DATA SCIENCE METHODOLOGY

Technical Communication Project By:

Dharshan Kumar K S	CB.EN.U4AIE19024
Dinesh S	CB.EN.U4AIE19025
Harish K	CB.EN.U4AIE19029
Kabilan N	CB.EN.U4AIE19033
Sivamaran M.A.C	CB.EN.U4AIE19061
Roshan Tushar S	CB.EN.U4AIE19071

Acknowledgement

For the completion of this assignment, we have taken the guidance and trust of many respected faculty who have our utmost gratitude. As the completion of this assignment gave me much pleasure and more knowledge, we would like to show our gratitude for **Mr. Sarvana Prabhu**, English teacher, in Amrita university for giving us good guidelines for the assignment throughout numerous consultations. We would also like to expand our gratitude to all those who have directly and indirectly guided us in helping us with this assignment.

I also thank Amrita University for consent to include copyrighted pictures as a part of our paper without which the explanations and theories stated in this document could not have been visualised.

Many people, especially my team mates have been a very vital part of this assignment and their opinions have been of the utmost importance for the completion of this assignment.

Declaration

We hereby declare that the project entitled **"Data Science Methodology"** submitted to Amrita Vishwa Vidyapeetham, Coimbatore is the record of original work done by us under the guidance of **Mr. Sarvana Prabhu**, Department of English, Amrita Vishwa Vidyapeetham and this project work has not been submitted for any degree, diploma or other similar titles elsewhere. However, extracts of any content which has been used for this project have been duly acknowledged by providing the details in the references.

5th May 2020

Dharshan Kumar K S

Dinesh S

Harish K

Kabilan N

Sivamaran M.A.C

Roshan Tushar S

Table of Contents

Abstract

A Data Scientist always needs a methodology to solve data science's problems which are applicable in our daily life. Let's suppose that you are a Data Scientist and your first job is to increase sales for a company, they must know what product they should sell on what period. You will need the correct methodology to organize your work, analyze different types of data, and solve their problem. Your customer doesn't care about how you do your job; they only care if you will manage to do it in time.

Despite the recent increase in computing power and access to data over the last couple of decades, our ability to use the data within the decision making process is either lost or not maximized at all too often, we don't have a solid understanding of the questions being asked and how to apply the data correctly to the problem at hand.

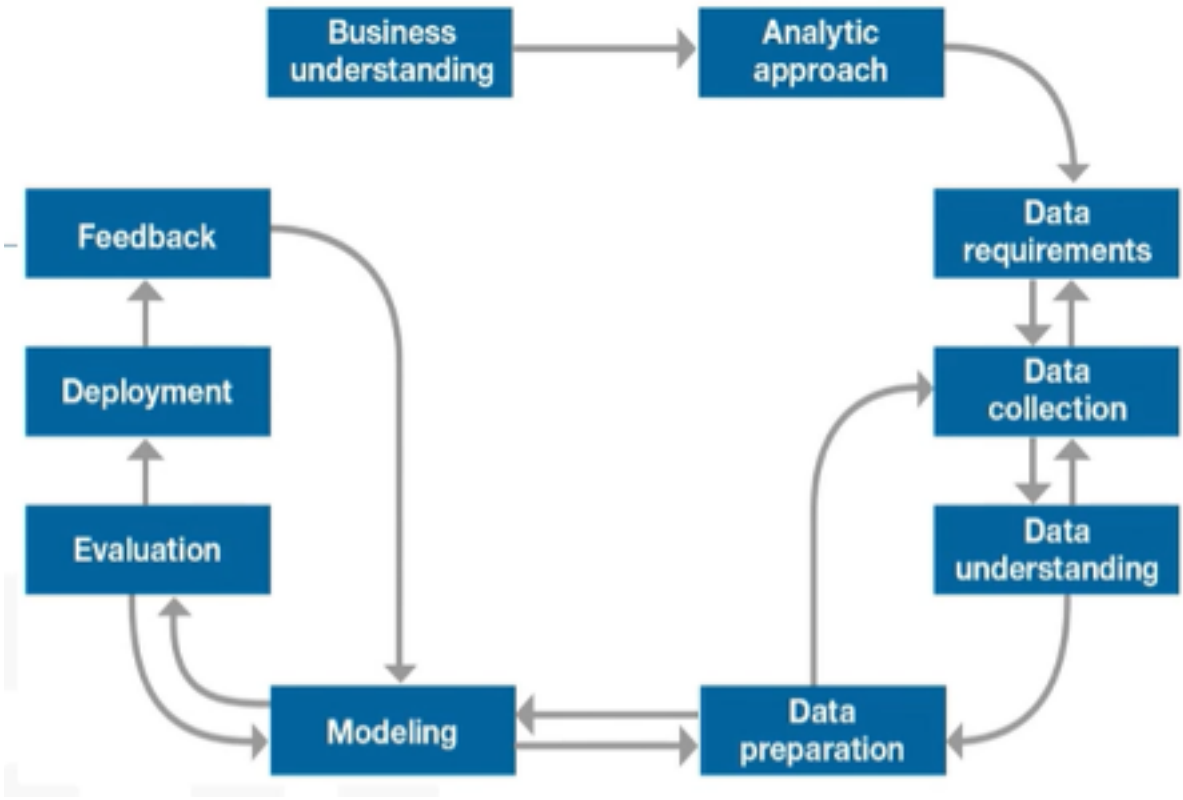
This assignment has one purpose, and that is to share a methodology that can be used within data science, to ensure that the data used in problem solving is relevant and properly manipulated to address the question at hand. Accordingly, in this course, you will learn: - The major steps involved in tackling a data science problem. - The major steps involved in practicing data science, from forming a concrete business or research problem, to collecting and analyzing data, to building a model, and understanding the feedback after model deployment. - How data scientists think!

