CAPSTONE REPORT

S. CHETTIH

1. Definition

1.1. **Project Overview.** Educational institutions collect large volumes of data, but data that is publicly shared often contains only information at the institutional level. The inspiration for this project came from my discovery of the Parent and Family Involvement in Education survey, collected by the National Center for Education Statistics, which provides anonymized data on a variety of students from kindergarten through high school. I was inspired for the capstone project to apply machine learning techniques to this dataset.

School administrators need accurate, automated, and simple methods to identify students who are at risk of chronic absenteeism. The goal of this project is to build a machine learning model which can reliably identify students who are likely to be chronically absent. To my knowledge, this is the only attempt to apply machine learning techniques to the PFI survey data, and the only attempt to predict student attendance over the course of an academic year.

1.2. **Problem Statement.** I define a student as chronically absent if, by the end of the school year, they have missed at least 15 days of school[1]. The goal is to identify those students who, by the end of the year, will have missed at least 15 days of school, using information available at the beginning of the school year. To this end, I build a model to predict chronically absent students among the PFI survey data.

1.3. **Metrics.** Accuracy is a common metric for classification problems, but I don't believe it's appropriate for this problem, as chronically absent students are relatively rare. I will use both the F_1-score and the ROC AUC score instead.

2. Analysis

2.1. **Data Exploration.** The initial dataset of the 2016 results of the Parent and Family Involvement in Education survey[2] was downloaded in the CSV format. Further operations were done in Python with the pandas package. 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.

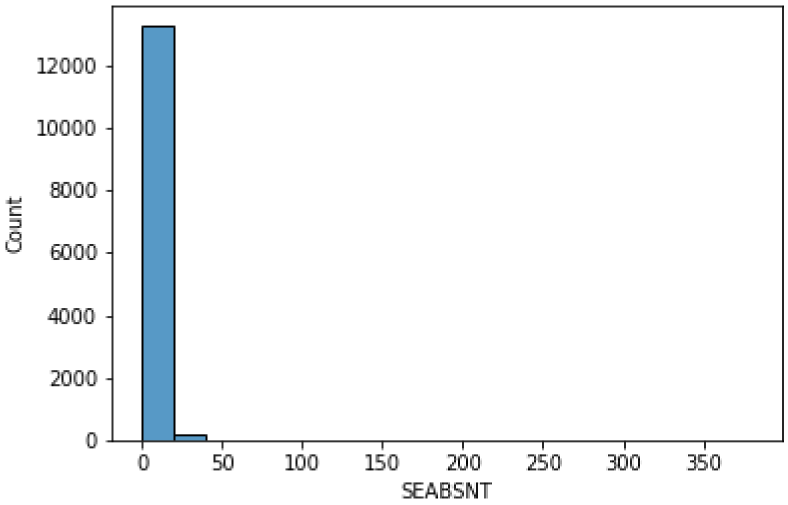
[1] `https://www2.ed.gov/datastory/chronicabsenteeism.html`, accessed 1/11/2022.

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

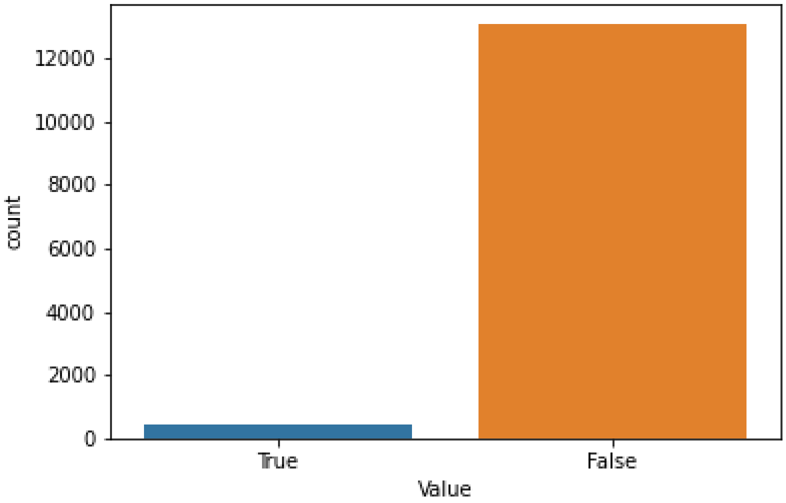
Table 1 explains the most relevant features from the dataset. The feature ‘SEABSNT’ (number of days the student has been absent from school) is transformed into a boolean, True if ‘SEABSNT’ is greater than or equal to 15, and False otherwise. This boolean is the target variable. Since the surveys were filled out by parents and/or guardians of the students the number of absences for the school year may not be exact.

