

Predicting Academic Performance of NYU International Students Using Machine Learning and Visual Analytics

Ananya Agarwal, Naman Tyagi
CS-GY 9223 – Visualization for Machine Learning
Instructor: Prof. Claudio Silva

1 Problem Statement

International students often face distinct challenges compared to domestic students. These challenges include cultural adjustment, financial pressures, language barriers, visa restrictions, and limited access to support networks. Such factors can directly or indirectly influence academic performance, mental health, and overall success.

Traditional academic performance prediction models typically rely on simple features such as test scores and parental education but fail to capture the broader human and environmental factors that affect learning outcomes.

This project aims to predict academic performance among international students at NYU by analyzing demographic, socioeconomic, behavioral, and well-being factors using machine learning. By incorporating such variables alongside academic data, we hope to provide a holistic understanding of what drives success or struggle among NYU's international students. These insights can guide academic advisors, student support offices, and policymakers in designing personalized interventions that improve academic retention and well-being.

In addition, the project explores the impact of variables like sleep duration, work hours, financial aid, internet accessibility, homesickness, and use of support services. These variables are uniquely relevant to the international student experience and are often underrepresented in academic success prediction literature. Our aim is to create not only a predictive model but also a framework for visual diagnostics that institutions can use to proactively intervene and provide tailored support.

2 Data Collection

To ensure the model reflects the real-world context of New York University's international student community, we will gather primary data through a Google Form survey.

This form will collect authentic insights directly from international students at NYU, including:

- Academic habits
- Socioeconomic conditions
- Well-being and lifestyle habits
- Campus engagement

In addition to this newly collected data, we will integrate the Kaggle dataset titled “Students Performance in Exams” to align with university-level performance studies. Combining real-world and existing sources will help enhance the diversity, representativeness, and robustness of our model.

All data will undergo preprocessing including:

- Cleaning and normalization
- Encoding categorical variables
- Imputing missing values

Data Types

- **Categorical:** Gender, Visa type, Financial support, Major, Internet reliability
- **Ordinal:** Stress level (1–10), Social connectedness (1–10), Homesickness frequency
- **Numerical:** GPA, Study hours, Sleep hours, Work hours, Commute time

3 Machine Learning Models

To predict academic performance, we will use a combination of linear, regularized, and ensemble models:

- Linear Regression
- Ridge and Elastic Net Regression
- Random Forest Regressor
- XGBoost and CatBoost
- Stacking Ensemble
- K-Means Clustering (for unsupervised profiling)

These models provide both interpretability and predictive power. Stacking will allow us to combine multiple algorithms for optimal performance. Clustering techniques will help identify distinct student profiles based on non-academic attributes.

4 Evaluation Metrics

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- R^2 score and Adjusted R^2
- k-Fold Cross-Validation
- Residual analysis and learning/validation curves

These metrics ensure robustness, interpretability, and generalizability of the model. By combining traditional metrics with diagnostic plots, we will monitor the model’s performance in terms of both underfitting and overfitting.

5 Visualization Approach

Since this is a Visualization for ML course, we will emphasize explainability and insight generation through multiple visualization layers.

Tools

- **SHAP:** Feature attribution and global importance
- **LIME:** Local interpretability for individual predictions
- **D3.js:** Interactive browser-based visualizations
- **Seaborn / Plotly:** Exploratory data analysis and static visuals

Planned Visualizations

Visualization	Purpose
SHAP Summary Plot	Global feature importance
LIME Explainers	Local student-level interpretability
Radar Charts	Compare well-being and academic indicators
UMAP / t-SNE Projections	Cluster behavior and risk profiles
D3.js Dashboard	Interactive filtering by GPA, stress, sleep, etc.

Each visualization will offer tooltips, filters, and interactive controls to aid advisors in exploring high-risk students and identifying key intervention areas.

6 Contributions Beyond Prior Work

Previous studies using similar datasets (like the Kaggle one) primarily:

- Focused on high-school or general student populations
- Used limited academic and demographic features
- Relied on simple models with little interpretability

Our project provides the following novel contributions:

1. Introduction of behavioral, well-being, and visa-related features into prediction
2. Custom survey targeting NYU international students
3. Use of ensemble and stacked models for improved generalization
4. Explainability using SHAP and LIME
5. Interactive visualization layer tailored for university use

Conclusion

This project blends predictive modeling with visual storytelling to better understand and support NYU’s international student population. Our framework aims to provide not only accurate predictions of GPA but also a human-centered interface that highlights actionable insights. The long-term vision is to enable NYU’s support staff to identify risk factors early, customize interventions, and improve student success in both academic and well-being domains.