Massachusetts Institute of Technology

Predicting Student Performance for Targeted Intervention

15.072: Advanced Analytics Edge

Professor Rahul Mazumder

N. Nerur, B. Olafsson, J. Pinter, A. Xie, & Y. Yu

Massachusetts Institute of Technology

Meet Jimmy

A high school junior at risk of failing

- **Jimmy's Challenges:**
  - Struggles to focus during lectures and falls behind in class
  - Lacks access to tutoring and a quiet study space due to financial difficulties
- **Impact on his Academic Performance:**
  - Struggles with foundational concepts, especially in math
  - Dreams of attending college, a dream difficult to realize without support
- **Relevance to our Project:**
  - Identify key factors in academic success and struggles like Jimmy's
  - Guide interventions for support – tutoring, counseling, additional resources
  - Focused support for students like Jimmy to succeed

Existing Methods

- **Traditional Methods:**
  - Descriptive statistics and regression analysis
  - Identify correlations between variables and academic outcomes
  - Limitations: Overlook non-linear relationships and complex patterns
- **Recent Advancements:**
  - Machine learning algorithms show promise e.g., Naïve Bayes and K-Nearest Neighbor
  - Enhanced prediction accuracy on educational datasets
- **Limitations of Current Approaches:**
  - Data challenges: Limited sample sizes, data acquisition difficulties
  - Methodological issues: Reliance on few algorithms, single-point data collection
  - Interpretability concerns: Complex models sacrifice interpretability

Datasets – Original & Augmented

Kaggle dataset

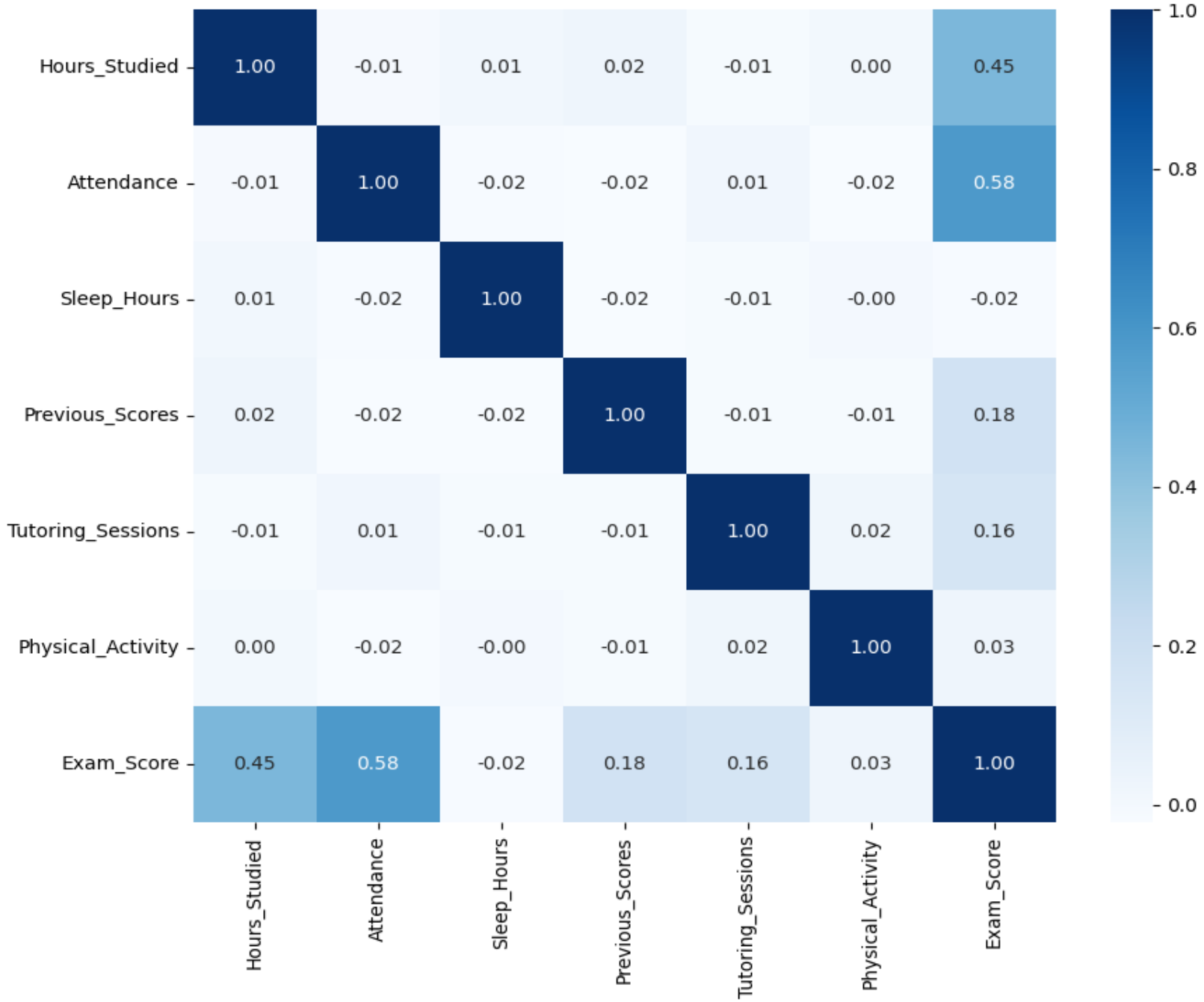
- **Target Variable:**
  - Final exam score (Regression)
- **Feature groups:**
  - Classroom discipline: Hours_Studied, Attendance
  - Home environment: Parental_Involvement, Family_Income, Internet_Access
  - Extracurricular: Extracurricular_Activities, Physical_Activity

- **Data Augmentation**
  - Balanced class distribution using Synthetic Minority Oversampling Technique (SMOTE)
  - Generate new datapoints that resemble distribution of existing data with Gaussian Mixture Models (GMMs)

