USING MACHINE LEARNING MODELS TO PREDICT STUDENT RETENTION: BUILDING A STATE-WIDE EARLY WARNING SYSTEM

A Thesis

Presented to

the Faculty of the Elmer R. Smith College of Business and Technology

Morehead State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Travis Muncie

October 30, 2020

Accepted by the faculty of the Elmer R. Smith College of Business and Technology, Morehead State University, in partial fulfillment of the requirements for the Master of Science degree.

____
Dr. Nilesh N. Joshi
Director of Thesis

Master's Committee: ______________________________, Chair
Dr. Nilesh N. Joshi

____
Dr. Ahmad Zargari

____
Dr. Sam Nataraj

____
Date

USING MACHINE LEARNING MODELS TO PREDICT STUDENT RETENTION: BUILDING A STATE-WIDE EARLY WARNING SYSTEM

Travis Muncie
Morehead State University, 2020

Director of Thesis: ____
Dr. Nilesh N. Joshi

The state of Kentucky lags significantly behind the national average in terms of student retention rates at its 4-year public institutions. To exacerbate the issue, Kentucky high school students continue to enter in-state postsecondary institutions at a lower rate each year. The study of student persistence is prevalent in the academic community, but after reviewing the literature, most studies focus on one-off statistical analyses that are not very actionable in practice. The use of machine learning algorithms to build models is also not very prevalent in the literature even as other sectors have quickly adopted the technology. This study looks to build a system, using machine learning algorithms to predict retention, and provide risk scores to institutions for at-risk students. The data used was obtained through the Kentucky Postsecondary Data System and transformed for analysis using SQL Server Management Studio. Ten years of student records were used including the periods of 2008 to 2018. The records were imported to Python where the scikit-learn package was used to build three separate classification models (naive Bayes, decision tree, and logistic regression). Features were chosen based on independent variables that had predictive power in the literature. The cost-sensitive weighted logistic regression model provided the best results for correctly recalling the highest percentage of students who were not retained on unseen data. The model was able to recall 67% of students who were not retained. It was also able to recall 38% of not retained students in the .08 to 1 probability range with only a 3% false-positive rate. The results suggest that using supervised machine learning classification models are effective in predicting students who will not be retained and can act as a foundation for a system-wide early warning system.

Accepted by: ________________________________, Chair
Dr. Nilesh N. Joshi

____
Dr. Ahmad Zargari

____
Dr. Sam Nataraj

Acknowledgements

First, I would like to thank my thesis director, Dr. Nilesh N Joshi who, even in a time of high uncertainty and change, accepted to act as my thesis director and gave me the guidance and feedback necessary to complete this research. Also, I give my highest appreciation to Dr. Ahmad Zargari and Dr. Sam Nataraj. Dr. Zargari for organizing my thesis committee and providing me with the information needed to meet the standards for thesis completion, and Dr. Nataraj for serving on my thesis committee and guiding me over the last two years as my graduate advisor. The Department of Engineering and Technology Management faculty at Morehead State University are very talented and experienced, and I am blessed to of had the opportunity to learn and grow from their knowledge over the last two years.

I also want to thank the Kentucky Council on Postsecondary Education for allowing me to pursue my graduate degree and my co-workers at the Council for giving me the support and advice during my studies. Last, I want to thank my wife, Mrs. Lisa Muncie for providing me with the encouragement and patience that I needed when my work and school life became difficult. Thanks to all who supported me and provided me with the resources I needed during this journey.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Importance of Higher Education

In the Higher Education's Return on Investment Report (2020) recently published by the Kentucky Council on Postsecondary Education, the value of a postsecondary credential is made evident. In this longitudinal study, 2010 Kentucky high school graduates were tracked through their postsecondary education and then workforce outcomes calculated by their credential level obtained. It found that a person earning an associate degree would earn $442 thousand more than a high school graduate, and a bachelor's degree earner would earn $1.2 million more over a lifetime. These benefits extended to the state as well showing that for the $620 million invested in higher education, the class of 2010 will contribute $43.8 billion to the economy through taxes and spending. With these statistics in mind, one can quickly see the importance of graduating our higher education students at a high rate to improve the workforce talent pool and build a more fruitful economy.

With the understanding of the significance of an educated workforce to the economy, it is important to look at the high school to the postsecondary pipeline, where there are discouraging trends. According to the Kentucky Center for Statistics, in their High School Feedback Report (2019), the rate in which Kentucky high school students are enrolling in a state (private and public) postsecondary institution is slowly declining. In fact, the rate has dropped from 55% in 2014 to just 51.7% in 2018. With the number of students enrolling in postsecondary education declining, and the importance of postsecondary credentials to our economy evident, the question of what to do next should be asked. One answer is that higher education institutions need to do a better job graduating the students that they have, and one of the strongest predictors of earning a credential is the retention of students from their freshman to sophomore year. In my own analysis, using records from the Kentucky Postsecondary Education Data System in SQL Server Management Studio, I found that Kentucky students at public 4-year institutions who were retained from fall 2012 to fall 2013 had a 6-year graduation rate of 70% compared to an overall rate of only 55%. With little hope of the percentage of high school students enrolling in a postsecondary institution rising, the state needs to look for other avenues to increase the proportion of Kentucky residents with a postsecondary credential, and increasing the retention rate seems to be an effective way to do so.

1.2 Problem Statement

The state of Kentucky lags significantly behind the national average regarding retention of students (calculated fall enrollment to following fall enrollment) at its 4-year public institutions. According to the National Center for Education Statistics (2020), the retention rate for first-time, full-time, bachelorette seeking students at public universities (enrolling in the fall of 2017) is 81%. In contrast, the retention rate for that same population in Kentucky is just 77%. A gap also exists when comparing Kentucky's graduation rate to the national average. From the same report, the national 6-year graduation rate for first-time, full-time, bachelorette seeking students at public universities (enrolling in the fall of 2012) stands at 62%. Kentucky was only at 55%. In practical terms, this means a less educated population (when compared nationally) who, in most cases, incurred debt with no earned credential. Many of the state's postsecondary institutions have programs in place to increase retention and persistence. For example, the University of Kentucky has an Office of Retention and Analytics completely dedicated to the subject (https://education.uky.edu/retention), but not all institutions have the analytical resources or staff for such an effort. This is where the state government and particularly the Kentucky Council on Postsecondary Education is uniquely positioned to help.

1.3 Significance & Research Question

The Kentucky Postsecondary Education Data System (KPEDS) contains millions of unit-level student records dating all the way back to 1995. These data include enrollment, course, grades, class inventory, credential, transfer, entrance exam, and financial aid information. This allows the Kentucky Council on Postsecondary Education (CPE) access to a vast amount of information in which to do research. In the case of retention, it allows many variables, across different dimensions, to include in modeling. It also allows for the use of data across institutions and sectors. The CPE's information infrastructure also provides the necessary tools to conduct research. With direct access to KPEDS with open source tools like Python, datasets can quickly be created and analyzed with less labor-intensive methods like machine learning. Even with robust information systems and analysis tools, the CPE Data and Advanced Analytics team is still very small. Therefore, effective and efficient methods like machine learning are important so that automation can be used and repeated without much intervention. Using KPEDS in tandem with machine learning, this research can provide value to Institutional Research Units at the state's 4-year public universities to identify students who are at a high risk of not being retained and then use that information to implement intervention strategies.

The research questions are, using data from the Kentucky Postsecondary Data System, and a machine learning model:

- Can fall to fall retention be effectively predicted for full-time, first-year students of 4-year public postsecondary institutions?
- Can the model then be used to identify risk categories for each student? By extension, will the availability of better information lead to a rise in retention rates for all 4-year public institutions?

1.4 Research Objectives

The primary objectives of this research are as follow:

- To identify and create features that are significant to student retention of the chosen population.
- To create multiple machine learning models and choose the most effective at recalling/predicting students who were not retained based on unseen data.
- To analyze the individual student probabilities created from the model over multiple years and use the frequency distribution to create risk categories based on probability thresholds.
- Recommend a system that can easily be replicated each fall semester, that provides a list of at-risk students to institutions.

1.5 Assumptions and Limitations

The main assumption is that the data used in this research was accurately reported by each institution during the collection process. Data is collected throughout the academic year through multiple collection periods. Reporting guidelines are provided (http://cpe.ky.gov/policies/data/2020-21guidelines-all.pdf) that have extensive definitions and instructions, but there is still a possibility that information was misreported. There are also many limitations on what feature variables can be used in the analysis. This is due to time constraints on when the data is collected and the focus on actionable research. Also, some important features identified in the literate were not available. These variables are mostly psychological and qualitative in nature (surveys, etc.) Limitations also exist with the type of machine learning algorithm that can be used. Since the research is based on creating a prediction of a classification (retained, not retained) a classifier algorithm was chosen.

1.6 Definition of Terms

This section contains a list, along with their definitions, of frequently used terms and acronyms that will be used throughout the thesis. Along with basic terms, it also contains definitions of some of the features that will be used in the research.

Retention: Determined by a student's re-enrollment, in the subsequent fall term of their original enrollment, at the same institution.

Graduation Rate: The proportion of first-time, full-time, fall semester students who graduate with a bachelorette degree within 6 years of initial enrollment.

First-time Student: A student who has graduated from high school and has no prior postsecondary experience attending any institution for the first-time at the undergraduate level.

