WGU C951

Task 3

MACHINE LEARNING PROJECT PROPOSAL

Daniel LaForce

001119118

2/22/23

Table of Contents

A. Project Overview

This project helps solve the problem of disengaged students and student retention within higher education using artificial intelligence and machine learning. This proposal aims to develop a machine-learning solution that will help improve student engagement and retention at Western Governors University by analyzing student behavior and opinion throughout each semester.

A.1. Organizational Need

A considerable challenge facing today's higher education institutions is student engagement and retention. The low retention rate of students is a problem that affects the institution's success, and it is essential to find a solution. Western Governors University could see huge benefits from a machine learning solution that promises to improve student engagement and retention using cost-effective artificial intelligence to gather and analyze data on student behavior.

A.2. Context and Background

Artificial intelligence in higher education is becoming increasingly prevalent, and many examples of artificial intelligence are being used to improve student engagement and retention. Artificial intelligence can analyze data from various sources, such as student behavior, demographics, academic performance, and feedback. We can use this information to personalize the learning experience and give students the support they need to succeed.

A.3. Outside Works Review

The following are three outside works that explore machine learning solutions that apply to the need described in part A1.

a. "A Machine Learning, Artificial Intelligence Approach to Institutional Effectiveness in Higher Education" by John N. Moye. The author examines data sets that include measures of interactions observed during student-instructor interactions and average GPA relevance to withdrawn students. The author believes that a machine learning approach to Higher Education is a practical, effective, and systematic approach to twenty-first-century education.

b. “Higher Education Transformation for Artificial Intelligence Revolution: Transformation Framework” by Rawan Ghnemat, Adnan Shaout, and Abrar M. Al-Sowi. The authors discuss how Artificial Intelligence in Education (AIED) will help shape education in the foreseeable future and reinvent education as we know it. They discuss using "Intelligent educational recommendation systems," which utilize machine learning, data analytics, and business solutions to help determine student behavior and performance while personalizing learning on an individual basis. The authors believe this will help student retention and reduce student dropout rates. They believe the fundamental problem of current student engagement lies with the use of aging teaching methods use in direct lectures, which are, in turn, based on a list of predefined learning outcomes.

c. “The Impact of AI on Teaching and Learning in Higher Education Technology.” By Satya Vir Singh and Kamal Kant Hirnan. Satya and Kamal believe that students have artificial intelligence to thank for being able to learn and attend higher education whenever and wherever they like. With personalized feedback on assignments, quizzes, and other assessments, students will have the tools required to succeed. Furthermore, student engagement could be better within higher education now than at any time in the past.

A.4. Solution Summary

By utilizing artificial intelligence to gather and analyze data on student behavior, the detailed proposed machine learning solution will allow Western Governors University to gather and analyze student behavior data. Personalized support will be offered to students, helping to increase both engagement and retention.

A.5. Machine Learning Benefits

The proposed machine learning solutions will provide several benefits to Western Governors University by using artificial intelligence to gather and analyze data on student behavior. The solution will provide insights into student engagement and retention and, in turn, help fund the future growth of our institution. These solutions will also provide personalized support to students, helping to increase engagement and retention rates.

B. Machine Learning Project Design

B.1. Scope

In-scope items:

- Select a suitable artificial intelligence/machine learning solution algorithm we will employ.
- Develop a machine-learning solution and test for effectiveness with predefined test cases.
- Prepare the data for the Artificial Intelligence/machine learning algorithm(s).
- Determine how to measure the performance of the Artificial Intelligence/machine learning solution.
- Measure the performance of the proposed Artificial Intelligence/machine learning solution.
- Evaluate the project's success based on predefined criteria.

Out-of-scope item:

- Deploy and integrate the proposed solution into the existing system of Western Governors University's student-facing system.

B.2. Scope

Goals:

- To improve student engagement and retention at Western Governors University using machine learning and artificial intelligence.

