Personal Manifesto

By: Alison Huang

Table of Contents

Week 1: Problem Formulation Stage

Informational Interview - Planning

I plan to have an interview with **Joma Tech**, who is a data scientist and a Youtuber. The main reason why I chose him as my interviewee is that he gave me a basic idea about what data science is and how it works in the real world. His video type is humorous and attractive (I have subscribed to his youtube channel for years). He has a computer science education background and enriched work experience in data science at FANNG companies. I believe he is the perfect candidate for an interview.

I will get in touch with him by sending an email and ask if he has interests and available time for the interview. If he accepts for the live interview, I plan to have a 40-minute interview by asking some general questions about how the data scientist likes in the real world and what issues or problems we will face and never happen in the college class. Then, I will ask for some maxims and advice to current data science students.

If he doesn't have time to do the live interview, then I will send out the question list (no more than 10 questions) by email and he just needs to write down those answers.
If he doesn't want to do any kinds of interviews, I will watch the video on youtube he created, called *What REALLY is Data Science? Told by a Data Scientist* and finishing the assignment.

Reading Responses

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

    1. “Think carefully about the problem to be solved and about the use scenario.” Remember that each business decision-making problem or concern is unique. They come with diverse goals, constraints, and considerations. Don’t try to use only one tool or way to solve a problem. ***Problem Formulation - Maxim***

    2. Ask “why?” and “so what?” after every answer you’ve got. The more meaningful questions you raised, the more accurate tool or model you will make. In the real world, there are profusion variables and keys waiting to be used. It is impossible to use them all and critical to keep asking “why?” and “so what?” when I screen them. For example, if I decided to use “the length of the personal statement” to the project, I will ask “why do I believe this variable is essential to the project?” If I don’t use it, how does it affect the model? Moreover, these questions help you to shrink the cost of time of data collection and cleaning. ***Data collection and cleaning - Question***

- **Chris Wiggins interview**
    1. “People, ideas, and things in that order.” Remember as data science, we work with people, the stakeholders, we train data and use data, but we don’t work with data. All of the tools and methods we choose should be fit with the people we work for. Only we get what our stakeholders want and concern about, then we can start to brainstorm and get ideas about the project accurately. ***Problem Formulation - Maxim***

    2. “So what?” Keep asking questions to determine we are on the right track. We, as data scientists, need to have a clear and calm mind before opening the Juypter notebook. ***Problem Formulation - Question***

- **Erin Shellman interview**
    1. “Add more tests to the score.” Don’t overestimate your model. Put your model to the real world and test it again and again. The model we created is for solving real-world problems, so you never have too many tests before finishing the project. Moreover, be aware of deadlines and budgets because there’s no unlimited time and money staying on analysis and tests. It requires me to have time management, project management, and collaborative skills to maintain the smooth operation of projects. ***Modeling & Analysis - Goal***

2. “So what?” Again ask yourself and figure out if you are solving the right problem. Don’t go too far away and communicate with team members and stakeholders. ***Problem Formulation - Question***

- **Jake Porway interview**
    1. Jake Porway founded DataKind as a non-profit organization because people are sensitive to their personal data and always question what data scientists ask their data for and their motivation. This is a big ethical problem right now because everyone has social media accounts and we leave our data on the Internet. A data scientist has the responsibility to keep all of the personal data in secret and not share them with anyone other than team members and stakeholders. ***Data collection and cleaning - Ethical commitment***

    2. “Question everything, but be an optimist.” Asking the right question is lesson one for a data scientist. It is definitely not the employee who always sits in the office and programming all the time. The significant soften skill for a data scientist is communication, which means asking appropriate questions to make the model optimize. ***Modeling and analyzing data - Maxim***

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 look forward to strengthening this capability. First of all, when a data scientist starts to formulate a good problem, he/she needs to **communicate** with stakeholders and figure out what they really want and are concerned about. During this stage, it requires the data scientist must have excellent communication skills and **critical thinking** which allows him/her to ask the right questions. Besides that, the data scientist has the ability to divide one general problem into separate subtasks. (**SIADS 501 course**)

2. **a repertoire of problem types**

   I need to improve this capability. Since I don't have any background in computer science or IT, I need **more practice and exercises** on this capability. Besides doing assignments of SIADS courses, I will find more volunteer opportunities on data science projects to train the ability of repertoire of problem types. (**SIADS 505; 511; 515 courses**)

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

   I already have this capability.

   Through Python 3 specialization and SIADS 501, I have gained knowledge and experience in identifying types of problems: regression, profiling, causal, classification, and similarity matching.

   - ➔ Regression is for predicting.
   - ➔ Classification is to segregate items based on their labels or features.
   - ➔ Similarity matching is for recommendation pushing. For example, Alice and I are both from Mainland China and if I am obsessed with dumplings and chow mein, then she probably has the same food habits as me.
   - ➔ Causal is to determine and describe the causal relationship among several variables. For example, if I turn down the thermometer of AC in winter at my apartment, it will reduce my cat's quantities and times of eating because he would like to stay at his pet-bed all day long instead of eating. Only when he's starving, he will leave the bed and eat.
   - ➔ Profiling is to determine the target customers for the business. In my previous work as customer relations manager, one of my jobs is to determine the target customers. For instance, they graduated from high school. Their family annual income is above 100,000RMB. They have taken TOEFL at least once. It helps the marketing and salesperson to precisely localize which area they should pay more attention to.

