**DS-SF-34 | Final Project | 2 | Experimental**

**PROMPT**

**Objectives:**

* Create an outline of your research design approach, including hypothesis, assumptions, goals, and success metrics.
* Confirm your data and create an exploratory data analysis notebook with statistical analysis and visualization.

In this project, you will complete the problem statement and research design outline for your final project. This will serve as the starting point for your analysis. Make sure to include a specific aim and hypothesis, well-defined risks and assumptions, and clearly articulated goals and success metrics. You will create a Jupyter notebook that explores your dataset mathematically.

**DELIVERABLES**

**Project Design Write-up**

* **Requirements:**
  + Well-articulated problem statement with "specific aim" and hypothesis, based on your lightning talk.
  + An outline of any potential methods and models.
  + Detailed explanation of the available data. (i.e., build a data dictionary or link to pre-built data dictionaries)
  + Describe any outstanding questions, assumptions, risks, and caveats.
  + Demonstrate domain knowledge, including specific features or relevant benchmarks from similar projects.
  + Define your goals and criteria, in order to explain what success looks like.

Below is a guide to help you complete the project design write-up. The questions for each section are merely there to suggest what the baseline should cover; be sure to use detail as it will make the project much easier to approach as the class moves on.

* Project Problem and Hypothesis
  + What's the project about? What problem are you solving?
    - ***Predicting where user groups are traveling more frequently based on previous purchasing behavior.***
    - ***Problem: I work at a company in the travel industry. Many of our clients are concerned about how to spend their marketing budget. Since we collect a lot of data on their travelers, we should advise them on where certain groups of travel often visit.***
  + Where does this seem to reside as a machine learning problem? Are you predicting some continuous number, or predicting a binary value?
    - ***This is a classification problem. I am trying to predict the top locations for a cluster (user types).***
  + What kind of impact do you think it could have?
    - ***This will help clients optimize their marketing campaign. Gain insights on user behavior, knowing who to target will reduce marketing uncertainty (minimize on marketing costs). Better marketing conversion, potential to keep users engaged so they do not search on other competitor sites.***
  + What do you think will have the most impact in predicting the value you are interested in solving for?
    - ***Number of travelers in a single booking, especially children vs. adult***
    - ***Loyalty member, if a user is using points/miles to purchase a flight***
    - ***Booking Window, how much time is between purchasing a flight and travel date***
* Datasets
  + Description of data set available, at the field level.
  + If from an API, include a sample return. (this is usually included in API documentation!)
* Domain knowledge
  + What experience do you already have around this area?
    - ***Yes. I work in the travel industry, directly involved with pulling data.***
  + Does it relate or help inform the project in any way?
    - ***I have the background in what clients are looking for and how our company is trying to advise them on marketing strategies***
  + What other research efforts exist?
    - Use a quick Google search to see what approaches others have made, or talk with your colleagues if it is work related about previous attempts at similar problems.
    - This could even just be something like "the marketing team put together a forecast in excel that doesn't do well."
    - Include a benchmark, how other models have performed, even if you are unsure what the metric means.
    - ***Travel industry has a ton of data and there are many models available. I work closely with many client data analytics team and we are starting to compare our analysis to see what would help both our businesses.***
* Project Concerns
  + What questions do you have about your project? What are you not sure you quite yet understand? (The more honest you are about this, the easier your instructors can help)
    - ***What is the best way to cluster a group of people?***
    - ***There is a large number of locations people can go to, how to find out the classification accurately?***
  + What are the assumptions and caveats to the problem?
    - What data do you not have access to but wish you had?
    - What is already implied about the observations in your data set? For example, if your primary data set is twitter data, it may not be representative of the whole sample. (say, predicting who would win an election)
    - ***I am not looking into seasonality.***
    - ***I am making assumptions about attributes that make up a user type, but this could be an inaccurate assumption***
  + What are the risks to the project?
    - What's the cost of your model being wrong? (What's the benefit of your model being right?)
      * ***If the outcome is inaccurate and we present this to a client, we be provided the bad suggestions, hurting our relationship***
      * ***If the outcome proves to be accurate and we present this to a client, we have provided strategies that could strength their marketing campaigns, thus strengthening our relationship.***
    - Is any of the data incorrect? Could it be incorrect?
      * ***Since we collect the data ourselves, the data cannot be incorrect***
      * ***However how I assume the data could cause issues***
* Outcomes
  + What do you expect the output to look like?
    - ***There will be a label of potential location for each user***
  + What does your target audience expect the output to look like?
    - ***A summary chart showing the user type, top 5-10 locations, and expected number of travelers***
  + What gain do you expect from your most important feature on its own?
  + How complicated does your model have to be?
  + How successful does your project have to be in order to be considered a "success"?
  + What will you do if the project is a bust (this happens! but it shouldn't here)?

**Exploratory Analysis Write-up**

* **Requirements:**
  + A well organized Jupyter notebook with code and fully ran top to bottom.
  + At least one visual for each independent variable and, if possible, its relationship to your dependent variable.
    - It's just as important to show what's not correlated as it is to show any actual correlations found.
    - Visuals should be well labeled and intuitive based on the data types.
      * For example, if your x variable is temperature and y is "did it rain," a reasonable visual would be two histograms of temperature, one where it rained, and one where it didn't.
    - Tables are a perfectly valid visualization tool! Interweave them into your work.
  + Provide insight about dataset and its impact on your hypothesis.

**RESOURCES**

**Suggestions for Getting Started**

* The more time you spend researching, the less time you'll likely spend writing; this is a positive sign!
* While researching, keep track of all of your resources. Make sure they're trustworthy.
* If you've seen similar work online, see if you can find the code that implemented the data munging. It might come in handy.
* If your project requires using an API, make sure you can get access to it. Not everyone gives away API keys immediately, and you don't want to be caught with no data with one week left to work!
* Consider building some helper functions that help you quickly visualize and interpret data.
  + Exploratory data analysis should be formulaic; the code should not be holding you back. There are plenty of "starter code" examples from class materials.
* **DRY:** Don't Repeat Yourself! If you see yourself copy and pasting code a lot, turn it into a function, and use the function instead!

**Specific Tips**

* Provide a sense of depth and scale to the project, which can be used to guide where the majority of your time should be spent working on the project.
* Show a clear connection between the datasets and the problem presented. The project should avoid working with independent variables (or features) that would not ordinarily be available in order to predict your target.