Project Name

Alex West, Anusha Praturu

WHY

Problem Statement (What problem is your project/product solving?)

- Describe the problem you are solving.

One of the most ubiquitous datasets used in teaching supervised machine learning is the 1996 Census Income dataset. It includes 14 attributes and is used to predict if someone earns more or less than $50,000. It is an extremely useful teaching tool with a major flaw: the inferences made from the machine learning models are erroneous. Typically, the models show that education is the greatest predictor of income. However, due to recent research, it has been shown that education is not a predictor of income for people of color in the United States. We seek to improve this dataset to make it a more accurate predictor of income, and also improve the output to be more actionable.

- Why is this a compelling and impactful problem to solve? Why is this a big opportunity?

The dataset is outdated and misleading, and thus is a poor predictor of income. Models trained on this data are thereby rendered inaccurate and biased. Further, the arbitrary cutoff of $50,000 is not useful.

- What assumptions are you making about the problem / opportunity? (these assumptions would directly influence the key elements and features you would build in the MVP)

Attempting a solution assumes that additional data exists publicly that would allow us to improve the dataset. It also assumes that we will easily be able to join this data to the existing data with a reasonable amount of effort. It assumes that we will be able to collect data from the same time frame to supplement our existing data source, which itself will be updated.

Target Customer

Who is the primary customer/user? What is the use case? What are key assumptions you are making about the primary customer/ user and use case?

- Be specific in identifying your targeted customer segment for the MVP.

- What is the use case enabled by your MVP for this target customer segment?
- When will you be doing initial customer validation?
- How would you go about to identify potential users / customers?
- How would you go about validating the key assumptions?

Our primary customer is anyone using the existing 1996 US census income dataset in order to train their machine learning models. This includes researchers, data science students, hobbyists, and the like. For ease, we will focus specifically on data science students, primarily those exploring machine learning models for the first time. Our MVP will improve the model accuracy and model utility of any/all projects being developed by these users. We will take an iterative approach to customer validation, soliciting feedback with each step of dataset improvement.

How big would be the impact?

- How would you go about quantifying impact, market opportunity for this particular problem you are trying to solve by building a data science product?

We will quantify the impact of our MVP once we better understand the scale at which the 1996 Income dataset is currently being used.

WHAT

Minimal Viable Product (MVP)

- What is the minimal viable product that you are building that will specifically test the fundamental assumptions you have about the problem and the value of your solution?
    - What is the feature set and why? Did you do customer discovery?
    - What is the user experience and why does it matter to your customer?
- What data science approach would you intend to use for the MVP? (this is NOT UI / UX but technical discussion)
- How would you potentially test the efficacy of the MVP? When would you start testing?

The MVP of our project is an updated and improved training dataset for income prediction. It also includes improved labels to allow for more useful output than a simple > or < $50K indicator. At minimum, it will be an open-source downloadable csv file with all appropriate data documentation. The user experience will include an informative and useful landing page that will explain the utility of the dataset,

how to use it, and sample applications. It will also include documentation, including a whitepaper describing the purpose, methodology, and theory behind the improved dataset. This user experience is important because it mimics the existing open-source datasets currently available. Our goal is to surpass the bare minimum and create a landing page that educates on *why* as well as *what* and *how.*

The data science approach we intend to use for the MVP includes the widely available machine learning tools such as sci-kit learn, jupyter notebooks, etc, that most beginning ML students will have access to. We will create a baseline model using the existing UCI dataset, then iteratively test our new datasets to improve performance on the baseline.

Finally, we will conduct an experiment that quantifiably demonstrates the value of the improved dataset, by training several models before and after our efforts and measuring accuracy scores.

What is the key **differentiation** in how you solve this problem compared to existing solutions and/or approaches?

What is the **competitive landscape** (existing solutions or approaches)? How are these potentially competitors approaching the problem and solution? What is the state of art technical benchmark for this type of problem?

To our knowledge, there are no existing solutions that improve the 1996 income dataset by joining to other publicly available data.

Value Proposition (what value/utility does your project/product provide to your intended users?)

- What value would your data science based MVP bring to the target customer segment?
- How is your solution better and more differentiated than current solution(s) ?
- Why should the market and your target customers care? In other words, why is this compelling to them?

An updated open-source machine learning dataset provides value on multiple levels for our users. One, it provides a newer income prediction dataset, with recent data and more applicable results. Two, more gradation in the income

bucket allows for greater fine tuning of specific algorithms (more teaching opportunities). And finally, including a social justice component allows for a more nuanced discussion of machine learning models from the beginning; as it stands today, most data science students (either formal or self-taught) seek out open-source data to hone their skills. The value in our MVP is that it sets the bar higher for open-source data; adding documentation that takes into account newer research and race-related issues while also teaching machine learning concepts. There is value in learning about the principles of machine learning while also learning about the potential for bias and inaccurate predictions.

Mission

**

HOW

Data sets

- What datasets do you intend to use?
- Are the datasets public?
- What are the datasets attributes / metadata that could make the exploratory data analysis easier / harder?

Our primary dataset will be the US Census - specifically the PUMS (public use microdata sample). These are individual census responses, so we can replicate the original dataset as closely as possible. This data is painfully public :) meaning it is available in a multitude of formats.

Our secondary goal is to add data from 3rd party sources that add other attributes to improve the model. This will require more

Project Management

- What is the role of each member (who will do what specifically)? Who is the project manager and chief facilitator? Who is the resident SME? Who is the product manager? Who is the lead / who are the leads on infrastructure, data engineering, EDA, model exploration and identification, etc?

	Role

Alex	Being the SME on the Census data, supporting model building and dataset creation, writing whitepaper, creating data visualization website
Anusha	Model building and evaluation, writing whitepaper, researching, updating, and supplementing dataset

**Some teams have used pairing principles to assign two members of the team to main tasks.

- What are the strengths and weaknesses of each team member?

	Strengths	Weaknesses
Alex	Subject matter expertise Writing Data viz	Not a native coder
Anusha	Computer science Writing, interpretation	Unfamiliar with census data

- What are the guiding values of the team? How would the team work together? What's the operating rhythm? How would you resolve conflicts? How do you communicate?
- How would you plan to implement agile methods, and what tools will you use?

Guiding values and shared commitments are outlines in Team Process Agreement

Technical Approach

- What methodologies would you use for initial data exploratory analysis to ensure your datasets are sufficient and meaningful?
- What data science algorithms are you intending to develop and build for the project? What challenges do you potentially foresee?
- What help do you need?

We'll be using the machine learning capabilities that we developed in W207 in order to develop models to evaluate the success of our dataset creation. We will also be

conducting a thorough EDA on all datasets that we use and include in our endeavor. We anticipate facing challenges on developing a strategy to join third party datasets to our main dataset, and may solicit help in this task.