OPIM 5770: ADVANCED BUSINESS ANALYTICS AND PROJECT MANAGEMENT

FALL 2023

TRACKING STUDENT JOURNEY:

ANALYZING RETENTION AND GRADUATION RATES

STUDENT: VENKATA SIVA LINGA SAI KRISHNA BONDALAPATI

TABLE OF CONTENTS

I. EXECUTIVE SUMMARY:

The University of Connecticut is facing challenges with student retention and graduation rates and sought out assistance from the students enrolled in the Master of Business Analytics and Project Management (MSBAPM) program. In the Fall of 2023, UConn's Department of Budgeting, Planning, and Institutional Research (BPIR) presented this problem to UConn's Advanced Business Analytics and Project Management course students for their capstone. They provided several datasets in hopes that students could analyze and identify trends within them, and receive recommendations on how to improve. Group 4 decided to focus on retention rates.

The initial project phase involved meticulous data cleaning, consisting of fixing data types of columns, editing data entries to ensure data is consistent, addressing missing values, and dropping the unnecessary columns for further analysis. Following this, each dataset underwent merging based on a common variable, specifically the dummyID representing student ID numbers, ensuring the anonymized association of data with individual students.

Subsequently, the best practice of exploratory data analysis (EDA) was employed to gain insights into the dataset. This involved exploring summary statistics like mean, median, mode, interquartile ranges, and value counts for unique column values. Basic data visualizations were incorporated during this phase to present the data's shape, preparing it effectively for subsequent modeling.

For modeling, the target variable used was retention_rate_2_0, a market to indicate whether a student was enrolled in their fourth semester. Data was split into Training, 80%, and Testing, 20%. Seven models were evaluated but only four were viable and remained the main focus. The best model of the four was Gradient Boosting (XGBoost). Accuracy, Precision, and F1-score were used to evaluate the performance metrics of the models. From there, some feature importance was done to show which columns in the dataset had the best predictive power for predicting retention rates. Microsoft's PowerBI was then used to create interactive dashboards for the dataset. These dashboards were made to present research findings to the project sponsors (BPIR).

Some of the business recommendations are:

- Strategic recommendations for improving retention rates through student engagement in clubs, activities etc.
- Addressing retention challenges through financial support solutions.

- Guiding strategies for students encountering academic challenges within their respective departments.

II. PROBLEM STATEMENT:

UConn is a R1 public research university, making it a highly competitive, prestigious institution with a global reputation for quality education and academic success. The reputation of all universities hinges on student enrollment, graduation, and retention. UConn's BPIR Department says the university's undergraduate program needs improvement in their retention and graduation rates because they are crucial indicators of institutional effectiveness and student success in higher education. If lower retention and graduation rates persist, there's a potential impact on the university's reputation and national rankings. This is a cause of concern for current students, as it could influence the value of tuition for them, but beyond them, it may affect the university's appeal to prospective enrollees. Addressing these challenges is crucial for upholding the university's standing and ensuring students receive the full value of their tuition.

III. METHODOLOGY:

APPROACH:

The primary focus of Group 4's project centers around the analysis of student retention. The idea is to seamlessly integrate the SEMMA (Sample, Explore, Modify, Model, and Assess) process into this complex project. This structured data mining approach has proven highly effective in guiding data analysis, from initial data sampling and exploration through data

modification, modeling, and final assessment. It has played a crucial role in enhancing the quality and depth of the project's data analysis and decision-making processes.

There are six datasets where each dataset had redundant columns, which brought more complexity in identifying and cleaning the columns. The best approach followed was dividing each dataset into three categories i.e. categorical data which contained categorical columns, numerical data which contained numeric columns, and float data which contained decimal value columns. All Data Wrangling techniques were implemented, such as missing values counts, similar columns, and similar distribution which can affect the modeling analysis later on. After removing the insignificant columns from the retention rate dataset, a combining of pell grant columns was done from the first to fourth semesters, into a single Pell Grant column to reduce the duplication of the data. The Qsigns and LC enrollment dataset had few records which can cause a potential issue of merging this dataset to retention dataset can lead to a loss of 75% of the original dataset. Because of this issue the datasets are excluded from further analysis.

