Balancing Act: Sensitive Data and Accuracy in University Dropout Prediction

Bryan Chi Fai Pang
Student ID: 50120081
CIND820 Project Presentation
Supervisor: Dr Ceni BABAOGLU

University Student Dropout

Global dropout rate ranging from 30% OECD countries to 50.9% in Costa Rica

-   North American Context:
-   Canada: Up to 20% of students quit, 20%-50% of students change initial program.
-   Canadian workers in their 40s with degrees earn 53% (Cdn $13) per hour more
-   In US, eight year difference in term of life expectancy between degree holder and non degree holders

The mortality gap between Americans with and without four-year degrees is widening

Average years of life remaining for 25-year-old Americans
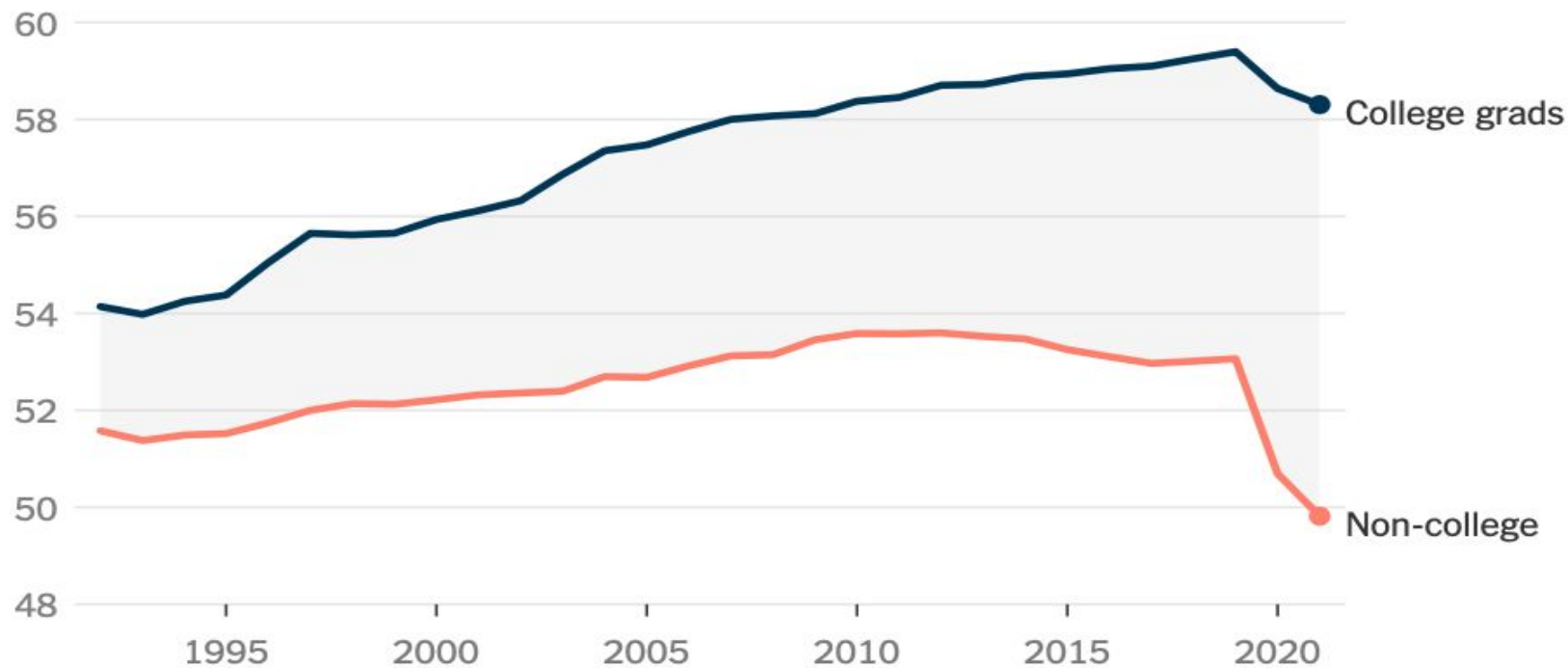
College grads

Non-college

States and shortcomings of University Student Dropout Prediction in Machine Learning

While all the studies confirm the effectiveness of data mining and ML approach in predicting dropout...