Objectives:

- Utilize supervised learning to train the machine learning algorithm.
- Provide personalized learning experiences for each student.
- Reduce student withdrawals.

- Accurately predict outcomes based on the relationships between features and target variables.
- Identify problems, including lack of retention and understanding of the information or falling behind ideal progression time schedules.
- Continuously monitor and update the algorithm to ensure effectiveness.

Deliverables:

- Labeled datasets for training the system.
- Personalized recommendations for each student.
- Accurate predictions of student outcomes.
- Identification of problems affecting student engagement and retention.
- Ongoing monitoring and updates to the algorithm.

B.3. Standard Methodology

This project will follow the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology to ensure a structured and organized approach to implementing the machine learning solution. The following six-step process will be employed for our data mining projects by our Data Scientists:

- business understanding
- data understanding
- data preparation
- modeling
- evaluation
- deployment.

We will utilize this methodology to ensure a thorough and organized approach to developing the machine learning solution.

B.4. Projected Timeline

The projected timeline for this proposed project is as follows:

Week 1-2: The first two weeks will be spent on project planning, defining objectives, working on the deliverables, detailing the methodology, and identifying the required resources.

Week 3-4: We will start data collection and preparation, identification of sources and cleaning, transforming and integrating data sets, and finally, verifying data quality.

Week 5-6: After we verify data quality, we will begin exploratory data analysis and data visualization, feature engineering, and determining and identifying relationships, patterns, and any anomalies in the data.

Week 7-8: At the end of 8 weeks, we will have completed model development and the selection of the appropriate algorithms, designing and training the models, and evaluating the model’s performance.

Week 9-10: Sprint five will include model deployment and testing, integration of the model into our system, testing the model for accuracy, reliability, and scalability, and optimizing the performance.

Week 11-12: We conclude the project with sprint six by documenting the project, evaluating its effectiveness, and creating a plan for ongoing maintenance and improvement.

Sprint Schedule

Sprint	Start	End	Tasks
1: Project Planning	March 1, 2023	March 14, 2023	Scoping, defining objectives, deliverables, methodology, identifying required resources, and project planning.
2: Data Collection and Preparation	March 15, 2023	March 28, 2023	Identifying sources, collecting data, cleaning, transforming, integrating data sets, and verifying data quality.
3: Exploratory Data Analysis	March 29, 2023	April 11, 2023	Data visualization, feature engineering, and identifying relationships, patterns, and anomalies in the data.
4: Model Development and Evaluation	April 12, 2023	April 25, 2023	Selecting appropriate algorithms, designing and training models, and evaluating model performance.
5: Model Deployment and Testing	April 26, 2023	May 9, 2023	Integrating the model into the system, testing accuracy, reliability, scalability, and optimizing performance.

6: Finalizing the Project	May 10, 2023	May 24, 2023	Documenting the project, evaluating its effectiveness, and creating a plan for ongoing maintenance and improvement.

B.5. Resources and Costs

Resource	**Description**	**Cost**
Data Scientist	Responsible for data collection, cleaning, modeling, and deployment	$15,000/
Software Engineer	Responsible for system integration, testing, and deployment	$12,000
Project Manager	Responsible for project planning, management, and communication	$8,000
Cloud Infrastructure	Cost of hosting and deploying the machine learning system on a cloud platform	$5,000
Hardware	Cost of purchasing and maintaining hardware, such as servers, computers, and storage devices	$3,000
	Total	$40,000 monthly + $3,000

B.6. Evaluation Criteria

Describe the criteria used to evaluate and measure the success of the completed project.

Objective	Success Criteria
Develop a machine learning model that accurately predicts student withdrawal rates.	The model achieves at least 85% accuracy on test data
Improve student retention rates	Achieve at least a 10% reduction in student withdrawal rates
Reduce marketing costs by targeting high-risk students	The model identifies the top 20% of students with the highest withdrawal risk with at least 75% accuracy

C. Machine Learning Solution Design

C.1. Hypothesis

Western Governors University can substantially improve student engagement and retention using a machine learning algorithm.

