Legalist

Goal

We are interested in building a machine learning model that would help us predict the outcome of legal cases given historical data.

In order to do that, we want to use FJC (federal judicial center) as a data source in order to build prospective models. You can find the resource here:

- https://www.fjc.gov/research/idb i

For simplicity, we’re only interested in **“Civil cases filed, terminated, and pending from SY 1988 to present“.** The following resources will also be helpful, please review.

- Research Guide
- Cookbook

General Tips

This is an open-ended problem, there are no right or wrong answers. We want to give you the flexibility to dive deeply into the technicalities of things and showcase your expertise and background.

Expectations: we’re looking for you to deliver the following items

- Code for the ETL pipeline
- PDF Report Outlining :
    - Steps and decisions that were taken to build this project
    - The basic data analysis you performed and prospective features you identified
    - Strategy in building the machine learning model
    - Along with anything else you would like to add!

If assumptions or educated guesses are performed, please give context. For example:

- Why these assumptions were made
- How the guesses/estimations were calculated
- Pros and cons of the suggested assumptions vs other assumptions taken into consideration

Project Requirements

- You are required to write code that would extract the data highlighted previously
    - Create a plan for the ingestion of future data strategy (volume + frequency)

- You are required to transform and load the data into a data store of your choice.
    - The data (as the majority of data) is a bit messy, what are some of the cleaning/transformation strategies you'd like to apply to fix that?
    - As you can see from the cookbook, some of this data is already encoded. What is the best way to store it for querying? What is the best way to store it for building the model?
- Using basic data analysis, you are required to submit a report of simple features of the data. What metrics would you include in your report?
    - Basic metrics like total number of cases, number of cases by nature of suit, distribution of cases filed over time
- Outline the strategy, features, tools and algorithms you are planning on using for building the predictive model
    - How would you test the accuracy of that model?
- This is optional, but if you submit a working model for prediction with analysis of its accuracy and precision you can score up to +15%

Submission

Your submission should include the following:

- Code for the ETL pipeline
- PDF Report Outlining :
    - Steps and decisions that were taken to build this project
    - The basic data analysis you performed and prospective features you identified
    - Strategy in building the machine learning model
    - Along with anything else you would like to add!

We want to give you the flexibility to dive deeply into the technicalities of things and showcase your expertise and background.

Before uploading your submission, **please .zip all your files.** Once completed, please upload the entire folder to the provided Greenhouse Link in your email correspondence. If you have any questions, please contact parth@legalist.com

Scoring Rubric

- Quality of business solution (35%)
    - Clear definition of business goals and understanding of the scope(10%)
    - Data-driven decision making (numbers, graphs and charts) (10%)
    - Expertise in understanding data problems and underlying business problem(5%)
    - Clear identification of business constraints (5%)
    - Alternative solutions visited (pros and cons) (5%)

- Quality of tech solution (35%)
    - Clean and scalable solution, quality of ETL pipeline (10%)
    - Quality of data management, cleaning and transformation (5%)
    - Quality of the model strategy (5%)
    - Clear definition of system features and limitations (5%)
    - Identification of potential issues(5%)
    - Ease of use and implementation(5%)
- Quality of the Report (20%)
    - Creativity and expertise in the metrics reported (10%)
    - Communication style and readability of the document (10%)
- Testing & Safeguards (10%)
    - What tests or analysis were set in place to avoid reporting wrong metrics
- Bonus: Submission of a running model and analysis about its accuracy (up to +15%)