Predicting Student Dropout and Academic Success

EMPOWERING EDUCATIONAL INSTITUTIONS WITH PREDICTIVE INSIGHTS

Hemanth Rayudu, Graduate Student

December 3, 2024

Northeastern University

MOTIVATION

Why This Project?

- High dropout rates negatively impact institutions, students, and society.Early identification can save resources and improve student outcomes.

Relevance:

- Globally, dropout rates remain a challenge across educational institutions.
- This project highlights actionable insights to tackle these issues.

Impact:

- Helps institutions improve graduation rates and provide targeted support.

LVX VERITAS VIRTVS Northeastern University

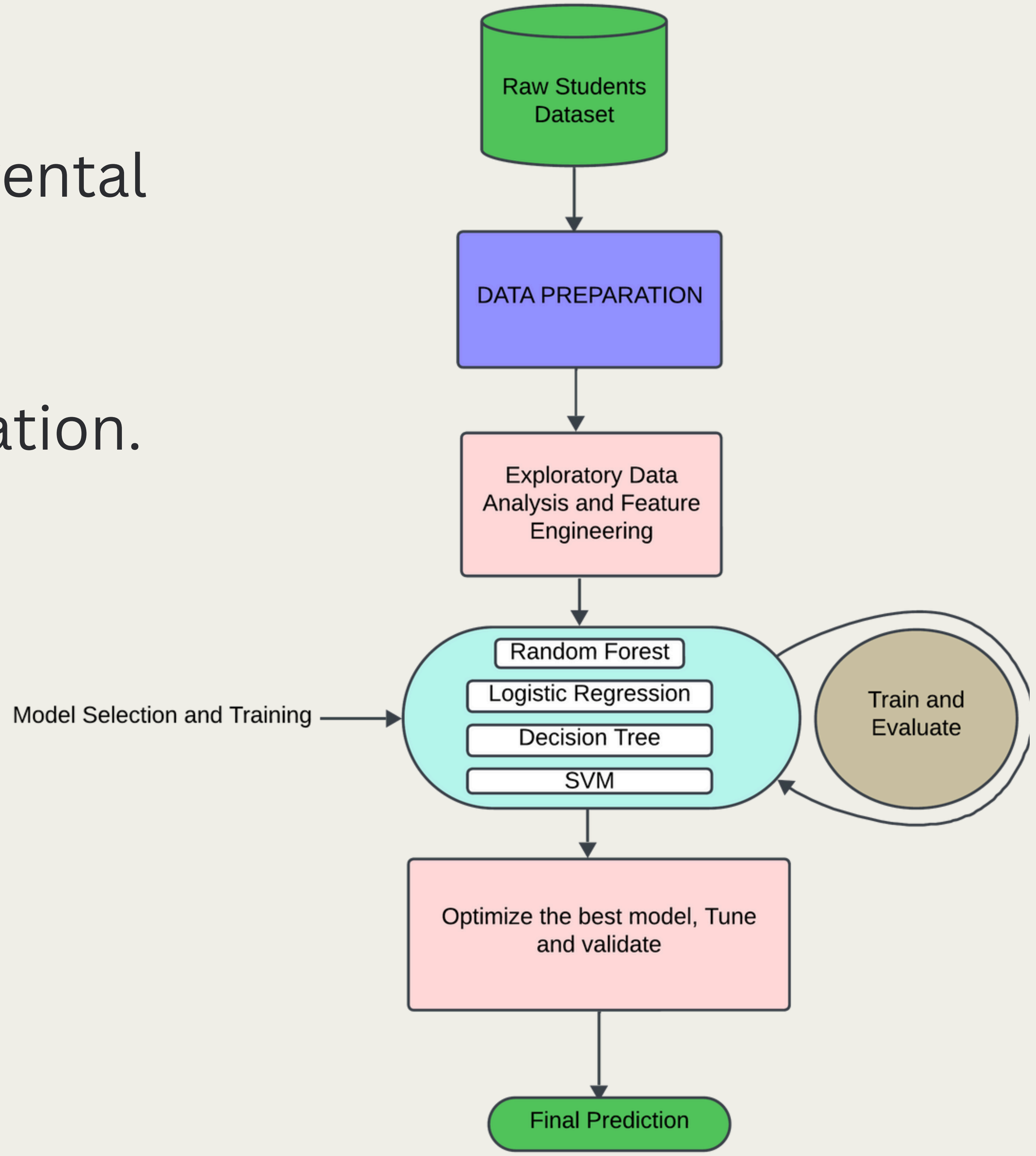
PROJECT OVERVIEW

Dataset:

- Source: UC Irvine Machine Learning Repository.
- Size: 4,424 records, 36 features (e.g., grades, parental education, socio-economic factors).

