

in data mining, the Cross Industry Process for Data Mining (CRISP-DM) methodology is widely used.

What is CRISP-DM?

The CRISP-DM methodology is a process aimed at increasing the use of data mining over a wide variety of business applications and industries. The intent is to take case specific scenarios and general behaviors to make them domain neutral. CRISP-DM is comprised of six steps with an entity that has to implement in order to have a reasonable chance of success. The six steps are shown in the following diagram:

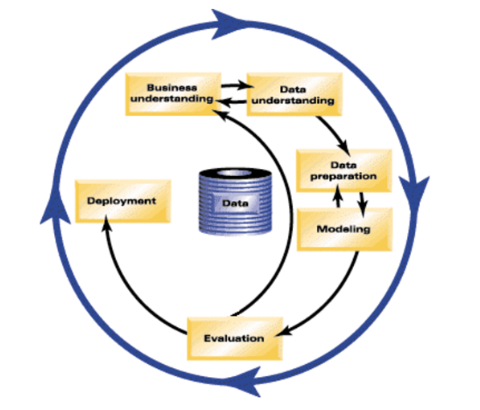


Fig.1 CRISP-DM model, IBM Knowledge Center, CRISP-DM Help Overview

**Business Understanding** This stage is the most important because this is where the intention of the project is outlined. Foundational Methodology and CRISP-DM are aligned here. It requires communication and clarity. The difficulty here is that stakeholders have different objectives, biases, and modalities of relating information. They don’t all see the same things or in the same manner. Without clear, concise, and complete perspective of what the project goals are resources will be needlessly expended.

**Data Understanding** Data understanding relies on business understanding. Data is collected at this stage of the process. The understanding of what the business wants and needs will determine what data is collected, from what sources, and by what methods. CRISP-DM combines the stages of Data Requirements, Data Collection, and Data Understanding from the Foundational Methodology outline.

**Data Preparation** Once the data has been collected, it must be transformed into a useable subset unless it is determined that more data is needed. Once a dataset is chosen, it must then be checked for questionable, missing, or ambiguous cases. Data Preparation is common to CRISP-DM and Foundational Methodology.

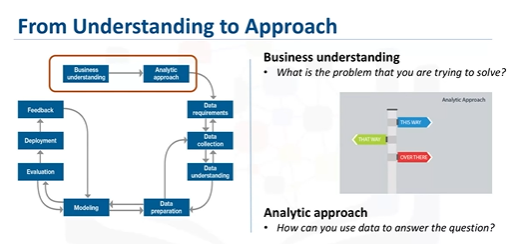
**Modeling** Once prepared for use, the data must be expressed through whatever appropriate models, give meaningful insights, and hopefully new knowledge. This is the purpose of data mining: to create knowledge information that has meaning and utility. The use of models reveals patterns and structures within the data that provide insight into the features of interest. Models are selected on a portion of the data and adjustments are made if necessary. Model selection is an art and science. Both Foundational Methodology and CRISP-DM are required for the subsequent stage.

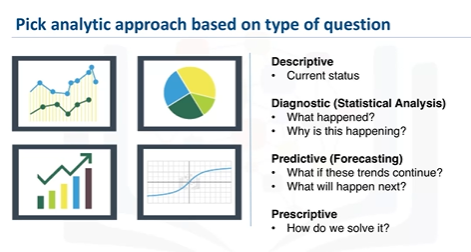
**Evaluation** The selected model must be tested. This is usually done by having a pre-selected test, set to run the trained model on. This will allow you to see the effectiveness of the model on a set it sees as new. Results from this are used to determine efficacy of the model and foreshadows its role in the next and final stage.

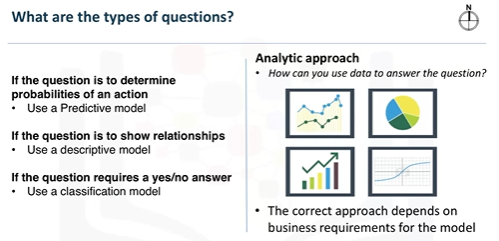
**Deployment** In the deployment step, the model is used on new data outside of the scope of the dataset and by new stakeholders. The new interactions at this phase might reveal the new variables and needs for the dataset and model. These new challenges could initiate revision of either business needs and actions, or the model and data, or both.

CRISP-DM is a highly flexible and cyclical model. Flexibility is required at each step along with communication to keep the project on track. At any of the six stages, it may be necessary to revisit an earlier stage and make changes. The key point of this process is that it’s cyclical; therefore, even at the finish you are having another business understanding encounter to discuss the viability after deployment. The journey continues.

