Personal Manifesto

By: Rahul Upadhyay

Table of Contents

Week 1: Problem Formulation Stage

Informational Interview - Planning

Person: Ye Yan, Senior Software Engineer at Magna International

About Ye Yan: Ye Yan utilizes Data Science methods such as Data Mining and Deep Learning in the field of automotive engineering. He has multiple years of experience in the field and is also well versed with the automotive industry's product development cycle. He utilizes his knowledge of Data Science to come up with solutions to better and more efficiently assess structural durability of automotive components.

Motivation for Interview: I work in the field of automotive engineering and aspire to learn data science. Ye Yan is a Data Scientist in the automotive engineering domain. Thus by interviewing him I will get to know about his experience as a Data Scientist in the field. This will guide me in my own journey.

Mode of Interview: Via Phone or Zoom Call.

Reading Responses

Readings:

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

  Excerpt from Paragragh-2, Business Understanding: *"but often the key to a great success is a creative problem formulation by some analyst regarding how to cast the business problem as one or more data science problems"*

  Response/Inference: The literature upto this paragraph outlines various methods in Data Science and various classes of techniques to find meaningful information from data. But, it was then emphasized that understanding of the end goal and formulating a problem statement to best address the end goal is as critical, if not more, than any step after that. An analyst's creative ability to formulate the given real world problem into a data science problem is instrumental in deciding the success of the problem solving exercise. The data science algorithms, discussions about computational efficiency etc. are all secondary and are driven by how the problem was stated and how it addressed the real world problem.

  Excerpt from Paragraph-4, Data Understanding: *"as we consider the relationship of the business problem to the data, we realize that the problem is significantly different."*

  Response/Inference: In the given paragraph, the writer alludes to how fraud detection in the medicare system is so different from how it is in the credit card industry. This is very interesting to note that problem statements that look seemingly identical in real world may need to be dealt with completely differently approaches in the data science world. To convert a real world problem into a crisp data science problem, its essential to understand the available data (or the lack thereof) and what information can be derived from it using the different Problem Formulation techniques. It may also be put like: Millions of lines of beautifully labeled data in csv formats should not entice the data scientist to blindly go for Supervised Learning.

- **Chris Wiggins interview**

  Excerpt: “but I didn’t think that they were using, or let’s say stealing, the appropriate tools for answering the questions they had”

  Observation: Although Dr. Wiggins does not assert that it is an atypical behaviour in academia or industry but he points out that often in academia, researches tend to limit there tools to those found in there own field. He points that Einstein did not constrain himself to methods in theoretical physics and employed techniques from mathematics and other fields in his scientific journey. Its essential to note that in the pursuit to solve a problem, one does not get fixated on one method or tool but makes effort to identify the best toolset to address the problem irrespective of which field or domain it comes from.

  Excerpt: “not everything that can be counted counts, and not everything that counts can be counted”

  Observation: In the field computational engineering there is a famous saying that “garbage in, garbage out’, which suggests that an ill defined model will lead to zero-value non sensical results. This thought can be extended to what Dr. Wiggins is alluding to. Even if the results are accurate and correct, they can still be adding zero or non-significant value compared to the applied effort. In any field of science and engineering, its imperative to understand that the effort should not be towards generating copious amounts of graphical plots, animations and results. The goal should be to divert attention and resources towards solving questions that appear most relevant in context of the Big Picture.

- **Erin Shellman interview**

  Excerpt: *“I typically start from the finish line. Assume that you’ve built the thing you’re considering, then ask “so what?”*

  Reflection: A lot of times in the pursuit of developing a product, the team gets too deep into the nitty gritty details of the product. From design to managing the logistics of manufacturing, its easy to forget the end goal. It may be prudent to ascertain way early in the journey: *why to even take up the journey?* Chalking out the minute details and timelines of the project in advance is good but having a strong understanding of the Big Picture is as essential in ensuring success as is any other element of project planning.

Excerpt: *“Plotting data in scatterplot matrices with R is really helpful because you can quickly start discovering relationships in your data without much work on your end”*

Reflection: This thought and the subsequent comments of Erin in the said interview resonated with me strongly. I feel she put it very well. I work in the field of computational automotive engineering and a lot of times I have saved hours of analysis by just looking at the data closely. Firing large CPU runs is important in lots of case but so is to employ basic engineering judgment in many other cases. One does not always need to use a Bazooka to kill a fly. In the same way, and as I have learned, if basic data analysis can give an answer to an obvious question, its not wise to fire overnight runs and block CPUs.

—----------

- **Jake Porway interview**

Excerpt: “I think an interesting thing that people may not think about when just getting into data science is that you always need to question your assumptions.”

Reflection: Jake emphasizes that a data scientist must revisit and question the sanctity of the assumptions their model is built on. This reminds me of a very interesting event of World War II. In WW2 a group of engineers were working on improving the armor strength of US fighter planes. Based on the bullet patterns they saw on the planes, they ascertained that the fuselage needs to be armored more. However, Abraham Wald who was a statistician in the team suggested that more than fuselage, its the engines that need to be armored. He argued that the planes hit on the engine never returned back to be part of the data pool.

Abraham Wald challenged the assumption that was drawn from the data. Its interesting to note that the data was correct, inference drawn from the data was correct and the assumptions built on the data were also correct. But, Abraham Wald’s understanding of the data helped him in identifying the flaws in the assumption and course correct the project effort. To sum it up: once model is as accurate as the assumptions it was built on.

—----------

Excerpt: “Data is new eyes.”

Reflection: The writer is putting forth a very interesting argument that Data Science allows one the ability to look at the world in a way that was never looked before. In this world, which is seemingly run by free will and absolute randomness, Data science can show patterns and trends that were never seen before.

