Analytics Startup Plan

Synopsis: ***This document provides a high-level walkthrough of the activities required to guide completion of the analysis.***

Project	*Predicting student's dropout, academic success*
Requestor	*OPORAJITA TAMANNA 301253651*
Date of Request	*JULY 15, 2024*
Target Quarter for Delivery	*SECOND HALF OF THE SEMESTER*
Epic Link(s)	*https://www.kaggle.com/datasets/naveenkumar20bps1137/predict-students-dropout-and-academic-success*
Business Impact	*Identifying and addressing factors leading to student dropout can improve retention rates and institutional effectiveness.*

1.0 Business Opportunity Brief

Clearly articulated business statement of the Ask, opportunity, or problem you are trying to solve for. A crucial step is to understand the nature of the business, system or process and the desired problems to be addressed. This will be communicated back to All stakeholders for alignment.

The specific ask:

The task is to analyze the dataset to predict which students are at risk of dropping out and identify the factors contributing to this risk.

1.1 Supporting Insights

Define any supporting insights, trends, and research findings. Where relevant, list key competitors in the market. What are their key messages, products & services? What is their share of the market, nationally and regionally?

Trends and research findings: Dropout rates have significant impacts on educational institutions. Identifying at-risk students early can help implement interventions to improve retention.
Key competitors: Other educational institutions using advanced analytics to improve retention.
Share of market: Understanding how retention rates compare nationally and regionally.

1.2 Project Gains

Describe any revenue gains, quality improvements, cost, and time savings (as applicable). What will you do differently and why would our customers care? What are the implications if we do nothing? This section is particularly key for prioritization against company goals and KPI's.

Revenue gains: Increased retention rates lead to higher revenue from continued tuition.
Quality improvements: Enhanced support services for at-risk students.
Cost and time savings: Reducing dropout rates decreases the need for recruitment to replace lost students.

Note: Completion of the following sections is possible only after a careful assessment and triage of the Ask. This is required to determine scope, resources, time, priority, and data availability.

2.0 Analytics Objective

List the key questions, assumptions and define the hypotheses. Often the deliverable may not just be an analysis output, however a recommended operating model or blueprint for a pilot etc.

Note: Asking the right questions and understanding the problem will lead to the right data, right mathematics, and right techniques to be employed.

Key questions: What factors are most predictive of student dropout? How can we accurately predict which students are at risk?
Assumptions: The dataset accurately reflects relevant factors influencing dropout.

2.1 Other related questions and Assumptions:

List any assumptions that may affect the analysis
The dataset includes comprehensive and accurate records of student demographics, academic performance, attendance, and other relevant factors that can influence dropout rates. It is assumed that all necessary variables have been consistently recorded and that there are no significant missing values or errors in the data.

Related Questions:
How do socioeconomic factors, such as family income and parental education levels, impact student dropout rates?
What role do academic performance and attendance play in predicting which students are at risk of dropping out?
Are there specific periods during the academic year when students are more likely to drop out?
Can we identify any patterns or trends in the data that correlate with higher dropout rates?

2.2 Success measures/metrics

What does success look like? Define the key performance indicators (success definition/indicators, drivers, and key metrics) against which the objectives will be analyzed. These should be drawn from the interlock meeting with key stakeholders and will inform them of the approach and methodology for the analysis.

Key performance indicators:
Accuracy of the predictive model
Precision and recall of identifying at-risk students
Reduction in dropout rates after implementing interventions

o

2.3 Methodology and Approach

Now that you have a good understanding of the Ask and deliverable, detail the recommended approach/methodology.

Type of Analysis: *Logistic regression, decision trees, and other classification methods.*

The initial approach will be to use a classification model to determine which student-level variables (e.g., demographics, academic performance, attendance, socioeconomic status) are most significantly related to a student's likelihood to drop out. We will also use other techniques, such as logistic regression and decision trees, to verify our findings and ensure robustness.

Methodology: *The key questions from the 'Analytics objective' section will guide our analysis. Specifically, we will investigate which factors are most predictive of student dropout and how we can accurately predict which students are at risk.*

We will start by identifying all students enrolled at the start of the academic year. We will then define the response variable to be a 1 if they dropped out during the year, and 0 otherwise. We will build a decision tree based on this sample and observe which variables are the most important in determining whether students drop out. We can then repeat this analysis using different subsets of the data (e.g., by semester or grade level) to check if the same variables are identified as the most important drivers of dropout, or if the importance of variables changes over time or across different groups of students.

Output: *The output will include a detailed report on the factors most predictive of student dropout, visualizations of the decision trees and other models used, and recommendations for interventions based on the findings. Additionally, we will provide a predictive model that can be used by the institution to identify at-risk students in future academic years.*

.

3.0 Population, Variable Selection, considerations

Capture learning about the data available today location, structure, and reliability; this would include data in operational systems including dealer sourced, data warehouse and any CRM or email marketing systems available today.

Audience/population selection: All enrolled students

Observation window: Data from the past academic year (e.g., September 2022 to August 2023)

Inclusions:

- Demographic information (e.g., age, gender, nationality)
- Academic performance (e.g., grades, evaluations)
- Socio-economic status (e.g., parental education, occupation)
- Attendance records
- External factors (e.g., unemployment rate, inflation rate)

Exclusions: Personally identifiable information (e.g., student names, addresses)

Data Sources:

- Academic records database
- Student Information System (SIS)
- Socio-economic surveys
- External economic data sources

Audience Level: All enrolled students in the institution

Variable Selection: Features such as age, gender, nationality, marital status, parental education and occupation, grades, attendance records, and external economic indicators (e.g., unemployment rate, inflation rate) will be used.

Derived Variables:

- Attendance percentage
- Composite scores for academic performance
- Socio-economic status index

Assumptions and data limitations: Data is assumed to be complete and accurate. Missing values or inaccuracies in data recording might affect the model's performance.

4.0 Dependencies and Risks

Identification of key factors that may influence the outcome of the project and likelihood of it happening:

Risk	Likelihood (based on historical data)	Delay (based on historical data)	Impact

Data Quality Issues	*Medium*		*Missing or inaccurate data can significantly affect the accuracy of predictive models.*
Model Accuracy	Medium		*Misclassification of at-risk students could lead to inappropriate interventions.*
Data Privacy Concerns	*High*		*Handling sensitive student data requires strict adherence to privacy regulations.*
Changes in Institutional Policy	*Low*		*Policy changes can alter key variables impacting dropout rates.*

5.0 Deliverable Timelines

List key dates and timelines as a work-back schedule. Activate line items based on complexity and line-of-sight required. Will set the stakeholder expectations for the process.

Item	Major Events / Milestones	Description	Scope	Days	Date
1.	Kick-off / Formal Request	*Initial project briefing*	*Project Overview*	*3*	
2.	EDA	*Assess data quality, perform in-depth analysis of statistical measures*	*Identify Issues*	*7*	*July 8th, 2024*
3.	Modeling	*Develop and validate predictive models*	*Model Training & Validation*	14	*July 15th, 2024*
4.	Governance	Documentation, Governance Plan, presentation peer review		7	*July 29th, 2024*
5.	Presentation	Presentation of refined findings	*Internal Validation &* *Stakeholder Feedback*	*7*	*August 5th, 2024*

6.	QA Output	*Quality assurance of analysis*	*Review and Corrections*		*August 5th, 2024*
7.	Portfolio	Presentation, Portfolio		*7*	*August 12th, 2024*
8.	Go/No Go	Decision point for final steps	*Stakeholder Decision*		*August 15th, 2024*
9.	Delivery & sign-off	Final delivery of the project	*Completion*		*August 15th, 2024*