**MORNING CHALLENGE- WK3**

**QUESTION 1: What is the difference between exploratory data analysis and predictive data analysis?**

What is Exploratory data analysis?

Data Exploration is essentially the first step in any data analysis. It involves summarising the main characteristic of any database or dataset. It’s mostly done in statistical software varying in advancement levels, depending upon the complexity of the dataset. It can be conducted using visual analytics tools as well; this is more preferred, since it enables users to have a quick and simplified look at the most important featured and variables of the dataset. It helps the user quickly determine if the variables on screen are good enough for further analysis.

What is Predictive data analysis?

Predictive Analysis falls under advanced analytics, and is used to make predictions about unknown events that might unfold in the future. In the simplest manner possible, Predictive analysis uses a host of different software, pairing them with many techniques ranging from artificial intelligence (AI), and machine learning, to analyze already existing data and make predictions regarding its course in the future.

The difference, though not immediately obvious, is quite great. Data Exploration uncovers the complex, intricate, and oftentimes invisible relationships between measurable variables. Predictive analysis on the other hand offers outcomes and possibilities in the future of the variables, from the variables.

**QUESTION 2: How would you define the role of a Data Scientist in Product Development Team**

* Selecting features, building and optimizing classifiers using machine learning techniques
* Data mining using state-of-the-art methods
* Extending company’s data with third party sources of information when needed
* Enhancing data collection procedures to include information that is relevant for building analytic systems
* Processing, cleansing, and verifying the integrity of data used for analysis
* Doing ad-hoc analysis and presenting results in a clear manner
* Creating automated anomaly detection systems and constant tracking of its performance

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

**Step 1: Frame the problem**: The first thing you have to do before you solve a problem is to define exactly what it is. You need to be able to translate data questions into something actionable.

**Step 2: Collect the raw data needed for your problem**

Once you’ve defined the problem, you’ll need data to give you the insights needed to turn the problem around with a solution. This part of the process involves thinking through what data you’ll need and finding ways to get that data, whether it’s querying internal databases, or purchasing external datasets.

**Step 3: Process the data for analysis**

Now that you have all of the raw data, you’ll need to process it before you can do any analysis. Oftentimes, data can be quite messy, especially if it hasn’t been well-maintained. You’ll see errors that will corrupt your analysis: values set to null though they really are zero, duplicate values, and missing values. It’s up to you to go through and check your data to make sure you’ll get accurate insights.

**Step 4: Explore the data**

When your data is clean, you should start playing with it!

The difficulty here isn’t coming up with ideas to test, it’s coming up with ideas that are likely to turn into insights.

**Step 5: Perform in-depth analysis**

This step of the process is where you’re going to have to apply your statistical, mathematical and technological knowledge and leverage all of the data science tools at your disposal to crunch the data and find every insight you can.

**Step 6: Communicate results of the analysis**

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

* Microsoft Excel
* Python
* R
* Tableau Software