendency, while mean is not
- Median absolute deviation is also more robust than the standard deviation

Model fitting:
- How well a statistical model fits a set of observations
- Examples: AIC, R^2, Kolmogorov-Smirnov test, Chi^2, deviance (glm)

Design of experiments:
The design of any task that aims to describe or explain the variation of information under conditions that are hypothesized to reflect the variation.
In its simplest form, an experiment aims at predicting the outcome by changing the preconditions, the predictors.
- Selection of the suitable predictors and outcomes
- Delivery of the experiment under statistically optimal conditions
- Randomization
- **Blocking:** an experiment may be conducted with the same equipment to avoid any unwanted variations in the

input
- **Replication:** performing the same combination run more than once, in order to get an estimate for the amount of random error that could be part of the process
- **Interaction:** when an experiment has 3 or more variables, the situation in which the interaction of two variables on a third is not additive

80/20 rule:
- Pareto principle
- 80% of the effects come from 20% of the causes
- 80% of your sales come from 20% of your clients
- 80% of a company complaints come from 20% of its customers

CLASSIFICATION PROBLEMS (Questions without answers):

1. What was the problem statement? Why were you creating the attrition model? What technique has been used? Initial questions were judge knowledge about requirements understanding.
2. Can you explain few situations where you used Logistic regression?
3. Can you explain few situations where you can use both linear & Logistic regression?
4. Why did you choose logistic regression for this problem? Why can't linear regression model work in its place?
5. At what level did you build the model? (like customer level, branch level etc)
6. What was the hypothesis and how did you check whether you can accept the hypothesis?
7. Explain the steps used in modelling?
8. How did you define model target variable? Explain what steps, like for problem definition how did you arrive at attrition definition? What was the strategy you used etc. How much time did each step take?
9. What do you mean by observation period, performance period and lag period? Why is a lag period needed?
10. What is the dependent variable you used in your project? How did you define it?
11. What are your independent variables? How did you define them?
12. How many variables you have used?
13. Have you created any derived variables? What are they?
14. Have you transformed the variables? What situations you have used transformations? Why do you require variable transformation?
15. Explain the steps involved in building logistic regression?
16. Explain the steps data preparation?
17. Explain the steps in variable reduction?
18. Which variable reduction technique did you use and what all are there?
19. Explain the model building & validation?
20. How do you sample the data for training, validation and out of time validation?
21. What is the equation of a logistic regression model?
22. How did you check correlation between your variables? What is multiple regression?
23. What are p values and why are they important in determining whether a variable is important for the model or not, explain the concept of p value in non-technical way.
24. What are the parameters that determine whether a logistic model is good or not?
25. If the model is not validating in out of time validation. What are the possible reasons for that?
26. What is misclassification matrix?
27. What is concordance/Discordance? How do you calculate?
28. What is somer's D and how it is different from Gini?
29. What is p-value and type-1 and type-2 error with respect to logistic regression modelling?
30. What is AIC/BIC? How these will be helpful in model building?
31. What is boot strapping and use?
32. What is Hosmer Lemoshov test and what is the use?
33. What is multicollinerity? What are consequences with multicollinerity in logistic regression?
34. What is ROC? What is use of it?
35. What is sensitivity &specificity? What is the use?
36. What is the difference between in-time validation and out of time validation?
37. What is over fitting & consequence of it?
38. How do you implement the model/Scoring the population or database?

MODEL BUILDING

1. Checking Avaibility of observations
2. Descriptive Statistics
3. Outlier Treatment
4. Missing value Treatment
5. Create New variable
6. Variable Analysis and Reduction
7. Splitting the data set into development and validation
8. Model building
9. Analysis

Checking Avaibility of observations:

- Total number of observations
- Number of available observations
- Percentage of Available observation
- Number of missing observations
- Percentage of missing observations
- Number of positive values
- Number of negative Values
- Number of observations with zero value

Descriptive Statistics

- Minimum value
- Maximum value
- Mean
- Median
- Standard deviation
- Skewness
- Kurtosis

Outlier Treatment

How to Detect Outliers

- Dot plot or Scatter plot
- Box plot

What Should We Do About Them?

- Transformation
- Deletion
- Winsorised Mean
- Trim med Mean

Missing value Treatment

- Dropping variables (when more than70 % of the data are missing, if the variable is very important we will go up to 60%)
- List wise/Case wise deletion

- Business Ratios and using ranges to cap/delete observation
- Nominal variables: Treat missing data as just another category
- Substituted (plugged in) values, i.e. (Single)Imputation
  1. Mean
  2. Subgroup Mean
  3. Median
  4. Subgroup Median
  5. A regression estimate
- Maximum Likelihood Estimation and Multiple Imputation

Variable Analysis and Reduction

- Chi-squared statistic
- Information value
- Spearman rank order correlation co efficient
- Clustering techniques
- Multicollinearity check
  1. Correlation Matrix(R value must lie between-0.4and 0.4,ideally)
  2. Tolerance(shouldbemorethan0.4,ideally)
  3. Variance Inflation Factor(should belessthan2.5,ideally)

Splitting the Data set into Development and Validation Sample

Development Sample (80%ofthe full file) Validation sample (20%ofthefull file)

*Situation may change if the number of observation in the full dataset is small

Regression Modelling

1. Transformation of variables (both dependent and independent variables, if required)
2. Transforming back to get original values(if dependent variable is transformed)
3. Create dummy variables for categorical variables

