2Week 2

From Data Understanding and Data Preparation

Data Understanding

Data Understanding

All Activities in the process of constructing the dataset.

It means prepare or clean data.

Is the data collected representative of the problem being solved?

Case Study:

For understanding data related to CHF we need to apply descriptive statistics

Univariate Stats - on each column

mean, median so on

Pairwise correlations - relation between two variables

if the correlation is high they are redundant.

Histogram - understand distibutions

Helps how to consolidate too many distinct categorical values.



Looking at data quality

Data quality (Univariates and histograms can be used.)

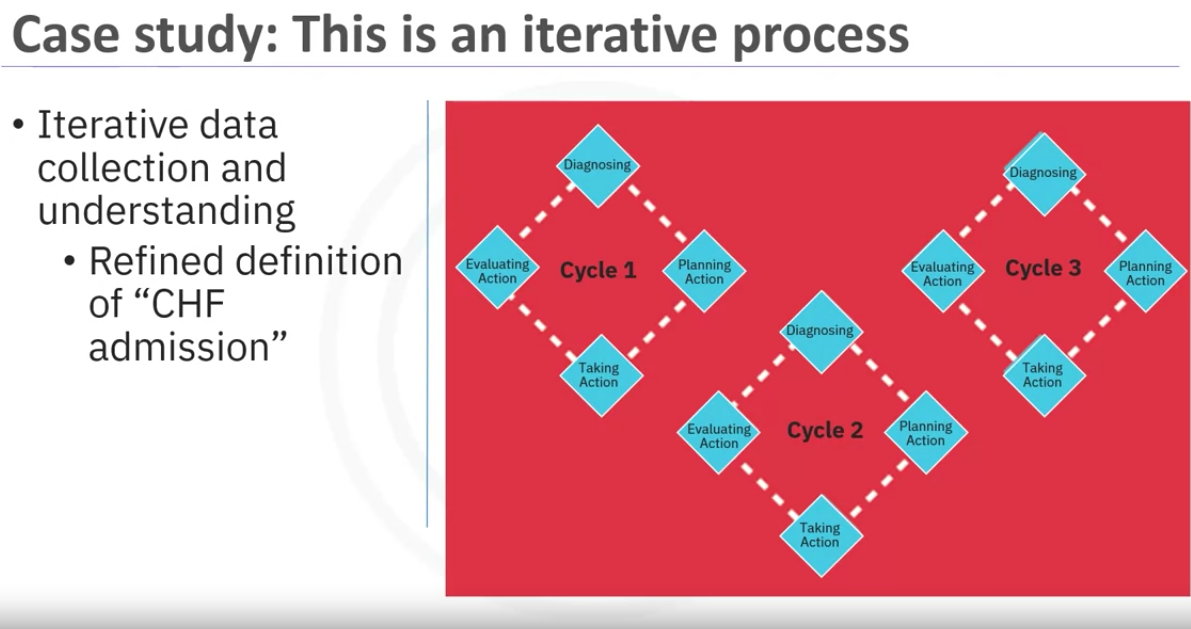
missing values (change or drop value)

Invalid or misleading values (age cannot have a value 999 which generally means data is missing but we have to correct it. if not we get skewed results which are not correct) reducing quality drastically.

The whole process is iterative.

The initial definition of CHF admision was not satisfying. upon more clinical experience the definition was redefined.

This means the process has to revisit the data collection stage and follow along from the beginning by adding second and third diagnosis and building a better definition of CHF admission.



Data Preparation-Concepts

Data Preparation - Cleaning:

Similar to cleaning dirt on veggies

Most time consuming.(70-90%)

Automating will reduce the time taken

Transformation:

Converting parts of data into specific format where it will be easy to make use of the data for the project.

What are the ways in which data is prepared?

