Question 1: What is difference between Exploratory Data Analysis and Predictive Data Analysis?

Answer – Exploratory data analysis (EDA) is an approach to analysing data sets to summarize their main characteristics, often with visual methods. EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task. Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns to spot anomalies.

Predictive analytics is the branch of the advanced analytics which is used to make predictions about unknown future events. Predictive analytics uses many techniques from data mining, statistics, modelling, machine learning, and artificial intelligence to analyse current data to make predictions about future. Predictive analytics can also be the use of data, statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data

Question 2: How would you define the role of a Data Scientist in Product Development team?

Answer - In the early stages of product development, all data scientists have similar functions, and they are primarily focused on setting up the computational and analytical infrastructure. As the product evolves, data scientists' roles change depending on the needs of the product team. Therefore the role of a data scientist in a product development team will include the following:

- Product generalists - who are generic problem solvers working across product issues you may encounter
- Early product analyst - to determine product market fit for a nascent product
- Growth analyst - to move a metric
- Core marketplace analyst - to ensure the healthy liquidity on your platform
- Ecosystem analyst - to identify competitive threats and strategic opportunities
- Machine-learning analyst - to ensure healthy operation of the algorithms that power your product

Question 3: Outline the various phases of a typical data science methodology

Answer - The Data Science Methodology is an iterative system of methods that guides data scientists on the ideal approach to solving problems with data science, through a prescribed sequence of steps.

1) Data Acquisition - For doing Data Science, you need data. The primary step in the lifecycle of data science projects is to first identify the person who knows what data to acquire and when to acquire based on the question to be answered. The person need not necessarily be a data scientist but anyone who knows the real difference between the various available data sets and making hard-hitting decisions about the data investment strategy of an organization – will be the right person for the job.
2) Data Preparation - Often referred as data cleaning or data wrangling phase. Data scientists often complain that this is the most boring and time consuming task involving identification of various data quality issues. Data acquired in the first step of a data science project is usually not in a usable format to run the required analysis and might contain missing entries, inconsistencies and semantic errors.
3) Hypothesis and Modelling -This is the core activity of a data science project that requires writing, running and refining the programs to analyse and derive meaningful business insights from data. Diverse machine learning techniques are applied to the data to identify the machine learning model that best fits the business needs. All the contending machine learning models are trained with the training data sets.
4) Evaluation and Interpretation - There are different evaluation metrics for different performance metrics. For instance, if the machine learning model aims to predict the daily stock then the RMSE (root mean squared error) will have to be considered for evaluation. If the model aims to classify spam emails then performance metrics like average accuracy, AUC and log loss have to be considered. Machine learning model performances should be measured and compared using validation and test sets to identify the best model based on model accuracy and over-fitting.
5) Deployment - Machine learning models might have to be recoded before deployment because data scientists might favour Python programming language but the production environment supports Java. After this, the machine learning models are first deployed in a pre-production or test environment before actually deploying them into production.
6) Operations/Maintenance - This step involves developing a plan for monitoring and maintaining the data science project in the long run. The model performance is monitored and performance downgrade is clearly monitored in this phase. Data scientists can archive their learnings from a specific data science projects for shared learning and to speed up similar data science projects in near future.
7) Optimization - This is the final phase of any data science project that involves retraining the machine learning model in production whenever there are new data sources coming in or taking necessary steps to keep up with the performance of the machine learning model. Having a well-defined workflow for any data science project

is less frustrating for any data professional to work on. The lifecycle of a data science project mentioned above is not definitive and can be altered accordingly to improve the efficiency of a specific data science project as per the business requirements.

Question 4: Mention 4 tools that a data scientist can rely on to effectively deliver his/her work

Answer

1) Python

2) PowerBI

3) Excel

4) R