Data Science Methodology

IBM Case study

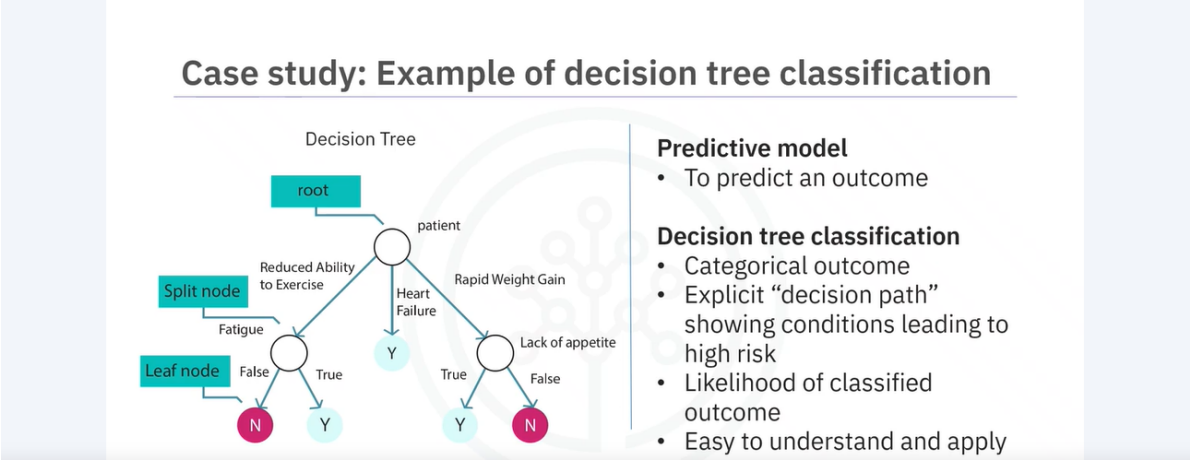
▼ Business Case study

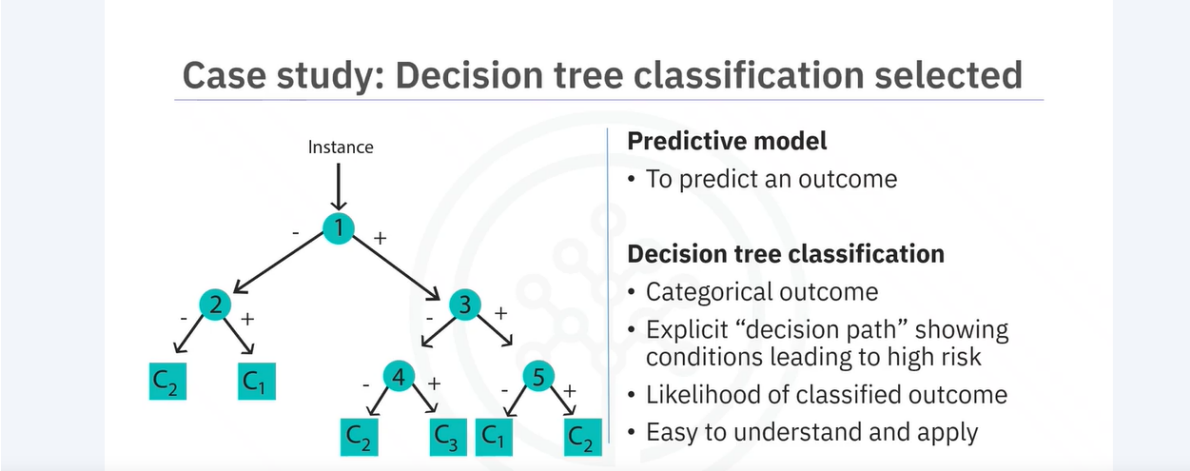
case study related to applying "Business Understanding" In the case study, the question being asked is: What is the best way to allocate the limited healthcare budget to maximize its use in providing quality care? This question is one that became a hot topic for an American healthcare insurance provider. 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: Set overall direction Remained engaged and provided guidance. Ensured necessary support, where needed. Finally, four business requirements were identified for whatever model would be built. Namely: Predicting readmission outcomes for those patients with Congestive Heart Failure Predicting readmission risk. Understanding the combination of events

that led to the predicted outcome Applying an easy-to-understand process to new patients, regarding their readmission risk.

▼ Analytical approach

case study related to applying Analytic Approach. 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.





