QS World University Rankings 2025: Machine Learning Analysis and Prediction

Presented by Youssef Masoud, Arnav Kucheriya, and Kushagra Verma

Agenda

- Dataset Overview
- Data Cleaning and Preprocessing
- Exploratory Data Analysis
- Model Selection and Training
- Model Evaluation and Visualizations
- Feature Importance
- Conclusions

Project Overview

- Understand factors influencing university rankings
- Predict Academic Reputation Scores using machine learning
- Compare the performance of Linear Regression and Random Forest models
- Identify key predictors for academic reputation

Dataset Overview

- Source: QS World University Rankings 2025
- 27 Columns, 1,500+ Universities
- Key Features: Reputation scores, research impact, faculty ratios

Dataset Details

Feature Types:

- Categorical: Institution Name, Location
- Numerical: Various score features (e.g., citations, employment outcomes)

Special Notes:

- Missing values represented as '-'
- Some features highly correlated with each other

Data Cleaning and Preprocessing

- Convert score columns to numeric
- Handle missing values
- Label encode categorical features

Detailed Preprocessing

Missing Values:

- ~5% of records dropped due to incomplete data

Label Encoding:

- Institution Name and Location turned into integers

Rationale:

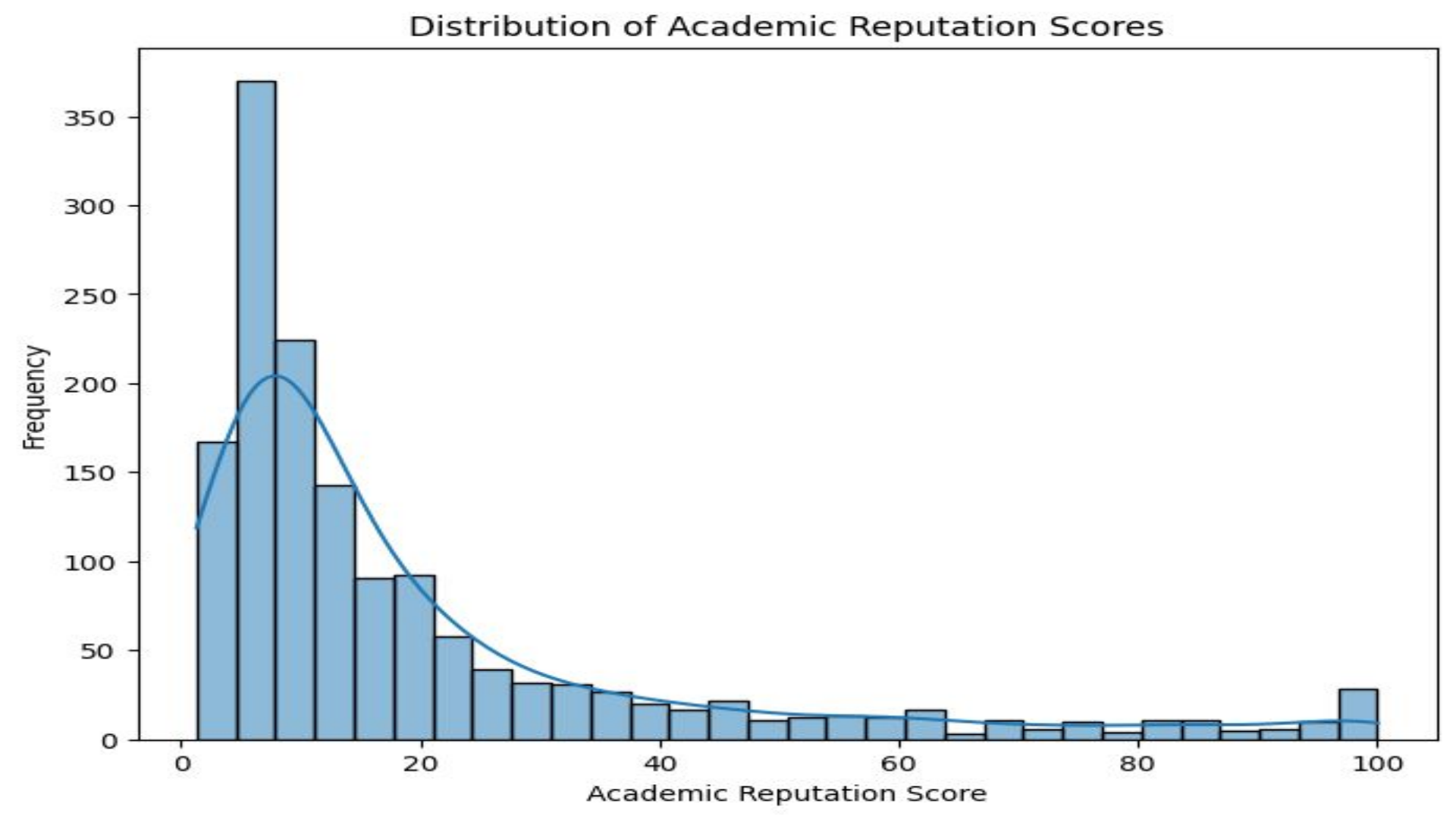
- Ensured clean numeric matrix input for modeling

