Answers to Case Studies
Chapter 2

Study tips

- Make sure to have given the questions a try before reading my answers
- The provided answers are one of many possible and correct answers that can be given to the raised questions.
- It will be great to compare your answers with my answers and use common sense and logic to find the better parts of each answer.

Now, let's look at the questions.

Show your Business Understanding using the five questions provided in this chapter.

Case Study 2: Predicting Solar Energy Generation

Q1

Do you have an acceptable Business Understanding of this case study? How can you make certain?

We can check our Business Understanding by the following five questions. These questions were provided in the chapter.

1. Who owns the risk and benefits? In other words, who is the decision-maker?
2. What is the Decision?
3. What is the challenge of decision-making?
4. What are the risks and benefits?
5. How can data analytics create value?

Q2

Show your Business Understanding using the five questions provided in this chapter.

Who owns the risk and benefits? In other words, who is the decision-maker? In this case, study the decision maker is the electric supplier. If they overestimate the electricity production of solar panels, they will have to pay the cost by producing more unplanned electricity which is expensive. If they underestimate, they will have to pay the cost of saving electricity or miss the opportunity of selling it. On the other hand, if they are reasonably accurate, they will be able to meet the electric demand at the lowest possible cost.

What is the Decision? How much extra electricity needs to be produced that will not be satisfied by just the amount generated by the solar panels?

What is the challenge of decision-making? If they overestimate the electricity production of solar panels, they will have to pay the cost by producing more unplanned electricity which is expensive. If they underestimate, they will have to pay the cost of saving electricity or miss the opportunity of selling it. Finding the balance will be difficult.

What are the risks and benefits? The are two risks: 1) not having produced enough low-cost electricity and having to produce high-cost electricity, 2) having electricity excess. The benefit is the opportunity to produce just enough electricity, and pay the minimum cost.

How can data analytics create value? By providing more accurate and reliable estimates. Furthermore, data analytics can give us insights as to whether the prediction tends to be overestimation or underestimation and we can err on the side of whichever is costlier.

Q3

Do you have an acceptable Data Understanding for this case study? How can you make certain?

We can check our Data Understanding by the following nine questions. These questions were provided in the chapter.

1. What are possible sources of data useful for this case study?

2. How is each source of data collected?
3. What are the Volume, Variety, and Velocity levels of each source of data?
4. What are the assumptions made in recording each value in each source of data?
5. What are the legal and ethical implications of using each source of data?
6. What kind of values do we normally get in each column in every source of data?
7. Where is the data source in the spectrum of raw to fully processed data?
8. How each source of data could be useful?
9. Have you been given access to meaningful samples of the data sources, and spent time getting to know them?

Q4

Mention at least one possible source of data that can be useful for this case study.

- Historical weather data

Q5

provide an answer to the nine questions provided under Data Understanding regarding the three data sources you mentioned in Q4.

How is the source of data collected? The weather data is collected in multiple ways. Here is a good source: https://flexbooks.ck12.org/cbook/ck-12-middle-school-earth-science-flexbook-2.0/section/11.14/primary/lesson/collecting-weather-data-ms-es/

What are the Volume, Variety, and Velocity levels of each source of data? The variety of data is limited, and there are only a handful of metrics we can measure the weather such as temperature and humidity. The Volume of the data depends on how far back we will include historical data. For instance, the volume will be 5-fold more if we want to use 10 years of historical data versus 2 years. Lastly, the velocity depends on the definition of data objects. If we set to analyze at a daily interval, every day we will get one new data object, and if we set out to do it every 30 minutes, every thirty minutes we will get a new data object.

What are the assumptions made in recording each value in each source of data? Besides the general assumptions of accuracy, and reliability, the biggest assumption is that it is meaningful to have one value for each metric in the period that is in the

definition of the data objects. For instance, If we are defining our data objects as one day, the implicit assumption is that it is meaningful to have one value for the temperature of one of the days. These assumptions simplify the data as every second during that day the temperature may have been different. Moreover, the temperature of different parts of the area that the temperature is attributed to may also have been different.

What are the legal and ethical implications of using each source of data? We have to be clear on who owns the data and make sure to abide by their terms. For instance, if a private company has collected the data for us, they might only allow us to use the data for personal matters and not commercial, and if we want to use the data commercially, they would like to charge us.

What kind of values do we normally get in each column in every source of data? Under each of the weather metrics, we would expect the range of the values we would expect to see. For instance, in Southern California, we expect the temperature to be between 35° and 100° Fahrenheit.

Where is the data source in the spectrum of raw to fully processed data? It is very unlikely for us to access the raw temperature data unless we go about collecting it ourselves. For instance, raw temperature data would be the datasets of temperature readings from thermometers from different locations. The data will be processed by averaging over the desired time and locations.

How the source of data could be useful? The amount of electricity that solar panels will be able to generate will have a direct correlation with the amount of exposure to the sun. Weather data can help us get an estimate for that.

Have you been given access to meaningful samples of the data sources, and spent time getting to know them? This has to be answered for real situations.

Q6

Try to imagine a situation in which Data Understanding can contribute to your Business Understanding for this cast study.

For instance, after going over the data we might see that the day of the year is a better predictor of electricity production than temperature. That is because the day in a calendar will tell us how many hours the sun will be in the sky.

Q7

Explain how Business Understanding can contribute to your Data Understanding for this case study.

Without business understanding and problem understanding, we would not know what patterns to look for in the data. We want to find patterns that will allow us to predict the amount of electricity that will be generated in the future.

Q8

What model can solve this case study, and describe at a high level the Data Preparation needed for such a model?

We can model this problem as a prediction and solve it using prediction algorithms such as linear regression, regression trees, or neural networks.

All these algorithms need the prepared data to have the following characteristics.

- The data must be processed into one table
- The definition of the data objects for the table will have to be decided. Let's say we will define each day as one data object.
- The dependent attribute will be the numerical value representing the amount of electricity generated by solar panels.
- The independent attributes that come from one or combinations of the data sources must be transformed so they describe the data objects (one day)

Q9

How would you evaluate the model you described in the previous question?

Evaluating prediction models may sound as simple as calculating R squared or RMSE, however, if we want to incorporate the cost of overestimation and underestimation we have to use alternative and custom metrics.

Q10

If the evaluation showed that the model is successful, how could the model be deployed? We didn't discuss development in this chapter, we will in the next chapters, so just give it a shot.

For the successful deployment of such a model, we would need data engineering skills so the trained models and the required data are available when needed.

Case Study 3: Ambulance Priority Management

Q1

Do you have an acceptable Business Understanding of this case study? How can you make certain?

We can check our Business Understanding by the following five questions. These questions were provided in the chapter.

1. Who owns the risk and benefits? In other words, who is the decision-maker?
2. What is the Decision?
3. What is the challenge of decision-making?
4. What are the risks and benefits?
5. How can data analytics create value?

Q2

Show your Business Understanding using the five questions provided in this chapter.

Who owns the risk and benefits? In other words, who is the decision-maker? The decision maker will be whatever organization is responsible for making the decision on which ambulance will be dispatched, and where the ambulance will go. Now, this governing organization may be a city's government or a healthcare system. Moving forward let us assume in this case study the decision maker is a healthcare system.

What is the Decision? Which ambulance will be dispatched, and where the ambulance will go?

What is the challenge of decision-making? First, A need for a quick decision. From the time that the need for an ambulance is reported, till the time that the decision is made, should not be more than a few seconds. Second, there are not a lot of data available that can give us the insight to compare the cases that need a limited number of ambulances.

What are the risks and benefits? Bad, or even good, decisions may lead to the loss of life. On the other hand, good decisions can maximize the number of individuals that are saved.

How can data analytics create value? If there are any meaningful patterns in the limited amount of data, an analytic solution may help improve the decision-making from arbitrary to insightful.

Q3

Do you have an acceptable Data Understanding for this case study? How can you make certain?

We can check our Data Understanding by the following nine questions. These questions were provided in the chapter.

1. What are possible sources of data useful for this case study?
2. How is each source of data collected?
3. What are the Volume, Variety, and Velocity levels of each source of data?
4. What are the assumptions made in recording each value in each source of data?
5. What are the legal and ethical implications of using each source of data?
6. What kind of values do we normally get in each column in every source of data?
7. Where is the data source in the spectrum of raw to fully processed data?
8. How each source of data could be useful?
9. Have you been given access to meaningful samples of the data sources, and spent time getting to know them?

Q4

Mention at least one possible source of data that can be useful for this case study.

1) Call Center Data
2) Hospital Records

Q5

provide an answer to the nine questions provided under Data Understanding regarding the three data sources you mentioned in Q4.

How is the source of data collected? 1)The call center data is collected live. When the emergency call center (911 in the US for instance) receives a call, the agent gets information while on the phone. 2) The hospital records will be collected continuously while the patient will go through different medical exams and procedures.

What are the Volume, Variety, and Velocity levels of each source of data? 1) The volume and velocity of the call center data are obvious. Each emergency report is a data object. The variety of the call center data is limited to the information collected by the call center. 2) The volume and velocity of hospital records will probably be higher than the call center data. Each time a patient is visited under emergency or non-emergency situations hospital records are generated. Regarding the variety of the hospital-recorded data, we should also expect much more variety as there is much more opportunity for data collection. Having said that, only a small part of the hospital records will be relevant to this case study.

What are the assumptions made in recording each value in each source of data? 1) for call center data we will assume that the reporter is honest, and also we will assume that there has not been a miscommunication between the incident reporter and the agent. Lastly, we also assume that the date entered has been done correctly. 2) There are three assumptions for the hospital record data. First. patients are honest when they provide the recorded information. Second, the medical exams are valid and reliable. Third, the doctor's interpretation of the medical situation is objective.

What are the legal and ethical implications of using each source of data? The patients whose data we will use may be protected by legal frameworks such as GDPR.

What kind of values do we normally get in each column in every source of data? 1) in the call center data we will get the call transcripts, and the columns such as the name of the reporter, along with other details about the emergency. 2) the hospital records data will be various such as numerical values of the medical exams, the descriptions of patient information, and the doctor's diagnosis. Moreover, medical images such as xrays and MRIs will also be available.

Where is the data source in the spectrum of raw to fully processed data? The data that we'll be given access to will be as raw as they could be meaning we will get the same data as it was collected.

How the source of data could be useful? 1) the call center data will be useful in the light of the fact that if we should find any actionable pattern that will give us any insight about the future it will have to be in the call center data, finding any pattern in the medical data will not help with this particular problem. 2) the hospital records could help give our models a glimps of the future so the pattern in the call center data can be mapped to those in the medical records.

Have you been given access to meaningful samples of the data sources, and spent time getting to know them? This has to be answered for real situations.

Q6

Try to imagine a situation in which Data Understanding can contribute to your Business Understanding for this cast study.

Going over the call center data we will probably develop a classification of different types of emergencies that may happen and lead to the need for an ambulance. This will help us understand the problem better.

Q7

Explain how Business Understanding can contribute to your Data Understanding for this case study.

Without business understanding and problem understanding, we would not know what patterns to look for in the data. We want to find patterns that will allow us to predict the amount of electricity that will be generated in the future.

Q8

What model can solve this case study, and describe at a high level the Data Preparation needed for such a model?

Classification Algorithms. Classification algorithms need the prepared data to have the following characteristics.

- The data must be processed into one table
- The definition of the data objects for the table will have to be decided, and the one that makes the most sense will emergency situation; each row in the table will describe one emergency situation.

- The dependent attribute will be a category that divides the types of emergencies. For instance, the categorization could be the following four.
    - High time sensitive & Low chance of survival
    - High time sensitive & High chance of survival
    - Low time sensitive & High chance of survival
    - Low time sensitive & Low chance of survival
- The independent attributes that come from one or combinations of the data sources must be transformed so they describe the data objects (one day)

Q9

How would you evaluate the model you described in the previous question?

Metric: the number of quality-adjusted life years.

Q10

If the evaluation showed that the model is successful, how could the model be deployed? We didn't discuss development in this chapter, we will in the next chapters, so just give it a shot.

Once the patterns in the call center data are found, the trained model can be used to infer which of the four categories the emergency call is regarding and make decisions based on the results.