The people who work in Data Science and are busy finding the answers for different questions every day comes across the Data Science Methodology. Data Science Methodology indicates the routine for finding solutions to a specific problem.

DATA SCIENCE METHODOLOGY

From problem to approach:

Methodology is defined as a system of methods used in a particular area of study or activity.



Understanding the problem:

You have been called by your boss for meeting about an important task with a tight deadline. You both go back and forth to ensure that all aspects of the task have been considered and the meeting ends with both of you confident that things are on track. That day, however, you have spent time in playing, you realize that you need to ask several more questions in order to accomplish the task. Unfortunately, he won't be available until the next day. Now, with the tight deadline. what do you do? Do you risk moving forward or do you stop and seek clarification

Approach to the problem:

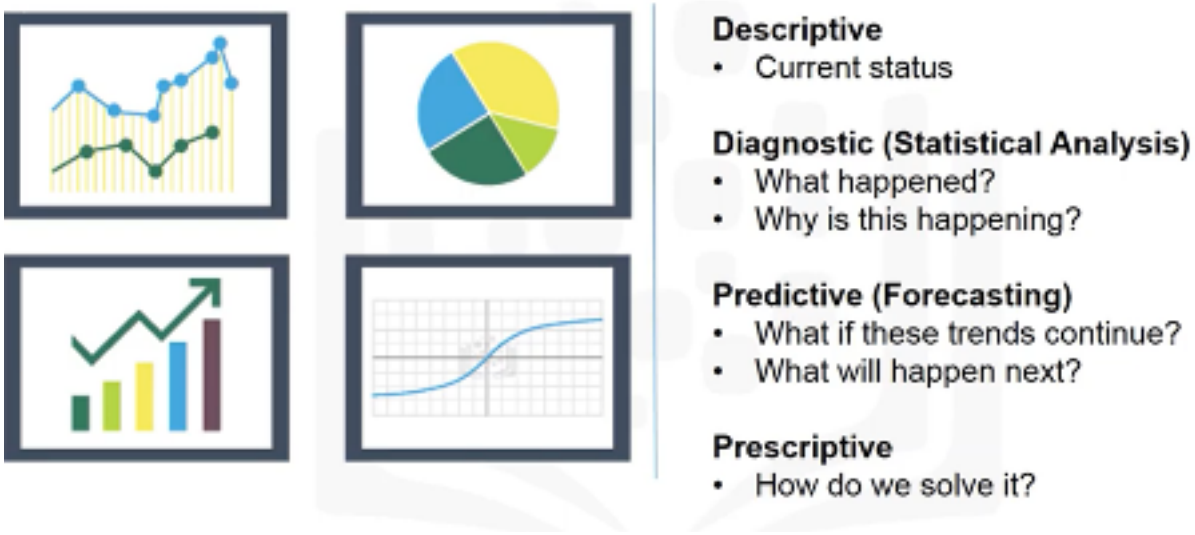
Data science methodology begins with spending the time to seek clarification, to attain what can be referred to as a problem understanding. Having this understanding the problem to be solved, allows you to determine which data will be used to answer the core question. Rollins suggests that having a clearly defined question is essential because it ultimately directs the analytic approach that will be needed to address the question

Seeking clarification:

Having a clearly defined question starts with understanding the goal of the person asking the question. For example, if your boss asks: "How can we reduce the costs of performing an activity?" We need to understand, is the goal to increase the profitability? Or is it to improve the efficiency of the activity? Once the goal is clarified, the next is to find the objectives that are in support of the goal. By breaking down the objectives, structured discussions can take place where priorities can be identified in a way that can lead to organizing and planning on how to tackle the problem.

Analytical approach to the problem:

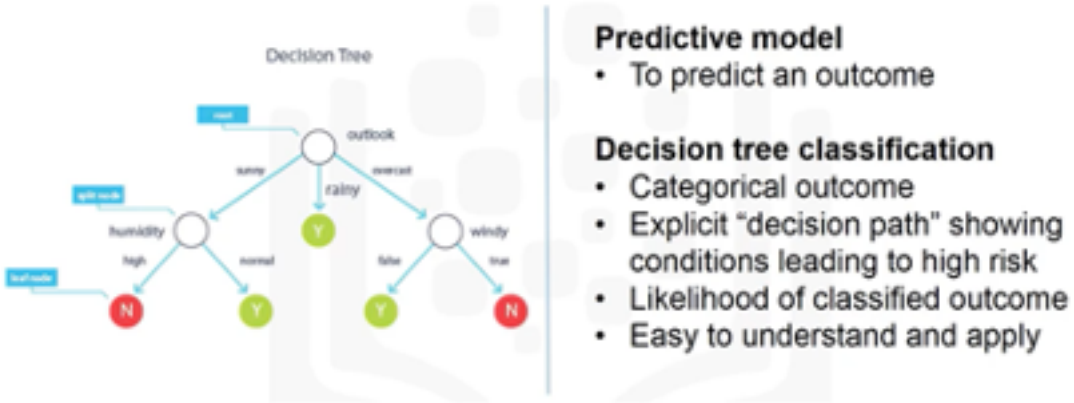
Once the problem is defined, the appropriate analytic approach for the problem is selected in the context of the requirements. Once the question is understand strongly, the analytic approach can be selected. This means identifying the type of patterns will be needed to address the question most effectively. The questions can be descriptive, diagnostic, predictive, prescriptive. If the question determines probability of an action, then a predictive model is used. If the question shows the relationships, a descriptive approach was required. Statistical analysis applies to problems that require counts.