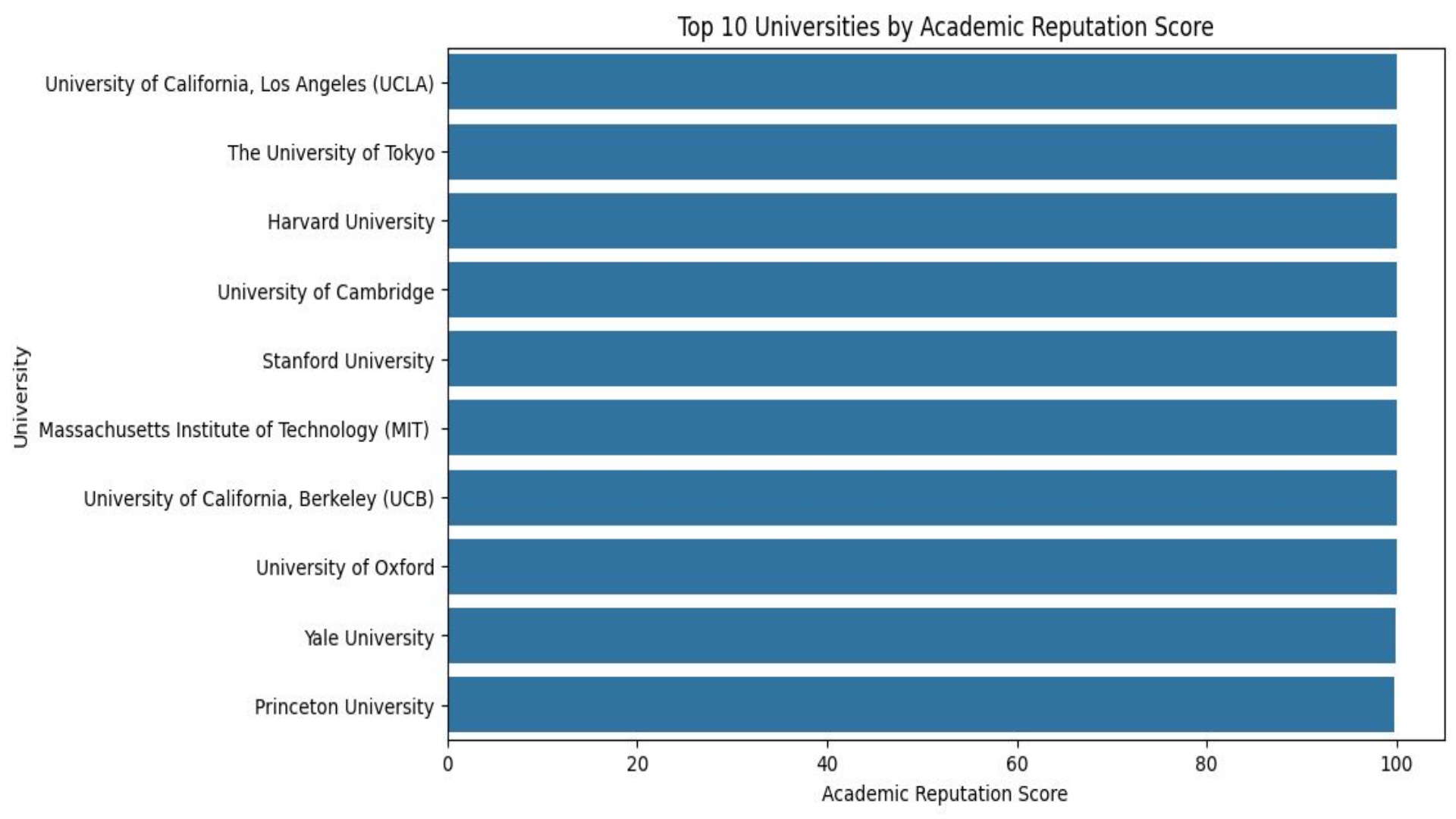
Why Data Preprocessing Matters

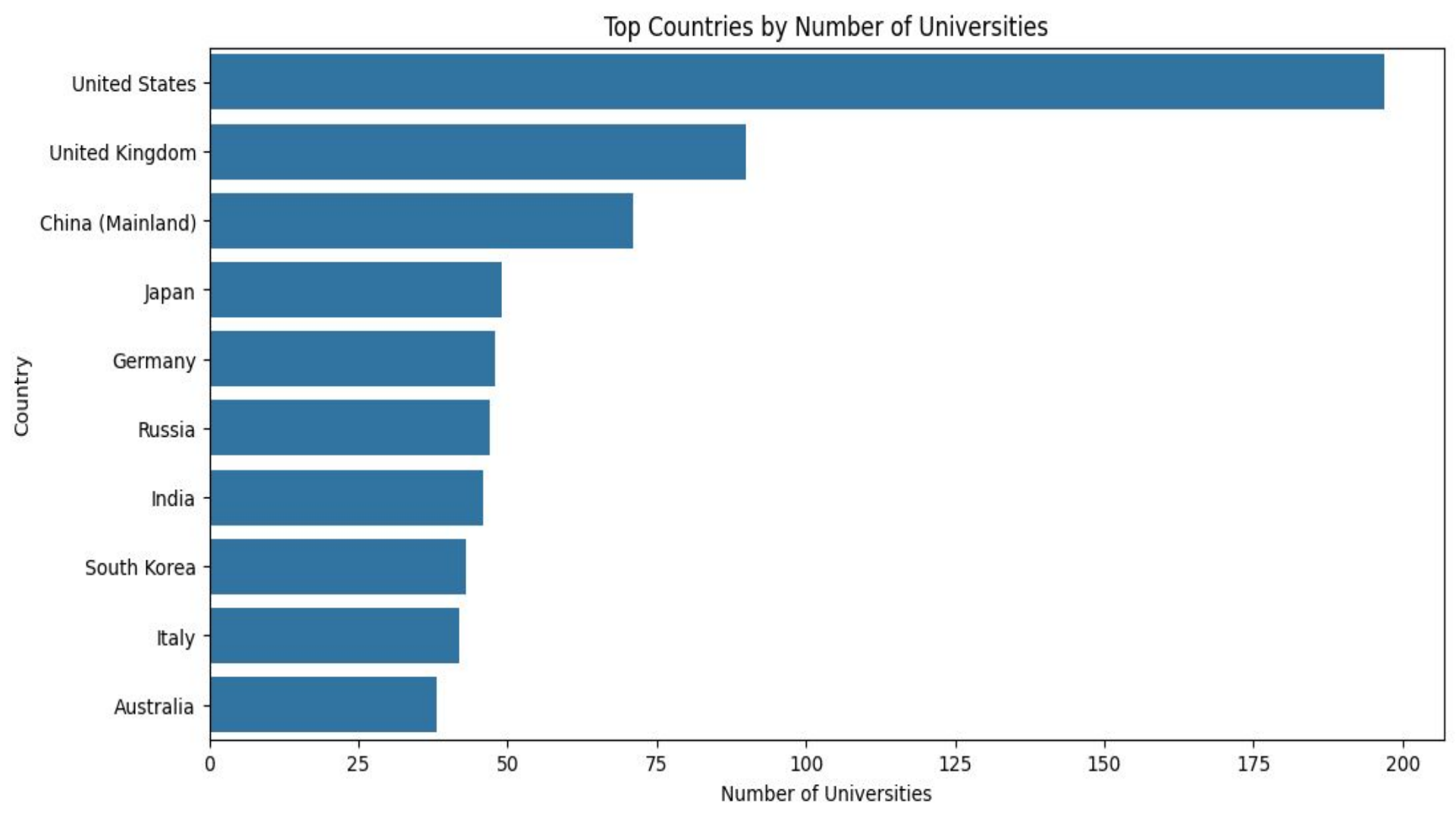
- Ensures model compatibility Improves accuracy
- Improves accuracy
- Reduces biases

