Personal Manifesto

By: Michael Wynn

Table of Contents

Week 1: Problem Formulation Stage

Informational Interview - Planning

- For this section of the assignment, I will be listening to an interview with Renée Teate who is currently a Data Scientist at higher ed analytics start-up HelioCampus, and she hosts the “Becoming a Data Scientist Podcast”.

- As someone who is looking to transition into being a data scientist, Renée is definitely the go-to person. I’d like to pick her brain vicariously through the datacamp podcast which can be found **here**.

- Renée is often lauded for her history of helping people of different backgrounds find the relevant learning tools to transition into data science, and although I’d like to be a data scientist in the finance industry, I still place the importance of being a generalist data scientist above all else.

Reading Responses

Chapter 2 - Business Problems and Data Science Solutions

- "...data scientists decompose a business problem into subtasks." Data scientists should not be dyed in the wool in only seeking a solution to answer the main business problem. They should also practice enough prudency in understanding what subtasks can be formed, and whether or not these subtasks can be broken down further into smaller subtasks. More often than not, solutions to subtasks can lead to human action though potential business applications. Through decomposing the main problem task, the data mining project/process can be more efficiently carried out without having to revert to the problem formulation stage repeatedly, "reinventing the wheel" as Foster and Tom puts it.
  Problem Formulation - Maxim

- "...data scientist must think about the comprehensibility of the model to stakeholders (not just to the data scientists)." Data scientists have to communicate the model explicably to the right people in order to implement the model and value-add to the organisation, and in order for management and other stakeholders to be convinced by the effectiveness of the proposed model, it has to be understandable to them. The idea of using simpler modelling techniques (if possible) becomes very relevant here, since most deep learning approaches do not allow most users for closer inspection of how the algorithm is working behind the scenes, the so-called "blackbox" effect. This was also echoed by Vicki Boykis in her interview - communication skill is key.
  Presentation and Integration into Action - Maxim

Chris Wiggins interview

- "How can I reframe this as a prediction task?" Data scientists should aim towards producing solutions that are more accurate and useful, and as such should employ supervised techniques over unsupervised ones wherever possible. This way, they can formulate the business problem more clearly by decomposing it further into sub-problems such as "what other attributes are important towards attaining an accurate target result?", which will impact the following stages in the pipeline and determine whether communicating the end result is well-judged to be explicable to stakeholders who may not possess the basic quantitative skills to comprehend the models easily.
  Problem Formulation - Question

- "The world's like that. The world doesn't hand you models" Not only should data scientists have the ability to determine how a business problem can be broken down into

smaller pieces, they should be able to also correctly identify how each of these smaller pieces can be matched to a specific data mining algorithm. Chris highlights the importance of creativity here. This involves experience and failing along the way too.
Analysis and Modeling - Maxim

Erin Shellman interview

- Shellman talks about trying to disprove initial assumptions formed when trying to understand a new dataset. This reinforces the data understanding and exploration stages and helps negate any unexpected surprises in the analysis and modelling stage. Ensuring quality of the data will help identify and/or possibly reveal underlying relationships more accurately.
  Data Collection & Cleaning - Expertise

- Shellman and the team went on to have a coffee chat with a beauty stylist to talk about their recommendation strategy of setting up an email reminder as a tool to inform customers about purchasing or restocking their supplies, to later find out that the stylist would not get commissions off these purchases should customers restock via the online channel. This is a great reminder to actively involve people with different domain expertise before investing the time and resources to carry out the data mining process. Not all data scientists know the ins-and-outs of how the commission structure of a company works and therefore it is crucial to collaborate with different business groups. This makes up part of the constraint that will guide the data mining process - what kinds of solutions should we use?
  Problem Formulation - Question

Jake Porway interview

- Porway explicitly cites data empathy as a desirable skill that data scientists should possess. Data scientists should be reflective and need to consider the context in which the data has been collected, in order to uncover any potential biases before delivering insights.
  Data Collection and Cleaning - Maxim

- “In short, learn statistics and be thoughtful.” Not only do data scientists have to validate the results of their model through statistical knowledge and common sense, they also have the responsibility to conscientiously pick the most appropriate algorithm and to communicate their results in a way that is most impactful to society, etc. With data comes power, and with power comes ethical responsibilities.
  Problem Formulation/Analysis and Modelling/Presentation and Integration into Action - Ethical Commitment

Plan for Knowledge Acquisition

Skills and Knowledge Inventory: Stage 1, Problem Formulation

How to conduct an inquiry in my application domain that leads to a good problem formulation

- I look forward to strengthening this capability. I previously worked on a project as a data analyst at EDHECInfra, which is an academic research institution to help investors benchmark private infrastructure investments in the debt and equity markets. The project was to predict the size of the global private infrastructure market (eg: privately-owned pipelines/ports/Public-private partnerships, etc) with the invested deal values we obtained from an external database provider, while matching these values to the fair value of total assets of each of these companies/SPVs. I was heavily mentored by the head of AI and the CEO in coming up with the business questions and what the regression models would look like (eg: what independent variables are to be included).

- How to attain this knowledge**:** I am aiming to, in the future, carry out a whole data mining project with little guidance by attaining the required comprehensive knowledge through the MADS program, specifically SIADS-501 and SIADS-505 which I will be taking this month, and SIADS-503 in the coming months. At the same time, I have followed influential data scientists on Linkedin who regularly share coding tips and interesting articles about data science. These courses and external content will enable me to better understand the data mining process and ideally, I'd like to revisit this project when I have attained the necessary foundation to see what we might have missed out, and take part in future projects more effectively.

A repertoire of problem types

- I look forward to strengthening this capability. I have only been really exposed to supervised learning techniques (Classification and regression) throughout my academic and professional careers, through the uptake of my undergraduate program via econometrics, the specialist diploma in business analytics and at work.

