Big_Data_Science_Assignment Name-Vinayak Raghupathy

NYUID-N11568565

Q1) Explain in your own words the concept of Hash Joins. Provide an example.

Describe the differences between Hash Joins and Normal SQL Join.

Explain this statement: the computational cost of a Hash Join depends on the cost of building the hash table.

Hash join is a type of join algorithm used in RDBMS(Relational Database Management Systems).The goal is to find for each value of the distinct attribute,the set if the tuples in each relation that have that value.Hash join requires one or more predicates of the form table1.columnA=table2.columnB such that the column types are same.

The hash join algorithm for inner join of two relations and has two phases-

BUILD PHASE

1)Prepare a Hash Table of the smaller relation.The entries in hash table consists of the join column and its row.Hash table is accessed by applying the hash function on the join attribute as it is quicker to find the rows corresponding to the join column then scanning the entire table.

PROBE PHASE

2)After the hash table is built,scan the larger relation and find the relevant rows from smaller relations by looking into the hash table.

For Example

Select ta.title_id

from titleauthor ta,authors a

where a.au_id=ta.au_id

and au_lname='Bloom'

In this example, the source of the build input stream is an index scan of author.Only rows with an au_lname value of "Bloom" are returned from this scan. These rows are then hashed on their au_id value and placed into their corresponding hash bucket. After the initial build phase is completed, the probe stream is opened and scanned. Each row from the source index of titleauthor is hashed on the au_id column. The resulting hash value is used to determine which bucket in the build set should be searched for matching hash keys. Each row from the build set's hash bucket is compared to the probe row's hash key for equality. If the row matches, the titleauthor.au_id column is returned.

Difference between Hash Join and Normal Join

A "sort merge" join is performed by sorting the two data sets to be joined according to the join keys and then merging them together. The merge is very cheap, but the sort can be prohibitively expensive especially if the sort spills to disk. The cost of the sort can be lowered if one of the data sets can be accessed

in sorted order via an index, although accessing a high proportion of blocks of a table via an index scan can also be very expensive in comparison to a full table scan.

A hash join is performed by hashing one data set into memory based on join columns and reading the other one and probing the hash table for matches. The hash join is very low cost when the hash table can be held entirely in memory, with the total cost amounting to very little more than the cost of reading the data sets. The cost rises if the hash table has to be spilled to disk in a one-pass sort, and rises considerably for a multipass sort.

The computational cost of Hash join depends on the cost of building the hash table.

This means that the cost of a hash join largely depends on the cost of building a hash table on one of the input sides to the join and using the rows from the other side of the join to probe it.Whenever a hash join is performed it requires the hash table to be built thus cost of the query i.e. the hash join depends on the cost of building the hash table.

Q2) Twitter Predicts the Stock Market

In the work mentioned in class for correlating sentiment analysis with stock market data, the authors relied on a correlation/causality algorithm. Briefly explain the algorithm in your own words. Provide an example.

Ans 2) Granger Causality is an algorithm that is based on prediction. According to Granger causality, if a signal X1 "Granger-causes" (or "G-causes") a signal X2, then past values of X1 should contain information that helps predict X2 above and beyond the information contained in past values of X2 alone.
For example G-causality to identify causal interactions in neural data. For example, Bernasconi and Konig (1999) applied Geweke's spectral measures to describe causal interactions among different areas in the cat visual cortex, Liang et al. (2000) used a time-varying spectral technique to differentiate feedforward, feedback, and lateral dynamical influences in monkey ventral visual cortex during pattern discrimination, and Kaminski et al. (2001) noted increasing anterior to posterior causal influences during the transition from waking to sleep by analysis of EEG signals

Paper Summary
Two tools were used to measure variations in the public mood from tweets submitted to Twitter service. The first tool, OpinionFinder, analyses the text content of tweets submitted on a given day to provide a positive vs. negative daily time series of public mood. The second tool, GPOMS, similarly analyses the text content of tweets to generate a six-dimensional daily time series of public mood to provide a more detailed view of changes in public along a variety of different mood dimensions.
Three phases were followed-
In first phase-
1)OpinionFinder measures postivie vs negative mood from text content.
2)GPOMS measures 6 different mood dimension from text content.
3)Extraction of time series of daily DJIA values from YAHOO Finance.