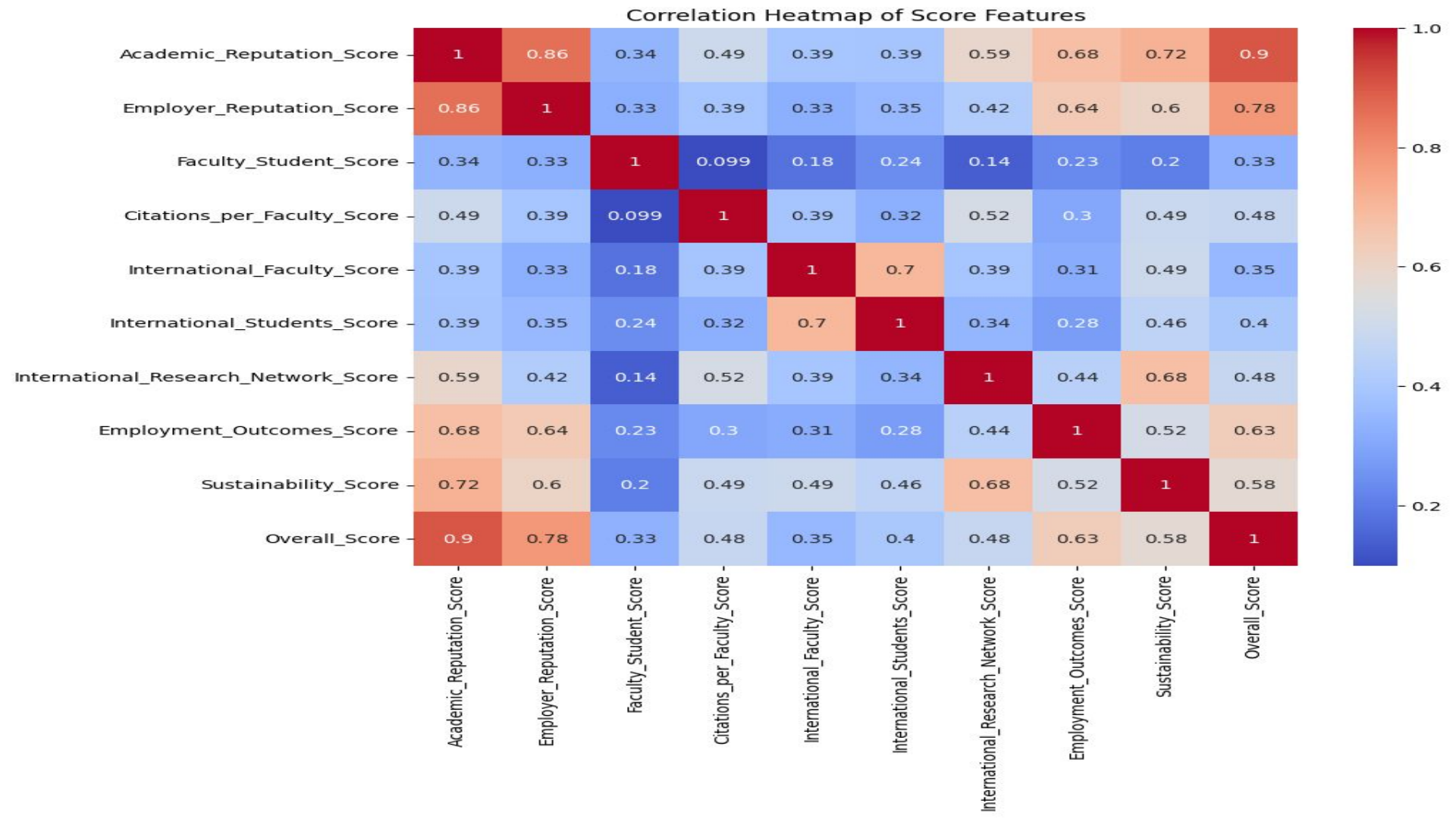
Exploratory Data Analysis (EDA)

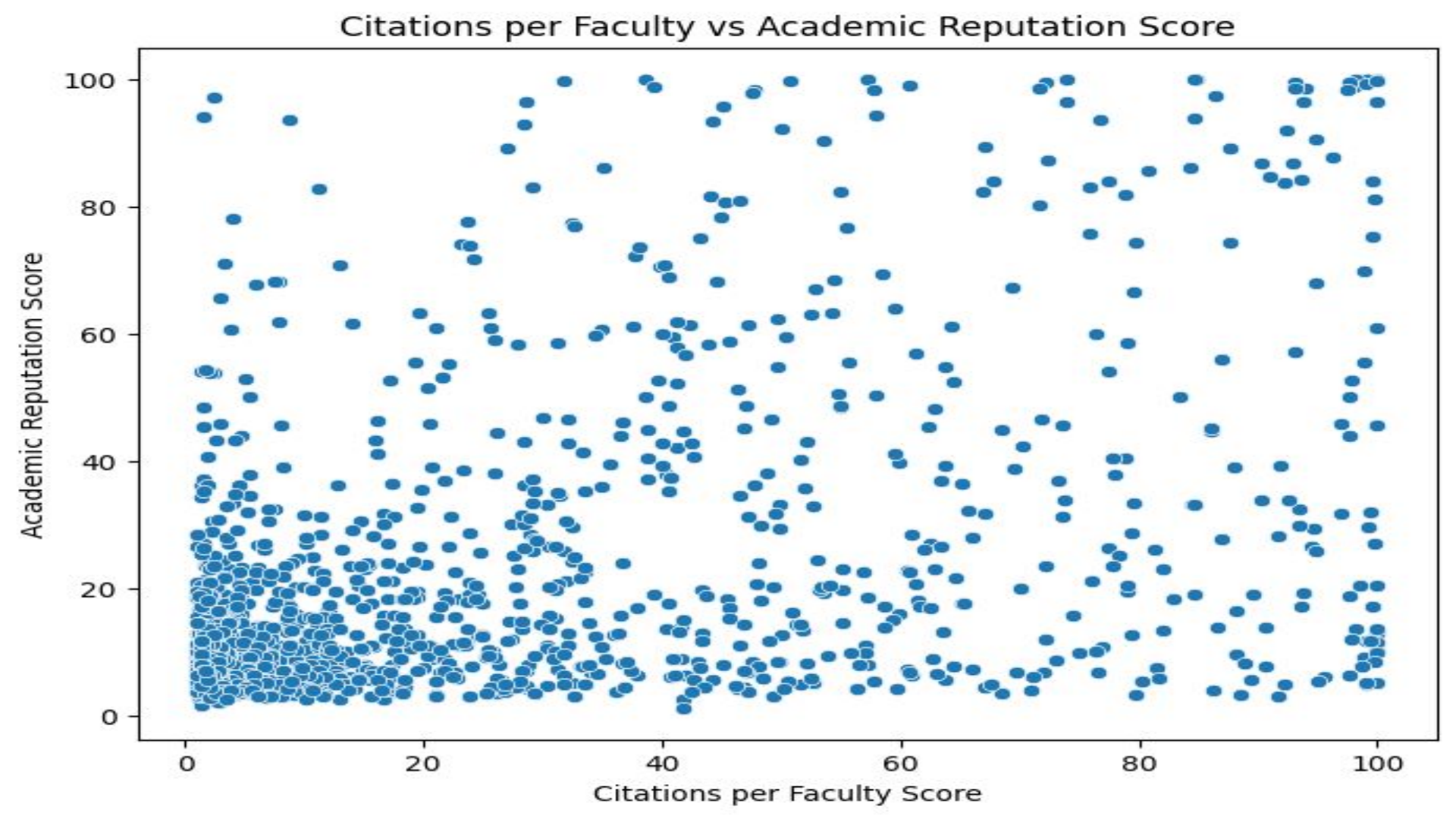
- Understand patterns and relationships
- Identify important variables
- Guide feature selection