YET:

They focus on algorithms and not on feature space...

CONCERNS:

Use of sensitive features: Is it necessary ?

Two studies focused on Algorithm Fairness and Model Performance

Deho, O. et al. (2023). Should Learning Analytics Models Include Sensitive Attributes? Explaining the Why.

Yu, Renzhe, et al. (2021). Should College Dropout Prediction Models Include Protected Attributes?

- **Protected attribute**
  Binary = Male, Female, High Income / Low Income, etc

- **Privileged Class**
  Male vs Female

- **Favorite Outcome**
  Likely to Dropout / Graduate

- **Comparison Based on Blind and Aware Dateses**

Features are considered as isolation instance, rather than part of a cluster.

Up to 4 features are excluded in these two studies and both study show no significant difference in model performance.

Research Question:

Attribute Types and Prediction Performance in University Dropout

How do different types of attributes impact the performance of student dropout prediction models?

What are the consequences of including or excluding specific classes of features on the accuracy of student dropout predictions?

		Set 1	Set 2	Set 3	Set 4	Set 5
Class of Attribute						
Demographic			x		x	x
Socioeconomic				x	x	x
Macroeconomic		x	x	x	x	
Academic		x	x	x	x	

Valentim Realinho's dataset: 4424 students and 35 attributes, presented as a paper and available in UC Irvine Machine Learning Repository

Rich in features
- **Demographic (6 features)**
- **Socioeconomic (8 features)**
- **Macroeconomic (3 features)**
- **Academic (17 features)**

**Target is three classes (Dropout, Enrolled, Graduate)**
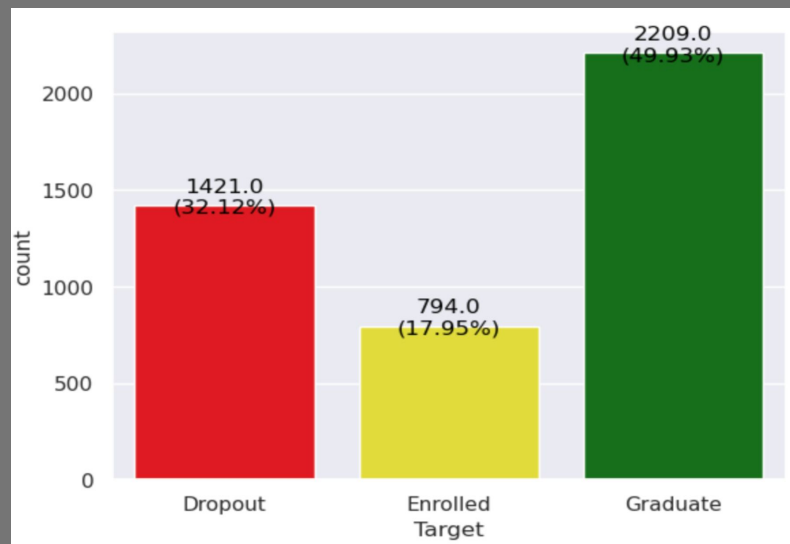
Methodology

- Dataset is split into Train/validation set (TV set) (80%) and Test Set (20%)
- Algorithms Selection: Cross validation with TV set (Random Forest n=10, RF n=100, Support Vector Machine, SVC kernel linear, Gradient Boosting, XGB Booster)
- Best three are used, based on Average F1 and Average Accuracy