Massachusetts Institute of Technology

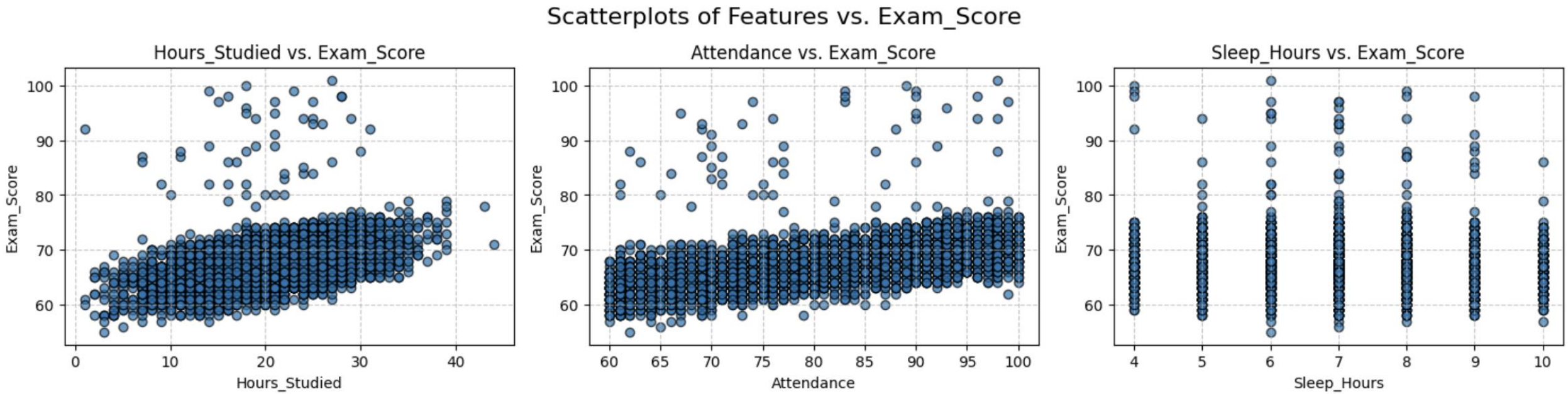
Feature Engineering

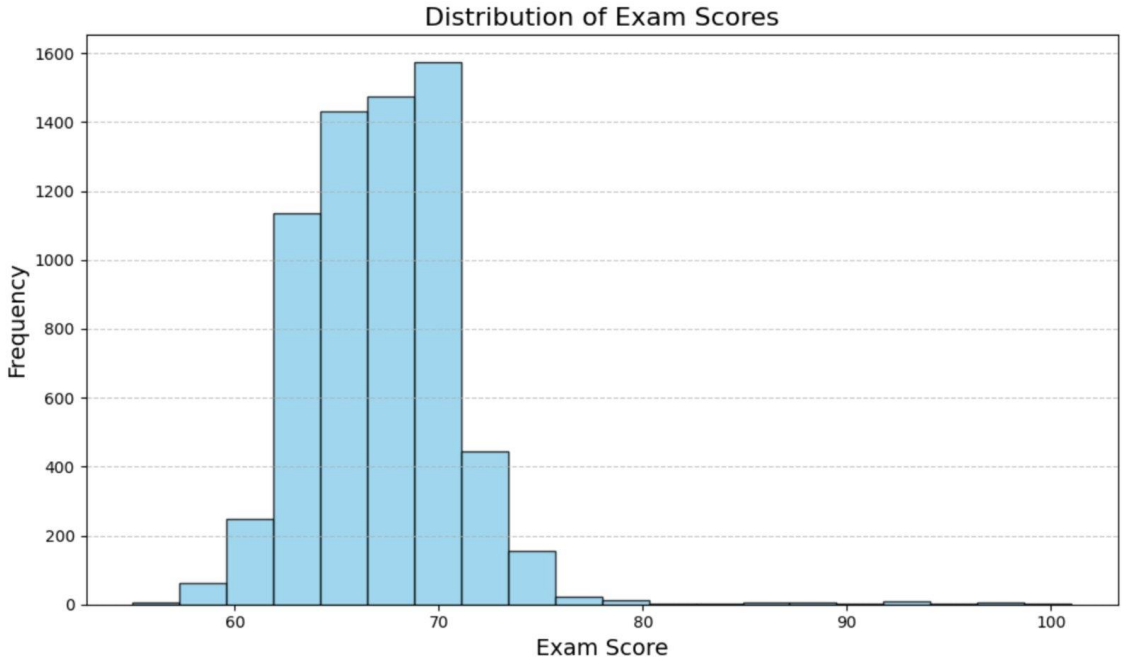
- **Class Engagement Interaction**
  - Hours Studied * Attendance
- **Home Support**
  - Parental Involvement * Access to Resources
- **Learning Engagement**
  - Motivation Level * Tutoring Sessions
- **Study Sleep Ratio**
  - Hours Studied / Sleep Hours
- **Health Balance**
  - Physical Activity * Sleep Hours


Correlation Matrix for Numeric Features
Hours_Studied
Attendance
Sleep_Hours
Previous_Scores
Tutoring_Sessions
Physical_Activity
Exam_Score
1.00 -0.01 0.01 0.02 -0.01 0.00 0.45
-0.01 1.00 -0.02 -0.02 0.01 -0.02 0.58
0.01 -0.02 1.00 -0.02 -0.01 -0.00 -0.02
0.02 -0.02 -0.02 1.00 -0.01 -0.01 0.18
-0.01 0.01 -0.01 -0.01 1.00 0.02 0.16
0.00 -0.02 -0.00 -0.01 0.02 1.00 0.03
0.45 0.58 -0.02 0.18 0.16 0.03 1.00
1.0
0.8
0.6
0.4
0.2
0.0

