Advit Agrawal
Ashoka Horizons Assignment 1

Homework Assignment

Question 1

1. The “Aha!” Moments from Data Science and Machine Learning Readings
After reading those articles and slides, a couple of ideas really stood out and made me look at data science and machine learning differently.
First, I was surprised to see how much raw data can matter compared to the kind of algorithm you use. I always thought the most advanced models were the secret to success. But the Google paper showed that if you have a lot of good data, even simple models can do just as well or better than fancy algorithms working with less data. That changed my perspective. It means collecting and managing data well is often more important than spending all your time tweaking models. If you have enough data, a straightforward approach can get you great results. That’s both surprising and practical.
Second, I realized machine learning isn’t some magical black box. Pedro Domingos explained that it’s really a mix of art and science. Success comes from understanding the basics, like how to avoid overfitting, how to pick good features, and how to balance between making a model too simple or too complicated. That made me see that machine learning is more about careful experimentation and common sense than about secret formulas or advanced math. It was reassuring, because it means anyone willing to put in the work can get good at it, not just people with advanced degrees.
Both of these ideas made me see that while machine learning is powerful, it’s also grounded in real-world practicality. Data is often the real star of the show, and good results come from paying attention to the fundamentals, not just chasing the latest trend.

Question 2

2) Let’s talk about how banks catch fraud using data science and machine learning. It’s a great example of why having lots of data matters so much.
Banks handle millions of transactions every day. Trying to spot the few that are fraudulent is like finding a needle in a haystack. But with enough data, not just transaction amounts and times, but also where people are, what devices they’re using,

and even how they usually behave; banks can train models to notice when something's off.
After doing some research I found that the whole idea of using big data for fraud detection isn't just theory. According to a Google Research article, "The Unreasonable Effectiveness of Data," having massive amounts of data can let even simple models catch fraud patterns that would be invisible otherwise (Google Research, 2009).
What's really interesting is that the data doesn't have to be perfect or all the same. In fact, real-world data is often messy. It comes from different places and might have gaps or errors. But that's actually helpful. The models learn to work with all kinds of information, so they can spot patterns humans would miss.
For example, if someone's credit card suddenly gets used in another country or there's a flurry of small purchases before a big one, the model can flag it as suspicious. It's not just about having a fancy algorithm. If you only have a little bit of data, even the smartest model won't be able to tell what's normal and what's not.
So, in fraud detection, the real power comes from having lots of data, even if it's a bit messy or comes from different sources. That's what makes it possible to catch fraud quickly and keep people's money safe. In this case, the actual data is very important.

Question 3

3) One challenge I found interesting in current machine learning is how it can be hard to understand why a model makes certain decisions. This is sometimes called the "black box" problem. After reading Domingos' "A Few Useful Things to Know About Machine Learning," I realized that even models that work well can be tough to interpret, especially as they get more complicated. If you can't explain how a model arrived at its answer, it's hard to trust it in important situations like medicine or law, where you need to know why a decision was made.
When I was reading about machine learning, I realized something important. These models are really good at finding patterns and making predictions, but they don't always explain how they got there. That means if you just trust the model and don't ask any questions, you might not catch mistakes or see when it's making decisions based on bad information. For someone learning data science, it's crucial to understand this. You don't want to just take the model's word for it. You need to dig in, see how it works, and make sure it makes sense for the problem you're trying to solve.
Humans have a big role to play here. We need to check and question what the models are doing. People with real-world experience can spot when something doesn't add up or when a model might be making decisions based on bad or biased data. In the future, I see humans working alongside machine learning systems, making sure the technology

is used responsibly and that decisions are fair and understandable. Machine learning is a powerful tool, but it works best when it's guided by human insight and common sense.

The Fun Question

(PS: I'm very fond of debating)

If I imagine how a Large Language Model like ChatGPT learns to have conversations, I think it's a bit like preparing for a debate by listening to hundreds of debates on all kinds of topics. Instead of just memorizing facts, you pick up on how people argue, the kinds of things they say, and how they respond to each other. The more debates you hear, the better you get at predicting what someone might say next or how to reply in a way that makes sense.
So, ChatGPT is like a debater who's watched and listened to countless conversations, debates, and discussions. It doesn't just copy what it's heard, but learns patterns and ways of putting ideas together. That's how it can join in a conversation or even debate a topic, sounding more and more human the more it "listens." It doesn't necessarily 'know' everything, it simply builds patterns from the large amount of data that it is trained using.

Side Note: How can I keep learning about data science and machine learning even after this course? I'm very interested in it, but I'm not able to identify the source which would help me learn the best. Are there certain YouTube channels, online courses, or projects I should check out to keep building my skills and understanding?

Github Exercise

I chose the kitchen because it's a busy place where a lot happens every day. From cooking to preparing ingredients and using different tools, the kitchen creates a lot of useful information. Collecting data from the kitchen helps me understand how things work there and also gives me a chance to practice organizing real-life data.

Why I Made Each File

1. ingredients.csv

I made this file to keep track of all the ingredients used in cooking. Knowing what ingredients are used and how much helps with:

- Keeping an eye on what's running low
- Making sure recipes can be repeated correctly
- Understanding what kind of food is being prepared

I included:

- The name of the ingredient
- How much of it was used
- What type of ingredient it is (like vegetable, spice, or meat)

2. utensils.csv

This file shows which kitchen tools were used and how often. Tracking this helps to:

- Know which tools are used the most
- Plan when to clean or replace utensils
- See how the kitchen workflow happens

I recorded:

- The name of the utensil
- How many times it was used

3. nutrition.csv

I added this file to keep basic nutrition information for each ingredient. This is useful for:

- Figuring out how healthy a meal is
- Planning meals based on nutrition

- Comparing different recipes

The information includes:

- Ingredient name
- Calories, protein, fat, and carbs

Why I Used Separate Files

I kept the data in separate files because it makes things simpler and clearer. Each file focuses on one type of information, so it's easier to update and analyze. Also, if I want to add more data later, like appliance use or photos, I can just add new files without mixing everything together.

How I Did the Assignment

Setting Up the Environment

- I installed WSL (Windows Subsystem for Linux) using the steps provided in the assignment pdf
- I downloaded Miniconda and installed it inside Ubuntu using the terminal.
- Then, I created a new Conda environment for this project with Python 3.12 and typed in commands like 'conda create' and 'conda activate' to establish the environment

Collecting and Saving Data

- I created an excel sheet with different tabs for each different CSV file
- Then, I saved these tables as CSV files.

Organizing Files

- I put all the CSV files and this write-up in a folder called = in my GitHub repository.
- I made sure the files are named clearly so anyone can understand what they contain