EDA & Assumptions:

Upon exploring the newly cleaned dataset, research was done on domain-specific online publications and reference articles, and there was a clear domination of semester GPA affecting retention rate. It was later found to heavily dominate predictive power, and was subsequently dropped. Based on details from the dataset, it was also assumed that Pell Grant can be a potential parameter to identify retention rate trends. Initial Data Exploration also provides strong support for this assumption in terms of identifying the difference, finding that about 64% students who are retained have received Pell Grants while 36% students are not retained. It is expected that financial needs will play a significant role later in the data modeling.

DATA CLEANING:

After the conclusion of the Fall 2023 semester, it is recommended to exclude retention data from the analysis since, at that point, past students will have already graduated and the journey of the incoming cohort cannot be tracked, rendering this information unnecessary for consideration.

Class-Taking Data:

The class-taking dataset, containing an extensive 991,849 entries and 23 features, encounters redundancy challenges due to there being 51,595 unique dummy IDs, each averaging 19 records. To mitigate this issue during dataset merging, a strategic pivot operation based on dummy IDs was implemented. This restructuring involved populating fields with counts, offering a concise representation that effectively resolves redundancy. This approach not only simplifies the dataset but also captures instances where the same dummy ID has identical grades across various subjects, enhancing the quality of subsequent analyses.

Sixty-Second Survey Data:

In the sixty-second survey data, the presence of numerous duplicate dummy IDs prompted the use of the groupby function. This function facilitated the grouping of data based on dummy IDs, followed by the calculation of the mean for each dummy ID's values across all columns. The result is a dataset where each dummy ID is associated with a singular, averaged set of values, eliminating redundancy. This streamlined dataset ensures that each dummy ID is represented uniquely, contributing to a more efficient and representative dataset for further analysis.

Q_Center Sign-ins Dataset:

The Q_Center sign-ins dataset, housing 6,175 unique IDs, presented a distinct challenge. Merging this dataset with the retention dataset was considered impractical,

foreseeing a substantial 75% loss of data. Prioritizing data integrity, the decision was made to exclude the Q_Center sign-ins dataset from the analysis. While this choice narrows the focus, it safeguards the reliability and representativeness of the retained data, ensuring that the analysis remains robust without compromising overall dataset quality.

Data Merging:

The model uses four datasets: retention dataset, GPA dataset, class taking dataset, and 60 seconds survey dataset. The datasets were merged into a single file using an inner join on dummyID. Then, the data is split into 80% training and 20% testing datasets.

In simplifying the visualization process, the data for Pell Grants was combined across the first four semesters. If a row had a value of 1, it means the student received Pell Grants for all four semesters. An average of 0.75 suggests Pell Grants for three semesters, 0.50 indicates two semesters, 0.25 means one semester, and 0 means no Pell Grants throughout the four semesters. This method helps in understanding how students benefited from Pell Grants during their time at the university.

FINAL DATASET:

Column Name	**Description**	**Importance**
gender	**Gender of the student (1 for male, 0 for female)**	**Demographic information**
pell_grant	**Indicates if the student receives a Pell Grant (1 for yes, 0 for no)**	**Financial aid status**
Level	**The conversion of categorical level names, including col_Freshman,**	**Academic year information**

	col_Sophomore , col_Junior , col_Senior into binary numerical representations, with '0' and '1' values,	
residency	**The conversion of categorical residency names, including col_connecticut, col_nonresident alien, col_out_of_state into binary numerical representations, with '0' and '1' values,**	**Geographic information**
entry_campus	**The conversion of categorical entry_campus names, including 'col_AVYPT,' 'col_HRTFD,' 'col_STMFD,' and 'col_STORR,' into binary numerical representations, with '0' and '1' values,**	**Academic program information**
entry_school	**The conversion of categorical entry_school names, including col_Engineering, col_Fine Arts, col_Liberal Arts & Sciences, into binary numerical representations, with '0' and '1' values,**	**Academic program information**
Semester_GPA	**Grade Point Average for the current semester**	**Academic performance in the current semester**
Cumulative_GPA	**Cumulative Grade Point Average for all semesters**	**Overall academic performance**
Semester_Enrolled_Credits	**Number of credits the student is enrolled in for the current semester**	**Academic workload for the semester**
Semester_Passed_Credits	**Number of credits the student has passed in the current semester**	**Academic progress for the semester**
Total_Enrolled_Credits	**Total number of credits the student is enrolled in**	**Overall academic workload**
Total_Passed_Credits	**Total number of credits the student has passed**	**Overall academic progress**