Checklist

- F-test (the relevant p-value must be less than 0.05).
- R-square and Adjusted R^2(should not be less than0.5,themore the better)
- Tolerance (should be more than 0.4,ideally)
- VIF (shouldbelessthan2.5,ideally)
- T-test for final model variables(the relevant p-value must be lessthan0.05)

Logistic Modelling

Transformation of variables (independent variables, if required) Create dummy variables for categorical variables

Checklist

- Concordant (should be between 65% and 85%)
- Discordant (should be between 15% and 35%)
- KS should not be beyond 3rd decile
- Rank ordering of bads
- Model must hold in k-cross validation samples
- No flipping of signs across samples
- Beta coeff should be close in dev and val samples

- Validation KS should be less than Dev KS by max 2%
- Variables should have logical trends
- "Somer's D, Gamma, Tau-a, C (Should vary between 0 and 1, with larger values corresponding to stronger association between predicted and observed values)
- Hosmer-Lemeshow Goodness of Fit Test(Large values indicates lack of fit of the model)

Decile Analysis

- Rank ordering
- KS Statistic (should be between 40 and 70)
- Lift Curve
- Divergence

VALIDATION CHECK

1. Check whether the same variables come in the model significantly(p-value)
2. Check whether the same variables come in the model with same sign
3. KS Statistic (should be between 40 and 70)
4. Lift Curve
5. Divergence

Information Value:

IV=woe*(%good-%bad)

- **<0.02 Not predictive**
  0.02-0.1 Weak
  0.1-0.3 Medium
  >0.3 Strong

Note: Variable with IV> 0.5 should be investigated for over predicting.

a. Weight of Evidence

Another consideration when defining the groups during formatting is the Weight of Evidence(WOE).The WOE measures the strength of each attribute or grouped attributes, in separating the 'good' and 'bad' accounts. It is a measure of the difference between the proportions of 'goods' and 'bads' in each attribute. The WOE is based on the log of odds calculation:

i.e. **WOE =ln(%good/%bad)**

FACTOR ANALYSIS

Q. Why do we do Factor analysis?

Ans: To reduce the number of variables for a better presentation of the key factors.

Q: What is understood by dimension reduction techniques?

Ans: Dimension (variable) reduction techniques aim to reduce the data set with higher dimension to one of lower dimension, without the loss of feature of information that is conveyed by the data set. The dimension here can be conceived as the number of variables that a data set contains.

With the advent of big data and the ability to process and store large amounts of data, organizations today try to store as much data as possible. This leads to an increase in the attributes that are stored. Think of it as a database table that increases not just in rows but also in terms of columns (variables).

For data scientists creating appropriate models, not all variables are relevant. In addition, large multicollinearity, which diminishes model performance, may be encountered. Variable reduction techniques help weed out this issue. Also, the model is much more crisp in terms of being able to be understood and explained and is less costly (uses less computational resources).

Dimension reduction techniques are almost always executed as a precursor to another technique, such as regression. It is a way to speed up model-building without compromising on the potential of a model.

Q: What are some commonly used variable reduction techniques?

Ans: Two commonly used variable reduction techniques are:

- Principal component analysis (PCA)
- Factor analysis

Q: Can you provide a brief overview of principal component analysis?

Ans: The crux of PCA lies in measuring the data from the perspective of a principal component. A principal component of a data set is the direction with largest variance. A PCA analysis involves rotating the axis of each variable to the highest eigenvector/eigenvalue pair and defining the principal components, i.e., the highest variance axis or, in other words, the direction that most defines the data. Principal components are uncorrelated and orthogonal.

Q: Can you provide a brief overview of factor analysis?

Ans: The key concept behind factor analysis is the presence of a latent variable that stores much of the information of a set of variables in a data set. For example, a group of respondents can answer questions relating to income, education, and spending similarly, because they are in the same socioeconomic category.
In factor analysis, we define factors that are the same in number as the number of variables in a data set. Each factor captures a certain amount of variance in each variable. The eigenvalue is the measure of how much variance of observed variables is captured by a factor.

All factors are sorted in their descending order of value. The factors with low value are discarded, and top factors are retained as factors that explain most variance in the observed variance. It is helpful to know the number of factors in advance.

Q. What is use of factor analysis?

Ans: Factor analysis is used to uncover the latent structure (dimensions) of a set of variables. It reduces attribute space from large number of variables to smaller number of factors and as such is a "non-dependent" procedure

(that is, it does not assume a dependent variable is specified). Factor analysis could be used for any of the following purposes:

- To reduce a large number of variables to a smaller number of factors for modeling purposes, where the large number of variables precludes modeling all the measures individually. As such, factor analysis is integrated in structural equation modeling (SEM), helping confirm the latent variables modeled by SEM. However, factor analysis can be and is often used on a stand-alone basis for similar purposes.
- To establish that multiple tests measure the same factor, thereby giving justification for administering fewer tests. Factor analysis originated a century ago with Charles Spearman's attempts to show that a wide variety of mental tests could be explained by a single underlying intelligence factor (a notion now rejected, by the way)
- To validate a scale or index by demonstrating that it's constituent items load on the same factor, and to drop proposed scale items which cross-load on more than one factor.
- To select a subset of variables from a larger set, based on which original variables have the highest correlations with the principal component factors.
- To create a set of factors to be treated as uncorrelated variables as one approach to handling multicollinearity in such procedures as multiple regression
- To identify clusters of cases and/or outliers.
- To determine network groups by determining which sets of people cluster together (using Q-mode factor analysis, discussed below)

