Personal Manifesto

Prepared By: Prashant Sanghal

Table of Contents

Application in Domain of Interest

Domain: Energy / Supply Chain & Procurement

Problem 1: Energy organizations operate in a high consumer-demand sector, which requires a continuous supply of complex materials and services to keep the lights on. Usually, these services are supplied by 3rd-parties (a.k.a suppliers/distributors/manufacturers) who often set their own price and profit margins based on multiple factors such as market difficulty, availability, and complexity.
As a buying organization, it is a challenge to benchmark supplier prices quickly and often requires going through a 6 to 12 months long bidding cycle, which is a problem. Hence, the question before us is, can we utilize organizational historical spend data to build a predictive model that can estimate should-cost prices (historical contract pricing) before we buy? Would it be beneficial for business users to know this before they negotiate contracts and sign off deals?

Problem Type: This would be considered a regression problem, which will use historical price as a target variable and other features such as order quantity, dates, supplier rebates etc. as an input to predict supplier pricing. The primary goal of this model would be to fit a regression line between input and target variables that gives the least variance between actual and predicted values. Upon model deployment, organizations would be able to use this as an internal benchmark to see if the newly offered price follows the same trend or deviates due to changed market conditions.

Problem 2: Another problem in this domain is inventory management, where business users in anticipation of maintenance, repairs or turnaround, order large inventory levels filling up the entire warehouse space, leaving excess inventory unused by operations. This results in disposing unused inventory for a fraction of the paid cost.

Problem Type: This would be considered a combination of co-occurrence grouping and profiling problem, where former approach would be used to group warehouse inventory levels in to High/Medium/Low baskets driven by usage, while the latter approach would be used to monitor usage shift when a high moving item downgrades to a lower basket.
The primary goal here is to free up inventory cash and effectively re-allocate organizational resources to manage most consumed items.

Problem 3: Finally, if we look at spend analysis, we would realize that organizations have multiple spending behaviors and contractual obligations with various 3rd parties but it is not very clear why one relationship approach is so different from another. Example: why consumable supplies such as pipes, fittings, chemical lab supplies, gases etc. are considered tactical purchases while chemicals ordered in bulk (which could also be classified as a consumable) falls under strategic purchase.

It seems classification of categories is decided based on certain business needs such as risk, frequency, volume, total dollars spent etc., which could be a user-defined target and might not be available with us when we train the model.

Problem Type: Firstly, this would be considered a multi-classification problem because there could be more than 2 spend categories such as strategic, tactical, complex, spot and other types. Furthermore, whether this problem would fall under supervised or unsupervised learning, will depend on availability of target variables.
In an instance, where we do not have the target variables predefined, the problem would be considered unsupervised, and we would need to obtain user-defined criteria to first label the targets, and then use clustering to group these targets and set appropriate strategic responses to sustain operational needs.

Problem Formulation (Week 1)

Plan for Knowledge Acquisition

Skills and Knowledge Inventory

For each item below, select one of the following:

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

Know:

- how to conduct an inquiry in my application domain that leads to a good problem formulation
- a repertoire of problem types
- how to map problems in my application domain to the repertoire of problem types

Answer: With reference to supply chain domain, I was able to acquire this capability by working directly with various business users. My primary role was to understand procurement scope and solicit competitive proposals to sustain operational needs.
Last year, I had an opportunity to apply some of these concepts using data science techniques on supply chain dataset I obtained from Kaggle platform, which got me interested in pursuing my higher education and further improving this capability. Through this experiment, I got the chance to consolidate multiple datasets in to a single table, analyze which input variables to keep Vs reject, identify variables we could have gathered more example inventory details etc. to think through business standpoint and understand which business problems (supplier prediction, inventory optimization, spend segmentation) we could solve using this data.

I applied few ML techniques such as regression, trees and clustering to explore above problem types and discovered business insights such as predicted price of assembly by features, trends of fast moving frequently purchased assembly and supply clusters, which was a great learning for me.
However, there are number of learning gaps I have identified in my knowledge base which I would like to improve in all 3 areas of knowledge acquisition.
Through MADS (Master of Applied Data Science) program, I am looking to develop advanced skills in data science and big data with ability to automate ML (machine learning) workflows and code re-use for faster model deployment. By applying six-stages of CRISP-DM (Cross Industry Standard Process for Data Mining) methods, I want to solve multi-dimensional business problems which results in better decision-making capability and higher cost efficiency.
The reason I mentioned big data, auto ML and code re-use is because during problem formation stage knowing the scope, scalability and code reproducibility of the problem we are solving, we can be better prepared in defining and solving the problem we are solving.

Maxims, Questions, and Commitments

I will always ask/say….

Question: What are we trying to predict and why?

Problem description: Reference Problem 1
As a buying organization, it is a challenge to benchmark supplier prices quickly and often requires going through a 6 to 12 months long bidding cycle, which is a problem. Hence, the question before us is, can we utilize organizational historical spend data to build a predictive model that can estimate should-cost prices before we buy? Would it be beneficial for business users to know this before they negotiate contracts and sign off deals?