—----------

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**

   Response: I feel that I have this ability but I am sure that this will only be strengthened manyfolds by the time I have completed my Masters in Applied Data Science. I have been in automotive engineering for over 7 years. I worked from design to the engineering analysis of automotive components which gave me a holistic understanding of automotive manufacturing from drawing board to the finished part. Due to the variety of work experience in this domain, I have a strong understanding of the milestones in the product development process where things pickup pace and also where the things slowdown.

   As an engineer I try to quantify the product cycle progress and explore the variables that directly impact the rate of progress. I even try to quantify the qualitative aspects of the Product Cycle such as work efficiency of designers and engineers and attempt to identify the roadblocks impeding the progress. I like to develop data and I strongly believe that sometimes the best questions can be posed only after spending time with the non-sensical mountain of data. Embracing the chaos allows one the creativity to pose questions and find methods to answer the questions to make some sense of the chaos.

2. **a repertoire of problem types**

   Response: I think that I have a very limited number of Problem Types to formulate real world problems into data science problems. I need to expand this repertoire of problem types by taking courses related to collection, analysis and post processing of data. With that said, I have strong familiarity with two problem formulation types: dimension-reduction and multi-variate regression. I often use Singular Value Decomposition to denoise and make sense of sensor data from vehicle testing. I also use regression methods to develop hyperspace with given variables to setup optimization models later.

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

Response: I need to expand my repertoire of problem formulation types but I feel that I have decent grasp over how to formulate many real world problems into the two problem formulation types that I am aware of. I use regression modeling to map discrete data points in multiple domains using polynomial and non-polynomial regression models. A lot of prediction problems in real world can be addressed by this technique. I also use dimension reduction techniques to identify and remove the non essential variables from the data which allows me to reduce the problem size. Again, a lot of chaos in data starts making sense when non essential variables are subtracted from data. I look forward to expand my ability to map real world problems into one of the many problem formulation types that exists in Data Science.

Application in Domain of Interest

Domain: Automotive Engineering

Project 1 Description:

Name: Segregation of CAD parts into ‘Structural’ and ‘Non-Structural’ domains.

Premise: In automotive computational analysis, design engineers provide CAD models to computational engineers to carry out structural analysis. The computational engineers spend hours to days segregating the structural parts from the CAD. Parts such as rubber, pipes, plastics etc. are deleted and the metallic frames and other structural components are retained. This is a manual process and most automated filtering techniques to segregate the parts are often found to be impractical in real use.

Problem Description in Real Life: The objective is to develop a robust method, tool or protocol that reduces the manual effort and time needed to segregate structural parts from non-structural parts.

Project 1 Problem Type (with explanation):

Enquiry into use of Data Science: There are volumes of pre-seggregated and post-seggregated CAD models that can be used to identify patterns in how engineers deem a part structural or non-structural. Thus the solution can be date driven.

Enquiry into Classification of Problem: Supervised Learning.

Type of Data Required: Database documenting the physical attributes such as density, volume, material, surface-area etc. of each CAD part along with how that part was earlier manually identified: ‘Structural’ or ‘NonStructural’. This type of data is usually readily available or easy to create.

Problem Description in Data Science: The Supervised Learning algorithm would be first trained on the input data. New CAD part information would then be fed into the system to identify the accuracy of the system to correctly identify if the part is structural or not.

Objective of Data Science Model: The trained Supervised Learning system would then be packaged as a desktop program. The intention is that it will facilitate the quick segregation of CAD parts and reduce human effort and the involved financial resources.

Project 2 Description:

Name: Blackbox tool to predict strain value.

Premise: Fatigue life assessment of automotive components are done by performing calculations on data obtained from multiple DAQ sensors mounted on those components. One such data is strain. There are over 100 documented variables that impact the strains developed at a given point such as temperature, density of material, acceleration of vehicle etc.

Problem Description in Real Life: The process of setting up the data acquisition is costly. Its not viable to rerun the experiment to study the impact of one or few variables on strain value. The intent is to develop a method, tool or protocol to approximate strains that may develop for given variables.

Project 2 Problem Type (with explanation):

Enquiry into use of Data Science: There are volumes of sensor data available that can be employed to develop a tool to predict strain values. Regression analysis appears to be a viable approach. However, a simple polynomial regression method is not practical.

Classification of Problem: Dimensionality Reduction followed by Regression.

Type of Data Required: Values of Strain with corresponding scalar values of variables

Problem Description in Data Science: Dimensionality reduction methods will be used to first identify the variables that are the primary drivers of the strain. Variables that are not significant will be removed. Non polynomial regression model will be setup that takes a vector of variables as input and outputs one scalar value for the strain.

Objective of Data Science Model: The end goal is to develop a black-box that engineers can use to obtain strains corresponding to arbitrary set of input variables.

Questions, Maxims, and Commitments

Question (I will always ask…)
How accurate is the training data?

1-Sentence Project Description: Segregation of CAD parts into ‘Structural’ and ‘Non-Structural’ domains.

Meaning in Context
The database provided with the CAD parameters and labels were developed manually. If this database is to be used to train the Supervised Learning system, it is essential to ensure the correctness of the data. It must be ensured that the labels and parameters for each line item in the database is correct. If not so, then a rough estimate of the incorrectness of data must be documented.

Importance for this stage of the project
The accuracy of the Supervised Learning system will be dependent on the training data. Historically, the segregation of structural parts from non-structural parts in the CAD assembly is done manually. Sometimes, engineers label parts incorrectly. This incorrectness of data in the database is hard to identify. Having a rough estimate in percentage about the incorrectness of the data allows the data scientists to predict the accuracy of there Supervised Learning Prediction tool better.