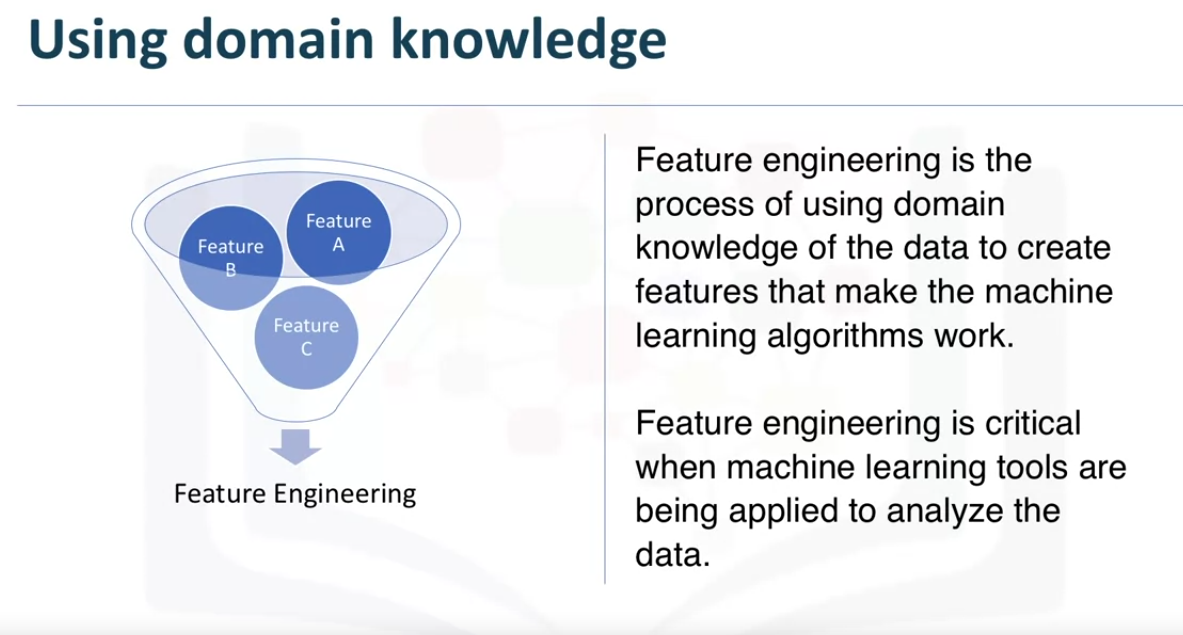
Finding missing and invalid values

remve duplicates

Formatting

Feature Engineering is also part of data transformation

Features are important characterstics in data that helps ML process.



When working with text, text analysis steps for coding the data are required to be able to manipulate the data.

The data scientists need to know what they are looking for within their dataset to address the question.

Text analysis is critical to ensure that the proper groupings are set and that the programming is not overlooking what is hidden within.

Reading: Correlation

Please note that the phrase "literary review" in the next video:

Data Preparation - Case Study, is supposed to be "literature review"

Case Study:

Define CHF - sounds easy but not but defining it was not straight forward

First, the set of diagnosis-related group codes needed to be identified as CHF implies certain kind of fluid buildup.

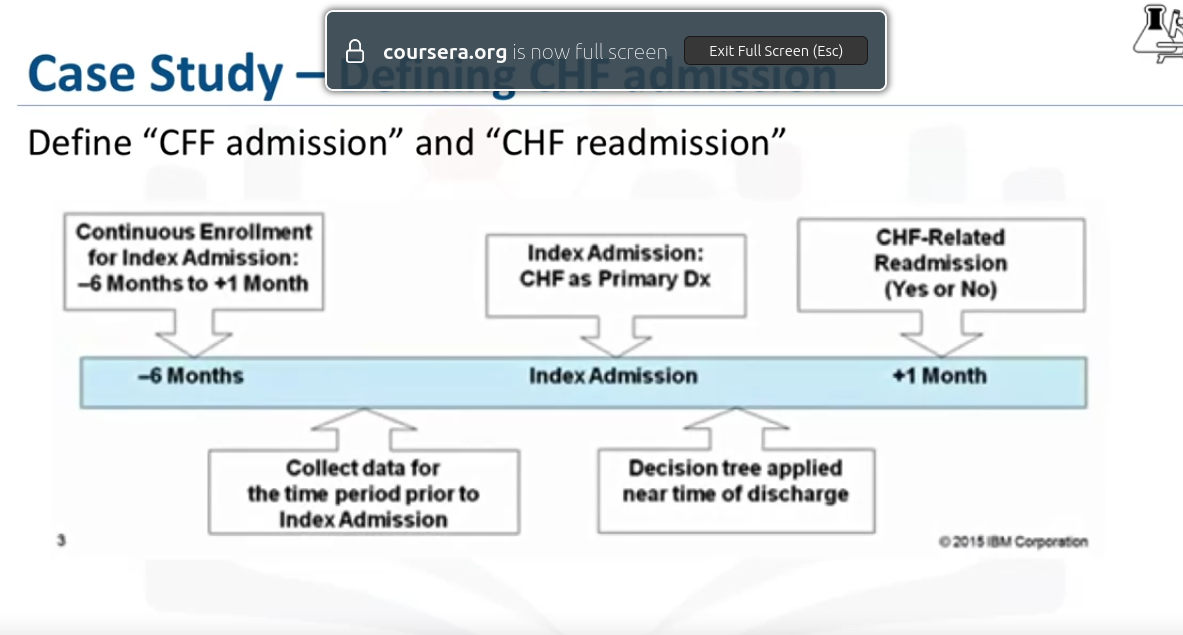
Also CHF is one type of heart failures.

