MBA ADMISSION CLASSIFICATION USING LOGISTIC REGRESSION WITH BACKWARD ELIMINATION

By NEXUS

INTRODUCTION

Background

Executive Summary

BACKGROUND

- **Context and Motivation**: Highlights the significance of trends in MBA admissions, placements, and factors influencing success metrics, stressing the value of analyzing this data.
- **Objective**: Predict outcomes (admissions, placements) or identify key success factors for actionable insights.
- **Dataset Overview**: Includes demographic data, academic performance, test scores, and placement details.
- **Key Challenges**: Addresses missing data, imbalanced classes, and complex feature relationships.
- **Impact**: Supports evidence-based decisions for students, institutions, and recruiters

EXECUTIVE SUMMARY

Project Purpose: Utilize machine learning, specifically logistic regression, to predict MBA admissions outcomes and identify key success factors.

Methodology:

- Data preprocessing: Handled missing values, **SMOTE** for class imbalance.
- Model building: **Logistic Regression** with **backward elimination**.

- **Key Results**:
  - High-performing features: GPA, GMAT, and work experience.
  - **AUC: 0.921**, indicating excellent predictive power.
  - Model interprets the influence of predictors effectively.

DATASET

Descriptive Analysis

Data Preprocessing

Exploratory Data Analysis

DESCRIPTIVE ANALYSIS

- The dataset contains 6,194 MBA application records of 2024 with details such as demographics (gender, race, international status), academic performance (GPA, GMAT), work experience (years, industry), and admission results.

	application_id	gpa	gmat	work_exp
count	6194.000000	6194.000000	6194.000000	6194.000000
mean	3097.500000	3.250714	651.092993	5.016952
std	1788.198115	0.151541	49.294883	1.032432
min	1.000000	2.650000	570.000000	1.000000
25%	1549.250000	3.150000	610.000000	4.000000
50%	3097.500000	3.250000	650.000000	5.000000
75%	4645.750000	3.350000	680.000000	6.000000
max	6194.000000	3.770000	780.000000	9.000000

