DSR - Assignment-1

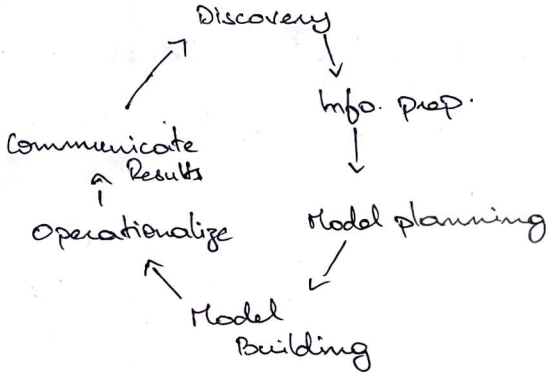
Haren A
2451-18-733-162
CSE-3

1) Data science is the field of study that combines domain expertise, prog. skills & knowledge of math & statistics to extract meaningful insights from data.

Business Intelligence (BI)	Data Analytics (DA)
• Concerned w/ collecting raw data and eval. historic growth of a business.	• Concerned w/ converting raw data and analysing it so as to set future trends & patterns.
• Emphasizes studying data based on situation that already took place	• Tends to highlight future patterns likely to occur in future.
• BI is related more to questioning the existing trends.	• DA helps make a decision based on past data.
• BI consists of tools that deal in data collection, producing reports.	• DA consists of tools that analyze raw data & turn it into useful info
• Concerned with achieving existing goals.	• Data analysis leads to addition of goals

> Business Intelligence helps interpret past data & used for reporting or descriptive analysis, whereas Data Analytics is designed to uncover the specifics of extracted insights. Data Science differs from abv. two in the aspect that it helps find meaningful correlation b/w large datasets & not related with predictive or descriptive analysis.

2) Data science process:



· Discovery:

It's very exceptionally imperative to get the diff. determinations, prerequisites, needs & required budget-related with venture. You must have the capacity to inquire the correct questions like do you have the desired assets or not. In this stage you get to outline thode issue & define starting hypotheses to test.

· Information preparation:

In this stage, we investigate, preprocess & condition data for modeling. We can perform info cleaning, changing and visualization, this helps in finding exceptions and find a relationship among datasets.

· Model planning:

Here, we decide strategies & methods to draw conn. b/w factors

· Model building:

In this stage, we'll create datasets for training & testing purposes.

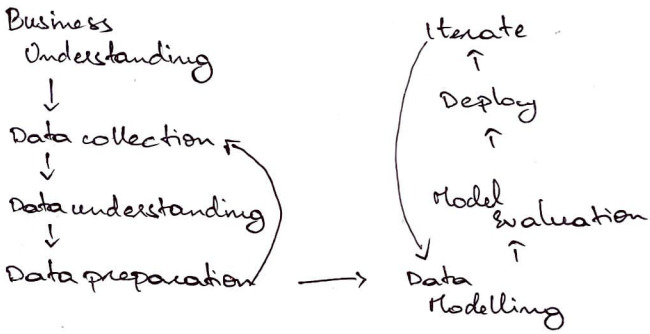
· Operationalize:

In this stage, we convey the last briefings, code and specialized reports, this gives a clear picture of execution and other related limitations.

· Communicate Results:

It's imp. to assess outcome of the objective. So, within the final stage, we ~~so~~ recognize all the key discoveries, communicate to the others & decide in the event that the outcomes about venture are a victory or not.

(6)



· Business understanding:

It's imp. to understand the problem, to be able to solve it. ~~of a~~ It's recomanded to take consultation from domain experts for a better understanding of the problem. After asking required question we move on to the next phase.

· Data collection:

After gaining clarity on pbm. stmt., we need to collect relevant data to break pblm. into smaller components. A data science project starts w/ identification of various data sources, which may include web server logs, social media posts, data from digital libraries or any datasets. A major challenge faced by data professionals in this phase is to understand where is it coming from.

· Data preparation:

Collected data may or may not be in required format. Therefore collected data should be cleaned before processing it anyfurther. Thus, this step is also called Data cleaning or Data wrangling. Data acquired in previous step might not give clear analytical picture or patterns in the data. So, to understand this data needs to be structured and cleaned. Apart from this, there could be any missing values which might cause obstruction in model building and analysis. Exploratory Data Analysis (EDA) plays an imp. role at this stage as summarization of clean data helps in identifying the structure, patterns or anomalies present.

· Data modelling:

In this stage, feature selection is one of the first things that should be done. Here we should try to reduce the dimentionality of the dataset. It's essential to identify what is required, is it classification, regression or a prediction problem, once this is sorted we can implement the model. After modelling, model performance measurement is required. Model should be robust and not overfitted as it'll not give accurate results for future data if overfitted.

· Interpreting data:

This is last step of a Data Science project and also the most important step. This step should be executed such as ~~even a~~ anyone can understand the projects outcome. Actionable insight from the model shows how data science has the power of doing predective or prescriptive analysis.

7)

Probability Density Function	Probability Mass Function
• Used when there's a need to find a sol. in a range of cont. random variables.	• Used when there's a need to find a sol. in a range of discrete random variables.
• Uses continuous random variables	Uses discrete random variables.
• Def. by ~~$f(x) = P(a < x)$~~ $\int_a^b f(x)dx = P(a < X < b)$	• Def. by $f(x) = P(X = x)$
• Sol. falls in the radius range of cont. random variables.	• Sol. falls in radius b/w numbers of discrete random variables.
• Eg: normal distribution	• Eg: Bernoulli distribution.

