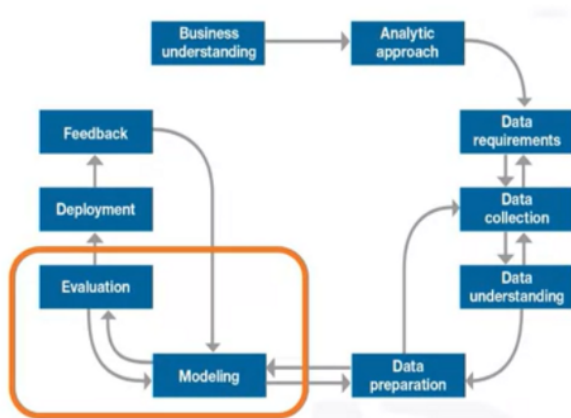




Follow



Welcome to the data science methodology. Till now we have seen all *3 stages of data science methodology* from *Problem to approach, Requirement to collections, Understanding to preparation*. We have discuss amazing example with case study approach if you haven't read this article series read from below links. and already read that go directly with this articles. In this article, You can learn about how to select the model and how to evaluate that model or this model is ready for deployment or not.

Article Series :

1. [Overview of Data Science Methodology](#)
2. [Part-1 Data Science Methodology- From Problem to Approach](#)
3. [Part-2 Data Science Methodology From Requirement to Collection](#)
4. [Part-3 Data Science Methodology From Understanding to Preparation](#)
5. [Part-4 Data Science Methodology From Modelling to Evaluation](#)
6. [Part-5 Data Science Methodology From Deployment to Feedback](#)

#1) Modeling



Modeling is the phase of the methodology of data science in which the data scientist has the opportunity to taste the sauce and determine if it needs more seasoning or if it needs more seasoning !

This part of the course is designed to answer two key questions:

- *First, what is the purpose of data modeling, and*
- *Second, what are the characteristics of this process?*

Data modeling focuses on the development of *descriptive or predictive models*.

Data Modeling – Using Predictive or Descriptive?

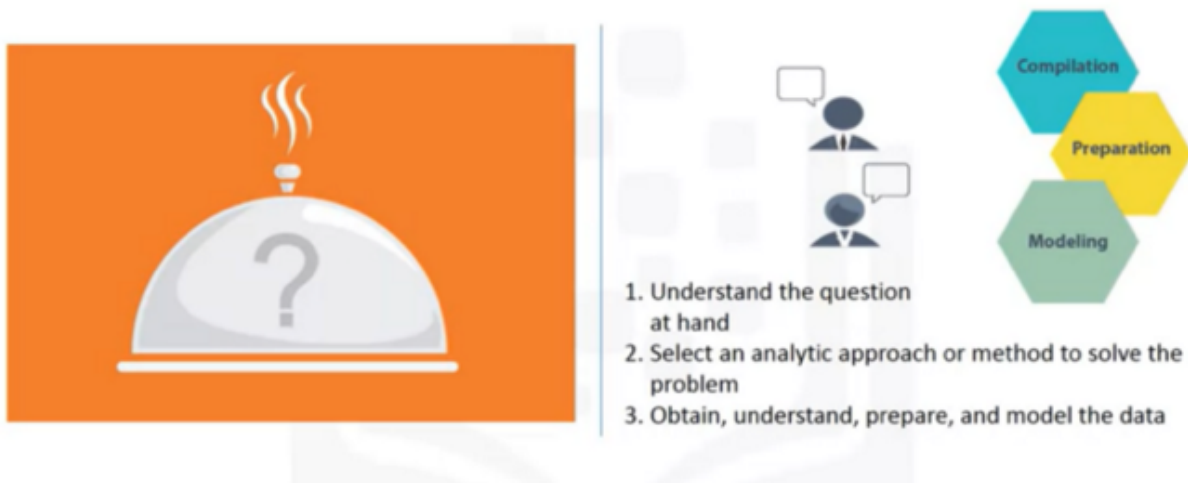


- An example of a descriptive model might be the following: if someone did it, they probably prefer it.
- A **predictive model** attempts to provide **yes / no** results or to **stop / continue**. These models are based on an analytic approach learned either statistically or Machine learning. The Data Scientist will use a training set for predictive modeling.
- *A training set is a set of historical data in which the results are already known.* The training set serves as an *indicator to*

determine if the model needs to be calibrated.

- At this point, the data scientist will use several algorithms to ensure that the variables involved are really needed.