Machine Learning is a field of study that teaches computers to learn without being programmed. Machine Learning can be used to identify relationships and trends in data. In the case where the question is to learn about human behaviour, then an appropriate response would be to use Clustering Association approaches.

Case study- Decision tree:

A decision tree classification model is easy for non-data scientists to understand and apply. Clinicians can readily see what conditions are causing a patient to be marked as high-risk and ML models can be built and used at various points. This gives a representation of the patient's risk and how it is evolving with the various treatments being applied. This is the reason, the decision tree classification approach was chosen for building the Congestive Heart Failure re-admission model.



From Requirements To Collection:

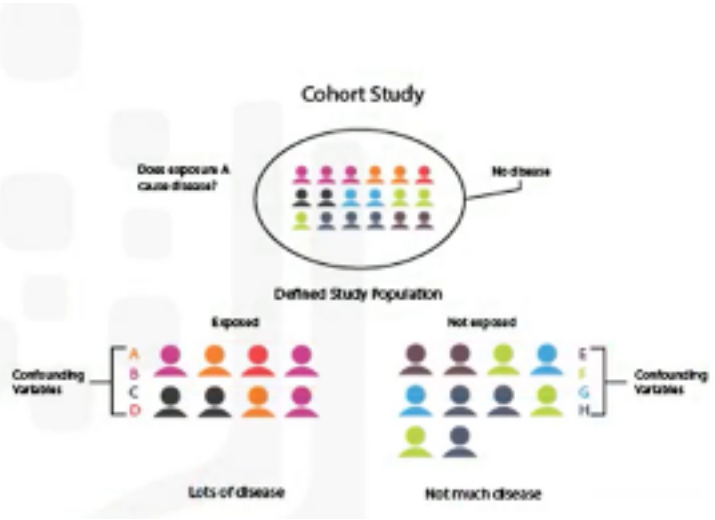
Data requirements:

Think data science methodology as cooking with data. Each step is essential in cooking. So, if the problem that needs to be solved is the recipe and data is an ingredient, then the data scientist needs to identify how to collect them, how to understand or work with them, and how to prepare the data to get the desired outcome. understanding of the problem , and using the analytical approach elected, the Data Scientist is ready to get started.

Requirement to collection:

Prior to undertaking the data collection and data preparation stages of the methodology**,** it's vital to define the data requirements for decision-tree classification. It includes identifying the necessary data content, formats and sources for data collection

Case study- Selecting the cohort:



In the case study, at first we have to define the data requirements for the decision tree classification approach . This included selecting a suitable patient cohort from the health insurance providers. In order to compile the complete clinical histories, patient needed to be admitted as in-patient within the provider service area, so they'd have access to the necessary information. they focused in diagnosis of congestive heart failure during one full year. patient

must have had continuous enrollment for at least six months, prior to the primary admission for congestive heart failure, so that complete medical history could be compiled. Congestive heart failure patients who have other medical conditions, were excluded from the cohort because it causes re-admission rates and, thus, could skew the results.

Defining the data:

We have to define the content, format, and representations of the data needed for decision tree classification. This modeling technique requires one record for each patient. To model the re-admission outcome, the data need to be covering all aspects of the patient's clinical history. This content include admissions, primary, secondary, and tertiary diagnoses, procedures, prescriptions, and other services provided either during hospitalization or throughout doctor visits. Thus, a patient have thousands of records, representing all the attributes. To get to the one record for each patient format, the data scientists rolled up the transactional records to the patient level, creating a number of new variables to represent that information. thinking ahead and anticipating subsequent stages is important.

Data collection:

when the initial data collection is completed by the data scientist takes place to determine whether they have what they need. the data requirements are revised and decisions are made as to whether the collection requires more or less data. Once the data is collected, the data scientist will have a good understanding of what they will be working with. Techniques such as descriptive statistics and visualization can be applied to the dataset, to assess the content, quality, and initial insights about the data. Unfilled places in data will be identified and either filled or omitted.

Gathering available data:

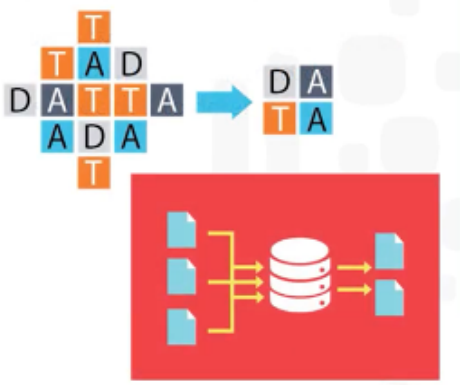
Collecting data requires where to find the data elements that are needed. In the context of our case study, these can include: demographic, clinical and coverage information of patients, provider information, claims records, as well as pharmaceutical and other information related to all the diagnoses of the patients.