C.2. Selected Algorithm

Supervised learning is the machine learning algorithm we will use in this project.

C.2.a Algorithm Justification

Because supervised learning will allow us to use labeled datasets to train the system and to use regression and classification, we can selectively feed the knowledge we want it to learn and throw out the dreck data. Supervised learning will prevent unwanted data from influencing artificial intelligence and its decisions.

C.2.a.i. Algorithm Advantage

By utilizing supervised learning to train the system, we can accurately predict outcomes based on the relationships between the features and target variables within the training data. We can also use supervised learning to create a personalized learning experience for each student by providing targeted recommendations based on their learning style and monitoring their continued progress through the coursework. Supervised learning can increase engagement and retention rates as students feel more supported and motivated to succeed. It can also accurately identify if students are suffering from problems, including but not limited to lack of retention and understanding of the information or falling behind ideal progression time schedules.

C.2.a.ii. Algorithm Limitation

One limitation we foresee in supervised learning is that it may not effectively capture complex relationships between the different variables. However, this limitation of supervised learning is its reliance on labeled data, which may be limited, biased, or otherwise tainted somehow. If the training data does not represent the entire student population or does not include relevant variables, the algorithm's predictions may cause some undesirable behaviors due to being inaccurate or incomplete. Additionally, the model may struggle to adapt to changes in student behavior, family life, or financial circumstances. It may also need help with unforeseen external factors, such as a pandemic, economic downturn, or problems within our system and coursework that can negatively impact student engagement and retention. Therefore, it is crucial to continuously monitor and update the algorithm to ensure it remains effective over time regardless of these external factors and unforeseen circumstances.

C.3. Tools and Environment

The tools and environment we will use are as follow:

- Programming language

- Machine learning libraries
- Data storage
- Data visualization tools
- Cloud computing service
- Integrated Development Environment

The selection of a programming language is vital to the project as we will utilize this language to implement the algorithms and models for supervised learning of the system. Because the language will help determine what machine learning libraries we must select both of these together. We will use machine learning libraries to implement and optimize the learning models. We have decided to use Python as our language, with PyCharm as our IDE and the Scikit-Learn library to predict user behavior using its classification algorithms.

We will store the student data used to implement the supervised learning models in a POSTGRESQL database stored on the Amazon Web Services platform. To visualize the data that we have mined and will be used to train the learning models, we will use Tableau, Matplotlib, and Seaborn.

As a part of the backend of our solution, we will employ a database management system to store student data and allow for the data's manipulation. This data will be used also be used to train the learning models.

C.4. Performance Measurement

Measuring the quality and performance of an educational institution such as Western Governors University is multifaceted. One way to assess performance is by tracking the number of students who complete their courses, graduate from their programs, and find employment in their fields of study. According to the National Center for Education Statistics (NCES), we can use this data to benchmark the performance of WGU's students (NCES, 2022).

Quality, on the other hand, can be measured by collecting feedback from students on their satisfaction with the learning experience and from employers on the job readiness of Western Governors University graduates. The Council for Higher Education Accreditation (CHEA) explains the importance of quality assurance and accreditation in higher education, highlighting the role of student and employer feedback in assessing the quality of educational programs (CHEA, 2019).

In addition to these metrics, other indicators, such as student retention rates and time to completion, can also be used to evaluate the effectiveness of WGU's artificial intelligence system for improving student engagement and retention. The United States Department of Education (USDE) provides data on various outcomes of college education, including retention rates, time to completion, and GPA. Our organization can use all of these to evaluate the impact of WGU's artificial intelligence system on student performance (USDE, 2021). Though Western Governors University does not use a GPA scoring system, we can utilize this information as a performance metric when analyzing data sets not mined locally.

