Midterm Review

1. **Data Science**
2. What is Data Science?

Data science continues to evolve as one of the most promising and in-demand career paths for skilled professionals. Today, successful data professionals understand that they must advance past the traditional skills of analyzing large amounts of data, data mining, and programming skills. In order to uncover useful intelligence for their organizations, data scientists must master the full spectrum of the data science life cycle and possess a level of flexibility and understanding to maximize returns at each phase of the process.

Effective data scientists are able to identify relevant questions, collect data from a multitude of different data sources, organize the information, translate results into solutions, and communicate their findings in a way that positively affects business decisions. These skills are required in almost all industries, causing skilled data scientists to be increasingly valuable to companies.

1. Describe in your own terms the technological advances and other phenomena that have caused the Big Data phenomenon.

* There are many factors that contributed to the emergence of today's big data ecosystem, but there's a general consensus that big data came about because of a range of hardware and software designs that simply allowed big data to exist.

Big data sets are often identified as data sets that can't fit into a simple database network, because their analysis requires too much work on the part of the servers handling the data.

* Another advance accommodating big data was changes in the ways that handlers processed data sets. Rather than linear processing through a conventional relational database design, handlers started to use tools like HADOOP and related hardware management pieces to eliminate bottlenecks in data processes.
* The result is the big data world that we live in, where massive data sets are stored and maintained in data centers, and increasingly accessed by a wide range of technologies for a wide range of uses. From commerce to ecology, from public planning to medicine, big data is becoming more and more accessible. Meanwhile, government agencies and other larger organizations are still pushing the boundaries of big data sizes and implementing even more advanced solutions.

1. Describe in your own terms what you think are the important skill sets needed to be a data scientist. Elucidate this by drawing on your own experience in carrying out the assignments in this course or other experience you have had outside this course.

Data Visualization will be the major set of skills for the data science process, which basically means that after collecting and modelling the data within various data frames, it will then be visualized for conclusions to be made.

Statistics will help find meaningful trends amongst the data. The main goal to be identifying patterns and trends within amongst data received.

Machine learning will be using the statistical models to perform tasks with instructions. Understanding the instructions and algorithms within Machine learning models will come from skills obtained in the Computer Science and Math areas.

These computational methods may become too complex to explain to non-technical audiences. The authority to articulate and present such models of predictive power is constituted in the Communication and Presentation Skills.

Domain Expertise is given the least priority as it will only broaden its horizons once the rest of the skillsets have been obtained and mastered.

(A skill I would like to add into this Data Science Profile would be: Cloud Computing. When dealing with such immense datasets, Cloud Computing would enable users to rent and incorporate data servers, storage, and computing power from providers. In my perspective, although this entity would rather come under Computer Science and Data Visualization principles, knowledge, and implementation regarding Cloud Computing too, should be classified as a skill. It is evident to keep up with the necessary features that circulate the field of Data Science.)

1. **Exploratory Data Analysis and the Data Science Process**
2. Describe in plain terms what Exploratory Data Analysis is. List the kind of things that can be achieved by carrying out EDA.

* State of Flexibility

**John Tukey – Father of Data Analysis (statistician)**

It may mean a technical way but EXPLORE what data could have and what it could be (hidden). Willing to find out things that are believed to not be there and we look for them.

* An approach for data analysis that employs a variety of techniques to maximize insight of a dataset (explore relationships and patterns, distribution)
* To uncover an underlying structure (distribution of structure of the data). Feature extraction. (detect important variables)
* Detect outliers and anomalies. (things that stand out)
* Develop model: come up with sensible models based on the behavior of data.

1. Contrast EDA against Confirmatory Data Analysis (e.g. in terms of what the focus is, what the techniques are, whether or not an assumption is made.)

* CDA: Concerns itself with modelling and hypothesis. And then statistical confirm things. (we have models and things to confirm). It focuses on the model. Quantitative analysis. ASSUMPTIONS ARE MADE

Problem -> Data -> Model -> Analysis -> Conclusions

* EDA: There is no model or hypothesis (the understanding of the problem is changing as we go). Understanding the flexibility and broadening our willingness to find out things that are not directly visible. (flexible to change your mind too.) (ask better questions based on the changing question). It focuses on the data. Graphical analysis. ASSUMPTIONS ARE NOT MADE.

Problem -> Data -> Analysis -> Model -> Conclusions

1. Discuss some of the basic tools that can be employed to carry out EDA.

* Plots: Plotting distributions of all variables (box plots), plotting time series data (economics business side of data based on daily, monthly updates of data e.g. electricity usage), transforming variables(that would depict something, that would have crossed the mind nevertheless), looking for pairwise relationships between variables using scatter plots(look for correlation of data in a visual manner). It will help guide to MODEL the data in the later stages.
* Graphs:
* Summary Statistics: Reduce the messy data into a few numbers that will tell you a great deal about the data. Min/Max/Median/Mean/S. D/Identify outliers/quartiles. R programming is good for this^
* EDA IS ALSO A MINDSET: Goal is to understand the data- gain intuition (understand the process that generated the data)

1. Give a schematic description of the Data Science Process and discuss how its components interact. Explain the significance of EDA in the data science process.