Understanding the question



- The **success of data collection, preparation and modeling depends on an understanding of the problem in question and the appropriate analytical approach.**
- The data support the answer to the question and the quality of the ingredients in the kitchen is the basis of the result.
- Each step requires constant improvements, adjustments and tweaking to ensure the strength of the result.

In the descriptive data science methodology of John Rollins, the framework is designed for three things:

- *First, understand the question that concerns you.*
- *Secondly, choose an analytical approach or method to solve the problem.*
- *Thirdly, obtaining, understanding, preparing and modeling data.*

The ultimate goal is to bring the data scientist to a point where it is possible to create a data model to answer the question.

Was the question answered?



- While dinner is being served and a hungry guest sits at the table, the key question is: have I prepared enough to eat? *We hope that at this stage of the methodology, model evaluation, deployment and feedback cycles of the models will ensure that the response is relevant and near to the result.*

- This relevance is essential for the whole field of data science, as it is a relatively new field and we are interested in the possibilities it offers.
- The more people benefit from the results of this practice, the more the field develops.

Case study:

- The modeling is the phase of the methodology of data science during which the data scientist has the opportunity to taste the sauce and determine if it breaks or if it needs additional seasoning! Now apply the case study to the modeling phase as part of the data science methodology.

Here we will discuss one of the many aspects of model construction, in this case optimizing the parameters to improve the model.

Case Study – Analyzing the 1st model



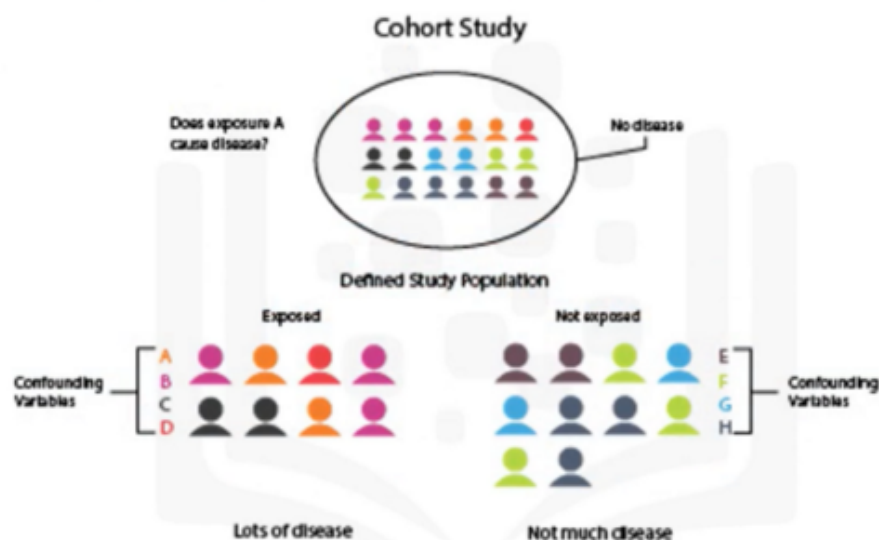
Initial decision tree classification model

- Low accuracy on “Yes” outcome

Model	Relative Cost Y:N	Overall Accuracy (% correct Y & N)	Sensitivity (Y accuracy)	Specificity (N accuracy)
→ 1	1:1	85%	45%	97%
2	9:1	49%	97%	35%
3	4:1	81%	68%	85%

- With a set of prepared training data, it is possible to construct the first classification model of the decision tree for congestive readmission for heart failure. We are looking for patients with high risk readmission. The result that will interest us will be a congestive readmission for heart failure equivalent to “yes”. In this first model, the overall accuracy of the classification of the results was 85% and not 85%. It sounds good, but represents only 45% of the “yes”. Actual readmission are ranked correctly, which means that the model is not very accurate.
- The question is : *how to improve the accuracy of the model to predict the outcome itself?* For the classification of the decision tree, the best parameter to adjust is the relative cost of the results yes and not classified incorrectly.

Case Study – How to improve the model?



- *Think of it this way:* When a true *non-readmission* is *misclassified* and actions are taken to reduce the risk of *this patient*, the *cost of this error* is a *wasted intervention*.