Analytic Approach

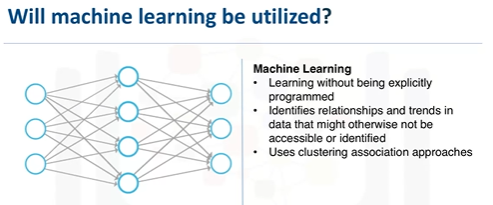


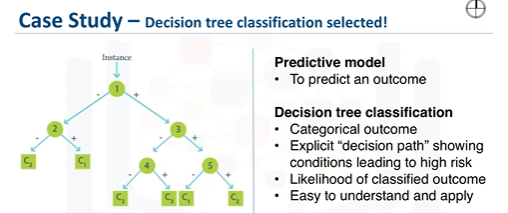




descriptive approach: This would be one that would look at clusters of similar activities based on events and preferences. Statistical analysis applies to problems that require counts.

Machine Learning can be used to identify relationships and trends in data that might otherwise not be accessible or identified. In the case where the question is to learn about human behavior, then an appropriate response would be to use Clustering Association approaches.





For the case study, a decision tree classification model was used to identify the combination of conditions leading to each patient's outcome. In this approach, examining the variables in each of the nodes along each path to a leaf, led to a respective threshold value. This means the decision tree classifier provides both the predicted outcome, as well as the likelihood of that outcome, based on the proportion at the dominant outcome, yes or no, in each group. From this information, the analysts can obtain the readmission risk, or the likelihood of a yes for each patient. If the dominant outcome is yes, then the risk is simply the proportion of yes patients in the leaf. If it is no, then the risk is 1 minus the proportion of no patients in the leaf. A decision tree classification model is easy for non-data scientists to understand and apply, to score new patients for their risk of readmission. Clinicians can readily see what conditions are causing a patient to be scored as high-risk and multiple models can be built and applied at various points during hospital stay. This gives a moving picture of the patient's risk and how it is evolving with the various treatments being applied. For these reasons, the decision tree classification approach was chosen for building the Congestive Heart Failure readmission model.

IBM Watson Studio lets you build and deploy an AI solution, using the best of open source and IBM software and giving your team a single environment to work in.

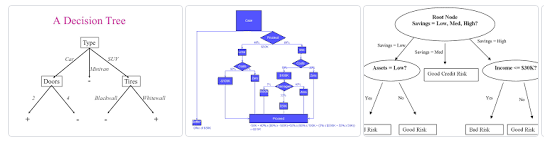
Why are we interested in data science?

Once the business problem has been clearly stated, the data scientist can define the analytic approach to solve the problem. This step entails expressing the problem in the context of statistical and machine-learning techniques, so that the entity or stakeholders with the problem can identify the most suitable techniques for the desired outcome.

a machine learning algorithm, decision trees, and see if it is the right technique to automate the process of identifying the cuisine of a given dish or recipe while simultaneously providing us with some insight on why a given recipe is believed to belong to a certain type of cuisine.

What is a decision tree?

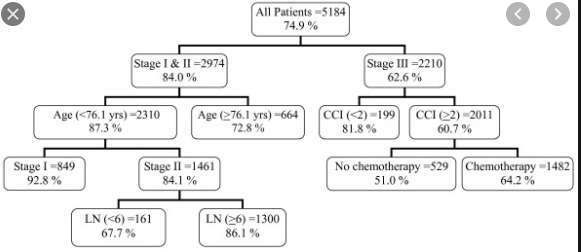
A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.



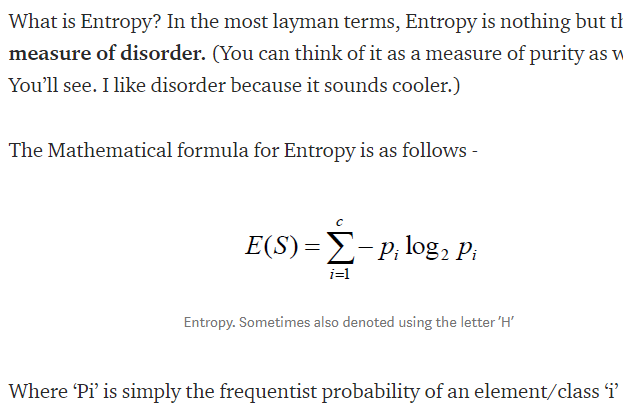
In order to build a very powerful decision tree for the recipe case study, let's take some time to learn more about decision trees.