Maxim (I will always say…)
‘One does not need to be correct to be right’

Which Project
Blackbox tool to predict strain value.

Meaning in Context
In the said project, the goal is to develop a predictive tool for strain values using only the available data. Depending on the variable sensitivity, in engineering calculations varied levels of approximations are acceptable in values of variables. The data scientist should be aware of the kind of accuracy that is being expected from the predictive tool.

Importance for this stage of the project
Its essential for a data scientist to ascertain with the potential end user about the accuracy that is being seeked. A more accurate prediction needs higher quality datapool. Also, if a prediction of 2.5 works as good as 2.555 then one must strive for 2.5. This will reduce cost of development and cost of data.

Professional/Ethical commitment (I will always/never...)
I will always include the Subject Matter Experts in all stages of development

Which Project
Segregation of CAD parts into ‘Structural’ and ‘Non-Structural’ domains.

Meaning in Context
The said project needs a lot of feedback at every stage of the development. Instead of separating them from the development process, it will be prudent to share the progress and preliminary results of the project as it unfolds. Course correcting is important and should be driven by experts on the data.

Importance for this stage of the project
Keeping the SMEs in loop at all stages of this project will ensure that data parsing, cleaning, modeling, analysis and finally integration into action is driven by the experience of the SMEs. This ensures that the final product is in conjunction with the needs of the end users.

Week 2: Data Collection and Cleaning Stage

Potential Personal Project Tweet

I noted my everyday mood for 6 months → I also recorded my hours of sleep for those days → Crunched the numbers and asked → More Sleep More Happy 🤯? → Yes! → **Link**

Number of Characters: 175

Reading Responses

- **Law of Small Numbers**

  Excerpt: *“Naturally, you focus on the story rather than on the reliability of the results.”*

  Reflection: The writer alludes to a very common occurrence in data evaluation where in the observer of the data is more focused on the inference of the data than the means, methods and assumptions on the basis of which the data was collected. It is to be understood that a very correct inference can be drawn from a data set which is completely incorrect. How the data is obtained and under what assumptions it is obtained are as important questions as what the data is trying to convey.

  Excerpt: *“..far too willing to reject the belief that much of what we see in life is random.”*

  Reflection: Just before this excerpt, the writer alludes to the “*hot hand*” fallacy in which one tends to believe that there is a pattern in chaos and what is happening will continue to happen. This was also well explained by Prof. Thaler and Selena Gomez in the 2015 movie, *The Big Short.* The writer also reflects that this behaviour may be driven by our hunter gatherer brain which wants to identify and predict natural occurrences. Some stock traders after generating huge profits in quick succession start believing that they have identified the *pattern* in which the stock market works. These stock traders learn the “*hot hand*” fallacy by practice. I was that stock trader. From my experience and through this reading, I infer that in data science and life, accepting that most events are naturally random and tend towards more chaos may do more good than harm.

- ***Statistical Biases Types Explained***

  Excerpt: “*..95% “no” and 5% “yes” answers, what does it mean? Exactly nothing*.”

  Reflection: The writer discusses about the phenomenon of *Selection Bias* wherein how incorrect, biased and skewed selection of data leads to incorrect, biased and skewed conclusions and have no truth or value. I have noticed selection bias in datasets where certain data-points are omitted by ‘experience’ engineers suggesting that it is an outlier. There are several checks and balances that have come recently to ensure acquisition and cleaning of data is done with as much effort and prudence as the subsequent engineering or analysis using that data.

Excerpt: "*.But remind yourself all the time that only success stories are published!."*

Reflection: The writer discusses about the *Survivorship Bias* wherein the data that never made it to the dataset is completely ignored during the analysis. The method and assumptions that dictate what data will be omited from the dataset needs to be carefully examined. Its essential to identify potential points of biases that may lead to non inclusion of certain kinds of data in the study. In World War II, Abraham Wald suggested that the American fighter planes must be provided with more armor near the engine area than the fuselage. Even though the documented data showed more bullet holes on the fuselage than the engine, Abraham Wald recognized that the planes which were hit heavily on the engine never returned back to be part of the database.

- ***Data Cleaning 101***

Excerpt: "However, some of the corrections may not be obvious."

Reflection: The writer suggests that in the data pool one must correct data only when enough information is available. There are times when it is obvious that a certain data element should be **this** than **that**. But, many a times, ruling **that** over **this** is not as obvious. Cleaning the data to update, fix or remove labels and values should have a thorough and documented method. It should also be flexible and allow case by case decision seeming *in-correct* data elements.

Excerpt: "..*you are entitled to clear information"*

Reflection: I have noticed this unfortunate practice where engineers believe that questions that can be answered by inferences drawn from other previously answered questions should not be put forth. It may be just my experience which is drawn from my line of work but I think that this is more common than we would want to admit. I have seen bosses and graduate advisors getting upset with employees and students asking questions that have seemingly *obvious* answers. If this is a norm, it must change. One must not be afraid to ask seemingly *stupid* questions. A data scientist should feel free to ask all the questions that they deem relevant to there study. They are entitled to it and they are also responsible to ensure that relevant information is seeked.

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

  Excerpt: "..data scientists often fail to do is set the seed values for their analysis."

  Reflection: The writer after the said excerpt emphasizes the importance of using *seed values* in studies involving some kind of stochastic process. For analysis runs which include random value generation, it is essential to create these values that may allow one to generate same or similar distribution of random numbers to test if the results are reproducible.