Application in Domain of Interest

Domain:

Higher education

I have more than 7 years of work experience in higher education since I was an undergraduate student. In my previous role, customer service manager, I served international students who got admission letters and decided to study abroad in the US. I found most of them had been worried about their applications and submitted at least five applications at one time. I even had a student client who submitted ten applications! Nevertheless half of these applications are backup plans and one student can only accept one admission offer. It raised my confusion and question, which is this phenomenon creates wastes of resources, human, money, and time. It drives me to want to do research and analyst in higher education, especially the application approval prediction.

Project 1: Predicting university application approvals

Project 1 Description: Universities receive a lot of applications all over the world. Many of them get rejected for various reasons. Usually, one applicant applies to more than one university at one time and he/she spends pretty money and time on their applications, but he/she only enrolls in one university. Most of the applicants use limited resources and cases to analyze their application approval rates manually which is low effective. I want to build a machine learning model to predict if a university application will get approved for saving the applicant's money and time.

Project 1 Data source: GTER platform (http://www.gter.net/offer/index.html). This is a social platform for Chinese students to post their application status voluntarily.

Project 1 Variables:

1. Shifts level of high school/undergraduate school (required);
2. Cumulative GPA (required);
3. Language test score (required);
4. Standardized test score (if applicable);
5. Type of degree (required);
6. Categories of major[1] (required);
7. Quantities of recommendation letters (affirmative only; if applicable);

[1] Art-related; STEM; Business; Economy; Literature, Language, and Social Science (https://www.thebalancecareers.com/choosing-a-college-major-by-field-3570279)

[2] International Math/Physics/Chemistry/Biology/informatics Olympiads Medalists; Regeneron Scholar; Intel Science and Engineering Fair (ISEF) 1st to 5th winners; The Mathematical Olympiad Summer Program participants; Physics/Chemistry/Biology Olympiads national finalists; MIT Research Science Institute participants; Telluride Association Summer Program participants; Summer Science Program participants; Stanford University Mathematics Camp (SUMaC); CERN Internship; High School Honors Science Program at Michigan State (HSHSP); Ross Program in Mathematics at Ohio State University; Google Science Fair Global Finalists; Publish a history research article in The Concord Review; ISSYP; UPenn Management & Technology; Math Prize for Girls; MIT women's technology program (https://zhuanlan.zhihu.com/p/55516201)

8. Quantities of global/national awards[2] (major-related only; if applicable)

Project 1 Problem Type:

- **Data reduction:** for this project, data reduction is a major duty because students post their offer information voluntarily and it probably contains irrelative information or emotional words. It requires me to eliminate the noise and leave the signal only;
- **Causal:** there are plenty of variables I can use, but I only choose eight variables and only five of them are required. This is because I have to make sure there are causal relationships among variables and outputs.;
- **Classification:** for this project, I want to produce outputs, like *approved, not approved,* and *need more information to determine*.

[2]

Project 2: College study plan for international students

Project 2 Description: As a former international student in the US, I realized that most universities and colleges in the US lack advisors who fully understand the situation and concerns of international students with different cultures and backgrounds. Those advisors can't give effective and reasonable advice to international students because language barrier advisors don't know students' disadvantages and advantages. I want to build a machine learning model to create an annual study(course) plan, focusing on academic tutorship and training, for international students based on historical data and their cultural background.

Project 2 Data source and Variables: Since the difficulty of collecting data from institutions and universities, I will conduct an online survey for data collection.

School name	Major: Year:
Education level	Current GPA
Cumulative GPA	Expected GPA
Language proficient (1-10)	Budgets/yr (USD)
Course with the lowest GPA	Course with the highest GPA
Rank programs you would like to attend	1)1-1 tutorship 2)Small class tutorship 3)Online academic videos 4)Online exercises/tests/assessments 5)Others

Project 2 Problem Type:

Classification: I will classify universities by US News ranking, like 10: top10; 9: top 11-20; 8: top 21-30; 7: top 31-40; 6: top 41-50; 5: top 51-60; 4: top 61-70; 3: top 71-80; 2: top81-90; 1: others.

Moreover, I will determine which aspects the student needs to improve. Language proficiency lower than 5: language training; The lowest GPA course: Math-related course; Writing-related course; Art-related course; Business-related course.

Furthermore, I will provide different types of training/tutorship based on budget level. (all time durations are cumulative)

- ➔ $10,000 above/yr: 200 hours 1-1 tutorship/48weeks, mentoring, exercises, and tests (include unlimited 24/7 live QA);

- ➔ $9,999 - $5,000/yr: 144 hours 1-1 tutorship/48weeks, exercises, and tests (include unlimited 24/7 live QA);
- ➔ $4,999 - $2,500/yr: 72 hours 1-1 tutorship/48weeks, exercises, and tests (include unlimited weekday live QA);
- ➔ $2,499 - $1,000/yr: 80 hours small class tutorship/48weeks, exercises, and tests (include 5 hours weekday live QA);
- ➔ $999 - 500/yr: 200 hours online videos/48weeks, exercises, and tests (include 1 hour weekday live QA);
- ➔ $499 and below/yr: 120 hours online videos/48weeks and tests (answers provided)

Each tutorship/training is created based on the ranking of the university, differences between current GPA and expected GPA, language proficient, the lowest GPA course, and the budget.