Clinical guidance was needed to get the right codes for CHF.

Next Defining re-admission criteria:

The timing of events needed to be evaluated in order to define whether a particular CHF admission was first time(index admission) or CHF related re-admission.

Based on clinical expertise, a time period of 30 days was set as window for readmission.



Next transactional records were aggregated

data had multiple records for each patient and it was combined to one record.

Transactional records had

claims : professional provider, facility, pharmaceutical

Inpatient and outpatient records: diagnoses, procedures, prescriptions

Possibly thousands per patient, depending on clinical history.

All had to be aggregated to single record per patient

create new columns representing the transactional records

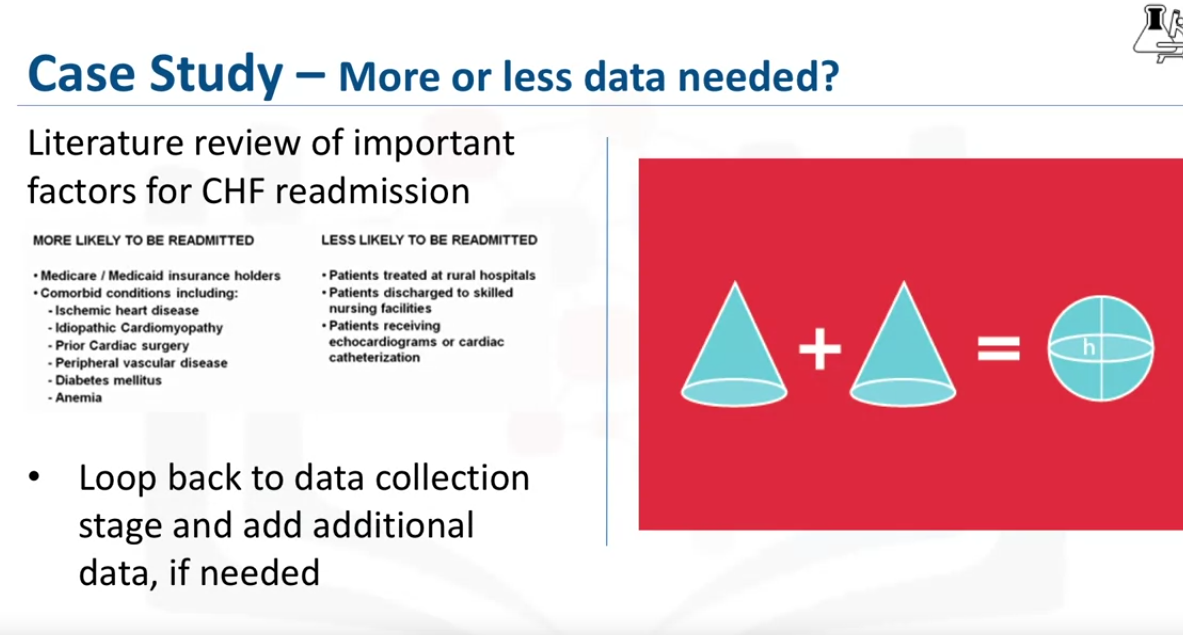
Inpatient and outpatient details: frequency, recency, diagnoses/length of stay, procedures, prescriptions

Comorbidities with CHF: like diabetes, hypertension, so on.

This was needed for Decision Tree Classification.

More or Less Data is needed?

A literary review on CHF review was conducted to check if anyting was missed when collecting all the data and preparing it.



