Communications of the Association for Information Systems

Volume 50 Article 1

3-10-2022

When to Use Machine Learning: A Course Assignment

Varol Kayhan
University of South Florida, vkayhan@usf.edu

Recommended Citation
Kayhan, V. (2022). When to Use Machine Learning: A Course Assignment. Communications of the Association for Information Systems, 50, pp-pp. https://doi.org/10.17705/1CAIS.05005

Communications of the Association for Information Systems

Education Article DOI: 10.17705/1CAIS.05005 ISSN: 1529-3181

When to Use Machine Learning: A Course Assignment

Varol O. Kayhan

School of Information Systems and Management, University of South Florida

vkayhan@usf.edu

Abstract:

The number of institutions that offer machine learning courses continues to increase. However, supplementary materials that help instructors teach these courses fail to address an important step in the machine learning process; that is, conceptualizing a problem using a valid input-output relationship. To address this issue, I first review frameworks in extant work before proposing a decision flow. After discussing steps in the decision flow, I present a course assignment that reinforces the concepts in the decision flow. I conclude by discussing the lessons learned after using this assignment in a graduate course at a university in the United States.

Keywords: Machine Learning, Course Assignment.

This manuscript underwent editorial review. It was received 03/28/2021 and was with the authors for six months for two revisions. Craig van Slyke served as Associate Editor.

1 Introduction

Based on responses from 2,000 organizations across 10 industries, a recent McKinsey report found that nearly half the organizations (47%) had at least one artificial intelligence (AI) capability in their business processes—up from 20 percent in 2017 (Wladawsky-Berger, 2019). The same survey also found that 71 percent of the respondents expected to increase their investments in AI significantly in the coming years. However, most respondents also admitted that they lacked essential AI skills in their workforce. As a result, many higher education institutions have begun trying to fill the skills gap by offering new programs such as data science, data analytics, and business analytics (Perry, 2018). Even though these programs follow department- or college-specific curricula, they all include at least one or more machine learning (ML) course(s). These courses usually become popular in many programs, such as massively open online course platforms (Coursera, 2020), because they not only provide students with a gateway to AI but also allow them to build innovative applications so that they can apply for high-paying jobs in the analytics field. In this paper, I discuss how students in higher education institutions can develop and propose ML projects.

Due to ML's popularity, many educators, practitioners, and enthusiasts have created easily accessible, inexpensive, and sometimes freely available content such as lectures, textbooks, tutorials, blog posts, and videos (see Table A1 of Appendix 1 for a few samples). This content explains popular models and algorithms in addition to knowledge about applying them on sample data sets. However, this content mostly fails to address an important step in the ML process; that is, conceptualizing a problem using a valid input-output relationship. In fact, most supplementary and teaching materials—from textbooks to online tutorials—elaborate on the importance of using ML to solve the right problems, though they do not offer a methodology on how to conceptualize a problem using an input-output relationship and check the relationship's validity. As a result, many students lack the preparation for the workforce because they cannot conceptualize valid input-output relationships for the problems they propose to solve using ML even though they could be proficient in many ML algorithms and their applications. I address this problem by providing educators with a simple framework that comprises a decision flow and course assignment that can help students check the validity of input-output relationships of problems they propose to solve using ML.

This paper proceeds as follows: in Section 2, I briefly discuss the ML process model and existing frameworks for conceptualizing problems that ML can solve. In Section 3, I propose my framework. In Section 4, I present instructions for the course assignment and provide a proof of concept for it. In Section 5, I discuss the lessons I learned from using this assignment in several graduate-level courses at a university in the United States. In Section 6, I conclude the paper.

2 Background

Machine learning allows computers to learn from and identify patterns in data (Mitchell, 1999). In this sense, it closely resembles the definition for data mining (Linoff & Berry, 2010), which explains why earlier work has often used the two terms interchangeably (see Kohavi & Provost, 1998; Mitchell, 1999). For this paper's purposes, I do not formally distinguish between DM and ML; therefore, I only use the term ML hereafter to refer to learning from data.

One can use different types of learning paradigms in ML. The most widely used ML paradigm, supervised learning, uses labeled data; in contrast, the second most widely use ML paradigm, unsupervised learning, uses unlabeled data. For example, cluster analysis, dimensionality reduction, and outlier detection all constitute unsupervised methods because they involve finding patterns and rules in unlabeled data. The third paradigm, reinforcement learning, has gained more popularity in recent years and involves learning a task through practice (Mnih et al., 2015). In this case, ML models learn how to complete a task by gaining rewards and avoiding penalties. Hybrid paradigms also exist, such as semi-supervised learning and active learning in which the algorithms attempt to label the data on their own. In semi-supervised learning, algorithms do so by using the few available labels in the data set, while, in active learning, they do so by asking the user or source to provide the labels one at a time (Jordan & Mitchell, 2015, Olivier et al. 2006). In this paper, I focus specifically on supervised learning. I refer to the label as output (or y) and the data to predict the label as inputs (or x). The other learning paradigms—such as unsupervised, reinforcement, semi-supervised, and active learning—lie beyond the paper's scope.

For machine learning, one usually needs to complete certain tasks systematically in succession. The ML literature usually refers to these tasks in aggregate as a process model. The cross-industry standard process for data mining (CRISP-DM), which a consortium of companies in Europe developed in the late 1990s, represents one popular process model (Wirth & Hipp, 2000). Even though the consortium developed it for DM, one can easily adopt it to ML.

According to CRISP-DM, a typical project comprises six phases: 1) business understanding, 2) data understanding, 3) data preparation, 4) model building, 5) testing and evaluation, and 6) model deployment. All phases, except the first, are structured, which means that one can learn them simply enough from textbooks and tutorials. The first phase's highly unstructured and difficult-to-master nature means that, to learn it, one needs to not only solidly understand the context and business in question but also have the acumen to conceptualize the problem correctly in that context. Even if one solidly understands the context or business, one still might not know how to conceptualize a valid input-output relationship for the problem even after taking an ML course.

To address this gap, one framework—in the supervised learning context—suggests that one can conceptualize any problem using an input-output (or x-y) relationship and, thus, use ML to solve it (Brynjolfsson & Mitchell, 2017; Ng, 2016). Accordingly, one can use a set of inputs to explain or predict a single output. Researchers usually represent this conceptualization using A→B in which A constitutes the set of inputs and B the output. Therefore, if one can conceptualize a problem using inputs and an output, one can solve it with ML.

Even though seasoned ML professionals find this input-output conceptualization natural and intuitive, novices might find it harder to conceptualize because it comes with very few stipulations or boundary conditions. As such, novices often resort to the "law of the hammer" (Brislin, 1980) in which ML becomes the hammer, and any A→B relationship needs hammering. To make matters worse, novices might conceptualize an A→B relationship not necessarily for the problem at hand but rather from an easily accessible data set with multiple columns. As a result, novices might propose invalid A→B relationships that solve moot problems.

