Student Dropout Prediction

Galuh Alifani

Agenda

1. **Background & Objective**
2. Dataset Context & Methodology
3. Exploratory Data Analysis
4. Feature Engineering
5. Model Summary
6. Web Application

Background & Objective

The problem:

Academic success plays a crucial role in shaping an individual's future. However, many students face challenges that leads to potential dropout.

The objective:

Create a predictive model and a corresponding web-app to identify whether a student is likely to drop-out, so that the school can offer early interventions.

We mostly will track recall rate to ensure we minimize false negatives.

Target users:

Mainly for school academic department, teachers and academic counselors.

Agenda

1. Background & Objective
2. **Dataset Context**
3. Exploratory Data Analysis
4. Feature Engineering
5. Model Summary
6. Web Application

Dataset Context

Provider	
Dataset Title	Predict Students' Dropout and Academic Success

Description & Content	This dataset contains 4,000 records of students from a higher education institution from various undergraduate degrees. Information in the dataset includes:
	• Information known at the time of student's enrollment
	  • Education background
	  • Family & financial background
	  • Macroeconomic conditions;
	• Students' academic performance in 1st & 2nd semesters
	• Students' enrollment status (Graduated, Dropout, Enrolled) at the end of 2nd Semester

Methodology: Step by Step Approach

Several processes taken during data processing, model building, and app-deployment

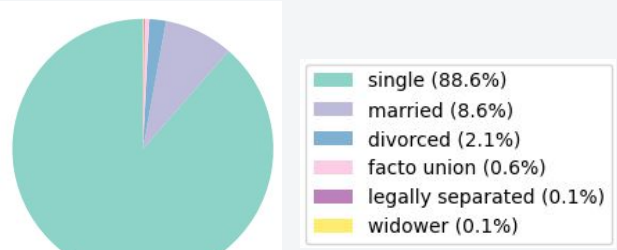
1. **Data clean-up & pre-processing**
2. **Exploratory Data Analysis**
3. **Feature Engineering**
4. **Model Definition & Training**
5. **Model Saving & Inference**
6. **Web App Development & Deployment**

Agenda

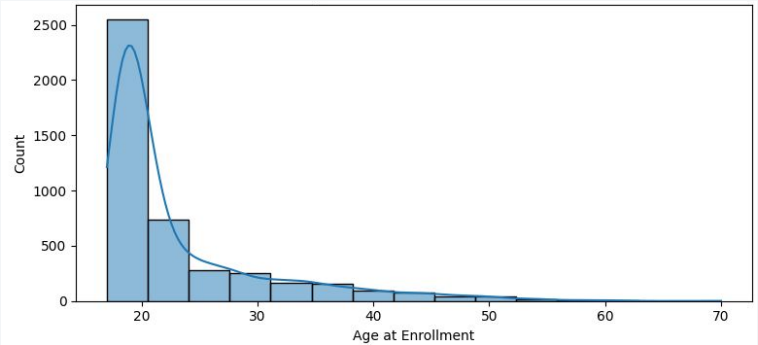
1. Background & Objective
2. Dataset Context & Methodology
3. **Exploratory Data Analysis**
4. Feature Engineering
5. Model Summary
6. Web Application