Table 2 gives statistical measures for selected features, as well as the target variable. Our target variable is relatively rare, at around 3.2%, which will pose significant difficulties for our model. The distribution of number of days absent is no better, as the median is 3, well below our threshold of 15. The dataset is relatively evenly divided among grade levels and ages, and includes enough students with disabilities, or who have repeated grades, to be able to draw conclusions.

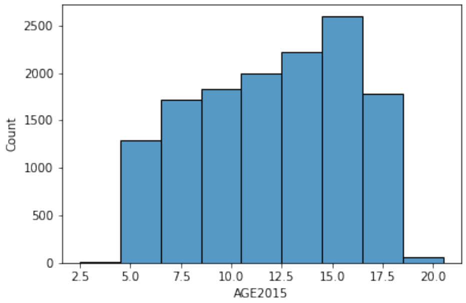
2.2. **Exploratory Visualization.** The distribution of days absent is heavily concentrated to the left, as seen in the histogram below.

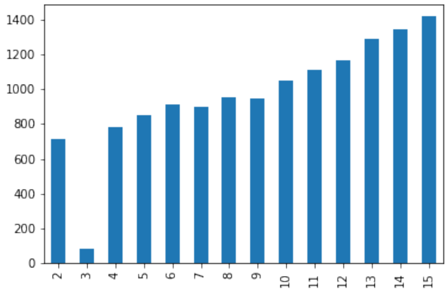


As expected, the number of chronically absent students in the database is also relatively small, though I hoped it was large enough to train a model.

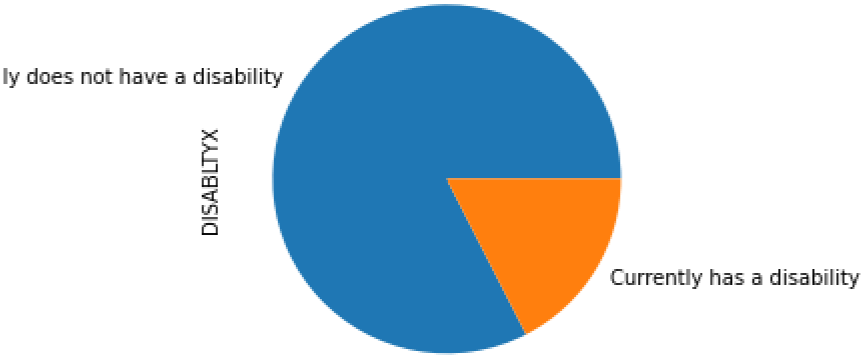


The ages and grades are relatively evenly distributed, as shown in the histograms below. A grade of ‘2’ corresponds to full-time kindergarten, ‘3’ corresponds to part-time kindergarten, ‘4’ corresponds to 1st grade, and so on through ‘15’, which corresponds to 12th grade.

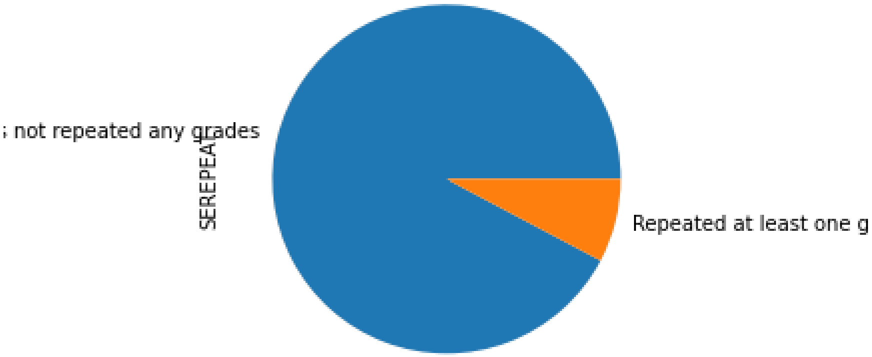




The number of students who currently have a disability are not in the majority, but they are present in non-trivial numbers.



Similarly, a non-trivial number of students in the dataset have repeated at least one grade.



2.3. **Algorithms and Techniques.** Decision tree algorithms are appropriate for this classification problem, and they have the added benefit of being interpretable[3], though they should be compared to other algorithms such as random forests, k-NN and logistic regression for thoroughness. To this end, I used AutoGluon-Tabular to compare models and ensembles rather than sticking to a single type of algorithm.

[3] 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.