To provide more context around the A→B relationship, extant literature adopts an automation perspective by suggesting that ML can help one automate tasks (see Brynjolfsson & Mitchell, 2017). Therefore, one should conceptualize a problem not only as an A→B relationship but also approach it as an automation issue. For example, Ng (2016) suggests that "[if] a typical person can do a mental task with less than one second of thought, we can probably automate it using [ML] either now or in the near future". Brynjolfsson and Mitchell (2017) further extend the automation perspective by bringing cost into the picture. Accordingly, if an ML system becomes more cost-effective than a human for performing the same task, this system will replace the human (p. 1531). They further propose a 21-item rubric to determine whether one can automate a task using ML. Despite their usefulness, these 21 items do not offer in-depth insights into the nature of A or B.

Even though the automation and cost perspectives provide context into an A→B relationship, they do not discuss A or B in depth. Furthermore, one could use ML to solve other problems besides automation. To address this gap, I propose a framework that can enable novices to identify valid A→B relationships for problems that ML can solve. I discuss this framework in detail in Section 3.

3 Proposed Framework

3.1 Identify an A→B Relationship

To determine whether one can solve a problem using ML, one first needs to identify an A→B relationship about this problem where A represents the set of inputs and B the output. Please note that both A and B must have been observed already and captured in a data set because novices can sometimes suggest that only A should exist in the data set so that, when built, the ML model will generate or predict B. However, I need to reiterate that ML models first find patterns between A and B and then use these patterns to make predictions. They can only do so if both A and B exist in the same data set; therefore, proposing an ML project requires one to identify an A→B relationship and to ensure that one has a data set that includes both A and B.

3.2 Ensure that B is not Derived from A based on a Known Rule

Second, one needs to ensure that B is not derived from A based on a known rule or set of rules. Otherwise, an ML model that examines this A→B relationship will rediscover the same rule or set of rules. For example, consider an inventory system in which A represents how much product a company has in stock and B whether the company needs to restock the product (coded as 1 for yes and 0 for no). Also, consider that the inventory system uses a rule such that it recommends the company to restock a product (e.g., B = 1) if the quantity in stock drops below a certain pre-determined value (e.g., if A <= 10). A novice might obtain values for A and B from this system and suggest building an ML model that predicts when to restock a product. However, such an ML model would rediscover the same rule already in the system.

However, I do not mean to suggest that an A→B relationship cannot be based on rules. To the contrary, earlier work suggests that the purpose of using ML is to extract all rules between A and B so that B can be predicted using A (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Therefore, we must distinguish between already known rules that systems, devices, or mechanisms codify and unknown rules. If one already knows rules and a system, device, or mechanism codifies them, any data that one captures from this system, device, or mechanism in the form of A and B will lead one to rediscover the same rules. Otherwise, if no known rules exist between A and B, one can use ML to identify new ones. Note that one still can choose to use ML even when one knows the rules between A and B and wants to convert them into a machine-readable format to embed them in a system, device, or mechanism.

3.3 Ensure that B is not Derived from A based on a Known Equation

Third, one needs to ensure that B is not derived from A using a known equation. If one can compute the value of B by inserting the values of A into a known equation or formula, this A→B relationship does not warrant ML. For example, consider that B represents a student's overall course grade and A the grades that the student earned on individual assignments, quizzes, and exams. Unless indicated otherwise, B represents the weighted average of all grades in A. However, novices will often collect data on A and B to suggest that an ML model can predict the course grade based on individual grades on assignments, quizzes, and so on. If built, this ML model would rediscover the weighted average formula as the underlying pattern between A and B.

However, I do not mean to suggest that one cannot conceptualize an A→B relationship as an equation. In fact, most ML algorithms assume that the underlying pattern between A and B is an equation such that one calculates B by entering the values of A into it. For example, consider a regression model. As one of the most commonly used ML models, regression models calculate the necessary beta coefficients for each variable in A and use them to calculate B. Therefore, one needs to distinguish between whether one knows or does not know an equation between A→B. If one already knows an equation that computes B using the values of A, one does not need to use ML. If one does use ML in this situation, the model will rediscover the same equation.

In summary, if B is not derived from A using a known rule or an equation, then one can proceed to the third step to check whether one can use ML to examine the A→B relationship.

3.4 Check whether B is Observed at a Future Time Relative to A

Fourth, one needs check whether B is observed at a future time relative to A. If such a time difference exists between A and B, then one can use ML to examine this A→B relationship. Many interesting ML applications, particularly in the healthcare field, take advantage of this time difference. For example, consider that A refers to the biological and clinical variables of a patient collected at the time of discharge from a hospital and B refers to whether the patient is readmitted to the hospital within the next 30 days after discharge (see Mortazavi et al., 2016). Because A is observed at an earlier time (referred to as T1 hereafter) and B at a later time (referred to as T2 hereafter), algorithms that examine this A→B relationship can identify patterns between A and B and, thus, predict a future outcome based on data collected at an earlier time.

Similarly, consider another model in healthcare that can predict who has a higher chance to have a severe case of the coronavirus disease of 2019 (COVID-19) based on their biological and clinical variables collected during admission to a hospital (see Lassau et al., 2021). In this example, B refers to COVID-19 severity (observed at T2) and A refers to the biological and clinical variables (observed at T1). Due to the

time difference, an ML model built on this A→B relationship can identify patterns that can determine a future outcome based on variables collected at an earlier point in time.

The time difference between A and B can lead to interesting ML opportunities in other contexts as well. For example, in the higher education context, A (measured at T1) could be a student's sociodemographic variables—such as age, gender, and marital status—and curricular variables, such as a grade in an assignment, while B (measured later at T2) could be whether the student will drop out (see Kotsiantis et al., 2003). As before, a model built on this A→B relationship can identify patterns and determine a future outcome based on variables collected at an earlier time.

The marketing field frequently takes advantage of this time difference between A and B to acquire new customers or engage in targeted advertising. For example, consider that A refers to a customer's sociodemographic variables, such as age and income (collected at T1) and B refers to whether the customer used a coupon or bought a product at T2 (see Treiber, 2021). Based on previous transactions that capture both A and B, an ML model can identify patterns and determine B based on inputs collected at an earlier time.

Please see Table B1 of Appendix B for some example ML projects that involve a time difference between A and B from the popular media. Even though a time difference between A and B leads to interesting ML applications, it does not constitute a requirement. In fact, other interesting ML applications in which one observes both A and B simultaneously with no time difference exist. To determine the validity of these A→B relationships I propose a fifth and final step.