Deferring inaccessible data:

When certain data was needed, but that data source was not yet integrated with the rest of the data sources. It is alright to defer decisions about unavailable data, and attempt to acquire it. For example, this can be done after getting some intermediate results from the predictive model. If those results suggest that the data might be important in obtaining a better model, then it is to be done. there we were able to build a reasonably good model without that data.

Merging data:

DBAs and programmers work together to extract data and then merge it. This allows for removing redundant data, making the data available for the following stages of the methodology, which is data understanding. At this stage, data scientists and analytics team members can discuss various ways to manage their data, including automating certain processes in the database, so that data collection is easier and faster.



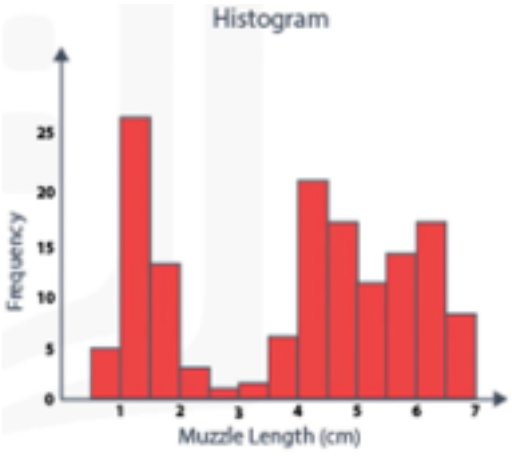
Data understanding:

Data Understanding is the fifth most important process in designing a data model. Before analysing the data, it should be converted into a structured form and create a proper data set. Data understanding mainly enforces on finding and understanding which set of data values or attributes are representative of the data. That is, the attributes which affect the given problem need to found. A case study has been discussed in detail which clearly explains the 'Data Understanding' process.

There are many methods by which we can understand the data. Some of those methods are descriptive statistics, Univariate statistics, Pairwise correlations and Histogram. It is required to choose the appropriate method based on the given problem. Among these methods, the histogram is a good way to understand how values or a variable are distributed, and what sorts of data preparation may be needed to make the variable more useful in a model.

Case Study – Understanding the data:

As part of the case study, we can use data related to congestive heart failure admissions. To understand this data, descriptive statistics needed to be run against the data columns that would become variables in the model. There are three important steps to be followed to understand this data. First, these statistics should include Hearst, univariates, and statistics on each variable, such as mean, median, minimum, maximum, and standard deviation. Second, pairwise correlations were used, to find how closely certain variables were related, and which of them, if any, were very highly correlated, meaning that they might be essentially redundant, thus making only one relevant for modelling. Third, the histograms of the variables were examined to know their distributions.

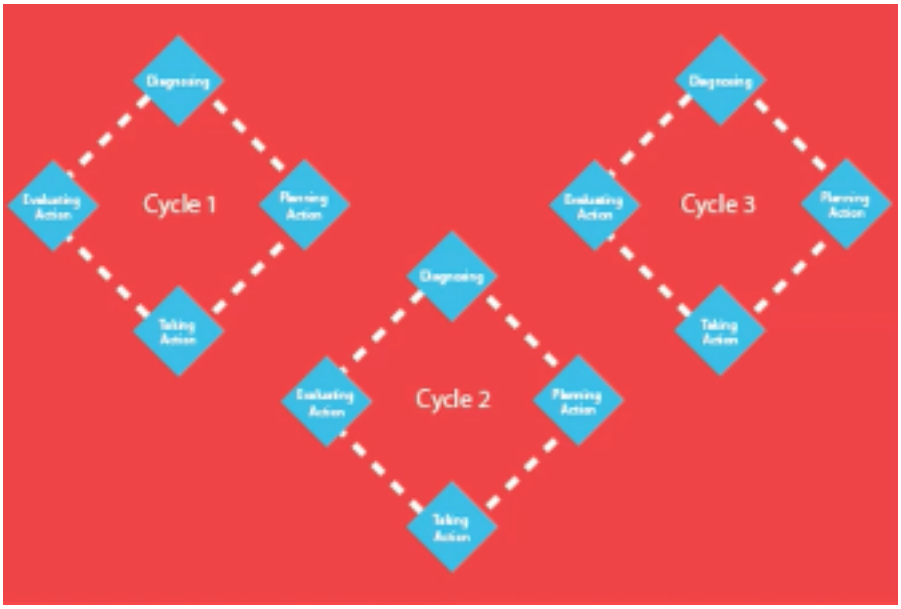


Case Study – Looking at data quality:

A case study has been illustrated to know about data quality. The data quality can be defined as the data set with a minimum number of missing values and invalid or misleading values. The data quality can also be analysed using univariates, statistics and histograms. From the information provided, certain values can be re-written or even dropped if necessary, such as when a certain variable has many missing values. Sometimes a missing value might mean "no", or "0" (zero), or means "we don't know". Or, if a variable contains invalid or misleading values, like a numeric variable called "age" that contains 0 to 100 and also 999, where that "triple-9" actually means "missing", but would be treated as a valid value unless those values are corrected.

