Scientific Report: Predicting Students’ Final Grades in an Online Course

Nurul Amin

Introduction

The increasing shift towards online education platforms has created vast amounts of learner data, offering a rich source of insight into student performance. In this project, the goal is to leverage machine learning to predict students' final grades in an online course using data collected from quizzes, mini-projects, peer reviews, and interaction logs. By applying two supervised learning algorithms, we aim to determine which model better predicts final grades and analyze which features contribute most to the prediction.

Data Processing

The dataset contains anonymized information from 107 students enrolled in a nine-week online machine learning course hosted on Moodle. The dataset includes:

- **Grades** from 3 quizzes, 3 mini-projects, and 3 peer reviews.
- **Course interaction logs** categorized into:
  - **Status0**: course-related activities.
  - **Status1**: assignment-related activities.
  - **Status2**: grade-related actions.
  - **Status3**: forum-related actions.

After loading the data, I confirmed there were no missing values. Each student has 9 grades and 36 log-based features, along with the final grade, which serves as the target variable.

- **Features Chosen**: I retained all features except the "ID" column, which is irrelevant for prediction purposes.
- **Method of Choice**: Two supervised learning approaches were chosen for the task:
  1. **Linear Regression**: A simple model that provides interpretable results.
  2. **Random Forest Regressor**: An ensemble learning method that captures non-linear relationships and provides insights into feature importance.

Data was split into 80% for training and 20% for testing, ensuring that the models can generalize to unseen data.

Data Analysis

Model Performance:

Two models were trained on the data to predict students' final grades.

1. **Linear Regression**:
    - Mean Squared Error (MSE): 0.92
    - R-squared (R²): 0.78
2. **Random Forest Regressor**:
    - Mean Squared Error (MSE): 0.05
    - R-squared (R²): 0.99

Data Visualization:

To assess the models' performance, I plotted predicted versus actual final grades for both models:

- **Linear Regression**: The scatter plot revealed that while the predictions followed the general trend of actual grades, there was more deviation, especially at the extremes of the grade range.
- **Random Forest**: The Random Forest model predicted much closer to the actual values, with the data points almost perfectly aligned along the diagonal, indicating near-perfect prediction accuracy.

Interesting Observations:

- The **Random Forest** model significantly outperformed **Linear Regression**, with an almost perfect R² score of 0.99. This suggests that student grades depend on complex, non-linear relationships between features, which the ensemble-based Random Forest was able to capture.
- **Linear Regression**, while interpretable, struggled to capture the full complexity of the data, reflected in the relatively higher MSE and lower R² score.

Conclusion

This project demonstrates that predicting students' final grades using course interaction logs and assessment data is feasible with machine learning. However, several bottlenecks were encountered and resolved:

- **Model Selection**: Initially, Linear Regression was chosen for its simplicity, but it became clear that more sophisticated models like Random Forest were needed to capture the non-linear interactions between features. By employing a Random Forest Regressor, we achieved a much better performance.
- **Feature Importance**: One challenge was determining which features were most predictive. Random Forest provided a natural solution by offering feature importance. The three most important features identified were:
    1. Week 5 Mini Project 2 Grade (`Week5_MP2`).
    2. Week 7 Mini Project 3 Grade (`Week7_MP3`).
    3. Week 5 Peer Reviews 2 Grade (`Week5_PR2`).

These results suggest that students' performance on later assessments (especially the mini-projects) strongly correlates with their final grades, which could be indicative of cumulative learning throughout the course.

Improvements:

- **Model Complexity**: A more complex model like Random Forest worked well, but further improvements might be possible with techniques like boosting (e.g., XGBoost) or using deep learning models for very large datasets.
- **Cross-validation**: Adding cross-validation could provide more robust estimates of model performance and help avoid overfitting.

In conclusion, using machine learning models to predict student grades provides valuable insights that could aid educators in identifying students at risk of underperforming early in the course, allowing for targeted interventions and support.