(In what ways does the question apply to the application or problem context you describe?):
From business standpoint, it is important to know which input features will be selected as a part of this model and how it will correlate with the target variable (should-cost price), which we are trying to predict. Any disconnect in this understanding will result in biased outcomes. Here should-cost price refers to contract pricing paid to the supplier historically. I am assuming that during contracting phase, we have looked at input should-cost factors such as direct costs (cost of raw material, labor, transportation), indirect costs such as (cost of maintenance, overheads, training) and supplier margins, and care was taken through bidding to keep should cost prices competitive.

(Why is the question important to ask in the specified problem context?):
This question is important because without this knowledge we will not be able to build a predictive model with relevant input features. As stated above, we would need to know which cost elements make up the should-cost target variable. For example, if the target cost only

includes the cost of direct material/labor and not other costs such as business expenses, profit margins, overheads etc., then model prediction will not align with the received supplier pricing, resulting in significant price difference even if the model prediction is 100% accurate. The model will be not be a good fit for business use-case and future implementation.

Maxim: What are the business requirements?
(***I posed this maxim as a question, because this is the big idea I have used in supply chain to uncover customer business needs. Posing this as a question, helped my customers connect business needs at the organizational level. Example: How far suppliers should be located from the business location to improve supply assurance, how much budget should we allocate for procurement to successfully complete plant turnaround and so on ***)

Problem description: How should we segment our spend type? Should we follow a naming standard to do this or actually talk to a business user to understand what they want for the plant and how they want it supplied?

(In what ways does the maxim apply to the application or problem context you describe?):
Business requirements define how the product/service should be supplied to the business users. For example: Should the product category meet the highest quality standards and always be delivered on-time or is there some tolerance built-in to accommodate for lower prices. While doing category segmentation, usually when the focus is based on the naming standard, user-defined needs are often missed. Organizations receive whatever pricing and the potential to realize cost efficiencies are missed. Example: For more information on naming standards, we could look at UNSPSC codes to see how different products are classified but these names are not necessarily based on how business users (maintenance, operations) use these categories in their business. Example: As per naming standard 'Tools' would fall under MRO supplies (low dollar high volume supply) while as per 'user-defined' need 'Tools' would fall under specialized category (high dollar high volume supply) critical for precision cutting.
By understanding business requirements as a maxim, we seek to understand our customers beyond a data transaction, which in this case would be a user-defined need.

(Why is the maxim important to follow in the specified problem context?):
Hence, this maxim is important to establish our pre-classification needs i.e. seeking to understand how business intends to use the product we are trying to classify first. As an example, we can reach out to our end users (procurement, maintenance, operations) to understand frequency of buy, volume or demand, total dollars spend per category, as a basis for analyzing category clusters. Example: We could define Cluster 1- to represent strategic purchase because products procured under this cluster are frequently procured, in high volume and has maximum spend compared to cluster 2- which could be a slow-moving category and business does not buy it that often. By classifying our problem using naming standard convention (UNSPSC) alone, we would miss the opportunity to learn user-defined business requirements which is extremely crucial for building the product (classification model in this example) that my customer intends to see.

I will never do (without registering a protest) or I will always...

Ethical commitment: I will never work on projects where data is collected unethically from unreliable sources.

Problem description: Can we utilize organizational historical spend data to build a predictive model that can estimate should-cost prices (historical contract pricing) before we buy? Would it be beneficial for business users to know this before they negotiate contracts and sign off deals?

(What does the ethical commitment mean in the context of the problem?):
Supply chain data is usually confidential and should not be collected without proper consent of the supplying organization or responsible team. As an example: Contractual information such as supplier-should cost price (contract price) is confidential to the party that is supplying as a part of the bidding obligation. If this information is leaked or shared without unauthorize use by anyone, it can lead to loss of revenue and trust issues between parties.

Ethical commitment in above context means having consent to collect data for the purpose of business use-case from all the parties who are involved as part of the confidential agreement. Collecting this information without proper consent from all involved parties is considered stealing information, which would be serious offence. Example: Having unauthorized access to contract pricing of one supplier and sharing it with another competitor to negotiate a better deal. This practice is not acceptable. The same could apply in data science context, where contract pricing was collected from various sources (without proper consent) and then model was shared with various other parties, where target variable (should cost/contract pricing) was left exposed.

(Why is it important to always/never do this in the context of the problem?):
It is important to never do this from the standpoint of avoiding litigation against business code of ethics and ensuring confidential information of all parties is kept confidential under all circumstance, whether supply chain, data science or other use cases business might have. By exposing pricing information to other parties (as an example), we are in breach of contract terms and conditions and hence, a professional care should be taken to always prevent this from happening.
As data scientists, we should always look out for confidential information and should always ask our end users to clarify if we are dealing with any confidential information, which we are in the above problem description.