Case Study – This is an iterative process:

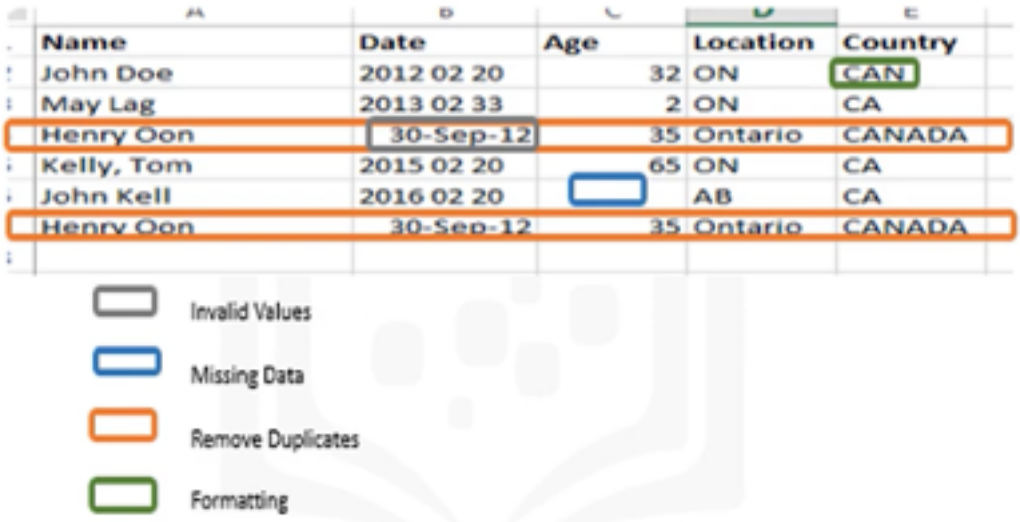
Now the case study explains about the iterative process involved in processing the data. Initially based on the primary diagnosis of congestive heart failure, the meaning of congestive heart failure admission was formed. But working through the data understanding stage revealed that the initial definition wasn't capturing all of the congestive heart failure admissions that were expected, which was based on clinical experience. This meant looping back to the data collection stage and adding secondary and tertiary diagnoses and building a more comprehensive definition of congestive heart failure admission. The more one works with the given problem and its related data, the more one learns and thus the more refinement which will be done within the model, ultimately resulting in a better solution to the problem.



Data preparation:

Data preparation can be defined as cleansing the data. We can relate the data preparation to that of washing freshly picked vegetables in so far as unwanted elements, like dirt or imperfections, are removed. Together with data collection and data understanding, data preparation is that the most time-consuming phase of a data science project, typically taking seventy per cent and even up to even ninety per cent of the overall project time. Automating a number of the data collection and preparation processes within the database can reduce this time to as little as 50 per cent. This time savings translates into increased time for data scientists to concentrate on creating models.

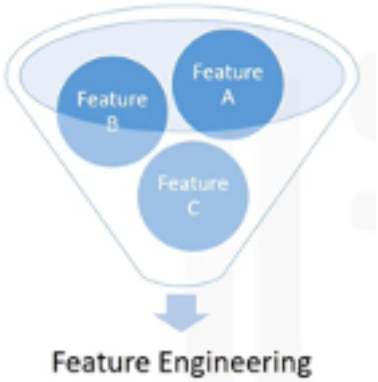
Now we can see about transforming the data into a well-defined format with a smaller number of invalid terms. To continue with our cooking metaphor, it is known that the method of chopping onions to a finer state will leave its flavours to spread through a sauce more easily than that might be the case if we were to drop the entire onion into the sauce pot. Similarly, transforming data within the data preparation phase is the process of getting the data into a state where it's going to be easier to deal with. An example of data cleansing is given below



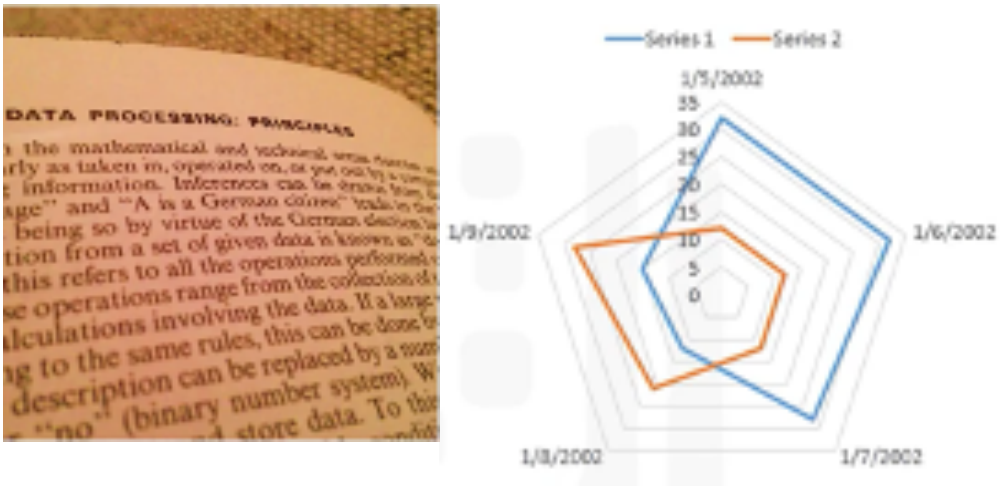
This table illustrates the Data cleansing process

To work effectively with the data, it must be prepared in such a way that addresses missing or invalid values and removes duplicates, toward ensuring that everything is formatted in a proper manner.