Distribution of Academic Reputation Scores
Frequency
0
50
100
150
200
250
300
350
0
20
40
60
80
100
Academic Reputation Score


Top 10 Universities by Academic Reputation Score
University of California, Los Angeles (UCLA)
The University of Tokyo
Harvard University
University of Cambridge
Stanford University
Massachusetts Institute of Technology (MIT)
University of California, Berkeley (UCB)
University of Oxford
Yale University
Princeton University
University
0
20
40
60
80
100
Academic Reputation Score


Top Countries by Number of Universities
United States
United Kingdom
China (Mainland)
Japan
Germany
Russia
India
South Korea
Italy
Australia
Country
0
25
50
75
100
125
150
175
200
Number of Universities


Correlation Heatmap of Score Features
Academic_Reputation_Score
Employer_Reputation_Score
Faculty_Student_Score
Citations_per_Faculty_Score
International_Faculty_Score
International_Students_Score
International_Research_Network_Score
Employment_Outcomes_Score
Sustainability_Score
Overall_Score
1 0.86 0.34 0.49 0.39 0.39 0.59 0.68 0.72 0.9
0.86 1 0.33 0.39 0.33 0.35 0.42 0.64 0.6 0.78
0.34 0.33 1 0.099 0.18 0.24 0.14 0.23 0.2 0.33
0.49 0.39 0.099 1 0.39 0.32 0.52 0.3 0.49 0.48
0.39 0.33 0.18 0.39 1 0.7 0.39 0.31 0.49 0.35
0.39 0.35 0.24 0.32 0.7 1 0.34 0.28 0.46 0.4
0.59 0.42 0.14 0.52 0.39 0.34 1 0.44 0.68 0.48
0.68 0.64 0.23 0.3 0.31 0.28 0.44 1 0.52 0.63
0.72 0.6 0.2 0.49 0.49 0.46 0.68 0.52 1 0.58
0.9 0.78 0.33 0.48 0.35 0.4 0.48 0.63 0.58 1
1.0
0.8
0.6
0.4
0.2


Citations per Faculty vs Academic Reputation Score
Academic Reputation Score
Citations per Faculty Score
0
20
40
60
80
100

Model Training Summary

- Linear Regression: Quick training, simple assumptions
- Random Forest: Handles complexity, avoids overfitting
- Purpose: Predict Academic Reputation Scores accurately

Training Code: Linear Regression

```
from sklearn.linear_model import LinearRegression
lr_model = LinearRegression()
lr_model.fit(X_train, y_train)
```

Training Code: Random Forest

```
from sklearn.ensemble import RandomForestRegressor
rf_model = RandomForestRegressor(n_estimators=100, random_state=
rf_model.fit(X_train, y_train)
```

Model Evaluation Code

```
from sklearn.metrics import mean_squared_error, r2_score

# Predictions
y_pred_lr = lr_model.predict(X_test)
y_pred_rf = rf_model.predict(X_test)

# Evaluation Metrics
rmse_lr = mean_squared_error(y_test, y_pred_lr, squared=False)
r2_lr = r2_score(y_test, y_pred_lr)

rmse_rf = mean_squared_error(y_test, y_pred_rf, squared=False)
r2_rf = r2_score(y_test, y_pred_rf)
```

