CAPSTONE PROPOSAL

S. CHETTIH

1. Domain Background

Public education is a cornerstone of American democracy, and education is seen as the great equalizer of American society. However, it cannot fulfill that role for students who are chronically absent, which I define as missing at least 15 days of school in a year[1]. The US Department of Education describes chronic absenteeism as a hidden educational crisis, which has surely worsened during the COVID-19 pandemic.

2. Problem Statement

School administrators need accurate, automated, and simple methods to identify students who are at risk of chronic absenteeism. Identifying students at risk of chronic absenteeism is important for realizing that promise of equalization, and for ensuring that students, in their wildly varying circumstances, are given the support they need to stay in school.

3. Solution Statement

I propose to build a machine learning model which can reliably identify students who are likely to be chronically absent, based on information that most school administrators would have access to. Decision tree algorithms are appropriate for this classification problem, and they have the added benefit of being interpretable[2], though they should be compared to other algorithms such as k-NN and logistic regression for thoroughness. To this end, I will use AutoGluon-Tabular to compare before tuning a particular model. Accuracy is a common metric for classification problems, but I don't believe it's appropriate for this problem, as very few students, relatively, are at risk of chronic absenteeism. Instead, I will use the F_1-score to compare the models produced by AutoGluon-Tabular, and to determine the overall success of my solution.

I plan to use AWS SageMaker to create an instance on which to run my code, and I will train my model on a separate instance. The training and testing datasets will be stored in S3.

---

[1] `https://www2.ed.gov/datastory/chronicabsenteeism.html`, accessed 1/11/2022. Some states define chronic absenteeism as missing 10% of the days in a school year, which could mean anywhere from 15 to 22 days depending on length.

[2] Muzaferija, Ibrahim & Mašetić, Zerina & Jukic, Samed & Kečo, Dino. (2021). Student Attendance Pattern Detection and Prediction. Journal of Engineering and Natural Sciences. 3. 10.14706/JONSAE2021313.

4. Datasets and Inputs

The dataset I will use is the 2016[3] results of the Parent and Family Involvement in Education survey, collected by the National Center for Education Statistics[4]. This dataset provides student-level information about what kind of school they attend, which grade they are in, their absences for the year, and demographic information about the student. Since the surveys were filled out by parents and/or guardians of the students, the number of absences for the school year should not be treated as exact. After removing students for whom no absence data is provided, we are left with 13,523 rows (each corresponding to a particular student) and 822 columns (mostly weights & imputation flags for other columns), with 434 students who were chronically absent (about 3.2%). The dataset includes students from kindergarten through 12th grade, in public and private schools.

5. Benchmark Model

I am unaware of other attempts to predict attendance of individual students with machine learning (many machine learning applications focus on image-based attendance systems) or of other attempts to perform machine learning on this dataset. The paper "Student Attendance Pattern Detection and Prediction" published in the Journal of Engineering and Natural Sciences predicts student attendance in classes at the university level on a day-to-day basis using a decision tree model[5], and attains an F_1-score of 66.55%. The confusion matrix for their model is reproduced below (0 indicates a student marked absent, 1 a student marked present). These will serve as benchmarks for my project.

	true 0	true 1	class precision
predicted 0	31878	8291	79.36%
predicted 1	4814	13036	73.03%
class recall	86.88%	61.12%	

6. Evaluation Metrics

As mentioned above, I will use the F_1-score to select the best model produced by AutoGluon-Tabular, and to compare the result to the benchmark model. I will also consider the confusion matrices, as fewer than 5% of students in the dataset are in the target class.

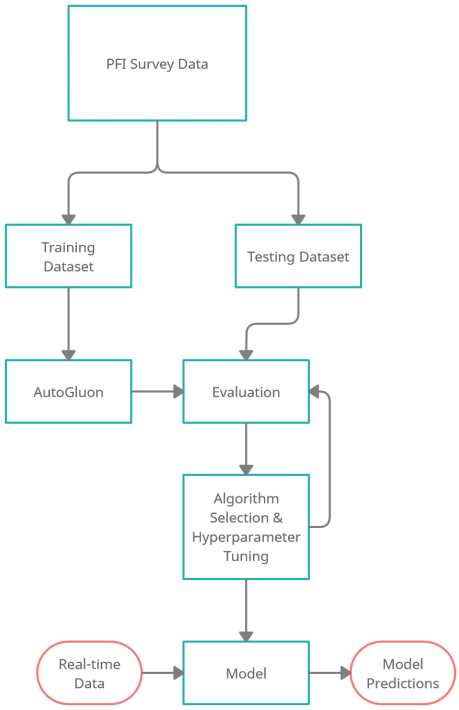
[3]In my original proposal, I mistakenly thought the data I was using was from 2019. It is instead from 2016.

[4]`https://nces.ed.gov/nhes/data/2016/pfi/pfi_pu.csv`, accessed 1/11/2022

[5]Muzaferija, Ibrahim & Mašetić, Zerina & Jukic, Samed & Kečo, Dino. (2021). Student Attendance Pattern Detection and Prediction. Journal of Engineering and Natural Sciences. 3. 10.14706/JONSAE2021313.

7. Project Design

To complete my project, I will upload the dataset to an S3 bucket and use a Jupyter notebook in SageMaker perform exploratory analysis on the data set, as well as feature engineering. Then I will split the data into separate testing and training sets, being careful to include a proportional amount of students at risk of chronic absenteeism in each set. I will write a SageMaker script to train a model on my training dataset, and compare its performance on the test set to my benchmark. If necessary, I will perform further feature selection, hyperparameter tuning and analysis. The deployed model will be able to take in current data about a particular student and predict whether their total absences by the end of the school year will total at least 15, thereby putting them at risk of chronic absenteeism.



Thank you for your time!