A Model of College Admission Decision-Making Using Machine Learning

Kanadpriya Basu, Treena Basu, Ron Buckmire, Nishu Lal

March 29, 2019

1 Abstract

Every year academic institutions invest considerable effort and substantial resources to influence, predict and understand the decision-making choices of the applicants who have been offered admission. In this paper we explore and compare several supervised machine learning classification techniques to develop a mathematical model to help predict whether a student who has been admitted to the college will accept that offer. Four years of data on students admission to a small liberal arts college are used to build the classifiers and we evaluate the performance of these algorithms using the metrics of accuracy, precision, recall, F_β score and area under the receiver operator curve (AUC). The results from this study indicate that the logistic regression classifier performs very well in determining whether the student will accept the offer extended by the college. This algorithm can prove to be invaluable by helping institutions target individuals with low chances of following through on an admission offer through emails, texts and counseling to reach target enrollment.

2 Introduction

Occidental College is a small, highly selective liberal arts college in North East Los Angeles with a mission "to provide a gifted and diverse group of students with a total educational experience of the highest quality–one that prepares them for leadership in an increasingly complex, interdependent and pluralistic world." Each year the college receives an average of 6,292 applications of which around an average of 44% applicants are offered admission and an average of 535 students enroll.

Occidental College collects various kinds of data about its admitted students such as Gender, Ethnic Background, Top Academic Interest (categorical variables), First Generation to College and Campus Visit Indicator, (binary variables), High School GPA, ACT score or SAT score (numerical variables) just to name a few.

3 Problem Statement

With the four years of confidential data the college has provided about the annual admitted pool of students the goal of the project described in this paper is to make a prediction regarding each student and classify them into one of two categories: "admitted student who will accept the admission offer" and "admitted student who will reject the admission offer". In other words, we are working on a *binary classification problem* using *supervised machine learning* with the ultimate goal of being able to classify new instances, i.e. when we have a new admitted student we would like to able to predict whether or not that student will accept or reject the admission offer.

Below in Figure 1 is a flowchart that represents the admission whittling process with the statistics for the classes of 2018 through 2021 (students admitted in 2014 through 2017):

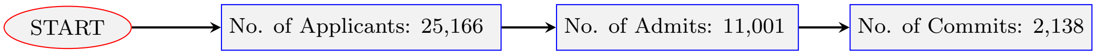

Figure 1: Summary Numbers for Class of 2018 through Class of 2021

We will use the term 'admits' to represent students whom the college has extended an admissions offer to and 'commits' to represent the students who have accepted the offer. Note that international students and early decision students have been omitted from our dataset and thus we have reduced our dataset from $11,001$ to $9,626$ admits.

The eight machine learning techniques implemented in this paper are:

- *Logistic Regression* (LG)
- *Naive Bayes* (NB)
- *Decision Trees* (DT)
- *Support Vector Machine* (SVM)
- *K-Nearest Neighbors* (kNN)
- *Random Forests (RF)* (parallel ensemble),
- *Gradient Boosting* (GB) (sequential ensemble),
- and one deep learning technique: *Multi-Layer Perceptron (MLP)*

4 The Data

Preprocessing is an important step in any data science project; the aim is to cleanse the dataset and prepare it to be further used in a prediction algorithm. The confidential data received from Occidental College was organized and fairly clean and thus we needed to make only a few changes in order to make the data suitable for our chosen machine learning algorithms.

The data consists of 36 pieces of personal information about admitted applicants. Using these available features, our main task in this study will be to focus on the binary response of whether an admitted applicant to Occidental College will follow through on the offer, i.e., will they accept or reject the admission offer.

A standard challenge in data cleaning is to determining how to deal with missing data. Simply disregarding data with missing entries results in loss of valuable information for the model to learn from and incorporating data with missing entries may compromise prediction results. It is important to identify the features with missing entries, locate such entries and implement some sort of treatment that allows us to incorporate the data into the model, since the feature in question may be a strong predictor in determining the algorithms outcome. For example, treatments may include replacing missing entries of a numerical variable with its median value and performing imputation technique to replace missing entries of categorical variables. The cleansing of the dataset results in $7,976$ admits.

Below in Table 1 we present the comparative statistics for the last four years starting from 2014 (Class of 2018) to 2017 (Class of 2021).

Table 1: First-Year Commit Data (Raw)

Applications	Class of 2018	Class of 2019	Class of 2020	Class of 2021
Number of Applicants	6,071	5,911	6,409	6,775
Admit Number and Admit %	2,549, 42%	2,659, 45%	2,948, 46%	2,845, 42%
Number Enrolled and Enrolled %	547, 9.01%	518, 8.76%	502, 7.83%	571, 8.42%

We observe and address the inherent class imbalance of this dataset. Since only 17% of all admits accept the admission offer (as calculated from the cleansed dataset), simply using the *accuracy* score could result in a false impression regarding the model performance. Instead we use the more suitable metrics of *precision, recall, F_β score* and *area under the receiver operator curve* to overcome the challenges caused by class imbalance.

5 Prediction Techniques

The goal of this paper is to compare the eight classification techniques in the context of college commitment of admitted students. Below in Flowchart 2 we illustrate the main steps executed to arrive at a prediction of a new instance.