In the second phase, investigation was done whether public mood as measured by GPOMS and OpinionFinder is predictive of future DJIA values. A Granger causality analysis is used in which DJIA values are correlated to GPOMs and OF values of the past n days.

In the third phase, Self-Organizing Fuzzy Neural Network model is used to test whether prediction accuracy of DJIA prediction models could be improved by including measurements of public mood.

Q3) Flu Detection by Google

Describe the overall architecture of the analytics behind Google trends and its prediction to Flu Outbreaks.

Explain in your own words why the Flu Outbreak detection analytics platforms by Google failed.

Ans 3)Google Flu Trends was developed from a generic GT(Google Trends) tool.Users can enter queries on the web to see the relative volume of these queries for example queries related to flu.GT does analysis on the fraction of google web searches over a period of time and extrapolates the data to estimate the search volume.The information is then displayed on search volume index graph which is updated daily.Below it is the news reference volume graph which shows the frequency with which web search queries appeared in Google News Stories.If a spike is detected in the News Reference volume graph GT labels the search volume index graph with the random but relevant Google news headline which was published near the the time of spike towards the right.While region,cities and languages with highest volume are shown at the bottom of the page. In ranking the top regions, cities, and languages, GT takes a sample of all Web search queries and determines the areas or languages from which the most searches for the entered terms originate.Origin of web search queries can be established from the IP address of query logs. The algorithm then calculates the ratio of a variable's search volume from each city and the total search volume from those cities.City name and bar charts represent this ratio and can be less meaningful if they are close together.

However data for GT can have inaccuracies for a number of reasons like data sampling issues and the approximation method used.The search criteria is not standardized, a user can enter symptoms differently depending on their education level and cultural and language backgrounds.For example user can enter fever,pyrexia and rigors for the same symptom.GT is available only in limited number of languages and region specific versions.People can search google for information related to Influenza when they do not actually have the flu and are enticed by the factors such as increased media attention related to the illness.So due to these reasons Flu Outbreak detection analytics platforms by Google failed.

Q4 Explain and define the following concepts in **your own words**

1. Big Data

   Big data is an evolving term that can be used to describe any voluminous amount of structured,unstructured or semi structured data that has the potential to be mined for information.It is characterized by the three V's-

   Volume(Extreme Volume of Data)

   Enterprises are flooded with ever growing data of all types easily amassing terabytes or Even petabytes of information.For example 12 Terabytes of Tweets generated each day are turned into improved product sentiment analysis.

   Velocity(Velocity at which data must be processed)

   For some applications even 2 minutes is too late.Time sensitive processes such as catching fraud here big data can be used as streams in the enterprise to maximize value.For example Analyze 500 million daily call detail records in real-time to predict customer churn faster.

   Variety(Wide Variety of Types of Data)

   Big data can be any type of data-Structured or unstructured like text,audio,video.We can get new insights on analyzing these data types.For example monitoring 100's of live video feeds from surveillance cameras to target points of interest.

2. Predictive Analytics

   It is the practice of extracting information from exisiting data sets in order to determine the patters and future trends and outcomes.It does not let you know what will happen in the future.It only forecasts what can happen in the future with a minimum level of reliability and also includes risk assessment and what if scenarios.

   They can be applied to business to analyze current data and historical facts in order to better understand customers,products and partners and to identify potential risks and oppurtunities for a company.A number of techniques like data mining,statistical modeling and machine learning can help in make future business forecasts.

   For example,stores that use data from loyalty can analyze past buying behavior to predict promotions or coupons a customer may participate or buy in the future.

3. Analytic

   Discovery and communication of meaningful and patterns in data.
   It is valuable in areas that are rich in recorded information and rely on simultaneous application of statistics,computer programming and operations research to quantify performance.It also favours data visualization to communicate insight.

   Specifically, areas with analytics include predictive analytics,prescriptive analytics,web analytics,credit risk analysis ,fraud analytics.Web Analytics allows marketers to capture session-level information about interactions on a website using an operation called sessionization.For Example Google Analytics is a popular free analytics tool.These interactions can provide information to track the referrer,search keywords,ip address and track the activities of the vistitor.So marketer can use this to improve his website with creative content and information architecture and marketing campaigns.

4. Data Science
   To define Data Science we need to first define the following terms

   Data- The facts from which conclusion can be drawn.Information in raw and unorganized form that refer to or represent conditions,ideas or objects.

