Personal Manifesto

By: Varshini Rana

Table of Contents

Week 1: Problem Formulation Stage

Informational Interview - Planning

For the Informational Interview, I will be collecting an interview transcript from the book Data Scientists at Work by Sebastian Gutierrez, which is a collection of sixteen interviews of renowned data scientists. The interviewee shall be Eric Jonas, a research scientist in computational neuroscience and signal processing, who is currently a Professor in the Department of Computer Science at the University of Chicago. I chose him as the interviewee for this task because the field of computational neuroscience is something that I have always been intrigued by, and I believe that his industrious work in this field makes him an exemplary authority to talk about it. I am very keen on understanding how he leverages neuroscience data to determine how the brain learns, forms memories and assimilates new information.

Reading Responses

Readings:

- **Chapter 2 - Business Problems and Data Science Solutions**

1. “Data mining results should influence and inform the data mining process itself, but the two should be kept distinct.” The data mining process to build a model and the way the model (the results of data mining) is used are distinct in that the process of building a model is often with historical data, while the model so built takes existing or new data points as an input to provide a prediction.
2. “Data mining is a craft.” While data mining involves a fair amount of science and technology, the most effective insights from data mining require quite a bit of creativity to bring it all together. Like any craft, a well-structured approach to data mining with some leeway to allow for creative conceptualization would go a long way in solving the problem in question.

- **Chris Wiggins interview**

1. “The key is usually to just keep asking, “So what?”” Thinking through how your actions impact something that you value by asking “so what?” after every action that you undertake enables you to streamline your thought processes to be clear about what you’re working towards and what you want to accomplish.
2. “The way you become an expert in a field is to make every mistake possible in that field.” Making every possible mistake gives you exposure to every possible scenario where failure occurs so that you’re better prepared for the next time you encounter such scenarios and rise up to the challenge by providing an optimum solution.

- **Erin Shellman interview**

1. “Data’s just the world making noises at you.” Raw data is often messy, crude, and unintelligible. The insights that you can glean from it after massaging it in an appropriate way is what makes the data valuable.
2. “If you talk to somebody who has something you want, follow up.” The field of Data Science often demands the ability to ask for what you want and to be persistent in the pursuit of it if you truly want to obtain it. This would impress upon the mind of the person whose help you wanted about your seriousness and commitment to seeing your project through, in which case they would be more inclined to give you what you want.

- **Jake Porway interview**

1. “I still believe that a data scientist is just a statistician who can program well.” A statistician is able to leverage mathematical techniques to analyse and draw key conclusions from data. A data scientist just extends that thinking by collecting, modeling,

visualising, and drawing meaning from data through the judicious use of statistical formulae and computer algorithms, among other things.

2. “There’s almost no limit to where data and data science can be applied.” Every single action (or sometimes even inaction) of ours, who we are, what we do, how we do it, and everything else is a possible data point for someone who looks closely enough and wants to achieve a certain goal related to it.

Plan for Knowledge Acquisition

Skills and Knowledge Inventory: Stage 1, Problem Formulation

1. **how to conduct an inquiry in my application domain that leads to a good problem formulation**

I already have this capability. I acquired this capability during my two years of work as an Associate Data Scientist at my start-up, most notably on a fraud detection project I was working on in the domain of Health Insurance. During the initial stages of project development, I had to pick the brains of the leaders and members of the fraud investigation team — many of whom were former practicing physicians — of the health insurance companies who had partnered with my start-up. Long discussions were had on the numerous pain points they experienced during the course of their investigations, the available data and any other data they could start collecting during the course of claim processing which they thought would help me in developing an intelligent fraud detector, insights gleaned from their domain expertise, and tiny nuances they pointed out which I would've missed out on by just looking at the data alone. All of these discussions were immensely valuable and indispensable in helping me formulate an actionable problem statement.

2. **a repertoire of problem types**

I do not have this capability yet as I'm still at the budding stage of my Data Science career with exposure to mainly Classification (or Class Probability Estimation) problem types, although I am aware of the existence of other problem types. I plan to acquire this capability through a new job I accepted in an MNC starting in February 2022, which shall expose me to various domains such as Finance, Retail, and Manufacturing. I expect to encounter and work on a variety of problems such as those related to causal modeling, similarity matching, profiling, and co-occurrence grouping among others. I also plan to develop my repertoire of various problem types through MADS coursework, which includes courses such as SIADS 543: Unsupervised Learning, SIADS 644: Reinforcement Learning Algorithms, SIADS 652: Network Analysis, and SIADS 685: Search and Recommender Systems. The Milestone and Capstone project-based MADS coursework shall also go a long way in helping me develop my repertoire of various problem types in the field of Data Science.

