2023

Predictive Analytics Case Study

DSC630-T301 PREDICTIVE ANALYTICS

JOSHUA GREENERT

Introduction

What was the problem being solved?

The purpose of the predictive analytics case study was to determine the reasoning for first- and second-year dropouts for a specific South African University; afterwards, the objective was to provide support through potential outreach programs to ensure the students suspected of dropping out remained for the duration of their education.

Why was this problem important to solve?

The university was unaware as to why their first-year students were dropping out at an alarming rate. E-Learning materials, such as Blackboard, were already implemented but didn't have the capability to show data or results until the year had ended; therefore, the university needed a method to understand why dropouts were occurring to proactively improve student retention while ensuring student success (Lourens & Bleazard, 2016).

How was the data acquired?

The data were obtained ethically with the permission of the university from the Higher Education Data Analyzer (HEDA) using SQL, then placed into an open-source application called Konstanz Information Miner (KNIME). The information included a variety of variables such as background, demographics, performance-linked data along with other university-related values such as degree program and whether students were attending the school using financial aid. While the initial case study included approximately twenty-seven unique variables, the number of variables were reduced to eight after accounting for redundancies and possible overfitting (Lourens & Bleazard, 2016).

Methods and Results

What steps were taken to prepare the data?

From the total 1,593 records that were obtained, these were reduced to approximately 452 records after removing data with missing values and students that weren't either first-year attendees or second-year dropouts from 2008 to 2014. The variables within the dataset were transformed for certain data to avoid overfitting, and the remaining values were clustered and binned. All the data was split into three separate groups: the training set (used to train the data), the test set (used to test the findings after the model was completed), and the validation set (an additional set of data to further confirm that the test data wasn't overfitted).

How was this problem solved?

Three different modeling techniques were used to find the greatest accuracy for the remaining variables while avoiding overfitting as well. A confusion matrix was made for all modeling techniques separately to compare the p-values with interests to retain values that held a significant result ($p < 0.05$). After discovering that the Logistic Regression model provided the best accuracy, the model was then applied to the test set, then the validation set to confirm the training set's conclusion. The results of the predictions provided a list of students that were at risk of dropping out; these results were exported from KNIME back into the HEDA management system for the university to complete their student retention efforts.

What modeling techniques were used?

The main modeling techniques used for this predictive analytics case study were the Logistic Regression, Decision Tree, and Naive Bayes models. An important note about Naive

Bayes is that it accepts both categorical and continuous data; therefore, the option to make dummies wasn't essential for this specific modeling technique.

Why did the team choose the methods/models they did?

The reason these models were chosen was primarily because they are supervised modeling techniques (i.e. there is a target to check against for the results) and the goal of the case study was to predict future student expectations.

What metrics were used to evaluate the results? Why was this metric chosen?

The metrics used to compare all three methods were Area Under Curve (AUC), Percentage Correctly Classified (PCC), and the error percentage based on each model's performance. The advantage of using AUC is that it can provide a score that is based on the model's ability to differentiate between positive and negative classes (Bhandari, 2020). PCC and the error rate provide a similar view as the AUC, but provide additional methods to check against the outcome of the AUC. By using these particular metrics to compare the results, the reviewer can ensure that their conclusions have been correctly reached.

Conclusion

How were the results or model implemented?

The results of the case study were sent back to the university to perform an outreach program to their students with a high percentage of dropping out before the end of the first year. With the model already established, it can be updated and reviewed when additional data is made available to contain the most accurate predictions of potential student drop-outs. A recommendation from the reviewers requested that the university also provide first-year students with surveys to increase the accuracy of future predictions.

What were the actionable consequences of the case study?

Directly after the completion of the case study, the university was provided access to a list of students that rated medium to high on the possibility of dropping out. These students can be contacted by the university to retain them for future years while increasing the total revenue of the school. Moreover, surveys can be taken from these students to obtain details on why they might be considering dropping out, or what changes can be made for further retention.

What did the team learn from the case study?

Obtaining the data to begin the predictive analysis was easy. However, returning the data back to the university for the predictions in real time is not feasible.

How should or would the team approach the problem differently in the future?

For this case study, the team used supervised learning methods since there was a target and features to select from for the best predictions. Yet, they chose one modeling technique that could use both categorical and continuous data simultaneously, Naive Bayes. This problem could be approached differently in the future by making dummies for all the categorical variables that persisted in the twenty-two remaining. Additionally, students with the highest ratings for dropping out could have been verified with the school to check for feedback from the professors on the students' performance. This additional check could clarify whether the accuracy of the predictions is working as expected.

References

Bhandari, A. (2020, June 15). AUC-ROC Curve in Machine Learning Clearly Explained. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/#:~:text=The%20Area%20Under%20the%20Curve

Lourens, A., & Bleazard, D. (2016). Applying predictive analytics in identifying students at risk: A case study. South African Journal of Higher Education, 30(2). https://doi.org/10.20853/30-2-583