Personal Manifesto

By: Adam Bakopolus

Table of Contents

Week 1: Problem Formulation Stage

Informational Interview - Planning

To further engage with the data science community and gain practice-based insights, I will be reading **Caitlin Smallwood**'s interview in the "Data Scientists at Work" book by Sebastian Gutierrez. At the time of the interview, which is already readily collected and available through this course textbook, Smallwood was the Vice President of Science and Algorithms at Netflix. I was particularly interested in this interview as Smallwood, as part of her work at Netflix, would be heavily involved in the implementation and enhancement of their show/movie recommender algorithms. I'm very interested in learning more about the statistical models and approach that went into generating this process not only just as a user of Netflix but also as a current data analyst that would be able to leverage these procedures in my own work. As I work within the healthcare field, I see a clear and immediate need for solutions that allow decision makers and practitioners to clearly identify various populations in need of intervention or additional care to improve outcomes. Understanding the statistical models used and how the solution was rolled out would be invaluable to better understand how this can be leveraged across different fields.

Reading Responses

Readings:

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

Chapter 2's reading made the important note that "iteration is the rule rather than the exception" when it comes to the data mining process. While there may be an initial plan that is reviewed and signed off on for a particular data project for either school or work, the ability to be flexible and change the approach when roadblocks or new findings are discovered is immensely valuable. There will very rarely be a linear path from start to finish for a project, so the willingness to re-tool approaches and rethink strategies will be a great skill to obtain to allow for the end work to be both accurate and meaningful.

This reading also did a really great job of highlighting the key difference between "mining the data to find patterns and build models" and "using the results of data mining". A very effective data mining process and an extremely meaningful finding can all be rendered useless due to a data scientist's poor presentation or inability to relay insights in a digestible way to the audience. A data scientist must be able to effectively bridge the gap between both technical and non-technical audiences, and discuss the process and findings in a way that is both understandable and actionable for the audience to allow for the appropriate next step of utilizing the finding to begin.

- **Chris Wiggins interview**

I really liked Wiggins' note that "the key [when beginning a data problem] is usually to just keep asking, 'So what?' You've predicted something to this accuracy? So what?". This is a very important thought to keep in mind as a data scientist as there is typically a myriad of different paths that can be taken when grappling with a certain problem or question. Before delving in too deeply into the weeds programming-wise or research-wise, it's important to also remain cognisant of the high-level goal of a project. Busy work is not always productive if the discovered insights are meaningless or not actionable.

In Wiggins' interview as well I really liked his discussion around a thought attributed to Einstein that "not everything that can be counted counts, and not everything that counts can be counted." In my own day to day work in the healthcare space, I'm dealing with big data from medical and pharmacy claims and enrollment data that can at times be stored in tables with billions of rows. The ability to reduce data is critical as a data scientist to gather meaningful insights from the often overwhelming datasets that are available for certain projects, and it is very important to know either upfront or early on in the process what the key fields are to quickly aggregate and reduce the data source to a more reasonable level.

- **Erin Shellman interview**

Shellman had an interesting remark that “the most interesting types of data are those collected for one purpose and used for another.” This ties in really well with one of our maxims discussed in the lectures that the original formulation is rarely the right formulation. A framework that was first implemented for one question but was better suited for another is definitely a common case in my own day to day work, but this is a valuable note that I agree with. The ability to be flexible and not abandon work even if it doesn’t quite fit the original purpose is valuable, and the creativity to find alternative uses for these types of solutions is a desirable skill for a data scientist.

Similarly to Wiggins, Shellman also discussed the importance of considering the analysis’ end goal itself prior to really getting into the weeds development-wise as she “typically start[s] from the finish line. Assum[ing] that you’ve built the thing you’re considering, then ask ‘so what?’” Do customers want your product and can they use it? Not only is this consistent with the Wiggins interview, but it also ties in nicely with our own lecture discussions. A successful data analysis when grappling with a problem will continuously ask “why is that”. This approach ensures that the time spent developing a model will not be wasted as there is a true need or role for the finished product.

- **Jake Porway interview**

Porway made a great note that the next great discovery that is in line with the microscope (viewing of the small) and telescope (viewing of the large and far away) is the “macroscope that lets us look at the complicated patterns of society and nature and the ways that they interact.” With the many fields, like healthcare, for example, constantly changing and becoming more and more complex, the goal of a good data scientist is to be the tool for a project or organization that is finally able to evaluate the wide range of data and knowledge available in a meaningful way that will unlock solutions to the increasingly complex problems facing domains today.