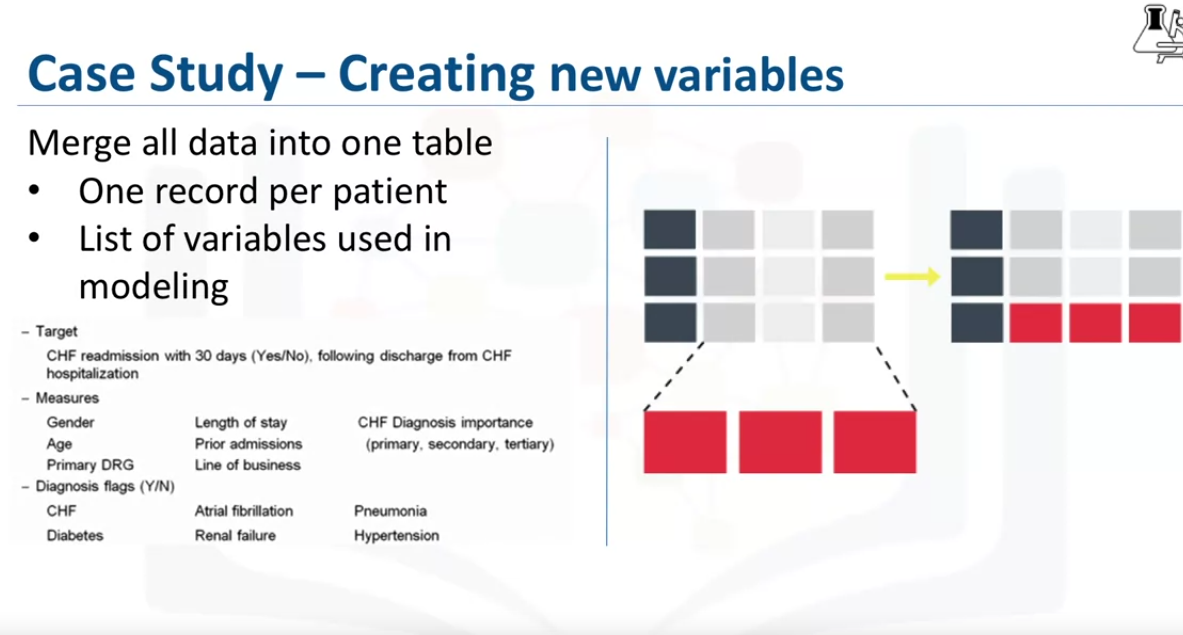
Aggregating meant merging tables at transactional level with demographic data and other data.

One table was generated with single record for each patient with many column.

The columns represented details related to the differnt properties.

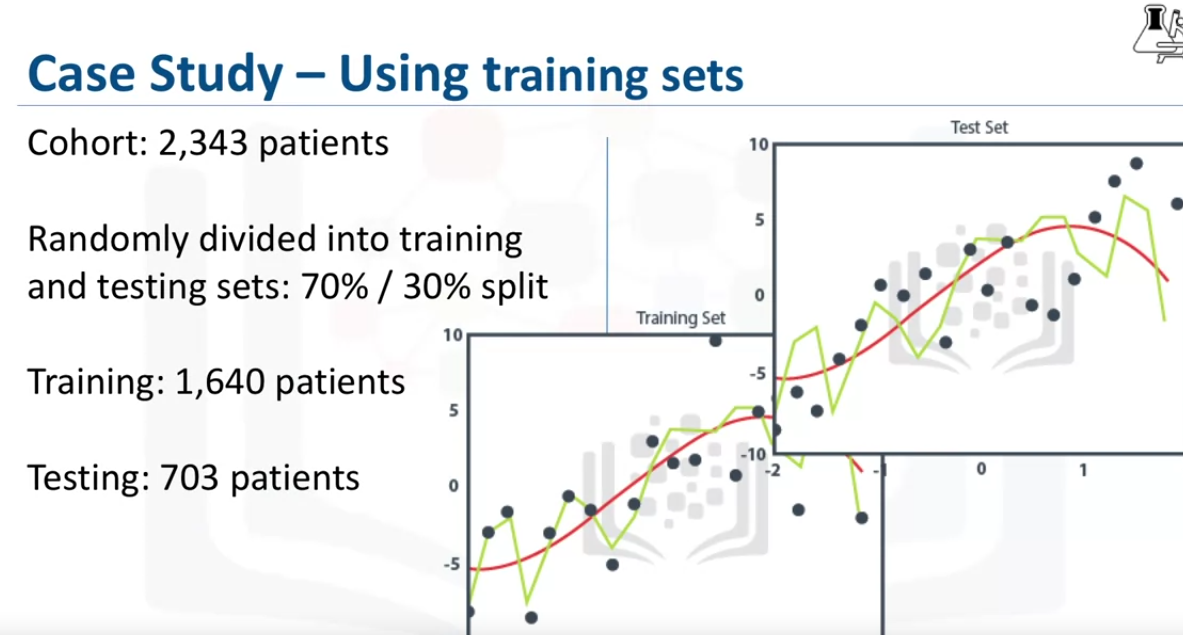
The final goal was to predict a target column wich labels Yes or no sho if there might be readmission or no.

List of variables used in the casestudy:



Using Training set

The data was split into training and test data for the ML Decision Tree.



