Student Fee/Expense prediction Using Machine Learning

Sunny Kumar, Kandula.

University of North Texas at Denton

CSCE 5215 Introduction to Machine Learning

Professor: Solomon, Ubani.

Introduction

Today, fresh high school graduates and people who would want to pursue higher education or probably go to college, have a wide variety of options in terms of Universities to choose from. One primary factor that plays a huge role in making a decision while choosing a place to study or pursue education is finance; i.e. The sum of money that one would have to pay to the University or the other expenses that might incur while having to pay for education. Although there are many credible sources available that can provide with useful information regarding a particular college or university of interest, I would like to create a Machine Learning Model, which will not only provide this financial sum as a prediction based on just the University but also some intrinsic parameters that would pertain to the type of study and course of choice.

Today, this problem most of it is still addressed in person, i.e. students seek help from consultants who can then guide them choosing a University depending on their choice of study and Financial standing. This approach is still very relevant today, according to an article published online by *Technavio* the total revenue or market cap of such Consulting firms is Five Hundred and Fifty Million USD as of 2022, and is expected to grow by Six Hundred Million by 2027. This is no small number considering that these services come at a cost to those seeking help.

If there was a **Free tool** that could provide an rough estimate of how much it would cost to attend particular University of interest, it would then become easier for students to have clarity on being able to attend that University and then start their research about other things pertaining to that institution, in-order to make the decision of choosing a place of study. Although there are many online tools that can perform the task at hand (Estimating student expenses, they still do charge money for their services.). I intend to provide the same service, or provide a Machine

Learning Model, that provides this financial information for free. I intend to help those students who try and perform their own study in-order to decide which college they can or would like to attend.

On the basic outline of this project, I see it as a Regression problem. Initially I intend to use a baseline model that performs the same task of taking student inputs and predicting the total cost. Starting there I would like to work on various other models and compare which of them would work the best for the Dataset at hand. I do not see it as a simple task of fitting a model but rather as Analysis of Regression models and conclusion based on the Analysis performed inorder to choose the model that performs optimally. I would also like to share this work to the field if it might ever be useful to somebody to look up regression models that provide an estimate on target variable being Cost or Money.

Literature Review

Upon performing a Literature Review of the problem that I intend to solve, or provide a solution for I stumbled upon various papers that were pretty interesting but not pertaining to predicting the finances or cost estimate for a University.

One such study was about using Machine Learning for Graduate Admission Prediction, which is a model that can help students predict the probability and chances of students actually getting an admission into a particular University of their interest, this model in turn helps students understand their chances and then target Universities with higher chances of admission. However, I find it amusing but a useful approach since most of the college admissions involve a holistic approach along with the statistics of a student over his/her study in the past. Within this paper, the authors: ***AlGhamdi, (et. al.)*** train and test regression models. They used Linear Regression, Logistic Regression, Decision Tree models in order to perform some experiments, their dataset was also retrieved from Kaggle! Their data consisted of both categorical and continuous features similar to mine.

My second review involves a paper that predicts the price of Diamonds, although this is far from my field of work at the moment, involves the same context on the overview of having to predict a floating value as output that represent sum/money, my idea was to study and understand how they approached to having work with data in order to predict a real world numerical value. This paper summarizes various aspects of having to predict the cost of Y entity, influenced by X features. The authors: ***A. Waad, Al-H. Ekram, K. Bawazeer, Al. Hanan used*** different ML algorithms to predict the price of Diamonds, once again most of the models used were Regression models and Neural Networks. They provided a conclusion along with several metrics for each of this model to conclude that Random Forest Regression performs the best out of all. This is an important outcome to my use, I now intend to use RFR as one of my models while predicting the cost for students.

Having understood their approach of using Machine learning models and tuning them in order to predict prices/costs I was still uncertain about using categorical data like state (location within a country.) and understanding how it influences the price? Since in the real world the entity prices such as Land, Houses and Education can vary from location to location, I would want my model to grasp this as an important feature. The data set used by the authors: Raul-Tomas Mora-Garcia, CL. Maria Francisca, V. R. Perez Sanchez contains a mix of both Categorical and Continuous data related to the city of Alicante, which is a Valencian Community located in Spain. The authors then try to use this data set to predict prices of Houses in that location during Covid. And also understanding how they split the dataset inorder to search for hyperparameters while training was thought provoking.

After reading these papers, my approach has changed a bit into having first understand the influence of each of the categorical features, and then map them using encoding to perform similar analysis to that used in Diamond Prediction by the Authors of paper Algorithms for Diamond Price Prediction. Infuse this understanding, find hyperparameters for my models and then use the best model to make predictions for Fee.

Data Exploration

The dataset I chose for this task is available on an open source platform Kaggle. This dataset titled, ‘Average Cost of Undergrad by State in United States’ is posted by a user named: Kenmoretoast. According to the user who had posted this dataset on Kaggle, they have compiled this dataset from the National Center of Education Statistics Annual Digest.

Looking at the dataset it is clear that this data set is a mix of both Categorical and Continuous data. This data set needs to be preprocessed inorder to be fitted to a Regression model. Although this dataset only contains Six features including the target variable. It has an abundant count of rows for proper having the ability to train and test a model by splitting this entire set into Train and Test sets. One challenge would be bringing down the variance of each categorical feature after encoding, this could lead to overfitting the model. For example this data set contains a feature called state, which indicates the state of study there 50 unique values to this feature and even after encoding this results in increased dimensionality, which makes the training a bit more complex.

I tested this dataset to have some assumptions and then verify them:

- Private Institutions cost more than Public Universities.
- State of study also has a significant influence on the cost of education.
- Four year degrees take over in terms of cost incurred to that of a 2 year degree.

These are my initial observations without performing any data cleaning or visualization. I will speak more on this in coming instances of reports, with visualization.

The categorical fields within this data are, ‘Year’, ‘State’, ‘Type’, ‘Length’, ‘Expense’, and there is only one continuous value which happens to be the target variable, ‘Value’. And has close to 3550 data points/rows.

References

Technavio. (2022, October). Education consulting market industry analysis. Retrieved March 18, 2023, from https://www.technavio.com/report/education-consulting-market-industry-analysis

Alghamdi, A. M., Barsheed, A., AlMshjary, H., & Alghamdi, H. M. (2020). A Machine Learning Approach for Graduate Admission Prediction. Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing. https://doi.org/10.1145/3388818.3393716

Al-Suraihi, W., Al-Hazmi, E., Bawazeer, K., & Alghamdi, H. M. (2020). Machine Learning Algorithms for Diamond Price Prediction. Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing. https://doi.org/10.1145/3388818.3393715

Mora-Garcia, R., Cespedes-Lopez, M., & Perez-Sanchez, V. R. (2022). Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times. Land, 11(11), 2100. https://doi.org/10.3390/land11112100

Avg Cost of Undergrad College by State. (2023, March 17). Kaggle. https://www.kaggle.com/datasets/kfoster150/avg-cost-of-undergrad-college-by-state