**Problem to Approach Business Understanding!**

Has this ever happened to you? You've been called into a meeting by your boss, who makes you aware of an important task one with a very tight deadline that absolutely has to be met. 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. Later that afternoon, however, after you've spent some time examining the various issues at play, you realize that you need to ask several additional questions in order to truly accomplish the task. Unfortunately, the boss won't be available again until tomorrow morning. Now, with the tight deadline still ringing in your ears, you start feeling a sense of uneasiness. So, what do you do? Do you risk moving forward or do you stop and seek clarification. Data science methodology begins with spending the time to seek clarification, to attain what can be referred to as a business understanding. Having this understanding is placed at the beginning of the methodology because getting clarity around 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 vital because it ultimately directs the analytic approach that will be needed to address the question. All too often, much effort is put into answering what people THINK is the question, and while the methods used to address that question might be sound, they don't help to solve the actual problem. Establishing a clearly defined question starts with understanding the GOAL of the person who is asking the question. For example, if a business owner asks: "How can we reduce the costs of performing an activity?" We need to understand, is the goal to improve the efficiency of the activity? Or is it to increase the businesses profitability? Once the goal is clarified, the next piece of the puzzle is to figure out 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. Depending on the problem, different stakeholders will need to be engaged in the discussion to help determine requirements and clarify questions. So now, let's look at the case study related to applying "Business Understanding" In the Case Study, the question being asked is: What is the best way to allocate the limited healthcare budget to maximize its use in providing quality care? This question is one that became a hot topic for an American healthcare insurance provider. As public funding for readmissions was decreasing, this insurance company was at risk of having to make up for the cost difference, which could potentially increase rates for its customers. Knowing that raising insurance rates was not going to be a popular move, the insurance company sat down with the health care authorities in its region and brought in IBM data scientists to see how data science could be applied to the question at hand. Before even starting to collect data, the goals and objectives needed to be defined. After spending time to determine the goals and objectives, the team prioritized "patient readmissions" as an effective area for review. With the goals and objectives in mind, it was found that approximately 30% of individuals who finish rehab treatment would be readmitted to a rehab center within one year; and that 50% would be readmitted within five years. After reviewing some records, it was discovered that the patients with congestive heart failure were at the top of the readmission list. It was further determined that a decision-tree model could be applied to review this scenario, to determine why this was occurring. To gain the business understanding that would guide the analytics team in formulating and performing their first project, the IBM Data scientists, proposed and delivered an on-site workshop to kick things off. The key business sponsors involvement throughout the project was critical, in that the sponsor:

1. Set overall direction 2. Remained engaged and provided guidance 3. Ensured necessary support, where needed. Finally, four business requirements were identified for whatever model would be built.

Namely:

1. Predicting readmission outcomes for those patients with Congestive Heart Failure

2. Predicting readmission risk

3. Understanding the combination of events that led to the predicted outcome

4. Applying an easy-to-understand process to new patients, regarding their readmission risk.

**From Understanding to Preparation 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?**

Let's apply the data understanding stage of our methodology, to the case study we've been examining. 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. For example, for a categorical variable that has too many distinct values to be informative in a model, the histogram would help them decide how to consolidate values. 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.

**From Understanding to Preparation Data Preparation -Concepts!**

In a sense, data preparation is similar to washing freshly picked vegetables insofar as unwanted elements, such as dirt or imperfections, are removed. Together with data collection and data understanding, data preparation is the most time-consuming phase of a data science project, typically taking seventy percent and even up to even ninety percent of the overall project time. Automating some of the data collection and preparation processes in the database, can reduce this time to as little as 50 percent. This time savings translates into increased time for data scientists to focus on creating models To continue with our cooking metaphor, we know that the process of chopping onions to a finer state will allow for its flavours to spread through a sauce more easily than that would be the case if we were to drop the whole onion into the sauce pot. Similarly, 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. Feature engineering is critical when machine learning tools are being applied to analyze 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 needs to know what they're looking for within their data set 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.