Data Collection and Cleaning (Week 2)

Plan for Knowledge Acquisition

Wk 2 - Collect Sources for Data Science News: Please see my response under ‘’Sources for Data Science News” at the end of the report.

Skills and Knowledge Inventory

For each item below, select one of the following:

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

Stage 2: Data Collection and Cleaning

Know:

- common problems with data sets that can lead to misleading results of analyses:

Answer: I have some experience in this area such as treating missing values, transforming categorical values to binary form, identifying outliers etc. but I am looking forward to strengthening this skill by taking courses in SIADS 505, 532, 632 which deals with data mining and data manipulation techniques.

- potential data sources in my application domain

Answer: So far, I have used Kaggle platform to apply my previous learnings in data science. In general, supply chain dataset contains sensitive information about supplier pricing so, getting it from the open web without proper authorization would be difficult. I look forward to strengthening this capability (example using API to retrieve information) by taking courses in SIADS 655, 685, 503 which will help me manage ethical issues, extract information from text documents using NLP processing as well as learn other data retrieval techniques.

- how to understand and document data sets

Answer: I believe this would deal with reading data attributes correctly such as: does data attributes contain numerical values (integer, float), Booleans (True/False), categorical text or values describing frequency and dates etc. I have some experience exploring tabular data but would like to strengthen my capability further by taking courses in SIADS 505, 515, which will help me learn various data structures, exploration techniques and how to efficiently process each data structure.

- how to write queries and scripts that acquire and assemble data

Answer: I have some knowledge of SQL (example writing basic queries) but this is an area of development for me. I look forward to strengthening this capability by taking courses in SIADS 511, 611 which will help me learn advanced concepts in SQL, interact with databases and understand database architecture.

- how to clean data sets and extract features:

Answer: I have some experience in this area and I look forward to strengthening this capability by taking courses in SIADS (501, 505) enrolled, 515, 521, 522 which will help me visualize and efficiently see features in dataset that would make sense to keep, create or eliminate features with caution.

Maxims, Questions, and Commitments

I will always ask/say….

Question: What are top 3 issues with the collected data?

Problem description: Reference Problem 3 Week 1:
Finally, if we look at spend analysis, we would realize that organizations have multiple spending behaviors and contractual obligations with various 3rd parties but it is not very clear why one relationship approach is so different from another. Example: why consumable supplies such as pipes, fittings, chemical lab supplies, gases etc. are considered tactical purchases while chemicals ordered in bulk (which could also be classified as a consumable) falls under strategic purchase.
It seems classification of categories is decided based on certain business needs such as risk, frequency, volume, total dollars spent etc., which could be a user-defined target and might not be available with us when we train the model.

(In what ways does the question apply to the application or problem context you describe?):
This is the starting point of gathering data and ensuring that what we gather is relevant to the business problem we are trying to solve. For example: Category segmentation is a multi-classification problem. It can have multiple labels which can be either described by a naming standard or user-defined business needs. Benefit of using naming standard (UNSPSC standard) would be that it will offer a standard way of describing a category, while user-defined business needs (business end users example maintenance, operations etc.) could vary across various business functions. Example: they could call Tools not MRO supplies but a specialized category discussed earlier. Having said that, I chose to define it based on user-defined business needs to address user's problems directly and classify each spend type by 4 main response types which can handle any kind of procurement. That would be

1) Strategic
2) Leveraged or tactical
3) Complex,
4) Spot purchase

Hence, keeping this in mind, the top 3 data collection and cleaning problems that we could encounter in this domain would be:

1) Imbalanced data set i.e. not having enough target variables for each response type to train a model to predict unbiased labels.
2) Low quality gathered data i.e. collected data has multiple missing values (who supplied it when, how much), contains outliers (high order values entered manually), multiple target labels defined differently by various users, and so on.
3) Survivorship bias arising out of categories which were pre-screened to be selected based on high risk, high volume, specific location and limited to tier 1 and 2 suppliers only. Example of Tier 1 supplier would be an OEM manufacturer and Tier 2 would be a global distributor.

(Why is the question important to ask in the specified problem context?):
The success of this model would depend on how clearly each category was classified accurately. Which means, did the model output contained mixed labels? If yes, does it make sense to have tactical response type listed under strategic label? How does business feel about mixed classification? Is there a learning or opportunity to improve?
At the end of it, as data scientists, if we can communicate clarity and reason why model produced these mixed labels, and assist in productive decision making, we are good. Hence, addressing these issues upfront is extremely useful for delivering confidence in your model results and model deployment in real-time.

Maxim: Garbage in, Garbage out
Problem description: Reference problem 2 week 1:
Another problem in this domain is inventory management, where business users in anticipation of maintenance, repairs or turnaround, order large inventory levels filling up the entire warehouse space, leaving excess inventory unused by operations. This results in disposing unused inventory for a fraction of the paid cost.

