COMS - W4995 Applied Machine Learning Project Proposal

Gokul Sunilkumar	Lei Liu	Rachel Peng	Xiaolin Sima	Ziyao Zhou
gs3214	ll3442	xp2197	xs2483	zz2915

October 10, 2022

1. **Background and Context to the Problem Statement:** In this project, we are exploring a dataset that includes K-12 student demographics, achievements, and state financial data. Our main goal is to identify demographic and state-level patterns, and infer factors predictive of academic performance (Supervised Regression Task). To give some background, education in the U.S. is provided in public and private schools and by individuals through homeschooling. The funding of public elementary and secondary schools in the U.S. involves a combination of local, state, and federal government revenues, in proportions that vary substantially both across and within states. Moreover, state governments set overall educational standards, often mandate standardized tests for K-12 public school systems and supervise. A common measure for academic performance is National Assessment of Educational Progress or NAEP scores, the largest continuing and nationally representative assessment of what U.S. students know and can do in various subjects.

2. **Identification and Description of the Dataset:** We will use the data from U.S. Education Datasets: Unification Project. The data is scripted from the U.S. Census Bureau and the National Center for Education Statistics (NCES), recorded from 1986 to 2019. The dataset size is 1.3MB, containing 1715 rows and 266 columns with both numerical and categorical data types. Due to the high column number in this dataset, it's necessary to extract and encode data from column names before training models. In the dataset, demographics and achievement data, such as the number of students and the average NAEP math/reading exam scores, are categorized by year, state, grade, ethnicity, and gender. The financial data contains revenue and expenditure by the local, state, and federal government. From examining this dataset without doing too much data exploration work, we are confident to say that it's ideal for performing various data visualizations, cleaning missing values, scaling, as well as experimenting with different machine-learning models to predict academic performances.

   Source: https://www.kaggle.com/datasets/noriuk/us-education-datasets-unification-project

3. **Proposed ML Techniques:**

   Track 1: Regression Techniques

Linear	Baseline
Ridge	L2 regularization. Yields parameters that are small values. Combats overfitting.
Lasso	L1 regularization. Yields parameters that are 0 values. Combats overfitting. Eliminates features with minor importance by performing feature selection.
Elastic Net	Can outperform lasso on data with highly correlated predictors. Allows for a balance of both penalties (L1, L2). Can perform feature selection as there are a large number of features (266) in this dataset.

   Track 2: Tree Regression Techniques

Random Forest	Works well with both numerical and categorical variables. Not sensitive to missing values and can handle outliers to some extent. Can also exploit its feature_importance attribute and tune its hyperparameters to better interpret the independent variables and their contribution to the results.
XGBoost	Handles missing values in a better fashion. As multiple weak learners are combined in a sequential manner, it iteratively improves observations.

   After comparing the performance metrics, the best performing model out of the above 6 models across the two tracks is chosen.