Demographic Overview of Dataset

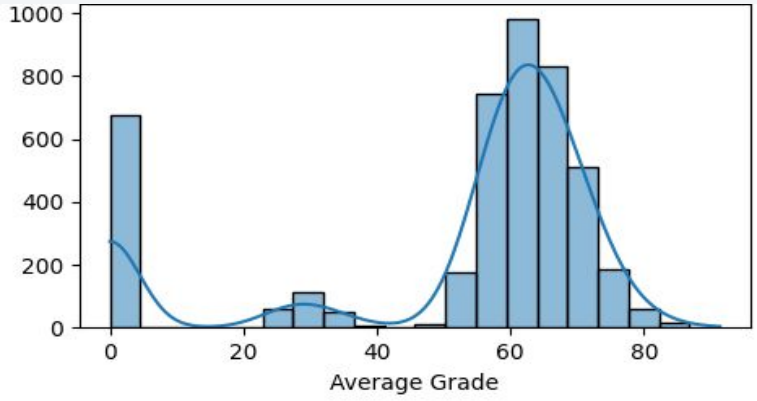
88% are single



Wide age range, with outliers at age 70

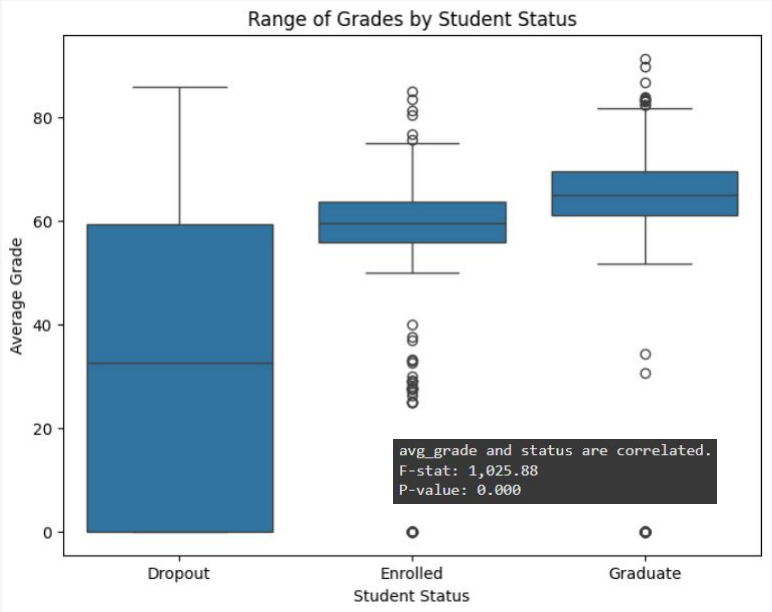


Average grade is 60-70

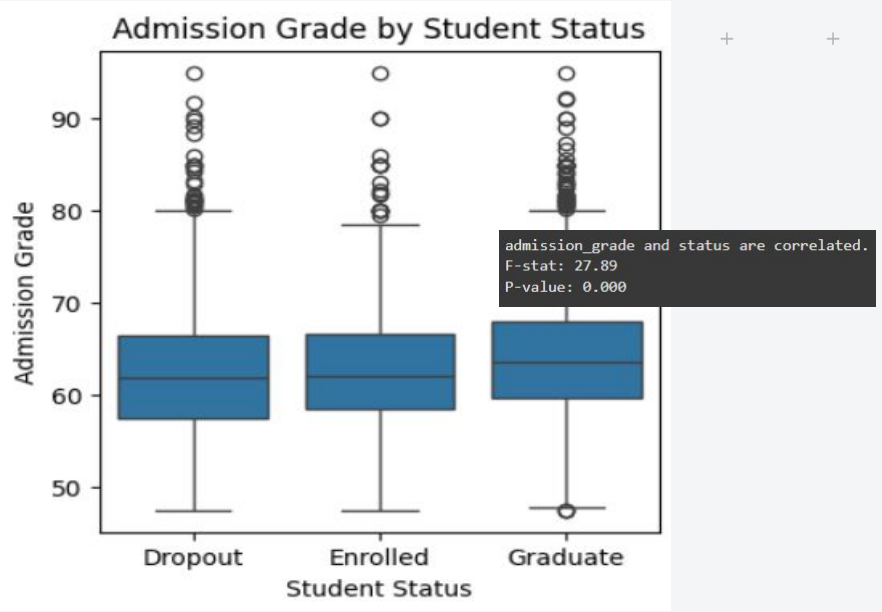


Do dropout students tend to have lower semester grades & admissions grade?

Average grade differs by status



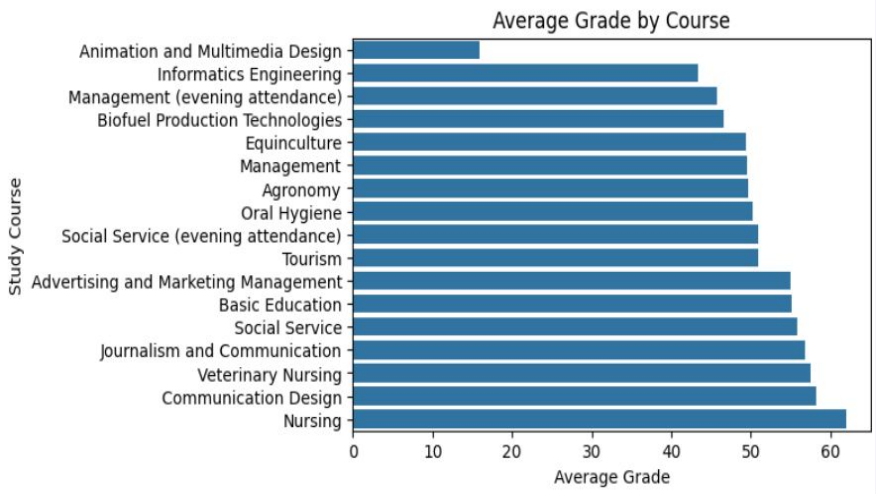
...with less difference on admissions grade