Q. What are the different types of rotation in Factor loading?
Ans:

a. Varimax rotation is an orthogonal rotation of the factor axes to maximize the variance of the squared loadings of a factor (column) on all the variables (rows) in a factor matrix, which has the effect of differentiating the original variables by extracted factor. Each factor will tend to have either large or small loadings of any particular variable. A varimax solution yields results which make it as easy as possible to identify each variable with a single factor. This is the most common rotation option.

b. Quartimax rotation is an orthogonal alternative which minimizes the number of factors needed to explain each variable. This type of rotation often generates a general factor on which most variables are loaded to a high or medium degree. Such a factor structure is usually not helpful to the research purpose.

c. Equimax rotation is a compromise between Varimax and Quartimax criteria.

d. Direct oblimin rotation is the standard method when one wishes a non-orthogonal (oblique) solution–that is, one in which the factors are allowed to be co-related. This will result in higher eigen values but diminished interpretability of the factors. See below.

e. Promax rotation is an alternative non-orthogonal (oblique) rotation method which is computationally faster than the direct oblimin method and therefore is sometimes used for very large datasets.

Q: What is factor loading?
Ans: Each factor in a data set defines the latent variable, in other words, the underlying variable that defines a set of variables in a data set. Factor loading describes the relationship or association between each variable and each factor. Higher association indicates that the factor can be used to describe that variable.
An example: While analyzing 70 variables that affect customer churn, a factor analysis was run, and because 70 variables were used for this analysis, the algorithm gave 100 factors. On observing the factor loading, it is found that demographic variables have high loading for one specific factor. So, these variables are combined into one factor, and so on, for other variables.
Factor loading is an important parameter to assess factor-variable dependence.

SEGMENTATION

Q: What is Segmentation?
Ans: Segmentation is the process of dividing a data set into clearly differentiated groups, relevant to a particular business. Companies segment their customers into different clusters to decide how to create differentiated strategies for each cluster and to maximize the value of the business.

But segmentation is not just used for customer analytics; it can be used for myriad other solutions. We might want to segment geographical areas based on their population density, or employees on their propensity to attrite.

Segmentation algorithms (also known as clustering algorithms) are very common, and the chances are extremely high that you will be tested on these concepts in an interview. In this chapter, I will go through some common interview questions related to segmentation and clustering.

Q: What are supervised and unsupervised learning algorithms? How are they different from each other?
Ans: Supervised and unsupervised learning algorithms are the two broad classifications for all statistical algorithms. The major difference between the two is how outputs to a model are defined. Keeping this segregation in mind helps an analyst to better choose which kind of problem-solving is best suited to a situation.

In supervised learning, model defines the cause and effect of inputs on given outputs. In other words, the inputs define what we are looking for in a model. So, in supervised learning models, we focus the model on existing relationships between inputs and outputs, to define and predict the unknown.
For example, in all classification techniques, we know from historic data what the different categories in dependent variables are. The goal of these techniques is not to come up with categories but to define them and how they are dependent on independent variables.

In unsupervised learning, the output of the model is not defined. The unsupervised learning models are used to define what our output looks like. Clustering techniques are a classic example of unsupervised learning. Here, we do not have the clusters (output) beforehand; rather, the model comes out with the clusters based on the input and criteria.

Q: Can you give an example to differentiate between supervised and unsupervised learning algorithms?
Ans: The example I'll use here is face recognition.

- Supervised learning: Learning, from examples, what a face is, in terms of structure, color, etc., so that after several iterations, the algorithm can define a face.
- Unsupervised learning: Because the example provided does not yield a desired output, categorization is undertaken, so that the algorithm differentiates correctly between the face of a horse, cat, or human (clustering of data).

Q: What are some of the supervised and unsupervised algorithms?
Ans: All classification algorithms fall under the supervised category. Following is a list of a few classification techniques:

- Naïve Bayes
- Support vector machine
- Ensemble Learning (Random forest, Bagging, Boosting)
- Decision tree
- Logistic regression
- K-Nearest Neighbors

Etc.

All Regression algorithms fall under the supervised category. Following is a list of a few regression techniques:

- Linear regression
- Decision trees (Regression)
- Ensemble learning (Random Forest, Bagging, Boosting)
- K-Nearest Neighbors

Etc.

All segmentation algorithms and variable reduction algorithms (such as those in the following list) fall under the unsupervised category.

- K-means
- Fuzzy clustering
- Hierarchical clustering
- DBSCAN
- Spectral Clustering
- Factor analysis
- Association Analysis (Market Basket Analysis)

Q. Tell us something about cluster analysis?

Ans: Cluster analysis or clustering is the assignment of a set of observations into subsets (called clusters) so that observations in the same cluster are similar in some sense. Clustering is a method.

- To validate a scale or index by demonstrating that its constituent items load on the same factor, and to drop proposed scale items which cross-load on more than one factor.
- To select a subset of variables from a larger set, based on which original variables have the highest correlations with the principal component factors.
- `To create a set of factors to be treated as un correlated variables as one approach to handling multi collinearity in such procedures as multiple regression
- To identify clusters of cases and/or outliers.
- To determine network groups by determining which sets of people cluster together (using Q-mode factor analysis, discussed below)

Q: How is clustering defined?

Ans: Clustering (or segmentation) is a kind of unsupervised learning algorithm in which a data set is grouped into unique, differentiated clusters.

Let's say we have customer data spanning 1,000 rows. Using clustering, we can group the customers into differentiated clusters or segments, based on the variables. In the case of customers' data, the variables can be demographic information or purchasing behavior.

Clustering is an unsupervised learning algorithm, because the output is unknown to the analyst. We do not train the algorithm on any past input-output information, but let the algorithm define the output for us. Therefore (just like any other modeling exercise), there is no right solution to a clustering algorithm; rather, the best solution is based on business usability.