Additionally, Porway also makes an important note that “we shouldn’t underestimate how much little things like that can transform an organization”. This again ties back into the importance of flexibility as a data scientist. While there may be an overarching end goal that a project is moving towards, little findings and insights along the way may also have an impact and time should be allocated to thoroughness and ensuring that even non-anticipated findings can have a meaningful impact.

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**

Yes, I already have this capability in the healthcare domain. In my day to day work, medical, pharmacy, and dental claims and insurance enrollment data are readily available and serve as a huge database to pull and glean insights from. With all of this data available, inquiries that investigate the quality and cost of care for certain portions of the population can easily be completed through a lot of the problem formulation types discussed in this lecture.

2. **A repertoire of problem types**

Yes, I already have this capability, as well, as my work heavily involves data reduction, which, in turn, allows for many of the other problem types and approaches discussed in this week's lectures to be more easily completed as insights can more easily be gathered from simplified datasets. However, I am very excited to continue building my skills through the MADS program, as Data Manipulation and Math Methods for Data Science have already greatly improved my statistical knowledge and python skills. I'm excited for future courses that continue to build this knowledge set to allow for improved skills and machine learning techniques to enhance a lot of the problems I encounter on a day-to-day basis, like regression analysis and similarity matching, for example.

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

Yes, I have mapped problems in my application domain to these various problem types. Following the data reduction of the various healthcare claims and eligibility data, regression analysis to predict certain scores or values for a population or quality measure, clustering certain events or portions of the population together, classifying medical events or demographic details into certain categories or buckets, similarity matching enrolled members together based on certain characteristics like name and date of birth, profiling to identify various outliers for quality measures or cost, causal modeling to determine if a certain event or medical event can be used as a predictor for other medical events, etc. are all various problem types that I have daily familiarity with.

Application in Domain of Interest

Domain: Health care

Project 1 Description: Identify individuals within certain provider organizations or physician groups with a high total cost of care for a measurement period as a means of intervention and providing more preventative care to drive costs down moving forward.

Project 1 Problem Type: This is a profiling problem type that evaluates the total cost of care associated with all patients tied to various provider groups and flags high outlier costs. The high outlier cost patients would be those targeted with additional preventative care.

Problem 2 Description: Create an Adverse Childhood Experiences (ACEs) flag through use of available demographic and medical claims data to identify a portion of the population in need of additional community and social resources to ensure long term opportunities and health are not impacted.

Project 2 Problem Type: This is a classification problem. Through demographic data like first name, last name, date of birth, address, subscriber relationship, etc. and medical claims data around alcohol or drug abuse, for example, an at-risk ACEs child population can be identified and intervened with moving forward. This essentially would be a “Yes”, “No” flag that would make a determination on whether the child is in an at-risk environment, which would allow for early preventative medical and social care.

Questions, Maxims, and Commitments

Question (I will always ask…)
Will the final product or model be actionable for the downstream care providers?

1-Sentence Project Description
Create an Adverse Childhood Experiences (ACEs) flag to identify an at-risk child population to appropriately allocate social and medical resources to ensure long term opportunities and health are not impacted.

Meaning in Context
There will likely be multiple different types of downstream users of this ACEs flag, such as medical care providers, foster care care organizations, social workers, etc. so an overcomplicated model or approach does not need to be shared with these stakeholders as the details/approach to arrive at this flag are not necessary. To be a useful and actionable end product, the flag just needs to be a simple yes or no classification that will allow for tailored care for those who most need it.

Importance

This ties in nicely with a lot of the discussion in both lecture and in this week's selected interviews. An incredibly desired skill of a data scientist is not just to be able to code effectively and create a very advanced model but to have the ability to understand an end user's need for the end product and be able to bridge the gap between the logic and approach to generate the model and how to use it in a day to day application. A data scientist, in not just a classification problem type but in all problem types, needs to be an effective presenter across all knowledge levels to ensure that the work is not "wasted" and can have a meaningful impact. As part of this, this again reiterates the importance of continuously asking "so what" as work progresses towards an end goal.

Maxim (I will always say…)
Data beats emotions.

Which Project
A profiling problem type that evaluates the total cost of care associated with patients, flagging high outlier patient costs for more targeted preventative care.

