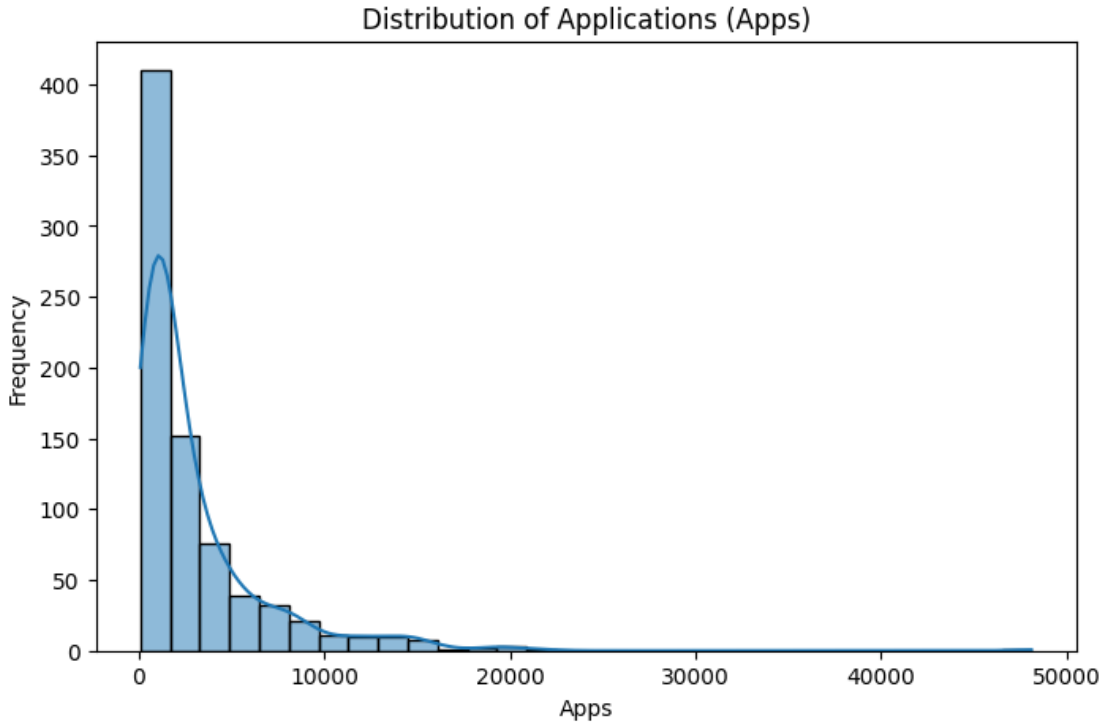
1. Introduction

This project aims to predict the number of applications received by U.S. colleges using machine learning techniques. The analysis follows the CRISP-DM methodology, including data understanding, preprocessing, exploratory analysis, model development, and evaluation. The dataset contains admissions-related variables such as acceptance numbers, enrollment, academic quality indicators, expenditures, and demographic information.