Diagram

Description automatically generated

This Data Science process is extracted from seeing what people/companies/industries do i.e. seeing their day to day activities and some of them are industry driven.

The broken line from BDP to The Real World is a feedback. Ethics: E.g. Weather vs Hiring app machine learning model.(that’s why we need feedback to ethically stay on track).

^Visualization in this process can come either directly from the raw data, or after Machine learning algorithms.

^When you are done with the machine learning model, it can/will become a product(data). And that data product can be used as a recommender system.

Patterns observed in data may indicate something wrong or right in the logging process that needs to be fixed. EDA makes sure that the product is performing as intended. Understand user behavior, customer behavior. Determine the obvious outliers.

Using EDA can make you go back to collect the raw data or can help you move forward.

Use the data that is left behind to benefit the company.

1. Describe in what ways EDA helped you to either (i) define a problem or (ii) develop a model in the project or the assignments you carried out in this course.

Ethics: E.g. Weather vs Hiring app machine learning model.

EDA HELPED PREVENT OUTLIER DETECTION BY DATA ENTRY MISTAKES, ERRORS IN SCRAPING.

1. **Data Wrangling**
2. Use and explain the following functionalities of the data transformation tool dplyr: – the five “verbs” – filer, select, arrange, mutate, summarize – Group-by and ranking functionalities

* Dplyr is a tool or library to make data exploration and data manipulation faster. Lesser code (especially using chaining syntax) and More efficient performance. Fast on data frames
* Chaining or pipelining means cntl + shift + m (more convenient way to write when there are many lines, readability**) (%>%)**
* Filter: Find, among the rows, the ones that would satisfy certain criteria. (create subsets) (row selection).

filter (flights, Month == 1, DayofMonth== 1)

* Select: When we want to select only certain variables. (column selection)

select (flights, DepTime, ArrTime)

select (flights, Year, contains(“Taxi”), contains(“Delay”))

* Arrange: Reordering the rows depending on certain criteria.

Flights %>%

select (UniqueCarrier, DepDelay) %>%

arrange (DepDelay)

* Mutate: It will add new columns at the end of your dataset. Does not change the original data frame. (if you only want to keep the new ones, use transmute)

mutate (flights, gain = dep\_delay – arr\_delay)

* Summarise: it is typically paired with group by

byday <- groupby (flights, year, month, day)

summarise (by\_day, delay = mean(dep\_delay)

* Ranking: WINDOW FUNCTIONS

1. Explain what the notion of “tidy” data means

Data is untidy when: One variable is spread across many columns, or observation is spread across many rows, or one column contains two variables, or NEED TO COMBINE VARIABLES to for the welfare of the data.

TIDYING DATA helps organize tabular data. Each variable (something we would observe) takes its own column. Each observation has its own row. It makes variables easy to access as vectors (which R is really good at). It also preserves cases during vectorized operations.

When dealing with missing values. (methods -> Mean Value Imputations, Heuristic based Imputation, Random Value Imputation (allowing statistical evaluation, random prevents any bias), Imputation by nearest neighbor (in classification), Imputation by interpolation (linear regression, better off estimating values)

* + - drop\_na (x, x2)
    - fill (x, x2)
    - replace\_na (x, list (x2 = 2)

1. Use and explain the following functionalities of the tool tidyr – gather, separate, and spread

Tidyr: Changing the shape of the data. (python would use pandas)

* Gather: moves column names into a key column, gathering the column values into a single valued column.

gather (data, key, value)

gather (table1, ‘1999’, ‘2000’, key = year, value = ‘cases’)

* Separate: separate/split each cell in a column to make several columns. (the opposite is unite)

separate (table3, rate, into = c (“cases”, “pop”))

* Spread: moves the unique values of a key column into the column names, spreading the values of a value column across the new columns.

spread (table2, type, count)

1. Explain and demonstrate how to work with multiple tables (joins) in dplyr

Collective multiple tables of data are called relational data. These entities have to be modelled in a certain manner in order to be consistent. Manipulate and work with them.

* Mutating Joins:
  + Inner Join: First matches by their keys, then copies across variables from one table to the other
  + Outer Join: left join (keeps all observations in x), right join(keeps all observations in y), full join(keeps all observations in x and y)
* Filtering Joins:
  + Semi join (x, y): keeps all observations in x that have a match in y
  + Anti-join (x, y): drops all observation in x that have a match in y

1. **Data Visualization**
2. Describe visualization principles
3. Tufte’s visualization aesthetic
4. Design and color
5. Telling story with data
6. Visualizing patterns over time; visualizing proportions; visualizing proportions; visulaizing relationships; visualizing text information
7. Mapping