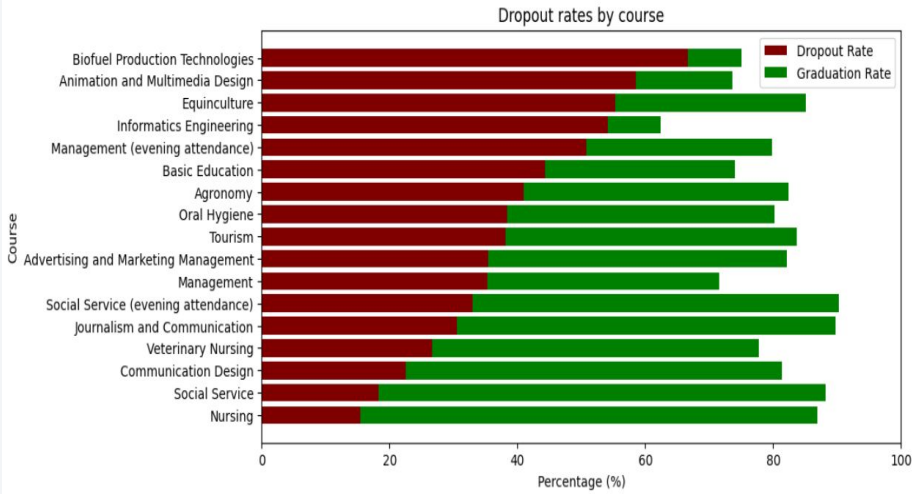
- Dropout students tend to have lower average semester grades compared to students with other statuses
- Dropout student also have slightly lower average admissions grade, especially compared to graduated students

Do different courses have different average grade and dropout rate?

Some of courses that have lower avg grade..



...also have higher dropout rates

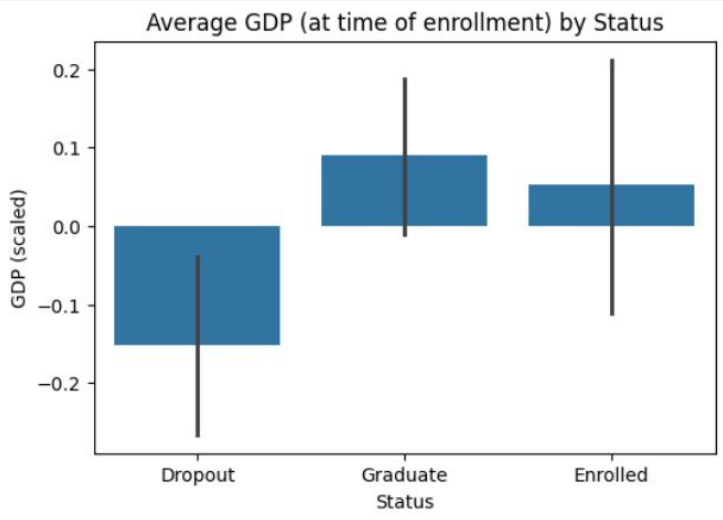


Yes, in average, different courses have quite sizeable differences in average grade and/or dropout rates.