3. **how to map problems in my application domain to the repertoire of problem types**

I already have this capability. I acquired this capability during my two years of work as an Associate Data Scientist at my start-up, involved in building applications mainly in the Healthcare and Health Insurance domains. While my hands-on exposure is currently limited to a

small subset of the repertoire of problem types, I am aware of how to map them with the problems in my domain thanks to various conversations with mentors involved in this line of work. For example, the prediction of the probability that a person with Chronic Kidney Disease would succumb to complications based on lab test values and frequency of dialysis treatments is a class probability estimation problem. In this case, preventive interventions can be put forth to forestall the deterioration of health. Another example would be link predictions in computerized drug discovery, wherein large volumes of biomedical data represented as a network become useful for predicting results from drug-drug and drug-disease interactions, which leads to faster and more efficient drug discovery.

Application in Domain of Interest

Domain: Healthcare

Project 1 Description: The creation of an image classification model of chest X-rays for the detection and accurate diagnosis of pneumonia.

Project 1 Problem Type (with explanation): This is a classification (or class probability estimation) problem. The image classification model would predict either the presence or absence of pneumonia based on lung opacity (an indicator of the presence or absence of lung inflammation) as seen on chest x-rays of pneumonia patients and healthy control respectively.

Problem 2 Description: Information extraction (named entity recognition) of clinical words, phrases and standard medical codes (such as those for diseases and procedures) from unstructured clinical text such as discharge summaries to make health insurance claim processing faster and more efficient.

Project 2 Problem Type (with explanation): I would classify this as an optimization problem. Currently, due to the lack of availability of Electronic Health Record (EHR) resources in my home country, the process of perusing through discharge summaries to determine the cause of admission of the patient to the hospital is being done manually by health insurance claim processors. This is a very cumbersome and time-consuming process. This results in an increase in health insurance claim processing time, money spent on resources, and also the probability of human error. Automatic information retrieval would do away with all of these pitfalls.

Questions, Maxims, and Commitments

Question (I will always ask…)
Whom will these results benefit?

1-Sentence Project Description
The creation of an image classification model of chest X-rays for the detection and accurate diagnosis of pneumonia.

Meaning in Context
In this case, it is possible that these results will benefit radiologists or doctors in making accurate diagnoses, which they would've otherwise missed out on due to microscopic subtleties in the X-rays which are invisible to the naked eye and the possibility of human error. These results also have direct implications and benefits for patients undergoing this radiological procedure as these ensure that they receive timely and faultless treatment for the ailment so detected.

Importance for this stage of the project
Knowing who is going to consume these results and how it will benefit them is probably the most important consideration in the problem formulation stage as it will give the data scientist an idea of the return on investment he will obtain if he undertakes this project. By knowing the target market for the results of his project, the data scientist can ensure that he directs his attention towards the goal of making things easier for them (radiologists and doctors in this case) and can anticipate the potential compensation he might receive if he chooses to undertake this project.

Maxim (I will always say…)
A good data science problem will aim to provide actionable insights, not just make predictions.

Which Project
The creation of an image classification model of chest X-rays for the detection and accurate diagnosis of pneumonia.

Meaning in Context
If it is determined that the results of the predictive model point to a patient having pneumonia, he should be provided the appropriate medical treatment for it. The results of this particular project are such that they demand immediate action and empower the stakeholders to make informed decisions, as opposed to a prediction with just theoretical implications.

Importance for this stage of the project
Formulating a problem statement to work on by considering the possible actions that would be taken based on the potential results would give the data scientist an assurance that the results of his work would have the impact that he intended (i.e. in this case, his prediction of pneumonia or lack thereof are useful and are being taken seriously), since appropriate actions are being taken based on the insights that he generated.

Professional/Ethical commitment (I will always/never...)

I will always be aware of and respect the licensing status of the data that I want to use in my project.

Which Project

The creation of an image classification model of chest X-rays for the detection and accurate diagnosis of pneumonia.

Meaning in Context

In this case, the project would make use of chest X-rays from patients for training an image classification model. Protecting the privacy of these patients and adhering to the terms of fair use of this data is the ethical responsibility of the data scientist who intends to use this data. Misuse of this data or non-compliance to the specifications of the data license would cause a breach of confidentiality, which could have serious legal, social and financial consequences.

Importance for this stage of the project

The stage of problem formulation is the stage where the data scientist starts thinking about the kind of data he would require for his project and where to obtain it from. It is important that he ensures that the data he chooses to use is legally authorised for his use. In this case, since the data in question is PHI or Protected Health Information, the sensitive nature of the data makes it imperative that he obtain the necessary permissions to use the data as intended beforehand. If he isn't able to obtain the necessary permissions, he may have to source permitted data from elsewhere or rethink the entire project. Hence, being aware of the licensing status of the data and how he hopes to adhere to it also enables him to plan for the project.

Week 2: Data Collection and Cleaning Stage

Potential Personal Project Tweet

“Can I help prevent pneumonia misdiagnoses by guiding doctors with the power of artificial intelligence?” This was the thought that ignited my endeavor to create an image classification model using chest X-ray images for pneumonia detection, which I would obtain from Kaggle.