Full-time Student: An undergraduate student who has enrolled in 12 or more semester credit hours.

High School GPA: The final weighted high school grade point average determined at a student's graduation from high school.

Residency Status: Residency of a student determined by the Council on Postsecondary Education's "Policy on Classification of Residency for Admission and Tuition Assessment Purposes".

URM: Underrepresented minority defined as a student who self-reports their race/ethnicity as Black, American Indian or Alaskan Native, Hispanic or Latino, Native Hawaiian, or Other Pacific Islander, and two or more races.

Low Income: A student is designated low income if they received the Pell Grant in their first semester of enrollment.

Underprepared: A student is designated underprepared if they do not meet the minimum threshold of readiness as determined by the Council on Postsecondary Education's College Readiness policy. (http://www.cpe.ky.gov/policies/collegereadiness.html)

Chapter 2: Literature Review

This literature review will be split into two parts. The first will focus on journal articles and research around the subject of postsecondary student retention theory and frameworks. The second will be focused on predictive modeling for retention of students including studies that used machine learning models. The overview and explanation of machine learning and machine learning algorithms will be discussed in the "Methods" section of this thesis.

2.1 Retention Theory

To act as a foundation for the quantitative research focused on postsecondary student retention, it is very important to first look at the theories and frameworks that a large proportion of the research is predicated on. One person that has had an immense impact on this subject is Dr. Vince Tinto. Dr. Tinto worked as the chair of higher education for the Syracuse University and is best known for his book "Leaving College" (1993) where he created a theory around student retention. It is based on the previous work of Emile Durkheim that focused on suicide. Dr. Tinto believes that student retention outcomes are primarily community-based and therefore the problem of student attrition is mainly dependent on the university community. The reason he felt student attrition as analogous to suicide is because, in both instances, an individual decides to leave their society/community.

In Tinto's paper written for the Maryland College Personnel Association (1987), he goes over the primary principles of student retention and how institutional retention programs can be made more effective. Tinto (1987) frames seven main principles of student retention. These principles are as follows: academic difficulty, adjustment, goals, uncertainty, commitments, incongruence, and isolation. To give a good foundation of Tinto's framework, the next section will be dedicated to his work on the Principles of Effective Retention. A brief explanation of each principle will be given as defined by Tinto (1987).

Academic difficulty focuses on the ability of the student to meet the academic standards of the institution. Tinto (1987) attributes most student attrition of academic difficulty to the prior preparedness of the student for college. Although an important factor, Tinto points out that academic difficulty only attributes to around 20% of all dropouts nationally.

Adjustment is centered around the social transition that a student must make when entering higher education. According to Tinto, this transition is much more difficult for students that do not have backgrounds familiar with college life. These populations include low income, minority, and first-generation students. Tinto views these populations as particularly vulnerable and thinks that without identification and assistance they are very likely to leave the academic institution.

Goals reflect the internal motivation of an individual student. Tinto (1987) points out that many students enter college without intention to graduate. This is also true for students who enter college with the intention to transfer. This is most prevalent in 2-year academic institutions where a student may intend to transfer to a university. Tinto states that a student's goals can either be more limited or less intensive than that of the institution and that goal misalignment can lead to student attrition. Lack of goal clarity will lead to uncertainty if allowed to persist. Without clear goals, a student can become dismayed by the difficulty of college and is more likely to leave.

Even with clear goals, a student's commitment to the goal of college completion is also very important. Tinto (1987) views a student's commitment mostly in the context of student experience after entry. In his view, commitment is mostly a function of individual experience and that relationships with other members of the institution, faculty staff, and other students is the primary mechanism of whether a student will be committed to their education.

Incongruence occurs when the social and intellectual life of an institution does not serve a student's needs and interests. Tinto (1987) uses the example of a student who is bored and not intellectually stimulated by the demands of the academic institution. This, in most cases, leads to transfer and not dropout, but it still counts against an institution when looking at most metrics that are designed to track student persistence.

Isolation is centered around a student's assimilation into the institution's community at large. Tinto explains this phenomenon not as misalignment of goals or intellectual pursuit, but as the failure of a student to build relationships and being separated from college communities.

In his paper "Taking Student Retention Seriously" (1999), Tinto expands on his framework and introduces the concept of "Learning Communities". This strategy attempts to create small communities through block course scheduling ensuring that a group of students are taking their courses together. Tinto explains that this leads to more engagement and learning both in and out of the classroom. Building learning communities builds on the foundation of the principles detailed in Tinto (1987) and attempts to address many of the community-based causes of student attrition.

Outside of Tinto's work, another commonly used theory in the research was that of Dr. Alexander Astin (1984) called the theory of Student Involvement. The theory is rooted in Astin's previous empirical work studying college dropouts and student retention. The theory is based on the importance of student involvement and co-curricular activities. In its simplest form, it looks at desirable higher education outcomes (persistence, graduation, etc.) as a product of both inputs, previous experience, race, demographics, academic achievement, etc. and environment which is the experience of the student in college. The quality of the experience, and by extension desirable outcomes, are predicated on student involvement. Student involvement as defined by Astin is "the quantity and quality of the physical and psychological energy that students invest in the college experience" (Astin, 1984, p. 528). The basic hypothesis behind the theory is that the effectiveness of any educational policy or practice is directly related to the ability of that policy or practice to increase student involvement.

2.2 Empirical Studies

While reviewing the empirical studies focused on student retention, the influence of the theories and frameworks discussed in the previous section was evident. Most studies took a similar approach and focused on predictor variables derived from the theory literature. One of the original empirical studies was done by Dr. Alexander Astin. In the study "Preventing Students from Dropping Out" (1975), Astin used datasets obtained from freshman questionnaires to attempt to predict student attrition. Using a dichotomous (dropout versus nondropout) dependent variable, Astin used 53 independent variables for prediction in a logistic regression model (logistic regression will be explained in the "Methods" section). 110 variables were initially used but reduced to 53 using a stepwise regression technique until variables were no longer significant ($p < .01$). The 53 independent variables were categorized into 6 separate groups. These groups were academic background (high school GPA, high school rank, college admission tests, etc.), family background (religion, race, parent's education level, income, etc.), educational aspirations (terminal degree pursued), study habits (student responses to question about study habits), expectations about college (student's self-prediction of drop out), and other characteristics (containing such things as smoking habits). The population consisted of 38,708, first-time, bachelorette seeking students, but was reduced to 9,750 (randomly selecting every 4th student) due to computational limitations at the time. The population was also split into 4 subgroups: nonblack men, nonblack women, blacks in black colleges, and blacks in white colleges, and then ran through the model individually. The output was a classification (dropout, nondropout) and the likelihood that a student would persist described as “dropout-proneness” by Astin. The variables with the greatest predictive value were all associated with the student’s academic record and academic ability for all subgroups. Astin concluded that the most drop out prone students had poor high school academic records, low aspiration, poor study habits, parents with low educational attainment, and came from small towns.

Marc Scott, contributing to the book “Retention, Persistence, and Writing Programs” (2017) used Big Data Analytics to monitor the effects of his institution’s writing composition assistance program on retention rates and developmental writing composition course pass rates. Scott, the director of the program, collected student IDs and provided them to his institutional research office for evaluation. Using Tableau (data mining and visualization software) he was able to quickly find trends for the students who visited his assistance center. He also used the data provided in Tableau to assess curriculum, which led to changes to the pedagogy of the developmental composition courses. After the change, and with the help of his assistance program, Scott (2017) saw an 8 percent increase in the pass rate in developmental composition courses. He also performed a chi-square test, controlling for the student’s ACT composite score, and determined that the change was statistically significant. This study, although mostly descriptive in nature, showed the importance of data mining and data visualization as a companion to more sophisticated statistical methods.

Garrett, Bridgewater, and Feinstein (2017) studied the impact of first-year composition courses on student retention and future course success using a sample of 2,068 first-time students (both freshman and transfer) obtained from a small university. To analyze the data, they used an Association Rule Mining technique that allowed them to view relationships from many first-year courses and examine their predictive value. They found that failing a first-year writing course was a strong predictor of eventual graduation. They also found that performance in general education courses, particularly a cluster of courses designed for research and language skills (public speaking, writing, and information literacy) had a strong correlation with student retention.

Giani, Alexander and Reyes (2014) studied the effect of dual-credit participation (taking a course in which you receive both high school and postsecondary credit) on postsecondary outcomes that included second year persistence. The dataset contained both high school and postsecondary records provided by the Texas Education Research Center which maintains a longitudinal data system for the state of Texas. Students who graduated on time and matriculated to a postsecondary institution in the year 2004 were then used as the primary population. Propensity Score Matching was implemented to control for self-selection bias. A dichotomous variable was created for each student to identify if they participated in dual credit (1) or did not (0). A logistic regression model was created, and the odds ratios used for comparison. A student who participated in dual credit was shown to be 1.57 times more likely to persist to the second year of college at a $p > .001$ significance level.

Stewart, Lim and Kim (2015) used Tinto's longitudinal model of institutional departure as a theoretical framework to evaluate demographic variables, family characteristics, academic performance factors, and remedial course placement as predictors of persistence at a public research institution. With a sample of 3,213 students obtained through a longitudinal database, the study used ANOVA (factorial analysis of variance), Pearson's product-moment correlations, and multiple regression to analyze the relationship of the variables to college persistence. The study found that High school GPA and first-semester college GPA were the two strongest predictors of persistence. They also found that non-remedial students were more likely to be retained than their counterparts who took remedial courses.