Meaning in Context
In the case of a profiling problem type that is focused around outliers in medical care costs, the data should override the emotions of perhaps a care provider or a stakeholder involved in the care delivery process. Oftentimes if not always, data is incredibly convincing and clear in what it is portraying and should be weighed more heavily than emotions. If the data, in this profiling case, is showing a patient with extreme ($50k a year, for example) costs, the profiling would quickly flag this as an outlier. The outlier analysis would make clear that a change would be needed, which perhaps would not have been the case if the data was not available and the care decisions were influenced by emotions of a family member or a care provider continuing to stick to the same process of care despite it not being perhaps the optimal approach.

Importance
Oftentimes, it is possible for emotions (whether you have a vested interest in a particular outcome, company, person, etc.) to influence a decision. As a result, this decision may not lead to the optimal outcome, as this optimal outcome may have required actions that were uncomfortable or difficult to make due to this vested interest. This maxim is valuable to keep in mind as it makes clear that data is so valuable and can often be so convincing and clear in what it's presenting that it should be weighed heavily against the raw emotion tied to a preexisting decision or approach. This is often difficult to carry in day to day work as many emotions and decisions from many different coworkers and stakeholders need to be considered, but I believe a reliance on the data is a great way to steer a conversation to the optimal outcome.

Professional/Ethical commitment (I will always/never...)
I will not sacrifice the integrity of the data itself to ensure hitting a quota or goal set by a client.

Which Project
A profiling problem type that evaluates the total cost of care associated with patients, flagging high outlier patient costs for more targeted preventative care.

Meaning in Context
In identifying a high cost outlier, it could be possible to define an outlier in a wide number of ways, either a certain cost threshold ($50k+, for example) or a certain cost category being higher than expected ($5k+ cost in an emergency department, for example). If there was an incentive to identify as many outliers as possible by a client, I would ensure that the definition of an outlier is very clear early on in the formulation process to ensure that all parties involved are on the same page and that there is reliability and consistency in regards to how an outlier patient was flagged.

Importance
This is a constant in my own current day to day work, as a lot of my time is spent not necessarily coding but also meeting with clients or internally to develop specifications for how the reporting will be generated. To make sure that there are a limited number of questions when the final product is generated and that the client or manager is getting the final output that they expected, these types of specification meetings are vital, especially if there are incentives tied to the number of individuals that may be bucketed into various categories. There are a wide number of ways to profile the data, so the problem formulation stage always should include transparency between all involved parties to ensure that the outcome is reliable and in line with expectations.

Week 2: Data Collection and Cleaning Stage

Potential Personal Project Tweet

Using enrollment and claims data from health insurance companies, we created a regression model identifying members with a high probability of emergency department usage! This model will allow providers to more effectively target their patients with preventative care.

Reading Responses

- **Law of Small Numbers**

The author offered a recommendation to readers and researchers alike to “replace impression formation by computation whenever possible.” This ties in well with last week’s maxim, as well, in that data beats emotions. It is important as a data scientist to not just settle on a finding that seems reasonable but instead validate and confirm the finding through well formulated analyses. ***Data Collection and Cleaning - Maxim***

The author also asserts that “people are not adequately sensitive to sample size.” The article specifically used a sample of 150 versus 3,000 to make the point that few would be able to determine what the “appropriate” size would need to be for the study to be considered statistically significant and thus meaningful. This is an important note for my own work moving forward, as well, to better understand what is needed for work to be considered significant and ensure that the data collection and cleaning processes are in place to allow for this to happen. ***Data Collection and Cleaning - Maxim***

- ***Statistical Biases Types Explained***

“[Observer bias] can come in many forms, such as (unintentionally) influencing participants...or doing some serious cherry picking.” This is a very important note that is especially relevant currently in my own day to day work, as, in the healthcare space and with COVID, healthcare utilization and costs decreased between 2019 and 2020 almost across the board. However, one should definitely not suggest that there has been improvement in the system due to this trend and instead evaluate a multi-year trend. This portion of the reading was an important reminder to be aware of this bias and always provide appropriate context where it is needed. ***Data Collection and Cleaning - Expertise***

A relatively new form of bias that was introduced to me in this reading was Survivor Bias, or when “the researcher focuses only on that part of the data set that already went through some kind of pre-selection process.” This was an important reminder and acknowledgment that data from outside sources or from a relatively new process should be evaluated and quality checked thoroughly before incorporating into an analysis. This would be an effective step in ensuring completeness and that there were no omissions, intentional or not, that could influence what the data would eventually show in reporting. ***Data Collection and Cleaning - Expertise***