Character count (including spaces): 275

Reading Responses

- **Law of Small Numbers**

1. “Sustaining doubt is harder work than sliding into certainty.” Being unsure about something often requires more mental work than definitively coming up with a decision, as our minds have a natural inclination to gravitate towards conclusiveness.
2. “We are far too willing to reject the belief that much of what we see in life is random.” We often try to find patterns in seemingly random events and are unable to accept the fact that most of what we see in life is purely coincidental, leading to preconceived notions.

- **Statistical Biases Types Explained**

1. “Biased statistics are bad statistics.” Statistical analyses done with biases are not going to be generalizable for the cohort the study was intended to gain insights into, since the underlying assumptions the study was based on itself is erroneous.
2. “If you let the subjects of your analyses select themselves, that means that less proactive people will be excluded.” The subjects who volunteer for the analysis exhibit a certain type of behaviour which may not be representative of the entire population under scrutiny for that particular study.

- **Data Cleaning 101**

1. “Unless you have a good level of confidence in your assumption, you probably should not make the correction or else you may be creating a typo instead of correcting one.” While corrections we make in the data before using it may be well-intentioned, we shouldn’t be making undue modifications without certainty just because we can, as doing this may defeat the purpose of it by introducing erroneous assumptions.
2. “Don’t be afraid of picking up a phone or shooting an email to the source of the data.” It is necessary to be absolutely sure about the authenticity or correctness of the data your project will be based off of, which is why the process of reaching out to the sources of the data to confirm this is paramount.

- **10 Rules for Creating Reproducible Results in Data Science**

1. “All data science is research.” All science requires systematic inquiry and investigation to generate insights; data science encompasses the same endeavor, but with data.
2. “Manual data manipulation is hidden manipulation.” Manual data manipulation steps often aren’t reproducible. They are difficult to document, track, and repeat. Hence, they should be avoided and a script with general data manipulation steps can be used instead.

Plan for Knowledge Acquisition

Skills and Knowledge Inventory: Stage 2, Data Collection & Cleaning

1. **common problems with data sets that can lead to misleading results of analyses**

I already have this capability. I acquired this capability during my two years of work as an Associate Data Scientist at my start-up, where one of my responsibilities was to develop a fraud detection product in the domain of Health Insurance using structured claim data. The product consisted of a classification algorithm which can predict whether an insurance claim is “fraudulent” or “legitimate”. The initial results I obtained were extremely misleading. On closer inspection of the data, I realised that one of the features used for model training was a field for the “balance sum insured”, which I would typically not be aware of before a claim has been processed. By including this field, I was essentially “cheating” since “fraudulent” claims in the training data had a higher “balance sum insured” value than the “legitimate” claims due to not being used up since these claims were not successfully processed. Another issue I encountered with the data on initial inspection was that all the columns in the tabular data had been shifted by one column to the right, and so all the fields had incorrect headers. Extraneous commas in the CSV file were the reason for this issue and were promptly fixed. Other issues, such as missing values, duplicates, and inconsistencies within fields are all extremely common based on my experience.

2. **potential data sources in my application domain**

I already have this capability. During my tenure of two years as an Associate Data Scientist in a start-up, I was encouraged to explore and obtain structured data from the in-house OMOP Common Data Model, hosted on AWS Redshift, which contained anonymised healthcare data (received from clients) devoid of identifiable information. I learnt how to query the databases with SQL via tools such as Aginity Workbench and SQL Workbench, and export the structured data which I required for my analyses in the form of CSV files. I was also able to programmatically access and connect to these databases with Python libraries such as psycopg2. Moreover, I was able to use proprietary APIs to download unstructured data received from clients in the form of bills, lab reports, discharge summaries, and radiology reports as PDF files for my analyses. Externally, I have used and am familiar with several data sources in my application domain, such as Kaggle, data.gov, and the UCI Machine Learning Repository.

3. **how to understand and document data sets**

I already have this capability, owing to my tenure of two years as an Associate Data Scientist in a start-up, wherein I spent a huge chunk of my time scrutinizing datasets, playing around with

them, and trying to properly understand them before going ahead with my analyses. I have found that having a high-level overview of the data types and number of values of all of the fields in my datasets go a long way in determining whether everything is as expected. With pandas, I would just use a simple function such as pandas.DataFrame.info(), which would give me a quick summary of the data types and also the number of values (and missing values, if any) of each column in the dataset. This would help me figure out to some extent if each column has the intended values, and what intervention I would need to bring about, if any. Using a function such as pandas.DataFrame.head() would also show me the first 5 rows (along with column names) of the dataset, so that I can quickly see whether everything is as it should be. Similarly, other pandas capabilities would help me summarise, aggregate, and visualise the data so that I could get a glimpse of the data from other perspectives. During the course of my work, I have also had to document some datasets which were given to me by the client, so that others in my organization could use it for their own purposes. I have had to include comprehensive descriptions of the fields in the datasets, along with short accounts of how they were collected, and the possible values (or range of values) that would go into each field. I have also had to include a Data License which would authorise my associates to use this data with the client's consent.