▼ Data Requirements

In the case study, the first task was to define the data requirements for the decision tree classification approach that was selected. This included selecting a suitable patient cohort from the health insurance providers member base. In order to compile the complete clinical histories, three criteria were identified for inclusion in the cohort. First, a patient needed to be admitted as in-patient within the provider service area, so they'd have access to the necessary information. Second, they focused on patients with a primary diagnosis of congestive heart failure during one full year. Third, a 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 also had been diagnosed as having other significant medical conditions, were excluded from the cohort because those conditions would cause higher-than-average re-admission rates and, thus, could skew the results. Then the content, format, and representations of the data needed for decision tree classification were defined. This modeling technique requires one record per patient, with columns representing the variables in the model. To model the readmission outcome, there needed to be data covering all aspects of the patient's clinical history. This content would include admissions, primary, secondary, and tertiary diagnoses, procedures, prescriptions, and other services provided either during hospitalization or throughout patient/doctor visits. Thus, a particular patient could have thousands of records, representing all their related attributes. To get to the one record per patient format, the data scientists rolled up the transactional records to the patient level, creating a

number of new variables to represent that information. This was a job for the data preparation stage, so thinking ahead and anticipating subsequent stages is important.

def: A **cohort** is a group of individuals who share a common characteristic or experience within a defined period. In this context, it refers to a specific set of patients selected for study based on certain criteria.

▼ Data Collection

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 congestive heart failure patients. For this case study, certain drug information was also needed, but that data source was not yet integrated with the rest of the data sources. This leads to an important point: 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. As it turned out though, they were able to build a reasonably good model without this drug information. 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.

▼ Data Understanding

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 Preperation

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 needed 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.

▼ Modelling

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 I 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 II 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 behaviour. 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.

▼ Evaluation

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, four models were built with four 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 cost, the ROC curve helped in selecting the optimal model.

▼ CRISP-DM(not casestudy)

Welcome to an Introduction to CRISP-DM. After watching this video, you'll be able to, define CRISP-DM, list and describe the six stages of the CRISP-DM model, and explain what happens after the final CRISP-DM stage. CRISP-DM, which stands for Cross-Industry Standard Process for Data Mining, is an industry-proven way to guide your data mining efforts. CRISP-DM is an iterative data mining mode and is a comprehensive methodology for data mining projects which provides a structured approach to guide data-driven decision making. As a data methodology, a study of the CRISP-DM model includes six data mining stages, their descriptions and provides explanations of the relationships between tasks and stages. And as a process model, CRISP-DM provides high-level insights into the data mining cycle. Like other data mining science methodologies, CRISP-DM requires flexibility at each stage, and communication with peers, management, and stakeholders to keep the project on track. After any of the following six stages, data scientists might need to revisit an earlier stage and make changes. The business understanding stage is the most important because this stage sets and outlines the intentions of the data analysis project. This stage is common to both John Rollins data science methodology, and CRISP-DM methodology.

This stage requires communication and clarity to overcome stakeholders'differing objectives, biases, and information related modalities. Without a clear concise and complete understanding of the business problem and project goals, the project effort will waste time and resources. Then, CRISP-DM combines the stages of data requirements, data collection, and data understanding from Johns Rollins methodology outline into a single data understanding stage. During this stage, data scientists decide on data sources and acquire data. Next during the data preparation stage, data scientists transform the collected data into a usable data subset and determine if they need more data. With data collection complete, data scientists select a dataset and address questionable missing or ambiguous data values. Data preparation is common to foundational data methodology in CRISP-DM. The modeling stage fulfills the purpose of data mining and creates data models that reveal patterns and structures within the data. These patterns and structures provide knowledge and insights that address the stated business problem and goals. Data scientists select models based on subsets of the data and adjust the models as needed. Model selection is an art and science. Both foundational methodology and CRISP-DM focus on creating knowledge information that has meaning and utility. During the evaluation stage, data scientists test the selected model. Data scientists usually prepare a pre-selected test to run the trained model. The test platform sees the data as new and data scientists then assess the model's effectiveness. These testing results determine the model's efficacy and foreshadow the model's role in the next and final stage. Finally, during the deployment stage, data scientists and stakeholders use the model on new data outside of the scope of the dataset. New interactions during this stage might reveal the new variables and need for a different dataset and model. Remember that the CRISP-DM model is iterative and cyclical, deployment results might initiate revisions to the business needs and actions, the model and data, or any combination of these items. After completing all six stages, you'll have another business understanding meeting with the stakeholders to discuss the results. In CRISP-DM, the stage is not named. However, in John Rollins Data Science methodology model, the stage is explicitly named the Feedback stage. You'll continue the CRISP-DM process stages until the stakeholders, management, and you agree that the data model and its analysis provide the stakeholder with the answers they need to resolve their business problems and attain their business goals. In this video, you

learned that CRISP-DM stands for Cross-Industry Standard Process for Data Mining. The CRISP-DM model consolidates the steps outlined in foundational data methodology into the following six stages, business understanding, data understanding, data preparation, modeling, evaluation, and deployment. You'll continue the CRISP-DM process until the stakeholders, management, and you agree that the data model and its analysis answer the business questions.