Westminster College: A Retention Prediction Model

Ryan P. Smith

Northwest Missouri State University, Maryville MO 64468, USA
S547012@nwmissouri.edu

Abstract. The abstract should briefly summarize the contents of the paper in

Keywords: higher education · retention · data analytics · data science

1 Introduction

At Westminster College, the Office of Institutional Effectiveness is responsible for compiling and analyzing data at our institution. This project will focus on the domain of higher education, and more specifically, we will be examining retention. The domain of higher education is being examined in this project as it can be utilized within our department in the future, and hopefully help provide more insight into retention factors at our college.

The data for this project will be sourced from Westminster College. We will be utilizing a couple of our databases from both our Enrollment Customer Relationship Management tool (CRM) Slate by Tehcnolutions, and our main SIS (Student Information System) which is provided by Jenzabar. All the student data used will be anonymized to ensure that there is no personal identifiable information used. The data will be compiled from the many tables into a useable dataset for this project. Data features that will be included in the dataset will be some biographical such as gender, commuter/resident, athlete. Some other data features will be items such as ACT score and high school GPA. This project will look to take those factors, create a multi-variable regression model to predict the likelihood that a student at Westminster College will be retained or not.

1.1 Defining the Problem

For this project, we want to see if retention can be predicted at Westminster College based on the selected data points and a machine learning model. There are multiple reasons why this is an important issue to look at and examine. First there appears to be a shrinking pool of high school graduates as we approach 2030. [5] This means that if there are fewer students in the graduating high school classes to recruit, there will be more competition between colleges and universities to enroll these students. If this leads to lower enrollment at an institution, that means that retention will become even more important to retain the students that you are able to enroll.

We can also see that enrollment has currently been dropping overall for the past several years. [4] The National Student Clearinghouse also illustrates this. We can see the enrollment declines among the different higher educations types, with a slight uptick actually showing in community colleges. [2] Since Westminster College has two prominent sources of revenue, donations and enrollment, it will be important to retain students at a higher percentage than before, if the enrollment of students continues to decline.

1.2 Goals of this Research and Methods to explore

This project will be looking to successfully predict if a student is at risk of being retained. We can use this data with our Student Success Center so that they can set up an intervention, or take the appropriate actions to ensure they can help the student persist. After gathering all the data to analyze, a model will be built and trained on train data, and then tested. After the model is completed, we can discuss the results and see how the model performed.

This project will be following the traditional Data Science Life Cycle. The steps are outlined below as:

1. Business Understanding - as seen above, this will be Predicting Retention at Westminster College.
2. Data Collecting - Data set to be produced from our CRM and SIS.
3. Data Cleaning - Missing data will be handled as well as mismatched data types.
4. Exploratory Data Analysis - Here we will begin to find trends and groups in the data.
5. Model Building - Here the model method will be selected, and then trained and tested.
6. Results - Visualizations will be created, conclusions will be formed and discussed.

The key components that will be focused on for this project will be data from previously enrolled students. We will be focusing only on students that attended Westminster College, and utilize the data that we have access to for each student. Some items may be missing such as ACT from recent years. Enrollment recently went to ACT optional, so there are more missing ACT scores than from the past, as more students are choosing not to submit those for consideration. We will look at including and not including ACT scores as another trend has been the decline in overall ACT scores across the country [3], and the state of Missouri [1] (19.5 and 19.8 respectively). This could be an interesting limitation, so we will try looking at retention with and without this feature.

2 About the Data

Data was collected, cleaned, and analyzed for this project. It came directly from two databases from Westminster College. These two databases come from the

Student Information System (SIS) which is provided by Jenzabar. The other is our Enrollment Customer Relationship Management (CRM) called Slate, by Technolutions. The data was pulled directly from a combination of these two databases, compiled into a usable dataset within a csv file, cleaned using python, all before being analyzed.

2.1 Data Sourcing

The data was sourced from two databases on campus, the SIS and our Enrollment CRM. This data is structured data stored within these two databases. All tables being used to source our data are able to be joined together based on the student ID. The data was sourced directly from the databases, and the results were written to a .csv file to be imported into a python script for cleaning and initial data exploration. There was no scraping needed to gather this data, it was collected with a SQL script using SQL Server. All tables and the two databases were able to be joined together with the primary key of student ID.

When looking back at students entering from the year 2010, there are 2605 records in the data set, when expanded to the year 2000, the amount of records increased to 5265. This project utilized the latter to have more records for analysis. For this project, the data fields were limited to 7 total. The fields in the data set along with their respective data type will be as follows:

1. Unique Identifier: Numeric Integer - Used to identify a unique student within the data set.
2. ACT Score: Numeric Integer - The highest ACT score that a student submitted with their enrollment.
3. Resident/Commuter: Character - denotes is a student resides on campus or commutes to campus
4. Gender: Character
5. Athlete: String - A sport will denote whether the student is an athlete or not (converted to a Y or N)
6. High School GPA: Numeric Float - a float numeric value to show the GPA from the students graduating High School.
7. Graduation Data/Exit Date: date - Currently a data field for both columns, these were combined along with exit reason to establish if a student graduated and was retained or withdrew and was not retained.

Overall, the data is very clean. There are some missing values in ACT scores, which was expected since we became an ACT optional reporting school in terms of enrollment. Looking through the data there are several other missing values, but not enough to have to cut it out. During the cleaning phase those issues will be addressed. There are two date fields that are formatted as YYYY,mm,dd, however, those fields were combined into a single field to determine if the student was retained (graduated) or not (withdrew). There were no bad characters found in the data set.

All of the data compiled from the databases was executed and written to a .csv document. With this format, the data was taken and using various python libraries, cleaned and explored (both examined more in the upcoming sections). Also, as mentioned there are a couple items where were combined from the elements from the database to make a single feature within the dataset. After collecting and compiling all the data, the next section will cover how the data was cleaned.

2.2 Cleaning the Data

Steps taken to clean the data.

2.3 Exploring the Data

A section on expository data analysis to follow.

3 Model Building

3.1 Choosing the Model

Decide which model to use with our data.

3.2 Building the Model

Walk through the building of the chosen model.

3.3 Training and Testing the Model

Create a training and test data set to use with the model.

4 Results

4.1 Clear Summary of Results

What did we find out from our model?

4.2 Visualizations

Share some visuals from the results and show what the data says.

5 Discussion

5.1 Conclusion of Results

5.2 Limitations

5.3 Future Recommendations and Work

References

1. Act scores drop for 6th straight year, https://www.northwestmoinfo.com/local-news/act-scores-drop-for-6th-straight-year/
2. Current term enrollment estimates, https://nscresearchcenter.org/current-term-enrollment-estimates/
3. Castillo, E.: Act scores hit 30-year low, https://www.bestcolleges.com/news/act-test-scores-hit-30-year-low
4. Knox, L.: Leveling off on the bottom, https://www.insidehighered.com/news/admissions/traditional-age/2023/05/24/leveling-bottom
5. Seltzer, R.: Birth dearth approaches, https://www.insidehighered.com/news/2020/12/15/more-high-school-graduates-through-2025-pool-still-shrinks-afterward