4. **how to write queries and scripts that acquire and assemble data**

I already have this capability. I acquired this capability during my tenure of two years as an Associate Data Scientist in my start-up. I obtained structured data from the in-house OMOP Common Data Model, hosted on AWS Redshift, which contained anonymised healthcare data (received from clients) devoid of identifiable information. I learnt how to query the databases with SQL via tools such as Aginity Workbench and SQL Workbench, and export the structured data which I required for my analyses in the form of CSV files. I was also able to programmatically access and connect to these databases with Python libraries such as psycopg2. The Python requests module helps me to connect to proprietary APIs to download unstructured data received from clients in the form of bills, lab reports, discharge summaries, and radiology reports as PDF files for my analyses. I am also familiar with a pandas functionality pandas.read_html() which can extract relevant tables from the webpage in the URL that would be provided as an argument to the function.

5. **how to clean data sets and extract features**

I already have this capability, owing to my tenure of two years as an Associate Data Scientist in a start-up, wherein I spent a huge chunk of my time scrutinizing datasets, playing around with them, and trying to properly understand them before going ahead with my analyses. I am quite used to finding missing values, duplicates, and inconsistencies in datasets, and employing pandas functionalities such as pandas.DataFrame.dropna(), fillna(), drop_duplicates(), etc. to clean them efficiently. I am also well-versed in feature engineering, which involves creating new features from existing ones, and dropping redundant features. An example of this during the development of a fraud detection product was that I created a feature called "length of stay in

the hospital” by computing the difference of two features — “admission date” and “discharge date”. I have also employed Principal Component Analysis (PCA) and feature selection techniques such as forward selection and backward elimination to arrive at the most optimum set of features sans redundancy for model training.

Maxims, Questions, and Commitments

Question (I will always ask…)
How much is too much data cleaning?

Which Project
The creation of an image classification model of chest X-rays for the detection and accurate diagnosis of pneumonia.

Meaning in Context
The entire premise of the image classification model rests on the model being able to detect microscopic subtleties in chest X-ray images, which would otherwise be invisible to the naked eye. While we have a variety of computer vision techniques at our disposal to augment the images, sharpen them, modify the contrast, brightness, rotation, etc. while cleaning or correcting the images, we should be wary of knowing when it becomes too much, as these actions could have unintended repercussions and significantly affect the results, i.e. the microscopic subtleties in the X-ray images could get drastically altered and give the wrong impression.

Importance for this stage of the project
Data cleaning should be an extremely conscientious process. It should be done only when and only to the extent that there is some certainty that the resulting dataset would be accurate, free of bias, and without inconsistencies. This is done not only to ensure that the data is easy to work with and train, but also to have some level of confidence in the results of the model trained on it. Just because we have all the capabilities to correct or clean data, it doesn't mean that we absolutely must use all of them, at all times, and in all situations.

Maxim (I will always say…)
Manual data manipulations are not always reproducible.

Which Project
The creation of an image classification model of chest X-rays for the detection and accurate diagnosis of pneumonia.

Meaning in Context
Suppose I browse the dataset of chest X-ray images and find a bright red spot on one of the images. I would be tempted to remove this anomaly right then and there using an image editing tool, such as Photoshop, since it's quick and convenient. However, I may have missed fifty other images with similar anomalies since I didn't notice them while idly browsing through the images. This would significantly mess up my classification model as the wrong conclusions would be drawn. Moreover, any person picking up this project after me would be unaware of this anomaly since the correction was undocumented and done on the fly. Hence, it is necessary that any changes I make be done using a scripting language for better coverage, ease of documentation, and reproducibility.

Importance for this stage of the project
It is important to ensure that any modifications I make to the raw data at the data collection and cleaning stage of the project be reproducible so that there is some semblance of consistency in the process should it be applied on a repeat basis, and also to enable ease of documentation for the next person to follow, something that would next to impossible were the modifications done manually on the fly. Using a scripting language would help me with this and also make the data cleaning step generalizable for that particular type of data and the issues commonly observed in that data.

Professional/Ethical commitment (I will always/never...)

I will never settle for working with data whose only merit is that it just happens to be easily available, even though it isn't representative of the population I'm considering for my study.

Which Project

The creation of an image classification model of chest X-rays for the detection and accurate diagnosis of pneumonia.

Meaning in Context

Pneumonia could be of three different types — bacterial, viral, or mycoplasma pneumonia. Suppose I obtain data just for bacterial pneumonia and healthy control, and fail to find data for the other two types. While it was convenient to have this sample of pneumonia, it isn't representative of the entire gamut of pneumonia conditions which exist in reality. This could bias my model into detecting only bacterial pneumonia and failing to detect the other two types, which could lead to erroneous interpretations of whether a patient has pneumonia or not.