3.5 Check whether the Original Process that Generates or Captures B is Costly, Labor Intensive (i.e., Manual), or Subjective

Fifth, one needs to determine whether the original process that generates or captures B is costly, labor intensive (i.e., manual), or subjective. In many cases, different mechanisms, devices, or individuals generate A and B. Therefore, one can build an ML model that can function as a proxy for the mechanism, device, or individual that generates B and, thus, reduce costs, manual effort, or subjectivity. One also can consider this function as automation such that an ML model that one builds on this data set allows one to automate the process that generates B. Please note that the values of B must have been captured in this data set using the original costly, labor-intensive, or subjective process.

One can consider a process costly if, for example, it requires a highly specialized expert or expensive equipment to generate the value of B. Consider the process of determining how much risk an everyday contract between two parties poses to each party. In this case, A refers to the contract and B to the risk level. To identify B, one usually hires an experienced lawyer who might need to review the contract line by line. Naturally, hiring a lawyer for this task would prove quite costly because the lawyer would likely charge per hour. Instead, one can build an ML model on existing A and B to determine the value of B (for a new A) without a lawyer (e.g., LawGeex).

Similarly, one can consider a process labor intensive if, for example, it requires one or more individuals to work long hours with attention to detail while generating the value of B. Consider the process of transcribing speech in a video (i.e., closed captioning). In this case, A refers to the speech and B to its transcription. To identify B, one or more individuals might analyze the speech second by second to transcribe it. Instead, one can build an ML model on existing A and B (also called speech recognition) and, thus, generate transcription (i.e., B) for any kind of speech without any help from humans (e.g., IBM Watson's speech-to-text application).

Finally, one can consider a process subjective if, for example, multiple individuals look at the same A but generate a different value for B. Consider the process of sentencing in a court of law. A refers to the facts of a case and B to the sentence. Each judge who examines the same facts might come up with a different sentencing recommendation based on how they assess public safety or on their own beliefs and prejudices. Instead, one could build an ML model on existing A and B that judges could use to make more objective sentencing recommendations (e.g., Equivant). However, one must take care in these situations because such ML models can perpetuate existing biases and subjectivity already embedded in A and B.

In summary, one can use ML for an A→B relationship that someone generated using a costly, labor-intensive, or subjective mechanism, device, or individual. When one builds an ML model on this A→B relationship, patterns identified between A and B can allow the ML model to function as a proxy for the device, mechanism, or individual that originally captured or generated B. As a result, the model can

determine the value of B automatically and, thus, reduce costs, effort, or subjectivity. Note that, while one might build an ML model to reduce costs, effort, and subjectivity simultaneously, one can certainly focus on just one goal. Therefore, I urge readers to notice the OR operator, which generates *true* if at least one end goal is *true*. For example, consider an ML model that one builds to reduce costs. Even if this model has no bearing on the other goals (i.e., effort or subjectivity), it could still be viable.

Please see Table C1 of Appendix C for some example ML models from the popular media that function as proxies for the mechanisms, devices, or individuals that observe and generate B and, thus, reduce the costs, manual effort, or subjectivity involved in generating B.

Based on the above discussion, I propose the decision flow in Figure 1. As the figure shows, the decision proceeds as follows:

1) Identify an A→B relationship for the problem at hand.
2) Ensure that B is not derived from A using a known rule.
3) Ensure that B is not derived from A based on a known equation.
4) Check whether B is observed at a future time relative to A
5) If not (i.e., A and B are observed simultaneously), check whether the original mechanism, device, or individual that captures B is costly, labor intensive, or subjective such that one could build a model to function as a proxy for it.

Note that, even if the decision flow presented in Figure 1 might help one establish the validity of an A→B relationship, it does not ensure that this ML model can create value or be useful. Therefore, one must ensure that the model also generates value for stakeholders, which falls outside this paper's scope.

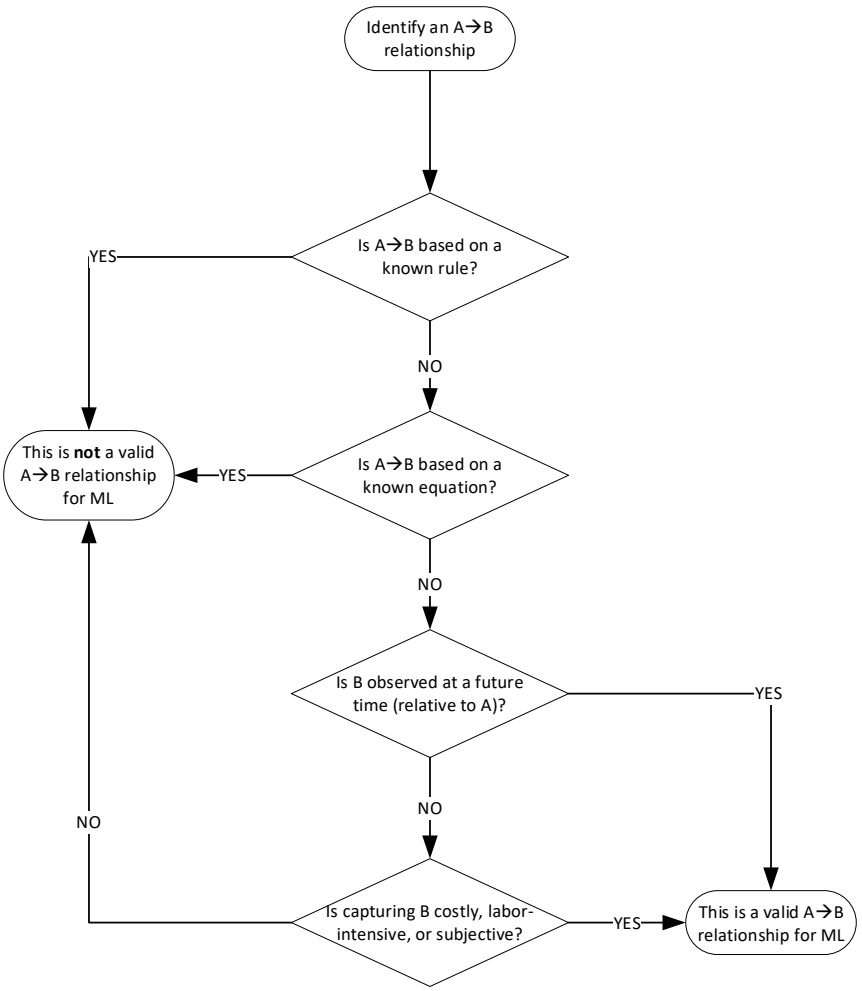

Figure 1. Decision Flow to Check the Validity of an A→B Relationship in the Supervised Learning Context

4 Course Assignment

To reinforce the concepts that the framework I propose highlights, I developed a course assignment that can cater to both undergraduate and graduate students enrolled in ML courses in higher education institutions. Preferably, instructors should use the assignment toward an ML course's end so that students have enough exposure to different data sets and models.