Project 2 Example - Student info

Student Name	Alison	**School**	University of Michigan
Major	Accounting	**Year**	Freshman
Education level	Bachelor of Science	**Current GPA**	2.3 / 4.0
Expected GPA	3.8	**Lowest course**	Intermediate financial acct
Highest course	College essay	**Language prof.**	6 / 10
Budget	$11,000/yr	**Rank**	1, 3, 4, 2, 5

Project 2 Example - Study Plan (weekly, 4 weeks)

1. Financial accounting tutorship
2. Teacher: TOP10 University Accounting Major Master's degree or 5+ years related work experiences background
3. Extra: recommendation letter by tutor/mentor or referral opportunity to financial institutions
4. Academic performance monitor
5. Term evaluation report

	Mon	Tue	Wed	Thur	Fri	Sat/Sun
09:30-10:30	Tests				Live QA	1-1 tutorship
10:40-12:40	Live QA	Exercises				
14:40-16:40		Live QA	1-1 tutorship			
18:40-20:40			Live QA	Exercises		

Notes:

1. Curriculums need to be normalized.
2. Exercises need to be normalized.
3. Evaluation/performance report needs.

Maxims, Questions, and Commitments

Question (I will always ask…)

Who are the stakeholders?

Which Project

- Predicting university application approvals

Meaning in Context

For this project, the stakeholders are Chinese students and their parents. Most of them lack knowledge about data science and statistics. I will make the model easy to understand and the presentation will be more visualized and avoid too many professional terminologies. Moreover, I will make the presentation and report in the Chinese language because the majority of parents don't understand English.

Importance

This question helps me to understand the demands and concerns of stakeholders and do not go too far away from the main point of the project. Determining the stakeholders helps me **present the project and results in proper ways**, like word choice and the way of presentation. Moreover, it allows stakeholders to understand the model quickly and readily.

Maxim (I will always say…)

"It's just something that you need to think through clearly before you ever pick up a pencil or touch a keyboard"

Which Project

- Predicting university application approvals

Meaning in Context

For this project, I need to figure out which variables I must use and what else I might need to use before starting to code. Besides that, I spend a lot of time researching the data source I plan to use to find out how I can extract data from websites and what kind of noise is hiding in those data and how to clean them. All in all, I always start at creating an SOP or workflow before starting opening Python.

Importance

You can't build a house without blueprints. Same as the data science project. It is important to have a clear understanding of the project we're working on before we start hitting the keyboard.
I have seen a terrible self-developed (team-developed) CRM system without blueprints and this IT team only has a scratchy, disordered, outdated, and confusing so-called SOP. This CRM is just likely an elementary version of Excel without any UI and UX designs, all menus are one-level menu, and very hard to use. They literally move (or you can say copy and paste) ten more huge Excel worksheets to web pages (actually it's Share My Works). This CRM lacks the balance between design and content. It offers a wide range of unnecessary features because the IT team is too lazy to delete abolished "worksheets". This lacks up-to-date technology definitely. I have developed and improved the CRM system for my previous company and I realize that a CRM system integrating poor plugins and using old school design can hurt a business in terms of functionality and brand image because I am not the only user and this system should be perfectly used by other co-works and partners. **This case gives me a negative example perfectly.** If I start coding without a clear understanding and feasible workflow/SOP, I will never produce any projects with acceptable quality.

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

Never share your data with the public without permission

Which Project

- Predicting university application approvals

Meaning in Context

When I collect data from online sources, it probably contains some personal information. Even though the data source website I am using for this project does a great job of privacy protection, for instance, their rule stipulates to submit admission information by a normalized and anonymous form. It is still, however, trackable to users' personal information by their users' name or ID on the website. For this project, I should irritate those data and never share them with the public because someone might utilize these to backward track personal info.

Importance

This is a sensitive topic and everyone worries about their personal privacy nowadays. As data scientists, we work with lots of data and we need to get personal data from stakeholders. Keeping data in private is the fundamental career ethic for us. Protecting privacy data helps the project be more reliable and builds a reputation for me.
When the clients or stakeholders don't believe the data scientist has a career ethic, they won't share or provide personal data to us. It is the worst thing for any project. It means the data scientist is not reliable and untrustworthy anymore. We need to build a community with trust. This is our responsibility.

Week 2: Data Collection and Cleaning Stage

Potential Personal Project Tweet

Use admission data to predict the top universities' acceptance rates for international students. It helps them to predict rates and reduces costs and time commitment. It's hard to collect and clean data in a short time since the resource is from volunteered info containing unrelated info.

Reading Responses

- **Law of Small Numbers**
    - Collecting large amounts of data as much as possible is significant to reduce the risk of biases and extreme results because “large samples are more precise than small samples”. If the data we use for modeling has biases, such as selection bias or observer bias, the model is not persuasive and shouldn’t be used by stakeholders. Besides that, we can apply cross-validation to a large amount of data readily and not create an overfitting model. ***Data Collection and Cleaning - Goal***
    - The intuition of human beings is not reliable and if we depend too much on it, we will more likely categorize a single random data or event as a whole. We are “far too willing to reject the belief that much of what we see in life is random”. It taught me that when I start a project, look for data carefully and ask questions all the time. Are these data large enough? Are there a lot of missing values? Are these data related to my project? ***Data Collection and Cleaning - Maxim***

- **Statistical Biases Types Explained**
    - Researchers subconsciously select data, which is favorable to the expectant result, and work with a specific subset of subjects instead of the whole. It brings the model a self-selection bias. It makes the result look nice and beautiful, but actually harmful to the project and outcomes. ***Data Collection and Cleaning - Goal***
    - Be careful to set up questionnaires to collect because some of the questions might influence “participants or doing some serious cherry-picking”. Make sure the survey we create is feasible and reliable. ***Data collection and Cleaning - Maxim***

- **Data Cleaning 101**
    - “Does what you’re looking at making sense?” As data scientists, we can’t build up a rerunnable model without a clear mind. It requires us to keep asking questions to ourselves to make sure the model has strong connections with questions. Besides that, we can make an easy and straightforward way to clean data in the course of asking questions. ***Modeling & Analysis - Question***
    - Improving communication skills is always necessary. Asking questions to stakeholders or any data resources. Talking to our team members or anyone working on the project. Make everything clear and leave nothing to confuse people. “This is especially true if you are the client of the data source because you are entitled to clear information”. ***Data Collection and Cleaning - Maxim***