Karen Leppel (2001) used data from the 1990 survey of Beginning Postsecondary Students obtained from the National Center for Education Statistics (n= 4,947) to examine the impact of chosen majors on retention for a freshman. She used a combination of least squares regression and logistic regression for her analysis. Splitting the population by gender, she found that men that entered education and undecided majors were less likely to be retained while men in business majors were more likely to be retained. For women, those who enrolled into business and undecided majors were less likely to be retained while those who entered health majors were more likely to be retained. Retention was defined as a returning enrollment the following year.

Allen and Robbins (2008) used a sample of 48,232 from 25 public universities to attempt to predict if a student would persist to the third year of their chosen major. Using hierarchical logistic regression, the sample was split into an estimation and validation sample. The estimation sample was used to create coefficients and an interest-major composite score derived from the likelihood of persisting in the initial major. Their results suggest that interest-major fit and first-year academic performance both independently predict persistence in a major in the third year of enrollment.

Porchea et al. (2010) looked at how academic preparation, psychosocial, socio-demographic, situational, and institutional factors influenced student success outcomes of students entering community college. Using a sample of 4,481 students who participated in the 2003 Student Readiness Inventory validity study, the researchers used a multinomial logistic regression model to analyze the independent variables to multiple classifications of dependent variables (including dropout) to determine their effect on the likelihood to succeed. Of the variables, High school GPA and the academic discipline had the greatest positive effect on remaining enrolled and obtaining a degree for non-transfer students.

Hu and St. John (2001) used the Indiana Commission for Higher Education's Student Information System to analyze the effect of student financial aid on persistence. The sample consisted of three cohorts of full-time, resident undergraduate students enrolled in Indiana public 4-year institutions. Logistic regression was used and a dichotomous outcome variable of 1 or 0 (retained, not retained). The sample was split into subgroups by race and gender to identify differences in outcomes. The study found that receiving financial aid positively affected the likelihood of persistence in all race groups with a larger effect on the African American and Hispanic groups.

Bogard et al. (2011), of the Western Kentucky University Institutional Research unit, created a system using data mining and machine learning algorithms to predict students who would not be retained. The goal was to create a system that could predict student retention and then create risk indicators to be used for intervention efforts. The population was first-time, first-year, degree-seeking students from three consecutive academic years obtained through their student information system. SAS Enterprise Miner was used for the analysis and SAS BI (decision support system) was used to visualize the results and risk indicators for retention staff. Four models were evaluated on their validation misclassification rates and robustness. The four models were logistic regression, decision trees, neural networks, and ensemble models. A decision tree model was eventually chosen and then trained to predict retention based on three time periods (pre-enrollment, 5th week of term, and end of term) at which point new variables were introduced to the model as they became available. The model gradually improved as new variables were introduced. By the full semester time period (all variables included) the model had a 79% overall accuracy and 75% recall of students who were not retained.

2.3 Summary

In summary, there were many commonalities in both the frameworks employed and the variables used for statistical analysis. Dr. Tinto's work was cited in almost all the journal articles and studies reviewed. Among the independent variables employed to predict student retention, many studies used a mix of academic performance, demographic, financial aid, and student survey data. In almost all studies variables related to academic performance stood out as having the most predictive power regarding retention. Specifically, high school GPA and first-semester GPA were consistently shown to have a close relationship with student success. Logistic regression with a dichotomous dependent variable was the most common statistical analysis technique. Using the information obtained from the literature, this research will attempt to use features (independent variables), or close proxies, that have been shown to have a strong relationship with student retention.

Although there was an abundant amount of empirical studies focused on student retention, most were one-off statistical analyses. What was not very common, with the exception of Bogard et al. (2011), were studies that used machine learning algorithms or created systems to make their research actionable. Therefore, the research performed in this thesis can add value to the larger student retention conversation.

Chapter 3: Methodology

3.1 Research Design

This research is quantitative in nature. Using supervised classification machine learning models, and administrative student data, the research is designed to predict if a student would be retained the following fall semester of entry on a set of features. Machine learning is a subset of artificial intelligence and is used in the data science world to build mathematical models to understand data (Vanderplas, 2017). The focus is on supervised models which revolve around classification or regression algorithms to predict unseen data (Vanderplas, 2017). The data is used to classify students into risk categories to act as an early warning system for academic institutions. This chapter is divided into six sections. These sections are Research Design, Population, Data Collection Method, Tools, Model Features and Procedure, and Data Analysis. The Population section details the count and characteristics of the subjects. The Data Collection Method section details how the data was obtained. The Tools section summarizes the applications and libraries used for analysis. The Model Features section gives a list and definition of the features used to train the models and a description of the dependent variable. Last, the Procedure and Data Analysis section gives a detailed walkthrough of the full research procedure and analytical methods used.

3.2 Population

The population used in this research is first-time, full-time, baccalaureate-seeking students who enrolled in each fall semester at Kentucky public 4-year universities. This includes first-time enrollees at the University of Kentucky, University of Louisville, Western Kentucky University, Morehead State University, Murray State University, Northern Kentucky University, and Kentucky State University. Ten years of student records were used starting from the year

2008 to 2018. Student characteristics were also used including gender, underrepresented minority status, low-income status, college preparedness, and academic achievement. A full list of features and definitions will be provided in the Model Features section. The total number of student records used in the research is N=179,517. 793 student records were excluded from the dataset because their weighted high school GPA did not fall in the GPA range of 0.1 to 6. Figure 3-1 shows the full list of student characteristics and frequency counts.

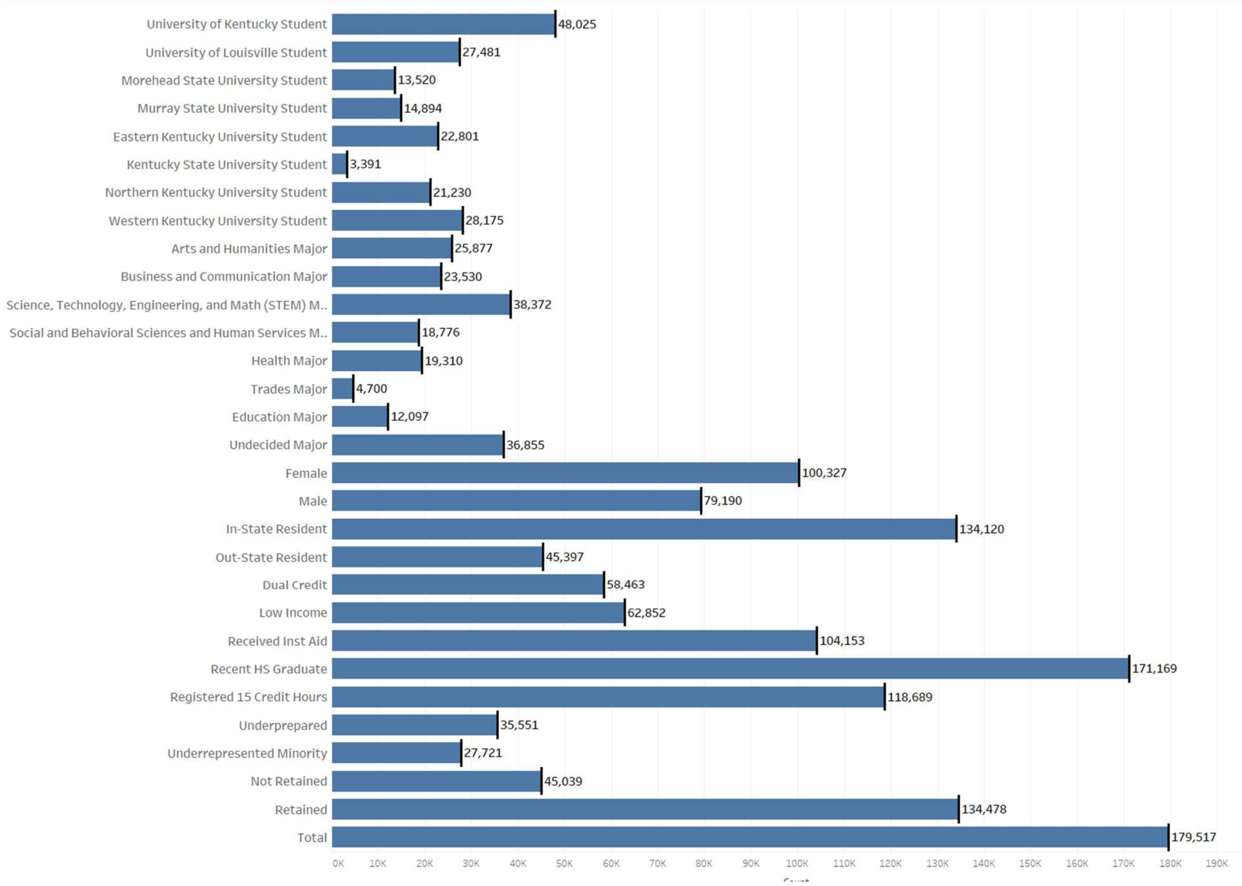


Figure 3-1 *Student Characteristics and Frequency*

3.3 Data Collection Method

All data used in the research was obtained from the Kentucky Postsecondary Education System (KPEDS). KPEDS was created and is maintained by the Kentucky Council on Postsecondary Education (CPE). The CPE is a Kentucky state agency tasked with coordinating the Kentucky higher education system. KPEDS is a vast database that contains millions of student records across many dimensions. The database is populated from multiple collections that occur throughout each academic year. The primary data is pulled from individual institution's student information systems and provided to the CPE based on published reporting guidelines. Specifically, this research used records from the student enrollment table and matched them to student exam, student financial aid, and course completion records to create the research dataset. Direct access to the KPEDS warehouse and custom SQL scripts were used to match the student records.