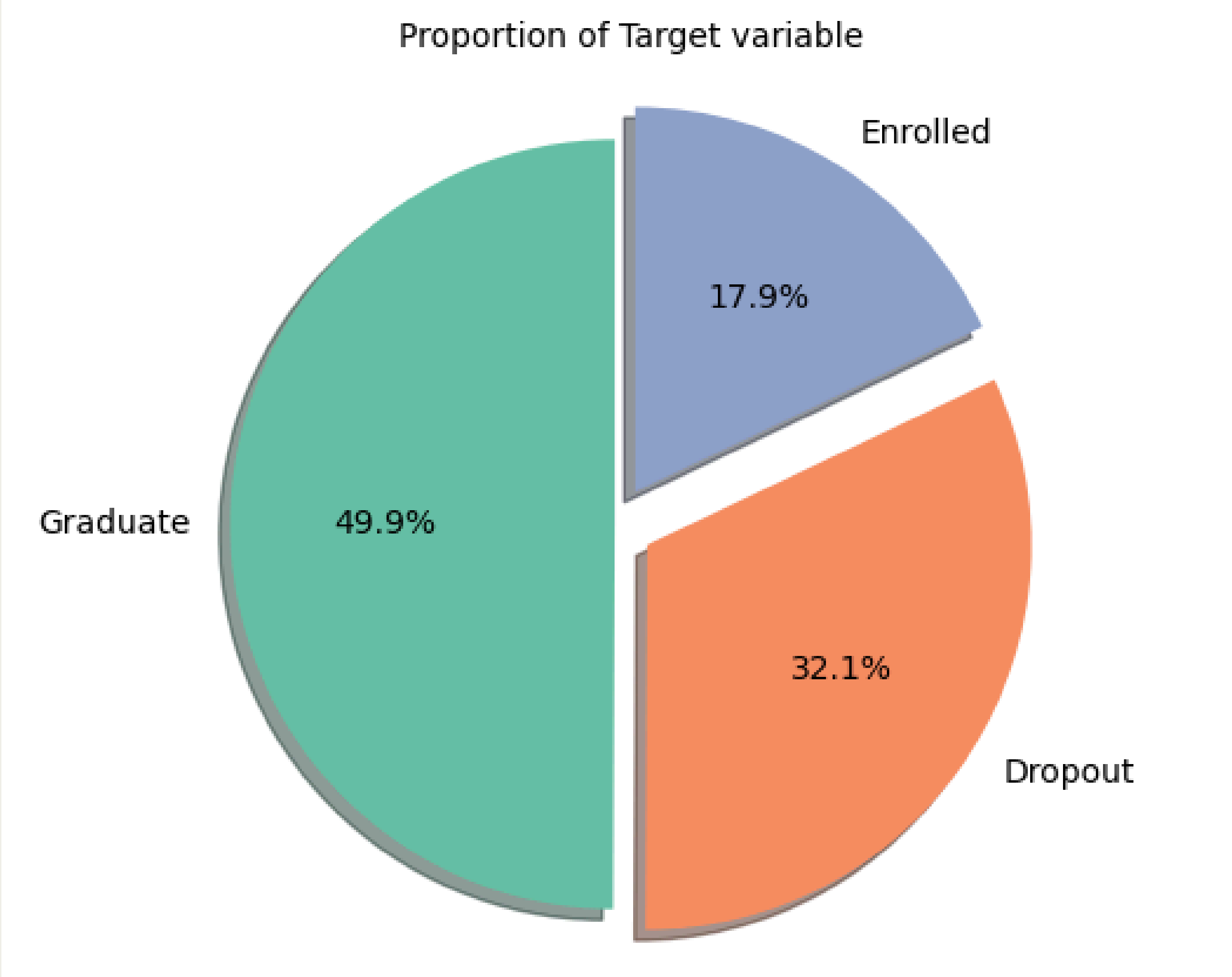
Languages and Tools:

