Model Governance Framework for Predictive Student Outcomes

Validation, Monitoring and Governance

Centennial College

Business Analytics Capstone Project

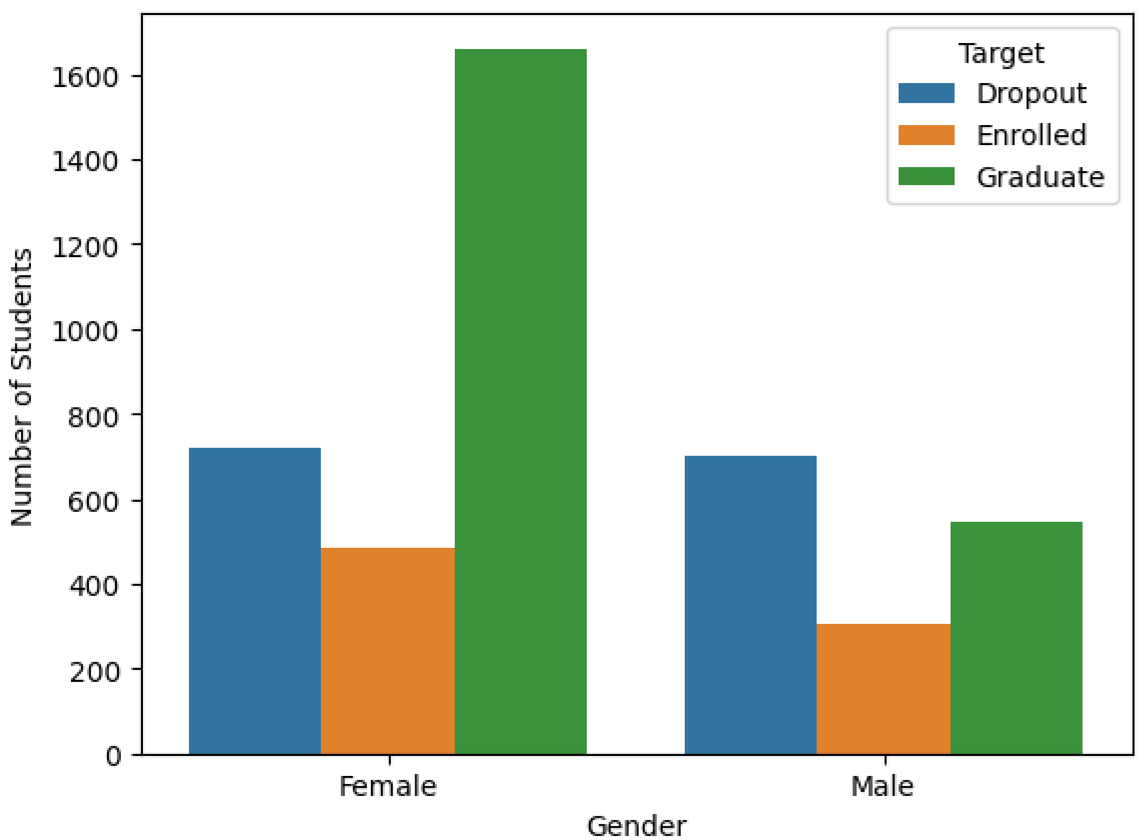
Oporajita Tamanna
August 15th, 2024

1.0. Variable Level Monitoring

1.1. Model Build Variable Level Statistics

In the initial stages of building the student dropout prediction model, key variables were scrutinized to ensure data quality and relevancy. Descriptive statistics such as mean, median, standard deviation, and distribution of categories were computed for all numerical and categorical variables.

Numerical Variables: The numerical variables, such as age and unemployment rate, were summarized. For instance, the average age of students was calculated alongside its distribution across the dataset.
Categorical Variables: Categorical variables like 'Marital status,' 'Gender,' and 'Target' (which indicates whether a student drops out, remains enrolled, or graduates) were also analyzed. The frequency distribution for each category was computed.



1.2. Acceptable Ranges

Based on the descriptive statistics, acceptable ranges for each variable were identified:

Age: Typically ranges between 18 and 35 years. Values falling outside this range were considered anomalies and flagged for further inspection.

Unemployment Rate: Expected to range between 0% and 15%. Values outside this range may indicate data entry errors or exceptional cases that require additional validation.