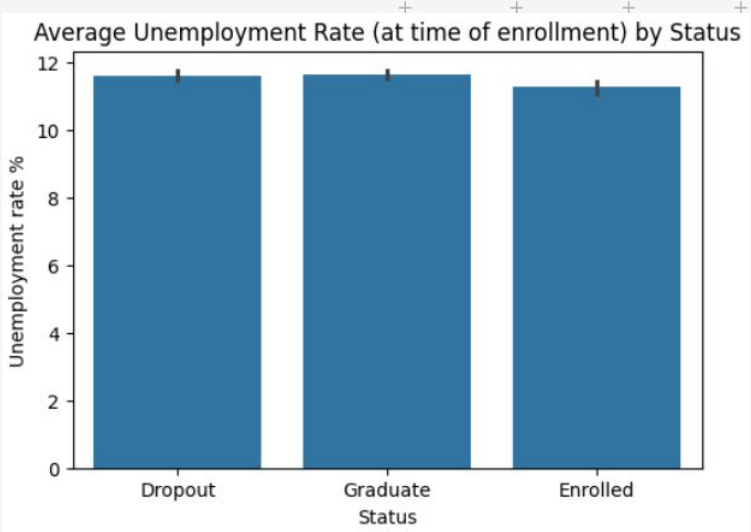
- Courses with technical or specialized content tend to have lower average grade and higher dropout rates, which could indicate that **students face more challenges in these areas** or large variety between students' aptitude.

Do macroeconomic condition impact student dropout rates?

Statuses have different avg GDP



But minor difference in unemployment rate

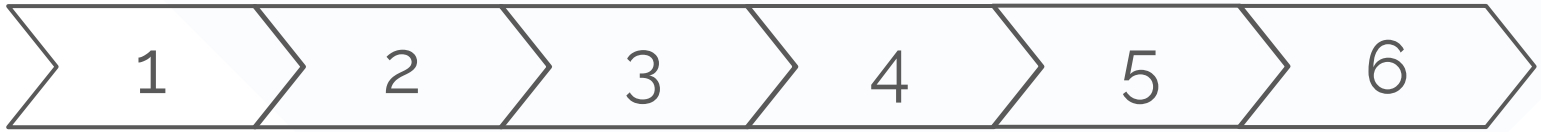


Economic stability and financial resources likely play a significant role in student dropout risk, where those who enrolled in study program during the times when the country have stable or high GDP, tend to perform well, whereas those who enrolled during recession might have increased risk of dropout.

Agenda

1. Background & Objective
2. Dataset Context & Methodology
3. Exploratory Data Analysis
4. **Feature Engineering**
5. Model Summary
6. Summary & Conclusion

Feature Engineering Approach



Data Splitting

- 75%-25% Train-Test Split
- Use stratify parameter to maintain class balance

Feature Selection

- Feature Importance
- Cardinality
- Multicollinearity:
- Phik correlation

Encoding

With One-Hot Encoding due to nominal categorical data

Outlier Handling

- Skewness analysis
- Calculate outlier % and boundaries
- Capping (winsorizing) at 99th percentile