Lab: From Understading to preperation

Summary:

Glossary

Modeling to Evaluation

Concepts:

Modeling - Sampling the data

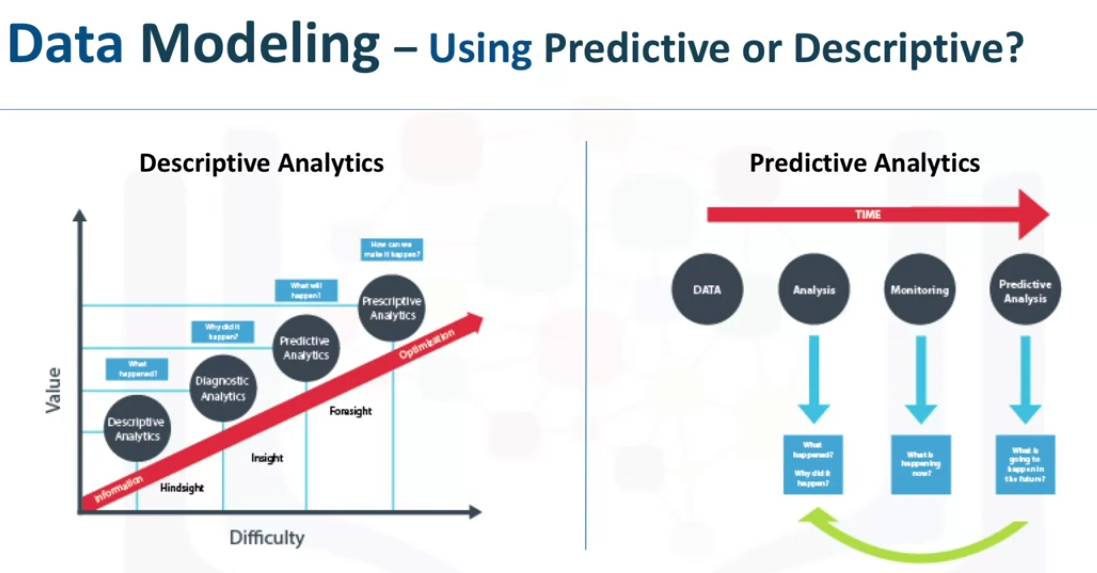
In what way can the data be visualized to get the answer that is required?

This section answes two question

What is the Purpose of Data Modeling?

What are the key characterstics of this process?

Data modelling focuses on developing models that are either descriptive or predictive analysis



Descriptive model example would be like

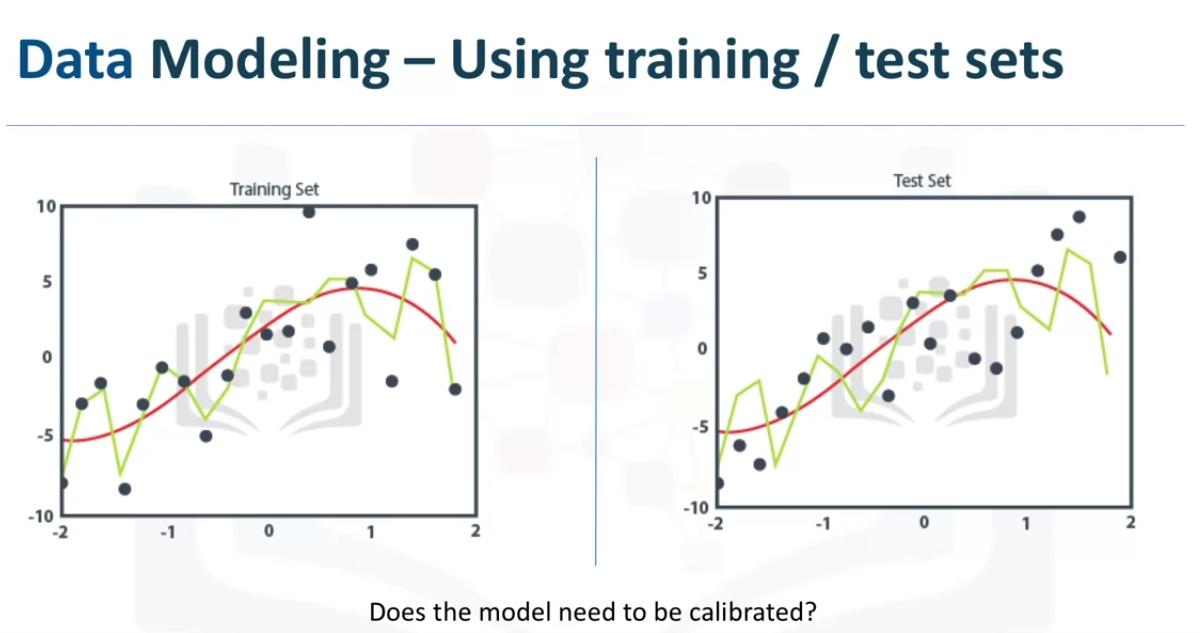
if a person did a certain thing they are likely to prefer something related.