Q: What are the two basic types of clustering methods?

Ans: There are two basic types of clustering techniques:

- Hierarchical clustering
- Partitional clustering

Q: How is hierarchical clustering defined?
Ans: Hierarchical clustering attempts to either merge smaller clusters into larger ones or break larger clusters into smaller ones. The basic rule at the core of this technique is deciding how two small clusters are merged or which large cluster is split. The final outcome of the algorithm is a tree of clusters called a dendrogram, which displays how the clusters are related. By splitting the dendrogram at a chosen level, a clustering of the data set into separate groups is achieved.

As shown in Figure, the dendrogram is split at various levels to come up with the required number of clusters.

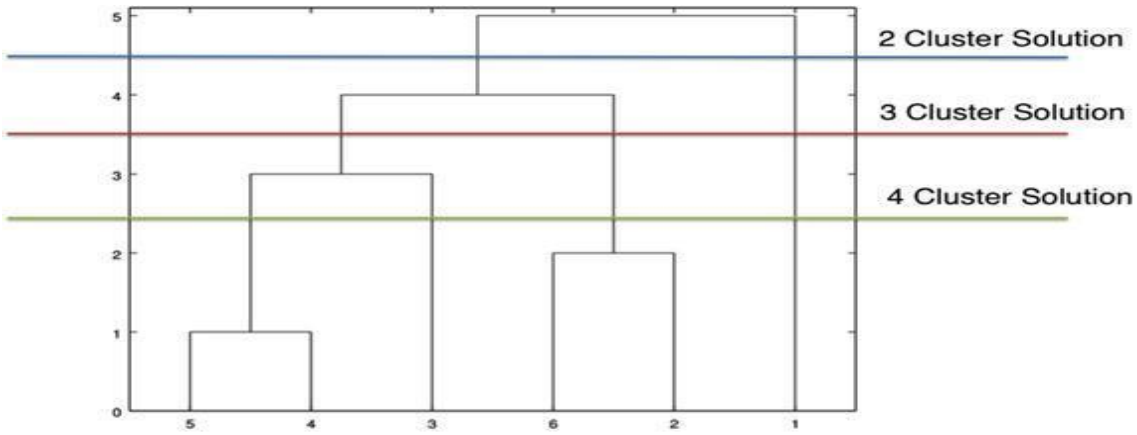


Figure: Dendrogram and cluster solutions

Q: How is partitional clustering defined?
Ans: Partitional clustering works to directly decompose the data set into a set of differentiated clusters. The core rule here is to minimize some measure of dissimilarity in the samples within each cluster, while maximizing the dissimilarity of different clusters.
For example: At most times, researchers try to reduce the within-cluster variance and increase variance between the cluster. A good measure is to take the ratio of the 2 measure and maximize it to ascertain the right number of clusters.

Q: What is meant by "exclusive clustering"?
Ans: This is the most common type of clustering, in which each object or data point belongs exclusively to only one cluster. This is also the most desired form of clustering, in most cases. For example, it would be necessary for a customer to be part of only one segmentation group, so that a unique, dedicated, and exclusive marketing effort could be formulated as part of a campaign.

Q: What is non-exclusive or overlapping clustering?
Ans: Often, if not always, an object can be part of more than one cluster. These are mostly borderline objects, in which we define the boundaries of clusters to overlap each other.
An example is demographic clustering, in which students can be part of both a student cluster and a high-spender cluster, which would be rare.

Q: What is the concept behind fuzzy clustering?
Ans: Rather than an object being part of clusters only (one-to-one mapping), an object can be part of all clusters, with varying degrees of membership. We call this type of clustering fuzzy clustering.

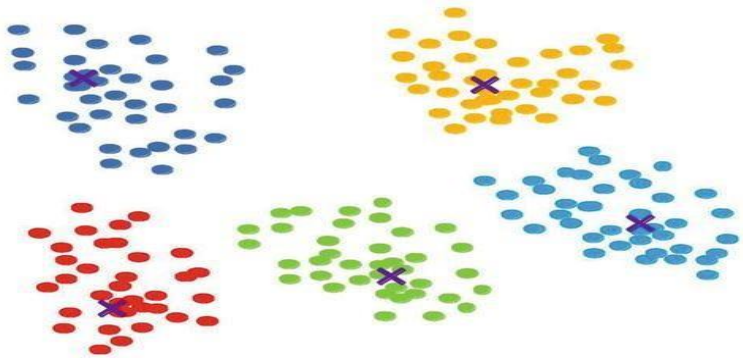
Each object is given a score (between 0 and 1) that depicts the degree to which an object is part of a specific cluster. An example is the cluster of employees on the basis of their skill sets. An employee can possess all skill sets but exhibit varying degrees of competency in each.

Q: Can you differentiate between a complete vs. a partial clustering?
Ans: In a complete clustering, all objects in the data sets are forced to be part of a cluster. Even when there are outliers in the data sets, they are definitely attached to a cluster or are clusters in themselves.
On the other hand, in a partial cluster, all data points are not necessarily part of a cluster. An example of this can be employees with different skill sets. An employee can be part of more than one cluster of skills.

Q: What is meant by "k-means clustering"?
Ans: K-means is the most widely used clustering algorithm in industry. The reason for its popularity is based on the fact that it's both easy to execute and understand.

At the crux of it, k-means clustering identifies random means centers in data sets and attributes cluster membership around those means.

As shown in Figure, random points are determined which later form the cluster centers.