Marital Status, Gender, and Course: Each categorical variable has predefined categories (e.g., Male, Female for Gender) with acceptable distributions based on historical data.

```
student_df.describe()
```

	Marital status	Application mode	Application order	Course	Daytime/evening attendance	Previous qualification	Mother's qualification	Father's qualification	Mother's occupation	Displaced	...	Debtor	Tui fee to
count	4424.000000	4424.000000	4424.000000	4424.000000	4424.000000	4424.000000	4424.000000	4424.000000	4424.000000	4424.000000	...	4424.000000	4424.0
mean	1.178571	6.886980	1.727848	9.899186	0.890823	2.531420	12.322107	16.455244	7.317812	0.548373	...	0.113698	0.8
std	0.605747	5.298964	1.313793	4.331792	0.311897	3.963707	9.026251	11.044800	3.997828	0.497711	...	0.317480	0.3
min	1.000000	1.000000	0.000000	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	0.000000	...	0.000000	0.0
25%	1.000000	1.000000	1.000000	6.000000	1.000000	1.000000	2.000000	3.000000	5.000000	0.000000	...	0.000000	1.0
50%	1.000000	8.000000	1.000000	10.000000	1.000000	1.000000	13.000000	14.000000	6.000000	1.000000	...	0.000000	1.0
75%	1.000000	12.000000	2.000000	13.000000	1.000000	1.000000	22.000000	27.000000	10.000000	1.000000	...	0.000000	1.0
max	6.000000	18.000000	9.000000	17.000000	1.000000	17.000000	29.000000	34.000000	32.000000	1.000000	...	1.000000	1.0

8 rows × 21 columns

1.3. Missing Values

Assessment:
A thorough examination of the dataset was conducted to identify any missing values across all variables. The isnull().sum() function was employed to quantify missing data.

Findings:
There were no missing values identified in the dataset. This ensures the integrity of the dataset for subsequent modeling processes.

1.4. Variable Drift Monitoring Tolerance

To ensure the model remains relevant over time, variable drift (i.e., changes in the statistical properties of input variables) was monitored with a tolerance level set at 5% deviation from the original data distribution. This involved comparing new data distributions with the original model build statistics to detect significant changes.

2.0. Model Monitoring, Health & Stability

2.1. Initial Model Fit Statistics

The model was evaluated using several key fit statistics during the training phase:

Logistic Regression Model: Achieved a training accuracy of 87.87% and a validation accuracy of 90.08%.
Decision Tree Model: Provided an interpretable model with a focus on feature importance and tree structure.
Random Forest Model: The baseline model achieved an accuracy of 92% on validation data, indicating robust performance across different student profiles.

2.2. Parameter #1

The model's accuracy was consistently monitored across different stages:

Training Accuracy: The Logistic Regression model's training accuracy was 87.87%, while the validation accuracy reached 90.08%, indicating good generalization.

Validation Accuracy: Random Forest with hyperparameter tuning achieved a validation accuracy of 92%, outperforming other models. These metrics indicate that the model performs well on unseen data.

2.3. Parameter #2

Confusion Matrix & Classification Report

The confusion matrix and classification report provided deeper insights:

Logistic Regression: The model was trained on 70% of the dataset and validated on the remaining 30%. The model showed high precision (0.91 for graduates) and recall (0.93 for graduates), with an overall F1-score of 0.92 for graduates and 0.87 for dropouts.

Decision Tree: Provided a balanced precision and recall across both classes, though slightly lower than the Logistic Regression model.

Random Forest: After tuning, the Random Forest model demonstrated superior classification performance, particularly in correctly identifying graduates (class 2).

Random Forest Hyperparameter Tuning:

In the tuning process, key hyperparameters such as the number of trees (n_estimators), the maximum depth of the trees (max_depth), and the number of features considered for splitting (max_features) were optimized to enhance the model's performance.

3.0. Risk Tiering

The governance framework integrates automation tools to enhance the efficiency of monitoring and reporting for the student dropout prediction model:

3.1.1 Automated Alerts and Dashboards

Real-time dashboards display critical metrics such as model accuracy, dropout prediction rates, and feature importance rankings. Automated alerts are triggered when predefined thresholds, such as a sudden drop in accuracy or a significant change in feature importance, are breached, ensuring prompt intervention and corrective actions.

3.1.2 Quarterly Performance Reviews

Comprehensive reports are generated quarterly, summarizing the model's overall performance, trends in dropout predictions, any observed drift in data or model behavior, and corrective actions implemented. These reports provide stakeholders with clear, actionable insights, enabling proactive decision-making to maintain and improve the effectiveness of the dropout prediction model.