Predictive Models on the other hand:

Tries to yield yes or no outcomes.

The models can stats driven or ML driven

Data is divided into training and test data.



Training Data: Historical data for which the outcomes are already known.

it acts like a gauge to let us know if the model has to be calibrated.

In this stage data scientists will play around with different algorithms to ensure variable in play are actually required.

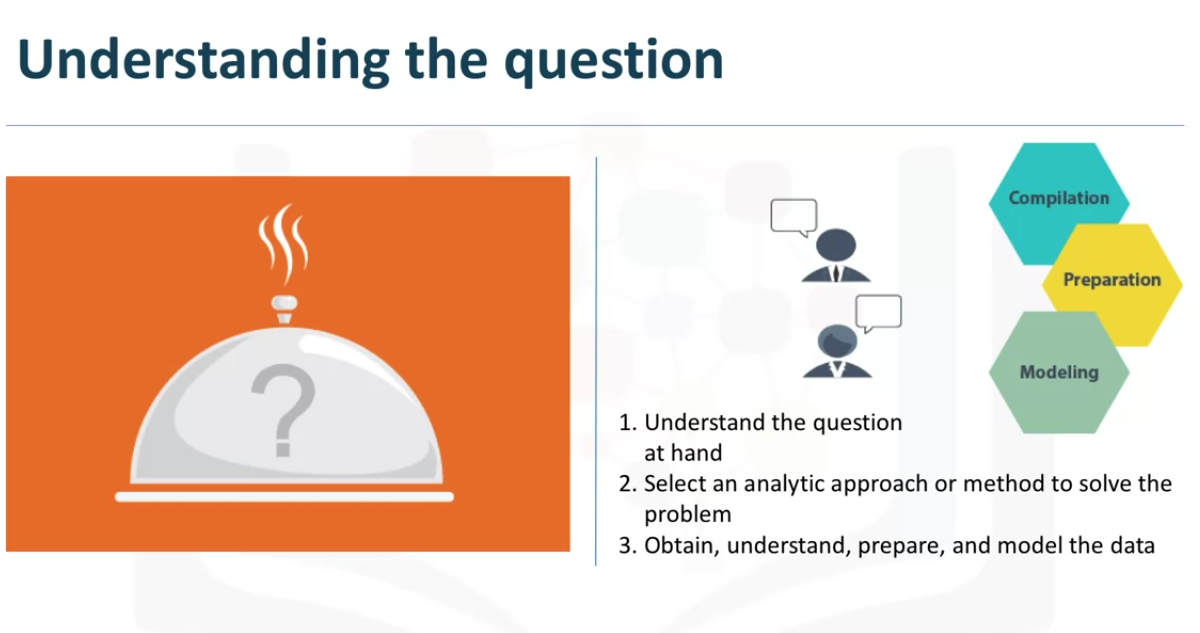
The success of data compilation, preparation and modelling, depends on

1. the understanding of the question
2. Analytical approach selected to solve the problem.

Quality of the data is very important.

Constant refinement is necessary

John Rollin’s Data Science Methodology is geared to do 3 things



After the model is built, we need to ask if the question is answered?(Evaluation stage)

In this stage of methodology, model evaluation, deployment and feedback loops ensure that the model is working properly and generates the relevant answers.

Case Study:

Checking if sauce for the food is correctly seasoned or not.

With a training set picked from the data we got from the Data preparation stage, a first decision tree classification model can be built for CHF readmission .

We are looking for patients with high-risk of readmission, so the outcome of the interest will be an yes for CHF readmission,

In the first model the accuracy was 85%.

This sounds good but it represents only 45% of yes, the actual readmissions are correctly classified meaning that the model is not very accurate.



Although the overall accuracy is 85%, the accuracy of yes is only 45%.

How can the increase accuracy of yes?

The best parameter to adjust here is the relative cost for misclassified yes and no outcomes.

When a true, non-readmission is misclassified, and actionis taken to reduce that patient’s risk , the cost of that error is the wasted intervention.

A stats person calls it a type 1 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 ll the attended costs plus the trauma to the patient

This is type 2 error or false negative.

So its reasonable to adjust the weights of misclassification.

The default is 1 to 1 but the decision tree algorithm allows to set higher value for yes coutcomes.