		True condition			
Total population		Condition positive	Condition negative	$Prevalence = \frac{\sum \text{Condition positive}}{\sum \text{Total population}}$	$Accuracy (ACC) = \frac{\sum \text{True positive} + \sum \text{True negative}}{\sum \text{Total population}}$
Predicted condition	Predicted condition positive	True positive , Power	False positive , Type I error	$Positive \text{ predictive value (PPV), Precision} = \frac{\sum \text{True positive}}{\sum \text{Predicted condition positive}}$	$False \text{ discovery rate (FDR)} = \frac{\sum \text{False positive}}{\sum \text{Predicted condition positive}}$
	Predicted condition negative	False negative , Type II error	True negative	$False \text{ omission rate (FOR)} = \frac{\sum \text{False negative}}{\sum \text{Predicted condition negative}}$	$Negative \text{ predictive value (NPV)} = \frac{\sum \text{True negative}}{\sum \text{Predicted condition negative}}$
		$True \text{ positive rate (TPR), Recall, Sensitivity, probability of detection} = \frac{\sum \text{True positive}}{\sum \text{Condition positive}}$	$False \text{ positive rate (FPR), Fall-out, probability of false alarm} = \frac{\sum \text{False positive}}{\sum \text{Condition negative}}$	$Positive \text{ likelihood ratio (LR+)} = \frac{TPR}{FPR}$	$Diagnostic \text{ odds ratio (DOR)} = \frac{LR+}{LR-}$
		$False \text{ negative rate (FNR), Miss rate} = \frac{\sum \text{False negative}}{\sum \text{Condition positive}}$	$Specificity (SPC), Selectivity, True negative rate (TNR) = \frac{\sum \text{True negative}}{\sum \text{Condition negative}}$	$Negative \text{ likelihood ratio (LR-)} = \frac{FNR}{TNR}$	
				$F_1 \text{ score} = \frac{1}{\frac{1}{Recall} + \frac{1}{Precision}}$	

- A statistician calls this a **Type I error** or a **false positive**. But *when a real readmission is misclassified and no action is taken to reduce this risk*, the cost of such an error is *readmission* and all associated costs, as well as trauma to the patient.
- It's a **Type II error** or a **false negative**. Then *we can see that the costs of the two different types of incorrect classification errors* can be very different. For this reason, it is reasonable to adjust the relative weights of the incorrect classification of the results yes and no.
- The default is between 1 and 1, but the decision tree algorithm allows you to set a higher value for yourself.

Case Study – Analyzing the 2nd model



Second model

- High accuracy on “Yes” but poor on “No”

Model	Relative Cost Y:N	Overall Accuracy (% correct Y & N)	Sensitivity (Y accuracy)	Specificity (N accuracy)
→ 1	1:1	85%	45%	97%
→ 2	9:1	49%	97%	35%
3	4:1	81%	68%	85%


- For the **second model**, the relative cost was set at **9/1**. This report is very high, but provides more information about the behavior of the model. This time, the **97% model worked well**, but at a **very low cost, with a general accuracy of only 49%**. Obviously, **this is not a good model**.
- The **problem** with this result is the **large number of false positives, suggesting unnecessary and costly interventions** for patients that have never been re-admitted.
- Therefore, the data scientist must try again to get a better balance between the **yes and no data**.

Case Study – Analyzing the 3rd model



Third model

- Better balance on “Yes” and “No” accuracy



Model	Relative Cost Y:N	Overall Accuracy (% correct Y & N)	Sensitivity (Y accuracy)	Specificity (N accuracy)
→ 1	1:1	85%	45%	97%
→ 2	9:1	49%	97%	35%
→ 3	4:1	81%	68%	85%

- For the **third model**, the relative cost was set to a more reasonable **4: 1 ratio**. This time, **68% was obtained yes**, but statistician called it **sensitivity**, and **85% accuracy for the no**, called **specificity**. , with an overall **accuracy of 81%**.
- This is the best balance that can be achieved with a relatively limited training set of workouts by adjusting the relative cost of the misclassified yes and no result parameters. Of course, modeling requires much more work, including an iteration in the data preparation phase, to redefine some of the other variables to better represent the underlying information and thus improve the model.

#2) Model Evaluation





Evaluation

- *Does the model used really answer the initial question or does it need to be adjusted?*

A model evaluation goes hand in hand with the creation of models. The modeling and evaluation steps are performed iteratively. The evaluation of the model is carried out during the *development of the model and before deployment.*

- *The evaluation evaluates the quality of the model, but also provides the opportunity to determine if it meets the initial requirements.*

The evaluation answers the question:

- *Does the model used really answer the original question or should it be adapted?*

The evaluation of the model can have two main phases.

When and how to adjust the model?