Importance for this stage of the project

Knowingly collecting only a specific subset of the population as a sample just because it was easy to access points to an underrepresented population, which could drastically affect the results of the model the data would be trained on when applied in a real-world scenario, and defeat the entire purpose of its development and deployment. Hence, it is important to avoid selection bias by either working to collect the other known representatives of the population, or to modify the problem statement to include and predict for only the types of data you have (for example, in this case, the labels for classification can be "bacterial pneumonia" and "healthy control", as opposed to just "pneumonia" and "healthy control"), in which case the results may or may not be useful depending on the situation.

Week 3: Data Analysis and Modeling Stage

Informational Interview - Reflection

The interview transcript of Eric Jonas was collected from the book Data Scientists at Work by Sebastian Gutierrez. Eric Jonas is a research scientist in computational neuroscience and signal processing. He is currently a Professor in the Department of Computer Science at the University of Chicago.

What intrigued me most about Jonas is that he is a forward-thinker. He is always on the lookout for the next problem that would keep him on his toes, and the problems that he chooses to contend with are often those which would have far-reaching consequences into the future. As a budding data scientist, this attitude is something that really inspired me into adopting this way of thinking.

As an industrious person with extensive experience not only in academia but also in the industry, Jonas provides a unique perspective of what success looks like from the lens of a PhD student, an entrepreneur, and an academic researcher. While the nuances differ, he maintains that it all really boils down to "what value are we providing?" **[question - W1 problem formulation]**. This added value could be monetary, related to ROI, something technical such as improving the predictive accuracy of a model for purposes downstream, or even an idea of what necessitates this venture in the first place, which is important to consider before choosing to work on a problem.

On being asked whether Jonas approaches a data problem model-first or data-first, Jonas insists on not going ahead with the project until he familiarises himself with the data first. He believes that knowing the right questions to ask helps in understanding the data **[maxim - W2 data collection and cleaning]** well enough to start thinking about further analysis and figuring out which modeling approach to employ. He believes that with modeling it is easy to change the approach and try different things, but if the data hasn't even been looked at and played around with first, it is easy to be led astray with fallacious assumptions.

Jonas also is of the opinion that a data scientist cannot just be someone who comes in, does the math, and leaves. He professes that the data scientist actually needs to care about the domain of the project **[professional commitment - W1 problem formulation]**. He states that a data scientist who just wants to apply the tools he has to solve the problem in a domain, but doesn't really have the knowledge or care to learn about the domain of its application would not only have a poor understanding of the problem, but would also not be willing to make the compromises necessary to understand how to guide his own work.

Three follow-up questions I would have liked to ask Jonas are:

1. Hypothetically, if you weren't a man of science, what would you be doing professionally instead?
2. What is the best challenge that you've faced so far, and how did you overcome it?
3. What is one advice that you would give to your past self of ten years ago?

Word count: 496

Reading Responses

- **Overfitting in Machine Learning: What is it and how to prevent it**

1. “Noise interferes with signal.” The machine learning model would start thinking of the irrelevant information (i.e. noise) in the dataset as a significant pattern and would make erroneous generalizations from it, masking the actual insights which could be gleaned from it.
2. “Simple learners tend to have less variance in their predictions but more bias towards wrong outcomes.” Algorithms with a simple, inflexible underlying structure tend to train models which are consistent but are more likely to be inaccurate on average.

- **Common pitfalls in statistical analysis: The perils of multiple testing**

1. “In any study, when two or more groups are compared, there is always a chance of finding a difference between them just by chance.” When it is hypothesized that no difference exists between the groups being compared (i.e. the null hypothesis), there is some probability of falsely rejecting it (i.e. the False-Positive Error rate or Type 1 Error) by finding some difference between them by chance.
2. “Results from single studies should not be used to make treatment decisions.” The researcher should try to gain insights from supporting data and other studies published by the scientific community to validate the results of his study for his insights to be actionable.

- **P-Hacking and the problem with Multiple Comparisons**

1. “So, to me the answer is (to) replicate yourself.” The initial results of the study, deemed to be exploratory and reported as such, can be attempted to be replicated with a new sample for validation and to exhibit the robustness of the results.
2. “Just remember–fishing is fine for tuna, bad for data analysis.” Seeking to obtain a particular result from data analysis which is unfounded or statistically unsupported in practice is a disingenuous way of “hacking” or “gaming” the system, which leads to unsubstantiated assumptions.