Massachusetts Institute of Technology

Feature Relationship with Target





Bin	Count
Excellent	2131
Good	1625
Average	1468
Low	1383

Regression

Classification

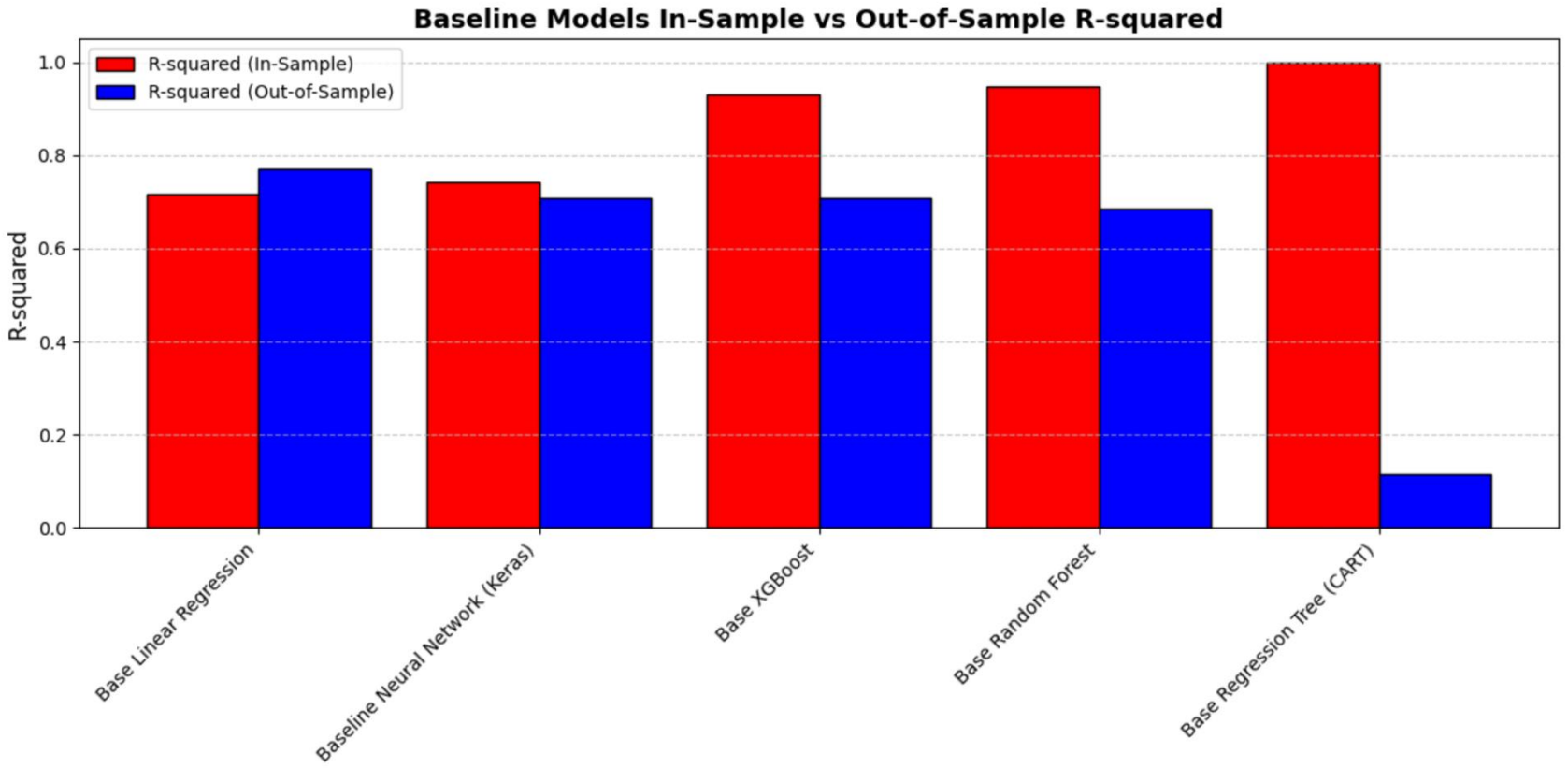
Massachusetts Institute of Technology

Regression Models Overview

Model Groups	Interpretability	Drawbacks
Linear Models (Linear Regression, Lasso L1 Regularization, Ridge L2 Regularization, Elastic Net L1-L2)	• Coefficients have direct and intuitive interpretations. • Easy to explain to non-technical stakeholders.	• Assumes linear relationships, which may not capture complex patterns. • Sensitive to outliers and multicollinearity.
Individual **Regression Decision Tree**	• Provides visual explanations through the tree structure. • Easy to interpret feature splits and decision rules.	• Prone to overfitting without pruning or regularization. • Less stable (small changes in data can lead to different trees).
Ensemble Black Box Methods (Random Forest, Gradient Boosting, Neural Networks)	• Difficult to interpret as they combine multiple models. • Techniques like SHAP or feature importance can help explain predictions.	• Computationally intensive. • Require more tuning and may lack transparency for decision-making.

