Estimating Influential Samples in the Fragile Families Challenge

The Fragile Families Challenge (FFC) is a crowd-sourced machine learning (ML) competition to predict sociological outcome variables using the Fragile Families Childhood Wellbeing Study (FFCWS) [1]. The FFCWS is an expansive, longitudinal study following a cohort of nearly 5,000 ethnically diverse and mainly low socioeconomic status children and their families from birth through age 15. These families are called "fragile families" because they consist of mostly unmarried, low-income parents and are likely to live in unstable circumstances. Using ML, we can evaluate how predictable social traits are in order to improve targeted policy-making for diverse populations.

Using the common task method, the FFC participants were given background information for each child from birth to age 9. With this information, they were asked to predict six variables at age 15: household material hardship, child grade point average, child grit, household eviction, primary caregiver participation in job training, and primary caregiver layoff. All outcomes are based on self-reported data, and the participants had access to a subset of outcomes at age 15 to train and validate their models. There is a set of holdout data that was used to test the models for accuracy and to evaluate the final approaches.

Using ML to predict sociological variables is an inherently challenging task [2]. Additionally, this dataset follows participants from an underrepresented population over the course of 15 years which makes this task more difficult. For the FFC, 160 teams submitted predictions of the six outcome variables using myriad ML approaches. Despite the number of teams, none were able to predict any of these variables more accurately than the simple benchmark model. Upon inspection, we find that there are consistent problems with the FFC participants' data handling procedures and prediction methods. We identified five major issues: feature selection and dimensionality reduction; missing or sparse data imputation; model evaluation and interpretability of methods; varied data types and complexity of responses recorded; and the presence of influential samples.

Our research focuses on the last of these issues: influential samples that disproportionately affect the generalization error of the prediction methods in a negative way. We use influence functions to calculate each sample's contribution to a method's performance [3]. An *influence function* estimates the influence of an individual on the parameters of a model. By studying the influence of each sample, we are able to identify if subgroups of samples substantially degrade prediction performance. Furthermore, we aim to understand why a particular sample subgroup negatively affects prediction performance.

In order to avoid restricting the submission space we can evaluate, we will not be model-dependent when finding problematic individuals. We estimate the influence of each individual using a standard statistical measure known as Cook's distance[3]. Cook's distance utilizes linear regression to generate influence estimates for each individual. Our work focuses on understanding the effect of feature selection and imputation methods rather than model architectures. As such, we use the imputed, reduced datasets as the input to the Cook's distance function to identify influential samples. This allows us to compare feature selection and imputation strategies across submissions regardless of the model used in the submission.

We apply Cook's distance in the languages used by participants for the FFC: Python and R. For Python, we use Sklearn's *Yellowbrick* package (Figure 1), and in R, we use the *cookd* function from the *car* package.

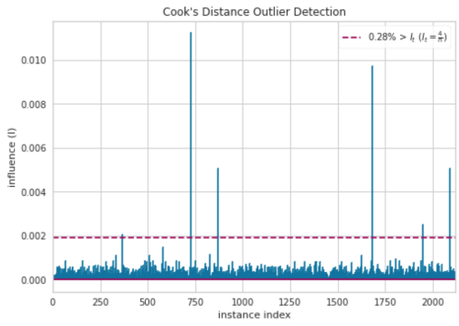

Figure 1: The output of Sklearn's *Yellowbrick* package on an FFC submission predicting child GPA. The instance index corresponds to FFCWS individuals. As visualized here, only a few samples exceed the threshold for influence.

```
net_difference = {}
for submission s in FFC:
    data = imputed and feature selected output from s
    threshold, influence = cooks_distance(data)
    for sample i in influence:
        net_difference[i].append(threshold - influence[i])
```

Figure 2: Our method for identifying the distribution of influence for each sample across submissions.

We focus on a subsample of the FFC submissions over which we aim to identify influential samples across a wide variety of relevant features and imputation methods. We identify these influential samples using an iterative method that calculates the distribution of each sample's influence across submissions (Figure 2) . Then, we test each sample for significance to identify those that we will examine more closely. We first used the Cook's distance threshold of $D(i) < \frac{4}{n}$, where n is the number of samples and $D(i)$ refers to the Cook's distance of the ith sample. We vary this threshold and investigate filtering samples near this threshold. We also consider a threshold of $\mu + \sigma$ where μ is the mean and σ is the standard deviation of $D(i)$ across samples.

The submissions we have investigated so far seem to agree on a few individuals that are influential. With additional experimentation, we will confirm our initial findings. In the event that we find significant discrepancy in influence across submissions, we will consider integrating model-specific measures and explore other methods for finding influence.

The ability to determine influence across samples in the FFCWS cohort will allow us to build better machine learning-based models to make sociological predictions. While our approach only addresses one aspect of the issues with using ML methods for social science predictions from longitudinal data, it is an important step missed by the FFC. Understanding the inaccuracy of the original predictions and subsequently improving them could direct policy and resource allocation by enabling policymakers to craft more effective interventions for at-risk children. [4]

[1] Salganik, M. J. et al. (2020). PNAS. 201915006.
[2] Liu, D., Salganik, M. (2019). SocArXiv.
[3] Cook, R. Dennis (February 1977). Technometrics. American Statistical Association. 19.
[4] Kleinberg, J., Ludwig, J., Mullainathan, S., Obermeyer, Z. (2015). American Economic Review. 105.