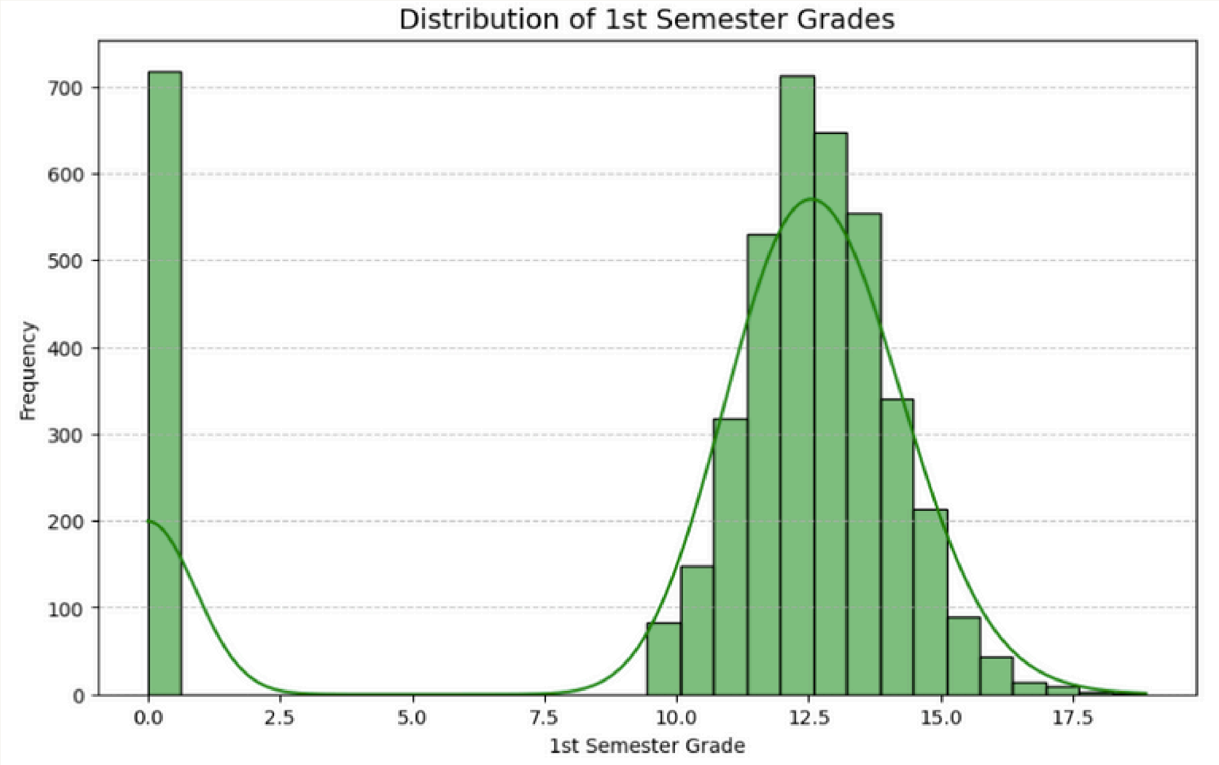
- Python for data analysis, modeling, and visualization.
- Jupyter Notebook for step-by-step execution.
- Libraries: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, Imbalanced-learn.
- Techniques: EDA, Feature Selection, PCA, SMOTE.
- Models: Random Forest, Logistic Regression, SVM, Decision Tree.