Massachusetts Institute of Technology

Regression: Baseline Results



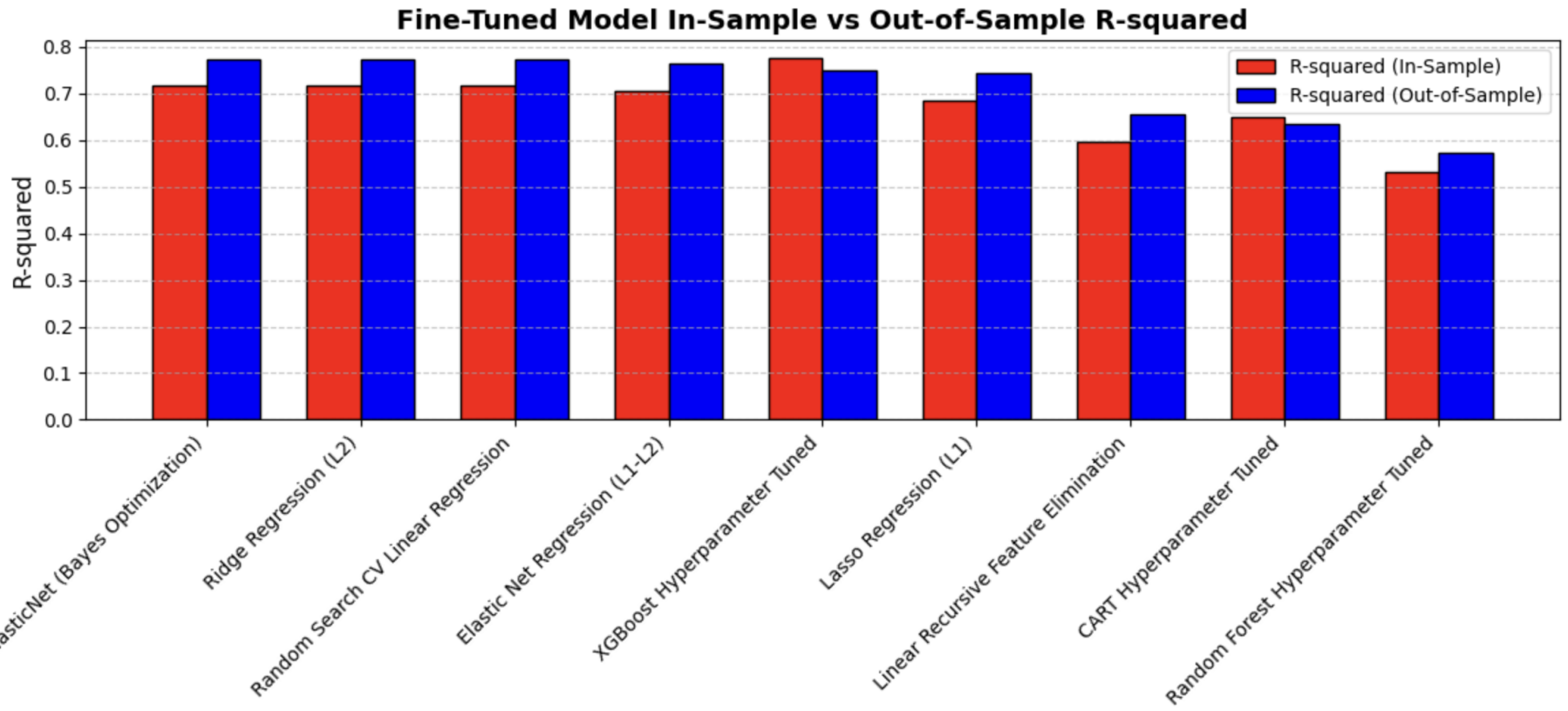
Best Baseline Model:
Linear Regression

Out-of-Sample
R-Squared:

77.11%

Massachusetts Institute of Technology

Regression: Hyperparameter Tuning



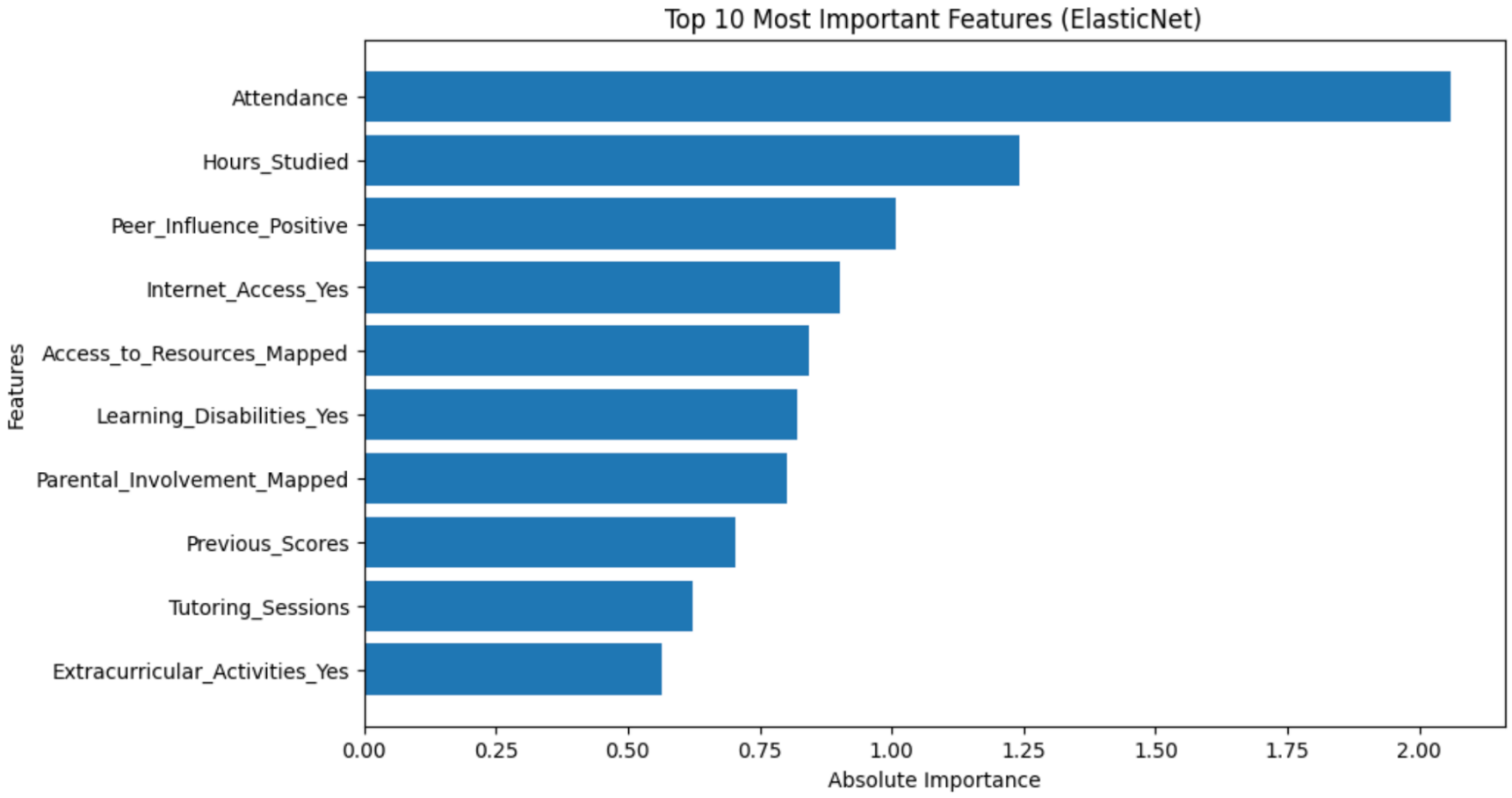
Best Model: Elastic Net

Out-of-Sample R-Squared:

77.14%

Massachusetts Institute of Technology

ElasticNet: Key Features



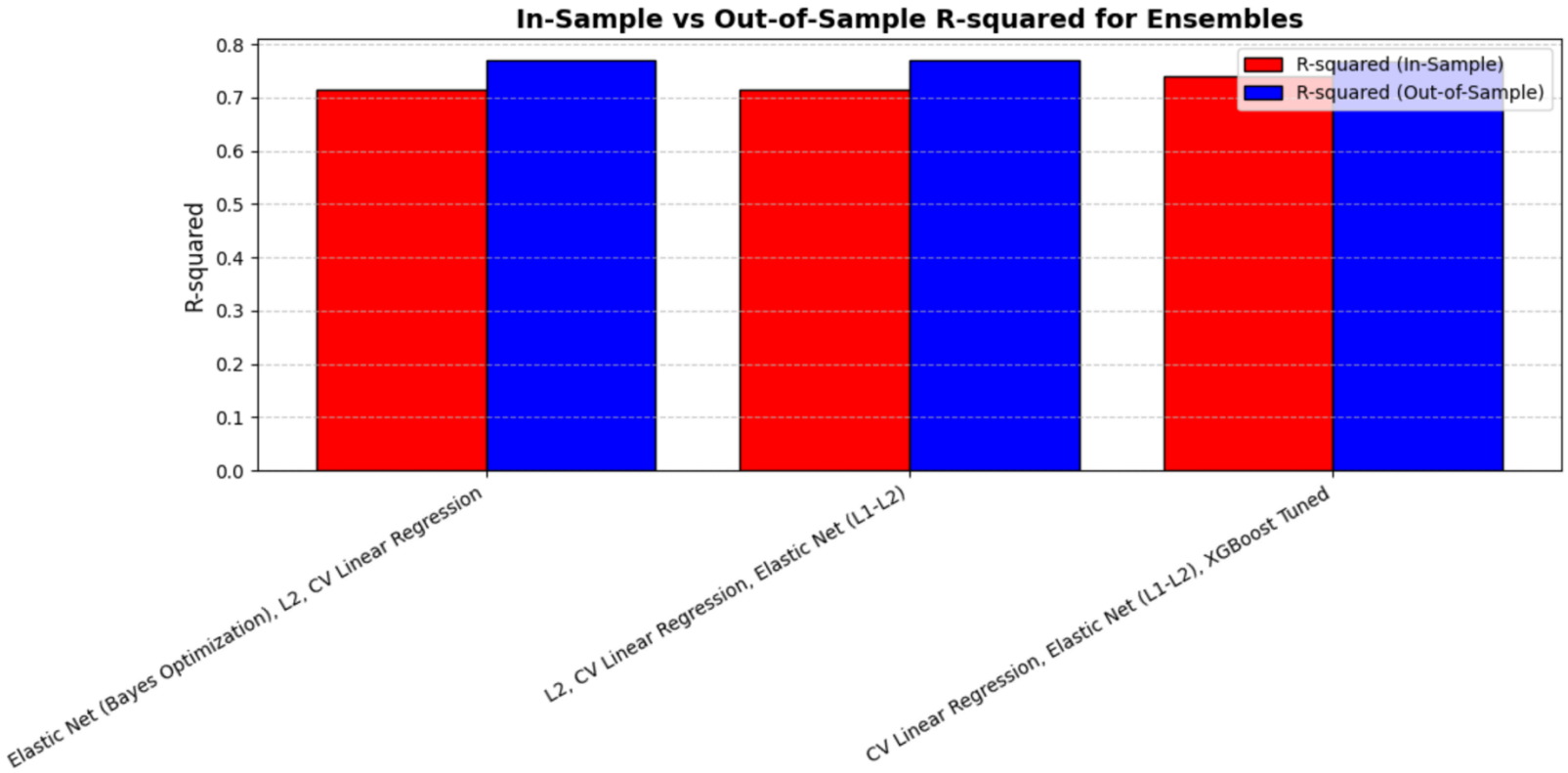
Regression: Hyperparameter Tuning

Bayesian Optimization: Uses probabilistic models to predict the performance of hyperparameters and selects the most promising ones to test

Best Hyperparameters

- **Best α ≈ 0.004**
  - Controls the overall strength of regularization. A smaller α indicates weaker regularization, meaning the model relies more on the data itself and less on regularization penalties
  - Makes sense because linear regression performs very well already without tuning
- **Best L1 Ratio (ρ) ≈ 0.5**
  - Specifies the balance between L1 (Lasso) and L2 (Ridge) regularization
  - The model is applying an equal mix of Lasso (sparsity) and Ridge (shrinkage)

Regression: Ensembling (Tuned Models)