- ***Data Cleaning 101***

A perhaps simple but valuable question posed in this article when evaluating a data set for cleanliness is “does the data match the column label?” Being in the big data healthcare space, this can often be a challenge, albeit a very important one, to ensure. Prior to sending over any tables or datasets to a client, there are thorough quality checks to validate that Personal Identifiable Information (PII) did not sneak into a delivered field (an email address or phone number in a city or zip code field, for example). This type of data cleaning is a significant portion of the day to day role to allow for confidence in a delivered product. ***Data Collection and Cleaning - Question***

Another important note that the author makes in this article is to “communicate with the source.” As I’ve seen again in my own day to day work. It’s very common for there to be data errors or omissions in submitted files or data sets. As opposed to pushing through with perhaps an inaccurate or incomplete dataset, it is a valuable skill as a data scientist to be proactive and follow up with a client or data partner to ensure receipt of the cleanest data possible. Further, it can often, as well, to not just follow up when questions arise but push for documentation at the start to again ensure that both parties are on the same page. ***Data Collection and Cleaning - Maxim***

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

A valuable first rule by the author was that “for every result, keep track of how it was produced.” This is a very important part of the data collection and cleaning as this is a step in the process that is often repeated. Implementing an automated process and providing documentation that would allow for anyone to run the program and reproduce the results are common practices in my own day to day work, as well. ***Data Collection and Cleaning - Expertise***

In a similar way, the author notes the value in “record[ing] all intermediate results, when possible, in standardized formats”. Often when collecting and cleaning data, there are many temporary or base tables that I create as part of the overall build to the final “clean” product. Saving these tables allows for both myself and others to quality check and validate that each step is working as expected and allows for much easier backtracking when trying to identify a particular bug. ***Data Collection and Cleaning - Expertise***

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**

Yes, this is a capability I am familiar with in my own line of work. Oftentimes data being submitted to my organization on a day to day basis has issues within certain fields (either inappropriately left blank or null, having an incorrect value, etc.) so I have become adept at establishing workarounds for this incorrect data. However, I also have become comfortable communicating with the source of the data to confirm that the issue is present and whether the workaround is appropriate or a re-submission would be required.

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

Yes, I am also familiar with the data sources in the healthcare domain. Predominantly in my own day to day work we handle millions to billions of rows of medical, pharmacy, and dental claims and enrollment data. Additionally, other data sources like death and birth certificate data, Medicaid capitation files, and electronic medical records are additional files that are available for certain projects depending on the particular client.

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

In addition to my day to day work with large data sets, we also are responsible for documenting our processes and how the data sets are pulled in, cleaned, etc. For this portion of the role, we rely heavily on JIRA and Confluence to track the steps taken and ensure that the process from start to finish is well documented to ensure that any analyst or colleague would be able to pick up the project or dataset and be able to complete the work.

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

This is something that I have become very comfortable with throughout my time at my current company. There are many processes in place that we leverage that utilize Spark SQL as a way of loading claim and enrollment data not only into our system but also transforming and cleaning said data in a predefined way depending on a particular client. More downstream, I also heavily rely on SAS to import and export data from disparate sources and again clean and transform various fields as needed depending on how dirty the data may be (inconsistent field population, errors within certain fields, etc.).

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

A large part of my day to day work is to also use SQL to query large datasets and glean meaningful insights and create perhaps more condensed datasets that are more practical and manageable to work with. SQL is a great tool to not only query against tables with perhaps millions and billions of rows, but is also very useful when there are known data issues. SQL logic is very effective at cleaning data essentially “on the fly” and is something that I have become very comfortable with.

Maxims, Questions, and Commitments

Question (I will always ask…)
Does the available data for a particular table or column make sense?

Which Project
Create an Adverse Childhood Experiences (ACEs) flag to identify an at-risk child population to appropriately allocate social and medical resources to ensure long term opportunities and health are not impacted.

Meaning in Context
For a project around Adverse Childhood Experiences, there would need to be very clean demographic data around first and last name, subscriber relationship data on medical enrollment files, address data, etc. in order for a child to be appropriately linked to an adverse experience. If there are phone numbers in an address field or inconsistent or incomplete population of some of these fields, the generation of the flag would not be as successful if there were a more complete file to draw from.