**Understanding to Preparation Data Preparation**

Case Study!

In a sense, data preparation is similar to washing freshly picked vegetables insofar as unwanted elements, such as dirt or imperfections, are removed. So now, let's look at the case study related to applying Data Preparation concepts. In the Case Study, an important first step in the data preparation stage was to actually define congestive heart failure. This sounded easy at first but defining it precisely, was not straightforward. First, the set of diagnosis-related group codes needed to be identified, as congestive heart failure implies certain kinds of fluid buildup. We also need to consider that congestive heart failure is only one type of heart failure. Clinical guidance was needed to get the right codes for congestive heart failure. The next step involved defining the re-admission criteria for the same condition. The timing of events needed to be evaluated in order to define whether a particular congestive heart failure admission was an initial event, which is called an index admission, or a congestive heart failure-related re-admission.

Based on clinical expertise, a time period of 30 days was set as the window for readmission relevant for congestive heart failure patients, following the discharge from the initial admission. Next, the records that were in transactional format were aggregated, meaning that the data included multiple records for each patient. Transactional records included professional provider facility claims submitted for physician, laboratory, hospital, and clinical services. Also included were records describing all the diagnoses, procedures, prescriptions, and other information about in-patients and out-patients. A given patient could easily have hundreds or even thousands of these records, depending on their clinical history. Then, all the transactional records were aggregated to the patient level, yielding a single record for each patient, as required for the decision-tree classification method that would be used for modeling.

As part of the aggregation process, many new columns were created representing the information in the transactions. For example, frequency and most recent visits to doctors, clinics and hospitals with diagnoses, procedures, prescriptions, and so forth. Co-morbidities with congestive heart failure were also considered, such as diabetes, hypertension, and many other diseases and chronic conditions that could impact the risk of re-admission for congestive heart failure. During discussions around data preparation, a literary review on congestive heart failure was also undertaken to see whether any important data elements were overlooked, such as co-morbidities that had not yet been accounted for. The literary review involved looping back to the data collection stage to add a few more indicators for conditions and procedures.

Aggregating the transactional data at the patient level, meant merging it with the other patient data, including their demographic information, such as age, gender, type of insurance, and so forth. The result was the creation of one table containing a single record per patient, with many columns representing the attributes about the patient in his or her clinical history. These columns would be used as variables in the predictive modeling. Here is a list of the variables that were ultimately used in building the model. The dependent variable, or target, was congestive heart failure readmission within 30 days following discharge from a hospitalization for congestive heart failure, with an outcome of either yes or no.

The data preparation stage resulted in a cohort of 2,343 patients meeting all of the criteria for this case study. The cohort was then split into training and testing sets for building and validating the model, respectively.

**From Modeling to Evaluation 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! This portion of the course is geared toward answering two key questions:

* First**, 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. In John Rollin's descriptive Data Science Methodology, the framework is geared to do 3 things:

* Understand the question at hand,
* Select an analytic approach or method to solve the problem, and
* Obtain, understand, prepare, and model the data.

The end goal is to move the data scientist to a point where a data model can be built to answer the question. With dinner just about to be served and a hungry guest at the table, the key questions is: Have I made enough to eat? Let's hope so. 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. The more people that benefit from the outcomes of this practice, the further the field will develop. This ends the Modeling to Evaluation section of this course, in which we reviewed the key concepts related to modeling. Thanks for watching!