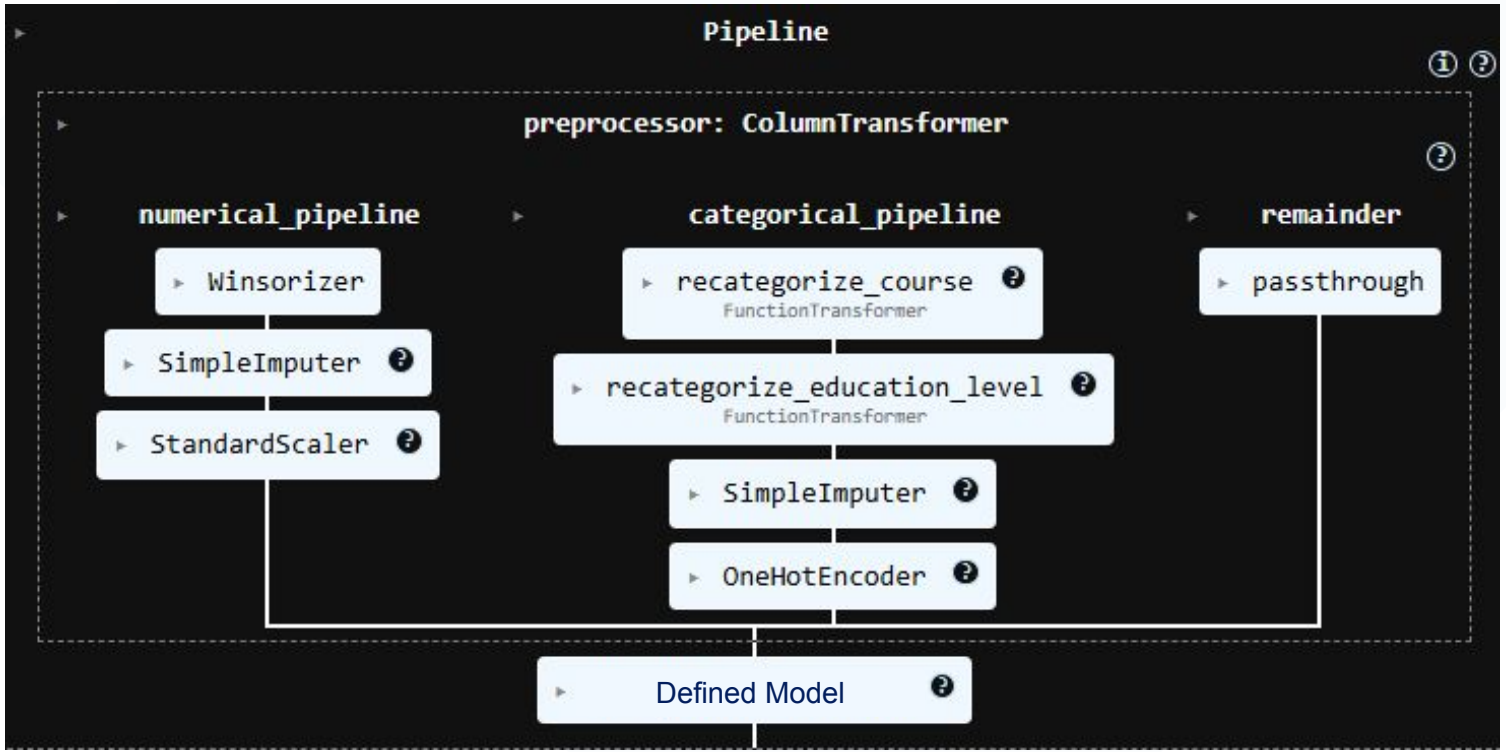
Feature Transformation

- Transform features of high multicollinearity
- Transform high-cardinality features

Scaling

Using standar scaler

Pipeline Overview



Agenda

1. Background & Objective
2. Dataset Context & Methodology
3. Exploratory Data Analysis
4. Feature Engineering
5. **Model Summary**
6. Web Application

Step1: 5 models are trained with standard/default parameters

KNN

SVM

Decision Tree

Random Forest

CatBoost

Step 2: Cross-Validation Best model is CatBoost with 0.83% avg recall rate

	Model	Train Recall Mean	Train Recall Range	Test Recall Mean	Test Recall Range	Average Score Variance	Train-Test Diff
0	KNN	0.7467	0.7104 - 0.7831	0.6986	0.627 - 0.7702	0.107955	0.0341
1	SVM	0.8415	0.8152 - 0.8677	0.8141	0.7658 - 0.8624	0.074560	0.0197
2	DecisionTree	0.8574	0.8373 - 0.8776	0.8085	0.7625 - 0.8544	0.066089	0.0263
3	RandomForest	0.8358	0.8097 - 0.862	0.8310	0.7989 - 0.8631	0.058254	0.0049
4	CatBoost	0.8574	0.8321 - 0.8827	0.8338	0.7946 - 0.873	0.064546	0.0145

Step 3: Hyperparameter Tuning on CatBoost with GridSearchCV

Parameters to tune:

- Iterations (100,500);
  - Iterate to boost fit
- Depth (4,6,8,10);
  - Max depth of each tree
- Auto Class Weights (Balanced);
  - Adjust class weight to handle imbalance
- l2_leaf_reg (5,20)
  - Regularize to reduce overfitting

Best hyperparameters for CatBoost:

{ 'classifier__auto_class_weights': 'Balanced',
'classifier__depth': 4,
'classifier__iterations': 500,
'classifier__l2_leaf_reg': 20}

Best recall for CatBoost: 0.8677137732763623

Step 3: Parameter Tuning

Cross-Validation Post-Tuning

```
Best Model: CATBOOST (POST-TUNED)
Avg Test Recall from Best Model: 0.8592
```