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

- “Avoid manual data manipulation steps”. Don’t hard-code your model. Otherwise, it loses the original goal of building up the model. We create the model for reducing manual-workload, not increasing it. “Link to the data used to calculate the summary” and allows your result to be inspected and traceable. This is our job to summarize the data in a readable and reliable form. ***Modeling & Analysis - Goal***
- “Providing public access to scripts, runs, and results.” Always remember teamwork is the key to success. You can’t do everything in a real-world project or problem. Don’t treat the model like your baby. It is not your personal property. “Provide access to others in your team and organization”. But, remember never to share data with the public without permission. ***Ethical Commitment - Maxim***

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. There are several common problems leading to misleading results I should avoid: Choose irrelevant data. Cherry-picking results. Choose biased datasets. Not enough data. Flawed correlation. I collect data as much as possible and compare the available quantity of dataset with the ideal quantity of dataset to determine if I have enough data. I am respectful of any results I got. When I was building the CRM system for my previous company, I sometimes got unexpected results, but never ignored them. Co-workers and I discussed and figured out how to improve the system.

2. **potential data sources in my application domain**
   I already have this capability. Identifying appropriate data sources to build up an effective model is challenging. I use the data from GTER(Ji-Tuo TianXia) platform as the primary data source, which are data collected expressly for research purposes to address a specific problem or hypothesis. This platform owns more than one million members, who are mostly prospective university students from Mainland China and volunteer to share their offer/admission information online.
   Secondary data refer to data that was collected for other purposes and are being used secondarily to answer some of the research questions. I can use secondary data to test the model's effective performance and this data helps me to determine if my primary data is biased. **Course 532** helps me to characterize each type of data through pattern extraction and similarity measures.

3. **how to understand and document data sets**
   I already have this capability. I store raw data once I have it as a backup. During any stage of data science, I script my analyses and record all intermediate results to make my model rerunnable and trackable. If the dataset is maintained in a version control system, I use a markdown file. I use the built-in database metadata features and software tools to create a data dictionary if the dataset is maintained as a database. **Course 503** will provide me a systemic and comprehensive knowledge of data privacy, provenance, and accountability, which means, in the data science field, there are policies and regulations about how to understand datasets and to store (or document) them safely.

4. **how to write queries and scripts that acquire and assemble data**
   I look forward to strengthening this capability. I will make notes or scripts during coding and record all intermediate results, which makes it easier to identify questionable

analysis when wrong. **Course 511** will assist me to increase my skills to write queries and scripts. I will improve it with plenty of exercises and practices during the study.

5. **how to clean data sets and extract features**
   I already have this capability. During the data cleaning, we are losing some possible features, so we need feature extraction before starting cleaning the data. For example, in some cases, we need to pay attention to numerical characters because those characters, presenting in the reviews, can be useful. Besides that, we need to determine the number of uppercase words or sentences because when people are angry or frustrated, they usually express their emotions by writing in UPPERCASE. During cleaning, I make all text lower case, removing punctuation, stop words, emojis, URLs, and HTML tags, and correct spellings. **Course 505** trained me a lot on how to clean and manipulate data in different methods.

Maxims, Questions, and Commitments

Question (I will always ask…)
Does what you're looking at make sense?

Which Project
Predicting university application approvals for Chinese international students

Meaning in Context
Can the proxies, such as GPA in high school, length of the personal essay, number of awards, and so on, represent the accurate admission rate? If not, which proxies I missed or calculate them wrong or put the incorrect weight to the proxy? I will collect data through the GTER website, a social platform for Chinese international students, sharing their university admission information, and IPEDS (American University data) data from Kaggle to find out top-50 American universities acceptance rates for Chinese international students, not including Taiwan and Hongkong. However, there is a bunch of irrelevant information from the social platform, which requires me a **high-proficient skill of data manipulation**. I believe that the question I will always ask makes me **have a distinct understanding and way of collecting and cleaning data.**

Importance
Is my data related to the project closely? Does my data explain or answer questions correctly? Is the data I have collected feasible or reasonable? All questions are about making sense of what I am looking at. When I collect data, I sometimes get lost in the ocean of datasets. This question reminds me to work around the project or hypothesis compactly and allows my **data to have a strong connection with the model**, which should **be reproducible and others can reproduce the results** of my study or model. Only the data related to the main point of the project closely can **produce unbiased results**.

Maxim (I will always say…)
Avoid manual data manipulation steps

Which Project
Predicting university application approvals for Chinese international students

Meaning in Context
The meaning is manifest to the project. While collecting data through online resources, I definitely have many irrelevant or unrepresented data. For example, people express their negative emotions by emotional words or sentences when they get a rejected admission letter. These words are unrelated to my project, which means I need to determine those emotional words and eliminate them. Moreover, it requires **a mass workload**, maybe even more than a whole week. It looks like it is easier to do it manually, but actually, it is **harmful to the model and result.**

Importance
When I started to study Python by myself last year, I mostly liked hard-coding when I was having trouble cleaning data. It just shows that I am not confident with my ability of Python coding and need more practice. Now, when I am facing the same trouble collecting and cleaning data again, I try to look for a way of coding to fix the problem instead of manipulating data manually which is meaningless and a waste of time. This maxim helps me to **make my coding and model clear to read and understand, and also makes stakeholders or team members rerun them without mistakes and confusion.**

Ethical commitment (I will always/never...)
I will never make hypotheses after getting outputs.