   Science-It is the pursuit and application of knowledge and understanding of the natural and social world following a systematic methodology based on evidence.

   Data Science is an interdisciplinary field about processes and systems which can be used to extract knowledge or insights from data in various forms either structured or unstructured,which is a continuation of some of the data analysis fields like statistics,data mining and predictive analytics similar to Knowledge Discovery in Databases(KDD).It uses the above fields to investigate problems in various domains like agriculture,marketing optimization,fraud detection ,public policy etc.

5. Data Mining
   It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence,machine learning,statistics and database systems.Overall goal is to extract information and transform into an understandable structure for further use.It is the analysis step in KDD(Knowledge Discovery in Database).

   Example of Data Mining of government records-particularly records of the justice system helps us to identify systemic human rights violations in connection to generation and publication of invalid or fraudulent legal records of various governmental agencies.

6. Machine Learning
   It is a type of Artificial Intelligence that provides computers with the ability to learn without being explicitly programmed.It's area of interest is in the development of computer programs that can teach themselves to grow and change whenever they are shown new data.
   Its process is similar to data mining.Both search through data to look for patterns.However while data mining extract data for human understanding, machine learning uses that data to improve

program's own understanding.For example Facebook's news Feed changes according to the user's personal interactions with the other users.If a user frequently tags his friend in photo or writes on his wall,the News Feed will show more of the friend's activity in the user's News Feed due to presumed closeness.

7. Statistics

   It is the study of collection,analysis,interpretation and organization of data.Statistics can work with all kind of data like planning data collection in terms of design of surveys and experiments.Two kinds of methods are used in statistical analyzing-

   Descriptive Statistics-Summarize the data from sample using indexes like mean and standard deviation.

   Inferential Statistics-Conclusion are drawn from data having random variation.

   Some well known examples of statistical tests and procedures are –

   Correlation which is the extent to which two variables have linear relationship with each other.

   Regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables

8. Business Intelligence

   It is a process involving technology for analyzing data and presenting actionable information to help the corporate executives,business managers and other end users make more informed business decisions.

   A variety of tools,methodologies and applications are used to collect data from both internal and external sources,analyze it,develop and execute queries on it and create reports,dashboards to make the results available to corporate decision workers and also the operational workers.

   It involves ad hoc analysis and querying, enterprise reporting, online analytical processing (OLAP), mobile BI, real-time BI, operational BI, cloud and software as a service BI, open source BI, collaborative BI and location intelligence.

9. Cross-validation

   It is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice.

   The aim is to define a dataset to test the model in training phase and reduce the problems like overfitting and give an insight how the model will generalize to independent dataset.

   Cross-validation can also be used in variable selection.Suppose we are using the expression levels of 20 proteins to predict whether a cancer patient will respond to a drug.

   A goal would be to determine which subset of 20 features will help us generate the best predictive model.

10. Confusion Matrix

    It is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning.

    Assuming a sample of 27 animals — 8 cats, 6 dogs, and 13 rabbits, the resulting confusion matrix could look like the table below:

		Predicted		
		Cat	Dog	Rabbit
Actual class	Cat	5	3	0
	Dog	2	3	1
	Rabbit	0	2	11

In this confusion matrix, of the 8 actual cats, the system predicted that three were dogs, and of the six dogs, it predicted that one was a rabbit and two were cats.We faced a problem as the system can't distinguish between cats and dogs although it can differentiate between rabbits and other animals.All correct values lie in the diagonal and the errors lie outside the diagonal in table.

11. Unstructured Data

    All those things that can't be so readily classified and fit into a neat box: photos and graphic images, videos, streaming instrument data, webpages, PDF files, PowerPoint presentations.

    It doesn't reside in a traditional row-column database and is the opposite of structured data — the data stored in fields in a database.

12. Structured Data

    Data that resides in a fixed field within a record or file. This includes data contained in relational databases and spreadsheets. It depends on creating a data model. This includes defining what fields of data will be stored and how that data will be stored: data type (numeric, currency, alphabetic, name, date, address) and any restrictions on the data input (number of characters; restricted to certain terms such as Mr., Ms. or Dr.; M or F).

    Structured data is often managed using Structured Query Language (SQL) and has has the advantage of being easily entered, stored, queried and analyzed.