(In what ways does the maxim apply to the application or problem context you describe?):
Let's consider a worst-case scenario, where model incorrectly classified infrequently/unused inventory type (labeled as 'Low basket' described in week 1) as frequently/highly used item by operations. This would signal business to stock up large inventory of 'low basket' items, encouraging more cash resources to be spent on wrong items. We as data scientists, need to be prepared to take responsibility of our model outcomes and hence, cannot expect the model to function well if we let incorrect data feed in to our model.
Garbage in, could mean two things:

1) Bad data quality and unreliable source of gathering.
2) Running ahead of milestones without fully understanding the business problem.

(Why is the maxim important to follow in the specified problem context?):
Let's discuss the consequences of 'garbage out' on business. We already know, we would end up filling up our warehouse with wrong inventory items, we don't need. But then, bigger question here is how are we going to handle 'garbage out' scenario? Do we need to allocate resources

such as task force to reach out to various customers or suppliers who could buy back this inventory at lower price than what we paid for? Or are we looking at telling our investors/shareholders in quarterly call that due to incorrect model prediction we could not meet your ROI expectations?
There could be number of such examples that an organization might have to manage. As an example: A purchase requisition with incomplete service specifications would be difficult to source from suppliers. No one would know what we are trying to source, including suppliers resulting in un-replenished orders

I will never do (without registering a protest) or I will always...

Ethical commitment: I will never discard a missing column without fully investigating non-missing values it contains.

Problem description: Reference Problem 1 Week 1
Can we utilize organizational historical spend data to build a predictive model that can estimate should-cost prices before we buy? Would it be beneficial for business users to know this before they negotiate contracts and sign off deals?

(What does the ethical commitment mean in the context of the problem?):
Based on my learnings, it simply means preserving data as much as possible as long as it helps us identify the target variables. For example: I encountered a dataset where non-missing values inside the missing columns, contained 23.5% available information, which I used in predicting the model output. Initially, I was tempted to discard this data right away but after doing some research, I learned that deleting missing columns would have impacted my model very much.

(Why is it important to always/never do this in the context of the problem?):
In many instances, data is gathered through manual entry by field personnel's, who are directly involved with supporting or assessing the business function. For example: field operator reporting supplier incident or responsible for raising service notification, and procurement team processing order replenishment of goods and services are examples of different roles and responsibilities we encounter in various organizations. These data entries could all be used in developing detailed spend report but it may not have a prescribed method for recording these transactions under a designated column label. We could have part of information in one column and the rest in another column resulting in missing values in both columns. If we were to delete one of these columns, there is a potential of losing important information that could be used by the model. Hence, it is always important to do this in the context of the problem. Another learning I had from Airbnb blog was to consider each transaction not as merely a data transaction but a series of events involving people making decisions. We want to preserve these decisions as much as possible, and use in trend analysis wherever we can.

Data Analysis and Modeling (Week 3)

Plan for Knowledge Acquisition

Skills and Knowledge Inventory

For each item below, select one of the following:

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

Stage 3: Data Analysis and Modeling

Know:

- common mistakes in data analysis that lead to misleading results:

I have some experience in this area which I acquired by experimenting with Kaggle dataset available on public website (example: identifying & replacing missing values, regularizing outliers and dataset etc.) but I am looking forward to strengthening this skill capability by taking courses in SIADS 501-enrolled (learning modelling and analysis framework), 503 (which will help me explore data science ethics), 505-enrolled & 515 (learning data cleaning techniques and looking forward to strengthening programming skills), 542, 543, 642, 643 & 655 (where I will learn various modelling techniques depending upon the type of problem I will be solving). I also feel that data mining concepts offered through SAIDS 532, 632 will be very helpful in identifying and solving data gathering problems upfront (example: does collected data have confounding or conditional on collider attributes that can lead to misleading results?).

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

I have some experience in this area and had a chance to experiment with regression, trees, clustering and neural net modelling techniques on Kaggle dataset. But I am looking forward to strengthening my skills in this capacity by taking courses in SIADS 542, 543, 642, 643, 655 & 523 (which will help me develop understanding of best modelling approach by data type and business problems, help me automate model selection & hyperparameter tuning with Cross-validation and multi-model evaluation via pipelines. It will also help me interpret various metrices such as AUC ROC, RMSE, Bias-Variance trade off analysis, and communicate results to wider audiences, in simple way and more effectively). I learned from my recent informational interview that ‘keeping it simple’ is the key to understanding your customer.

Having said that, before strengthening modelling skill, I would like to develop strong skills in data gathering, preparation and how to deal with data at scale (example: SIADS 516, 511 & 611 which deals with big data, SQL & databases will be a critical skill to add to my toolbox, where I lack background).

Maxims, Questions, and Commitments

I will always ask/say….

Question: What are some possible confounds?