  Excerpt: "And words are imprecise*"*

  Reflection: With no data to back the claim but with full confidence along with abundant humility I would argue that the *Conclusions* section of a research or white paper is the most read part of the paper. Also the **rudest,** as argued ahead.

  I think that Andrew Tait would agree with the line in **Law of Small Numbers** – *Naturally, you focus on the story rather than on the reliability of the results.* Andrew Tait suggests, as I infer, that it is imperative that one while presenting the conclusions should also point to the results or body of work that supports that conclusion. He also suggests that it should be possible to trace the conclusions all the way back to the raw data. This generates confidence in the results and allows for its reproducibility.

  If one lives in a utopian world far far away from earth where raw data is pristine and void of all biases, one can argue that the results obtained from that data is the purest form of truth. **However,** even from the purest form of truth various conclusions that are drawn can be absolutely untrue. Conclusions, deductions and inferences are influenced by the inherent biases that exists in individuals. Results may be true but conclusions may not necessarily be. I also humbly argue that the *Conclusions* section of the paper is the rudest part of the paper because the author tries to advocate there inferences drawn from results based on there own beliefs and biases which may not be true for others. By pointing to results and raw-data in the *Conclusions* section, the author allows the reader to draw there own conclusions. This is a good practice for reproducibility. It is also kind.

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

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

   Comment on Capability: I already have this capability.

   Acquirement of Capability: I have worked on various kinds of data sets in the automotive industry such as DAQ readings obtained from sensors in the vehicle frame, strain data obtained from transistors on a crash dummy etc. One of the major problems with such data is to gauge the sensitivity of the sensors which record the data. There are set protocols to calibrate the data to ensure that measurement errors are factored in for. Removing the outliers from the data is fairly easier but understanding that the entire data may be an outlier needs experience and understanding of similar data from other experiments. My work experience has helped me come up with my own laundry list of methods, in conjunction with industry standards, to identify problems in data before any analysis is done.

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

   Comment on Capability: I already have this capability.

   Acquirement of Capability: In automotive industry, as is in other fields, experimental data for newer variants of vehicle is not available in the public domain for proprietary reasons. However, there are volumes of data related to strength, fatigue and crash performance of relatively older vehicle variants available via government and private institutions like NHTSA, IIHS etc. Besides these, several countries also publish similar data for there own vehicles on there websites. I have a good understanding of the places where most of the data in my application domain can be obtained from. I acquired this skill by my work experience and self interest.

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

   Comment on Capability: I look forward to strengthening this capability.

   Acquirement of Capability: In automotive field, most of the data is available in ASCII format which is directly obtained from sensors and instruments. Each instrument has its own output format. Currently, there is no standard for database architecture to curate this data for uniformity. am not well versed in how to most effectively document and curate data sets and plan on taking SIADS532 to learn best practices in data characterization and SIADS611 to get acquainted with database architecture.

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

   Comment on Capability: I already have this capability.

   Acquirement of Capability: I have acquired this capability on my own. I felt that datasets in my field are difficult to acquire and assemble using basic desktop spreadsheet softwares. I learned javascript and python to be able to clean and assemble data. I learned how to use *Puppeteer,* a library to control chrome, for data-mining. These tools and methods allow me to obtain, curate, clean and assemble data before any analysis can begin.

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

   Comment on Capability: I **partially** have this capability but needs improvement.

   Acquirement of Capability: I learned dimension reduction methods using PCA and SVD in my previous graduate research work. I also use it in my field. I know basic techniques to identify the primary variables in a multivariate system that drive the output. Conversely, I know how to detect and remove outliers. However, for pure data science, my skills are insufficient. I plan on taking SIADS-505 (Data Manipulation), SIADS-515 (Efficient Data Processing) and SIADS-516 (Scalable Data Processing) to further my knowledge and skill set.

Maxims, Questions, and Commitments

Projects: Segregation of CAD into ‘Structural’ and ‘Non-Structural’ parts
Projects: Blackbox tool to predict Strain Value.

Question (I will always ask…)
How diverse is the data pool?

Which Project
Segregation of CAD into ‘Structural’ and ‘Non-Structural’ parts

Meaning in Context
In automotive engineering, there is a wide array of CAD parts in a wide array of sub-assemblies such as suspension-system, cradles, chassis, powertrains, etc. It's essential that before the ML tool is trained, it is ascertained that CAD from a diverse pool of sub-assemblies is included.

Importance for this stage of the project
A large and also diverse pool of CAD items will help build a robust ML algorithm for predictive purposes. This will ensure higher confidence in the predictions made by the algorithm.

—----------

Maxim (I will always say…)
Cleaner the data, Clearer the Predictions

Which Project
Segregation of CAD into ‘Structural’ and ‘Non-Structural’ parts

Meaning in Context
It is important that each of the CAD items in the data pool are properly labeled and a fixed nomenclature is used to label them which is standardized across all automotive subassemblies CAD parts in the training data pool. Duplicate entries should be removed, ambiguous data points must be deleted and proper organization of data is done before training is commenced.

Importance for this stage of the project
An appropriate amount of effort should be put into cleaning and preprocessing the data. It is important that the training data is of high quality. High-quality data is unambiguous, well-labeled, and organized. Doing so ensures the robustness of the ML algorithm built later.

—----------

Professional/Ethical commitment (I will always/never...)
I will always keep a record of all the changes to the prediction and validation data set.

Which Project
Blackbox tool to predict Strain Value.

Meaning in Context
During the course of cleaning and preprocessing data, the data analyst adds, changes, or removes data elements from the data pool. Updates to the data have an impact on the output of the ML algorithm developed using this data. The data analyst should ensure that they document all the updates done to the data set during the course of the project.