2.4. **Benchmark.** 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[4], and attains an AUC score of 0.812 and 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%	

3. Methodology

3.1. **Data Preprocessing.** This dataset does not need preprocessing. It has no missing values, and all categorical features have been encoded as integers.

3.2. **Implementation.** The implementation process has three main stages:

(1) Training using AutoGluon on large dataset
(2) Using feature importances to pare down dataset and select best algorithm
(3) Retraining using small dataset

In the first stage, I ran AutoGluon-Tabluar on a separate instance, which required separate Python scripts and YAML configuration files (found in the scripts and config folders). For these, I am indebted to the AWS Amazon SageMaker examples provided on GitHub[5]. The first time I used AutoGluon, I used the F-score as my evaluation metric. This was a disaster, as most models scored 0.0 by this metric. Therefore, I ran AutoGluon a second time using ROC AUC as my evaluation metric. The results were modestly promising, with CatBoost as my top model with a score of approx. 0.67. However, looking at the predictions on my test set revealed that the model never put out a prediction of 'chronically-absent'. I chose to forge ahead with paring down the dataset, hoping that this might improve the model somewhat.

In the second stage, I selected only the top 5 most important features. Even the most important feature, 'DISABLTYX', had an importance of less than 0.02. I chose to keep

[4] 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.

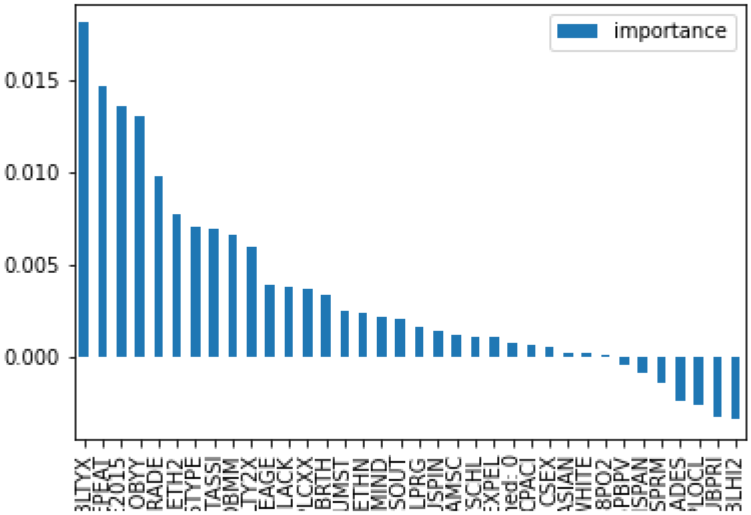
[5] found at `https://github.com/aws/amazon-sagemaker-examples/tree/master/advanced_functionality/autogluon-tabular-containers`

only a few features because all the importance scores were so low. Additionally, I removed ‘CDOBYY’, since this feature is the year of birth of the students, and highly correlated to ‘AGE2015’, their age on Dec 31st, 2015. Looking at the leaderboard, there was no clear algorithmic winner. Most of the top algorithms were ensemble methods, and the bottom algorithms were k-NNs. In the interest of completeness, I decided to train a simple Decision Tree Classifier and a Random Forest Classifier for comparison. Both of these models were obviously overfit, judging from the differences between their cross validation scores and their scores on the test set. In fact, neither model was better than random guessing, as shown by their ROC AUC scores of 0.5, which highlights how difficult this prediction problem is.

Finally, I once again ran AutoGluon-Tabular on a separate instance, using the pared-down train and test sets and ROC AUC as my evaluation metric. The most important feature was again ‘DISABLTYX’, with a score of approx. 0.054. The position on the leaderboard of algorithms did not significantly change, and neither did the ROC AUC scores.

3.3. **Refinement.** My focus for improving the algorithms was on feature selection. The original dataset had hundreds of features, which made it difficult to select how many to train AutoGluon on. Dropping students who were homeschooled, and features related to homeschooling, was the obvious first choice. Beyond that, I chose to keep the features which would be readily available to a school administrator, such as whether a student has been in the same school the whole year. I chose to drop the features which focused on parent’s opinions about the school, parent involvement such as reading books or having meals together, and demographic information on the student’s household, by the same logic.

The low feature importance scores made it difficult to pare the features down further. See the graph of feature importances below.



From the shape of the graph, there is a natural cutoff after the first four features, but I decided to include 'GRADE' knowing that the feature 'CDOBYY' was almost perfectly correlated with 'AGE2015' (as described in the previous section), and should therefore be removed from the dataset. However, the low importance scores overall were a much larger problem, and one that cannot be solved by feature engineering.