Problem description: Reference problem 2 week 1:
Another problem in this domain is inventory management, where business users in anticipation of maintenance, repairs or turnaround, order large inventory levels filling up the entire warehouse space, leaving excess inventory unused by operations. This results in disposing unused inventory for a fraction of the paid cost.

(In what ways does the question apply to the application or problem context you describe?):
Confounder is a variable that influences both dependent and independent variables. In our above problem, let's consider an example where 'buying' at a given time is dependent on the 'order controls' that are set based on historical demand usage. Here it would seem that 'buying' and 'order controls' are dependent on historical usage but in fact there is another variable 'change in maintenance frequency' which would determine how 'buying, order controls and historical usage' should change with the increase in maintenance frequency. It is like the more we need to maintain, the more we would need to buy to replenish inventory levels.
Hence, 'change in maintenance frequency' would be considered a confounder which would need to be controlled in order to avoid distortion in 'buying' and 'order controls' beyond previous usage.
In addition, there could be other confounders that could impact 'buying' and 'order controls' independent of the usage. As an example, increase in the supply of defective goods could impact 'order controls and buying frequency' because suddenly we are now finding ourselves ordering same stuff from other suppliers more often, and returning defective goods against initial orders. In this case, controlling quality as a confounder would be an important factor.

(Why is the question important to ask in the specified problem context?):
Identifying and controlling confounders is important because otherwise it could lead to bias and distortion in our analysis (high variance). As an example, we would find that uncontrollable 'change in maintenance frequency' or 'increase in the supply of defective goods' has skewed our previously set 'buying' and 'order controls' making our model irrelevant, non-dependable and no longer supporting business needs. It could also lead to unaccounted increase in procurement cost (overhead expenses) because of lack of planning time to prepare for volume spikes or cost of switching suppliers.

Maxim: Fishing is fine for tuna, bad for data analysis.
(Here I will discuss the effect of multi-testing on the same data)

Problem description: Reference Problem 1 Week 1

As a buying organization, it is a challenge to benchmark supplier prices quickly and often requires going through a 6 to 12 months long bidding cycle, which is a problem. Hence, the question before us is, can we utilize organizational historical spend data to build a predictive model that can estimate should-cost prices before we buy? Would it be beneficial for business users to know this before they negotiate contracts and sign off deals?

(In what ways does the maxim apply to the application or problem context you describe?):
Let's consider a scenario. Say, we did our predictive modelling and established a pricing benchmark for various supplier products and are ready to share our model findings with the business. The test score we got after running the trained model on unseen data, was 70%. However, when we went to discuss the results with the business users, we got the sense that they want a model that can predict supplier price with at least 85% accuracy, which I agree should have been identified earlier, during the problem formulation stage.
So, what would be our options now, considering we need to improve our model accuracy?

1) We can gather more data, clean some noisy features (unwanted), re-train our model and hopefully that should get us to 85% accuracy. Or,
2) We can go back and refine our model parameters, try out a different modeling approach (example trees instead of regression) to improve accuracy score. Or,
3) Least preferred, we can run multiple tests on the same data and keep adjusting the model hyperparameters until we get 85% accuracy to align with business expectations.

Let's say, due to time constraints, we decided to go with option 3 "run multiple test" to reach accuracy goals. We handed-off the project to the business. Everybody is excited to try this model, and now they run this trained model on unseen data.

Our results are not what business was expecting. Model is not performing how everyone on the team had anticipated. The contract negotiation team cannot rely on this model to sign off deals or use benchmarked price to estimate supplier margins due to high variance. The main point here is that, when doing analysis, it is important to know how our model could impact business needs. Hence, I found this maxim critical to ensure that we are talking to people regularly, asking questions to understand what a good model looks like to our customers. This maxim, almost sounds like it is an ethical commitment but before that comes "due diligence" making sure that we have done what we are required to do before we say "I will not do that, it's ethically unfit to do so".

(Why is the maxim important to follow in the specified problem context?):
For me, this maxim is important from the standpoint of implementing 'due diligence' and making sure that we are looking out for red flags in our data story as we continue to unfold insights in each phase.
Examples:

1) Seems we don't have enough data to address the business problem
2) We have lots of missing data including target variables. How should we train?
3) Input features are non-linear, I see correlation.
4) There is presence of confounders, we have not addressed, and so on.

But we continue to run our analysis and somehow achieve results (using multi-testing in this case), which clearly our business users will not be able to depend on, unfortunately.

I will never do (without registering a protest) or I will always...