Which Project
Predicting university application approvals for Chinese international students

Meaning in Context
It is difficult and time-consuming when I collect and manipulate data, especially for doing research on the topic of a foreign country. I believe there are some preconceived notions about university admission. For example, Chinese students pursue language tests and standardized test scores blindly. It causes that even though most of them get their dream schools admission letter and have extraordinary test scores, they lack independent critical thinking and fear to speak publicly. Nevertheless, these are the features TOP universities regard as significant. **If I make a hypothesis after getting outputs, I will be deceived by my prejudices and the prediction will definitely be incorrect and unreliable.**

Importance
Never jump to the conclusion in one step. The model is not made for one time and it's made for the whole team or stakeholders to allow them to rerun and update the model all the time. I will create and train the model to stay true to the datasets and outputs. Despite the outputs that may reject my hypothesis, I should not self-deception and change the hypothesis. It also helps us to identify where a problematic result is wrong.

Week 3: Data Analysis and Modeling Stage

Informational Interview - Reflection

1. **Insight.** Data science is not about making complicated models. It's not about making awesome visualizations and it's not about writing code. Data science is about using data to create as much impact as possible for your company. The impact could be in the form of the future vision of the company or the form of products. The data scientist is the field/company influencer using tools, like Python, SQL, or R, to make complicated models or data visualizations. The real job of a data scientist is to solve real company problems and what kind of tools you use does not really matter. ***Modeling & analysis***
2. **Maxim.** "Collecting, storing [and] transforming all of these data engineering effort[s] is pretty important." He talked about what data scientists do in the real world by breaking down three types of companies, start-up, medium-size, corporation. No matter what company, data collection and manipulation are the required and essential skills for data scientists. ***Data collection and cleaning***
3. **Ethical statement.** As a data scientist, you are not a data cruncher, you are a problem solver. This might not be related to ethical commitment very close, but when I rethink this statement, I realize that setting up a subject and macroscopical concept before learning or digging into the field of data science is much essential. We want to be and should be the driver or a guide to lead the company in the right direction by using models we create. We are not the tool. ***Problem formulation***
4. **For additional three questions.**
    a. During previous data science work, have you dealt with overfitting models? If so, how did you figure it out? If not, how could you avoid overfitting your model?
    b. For the FANNG company, how many teammates in the data science team, what are their backgrounds, what are their roles, and why does the company set up like this?
    c. How many variables are you using for a general model in the real-business project? How did you and your team determine which variables are essential and which else are not?

Reading Responses

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

1. When we build a model, we need to pay attention to where and how the model learns from. Does it learn from the noise? Or the signal? It is because once the model learns from the noise, the model does not predict values accurately. It requires the data scientist to comprehensively understand how overfitting happens and what the solutions are. Sometimes, we need more data and different methods to train our models. ***Modeling & Analysis; Expertise***
2. We can split the datasets into separate training and test subsets. We build and train our model by using the training dataset and evaluate the performance of the model by using the test subsets. We will know how well or bad our model works. If the model works worse on the test subsets than on the training subsets, it means the model is overfitting. ***Modeling & Analysis; Expertise***

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

1. One of the significant standards or milestones of a model is reproducible. Sometimes, too many variables get involved in the model which results in incorrect predictions and unrunnable models. For the project of predicting university approval rates, if I use gender, location, GPA, awards quantity, and volunteer experiences as variables, I would have a high probability of observing the effect of a statistically significant feature in one of them. ***Modeling & Analysis; Expertise***
2. There are two approaches to avoid this kind of problem: one if the family-wise error rate (FWR) and the other is the false discovery rate (FDR). The first one is the probability of making at least one Type I error. The FDR has greater power, at the cost of increased numbers of Type I error, because the procedure of FDR is less severe control of Type I error compared to the FWR procedure. ***Modeling & Analysis; Expertise***

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

1. **“**The more models/analyses you run on the same data, …, the greater the odds that when you observe a p-value less than .05 it is happening purely by chance.” This article, especially this sentence, emphasizes the topic of week 3, the detrimental effect of running and testing the model with the same datasets. The best way to reduce the odds is to replicate the model, or split the datasets into separate training and testing subsets. ***Modeling & Analysis; Expertise***
2. The worst way of data analysis is to hypothesize after results are known, aka HARKing. It violates the regular sequence of analysis and the ethical commitment of data science. For me, it is the same as cheating just like checking the answer before taking the test. Again, how to detect it or avoid it? Replicate the model. ***Modeling & Analysis; Ethical Commitment***

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

1. Sometimes, we focus too much on correlated trends and neglect the inner natural relationship between these trends. At the first sight, it makes sense that students who have study abroad experiences have high rates of graduation. However, the real causal relationship is that those students had better academic performance than others. We are likely to forcibly create causal relationships with non-correlated trends. ***Modeling & Analysis; Expertise***
2. We need randomized controlled trials to avoid a forceful causal relationship happening on our models. It is hard to detect, but easy to avoid. This is how some graduate students or even professors were shadowed by this kind of statistical conclusion. ***Modeling & Analysis; Expertise***

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

1. **"**Simpson's paradox is not a contrived pedagogical example, … can easily bewilder the statistically naive observer." Simpson's paradox gives me fresh knowledge and understanding of statistics. It shows that sometimes there is an inequality situation happening in comparison. Besides that, Simpson's paradox can increase correlations. ***Modeling & Analysis; Maxim***
2. Simpson's paradox arises the question, Which data should we consult in choosing an action, the aggregated or the partitioned? In the smoking habits and 20-year survival case, the result of smokers has a higher rate of living contrary to common sense. It is because we ignore some common senses in this case, such as the older people are less likely to live 20 more years no matter if they smoke. ***Modeling & Analysis; Question***