13. Semi-Structured Data

    Semi-structured data is a cross between the structured and unstructured data. It is a type of structured data, but lacks the strict data model structure. With semi-structured data, tags or other types of markers are used to identify certain elements within the data, but the data doesn't have a rigid structure. For example, XML documents and NoSQL databases.

14. Data Clustering

    The task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups. It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

    Grouping of shopping items

    Clustering can be used to group all the shopping items available on the web into a set of unique products. For example, all the items on eBay can be grouped into unique products.

15. Stream Mining

    The process of extracting knowledge structures from continuous, rapid data records. A data stream is an ordered sequence of instances that in many applications of data stream mining can be read only once or a small number of times using limited computing and storage capabilities. Examples of data streams include computer network traffic, phone conversations, ATM transactions, web searches, and sensor data.

Goal is to predict the class or value of new instances in the data stream given some knowledge about the class membership or values of previous instances in the data stream.RapidMiner is a commercial software for stream mining.

16. Data Classification

    In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. An example would be assigning a given email into "spam" or "non-spam" classes or assigning a diagnosis to a given patient as described by observed characteristics of the patient (gender, blood pressure, presence or absence of certain symptoms, etc.).

    Classification can be thought of as two separate problems – binary classification and multiclass classification. In binary classification, a better understood task, only two classes are involved, whereas multiclass classification involves assigning an object to one of several classes.

17. Supervised Learning

    The machine learning task of inferring a function from labeled training data.The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value.

    Analyzes the training data and produces an inferred function, which can be used for mapping new examples.

    Supervised learning includes two categories of algorithms:

    Classification: for categorical response values, where the data can be separated into specific “classes”

    Regression: for continuous-response values

    Common classification algorithms include:Support vector machines (SVM),Neural networks

    Common regression algorithms:linear regression,non-linear regression.

18. Correlations

    In statistics, dependence is any statistical relationship between two random variables or two sets of data. Correlation refers to any of a broad class of statistical relationships involving dependence, though it most in common usage often refers to the extent to which two variables have a linear relationship with each other. Familiar examples of dependent phenomena include the correlation between the physical statures of parents and their offspring, and the correlation between the demand for a product and its price.

    Correlations are useful because they can indicate a predictive relationship that can be exploited in practice. For example, an electrical utility may produce less power on a mild day based on the correlation between electricity demand and weather. In this example there is a causal relationship, because extreme weather causes people to use more electricity for heating or cooling.

19. Distributed Systems

    A software system which has components located on networked computers communicate and coordinate their actions by passing messages.Three important features are-

    1)Component Concurrency

    2)lack of global clock

    3)Independent failure of components

A common goal can be solving a large computational problem.Alternatively, each computer may have its own user with individual needs, and the purpose of the distributed system is to coordinate the use of shared resources or provide communication services to the users

Examples include-telecommunication networks like telephone and cellular networks,network applications like world wide web and peer to peer networks.

20. Unsupervised Learning

A machine learning task of inferring a function to describe hidden structure from unlabeled data.Since the data is unlabeled no error or reward signal is there for finding the solution.

Many methods employed in unsupervised learning are based on data mining methods used to preprocess data.

One of the approaches in unsupervised learning is the method of moments. In the method of moments, the unknown parameters (of interest) in the model are related to the moments of one or more random variables, and thus, these unknown parameters can be estimated given the moments. The moments are usually estimated from samples in an empirical way. The basic moments are first and second order moments.

21. Training Data

A set of data used to discover potentially predictive relationships, used in intelligent systems, machine learning, genetic programming and statistics. Searching through training data for empirical relationships tend to overfit the data, meaning that they can identify apparent relationships in the training data that do not hold in general.

22. Test Data

A set of data used to assess the strength and utility of a predictive relationship, used in intelligent systems, machine learning, genetic programming and statistics.It is independent of training data but follows same probability distribution.

23. Recommender Systems

These systems predict the 'rating' or preference a user will give to an item.For example they are used in movies,music,books,search queries and products.Rating is given in two ways

Collaborative-Model is built based on user's past behavior and similar decisions of other users.For example user based algorithm and kernel-mapping recommender.

Content-Based-Discrete Characteristics of items are used to recommend additional items with similar properties.For example Rotten Tomatoes, Internet Movie Database.