AutoGluon is already a state-of-the-art algorithm, and it is capable of squeezing nearly all the performance possible out of the data I provided. Hyperparameter tuning, such as tweaking the learning rate and limiting the maximum depth of trees, yielded nearly identical results to the original.

4. Results

4.1. **Model Evaluation and Validation.** The final model was a CatBoost ensemble model, which is made of symmetric trees. It achieved a ROC AUC score of approx. 0.67 on the test set, but further analysis reveals that the maximum value given to a True prediction was approx. 0.138. That means that the model will never predict that a student is chronically absent, which is not acceptable. The classification report for the model shows high precision and recall on the not-chronically absent class, as expected, and a divide-by-zero error when calculating the same statistics for the chronically absent class.

4.2. **Justification.** The benchmark model, from the paper "Student Attendance Pattern Detection and Prediction," predicts student attendance on a day-to-day basis, which avoids the issue of highly imbalanced target classes. It achieves an AUC score of 0.812 and an F_1 score of 66.55%, neither of which are matched by my model. In particular, their model has class precisions of 79.36% (for predicted absent) and 73.03% (for predicted present), which indicate that the model is reasonably precise with its class predictions.

I sought a simple, interpretable, reasonably accurate model for predicting whether a particular student is likely to be chronically absent. It is obvious that I was not successful, but it is worth interrogating where and how I failed. Firstly, imbalanced classification is a difficult problem. This was compounded by very low importance scores among the available features. This leads me to believe there is high variance in the target variable which cannot be explained by the variables in the PFI dataset. Secondly, the target variable effectively created a sharp cutoff between students who were absent 14 days instead of 15, a cutoff which was not present in the original. Additionally, the survey was filled out by parents and/or guardians, who may not accurately recall the number of days that a student was absent over the course of a year. Together, I believe these issues show that the PFI survey data is not well suited to predicting chronic absenteeism, regardless of the machine learning methods used.

Thank you for your time!

Feature Name	Notes
GRADE	Student's current grade or year of school (Kindergarten through Twelfth grade, or none of the above)
SCPUBPRI	Type of school student attends (Public, private but not religious, private Catholic, private religious but not Catholic, none of the above)
DISTASSI	Whether the school is the student's district-assigned school
SSAMSC	Whether the child has been in the same school since the beginning of the school year
SEGRADES	Parent/Guardian's estimate of the student's grades (Mostly A's, mostly B's, mostly C's, mostly D's or lower, School does not give these grades, or skip)
SEADPLCXX	Whether the student is enrolled in any high school Advanced Placement classes
SEREPEAT	Whether the student has repeated any grades
SESUSOUT	Whether the student has had an out-of-school suspension
SESUSIN	Whether the student has had an in-school suspension
SEEXPEL	Whether the student has been expelled from school
DISABLTYX	Whether the student currently has a disability
AGE2015	Student's age on Dec 31st, 2015
CDOBYY	Student's birth year
CDOBMM	Student's birth month
S16TYPE	Type of school on the Common Core of Data or Private School Survey (Catholic, other religious, Nonsectarian, Public, Data are missing for school, Homeschooled student)
RACEETH2	Detailed race and ethnicity of child (White non-Hispanic, Black non-Hispanic, Hispanic, Asian or Pacific Islander non-Hispanic, All other races and multiple races non-Hispanic)
SEABSNT	How many days student has been absent from school, since the beginning of the school year

Table 1. Dataset Features

Feature	Data Type	Statistics
SEABSNT	Int	Mean: 4.2, Median: 3, IQR: 4, Min: 0, Max: 364
DISABLTYX	Cat	Currently has disability: 2,364 (17.5%), Currently does not have a disability: 11,159 (82.5%)
SEREPEAT	Cat	Repeated at least one grade: 1,045 (7.7%), Has not repeated any grades: 12,478 (92.3%)
GRADE	Cat	Full-time K: 712 (5.3%), Part-time K: 85 (0.6%), 1st: 779 (5.8%), 2nd: 848 (6.3%), 3rd: 915 (6.8%), 4th: 896 (6.6%), 5th: 955 (7.1%), 6th: 946 (7.0%), 7th: (7.8%), 8th: 1115 (8.2%), 9th: 1168 (8.6%), 10th: 1290 (9.5%), 11th: 1342 (9.9%), 12th: 1421 (10.5%)
AGE2015	Int	Mean: 12.1, Median: 12, Std Dev: 3.79, Min: 3, Max: 20
target	Bool	True: 434 (3.2%), False: 13,089 (96.8%)

Table 2. Feature statistics