Q: What is the basic algorithm of k-means clustering, in layperson's terms?
Ans: We first choose the k random means from our data sets. K is the number of clusters that we would like to finally extract from our data set.

Now, each data point is attached to each mean that we have chosen, based on the proximity of that data point to the mean. The group of all these data points with their respective mean will form a cluster.
We then re compute the mean for each computed cluster. The preceding steps are rerun until the recomputed means of all clusters are correctly determined.

Q: What is the proximity measure that you take in k-means clustering?
Ans: There are various proximity measures that can be employed. Euclidean distance is one such measure that is heavily used. It is simply the ordinary distance between two points. Other measures include Taxicab metric, Manhattan distance, and Jaccard measure.

Q: What differentiates a k-means from a k-median? When would you use k-median as opposed to k-means?
Ans: A k-median employs median as the centroid metric, as opposed to the means used in a k-means technique. The basic reason someone would use a k-median rather than a k-means is generally the same as that for using a median. To a large extent, the presence of outliers tends to skew the centroid of a data set. An analyst should be able to judge whether outliers are true representatives of a data set. If they are, then using k-means would be preferred; otherwise, a k-median is used.

Q: What are some of the limitations of the k-means clustering technique?
Ans: The biggest limitation with the k-means technique is inherent in the way it is calculated. The user is required to know beforehand the number of clusters that he or she intends to extract from the data set. This can be both a positive and a negative. It can be positive, because the algorithm is forced to give out the number of clusters that the user requires for business execution, irrespective of whether there is a better cluster solution.

On the other hand, it is a limitation, because the user is not informed whether there is a better cluster solution for the data set. A seven-cluster solution might be better than a four-cluster solution. Analysts usually run all cluster solutions and then pick the one that is most efficient or makes most business sense.

The other biggest issue with the k-means technique is the fact that the algorithm can give different results in different iterations. This is because of the way this technique is designed. Because the first step involves identifying random centroid values, each iteration would have different values and thus can give different results.

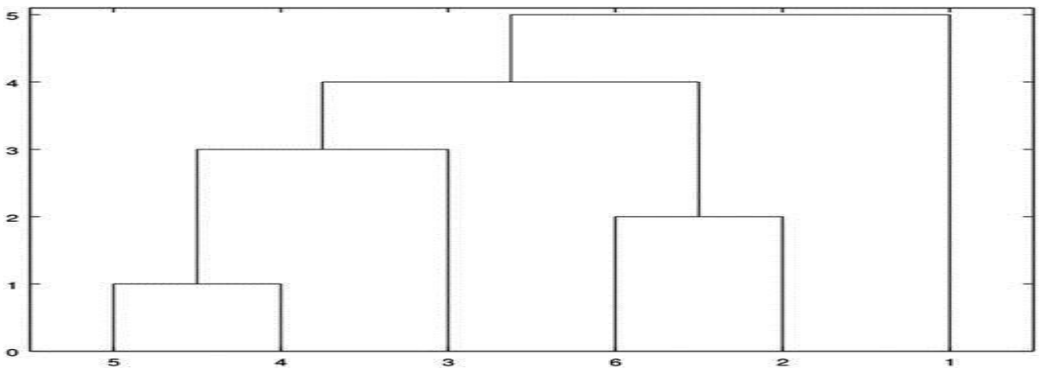
Q: Given that each iteration of k-means gives us different results, how would you ensure picking the best results?
Ans: This is done by calculating the SSE for each iteration. SSE stands for sum of squared error. SSE is calculated by first determining the distance between each data point and closest computed centroid and then summing all these distances. A smaller SSE represents a better solution.

Q: What are the two types of hierarchical clustering?
Ans: The two types are

- Agglomerative clustering
- Divisive clustering

Q: What is the difference between agglomerative vs. divisive clustering?
Ans: The main difference lies in how the initial group is defined. In agglomerative clustering, each data point is considered a cluster of its own. In each iteration, the data points are merged to form clusters that eventually form one big cluster containing all data points. Consider this a bottom-up approach.
On the other hand, divisive clustering is a top-down approach, in which all data points are initially considered part of one big cluster and then eventually broken into sub-clusters. Finally, the optimal number of clusters is derived.

Q: What is a dendrogram?
Ans: A dendrogram is a graphical representation of data sets and their cluster membership, using a tree-like diagram. Each vertical line represents either a data point or a cluster. The bottom-most vertical lines represent a data point. As we subsequently move up the diagram, the vertical lines merge (finally, into one), to reveal the cluster formation.



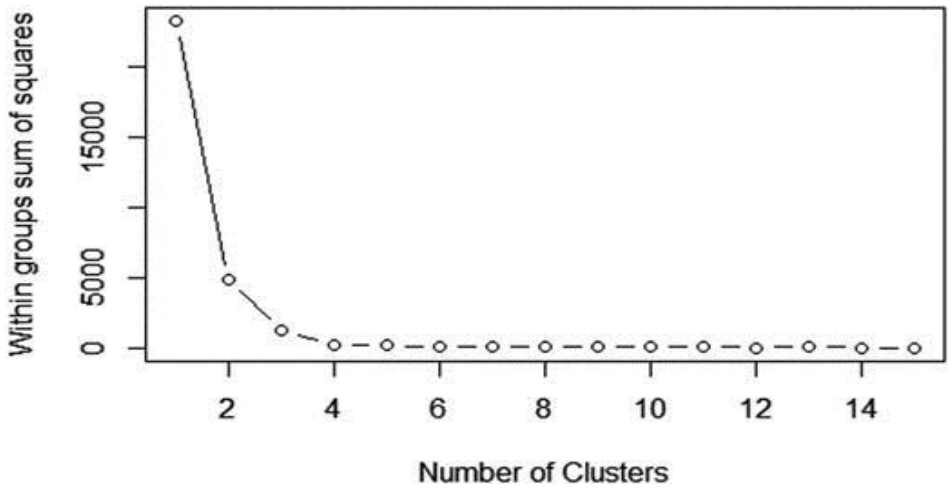
Q: What is the basic algorithm behind the agglomerative clustering technique?
Ans: First, consider all data points as a separate cluster. Then, using a proximity measure, define the proximity between all clusters. Combine clusters that are closest. Repeat this until only one cluster is left.