- **Correlation vs. Causation: An Example**

1. “Without randomized controlled trials, we cannot say one activity caused another.” A randomized controlled trial would include a randomized population being subjected to the activity under scrutiny and a control population to determine significant differences between the two groups. By not conducting this analysis, we cannot say indubitably that one trend causes another. At the most, we can claim that the trends are correlated.
2. “Humans naturally see patterns where they don’t exist, and we like to tell a cohesive story about what we think is going on.” We often try to find patterns in seemingly random events. We tend to crave structure and reject chaos in favour of something explainable, regardless of whether it is an incorrect assumption to make.

- **Simpson's Paradox in Real Life** or **Ignoring a Covariate: An Example of Simpson's Paradox**

1. "If the populations are separated in parallel into a set of descriptive categories, the population with higher overall incidence may yet exhibit a lower incidence within each such category." Although the overall performance metrics of a population considering all categories under scrutiny may be high, the performance of the individual categories may be lower than those of the other population being considered in the study.
2. "Simpson's paradox is not a contrived pedagogical example." The Simpson's paradox is something that occurs in real-life situations and is not just something that materialized as a teaching contrivance.

- **Conditioning on a collider**

1. "Causal inference from observational data boils down to assumptions you have to make and third variables you have to take into account." Observational studies cannot prove cause and effect by themselves, unlike randomized controlled trials. Hence, certain assumptions and other variables have to be considered while attempting to answer causal questions with only observational data.
2. "If you really care about a cause, don't give mediocre studies an easy time just because they please you." Don't give a lot of weight to studies that seem plausible on the surface but are actually just glorified studies which condition on colliders, which instigate a misrepresented association between two entities where none exists.

Plan for Knowledge Acquisition

Skills and Knowledge Inventory: Stage 3, Data Analysis & Modeling

1. **common mistakes in data analysis that lead to misleading results**

I already have this capability. I acquired this capability during my two years of work as an Associate Data Scientist at my start-up, where one of my responsibilities was to develop a fraud detection product in the domain of Health Insurance using structured claim data. The product consisted of a classification algorithm which can predict whether an insurance claim is “fraudulent” or “legitimate”. The initial results I obtained were extremely misleading. On closer inspection of the data, I realised that one of the features used for model training was a field for the “balance sum insured”, which I would typically not be aware of before a claim has been processed. By including this field, I was essentially “cheating” since “fraudulent” claims in the training data had a higher “balance sum insured” value than the “legitimate” claims due to not being used up since these claims were not successfully processed. I also made the rookie mistake of including all the available features in the training of my model, not realising that they were introducing a considerable amount of noise into the system, which resulted in overfitting. Some well-planned feature engineering and dimensionality reduction techniques fixed this issue. In a different project involving the prediction of the probability that a person with Chronic Kidney Disease would succumb to complications based on lab test values and frequency of dialysis treatments, I initially made the mistake of excluding outlier values based on misguided assumptions, which for this problem, had dire consequences. I realized that outliers need to be considered in the broader context of the analysis depending on the problem at hand.

2. **a repertoire of models and how to estimate, validate, and interpret each of them**

I look forward to strengthening this capability. I’m still at the budding stage of my Data Science career with exposure to mainly Classification (or Class Probability Estimation) problem types, which in turn exposed me to a variety of classification algorithms such as Logistic Regression, Decision Tree Classifier, Random Forest Classifier, XGBoost Classifier, K-Nearest Neighbors Classifier, Support Vector Machines, etc. Naturally, I learnt how to effectively interpret, estimate, and validate them in various ways. While I am aware of the existence of other model types, I do not have hands-on experience with training, evaluating, and interpreting them. I plan to acquire this hands-on experience through a new job I accepted in an MNC starting in February 2022, which shall expose me to various domains such as Finance, Retail, and Manufacturing. I expect to encounter and work on a variety of problems such as those related to causal modeling, similarity matching, profiling, and co-occurrence grouping among others, which shall naturally expose me to a variety of models. I also plan to develop my repertoire of various models through MADS coursework, which includes courses such as SIADS 542: Supervised Learning,

SIADS 543: Unsupervised Learning, SIADS 642: Deep Learning, SIADS 644: Reinforcement Learning Algorithms, SIADS 652: Network Analysis, and SIADS 685: Search and Recommender Systems. The Milestone and Capstone project-based MADS coursework shall also go a long way in helping me develop my repertoire of various models and algorithms in the field of Data Science. In addition to this, I shall also be perusing through newsletters, blog posts, and tutorials from online resources such as Machine Learning Mastery and Towards Data Science to keep myself abreast of the exciting developments in the field of Data Science.

Maxims, Questions, and Commitments

Question (I will always ask…)
What is the best algorithm to use for this problem?

Which Project
The creation of an image classification model of chest X-rays for the detection and accurate diagnosis of pneumonia.

Meaning in Context
There are many things to consider while choosing an appropriate algorithm in the context of this project. The underlying structure of the algorithm used in training the model should be such that it is able to support and work well with images (radiological images of chest X-rays in this case), and effortlessly detect subtleties in them which are invisible to the naked eye. Neural networks, such as Convolutional Neural Networks (CNNs) are the widely-accepted standard for image classification. Then comes the question of the architecture of the CNN, which we can build ourselves. If we are using one of the many pretrained models out there, such as VGG19, ResNet50, etc., we should think about fine-tuning them with our own data for more accurate predictions.