24. Trust Based Recommender Systems

A recommender system is of little value for a user if the user does not trust the system.Trust can be built by a recommender system by explaining how it generates

recommendations, and why it recommends an item.

25. Biologically Inspired Data Mining

Data Mining techniques that are inspired from biological phenomenon for example

An Ant Colony Optimization algorithm (ACO) is essentially a system based on agents which simulate the natural behavior of ants, including mechanisms of cooperation and adaptation. Each path followed by an ant is associated with a candidate solution for a given problem.

• When an ant follows a path, the amount of pheromone deposited on that path is proportional to the quality

of the corresponding candidate solution for the target problem.

• When an ant has to choose between two or more paths, the path(s) with a larger amount of pheromone have a greater probability of being chosen by the ant.

As a result, the ants eventually converge to a short path, hopefully the optimum or a near-optimum solution for the target problem.

26. Knowledge Discovery(KD)

It is the process of automatically searching large volumes of data for patterns that can be considered as knowledge about data.It is basically deriving knowledge from input data.It developed out of the data mining domain.

Abstraction of Input data is generated.Knowledge obtained from input data can become input data for more knowledge and discovery.An important application of KD is in the software modernization which considers understanding existing software artifacts.

Here the knowledge obtained from existing software is used to present models on which specific queries can be made.An entity relationship is a frequent model of representing knowledge obtained from existing software.

27. Class Label (in Data Classification)

A set of training objects each with a number attribute values are given to the classifier which makes rules for each class in training set.After that rules are applied to classify objects and determine which class the object belongs to.Classification is used to predict the class labels of new data objects.

28. Standard Deviation

In statistics, the standard deviation (σ) is a measure that is used to quantify the amount of variation or dispersion of a set of data values. The standard deviation of a random variable, statistical population, data set, or probability distribution is square root of variance.A useful property unlinke variance it is expressed in same units as data.

The standard deviation is commonly used to measure confidence in statistical conclusions. For example, the margin of error in polling data is determined by calculating the expected standard deviation in the results if the same poll were to be conducted multiple times.

Let X be a random variable with mean value μ:

Standard Deviation $\sigma = \sqrt{(X-\mu)^2}$

29. Variance

In Probability and Statistics,variance means how far the numbers are spread out.A variance of 0 means the numbers are identical.It is always non-negative. A small variance indicates that the

data points tend to be very close to the mean and also to each other, while a high variance indicates that the data points are very spread out around the mean and from each other.

The variance of a set of observed values that is represented by random variable X is its second central moment, the expected value of the squared deviation from the mean μ = E[X]:

$$Var(X) = E[(X - \mu)^2]$$

30. ETL Jobs

ETL(Extract,Transform and Load) refers to a process in database usage and especially in data warehousing that:

Extracts data from homogeneous or heterogeneous data sources.

Transforms the data for storing in proper format or structure for the purpose of querying or analysis

Loads the data into final target typically a database.

For example, a cost accounting system may combine data from payroll, sales, and purchasing.Some commercial ETL Tools are Informatica PowerCenter,IBM Infosphere DataStage.

31. SQL

SQL(Structured Query Language) is a database computer language designed for retrieval and management of data in relational database.When SQL command is executed in RDBMS the system determines the best way to carry out your request and SQL enginer figures out how to interpret the task.

Standard SQL commands are CREATE,SELECT,INSERT,UPDATE,DELETE and DROP.

Example of SQL SELECT statement:

To select all columns from a table (Customers) for rows where the Last_Name column has Jackson for its value, you would send this SELECT statement to the server back end:

Select * from Customers where Last_Name='Jackson'

32. NoSQL

NoSQL(non SQL or Non Relational) database provides a mechanism for storage and retrieval of data which is modeled in means other than the tabular relations used in relational databases. The data structures used by NoSQL databases (e.g. key-value, wide column, graph, or document) make some operations faster and are also more flexible then the relational data structures.

NoSQL databases are increasingly used in big data and real-time web applications.NoSQL systems are also sometimes called "Not only SQL" to emphasize that they may support SQL-like query languages.

For Example Graph database(Neo4J) is designed for data whose relations are well represented as a graph consisting of elements interconnected with a finite number of relations between them. The type of data could be social relations, public transport links, road maps or network topologies.