Importance
This ties in really well with the reading around data cleaning and Professor Resnick's note that most of a data scientist's day to day work will revolve around cleaning data and ensuring that it is sufficient for use in any downstream reporting. Before beginning any project, it is key to evaluate the data sets that will be utilized to ensure that you have what is needed to complete a project satisfactorily for a client or colleague. Taking these data cleaning steps early in the process ensures that sufficient workarounds can be put in place and, if needed, a data provider can be reached out to for any clarification or perhaps even a resubmission of a file in question.

Maxim (I will always say…)
Don't treat the symptom, figure out what the root cause is.

Which Project
Create an Adverse Childhood Experiences (ACEs) flag to identify an at-risk child population to appropriately allocate social and medical resources to ensure long term opportunities and health are not impacted.

Meaning in Context
As noted earlier, when dealing with large amounts of enrollment and medical claims data, there is a high possibility that data may perhaps be incomplete or submitted to our organization incorrectly. While it would be straightforward to perhaps just establish a workaround (if zip code, for example, is not submitted with its leading 0s, it would be very easy to just add a cleaning step that adds the zero in), it would be best to instead reach out to the data provider or source and bring this issue to their attention. Cleaning up the pipeline itself is a better approach than creating workarounds as, if the latter approach is taken, it will not be long until there are perhaps an overwhelming number of steps that are needed to ensure that the data is as clean as possible.

Importance
As noted above and in the readings and lectures for the week, it is critical to not only just develop the skills needed for data cleaning and manipulation, but a data scientist must become comfortable with the communication portion of their role. With so much data available, especially within the healthcare domain, there will always be situations where there is dirty data. With a lot of these issues often being data entry or population driven, reaching out to the organization that provided the data is often the best way to ensure that the data is not just corrected but is correct for all future iterations. This ties in well as well with the importance of carefully documenting steps within the data cleaning process. Noting the history that a field was previously wrong but since corrected is valuable as you may not always be the analyst working a particular project, and it's important to have the data as accurate as possible to avoid a case where a role is transferred and the cleaning/workaround process is not well fleshed-out.

Professional/Ethical commitment (I will always/never...)
I will always keep track of how a process is implemented to allow for any analyst or colleague to pick up a particular project

Which Project
Create an Adverse Childhood Experiences (ACEs) flag to identify an at-risk child population to appropriately allocate social and medical resources to ensure long term opportunities and health are not impacted.

Meaning in Context
As noted earlier, while data collection and cleaning was the major objective of this week's lectures, it is also very important to be able to carefully document the steps that were taken to arrive at a "clean dataset". If there were files that needed to be resubmitted, due to perhaps a data entry error, or there were certain fields that required a workaround to allow for appropriate flagging of children with adverse experiences (backfilling the zeros for a zip code, for example) this would all be important to note. Ideally, anyone is able to pick up a project and complete it successfully if the appropriate level of documentation is available.

Importance
This week's lectures and readings generally showed valuable tips and skills to acquire to clean and collect data. However, I believe below the surface that it also highlighted very good practices that should be implemented by all data scientists. Although not specifically related to data collection and cleaning, documentation and relaying findings are part of an invaluable process that ensures a project will not just be dependent on one person and that the minute details are accounted for during any iteration of the collection process.

Week 3: Data Analysis and Modeling Stage

Informational Interview - Reflection

I completed reading Caitlin Smallwood's interview in the "Data Scientists at Work" book. At the time of the interview, Smallwood was the Vice President of Science and Algorithms at Netflix.

A really great maxim that was reinforced during this interview was that "correlation does not imply causation". Smallwood noted that Netflix's CEO "believes in causality over correlation and is really strong on making decisions with as much data as possible rather than just judgment." This tied in really nicely with a lot of the lecture and reading material from this week that focused on exactly this concept. Very "interesting" correlations can be found when evaluating certain datasets together, but as data scientists it is critical when performing analyses or generating a model to not assume causation even if there is a clear correlation. Additionally, it is equally important to present the findings in a way that does not lead on your audience, whether it be stakeholders or other team members, and clarify what a correlation may mean.

An important question from this interview was: What are the next steps that can be taken if an analysis or approach does not end up working or does not yield any meaningful or statistically significant results? Smallwood also grappled with this as she noted that "if [a model is] not working, what are some of the ideas that are around that we could try differently?" I think this very valuable skill of reviewing methodology and approach for an analysis is a necessity for a data scientist, as flexibility and willingness to adapt often will lead to the best final solution. It is important to note here that this is not an example of p-hacking scene from the lectures and materials for this week. Smallwood and Netflix are not just running multiple different analyses to find a meaningful correlation to report on, but instead carefully grapple with company questions and arrive ethically at a meaningful insight at the end because of this approach.

