correlation one

Analyzing Natesh's Coffee Consumption

New York City | Team #1

Rasheed Sabar, Sham Mustafa, Haris Jaliawala, John Benjamin

Business Problem

Correlation One is a company that assesses, connects, and trains advanced data scientists across the globe. It is crucial for Correlation One to gain a better understanding of Natesh's coffee consumption in order to deliver better outcomes to its students. In particular, we are interested in understanding what are patterns in Natesh's coffee consumption, and ultimately, what levels of coffee provide optimal lecture delivery?

[This business problem description example is a little short. Ideally we want you to provide more detail]

Business Impact

With optimal coffee consumption, experts at Correlation One estimate that Natesh can speak up to 5x faster, thereby delivering 5x more jokes and content, on average. Experts say Natesh's jokes are expected to increase student satisfaction and participation by 20%, on average. Delivering 5x more content will also allow Natesh to review previous concepts more frequently and cover more advanced topics. Therefore, understanding Natesh's coffee consumption and predicting the optimal amount prior to each lecture could have a substantial impact on student outcomes.

[It is useful if you can provide an estimated impact of your target/moonshot analysis. If things go exactly as planned, how much does your company benefit? If things go much better than planned, how much does your company benefit?]

Data

The dataset consists of four data points with the following fields:

`date (DateTime)`: date of lecture

`num_coffee (Int)`: number of coffee cups Natesh consumed prior to lecture
`avg_wpm (Float)`: average number of words/minute during lecture
`live_chat (Dict[String])`: a dictionary of all the student chats in #live-chat during that lecture

One advantage of this dataset is that it provides rich information on student chats for every single lecture. One disadvantage of this dataset is that it contains relatively few - only four - data points, which is a very small sample size. Another complication is that #live-chat was created during the fourth lecture, which means that information is missing for the previous lectures. This means that the advantage of the dataset actually isn't an advantage at all, and this is just all in all a pretty mediocre dataset.

[Tell us (at least approximately) how large your dataset is. Highlight both the strengths and weaknesses of your dataset(s).]

Methods

Visualizations

One key component of understanding the patterns in Natesh's coffee consumption is providing high-quality visualizations of his historical coffee consumption. Here are some of the static and interactive visualizations we will provide:

- Natesh's coffee consumption over time (univariate)
- Student outcomes of interest over time (univariate)
- Correlation plot of Natesh's coffee consumption vs. student outcomes
- Heatmap of where Natesh's coffee was sourced from

[Remember that exploratory data analysis (appropriate visualizations) is the first, crucial step in modeling! So you should be describing the EDA that you will be performing on your datasets, not just the "models" you will apply.]

Models

Another key component of our project is determining the association between Natesh's coffee consumption and key student outcomes (satisfaction, learning, engagement… etc.). Ideally, we will be able to identify the optimal amount of coffee to maximize these outcomes of interest. These are some of the methods we will be exploring:

- polynomial regression: since we expect a non-linear relationship between coffee consumption and student outcomes (i.e., too much coffee might impact Natesh's ability to focus).

- heterogeneous treatment effect: since different students may respond to Natesh's coffee consumption in different ways, we want to understand both the outcomes on average, but also within different subpopulations of students.
- neural network: eh, why not (obviously be more descriptive this and have good reason to include in your model).

[If you are going to explore a model, please provide a brief description why you think it is useful for your particular project. You must demonstrate you have a high-level understanding of the method.]

Interface

The final front-end product will feature two landing pages: an Analytics page with visualizations of the historical data, and a Recommendation page, where the results are summarized with a recommendation for the number of coffee cups Natesh should consume prior to lecture. The interface will allow for interactive visualizations of the historical data, so that users can click on a particular lecture day and see the results for that day, or all outcomes over time. If time permits, we will also try to include a feature in the Recommendation section where a user can see what the effect is predicted to be if Natesh consumes X% more or less coffee than recommended.

[You can also include a picture of a hand-drawn sketch of the dashboard to show us that you have thought about what your final product will look like]

This is an example of what our dashboard could look like: (1) a panel with key metrics on student engagement, (2) a plot of Natesh's coffee consumption over time, (3) a plot summarizing the metrics of student engagement for each lecture day, (4) plots highlighting key features of messages in the #live-chat channel.

[This is an example of a pretty good dashboard because it is clear, professional, well-labeled, and relatively simple. Ideally your visualizations will also have an interactive component.]

Bad Examples
This is an example of a BAD dashboard (unprofessional, bad labeling)
This is an example of a BAD dashboard (confusing, unclear)

Milestones

In this section, we provide details on the milestones we intend to achieve in our project. In particular, we have outlined four different versions: we expect to finish Version 1 with 100% probability, Version 2 with ~70% probability, Version 3 with ~20% probability, and Version 4 with ~5% probability (if things go extremely well). We have color-coded our versions as follows: data, analysis, and visualizations, to make it clear that our versions advance on all levels.

Version 1: Build simple dashboard (2 static plots, 2 interactive plots) to understand Natesh's historical coffee consumption and its correlation with 3 student outcomes of interest (engagement, satisfaction, and performance)
Version 2: Build prediction model (+ 1 static plot, 1 interactive plot) and determine what level of coffee consumption is optimal
Version 3: Build appropriate causal inference model to assess the strength of the causal relationship between the factors of interest (include key assumptions in the dashboard)
Version 4: Use additional data from Natesh's classes at Harvard and/or DS4A 2019 to stress-test the model

[Your Version 1 must be achieved with 100% probability. Notice your Version 4 should be very difficult for you! We want to see you really challenge yourself in what the team could achieve. You should have minimum 3 versions and maximum 5 versions for your project. If you are a larger team with a more complicated project, it may be helpful to have more versions to stay organized.]

[You do not have to color-code like this, but your project milestones should ideally contain clear milestones on data, analysis, and visualizations.]

Timeline

Date	Deliverable	Details
Week 1	Team Formation	
Week 2	Work on idea formation	
Week 3	Idea should be finalized & Start on Scoping	
Week 4	Project Scoping Completed	
Week 5	Datasets sourced	
Week 6	Basic EDA/Cleaning of datasets completed	
Week 7	In-depth EDA, jupyter analysis, mockup of frontend	
Week 8	Frontend Design	
Week 9	Backend Design	
Week 10	Front End Infrastructure Complete	
Week 11	Application infrastructure complete	
Week 12	Project complete	

[Teams should provide more detail on your timeline than this example. Ideally tell us who will do what, especially if you're part of a larger team (>4 people).]

Concerns

The primary concerns with our project are (1) that some of our team members do not know anything about data analysis, and (2) that we have basically no data.

[This section should be longer than this and provide more detail, such as what will you do to address the concerns you have.]