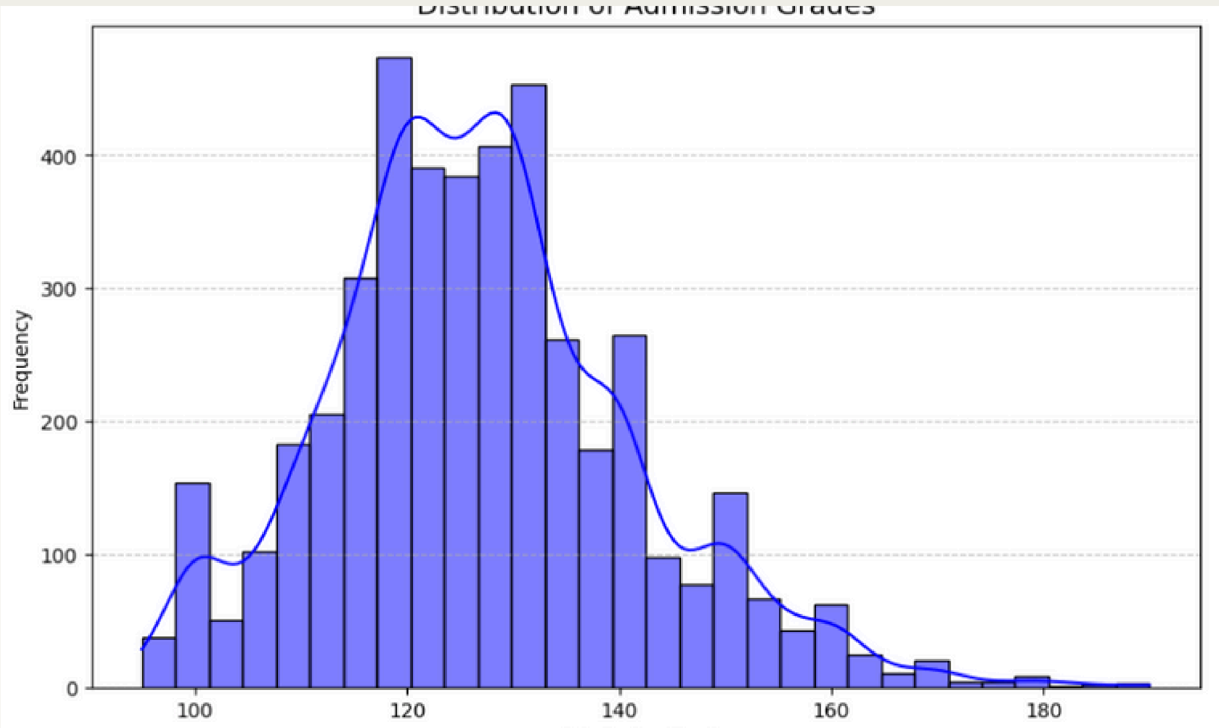

Raw Students Dataset
DATA PREPARATION
Exploratory Data Analysis and Feature Engineering
Model Selection and Training
Random Forest
Logistic Regression
Decision Tree
SVM
Train and Evaluate
Optimize the best model, Tune and validate
Final Prediction

EXPLORATORY DATA ANALYSIS

- The Pie-Chart indicates the need for targeted support for at-risk groups.







- Most students score between specific ranges, and outliers exist in grades.