3.4 Tools

Three primary tools were used in the dataset creation, dataset analysis, model creation, and data visualization processes. The three tools are SQL Server Management Studio, Python through Jupyter Notebook, and Tableau.

3.4.1 SQL Server Management Studio (SMSS)

SQL Server Management Studio is an integrated environment for managing SQL infrastructure ("SQL Server Management Studio", n.d.). With full graphical interface support, the SQL script language is used to access and manipulate database objects that are contained in a relational database. The application was used in this research to create the research dataset, create dummy variables from categorical variables, impute data, and do basic analysis.

3.4.2 Python & Jupyter Notebook

Python is an open-source, object-oriented, high-level coding language that is used for multiple different tasks ("What is Python", n.d.). It is very popular in both application development and data science where it is used primarily in the field of machine learning. It has many libraries that are built specifically for data manipulation and machine learning with many packaged statistical algorithms and functions. Jupyter Notebook is a web-based development environment that is targeted to the data science community because of its design around a scientific notebook ("Project Jupyter", n.d.). The combination of Jupyter Notebook, with a Python kernel, were used to create and test the machine learning models. Four primary Python libraries were used. Pandas was used to create data frames and do basic data analysis, Scikit-learn was used to create and test models, Seaborn was used for visualization, and Pyodbc was used to connect the notebook to the SQL server and move the dataset between the two environments.

3.4.3 Tableau

Tableau is a business intelligence tool that is used for data analytics and visualization. The product is designed to make data more accessible through visualization ("What is Tableau", n.d.). It was used in this research to analyze probability frequencies and visualize research outcomes.

3.5 Model Features

Features, in the context of machine learning, are the independent variables used to train a model to predict an outcome (dependent) variable. In this research, the models will be trained on independent features to predict a binary classification variable.

3.5.1 Feature Selection

Initial features were selected based on literature and variable availability in the database. All features are either basic student characteristics variables or fall into one of Tinto's principal categories (1987). Many of the features were also chosen based on Bogard et al. (2011) where they were shown to have predictive value to student retention. Each feature's relationship to the outcome variable was also checked using their Pearson correlation coefficient and their decision tree feature importance score. Only the top 10 features were chosen for the final models, but all features were used in the initial run.

3.5.2 List of Features

The full list of features, their data types, and descriptions are shown in Table 3-1.

Table 3-1 *Feature List and Descriptions*

Feature	Feature Type	Description
Retention 2nd Fall	Binary (0,1) 1= Not retained 0= Retained	Dependent outcome variable.
Institution	Binary (0,1) 1= Enrolled at institution 0= Not enrolled at institution	Split into 8 dummy variables. One for each institution.
Enrolled Major	Binary (0,1) 1= Enrolled in major 0= Not enrolled in major	Split into 8 dummy variables. One for each major group.
Underrepresented Minority Status	Binary (0,1) 1= Race in URM category 0= Race not in URM category	Determined from a self-reported race variable in the Enrollment table.
Low Income	Binary (0,1) 1= Received Pell Grant first semester. 0= Did not receive Pell Grant first semester.	Determined by the Pell Grant award in the semester of matriculation. Obtained from Student Financial Aid table.
Gender	Binary (0,1) 1= Female 0= Male	Obtained from the Enrollment table.
Residency	Binary (0,1) 1= In-State	Obtained from the Enrollment table.

Feature	Feature Type	Description
	0= Out-of-State	
College Readiness	Binary (0,1) 1= Underprepared 0= Prepared	Obtained from the Student Exam table.
Dual Credit Participation	Binary (0,1) 1= Enrolled in a dual credit course in high school. 0= Did not enroll in a dual credit course in high school.	Calculated field based on student's past enrollment records.
Registered 15 Credit Hours	Binary (0,1) 1= Enrolled in 15 credit hours first semester of entry. 0= Did not enroll in 15 credit hours in the first semester of entry.	Obtained from the Enrollment table.
Institutional Financial Aid	Binary (0,1) 1= Received institutional grant or scholarship. 0= Did not receive institutional grant or scholarship.	Obtained from the Student Financial Aid table.
Recent Graduate	Binary (0,1) 1= Graduated from high school in the same year as postsecondary enrollment. 0= Did not enroll in the same year as high school graduation.	Obtained from the Enrollment table.
ACT Composite Score	Continuous	Obtained from the Student Exam table.
High School GPA	Continuous	Obtained from the Enrollment table.
First Semester GPA	Continuous	Calculated field created from course grades in the first semester.

3.6 Procedure and Data Analysis

3.6.1 Dataset Preparation and Initial Analysis

The research dataset was created using SQL Server Management Studio. Custom SQL scripts were created to transform the records and create calculated fields to act as features. A database table was created to hold the research records and ensure that each student had only one row. Outliers were excluded during the table insert process by including ranges in the "Where" statement. This only affected records with out of range weighted high school GPAs (number of records excluded show in "Population" section). Missing ACT Composite scores that had a SAT Total score (this is common for out-of-state students) were imputed values based on the SAT to ACT concordance table. The concordance values are based on a study to examine relationships between ACT and SAT scores done by the College Board ("ACT/SAT Concordance – Scores", n.d.) Any remaining missing ACT Composite scores were imputed with the mean value. Missing High School GPA records were also imputed with the mean value. The mean value for both were determined using the AVG function in SQL against the entire dataset.

Using the Python coding language within the Jupyter Notebook development environment, the PYODBC library was used to import the research dataset into Jupyter Notebook. PYODBC creates a secure Open Database Connectivity connection to the database management system. Once successfully imported, the dataset was moved to a data frame using the Pandas library. The rows and columns were then counted to make sure they matched the database table and that all data were available. The dataset was then split into two lists. One that contained the features and the other that contained the predicted variable. These were then passed into X and Y data frames. Features were then checked for their relationship to the predicted variable using a heatmap correlation matrix and the decision tree feature importance function. Any features with no relationship were removed from the X data frame. The correlations were created from the correlation function in Pandas and the heatmap created using the heatmap function in Seaborn. Last, using the train, test, split function in the Scikit-learn library, the dataset was split into a 70% training dataset and a 30% testing dataset that was used for cross-validation. The two datasets were created with a random state of 42 (selected every 42nd record) so that the datasets could be replicated between models.

3.6.2 Building and Validating Models

Scikit-learn has many machine learning algorithms as part of its library. In the case of this research, the focus was on algorithms that are able to predict a binary classification variable. The three classification models used were logistic regression, decision tree classifier, and Gaussian naive Bayes. To ensure that the best possible model to predict the outcome variable was used, each model was trained against the X and Y dataset and then the performance compared. Below, is a brief description of how each algorithm is used.

Logistic Regression:

A statistical method that is used to predict binary classes. The algorithm uses the probability of occurrence of a binary event using the logit function (Navlani, 2019). The model makes a prediction based on a threshold. This is set by default at 0.5 or above. It predicts the 1 outcome (in this case that would be not retained), so a probability of 0.5 or above would be predicted as not retained.

Decision Tree Classifier:

A structure that contains decision nodes that represent each feature and outcome. It can be easily visualized with a flowchart diagram (Navlani, 2018). This model is based on decision rules that are inferred from the data features. Decisions are based on Information Gain (IG).

Gaussian Naive Bayes:

Naive Bayes is a statistical classification technique that is based on the Bayes Theorem and is considered one of the simplest supervised machine learning algorithms (Navlani, 2018). It utilizes the concepts of probability and maximum likelihood.

Each model was trained by using the fit function in Scikit-learn against the training split of the dataset (70%). They were then cross-validated against the test split of the dataset (30%). Since the population was so large, there was no need to do other forms of cross-validation.

3.6.3 Model Selection

Each model was evaluated using a confusion matrix, classification outcome report, and Receiver operating characteristic area under the curve (ROC/AUC). All are part of the metrics package in Python. A brief description of the metrics is provided below.

Confusion Matrix:

Model aggregate output of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN).

Accuracy:

(TP+TN)/(TP+TN+FP+FN)

Precision:

TP/(TP+FN)

Recall:

TP/(TP+FN)

Receiver operating characteristic curve (ROC)

Plot of true positive rate against false positive rate at different threshold settings.

The focus was on the model with the highest recall and ROC/AUC. This would represent the model that was best at predicting students who would not return.

3.6.4 Model Tuning

The initial run of the three models gave a good baseline of performance, but machine learning models suffer if the outcome classification is imbalanced. In the case of this research, this was true. The training dataset contained 75% 0 classification (retained) and only 25% 1 classification (not retained). This causes the model to become biased and underpredict the minority classification. To attempt to negate this, two balancing methods were employed and compared on the training dataset (the test dataset was not changed). The two methods are below.