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6194 entries, 0 to 6193
Data columns (total 10 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   application_id  6194 non-null   int64
 1   gender          6194 non-null   object
 2   international   6194 non-null   bool
 3   gpa             6194 non-null   float64
 4   major           6194 non-null   object
 5   race            4352 non-null   object
 6   gmat            6194 non-null   int64
 7   work_exp        6194 non-null   int64
 8   work_industry   6194 non-null   object
 9   admission       1000 non-null   object
dtypes: bool(1), float64(1), int64(3), object(5)
memory usage: 441.7+ KB
```

DATA PREPROCESSING

Handling Missing Values:

- Addressed missing data in "race" and "admission" columns via imputation, removal, or encoding as per project needs.

Data Cleaning:

- Removed irrelevant columns and standardized categorical variables.

Encoding Categorical Variables:

- Applied one-hot or label encoding to "gender," "major," "work_industry," and "race“.

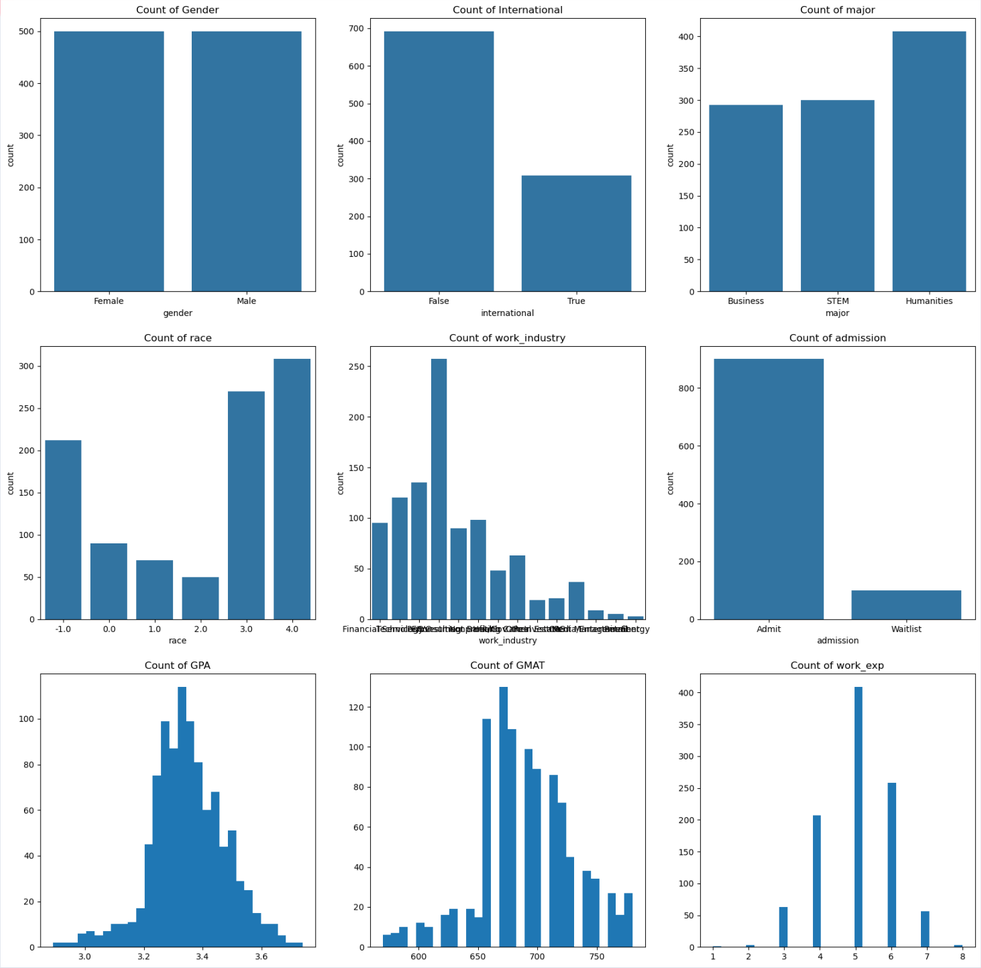
Feature Scaling:

- Scaled numerical features ("GPA," "GMAT," "work_exp") using standardization/normalization.

Imbalanced Data Handling:

- Balanced the target variable "admission" using SMOTE for better model performance.

EXPLORATORY DATA ANALYSIS



Demographic and Academic Distributions:

- Balanced gender distribution; domestic applicants outnumber international.
