URDAD Design Doc [Draft]

Maria Berova, Archit Ganapule, Aileen Kuang, Arjun Naik, Maitreyi Parakh

February 19, 2023

1 Overview

This document aims to document the design of URDAD (Utilizing Regression for Date Aggregate Datasets). The purpose of our implementation was to make a model based off of extensive past Seattle crime data to predict future crime rates at a given date.

URDAD is structured around "layers" of date specifics.

- The **top layer** is the year. A polynomial regression is used to predict the total amount of a specific crime committed in a given future year.
- The **second layer** is the month. We calculated each month's average contribution (given as a fraction) to the aggregate crime count for the given year.
- The **bottom layer** is the day. We calculated each day's average contribution (given as a fraction) to it's month's aggregate crime count.

The outline for calculating the crimes committed on a given date (year, month, day) is given as follows:

1. Use the top layer to predict the total number of crime committed that year. For sake of convenience, call this number N.
2. Obtain the given month's average contribution from the second layer (call this fraction M). $N \times M$ will then be the amount of crime predicted to occur in the given month.
3. Obtain the given day's average contribution from the bottom layer (call this fraction D). $N \times M \times D$ will then be the amount of crime predicted to occur on the given day.

2 Layers of Time Data

2.1 Motivation

We decided to model trends over the years, months, and days separately from each other, then combine these trends to make an accurate prediction. This allows us to reflect specific trends of crime over months, and over days of the month, in our predictions. (For example, there could be more robberies on Black Friday, which would be reflected in how we model days in a month). It also prevents overfitting data by making general models for different layers.

2.2 Implementation

The "top layer", or the year, is crucial to making future predictions. We fit a polynomial model to past crime occurrence data by gradient descent. [After subsequent train-test splits and refinements, we determined that an n-degree polynomial was best suited for our model.]

The "second layer", or the month, is essentially a model of each month's contribution to the total crime count of a year. We first calculated the percent of each year's crimes committed in a given month for every month. Then, we averaged the percents we obtained over all the years sampled. Overall, this yielded the **average percentage of a year's crimes committed in each given month**.

The "bottom layer", or the day, is essentially a model of each day's contribution to the total crime count of a month. We first calculated the percent of each month's crimes committed on a given day for every day of every month, over all of the years. Then, we averaged the percents we obtained over all the years sampled. Overall, this yielded the **average percentage of a month's crimes committed on a given day**.

3 Scalability

An important consideration when designing our model was scalability. We created separate trends for days, months, and years in order to avoid overfitting for specific dates. The top layer uses a carefully considered model to make future predictions, so it is viable for many years in the future. Thus, it can be extended beyond the required "3 months in the future" requirement given in the Datathon prompt.

[Add some stuff about how our implementation is fast, idk]

4 4th Datathon-Specific Implementation

[Add some very specific stuff on how we made the algo, e.g., using pandas to create different dataframes, using tensorflow and sklearn]