

Lesson 8: Model Cards

45 minutes

Overview

In this lesson, students will investigate a model for bias and be introduced to a Model Card, which is a way of representing important information about a trained model that could help uncover bias. They will be investigating a Medical Priority app, which helps a hospital decide how soon to view patients based on their symptoms. As students go through the activity, they realize that the app is biased based on personal information and examine how this could happen.

Question of the Day: How can we evaluate machine learning models once they've been trained?

Standards

Full Course Alignment

AI4K12 National Guidelines 2021

- **BI-3** - Computers can learn from data

CSTA K-12 Computer Science Standards (2017)

- **IC** - Impacts of Computing

Agenda

Warm Up (5 minutes)

Journal

Activity (35 minutes)

Medical Priority App

Prompt 1

Prompt 2

Analyzing Model Cards

Wrap Up (5 minutes)

Journal

Objectives

Students will be able to:

- Read a model card and use it to evaluate a model

Preparation

- Review the Code Studio levels before the lesson
- Anticipate how you will guide the discussion on racial and gender bias, considering that some students may have experiences similar to those discussed in this lesson.
- Check the **"Teacher's Lounge"** forum for verified teachers to find additional strategies or resources shared by fellow teachers

Links

Heads Up! Please make a copy of any documents you plan to share with students.

For the teachers

- **Model Cards** - Slides

▼ Make a Copy

For the students

- **AI: Training Data & Bias** - Video (**Download**)
- **Model Cards** - Activity Guide

▼ Make a Copy

Vocabulary

- **Bias** - When a decision favors some things and de-prioritizes or excludes others

Teaching Guide

Warm Up (5 minutes)

Journal

 **Discuss:** *Imagine you are babysitting a young cousin and they run up to you from playing looking upset and asking for help. They show you that they skinned their knee and it isn't bleeding, but it's very pink. If you had to choose only one way to react, which would you pick?*

(1) Tell them to deal with it and go back to playing

(2) Tell them to wait a minute while you go get some band-aids

(3) Call 911 and ask for an ambulance

Have students journal individually, then have students share by showing on their fingers which option they chose. This situation is purposefully vague, and the class may find itself split between option (1) and option (2).

Teaching Tip

Not A Realistic Choice: Students may point out that this is not a realistic way to choose your reaction. There are plenty of other choices you could make, and the situation doesn't provide any additional context that might be important - for example, perhaps the location where they were playing has an important role in the decision they would make. These feelings are valid and can lead to a nuanced conversation about how difficult it is to make decisions without full context, especially for a machine. For now, encourage students to choose the best of the possible options while acknowledging that the realistic version of this choice can be much more complex.

 **Discuss:** *What kind of injury would lead you to make a different choice?*

Have students think individually first, then share with a partner before having a full-group discussion.

Discussion Goal: Students should describe how different situations call for different decisions. As students share, highlight the features that they are using to make these decisions - for example, whether or not there is blood or whether or not their cousin is crying. Allow several students to share their answers so a wide variety of ideas are heard. Use some of what you heard as examples in your remarks to transition to today's activities.

Remarks

Thank you all for sharing your ideas! I heard a lot of different features that helped you make a decision, but I also heard some people make different decisions even in the same situation. This same thing can happen with computers too - the decisions it makes can be different from the ones we expect, even in very similar situations. Today, we'll look at how we can test a computer model to see what kinds of decisions its making and make sure those decisions are fair.

 **Question of the Day:** How can we evaluate machine learning models once they've been trained?

Activity (35 minutes)

Medical Priority App (15 minutes)

 **Code Studio:** Have students log into Code Studio. The second level contains an app that they will use for the first part of this lesson.

 1

Medical Priority App

Display: Show students the information about this app and read it aloud. Demonstrate how to use the app by entering different feature information then pressing predict and seeing the result. Clarify for students what the racial abbreviations mean:

- BAM: Black / African American
- W: White
- A: Asian
- AIAN: American Indian / Alaskan Native
- NHPI: Native Hawaiian / Pacific Islander

Once students see how the app works, continue to the next slide with specific tasks to investigate

Prompt 1

 **Display:** *Change the features of the app and test what happens when different people enter the hospital.*

- *Find a person who will be admitted as “priority”.*
- *Find a person who will be admitted as “normal”*
- *Find a person who will be admitted as “return later”*

Circulate: Circulate around the room as students engage with the app. Ask them what features they’re selecting to try and get a “priority” recommendation, or a “return later” recommendation. Students will probably notice that certain features are more likely to lead to a priority recommendation, specifically: a patient that is bleeding, is short of breath, and has pain in their head or chest is highly likely to have a priority recommendation. Students may also notice that certain demographic information also seems to affect the recommendation, but it’s okay if they don’t bring this up.