HEATMAP AND CORRELATION ANALYSIS

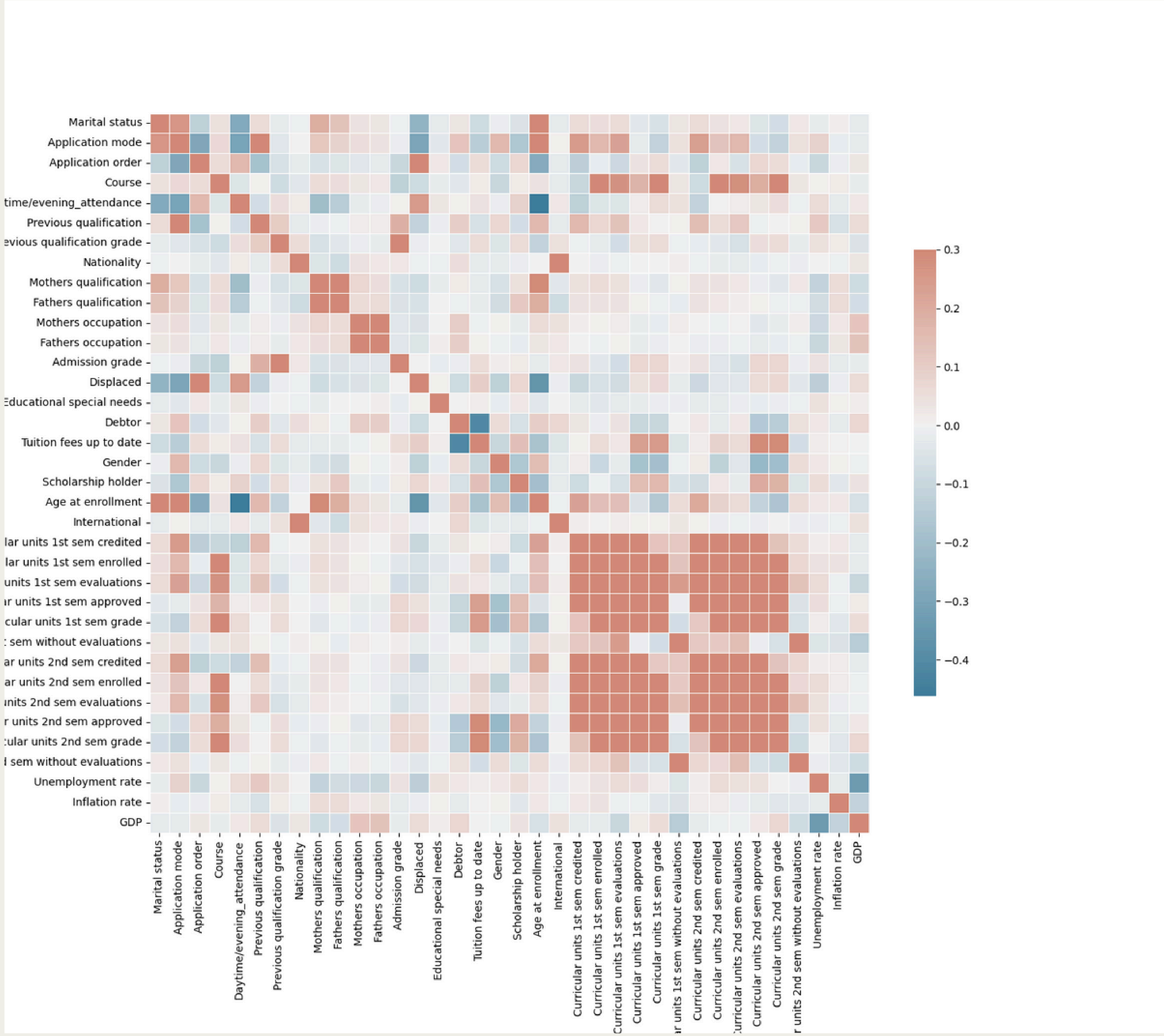
- The correlation matrix shows that most features are strongly positively correlated (bright orange), a few are strongly negatively correlated (blue), and only a small number have weak positive or negative correlations.

Strong correlations:

- Semester grades with dropout likelihood.
- Tuition fees and timely payments with graduation rates.