4.1 Learning Objectives

At the end of this assignment, students should be able to:

- Propose a machine learning project
- Justify the need for this project
- Assess the project's viability
- Identify the project's data set requirements by constructing a fictitious data set

4.2 Prerequisites

Before completing this assignment, students must demonstrate proficiency in:

- Defining machine learning
- Distinguishing between supervised and unsupervised learning
- Explaining the CRISP-DM process model
- Examining a data set necessary to build a machine learning model
- Building a machine learning model using at least one algorithm

4.3 Instructions

- In this assignment, you will propose a new ML project that no one has done before. To identify an idea, consider your current (or previous) workplace. You may want to think about some challenges you face in this workplace and determine whether you can tackle them using ML. You also may want to think about some new products or services that you can offer to internal/external customers in this workplace using ML.
- If your current (or previous) workplace does not lend itself to an ML project, then you can think about a project that might interest you personally or might benefit an organization (such as a nonprofit or student club) in which you participate. For example, you can think about how a nonprofit or student club can provide a new product/service (or improve an existing product/service) for its stakeholders. Again, the project must be new.
- This assignment will test your ability to come up with a project idea from scratch. Your grade will depend on how well you develop your project's conceptualization.
- For an example project, please see the one posted in your learning management system.
- Do not try to identify a project idea using one of the following:
  - Online search engines
  - Data-hosting websites (such as kaggle.com or github.com)
  - Blogging websites
- If you propose a project from one of these sources, you will not receive any credit. Furthermore, you will not receive any credit if you propose a project that you already currently participate in. Please do not propose a project that you worked on in another course either. The best way to identify a new project involves using introspection; that is, think about what problem you would want to solve using ML (such as one you observed in your current/previous workplace, your personal life, a nonprofit, or a student club).
- After you decide on a project, please address each of the following items:

1) **Goal (15%)**: describe your ML project's goal as specifically as possible (i.e., avoid ambiguous or general goal statements). For example, avoid overly general goal statements such as "learning new insights about churning customers". A more specific statement would include something such as "identify customers who are likely to churn in the next month".
2) **Benefits (15%)**: describe the project's benefits, which includes who will benefit from it and how. Please be as specific as possible. For example, will it lead to making/saving money, gaining reputation, increasing customer satisfaction, improving services, saving time, automating a task, or serving more people?
3) **Unit of analysis (5%)**: define your unit of analysis, which concerns what one row in your data set stores. For example, if your unit of analysis is a customer, an entire row would capture information about a single customer.
4) **A→B relationship (20%):** first, list each input variable that makes up the A needed for this project, describe each variable briefly, and identify each variable's data type (i.e., whether continuous or categorical). Next, identify the output variable that makes up the B. Describe this variable briefly and identify its data type (i.e., whether continuous or categorical).
5) **Self-check (20%)**: check the decision flow in Figure 1 and ensure the validity of the A→B relationship. Discuss why based on Figure 1.
6) **Data sources (10%)**: briefly discuss where the data on A and B exist and how you plan to retrieve them. If data on A and/or B do not exist, briefly explain how you can capture them.
7) **Fictitious table (10%)**: create a fictitious table comprising all input and output variables and include five fictitious records.
8) **Data set (5%):** discuss the size of the data set (in number of records) you can find for this ML project. Also, discuss whether the data set's size is adequate to build ML models (if not, discuss what you can do about it).

4.4 Grading Rubric

Table 1. Question 1: Goal

Grade	Description
Excellent (100%-90%)	The goal is well defined. Its description is coherent and organized. It is specific such that it can be achieved using the proposed data.
Satisfactory (89%-70%)	The goal lacks clarity. It is coherent but not specific. It is not sufficiently aligned with the proposed data.
Unsatisfactory (69% or less)	The goal is not appropriate for one or more of the following reasons: it is not defined well, it is incoherent, it is too general, and/or it cannot be achieved using the proposed data.

Table 2. Question 2: Benefits

Grade	Description
Excellent (100%-90%)	The benefits are well articulated. The beneficiaries are identified. The importance of the benefits is discussed clearly.
Satisfactory (89%-70%)	The discussion of benefits needs more clarity. The project's beneficiaries and/or the benefits' importance are ambiguous.
Unsatisfactory (69% or less)	The discussion of benefits is not appropriate for one or more of the following reasons: it is not articulated well, it is incoherent, the beneficiaries are not identified, and/or the benefits' importance is not identified.

Table 3. Question 3: Unit of Analysis

Grade	Description
Excellent (100%-70%)	The unit of analysis is identified correctly.
Unsatisfactory (69% or less)	The unit of analysis is not identified correctly.

Table 4. Question 4: A→B Relationship

Grade	Description
Excellent (100%-90%)	A comprehensive list of the input variables is provided. For each input variable, its description and data type are articulated clearly. The output variable is described well and its data type is indicated clearly.
Satisfactory (89%-70%)	A comprehensive list of the input variables is provided; however, their descriptions and/or data types are not clear. The output variable and its data are identified, but the description is ambiguous.
Unsatisfactory (69% or less)	The list of input and output variables is insufficient for one or more of the following reasons: the input variables and/or output variable are missing; descriptions are incoherent or missing; and/or the data types are ambiguous, incorrect, or missing.

Table 5. Question 5: Self-check

Grade	Description
Excellent (100%-90%)	The discussion is well organized, insightful, and coherent. Evidence from Figure 1 is provided to demonstrate that A→B relationship is valid.
Satisfactory (89%-70%)	The discussion lacks clarity. It is coherent but mostly descriptive. Evidence from Figure 1 is not used effectively.
Unsatisfactory (69% or less)	The discussion is insufficient for one or more of the following reasons: it is not articulated well, it is incoherent, it does not provide any evidence from Figure 1, and/or the evidence provided from Figure 1 is not appropriate.

Table 6. Question 6: Data Source

Grade	Description
Excellent (100%-90%)	The data sources are identified clearly. If they do not exist, a clear explanation is provided on how to capture A and B.
Satisfactory (89%-70%)	The data sources are ambiguous, and it is not clear whether data already exist or will be captured.
Unsatisfactory (69% or less)	The data sources are not identified, and no discussion is provided on data availability.

Table 7. Question 7: Fictitious Table

Grade	Description
Excellent (100%-90%)	The fictitious table includes all input and output variables. The row values provided in the table are compatible with the variables' description and data types.
Satisfactory (89%-70%)	The fictitious table includes most of the input and output variables. Some of the row values provided in the table are incompatible with the corresponding variables' description and data types.
Unsatisfactory (69% or less)	The fictitious table is insufficient for one or more of the following reasons: most of the variables are missing, most rows or their values are missing, and/or most row values are incompatible with the corresponding variables' descriptions and data types.