 Teaching Tip

What's My Priority? Students may try to use their own demographic information in the app, but this is explicitly *not* the purpose of this exercise. Remind students that we want to test this in a wide variety of scenarios, which means testing with a wide variety of people and not just people who match their age, gender, and race.

 **Share Out:** Ask a few students to share what features led to priority recommendations, and if that recommendation seems to make sense.

Prompt 2

 **Display:** The hospital wants to check the result under very specific situations. The table in the slide shows four different groups. Each column represents a different person that the group will test.

Assign students to different groups: A, B, C, or D. Each person will only test the person in their group. Everyone in the same group should get the same result.

 **Share Out:** Give students a few minutes to find their results, then ask a person from each group to share the results with the class. Record the results somewhere prominent where students can see them. Students should report out the following answers:

- Group A: normal
- Group B: priority
- Group C: priority
- Group D: priority

 **Display:** Go to the next slide with Prompt 2A and read the directions aloud. Have students re-run the app on this new person. Students should only need to change one or two features to generate the new priority, so this shouldn't take very long.

 **Share Out:** Give students a few minutes to find their results, then ask a person from each group to share the results with the class. Draw a line through the previous answers and write the new ones next to them - however, make sure both answers are clearly visible for the next discussion. Students should discover the following answers:

- Group A: normal --> priority
- Group B: priority --> normal
- Group C: priority --> return later
- Group D: priority --> return later

 **Discuss:** *All of the results changed. Do those changes seem like they make sense? Or do some of them seem surprising?*

Have students talk with a neighbor first before asking students to discuss as a full group. Give students space to share their own observations and reflect on each other's responses, and try to avoid validating responses as "right" or "wrong".

Discussion Goal: Students may try to rationalize the changes to Group A and Group B as making medical sense. Students may struggle to justify the changes to Group C and Group D, since none of the medical information changed - only the personal information. Encourage students to consider what this means in the real world - for example, with Group D: a man with these symptoms can walk into a hospital and be seen as a priority, but a woman with the same symptoms would be told to return later.

Teaching Tip

Lived Experience: Students may have strong reactions to this prompt as it highlights how a decision is being made based on personal information rather than objective information. They may have their own stories and experiences where they've seen different decisions being made based on someone's age or gender or skin tone. This personal connection is an important part of the lesson, and letting students share these stories can be a powerful bridge to the next activity.

When facilitating this discussion, a good mantra to keep in mind is: **assume there's someone in the room who is your data point.** Sometimes a discussion can slip into focusing on data and numbers and hypotheticals without acknowledging that there may be students with firsthand experience with medical discrimination in the room. Consider how you would like to address this and create an inclusive space for discussion in your classroom.

Remarks

It's a good thing we were hired to investigate this app, because it looks like we've uncovered some disturbing trends in how it was designed. We've now seen a few situations where the app is making different decisions based on your personal information rather than the medical condition you're in. For something as important as a hospital visit, this is unacceptable. This is an example of how machine learning can be biased

Medical Priority in the Real World: The Medical Priority example is based on a real example of algorithmic bias in how people were recommended for treatment. You can hear Dr. Ruha Benjamin explain the real-life example in this clip from the Glad You Asked show, from 15:48-17:52: [Click Here for Video](#). Rather than voicing over the data in the slide deck, you may decide to show this video clip to your students instead. This video isn't intended for a middle school audience, so you may have to follow-up with a discussion prompt or additional explanations to help clarify the clip.

Video: Watch the [AI: Training Data & Bias](#) video

Key Vocabulary:

- **Bias:** When a decision favors some things and de-prioritizes or excludes others

To encourage active engagement and reflection, use one or more of the strategies discussed in the [Guide to Curriculum Videos](#).

Analyzing Model Cards (20 minutes)

Remarks

Having a biased model is clearly a problem, and it's another reason our role as evaluators and testers of machine learning models are so important. This is also a problem that people in the machine learning community are trying to solve. Let's listen to one possible solution, then we'll talk about how we can use this idea when creating our own machine learning models.

Video: Have students watch the Model Card video in the slides. This video is part of a longer show - the clip in the slides is pre-set to only play the relevant clip from 18:33 - 19:26

Model Cards: Model Cards were initially proposed as part of the academic paper [Model Cards for Model Reporting](#). One of the co-authors of this paper, Deb Raji, is featured in this video clip discussing Model Cards.

