CARLETON UNIVERSITY

COMP 4107 Project

Gabrielle Latreille (101073284)
Connor Stewart (101041125)
Yves Ganza (101053186)

Abstract

Contained in the report herein, we discuss the design and implementation of a K-fold cross-validation neural network model on a set of College Scorecard graduation statistics. The purpose of the model is to build a testing network to determine if a connection exists between college-major choice and starting salary, student debt, school choice, and more. As per the warnings of other researchers, we carefully looked to verify if the randomly selected set of samples – the College Scorecard CSV spreadsheet – represents the unseen data correctly or if the sampling methods used by the researchers add too much variation to the dataset. Ultimately, the dataset itself was cleared of unneeded columns and suppressed rows of data were removed from consideration. The experiments indicated that a relationship between college major and starting salary, student debt, etc., is weak and likely causal. In conclusion, we see that the results of the experiment were somewhat successful.

Introduction

There are commonly assumed statistically significant correlations between college majors and school choice with graduate employment outcomes. Many people commonly believe that graduates from university commonly enter various occupations related to what they chose to study in college and are compensated in proportion to the degree they earned. The purpose of this writeup is to verify the claim that there is a causal correlation between college major, school, and graduate outcomes. Primarily, the goal is to determine the institution someone obtained their degree from by analyzing a College Scorecard Dataset. Given user input indicating starting salary after graduation, monthly earnings, median total debt after graduation, monthly loan payment, average annual cost to enroll in the program, whether the student is receiving a federal loan, the student's race/ethnicity, and SAT/ACT scores upon entering college, we are to determine from which school they obtained their degree and what major they chose. Furthermore, the student will input their degree level between a certificate, associate's, bachelors, masters, first professional, and doctoral degrees to control for education level.

A primary goal is not just to predict a connection between a student's college major or school and the starting salary after graduation, but to find if a connection exists at all. In this sense, if the project results turn up negative, and no neural network model can deduce a valid accuracy range, we can conclude that no relationship was found. Therefore, failing to build an accurate model can be interpreted as evidence for the claim that no genuine underlying relationship exists. However, finding a good relationship would provide strong evidence for the claim that the choice of college major and the school does indeed have a valid correlation with the starting salaries of recent graduates. In this neural network, a high accuracy score would indicate a strong correlation between college major and college with starting salary, whereas results close to zero indicate no relationship. Accuracies around fifty percepts correct indicate that a causal relationship may exist because we have a good chance of guessing the correct college major or school-based off starting salary. In fact, due to there being so many different schools and college majors to choose from, accuracies as low as the twenty-to-thirty percent could indicate a causal relationship exists, as that still represents a strong chance of a correct prediction when we account for the dozens of majors and schools included in the spreadsheet.

Related Works

Our data is sampled from the U.S. Department of Education's College Scorecard. The dataset is publicly available online at https://data.ed.gov/dataset/college-scorecard-all-data-files-through-6- 2020/resources [3]. Note that the College Scoreboard also has a functional graphical user interface at https://collegescorecard.ed.gov/. The dataset consists of a limited number of students from a limited number of schools placed into a CSV spreadsheet [3].

The research concept presented in the report herein is akin to the one presented in A closer look at cross-validation for assessing the accuracy of gene regulatory networks and models by Tabe-Bordbar et al. This journal article depicts cross-validation as a technique to assess the generalizability of a model to unseen data which relies on assumptions that may not be satisfied when studying genomics datasets [2]. A problem discussed in the paper is that cross-validation assumes a randomly selected set of samples – the test set – represents the unseen data correctly [2]. In general, this is a similar underlying problem for the data in the report. We do not know if the data truly represents the underlying assumption that there is a connection between college majors and starting salary. As per the paper, the assumption does not hold where samples are obtained from different experimental conditions [2]. This is directly analogous to the College Scorecard dataset, we do not know who chose to participate and send their data during the survey process, and we do not know how schools chose to include or suppress information in the College Scorecard statistics. In general, we can say that we do not know if all the schools sent and collected data in the same manner or represented data the same way. We can see how this could skew the dataset and muffle the statistics or determine the correlation between college majors and salaries since there might be jumps in data between schools due to how the data was initially collected.