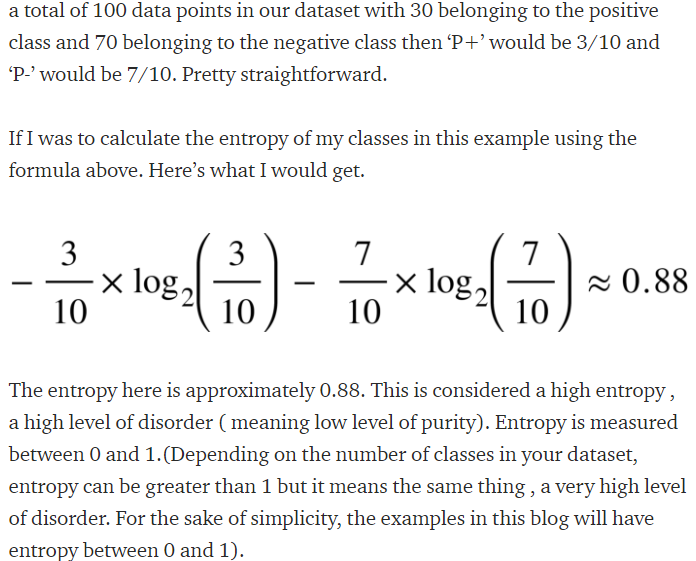
* Decision trees are built using recursive partitioning to classify the data.
* When partitioning the data, decision trees use the most predictive feature (ingredient in this case) to split the data.
* Predictiveness is based on decrease in entropy - gain in information, or impurity.

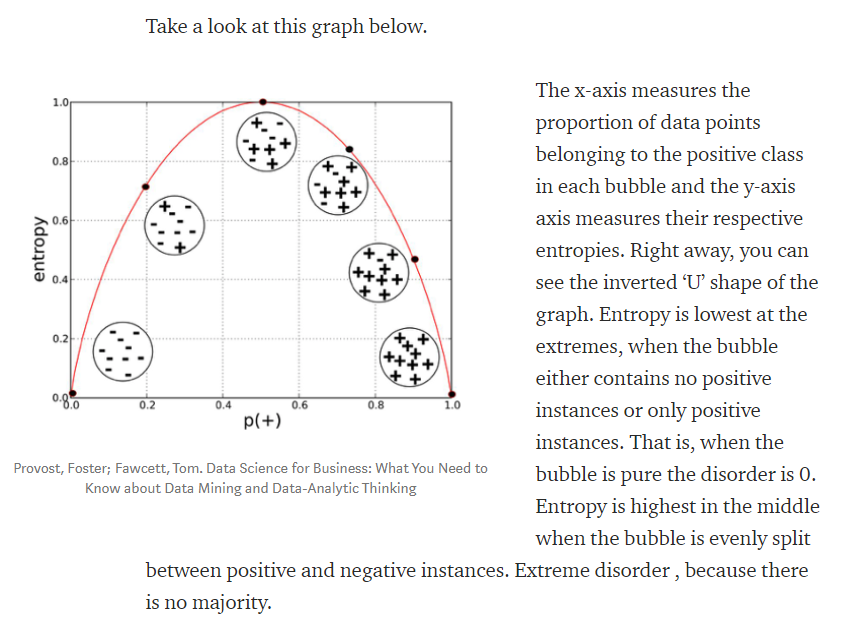
Recursive partitioning is a statistical method for multivariable analysis. Recursive partitioning creates a decision tree that strives to correctly classify members of the population by splitting it into sub-populations based on several dichotomous independent variables.

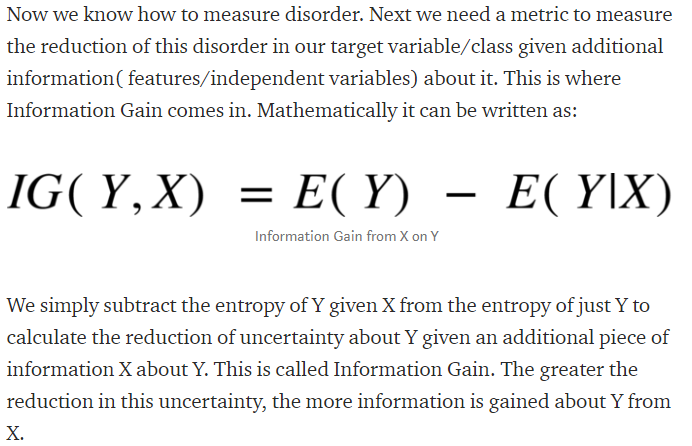


Entropy is a measure of disorder or uncertainty and the goal of machine learning models and Data Scientists in general is to reduce uncertainty.