- How to attain this knowledge**:** I am looking to expand my knowledge into other problem types through future courses in the MADS program, specifically SIADS-642, SIADS-542 and SIADS-543. I googled each of the problem types after the video lecture and have found a lot of literature and implementation of these techniques online (Eg: Kaggle), and will aim to delve into them in the coming months.

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

- I look forward to strengthening this capability. As mentioned, the previous data mining project “market sizing” was heavily guided and implemented by senior members of the team, so I had little involvement in deciding or formulating the problem types to be explored and implemented at my current workplace.

- How to attain this knowledge**:** I believe as I move along the MADS program, both courses SIADS-542 and SIADS-543 will allow me to implement and fine-tune the right supervised and unsupervised machine learning problems to solve all types of business problems formulated, beyond my current domain in infrastructure project finance, with the aim that I may hopefully transition towards being a generalist data scientist who can extensively apply different problem types to different datasets.

Application in Domain of Interest

Domain: *Finance - Banking*

Project 1 Description:

- Bank XYZ has instructed you and your data science team to come up with a credit card fraud detection model in order to detect fraudulent transactions as early as possible so that the bank may minimize its losses. At the same time, upper management would like to know if it is possible to improve the level of security for its customers in order to improve on the accuracy of determining whether a fraud has most likely occurred for each of its customers.

Project 1 Problem Type:

- This is a time-series anomaly detection problem with the use of the time variable (seconds for eg) and transaction amount (checking to see whether these transaction amounts are outliers) for each customer. Data on each customer such as age, sex, job, saving accounts ("little" to "rich" bins), type of transaction, item of transaction, and whether the transaction is an inflow or an outflow can determine each behavioural pattern, which can help improve customers' security.

- Additionally, since the data is immense, data reduction through Principal Component Analysis (PCA) should also be used for easier visualisation and analysis without too much informational loss. With PCA, machine learning algorithms can operate faster to keep the data mining project timeline intact.

Problem 2 Description:

- Your manager at Bank ABC is trying to understand why there is an increasing trend in the number of customers abandoning the credit card service. You have been tasked to predict which customer is going to churn so that the relevant customer service team can improve on better service delivery in hopes that these customers change their minds.

Project 2 Problem Type:

- This problem specifies a target of whether or not a customer will likely churn. Therefore, this requires a supervised method using classification (probability estimation) technique which will produce a result that is binary; Will churn or will not churn, based on available customer data such as customer demographic, type of credit card, credit limits, credit card usage, customer dissatisfaction levels, etc.

Questions, Maxims, and Commitments

Question (I will always ask…)

- How can the data mining result be implemented?

Which Project

- Introducing a credit fraud detection model in order to minimize losses and maximise customer security

Meaning in Context

- In this context, the objectives in the problem formulation stage are straightforward (to come up with a credit card fraud system and at the same time ensure customer security). It is, however, important to understand how this model can be used for the bank. For instance, once the model determines an anomaly, it can trigger an automated call to the affected customer to verify the transaction, and flags the fraud detection team immediately once the customer indicates that this was in fact a fraudulent transaction in order for them to freeze the transfer of funds.

Importance

- By asking this question at the first stage of the data mining project, the data science team can understand who the stakeholders are and to include them in the problem formulation stage in order to maximize time and resources efficiently, and to ensure that the data and the right model can be obtained and used respectively.

Maxim (I will always say…)

- I will always say “avoid making work for others”

Which Project

- Predicting the likelihood of customers abandoning their credit card services

Meaning in Context

- In this context, it is inevitable that the data science team requires coordination with the credit card and customer service department (and possibly even the call centre who receive calls of abandonment by these customers) in order to get the data they need, communicating and agreeing on how the final result should look like. In order to efficiently do this, teams have to understand what features are available in the datasets (eg: pre-empt any quality issues, etc), and what model should be reliably used to decrease the likelihood of churn.

Importance

- With different teams agreeing on how the final result should look like, the data science team can initiate the data mining process without much interruptions of having to continually revert to the problem formulation stage (setting up many meetings with the different departments or requiring them to provide data in another format).

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

- I will always minimise inflicting potential harm

Which Project

- Predicting the likelihood of customers abandoning their credit card services

Meaning in Context

- In this context, it is crucial to ensure that the results yielded with the modelling technique used do not generalise the insights to certain demographic variables such as racial background, age or income. If such insights are released to the wealth management team, this might inadvertently promote discriminatory sales practices and cause reputational damage to bank ABC, despite how successful the bank has become in decreasing likelihood of churn.

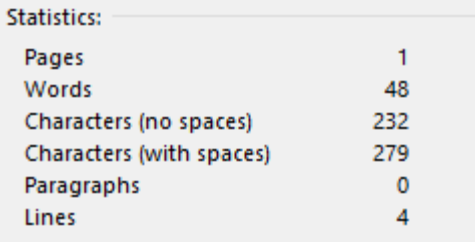
Importance

- It is important that the data science team instill practices that are best suited to battle this by planning in advance to make sure the assumptions used do not give the model the ability to segregate the insights in this manner. At the same time, setting out clear lines of responsibility for owners of the model to monitor the performance and have regular reviews of the model throughout the data mining project.

Week 2: Data Collection and Cleaning Stage

Potential Personal Project Tweet

- According to a loan dataset from Kaggle, if we cluster customers based on whether they have deposit accounts, housing loans or both, the odds of taking up personal loans are much higher for those who possess both. This should be their target when banks market for personal loans.