Furthermore, we can utilize several metrics such as accuracy, precision, recall, F1 score, AUC-ROC, and confusion matrix to measure the effectiveness of the artificial intelligence system. These metrics will help to evaluate the performance of the artificial intelligence system in identifying students who need intervention and predicting their outcomes.

D. Description of Data Sets

D.1. Data Source

The data source for this proposal will be extracted from the National Center for Education Statistics (NCES), a primary federal entity for collecting and analyzing educational data in the United States (National Center for Education Statistics, 2022). Specifically, we will obtain graduation rates from the Fast Facts section of the NCES website (National Center for Education Statistics, 2022). Additionally, data on student retention rates, time to completion, and though Western Governors University doesn't have an equivalent metric, we will use the GPA metric collected from the College

Scorecard, a tool provided by the United States Department of Education that provides data on various outcomes of college education (United States Department of Education, 2021). We will assess the quality of educational programs by collecting feedback from students on their satisfaction with the learning experience, as well as from employers on the job readiness of Western Governors University graduates (Council for Higher Education Accreditation, 2019).

D.2. Data Collection Method

The implemented data collection may involve numerous methods, including but not limited to surveys at different times during the students' term, interviews, and analysis of existing data. In the case of measuring the quality and performance of an artificial intelligence model developed for improving student engagement and retention at Western Governors University, data collection will also involve using various metrics to evaluate the model's effectiveness.

One possible way to collect student engagement and retention data is through surveys or interviews with students, faculty, and staff. We can use these surveys to collect feedback on the learning experience, satisfaction with the program, and the effectiveness of the artificial intelligence model in improving engagement and retention.

Another way to collect data is by analyzing existing data, such as student records and performance metrics. This data can include data may include student retention rates, time to completion of a course or semester, GPA, and other relevant outcomes. The United States Department of Education's College Scorecard provides data on various outcomes of college education that we can use to evaluate the impact of Western Governors University's artificial intelligence system on student performance (United States Department of Education, 2021).

D.2.a.i. Data Collection Method Advantage

There is a massive advantage to our selected data collection method. One such advantage of utilizing the method used in this project is that it allows for collecting vast amounts of data in a relatively short time. This method can give us a quick look into the results and provide a good understanding of student engagement and retention. According to Johnson and Christensen (2010), online surveys are an efficient and effective way to collect data from many participants. This method can save time and University expenses compared to traditional survey methods, such as paper surveys or in-person interviews.

Using this data collection method also provides a level of anonymity to participants. In today's world, privacy and anonymity are essential to almost everyone. Our data collection method will respect the student's anonymity and encourage honest and open responses. According to Kessler (2002), anonymous surveys can reduce social desirability bias and increase response rates, leading to more accurate data. The online survey method used in this project allows participants to respond to questions without fear of judgment or consequences, increasing the likelihood of honest responses and substantially improving the collected data.

D.2.a.ii. Data Collection Method Limitation

While our data collection method is most advantageous, there is a potential bias in self-reported data. As noted by Dillman (2014), individuals may only sometimes provide accurate or complete information when responding to surveys or questionnaires. This potentially tainted information can result in a skewed dataset that does not accurately reflect our student base.

Also, there is the possibility of non-response bias, where specific individuals or groups may be more likely to opt out of the survey or not respond, leading to a biased sample.

Likewise, it is also possible that the data collection method might be limited by the quality and availability of data sources; as McLeod and Lewis (2012) noted, secondary data sources may not always be reliable or complete, which can impact the validity of the findings.

Limited resources and expertise are also required to collect and analyze the data. Data collection can be time-consuming and costly and may require specialized skills or software to effectively analyze the data (Dillman, 2014). Unfortunately, we are charting new waters in utilizing artificial intelligence to improve our student engagement and retention and need a way to accurately determine if we need to hire more than one data scientist to collect and analyze the data. Exciting and advantageous software suites and tools will no doubt enter the world of artificial intelligence as time progresses and needs to be taken into account in the budget and is outside the scope of this proposal.