Average F1 Scores and Average Accuracies by Model
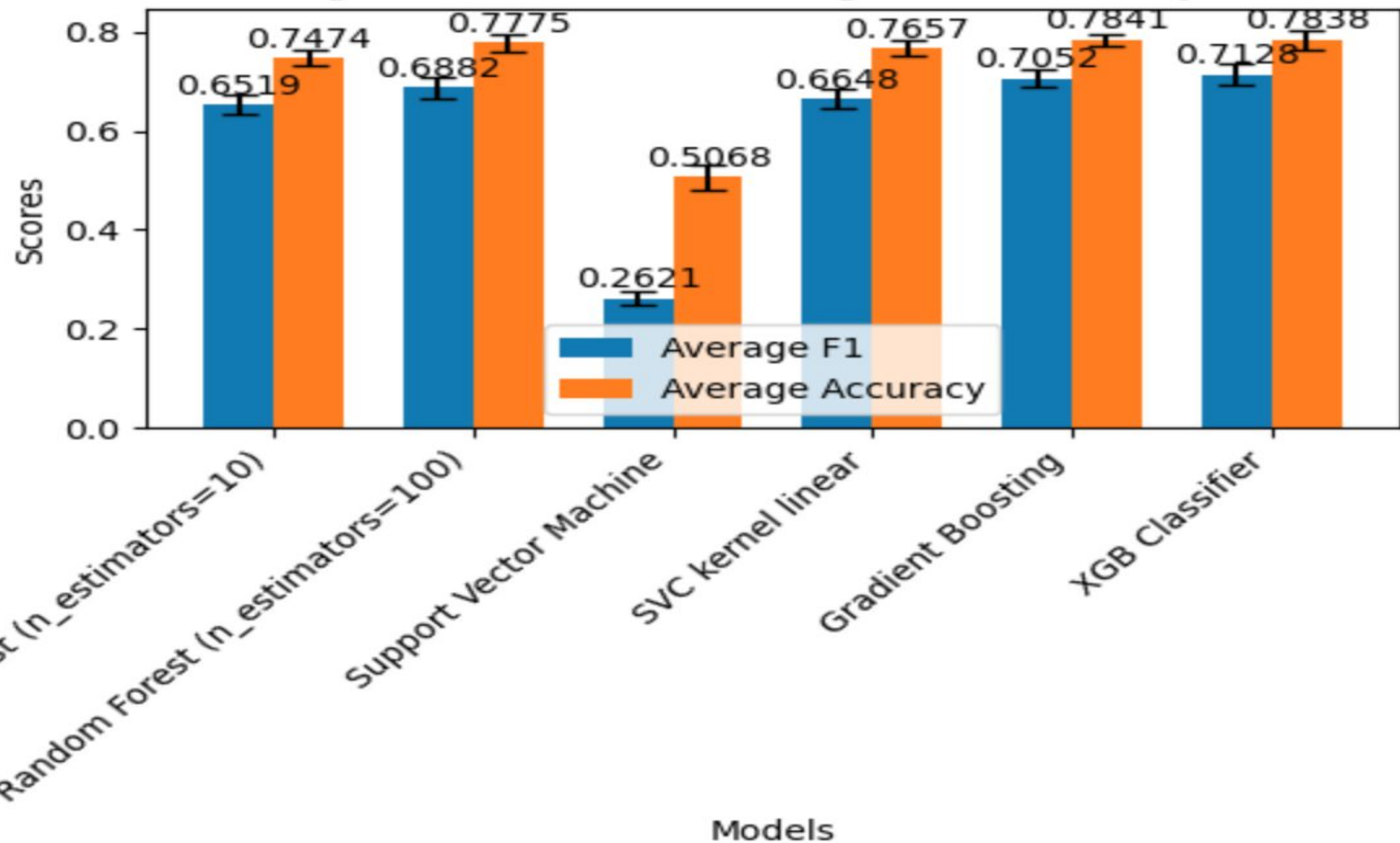
XGB: Metrics by Subset
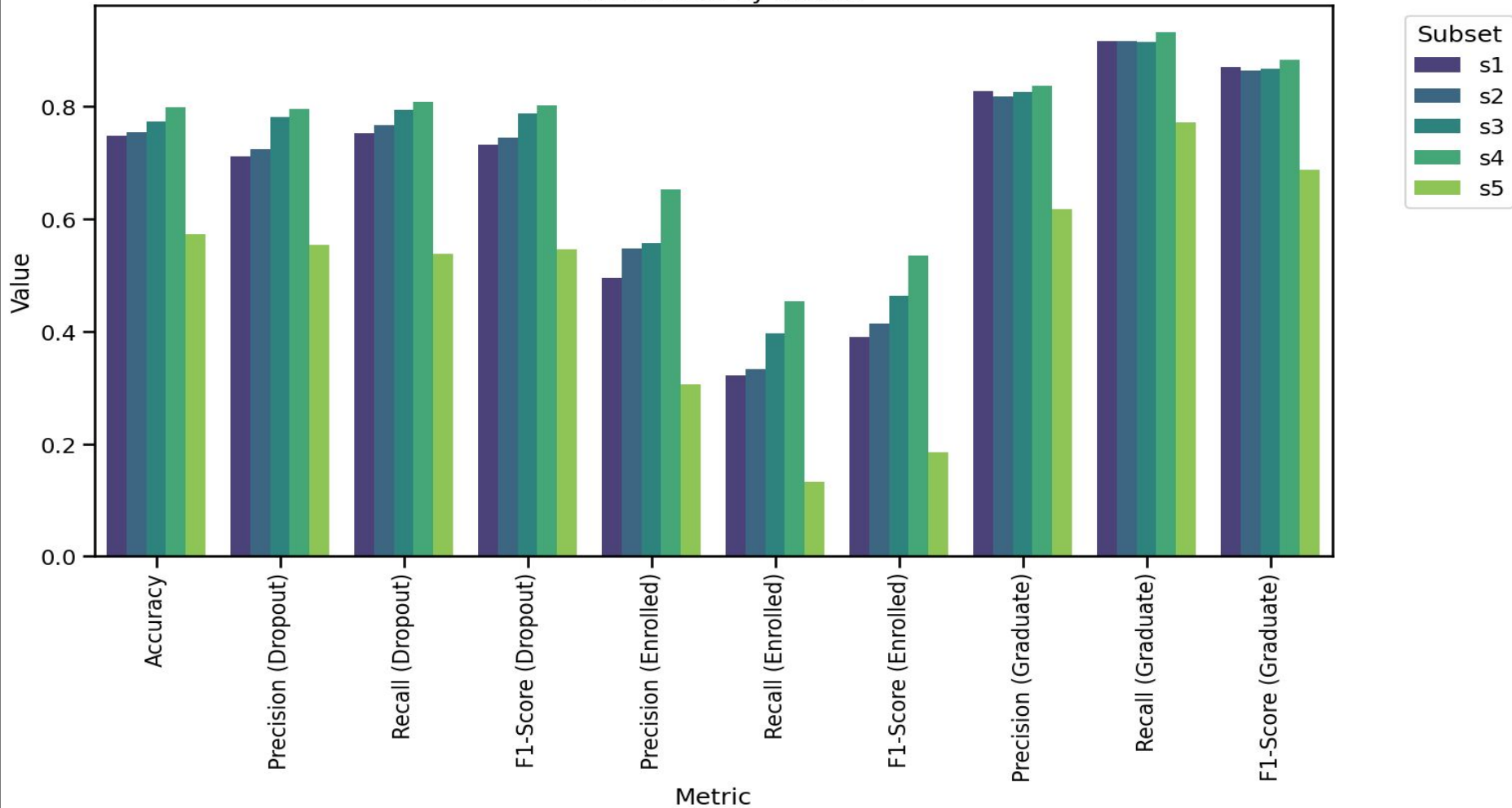
Results:

Model trained using Academic and Macroeconomic data (s1) performs well at 74.71 % accuracy.

Model with additional Demographic data (s2) shows slight improvement at 75.40% accuracy (0.69 %)

Model with baseline and additional socioeconomic data (s3) shows (2.65% increase) at 77.36% accuracy.

Model with all the data (s4) performs best, with 79.89% accuracy (5.1% increase)

Model with just demographic and socioeconomic data shows worst performance 57.24% accuracy (decrease in 17.47%)

Precision and Recall and Classes

- **Graduate: 80% Precision and 90% in Recall throughout s1-s4**
- **Dropout: 70-80% both Precision and Recall s1-s4**
- **Enrolled: 50-65 % Precision, 32-45% in Recall s1-s4**

Discussions

Use of demographic and socioeconomic data increases model accuracy by 5% in XGBoost Model, comparing to using to using only academic and macroeconomic data alone.

However, the increase of performance is achieved through addition of 14 features, many of them can be considered sensitive and unrelated to academic performance.

Achievement of the study

Refocus on the feature space and features rather than algorithms used.

Demonstrate that reasonable good model can be developed with academic data alone.

Contribute to study of use of AI Fairness and ML-based decision making in Dropout Prediction and informed decision in the use of sensitive features in dropout prediction.

Limitations of study and scope for further research

The target of the dataset is three classe, whereas all of the studies reviewed are binary classification.

Complexity of the methodology: three models, 5 subsets and three classes. A simpler approach (two subset and 2 class models) would be easier to understand and execute.