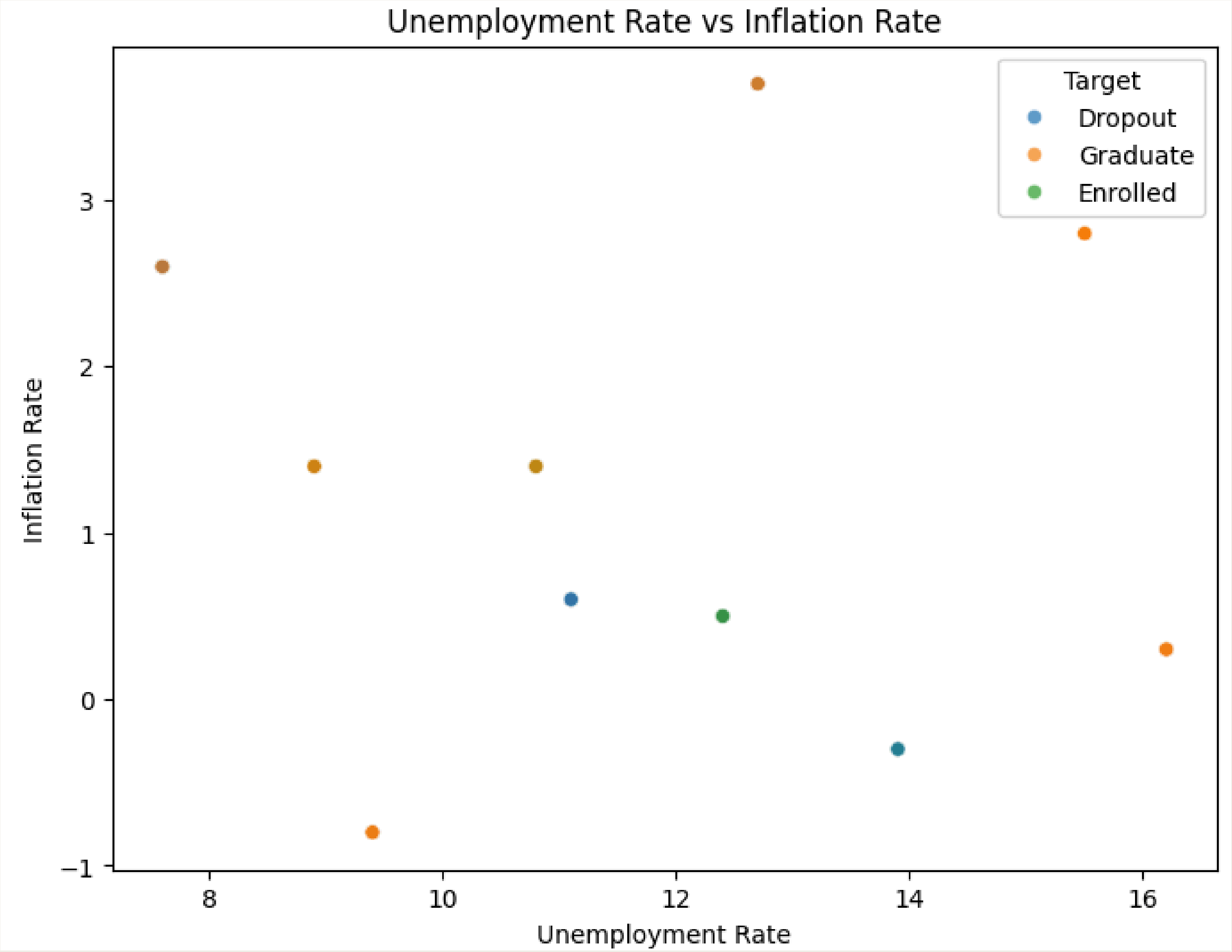
Weak correlations:

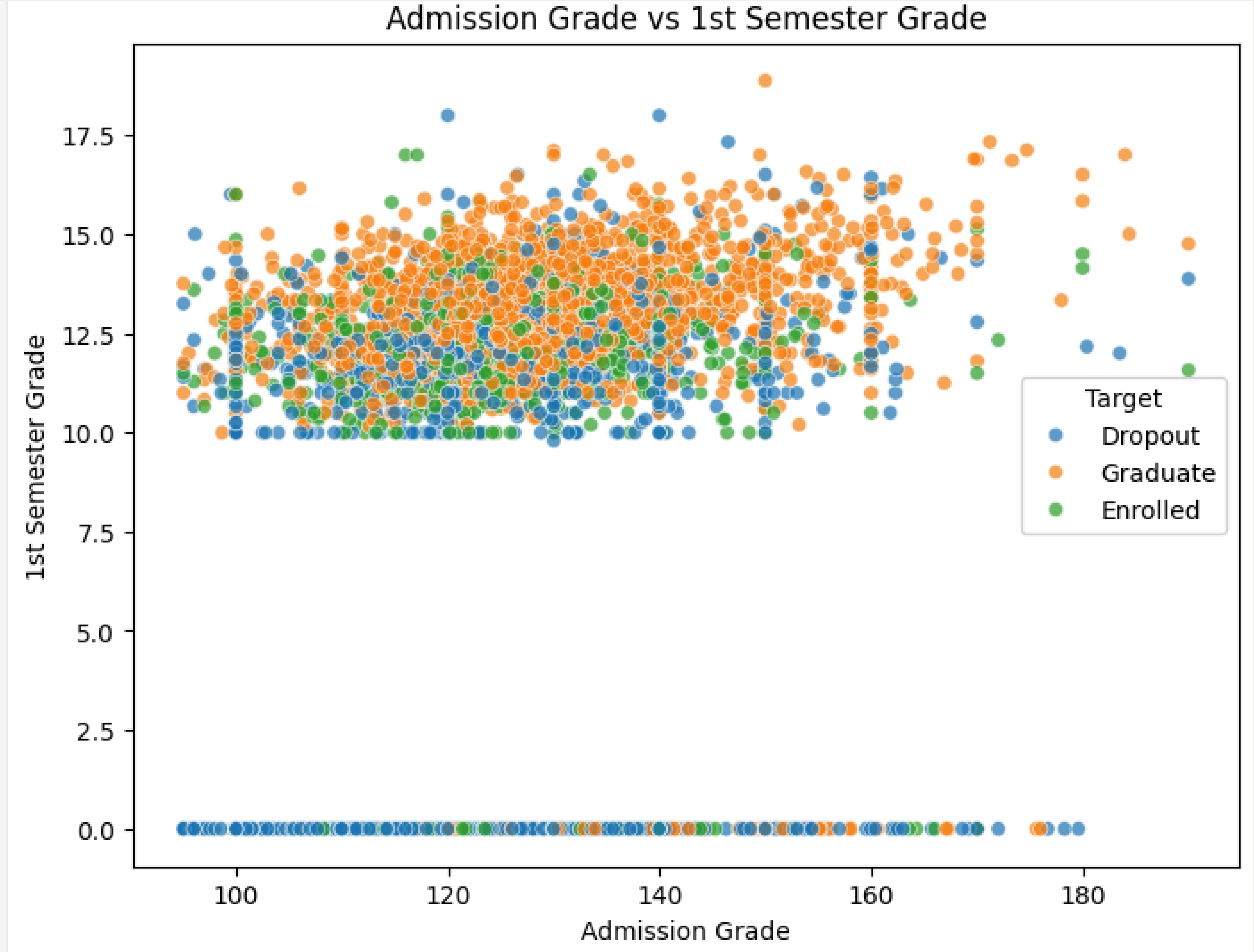
- Age of enrollment and student outcomes.



SCATTER PLOTS AND RELATIONSHIPS

- Students with high admission grades tend to perform well in the first semester.