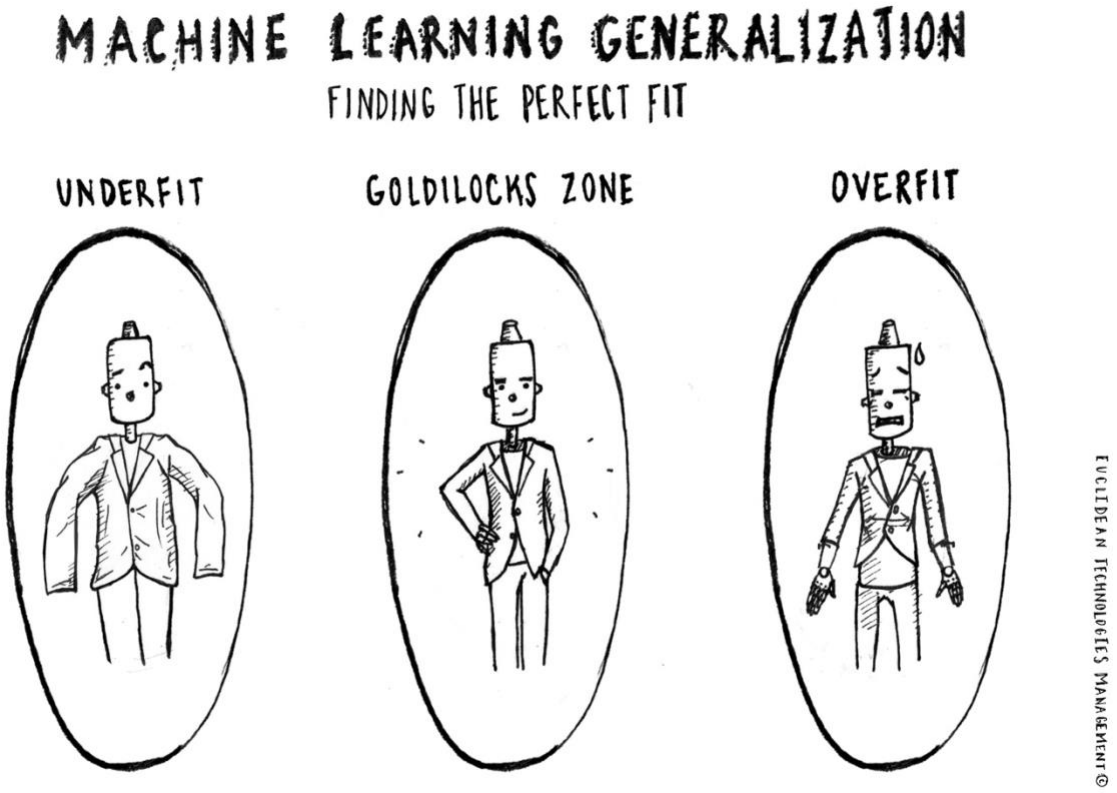
Ethical commitment: I will always hold out some data as a test set, to avoid overfitting.

Problem description: Reference Problem 1 Week 1
As a buying organization, it is a challenge to benchmark supplier prices quickly and often requires going through a 6 to 12 months long bidding cycle, which is a problem. Hence, the question before us is, can we utilize organizational historical spend data to build a predictive model that can estimate should-cost prices before we buy? Would it be beneficial for business users to know this before they negotiate contracts and sign off deals?

(What does the ethical commitment mean in the context of the problem?):
Let's build further on the same example discussed above. Now, that we know the impact of doing multi-testing on the same dataset. It has now become our ethical commitment to ensure that we hold out some dataset for validation testing instead of running multiple tests on the unseen data.
As an example: I would always use cross-validation approach on training dataset, where I would create equal size bins of train-test splits on training samples and use 'holdout' portion in each bin to hyperparameter tune my training model. The benefit is, that way I would be able to build a well generalized model which gets trained and optimized on each 'holdout' portion a.k.a validation set, before it is tested on unseen data for the first time. This will ensure that model parameters are well trained and have been validated in each iteration to obtain least difference between training and test scores, thus avoiding model from overfitting.
In other words, it's like training a child in a classroom and asking him/her to appear in a series of pre-tests before taking the final exam which will happen only once and only the final score will be added to the report card.

(Why is it important to always/never do this in the context of the problem?):
This ethical commitment is important because we want to avoid overfitting i.e. we don't want our training and test scores too far apart because then it will make our model irrelevant for commercial use or supporting business needs. It could also mean that the features we used to train our model does not represent the test data. Hence, we might have to gather more data and ask questions from business users to see how new data we are collecting is representative of the overall population.
For example: Let's say, the same child who took series of pre-tests to do well in the final exam, did extremely well at the school level but scored lower marks at the state level school competition. It turns out, students who participated at the state level school competition had taken additional courses, which were not covered by some schools in the classroom. Requiring, some students to take additional training, improve learning through pre-tests before they could perform well in the external competition.

Below is a picture I found on https://www.euclidean.com/overfitting-underfitting-models which shows difference between overfitting, underfitting and perfect fit, which was very interesting.