---

2. Exploratory Data Analysis (EDA)

2.1 Distribution of Applications (Apps)

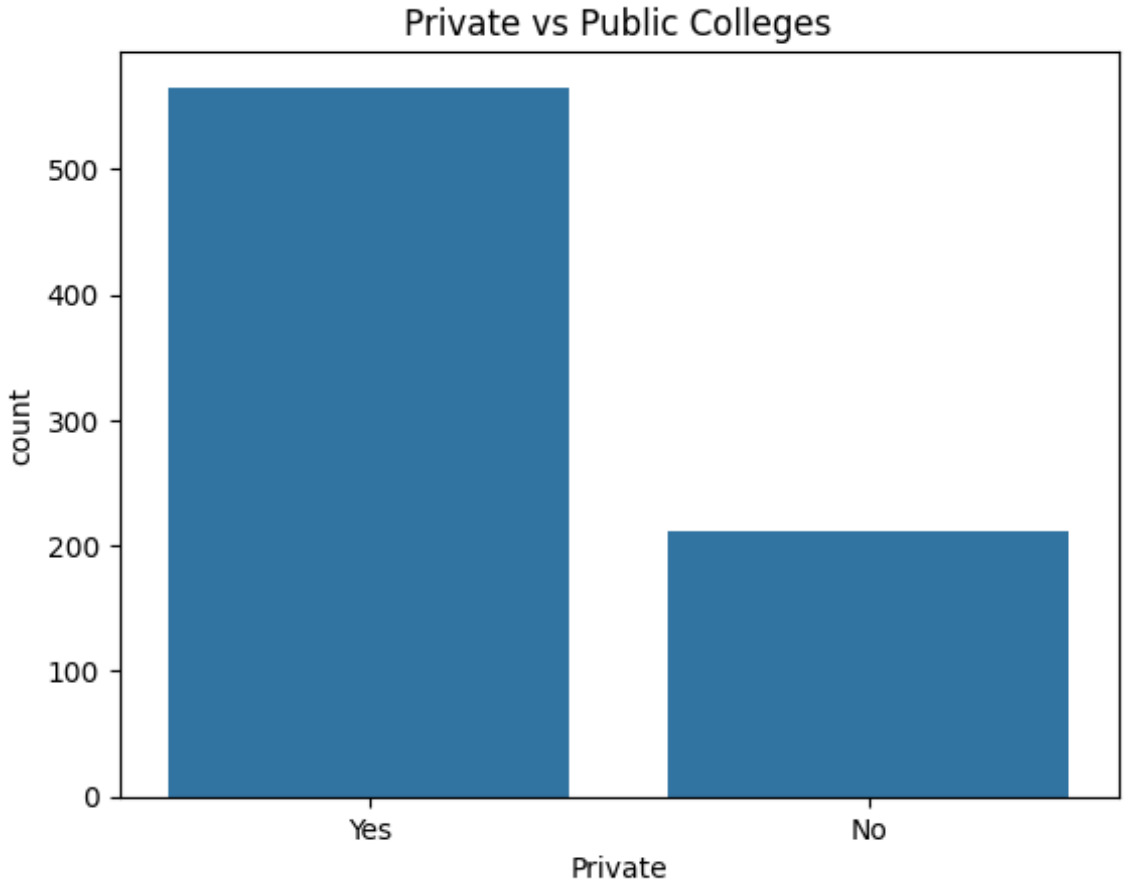


The distribution of the target variable *Apps* is highly right-skewed.
Most colleges receive relatively few applications, whereas a small number of highly competitive institutions receive extremely high volumes.
This indicates:

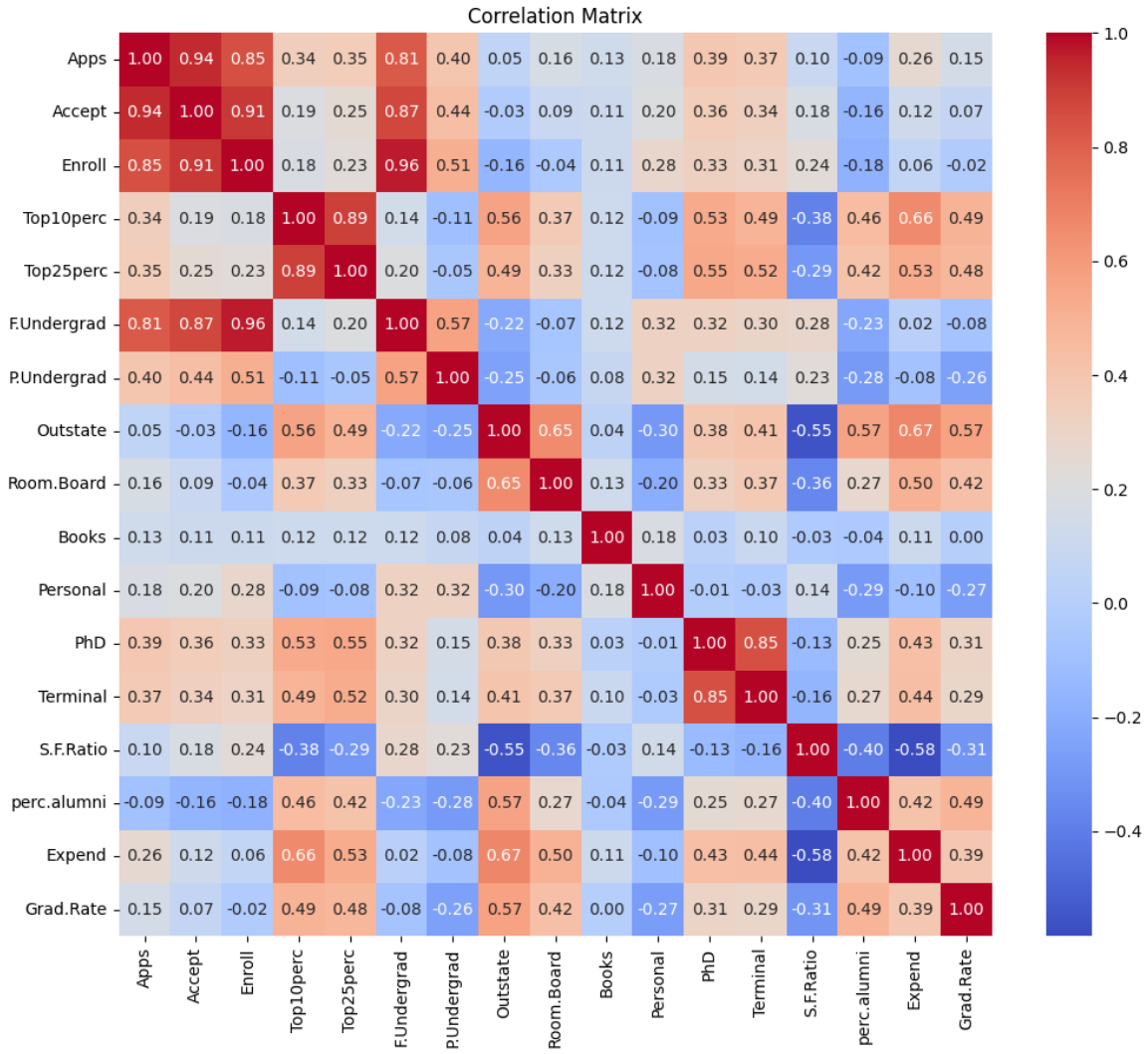
- The presence of *outliers*
- High variance
- A need for robust models that handle skewed distributions well (e.g., Random Forest, Gradient Boosting)

2.2 Private vs Public Colleges

A bar plot shows that the number of private colleges in the dataset is substantially higher than public colleges.
Further analysis suggests that institutional type does influence application numbers, although it is not the strongest predictor compared to numerical academic and financial variables.



2.3 Correlation Analysis



The correlation matrix reveals several important patterns:

- **Accept** has the strongest correlation with **Apps**, suggesting that colleges receiving more applications tend to accept more students.
- **Enroll**, **Top25perc**, and **Expend** also show meaningful correlations.
- Academic quality indicators (Top10perc, Top25perc) are moderately correlated, indicating that more selective or academically strong colleges tend to attract more applicants.
- Variables such as **Books** and **Personal** expenses show very weak correlations and contribute little to prediction.

Overall, the correlation matrix confirms that:

Application volume is driven largely by acceptance behavior, student enrollment patterns, and college academic profile.

---

3. Model Development & Comparison

Three predictive models were developed:

- **Linear Regression**
- **Random Forest Regressor**
- **Gradient Boosting Regressor**

Each model was evaluated using **MSE, MAE, and R^2** on test data.

---

4. Model Performance

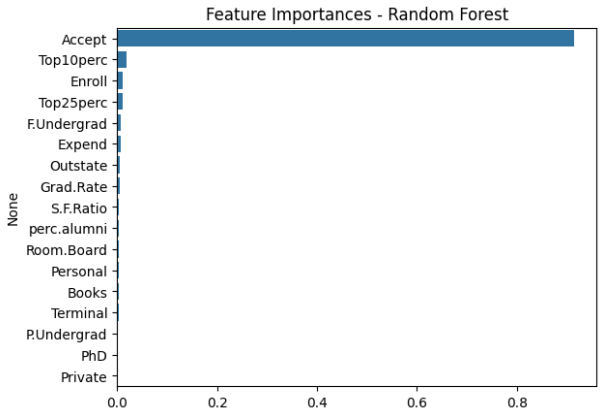
Model	MSE	MAE	R^2
Linear Regression	1,492,443.38	744.86	0.89
Random Forest	1,029,753.23	591.57	0.92
Gradient Boosting	1,316,452.54	582.13	0.90

