Feature Selection Options by Scoring Metric

Each scoring metric offers a distinct perspective on feature importance. The following outlines the selected features for each metric along with a brief explanation of how these features contribute to understanding student retention.

1. Feature Selection Options

1.1. Scoring Metric: “accuracy”

- **Selected Features:**
  - **Tuition fees up to date:** Indicates financial reliability, a key predictor of retention.
  - **Age at enrollment:** Reflects student maturity and readiness.
  - **2nd_sem_perf_ratio:** Represents performance consistency.
  - **Failed units ratio:** Serves as an indicator of academic challenges.
  - **Course_Graduate_prob:** Estimates the likelihood of course completion.

1.2. Scoring Metric: “precision_macro”

- **Selected Features:**
  - **Tuition fees up to date**
  - **Age at enrollment**
  - **Curricular units 2nd sem (grade):** Provides direct insight into early academic success.
  - **2nd_sem_perf_ratio**
  - **Failed units ratio**
  - **Late_enrollment:** Captures risks associated with delayed enrollment.
  - **Parental_influence:** Reflects the impact of external support.
  - **Course_Graduate_prob**
  - **Application mode_Enrolled_prob:** Indicates the influence of the mode of application.

1.3. Scoring Metric: “recall_macro”

- **Selected Features:**
  - **Mother's qualification:** Acts as a proxy for educational support.
  - **Father's qualification**
  - **Tuition fees up to date**

    - **Age at enrollment**
    - **Inflation rate:** Introduces an economic dimension.
    - **2nd_sem_perf_ratio**
    - **Failed units ratio**
    - **Late_enrollment**
    - **Course_Graduate_prob**
    - **Application mode_Enrolled_prob**

1.4. Scoring Metric: “f1_macro”

- **Selected Features:**
    - **Mother's qualification**
    - **Father's qualification**
    - **Tuition fees up to date**
    - **Age at enrollment**
    - **Inflation rate**
    - **2nd_sem_perf_ratio**
    - **Failed units ratio**
    - **Late_enrollment**
    - **Course_Graduate_prob**
    - **Application mode_Enrolled_prob**

1.5. Scoring Metric: “roc_auc_ovr”

- **Selected Features:**
    - **Tuition fees up to date**
    - **Curricular units 1st sem (grade):** Captures initial academic performance.
    - **Curricular units 2nd sem (grade)**
    - **2nd_sem_perf_ratio**
    - **Failed units ratio**

---

2. In-Depth Analysis of Feature Impact

2.1. Strongly Relevant Features Across Metrics

- **Tuition fees up to date:**
  Consistently selected across all metrics, underscoring the critical role of financial stability.

- **2nd_sem_perf_ratio and Failed units ratio:**
  Their repeated presence indicates that academic performance and failure rates are pivotal in predicting retention.

2.2. The Role of Parental and Background Factors

- **Mother's and Father's qualifications:**
  Their selection in the precision, recall, and f1 metrics highlights the influence of parental education and support.
- **Parental_influence:**
  Specifically chosen under the precision metric, this feature suggests that parental expectations or direct support significantly affect retention outcomes.

2.3. Enrollment & Course-Related Influences

- **Age at enrollment:**
  Consistently selected, indicating that the timing of enrollment is a vital predictor.
- **Late_enrollment:**
  Appearing in the recall, f1, and precision metrics, this feature flags the potential risks associated with delayed enrollment.
- **Course_Graduate_prob:**
  Serves as a direct indicator of the likelihood of course completion, reinforcing its importance in the retention model.