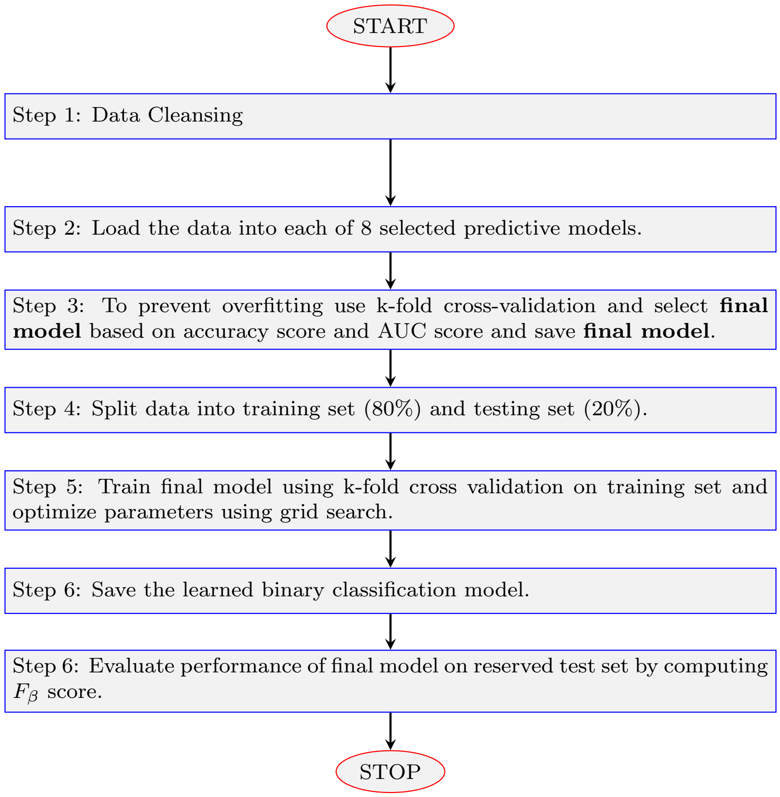
START
Step 1: Data Cleansing
Step 2: Load the data into each of 8 selected predictive models.
Step 3: To prevent overfitting use k-fold cross-validation and select **final model** based on accuracy score and AUC score and save **final model**.
Step 4: Split data into training set (80%) and testing set (20%).
Step 5: Train final model using k-fold cross validation on training set and optimize parameters using grid search.
Step 6: Save the learned binary classification model.
Step 6: Evaluate performance of final model on reserved test set by computing F_β score.
STOP

Figure 2: Workflow

6 Conclusions and Future Work

Our studies indicate promising results for the role of supervised machine learning algorithms in determining student outcomes in the college admission process. This paper analyzed and compared the results of applying eight classifiers to a data set containing $7,976$ samples representing admitted applicants to Occidental College and their personal attributes. Each prediction technique was trained on the entire data set, and their performance was measured by calculating the accuracy and AUC score through k-fold cross validation. The model with the best AUC score (Logistic Regression) was then selected. To measure the effectiveness of the best model, we divide the data into two parts, and train the model on 80% of the data and finally test it on 20% of the data that was held out. Finally, we optimize the performance of the final model with grid search technique.

We conclude this section with a discussion about some avenues for future research in this area as a result of these findings. Firstly, "better and cleaner data" often beats better algorithms, and designing good features goes a long way so an aspect to consider is how to collect "better and cleaner data", i.e., collect data on features that have more predictive powers than the features we had for this study. This is a difficult question that may require deep knowledge of the problem domain. This might be achieved by designing arrival surveys for incoming students to better understand their reasons for committing to the college. Another approach may be to establish a pipeline for collaborative research between local and district schools with institutions of higher education to detect the extent of college intentions of graduating students. Further studies on feature selection could result in simplification of models and shorter training times in some cases.

7 The Impact of this Research

The future and sustainability of the traditional higher education business model is up for much debate recently, especially with rising costs of tuition, concerns about the debt incurred by students and their family, unsure job prospects and the availability of cheaper options to specialized education such as massive open online courses (MOOCs) [1].

This business model is of course highly dependent on the type of academic institution being considered, such as public colleges and universities, private non-profit colleges and private for-profit colleges. Primary sources of income for institutes of higher education include: student tuition, government funding, endowments, alumni donations and fees-for-service. In order to ensure financial stability, many colleges and universities are trying to operate at full capacity which means increasing their target enrollment.

The cost (sticker price) of attending a small liberal arts college such as Occidental College is $70,182 for the 2018-2019 academic year [2]. If the college falls below its target enrollment by only 10 students this could result in a potential financial loss of $701,820 per year for four years, amounting to a total loss of more than $2.8 million. The reader should note that this only represents a potential financial loss and that some students receive a discount rate. However, this discount rate is not determined until after a student has enrolled. Not to mention the loss of future revenue that these admitted students who don't attend could have generated through alumni donations. Thus, it is very important for an academic institution to definitively know how many incoming students they can expect to enroll, especially if they are relying on the revenue generated from the tuition the students bring in.

The authors acknowledge that data science has not been used very often in the academic business model of determining whether a student will accept or reject an admissions offer. The mathematical model developed in this paper simply requires knowledge of personal traits of individual applicants so that it can be fed into the applied algorithms. Being able to accurately predict whether or not a student admitted to an academic institute will accept or reject the admissions offer can help institutions of higher education reach target enrollment. For many institutions, enrollment management is a key component of academic administration.

For example, suppose the model predicts an applicant will reject the admissions offer. Under these circumstances the institution can focus on increasing communications (e.g.through emails, texts and counseling) with such individuals and perhaps allocate resources to provide them the support they need to help make a decision. Or, alternatively, if our model predicts certain students will accept an offer of admission then the admissions office can focus resources on recruiting these students to actually attend. In the parlance of admissions, this is known as increasing the yield rate of admitted students.

References

[1] L. Lapovsky, TIAA-CREF Institute, The Higher Education Business Model, Innovation and Financial Sustainability. https://www.tiaa.org/public/pdf/higher-education-business-model.pdf

[2] https://www.oxy.edu/admission-aid/costs-financial-aid