Q: Can you briefly explain some of the proximity measures that are used in hierarchical clustering techniques?
Ans: There are numerous proximity measures used in the clustering techniques. Min defines cluster proximity as the minimum distance between the closest two points in the clusters, whereas max defines cluster proximity as the maximum distance between any two points.
Group average refers to the average of all pair-wise distance between the points in the clusters.
An alternative technique, Ward's method, is more widely used.

Q: What is Ward's method of defining cluster proximity?
Ans: Ward's method is very similar to the k-means method of finding optimal cluster numbers. It measures the proximity by the increase in SSE when two clusters are merged. In other words, it reduces the sum of squared errors while clusters are joined together.

Q: How do you determine the optimal number of clusters for a data set?
Ans: A clustering technique is both an art and a science. Determining the optimal number of clusters is a crucial part of a clustering technique, and it is different from the actual clustering itself.
Determining optimal clusters requires consideration of both the technical aspects as well as the business aspects.
Technical Methods: Dendogram, Elbow method
Statistical Metrics: Silhouette coefficient (SC value), Pseudo F-value, Cubic Clustering Criteria etc
Business Aspects: Using Profiling, you can check which solution is differentiating
Best Practices: Looking at cluster size (should not have very big or very small clusters)

Q: Can you briefly explain the elbow method to determine the optimal cluster solution?
Ans: An elbow plot is a graph drawn from a number of clusters on an x axis and the SSE of each cluster solution on a y axis. We also call this a hockey stick graph, because the curve bends sharply, like an elbow point. We consider the point on the bend of the x axis as the optimal cluster solution.



In the above figure, we notice that the curve bends sharply at cluster number 4.The incremental reduction in error terms, while increasing the number of cluster solutions after 4, is very low. Thus, we consider 4 as the optimal cluster solution.

Q: What is the business aspect of determining the optimal cluster solution?
Ans: A clustering algorithm only segments a data set into various clusters, in which data sets are closer together, based on various attributes. Defining these clusters into meaningful definitions and identifying usage and business strategy around them is something that an analytical person or data scientist brings to the mix.

Each cluster identified by an algorithm should have a business meaning. For example, demographic segmentations can yield clusters that are meaningless for business usage. Also, clustering solutions cannot identify segments that actually have some meaning. For example, attributes with absolute zero values cannot be identified by the algorithm, as they almost always cluster data sets in a vicinity.

This is where the business aspect of the solution comes into the picture. Most of time, analysts come out with a high number of optimal cluster solutions, using technical aspects. Then they merge various clusters, using business rules. As stated previously, this is as much an art as a science.

Q: Can you explain, using a case study, the use of clustering techniques in the retail industry?
Ans: A mobile phone manufacturer would like to launch in a new geographical area. Traditionally, the phone manufacturer competes at all price levels and custom-creates phones to suit different buyers in a particular area. For the new geographical area, the manufacturer starts out by clustering the large sample data set of citizen demographics. This data set is then enriched using a primary survey of needs of various mobile phone users. Using this clustering methodology, the manufacturer is able to distinctly identify five clusters that behave in different ways and have unique needs. Thus, the manufacturer custom-creates a mobile phone for these five segments. Subsequently, marketing campaigns are also designed by keeping the behavior of these five segments in mind.

Q. I want to know for my product what my target population is. How do I do that using which statistical technique?
Ans: Need to do market segmentation using Cluster analysis. First do hierarchical clustering to identify no of segment of the market for similar product and then do K-means cluster to identify the levels of the parameters of the cluster to target.

Q. Usually how the analysis happens in a customer engagement analysis etc whether we do factor analysis first or cluster analysis or regression?
Ans: First we do factor analysis to reduce variables then with the factors we do regression to identify important factors then we do Cluster analysis based on the important factors.

Segmentation (Other questions without answers):

1. What is segmentation? What is the importance of segmentation?
2. Can you explain few situations where you have used segmentation?
3. Can you explain the methodology which you followed?(Steps)
    1. What is the data considered for this analysis?
    2. What level are you doing analysis (customer level/product level/category level/region level)?
    3. What variables you selected for this analysis?
    4. Why do you require missing value treatment? (if you don't perform, what are consequences?)
    5. Why do you require outlier treatment? (if you don't perform, what are consequences?)
    6. What is multicollinierity? What are consequences of multicollinierity in segmentation?
    7. How will you remove multicollinierity?
    8. What are different variable reduction techniques you have been used?
    9. Do you know about factor analysis?
    10. What is the factor analysis? What is importance of factor analysis?
    11. What is the difference between factor analysis and principle component analysis?
    12. Do you require standardizing variables for factor analysis?
    13. How many factors you got in your analysis? How did you finalize?
    14. What are the Eigen value/latent roots and their importance?
    15. What is importance of standardization? What are the different ways to standardize the data?
    16. How do you standardize in SAS/R/Python?
    17. What segmentation technique you have used in your project (K-Means/K-Medians/ Hierarchical/DBSCAN)?
    18. Is standardization required for Hierarchical segmentation?
    19. What is the difference between K-means and Hierarchical segmentation? Which one do you prefer?
    20. How do you finalize the segment solution? And optimize the solutions?
    21. How many segments you got in your analysis? What is their importance?
    22. How will you implement the solution?
    23. How do you score your entire population?