Statistics:	
Pages	1
Words	48
Characters (no spaces)	232
Characters (with spaces)	279
Paragraphs	0
Lines	4

Reading Responses

Law of Small Numbers

- "...results of large samples deserve more trust than smaller samples..." It is always important to be careful in the decision-making process that is reliant on a small data set. The insights derived might not necessarily give us an accurate account of what we should actually observe as there is an increased likelihood of more outliers in small rather than larger samples of data. Therefore when analyzing data, we should be mindful of the bias that lurks here to avoid making false conclusions.
  Data Collection & Cleaning - Maxim

- "We are pattern seekers." Although it is innate in us, built by the processes of evolution in order to survive, I must not settle on causal explanations in order to look for relationships when there is none. It is a vital part of analysing our modelling results to determine if there truly exists bi- or multivariate relationships within the data. As data scientists, it is our ethical responsibility to produce non-misleading results.
  Data Analysis & Modeling - Ethical Commitment

Statistical Biases Types Explained

- "Biased statistics is bad statistics." Intentional or not, providing misleading information to end users through misleading visualisations and flawed correlations from less-than-thorough analysis can have detrimental effects on the business. We must check whether these correlations and results from our analysis are reliable based on statistical knowledge, logic and investigative work.
  Data Analysis & Modeling - Maxim

- "...find trustworthy raw data and do your own analyses to learn a "truer truth."" Conducting several evaluations on the data, especially if it is given by "stupid" humans, is vital towards mitigating any potential biases that lurk within it. Do not underestimate exploratory data analysis - do this often and learn to appreciate it.
  Data Collection & Cleaning - Expertise

Data Cleaning 101

- "Does your data make sense?" It's an important practice to always check for outliers, whether the column data matches the label in its meaning and format type, and whether the data contains the variables needed to properly capture the phenomenon I am estimating or predicting. If anything seems out of place, always ask questions to clarify. As the article mentioned - "communicate with the source".
  Data Collection & Cleaning - Question

- "Use the tool that makes sense." There are tons of data science tools out there to be used but it ultimately relies on what kind of data I am working with and the task with which I have to carry out for a data mining project. At the same time, always be mindful of time constraints and rely on tools I am most comfortable with.
  Data Collection & Cleaning/Data Analysis & Modeling - Maxim

10 Rules for Creating Reproducible Results in Data Science

- Have the data cleaning process and modelling results documented at every step in a standardized manner for sake of replication and future-proofness (Eg through version control system). One way of ensuring quality of the data results is if the documented steps can be executed on another identical and larger dataset. Doing this also inculcates a culture of quality data in data mining projects of an organisation, and allows for higher overall productivity.
  Data Collection & Cleaning - Maxim

- Make your code comprehensible. It is impossible to remember all discussions or processes in a data mining project. Not only that, but very often a project may require scaling and can involve other departments with non-technical backgrounds to be added into the process. It is good practice to ensure that I always add footnotes and organized pieces of code so that it is readable regardless of individual domain expertise.
  Presentation and Integration into Action - 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 look forward to strengthening this capability.

- There are lots of opportunities to gain expertise in the data collection and cleaning phase through online articles (even stack overflow) written by experienced data scientists or analysts. However, it is worth noting that this knowledge can only be transformed into a habit through practice. Experimenting with datasets in Kaggle will give me ample opportunities to make certain data cleaning steps a habit. I am also concurrently taking SIADS 505 which is a great foundational course in numpy and pandas libraries in Python programming. In the future, I am looking forward to other courses in the MADS programme, specifically SIADS 511 & SIADS 631 to reinforce the skill sets acquired that can help eliminate certain biases and improve the data collection and cleaning process in my work.

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

- Although I have some previous experience in this, I look forward to strengthening this capability.

- In my work as a Data Analyst at EDHECInfra, it was an important step to try to get as much data as we can on private infrastructure companies that we track. Very often, this includes trying to get better deals with relationship managers (stakeholder management is essential in the data collection phase) of several database providers for a team of 6 users. Looking for as many reliable data sources is key towards executing an insightful project. I have also started a Github account and will look forward to following great coding content through others' repositories to see how they scrape for data and link different data sources up in their codes.

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

- Although I have some previous experience in this, I look forward to strengthening this capability.

- The market sizing project required constant revisions and it was paramount for me to document at which stage and where I attained the various datasets from. There was an instance when I was punished for not documenting steps for acquiring a few data

sources when the Director suggested changing the whole premise of the project, and it consumed a lot of time revisiting certain development branches to try to recall lines of thinking as to how and why we came up with several of these assumptions for cleaning the data. Having a confluence documentation that links to the github repository of the work helps with identifying or laying out a brief overview of how we should approach the data collection and cleaning phase of the project. Kaggle also has a lot of practical examples of how experienced coders and data scientists document their processes and I am hoping to pick up the best examples that are most effective and comfortable for me.

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

- Although I have some previous experience in this, I look forward to strengthening this capability.

- I have had experience writing scripts/codes that scrape data from websites through the specialist diploma course. However, I believe SIADS 505 will also supplement my current skills and I am also looking forward to taking SIADS 516, which will give me an idea of how to solve hard practical issues such as scanned PDF annual reports, etc. I am hoping to gain some insight into this. In the meanwhile, kaggle content providers will also give me a glimpse at how they solved extracting from hard-to-get data such as satellite images, etc.

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

- Although I have some previous experience in this in R, I look forward to strengthening this capability.

- SIADS 505 will teach me the basics of doing this through Python. There are also a few other courses in the MADS program that will help me improve on this skill, but gaining expertise on this requires more than just learning, but it also requires practice more than anything else. I am hoping to revisit the python 3 programming specialisation I completed prior to undertaking the MADS program as a refresher, relying on several youtube bootcamps that code for 5-7 hours straight to get my basics right and applying these skills on the many datasets that kaggle provides, all at the same time while undertaking SIADS 505. I have also followed a few influential data scientists or aspiring data scientists who regularly share coding tips or articles on how they can effectively handle complex data issues or problems whilst at the cleaning stage. There are also articles on google that effectively talks about all the essential data cleaning steps (a general guide of cleaning any kind of data) that eliminate biases and ensuring that proper EDA has to be performed in order to efficiently extract features that are meaningful.

Maxims, Questions, and Commitments

Question (I will always ask…)

- I will always ask “Is the data representative or is it biased?”