- ***Conditioning on a collider***

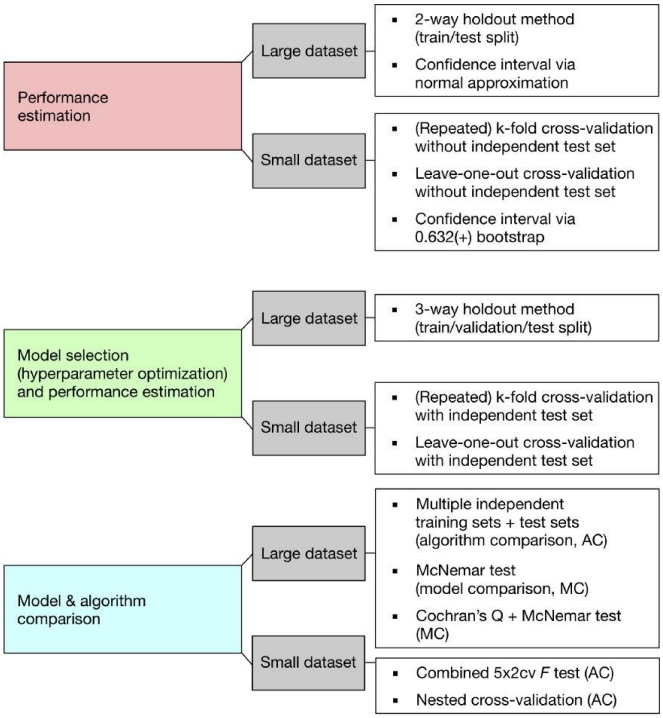
1. "If you really care about a cause, don't give mediocre studies an easy time just because they pleased you ..." This explains a lot of cases that have happened or are happening in the real world. In China, a graduate student has to publish at least one academic article to a professional journal to obtain his/her master's degree. This is a solid requirement for most majors and programs. However, the truth is not all graduate students have the ability or opportunities to publish his/her research outputs. Actually, most of them failed or found out extreme mistakes during their research. If they want to graduate on time, they either change their research topic or "manipulate" results. These mediocre studies or even worse ones are produced and the majority of these studies are useless and waste resources. ***Modeling & Analysis; Maxim***
2. In my opinion, this article leaves me profound significance not only about data science, but also about higher education, and even the original nature of human beings. Just like the article claims "people are very willing to speculate about confounding variables, so why not speculate a collider for a change?" It probably gives us an idea about how to develop policy, regulation, and the system of monitoring to prevent and detect this phenomenon. ***Modeling & Analysis; Maxim***

Plan for Knowledge Acquisition

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

- **common mistakes in data analysis that lead to misleading results**
  I look forward to strengthening this capability. The fundamental point of creating a model is to be aware of the inequality situation happening in comparison and control the number of variables. Make sure that the model is reproducible. **SIADS 532 and SIADS 630** will teach me how to solve real-world problems and learn how this kind of problem comes to embody and how to prevent and fix it. According to the article, *How not to lie with statistics: avoiding common mistakes in quantitative political science,* a researcher wanted to describe the number of apples eaten per week, the number of oranges eaten per week, the number of visits to the doctor per year. He got the multiple regression equation which is $\hat{y} = 10 - 1.5X_1 - 0.25X_2$. It means that the increasing quantity of apples or oranges eaten weekly will decrease the average number of visits to the doctor annually. The full story of this research is too long to describe here, but the final statement the researcher made is that "for one dollar spent on two apples, doctor visits would decrease by about three, whereas the same dollar spent on 20 oranges would decrease doctor visits by five on average." If you read the article along with the way of thinking and analyzing by this researcher, you might have the same conclusion as him. Nevertheless, we know that these variables don't have any causal relationships. This example explains again the importance of using explanatory variables in data science. Don't add too many variables to one model, even though they have correlations.

- **a repertoire of models and how to estimate, validate, and interpret each of them**
  I look forward to strengthening this capability. I need some practice and exercises on data analysis of real-world cases to enhance this capability. There is a good reference for determining this, which is *Model evaluation, model selection, and algorithm selection in machine learning* by Sebastian Raschka. **SIADS 630 and SIADS 631** will provide me with real-life examples to work on.



Reference

King, Gary. 1986. How not to lie with statistics: Avoiding common mistakes in quantitative political science. American Journal of Political Science 30(3): 666-687.

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*.

Maxims, Questions, and Commitments

Question (I will always ask…)
Which data should we consult in choosing an action, the aggregated or the partitioned?

Which Project
University admission approval rate for international students

Meaning in Context
For this project, I don't use "gender" as the main (single) variable to predict the application rate. It is because diverse application rates in one university exist in different majors (for international students). For example, the law school and the business school. For example, the law school's acceptance rate is only 9.2%. The business school's acceptance rate is 53.3% which is very high. The two kinds of gender distribution proportion of applicants. Females prefer to apply for business school, so the business school female acceptance rate is 83.3%. On the other hand, males prefer to apply for law school, so its female acceptance rate is just 16.7%. As a result, in terms of numbers, law schools with low acceptance rates have relatively few females who are not admitted because there are fewer female applicants. The Business school with high acceptance rates admits a lot of males, but not many applicants. When the final summary is made, the number of female students is actually superior.

Importance
I obtained a substantial amount of knowledge and examples from this week. Through lectures and readings, I realized that I usually make those mistakes before, which are harmful to the model and outputs. It helps me avoid creating a model with unequal situation variables or comparing too many variables. **To avoid Simpson's paradox in the project, I must consider different weights on various variables.** It reminds me multiple times that using different datasets tests the model to make it **reproducible**. Don't just pick up the nice-looking output only, but pay close attention to those odds outputs.