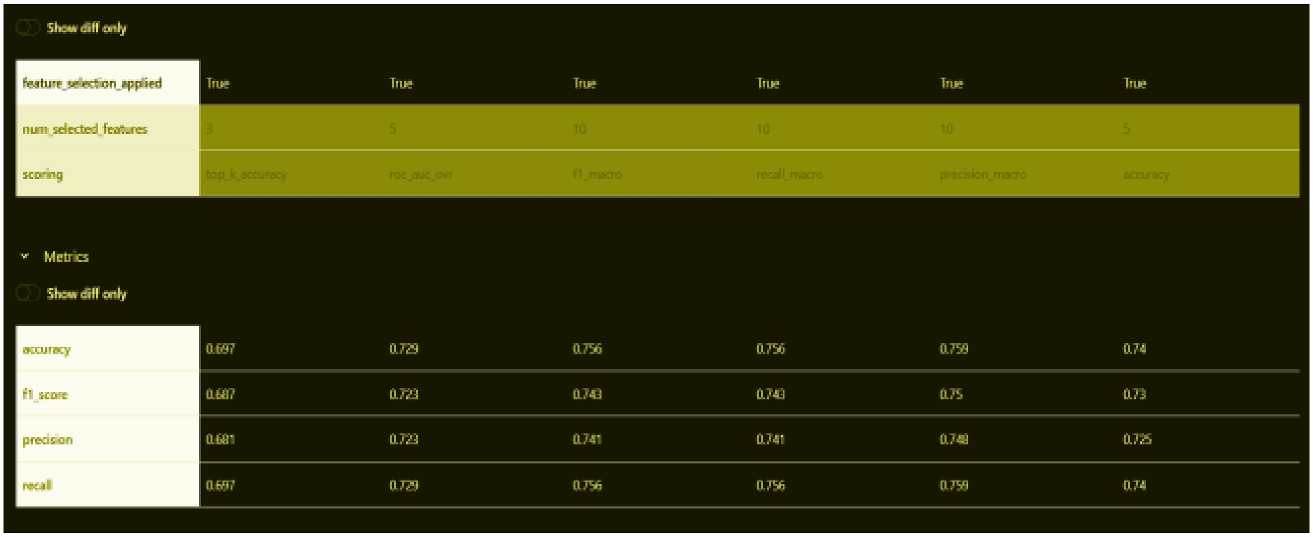
---

3. Additional Suggested Features (Not Present in the Dataset)

- **Engagements (e.g., watch time)**
- **Course Relevancy (Market Trend)**
- **Attendance**

These additional features could further enhance the model by incorporating behavioral and contextual factors.

MLFLOW TABLE



○ Show diff only

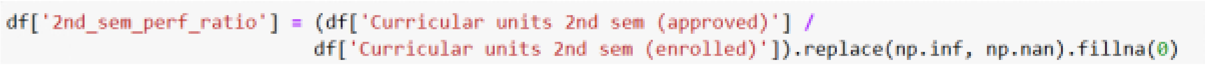
feature_selection_applied	True	True	True	True	True	True
num_selected_features	3	5	10	10	10	5
scoring	top_k_accuracy	roc_auc_ovr	f1_macro	recall_macro	precision_macro	accuracy

⌄ Metrics

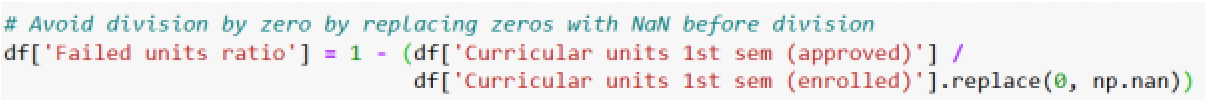
○ Show diff only

accuracy	0.697	0.729	0.756	0.756	0.759	0.74
f1_score	0.687	0.723	0.743	0.743	0.75	0.73
precision	0.681	0.723	0.741	0.741	0.748	0.725
recall	0.697	0.729	0.756	0.756	0.759	0.74

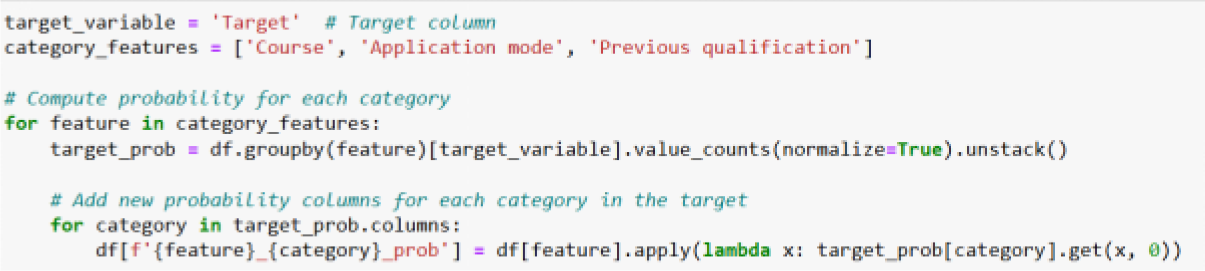
The definitions of some selected features which were derived from existing are as follows :



```
df['2nd_sem_perf_ratio'] = (df['Curricular units 2nd sem (approved)'] /
                            df['Curricular units 2nd sem (enrolled)']).replace(np.inf, np.nan).fillna(0)
```



```
# Avoid division by zero by replacing zeros with NaN before division
df['Failed units ratio'] = 1 - (df['Curricular units 1st sem (approved)'] /
                                df['Curricular units 1st sem (enrolled)'].replace(0, np.nan))
```



```
target_variable = 'Target'  # Target column
category_features = ['Course', 'Application mode', 'Previous qualification']

# Compute probability for each category
for feature in category_features:
    target_prob = df.groupby(feature)[target_variable].value_counts(normalize=True).unstack()

    # Add new probability columns for each category in the target
    for category in target_prob.columns:
        df[f'{feature}_{category}_prob'] = df[feature].apply(lambda x: target_prob[category].get(x, 0))
```

Link of dataset - https://zenodo.org/records/5777340#.Y7FJotJBwUE