Which Project

- Predicting the likelihood of customers abandoning their credit card services

Meaning in Context

- This data will have to be collected from other departments that keep a record of customer credit subscriptions, credit limits, and sales contact whether it was a face to face subscription or online, etc. Such data will include sensitive customer details such as age, racial background, job, etc. It is vital at this data cleaning stage to ensure that the data collected is representative according to the dependent variables I will be using to model. If the number of Indians are only a handful (or not proportionate to the number of Indians in Singapore for instance), whilst the number on chinese credit card subscribers are large, the data is definitely skewed and there exists a selection bias in the sample data. Another example is if the data collected only contains customers with over 100k as annual salary.

Importance

- Modelling and making churn predictions based on these modelling results will definitely segregate customers inadvertently based on racial backgrounds or income, and its accuracy will definitely be affected when used in deployment (detrimental effects to the business in terms of reputation). Therefore, we need to seek reasons why the data sample is not representative (by talking to the sales team, for eg) and/or how can the data scientist go about collecting the data. In this scenario, EDA is an especially important step in visualising potential biases that may already exist in the dataset.

Maxim (I will always say…)

- I will always say “Proper documentation will save me and my business.”

Which Project

- Introducing a credit fraud detection model in order to minimize losses and maximise customer security

Meaning in Context

- It is essential that the project is well-documented, specifically the data collection and cleaning phase. When it comes to sensitive data concerning customer details, it is imperative to ensure that this raw high dimensional data coming from various departments/sources have to be interpretable in order to maximise customer security based on their behavioural transactional patterns. This requires extensive cleaning, transforming and constant massaging of data so that meaningful information can be extracted. With such a multi-faceted workflow, proper documentation is key to ensuring accurate results.

Importance

- When it relates to real-time customer transactional data, it is only going to increase exponentially. More data requires that I document every process as I anticipate that the project will inevitably be scaling up as the bank grows and/or customer transactional patterns change. Proper documentation will save my company money through time, efficiency and ultimately from losing money by missing out on frauds due to poor documentation.

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

- I will never let any assumptions about the data go untested

Which Project

- Predicting the likelihood of customers abandoning their credit card services

Meaning in Context

- We have established our innate flaws as humans to always be pattern seeking. Indeed, at the problem formulation stage, we would have had discussions of what results we can expect based on theory and/or business operating history and insight. However when it comes to data that include sensitive information that may potentially have business reputational risk at stake, such as demographic variables or other variables that may be misconstrued to be socially divisive (eg: only rich people can afford to be sub-optimal in their credit card selection that they keep to the same service), it is always important to first check the quality of the data and to test the collective and/or innate assumptions made.

Importance

- I must never let my assumptions about the distribution and contents of the data go unexamined. In fact, should there be a change in business objective that requires a change in the supervised model used from random forest to logistic regression, the accuracy of these models depend on different assumptions that range from independence, no multicollinearity, etc. Running rigorous EDA through sequence plots, scatter plots and histograms can help test my assumptions in order to avoid bias to creep into the modelling and analysis phase.

Week 3: Data Analysis and Modeling Stage

Informational Interview - Reflection

- “How do you know if the results of your model are being used properly and it's not being misused or misinterpreted?” Ethical data science practices entail not just being able to interpret the outputs of a model but also interpreting the inner functionings of the modelling techniques used, to get around the “black box” problem, and to ensure due diligence at the workplace. For any misapplied black box models on sensitive data such as financial, health, hiring, etc, the results and insights drawn from the models can invariably lead to unforeseen consequences that harm societies. I must be able to anticipate that some models amplify biases and must balance modelling techniques based on complexity and interpretability.
  Data Analysis & Modeling - Ethical Commitment

- Renée, coming from a systems engineering background, talks about the biggest ethical challenges that data scientists face and that is the question of bias-ness in datasets, which she referred to as “..historical racism that’s baked into systems”. Machine learning techniques are really only doing pattern matching, and it’s a lot like stereotyping. I have to be aware of any biases present in the data at the collection and cleaning stage. It is necessary for me to ensure that this phase is done thoroughly; not just through proper data cleaning but going beyond and asking relevant questions from different departments to ensure understability on my part in order to derive justified assumptions, before moving forward to the modelling and analysis stage. Quite simply; “Is my data representative?”
  Data Collection & Cleaning - Question

- Renée emphasizes that beginners should approach their story-telling (eg: analysis of their results) in a way that makes people comfortable. She also emphasizes this point later in the podcast by saying: “communication skills are really important too, not just the tools and techniques.” I must present my results and recommendations of a data-mining project without getting too technical and to ensure the information I present is usable (value-adding component). This brings me back to the point of comprehensibility in week one, when presenting data-mining results to relevant stakeholders and as explicitly mentioned by Vicki Boykis; I must be a meaningful and effective explicator as a data scientist.
  Presentation and Integration into Action - Maxim

- This podcast is a great reinforcement of SIADS 501 and helped me home in on the maxims, questions and commitments that I have been documenting throughout these past 2 weeks. There are other questions that I would have liked to ask Renée;

  1) What is the most dreadful part about being a generalist data scientist?

  2) What kinds of models were better than others in predicting a university student's likelihood of enrolment?

  3) With so much emphasis on ethics in data science, do you think there is a need for a call for stronger regulation around data science to ensure more fairness and transparency?