Importance for this stage of the project
After familiarizing ourselves with the data, the next step is to think about what the best approach would be in the context of analysis and modeling. When thinking about modeling, knowing to employ the right algorithm to train our data on (keeping in mind the capabilities of the model and the bias-variance trade-off) is paramount, as it would not only streamline our efforts in the right direction (as evidenced by similar efforts by other members of the scientific community trying to solve similar problems), but would also help us anticipate the level of effort we can expect to put in to solve the problem at hand. It would also give us an idea of the effort taken to evaluate and explain the results in a manner that actually answers the question that the project was intending to address.

Maxim (I will always say…)
Fishing is fine for tuna, bad for data analysis.

Which Project
The creation of an image classification model of chest X-rays for the detection and accurate diagnosis of pneumonia.

Meaning in Context
Suppose I split up my training dataset of chest X-ray images into multiple smaller datasets, each with something different being done to them under the guise of “exploratory analysis”. For example, I could enlarge the images of one set by some factor, I could rotate the images of another set by some degree, I could completely change the brightness of the images of a set such that it is completely different from that of the other sets, and so on and so forth. I could then train a model with all of these different datasets in turn hoping for a signal or some significant finding to pop up from at least one of them. These findings are not something I planned for, and are most likely spurious since I was “fishing” for them.

Importance for this stage of the project
By splitting up the dataset into smaller different datasets with distinct characteristics, I am likely to find something statistically significant from at least one of the smaller datasets. However, we would have no way of knowing if these results were due to the change I made on the dataset to separate it from the other ones, or merely due to an artifact of the sample, or a legitimate judgment call I made after talking to subject matter experts. Since I was “fishing” for some finding, any finding, without knowing what to expect beforehand, I could not say with any certainty that the results I obtained were completely within expected parameters. Such an approach would render my analysis immaterial as it would lead to fallacious assumptions.

Professional/Ethical commitment (I will always/never...)

I will always use hold-out test datasets and cross-validation to mitigate the problem of overfitting.

Which Project

The creation of an image classification model of chest X-rays for the detection and accurate diagnosis of pneumonia.

Meaning in Context

The image classification model I would create should be generalizable to make accurate predictions on real-world situations, i.e. with chest X-ray images never seen by the model before. Creating a hold-out test set and using cross-validation would help me validate the model and would give me some confidence that the model has not memorized the noise in the training dataset so much so that it provides wrong predictions for images that it hasn't been exposed to while training.

Importance for this stage of the project

The results of the model created would likely lead to an intervention of some sort, as it identifies potential patients of pneumonia who would then be admitted, subjected to further tests and treatment, etc., which results in considerable expenditure of time and money for all parties involved. Hence, it is necessary to ensure that the predictions are as accurate as possible. Overfitting would result in the model only making accurate predictions for the images in the training dataset and giving incorrect predictions for most of the images it has never encountered before. Therefore, it is essential to prevent overfitting by using techniques such as the usage of hold-out test sets and cross-validation to mitigate this problem.

Week 4: Presenting and Integrating into Action

Sources for Data Science News

> ***Instructions (delete before submitting)***
>
> *You will write a brief plan describing what sources of information about data science you plan to follow outside of assigned readings from this program. This could include blogs, podcasts, newsletters, conferences, or other sources. Present it as a short bulleted list, with a sentence describing why you plan to follow that source.*
>
> *When listing which resources you will use, be mindful of how many you are including. Too many resources will be unreasonable to keep up with. Too few resources will not keep you up to date with the industry.*

I plan to follow the following sources of information about data science to keep myself up to date with the industry:

-

Grading rubric:

- 1 point: Provides list of data science news sources.
- 1 point: Each source is accompanied by a short description of why it was chosen, will be useful, what it is, etc.
- 1 point: List is of a reasonable size (e.g. too many resources will be unreasonable to keep up with; too few resources will not keep you up to date with the industry.)

Reading Responses

Instructions (Delete these in your submission)

For each required reading, identify and explain two insights that you extracted from it. For each, make sure that you reference something from the reading and that you explain it in your words (it's OK to quote, but then explain).