A tree stops growing at a node when:

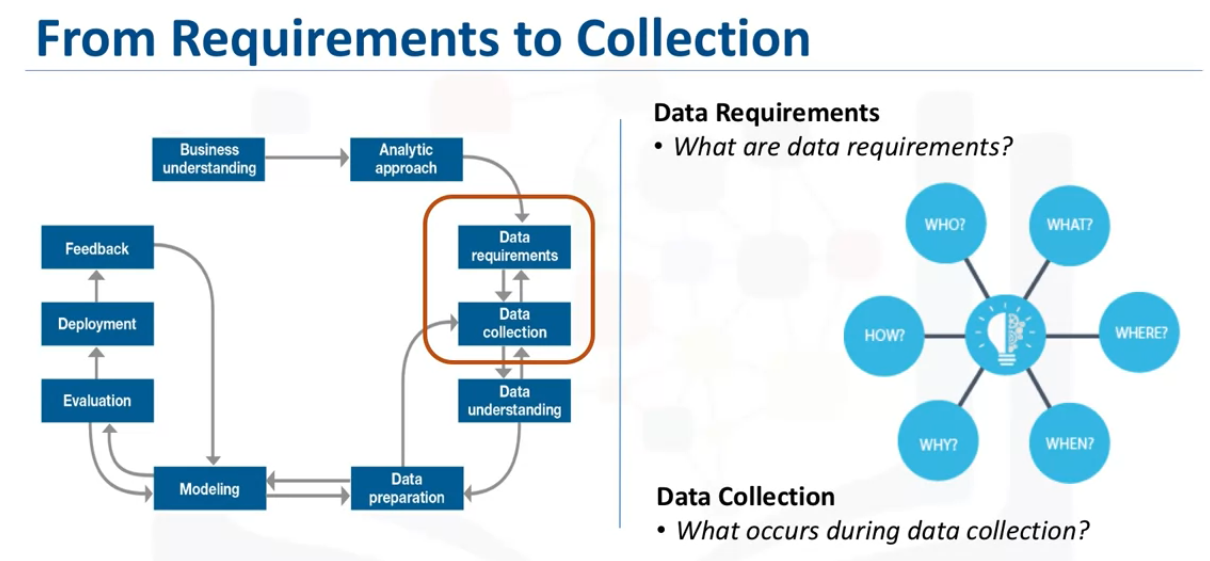
* Pure or nearly pure.
* No remaining variables on which to further subset the data.
* The tree has grown to a preselected size limit.

**Data Requirements**

Think of this section of the data science methodology as cooking with data. Each step is critical in making the meal. So, if the problem that needs to be resolved is the recipe, so to speak, and data is an ingredient, then the data scientist needs to identify: which ingredients are required, how to source or the collect them, how to understand or work with them, and how to prepare the data to meet the desired outcome. Building on the understanding of the problem at hand, and then using the analytical approach selected, the Data Scientist is ready to get started.

some examples of the data requirements within the data science methodology.

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. This includes identifying the necessary data content, formats and sources for initial data collection.



Data Collection

After the initial data collection is performed, an assessment by the data scientist takes place to determine whether or not they have what they need. As is the case when shopping for ingredients to make a meal, some ingredients might be out of season and more difficult to obtain or cost more than initially thought. In this phase the data requirements are revised and decisions are made as to whether or not the collection requires more or less data. Once the data ingredients are collected, then in the data collection stage, 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 data set, to assess the content, quality, and initial insights about the data. Gaps in data will be identified and plans to either fill or make substitutions will have to be made.

Collecting data requires that you know the source or, know where to find the data elements that are needed. It is alright to defer decisions about unavailable data, and attempt to acquire it at a later stage. For example, this can even be done after getting some intermediate results from the predictive modeling. If those results suggest that the drug information might be important in obtaining a good model, then the time to try to get it would be invested.

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

**Lab**

In the videos, we learned that the chosen analytic approach determines the data requirements. Specifically, the analytic methods to be used require certain data content, formats and representations, guided by domain knowledge.

Identifying the required data fulfills the data requirements stage of the data science methodology.