Statistics:	
Pages	1
Words	472
Characters (no spaces)	2,482
Characters (with spaces)	2,961
Paragraphs	7
Lines	46

Reading Responses

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

- "Is there a simpler model that I can pick?" I have to ask myself this when I decide on a modelling technique to use on the data. One of the main ways to reduce likelihood of overfitting is to constrain the complexity of the model. For example, if I am using neural networks, and the results are showing poor accuracy on the test dataset, I should consider removing layers to make the network smaller.
  Data Analysis & Modeling - Question

- Early stopping is effective and simple, do it often. As the model is evaluated iteratively through a large number of training datasets and as the model's performance starts to degrade, then put a halt to the training process. With k-folds validation set, training should be stopped at the point of the smallest error, which typically happens with overfitting as very often there is a decrease at first, then followed by an increase as the model overfits. Early stopping helps minimise generalisation error, thus decreasing the chance of overfitting .
  Data Analysis & Modeling - Expertise

Common pitfalls in statistical analysis: The perils of multiple testing

- Avoid the problem that comes with multiple testing. I should be making a well-informed decision about which hypotheses to test instead of testing all x number of hypotheses. Although testing more times seems like a more thorough examination of the data, it also comes with more erroneous occurrences (false-positive results) which will only complicate the analysis process even more.
  Data Analysis & Modeling - Expertise

- "Results from single studies should not be used to make treatment decisions." Interesting findings from a single test result can only be recommended to stakeholders if it makes meaningful business and economic sense, backed by other sources or studies that can validate this result. Communicating false results even though they are interesting, can lead to wrong business decisions that can cause unintended consequences to relevant stakeholders.
  Presentation and Integration into Action - Ethical Commitment

P-Hacking and the problem with Multiple Comparisons

- It is easy to fool myself and others if I do not adjust for multiple testing. Multiple comparisons are justified so long as I disclose it as exploratory and document the process. However, I should replicate myself using a new sample (the test data set) and include the replication in the same paper, to ensure robustness of my findings. In other words, do not report results that are just a "function of point and click statistics."
  Presentation and Integration into Action - Maxim

- "Of the three, this one is the worst." HARKing stands for Hypothesizing After Results Are Known, which as the name implies, goes against the principle of scientific research. An analyst is HARKing when he/she is testing a hypothesis, after an analytical study has already been made, as if it was set out from the beginning. As a data scientist, I should not give in to the HARKing temptation. Doing so increases the likelihood of committing a type 1 error. Additionally, this unethical practice produces a wastage of resources such as time and money since more replicated studies, with no true effect, are made. Even if I need to publish something, I must not change my hypotheses post hoc in hopes of producing project results that are merely eye-catching but are based on wrong insights.
  Data Analysis & Modeling - Ethical Commitment

Correlation vs. Causation: An Example

- "What are some possible confounding variables?" It is important for me to consider any variable Z that is causing an association or correlation between variables X & Y. If I can pin down Z, then I can report that the relationship between X & Y is just merely a spurious correlation. Only after controlling for Z, if I still observe a statistical significance between X & Y, then I should make a conclusion of a possible causal relationship. To control for confounders like Z, I can conduct randomized controlled experiments, multivariate regression analysis and causal inference assumptions backed by logic, business acumen and domain knowledge.
  Data Analysis & Modeling - Question

- "...create a skeptical community in which we make sound decisions for our benefit and not for a company's bottom line." I believe this will reinforce ethical standards in the field of data science. Data scientists will then be more rigorous in the insights derived in their modelling techniques, ensuring unbiased reporting and presentation of findings. Even though this may entail less interesting reads for the public, it will significantly curb fake news that is increasingly becoming prevalent in today's society. I must hold myself to a high standard of analysis and recommendation in order to deliver true and accurate findings, even if this means less citations or readership of my results. If it means just mere correlation, then I must report it as such.
  Data Analysis & Modeling/Presentation and Integration into Action - Ethical Commitment

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

- "...it can easily bewilder the statistically naive observer." As a data scientist, armed with knowledge in statistics, I should not always assume that statistical relationships are immutable. The strength of statistical relationships are more often than not, affected by controlling for a third variable (covariate). Simpson's Paradox reminds me that causal interpretations should be made with caution.
  Data Analysis & Modeling - Expertise

- "What factors are affecting the results I don't see?" During analysis, I should always look at the data from different angles through grouping and segmentation. Effectively, I can resolve Simpson's Paradox by stratifying the data to control for a third confounding variable. Just looking at the aggregated data alone can obscure the true relationship between the variables being studied.
  Data Analysis & Modeling - Question

Conditioning on a collider

- Never condition on a collider. Doing this will introduce bias when estimating the correlation or relationship between variables of interest. This will result in me mistakenly concluding an association between the two variables when in fact, there are none.
  Data Analysis & Modeling - Maxim

- I need to know how to deal with the third variable Z, whether it's a collider, mediator or a confounder, and the difference between all three depends on the direction of influence.

  1) A confounding variable Z is one that causes both X & Y (condition on it!).
  2) A mediator Z is one that causes mediation between X & Y, such as being caused by X which in turn, causes Z to influence Y (condition to get direct effect!).
  3) A collider variable Z is caused by both X & Y (as above maxim, never condition on it!)
  Data Analysis & 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

- I look forward to strengthening this capability.

- The week 3 content is new to me even though I went through a few complex statistical courses through econometrics in my undergrad study. I feel like it is very hard to retain the concept of different pitfalls in causal inference through confounders, colliders and mediators merely by understanding theoretically what they are. I should practice them on kaggle datasets, and also read extensively on the many articles and youtube videos to test my understanding of them and try to find real life examples of identifying such third variables. Hopefully, I will be able to identify them confidently as I progress in the MADS program, with the help of SIADS 521 & SIADS 522, when I embark on the milestone courses (SIADS 591 & SIADS 592).

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

- I look forward to strengthening this capability.

- I believe this is one area or stage in the data mining process that I need most improvement in. The MADS program's courses through SIADS 542 & SIADS 543 will expose me to the different techniques used in both supervised and unsupervised learning. Additionally, there are alot of sample scripts posted on the kaggle community that clearly enumerates why certain techniques were used and how they were validated and interpreting on a variety different types of datasets. Combining these two sources will give me a reasonably good grasp of a repertoire of models. I have also followed a couple of MADS program friends on Linkedin and they have also been sharing interesting articles on different modelling techniques which I find interesting.

