Ashley Nalikka

Final Project Report

Outcomes for Prospective and Current College Students

Problem

In this project, the first problem I was trying to discover was if there was a connection between college students dropping out and whether their mother had an occupation and whether or not they had educational needs. The second problem I was trying to discover if there was a connection between college students graduating and whether their mother had an occupation and whether or not they had educational needs. The third problem I was trying to investigate was if there was a connection between students planning to go to college and their parents’ income and their IQ.

Approach

My approach to solving these problems is using a Classification model. The target variables in each of the data sets did not come binary so I will use the category encoders to transform them into binary so I can use them in the model. I will use two X variables and one y variable. I will use standard scaler to fit the X variables and scaler to transform the X variables. Then I will perform the train, test, split on the X variables and y variable. Then I will find the test accuracy and the training accuracy. Finally, I will print the Classification report and the Confusion matrix report.

Findings

Drop Out vs Didn’t Drop Out

For the first model I performed a classification report using mother’s occupation and special needs as the x variables and whether or not they dropped out as the y variable.

Training Classification Report

Didn’t Drop Out

The classification report for training produced a precision scoring of 70% which means 30% of students in the model claimed to have dropped out when in reality they did not drop out. These 30% were false positives. The recall had a scoring of 98% which means 2% of students in the model said they stayed in enrolled in college but in reality they did drop out. These 2% were false negatives. The F1 score was 81% thus it says that the model accurately predicts students not dropping out.

Drop Out

This part of the classification report had a precision score of 71% meaning that 29% of students in the model said they didn’t drop out when in reality they did. These 29% were false positives. For the recall it had a score of 12% meaning that 88% of students in the model said that they did drop out when in reality they didn’t. These 12% were false negatives. The F1 score is relatively low with a score of 21% which means it does not accurately predict students dropping out.

Accuracy is 70% meaning this model is decent overall.

Test Classification Report

Didn’t Drop Out

The classification report for testing produced a precision scoring of 71% which means 29% of students in the model claimed to have dropped out when in reality they did not drop out. These 29% were false positives. The recall had a scoring of 97% which means 3% of students in the model said they stayed in enrolled in college but in reality they did drop out. These 3% were false negatives. The F1 score was 81% thus it says that the model accurately predicts students not dropping out.

Drop Out

This part of the classification report had a precision score of 64% meaning that 36% of students in the model said they didn’t drop out when in reality they did. These 36% were false positives. For the recall it had a score of 11% meaning that 89% of students in the model said that they did drop out when in reality they didn’t. These 89% were false negatives. The F1 score is relatively low with a score of 19% which means it does not accurately predict students dropping out.

The accuracy is 71% meaning this model is also decent overall as well.

According to the confusion matrix, this model is good at predicting people who have dropped out due to their mother’s occupation. But bad at predicting whether people dropped out based off of their mother’s occupation and educational special needs.

Graduate vs Didn’t Graduate

For the second model I performed a classification report using mother’s occupation and the age at enrollment as the x variables and whether or not they graduated at the y variable.

Training Classification Report

Didn’t Graduate

The classification report for training produced a precision scoring of 68% which means 32% of the model dropped out but in reality they didn’t drop out. The 32% were false positives. The recall had a score of 52% which means 48% of students did drop out but in reality they did. The 48% were false negatives. The F1 score is 59% meaning that this model does a mediocre job of predicting who didn’t graduate.

Graduate

Precision had a scoring of 61% meaning that 39% did graduate according to the model but in reality they did not graduate. The 39% were false positives. Recall had a scoring of 75% which 25% did not graduate according to the model but in reality they did graduate. The 25% were false negatives. The F1 score is 67% which means this model is okay at predicting who graduated from college.

The accuracy score is 63% meaning this model is average.

Test Classification Report

Didn’t Graduate

The classification report for testing produced a precision scoring of 64% which means 36% of the model dropped out but in reality they didn’t drop out. The 36% were false positives. The recall had a score of 44% which means 56% of students did drop out but in reality they did. The 56% were false negatives. The F1 score is 52% meaning that this model does a mediocre job of predicting who didn’t graduate.

Graduate

Precision had a scoring of 57% meaning that 43% did graduate according to the model but in reality they did not graduate. The 43% were false positives. Recall had a scoring of 75% which 25% did not graduate according to the model but in reality they did graduate. The 25% were false negatives. The F1 score is 65% which means this model is okay at predicting who graduated from college.

The accuracy score is 60% meaning it is also an average model as well.

Confusion Matrix

The model is really good at predicting that age in enrollment is not a factor for graduating. The model is not noticeably bad at predicting in other areas.

Planning to go to College vs Not Planning to go to College

For the third and final model I used parent’s income and IQ as the x variables and whether or not they planned to go to college as the y variable.

Training Classification Report

Don’t Plan

The classification report for training produced a precision scoring of 82% which means that 18% of students in the model said they weren’t going to college when in reality they were planning to go. The 16% were false positives. The recall had a score of 91% which means 9% of students in the model said they weren’t going to college when in reality they were planning to go. The 9% were false negatives. The F1 score is 81% meaning that this model perfectly predicts who does not plan to go to college.

Plan

The precision score is 77% meaning that 23% of students in the model said they were not going to college when in reality they were planning to go. The 23% were false positives. The recall had a score of 59% meaning that 41% of students in the model said they were going to college when in reality they were planning not to go. The F1 score is 67% meaning that this model is ok at predicting who plans to go to college.

The accuracy score is 81% meaning this model does an excellent job of predicting who plans or doesn’t plan to go to college.

Testing Classification Report

Don’t Plan

The classification report for testing produced a precision scoring of 82% which means that 18% of students in the model said they weren’t going to college when in reality they were planning to go. The 18% were false positives. The recall had a score of 89% which means 11% of students in the model said they weren’t going to college when in reality they were planning to go. The 11% were false negatives. The F1 score is 86% meaning that this model perfectly predicts who does not plan to go to college.

Plan

The precision score is 71% meaning that 29% of students in the model said they were not going to college when in reality they were planning to go. The 29% were false positives. The recall had a score of 58% meaning that 42% of students in the model said they were going to college when in reality they were planning not to go. The 42% was false negatives. The F1 score is 64% meaning that this model is ok at predicting who plans to go to college.

The accuracy score for this portion of the model is 80% meaning that this model accurately predicts whether or not people plan to go to college.

Confusion Matrix

This model is good at predicting parent income as a factor of planning to go to college. It is not necessarily bad at predicting other factors but the lowest score is 116 meaning that the model is the worst at predicting parent income and IQ as a factor of planning to go to college.

Further Research

For more in-depth studies on the topic, a data scientist could acquire another dataset that has data on the various colleges and universities to research possible outcomes for students that apply to these schools. It wouldn’t necessarily have to be machine learning. I think a detailed report with some visualizations could explain the outcomes that way people who are not experienced with data analytics or data science can understand the reports. I would use this as a continuation from my classification model on Planning to go to college. It was my best model out of the three and this study would complement the results very well.

Recommendations

First I would address who my clients would be: high school guidance counselor, health care providers for people with mental health or other disabilities and an admissions counselor to a university. For the high school guidance counselor, I would recommend that they pay extra attention to students with disabilities physical and mental and if the school cannot provide sufficient resources for them to succeed, then to recommend them to other facilities or organizations. For health care providers, I would suggest they give their patients proper documentation for accommodations so that they can do well in school and succeed. For the admissions counselor after showing them my model that deals with parent income and mother’s occupation, I would recommend that they look out for such students that have talent and good academics but are not necessarily from wealthy backgrounds and connect them to programs or financial aid that will help them succeed.