Predictive Modeling Analysis for Academic Success Classification Model

**1. Data Preparation**

* **Features and Target**:

- Features were extracted by dropping ‘id’ and ‘Target’ columns from the training data.

- The ‘Target’ column was label-encoded to numeric values for compatibility with machine learning models.

* **Train-Test Split**:

- The dataset was split into training and validation sets with an 80-20 ratio.

- Stratified splitting ensured the class imbalance in the ‘Target’ variable was preserved in both sets.

**2. CatBoost Classifier Performance**

* **Accuracy** : **83.11%**
* **Classification Report** :

- **Precision**: Highest for class ‘0’ (90%) and lowest for class ‘1’ (66%).

- **Recall**: Strong recall for class ‘2’ (93%) but weaker for class ‘1’ (60%).

- **F1-Score**: Balanced scores across all classes, with class `2` performing the best (F1=0.89).

* **Overall Performance**:

- CatBoost effectively handles the class imbalance and achieves the highest accuracy among the tested models.

**3. XGBoost Classifier Performance**

* **Accuracy** : **82.97%**
* **Classification Report**:

- **Precision**: Similar to CatBoost, highest for class ‘0’ (89%) and lowest for class ‘1’ (65%).

- **Recall**: Strong recall for class ‘2’ (92%) but weaker for class ‘1’ (60%).

- **F1-Score**: Consistent across classes, slightly lower than CatBoost.

* **Overall Performance**:

- XGBoost performed well but slightly underperformed compared to CatBoost in terms of accuracy and F1-scores.

**4. HistGradientBoostingClassifier Performance**

* **Accuracy** : **82.79%**
* **Classification Report**:

- **Precision**: Similar to other models, with the highest for class ‘0’ (90%) and lowest for class ‘1’ (65%).

- **Recall**: Strong recall for class ‘2’ (92%), while class ‘1’ remains challenging (60%).

- **F1-Score**: Comparable to XGBoost, slightly trailing CatBoost.

* **Overall Performance**:

- The model demonstrates strong performance, but like XGBoost, it slightly underperformed compared to CatBoost.

**5. Selected Model: CatBoost**

* **Reason for Selection**:

- CatBoost achieved the highest accuracy (83.11%) and performed consistently across precision, recall, and F1-scores.

- Its ability to handle categorical variables and class imbalance makes it a robust choice for this dataset.

* **Test Set Predictions**:

- CatBoost predictions on the test data were label-inverted to restore the original class labels (‘Graduate’, ‘Dropout’, ‘Enrolled’).

- A sample of the predictions includes:

- id: 76518, Target: Dropout

- id: 76519, Target: Graduate

- id: 76522, Target: Enrolled

* **Output Saved**: Predictions were saved in a ‘submission.csv’ file.

**Final Insights**

* **Model Performance**:

- All models performed similarly, with accuracy around 83%, indicating robust modeling and strong features.

- Class ‘1’ (Dropout) remains challenging across all models, suggesting room for feature engineering or resampling strategies.

* **CatBoost as the Best Model**:

- It exhibited the best balance between precision and recall and delivered the highest accuracy.

- Its built-in handling of categorical data and fast computation further support its selection.

* **Future Improvements**:

- Addressing the class imbalance for ‘Dropout’ could improve model recall.

- Hyperparameter tuning for CatBoost, XGBoost, and HistGradientBoostingClassifier may yield further gains.

This analysis showcases the potential of gradient boosting models in predicting academic success, leveraging strong features like grades, course details, and socioeconomic factors.