Maxims, Questions, and Commitments

Question (I will always ask…)

- I will always ask “Am I at risk of overfitting?”

Which Project

- Predicting the likelihood of customers abandoning their credit card services

Meaning in Context

- In this context, the aim is to predict likelihood of customers churning and this is dealt in terms of binary values that are operated under probabilities that depend on the independent variables I use in the model. It is possible to create a model that takes in many inputs, causing it to learn the customer training data set perfectly. However, this model will perform very poorly when deployed. The algorithm of a model that has too many inputs has incorrectly recognized noise as signals and the accuracy of such a model is very poor. There will be no way for me to discern why certain “yes” and “no” values pop up for certain customer profiles, once such an overfitted model is deployed.

Importance

- Given an inaccurately overfitted model, the results will no doubt be inaccurate. This will cause poor resource allocation on the part of stakeholders operating under this model. Although it is logical to assume that with more independent variables, a model’s performance can improve. However, this is not the case. Overfitting decreases a model’s ability to predict. The idea behind creating a model is to use it on customer profiles it hasn’t seen, and to produce an unbiased result, not to use it on customers who have already churn.

Maxim (I will always say…)

- I will always say “Correlation does not equal causality.”

Which Project

- Predicting the likelihood of customers abandoning their credit card services

Meaning in Context

- I should never assume that just because certain variables are correlated that one necessarily has a direct causal influence. In the context of this project, if the geographical location of customers happens to be a significant variable (For instance, likelihood of customers abandoning their credit cards with the bank is higher in one location than another), it is worth looking deeper into the reason why geographical location came up to be significant. Was it just by statistical chance? Was it due to a mediator variable such as a competing bank marketing more intensely and visibly in that area more so than others? It does not necessarily mean that the geographical location in and of itself as a feature, caused customers to abandon their credit card services. If I am able to identify the level of intensity of a competitor’s marketing campaign causing certain geographical variables to be statistically significant, this insight can be used to inform my own company’s marketing strategy, thereby ensuring less likelihood of churn in those areas.

Importance

- If studying for an exam does increase my grades, then there's every reason why one should study. However, in the business world, issues relate to sensitivity questions like how much should I study to get this grade? In order for me as a data scientist to recommend a course of action to stakeholders, I have to ensure that any intentional changes in one variable affects the outcome variable. Hence in determining a causal relationship, it is not just enough to have a statistically significant variable. It is also important to identify any confounding variables that create spurious correlations and ensure that any confounders found have to be controlled for.

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

- I will never succumb to producing inaccurate results despite it being favourable.

Which Project

- Predicting the likelihood of customers abandoning their credit card services

Meaning in Context

- Very often, management might want data scientists to produce surprising results that justify a stakeholder's or department's poor performance without absorbing too much blame. In this context, the customer service department might pressure me to produce results that are in favour of their performance such as poor customer service being an insignificant variable in predicting customer abandoning credit cards. If i give in to such pressure, the bank will continue losing money through customers leaving the bank's credit card services as no changes/reforms were made in the credit card customer service department (if in fact poor customer service was a significant variable in predicting customer dissatisfaction through credit card abandonment).

Importance

- It is crucial that I must not give in to pressures simply because it helps other departments or stakeholders of the business. As mentioned, given that I am a data scientist armed with statistical knowledge and a range of datasets, I must ensure that the analysis and outputs of my model are reported in a transparent, fair and accurate manner. Producing inaccurate results or creating models to fit pre-conceived notions of what's expected of any stakeholder (including myself) is unethical in any context.

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:

- Subscription to **Ken Jee's youtube channel**. Ken produces informative Data Science content through insights for the data science community, career advice, and also sports related analysis which is transferable in terms of technical skills (eg: coding). He is the Head of Data Science at Scouts Consulting Group and produces content on the youtube platform once a week. Ken's content includes interviews with data science leaders who generally talk about the evolution within the community, etc, which will keep me up to date with not just the finance industry, but also other industries as well.

- **Kaggle** is a great website that includes datasets that are actively worked on by the machine learning community to showcase and test their models, with a leaderboard for each competition and general coding tips and advice. Hopefully one day I can be confident enough to join the competition! The website will also keep me up to date in terms of the latest techniques or libraries used over time as new datasets are added in.

- **Blog - Towards Data Science** is a goldmine when it comes to data science articles ranging from the theoretical side in Bayes Theorem to practical uses through spam detection using logistic regression. The blog does allow for a filter called "Editor's pick" which filters for the best daily articles on data science. This will keep me up to date on data science in general ranging across different industries.

- **Linkedin** - I already have an account here and have followed many classmates through the MADS program who regularly share data science articles, experiences and even jokes. The social media platform also allows me to follow other influential data scientists such as Lex Friedman, etc. More importantly, it allows for data science job applications which details specific job descriptions and skills needed. Through constant monitoring, I can observe what skills and knowledge are needed in different industries and can keep myself up to date in that way.

Reading Responses

A History Lesson On the Dangers Of Letting Data Speak For Itself

- "...focus on identifying open-minded allies who can help build internal support and consensus for your ideas." If Semmelweis was able to garner at least some level of support from his colleagues, he would have not had such a hard response when presenting his recommendation. I must understand my stakeholders' existing attitudes to business processes, or at least try to, before presenting my recommendations on a company-wide level. This process requires active engagement with stakeholders (setting up mini-discussions, etc) and through time, I can get some level of agreement with my business recommendation, before disseminating my results publicly. Only from that standpoint can I deliver results that are more persuasive, backed by some level of colleague or peer support. This approach can also let me know if I am missing out on anything else. It would also allow for a more constructive discussion, rather than a flat out "no this is not feasible" in the decision making process.
  Presentation and Integration into Action - Expertise