PROCESS & MISCELLANEOUS

Q. How to optimize algorithms? (Parallel processing and/or faster algorithms). Provide examples for both?
Ans: "Premature optimization is the root of all evil"; Donald Knuth

Parallel processing: for instance in R with a single machine.
- doParallel and foreach package
- doParallel: parallel backend, will select n-cores of the machine
- for each: assign tasks for each core
- using Hadoop on a single node
- using Hadoop on multi-node

Faster algorithm:
- In computer science: Pareto principle; 90% of the execution time is spent executing 10% of the code
- Data structure: affect performance
- Caching: avoid unnecessary work
- Improve source code level

For instance: on early C compilers, WHILE(something) was slower than FOR, because WHILE evaluated "something" and then had a conditional jump which tested if it was true while FOR had unconditional jump.

Q. Examples of NoSQL architecture
Ans:
- Key-value: in a key-value NoSQL database, all of the data within consists of an indexed key and a value. Cassandra, DynamoDB
- Column-based: designed for storing data tables as sections of columns of data rather than as rows of data. HBase, SAP HANA
- Document Database: map a key to some document that contains structured information. The key is used to retrieve the document. MongoDB, CouchDB
- Graph Database: designed for data whose relations are well-represented as a graph and has elements which are interconnected, with an undetermined number of relations between them. Polyglot Neo4J

Q. Provide examples of machine-to-machine communications
Ans:
Telemedicine
- Heart patients wear specialized monitor which gather information regarding heart state
- The collected data is sent to an electronic implanted device which sends back electric shocks to the patient for correcting incorrect rhythms

Product restocking
- Vending machines are capable of messaging the distributor whenever an item is running out of stock

Q. Is it better to have 100 small hash tables or one big hash table, in memory, in terms of access speed (assuming both fit within RAM)? What do you think about in-database analytics?
Ans:
Hash tables:
- Average case O(1) lookup time
- Lookup time doesn't depend on size

Even in terms of memory:
- O(n) memory
- Space scales linearly with number of elements
- Lots of dictionaries won't take up significantly less space than a larger one

In-database analytics:
- Integration of data analytics in data warehousing functionality
- Much faster and corporate information is more secure, it doesn't leave the enterprise data warehouse
Good for real-time analytics: fraud detection, credit scoring, transaction processing, pricing and margin analysis, behavioural ad targeting and recommendation engines

Q. What is star schema? Lookup tables?
Ans: The star schema is a traditional database schema with a central (fact) table (the "observations", with database "keys" for joining with satellite tables, and with several fields encoded as ID's). Satellite tables map ID's to physical name or description and can be "joined" to the central fact table using the ID fields; these tables are known as lookup tables, and are particularly useful in real-time applications, as they save a lot of memory. Sometimes star schemas involve multiple layers of summarization (summary tables, from granular to less granular) to retrieve information faster.

Lookup tables: Array that replace runtime computations with a simpler array indexing operation

Q. What is the life cycle of a data science project?
Ans:

1. **Data acquisition:** acquiring data from both internal and external sources, including social media or web scraping. In a steady state, data extraction and routines should be in place, and new sources, once identified would be acquired following the established processes
2. **Data preparation**: Also called data wrangling: cleaning the data and shaping it into a suitable form for later analyses. Involves exploratory data analysis and feature extraction.
3. **Hypothesis & modelling**: Like in data mining but not with samples, with all the data instead. Applying machine learning techniques to all the data. A key sub-step: model selection. This involves preparing a training set for model candidates, and validation and test sets for comparing model performances, selecting the best performing model, gauging model accuracy and preventing over fitting
4. **Evaluation & interpretation:**
   Steps 2 to 4 are repeated a number of times as needed; as the understanding of data and business becomes clearer and results from initial models and hypotheses are evaluated, further tweaks are performed. These may sometimes include step5 and be performed in a pre-production.
5. **Deployment**
6. **Operations:** Regular maintenance and operations. Includes performance tests to measure model performance, and can alert when performance goes beyond a certain acceptable threshold
7. **Optimization:** Can be triggered by failing performance, or due to the need to add new data sources and retraining the model or even to deploy new versions of an improved model

Note: with increasing maturity and well-defined project goals, pre-defined performance can help evaluate feasibility of the data science project early enough in the data-science life cycle. This early comparison helps the team refine hypothesis, discard the project if non-viable, change approaches.