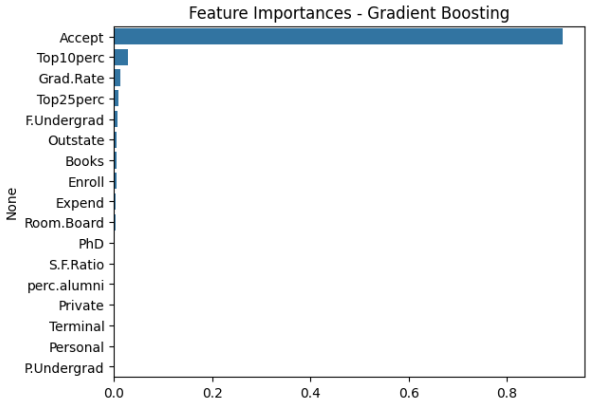
4.1 Performance Interpretation

- **Linear Regression** performs reasonably well ($R^2 = 0.89$), but its assumptions (linearity, homoscedasticity) limit performance on this nonlinear dataset.
- **Random Forest** achieves the best overall performance:
  - Lowest MSE
  - Substantial reduction in MAE
  - Highest R^2 (0.92)
- **Gradient Boosting** performs slightly worse than Random Forest but better than Linear Regression.

Conclusion:

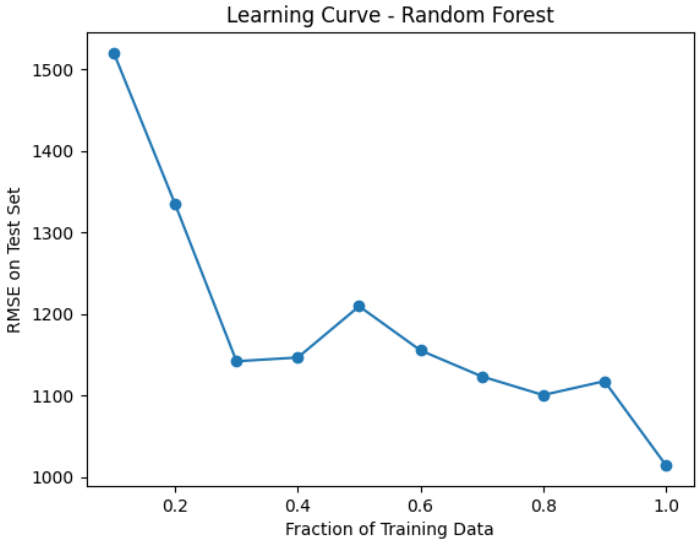
Random Forest is the most accurate and stable model for predicting college applications in this dataset.





---

5. Learning Curve Analysis



The learning curve for the Random Forest model shows:

- Higher error with very small training sizes (underfitting)
- Smooth and consistent improvement with more training data
- Convergence of training and validation error at higher sample sizes

This indicates:

- The model generalizes well
- Additional data could further improve performance
- The model is neither severely overfitting nor underfitting

---

6. Prediction Example

For a selected college:

- **Actual Applications:** 4877
- Linear Regression Prediction: 4222
- Random Forest Prediction: 3863
- Gradient Boosting Prediction: 4321

Random Forest slightly underestimates the value but remains close.
Gradient Boosting is the closest to the true value in this example.

---

7. Real-World Interpretation

This type of model can help institutions:

- Estimate expected application volumes
- Plan admissions capacity
- Allocate financial and staffing resources
- Adjust recruitment strategies
- Compare performance against similar institutions
- Forecast demand for future academic years

Predictive analytics in college admissions is highly valuable for budgeting, marketing, and strategic planning.

---

8. Conclusion

This project successfully demonstrates:

- Application of the CRISP-DM framework
- Use of EDA and visualization to understand dataset structure
- Development and comparison of machine learning models
- A real-world interpretation of institutional behavior and application trends

The **Random Forest** model provides the strongest predictive performance, affirming its ability to handle complex, nonlinear relationships in educational datasets.