33. RDBMS(Relational Database Management System)
    It is a program that lets you create,update and administer a relational database.Most commercial RDBMS use SQL to access the database.Leading RDBMS products are Oracle,IBM's DB2 and Microsoft's SQL Server.Data in RDBMS is stored in database object called Tables which are collection of related data entries and It consists of columns and rows.
    A single database can be spread across several tables while in a flat-file databases, each database is self-contained in a single table.

Sources

http://www.webopedia.com/TERM/P/predictive_analytics.html

https://en.wikipedia.org/wiki/Analytics

http://www.sciencecouncil.org/definition

https://en.wikipedia.org/wiki/Data_science

https://en.wikipedia.org/wiki/Data_mining

https://en.wikipedia.org/wiki/Examples_of_data_mining#Music_data_mining

http://whatis.techtarget.com/definition/machine-learning

https://en.wikipedia.org/wiki/Statistics

http://searchsqlserver.techtarget.com/definition/relational-database-management-system

https://en.wikipedia.org/wiki/Relational_database_management_system

http://www.webopedia.com/TERM/U/unstructured_data.html

http://www.tutorialspoint.com/sql/sql-rdbms-concepts.htm

http://www.tutorialspoint.com/sql/

https://en.wikipedia.org/wiki/NoSQL

https://en.wikipedia.org/wiki/Extract,_transform,_load

https://en.wikipedia.org/wiki/Variance

https://en.wikipedia.org/wiki/Knowledge_extraction#Knowledge_discovery

https://en.wikipedia.org/wiki/Regression_analysis

https://en.wikipedia.org/wiki/Correlation_and_dependence#Rank_correlation_coefficients

http://searchdatamanagement.techtarget.com/definition/business-intelligence

https://en.wikipedia.org/wiki/Cross-validation_(statistics)

https://en.wikipedia.org/wiki/Confusion_matrixs

http://www.webopedia.com/TERM/S/structured_data.html

https://en.wikipedia.org/wiki/Cluster_analysis

https://en.wikipedia.org/wiki/Data_stream_mining

https://en.wikipedia.org/wiki/Statistical_classification

https://en.wikipedia.org/wiki/Supervised_learning

http://in.mathworks.com/discovery/supervised-learning.html?s_tid=gn_loc_drop

https://en.wikipedia.org/wiki/Distributed_computing

https://en.wikipedia.org/wiki/Unsupervised_learning

https://en.wikipedia.org/wiki/Test_set

https://en.wikipedia.org/wiki/Recommender_system

http://sci2s.ugr.es/keel/pdf/algorithm/articulo/Ant-IEEE-TEC.pdf

Q5) Why is Data Science is considered a Science?
Ans5)
Scientific methodology includes the following:
1)Objective observation: Measurement and data (possibly although not necessarily using mathematics as a tool)
2)Evidence
3)Experiment and/or observation as benchmarks for testing hypotheses
4)Induction: reasoning to establish general rules or conclusions drawn from facts or examples
5)Repetition
6)Critical analysis
7)Verification and testing: critical exposure to scrutiny, peer review and assessment

Data science employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, chemometrics, information science, and computer science, including signal processing, probability models, machine learning, statistical learning, data mining,predictive analytics etc.

The data science life-cycle thus looks somewhat like:

1. Data acquisition

2. Data preparation

3. Hypothesis and modeling

4. Evaluation and Interpretation

5. Deployment

6. Operations

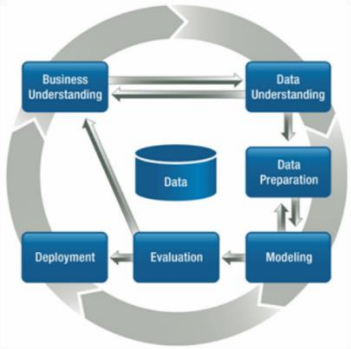
7. Optimization

Data scientists use their data and analytical ability to find and interpret rich data sources; manage large amounts of data despite hardware, software, and bandwidth constraints; merge data sources; ensure consistency of datasets; create visualizations to aid in understanding data; build mathematical models using the data; and present and communicate the data insights/findings. They are often expected to produce answers in days rather than months, work by exploratory analysis and rapid iteration, and to get/present results with dashboards (displays of current values) rather than papers/reports, as statisticians normally do

This follows the scientific method very closely, hence the reason for data science to be a science.