- I must aim to touch my audience emotionally. The author fleshes out human biases innate in us, and in particular, I thought of status-quo bias in decision-making through his statement "All great truths begin as blasphemies." Even with convincing evidence backed by sound reasoning, logic and substantial data, human nature can often get in the way by resisting changes in habits, or business leaders or management resisting changes in processes. One way that can help me negate this is to narrate and communicate my insights that can spark emotions that relate to the audience so that they follow along the story and become more emotionally invested. Indeed, visualisations will help significantly. Understanding such cognitive limitations will allow me to frame my presentation slides differently and more effectively.
  Presentation and Integration into Action - Maxim

Storytelling for Data Scientists

- "Data makes people think, emotions make them act." A recommendation taken in objectively by the audience requires just logic and reasoning backed by data. However, for the audience to act on it and put the results to deployment requires an active persuasion that connects with them emotionally. Only if my results can make the stakeholders care, will they act on it. To make them care, as the author described, my results should connect to at least one of the five emotions of fear, happiness, surprise, sadness, and love/hate.
  Presentation and Integration into Action - Expertise

- Do not bore my audience with minor details. I have to focus on what is important and keep it simple and straightforward. The audience has a short attention span and it is best that I use the limited time, focusing on the most important messages I want to convey. As Professor Paul Resnick mentioned in his lecture, I do not have to bring my audience along through the tedious data mining project in the presentation.
  Presentation and Integration into Action - Maxim

Interpretability is crucial for trusting AI and machine learning

- Will I be able to explain this model to different stakeholders? I, as a data scientist, will have to present recommendations and insights gathered from a data mining project to end users who come from different perspectives and have various needs. Accuracy at the expense of complexity can sometimes nullify my results, and therefore I must strike a balance between a model's interpretability and accuracy (which is very dependent on the purpose of the project) when picking a model.
  Presentation and Integration into Action - Question

- "Understanding your data set is very important before you start building models." In order to make the input-output model explicable to end-users, I have to first understand what independent variables I need that will accurately predict the dependent variable. This entails a lot of investigative work through data exploration, treatment and manipulation. It is very hard to select a model that will get the job done, reliably, without having a complete understanding of the limitations and characteristics of the data set.
  Data Collection & Cleaning/Data Modelling & Analysis - Expertise

The Signal and the Noise, Chapter 2

- Think probabilistically as a matter of principle. It is the bedrock of statistics that we cannot prove anything to a 100%. It is always important for me to communicate this, to allow for margins of error in the model's output, not just to be a reliable and honest data scientist but to also manage expectations of stakeholders.
  Presentation and Integration into Action - Maxim

- "Play enough poker hands, and you'll make your share of royal flushes." This reinforces the above maxim, in that, data scientists have to be open to a range of possible values including extreme outcomes. In other words, I will always deploy a range of possible values or outcomes to the audience when presenting my results, and not rule out the possibility of extreme predictions. Leaving room for a multitude of possibilities will help the business anticipate outcomes.
  Presentation and Integration into Action - Maxim

The Signal and the Noise, Chapter 6

- Decision-making also requires knowledge of the uncertainties. In any scientific process, especially one as dynamic as the business world, it is vital for me to communicate not just my results with varying degrees of certainty but also any uncertainties relating to the insights gathered. It is important to communicate all information, rather than bits and pieces of information that I am confident about because incomplete information, as the reading alluded to, can have the potential to mislead the audience, and can be detrimental.
  Presentation and Integration into Action - Maxim

- Making assumptions that lower the level of uncertainty isn't always preferable. Uncertainty isn't always an enemy and it is better to disclose and communicate uncertain outcome, causes, variables, etc right at the outset rather than minimising uncertainty just for the sake of minimising uncertainty. Reducing uncertainty by making bolder assumptions does not always improve a model's accuracy.
  Data Modelling & Analysis/Presentation and Integration into Action - Maxim

How Not to Be Misled by the Jobs Report

- "Human beings, unfortunately, are bad at perceiving randomness." As with week 3's reading, we are all pattern seeking. If we look hard enough, we may perceive signals in the data but they may not necessarily be right. We are bad at accepting the results of randomness. To this end, it is important to evaluate my results to see if there really is a pattern in the data. Communicating and presenting patterns that seemingly exist, but do not, will mislead the audience and there can be dire consequences for the business in terms of cost, reputation, etc.
  Presentation and Integration into Action - Maxim

- "The sampling error becomes magnified because those of us following the jobs report don't focus on the total number of jobs in the economy (more than 130 million). We focus on the relatively small change in the number of these jobs from month to month…" The article highlights the importance of framing my results when it comes to presenting them. I should put an emphasis on what the audience needs to see in order to decipher a meaningful trend and should flash out other contradicting and wrongly perceived trends in the presentation stage.
  Presentation and Integration into Action - Maxim

But what is this "machine learning engineer" actually doing?

- “It is often unclear how to ensure that your implementation actually works properly.” Before putting the model into production, it is always desirable that the machine learning engineers and data scientists plan on how to deal with different types of scenarios of failure. Machine learning engineers cannot possibly foresee all the possible solutions to every facet of a complicated task that come with efficient scaling and debugging. “Make friends with the Engineers” is an especially important maxim I borrow from Professor Paul in his lecture.
  Presentation and Integration into Action - Maxim

- Be humble about how much work I have done. Sitting at the intersection of data science and software engineering, machine learning engineers are the ones who will use their technical expertise to put the models I create into production. They feed data into the models and transform these models to production-level models that will handle real-time data. This, invariably, requires extensive technical skills, time and resources and oftentimes, this is much more work.
  Presentation and Integration into Action - Maxim

How we scaled data science to all sides of Airbnb over 5 years of hypergrowth

- “Over time, our colleagues on other teams have come to understand that the data team isn’t a bunch of Vulcans, but rather that we represent the very human voices of our customers.” The culture at Airbnb rests on a meaningful and collaborative cross-departmental mindset - one that ceases to think that the only thing data scientists care about are numerical figures which they can manipulate. Data scientists should not be defined as pattern-seeking robots, they are mediators that provide customers with what they need from the business. Therefore, the data that is gathered at a company shouldn’t be viewed as purely numbers but instead voices of customers. This takes customer service standards into serious consideration and decision-making to be geared towards driving the business (listening to what customers really need/want).
  Presentation and Integration into Action - Maxim

- I will always not be quick to toss my results of an analysis “over the wall”. Even when I am in urgency to move onto a next problem, simply handing over statistics to another department without clearly communicating an important insight can render the insight useless. Management can potentially lose out on an opportunity to act and this can mean losing out on cost-savings, opportunities to other revenue streams, etc. Hence, proactive communication on my part for all insights is key towards a value-added data mining solution.
  Presentation and Integration into Action - Ethical Commitment

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.