Since the publication of this paper, additional efforts have been made to focus on accountability and documentation in machine learning. A good overview of these efforts can be found in the [About ML](#) project, which includes real-world examples of Model Cards used in the machine learning industry.

Distribute: Pass out the [Model Cards](#) activity guide to each student.

Code Studio: Have students advance to the next level in Code Studio, which shows the Book Recommender app from yesterday and a partially completed model card.

Do This: Show students the directions for the first page of the activity guide. Students will help create a model card for the Book Recommendation app from yesterday

Circulate: Check in with students as they fill in the model card. Encourage them to consider how the data was collected and which features were used as the basis for filling in the Intended Uses and Limitations sections. For limitations, students may notice that more data should be collected since only 120 students were surveyed, or they

may notice that it was only collected in computer science classes and may not represent students from outside those classes.

 **Display:** Have students turn to the back of their activity guide and display the slides with the overview:

Even though we discovered bias in our original model, there is still a need to help nurses and medical professionals in the ER. Several other companies have created medical priority models that can be used to replace the biased one that we discovered. They've also sent along the Model Cards for us to evaluate. Looking at these model cards, can we find a model that we would recommend to the hospital to help decide medical priority?

Do This: Have students continue to Levels 3-5 in Code Studio. Each level has a model card that students can evaluate. Students should progress through each level independently, evaluating each model card and recording their recommendation on their activity guide.

 Teaching Tip

Stations: Instead of having students examine the model cards in Code Studio, you could consider creating a stations activity where each model card is printed around the room. Students would still complete the activity guide, but can choose which model card to examine and move around the room while doing so.

 3-5

Medical Dataset Model Cards

3

4

5

Circulate: Check in with students as they look through the model cards. Help students understand the format of the card and how the information can be used. Each model is intended to have a flaw in them that makes it difficult to recommend, but it's okay if not every student notices these flaws when first reading the card. Instead, students will have a chance later to discuss their findings with partners and possibly change their minds.

- **Level 3: ER Recommender V1** - this model is only 33% accurate which makes it hard to recommend.
- **Level 4: Northern Lights ER Priority** - this model has features that appear very specific. If students look at the Data Information section, they may notice that the model was trained in a rural hospital in Alaska. This means it probably won't do very well outside of Alaska and in other more diverse areas.
- **Level 5: AI Medical Recommender** - this model is only trained on 10 rows of data (or 10 people in the hospital). Even though it has high accuracy, it probably doesn't have enough data in a real-world scenario

 **Share Out:** Have students share their recommendations for each model. Encourage students to talk to each other and provide different justifications as to why a model should or shouldn't be recommended. Let students know there is a spot on the bottom of the activity guide to reflect, and they have permission to change their minds based on what they hear from their peers.

 Content Corner

Real-World Connection: These example model cards are exaggerated for pedagogical purposes, but similar phenomena can be seen in this article: **Hundreds of AI tools have been built to catch covid. None of them helped.** A few relevant quotes from the article are:

- "Many unwittingly used a data set that contained chest scans of children who did not have covid as their examples of what non-covid cases looked like. But as a result, the AIs learned to identify kids, not covid."
- "In yet other cases, some AIs were found to be picking up on the text font that certain hospitals used to label the scans. As a result, fonts from hospitals with more serious caseloads became predictors of covid risk."

As an extension of this lesson, you may choose to share this article with students. However, the article is not intended for a middle-school audience, so additional scaffolds and supports may be necessary to support student understanding and connecting back to today's lesson.

Wrap Up (5 minutes)

Journal

 **Prompt:** *If you could offer guidance to these companies training machine learning models, what advice would you give them to help create useful and unbiased machine learning models?*

This prompt is provided on the bottom of the activity guide that students were using in class. Students should reflect on the reasons why these models weren't great fits for the hospitals and translate that into advice. Some suggestions might include:

- Avoid demographic information in your model to help avoid bias
- Make sure the data represents the community that will be using the model
- Make sure enough data is collected to ensure it is accurate

Assessment Opportunity

Formative Assessment: The activity guide from today can be collected as form of formative assessment. The reflection question can provide insight into how well students understand how model cards can be useful for evaluating models.



This work is available under a **Creative Commons License (CC BY-NC-SA 4.0)**.

If you are interested in licensing Code.org materials for commercial purposes **contact us**.