For the second model the weight has been adjusted to 9 to 1. its a very khigh ratio but give insight to the mode’s behaviour,

now the yes accuracy is 97% but the no accuracy is down to 35%.

Overall accuracy is only 49% which is not good.

lot of false positives which increases the unnecessary costly intervention for patients

The tid model uses 4:1 ratio and the yes accuracy is 68% and no accuracy is 85% giving us a 81% overall accuracy. Which is fairly reasonable.

A lot of iterations can also happen where the data can be redefined and the process restarts.

Evaluation

Access model quality and ensures it meets the initial question or does it need readjustment.

Goes hand in hand with model building.

performed during model development and before deployment to avoid bugs and errors.

Has 2 main phases

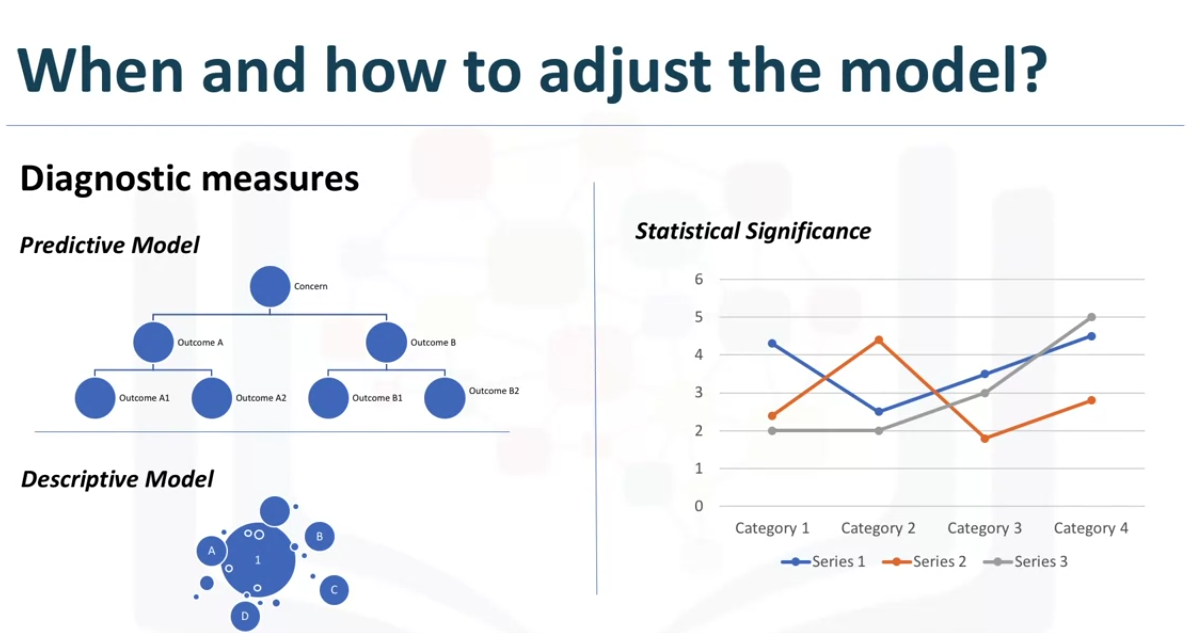
1.Diagnostic measures - Checks if the model is working as intended

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.