Presenting and Integrating into Action (Week 4)

Plan for Knowledge Acquisition

Skills and Knowledge Inventory

For each item below, select one of the following:

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

Stage 4: Presenting and Integrating into Action

Know:

- how to present results to domain experts who are not data scientists:

I have some experience in this area but I am looking forward to strengthening this skill by taking SIADS courses in 523, 521, 522, 524, 622 and 680 which I believe will help me focus on summarizing results, action plans utilizing visually interactive dashboards and communicating uncertainty with clarity. My goal is to keep it simple and work on delivering a product that customer wants to see and finds it sustainable.
Previously, as a part of my personal project, I had an opportunity to do a project walk through with ML engineer and data scientist who gave me valuable feedback and emphasized on developing real-time dashboarding capability.

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

I do not have experience in this area, and I am looking forward to developing my strengths by taking SIADS courses in 511, 515, 516, 611, 643 and 652. These courses will help me develop skills in SQL, databases and scaling big data technologies where I see myself integrating project work with advanced users in this area and be able to talk the same language. I, also see building machine learning pipelines, efficient data processing, network analysis and cloud computing (example AWS) as a key component in automating workflows and putting real-world models in to production.

Maxims, Questions, and Commitments

I will always ask/say….

Question: Did we deliver on customer needs?

Problem description: Reference problem 2 week 1:
Another problem in this domain is inventory management, where business users in anticipation of maintenance, repairs or turnaround, order large inventory levels filling up the entire warehouse space, leaving excess inventory unused by operations. This results in disposing unused inventory for a fraction of the paid cost.

(In what ways does the question apply to the application or problem context you describe?):
Here customer need refers to ‘never buying unused inventory’ for operations. This question is relevant because customer wants to free up this cash and probably want to use it to upgrade plant’s equipment. So, a critical success factor for us would be to monitor, how much this model was useful for the customer to save money (cut inventory cost) after deployment. Did analytics clearly show us unused stocked items and bin location? Were we able to see year-over-year reduction in unused supplier returns and inventory disposal? Was our model able to self-deal with new data changes in production and continue to generate relevant insights? Did it help decision makers make decisions?

These are just some of the questions which form the part of meeting customer needs. Others factors that could also play a vital role directly or indirectly would be the cost and risk of implementing a new model, project completion timeline and possibility of delay in troubleshooting problems, when something goes wrong.

As an example: When we buy a product from Wal-Mart, we want that product to deliver value for money as well as fulfill the reason why we bought it for. We also want that product to work safely and in easy to use manner. Otherwise, we move on.
I see customer needs no differently in data science context. It is our responsibility to deliver against these needs.

(Why is the question important to ask in the specified problem context?):
It is important in above problem context because we are ensuring that during each stage of data science cycle such as understanding business problems, gathering data & preparation, modeling, evaluation and while deploying models, we keep product deliverable as close and as aligned with customer needs as possible. Any gap in this area will leave our customers unhappy and not wanting to use the model, which would not be good.
As an example, in above problem, we are trying to build an inventory optimization model, which would require significant effort on the customer side to get project approved by multiple stakeholders. It may also require some revision in sourcing policy resulting in supplier negotiations and adjustment in pricing. It may also build big expectations among end users (such as warehouse personnel, procurement and field staff) who would want to see this model succeed and use allocated resources (personnel, equipment and storage) efficiently etc.
To accomplish this, customer would need to spend money in our expertise to save dollars down the road. For me, I see it as delivering against customer needs as well as expectations build-up on consistent basis.

Maxim: Data doesn't speak for itself
Problem description: Reference problem 2 week 1:
Finally, if we look at spend analysis, we would realize that organizations have multiple spending behaviors and contractual obligations with various 3rd parties but it is not very clear why one relationship approach is so different from another. Example: why consumable supplies such as pipes, fittings, chemical lab supplies, gases etc. are considered tactical purchases while chemicals ordered in bulk (which could also be classified as a consumable) falls under strategic purchase.

(In what ways does the maxim apply to the application or problem context you describe?):

Organizations spend money on goods and services every day. We collect this information in the form of spend report and try to analyze where we are spending money, why and with whom? When I first got the chance to look at this report, I was like it is pretty straight forward. We know our top supplier spend, where they are supplying and when. We just need to sustain that relationship and monitor business demand. But later I realized, it was more than that, way more. Few months later, when I pulled the spend report again, I noticed that our top supplier spend had changed. I spoke to various end users and came to know they found a better product with a different supplier, which was clearly not reflected in the spend report.
That day, I realized that data does not speak for itself. We need more checks and controls and gather information beyond traditional means to understand change in business priorities. We needed an alert system that could notify supply chain, procurement, finance, warehouse and other departments, when category spend shifts above or below certain levels with a specific supplier. Moreover, we needed it in real-time so that we could adjust our strategic response (i.e. ask rebates, reduce contract admin effort etc.) as appropriate and quickly.