Table 8. Question 8. Data Set

Grade	Description
Excellent (100%-70%)	The data set size is adequate for the project.
Unsatisfactory (69% or less)	The discussion is insufficient for one or more of the following reasons: the data's size is ambiguous, the data's size is insufficient, and/or the data's size is not discussed.

4.5 Example Solutions

I provide two example solutions in Appendices D and E for reference. The solution in Appendix D has a time difference between A and B, while the solution in Appendix E focuses on reducing costs, manual effort, or subjectivity in capturing B.

4.6 Proof of Concept

I used the assignment in an ML course for students in a master of science (MS) degree in management information systems (MIS) at a university in the United States. Overall, 18 students submitted an assignment. One submission did not earn any points because it failed the first step in Figure 1. Accordingly, it proposed a rule-based A→B relationship in which A could have derived the value of B using a set of rules despite the self-check that required the student to check the A→B relationship's validity using Figure 1. Out of the remaining 17 valid submissions, only two concerned automation and focused on reducing costs, manual effort, or subjectivity in generating the value of B. The remaining valid submissions (83%) involved a time difference between A and B and, thus, focused on predicting the future.

The average grade the valid submissions earned was 82.1 percent (with a standard deviation of 25.8%). The minimum grade was 50 percent and the maximum grade was 100 percent. Some common mistakes in these submissions:

- Had a misalignment between the project's goal and what the model predicted (i.e., B).
- Included date, time, or timestamp variables in A (rather than a time difference based on a reference date/time).
- Included certain variables in A that one could not observe before observing B.
- Included one or more variables in A that had unique values.
- Had a misalignment between variable descriptions and the values used in the fictitious table.
- Had a misalignment between the variables listed and the variables in the fictitious table.
- Conceptualized inherently multi-class B (i.e., having more than two categories) as binary-class B.
- Used an insufficient data set size.

While some mistakes above represent egregious ones, they also provide further justification for this assignment. Students, no matter how many ML courses they take in their curricula, have a hard time creating and conceptualizing a project from scratch. Educators usually provide students with assignments that require them to use ML algorithms on sanitized data sets. While these assignments prepare students for building predictive models and interpreting results, they do little to encourage critical and creative thinking. As a result, when students need to conceptualize a new project, they falter on some fundamental concepts that play a critical role in ML projects.

5 Lessons Learned

The assignment as it currently stands took shape over six years as I repeatedly used it in graduate-level data mining and machine learning courses. I have assigned it to both master of business administration (MBA) and MS in MIS students with success. Below, I share some lessons I learned while developing the assignment and using it in its current format.

5.1 Identifying a Project

In earlier assignment versions, several students misinterpreted the instructions and identified ML projects by conducting online searches. Their submissions borrowed from generic tutorials on data hosting or blogging websites. Therefore, I specifically ensured the instructions indicated that they should not identify an ML project through an online search. However, instructors may find it difficult to enforce this restriction given that many online ML tutorials exist. Furthermore, students could propose completed ML projects or ML projects that had already begun in their workplaces. Unfortunately, instructors also cannot easily detect such situations, which might undermine this assignment's learning objectives.

Furthermore, despite my expectations, certain students did not choose to propose ML projects related to their workplaces even though the nature of their jobs allowed them to do so. Therefore, the instructions provide flexibility for students to propose any project that interests them. Instructors who use this assignment should not expect all students with corporate jobs to propose ML projects related to their workplaces.

5.2 Implementing the Project

In earlier assignment versions, I provided students with the opportunity to earn extra credit if they obtained or collected the data proposed in the assignment and built at least two machine learning models after I positively assessed the validity of their A→B relationship. However, students ended up finding this task challenging because most students who chose this option could not obtain the data for personal reasons or due to difficulties in accessing or obtaining data from corporate databases. Therefore, the assignment's current version does not offer this option. Interested faculty members can include this option if they would like to convert this assignment into a capstone assignment (or project) that spans multiple semesters.

5.3 Group Work

Over the years, I have allowed students to complete the assignment in groups (each comprising two students) in addition to individual work. I have found that a well-functioning group usually proposes a much more thought-out ML project than individuals do. However, group work also limits the types of ML projects groups propose. Specifically, group members usually are reluctant to share information about their organizations with other group members. Therefore, group work mostly leads to ML project proposals concerning personal interests or hobbies.

5.4 Delivering Concepts

To help students achieve the assignment's learning objectives, I recommend that instructors discuss the framework and the accompanying decision flow that I propose in this paper with students either in person or online (such as a pre-recorded video lecture). Such a discussion might require anywhere between 30 minutes to an hour depending on its depth and breadth. I strongly recommend that instructors provide and dissect many example ML projects during this discussion.

5.5 Length of Student Submissions

I have found that students will often approach this assignment as an essay assignment and, thus, provide lengthy answers. Some submissions described goals and benefits on multiple pages. Even though students might need such descriptions to provide context or clarify business models or processes, they also make grading challenging. Therefore, interested instructors can provide a word/sentence/line limit for these questions to encourage students to provide concise but coherent descriptions. However, certain contexts require students to provide some background that renders some submissions vague or incomplete when omitted. To entertain such cases, I encourage instructors to add an optional "background" section so that students can provide essential details about a process or business model that their project discusses.

5.6 Fictitious Table

Earlier assignment versions did not require students to provide a fictitious table, but I found that adding this requirement benefitted both students and me in several ways. First, the table helped students visualize the data set and resolve any conceptualization errors. For example, students had an opportunity to refine variables and their descriptions after they entered values in the table. Second, the table allowed

me to clarify ambiguities in variable descriptions, which occurred when students provided concise variable descriptions, particularly if they had too many variables to discuss. Third, a fictitious table allowed a student to propose a work-related project without any concern for data privacy or confidentiality. Such confidentiality became particularly important when students wanted to know the feasibility of a project in their workplace using a specific data set. Therefore, rather than sharing the actual data, students could provide a table that contained fictitious values.

5.7 Class Discussion

Instructors can convert the assignment into a class discussion if they provide students with an opportunity to present their projects to their classmates. This opportunity might allow students to not only obtain feedback from their peers but also clarify any ambiguities. Class discussions also allow instructors to review general ML concepts as they arise.

5.8 Feedback from Students

Over the years, students who had corporate jobs responded positively to the assignment for two reasons. First, it allowed them to think about an ML project that could not only showcase their skills but also impact their work. In this regard, the assignment proved particularly helpful to students who wanted to incorporate ML into their job responsibilities but did not know where to start. Second, it allowed students to obtain feedback for potential ML projects that they had to develop in their workplaces. Given that many corporations today ask their employees to work as citizen data scientists to tackle business problems using business analytics and machine learning, the assignment provides a tremendous opportunity to obtain feedback from expert faculty on an ML project's viability.