As an ethical commitment, I will be open with my data to allow for colleagues to run a particular analysis without issue. I was very surprised by Smallwood's note around data openness at Netflix and how "people will be publicly called out if they are not being open with their data." I believe this is a valuable commitment to have especially when running analyses or generating models to allow for transparency into the process and to ensure that all appropriate steps to arrive at a meaningful conclusion have been taken.

For additional questions, the following would have been valuable:

1. With your team having a large amount of flexibility in the projects they choose to begin, are there any guardrails against p-hacking and ensuring that the insight gathered came from an ethical process?
2. With a large and comprehensive dataset available to you, are there any steps taken to avoid a model that is overfit?

3. When validating a model, what are some of the approaches taken to insure its performance (cross-validation, etc.)?

Reading Responses

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

The term and description of overfitting was not one I was familiar with prior to this reading and the author's note around a model "'memorizing the noise' instead of finding the signal" was a very helpful way of summarizing this issue. This will be a helpful limitation to keep in mind in my own day to day work that sample data should not be considered indicative of the larger population as a whole, and a perfect trend to this type of data (and the steps needed to accomplish this) may lead the analysis to stray from its goal of evaluating a larger group and gleaning a meaningful insight there. **Data Analysis and Modeling - Expertise**

Additionally, the concept of cross-validation was also new to me from this reading, and the author made clear that it's "a powerful preventative measure against overfitting."
Cross-validation is a great way to ensure that your model is not overfit and tailored to only handle a sample dataset as opposed to the larger population. My company at the moment is actually considering generating a model that will evaluate and flag patients that may be high emergency department utilizers in the next calendar year. I see a great use for k-fold cross-validation, for example, in testing the model and ensuring that it's able to effectively evaluate the withheld fold for each iteration, a great sign that the model is not overfit to the used sample data. **Data Analysis and Modeling - Expertise**

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

Again, despite being in the healthcare data analytics field, as a relative newcomer to modeling, the author's detail around Type 1 error, "the chance of finding a difference just by chance" and Type 2 error, or "failing to detect a difference that truly exists" was very helpful in grappling with some of the common issues a new data scientist may encounter. This ties in really with the issues detailed above around over and underfitting. A data scientist must ensure that the model is not over or underfit in a manner that would lead to Type 2 error and having the model be a false negative when evaluating the full population. Additionally, the ability to walk a fine line and not create a model that throws many Type 1, false-positive errors, is important to ensure that insights can be gleaned from the final results. **Data Analysis and Modeling - Expertise**

Similarly, I was interested in learning about the steps taken to avoid the dangers and false positives associated with multiple comparisons when conducting analyses. Specifically, the author's description of the "Bonferroni correction, [which] simply divid[es] the overall alpha level by the number of comparisons. If the p-value was set to 0.05, for example, and 100 different types of comparisons were run, there could be multiple situations that reject the null hypothesis just simply by chance. While the Bonferroni correction may definitely be a bit stringent,better understanding multiple comparisons and the high likelihood of creating a false positive from it is

valuable as I begin work as a data scientist. **Data Analysis and Modeling - Expertise**

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

I found very interesting this author's thoughts around how to best avoid a multiple comparison problem, with the solution being to "replicate yourself. You can report the initial study with the multiple comparisons but call it exploratory, and disclose what you did. Then, collect a new sample and test your results with a replication in the same paper." Of course, this type of fix may not always be possible, but it shows the importance of avoiding a multiple comparison problem and the false positives likely tied to it. By collecting a new sample, you're ensuring that the finding is not a false positive driven just by chance and a high number of investigations, but instead a meaningful insight that can be identified in an entirely new and independent sample. **Data Analysis and Modeling - Expertise**

This article was my first introduction into p-hacking and the author gave a very helpful description in that it's when "the analyst isn't really looking to test an hypothesis, but is 'letting the data speak' by running a model and just looking for statistically significant relationships."The process of p-hacking is unethical in that you're fishing for a relationship that likely may just end up being a false positive as detailed in the discussion above around multiple comparisons. This type of approach has bad optics and would also be very difficult to defend to either a stakeholder or colleague, so as an ethical commitment, in my own day to day work I will not engage in p-hacking. **Data Analysis and Modeling - Expertise**