In the initial data collection stage, data scientists identify and gather the available data resources. These can be in the form of structured, unstructured, and even semi-structured data relevant to the problem domain. Eg. Web Scraping of Online Food Recipes

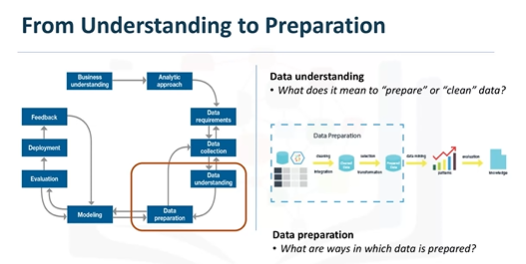
Now that the data collection stage is complete, data scientists typically use descriptive statistics and visualization techniques to better understand the data and get acquainted with it. Data scientists, essentially, explore the data to:

* understand its content,
* assess its quality,
* discover any interesting preliminary insights, and,
* determine whether additional data is necessary to fill any gaps in the data.

Week 2

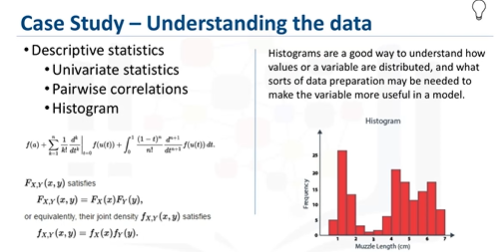
Data Understanding

Data understanding encompasses all activities related to constructing the data set. Essentially, the data understanding section of the data science methodology answers the question: Is the data that you collected representative of the problem to be solved?

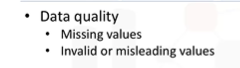


In order to understand the data related to congestive heart failure admissions, descriptive statistics needed to be run against the data columns that would become variables in the model.

* First, these statistics included Hearst, univariates, and statistics on each variable, such as mean, median, minimum, maximum, and standard deviation.
* Second, pairwise correlations were used, to see how closely certain variables were related, and which ones, if any, were very highly correlated, meaning that they would be essentially redundant, thus making only one relevant for modeling.
* Third, histograms of the variables were examined to understand their distributions. Histograms are a good way to understand how values or a variable are distributed, and which sorts of data preparation may be needed to make the variable more useful in a model.



The univariates, statistics, and histograms are also used to assess data quality. From the information provided, certain values can be re-coded or perhaps even dropped if necessary, such as when a certain variable has many missing values. The question then becomes, does "missing" mean anything? Sometimes a missing value might mean "no", or "0" (zero), or at other times it simply means "we don't know". Or, if a variable contains invalid or misleading values, such as 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 we corrected it.

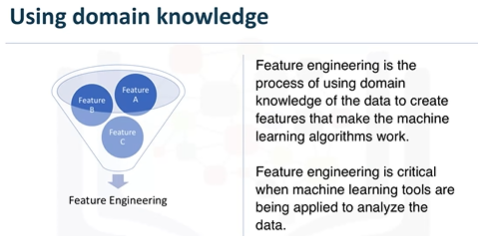


Initially, the meaning of congestive heart failure admission was decided on the basis of a primary diagnosis of congestive heart failure. But working through the data understanding stage revealed that the initial definition was not capturing all of the congestive heart failure admissions that were expected, 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. This is just one example of the interactive processes in the methodology. The more one works with the problem and the data, the more one learns and therefore the more refinement that can be done within the model, ultimately leading to a better solution to the problem.

Data Preparation – Concepts

Transforming data in the data preparation phase is the process of getting the data into a state where it may be easier to work with. Specifically, the data preparation stage of the methodology answers the question: What are the ways in which data is prepared? To work effectively with the data, it must be prepared in a way that addresses missing or invalid values and removes duplicates, toward ensuring that everything is properly formatted.

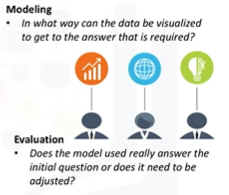
Feature engineering is also part of data preparation. It is the process of using domain knowledge of the data to create features that make the machine learning algorithms work. A feature is a characteristic that might help when solving a problem. Features within the data are important to predictive models and will influence the results you want to achieve.



When working with text, text analysis steps for coding the data are required to be able to manipulate the data. The data scientist needs to know what they're looking for within their dataset to address the question. The text analysis is critical to ensure that the proper groupings are set, and that the programming is not overlooking what is hidden within. The data preparation phase sets the stage for the next steps in addressing the question. While this phase may take a while to do, if done right the results will support the project. If this is skipped over, then the outcome will not be up to par and may have you back at the drawing board. It is vital 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 just one bad ingredient to ruin a fine meal.