Maxim (I will always say…)
“[I] should try to limit comparisons between groups and identify a single primary endpoint.”

Which Project
University admission approval rate for international students

Meaning in Context
For this project, choosing appropriate variables to compare is essential. Variables I consider are historical admission data, student’s high school GPA, and standardized test scores (if applicable, otherwise language test scores).

Importance
It is significant because abnegating the collider variable makes the model reasonable and reproducible. The purpose of this project is to shorten the time consumption of applications and reduce the financial costs of applications of back-up universities in case all applications fall through. Even though there is hearsay about TOP 20 universities that prefer students who are from specific cities, such as Beijing or Shanghai, this is the noise (or bias) in this project. I won’t take this as a variable.

Ethical commitment (I will always/never...)
I will never p-hack on my model.

Which Project
University admission approval rate for international students

Meaning in Context
While I was building and training the model, I don't know if the strength of the relationship among different variables, such as GPA, SAT, quantity of awards, etc. I definitely expect a p-value much lower than 0.05 which means that I should reject the null hypothesis, and further indicates that the variables I chose have a strong relationship with the university admission approval rate. I will always use unbiased datasets and never select them factitiously. I will limit the number of variables. Last but not least is never hypothesize after getting the results.

Importance
If I select biased datasets artificially and these data cannot explain the whole population of study-abroad students, I will get a perfect-looking p-value and the project will be submitted on time. Nevertheless, what is this for? This manipulated model cannot produce anything, but inaccurate prediction results. It is crucial to make it clear at the beginning of being a data scientist. Don't cheat yourself. Don't lie to yourself.

Week 4: Presenting and Integrating into Action

Sources for Data Science News

I plan to follow the following sources of information about data science to keep myself up to date with the industry. I always listen to podcasts in the early morning once I wake up. Actually, I use podcasts as the alarm clock. I chose them because they provide updated information and news about data science and I can listen to them anywhere and anytime. Moreover, I subscribe to **Joma Tech** on YouTube because the style of videos is casual, funny, and informative about data science. He uses different ways to perform how data scientists work in the real world.

- **Podcast - Data Skeptic**
  It is short (< 30 mins) and interviews data scientists in different fields and backgrounds.
- **Podcast - Linear Digressions**
  It updates weekly and host rapport makes each episode very accessible and easy to understand.

Reading Responses

- ***A History Lesson On the Dangers Of Letting Data Speak For Itself***
    1. **Presentation skill is the key to data scientists.** It is crucial what to show and how to show your predictions and outputs to stakeholders or the public. Not everyone is interested in the topic you're talking about and not everyone is familiar with statistical terminologies. If you choose the improper method to present the project, it probably reflects poorly on the feedback of audiences, like Dr. Semmelweis. ***Presentation & deployment; Expertise***
    2. **Know your stakeholders or audiences before presenting the outputs.** Sun Tzu said that knowledge precedes victory; confusion precedes defeat. Knowing your audience is the essential point of making your presentation an edged weapon increasing efficiency and acquiring expected feedback. ***Presentation & deployment; Expertise***

- ***Storytelling for Data Scientists***
    1. **Make your audience act** by blending your result into a convincing and empathic story. The purpose of the presentation is to implement your idea and make it come true instead of talking insipidly. ***Presentation & deployment; Goal***

    2. SUCCESs model optimizes the performance of the presentation. You don't need fancy tedious PowerPoint slides. Limit the quantity of length of PowerPoint and focus on the story and emotion. Keep catching the audience's attention and let them realize the model you create is valuable and reasonable. ***Presentation & deployment; Expertise***

- ***Interpretability is crucial for trusting AI and machine learning***
    1. **Accuracy or interpretability, we have to lean to one of them** in light of helping stakeholders without related background to make decisions or process. The black model box produces more accurate outputs but is difficult to understand. The main goal for some models is to make decisions or consummate workflow and this requires models are interpretable by decision-makers. ***Modeling & analysis; Goal***

    2. **Use the proper technology to achieve the interpretability of the model.** When you just start the project, you definitely need to summarize the main characteristics of datasets and make them interpretable because you need a clear understanding of datasets, like the relevant relationships between data and variables. If the project doesn't require pinpoint accuracy, you only need to build a white box model in view of presenting the model to your teammates. Otherwise, you want to use plentiful parameters to make a black-box model. For outside

users and further improvement, you need interpretability in post-modeling. ***Modeling & analysis; Expertise***

- ***The Signal and the Noise, Chapter 2***
    1. **Human beings are not designed for statistical analysis.** We, as human beings, have an innate sense of subjectivity. Through this book, when we do analysis, we need a voracious attitude to make better predictions. The author defines this type of expert as a foxy type. They are humble and have open minds to different opinions. They **don't dope subjective perspectives** to the prediction. ***Modeling & analysis; Ethical commitment***

    2. How to determine if you are a hedgehog or foxy type of expert? "**Do your predictions improve when you have access to more information?"** An outstanding model should be fitted with more information or data. Moreover, it can test if the model is biased or the datasets are manipulated (cherry-pick results) manually. ***Modeling & analysis; Maxim***

- ***The Signal and the Noise, Chapter 6***
    1. **"The importance of communicating the uncertainty in [their] forecasts accurately and honestly to the public."** None of us like uncertainty, but we have to coexist with it. It is not shameful to have error (margin of error) or uncertainty, but it's shameful to turn a blind eye to it. ***Modeling & analysis; Ethical commitment***