Synthetic Minority Oversampling Technique (SMOTE):

SMOTE increases the minority class by introducing synthetic examples through connecting all k minority class nearest neighbors using feature space similarity (Euclidean distance) (Swamynathan, 2017).

Cost-Sensitive Logistic Regression:

This method allows you to set weights to each classification to provide a "cost" when modifying the coefficients to reduce log loss (Brownlee, 2020).

All three models were re-trained after the training dataset was balanced with the SMOTE method and the default logistic regression model was replaced with a cost-sensitive logistic regression model that included weights that were equal to the proportion of each classification. The models were then re-run against the test dataset and the metrics evaluated against the original models.

3.6.5 Model Implementation

After the best performing model was chosen, the individual probabilities were written back to the test data frame so that each student had a probability of being retained. The predict probability function was used along with Pandas to write the data. Using the PYODBC library, the test dataset was exported back to the SQL Server table along with the predicted probabilities. Tableau, using an ODBC connection, was then connected to the database and the table set as its data source. Tableau was used to analyze the frequencies of the students at each probability threshold along with the distributions and the student's known retention status. Based on this information, risk categories were created based on the distribution to maximize the number of students who were not retained being captured in the medium and high-risk categories. The SQL table was then updated for each student with their risk category. This was done by creating a case statement that updated a new field in the table. The table was then used to report the students in the medium and high-risk categories.

Chapter 4: Results

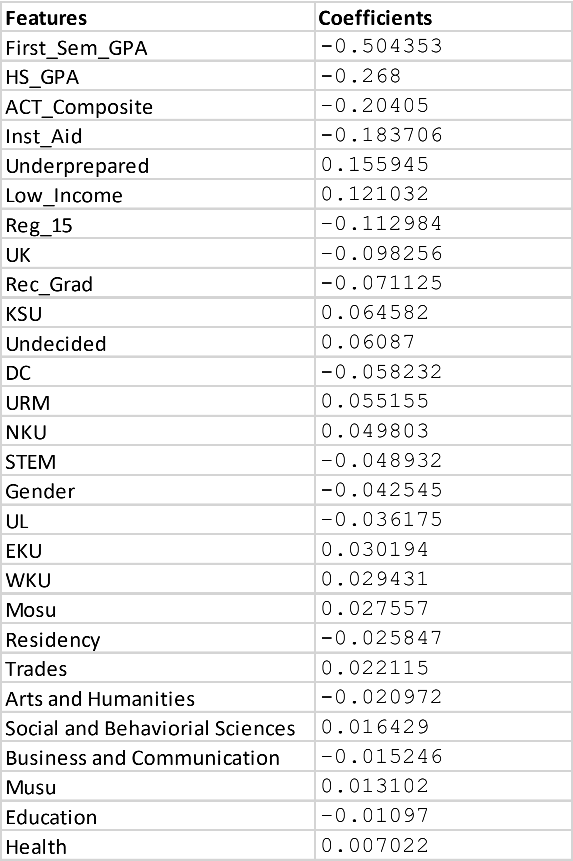
The primary metric used to evaluate model results is the model's ability to recall students who were not retained. This chapter will be split into 7 sections. These sections are Feature Correlation and Selection, Naive Bayes Model Results, Logistic Regression Model Results, Decision Tree Model Results, Model Selection, Probability Frequency and Visualization, and Summary. The Feature Correlation and Selection section will show the Pearson correlation coefficients for all features to the dependent variable and then choose the top 10 with the strongest linear relationship to use in the models. It will also show a correlation heatmap to check for relationships between independent variables. All three model results sections will show the confusion matrix, full metrics table (includes precision, recall and f1 for both the 0 and 1 outcomes), and the ROC AUC score obtained when running the model against the test (unseen) dataset. Results will be shown for all models with the imbalanced outcomes and then after the training dataset has been balanced using the SMOTE method. The logistic regression model will show results for both SMOTE and by using a weighted classification proportional to the dataset distribution. The Model Selection section will compare the three models, both before and after balancing, and choose the model with the best recall of the 1 outcome. Other metrics will also be considered such as overall accuracy and ROC AUC score to ensure the model's predictive ability. The Probability Frequency and Visualization section will show the distribution of the probabilities for all test dataset records and choose thresholds for risk categories. Last, the Summary section will summarize the results and give final recommendations.

4.1 Feature Correlation and Selection

Pearson correlation coefficients were created for all 28 independent variables against the Retained_2 outcome variable. The top 10 independent variables with the strongest (either positive or negative) linear relationship to the outcome variable were chosen. The number of variables used was chosen based on a preliminary check of model accuracy using different numbers of variables. All three models slightly lost accuracy when using both 9 independent variables and 11 independent variables.

4.1.1 Correlation Coefficients All Variables

Figure 4-1 shows a list of all variables and their correlation coefficient to the dependent variable.



Features	Coefficients
First_Sem_GPA	-0.504353
HS_GPA	-0.268
ACT_Composite	-0.20405
Inst_Aid	-0.183706
Underprepared	0.155945
Low_Income	0.121032
Reg_15	-0.112984
UK	-0.098256
Rec_Grad	-0.071125
KSU	0.064582
Undecided	0.06087
DC	-0.058232
URM	0.055155
NKU	0.049803
STEM	-0.048932
Gender	-0.042545
UL	-0.036175
EKU	0.030194
WKU	0.029431
Mosu	0.027557
Residency	-0.025847
Trades	0.022115
Arts and Humanities	-0.020972
Social and Behaviorial Sciences	0.016429
Business and Communication	-0.015246
Musu	0.013102
Education	-0.01097
Health	0.007022

Figure 4-1 *Variable Correlation Coefficients*

Based on the coefficients to the dependent variable First Semester GPA, High School GPA, ACT Composite Score, Institutional Financial Aid, College Readiness, Low Income, Registered 15 Credit Hours, University of Kentucky Enrollment, Recent Graduate, and Kentucky State University Enrollment were chosen for the models. For feature definitions, please see the Methods chapter.

4.1.2 Feature Coefficients

Relationships among the chosen predictor features were also checked using a heatmap of the correlation coefficients. This was done to check for multicollinearity. Figure 4-2 displays the heatmap with coefficients among all features. A darker shade indicates a stronger correlation.

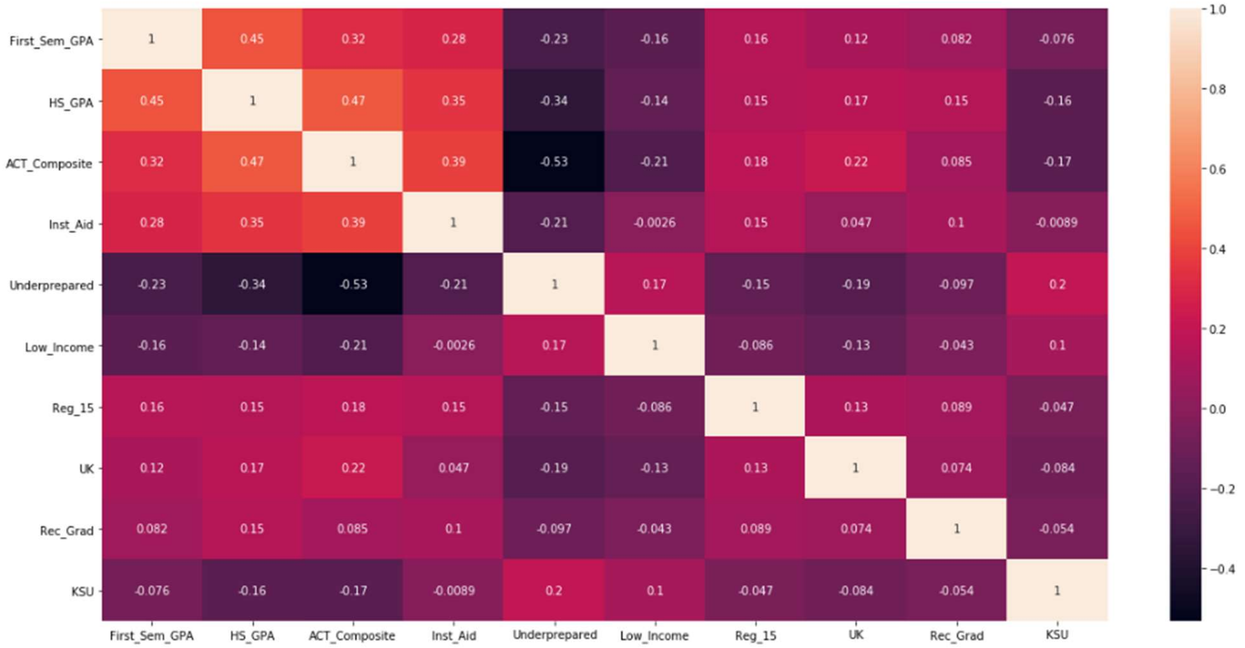
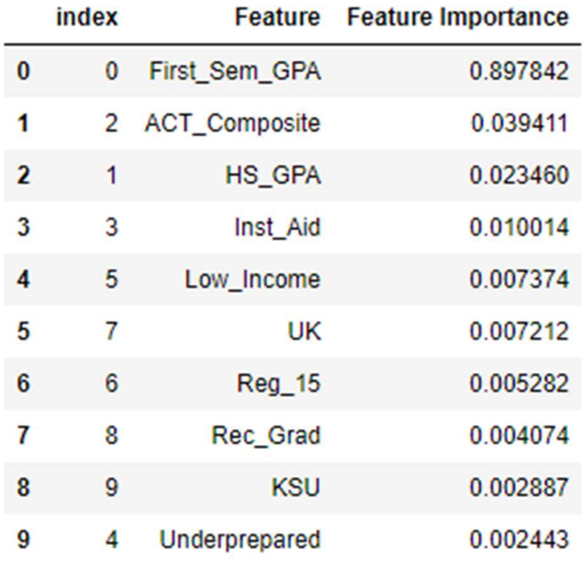


Figure 4-2 *Correlation Coefficients Heatmap All Features*

The features with the strongest relationship were College Readiness and ACT Composite Score at -0.53. Based on the matrix, no features were removed from the models based on multicollinearity.