Importance for this stage of the project
It's an ethical exercise to ensure that the data analyst records all the updates to the data set along with a justification of why the data was modified, renamed, or deleted. Journaling the modifications allows other data scientists and engineers to understand the nitty-gritty details of the input dataset. It also helps in the reproducibility of the exercise.

Week 3: Data Analysis and Modeling Stage

Informational Interview - Reflection

Note: In my week-1 assignment I mentioned that I wanted to take an interview with Ye Yan as his line of work is very similar to mine. Unfortunately, I was not able to get that interview. I reviewed the interview of Anna Smith from *Data Scientists at Work* and gained some interesting insights which are put forth.

About Anna Smith: She is an analytics engineer at *Rent the Runway,* an online/offline fashion company that provides high-end fashion dresses for rent. It's a NY-based company that manages and tracks over 50k inventory items. The goal of the interview was to learn the role of Anna Smith and how she employs data science in the company's logistics success.

Pertaining to: Question - W2 Data Collection and Cleaning

Anna puts forth that the first step pertaining to data collection and modeling is to employ basic summary statistics to understand what the data tries to convey. Once a cursory understanding is obtained, more statistical analysis is carried out to ascertain the statistical confidence that can be obtained from the data. Questions like *How much data do we need?* are floated. This helps the team figure out the number of logistical links that need to be generated to achieve the desired goals

Pertaining to: Maxim - W1 Problem Formulation

Anna begins by suggesting that in developing a model to accurately gauge a prospective client's *taste,* several variables such as shape, fit, style of fabric, color, and individual sensibilities have to be taken into account. She then comments: *"Oh, it's because they're pink and flirty-looking with ruffles."* and alludes to certain *latent variables* that represent someone's taste in dress. These *latent variables* are not explicitly known. In problem formulation, understanding input variables are essential and Anna emphasizes the importance of identifying these variables correctly.

Pertaining to: Commitment - W3 Analysis and Modeling

Anna talks about democratizing fashion and making it more accessible. She talks about the big picture and how her line of work in conjunction with data science is empowering to women. It is clear that Anna Smith feels committed to the cause of providing confidence to women not just by attire but also by the virtue of the desire to look good and have a great night!

Additional Questions I would want to ask?

- What parameters do you deem important to consider in the data collection stage?
- During data modeling how do you decide the best formulation technique?
- What is the future of Data Science in soft sciences like Fashion and Culinary arts?

Reading Responses

Instructions (Delete these in your submission)

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

  Excerpt: "*A well-functioning ML algorithm will separate the signal from the noise."*

  Reflection: A robust ML algorithm should work on only the data and not noise. An ML algorithm trained with noise leads to overfitting and would not be able to predict accurate results for a new set of input data different from the training data. It's imperative, that the ML algorithm segregates data and removes the noise before any training begins.

  Excerpt: *"Up until a certain number of iterations, new iterations improve the model."*

  Reflection: It is argued that *Error* vs *Number of iterations* is similar to a bell curve for the validation set. There is an optimal number of iterations up to the ML algorithm that should be trained and beyond that point, training the algorithm with the validation set yields in overfitting.

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

  Excerpt: Fortunately, much statistical research has been devoted to this problem, and "group sequential designs"

  Reflection: There is usually a 5% limit for alpha in significance testing for single comparison between groups (such as Student-T-Test). However when testing statistical difference among multiple groups, the probability of finding a difference just by chance increases manifolds. In such case, the Type-1 error with 5% limit would be an incorrect basis to gauge the difference between the groups.

  Excerpt: Worse still, if there was an overall statistically significant benefit (P <= 0.05)

  Reflection: If statistical comparison between 8 groups is performed, there would be 33% chance of observing a statistical significance with P<0.05. The writer points to the sensitivity in the significance testing when multiple subsets are present.

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

  Excerpt: *"...looking to test a hypothesis, but is 'letting the data speak'"*

  Reflection: The writer argues that one of the issues P-Hacking is that it's difficult to trust the predicted strength of a relationship based on just the reported data. A different set of data of the same class may or may not show similar strength of relationship. In Data Science related literature, many a time, researchers selectively present results which show statistical significance whilst holding back on all the data which suggested otherwise.

  Excerpt: *"under-powered studies have a higher chance of showing a substantially inflated effect size"*

  Reflection: The writer explains how in the presence of Type-I error in conjunction with low sample size, the calculated coefficients of correlation can be *substantially* higher than the actuality. It's important to identify if Type-I error exists and use a larger sample size for regression analysis to mitigate the issue of over-inflated correlation coefficients for underpowered studies.

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

  Excerpt: *"*The common problem in these articles is that they take two correlated trends and present it as one phenomenon causing the other.*"*

  Reflection: The writer points to a very common error of judgment where two independent trends with similar trajectories are considered correlated. He prefaces this thought with examples wherein he challenges the idea that taking music classes in school boosts academic performance and eating fish leads to violence. The writer argues that to establish a correlation, randomized studies are to be performed along with rigorous statistical analysis.

  Excerpt: *"..tendency marketers and companies take advantage of with regularity"*

Reflection: It is put forth that false correlations are often shown as causations in the marketing industry. It can be argued that this is not by lack of knowledge on the part of these companies but is done by choice. Without randomized controlled trials, causation cannot be suggested. The human tendency to draw conclusions and correlations based on trends is exploited by companies for commercial purposes.

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

Excerpt: different years (see Table 2). Between 1974 and 1978, the tax rate decreased in each income category, yet the overall tax rate increased from 14.1 percent to 15.2

Reflection: This excerpt is followed by tabulated tax rates in 1974 and 1978. It was shown that though the tax rates in each category decreased in 1978 compared to 1974, the overall tax rate still increased. This was owing to the increased weight factor due to higher inflation which is not that easily clear by cursory review.