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

The author's note that "observational studies cannot prove cause and effect" was not an insight I had previously considered but will be very valuable to keep in mind moving forward. In order to prove causation, more rigor, whether that be through an analysis with data that embodies the full population under investigation or some other approach, is needed than just simply observing a correlation. It is important as a data scientist to not lose sight of this well-known maxim that correlation does not equal causation, especially when relaying findings to a stakeholder or client. Ensuring that all parties understand the ramifications of a finding is critical in this field. **Data Analysis and Modeling - Expertise**

I also enjoyed this article's discussion around how selection bias can lead to a "skewed" result and show a correlation between two groups or events that are actually not correlated within the full population. If a study is limited to a certain demographic, this demographic may have qualities or experiences that drive a particular correlation between two variables. However, when the full population is evaluated, this correlation is lost as the quality or experience is not universally shared. This is a valuable warning to keep in mind as a data scientist to understand the population you're working with and how indicative it may be of the population at large. **Data Analysis and Modeling - Expertise**

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

The author's note that Simpson's Paradox "occurs at the level of a purely descriptive data analysis" was a helpful tidbit to keep in mind. This ties in really well with lecture discussions around the best way to present data to a client or stakeholder. Decisions around how to best present an analysis or model results are critical, and this paradox is important to keep in mind in future work as a final display can paint a picture for a client or stakeholder that is perhaps disingenuous to the data itself. **Analysis and Modeling - Expertise**

The author's note that "Simpson's paradox is the designation for a surprising situation that may occur when two populations are compared with respect to the incidence of some attribute" was a helpful reinforcement of the lecture discussions around confounding, mediating, and colliding variables. Viewing variables and analysis in this way was new to me, but this reading and the lecture materials have shown how critical it is so be mindful of not only an analysis but then how that analysis may be presented or displayed when conditioning on a 3rd variable. As a commitment, I'll be sure to be more mindful of variables that could be considered confounders, for example, in my own work and the impact these may have on insights from the work. **Analysis and Modeling - Expertise**

- ***Conditioning on a collider***

I liked the author's note that conditioning on a collider could be considered a case of "selection distortion effect." This again was a helpful reinforcement that conditioning on a 3rd variable (whether it be a confounder, mediator, or collider) has significant ramifications in regards to a final result from an analysis and may display a correlation that is not present when evaluating the population at large. As a data scientist, you must be mindful when conditioning on these 3rd variables and ensure that, in doing so, you're not artificially creating a correlation that only exists when this condition is applied due to the population that you have limited to. **Analysis and Modeling - Expertise**

The author also ended with a strong note that "if you really care about a cause, don't give mediocre studies an easy time just because they please you: At some point, the whole field that supports your cause might lose its credibility because so much bad stuff got published." This is an important note and reminder that as a data scientist you have not only an obligation to be ethical and diligent in presenting your own work, but to think critically about others work as well whether in your own company or for clients to ensure that the best possible final product can be reached that can withstand rigorous scrutiny. **Analysis and Modeling - Expertise**

Plan for Knowledge Acquisition

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

- Common mistakes in data analysis that lead to misleading results

Yes, I have a capability with this already in my own line of work. In dealing with healthcare data and attempting to evaluate cost and quality of care across age groups, Medicare vs. Medicaid, rural vs. urban, etc. it is critical to understand the effect conditioning on a third variable may have. In healthcare, Medicare costs per year are typically 2-3x the cost of a Commercial patient. However, this is due to Medicare members typically being 65+ years old and being more inclined to procedures and care. As a result, to avoid misleading results and provide a “fair” view of performance, a lot of the data used in reporting is normalized to provide a clearer snapshot. Scores above 1 indicate poorer performance relative to those less than 1 in regards to cost, for example. In data analysis, it is important to understand the populations you’re working with to ensure that the final reporting is impactful.

- A repertoire of models and how to estimate, validate, and interpret each of them

I only have a limited experience with generating and validating models as part of my company’s initial dive into creating a linear regression model to flag high emergency department utilizers, with the model being validated through k-fold cross-validation. Therefore, I am excited to learn more about both creating, validating, and generating predictive models as this course progresses. I am most excited about the Supervised Learning and Unsupervised Learning courses especially as they look to be the first courses that dive deeply into the model creation process and the many different approaches that can be taken to not only generate them but validate and ensure that the results can be interpreted correctly.

Maxims, Questions, and Commitments

Question (I will always ask…)
Is the analysis and reporting being affected by the noise within the dataset?

Which Project
Evaluate the total cost of care associated with patients, flagging high outlier patient costs for more targeted preventative care.