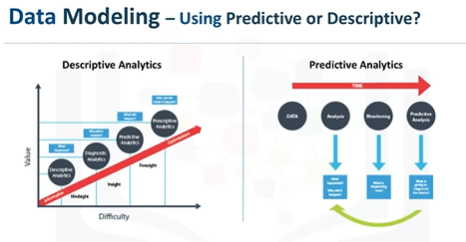
Modeling – Concepts

Modelling is the stage in the data science methodology where the data scientist has the chance to sample the sauce and determine if it's bang on or in need of more seasoning!



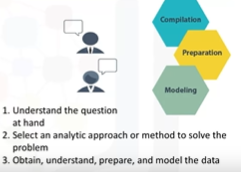
What is the purpose of data modeling, and second, what are some characteristics of this process?

Data Modelling focuses on developing models that 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 based on the analytic approach that was taken, either statistically driven or machine learning driven.



The data scientist will use a training set for predictive modelling. A training set is a set of historical data in which the outcomes are already known. The training set acts like a gauge to determine if the model needs to be calibrated. In this stage, the data scientist will play around with different algorithms to ensure that the variables in play are actually required. The success of data compilation, preparation and modelling, depends on the understanding of the problem at hand, and the appropriate analytical approach being taken. The data supports the answering of the question, and like the quality of the ingredients in cooking, sets the stage for the outcome. Constant refinement, adjustments and tweaking are necessary within each step to ensure the outcome is one that is solid.





The end goal is to move the data scientist to a point where a data model can be built to answer the question. In this stage of the methodology, model evaluation, deployment, and feedback loops ensure that the answer is near and relevant. This relevance is critical to the data science field overall, as it ís a fairly new field of study, and we are interested in the possibilities it has to offer.

Parameter tuning to improve the model (one of the many aspects of model building)

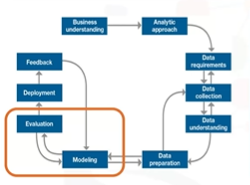
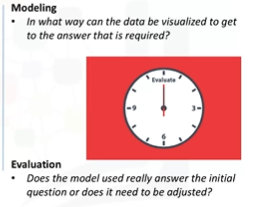
How could the accuracy of the model be improved in predicting the yes outcome? For decision tree classification, the best parameter to adjust is the relative cost of misclassified yes and no outcomes.

A statistician calls this a type I error, or a false-positive. type II error, or a false-negative

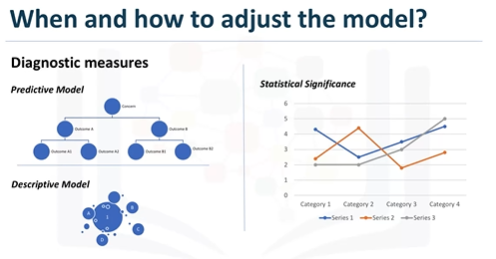
Evaluation

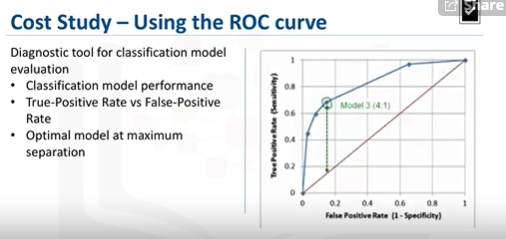
A model evaluation goes hand-in-hand with model building as such, the modeling and evaluation stages are done iteratively. Model evaluation is performed during model development and before the model is deployed. Evaluation allows the quality of the model to be assessed but it's also an opportunity to see if it meets the initial request.

Evaluation answers the question: Does the model used really answer the initial question or does it need to be adjusted?



Model evaluation can have two main phases. The first is the diagnostic measures phase, which is used to ensure the model is working as intended. If the model is a predictive model, a decision tree can be used to evaluate if the answer the model can output, is aligned to the initial design. It can be used to see where there are areas that require adjustments. If the model is a descriptive model, one in which relationships are being assessed, then a testing set with known outcomes can be applied, and the model can be refined as needed. The second phase of evaluation that may be used is statistical significance testing. This type of evaluation can be applied to the model to ensure that the data is being properly handled and interpreted within the model. This is designed to avoid unnecessary second guessing when the answer is revealed .