Here are some examples:

1. *Tait observes that it is important to "avoid manual data manipulation steps." When you clean data by hand, it is not a reproducible step that others can use in the future to validate/repeat your work.*
2. *"Outcome proxies will be gamed." When you define proxies for the outcomes you really care about, people may start behaving in ways that obscure the natural correlations between the proxy and the real outcome of interest.*
3. *"Who will be using the results and for what decisions?" Knowing who's going to use the results and how they're expecting to use it may shape data collection, analysis, and implementation.*

Grading rubric (for each of two insights, for each reading):

- .5 point: articulates a meaningful insight that makes reference to something in the reading
- .5 point: correctly interprets the insight in the reader's own words

- ***A History Lesson On the Dangers Of Letting Data Speak For Itself***
- ***Storytelling for Data Scientists***
- ***Interpretability is crucial for trusting AI and machine learning***
- ***The Signal and the Noise, Chapter 2***
- ***The Signal and the Noise, Chapter 6***
- ***How Not to Be Misled by the Jobs Report***
- ***But what is this "machine learning engineer" actually doing?***
- ***How we scaled data science to all sides of Airbnb over 5 years of hypergrowth***

Plan for Knowledge Acquisition

Instructions (Delete these in your submission):

For each item below, select one of the following:

- *I already have this capability. If so, describe how you acquired it.*
- *I look forward to strengthening this capability. If so, explain how. Mention specific courses where you think it will be covered or outside activities you intend to engage in.*

Note: you only need 1-3 sentences for each, though you are welcome to write more if you want.

Grading rubric (for each of the capabilities for the week):

- 2 points: Describes how capability was already acquired OR
- 2 points: Explains plan for how capability will be acquired
- (Note: maximum of 2 points total; it's OK to describe both how you already learned something about this capability and your plans to learn more, but you can only earn points from one or the other of the two rubric elements).
- Possible deductions
    - -1 Seems to misunderstand the capability
    - -1 Plan is vague; doesn't look ahead to the curriculum or other outside resources to make a guess about where it might be covered.
    - -1 Description of how capability was acquired is vague (e.g., "I have it from my job")

Skills and Knowledge Inventory: Stage 4, Presenting & Integrating into Action

- **how to present results to domain experts who are not data scientists**
- **how to work with software engineers to put models into production**

Maxims, Questions, and Commitments

Instructions (Delete these when submitting)

As with any professional, every data scientist has certain beliefs about their work that define how they conduct themselves on a daily basis. Based on what you learn each week about the profession, we will ask you to identify and share beliefs that resonate with you in the form of questions, maxims, and professional (or ethical) commitments. You will have to provide one question, one maxim, and one commitment each week.

For each, you will provide:

- *A **one-sentence statement** of the question, maxim, or commitment.*
    - *Please be sure that it is relevant to the project stage that was covered that week (e.g., problem formulation in week 1).*
- *Which of your two projects from your Application in Domain of Interest you will apply it to. Please just include a one-sentence summary of the project; the reader can refer back to the full description.*
- *One paragraph explaining **what it means.***
    - *Please be sure to explain with respect to the particular context of the hypothetical project**.***
- *One paragraph explaining **why it is valuable** to ask that question, make that statement, or state that commitment. How would it make the particular project go better, or help you avoid some pitfall?*

Question (I will always ask…)
Grading rubric:

- 1 point: Provides a one-sentence question
    - .5 point deduction: Multiple questions rather than a single one.
    - 1 point deduction: The question is specific to the particular project, rather than a generic one that could be asked of any project.
- 1 point: Provides good one-sentence description of the project, and it is one of the two described above.
- 1 point: Provides a clear explanation of what the question means in the context of the project specified
- 1 point: Explains why it is valuable to ask the question by suggesting how it would make the particular project go better.
- 1 point: Question, as applied, applies primarily to the current stage of the project (presentation and action).

Which Project

Meaning in Context

Importance for this stage of the project

Maxim (I will always say…)

Grading rubric:

- 1 point: Provides a one-sentence maxim
    - .5 point deduction: Multiple maxims rather than a single one.
    - 1 point deduction: The maxim is specific to the particular project, rather than a generic one that could be asked of any project.
- 1 point: Provides good one-sentence description of the project, and it is one of the two described above.
- 1 point: Provides a clear explanation of what the maxim means in the context of the project specified
- 1 point: Explains why it is valuable to apply the maxim by suggesting how it would make the particular project go better.
- 1 point: Maxim, as applied, applies primarily to the current stage of the project (presentation and action).

Which Project

Meaning in Context

Importance for this stage of the project

Professional/Ethical commitment (I will always/never...)
Grading rubric:

- 1 point: Provides a one-sentence commitment
    - .5 point deduction: Multiple commitments rather than a single one.
    - 1 point deduction: The commitment is specific to the particular project, rather than a generic one that could be asked of any project.
- 1 point: Provides good one-sentence description of the project, and it is one of the two described above.
- 1 point: Provides a clear explanation of what the commitment means in the context of the project specified
- 1 point: Explains why it is valuable to articulate the commitment by suggesting what, in the context of the particular project, might create an incentive not to take the action you've committed to.
- 1 point: Commitment, as applied, applies primarily to the current stage of the project (presentation and action).

Which Project

Meaning in Context

Importance for this stage of the project

Document update information

Updates to this document after the start of class: