Learning Response 1.1: Essential Elements of Predictive Analytics

Jason “Scott” Person

20250323

Hello Prof Ricky,

My first assignment is below. Given the elementary level of the subject matter, I was surprised at the amount of time and effort this submission took. Part of this is because I am totally unsure as to the expectations with regard to writing complexity. I would appreciate and get a lot of value out of any guidance you can provide as to whether or not the below is appropriate and if not, what I need to do to get it there.

Thank you!

Scott Person

1. What is Data Mining, according to McCormick?
   1. Break down McCormick’s definition into bullet points by key idea.
   2. Under each bullet point, indent sub-points and record the relevant details McCormick shares about those ideas. (For instance, explaining what it means that the data was “accumulated during the normal course of doing business”.)

Note: I may be overthinking this, but I’m having a bit of a dilemma regarding the plagiarism policy vs. you question to answer, “according to McCormick”. I’m going to interpret the question as you wanting a direct quote as rephrasing this would simply be the mechanics of the English language.

According to McCormick, “Data Mining is the selection and analysis of data, accumulated during the normal course of doing business, to find (and confirm) previously unknown relationships that can produce positive and verifiable outcomes through the deployment of predictive models when applied to new data.” (McCormick) Key points include:

* The data is “normal” business data. That means it isn’t specifically gathered for the analytical effort.
* A significant product of the effort is identifying unknown relationships in the data – we don’t start with a thesis about the data.
* The outcomes need to be supported by the data and provide business value.
* The relationships within the data need to work with new data.
* The resulting models must be deployable – I’m assuming he means the result needs to be a product or capability.

1. How does McCormick summarize the basic idea of predictive analytics?
   1. What are a couple of examples of actions that can be taken on the basis of
   2. What serves as *the basis* of a predictive model?
   3. How is a predictive model used in relation to new data?
   4. What are example actions that might be taken as a result?

McCormick summarizes the basic idea of predictive analytics as taking a model created with existing data and then applying new data to the model to generate scores that lead to business actions. A couple of examples I’ve seen:

* At the FDIC we used various metrics of consumer bank health to generate a score that is used to determine the amount of oversight the bank needs.
* Some of my former coworkers created multiple models to predict the relative performance of college basketball teams. These were then used to construct predictions NCAA tournament outcomes. These models are provided to subscribers of a gambling advisory service, who use the predictions to guide their wagering.

In these examples, the basis of the model is data from previous world events. Historical data is used to create the model, which is then applied against new data. In my above examples, oversight rules and betting decisions would be made based on the scores provided by the models.

1. Why does McCormick use the term “data mining”, and what is his goal for laying out “the essential elements” of it?

McCormick chose to use the term “data mining” because he uses the cross industries standard process for data mining. They use the term so he does as well. He does this for precision. He lays out the essential elements to define explicitly what is and is not data mining and to draw a distinction between it and other analytics terms like statistical analysis, hypothesis testing, and reporting.

1. Summarize the key guidance and recommendations McCormick gives in “Section 2. Problem Definition.” Outline these with a handful of main points with sub-points – all in full sentences.

In Section 2, McCormick makes several recommendations.

* The first step is to break the objective down into decisions to be made.
  + These decision points provide focus for the effort by guiding data selection.
  + Ultimately the model when deployed will help make the decisions identified here.
* A plan of action is required. Based on model output, decisions will be made. How these decisions are made defines an intervention strategy. That intervention is the goal of predictive analytics.
* A higher-level goal, what would probably be defined as The Goal is adding business value in excess of the project cost – positive Return on Investment (ROI).
  + It’s not a complete data mining project until deployment. Deployment is defined as the model output being used to guide business decisions. Just gaining insights, while important and often valuable, is not sufficient.
  + A high-level estimate of ROI should be determined at the beginning of the project. This justifies the investment in the project.
* Keep in mind that the project may evolve into a program. McCormick defines a program as multiple projects that collectively solve a larger problem. Often these will be multiple binary decision models.

1. Who is Tom Khabaza?
   1. What is his background?
   2. What is his relevance to CRISP-DM and the Nine Laws?
   3. Search the web for “Khabaza’s Nine Laws of Data Mining” to find Khabaza’s own blog post on this topic online. (The URL should include codimension.net.) Provide the URL to his post.
   4. OBSERVE: The post continues in parts two and three. There are links in the left-hand sidebar and at the end of the post to take you to the other parts.

Tom Khabaza is a data mining thought leader, one of the lead authors of CRIS-DM, and he wrote the nine laws of data mining. He has almost 20 years of experience in data mining going back to the initial definition of the term. Tom’s Nine Laws of Data Mining are discussed in his blog article here (<http://khabaza.codimension.net/index_files/9laws.htm>).

1. What is McCormick’s opinion of the CRISP-DM process? What are some reasons he gives for his opinion?

I could not find any explicit mention of McCormick’s opinion of CRISP-DM in the videos, but the fact that he uses it in his material is indicative of support. Further, I located an article comprised of interview experts with McCormick on TWDI’s website. In it, McCormick states that he is a “big fan” of CRISP-DM. His main reason for this is that the process provides a standard framework and structure for performing data mining (Upside Staff, 2024).

1. State and briefly summarize Khabaza’s Nine Laws of Data Mining in a numbered list.
   1. To prepare for this, scan Khabaza’s original three posts on the Nine Laws of Data Mining. Note that he states a law succinctly and then explains in some detail. You do not need to read all the details word for word, but it’s worth a quick scan.
   2. Then consult [Meta S. Brown’s summary of the laws in Forbes.com](https://www.forbes.com/sites/metabrown/2016/01/27/9-laws-for-data-analytics-profits/). Here is a direct link to [Meta S. Brown, 9 Laws for Data Mining –Forbes.com](https://www.forbes.com/sites/metabrown/2016/01/27/9-laws-for-data-analytics-profits/)   
      <https://www.forbes.com/sites/metabrown/2016/01/27/9-laws-for-data-analytics-profits/>
   3. Name and summarize — or clarify with an explanatory sentence or two — each of these laws, as an easy-to-review numbered list.

Note, here again I’m going to resort to copying from the resources. The laws come directly from Khabaza’s list. The summaries associated with each are mine.

1. Business objectives are the origin of every data mining solution.
   1. Know what problem you’re trying to solve.
   2. Data mining is a process as opposed to a technology
   3. No business objective 🡪 it’s not data mining
2. Business knowledge is central to every step of the data mining process
   1. If you don’t know the business, you won’t be able to solve the business’s problems well.
   2. Business objectives remain a focus of data mining throughout the process
   3. SMEs are important to have on the team
3. Data preparation is more than half of every data mining process
   1. A lot of time is spent getting the right data into the right place for the analytics.
   2. It’s not that data prep is x% of the project – the claim is that it’s ALWAYS more than half, which is a stronger statement.
   3. Data prep automation tools won’t make this go away
   4. Data prep is both getting the data into the right shape (OBT is preferred by most algorithms) as well as adding context and features that will aid in the data mining process
4. The right model for a given application can only be discovered by experiment or “There is No Free Lunch for the Data Miner”
   1. Expect (and appreciate) trial and error in getting to a valuable solution.
   2. Algorithms are tools, they don’t find the solution without the data miner (who incidentally does the work)
   3. Working through the various model algorithms is part of the process
5. There are always patterns
   1. This is referring to the data – there are always insights that can be made to get value from the data.
   2. “Watkins’ Law” – David Watkins
   3. Since the data came from the business, there are business patterns in the data
   4. Start with the business process/knowledge
   5. Quote 1 (response for 8b) McCormick: “I’ve never seen a project fail because there weren’t patterns in the data.” I included this one because I found it to be the most surprising of any statement made in the videos. Business patterns lead to data patterns is a logical statement, but McCormick goes further to say that he’s never seen a project fail because the patterns didn’t exist. I would appreciate a dialog about this one. Does this mean that patterns existed but couldn’t be used or possibly that the data miner wasn’t smart enough to recognize or leverage the patterns to create a model? More on this one as I’ll bring it up if we have a group discussion.
6. Data mining amplifies perception in the business domain
   1. You’ll learn from the process of data mining
   2. The process is different than normal business analysis and provides insights that would not be achieved with non-data mining techniques.
   3. Quote 2 (response for 8b) “Data mining provides a kind of intelligence amplifier, helping business experts to solve business problems in a way which they could not achieve unaided.” With data mining we can uncover patterns in data that are not otherwise observable because the human brain isn’t optimized to process huge amounts of text and recognize subtle patterns over a large data volume. The process of doing this augments (amplifies) human expertise.
7. Prediction increases information locally by generalization
   1. Using past data, we can fill in missing details for new situations
   2. “Prediction Law”
   3. The predictions are not certainties, they are scores that weight the chances of something happening.
8. The value of data mining results is not determined by accuracy or stability of predictive models
   1. Business value is what matters
   2. “Value Law”
   3. The technical quality of the model is important, but it’s not the goal. The goal is adding business value.
9. All patterns are subject to change
   1. Continually evaluate your model. Sometimes the past patterns doing apply.
   2. “Law of Change”
   3. It is both a changing world and a change in our understanding
   4. Models are certain to degrade
   5. Quote 2 (response for 8b) “Patterns discovered by data mining do not last forever.” The business and our understanding of the business changes. We discover new patterns, new algorithms, new technologies. All of this provides opportunities for evolutionary and revolutionary changes to the data mining solution.
10. Now view McCormick’s videos on the 9 laws and enhance your above bulleted list by adding 3-4 sub-points under each law, in which you report some of McCormick’s most relevant comments on each law.
    1. *It’s fine to be succinct here. I don’t need you to write a book, but I need to see evidence that you have viewed and understood McCormick’s excellent video summary of these laws.*
    2. McCormick includes some vivid examples and great quotes. (For example, “Data mining is a bicycle of the mind.”) *Include and explain at least three of these.*

Responses to question 8 are included inline with the question 7 responses.

References

Upside Staff. (2024, May 23). *Machine learning and AI: A conversation with Keith McCormick*. Transforming Data with Intelligence (TDWI). https://tdwi.org/Articles/2024/05/23/ADV-ALL-Machine-Learning-and-AI-with-Keith-McCormick.aspx