Best Model: Elastic Net, Ridge, and Linear Regression

Out-of-Sample R-Squared: 77.10%

Ensembling is unnecessary

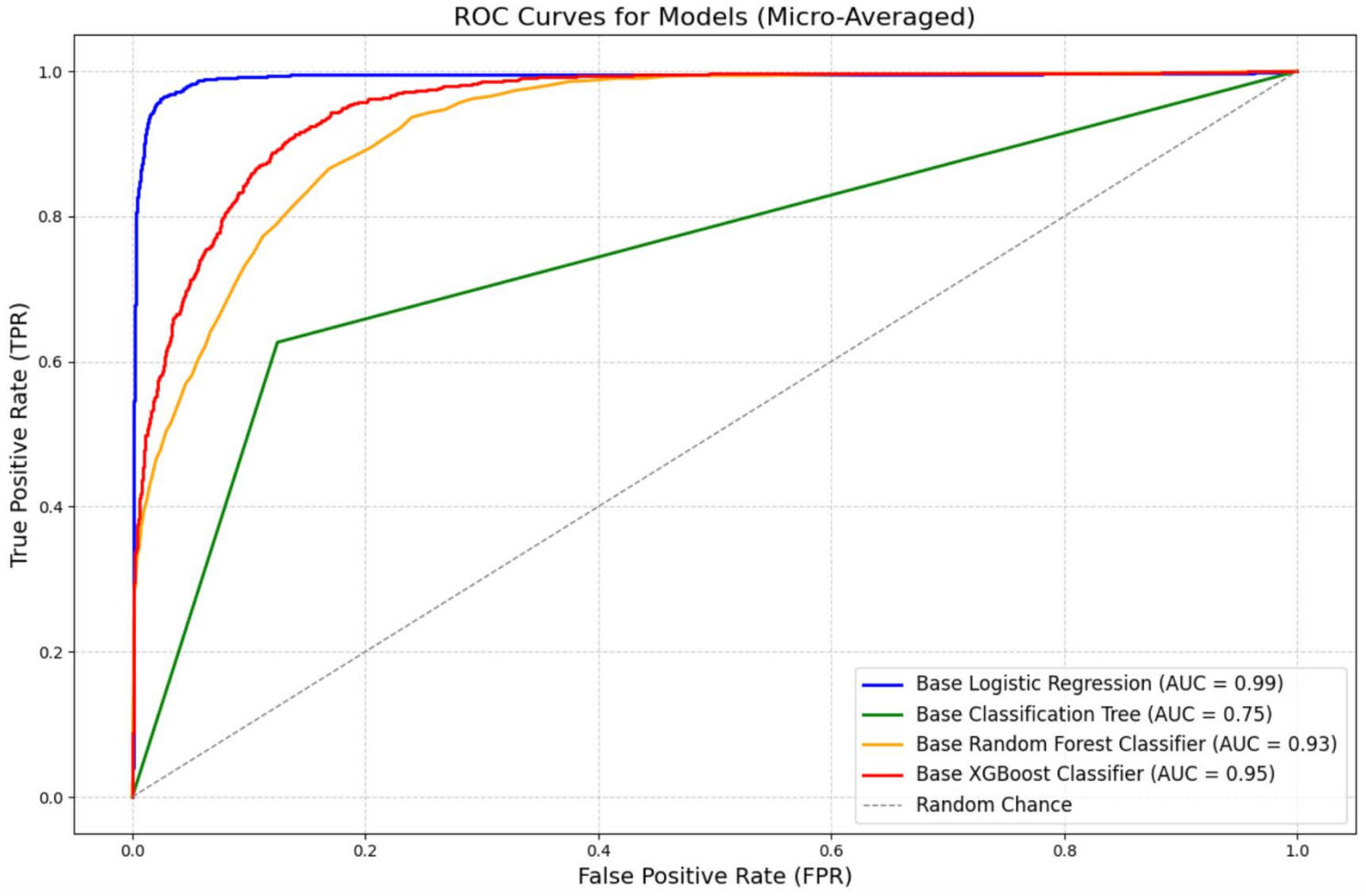
Massachusetts Institute of Technology

Classification Models Overview

Model Groups	Interpretability	Drawbacks
Linear Models (Logistic Regression)	- Coefficients provide intuitive, direct interpretations, easy to interpret	- Assumes linear boundaries between classes, sensitive to outliers.
Tree-Based Models (CART, Random Forest, XGBoost)	- Visualizations enable interpretability - Feature importance rankings aid understanding	- CART are prone to overfitting w/o pruning; Ensembles are less interpretable and computationally expensive.
Instance-Based Models (K-Nearest Neighbors)	- Simple, intuitive approach - No assumptions about distribution	- Computationally intensive - Sensitive to feature scaling and irrelevant features
Kernel-Based Models (Support Vector Machine)	- Effective for both linear and non-linear decision boundaries - Robust to high-dimensional datasets	- Difficult to interpret - Requires careful parameter tuning for optimal performance
Neural Network Models	- Flexible and capable of capturing complex non-linear relationships	- Black-box nature limits interpretability - significant computational resources and prone to avoid overfitting