Excerpt: Virginia, during the year 1910. Although the overall tuberculosis mortality rate was lower in New York, the opposite was observed when the data were separated

Reflection: Simpson's paradox usually shows up in the context of groups that involves multiple subgroups. The above excerpt is one such example where the paradox is witnessed. In the study of Tuberculosis related deaths in 1910, NY showed relatively lower mortality rates. However, when the study was segregated into subgroups based on race and other markers, it was learned that the inference drawn for the overall group does not hold true. This is a classic example of Simpson's paradox.

- ***Conditioning on a collider***

Excerpt: *Even though those two things are possibly positively correlated in the overall population, the correlation in my friend's sample is negative (X1 and X2 are negatively correlated conditional on Y).*

Reflection: Unlike mathematical *sets,* associativity does not hold in real life by default. If variable A is positively correlated to variable B and variable A is positively correlated to variable C, it does not mean that variable B and variable C are positively correlated or

even correlated at all.

—----------

Excerpt: *people are very willing to speculate about confounding variables, so why not speculate a collider for a change?*

Reflection: When there is a shared common basis between exposure and outcome, confounding variables occur. On the other hand, collider bias occurs when the outcome and the respective exposure have an effect on the third common variable. The writer argues that it's a natural tendency of people to assume confounding variables. They should develop skepticism and assume the occurrence of collider bias instead on some occasions.

—----------

Plan for Knowledge Acquisition

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

- common mistakes in data analysis that lead to misleading results

  Response: I have this capability

  Reflection: Setting up the model with wrong assumptions and pre-suppositions can lead to non-practical results in the end. The end goal is not to write codes but to develop a solution. If the assumptions in the modeling are wrong, then the solution building takes a back seat.

  Another common mistake in data modeling that leads to misleading results is the lack of statistical pre-processing of input. Sometimes the input data is either insufficient or spurious to develop a trustworthy model. A data analyst should have a toolset of statistical analysis tools such as significance testing, correlation testing, etc. They should employ these tools on the data set before any modeling is performed. I would attempt to expand my toolset of statistical analysis during my master's in applied data science.

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

  Response: I have the capability to interpret results but I need to improve my capability to validate results.

  Reflection: I want to expand my knowledge of statistical analysis tools that will allow performing tests on my input and validation data set. This will ensure that input data is

good to be used for any data science analysis. I am planning to enroll in SIADS-524 (Presenting Uncertainty) to expand my capabilities in stochastic analysis of data.

Maxims, Questions, and Commitments

Projects: Segregation of CAD into ‘Structural’ and ‘Non-Structural’ parts
Projects: Blackbox tool to predict Strain Value.

Question (I will always ask…)
Are the assumptions for the modeling correct?

Which Project:
Segregation of CAD into ‘Structural’ and ‘Non-Structural’ parts.

Meaning in Context
To create a classification model that identifies if a given component is ‘Structural’ or ‘Non-Structural’, several pre-processing steps need to be performed on the input CAD. This is a critical step and **how** the pre-processing/cleaning is performed needs to be defined accurately. The assumptions in the algorithm have to be driven by Subject Matter Experts. Before the analysis is carried out, all the assumptions and pre-suppositions involved must be presented and confirmed by the experts.

Importance for this stage of the project
The assumptions made during the modeling stage dictate the quality of the results that are obtained. By ensuring that the assumptions are accurate, the analyst can focus on the nitty-gritty details of the code and model with confidence. If assumptions are incorrect, the results will be unuseful for practical purposes.

Maxim (I will always say…)
Confidence in the model leads to confidence in the results

Which Project
Blackbox tool to predict Strain Value.

Meaning in Context
A data science tool for predictive purposes has several moving parts. The modeling assumptions and the quality of input data drive the quality of the predictive tool. Data analysis must spend most of the time on viewing, reviewing, questioning, and then again reviewing the assumptions and modeling details of the model. This gives confidence to the data analysis to stand behind there results.

Importance for this stage of the project
Data modeling is an iterative process. It takes several stages, inputs, and updates to get the most optimal model possible. All the models that are developed in the later stages are more or less dependent on the preceding model and modeling method. It's essential to spend enough time and resources on the initial modeling setup as if this is done right, the subsequent iterative stages would be more productive.

Professional/Ethical commitment (I will always/never...)
I will never cherry-pick validation data that show the best results for my predictive model

Which Project
Blackbox tool to predict Strain Value.

Meaning in Context
I will present the predictive capability of my model with random sets of validation data. I will present all the results and not just the results for which my model showed good results

Importance for this stage of the project
This ethical commitment will ensure my integrity as a data scientist which is as important for the project as the nitty-gritty details of my code and model. The end user should be able to trust my tool but also me as a data scientist.

Week 4: Presenting and Integrating into Action

Sources for Data Science News

I would like to use the following resources to expand my knowledge in Data Science”

- **Databricks**: It's an online portal that uploads Data Science case studies on a regular basis. Case studies also include links to raw data and other references which were used to carry out the study.

- **Knime Blog:** It's a repository of Data Science projects and visualization techniques pertaining to the automotive industry. They publish blogs on a weekly basis where data engineering is used in conjunction with FEA and other methods to solve real-world problems in the automotive industry.

- **Rapidminer White Papers:** Rapidminer is a company specializing in automotive data science engineering. Although it's not a blog site, they publish some very interesting white papers from time to time presenting cutting-edge projects in the automotive industry where data science was used to solve a given problem.