Meaning in Context
Similarly to many of the projects discussed this week in lectures and in readings, the context around certain trends is critical to keep in mind when reporting on a particular analysis or model. For example, in this project that evaluates costs, in 2020, due to the COVID pandemic, costs across the healthcare industry actually went down as elective procedures and other care was pushed back. However, despite this decrease in 2020, one should not assume that healthcare costs will continue to go down and I expect a sharp increase in 2021. If an analysis or model was looking at healthcare costs from 2017 - 2021, for instance, 2020 could definitely be considered noise and a model trying to evaluate these years may be overfit if it takes the steps to account for what was seen during this outlier year.

Importance

More generally, this question is focused around this week's discussions around overfitting and ensuring that a model eventually finds the signal while cutting out as much of the noise within a sample dataset as possible. Noise will likely always be present within a data set so it is critical as a data scientist to understand this and ensure that the model is not overfit to the trend. This often requires context and an understanding of the field's landscape but this again ties back in with a data scientist's responsibility to be in close communication with a data source or client or stakeholder to ensure that the correct path is being taken to lead to a final result that will be the most impactful.

Maxim (I will always say…)
Correlation does not equal causation

Which Project
Create an Adverse Childhood Experiences (ACEs) flag to identify an at-risk child population to appropriately allocate social and medical resources to ensure long term opportunities and health are not impacted.

Meaning in Context
There have been many studies that link children from adverse childhood experiences needing larger amounts of health and social care later in life. Thus, there is a clear correlation between ACEs flagged children and increased social and healthcare costs which has been vetted through numerous studies. Typically, the danger with correlation and causation is that studies will be strictly observational when noting a positive or negative relationship between two variables and the causation will be assumed, often incorrectly. In this project, causation was not assumed just due to the initial observation, but was heavily vetted through empirical studies, which lends confidence that these types of interventions are meaningful and can have a significant impact later in life. While this project does have strong evidence that the correlation is driven by the cause, this is a cautionary tale that one must be wary when viewing correlations as the cause may not be readily apparent, especially when the data may be conditioned on a 3rd variable.

Importance
There was significant discussion this week around the dangers of implying causation when a correlation is available. As data scientists, it is important to think critically prior to sharing any findings from analysis or model and ensure that there are not additional underlying factors that may be driving the relationship. Similarly, it is important to focus on any 3rd variables that are conditioned on and ensure that these conditionals do not affect the results in such a way that the findings or model can not be meaningfully rolled out to the larger, full population. These are all valuable questions to consider, especially between a client or stakeholder in order to ensure that the underlying question behind the study can be answered.

Professional/Ethical commitment (I will always/never...)
I will always hold out some data as a test set to avoid overfitting

Which Project
Evaluate the total cost of care associated with patients, flagging high outlier patient costs for more targeted preventative care.

Meaning in Context
As noted above, in the healthcare landscape, 2020 cost data is a clear outlier and costs, for the most part, decreased across many segments, like office visits, elective procedures, etc. To create a model that is not overfit and will not be strongly influenced by the year over year decrease seen between 2019 and 2020, a k-fold cross validation, or some other cross-validation that evaluates some test set would be critical. Having 2020 cost data as a fold or a major portion of one fold, would ensure that the model would not be overfit to a large cost dip and would instead be “forced” to look across the larger trend that is seen (where costs, historically, increase each year). Cross-validation is a critical portion of the model building process and would be invaluable for this particular project, as well.

Importance
While a relatively new concept for me, this week’s lecture has hit home the importance of not just building a model or analysis ethically (by avoiding multiple comparisons and p-hacking) but also the importance of taking the appropriate steps to ensure the model is properly validated. As a data scientist, there will rarely be a case where the dataset that is being worked with is 100% clean, so it’s critical to validate not only the data itself but then also the model to ensure that it has not been overfit due to some data quirk or outlier trend. This will be a process that I am excited to incorporate into my own day to day work moving forward.

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, in the form of a question, maxim, or professional (or ethical) commitment. For each insight, first describe it in 1-3 sentences and then, in bold, label it according to the following framework: ***{Stage the insight is relevant for: problem formulation; data collection & cleaning; modeling & analysis; presentation & deployment} - {which of the following it is: expertise; goal; maxim; question; commitment}***

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.* ***Data Collection and Cleaning - Maxim***
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.* ***Problem Formulation - Maxim***
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.* ***Problem Formulation - Question***

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

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

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