ACADEMIC_STRESS	**Self-reported academic stress level**	**Student's perceived stress level**
BELONGINGNESS	**Self-reported sense of belongingness**	**Student's sense of belonging to the university**
CLASS_INTEREST	**Self-reported interest in classes**	**Student's engagement in coursework**
CLASS_PARTICIPATION	**Self-reported class participation level**	**Student's active participation in classes**
FACULTY_ENGAGEMENT	**Self-reported faculty engagement level**	**Student's interaction with professors**
FINANCIAL_STRESS	**Self-reported financial stress level**	**Student's perceived financial stress level**
FOCUS	**Self-reported ability to focus on tasks**	**Student's ability to concentrate on tasks**
GRIT	**Self-reported level of grit**	**Student's determination and perseverance**
JOB	**Self-reported job status**	**Student's employment status**
MEMORIZATION	**Self-reported memorization ability**	**Student's ability to memorize information**
NOTE_TAKING	**Self-reported note-taking ability**	**Student's note-taking skills**
SATISFY_REQUIREMENTS	**Self-reported satisfaction with program requirements**	**Student's satisfaction with academic program requirements**
STUDENT_ORG	**Self-reported involvement in student organizations**	**Student's engagement in extracurricular activities**
SUPPORT_NETWORK	**Self-reported strength of the support network**	**Student's perception of available support network**

TIME_MANAGEMENT	**Self-reported time management skills**	**Student's ability to manage time effectively**
SIGN_INS_Q	**Number of times the student has signed into a specific location**	**Tracking student engagement with specific services**
retention_2_0	**Indicates if the student was retained within the first year (1 for yes, 0 for no)**	**Retention status within the university**

IV. RESULTS:

MODELING:

When analyzing student retention rates, a total of seven models were used to analyze the data. The Logistic Regression model achieved an accuracy of 70%, exhibiting higher precision (0.70) and recall (0.88) for retained students (class 1) compared to non-retained students (class 0) with precision and recall values of 0.68 and 0.41, respectively. The confusion matrix showed a trade-off between correctly identifying retained and non-retained students. The Random Forest model surpassed Logistic Regression with an accuracy of 86%, demonstrating balanced precision and recall for both classes, and achieved higher values for sensitivity (0.87) and specificity (0.84). In the context of student retention analysis, sensitivity refers to the model's ability to correctly identify students who are at risk of not being retained, indicating its effectiveness in capturing those instances. On the other hand, specificity measures the model's capability to accurately identify students who are likely to be retained. The confusion matrix showed the model's proficiency in correctly classifying retained and non-retained students. Gradient Boosting (XGBoost) showed really good performance at 85% accuracy. It maintained balanced precision and recall for both retained and non-retained students, with sensitivity and specificity falling between those of logistic regression and random forest.

Alternatively, the Naive Bayes (BernoulliNB) model showed lower overall performance with a 66% accuracy. While it showed higher sensitivity for retained students (0.74), it suffered from lower specificity for non-retained students (0.53). The K-Nearest Neighbors (KNN) model got 68% accuracy, displaying an imbalance in sensitivity (0.91) and specificity (0.33), showing a much stronger ability to correctly identify retained students.

Decision Trees achieved 77% accuracy, with balanced precision and recall for both retained and non-retained students. The confusion matrix underscored its effectiveness in

correctly classifying instances from both classes. Ensemble methods (AdaBoost) showed similar performance to Decision Trees with 77% accuracy and balanced precision and recall for both student groups, striking a middle ground between the individual models' performance.

In summary, Random Forest and Gradient Boosting demonstrated superior overall performance in predicting student retention, while logistic regression, decision trees, and ensemble methods (AdaBoost) showed competitive results. Naive Bayes and K-Nearest Neighbors exhibited lower accuracy and imbalances in sensitivity and specificity.