Massachusetts Institute of Technology

Classification: Baseline Results

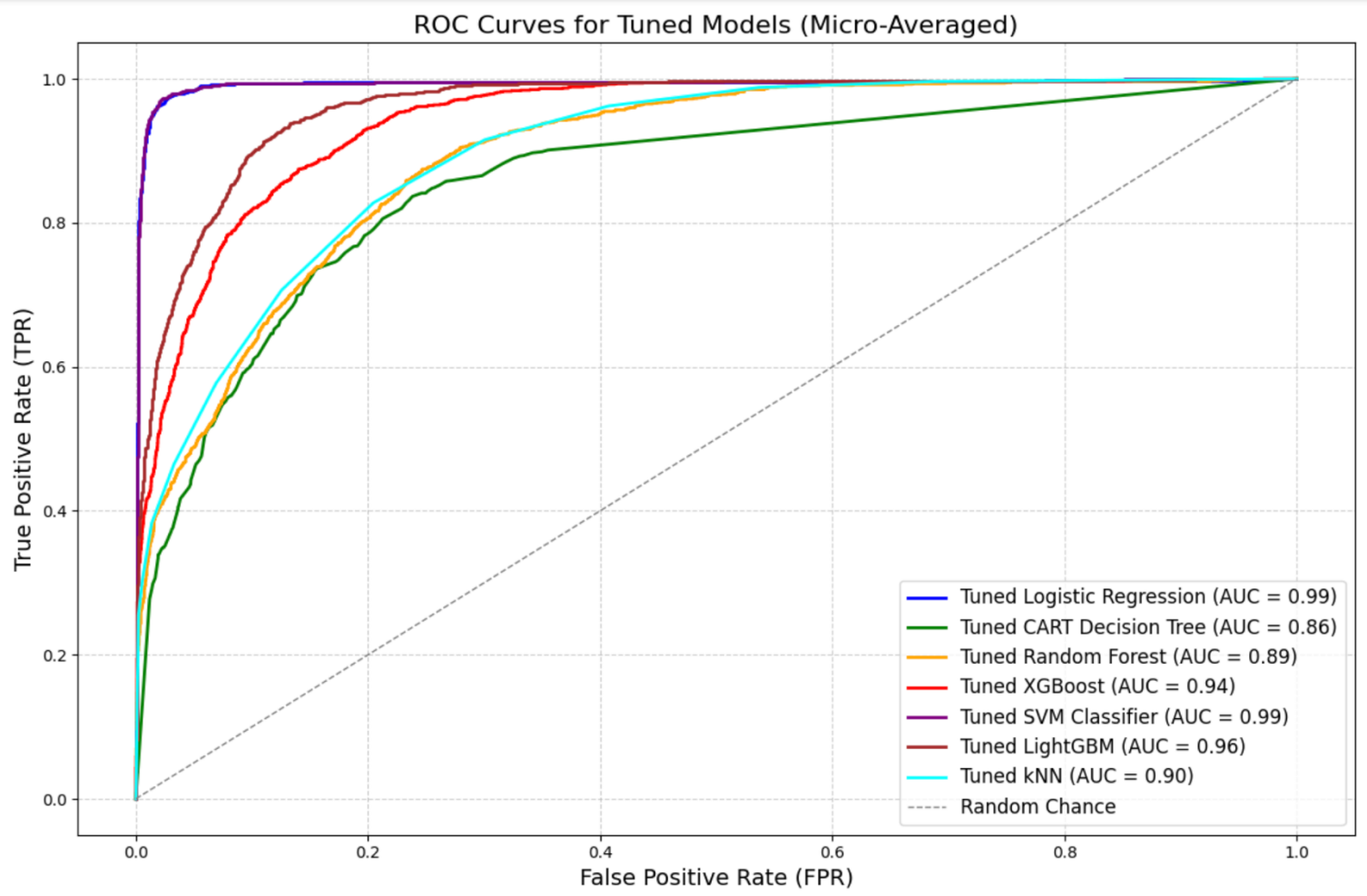


Key Insights

Best Baseline Model: Logit

Massachusetts Institute of Technology

Classification: Tuned Results



Key Insights

a clear improvement in model performance after fine-tuning.

Best Tuned Models:
SVC >= Logit
(achieving near-optimal AUC scores of 0.99)

Massachusetts Institute of Technology

Classification Recommendation

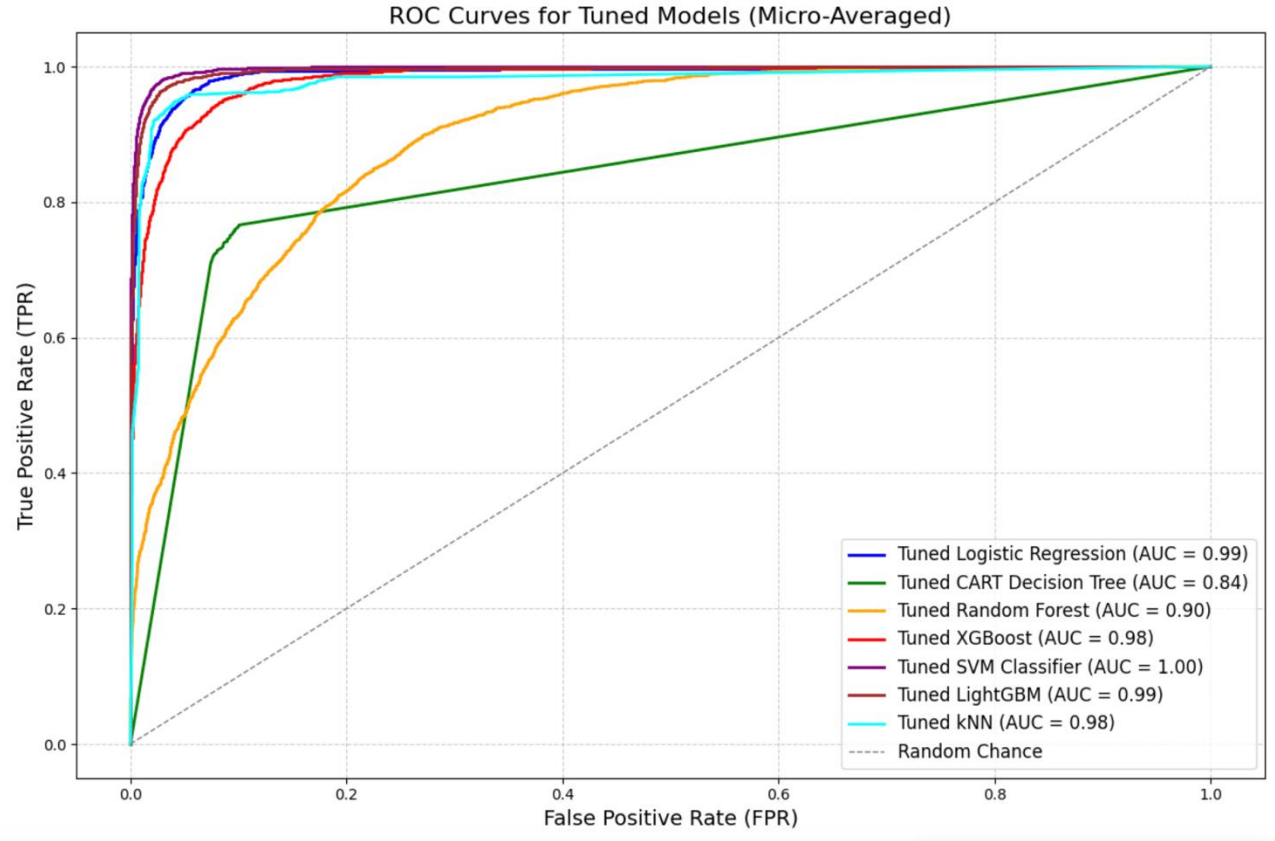
Table II: Fine-Tuned Model Performance Metrics

