CRISP-DM

1. These are the actions that need to be taken in order to built a robust contextual search that best fit the institution's dataset:
    a. Business understanding: We define the objectives by working with the customers and stakeholders that have complained about the inefficiency of the current search engine. Ask questions and define targets and then find relevant data that helps answer these questions using data science techniques..
    b. Data acquisition and understanding: We produce a clean, high quality version of the 100TB text dataset where its relationship with the target variables is understood.
    c. Modelling: We determine the optimal data features for the machine learning model to derive context. Create a model that gives the most accurate search results and then finalise the model that is suitable for production.
    d. Deployment: Deploy the search engine with the data pipeline to a production environment for final acceptance.
    e. Customer acceptance: We make sure that the search engine is satisfactory to the customers and stakeholder's needs.
2. The main data source for the model will be the 100TB dataset the institution holds. We start by preparing the data. We normalise, remove stop words, any unicode symbols and lemmatize it. We use ETL tools for hadoop(Apache sqoop or flume) to transform the data into an appropriate structure for training. We use the BERT nlp model to derive context by generating embeddings for the corpus as well as the query. The embedding vectors are used to find the cosine similarity between each text document and the search query. Based on the similarity, we sort and return the top *x* relevant documents. This helps us create a semantic search engine that takes context into account.
    a. To achieve this task we need a team that has an in-depth understanding of basic data science product cycles, statistics, data cleaning algorithms, and Hadoop ETL tools.
3. Bert uses transformers and is trained on huge datasets. It is also currently the state of the art model for text classification and clustering tasks. It takes both left and right context into account which it achieves by implementing mask language modelling and next sentence prediction techniques. Hence, the embeddings it creates should be most accurate. Cosine similarity also proves to be effective on textual datasets. NoSQL databases such as HBase can provide linear and modular scalability and block cache for real time data queries. Sqoop can be helpful to import data in HBase and control parallelism.

Customer Churn - Conceptual questions

1. Customer churn is when a customer stops using an organisation's products or services, i.e the customer ceases to be a customer. It is also called customer attrition.

2. Customer churn rate is the metric by which customer churn is measured. Essentially it is the number of people who stopped using the organisation's products/services during a period of time(month, quarters, years). It can be calculated as: [(customers lost during a time period / total customers at the beginning of that time period)*100].
3. Customer churn prediction is detecting which customers are likely to stop using a company's service/products. Businesses invest heavily in acquiring new customers and every time a customer stops using their service, it shows a loss on the investment. More resources are needed to replace the leaving customer with a new one, so being able to predict when a client is likely to leave and then offering them incentives to stay, can help businesses to save money and other resources. Understanding what keeps customers using a company's services is important and implementing suitable strategies to retain them can be helpful for the company. It can be stated as a binary classification problem where we predict whether a customer will stop using a given product/service within a particular period of time.
4. The two types of customer churn:
    a. Voluntary churn: Customers consciously decide to stop using a business's products/services. Companies focus on this type of churn since it is possible to retain these customers through incentives, improvement to product etc...
    b. Involuntary churn: This means customer churn happened due to reasons not related to your business. For example, payment failure etc...
5. The 3 main variables that one should track are the ones that affect the customer churn rate the most:
    a. Subscription length: It is the amount of time the customer spends paying for the company's products/services.
    b. Customer acquisition cost: the amount of money spent to gain one new customer.
    c. Customer's lifetime value: It is the customer's total value toward a company's business over the duration of their relationship.
6. Customer information type diversity is the distinctiveness in the customers of a given business like age, gender, location, expenditure, preferences etc..
7. The 4 common models are:
    a. Logistic regression
    b. Decision trees
    c. Random forests
    d. Gradient boosted trees
8. These are some of the methods we can implement:
    a. Check if any particular segment of customers are churning. Then analyse if their specific needs can be catered by your product/service.
    b. Creating proper communication channels with the customers can help immensely. This will help get real feedback on the company's products/services. The problems causing churn can then be identified and fixed.
    c. Analysing rival products/services to gauge whether customers are getting better value for their investment somewhere else.
9. Many customer churn analysis use Imbalanced datasets, past customer churn data and statistical methods like regression to predict churn and changes in churn rate. Underlying assumptions made

by these methods are inherently flawed and hence the models can show vast deviations from reality.

10. Customer segmentation is the process of dividing customers into groups based on common features i.e age-group, preferences, behaviours etc. This can be used with the other customer churn dataset and can be helpful in reducing churn rate by targeting specific groups of the customer base.