Analysis of Cost of Attendance for College Students using CRISP-DM

06.24.2021

—

Sri Lakshmi Tammineni - 11482201

SriLakshmiTammineni@my.unt.edu

Anusha Putta - 11500342

AnushaPutta@my.unt.edu

Overview

The aim of the project is to apply our design and understanding of various data extraction, cleaning, analysis and modeling techniques learnt during the length of our course and evaluate the dataset provided. For the purposes of our project we are analyzing the Cost of Attendance data set provided in our Data Sources section of the course project. The dataset consists of 1368 observations with 65 variables describing each observation. The data consists of observations both from undergraduate and graduate levels of education.

Goals

Below are the major goals for this project:

1. Successfully execute a data analysis project on the provided dataset using CRISP-DM process.
2. Prepare the data set for analysis (Describe and Clean Data and Verify data Quality).
3. Identify significant patterns and issues that are faced by the students in this dataset. Also, able to put forth a research problem statement.
4. Arrive at a significant conclusion(s) about the data and provide recommendation(s) to fix the issue(s). In this we will attempt to predict the cost of attendance using the CRISP-DM framework.

Specifications

- We will use Jupyter notebook (Python programming) and hosting on Google Colab.
- We will save a common dataset (post cleaning and preparation) to a shared folder in OneDrive.

Project Abstract

College education is expensive. College costs have increased over the last few decades and this trend is expected to continue. Typical college cost of attendance consists of several components and in this project we study a dataset consisting of 1368 students' cost of attendance data and all the various components. We would like to understand the trends of individual components of the cost and also the overall trend of cost of attendance. We would like to understand why college costs can or can't college costs be kept lower. We hope to analyze the data and understand the trends and hopefully predict the future costs.

This analysis might help future students with valuable insights into how much one should save for education, planning for education funding or what components of the costs can be managed effectively to keep the overall cost of going to college for higher education. While these are our proposed aims, we are open to let the data mould our research and conclude our findings.

As part of this project we apply our design and understanding of various data extraction, cleaning, analysis and modeling techniques learnt during the length of our course and evaluate the dataset provided. The dataset consists of 1368 observations with 65 variables describing each observation. The data consists of observations both from undergraduate and graduate levels of education.

Project Milestones

Business Understanding

This initial phase focuses on proposing project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

Imagine for a future student who needs estimates on what the cost for undergraduate and graduate could be. So we are trying to predict the cost of college based on its attributes. We try to find out the answers to the following 3 business questions:

1. *Can we predict the future cost of education?*
2. *Which components of college cost should students concentrate on to reduce expenses?*
3. *Can our prediction help future students in choosing a college with reasonable costs?*

Data Understanding

In the Data Understanding phase, we acquire within the project the data listed in the project resources. Possibly leads to initial data preparation steps.

- Examine the properties of the acquired data.
- Distribution of key attributes.
- Examine the quality of the data, addressing the question is it with data completed?
- Are there any missing values in data?

Data Cleansing and Preparation (Ms Excel and Python)

In this phase, we need to construct the final dataset from the raw dataset by formatting the data as below:

- Classify into sub-dataset for undergraduate and graduate students
- Calculate additional fields based on tuition and fees
- Evaluate and address missing values
- Evaluate and address erroneous values
- Evaluate and address duplicate values

Modelling (Supervised Approach)

Classification algorithms being considered for analysis:

1. KNN
2. RandomForest

We are planning to work with different model techniques to study the data better and achieve optimal results. In particular, we hope to use simple KNN classification and RandomForest modelling techniques. We chose the precision of classifying Undergraduate and graduate students' cost trends.

Evaluation

Thoroughly evaluate the model and review the steps executed to construct the model to be certain it properly achieves the business objective.

Using the dataset and the results of our machine learning and exploratory data analysis, we can produce graphs and images to show the details of cost trends for both undergraduate and graduate students.

Deployment

The knowledge gained will need to be organized and presented in a way. However, depending on the requirements, the deployment phase can be generating a report.

To this point, we have successfully studied the data, understood the various dimensions to each observation, cleaned the data to remove inaccuracies, and eliminated incomplete data points. We classified the dataset into two sub categories one for undergraduate and another for graduate students. Then we have been performing exploratory data analysis in Python, using graphs and plots to uncover patterns.

In project participation:

1. Business Understanding - These phases of the project were done individually and we had several brainstorming sessions to share our knowledge with each other.
2. Data Understanding - This phase of the project was done by Anusha.
3. Data Cleaning/Preparation - This phase of the project was done by Sri.
4. Mid Checkpoint Report - This document was created by Sri based on discussions with Anusha.
5. Modeling - We are both working on this phase individually with frequent Zoom meetings and knowledge sharing sessions.
6. Evaluation - We are planning to work on this together and summarize our observations, results as a part of the final report.