6 Conclusion

Machine learning courses have become a major component in many degree programs in higher education. Even though students who take ML courses might know how to build ML models, they still can face difficulties in conceptualizing an ML project from scratch. In many instances, students might not know how to conceptualize a valid A→B relationship for the problem at hand and, thus, propose incorrect or moot ML projects. In this paper, I address this issue by proposing a five-step decision flow so that students can assess the validity of an A→B relationship. I provide a course assignment that can reinforce this decision flow and allow students to propose an ML project from scratch. I discuss how instructors can use this assignment and some lessons I have earned over the years.

When tasking students with this assignment, instructors should encourage them to think about issues they encounter in their everyday lives or at work to identify potential A→B relationships. Another approach to identifying potential relationships could be to interview others. For example, one could ask friends, colleagues, or any stakeholder of salient goods and services questions about the issues they face in their day-to-day lives. Such interviews might also prepare students to run systematic and structured meetings, such as creativity or brainstorming sessions, in their current or prospective organizations. If they can bring subject matter experts to these sessions, they may increase the likelihood that they will identify potential ML ideas and, in particular, possible A→B relationships. They can then apply the decision flow that I present in this paper to those ideas to identify valid ones for ML.

References

Brislin, R. W. (1980). Cross-cultural research methods. In I. Altman, A. Rapoport, & J. F. Wohlwill (Eds.), *Environment and culture* (pp. 47-82). Springer.

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science, 358*(6370), 1530-1534.

Chauhan, G., Liao, R., Wells, W., Andreas, J., Wang, X., Berkowitz, S., Horng, S., Szolovits, P., & Golland, P. (2020). Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment. In A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, & L. Joskowicz (Eds.), *Medical image computing and computer assisted intervention* (LNCS vol. 12262). Springer.

Coursera. (2020). *2020's most popular courses*. Retrieved from https://www.coursera.org/collections/popular-courses-2020

CSIRO. (2020). *CSIRO and Microsoft partner to tackle plastic waste, illegal fishing, and efficient farming*. Retrieved from https://www.csiro.au/en/news/News-releases/2020/CSIRO-and-Microsoft-partner-to-tackle-plastic-waste-illegal-fishing-and-efficient-farming

Dave, P. (2021). Google phone cameras will read heart, breathing rates with AI help. *Reuters*. Retrieved from https://www.reuters.com/article/us-alphabet-google-health-idUSKBN2A427P

Davenport, F. V., & Diffenbaugh, N. S. (2021). Using machine learning to analyze physical causes of climate change: A case study of U.S. Midwest extreme precipitation. *Geophysical Research Letters, 48.*

Duporge, I., Isupova, O., Reece, S., Macdonald, D. W., & Wang, T. (2021). Using very-high-resolution satellite imagery and deep learning to detect and count African elephants in heterogeneous landscapes. *Remote Sensing in Ecology and Conservation*, 7(3), 369-381.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM, 39*(11), 27-34.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

Kaloop, M. R., Bardhan, A., Kardani, N., Samui, P., Hu, J. W., & Ramzy, A. (2021). Novel application of adaptive swarm intelligence techniques coupled with adaptive network-based fuzzy inference system in predicting photovoltaic power. *Renewable and Sustainable Energy Reviews*, *148, 111315*.

Khan, A. N., Ihalage, A. A., Ma, Y., Liu, B., Liu, Y., & Hao, Y. (2021). Deep learning framework for subject-independent emotion detection using wireless signals. *PLOS ONE, 16*(2), e0242946.

Kohavi, R., Provost, F. (1998) Glossary of terms. *Machine Learning*, *30*(2-3), 271-274.

Kotsiantis, S. B., Pierrakeas, C. J., & Pintelas, P. E. (2003). Preventing student dropout in distance learning using machine learning techniques. In *Proceedings of the 7th International Conference on Knowledge-Based Intelligent Information and Engineering Systems.*

Koutsouleris, N., Dwyer, D. B., Degenhardt, F., Maj, C., Urquijo-Castro, M. F., Sanfelici, R., Popovic, D., Oeztuerk, O., Haas, S. S., Weiske, J., Ruef, A., Kambeitz-Ilankovic, L., Antonucci, L. A., Neufang, S., Schmidt-Kraepelin, C., Ruhrmann, S., Penzel, N., Kambeitz, J., Haidl, T. K., Rosen, M., Chisholm, K., Riecher-Rössler, A., Egloff, L., Schmidt, A., Andreou, C., Hietala, J., Schirmer, T., Romer, G., Walger, P., Franscini, M., Traber-Walker, N., Schimmelmann, B. G., Flückiger, R., Michel, C., Rössler, W., Borisov, O., Krawitz, P. M., Heekeren, K., Buechler, R., Pantelis, C., Falkai, P., Salokangas, R. K. R., Lencer, R., Bertolino, A., Borgwardt, S., Noethen, M., Brambilla, P., Wood, S. J., Upthegrove, R., Schultze-Lutter, F., Theodoridou, A., Meisenzahl, E., & Consortium, P. (2021). Multimodal machine learning workflows for prediction of psychosis in patients with clinical high-risk syndromes and recent-onset depression. *JAMA Psychiatry, 78*(2), 195-209.

Kovalenko, E., Talitckii, A., Anikina, A., Shcherbak, A., Zimniakova, O., Semenov, M., Bril, E., Dylov, D. V., & Somov, A. (2020). Distinguishing between Parkinson's Disease and essential tremor through video analytics using machine learning: A pilot study. *IEEE Sensors Journal*, *21*(10), 11916-11925.

Kwon, Y. J., Toussie, D., Finkelstein, M., Cedillo, M. A., Maron, S. Z., Manna, S., Voutsinas, N., Eber, C., Jacobi, A., Bernheim, A., Gupta, Y. S., Chung, M. S., Fayad, Z. A., Glicksberg, B. S., Oermann, E. K., & Costa, A. B. (2021). Combining initial radiographs and clinical variables improves deep learning prognostication in patients with COVID-19 from the emergency department. *Radiology: Artificial Intelligence, 3*(2).

Landing AI. (2020). *Landing AI creates an AI tool to help customers monitor social distancing in the workplace*. Retrieved from https://landing.ai/landing-ai-creates-an-ai-tool-to-help-customers-monitor-social-distancing-in-the-workplace/

Lassau, N., Ammari, S., Chouzenoux, E., Gortais, H., Herent, P., Devilder, M., Soliman, S., Meyrignac, O., Talabard, M.-P., Lamarque, J.-P., Dubois, R., Loiseau, N., Trichelair, P., Bendjebbar, E., Garcia, G., Balleyguier, C., Merad, M., Stoclin, A., Jegou, S., Griscelli, F., Tetelboum, N., Li, Y., Verma, S., Terris, M., Dardouri, T., Gupta, K., Neacsu, A., Chemouni, F., Sefta, M., Jehanno, P., Bousaid, I., Boursin, Y., Planchet, E., Azoulay, M., Dachary, J., Brulport, F., Gonzalez, A., Dehaene, O., Schiratti, J.-B., Schutte, K., Pesquet, J.-C., Talbot, H., Pronier, E., Wainrib, G., Clozel, T., Barlesi, F., Bellin, M.-F., & Blum, M. G. B. (2021). Integrating deep learning CT-scan model, biological and clinical variables to predict severity of COVID-19 patients. *Nature Communications, 12*.