4.1.3 Decision Tree Feature Importance

Since a decision tree model was also tested, feature importance was checked using the feature importance function for the chosen variables. Figure 4-3 displays the results of the analysis.



	index	Feature	Feature Importance
0	0	First_Sem_GPA	0.897842
1	2	ACT_Composite	0.039411
2	1	HS_GPA	0.023460
3	3	Inst_Aid	0.010014
4	5	Low_Income	0.007374
5	7	UK	0.007212
6	6	Reg_15	0.005282
7	8	Rec_Grad	0.004074
8	9	KSU	0.002887
9	4	Underprepared	0.002443

Figure 4-3 *Decision Tree Feature Importance*

Based on the output, all features provide information gain to the model.

4.2 Naive Bayes Model Results

The training dataset, using the 10 selected features, was used with the default nb_model.predict function in Scikit-learn to fit the model. The model was able to achieve a 77.7% accuracy against the training dataset and maintain a 77.4% when used with the test (unseen) dataset. Table 4-1 shows the full performance results from the model when ran against the test dataset.

4.2.1 Results Imbalanced Dataset

Table 4-1 *Naive Bayes Imbalanced Results*

Confusion Matrix		
	Predicted: Retained	**Predicted: Not Retained**
Actual: Retained	34,637	5,665
Actual: Not Retained	6,474	7,080

Metrics Report			
	Precision	**Recall**	**F1**
Retained	0.84	0.86	0.85
Not Retained	0.56	0.52	0.54

Accuracy	**ROC AUC Score**
0.77	0.69

The results show that the model did much better predicting students who were retained than students who were not. This is possibly caused by the imbalanced training dataset. Using the imbalanced training dataset, the model can only recall 52% of the students who were not retained.

4.2.2 Results Balanced with SMOTE

The training dataset was updated using the SMOTE function from the Imblearn library. This oversampled the minority (not retained) outcome to be proportional with the majority (retained) outcome. This was done to give the model more data points to better predict the minority outcome variable. The model was re-run against the test dataset after the update. The updated results are shown in Table 4-2.

Table 4-2 *Naive Bayes SMOTE Balanced Results*

Confusion Matrix		
	Predicted: Retained	**Predicted: Not Retained**
Actual: Retained	31,043	9,259
Actual: Not Retained	4,870	8,684

Metrics Report			
	Precision	**Recall**	**F1**
Retained	0.86	0.77	0.81
Not Retained	0.48	0.64	0.55

Accuracy	**ROC AUC Score**
0.74	0.71

After training the model on the updated balanced training dataset, the accuracy fell slightly but the model did a much better job of recalling not retained students moving from 52%

to 64%. It also improved the ROC AUC score. The increase in recall appears to be at the cost of an increase in false positives.

4.3 Logistic Regression Model Results

The same process was repeated using the logistic regression model. An accuracy of 82% was achieved with the imbalanced training dataset and the model was able to maintain an 82% accuracy using the test dataset. The model results are displayed in Table 4-3.

4.3.1 Results Imbalanced Dataset

Table 4-3 *Logistic Regression Imbalanced Results*

Confusion Matrix		
	Predicted: Retained	**Predicted: Not Retained**
Actual: Retained	38,552	1,750
Actual: Not Retained	7,824	5,730

Metrics Report			
	Precision	**Recall**	**F1**
Retained	0.83	0.96	0.89
Not Retained	0.77	0.42	0.54

Accuracy	**ROC AUC Score**
0.82	0.69

The model performed better overall than the Naive Bayes model but did a worse job in recalling not retained students. It was able to recall retained students at a very high level (96%) suggesting that the logistic regression model is more sensitive to an imbalanced dataset.

4.3.2 Results Balanced with SMOTE

The model was retrained using the SMOTE training dataset used with the previous model. Table 4-4 shows the results against the test dataset after balancing.

Table 4-4 *Logistic Regression SMOTE Balanced Results*

Confusion Matrix		
	Predicted: Retained	**Predicted: Not Retained**
Actual: Retained	32,014	8,288
Actual: Not Retained	4,549	9,005

Metrics Report			
	Precision	**Recall**	**F1**
Retained	0.88	0.79	0.83
Not Retained	0.52	0.66	0.58

Accuracy	**ROC AUC Score**
0.76	0.73

Balancing the dataset appears to have had the same impact on the logistic regression model as the naive Bayes. The accuracy was dropped from 82% to 76% but the ability of the model to recall not retained students was greatly increased.

4.3.3 Results Classification Balancing

The Logistic Regression model was then refitted using the original training dataset, but also employing classification weights to each outcome. A 75% weight was added to the 1 (not retained) minority outcome and a 25% weight added to the 0 (retained) majority outcome. The weights account for the original proportions of the dataset which were 75% retained and only 25% not retained. The weights were implemented by hyperparameters and the updated results are displayed in Table 4-5.

Table 4-5 *Logistic Regression Classification Balancing Results*

Confusion Matrix		
	Predicted: Retained	**Predicted: Not Retained**
Actual: Retained	31,964	8,338
Actual: Not Retained	4,528	9,026

Metrics Report			
	Precision	**Recall**	**F1**
Retained	0.88	0.79	0.83
Not Retained	0.52	0.67	0.58

Accuracy	ROC AUC Score
0.76	0.73

Similar to the logistic regression model using the SMOTE training dataset, the overall accuracy dropped and the recall of not retained students was improved. Weighting the model with hyperparameters performed slightly better than the SMOTE trained model with a recall of 67% vs 66%.

4.4 Decision Tree Model Results

The decision tree model also did very well at predicting with unseen data scoring the exact same accuracy of 82% on both the training and test datasets. The model with imbalanced outcome data performed worse than all other models only recalling 41% of not retained students on unseen data. This was very similar to the imbalanced logistic regression model. This is shown in Table 4-6.

4.4.1 Results Imbalanced Dataset

Table 4-6 *Decision Tree Imbalanced Results*

Confusion Matrix		
	Predicted: Retained	**Predicted: Not Retained**
Actual: Retained	38,766	1,536
Actual: Not Retained	8,037	5,517

Metrics Report			
	Precision	**Recall**	**F1**
Retained	0.83	0.96	0.89
Not Retained	0.78	0.41	0.54

Accuracy	**ROC AUC Score**
0.82	0.68

4.4.2 Results Balanced with SMOTE

The Decision Tree model was retrained using the SMOTE balanced training dataset. Results are shown in Table 4-7.

Table 4-7 *Decision Tree SMOTE Balanced Results*

Confusion Matrix		
	Predicted: Retained	**Predicted: Not Retained**
Actual: Retained	32,331	7,971
Actual: Not Retained	4,735	8,819

Metrics Report			
	Precision	**Recall**	**F1**
Retained	0.87	0.80	0.84
Not Retained	0.53	0.65	0.58

Accuracy	ROC AUC Score
0.76	0.72

The model performed similarly to all other balanced models once retrained with the SMOTE training dataset. The overall model accuracy fell to 76% but the recall of not retained students increased to 65%.

4.5 Model Selection

Below, is a summary of all metrics reports for all models tested. This was used to make the final determination for model selection. Since the research is focused on the not retained student metrics, the values listed for precision, recall, and F1 are specific to the 1 or not retained outcome. Accuracy and ROC/AUC are for the entire model. The full summary of each model is displayed in Table 4-8.

Table 4-8 *Summary of Model's Metrics*

Summary of Model's Metrics					
Model	**Precision**	**Recall**	**F1**	**Accuracy**	**ROC/AUC**
Naive Bayes (Imbalanced)	0.56	0.52	0.54	0.77	0.69
Naive Bayes (SMOTE)	0.48	0.64	0.55	0.74	0.71
Logistic Regression (Imbalanced)	0.77	0.42	0.54	0.82	0.69
Logistic Regression (SMOTE)	0.52	0.66	0.58	0.76	0.73
Logistic Regression (Class Weight)	0.52	0.67	0.58	0.76	0.73
Decision Tree (Imbalanced)	0.78	0.41	0.54	0.82	0.68
Decision Tree (SMOTE)	0.53	0.65	0.58	0.76	0.72

Analyzing the model results together, the imbalanced logistic regression and decision tree models had the best overall accuracy but did a very poor job recalling not retained students. After balancing, the logistic regression model using the cost-sensitive class weighting technique performed the best in recalling not retained students. It also performed on par, or better, with all other balanced models regarding accuracy and ROC/AUC score. Because of this, the class weighted logistic regression model was chosen for student classification in the next section.