**From Modeling to Evaluation Modeling Case Study!**

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! Now, let's apply the Case Study to the modeling stage within the data science methodology. Here, we'll discuss one of the many aspects of model building, in this case, parameter tuning to improve the model. With a prepared training set, the first decision tree classification model for congestive heart failure readmission can be built. We are looking for patients with high-risk readmission, so the outcome of interest will be congestive heart failure readmission equals "yes". In this first model, overall accuracy in classifying the yes and no outcomes was 85 %. This sounds good, but it represents only 45 % of the "yes"; the actual readmissions are correctly classified, meaning that the model is not very accurate. The question then becomes: 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. Think of it like this:

When a true, non-readmission is misclassified, and action is taken to reduce that patient's risk, the cost of that error is the wasted intervention. A statistician calls this a type-one error, or a false-positive. But when a true readmission is misclassified, and no action is taken to reduce that risk, then the cost of that error is the readmission and all its attended costs, plus the trauma to the patient. This is a type 2 error, or a false-negative. So we can see that the costs of the two different kinds of misclassification errors can be quite different. For this reason, it's reasonable to adjust the relative weights of misclassifying the yes and no outcomes. The default is 1-to-1, but the decision tree algorithm, allows the setting of a higher value for yes. For the second model, the relative cost was set at 9-to-1. This is a very high ratio, but gives more insight to the model's behavior. This time the model correctly classified 97% of the yes, but at the expense of a very low accuracy on the no, with an overall accuracy of only 49%. This was clearly not a good model. The problem with this outcome is the large number of false-positives, which would recommend unnecessary and costly intervention for patients, who would not have been re-admitted anyway. Therefore, the data scientist needs to try again to find a better balance between the yes and no accuracies. For the third model, the relative cost was set at a more reasonable 4-to-1. This time 68% accuracy was obtained on only yes, called sensitivity by statisticians, and 85% accuracy on the no, called specificity, with an overall accuracy of 81%. This is the best balance that can be obtained with a rather small training set through adjusting the relative cost of misclassified yes and no outcomes parameter. A lot more work goes into the modeling, of course, including iterating back to the data preparation stage to redefine some of the other variables, so as to better represent the underlying information, and thereby improve the model. This concludes the Modeling section of the course, in which we applied the Case Study to the modeling stage within the data science methodology.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | 9-1 | 4-1 |  |
| Accuracy on Yes |  | 97% | 68% |  |
| Accuracy on No |  | 3 | 85% |  |
| Overall accuracy |  | 49% | 81% |  |
|  |  |  |  |  |

**From Modeling to 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. So now, let's go back to our case study so that we can apply the "Evaluation" component within the data science methodology Let's look at one way to find the optimal model through a diagnostic measure based on tuning one of the parameters in model building. Specifically we'll see how to tune the relative cost of misclassifying yes and no outcomes. As shown in this table, 4 models were built 4 with different relative misclassification costs. 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? For budgetary reasons, the risk-reducing intervention could not be applied to most or all congestive heart failure patients, many of whom would not have been readmitted anyway. On the other hand, the intervention would not be as effective in improving patient care as it should be, with not enough high-risk congestive heart failure patients targeted. So, how do we determine which model was optimal? 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. It has since been used in many other fields as well. Today it is commonly used in machine learning and data mining. 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 costs, the ROC curve helped in selecting the optimal model.

**From Deployment to Feedback 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 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. 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 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. This map is an example of a solution deployed through a Cognos application. In this case, the case study was hospitalization risk for patients with juvenile diabetes. Like the congestive heart failure use case, this one used decision tree classification to create a risk model that would serve as the foundation for this application. The map gives an overview of hospitalization risks nationwide, with an interactive analysis of predicted risk by a variety of patient conditions and other characteristics. This slide shows an interactive summary report of risk by patient population within a given node of the model, so that clinicians could understand the combination of conditions for this subgroup of patients. And this report gives a detailed summary on an individual patient, including the patient's predicted risk and details about the clinical history, giving a concise summary for the doctor.

**From Deployment to Feedback Feedback!**

Once in play, 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. Throughout the Data Science Methodology, each step sets the stage for the next. Making the methodology cyclical, ensures refinement at each stage in the game. The feedback process is rooted in the notion that, the more you know, the more that you'll want to know. That's the way John Rollins sees it and hopefully you do too. Once the model is evaluated and the data scientist is confident it'll work, it is deployed and put to the ultimate test: actual, real-time use in the field. So now, let's look at our Case study again, to see how the Feedback portion of the methodologies applied. 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.