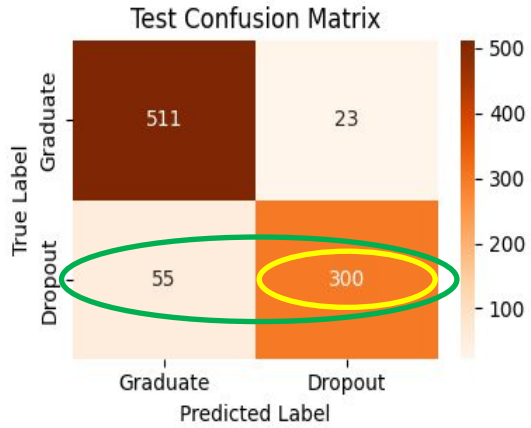
	Model	Train Recall Mean	Train Recall Range	Test Recall Mean	Test Recall Range	Average Score Variance	Train-Test Diff
0	CatBoost (Pre-Tuned)	0.8574	0.8321 - 0.8827	0.8338	0.7946 - 0.873	0.064546	0.0145
1	CatBoost (Post-Tuned)	0.8658	0.843 - 0.8886	0.8592	0.831 - 0.8873	0.050961	0.0063

Coefficient Analysis

```
Top 5 most important features ===
```

	Feature	Importance
6	s2_approved	58.056518
34	is_tuition_paid	7.647148
7	s2_grade	6.440061
3	s1_grade	2.976736
1	admission_grade	2.458692

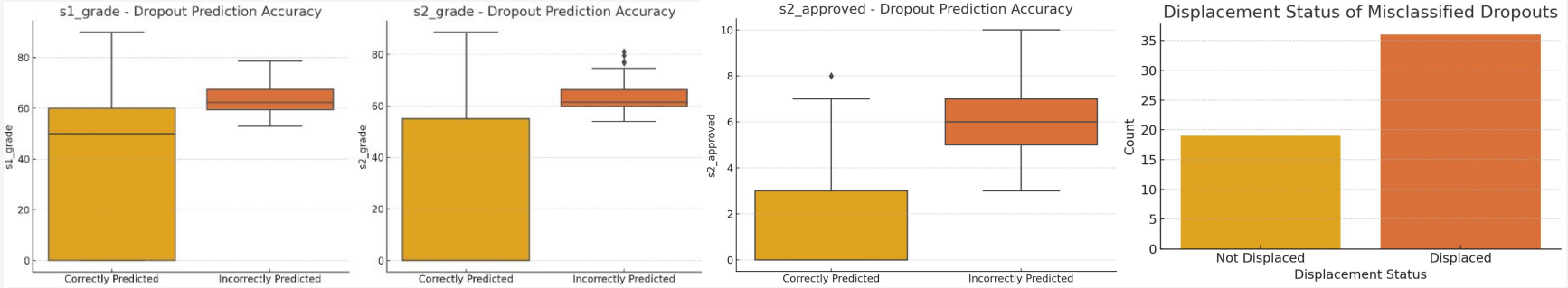
Test Prediction



Misclassification Analysis

Most of the misclassified students have high scores & financially OK

..but are displaced



In coefficient summary, most of the high-ranked coefficient are around academic result and financial condition, with `is_displaced` being in rank 23rd (med-low importance).

However, most of the misclassified students are those with high academic result, good financial condition (tuition is paid, not a debtor), but are displaced.

This means that the model might be overestimating the weight of academic result, while underestimating the weight/impact of feature `is_displaced`, contributing to incorrect predictions.

Summary & Conclusion

Most dropout students are potentially driven by low academic aptitude and financial distress. Our model is able to capture most of this (86% recall)

- Based on training 5 different models, tuned CatBoost has the best outcome:
  - ✔ 86% recall rate and 95% AUC score
  - ✔ No overfitting

- The highest-weighted parameter based on our best-trained model, are those parameters relating to academic aptitude, as well as financial condition.

However, there are high-performing & financially-stable students who are dropouts *and* misclassified

- Most of the misclassified students are “displaced”, which somehow the model missed to consider it
- The model might be overestimating the weight of academic and financial factors and underestimating the weight of psychological factors such as `is_displaced`.
- Imbalance dataset also plays a factor

Next steps: Evaluate weight of non-academic/financial parameters, increase complexity of model & add more relevant dataset of dropout classes

Agenda

1. Background & Objective
2. Dataset Context & Methodology
3. Exploratory Data Analysis
4. Feature Engineering
5. Model Summary
6. **Web Application**

Web App: Predict Student Dropout Status