Feature engineering is also part of data preparation. It is the method of using domain knowledge of the data to make features that make the machine learning algorithms work. A feature is a characteristic which may help when solving a problem. Features within the data are important to predictive models and can influence the results you would like to achieve. Feature engineering is critical when machine learning tools are being applied to analyse the data.

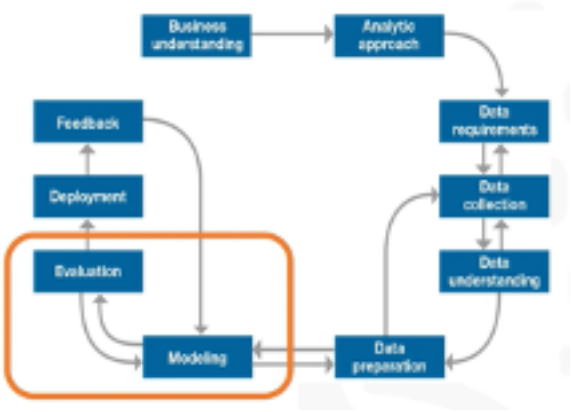


When working with text, text analysis steps for coding the data are required to be able to manipulate the data. The data scientist must know what they're trying to find within their dataset to deal with the question. The text analysis is critical to make sure that the right groupings are set, which the programming isn't overlooking what's hidden within. The data preparation phase sets the stage for subsequent steps in addressing the question. While this phase may take a short time to do, if done right the results will support the project. If this is often skipped over, then the result won't be up to par and should have you ever back at the drawing board. It is necessary to take your time in this area and use the tools available to automate common steps to accelerate data preparation. Make sure to pay attention to the detail in this area. After all, it takes only one bad ingredient to ruin a fine meal.



This picture illustrates the process of text analysis

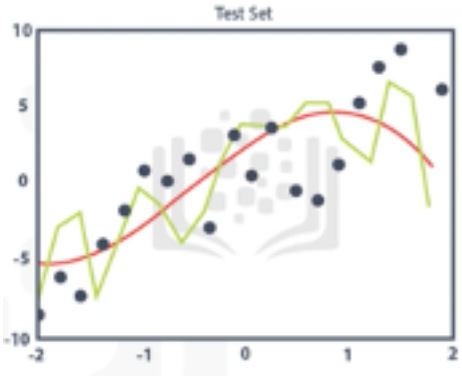
Modelling:



Modelling is the stage in the data science methodology where the data scientist checks the correctness of predicted result with that of the actual result. In this stage, the model can be modified to give the most accurate results. There are two models which are either descriptive or predictive. An example of a descriptive model might examine things like: if a person did this, then they're likely to prefer that. A predictive model tries to yield yes/no, or stop/go type outcomes. These models are dependent on the analytic approach that was taken, either statistically driven or machine learning-driven.

Using training / test sets:

The data scientist will use a training set for predictive modelling and test set for testing the results. A training set is defined as a set of historical data with known outcomes. The training set acts such as a gauge to find if the model needs to be calibrated. In this stage, the data scientist will perform different algorithms to make sure that the variables which are used in the model are required.



The success of data compilation, preparation and modelling, depends on the understanding of the problem which was given, and therefore the appropriate analytical approach being taken. The data supports the answering of the question, and just like the quality of the ingredients in cooking, sets the stage for the result. Constant refinement, adjustments and tweaking are necessary within each step to ensure the outcome is a solid one. In John Rollins' descriptive Data Science Methodology, the framework is designed to do 3 things. First, understand the question at hand. Second, select an analytic approach or method to solve the problem, and third, obtain, understand, prepare, and model the data. The end goal is to move the data scientist to an extent where a data model can be efficiently built to answer the question.

Overall, as Data Science is a fairly new field of study, and we are interested in the possibilities it has to offer. The more people that benefit from the outcomes of this practice, the further the field will develop.

Model evaluation:

Model evaluation can have two main phases the diagnostic measures phase and the statistical significance testing. The first is the diagnostic measures phase, which is used to ensure the model is working as intended. It can be used to examine the areas that require adjustments. Statistical significance testing can be applied to the model to ensure that the data is being properly handled and interpreted within the model.

