Data Science Methodology 101!

https://courses.cognitiveclass.ai/courses/course-v1:CognitiveClass+DS0103EN+v3/progress

Welcome to Data Science Methodology 101!This is the beginning of a story -one that you'll be telling others about for years to come.It won't be in the form you experience here, but rather through the stories you'll be sharing with others, as you explain how your understanding of a question resulted in an answer that changed the way something was done.Despite the recent increase in computing power and access to data over the last couple of decades, our ability to use the data within the decision making process is either lost or not maximized as all too often, we don't have a solid understanding of the questions being asked and how to apply the data correctly to the problem at hand.Here is a definition of the word methodology.It's important to consider it because all too often there is a temptation to bypass methodology and jump directly to solutions.Doing so, however, hinders our best intentions in trying to solve a problem.This course has one purpose, and that is to share a methodology that can be used within data science, to ensure that the data used in problem solving is relevant and properly manipulated to address the question at hand.The data science methodology discussed in this course has been outlined by John Rollins,a seasoned and senior data scientist currently practicing at IBM. This course is built on his experience and expresses his position on the importance of following a methodology to be successful.In a nutshell, the Data Science Methodology aims to answer 10 basic questions in a prescribed sequence. As you can see from this slide, there are two questions designed to define the issue and thus determine the approach to be used; then there are four questions that will help you get organized around the data you will need, and finally there are four additional questions aimed at validating both the data and the approach that gets designed.Please take a moment now to familiarize yourself with the ten questions, as they will be vital to your success.This course is comprised of several components:There are five modules, each going through two stages of the methodology, explaining the rationale as to why each stage is required. Within the same module, a case study is shared that supports what you have just learned. There's also a hands-on lab, which helps to apply the material. And finally, there are 3 review questions to test your understanding of the concepts.When you're ready, take the final exam.The case study included in the course, highlights how the data science methodology can be applied in context.It revolves around the following scenario: There is a limited budget for providing healthcare to the public. Hospital readmissions for recurring problems can be seen as a sign of failure in the system to properly address the patient condition prior to the initial patient discharge.The core question is: What is the best way to allocate these funds to maximize their use in providing quality care? As you'll see, if the new data science pilot program is successful, it will deliver better patient care by giving physicians new tools to incorporate timely, data-driven information into patient care decisions.The case study sections display these icons at the top right hand corner of your screen to help you differentiate theory from practice within each module.A glossary of data science terms is also provided to assist with clarifying key terms used within the course.While participating in this course, if you come across challenges, or have questions,then please explore the discussion and wiki sessions.So, now that you're all set, adjust your headphones and let's get started!

Welcome!
Welcome! (4:01)
No problem scores in this section
About this course
General Information
No problem scores in this section
Learning Objectives

No problem scores in this section

Syllabus

No problem scores in this section

Grading Scheme

No problem scores in this section

Change Log

No problem scores in this section

Copyrights and Trademarks

No problem scores in this section

Module 1 - From Problem to Approach

Learning Objectives

No problem scores in this section

Business Understanding (5:02)

No problem scores in this section

Analytic Approach (3:23)

No problem scores in this section

Lab - From Problem to Approach

No problem scores in this section

Graded Review Questions 0 of 3 possible points (0/3) 0%

Review Questions

Module 2 - From Requirements to Collection

Learning Objectives

No problem scores in this section

Data Requirements (3:28)

No problem scores in this section

Data Collection (2:54)

No problem scores in this section

Lab - From Requirements to Collection

No problem scores in this section

Graded Review Questions 0 of 3 possible points (0/3) 0%

Review Questions

Module 3 - From Understanding to Preparation

Learning Objectives

No problem scores in this section

Data Understanding (3:16)

No problem scores in this section

Data Preparation (7:16)

No problem scores in this section

Lab - From Understanding to Preparation

No problem scores in this section

Graded Review Questions 0 of 3 possible points (0/3) 0%

Review Questions

Module 4 - From Modeling to Evaluation

Learning Objectives

No problem scores in this section

Modeling (6:11)

No problem scores in this section

Evaluation (3:57)

No problem scores in this section

Lab - From Modeling to Evaluation

No problem scores in this section

Graded Review Questions 0 of 3 possible points (0/3) 0%

Review Questions

Module 5 - From Deployment to Feedback

Learning Objectives

No problem scores in this section

Deployment (3:31)

No problem scores in this section

Feedback (3:08)

No problem scores in this section

Graded Review Questions 0 of 3 possible points (0/3) 0%

Review Questions

Course Summary

Course Summary (3:25)

No problem scores in this section

Final Exam

Instructions

No problem scores in this section

Final Exam 0 of 20 possible points (0/20) 0%

Final Exam

Retake Exam

No problem scores in this section

GENERAL INFORMATION

- This course is free.
- It is self-paced.
- It can be taken at any time.
- It can be audited as many times as you wish.
- There is only ONE chance to pass the course, but multiple attempts per question (see the Grading Scheme section for details)

Prerequisites

- Data Science Hands-on with Open Source Tools (DS0105EN)

Recommended skills prior to taking this course

  - Passion for Data Science

Learning Objectives

In this course you will learn about:

  - The major steps involved in tackling a data science problem.
  - The major steps involved in practicing data science, from forming a concrete business or research problem, to collecting and analyzing data, to building a model, and understanding the feedback after model deployment.
  - How data scientists think through tackling interesting real-world examples.

Syllabus

Module 1: From Problem to Approach

  - Business Understanding
  - Analytic Approach

Module 2: From Requirements to Collection

  - Data Requirements
  - Data Collection

Module 3: From Understanding to Preparation

  - Data Understanding
  - Data Preparation

Module 4: From Modeling to Evaluation

- ○ Modeling
- ○ Evaluation

Module 5: From Deployment to Feedback

- ○ Deployment
- ○ Feedback

Grading Scheme

1. The minimum passing mark for the course is 70% with the following weights:
   - ○ 50% - All Review Questions
   - ○ 50% - The Final Exam
2. Though Review Questions and the Final Exam have a passing mark of 50% respectively, the only grade that matters is the overall grade for the course.
3. Review Questions have no time limit. You are encouraged to review the course material to find the answers. Please remember that the Review Questions are worth 50% of your final mark.
4. The final exam has a 1 hour time limit.
5. Attempts are per question in both, the Review Questions and the Final Exam:
   - ○ One attempt - For True/False questions
   - ○ Two attempts - For any question other than True/False
6. There are no penalties for incorrect attempts.
7. Clicking the "Final Check" button when it appears, means your submission is FINAL. You will NOT be able to resubmit your answer for that question ever again.
8. Check your grades in the course at any time by clicking on the "Progress" tab.

Change Log

This course was last updated on September 10th, 2020:

- labs updated

This course was last updated on September 15th, 2017:

- The review questions for all the modules were revised to address any typos or ambiguity in the questions.
- The final exam was also revised to include questions that are based on the course videos and that are free of any typos or ambiguity in the questions.
- New labs were created for Modules 1 - 4 and uploaded in the form of Jupyter notebooks for the convenience of students.
- For Modules 2, 3, and 4, Labs are available in Python as well as in R in order to accommodate the two popular programming languages in data science.
- Lab component removed from Module 5.

This course was previously updated on September 1st, 2016:

- The course content was completely redone based on the original content.

Copyrights and Trademarks

Learning Objectives

In this lesson you will learn about:

- Why we are interested in data science.
- What a methodology is, and why data scientists need a methodology.
- The data science methodology and its flowchart.
- How to apply business understanding and the analytic approach to any data science problem.

Video Business Understanding

Welcome to Data Science Methodology 101 From 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 directed 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 healthcare authorities in its region and brought in IBM data scientist 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 prioritize "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 that50% 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 directionRemained 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 FailurePredicting readmission risk.Understanding the combination of events that led to the predicted outcomeApplying an easy-to-understand process to new patients, regarding their readmission risk.This ends the Business Understanding section of this course.Thanks for watching!

Analytic Approach

Welcome to Data Science Methodology 101 From problem to approach Analytic Approach! Selecting the right analytic approach depends on the question being asked.The approach involves seeking clarification from the person who is asking the question,so as to be able to pick the most appropriate path or approach.

In this video we'll see how the second stage of the data science methodology is applied.Once the problem to be addressed is defined, the appropriate analytic approach for the problem is selected in the context of the business requirements.This is the second stage of the data science methodology.Once a strong understanding of the question is established, the analytic approach can be selected.This means identifying what type of patterns will be needed to address the question most effectively.If the question is to determine probabilities of an action, then a predictive model might be used.If the question is to show relationships, a descriptive approach maybe be required.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.For example if the question requires a yes/ no answer, then a classification approach to predicting a response would be suitable.Machine Learning is a field of study that gives computers the ability to learn without being explicitly programmed.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.So now, let's look at the 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 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 riskis 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.This ends the Analytic Approach section for this course.Thanks for watching!End of transcript. Skip to the start.

Lab - From Problem to Approach in Python
Lab Instructions
This course uses Cognitive Class Labs (formerly known as BDU Labs), an online virtual lab environment to help you get hands-on experience without the hassle of installing and configuring the tools. You will get access to popular open-source data science tools right inside your browser, like Jupyter Notebooks, which you will use to get hands-on practice with Spark in this lab.
Why is the business understanding stage important
A: it helps clarify the goal of the entity asking the questions

What are the two outstanding features of the data science methodology?
A: the flowchart is highly iterative
The flowchart never ends.

Can we predict the cuisine of a given dish using the name of the dish only
A: no

For example, the following dish names were taken from the menu of a local restaurant.

1. Beast
2. 2 pm
3. 4 minute

Are you able to tell the cuisine of the dishes?

A:
<details><summary>Click here for the solution</summary>

```python
        #The correct answer is:
        The cuisine is <strong>Japanese</strong>. Here are links to the images of the dishes:

        Beast: https://ibm.box.com/shared/static/5e7duvewfl5bk4317sna5skvdhrehro2.png

        2PM: https://ibm.box.com/shared/static/d9xuzqm8cq76zxxcc0f9gdts4iksipyk.png

        4 Minute: https://ibm.box.com/shared/static/f1fwvvwn4u8rx8tghep6zyj5pi6a8v8k.png

        Photographs by Avlxyz:
https://commons.wikimedia.org/wiki/Category:Photographs_by_Avlxyz

```

</details>

Q. What about by appearance only? Yes or no

A. No, especially when it comes to countries in close geographical proximity such as Scandinavian countries, or Asian countries.

Q. What about determining the cuisine of a dish based on its ingredients?

A. Potentially yes, as there are specific ingredients unique to each cuisine.

As you guessed, yes, determining the cuisine of a given dish based on its ingredients seems like a viable solution as some ingredients are unique to cuisines. For example:

- When we talk about American cuisine, the first ingredient that comes to one's mind (or at least to my mind =D) is beef or turkey.
- When we talk about British cuisines, the first ingredient that comes to one's mind is haddock or mint sauce.
- When we talk about Canadian cuisines, the first ingredient that comes to one's mind is bacon or poutine.
- When we talk about French cuisine, the first ingredient that comes to one's mind is bread or butter.
- When we talk about Italian cuisine, the first ingredient that comes to one's mind is tomato or ricotta.
- When we talk about Japanese cuisines, the first ingredient that comes to one's mind is seaweed or soy sauce.
- When we talk about Chinese cuisines, the first ingredient that comes to one's mind is ginger or garlic.
- When we talk about Indian cuisine, the first ingredient that comes to one's mind is masala or chillies.

Analytic Approach

So 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.

Why is the analytic approach stage important?

A: Because it helps identify what type of patterns will be needed to address the question most effectively.

Let's explore 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.

This is a decision tree that a naive person might create manually. Starting at the top with all the recipes for all the cuisines in the world, if a recipe contains rice, then this decision tree would classify it as a Japanese cuisine. Otherwise, it would be classified as not a Japanese cuisine.

Q. Is this a good decision tree? Yes or no and why?

A. No, because a plethora of dishes from other cuisines contain rice. Therefore, using rice as the ingredient in the Decision node to split on is not a good choice.

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.

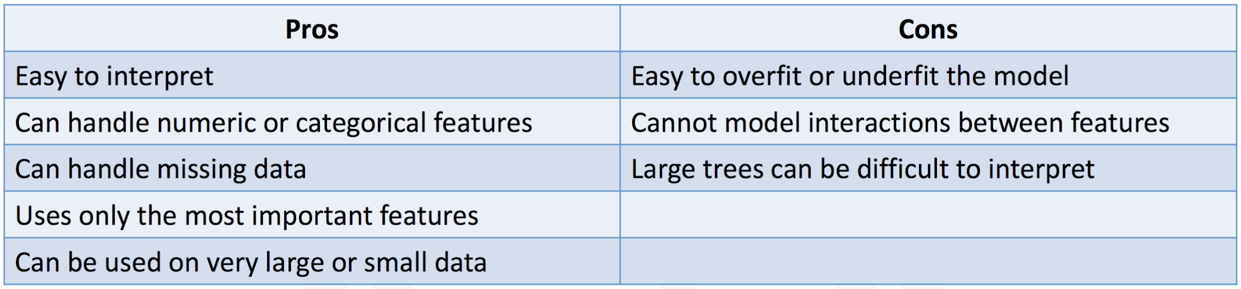
Suppose that our data consists of green triangles and red circles.

The following decision tree would be considered the optimal model for classifying the data into a node for green triangles and a node for red circles.

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.

Here are some characteristics of decision trees:



Pros	Cons
Easy to interpret	Easy to overfit or underfit the model
Can handle numeric or categorical features	Cannot model interactions between features
Can handle missing data	Large trees can be difficult to interpret
Uses only the most important features	
Can be used on very large or small data	

Now let's put what we learned about decision trees to use. Let's try and build a much better version of the decision tree for our recipe problem.

I hope you agree that the above decision tree is a much better version than the previous one. Although we are still using **Rice** as the ingredient in the first *decision node*, recipes get divided into **Asian Food** and **Non-Asian Food**. **Asian Food** is then further divided into **Japanese** and **Not Japanese** based on the **Wasabi** ingredient. This process of splitting *leaf nodes* continues until each *leaf node* is pure, i.e., containing recipes belonging to only one cuisine.

Accordingly, decision trees are a suitable technique or algorithm for our recipe case study.

Thank you for completing this lab!

This notebook was created by Alex Aklson. I hope you found this lab session interesting. Feel free to contact me if you have any questions!

This notebook is part of a course called *The Data Science Method*. If you accessed this notebook outside the course, you can take this course online by clicking here.

Graded Review Questions Instructions

1. Time allowed: **Unlimited**
   - We encourage you to go back and review the materials to find the right answer.
   - Please remember that the Review Questions are worth 50% of your final mark.
2. Attempts per question:
   - One attempt - For True/False questions.

- Two attempts - For any question other than True/False.

3. Clicking the "**Final Check**" button when it appears, means your submission is **FINAL**. You will **NOT** be able to resubmit your answer for that question ever again.
4. Check your grades in the course at any time by clicking on the "Progress" tab.

Quiz:
A methodology is a system of methods used in a particular area of study or activity.
The data science methodology described in this course is outlined by John Rollins from IBM.
The first stage of the data science methodology is business understanding.

Module 2 - From Requirements to Collection

Learning Objectives

In this lesson you will learn about:

- Data requirements and data understanding.
- What occurs during data collection.
- How to apply data requirements and data collection to any data science problem.

Video - Data requirements

Welcome to Data Science Methodology 101 From Requirements to Collection Data Requirements!If your goal is to make a spaghetti dinner but you don't have the right ingredients to make this dish, then your success will be compromised.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 desiredoutcome.Building on the understanding of the problem at hand, and then using the analytical approach selected, the Data Scientist is ready to get started.Now let's look at 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 datacollection.So now, let's look at the case study related to applying "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 readmission 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 diagnosis, procedures,prescriptions, and other services provided either during hospitalization or through outpatient/doctor visits.Thus, a particular patient could have thousands of records, representing all their relatedattributes.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.This ends the Data Requirements section for this course.Thanks for watching!

Video - Data Collection
Welcome to Data Science Methodology 101 From Requirements to 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 dataset, 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.In essence, the ingredients are now sitting on the cutting board.Now let's look at some examples of the data collection stage within the data science methodology.This stage is undertaken as a follow-up to the data requirements stage.So now, let's look at the case study related to applying "Data Collection".Collecting data requires that you know the source or, know where to find the data elements that are needed.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.Thanks for watching!End of transcript. Skip to the start.

Labs
From Requirements to collection

Introduction

In this lab, we will continue learning about the data science methodology, and focus on the **Data Requirements** and the **Data Collection** stages.

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.

In the **From Problem to Approach Lab**, we determined that automating the process of determining the cuisine of a given recipe or dish is potentially possible using the ingredients of the recipe or the dish. In order to build a model, we need extensive data of different cuisines and recipes.

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

Data Collection

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.

Web Scraping of Online Food Recipes

A researcher named Yong-Yeol Ahn scraped tens of thousands of food recipes (cuisines and ingredients) from three different websites, namely:

For more information on Yong-Yeol Ahn and his research, you can read his paper on Flavor Network and the Principles of Food Pairing.

Luckily, we will not need to carry out any data collection as the data that we need to meet the goal defined in the business understanding stage is readily available.

We have already acquired the data and placed it on an IBM server. Let's download the data and take a look at it.

Important note: Please note that you are not expected to know how to program in R. The following code is meant to illustrate the stage of data collection, so it is totally fine if you do not understand the individual lines of code. We have a full course on programming in R, R101, so please feel free to complete the course if you are interested in learning how to program in R.

Using this notebook:

To run any of the following cells of code, you can type **Shift + Enter** to execute the code in a cell.

Get the version of R installed

Using this notebook:

To run any of the following cells of code, you can type **Shift + Enter** to execute the code in a cell.

Get the version of R installed.

```
# check R version
R.Version()$version.string
```

Download the data from the IBM server.

```
# click here and press Shift + Enter
download.file("https://cf-courses-deloperSkillsNetwork-DS0103EN-SkillsNetwork/labs/Module%202/recipes.csv",
          destfile = "/resources/data/recipes.csv", quiet = TRUE)

print("Done!") # takes about 30 seconds
```

Read the data into an R dataframe and name it **recipes**.

```
recipes <- read.csv("/resources/data/recipes.csv") # takes 10 sec
```

Show the first few rows.

```
head(recipes)
```

Get the dimensions of the dataframe.

```
nrow(recipes)
ncol(recipes)
```

So our dataset consists of 57,691 recipes. Each row represents a recipe, and for each recipe, the corresponding cuisine is documented as well as whether 384 ingredients exist in the recipe or not, beginning with almond and ending with zucchini.

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.

Thank you for completing this lab!

This notebook is part of the free course on **Cognitive Class** called *Data Science Methodology*. If you accessed this notebook outside the course, you can take this free self-paced course, online by clicking here.

Lab Instructions

Please note that two versions of this lab were prepared: one in R and another one in python. You do *NOT* have to complete both labs as the content is identical. Complete the lab in your language of preference. This lab is in python and you can access the one in R by navigating to the other tab above.

Objectives

After completing this lab you will be able to:

- Understand Data Requirements
- Explore the stages in Data Collection

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.

In the **From Problem to Approach Lab**, we determined that automating the process of determining the cuisine of a given recipe or dish is potentially possible using the ingredients of the recipe or the dish. In order to build a model, we need extensive data of different cuisines and recipes.

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.

Web Scraping of Online Food Recipes

A researcher named Yong-Yeol Ahn scraped tens of thousands of food recipes (cuisines and ingredients) from three different websites, namely:

For more information on Yong-Yeol Ahn and his research, you can read his paper on Flavor Network and the Principles of Food Pairing.

Luckily, we will not need to carry out any data collection as the data that we need to meet the goal defined in the business understanding stage is readily available.

We have already acquired the data and placed it on an IBM server. Let's download the data and take a look at it.

Important note: Please note that you are not expected to know how to program in python. The following code is meant to illustrate the stage of data collection, so it is totally fine if you do not understand the individual lines of code. There will be a full course in this certificate on programming in python, Python for Data Science, which will teach you how to program in Python if you decide to complete this certificate.

Using this notebook:

To run any of the following cells of code, you can type **Shift + Enter** to execute the code in a cell.

Get the version of Python installed.

Get the version of Python installed.

```
# check Python version
!python -V
Python 3.7.12
```

Read the data from the IBM server into a *pandas* dataframe.

```
import pandas as pd # download library to read data into dataframe
pd.set_option('display.max_columns', None)

recipes =
pd.read_csv("https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloper
SkillsNetwork-DS0103EN-SkillsNetwork/labs/Module%202/recipes.csv")

print("Data read into dataframe!") # takes about 30 seconds
Data read into dataframe!
```

Show the first few rows.

So our dataset consists of 57,691 recipes. Each row represents a recipe, and for each recipe, the corresponding cuisine is documented as well as whether 384 ingredients exist in the recipe or not beginning with almond and ending with zucchini.

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.

Module 3 - From Understanding to Preparation

Learning Objectives

In this lesson you will learn about:

- What it means to *understand* data.
- What it means to *prepare* or *clean* data.
- Ways in which data is prepared.
- How to apply data understanding and data preparation to any data science problem.

Data Understanding

Welcome to Data Science Methodology 101 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, descriptivestatistics 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 those 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.This ends the Data Understanding section of this course.Thanks for watching!End of transcript. Skip to the start.

This section consists of two videos:

1. Data Preparation - Concepts.
2. Data Preparation - Case Study.

You can navigate through the videos using the tabs above!

In a sense, data preparation is similar to washing freshly picked vegetables in so far 0:14 / 3:03 Press UP to enter the speed menu then use the UP and DOWN arrow keys to navigate the different speeds, then press ENTER to change to the selected speed.Click on this button to mute or unmute this video or press UP or DOWN buttons to increase or decrease volume level.Muted Volume.Video transcript Start of transcript. Skip to the end. Welcome toData Science Methodology 101 From Understanding to Preparation Data Preparation - Concepts! In a sense, data preparation is similar to washing freshly picked vegetables in so far 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 flavors 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 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 details in this area. After all, it takes just one bad ingredient to ruin a fine meal. This ends the Data Preparation section of this course, in which we've reviewed key concepts. Thanks for watching! End of transcript. Skip to the start.

Data Preparation - Case Study (4:14)

Welcome to Data Science Methodology 101 From 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 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 readmission.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 outpatients.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 formodeling.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.Comorbidities with congestive heart failure were also considered, such as diabetes, hypertension,and many other diseases and chronic conditions that could impact the risk of readmission 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 comorbidities 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.This ends the Data Preparation section of this course, in which we applied the key concepts to the case study.Thanks for watching!

Lab

Objectives

After completing this lab you will be able to:

- Understand Data
- Prepare Data for analysis and inference

Introduction

In this lab, we will continue learning about the data science methodology, and focus on the **Data Understanding** and the **Data Preparation** stages.

Recap

In Lab **From Requirements to Collection**, we learned that the data we need to answer the question developed in the business understanding stage, namely *can we automate the process of determining the cuisine of a given recipe?*, is readily available. A researcher named Yong-Yeol Ahn scraped tens of thousands of food recipes (cuisines and ingredients) from three different websites, namely:

For more information on Yong-Yeol Ahn and his research, you can read his paper on Flavor Network and the Principles of Food Pairing.

We also collected the data and placed it on an IBM server for your convenience.

Important note: Please note that you are not expected to know how to program in python. The following code is meant to illustrate the stage of data collection, so it is totally fine if you do not understand the individual lines of code. There will be a full course in this certificate on programming in python, Python for Data Science, which will teach you how to program in Python if you decide to complete this certificate.

Using this notebook:

To run any of the following cells of code, you can type **Shift + Enter** to execute the code in a cell.

Get the version of Python installed.

Using this notebook:

To run any of the following cells of code, you can type **Shift + Enter** to execute the code in a cell.

Get the version of Python installed.

```
# check Python version
!python -V
Python 3.7.12
```

Download the library and dependencies that we will need to run this lab.

```
import pandas as pd # import library to read data into dataframe
pd.set_option('display.max_columns', None)
import numpy as np # import numpy library
import re # import library for regular expression
```

Download the data from the IBM server and read it into a *pandas* dataframe.

```
recipes =
pd.read_csv("https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0103EN-SkillsNetwork/labs/Module%202/recipes.csv")

print("Data read into dataframe!") # takes about 30 seconds
Data read into dataframe!
```

Show the first few rows.

Get the dimensions of the dataframe.

```
recipes.shape
(57691, 384)
```

So our dataset consists of 57,691 recipes. Each row represents a recipe, and for each recipe, the corresponding cuisine is documented as well as whether 384 ingredients exist in the recipe or not, beginning with almond and ending with zucchini.

We know that a basic sushi recipe includes the ingredients:

- rice
- soy sauce
- wasabi
- some fish/vegetables

Let's check that these ingredients exist in our dataframe:

Let's check that these ingredients exist in our dataframe:

```
ingredients = list(recipes.columns.values)
```

```
print([match.group(0) for ingredient in ingredients for match in
[(re.compile(".*(rice).*")).search(ingredient)] if match])
print([match.group(0) for ingredient in ingredients for match in
[(re.compile(".*(wasabi).*")).search(ingredient)] if match])
print([match.group(0) for ingredient in ingredients for match in
[(re.compile(".*(soy).*")).search(ingredient)] if match])
['brown_rice', 'licorice', 'rice']
['wasabi']
['soy_sauce', 'soybean', 'soybean_oil']
```

Yes, they do!

- rice exists as rice.
- Wasabi exists as wasabi.
- soy exists as soy_sauce.

So maybe if a recipe contains all three ingredients: rice, wasabi, and soy_sauce, then we can confidently say that the recipe is a **Japanese** cuisine! Let's keep this in mind!

In this section, we will prepare data for the next stage in the data science methodology, which is modeling. This stage involves exploring the data further and making sure that it is in the right format for the machine learning algorithm that we selected in the analytic approach stage, which is decision trees.

First, look at the data to see if it needs cleaning.

```
recipes["country"].value_counts() # frequency table
American        40150
Mexico           1754
Italian          1715
Italy            1461
Asian            1176
                ...
Indonesia          12
Belgium            11
East-African       11
Israel              9
Bangladesh          4
Name: country, Length: 69, dtype: int64
```

By looking at the above table, we can make the following observations:

1. Cuisine column is labeled as Country, which is inaccurate.
2. Cuisine names are not consistent as not all of them start with an uppercase first letter.
3. Some cuisines are duplicated as variations of the country name, such as Vietnam and Vietnamese.
4. Some cuisines have very few recipes.

Let's fix these problems.

Fix the name of the column showing the cuisine.

```
column_names = recipes.columns.values
column_names[0] = "cuisine"
recipes.columns = column_names
```

Recipes

Make all the cuisine names lowercase.

```
recipes["cuisine"] = recipes["cuisine"].str.lower()
```

Make the cuisine names consistent.

Remove cuisines with < 50 recipes.

```
# get list of cuisines to keep
recipes_counts = recipes["cuisine"].value_counts()
cuisines_indices = recipes_counts > 50

cuisines_to_keep = list(np.array(recipes_counts.index.values)[np.array(cuisines_indices)])
rows_before = recipes.shape[0] # number of rows of original dataframe
print("Number of rows of original data frame is {}.".format(rows_before))

recipes = recipes.loc[recipes['cuisine'].isin(cuisines_to_keep)]

rows_after = recipes.shape[0] # number of rows of processed dataframe
print("Number of rows of processed data frame is {}.".format(rows_after))

print("{} rows removed!".format(rows_before - rows_after))
```

```
Number of rows of the original data frame is 57691.
Number of rows of the processed data frame is 57394.
297 rows removed!
```

Convert all Yes's to 1's and the No's to 0's

```
recipes = recipes.replace(to_replace="Yes", value=1)
recipes = recipes.replace(to_replace="No", value=0)
```

Let's analyze the data a little more in order to learn the data better and note any interesting preliminary observations.

Run the following cell to get the recipes that contain **rice** *and* **soy** *and* **wasabi** *and* **seaweed**.

```
recipes.head()
```

```
check_recipes = recipes.loc[
        (recipes["rice"] == 1) &
        (recipes["soy_sauce"] == 1) &
        (recipes["wasabi"] == 1) &
        (recipes["seaweed"] == 1)
]
```

```
Check_recipes
```

Based on the results of the above code, can we classify all recipes that contain **rice** _and_ **soy** _and_ **wasabi** _and_ **seaweed** as **Japanese** recipes? Why?

#The correct answer is:

No, because other recipes such as Asian and East_Asian recipes also contain these ingredients.

Let's count the ingredients across all recipes.

```
# sum each column
ing = recipes.iloc[:, 1:].sum(axis=0)

# define each column as a pandas series
ingredient = pd.Series(ing.index.values, index = np.arange(len(ing)))
count = pd.Series(list(ing), index = np.arange(len(ing)))

# create the data frame
ing_df = pd.DataFrame(dict(ingredient = ingredient, count = count))
ing_df = ing_df[["ingredient", "count"]]
print(ing_df.to_string())
```

Now we have a dataframe of ingredients and their total counts across all recipes. Let's sort this dataframe in descending order.

```
ing_df.sort_values(["count"], ascending=False, inplace=True)
ing_df.reset_index(inplace=True, drop=True)

print(ing_df)
```

```
         ingredient  count
0               egg  21022
1             wheat  20775
2            butter  20715
3             onion  18078
4            garlic  17351
..              ...    ...
378  strawberry_jam      1
379  sturgeon_caviar     1
380      kaffir_lime     1
381           beech      1
```

```
382          durian      0

[383 rows x 2 columns]
```

What are the 3 most popular ingredients?
 #1. Egg with <strong>21,025</strong> occurrences.
# 2. Wheat with <strong>20,781</strong> occurrences.
# 3. Butter with <strong>20,719</strong> occurrences.

However, note that there is a problem with the above table. There are ~40,000 American recipes in our dataset, which means that the data is biased towards American ingredients.

Therefore, let's compute a more objective summary of the ingredients by looking at the ingredients per cuisine.

Let's create a *profile* for each cuisine.

In other words, let's try to find out what ingredients Chinese people typically use, and what is **Canadian** food for example.

```
cuisines = recipes.groupby("cuisine").mean()
cuisines.head()
```

As shown above, we have just created a dataframe where each row is a cuisine and each column (except for the first column) is an ingredient, and the row values represent the percentage of each ingredient in the corresponding cuisine.

For example:

- *almond* is present across 15.65% of all of the **African** recipes.
- *butter* is present across 38.11% of all of the **Canadian** recipes.

Let's print out the profile for each cuisine by displaying the top four ingredients in each cuisine.

```
num_ingredients = 4 # define number of top ingredients to print

# define a function that prints the top ingredients for each cuisine
def print_top_ingredients(row):
        print(row.name.upper())
        row_sorted = row.sort_values(ascending=False)*100
        top_ingredients = list(row_sorted.index.values)[0:num_ingredients]
        row_sorted = list(row_sorted)[0:num_ingredients]

        for ind, ingredient in enumerate(top_ingredients):
        print("%s (%d%%)" % (ingredient, row_sorted[ind]), end=' ')
        print("\n")

# apply function to cuisines dataframe
```

```
create_cuisines_profiles = cuisines.apply(print_top_ingredients, axis=1)
```

Module 4 - From Modeling to Evaluation

Learning Objectives

In this lesson you will learn about:

- What the purpose of data modeling is.
- Some characteristics of the modeling process.
- What it means to *evaluate* a model.
- Ways in which a model is evaluated.
- How to apply modeling and model evaluation to any data science problem.

This section consists of two videos:

1. Modeling - Concepts.
2. Modeling - Case Study.

You can navigate through the videos using the tabs above!

Video

Welcome to Data Science Methodology 101 From Modeling to Evaluation Modeling - Concepts!Modeling 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, andsecond, 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 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 modeling, 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 Rollins' descriptive Data Science Methodology, the framework is geared to do3 things: First,understand the question at hand. Second,select an analytic approach or method to solve the problem, and third,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: Have I made enough to eat?Well, 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!End of transcript. Skip to the start.

Video 2

Welcome to Data Science Methodology 101 From Modeling to Evaluation Modeling - Case Study!Modeling 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 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 traumato 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.This concludes the Modeling section of the course, in which we applied the Case Study To the modeling stage within the data science methodology.Thanks for watching!

Video 3

Welcome to Data Science Methodology 101 From Modeling to Evaluation - 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, thena 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 turning 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 misclassificationcosts.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, 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 separationbetween 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 classificationmodel.This curve quantifies how well a binary classification model performs, declassifying the yes andno 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 valuesof the relative misclassification cost, the ROC curve helped in selecting the optimalmodel.This ends the Evaluation section of this course.Thanks for watching!End of transcript. Skip to the start.

Lab

From Modeling to Evaluation

Estimated time needed: **20** minutes

Objectives

After completing this lab you will be able to:

- Create Models
- Evaluate the models

Recap

In Lab **From Understanding to Preparation**, we explored the data and prepared it for modeling.

The data was compiled by a researcher named Yong-Yeol Ahn, who scraped tens of thousands of food recipes (cuisines and ingredients) from three different websites, namely:

For more information on Yong-Yeol Ahn and his research, you can read his paper on Flavor Network and the Principles of Food Pairing.

Important note: Please note that you are not expected to know how to program in python. The following code is meant to illustrate the stage of data collection, so it is totally fine if you do not understand the individual lines of code. There will be a full course in this certificate on programming in python, Python for Data Science, which will teach you how to program in Python if you decide to complete this certificate.

Using this notebook:

To run any of the following cells of code, you can type **Shift + Enter** to execute the code in a cell.

Download the library and dependencies that we will need to run this lab.

import pandas as pd # import library to read data into dataframe

```
pd.set_option("display.max_columns", None)
import numpy as np # import numpy library
import re # import library for regular expression
import random # library for random number generation
```

We already placed the data on an IBM server for your convenience, so let's download it from the server and read it into a dataframe called **recipes**.

```
recipes =
pd.read_csv("https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0103EN-SkillsNetwork/labs/Module%202/recipes.csv")

print("Data read into dataframe!") # takes about 30 seconds
Data read into dataframe!
```

We will repeat the preprocessing steps that we implemented in Lab **From Understanding to Preparation** in order to prepare the data for modeling. For more details on preparing the data, please refer to Lab **From Understanding to Preparation**.

Download and install more libraries and dependencies to build decision trees.

```
# import decision trees scikit-learn libraries
%matplotlib inline
from sklearn import tree
from sklearn.metrics import accuracy_score, confusion_matrix

import matplotlib.pyplot as plt

!conda install python-graphviz --yes
import graphviz

from sklearn.tree import export_graphviz

import itertools
Collecting package metadata (current_repodata.json): done
Solving environment: \
```

Check the data again!

```
recipes.head()
```

[bamboo_tree] Only Asian and Indian Cuisines

Here, we are creating a decision tree for the recipes for just some of the Asian (Korean, Japanese, Chinese, Thai) and Indian cuisines. The reason for this is because the decision tree does not run well when the data is biased towards one cuisine, in this case American cuisines. One option is to exclude the American cuisines from our analysis or just build decision trees for different subsets of the data. Let's go with the latter solution.

Let's build our decision tree using the data pertaining to the Asian and Indian cuisines and name our decision tree *bamboo_tree*.

Decision tree model saved to bamboo_tree!

Let's plot the decision tree and examine how it looks.

The decision tree learned:

- If a recipe contains *cumin* and *fish* and **no** *yogurt*, then it is most likely a **Thai** recipe.
- If a recipe contains *cumin* but **no** *fish* and **no** *soy_sauce*, then it is most likely an **Indian** recipe.

You can analyze the remaining branches of the tree to come up with similar rules for determining the cuisine of different recipes.

Feel free to select another subset of cuisines and build a decision tree of their recipes. You can select some European cuisines and build a decision tree to explore the ingredients that differentiate them.

To evaluate our model of Asian and Indian cuisines, we will split our dataset into a training set and a test set. We will build the decision tree using the training set. Then, we will test the model on the test set and compare the cuisines that the model predicts to the actual cuisines.

