Machine Learning Engineer Nanodegree

Capstone Project- Predicting Graduate Admissions

Amal Imdad Alhaq

February 5th, 2019

Proposal

Domain Background

Graduate admissions are not quite accurate science and the admissions process took lots of time. Faculty committees, sitting around conference tables for hours on end, have abundant of data to decide about applicants, including their transcripts, personal statements, GPA, letters of recommendation and GRE scores. Problem is, it's not always clear just what the data mean. The admission systems vary widely from country to country, and sometimes from institution to institution. That makes it easy for biases to slip in undetected [1].

Concerns about diversity in graduate programs are well-founded. But standardized tests like the GRE are not what's holding the academy back from attaining greater diversity. The problem emerges instead in those long meetings in conference rooms. The faculty has plenty of applicants and limited time to make important decisions, all while navigating departmental politics and seeking to raise their program's prestige [2]. By machine learning, the whole decisions can be done faster with more accuracy to help the admission department to crack this problem.

Problem Statement

All the admission process is depending on faculty to deal with massive applications that they can accept in a limited time, that take more human endeavors. whereas the admission relies on many features, for instance, GRE score, TOEFL scores, university rating, Letter of Recommendation, Statement of Purpose and Research Experience, that all use numeric values, anticipating the graduate admission can decrease the time and helping the admission committee by giving them the possible candidates. On the other hand, helping students in shortlisting universities with their profiles. The predicted output gives them a fair idea about their chances for a particular university.

Datasets and Inputs

The dataset is obtained from Kaggle Repository. It is inspired by the UCLA Graduate Dataset. The dataset is owned by Mohan S Acharya [3].

Link of datasets: https://www.kaggle.com/shraban020/predicting-admission-by-logistic-regression/data

The dataset has 500 records updated month ago. It has 9 attributes are described as follows:

1. GRE Scores (out of 340)
2. TOEFL Scores (out of 120)
3. University Rating (out of 5)
4. Statement of Purpose "SOP" (out of 5)
5. Letter of Recommendation Strength "LOR" (out of 5)
6. Undergraduate GPA (out of 10)
7. Research Experience (either 0 or 1)
8. Chance of Admit (ranging from 0 to 1)

the characteristics of the dataset or input are a numerical value. And these features are appropriate given the context of the problem [3].

Solution Statement

Supervised learning can be used for this problem since it has numerical data. We can build a logistic regression model predicting the admission [3]. The most common solution to such problem is Regression methods.

1. Linear Regression
2. Polynomial Regression.
3. Logistic Regression

4. Regularization

Regression method is a form of predictive modeling technique which investigates the relationship between a dependent (target) and independent variables. This technique is used for forecasting between the variables. For example, the relationship between GRE Scores and Chance of Admit.

But We're not sure that the previous model would be fit our problem so I will pick off one the following algorithm.

- Ensemble Methods (Bagging, AdaBoost, Random Forest, Gradient Boosting)

- K-Nearest Neighbors

- Stochastic Gradient Descent Classifier (SGDC)

Graduate Admission would seem that using accuracy and F-score as a metric for evaluating a particular model's performance would be appropriate.

Benchmark Model

While I was searching for this problem, I found someone uses logistic regression for prediction. So I'll use another algorithm which is Adaboost method with regression as a weaker learner to predict the graduate admission. I will use AdaBoost mothed that are currently available in sci-kit-learn.

AdaBoost is one type of Ensemble. It takes many models of a weak learner and joined them to get a better model which is a strong learner. AdaBoost algorithm calls a given weak algorithm (regression) repeatedly in a series of rounds, it trained the data iteratively.

Evaluation Metrics

The Evaluation metrics that I will use for this problem F-score and the accuracy.

$$F\beta=(1+\beta 2)\cdot(precision\cdot recall/(\beta 2\cdot precision)+recall)$$

It uses true positive TP, true negative TN, false positive FP, and false negative FN. Where the accuracy measures the ratio of the number of correct predictions to the total number of predictions (the number of test data points) >> (TP+TN/(TP+FP+TN+FN)).

Where the recall is measure number the true positive TP over the sum of the true positive and false negative. While precision. Whereas the precision is defined as this formula (precision= TP/(TP+FP)).

Project Design

- Exploring the Data
    - loading necessary python libraries
    - loading the graduate admission dataset
    - note that the last column from this dataset "admit_chance" will be our target label, all other columns are features.
- Preparing and cleaning/ Preprocessing Data
    - shuffle and split data into features and target label
    - cleaning data if not cleaned
    - normalizing numerical features
    - split the data into training and testing sets
- Evaluating Model
    - Creating a Training and Predicting Pipeline
- Final Model Evaluation
    - Creating models for prediction
    - Creating a Training and Predicting Pipeline
    - Choosing the best model
- Model Tuning
- Extracting Feature Importance

References

[1] “The Problem in Graduate Admissions Is Culture, Not Testing - ETS Open Notes,” *Open Notes*. [Online]. Available: https://news.ets.org/stories/problem-graduate-admissions-culture-not-testing/.

[2] L. Cassuto, “Inside the Graduate-Admissions Process,” *The Chronicle of Higher Education*, 01-Feb-2016. [Online]. Available: https://www.chronicle.com/article/Inside-the-Graduate-Admissions/235093.

[3] *RSNA Pneumonia Detection Challenge | Kaggle*. [Online]. Available: https://www.kaggle.com/shraban020/predicting-admission-by-logistic-regression/data.