    2. **The way to avert uncertainties is never to ignore data.** You might want to involve many volatilities from the datasets as much as possible and it will increase the accuracy of the model. For example, for economic prediction, you need to pay more attention to data of recessions. ***Modeling & analysis; Expertise***

- ***How Not to Be Misled by the Jobs Report***
    1. **"... one month of jobs numbers doesn't tell you much of anything about how the economy is actually doing."** The meaning of this sentence echoes with my opinion talked above, which is that human beings have an innate sense of subjectivity. We usually hear that "the pick-up truck market is in short supply because the sales from the nearest vehicle dealer said I have to wait for 3 months to reserve the truck." Actually, according to Statista, the production of pick-up trucks in 2020 declined around US$778,000 compared to the one in 2019. We can't make a conclusion based on a salesperson's statement. We have to observe the whole data macroscopically. ***Modeling & analysis; Maxim***

    2. Uncertainty is everywhere and cannot be irritated. **We have to learn how to live with it and adapt to it.** Do not focus too much on "random statistical noise", but

on comprehensive trend analysis. When we are modeling and analyzing, do not "focus on the relatively small change in the number". ***Modeling & analysis; Expertise***

- ***But what is this "machine learning engineer" actually doing?***
    1. The machine learning engineer is **a link between software engineers and data scientists.** "You must stay up to date with the state of art technologies and constantly look for the places in which the overall product performance could be improved." The machine learning engineer is the one who undertakes the responsibility of "spot the potential area of improvement." ***Modeling & analysis; Maxim***

    2. Why are machine learning engineers essential to our society? How come this position was generated? It is because there is a gap between software engineers and data scientists. The former experts on programming and computer science knowledge and the latter is specialized in prescriptive analysis. Machine learning engineers are like interpreters between them and increase the efficiency of business operations. ***Modeling & analysis; Question***

- ***How we scaled data science to all sides of Airbnb over 5 years of hypergrowth***
    1. "[Data] is the voice of our customers." This is how data scientists treat data in the real world. Data is not cold and dead numbers on the screen. We have to put lives on it and read them like real people because the data represents the opinions of customers. ***Modeling & analysis; Maxim***

    2. Continuing on the above, the data scientist is not only an errand boy for analyzing but also a part of the decision-maker because they usually have expert knowledge and experiences on what those digits and the model mean and how it works. Being an obedient "errand boy" is not the goal of data scientists. **You should take on responsibilities on decision making as a key role in the business.** ***Modeling & analysis; Goal***

Plan for Knowledge Acquisition

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

- **how to present results to domain experts who are not data scientists**
  I look forward to strengthening this capability. I will take **SIADS 522** next month and believe I will gain knowledge of how to visualize information by using Altair. I am very much looking forward to taking this course and getting ready to learn new skills. Besides that, I will gain the oral and written communication skill of delivering results to stakeholders and clients efficiently through **SIADS 523**.
- **how to work with software engineers to put models into production**
  I look forward to strengthening this capability. In the real-world, data scientists usually work with one-off scripts that contain SQL queries. A software engineer can help with such a library and review new writing code and find opportunities to add new functionality to a data analysis toolbox. By working with software engineers, data scientists can focus on the research side and engineers focus on scalability, data reuse, and ensure that the input and output pipelines for each project are aligned with the global architecture. I will learn how to correctly apply, interpret results, and iteratively refine and tune supervised machine learning models to solve a diverse set of problems on real-world datasets through **SIADS 542 and SIADS 543.**

Reference

Stefanuk, A. S. (2020, November 24). *How software engineers and data scientists can collaborate together*. Big Data Made Simple. https://bigdata-madesimple.com/how-software-engineers-and-data-scientists-can-collaborate-together/#:%7E:text=Data%20scientists%20usually%20work%20with,%2C%20for%20example%2C%20SQL%20queries.&text=A%20software%20engineer%20can%20review,extract%20information%20from%20raw%20data.

Maxims, Questions, and Commitments

Question (I will always ask…)
How to present my model to the audience?

Which Project
University admission approval rate for international students

Meaning in Context
First of all, I understand the main audience of this project would be international students and their parents from Mainland China. Based on that, I will create some charts of the model to show the variables I am using and why. Moreover, I will make the presentation in the Chinese language. Last but not least, I will simulate the model and show how accurate the prediction is.

Importance
For this project, it is important to know what native language the audience speaks and how much data science or statistics background they have. Even though the majority of students understand English, their parents don't. My presentation needs to be adapted to the audience. If I use too many data science terminologies during the presentation, the audience would feel bored and not accept my model.

Maxim (I will always say…)
Data makes people think, emotions make them act.

Which Project
College study plan for international students

Meaning in Context
I create several case studies, just like short stories about international students, and use them during presentations. These stories are immersive experiences for the students and their parents and tend to bring more empathy to them.

Importance
For general audiences, data, digits, and formulas are cold and lifeless. Merging them into short stories will help these audiences understand the model readily-easily. Using the SUCCESs model will enhance the positive feedback of the presentation. People love stories and are easily convinced by emotional and unexpected stories.

Ethical commitment (I will always/never...)
I will always make interpretable models.

Which Project
University admission approval rate for international students

Meaning in Context
I am aware that the model for this project is not just for showing off. This means the model should be understandable for the audiences and the people who will use it. Once there are issues or anywhere that needs to be improved, other technical personnel is able to read my model and enhance it.

Importance
Interpretability is significant because it's beneficial for developers, audiences, and end-users. Black-box models are hard to understand and create hidden trouble for further development. For developers, they can improve the White-box models and end-users, they are willing to use these models for prediction because they somehow know the mechanism of models.