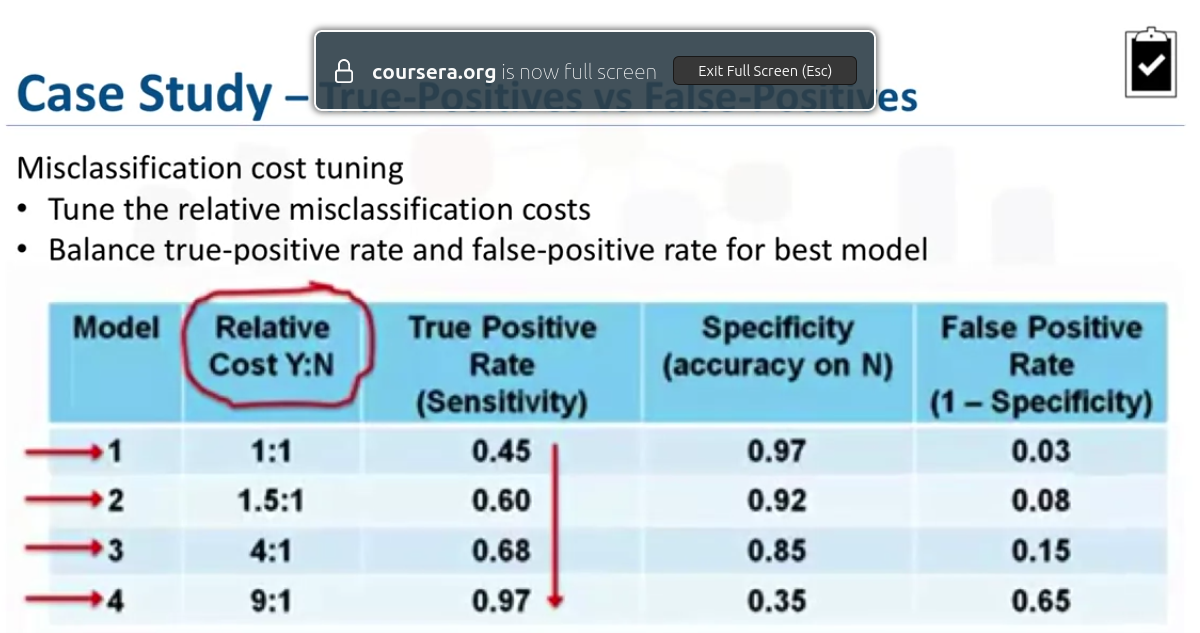
If the model is a descriptive model, one in which relationships are being assessed, then a testing set with known outcomes can be applied, and the model can be refined as needed.

2.Statistical significance - tests if the data is properly handled and interpreted within the model. This is designed to avoid unnecessary second guessing when the answer is revealed.



Case Study:

Lets build model by tuning parameters and and use the diagnostic tool to balacnce true- positive and false- postive.



How do we find the optimal model?

The optimal model is the one that gives the maximum seperation relative the the red base line.

We can see that the 3rd one with 4 to 1 is the best.

ROC = Receiver operating chatacteristic curve. developed during WW2 to detect aircraft on radar.

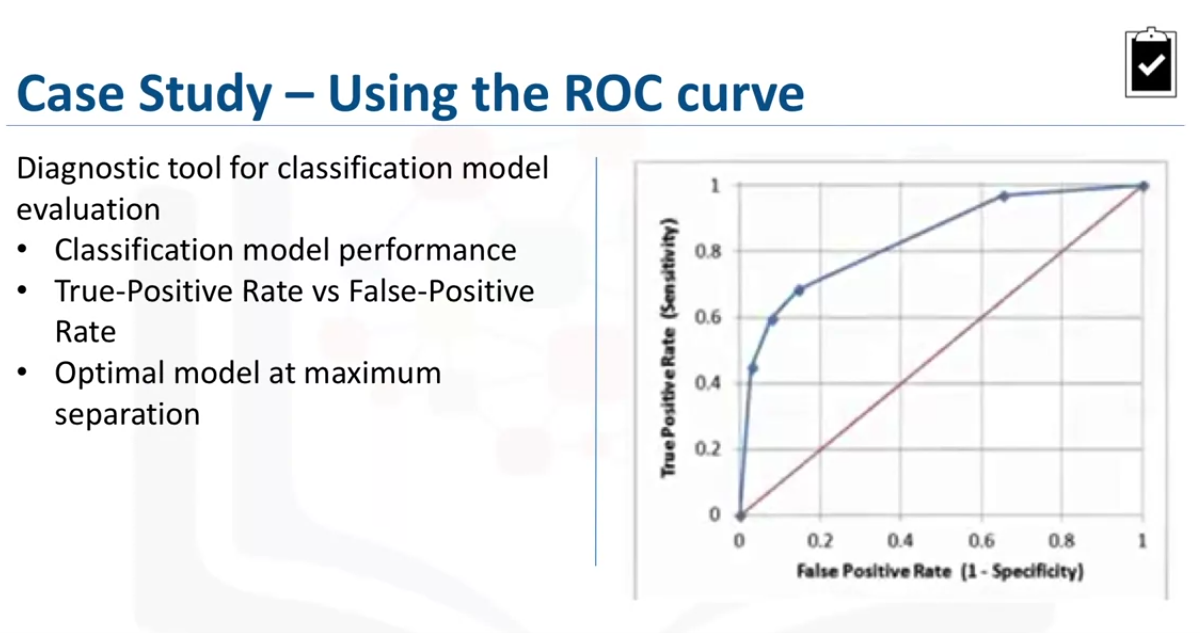
Its commonly used in ML and Data mining.

The ROC is a diagnostic tool that helps to find the optimal 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 true postive rate agains false positive rate for different values of the relative misclassificatin cost, the ROC helps in finding the optimal model.



Week2 lab2

Summary

Glossary