4.6 Probability Frequency and Visualization

To create the probability frequencies, probabilities were created against the test dataset using the weighted logistic regression model and the probability function. The probabilities were then joined back to the original data frame so that each student record in the test dataset contained their retention probability. The data frame was then exported back to SQL Server for analysis. This totaled 53,856 student records (30% of the original population). Tableau was used to analyze and visualize the results.

4.6.1 Probability Frequency Results

In Tableau, probabilities were grouped into 10% increments for easy analysis. A filter was then applied for both not retained and retained populations and the results shown as a percent of the total.

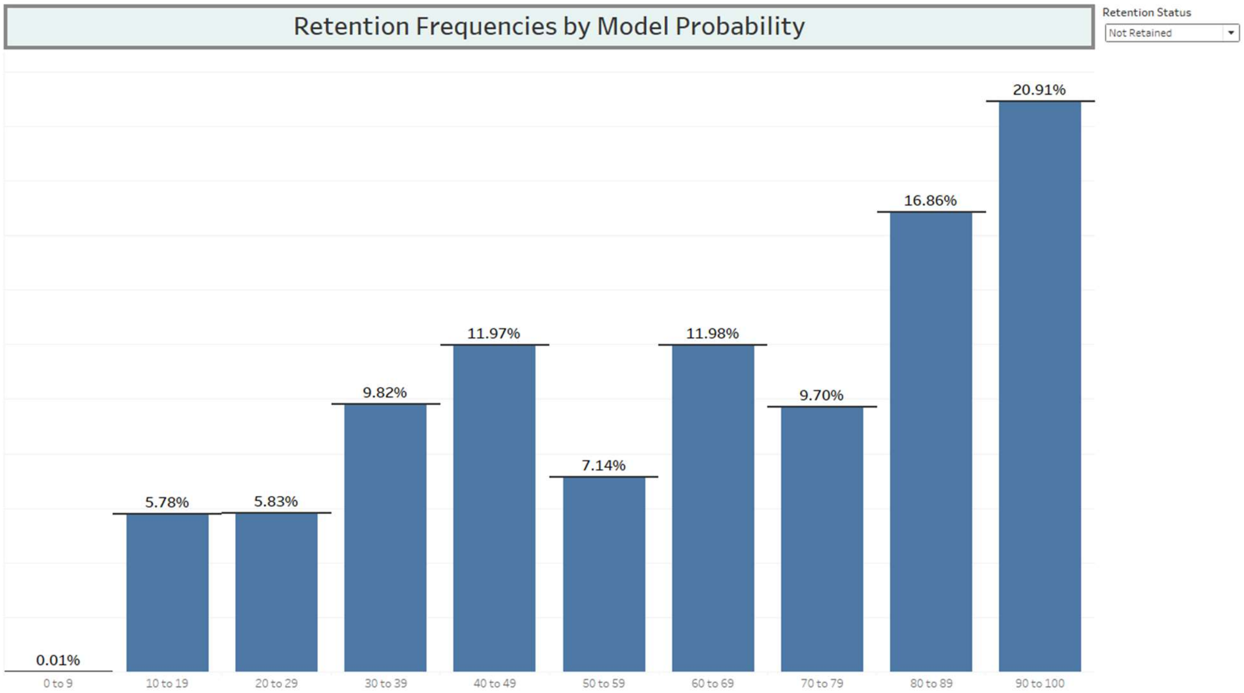


Figure 4-4 *Not Retained Student's Model Probability Frequencies*

The results for the not retained population, as shown in Figure 4-4, suggest that the model did well predicting the overall population but did very well in the upper probability range of 0.8 to 1. Of the students who were not retained in the test dataset, the model predicted almost 38% in this category. This suggests that the model is very accurate when predicting the highest risk students. The model recalls more students as the probabilities increase.

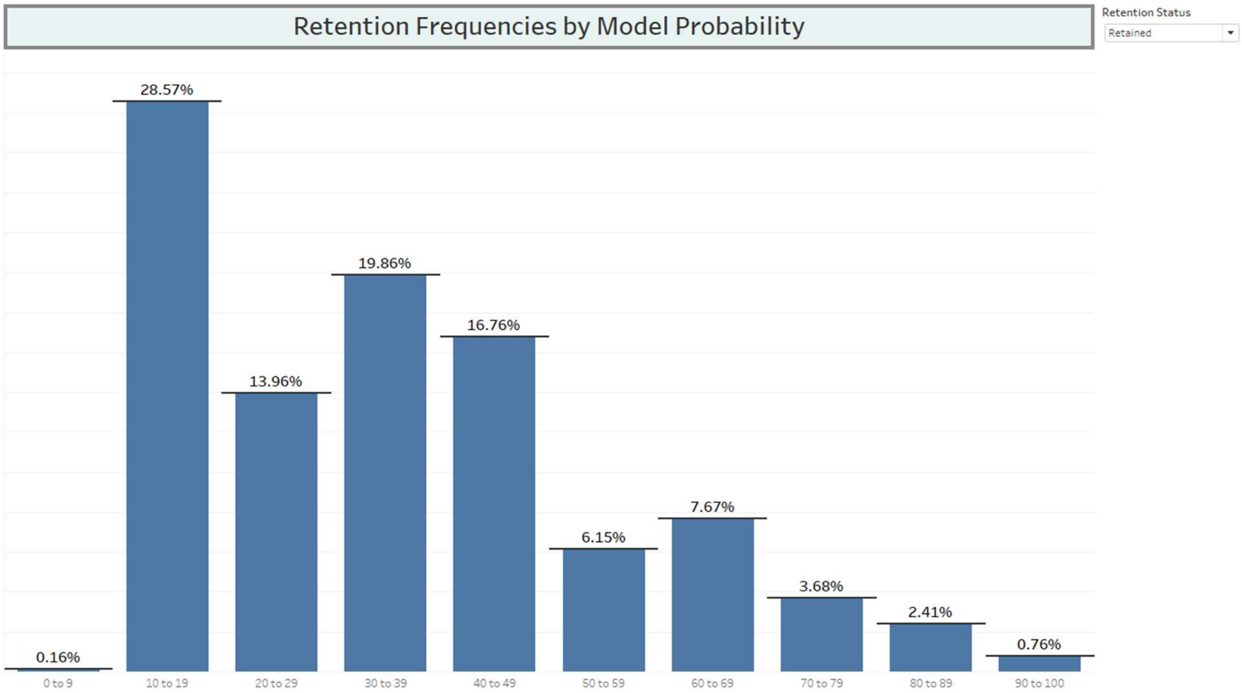


Figure 4-5 *Retained Student's Model Probability Frequencies*

The results from the retained population, as shown in Figure 4-5, suggest an inverse of the distribution of the not retained population. The largest proportion of the retained population were predicted in the 0.1 to 0.2 probability range which suggests the model was able to effectively reduce false-positive predictions. This was also apparent when looking at the highest ranges where only 3.17% of the students who were retained were given a probability of .8 to 1 of not being retained.

4.6.2 Risk Categories

The risk categories were split into four buckets based on the student's risk of not being retained for their following fall semester. These buckets are low-risk, medium-risk, high-risk, and very high-risk. To effectively assign risk categories, both population distributions needed to be considered to reduce false positives while also recalling as many not retained students as possible. With 12% of not retained students within the .4 to .49 probability category, this range was chosen as the start of the medium risk category to increase the proportion of not retained students captured from the model. This effectively moved the prediction threshold from .5 to .4 updating the recall of the population from 67% to 79%. The risk categories and the percentages in each category for retained and not retained students are shown in Table 4-9.

Table 4-9 *Risk Categories by Retention Status*

Risk	Threshold	Not Retained Percent	Retained Percent
Low	0.0 – 0.39	21.44%	62.55%
Medium	0.4 – 0.59	19.11%	22.91%
High	0.6 – 0.79	21.68%	11.35%
Very High	0.8 - 1	37.77%	3.17%

The risk categories were created as a calculated field and inserted into the SQL table so that they could be provided to institutions at the student level.

4.7 Summary

In this chapter, features were tested, different machine learning models compared, and risk categories created. The logistic regression model weighted by the proportion of the two outcome classifications proved to be the best model in recalling students who were not retained on the test dataset. Analyzing the probabilities created from the model against the actual retention results of the students in the dataset, the model appears to do a good job of recalling students at the highest risk of not re-enrolling in the following fall semester of their initial enrollment. The model was most effective in the very high-risk category. The model predicted almost 38% of students who were not retained into this category while only predicting a little over 3% of students who were retained into it. This is only a 3% false-positive rate of retained students and a 38% recall of not retained students. The results also suggest that implementation of the system could result in the effective reduction of non-persisting students even if priority is only given to the students who are in the highest risk category.

Chapter 5: Conclusion

This chapter is split into two main sections. The first section is dedicated to final thoughts and conclusions based on the research results and how they pertain to the research question. The second focuses on future research and how the research can be expanded.