Data

The dataset contains a large pool of information available in a CSV file format [3]. Note that some extra columns may be included in the algorithm given the size of the dataset if needed. Additionally, there is missing data due to corruption and suppressed data due to state and school laws, which we consider and remove. Data is organized and processed according to degree level to maintain controls over the correlation between primary and school with salary outcomes. For specific disciplines, a large number of people may complete only an undergraduate degree, whereas, in other fields, a large number of students may obtain graduate degrees. Therefore, comparing degrees only by the name of the major can cause mismatches in the predictions. For example, if half of the psychology majors in the spreadsheet obtained an undergraduate degree and the other half obtained a master's, Ph.D., or professional degree. The predicted salary values would fall in the middle between the undergraduate and graduate ranges, resulting in the program mispredicting both of the major degrees. Therefore, to find the actual correlation between major degrees and the school/salary on graduation, we should not mix different degree levels into the same category.

The college scorecard dataset consists of many different categories of data over different years of study. For this assignment, we decided to use data from the 2018 period. The spreadsheets were reduced to contain only columns with relevant information for the neural network. In general, multiple different datasets seem to be present in the college scorecard dataset over the years the study was conducted; however, the core information used for this application is present in all the graduate outcomes surveys on college scorecards [3].

Methodology

During the project, we decided to use K-fold cross-validation with a neural network model. The K-fold cross-validation statistical model is a method for evaluating and comparing learning algorithms by dividing data into two partitions, with the first K-fold partition used to train a model and the other K-fold partition being used to validate the model [1]. Typically, cross-validation training and validation sets must cross over in successive rounds so each data point can be checked for validation [1]. The form of cross-validation used in the project was K-fold, as it is the standard form of cross-validation [1]. In K-fold cross-validation, data is partitioned into k equally sized partitions, with k iterations of training and validation tests being performed such that within each iteration, a different fold of the data is reserved for validation with the remaining (k – 1) folds being used for learning [1].

The program is to use TensorFlow to obtain results from the dataset. TensorFlow is used to build the neural network model and evacuate the folds. The model generated uses fully connected layers and relu activation, followed by a final layer with dimensionality equal to the number of the program using softmax activation. The model is compiled using an adam optimizer with sparse categorical cross-entropy loss. The metric being analyzed is the accuracy metric for the model.

A second set of models to obtain results from the dataset for specific degree levels was also produced. One model consisted of three dense layers with relu activation and dimensionalities of thirty-two, eight, and two, respectively. A final layer of length equal to the number of programs with softmax activation was also used. Another model used three dense layers of dimensionalities of sixty-four, thirty-two, and eight, respectively, with the first two using relu activation and the last using softmax activation.

In this neural network model, we compare the results of student starting salaries to what the neural network *thinks* they majored in for their degree. The network attempts to find a correlation between a student's college major choice with starting salary, dept, etc., to find a trend. Therefore, the model checks if the starting salary by major matches the network's prediction by taking lines of information by the CSV file, extracting the starting salary, then cross-checking the network results with the major name in the original file. Since there are dozens of majors to choose from, even small accuracies could indicate a causal relationship between major and starting salary exists predictably.

Results

During testing, it was determined that more labels resulted in less accurate predictions for the model and that more models result in lower accuracy scores. Initial testing with a low number of models across all datasets and degree levels (i.e., undergraduate, masters, Ph.D., etc.) resulted in an accuracy rate average of approximately 60% averaged across ten K-folds. After breaking up the testing sets into four separate categories for undergraduate, master's, doctoral, and professional degree levels, the accuracy of 6%, 11%, 19%, and 35% was obtained. These abysmal results were primarily the result of using too many testing labels and programs, which resulted in less accurate results from the model.

Discussion

We were able to conclude that the neural networks were not remarkably accurate. The accuracy of the networks means that if there is a relationship between college major and salary, it may be a weaker relationship than we initially thought. We, however, note the existence of a causal relationship between college majors and student starting salaries/debts. The relationship is evident because it predicts the student's major much more accurately than random guessing.

Conclusion

In conclusion, the datasets possess a causal correlation between a student's starting salary/debt and the college major chosen. The results of the neural network model showed that we could predict a student's significance much better than randomly guessing. In conclusion, the neural networks prediction model was weaker than initially thought but still strong enough to allow predictions.

Citations

[1] P. Refaeilzadeh, L. Tang, and H. Liu, “Cross-Validation,” *Encyclopedia of Database Systems*, pp. 532–538, 2009.

[2] S. Tabe-Bordbar, A. Emad, S. D. Zhao, and S. Sinha, “A closer look at cross-validation for assessing the accuracy of gene regulatory networks and models,” *Scientific Reports*, vol. 8, no. 1, Apr. 2018.

[3] “Data Profiles - Department of Education Open Data Platform,” *Open Data Platform*, 02-Dec-2020. [Online]. Available: https://data.ed.gov/dataset/college-scorecard-all-data-files-through-6-2020/resources. [Accessed: 25-Apr-2021].