8. **How to efficiently scrape web data, or collect tons of tweets?**
Ans: Python example

- Requesting and fetching the webpage into the code: httplib2 module
- Parsing the content and getting the necessary info: BeautifulSoup from bs4 package
- Twitter API: the Python wrapper for performing API requests. It handles all the OAuth and API queries in a single Python interface
- MongoDB as the database
- PyMongo: the Python wrapper for interacting with the MongoDB database
- Cronjobs: a time based scheduler in order to run scripts at specific intervals; allows to bypass the "rate limit exceed" error

Q. How to clean data?
Ans: 1. First: detect anomalies and contradictions
Common issues:

- Tidy data: column names are values, not names, e.g. <15-25, >26-45…
  multiple variables are stored in one column, e.g. m1534 (male of 15-34 years' old age)
  variables are stored in both rows and columns, e.g. tmax, tmin in the same column
  multiple types of observational units are stored in the same table. e.g, song dataset and rank dataset in the same table
  *a single observational unit is stored in multiple tables (can be combined)
- Data-Type constraints: values in a particular column must be of a particular type: integer, numeric, factor, boolean
- Range constraints: number or dates fall within a certain range. They have minimum/maximum permissible values
- Mandatory constraints: certain columns can't be empty
- Unique constraints: a field must be unique across a dataset: a same person must have a unique SS number
- Set-membership constraints: the values for a columns must come from a set of discrete values or codes: a gender must be female, male
- Regular expression patterns: for example, phone number may be required to have the pattern: (999)999-9999
- Misspellings
- Missing values
- Outliers
- Cross-field validation: certain conditions that utilize multiple fields must hold. For instance, in laboratory medicine: the sum of the different white blood cell must equal to zero (they are all percentages). In hospital database, a patient's date or discharge can't be earlier than the admission date

2. **Clean the data using:**
   - Regular expressions: misspellings, regular expression patterns
   - KNN-impute and other missing values imputing methods
   - Coercing: data-type constraints
   - Melting: tidy data issues
   - Date/time parsing
   - Removing observations

Q. How frequently an algorithm must be updated?
Ans: You want to update an algorithm when:
- You want the model to evolve as data streams through infrastructure
- The underlying data source is changing
Example: a retail store model that remains accurate as the business grows
- Dealing with non-stationarity

Some options:
- Incremental algorithms: the model is updated every time it sees a new training example
Note: simple, you always have an up-to-date model but you can't incorporate data to different degrees.
Sometimes mandatory: when data must be discarded once seen (privacy)
- Periodic re-training in "batch" mode: simply buffer the relevant data and update the model every-so-often
Note: more decisions and more complex implementations

How frequently?
- Is the sacrifice worth it?
- Data horizon: how quickly do you need the most recent training example to be part of your model?

- **Data obsolescence:** how long does it take before data is irrelevant to the model? Are some older instances more relevant than the newer ones?
Economics: generally, newer instances are more relevant than older ones. However, data from the same month, quarter or year of the last year can be more relevant than the same periods of the current year. In a recession period: data from previous recessions can be more relevant than newer data from different economic cycles.

Q. What is POC (proof of concept)?
Ans:

- A realization of a certain method to demonstrate its feasibility
- In engineering: a rough prototype of a new idea is often constructed as a proof of concept

Q. Explain Tufte's concept of "chart junk"
Ans: All visuals elements in charts and graphs that are not necessary to comprehend the information represented, or that distract the viewer from this information

Examples of unnecessary elements include:
- Unnecessary text
- Heavy or dark grid lines
- Ornamented chart axes
- Pictures
- Background
- Unnecessary dimensions
- Elements depicted out of scale to one another
- 3-D simulations in line or bar charts

Q. How would you come up with a solution to identify plagiarism?
Ans: Vector space model approach

- Represent documents (the suspect and original ones) as vectors of terms
- Terms: n-grams; n=1 to as much we can (detect passage plagiarism)
- Measure the similarity between both documents
- Similarity measure: cosine distance, Jaro-Winkler, Jaccard
- Declare plagiarism at a certain threshold

Q. How to detect individual paid accounts shared by multiple users?
Ans:

- Check geographical region: Friday morning a log in from Paris and Friday evening a log in from Tokyo
- Bandwidth consumption: if a user goes over some high limit
- Counter of live sessions: if they have 100 sessions per day (4 times per hour) that seems more than one person can do

Q. Is it better to spend 5 days developing a 90% accurate solution, or 10 days for 100% accuracy? Depends on the context?
Ans:

- At the beginning: quick-and-dirty model is better
- Optimization later

Other answer:
- Depends on the context
- Is error acceptable? Fraud detection, quality assurance

Q. What is your definition of big data?
Ans: Big data is high volume, high velocity and/or high variety information assets that require new forms of processing.

- Volume: big data doesn't sample, just observes and tracks what happens
- Velocity: big data is often available in real-time
- Variety: big data comes from texts, images, audio, video…

Difference big data/business intelligence:
- Business intelligence uses descriptive statistics with data with high density information to measure things, detect trends etc.
- Big data uses inductive statistics (statistical inference) and concepts from non-linear system identification to infer laws (regression, classification, clustering) from large data sets with low density information to reveal relationships and dependencies or to perform prediction of outcomes or behaviours

Q. Explain the difference between "long" and "wide" format data. Why would you use one or the other?
Ans:

- Long: one column containing the values and another column listing the context of the value Fam_id year fam_inc
- Wide: each different variable in a separate column
  Fam_id fam_inc96 fam_inc97 fam_inc98

Long Vs Wide:
- Data manipulations are much easier when data is in the wide format: summarize, filter
- Program requirements

Q. Do you know a few "rules of thumb" used in statistical or computer science? Or in business analytics?
Ans:
Pareto rule:
- 80% of the effects come from 20% of the causes
- 80% of the sales come from 20% of the customers

Computer science: "simple and inexpensive beats complicated and expensive" - Rod Elder
Finance, rule of 72:
- Estimate the time needed for a money investment to double
- 100$ at a rate of 9%: 72/9=8 years

Rule of three (Economics):
- There are always three major competitors in a free market within one industry

Q. Name a few famous API's (for instance Google Search)
Ans: Google API (Google Analytics, Picasa), Twitter API (interact with Twitter functions), GitHub API, LinkedIn API (users data)