- **Two Minute Paper:** It's a youtube page where the creator presents the findings of the latest research in data science, ML, and deep learning in easy-to-understand ways. Excellent visualization methods and easy-to-understand presentation style make it possible to understand the entire paper within the duration of the video.

Reading Responses

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

  Excerpt: *The Hungarian physician stumbled in one essential area—the communication of his data.*

  Reflection: Although Semmelweis's data was truthful, valuable and actionable, it was not communicated well. If it was communicated in a scientific manner, there is a higher chance in it being accepted.

  —----------

  Excerpt: *He forgot what it was like to not know what he knew.*

  Reflection: The writer alludes to the concept of *curse of knowledge.* Semmelweis was not successful in building a common ground with the medical community. His impatience to prove himself correct alienated him from the community.

  —----------

- ***Storytelling for Data Scientists***

  Excerpt: *we draw generalizations from this single story instead of taking a step back and evaluating the data.*

  .

  Reflection: Jan Zawadski argues that humans are not rational beings and tend to think and reason emotionally. He then suggests frameworks of how this attribute can be used in presenting data. The goal is to present the results and inference in a form that captures the sensibilities of humans just how stories do.

  —----------

  Excerpt: *They all, to some degree, use elements of the SUCCESs model.*

  Reflection: The author presents the concept of SUCCESs for presenting ideas that sticks with the audience. The presentation should be **S**imple and easy to understand. It should have elements of surprise which comes up **U**nexpectedly. Results should be presented with **C**oncrete set of examples that the reader can relate to strongly. Data presented should be **C**redible such that the reader can rely on it and trust it. As humans are emotional beings, data should appeal to the **E**motions of the readers. And, finally data, problem-statement and results should all be woven together into a **S**tory.

  —----------

- ***Interpretability is crucial for trusting AI and machine learning***

  Excerpt: *It's hard to figure out what they learned from those data sets and which of those data points have more influence on the outcome than the others.*

  Reflection: The extraordinary predictive abilities of ML algorithm comes with a caveat: they are difficult to understand. Its hard to draw clear explanations regarding how and why the algorithm forms relationships between the variables and how the data points lead to the predictions. The ML algorithm is more so like a black box whose internal functioning is hard to interpret.

  —----------

  Excerpt: *While highly engineered features can boost the accuracy of your model, they will not be interpretable when you put the model to use.*

  .

  Reflection: The writer argues that its essential to develop a good understanding of the dataset which includes the characterisation and summarizing of the most important features of the dataset. To be able to best explain the input-output relationship one must begin with understanding the most **meaningful** features among the dataset. Highly engineered variables/features are hard to interpret even though they may lead to highly accurate predictions.

  —----------

- ***The Signal and the Noise, Chapter 2***

  Excerpt: *But our brains translate it into something more subjective.*

  .

  Reflection: The writer comments on the psychological evidence which suggests that humans tend to perceive estimates subjectively. A 90% probability of an event occurring does not automatically imply that there is 90% estimate of the event occurring but if the event is repeated enough times, there will be a 90% chance of that event occurring among all the events. The writer refers to the works of Amos Tversky in his argument and argues that we have trouble interpreting probabilities.

  —----------

Excerpt: *When the outcome is more predictable—as a general election is in the late stages of the race—the forecasts will normally be more stable.*

.

Reflection: The writer talks about the misconception that predictions are constants and should not change. Forecasts are subject to change as the data changes. In elections, the forecasts done close to the election times are the most accurate as those predictions are done based on the final set of voter polls. However, forecasts that are months older are less trustworthy. Its not due to the bad quality of the study but because people's decisions and interests swing heavily as the voting day approaches and so data from months before is not the most accurate representation of what the voter feels on the election day.

—----------

- ***The Signal and the Noise, Chapter 6***

Excerpt: *The forecasters later told researchers that they were afraid the public might lose confidence in the forecast if they had conveyed any uncertainty in the outlook.*

Reflection: The writer talks about the Grand Forks flood where the forecasters avoided communicating to the public regarding the uncertainty that existed in there estimations.This was done to avoid losing confidence among the public. I think this is a major ethical issue and by not abundantly making the estimations clear, the forecasters did not allow the city to be well prepared in advance for the upcoming floods. In data science, it is essential that estimations are prested along with uncertainties that exist in the estimations.

—----------

Excerpt: *…indicate a range of outcomes for where they see the economy headed.*

Reflection: The writer emphasizes that an essential part of scientific forecasting is to abundantly put forth the probabilistic considerationstions of the outcomes. Instead of putting forth a singe digit number estimating the probability, a sounder approach is to put a range of numbers and then presenting the respective probabilities associated with those estimated numbers.

—----------

- ***How Not to Be Misled by the Jobs Report***

  Excerpt: *Human Beings are unfortunately bad at perceiving randomness.*

  Reflection: One game performance of a team does not signify the performance of the team in coming weeks and stock performance of one week cannot predict the performance of the stock in coming months. As the writer suggests, humans are bad at perceiving randomness. Uncertainty and randomness are as important as the data itself. Accurate prediction requires understanding of the randomness.

  —----------

  Excerpt: *No one report can neatly summarize the health of a $17 trillion economy…*

  Reflection: The writer suggests that for data which is very large and diverse cannot be best predicted by a small subset of that data. Without a rigorous evaluation, very little can be discerned from studies with small data size.

—----------

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

  Excerpt: *Thankfully, as soon as you “get the grip” of what this fuzz is actually about …*

  Reflection: Best software craftsmanship does not reply on frameworks, APIs and the nitty gritty details of the code. It relies on the core idea of the problem solving approach. Once the programmer *gets a grip* of this they realize that most APIs and frameworks do not differ much. They are just means to an end. The means being the framework and the end being the solution.