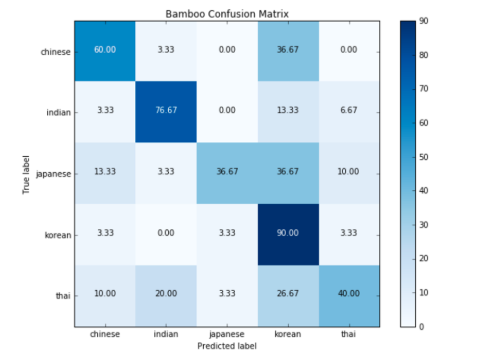
As you can see on this slide, the optimal model is the one giving the maximum separation between the blue ROC curve relative to the red base line. We can see that model 3, with a relative misclassification cost of 4-to-1, is the best of the 4 models. And just in case you were wondering, ROC stands for receiver operating characteristic curve, which was first developed during World War II to detect enemy aircraft on radar.

The ROC curve is a useful diagnostic tool in determining the optimal classification model. This curve quantifies how well a binary classification model performs, declassifying the yes and no outcomes when some discrimination criterion is varied. In this case, the criterion is a relative misclassification cost. By plotting the true-positive rate against the false-positive rate for different values of the relative misclassification cost, the ROC curve helped in selecting the optimal model.

LAB

To quantify how well the decision tree is able to determine the cuisine of each recipe correctly, we will create a confusion matrix which presents a nice summary on how many recipes from each cuisine are correctly classified. It also sheds some light on what cuisines are being confused with what other cuisines.

After running the above code, you should get a confusion matrix similar to the following:



The rows represent the actual cuisines from the dataset and the columns represent the predicted ones. Each row should sum to 100%. According to this confusion matrix, we make the following observations:

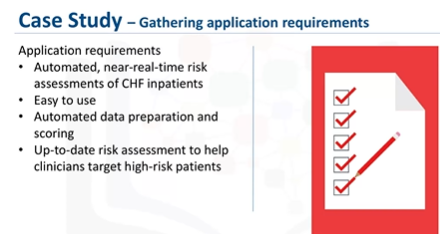
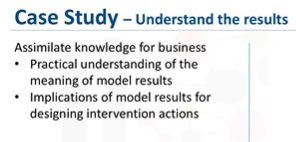
* Using the first row in the confusion matrix, 60% of the Chinese recipes in bamboo\_test were correctly classified by our decision tree whereas 37% of the Chinese recipes were misclassified as Korean and 3% were misclassified as Indian.
* Using the Indian row, 77% of the Indian recipes in bamboo\_test were correctly classified by our decision tree and 3% of the Indian recipes were misclassified as Chinese and 13% were misclassified as Korean and 7% were misclassified as Thai.

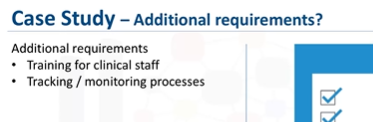
Please note that because decision trees are created using random sampling of the datapoints in the training set, then you may not get the same results every time you create the decision tree even using the same training set. The performance should still be comparable though!

Deployment

While a data science model will provide an answer, the key to making the answer relevant and useful to address the initial question, involves getting the stakeholders familiar with the tool produced. In a business scenario, stakeholders have different specialties that will help make this happen, such as the solution owner, marketing, application developers, and IT administration. Once the model is evaluated and the data scientist is confident it will work, it is deployed and put to the ultimate test. Depending on the purpose of the model, it may be rolled out to a limited group of users or in a test environment, to build up confidence in applying the outcome for use across the board.

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 also had to be easy for clinical staff to use, and preferably through browser-based application on a tablet, that each staff member could carry around. This patient data was generated throughout the hospital stay. It would be automatically prepared in a format needed by the model and 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. As part of solution deployment, the Intervention team would develop and deliver training for the clinical staff. 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.