FINAL MODEL CONTRIBUTIONS:

The cumulative impact of the top 10 feature importance columns in the Gradient Boosting model is substantial, collectively contributing to almost 79% of the predictive power (see Figure 6 in the Appendix). Earlier assumptions about GPA's weight on predictive power was found to be correct, so to ensure a better exploration of factors influencing student retention beyond the expected influence of GPA or grades, columns related to academic performance were deliberately excluded.

Notably, Pell Grant emerges as the most influential feature, holding the highest position and contributing significantly at 26.60%. This underscores the pivotal role of Pell Grant status in predicting student retention outcomes. This really drives home how having or not having a Pell Grant can strongly influence whether a student decides to stay in school. Pell Grants often reflect financial need and show just how much finances can shape a student's decision to keep going with their education. This insight showed the important connection between financial support and student retention.

Following closely, Student Organization represents 12.74% of the overall contribution, emphasizing the noteworthy impact of extracurricular involvement on predictions. Additionally, Academic Confidence emerges as a key factor with a contribution

of 6.01%, signifying the impact of self-assurance on predicted outcomes. This strategic approach allows for a more comprehensive understanding of the multifaceted factors affecting student retention beyond the conventional influence of academic grades.

VISUAL ANALYSIS:

In the data visualization, attention shifts to areas with the potential for improvement, particularly among students not receiving Pell Grants, who currently experience a higher percentage of non-retention (see Figure 1 in the Appendix). For every student not receiving Pell Grants for up to 4 semesters, there corresponds a retention rate of 2 students. Conversely, students receiving Pell Grants for up to 3 semesters exhibit an attrition of 2.5 students who are not retained. Notably, students receiving Pell Grants for up to 4 semesters showcase a 100% retention rate, indicating a successful retention outcome for all individuals in this category.

Directing attention to student organization involvement, it becomes apparent that students not engaged in clubs represent an area with the potential to enhance retention rates, given the observed higher non-retention rate in this group (see Figure 2 in the Appendix). Students actively participating in clubs demonstrate a higher retention rate of 64%, compared to the 36% retention rate for those not involved in clubs. Even a minimal investment of time in clubs contributes to an increased retention rate compared to non-participation.

Exploring retention rates across university departments reveals a combined retention rate of 17% for the Engineering and Business programs, signaling a notable decline in the number of students from these departments. Prioritizing retention efforts on departments such as ACES, Center for Excellence in Teaching & Learning, Ratcliffe Hicks, Social Work, and Education is crucial, as they exhibit nearly a 100% non-retention rate. Notably, the largest

department, Liberal Arts & Sciences, plays a pivotal role in student retention, offering an opportunity to elevate the overall retention rate.

V. CONCLUSIONS AND RECOMMENDATIONS:

In the face of the persistent challenge of student retention, targeted financial support measures must be implemented in order to effectively address the ongoing problem of student attrition, according to a thorough analysis. Even with Pell subsidies available, a sizable fraction—that is, 36 percent of students—still face retention challenges. The suggestion is made to include part-time jobs on campus in order to bolster current retention tactics. This proactive action aims to foster an environment that supports academic success in addition to providing students with the necessary financial help.

After the first round of Pell Grants is awarded, institutions should carefully shift their attention to long-term support systems. It is advised that institutional efforts be focused on providing all-encompassing support, which includes services like financial counseling and academic guidance. This all-encompassing method is purposefully designed to mentor students at every stage of their academic career.

An additional tactic to increase retention rates is proactive involvement via student clubs and organizations. Universities may create a dynamic and welcoming community by broadening their club offerings to meet the many interests of their student population. Financial support for these clubs is not just an investment in encouraging involvement but also a calculated move that improves the overall experience for students and has a favorable correlation with retention rates. Moreover, rewarding and recognizing students for their significant participation in these extracurricular activities serves to reinforce the connection between student engagement and academic perseverance.

