Name: Somesh Ramesh Ghaturle ID#: U01952080
Username: SG07981N

Assignment - 1

Conceptual Questions

Q1. What is data science? How does it relate to and differ from statistics?

Ans: - Data science is the application of scientific methods, procedures, algorithms, and systems to extract knowledge and insights from both organized and unstructured data sets. It analyzes and interprets complicated data using a combination of disciplines such as statistics, computer science, and information science. Statistics is a mathematical topic that provides programmatic tools and methods - such as variance analysis - for collecting data, designing experiments, and performing analysis on a given set of figures to assess a characteristic or determine values for questions.

The relationship between data science and statistics is that data science is an applied subset of statistics that employs statistical methods to analyze vast volumes of data and better comprehend the results. Statistics is essentially a theoretical discipline that develops tools for making sense of data and acting under uncertainty, but data science is a broader phrase that encompasses the application of techniques from statistics and many other fields to achieve goals.

Reference Links – 1 2

Q2. Identify three areas or domains in which data science is being used and describe how?

Ans: - Following are the few domains in which data science is being used.

1. **Fraud detection** is a vital use of data science in many sectors, as it thwarts financial losses and safeguards businesses and customers. Let's delve into data science's significance in fraud detection in various domains:

    - **Financial Sector:** In finance, data science is pivotal for spotting credit card fraud, detecting unusual transactions, and reducing unauthorized access risk. Advanced machine learning algorithms can identify suspicious activities in real-time by analyzing transaction history and customer behavior.
    - **E-commerce:** Online businesses depend on data science to identify fraudulent orders. They analyze customer behavior and transaction history to differentiate between legitimate and fraudulent purchases, protecting against credit card fraud and payment scams.

2. **Internet Search**: In the age of the internet's central role in our daily lives, search engines are crucial tools for accessing information. Complex algorithms and data science techniques underlie these search engines, making the internet more accessible and useful. We will explore how data science plays a vital role in internet search and impacts the content we find online.
    - **Search Engines and Data Science**: - Backbone of Search: Google, Bing, and Yahoo are the initial stops for people seeking information online. Data science is the foundation of these search engines, efficiently managing vast web data.

    I. **Data Processing**: Data scientists create systems to process and index web pages. They design algorithms to gather and assess web content for search engines to provide high-quality results.

    II. **Scalability**: Data science ensures search engines can handle billions of web pages. Techniques are developed to swiftly and accurately crawl and index immense amounts of data.

3. **Targeted Advertising** In the realm of modern advertising, precise audience targeting is crucial for successful marketing. Data science plays a pivotal role in this. Let's explore how it's changing the way brands connect with their customers.

    - **Data Science's Impact on Advertising**

    1. **Customer Segmentation:** Data science dissects audiences into detailed segments. Marketers analyze vast datasets to identify customer groups by demographics, interests, and behavior. This helps tailor ads for specific customer personas, making them more relevant.
    2. **Predictive Analytics:** Predictive analytics forecasts customer behavior and preferences. Machine learning predicts what products or services might interest a customer, allowing precise targeting.
    3. **Dynamic Content Optimization:** Data-driven platforms customize ad content in real time based on user profiles. Ad creative, copy, and formats adapt to maximize relevance and click-through rates.

Reference Links - 1

Q3. If you are allocated 1TB data to use on your phone, how many years will it take until you run out of your quota of 1 GB/month consumption?

Ans: - If I consume 1 GB of Data per month, it will take me lot of time to finish 1 TB because of the following points.

1. 1 TB is equal to 1024 GB so, if I choose 1 GB per month, it will take:
2. 1024 GB / 1 GB per month = 1024 months
3. To convert 1024 months into year we will divide by 12:
4. 1024 months / 12 months per year = ~ 85.33 years

So, it will take approximately 85.33 years to use up 1 TB of data at a rate 1 GB per month.

Q4. We saw an example of bias in predicting future crime potential due to misrepresentation in the available data. Find at least two such instances where an analysis, a system, or an algorithm exhibited some sort of bias or prejudice.

Ans: - Bias about crime prediction associated with predictive policing, specifically considering the effect of algorithmic bias on efficient technology installation in society and the legal implications of predictive policing. Or just being not fair and supporting one specific category without any solid justification.