Model	Accuracy (In-Sample)	Accuracy (OOS)	F1 (In-Sample)	F1 (OOS)
Support Vector Machine	0.962	0.963	0.962	0.963
Logistic Regression	0.952	0.955	0.952	0.955
Neural Network	0.964	0.937	0.964	0.938
XGBoost	0.996	0.870	0.996	0.871
Random Forest	0.997	0.759	0.997	0.756
Naive Bayes	0.679	0.695	0.683	0.700
K-Nearest Neighbors	1.000	0.685	1.000	0.688
CART	0.855	0.655	0.854	0.655

We recommend Logistic Regression as the primary model, offering a strong balance between performance and interpretability!

Massachusetts Institute of Technology

Using Augmented Data for Classification

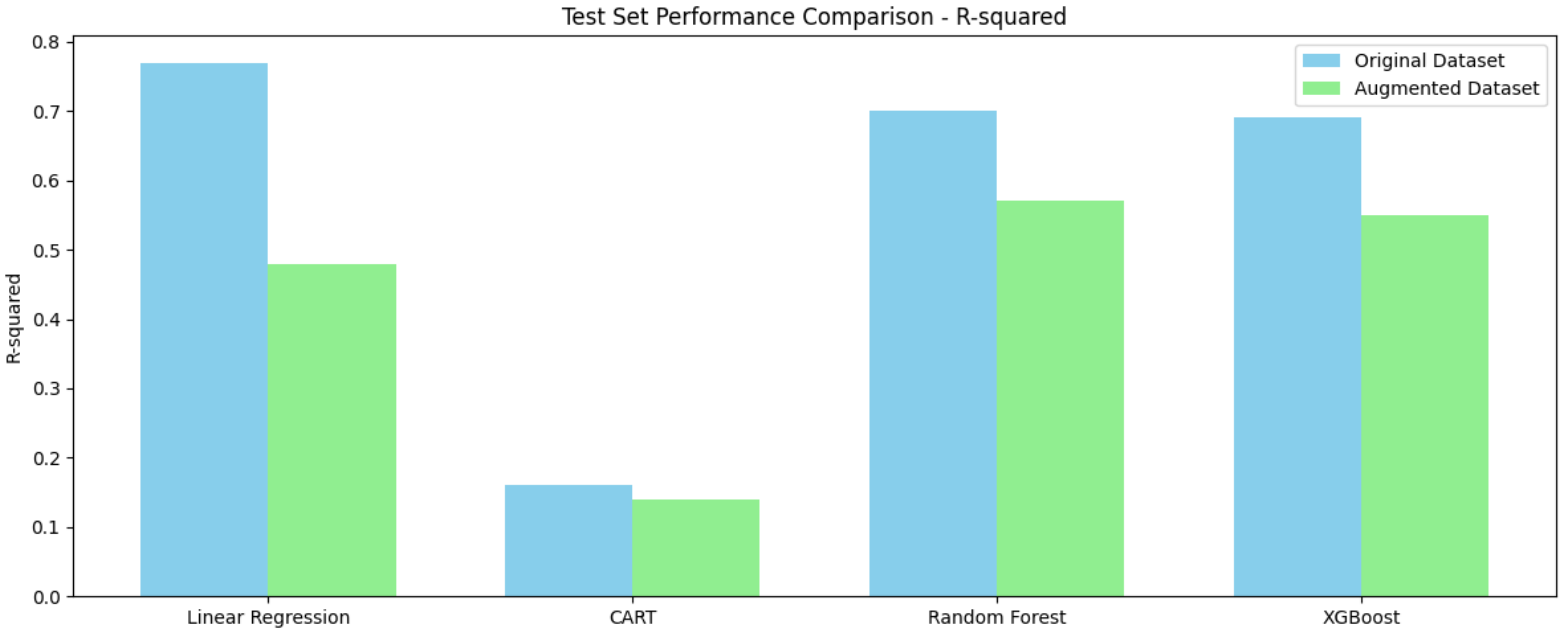


Key Insights

- AUC scores remained consistently high across models, suggesting good classification ability
- SVM achieved perfect (1.0) in-sample accuracy & 0.95 out-of-sample accuracy

Massachusetts Institute of Technology

Regression: Original vs. Augmented Data



Key Insights

- GMM augmentation unnecessarily added noise to numeric values
- Additional observations not worth the irreducible error

Massachusetts Institute of Technology

Key Results

Classification

- Binned students into 4 performance levels
- Achieved AUC of 95% on original test data with Logistic Regression
- Maintained AUC of 95% on augmented data with SVMs

Regression

- Predicted final exam scores
- Achieved OSR^2 of 77.1% on original test data with ElasticNet
- Augmentation decreased OSR^2 to 59.2%

Impact

- Enables targeted intervention of struggling students
- Equitable access to quality education