The University of Connecticut is recommended to give priority support to programs that are in high demand, such as business and engineering, when resolving academic issues within certain departments. The unique curriculum and encouraging placement rates linked to these programs are highlighted as part of this prioritization. It is also proposed that a good approach would be for each academic program to have successful alumni lead podcasts or workshops. The MSBAPM students can attest to this being a successful effort because former students in the program have led online workshops to showcase real-world use of industry tools like Tableau or Power BI. As students navigate the complexities of their chosen subjects, this investment not only provides insightful instruction but also fosters a sense of community and direction.

Additionally, it is considered necessary to focus specifically on departments that have high turnover rates, such as ACES, the Centre for Excellence in Teaching & Learning, Radcliffe Hicks, Social Work, and Education. The proposal to provide Pell Grants or other financial aid to students who choose to enroll in these departments is a focused intervention meant to alleviate financial strains. It is expected that these strategies will strengthen retention rates, and greatly enhance the general success of University of Connecticut students.

VI. REFERENCES:

In search of research regarding student retention rates, insights were drawn from various reputable sources. The study by Coleman (2022) highlighted the profound impact of mental health on student retention, shedding light on the importance of addressing psychological well-being in the overall retention strategy. This perspective aligns with the acknowledgment in the 4 Ps of Student Retention Framework by Kalsbeek et al., emphasizing the need to consider the broader factors influencing student persistence, including psychological aspects.

The analysis was further enriched by the Annual Reports on Retention and Graduation at the University of Connecticut, authored by Fuerst in both 2021/2022 and 2022/2023. These reports provided valuable institutional data and trends, allowing to contextualize the findings within the university's specific landscape. Understanding the local context is crucial for developing effective retention strategies, and these reports served as key references in this regard.

The work by Nieuwoudt (2022) and the related Sage Journals publication delved into the motivations behind students choosing to remain at university. This provided a qualitative dimension to the analysis, complementing the quantitative data that was worked with. Understanding the reasons students choose to stay can inform targeted interventions to enhance retention.

Additionally, Wiley University Services (2023) offered valuable strategies for student retention. By incorporating these strategies into the analysis, an aim to align the findings with practical, evidence-based approaches to improving retention rates, was made. The incorporation of external research findings into the project not only enriched the understanding but also allowed for the validation and contextualization of the analytical approach in the broader landscape of student retention studies.

Sources:

Coleman, M. (2022, April 19). HOW MENTAL HEALTH IS IMPACTING STUDENT RETENTION. Retrieved from The National Society of Leadership and Success: https://www.nsls.org/

David H. Kalsbeek, C. M. (4 Ps of Student Retention Framework). Improving Outcomes through the. DePaul University.

Fuerst, N. (2022). 2021/2022 Annual Report on Retention and Graduation. University of Connecticut.

Fuerst, N. (2023). 2022/2023 Annual Report on Retention and Graduation. University of Connecticut.

Nietzel, M. T. (2022, Dec 13). New Report: College Completion Rates Improve, But Disparities Remain A Problem. Retrieved from Forbes: https://www.forbes.com/sites/michaeltnietzel/2022/12/13/new-report-college-completion-rates-improve-but-disparities-remain-a-problem/?sh=a59246e59e55

Nieuwoudt, J. E. (n.d.). https://journals.sagepub.com/doi/10.1177/1521025120985228.

Nieuwoudt, J. E. (n.d.). Sage Journals Home. Retrieved from Student Retention in Higher Education: Why Students Choose to Remain at University: https://journals.sagepub.com/doi/10.1177/1521025120985228

University, W. (2023, May 29). Student Retention Strategies. Retrieved from Wiley University Services: https://universityservices.wiley.com/student-retention-strategies/