Q6) Provide an overview of the next steps in the predictive analytics lifecycle that you (the data scientist) would follow to predict the type of customer using the historical data collected by the business and the present data of the customer's behavior for a number of transactions.

Ans6)



The following will be the steps in the predictive analytics life cycle after the business understanding phase-

Data Understanding

In this phase we would try to understand the data collected from our internal clients i.e the marketers of the company like what are the different sources of data,the size of the data set,the frequency at which the data is generated per interval like a hour day or month ,whether the data is structured or unstructured.

Gain an understanding of customer related attributes like age,id,name,transaction history,number of items purchased in a specific interval and type of customer whether he is loyal,need based,discount ,wandering,high returned customers and other personal data.Data needs to be explored and verified for quality,check if data is complete.

The problems we may face while understanding data are data having missing value,noise in the data,data not available for some customers or attributes.

Data Preparation

This phase is the most important and time consuming of the life cycle.After gaining understanding of data we can handle the problems by either use a mean value for the missing values,clean the data and handle noise by smoothening bin by mean or median.We have to construct and integrate the data from multiple sources. An aggregation could include converting a table of customer purchases, where there is one record for each purchase, into a new table where there is one record for each customer. The table's fields could include the number of purchases, the average purchase amount, the percent of orders charged to credit cards, the

percent of items under promotion, etc..Data may also need to be formatted for example removing commas from within text fields
in comma separated data files.

Also we might do dimensionality reduction as a part of data reduction ,the algorithm we can use here is PCA(Principal Component Analysis) so that we can get a reduced representation of the dataset that is smaller in volume but still it preserves the structure and information with the data.So we can remove unimportant attributes like zip code from the customer data.In this phase ETL is perfomed for the data.Data can be visualized using tools like R,GnuPlot,Tableau.
If new problems are encountered and solutions are not available then it can mean that we need better understanding of data and can return back to the data understanding phase.

Data Modelling
In this phase we can select modelling techniques,generate test design and build predictive model for the loyal,need based,discount ,wandering,high returned customers .
Labelled Historical data of customers and the present data of the customer transactions is total data that we have.A predictive model can be generated that can predict the type of customer using this data and based on the type of customer marketers in the company can develop appropriate strategy.
We separate the Labelled historical data into train and test sets,build the model on test set and estimate its quality on the separate test set.

Historical data is used to train the model to arrive at the most suitable modelling technique. The modeling tool is run on the prepared data set to create one or more models.We can use modelling techniques like cluster analysis to find the pattern in transaction details in the historical data for similar type of customers like if the customers are buying lot of discounted product then prediction can be made about

new customers as discount customers.So if other customer buy those products they can also put in the same type and marketing strategy can be developed .
Classification models can also be used for predictive modeling.
This problem can be transformed to classification problem.Now the company has information on the payment behavior of their customers. By combining this financial information with other information about the customers, such as sex, age,income, etc., it is possible to develop a system to classify new customers into the various categories.Other techniques that can be used are Discriminant analysis , Rule induction methods, Decision tree learning etc. Ranking of the models can be done. Also ,The success of the application of modeling and discovery techniques technically should be judged.

Modelling techniques might require data in a specific format so it might also be possible that we need to go back to data preparation phase.

Evaluation

This phase consists of evaluating the results with respect to the business success criteria of the Project.Classifier will be run on the test data taken from the labelled historical data and accuracy will be checked.
After it is run the results will be compared to the business objectives and the results will be shown to internal clients.For example are we able to predict the customer type for any customer by looking at his current transactions.We can use cross validation here to find the subset of attributes of a customer that can help us build the best predictive model. If we have customer data which is grouped by observations, tools such as cluster analysis, association rules, and k-nearest neighbors usually provide the best results.

Deployment

Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to

be organized and presented in a way that is useful to the customer.Based on the evaluation results we need to develop deployment plan and document it.
Also develop a monitoring and maintenance plan to check that the model performance does not deteriorate.

In this phase final reports, briefings, code, and technical documents will be delivered to the internal clients using which they can develop marketing strategy.
In addition, the team may run a pilot project to implement the models in a production environment.Also we can have the meeting at the conclusion of the project,where the results are verbally presented to the customer.