Let's first create a new dataframe using only the data pertaining to the Asian and the Indian cuisines, and let's call the new dataframe **bamboo**.

```
bamboo = recipes[recipes.cuisine.isin(["korean", "japanese", "chinese", "thai", "indian"])]
```

Let's see how many recipes exist for each cuisine.

```
bamboo["cuisine"].value_counts()
```

```
korean      799

indian      598

chinese     442

japanese    320

thai        289

Name: cuisine, dtype: int64
```

Let's remove 30 recipes from each cuisine to use as the test set, and let's name this test set **bamboo_test**.

```
# set sample size
```

```
sample_n = 30
```

Create a dataframe containing 30 recipes from each cuisine, selected randomly.

```
# take 30 recipes from each cuisine

random.seed(1234) # set random seed

bamboo_test = bamboo.groupby("cuisine", group_keys=False).apply(lambda x:
x.sample(sample_n))
```

```
bamboo_test_ingredients = bamboo_test.iloc[:,1:] # ingredients

bamboo_test_cuisines = bamboo_test["cuisine"] # corresponding cuisines or labels
```

Check that there are 30 recipes for each cuisine.

```
# check that we have 30 recipes from each cuisine

bamboo_test["cuisine"].value_counts()
```

```
chinese     30

indian      30

japanese    30

korean      30

thai        30

Name: cuisine, dtype: int64
```

Next, let's create the training set by removing the test set from the **bamboo** dataset, and let's call the training set **bamboo_train**.

```
bamboo_test_index = bamboo.index.isin(bamboo_test.index)

bamboo_train = bamboo[~bamboo_test_index]
```

```
bamboo_train_ingredients = bamboo_train.iloc[:,1:] # ingredients

bamboo_train_cuisines = bamboo_train["cuisine"] # corresponding cuisines or labels
```

Check that there are 30 *fewer* recipes now for each cuisine.

```
bamboo_train["cuisine"].value_counts()
```

```
korean     769

indian     568

chinese    412

japanese   290

thai       259

Name: cuisine, dtype: int64
```

Let's build the decision tree using the training set, **bamboo_train**, and name the generated tree **bamboo_train_tree** for prediction.

Now that we defined our tree to be deeper, more decision nodes are generated.

Now let's test our model on the test data.

```
bamboo_pred_cuisines = bamboo_train_tree.predict(bamboo_test_ingredients)
```

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.

So let's go ahead and create the confusion matrix for how well the decision tree is able to correctly classify the recipes in **bamboo_test**.

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 data points 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! So don't worry if you get slightly different numbers in your confusion matrix than the ones shown above.

Using the reference confusion matrix, how many **Japanese** recipes were correctly classified by our decision tree?

#The correct answer is:

36.67%.

Also using the reference confusion matrix, how many **Korean** recipes were misclassified as **Japanese**?

#The correct answer is:

3.33%.

What cuisine has the least number of recipes correctly classified by the decision tree using the reference confusion matrix?

#The correct answer is:

Japanese cuisine, with 36.67% only.

Module 5 - From Deployment to Feedback

Learning Objectives

In this lesson you will learn about:

- What happens when a model is deployed.
- Why model feedback is important.

Deployment - Concepts & Case Study (3:31)

Welcome to Data Science Methodology 101 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 be rolled out to a limited group of usersor 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 interventionactions.The goal, of course, was to reduce the likelihood that these patients would be readmitted within30 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 assessment 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 overtime.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 risk 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.This ends the Deployment section of this course.Thanks for watching!End of transcript. Skip to the start.

Feedback (3:08)

Welcome to the Data Science Methodology 101 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 methodology is 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 readmission outcomes recorded.
Third, the intervention would then be measured to determine how effective it was in reducing readmissions.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 readmission 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.This is the end of the Feedback portion of this course.Thanks for watching!End of transcript. Skip to the start.

Summary video
Welcome to Data Science Methodology 101 Course Summary!We've come to the end of our story, one that we hope you'll share.You've learned how to think like a data scientist, including taking the steps involved in tackling a data science problem and applying them to interesting, real-world examples.These steps have included:forming a concrete business or research problem, collecting and analyzing data,building a model, and understanding the feedback after model deployment.In this course, you've also learned methodical ways of moving from problem to approach, including the importance of understanding the question, the business goals and objectives, andpicking the most effective analytic approach to answer the question and solve the problem.You've also learned methodical ways of working with the data, specifically,determining the data requirements, collecting the appropriate data,understanding the data, and then preparing the data for modeling!You've also learned how to model the data by using the appropriate analytic approach,based on the data requirements and the problem that you were trying to solve Once the approach was selected, you learned the steps involved in evaluating and deploying the model, getting feedback on it, andusing that feedback constructively so as to improve the model.Remember that the stages of this methodology are iterative!This means that the model can always be improved for as long as the solution is needed, regardless of whether the improvements come from constructive feedback, or from examining newly available data sources.Using a real case study, you learned how data science methodology can be applied in context,toward successfully achieving the goals that were set out in the business requirementsstage.You also saw how the methodology contributed additional value to business units by incorporating data science practices into their daily analysis and reporting functions.The success of this new pilot program that was reviewed in the case study was evident by the fact that physicians were able to deliver better patient care by using new tools to incorporate timely data-driven information into patient care decisions.And finally, you learned, in a nutshell, the true meaning of a methodology!That its purpose is to explain how to look at a problem, work with data in support of solving the problem, and come up with an answer that addresses the root problem.By answering 10 simple questions methodically, we've taught you that a methodology can help you solve not only your data science problems, but also any other problem.Your success within the data science field depends on your ability to apply the right tools, at the right time, in the right order, to the address the right problem.And that is the way John Rollins sees it!We hope you've enjoyed taking the Data Science Methodology course and found it to be a valuable experience one that you'll share with others!And, of course, we also hope that you'll

review and take other data science courses in the Data Science Fundamentals learning path!Now, if you're ready and up to the challenge, please take the final exam.Thanks for watching!End of transcript. Skip to the start.