Linoff, G. S., & Berry, M. J. (2010). *Data mining techniques: For marketing, sales, and customer relationship management* (3rd ed.). John Wiley & Sons.

Mitchell, T. M. (1999). Machine learning and data mining. *Communications of the ACM*, *42*(11), 30-36

Mnih V., Kavukcuoglu K., Silver D., Rusu A. A., Veness J., Bellemare M. G., Graves A., Riedmiller M., Fidjeland A. K., Ostrovski G., Petersen S., Beattie C., Sadik A., Antonoglou I., King H., Kumaran D., Wierstra D., Legg S., Hassabis D. (2015). Human-level control through deep reinforcement learning. *Nature, 518*, 529-533.

Mortazavi, B. J., Downing, N. S., Bucholz, E. M., Dharmarajan, K., Manhapra, A., Li, S.-X., Negahban, S. N., & Krumholz, H. M. (2016). Analysis of machine learning techniques for heart failure readmissions. *Circulation: Cardiovascular Quality and Outcomes, 9*(6), 629-640.

Murali, P., Hernandez, J., McDuff, D., Rowan, K., Suh, J., & Czerwinski, M. (2021). AffectiveSpotlight: Facilitating the communication of affective responses from audience members during online presentations. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems.*

Ng, A. (2016). What artificial intelligence can and can't do right now. *Harvard Business Review.* Retrieved from https://hbr.org/2016/11/what-artificial-intelligence-can-and-cant-do-right-now

Novak, J., Zarinabad, N., Rose, H., Arvanitis, T., MacPherson, L., Pinkey, B., Oates, A., Hales, P., Grundy, R., Auer, D., Gutierrez, D. R., Jaspan, T., Avula, S., Abernethy, L., Kaur, R., Hargrave, D., Mitra, D., Bailey, S., Davies, N., Clark, C., & Peet, A. (2021). Classification of paediatric brain tumours by diffusion weighted imaging and machine learning. *Scientific Reports, 11*(1).

Olivier, C., Bernhard, S., & Alexander, Z. (2006). Introduction to semi-supervised learning. In O. Chapelle, B. Scholkopf, & A. Zien (Eds.), *Semi-supervised learning* (pp. 1-12). MIT Press.

Parmar, H., Nutter, B., Long, R., Antani, S., & Mitra, S. (2020). Spatiotemporal feature extraction and classification of Alzheimer's disease using deep learning 3D-CNN for fMRI data. *Journal of Medical Imaging, 7*(5).

Perry, M. (2018). Data scientists in demand. *The Chronicle of Higher Education*. Retrieved from https://www.chronicle.com/article/data-scientists-in-demand/

Rogers, J. (2021). Moving beyond the self-reported scale: Objectively measuring chronic pain with AI. *IBM*. Retrieved from https://www.ibm.com/blogs/research/2021/01/nans-measuring-chronic-pain/

Shah, S. W., Kanhere, S. S., Zhang, J., & Yao, L. (2021). VID: Human identification through vein patterns captured from commodity depth cameras. *IET Biometrics, 10*(2), 142-162.

Silver, D. H., Feder, M., Gold-Zamir, Y., Polsky, A. L., Rosentraub, S., Shachor, E., Weinberger, A., Mazur, P., Zuki, V. D., & Bronstein, A. M. (2020). *Data-driven prediction of embryo implantation probability using IVF time-lapse imaging.* Retrieved from https://arxiv.org/abs/2006.01035

Sonawane, D., Miyapuram, K. P., Rs, B., Lomas, D. J. (2020). Guessthemusic: Song identification from electroencephalography response. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data.*

Treiber, J. (2021). How AI can be the key to loyalty and retention for retail in 2021. *Forbes*. Retrieved from https://www.forbes.com/sites/jonathantreiber/2021/02/02/how-artificial-intelligence-can-be-the-key-to-loyalty-and-retention-in-2021/?sh=4de2ce8f12aa

Ulloa Cerna, A. E., Jing, L., Good, C. W., vanMaanen, D. P., Raghunath, S., Suever, J. D., Nevius, C. D., Wehner, G. J., Hartzel, D. N., Leader, J. B., Alsaid, A., Patel, A. A., Kirchner, H. L., Pfeifer, J. M., Carry, B. J., Pattichis, M. S., Haggerty, C. M., & Fornwalt, B. K. (2021). Deep-learning-assisted analysis of echocardiographic videos improves predictions of all-cause mortality. *Nature Biomedical Engineering*, *5*,546-554.

Vieira, R. B., & Lambros, J. (2021). Machine learning neural-network predictions for grain-boundary strain accumulation in a polycrystalline metal. *Experimental Mechanics*, *61*(4), 627-639.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.*

Wladawsky-Berger, I. (2019). The current state of AI adoption. *The Wall Street Journal*. Retrieved from https://www.wsj.com/articles/the-current-state-of-ai-adoption-01549644400

Yala, A., Mikhael, P. G., Strand, F., Lin, G., Smith, K., Wan, Y.-L., Lamb, L., Hughes, K., Lehman, C., & Barzilay, R. (2021). Toward robust mammography-based models for breast cancer risk. *Science Translational Medicine, 13*(578).

Zeleznik, R., Foldyna, B., Eslami, P., Weiss, J., Alexander, I., Taron, J., Parmar, C., Alvi, R. M., Banerji, D., Uno, M., Kikuchi, Y., Karady, J., Zhang, L., Scholtz, J.-E., Mayrhofer, T., Lyass, A., Mahoney, T. F., Massaro, J. M., Vasan, R. S., Douglas, P. S., Hoffmann, U., Lu, M. T., & Aerts, H. J. W. L. (2021). Deep convolutional neural networks to predict cardiovascular risk from computed tomography. *Nature Communications, 12*.