Model Performance Comparison

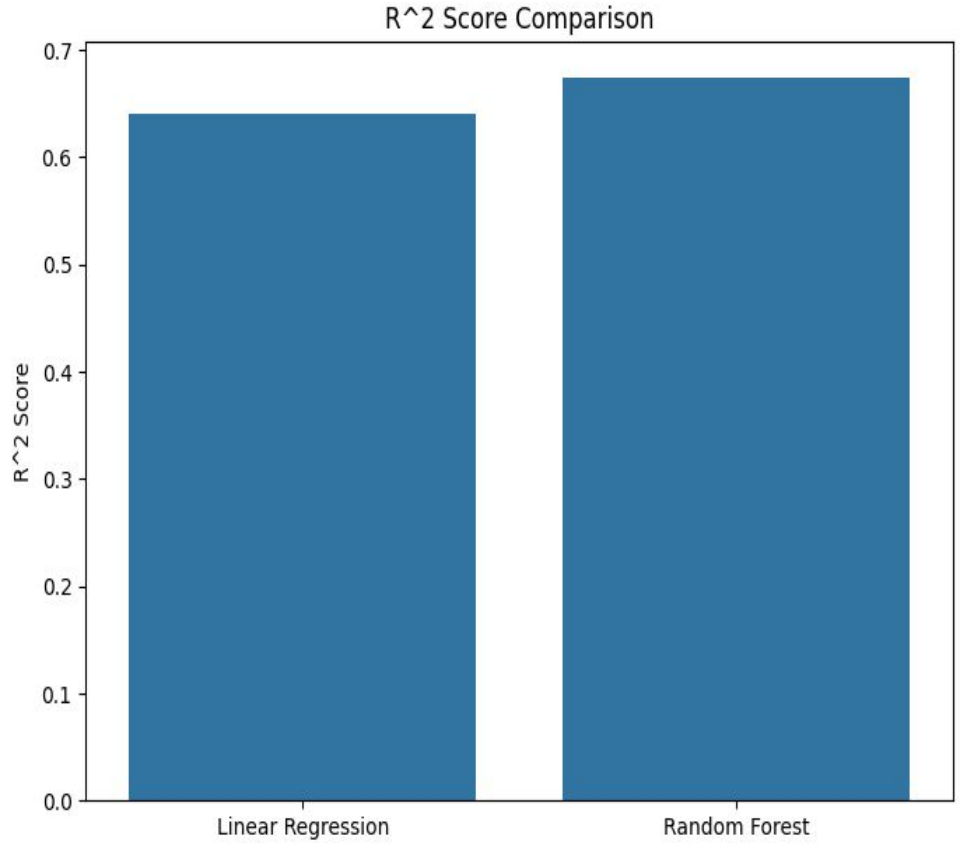
Linear Regression:

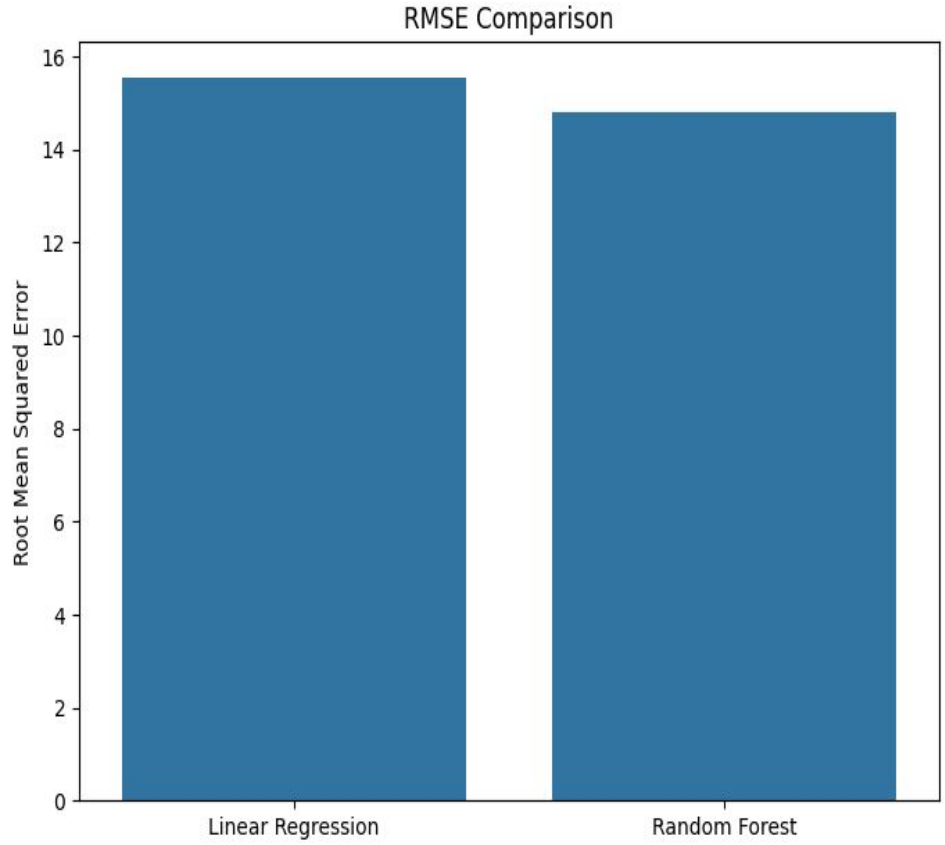
- RMSE: 15.53
- R^2 Score: 0.64

Random Forest:

- RMSE: 14.80
- R^2 Score: 0.67

RMSE and R² Score Comparison





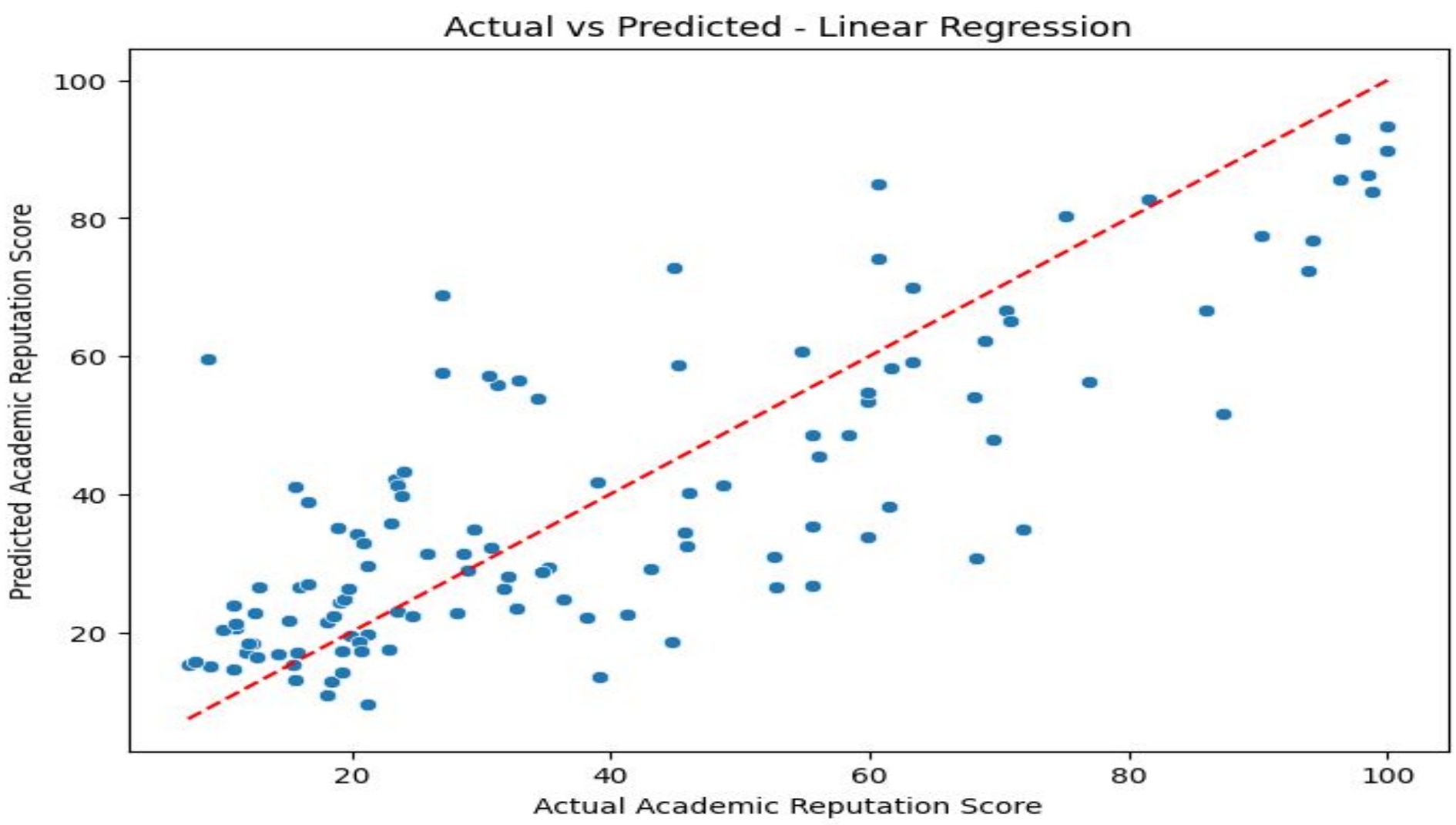
Why These Models?