VII. APPENDIX:

Figure 1 : Retention Rate Based on Pell Grant

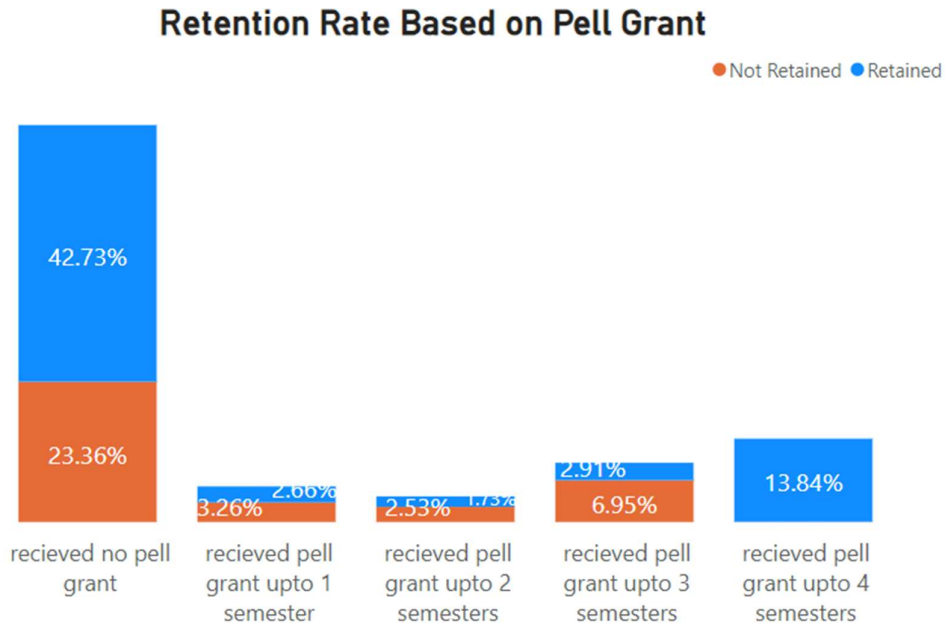

Figure 2 : Retention Rates of Student Participation in Clubs

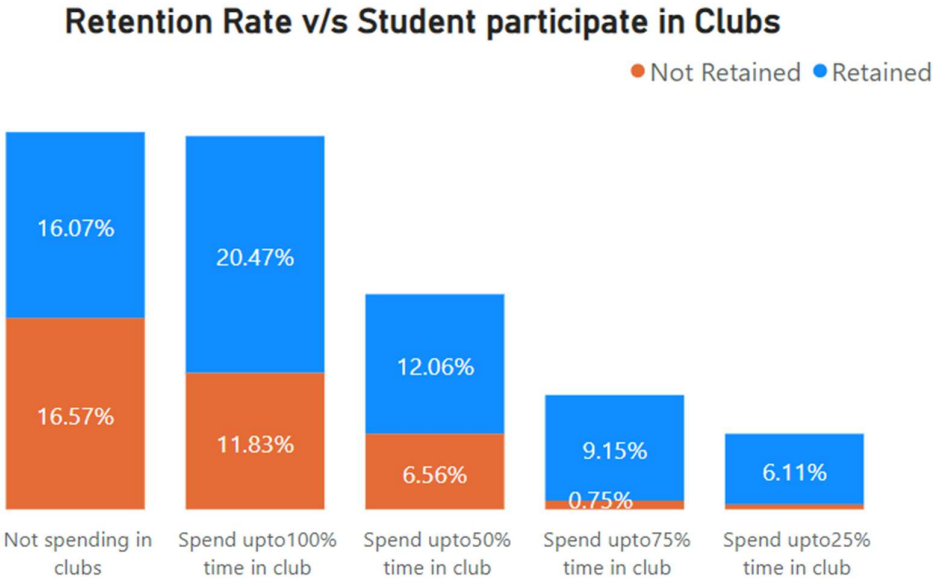

Figure 3 : Retention of University Departments