Diagnostic measures

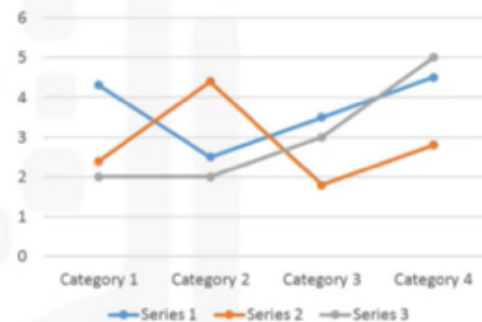
Predictive Model



Descriptive Model



Statistical Significance



- The first phase is the diagnostic measurement phase, *which ensures that the model works as intended*. If the model is predictive, a *decision tree can be used to assess whether the response provided by the model matches the original design*. This allows areas to be displayed where adjustments are required. If the model is a **descriptive model** that evaluates the relationships, a *set of tests with known results can be applied and the model refined as necessary*.
- The second evaluation phase that can be used is the **statistical significance test**. This type of *evaluation* can be applied to the *model to ensure that the model data is processed and interpreted correctly*. This is to *avoid a second unnecessary assumption when the answer is revealed*.

Case study :

- Let's go back to our case study to apply the *Evaluation component* in the data science methodology.

Case Study – Misclassification costs



Misclassification cost tuning

- Tune the relative misclassification costs
- Balance true-positive rate and false-positive rate for best model

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

- Let's look for a way to find the *optimal model* through a diagnostic measurement based on the configuration of one of the model's construction parameters. *We will examine more closely how the relative costs of misclassifying positive and negative results can be adjusted. As shown in this table, four models were constructed with four different relative misclassification costs.*

Case Study – Relative costs



Misclassification cost tuning

- Tune the relative misclassification costs

- Balance true-positive rate and false-positive rate for best model

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

- As we see, each value of this *model construction parameter* increases the true positive rate, or the *sensitivity*, of the accuracy in the prediction yes, to the detriment of a lower accuracy in the prediction no. that is, an increasing rate of false positives.
- The question is, *which model is best based on setting this parameter?* For budgetary reasons, the risk reduction intervention could not be applied to most patients with heart failure, many of whom would not have been readmitted anyway.
- On the other hand, the intervention would not be as effective as it should be to improve patient care, since the number of patients with high-risk heart failure was not enough.

Cost Study – Using the ROC curve

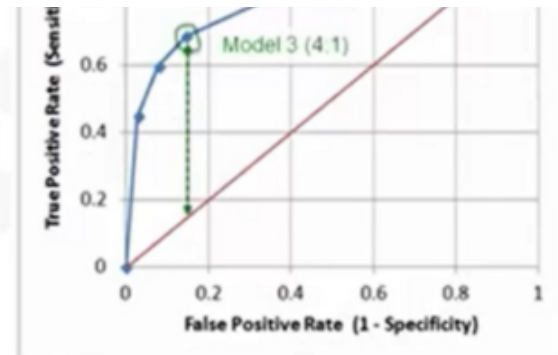


Diagnostic tool for classification model evaluation

- Classification model performance



- True-Positive Rate vs False-Positive Rate
- Optimal model at maximum separation



- *So how do we determine which model was optimal?* As you can see on this image above, the *optimal model is the one that provides the maximum separation between the blue ROC curve and the red baseline.*
- We can see that **model 3**, with a **relative cost of misclassification of 4 to 1**, *is the best of the 4 models.* And if asked, *ROC represents the characteristic operating curve of the receiver*, which was first developed during World War II to detect enemy aircraft on a radar.
- Since then, it has also been used in many other areas. Today, it is commonly used in machine learning and data mining. *The ROC curve is a useful diagnostic tool to determine the optimal classification model.*
- This curve quantifies the performance of a **binary classification model**, declassifying the results yes and no when a discrimination criterion is changed.
- In this case, the criterion is a relative cost of misclassification. By plotting the true positive rate against the false positive rate

for different values of the relative cost of misclassification, the ROC curve facilitated the selection of the optimal model.

Thanks for reading...!!!Happy Learning...!!!

References :

1. <https://www.coursera.org/learn/data-science-methodology>

Machine Learning

Datasciencemethodology

Data Science

Methodology

Modeling

Learn more.

Medium is an open platform where 170 million readers come to find insightful and dynamic thinking. Here, expert and undiscovered voices alike dive into the heart of any topic and bring new ideas to the surface.

[Learn more](#)

Make Medium yours.

Follow the writers, publications, and topics that matter to you, and you'll see them on your homepage and in your inbox. Explore

Share your thinking.

If you have a story to tell, knowledge to share, or a perspective to offer — welcome home. It's easy and free to post your thinking on any topic. Write on Medium

[About](#)

[Help](#)

[Legal](#)