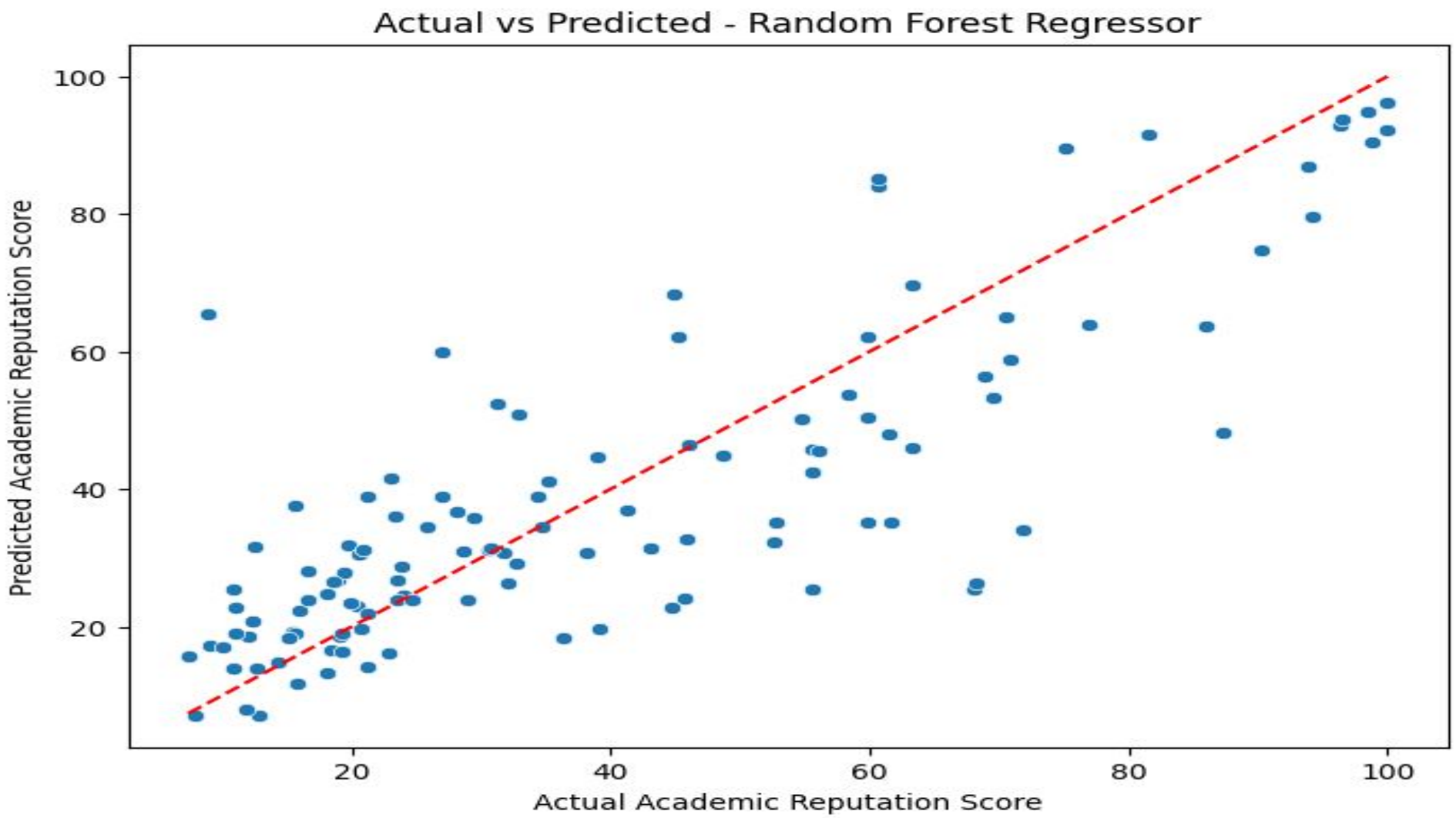
Linear Regression:

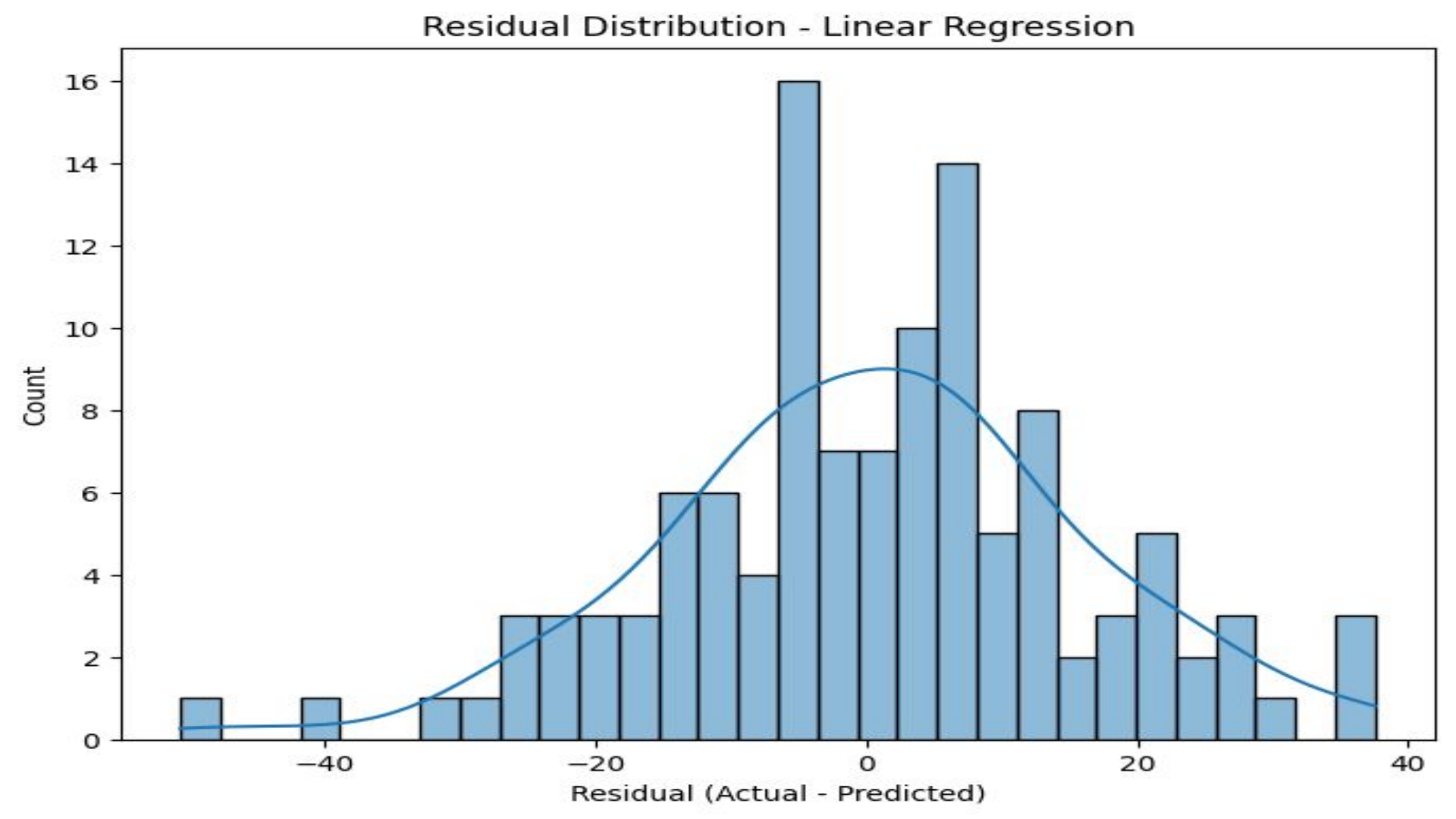
- Easy to interpret, good baseline

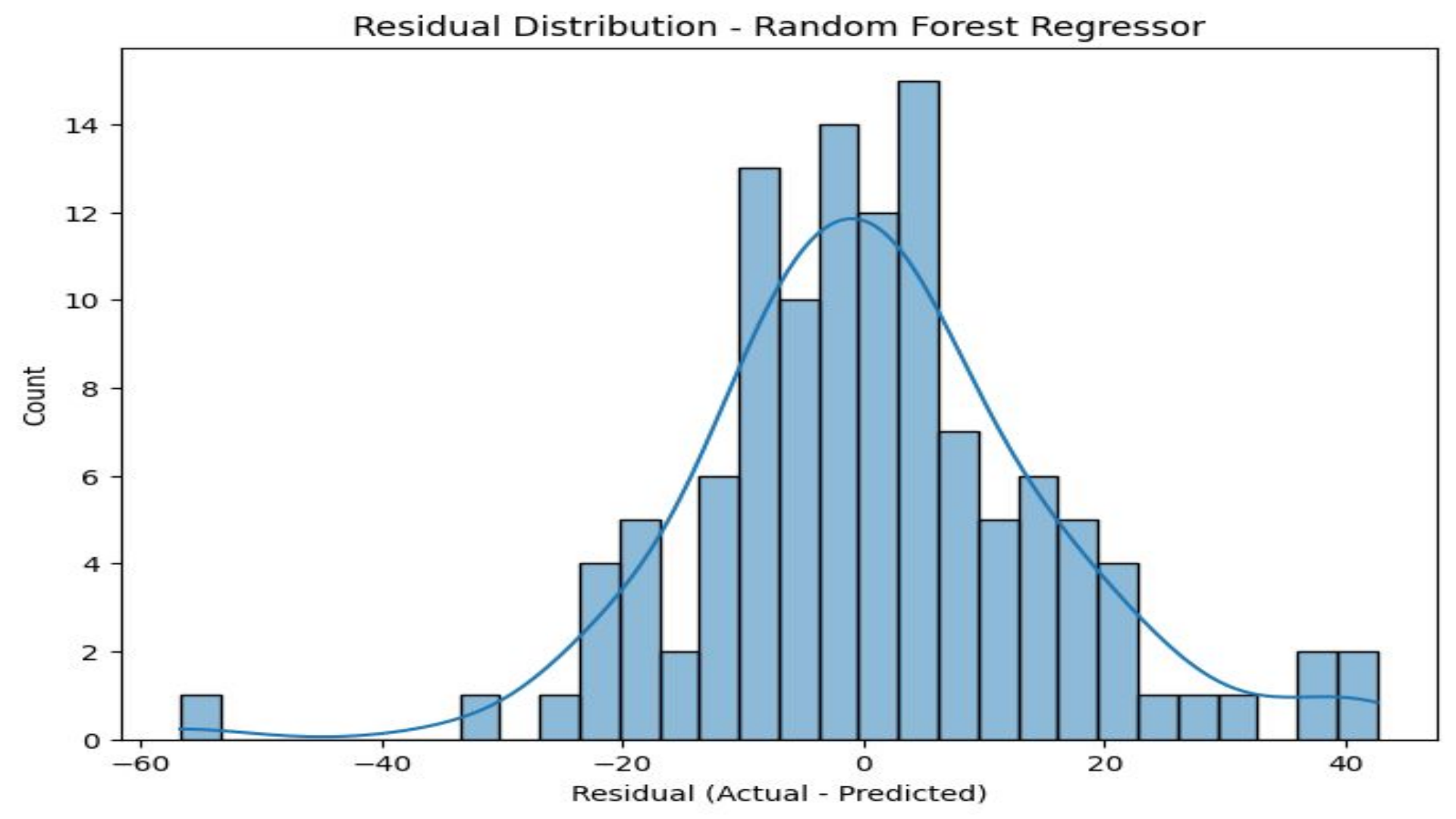
Random Forest:

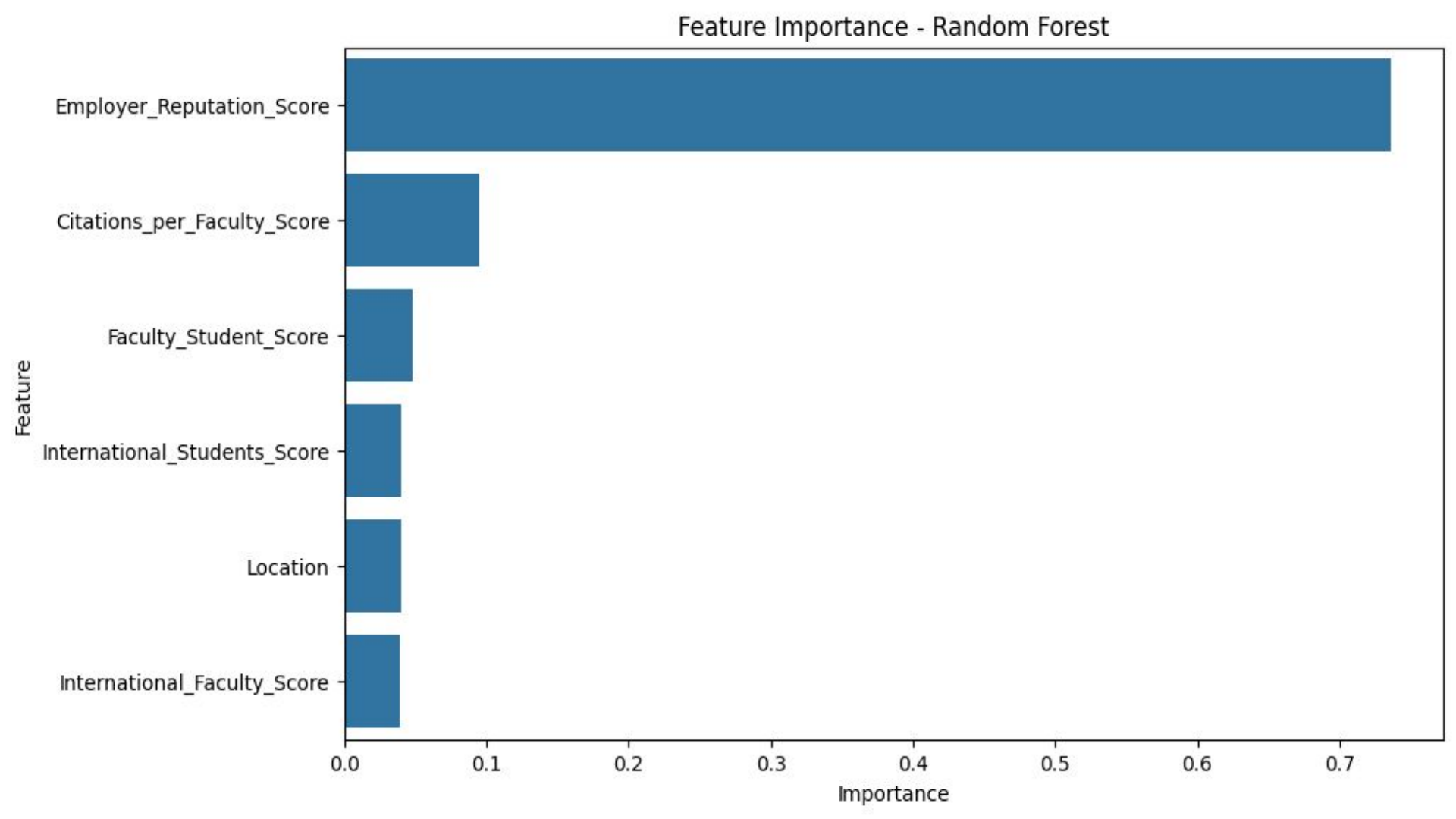
- Captures non-linearities, robust to overfitting
- Comparison provides balanced insight


Actual vs Predicted - Linear Regression
Predicted Academic Reputation Score
Actual Academic Reputation Score
100
80
60
40
20
20
40
60
80
100


Actual vs Predicted - Random Forest Regressor
Predicted Academic Reputation Score
Actual Academic Reputation Score
100
80
60
40
20
20
40
60
80
100


Residual Distribution - Linear Regression
Count
0
2
4
6
8
10
12
14
16
−40
−20
0
20
40
Residual (Actual - Predicted)


Residual Distribution - Random Forest Regressor
Count
0
2
4
6
8
10
12
14
−60
−40
−20
0
20
40
Residual (Actual - Predicted)


Feature Importance - Random Forest
Employer_Reputation_Score
Citations_per_Faculty_Score
Faculty_Student_Score
International_Students_Score
Location
International_Faculty_Score
Feature
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Importance

Key Insights

- Research impact (citations) is a critical determinant of academic reputation.
- Employer reputation plays a significant supporting role.
- Random Forest captures hidden patterns better than linear models.
- High-quality data preprocessing is crucial for model accuracy.
- Institutional strategies should prioritize research output and industry engagement.

Challenges Faced

Data Issues:

- Missing or corrupted entries
- Inconsistent data formatting

Modeling Issues:

- Linear Regression limited in capturing complex patterns

Mitigation:

- Careful data cleaning, choosing an ensemble model

Conclusion

This project analyzed the QS World University Rankings 2025 dataset to predict Academic Reputation Scores based on institutional features. Through a structured machine learning workflow, several important insights emerged:

- Random Forest outperformed Linear Regression
- Focus on research and employer links
- Machine learning provides strategic insights

Thank you