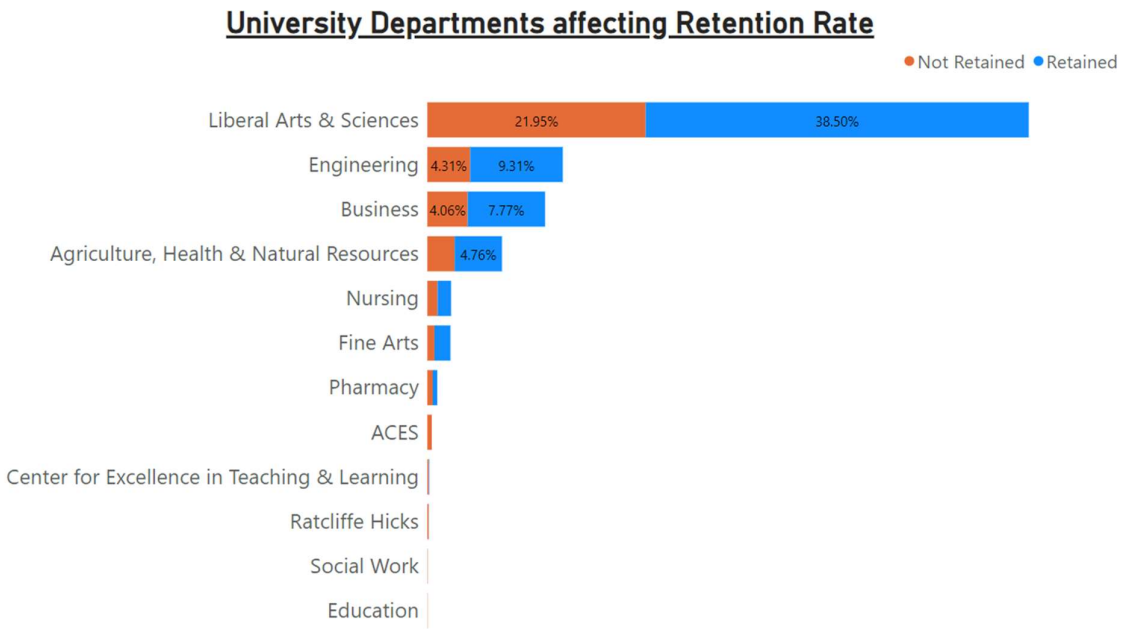

Figure 4 : Student Retention Rate Interactive Dashboard

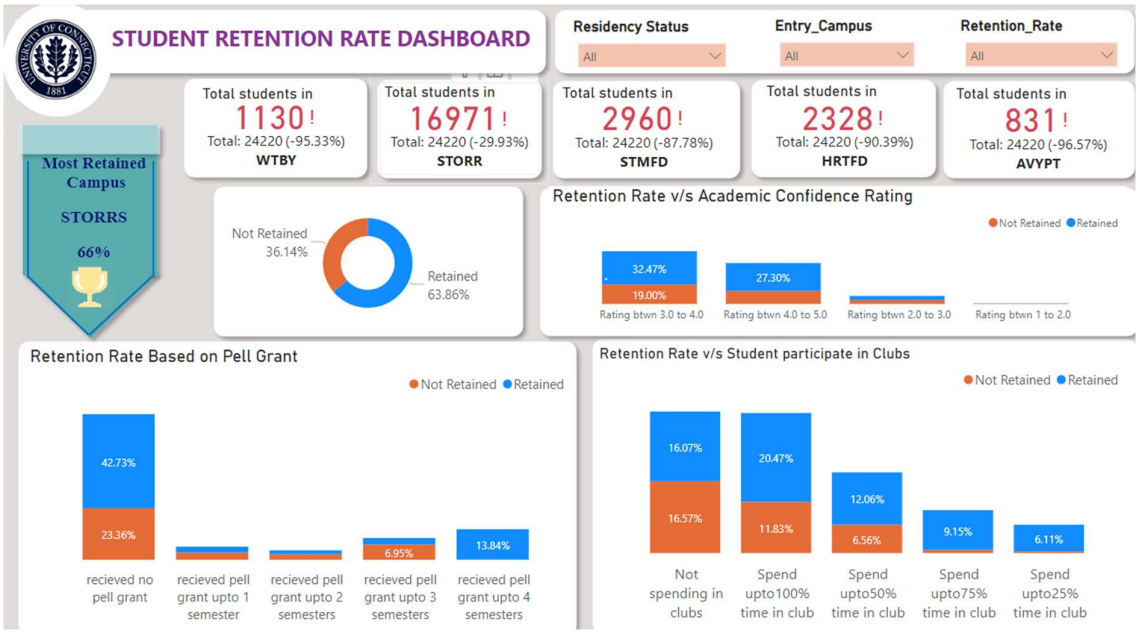

Figure 5 : Model Performance Chart

Figure 6 : Significant Column Contributions in the Modeling