Appendix A

Table A1. Sample Materials to Teach Data Mining and Machine Learning

Resource	Author	Type	URL
Tutorials	Google	ML	https://developers.google.com/machine-learning/crash-course
Tutorials	Amazon	ML	https://docs.aws.amazon.com/machine-learning/
Textbook & Tutorials	Aurelien Geron	ML	https://github.com/ageron/handson-ml2
Textbook & Tutorials	Jason Brownlee	ML	https://machinelearningmastery.com/
Tutorials	Varol Kayhan	DM	http://sas-book.com/
Course	Andrew Ng	ML	https://www.coursera.org/learn/machine-learning
Course	John C. Hart	DM	https://www.coursera.org/specializations/data-mining

Appendix B

Table B1. Use Cases with a Time Difference Between A and B

A	B	Source
Lab tests and chest X-rays at T1	Intubation or death within 30 days (T2) of hospital admission	Kwon et al. (2021)
Mammograms at T1	Development of breast cancer at T2	Yala et al. (2021)
Time-lapse videos of developing embryos at T1	Success of embryo implantation at T2	Silver et al. (2020)
GPS and heart rate data collected at T1	Identification of sports injury at T2	https://zone7.ai/case-studies/
Images obtained from echocardiographic videos at T1	Identification of mortality at T2	Ulloa Cerna et al. (2021)
Biological and clinical variables collected at T1	Development of psychosis at T2	Koutsouleris et al. (2021)
Large-scale atmospheric circulation at T1	Extreme precipitation (i.e., rain) at T2	Davenport & Diffenbau (2021)
Climatic variables captured at T1	Photovoltaic power generated by solar panels at T2	Kaloop et al. (2021)
Stainless steel's granular microstructure observed at T1	Stainless steel's behavior when strained at T2	Vieira & Lambros (2021)

Appendix C

Table C1. Use Cases in Which the Original Process that Captures B is Costly, Labor Intensive, or Subjective

A	B	Source
Video recordings of people performing specific motor tasks	Diagnosis of Parkinson's disease	Kovalenko et al. (2020)
Images of waste and trash taken at waste traps	Classification of trash and identification of plastics	CSIRO (2020)
Functional magnetic resonance imaging (fMRI)	Detection of Alzheimer's disease	Parmar et al. (2020)
X-ray image	Detection of the severity of pulmonary edema (excess fluid in lungs)	Chauhan et al.
Security camera footage	Detection of individuals and identification of social distancing	Landing AI (2020)
Diffusion-weighted imaging produced by magnetic resonance imaging (MRI)	Diagnosis and classification of pediatric brain tumor	Novak et al. (2021)
CT scans	Quantification of coronary artery calcification	Zeleznik et al. (2021)
High-resolution satellite images	Identification of African elephants	Duporge et al. (2021)
Biomarkers	Quantitative measurement of chronic pain	Rogers (2021)
Smartphone camera feed	Identification of heart rate and respiration	Dave (2021)
Faces captured in video feed	Identification of emotions	Murali et al. (2021)
Images of veins taken with a depth camera	Identification of an individual	Shah et al. (2021)
Changes in a person's heartbeat detected by radio waves	Identification of specific feelings	Khan et al. (2021)
Brainwaves recorded using an electroencephalography (EEG)	Identification of a song being listened to	Sonawane et al. (2020)

Appendix D: Example Project with a Time Difference between A and B

1) **Goal:** this project focuses on predicting how long it takes for someone to adopt a rescue animal admitted to an animal shelter (based on variables collected during admission). As a proof of concept, the project focuses only on cat and dog adoptions.
2) **Benefits**: this project's benefits include improved capacity planning and resource allocation. Predictions that a successful model make will allow caretakers to determine how much spare space they will have in subsequent days and how much food and supplies they will need. Furthermore, a successful model will help caretakers expedite animal adoptions.
3) **Define your unit of analysis**: the unit of analysis is a single rescue animal.
4) **A→B relationship:**

Table D1. Input Variables (A)

Variable	Description	Type
Species	The animal's species	Categorical
Breed	The animal's breed	Categorical
Color	The animal's color	Categorical
Age	The animal's age in years (at the time of admission)	Numeric (continuous)

Table D2. Output Variable (B)

Variable	Description	Type
Adoption time	The time it takes (in days) for the animal to be adopted	Numeric (continuous)

5) **Self-check**: in this relationship, a time difference exists between the inputs and output. While the inputs are observed at the time of admission, the output is observed at a later point in time.
6) **Data sources**: one can obtain the values of the inputs and output from the shelter's database.
7) **Fictitious table**:

Table D3. Fictitious Table

Species	Breed	Color	Age	Adoption time
Dog	Bulldog	Black	3	7
Cat	Maine Coon	Brown	5	10
Dog	Terrier	Brown	4	20
Cat	Persian	Black	2	3
Dog	Pitbull	White	5	15

8) **Data set:** the shelter processes roughly 300 animals per year. A year's worth of data would not be sufficient to build a model; therefore, data collected over several years should be sufficient to build a preliminary model.

Appendix E: Example Project with Costly, Labor-intensive, or Subjective B

1) **Goal:** this project focuses on predicting rescue animals' ages when a shelter admits them. As a proof of concept, the project focuses only on cat and dog adoptions.
2) **Benefits**: this project's benefits include automation and improved care. A trained veterinarian estimates each rescue animal's age by examining the animal's teeth when a shelter admits it—difficult, costly, and subjective process. Predictions that a successful model make can allow a shelter to automate the process and reduce costs and subjectivity. Furthermore, a successful model can allow caretakers to provide age-appropriate care to animals in the shelter.
3) **Define your unit of analysis**: the unit of analysis is a single animal.
4) **A→B relationship:**

Table E1. Input Variable (A)

Variable	Description	Type
Teeth	Image of teeth taken during admission	Image

Table E2. Output Variable (B)

Variable	Description	Type
Age	Animal's age in years	Numeric (continuous)

5) **Self-check**: in this relationship, capturing the output is costly, labor intensive, and subjective because a trained veterinarian determines animals' ages by examining the teeth for signs of age. If one can build a model to perform this process, it will function as a proxy for this veterinarian and make such determinations automatically, which might reduce costs, automate the task, and reduce subjectivity.
6) **Data sources**: the values of inputs and output can be obtained from the shelter's database.
7) **Fictitious table**:

Table E3. Fictitious Table

Teeth	Age
Image1	3
Image2	4
Image3	2
Image4	6
Image5	5

8) **Data set:** the shelter processes roughly 300 animals per year. A year's worth of data would not be sufficient to build a model; therefore, data collected over several years should be sufficient to build a preliminary model.

About the Author

Varol Kayhan is an Associate Professor of Information Systems in the Muma College of Business at the University of South Florida. His research focuses on decision making, machine learning, and organizational knowledge management, and has appeared in *Communications of the ACM, Information & Management, Behavior Research Methods, Big Data*, *Journal of Computer Information Systems,* and others. He authored an ebook on data mining that many universities have adopted as formal course material. He teaches data analytics and machine learning courses in undergraduate, graduate, and executive programs. He holds PhD and MS degrees from the University of South Florida and a BS degree from the Middle East Technical University in Turkey.