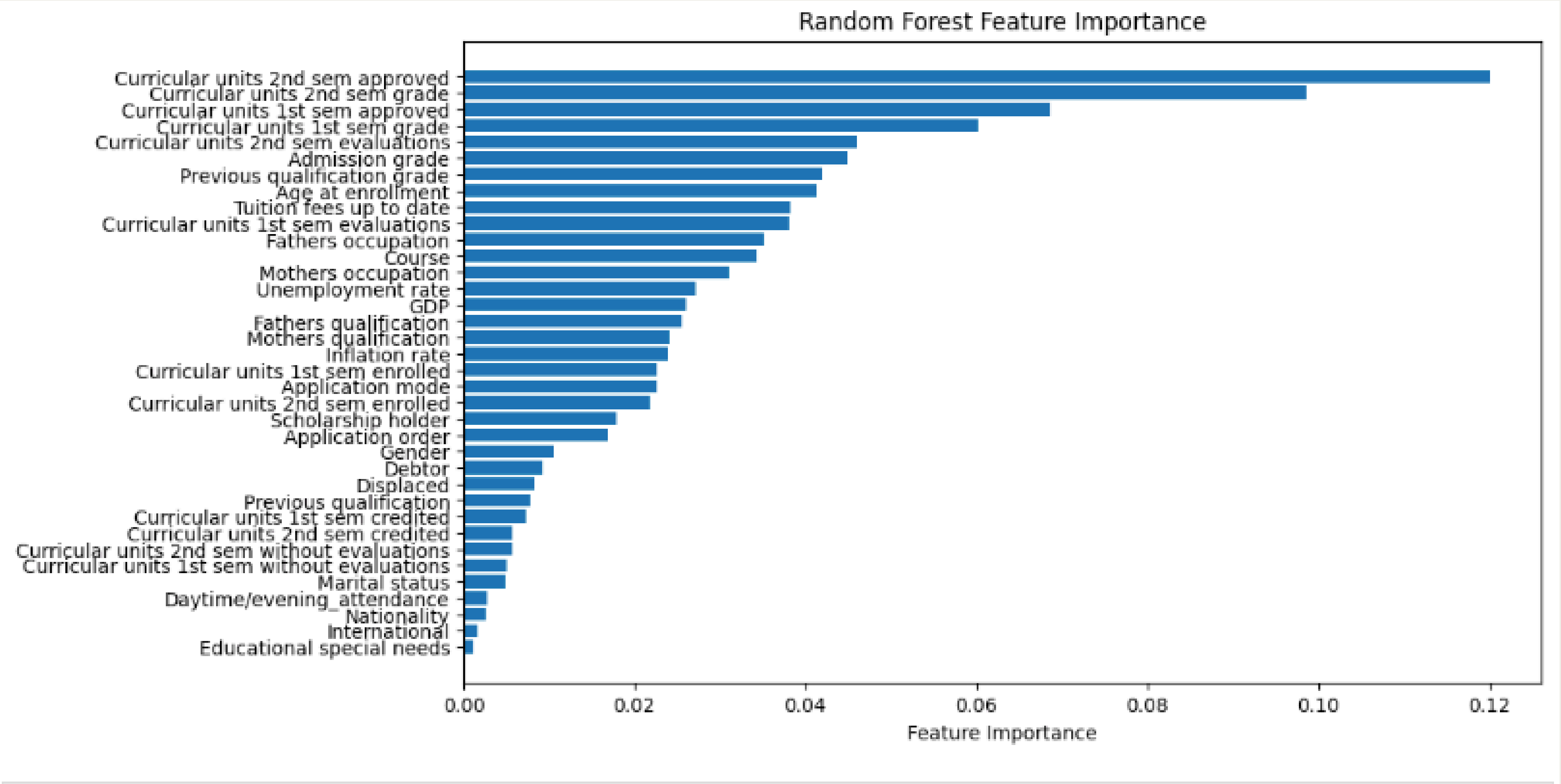


- High unemployment motivates students to stay in school longer.

RESULTS FROM FEATURE ENGINEERING

Key Predictors Identified:

- Curricular units 2nd sem approved: The most significant predictor of student success, emphasizing the importance of second-semester performance.
- Curricular units 2nd sem grade and 1st sem grade: Highlight the impact of academic consistency.
- Tuition fees up to date: Indicates that timely payment of fees correlates with better academic outcomes.



- Curricular units (1st and 2nd semesters) dominate the importance metrics.
- Financial factors such as tuition fees and scholarships also contribute significantly
- Features like Admission grade and Application mode also contribute moderately to predictions.
- Factors such as Nationality and Educational special needs have minimal impact on the predictions.

CONCLUSION

- “Curricular units 2nd sem approved" (Score: 2110.25) and "1st sem grade" (Score: 973.41) were identified as the most influential features for predicting student success.

Model Performance:

- The **Random Forest mode**l outperformed others with:Accuracy of **83%** and F1 Score of **84%.**

Data Balancing:

- SMOTE oversampling balanced the dataset, ensuring 2209 samples per class, which improved the reliability and fairness of predictions.
- Students with admission grades > 140 were more likely to graduate.
- Students with admission grades < 120 had higher dropout rates.

To prevent dropouts, institutions should offer personalized academic support, financial aid, and counseling services, while leveraging predictive analytics to identify and assist at-risk students early. Engaging students through mentorship, community-building activities, and redesigned courses can enhance retention and success rates.

References:

- UC Irvine Machine Learning Repository (Dataset): https://archive.ics.uci.edu/ml/index.php
- Scikit-learn Documentation (Machine Learning): https://scikit-learn.org/stable/documentation.html
- Imbalanced-learn Documentation (SMOTE): https://imbalanced-learn.org/stable/
- Matplotlib and Seaborn (Visualization): https://matplotlib.org/ & https://seaborn.pydata.org/

Thank you!

Hemanth Rayudu
Gradute Student, MSIS
Northeastern University
December 3, 2024

LVX VERITAS VIRTVS
Northeastern University