D.3. Quality and Completeness of Data

We will preprocess the data to prepare it for use in the algorithms. The process will involve several steps, including cleaning, normalization, and feature selection.

Cleaning the data involves removing any inconsistencies, errors, or missing values in the dataset. The data cleaning can be done by our Data Scientists using tools such as Python's Pandas library or Excel's Data Cleaning functions.

Normalization involves transforming the data so that it is on a consistent scale, which can help improve the performance of the algorithms. Normalization methods include Min-Max Scaling, Z-Score Scaling, and Robust Scaling.

Feature selection involves identifying the most critical variables in the dataset, which can help improve the accuracy and efficiency of the algorithms. Feature selection methods include correlation analysis, principal component analysis (PCA), and recursive feature elimination (RFE).

After our Data Scientists complete these preprocessing steps, the data will be ready to be fed into the machine learning algorithms.

D.4. Precautions for Sensitive Data

Working with sensitive data requires responsible handling, management, and storage. When working with sensitive data, it is vital to adhere to ethical and legal guidelines to protect the privacy and confidentiality of our students. According to the National Institutes of Health (NIH) guidelines, researchers should obtain informed consent from student participants and follow appropriate data handling and storage procedures to minimize the risk of data breaches (NIH, 2019).

Data should only be accessible by authorized personnel, and we should implement strict access controls and authentication mechanisms, as well as regular monitoring and auditing of data access logs (Jagadish, 2014). Only some within the organization should have such access, and we should restrict to our Data Scientists, Data Project Managers, and any others intimately working with the data collection team.

When communicating about sensitive data, it is essential to be open and transparent, especially concerning the purpose and scope of the data collection and analysis. Western Governors University should explain the potential benefits and risks to the student's privacy by participating in the study and ensure that participants accept the risks and clearly understand how their data will be used and protected (NIH, 2019).

Finally, it is essential to acknowledge that handling sensitive data requires a high level of responsibility and professionalism. Our data analysis team should know the potential consequences of mishandling sensitive data and take steps to minimize these risks through appropriate training, policies, and procedures (Jagadish, 2014).

References

Council for Higher Education Accreditation. (2019). Quality Assurance and Accreditation. https://www.chea.org/quality-assurance-and-accreditation

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). Internet, phone, mail, and mixed-mode surveys: The tailored design method. John Wiley & Sons.

Ghnemat, R., Shaout, A., & Al-Sowi, A. M. (n.d.). Higher Education Transformation for Artificial Intelligence Revolution: Transformation Framework. https://www.researchgate.net/publication/336760245_Higher_Education_Transformation_for_Artificial_Intelligence_Revolution_Transformation_Framework

Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big data and its technical challenges. Communications of the ACM, 57(7), 86-94. https://doi.org/10.1145/2611567

Johnson, R. B., & Christensen, L. (2010). Educational research: Quantitative, qualitative, and mixed approaches. Sage publications.

Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S. L., ... & Zaslavsky, A. M. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychological Medicine, 32(6), 959-976. DOI: 10.1017/S0033291702006074

McLeod, S. A., & Lewis, J. (2012). Secondary data analysis. Sage Publications.

Moye, J. N. (n.d.). A Machine Learning, Artificial Intelligence Approach to Institutional Effectiveness in Higher Education. https://scholarsarchive.byu.edu/etd/7159/

National Center for Education Statistics. (2022). Graduation rates. https://nces.ed.gov/fastfacts/display.asp?id=40

Pressman, R. S. (2014). Software engineering: A practitioner's approach. Palgrave Macmillan.

Sharma, P., Singh, P., & Singh, T. (2019). A review of machine learning algorithms. International Journal of Engineering and Advanced Technology (IJEAT), 8(6), 61-63.

United States Department of Education. (2021). College Scorecard. https://collegescorecard.ed.gov/