  —----------

  Excerpt: *Most importantly, people want you to gain new knowledge very quickly.*

  Reflection: The writer puts forth that in ML and software development there are a lof concepts that one must be aware of. However more than knowing a lot of concepts its more important to gain new knowledge very quickly. One of the ways to do it is by changing and breaking other people’s code and understanding the outputs and how those outputs were arrived at.

—----------

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

  Excerpt: *If you can recreate the sequence of events leading up to that decision, you can learn from it..*

  Reflection: The writer explains how AirBnb characterizes data in a human light and not just as metrics. Its considered as *Voice of Customers*. The writer follows it up by arguing that by studying the sequence of events that lead to a certain decision, data scientists can understand the sensibilities, the likes and dislikes of the individual.

  —----------

  Excerpt: *when data scientists are pressed for time, they have a tendency to toss the results of an analysis "over the wall"*

  Reflection: During data science projects, sometimes engineers get insights that help them understand the data in a way that was never done before. However when pressed with deadlines and the push to finish the project fast, these insights are often ignored, forgotten or never returned back to, to be acted upon. The writer argues that decision makers must understand the ramifications of of an insight not being acted upon.

  —----------

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**

  Response: I look forward to strengthening this expertise

  Reflection: To efficiently present data science results in a way which is clear to a layman or someone not familiar with the nitty gritty details of data engineering needs two kinds of skills: interpersonal communication skills and data science communication skills. The former is especially helpful in an active presentation format. The latter is important for both active and passive forms of communication such as white papers.

  To develop my data science communication skills I plan on taking *SIADS 523 - Communicating Data Science Results, SIADS 524 - Presenting Uncertainty and SIADS 522 - Information Visualization.* Besides MADS courses, I also plan on participating in poster and presentation competitions that will give me exposure to how other data scientists present there results.

- **how to work with software engineers to put models into production**

  Response: I have this capability

  Reflection: I work as a Senior Structural Engineer at Magna International, an automotive company. One of my responsibilities is optimizing the workflow of my automotive engineering team by identifying processes that can be automated. Once such a process is identified, I reach out to Software Engineering team and layout a detailed problem statement. I then work with the Software team to come up with proof of concept models. Through an iterative process, we develop a solution which is then released to my engineering team to test as a beta program. Based on the feedback, the software engineers update the code and final versions are released for use. I have worked on multiple projects where the above process(s) was followed. One of the key elements in this exercise is to understand the end goal and being clear in presenting the end goal along with the *Big Picture* to the software team.

Maxims, Questions, and Commitments

Projects: Segregation of CAD into ‘Structural’ and ‘Non-Structural’ parts
Projects: Blackbox tool to predict Strain Value.

Question (I will always ask…)
Which graphs/images in the report can be done away with?

Which Project
Blackbox tool to predict Strain Value.

Meaning in Context
I have noticed in my current work and previous graduate research projects that people like to add more than needed amounts of graphs/images and verbalogy to the report or white paper. This is done with the belief that by adding more graphs/images, the quality of the paper/report is enhanced. On the contrary, if a graph or image is not adding additional and also significant new information to the paper, it should be removed. The goal is to increase the interpretability of the paper. Adding lots of graphs/images/verbalogy sometimes adds more confusion than clarity.

Importance for this stage of the project
Presentation of the results is the last stage of the project. A report boils down all the effort, hard work and brainstorming that was involved in all the previous stages of the project. Thus it becomes essential that effort is put to ensure that the final set of results is presented in the most efficient way possible. Efficient presentation requires that the least amount of words, the least amount of graphs, and the least amount of images put forth the maximum results.

In the Strain Value prediction tool project, we may need to present several graphs to show how well the predictions match the expected value. But in doing so we will ensure that redundant or repeated information is not conveyed via graphs.

Maxim (I will always say…)
Easiest to understand explanation is often the correct explanation

Which Project
Blackbox tool to predict Strain Value.

Meaning in Context
There are various assumptions needed in this project. Different data scientists can formulate the same problem in different ways and thus use different methods to solve the same. However, for each method, the results should be presented with the starting aim in mind. The results should be presented whilst keeping the question in mind “does this present the solution in the simplest way possible?”

Importance for this stage of the project
The maxim alludes to the idea that the results should be presented in a way that is easy to understand. The report should not be filled with a convoluted maze of verbalogy that a layman cannot grasp. By ensuring easy interpretability of the results, the data scientist converts the data science results into a real-world solution to the starting problem.

Professional/Ethical commitment (I will always/never...)
I will never selectively present results to match a certain end inference

Which Project
Segregation of CAD into ‘Structural’ and ‘Non-Structural’ parts

Meaning in Context
In this project, the end data science tool should be able to efficiently differentiate between a *Structural Part* from a *Non-Structural Part* in the given CAD pool. During the tool testing, different sub-groups of the Validation Set would be employed. My ethical commitment is that I will not present results from only those Validation subsets which showed the best predictive capability of my tool.

Importance for this stage of the project
If my tool shows 99% confidence in predicting the CAD parts in the chosen validation sets and shows only 80% confidence in other sets, then my ethical commitment requires me to present the confidence interval of 80%-99% and not just state 99%. This ethical and professional commitment is necessary to ensure trust in the ML algorithm. One should not cherry-pick the validation sets which present their tools in good light.

Document update information

This is an administrative section for instructional team use only. It exists for the purpose of documenting any changes to the template made after the start of a class session. (This rarely happens, but when it does this is a transparency measure for documentation.) It’s just here to make sure you know we aren’t going to make changes to the document and therefore requirements without telling you.