- As I progress in the MADS program, I believe I will acquire the knowledge through SIADS 521, 522 and 523. Although I had some experience in this, I believe presenting the market sizing studies to industry leaders on private infrastructure investments gave me some pointers on what I should focus on. There are also articles online that will enable me to become a more effective data science communicator. At the same time, my job also requires me to do presentations from time to time and I am hoping to leverage on the MADS program, data science blogs, podcasts to keep me up to date on what is happening in the various industries, all in the process of making me filter what is important and what isn't so important when it comes to presenting any data science results.

how to work with software engineers to put models into production

- I look forward to strengthening this capability.

- I have some experience in this before in the project I was engaged with at EDHECInfra, as I worked directly with the Director of Index Production. I am hoping that SIADS 511 will give me a good understanding of what efficient data processing is, which will show me some foundational knowledge from the engineer's perspective. I am also anticipating that SIADS 642, 643 & 652 will give me a good grasp of the technical coding skills and knowledge required behind models, model maintenance and deployment.

Maxims, Questions, and Commitments

Question (I will always ask…)

- I will always ask “Will my audience be able to understand the input-output process of my model?”

Which Project

- Introducing a credit fraud detection model in order to minimize losses and maximise customer security

Meaning in Context

- In this context, customer security in banking is probably one of the most important factors that underlie the business foundation of any financial institution. Customers need to know that their banks have their privacy and account details kept safe and secured. The key word here is trust. Just as how customers trust banks, management will need to develop trust in the model I produce, and that model cannot be trusted if it is not fully understandable. If I am not able to explain how a similar buyer’s purchasing behaviour cannot be mapped to another because of differences in certain features or variables, this can cast doubt about the accuracy of the technique used, along with my reputation and viability as a data scientist. Furthermore, a company is not just risking her own reputation with false positive alerts of frauds, but also losing money. There are other dangers associated with trusting a “black-box” model, such as predicted results that point to unintended social biases that a bank must avoid at all costs.

Importance

- Interpretability of the model is key to finding out the value and accuracy of the results presented. As a data scientist, I have to present models that are useful to the business and this entails generating feature important values across all data points. Only through being able to explain the input-output process of the model, will the audience be able to glean into the insights of how they can produce specific results to value-add the business. In other words, for stakeholders to put any models I create into deployment, they have to understand that the model is a useful representation of a phenomenon such as consumer transactional patterns.

Maxim (I will always say…)

- I will always say “Speak for the results because they don’t speak for themselves.”

Which Project

- Introducing a credit fraud detection model in order to minimize losses and maximise customer security

Meaning in Context

- In this context, the credit card anomaly detection solution starts with many years of customer transactional data that themselves contain evolutions of purchasing behaviour. Figuring out the visualisations that communicate my results is a vital component of comprehension. It is obvious that through plots, consumer behaviour trends can be inferred but I have to be able to communicate any transactional correlation between different customers themselves against their previous purchasing patterns and also between customer groups, for example, and link that to the logic of how the model works in predicting fraudulent transactions. Communicating linkages within the data will help validate if the model works logically and in this context, whether it can be deployed to maximise customer security and minimise loss of money. It illustrates that simply presenting a model and its corresponding assumptions will not suffice. Good information delivery entails a presentation as a story with visualisations to aid the digestibility of the data.

Importance

- Without a good presentation design via framing (story-telling component) and visual plots that effectively communicate the analysis, the audience will not be able to fully grasp the information embedded in the data. Not many are well-versed in data science and data alone will not sufficiently communicate insights well enough to have a desired impact on a business. It falls on data scientists to actively communicate information in the form of a story that touches the emotions of the audience, using visuals that will help information to be delivered in a clear manner.

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

- Always communicate the uncertainty

Which Project

- Predicting the likelihood of customers abandoning their credit card services

Meaning in Context

- In this context, instead of saying "*a customer with at least $Y in the bank, who has been the bank's long standing client of over X number of years will not likely churn"*, I should deploy the result in terms of probabilities and frame it in the following manner; "*It is Z% likely that the client will not churn, given these attributes of the customer*." The other 100-Z% must be described to the audience as a margin of error that is dependent on the model. If I don't communicate this uncertainty to stakeholders, there can be wrong resource allocation decisions made. An example would be informing the credit card services to focus more on the wrong segment of a customer group. Another example would be providing discriminatory pricing (eg: discounts in the form of credit card fee waiver) to a customer segment who are not as likely to churn as compared to a segment that needs it most. Either decision is wrong. It results in a waste of resources and ultimately does not decrease customer churn. In fact, there might be an opposite effect and increase customer churn.

Importance

- Every model has some level of uncertainty. Rather than presenting results disguised under a classification algorithm that seems to perform perfectly, I must be upfront in letting my stakeholders know about the uncertainty surrounding my results. Delivering both certainty and uncertainty, in other words, delivering the full extent of information will help minimise inefficient decision-making for the management or stakeholders. Wrong decisions made allocationally for the business can sometimes lend themselves to undesirable consequences.