Machine Learning Proposal

Predict College Earning Potential

Domain Background

I propose to create a model for prediction for college selection based on earning potential. Students and parents have a tough time determining which colleges to apply. There are a lot of factors to consider and lots of conflicting information. Also there is lots of data available as well as lots of variables involved. But in general apart from SAT score and GPA that are used mainly for the admission process, several factors like University admission rate, public/private type of university etc., need to be considered.

References:
https://collegescorecard.ed.gov/
https://blog.prepscholar.com/does-sat-predict-your-college-success-and-future-income
https://pdfs.semanticscholar.org/1b87/6c546e7b715b17e893adf50d8911e99f7369.pdf

Problem Statement

Predict some of the major factors to be considered by a student when applying to Universities and help in the process of selecting Universities to apply.

Datasets and Inputs

Dataset:
https://collegescorecard.ed.gov/data/Most-Recent-Cohorts-All-Data-Elements.csv

Documentation:
https://collegescorecard.ed.gov/assets/FullDataDocumentation.pdf

https://collegescorecard.ed.gov/data/CollegeScorecardDataDictionary.xlsx

Solution Statement

This is a classification problem. If the student is hoping to earn at least $50,000 10 years after graduation, which universities might he plan on applying (without taking the degree major into consideration). Which of these factors available about Universities will matter most: SAT scores, size (number of students), spending per student by university, type (public/private), cost for students, rate of admission, rate of completion and rate of retention. I plan to use Ensemble methods like Ada Bost, Random Forest and Gradient Boost and select the one providing better accuracy.

Benchmark Model

Since there are numerous factors that could be used and hence not many standardised studies are available to provide an historic benchmark, I am planning to use either a Naive predictor where we assume everyone earns above $50,000 or Decision Trees, as a benchmark model and see how the above Ensemble models perform.

Evaluation Metrics

Planning to compare the accuracy score of the predictions of the Ensemble models specified compared to the Benchmark model.

Project Design

1. Data Exploration: Cursory investigation of the data to explore the dataset and the featureset and to find degree of correlations between variables.
2. Data Preprocessing: Cleaning the data including formatting and restructuring, normalizing the data and shuffling and splitting the data into training, validation and testing sets.
3. Feature selection: Extract feature importance, select relevant features and create new features if necessary/possible to improve accuracy.
4. Model selection: Experiment with Ensemble algorithms Ada Bost, Random Forest and Gradient Boost to find out the best algorithm for this scenario.
5. Model Tuning: Fine tune the selected model to increase accuracy and performance.
6. Testing: Test the model on testing dataset and do final model evaluation.