8) A hypotheses is an assumption or idea that's proposed for the sake of argument so that it can be tested to see if it's true.

There are four steps in hypotheses testing:

- the first step is for the analyst to state two hypotheses so that only one can be true
- next step is to formulate a plan, which outlines how the data will be evaluated.
- third step is to carry out the plan & physically analyze the sample data.
- fourth & final step is to analyze the results and either reject or accept the null hypotheses.

The null hypotheses is a baseline assumption that the treatments are equivalent & any diff. is due to chance. An alternative hypothesis is said to contradict the null hypothesis.

(9) Let's say variables in algebra as x, y, z. Here x can be no. of mobiles, y no. of heads and z is no. of students. A variable is something (alphabet) which represents an unknown number.

A random variable is diff. in the way that it has a whole set of values and it can take any of those randomly, while an algebraic variable can't have more than a single value at a time.

If a random var. $X = \{0, 1, 2, 3\}$, it could be 1 or 2 or 3 or 0 where each has a diff. probability. They can be either discrete or continuous. It's discrete if it has countable no. of distinct values, if it has infinite no. of values in an interval it's said to continuous.

(10) Properties of probability density func.:

- For a cont. random variable that takes some value b/w certain limits, say a & b, PDF is calculated by finding area under its curve & X-axis within the lower & upper limit

  $\Rightarrow P(x) = \int_a^b f(x)\,dx.$

- PDF is non-negative for all possible values, $f(x) \geq 0 \; \forall x$.
- Area b/w density curve & X-Axis is 1, $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

Properties of probability mass func.:

- The probabilities of all possible values of random variable should sum up to 1.
- All probabilities should be 0 or greater than 0.
- Probability of each event is b/w 0 & 1 [$\because P(x_k) \geq 0$].

③ Application of eigen vectors in data science

- communication system:
  eigen values were used by claude shannon to determine the theoretical limit to how much info can be transmitted thru a comm. medium like a telephone line or thru air. This is done by calc. eigen values & vectors of the communication channel & then waterfilling on the eigen values.

- designing bridges:
  the natural frequency of the bridge is the eigen value of smallest magnitude of a system that models the bridge

- designing car stereos:
  eigen value analysis is also used in design of car stereo, where it helps to reproduce the vibration of car due to music.

- Principle component analysis (PCA):
  This is performed for dimensionality reduction. With large datasets, finding significant features gets difficult. So, in order to check for the correlation b/w two variables & if they could be dropped off the table to make the ML model more robust.

4&5) linear algebra is a powerful tool for Data Science applications Analyzing the data is an imp. task in data management systems. linear algebra emerged as an optimal tool to analyze & manipulate the data. linear algebra helps us understand geometric terms in higher dimensions & do mathematical operations on them, this is the heart to almost all areas of math. Its concepts are a crucial prerequisite to understand linear algebra before getting started in Data Science and also to understand how the algorithms work.

Applications of linear algebra in Data Science:

- Loss function:

Consider how good a model is and fits a given data:

- some arbitrary prediction function, use it on independent features of data to predict the output, calc. how far-off is the o/p from actual o/p. Use this calc. values to optimize prediction func. It's difficult to calc. how diff prediction is from the expected o/p, this can be resolved using a loss function. loss func. is an application of vector norm, of which there are many types.

L1-Norm: Also called manhattan dist. or taxicab norm. This is the dist. travelled from the origin to the vector if only permitted directions are parallel to axes of the spaces.

L2-Norm: Also called euclidean dist, this is the shortest dist of the vector from origin.

- Regularizations:

This is an imp. concept we use to prevent overfitting.

- Covariance Matrix:

Bivariate analysis is an imp. step in data explanation to study the relation blw pairs of variables.

covariance or corelation is measured to study relation blw two cont. variables. Covariance indicates direction of linear relation blw variables.

- Dimensionality Reduction:

This method is to reduce the variables present in a dataset, so that we can perform some sort of coherence analysis.

- word embeddings:

This a way of representing words as low dimensional vectors of numbers while preserving their context in the document. These representations are obtained by training diff. neural networks on a large amount of text which is called a corpus.

- Computer Vision.
- Image representation as tensors.

12) A statistical hypothesis is an assumption about a population parameter, which may or maynot be true. Hypothesis testing refers to the formal procedure used by statisticians to accept or reject statistical hypothesis. Best way is to examine a population, as it's infeasible a sample from the population is tested.

There are two types of statistical hypothesis:

- Null hypothesis:
  Denoted by H_0, is usually the hypothesis that sample observations result is purely by chance.

- Alternative hypothesis:
  Denoted by H_1, states that sample observations are influenced by some non-random cause.

- Various test statistics

> P-value: The strength of evidence in support of a null hypothesis is measured by the P-value. Suppose the test statistic is equal to s, the P-value is the probability of observing a test statistic as extreme as s, assuming the null hypothesis is true, if p-value is less than the significance level, we reject the null hypothesis.

> Region of acceptance is a range of values, if the test statistic falls within the region of acceptance, null hypothesis isn't rejected. Set of values outside acceptance region is region of rejection, for which hypothesis has been rejected at the α level of significance.

> The statistical test, where the region of rejection is on only one side of sample distribution is called a one-tailed test. Eg, suppose H_0 states the mean ≤ 10, H_1 would be mean > 10. Region of rejection would consist of a range of numbers greater than 10.

> In the same way, where the region of rejection is on both sides is called two-tailed test. Eg, H_0 states mean = 10, H_1 states mean ≥ 10 or mean ≤ 10, which means region of rejection consists of no. both less than or greater than 10.