5.1 Research Conclusions

This research started with the problem of Kentucky's low retention rates compared to the national average and their possible effect on the state's workforce. It was also put into the context of a shrinking postsecondary pipeline with the college-going rate slowly declining year after year. The proposed solution was to build a system that could effectively predict student retention, using machine learning methods, that could also be easily operationalized and actionable. Based on the results of the proposed system, the research suggests that this could be done. The model, at the 0.5 decision threshold, was able to successfully recall 67% of those students who were not retained on unseen data after balancing the training dataset. This was further increased once probability frequencies were analyzed against retention results at which time 79% of students who were not retained were categorized into medium, high, and very high-risk categories. The most impressive result from the model was its ability to predict very high-risk students at the 0.8 probability and above. With a recall of 38% of not retained students and only a 3% false-positive rate for retained students, this could give academic institutions the information they need to successfully employ data driven retention strategies. It also allows institutions to focus on the students most at-risk and not waste limited resources on students who do not need assistance, showing why the 3% false-positive rate at this category is very important. The research suggests that the system proposed would act as an effective tool for reducing student attrition and reduce the student success gap that exists between Kentucky public 4-year academic institutions and national public 4-year academic institutions. The system also appears to be actionable and scalable to the entire academic postsecondary system.

5.2 Future Research

Although this research provided insight into the ability of machine learning models to inform retention strategies, there is still room to improve both the model and the process. Machine learning methods are constantly improving and are underutilized in the academic sector; therefore, much research opportunity remains in the space. This research should set a foundation and could be improved in four ways. This is through enhanced feature engineering, expansion of feature availability, system process improvement, and 2-year academic institution inclusion.

5.2.1 Feature Engineering with Unsupervised Models

This research used only supervised machine learning models to predict a classification of an outcome variable. Supervised models can be enhanced by using unsupervised machine learning models to engineer features that can then be used in model training. Unsupervised machine learning models focus on clustering algorithms to find hidden relationships among categories. This can create new categorical variables with high correlation to the outcome variable. For example, and the use case most prudent to this research, student enrollment patterns could be analyzed to find course enrollment clusters that have predictive value to retention or attrition. Those categories could then be engineered as features in the supervised model. The result would hopefully be improved model performance.

5.2.2 Feature Availability

The biggest limitations of this research were the timely availability of important model features and the exclusive use of postsecondary data. This limits the time available to both provide the information to institutions and to implement retention strategies. The actionability and performance of the system could be improved by including data elements from the student's high school career. If important features were identified in the high school data, the system could be implemented at the time of initial enrollment and not be bottlenecked by postsecondary data collection. The second phase of this research should focus on the participation of the Kentucky Department of Education to include feature selection and implementation from high school data.

5.2.3 System Process Improvements

The current system was created using many manual methods of data extraction, analysis, and movement between information systems. This could be greatly automated going forward. A stored procedure can be created with embedded Python syntax to automate feature selection, model selection, and risk category creation based on predetermined criteria from this research. This could all be triggered by data collection criteria being met within SQL Server. The reporting could also be automated using a combination of SQL Server Reporting Services and automatic data extraction updates available in Tableau Server. This would greatly reduce the hours and resources necessary to implement and scale the system in the future.

5.2.4 2-Year Academic Institution Inclusion

The current research only includes students from 4-year public institutions. Expansion of this research should include the prediction of retention of students who attend 2-year public institutions as well. Community and Technical Colleges make up 38% of public student enrollment in the state of Kentucky (Kentucky Council on Postsecondary Education, 2019) so the inclusion of 2-year public students into future research is vital to solving the state's postsecondary retention problem. This becomes even more apparent when looking at retention rates of the 2-year public sector which stands at only around 50%.

References

ACT/SAT Concordance - Scores - The ACT Test. (n.d.). Retrieved August 30, 2020, from https://www.act.org/content/act/en/products-and-services/the-act/scores/act-sat-concordance.html

Allen, J., & Robbins, S. (2008). Prediction of College Major Persistence Based on Vocational Interests, Academic Preparation, and First-Year Academic Performance. Research in Higher Education, 49(1), 62-79. Retrieved August 30, 2020, from http://www.jstor.org.msu.idm.oclc.org/stable/25704545

Astin, A. W. (1975). Preventing students from dropping out. San Francisco: Jossey-Bass

Astin, A. W. (1984). Student involvement: A developmental theory for higher education. Journal of College Student Personnel, 25(4), 297–308

Bogard, M., Helbig, T., Huff, G., & James, C (2011). A comparison of empirical models for predicting student retention. White paper. Office of Institutional Research, Western Kentucky University.

Brownlee, J. (2020, August 27). Cost-Sensitive Logistic Regression for Imbalanced Classification. Retrieved August 30, 2020, from https://machinelearningmastery.com/cost-sensitive-logistic-regression

Garrett, N., Bridgewater, M., & Feinstein, B. (2017). HOW STUDENT PERFORMANCE IN FIRST-YEAR COMPOSITION PREDICTS RETENTION AND OVERALL STUDENT SUCCESS. In RUECKER T., SHEPHERD D., ESTREM H., & BRUNK-CHAVEZ B. (Eds.), Retention, Persistence, and Writing Programs (pp. 93-113). Boulder, Colorado: University Press of Colorado. Retrieved August 30, 2020, from http://www.jstor.org.msu.idm.oclc.org/stable/j.ctt1kt830p.8

Giani, M., Alexander, C., & Reyes, P. (2014). Exploring Variation in the Impact of Dual-Credit Coursework on Postsecondary Outcomes: A Quasi-Experimental Analysis of Texas Students. The High School Journal, 97(4), 200-218. Retrieved August 30, 2020, from http://www.jstor.org.msu.idm.oclc.org/stable/43281031

Hu, S., & St. John, E. (2001). Student Persistence in a Public Higher Education System: Understanding Racial and Ethnic Differences. The Journal of Higher Education, 72(3), 265-286. doi:10.2307/2649332

Kentucky Council on Postsecondary Education. (2019). Student Enrollment [Dashboard]. Retrieved from https://reports.ky.gov/t/CPE/views/KentuckyPostsecondaryEducationInteractiveDataDashboard/Enrollment?%3AshowAppBanner=false&%3Adisplay_count=n&%3AshowVizHome=n&%3Aorigin=viz_share_link&%3AisGuestRedirectFromVizportal=y&%3Aembed=y

Kentucky Center for Statistics. (2019). *High School Feedback Report.* https://kystats.ky.gov/Latest/HSFR

Kentucky Council on Postsecondary Education. (2020). *Higher Education's Return on Investment: The Case for Why Higher Education Matters.* http://cpe.ky.gov/data/reports/ROIreport.pdf

Leppel, K. (2001). The Impact of Major on College Persistence among Freshmen. Higher Education, 41(3), 327-342. Retrieved August 30, 2020, from http://www.jstor.org.msu.idm.oclc.org/stable/3447979

Navlani, A. (2018, December 4). Naive Bayes Classification using Scikit-learn. Retrieved August 30, 2020, from https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn

Navlani, A. (2019, December 16). Understanding Logistic REGRESSION in PYTHON. Retrieved August 30, 2020, from https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python

Navlani, A. (2018, December 4). Naive Bayes Classification using Scikit-learn. Retrieved August 30, 2020, from https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn

National Center for Education Statistics. (2020). *Undergraduate Retention and Graduation Rates.* National Center for Education Statistics, U.S. Department of Education. https://nces.ed.gov/programs/coe/indicator_ctr.asp

Porchea, S., Allen, J., Robbins, S., & Phelps, R. (2010). Predictors of Long-Term Enrollment and Degree Outcomes for Community College Students: Integrating Academic, Psychosocial, Socio-demographic, and Situational Factors. The Journal of Higher Education, 81(6), 680-708. Retrieved August 30, 2020, from http://www.jstor.org.msu.idm.oclc.org/stable/40929572

Project Jupyter. (n.d.). Retrieved August 30, 2020, from https://jupyter.org/

Scott, M. (2017). BIG DATA AND WRITING PROGRAM RETENTION ASSESSMENT: What We Need to Know. In RUECKER T., SHEPHERD D., ESTREM H., & BRUNK-CHAVEZ B. (Eds.), Retention, Persistence, and Writing Programs (pp. 56-73). Boulder, Colorado: University Press of Colorado. Retrieved August 30, 2020, from http://www.jstor.org.msu.idm.oclc.org/stable/j.ctt1kt830p.6

SQL Server Management Studio (SSMS) - SQL Server Management Studio (SSMS). (n.d.). Retrieved August 30, 2020, from https://docs.microsoft.com/en-us/sql/ssms/sql-server-management-studio-ssms?view=sql-server-ver15

Swamynathan, M. (2017). Model Diagnosis and Tuning. In *Mastering machine learning with Python in six steps: A practical implementation guide to predictive data analytics using Python* (p. 214). Berkeley, CA: Apress. doi:10.1007/978-1-4842-2866-1

Tinto, V. (1987). “The Principles of Effective Retention.” ERIC. from eric.ed.gov/?id=ED301267.

Tinto, V. (1993). Leaving college: rethinking the causes and cures of student attrition. Chicago; London: University of Chicago Press.

Tinto, V. (1999). Taking retention seriously: Rethinking the first year of college. NACADA Journal, 19(2), 5-9.

Vanderplas, J. T. (2017). Machine Learning. In *Python data science handbook: Essential tools for working with data* (p. 332). Sebastopol, CA: O'Reilly.

What is Python? Executive Summary. (n.d.). Retrieved August 30, 2020, from https://www.python.org/doc/essays/blurb/

What is Python? Executive Summary. (n.d.). Retrieved August 30, 2020, from https://www.python.org/doc/essays/blurb/