In data science context, Let's take clustering example. If we group suppliers and products by business requirements such as total spend per product, frequency of purchase, market difficulty, length of relationship etc., we would be able to track the movement of suppliers and products from one cluster to another, when business conditions change. We can see this visually, in real-time and explain it to others easily by defining what each cluster represents for. Furthermore, we can run regression and classification models to answer other business questions.
However, this would require need for setting up this capability by modeling raw data, which would be difficult to decode it otherwise.

(Why is the maxim important to follow in the specified problem context?):
It is important from the standpoint of finding actionable insights and ensuring we are able to take business decisions for the benefit of organizational success.
As an example: Let's us consider we continue to respond to a tactical supplier, who recently won a strategic contract with our company. This supplier is now responsible for buying our products as well as supplying goods to our organization. What would be some of the dangers of treating this relationship as a one-off tactical relationship, instead of strategic? Example: we would not have a dedicated relationship manager and allocated resources to resolve business issues such as billing, contract disputes etc. in a timely manner. We would continue to operate at tactical level resulting in claims and loss in trust for each party.
In order to avoid this, we would need relationship management plan and team engagement for both organizations to work together as per effective date.

In data science context, we can deploy models in production to help organizations uncover actionable insights pre/post business changes. We can prepare for the unknowns, take actions to improve status quo by empowering data to make recommendations for us.

I will never do (without registering a protest) or I will always...

Ethical commitment: Thou shalt not hide uncertainty

Problem description: Reference Problem 1 Week 1
As a buying organization, it is a challenge to benchmark supplier prices quickly and often requires going through a 6 to 12 months long bidding cycle, which is a problem. Hence, the question before us is, can we utilize organizational historical spend data to build a predictive model that can estimate should-cost prices before we buy? Would it be beneficial for business users to know this before they negotiate contracts and sign off deals?

(What does the ethical commitment mean in the context of the problem?):
In above problem, our goal is to predict supplier price using historical spend. We could use multiple variables as an input to our model such as supplier quoted price, order quantity, frequency of purchase, product features, geographical location, stock availability, demand etc. Say, after gathering this data and pre-processing, we trained a random forest regressor to build a predictive model which gave an accuracy score of 90% on test data using top 10 features. Would it mean that this model will produce same results after deployment, in every scenario? Could there be new data features, trends or uncontrolled confounders in production that could influence our model accuracy?
I feel, yes there is a high probability that the model will not be able to replicate the same results and we will need to statistically simulate the model using different scenarios, before it can be implemented with some certainty.
Hence, it is important to always plan for uncertainty in our results. It is difficult to build a 100% accurate, weather proofed model, and our stakeholders/end-users/customers should always be made aware, so that they can plan for it when they implement model recommendations.

(Why is it important to always/never do this in the context of the problem?):
It is important to never hide uncertainty in our model outcomes because it could negatively impact business productivity and relationships. In the context of above problem, we were building supplier pricing prediction to establish a benchmark that could estimate difference between should-cost (actual cost) versus price paid to the supplier. If our model contains market indicators such as commodity price or consumer price index along with other input variables listed above, then our model would be able to predict increase in supplier overhead as market conditions change. Otherwise, we would need to keep these market indicators separately and call it out in our model prediction.

For example: The model summary could say, after simulating model results, we predicted supplier price to range between 85% to 90% accuracy using internal records with 95% confidence, and excludes market indicators such as CPI and commodity pricing adjustments, which could add 3% to 5% in supplier price this year.

By not hiding this uncertainty, we are helping business users such as cost estimation and procurement team build buffer, transparency and fairness in their negotiated deals.

Sources for Data Science News
I plan to follow the following sources of information about data science to keep myself up to date with the industry:

I am a member of Calgary meet up community and like to attend conferences/seminars from time to time to learn about different topics and network with fellow data scientists. Here is the list of meet up groups I am a member of:

- Calgary Artificial Intelligence Meet up
- Calgary Deep Learning Meet up
- Calgary Big Data Open Source Meet up
- Calgary Cognitive, AI & Data Science Meet up

Besides above, I like to read data science blogs from online publishing platforms. They cover many topics in data science such as new trends, examples of coding, how to visualize your results effectively and much more.

- medium.com
- towardsdatascience.com
- analyticsvidhya.com

I also like to refer to stack overflow, when I get stuck in coding.

Personal Project (Optimize Inventory Consumption)

Use company data to identify leftover inventory to cut back on excess, thereby freeing up cash, costly material scrap, and truck loads on highways.

Additional Comments / Final Summary:
In this paper we looked at 3 supply chain areas where we could apply data science and machine learning techniques to solve business problems:

1) Predicting supplier price
2) Optimizing inventory
3) Categorizing Spend

Using CRISP-DM (Cross industry standard process for data mining framework) we explored 6 main stages of data science cycle and saw what kind of questions, maxim and ethical commitments we should keep in mind while solving for above 3 problems:

1) Business Understanding
2) Data understanding
3) Data Gathering and Preparation
4) Modelling
5) Evaluation
6) Deployment

Taking this structured approach is the key takeaway for doing data science work.