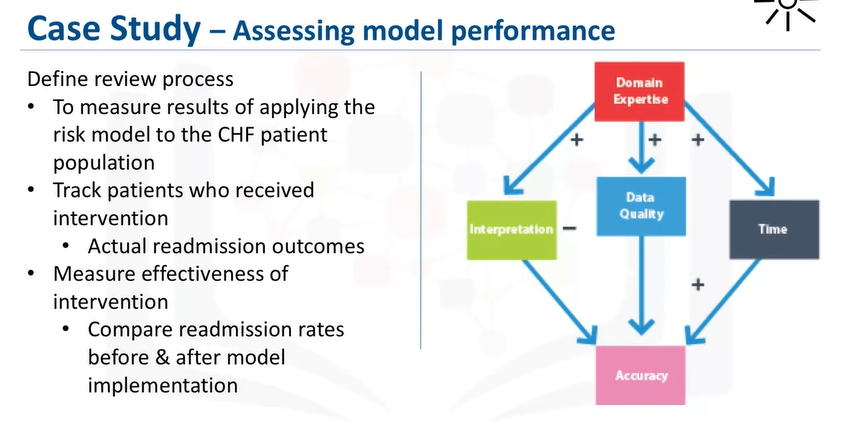


Feedback

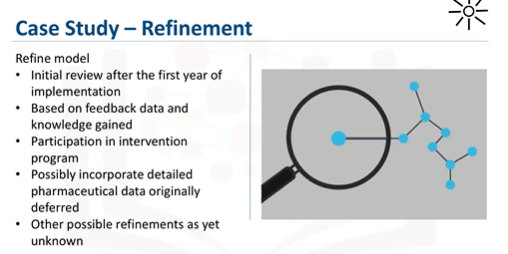
Feedback from the users will help to refine the model and assess it for performance and impact. The value of the model will be dependent on successfully incorporating feedback and making adjustments for as long as the solution is required.



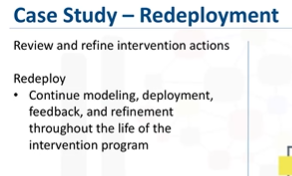
The plan for the feedback stage included these steps: First, the review process would be defined and put into place, with overall responsibility for measuring the results of a "flying to risk" model of the congestive heart failure risk population. Clinical management executives would have overall responsibility for the review process.



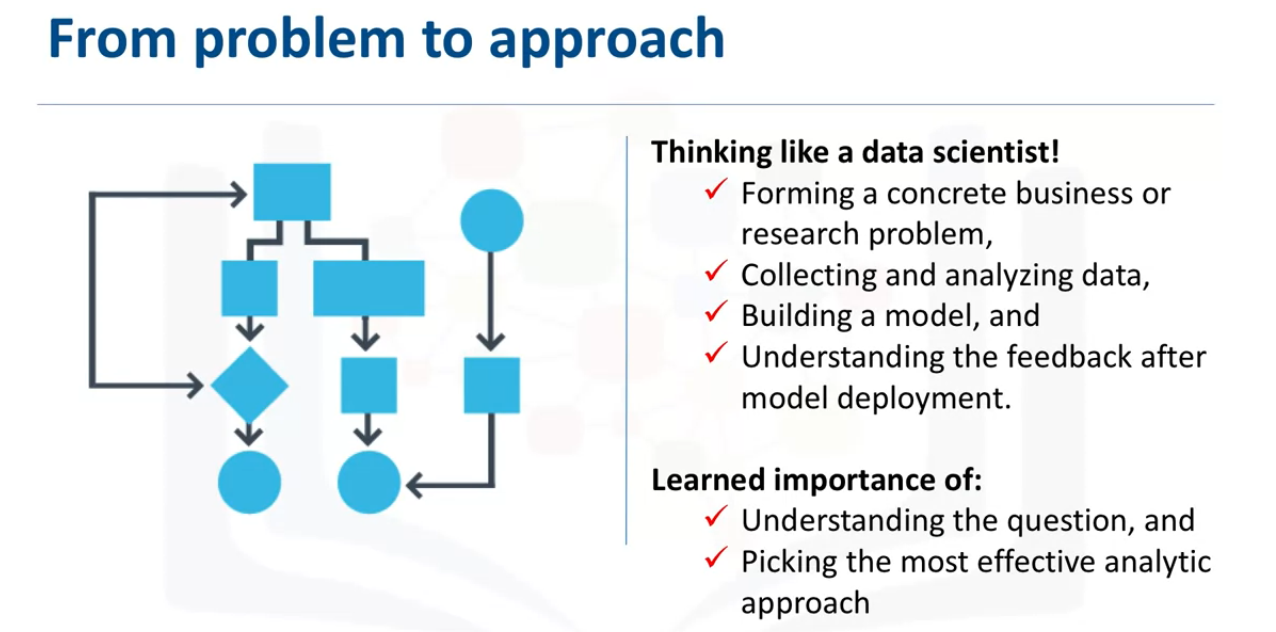
Second, congestive heart failure patients receiving intervention would be tracked and their re-admission outcomes recorded. Third, the intervention would then be measured to determine how effective it was in reducing re-admissions. For ethical reasons, congestive heart failure patients would not be split into controlled and treatment groups. Instead, readmission rates would be compared before and after the implementation of the model to measure its impact.



After the deployment and feedback stages, the impact of the intervention program on re-admission 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. If you recall, data collection was initially deferred because the pharmaceutical data was not readily available at the time. But after feedback and practical experience with the model, it might be determined that adding that data could be worth the investment of effort and time. We also have to allow for the possibility that other refinements might present themselves during the feedback stage.



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.

\