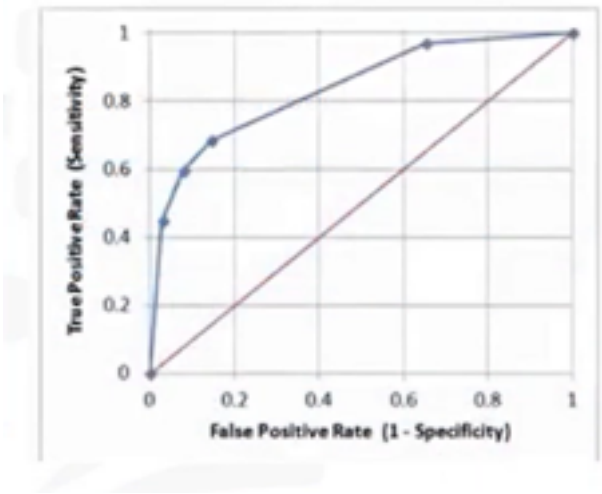
Model	Relative Cost Y:N	True Positive Rate (Sensitivity)	Specificity (accuracy on N)	False Positive Rate (1 – Specificity)
1	1:1	0.45	0.97	0.03
2	1.5:1	0.60	0.92	0.08
3	4:1	0.68	0.85	0.15
4	9:1	0.97	0.35	0.65

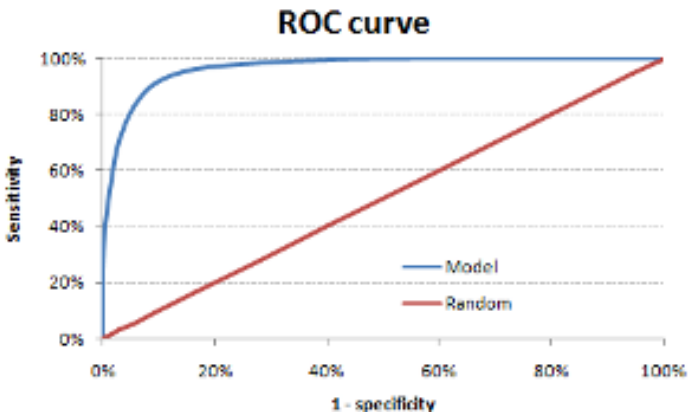
Case study-misclassification costs:

Misclassification costs is a metric often used to evaluate the performance of a model. As shown in this table, four models were built with four different relative misclassification costs. The true positive rate measures the proportion of actual positives that are correctly identified by the model, on the other hand, Specificity is defined as the proportion of actual negatives, which got predicted as the negative (or true negative).

As we see, each value of this model-building parameter increases the true-positive rate, or sensitivity, of the accuracy in predicting yes, at the expense of lower accuracy in predicting no, that is, an increasing false-positive rate. The question then becomes, which model is best based on tuning this parameter? There must be a balance between true-positive and false-positive rate for the best model.

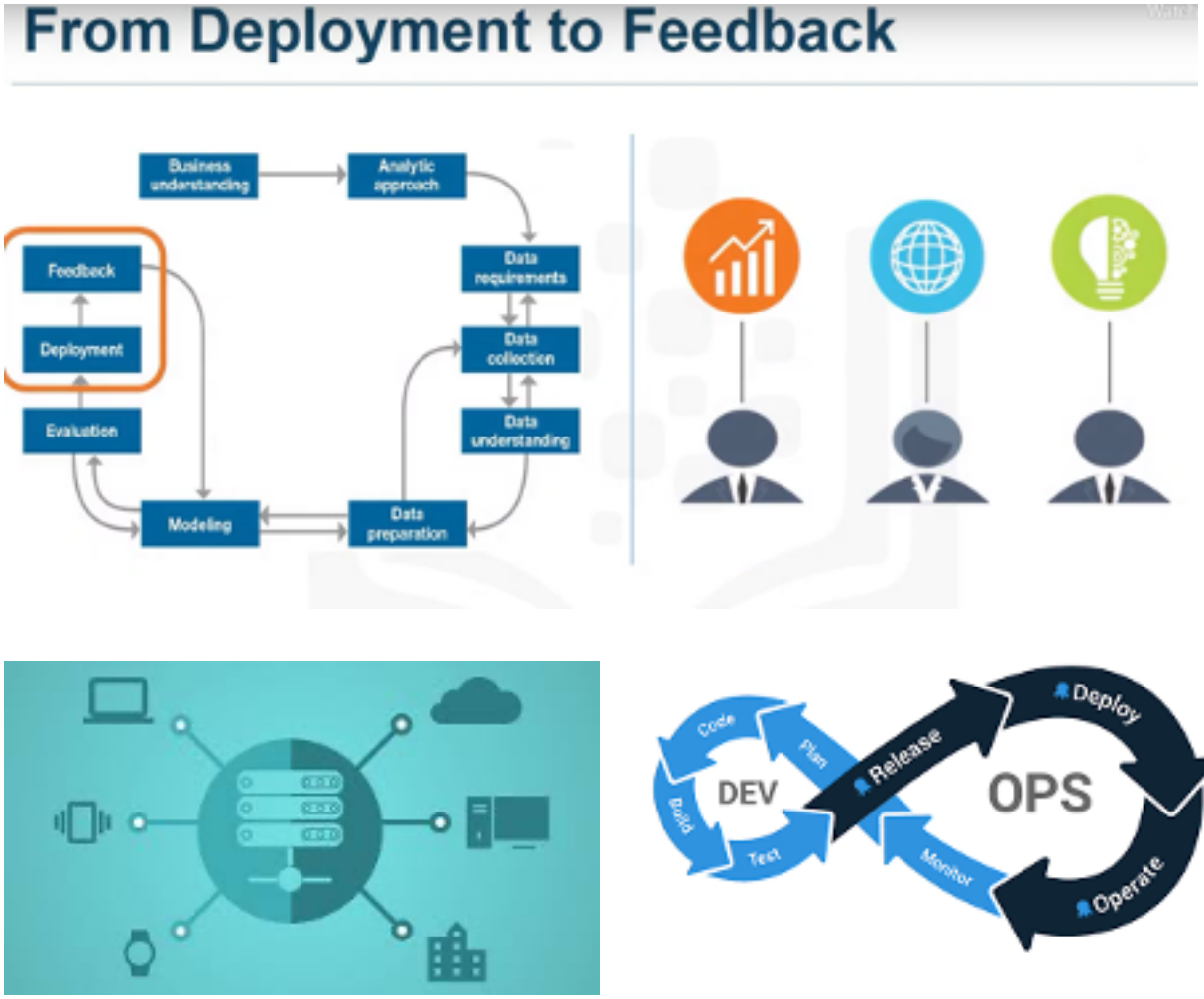
Case study – using the ROC curve:





ROC stands for receiver operating characteristic curve, which was first developed during World War II to detect enemy aircraft on radar. As you can refer to the above figure, the optimal model is the one giving the maximum separation between the blue ROC curve relative to the red baseline. This curve quantifies how well a binary classification model performs, declassifying the yes and no outcomes when some discrimination criterion is varied.

Deployment:



Deployment is all of the activities that make a system available for use. Once the model is evaluated and when the data scientist is confident that the model will work, it is deployed. The Deployment stage depends on the purpose of the model. In many instances, the model is rolled out or made available to a limited group of users, who are your target group.

It is important to practically understand the meaning of the model results. Assimilate knowledge for business implications of the model results for designing intervention actions. Finally, the model results get incorporated by the stakeholders of the business; by taking appropriate measures in their business model or making crucial changes in the business strategy.

So now, let's look at the case study related to applying Deployment. In preparation for solution deployment, the next step was to assimilate the knowledge for the business group who would be designing and managing the intervention program to reduce readmission risk. In this scenario, the business people translated the model

results so that the clinical staff could understand how to identify high-risk patients and design suitable intervention actions. The goal, of course, was to reduce the likelihood that these patients would be readmitted within 30 days after discharge.

During the business requirements stage, the Intervention Program Director and her team had wanted an application that would provide automated, near real-time risk assessments of congestive heart failure. It must easy to use such that every staff member can carry it around with ease. The patient data would be generated throughout the hospital stay. The data would automatically be fit in a format as per the ease and requirement of the model, then each patient would be scored near the time of discharge. Clinicians would then have the most up-to-date risk assessment for each patient, helping them to select which patients to target for intervention after discharge.

There would also be additional requirements like the intervention team would develop and deliver training for the clinical staff. There would be a requirement for tracking/monitoring processes. Also, processes for tracking and monitoring patients receiving the intervention would have to be developed in collaboration with IT developers and database administrators, so that the results could go through the feedback stage and the model could be refined over time.

Feedback:

Once the model is deployed, feedback from the users is collected and used to reassess the model. The changes are incorporated back in the modelling so that the performance can be improved. The model has more impact when the feedback is successfully incorporated into the model.

Throughout the data science methodology, the processes within are cyclical, ensuring refinement at each stage. Likewise, the process such as modelling, evaluation, deployment and feedback form a cycle in the listed order. This cycle can be ended when the adjustments made to the model has an insignificant change in the result.

In the case study, after the deployment and feedback stages, the impact of the intervention program on readmission rates would be reviewed after the first

year of its implementation. Then the model would be refined, based on all of the data compiled after model implementation and the knowledge gained throughout these stages. Other refinements included: Incorporating information about participation in the intervention program and possibly refining the model to incorporate detailed pharmaceutical data. But after feedback and practical experience with the model, it can be determined that adding that data is worth the investment of time and effort.

Also, the intervention actions and processes would be reviewed and very likely refined as well, based on the experience and knowledge gained through initial deployment and feedback. Finally, the refined model and intervention actions would be redeployed, with the feedback process continued throughout the life of the Intervention program.

References:

- https://medium.com/towards-artificial-intelligence/the-data-science-methodology-50d60175a06a
- https://www.geeksforgeeks.org/data-science-methodology-and-approach/
- https://en.wikipedia.org/wiki/Sensitivity_and_specificity
- https://medium.com/@chhavi.saluja1401/data-preparation-a-crucial-step-in-data-mining-dba35772f281
- https://en.wikipedia.org/wiki/Data_preparation
- https://www.ibm.com/support/knowledgecenter/SS3RA7_15.0.0/com.ibm.spss.crispdm.help/crisp_data_understanding_phase.htm
- https://cognitiveclass.ai/courses/data-science-methodology-2
- https://www.coursera.org/lecture/data-science-methodology
- https://www.sqlshack.com/introduction-to-data-science-data-understanding-and-preparation/
- https://medium.com/ml-research-lab/part-3-data-science-methodology-from-understanding-to-preparation-a666a8203179
- https://www.questionpro.com/blog/data-collection/
- https://www.informatica.com/in/services-and-training/glossary-of-terms/data-preparation-definition.html
- https://www.guru99.com/data-modelling-conceptual-logical.html