1. **Amazon's Recruiting Algorithm**: Amazon developed a machine learning system for recruiting, but it was found to be biased against women. The system was trained on resumes submitted to Amazon over a 10-year period, most of which came from men. This led the system to favor male candidates over female candidates. Reference Link - 1.

2. **Bias in Research**: Suppose you are researching whether a particular weight loss program is successful for people with diabetes. If you focus purely on whether participants complete the program, you may introduce bias into your research. For example, participants who become disillusioned due to not losing weight may drop out, while those who succeed in losing weight are more likely to continue. This could bias the findings towards more favorable results. Reference Link – 1.

Hands on problems 1.2, 1.3

Problem 1.2

The following table contains an imaginary dataset of auto insurance providers and their ratings as provided by the latest three customers. Now if you had to choose an auto insurance provider based on these ratings, which one would you opt for?

#	Insurance provider	Rating (out of 10)
1	GEICO	4.7
2	GEICO	8.3
3	GEICO	9.2
4	Progressive	7.4
5	Progressive	6.7
6	Progressive	8.9
7	USAA	3.8
8	USAA	6.3
9	USAA	8.1

Ans: - Based on Given Data. Let's, Calculate the average rating for the available Insurance provider as we have rating for each provider 3 times. So, we will be dividing the rating by 3.

Here we go,

3. GEICO:
   Available Ratings: 4.7, 8.3, 9.2
   Average Rating: (4.7 + 8.3 + 9.2) / 3 = 7.4

4. Progressive:
   Available Ratings: 7.4, 6.7, 8.9
   Average Rating: (7.4 + 6.7 + 8.9) / 3 = 7.6

5. USAA
   Available Ratings: 3.8, 6.3, 8.1
   Average Rating: (3.8 + 6.3 + 8.1) / 3 = 6.0

Based on the Average Rating Calculations, Progressive has the best and highest Average rating out of 10, followed by GEICO with an Average rating of 7.4, and USAA at last with an Average rating of 6.0.

Hence, If I wanted to choose an Auto Insurance Provider Based on the Available Rating, I will go for **Progressive** Insurance Provider.

Problem 1.3

Imagine you have grown to like Bollywood movies recently and started following some of the well-known actors from the Hindi film industry. Now you want to predict which of these actor's movies you should watch when a new one is released. Here is a movie review dataset from the past that might help. It consists of three attributes: movie name, leading actor in the movie, and its IMDB rating. [Note: assume that a better rating means a more watchable movie.]

Leading actor	Movie name	IMDB rating (out of 10)
Irfan Khan	Knock Out	6.0
Irfan Khan	New York	6.8
Irfan Khan	Life in a … metro	7.4
Anupam Kher	Striker	7.1
Anupam Kher	Dirty Politics	2.6
Anil Kapoor	Calcutta Mail	6.0
Anil Kapoor	Race	6.6

Ans: - Based on Given Data. Let's, Calculate the average rating for the available IMDB Rating as we have 3 Movies for Leading actor Irfan Kan and Anupam Kher 2 Movies and Anil Kapoor 2 Movies. **We have Different Number of data available for Rating just not to be Biased or being fair I am choosing best 2 Movies of Irfan Khan which is New York and Life in a … Metro**. Therefore, At End we can have an average and divide that by 2 as we are having 2 samples for other actors.

Here we go,

1. Irfan Khan
   Available Rating: 6.8, 7.4
   Average Rating: (6.8 + 7.4) / 2 = 7.10

2. Anupam Kher
   Available Rating: 7.1 + 2.6
   Average Rating: (7.1 + 2.6) / 2 = 4.85

3. Anil Kapoor
   Available Rating: 6.0 + 6.6
   Average Rating: (6.0 + 6.6) / 2 = 6.30

Based on the Average IMDB Rating Calculations, **Irfan Khan** has the best and highest Average rating 7.10 out of 10, followed by Anil Kapoor with an Average rating of 6.30, and Anupam Kher at last with an Average rating of 4.85. As, We Try to remove the biasness of data so there might be possibility we could expect different result whatever we are expecting right now.

Hence, If I wanted to choose as per my Decision on Based on the Available Rating and being fair, I will go with Irfan Khan Movies.