- Academic majors: evenly split between Business, STEM, and Humanities.
- GPA and GMAT scores follow a near-normal distribution

Professional Experience and Industry:

- Majority have 4–6 years of work experience.
- Diverse industries; Financial Services most common

Admissions Outcome:

- Majority classified as "Admit"; fewer in "Waitlist," indicating target imbalance.

EXPLORATORY DATA ANALYSIS

Relationships Between Features:

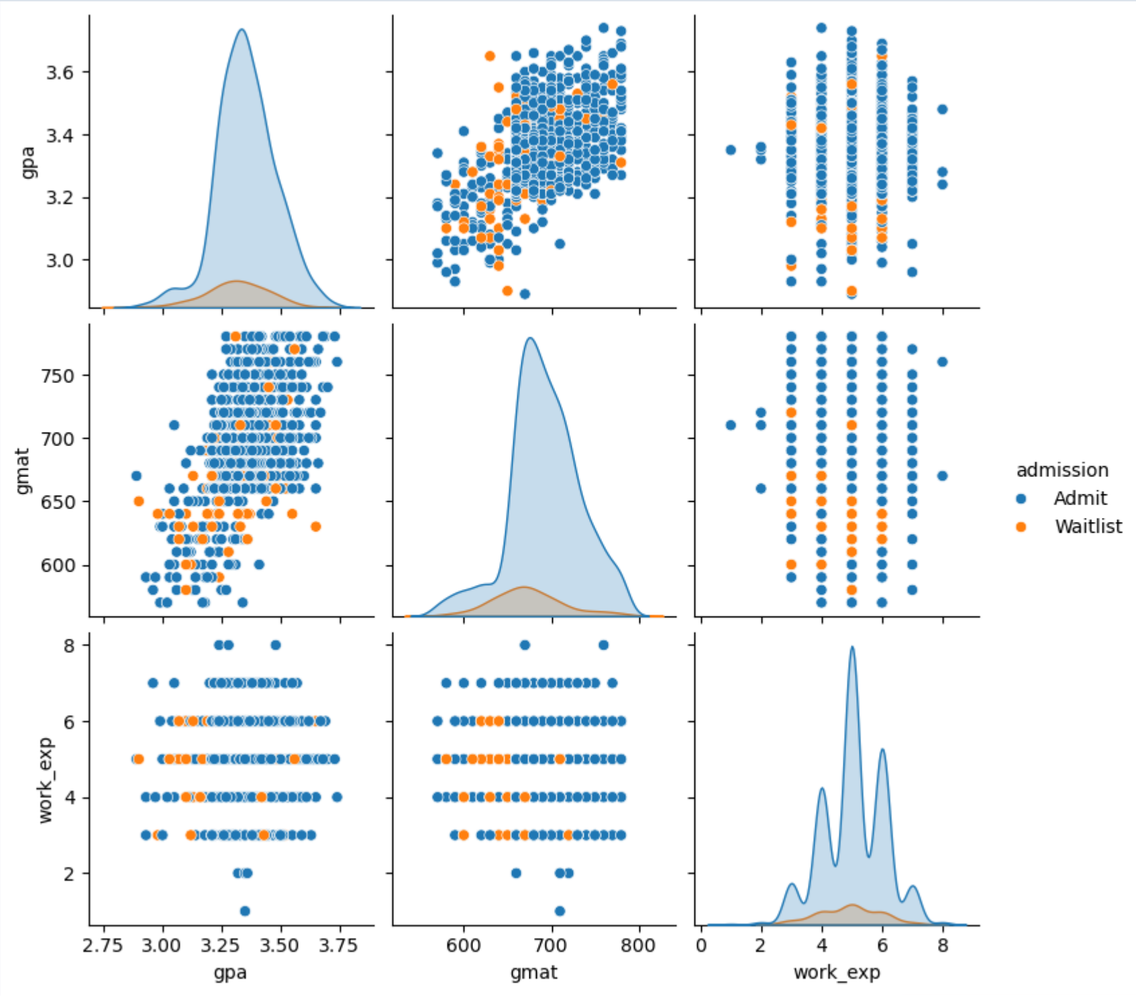
- Positive correlation between GPA and GMAT; no strong link with work experience

Admissions Insights:

- Admit" clusters around higher GPA and GMAT scores.
- "Waitlist" points are more scattered

Feature Distributions:

- GPA and GMAT: Bell-shaped distributions.
- Work experience: Skewed towards 4–6 years



MODEL AND RESULT

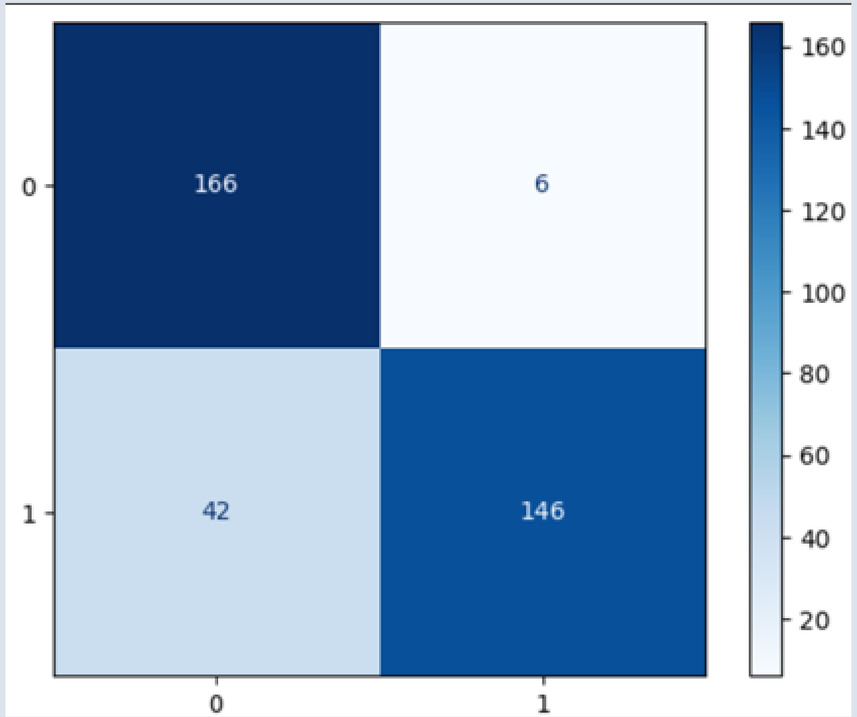
Result

Discussion

LOGISTIC REGRESSION

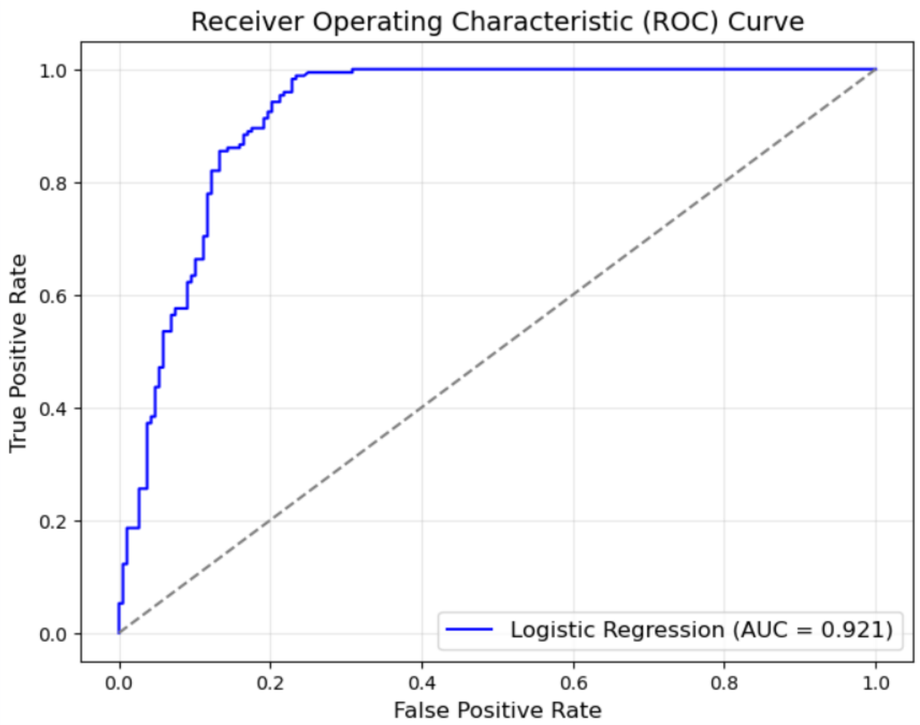
- **Simplicity & Interpretability**: Assumes linear relationship with log-odds, aiding feature impact interpretation.
- **Stakeholder Communication**: Provides clear insights into how features influence admissions outcomes.
- **Class Imbalance Handling**: Effective with weighting or resampling (e.g., SMOTE) for balanced predictions.
- **Baseline Evaluation**: Serves as a benchmark to confirm the need for more complex models

RESULTS



- **High Precision (96%):** Most predicted admissions are accurate.
- **Moderate Recall (78%):** Some applicants who should have been admitted were missed.
- The model has a strong **overall accuracy (~91%)**, but improvements are needed in reducing false negatives

RESULTS



- **AUC: 0.921** – Excellent discriminatory power.
- **Feature Selection**: Backward elimination removed statistically insignificant or highly correlated features to prevent multicollinearity.
- **Key Benefit**: Focused on impactful predictors, ensuring interpretability and avoiding overfitting

CONCLUSION AND RECOMMENDATION

Conclusion

Recommendation

CONCLUSION

- The project successfully predicts MBA admissions outcomes using Logistic Regression, with strong performance metrics such as an AUC of 0.921.
- The analysis emphasizes the importance of academic and professional metrics like GMAT, work industry and work experience in influencing admissions decisions.
- The results demonstrate the effectiveness of combining interpretable and complex models for predictive accuracy and actionable insights.

RECOMMENDATION

- To further enhance model performance and applicability, it is recommended to collect a more balanced dataset to address class imbalance.
- Additionally, incorporating external data such as interview performance or recommendation scores could provide a more holistic view of the admissions process.
- It would also be worth-while if we were to explore the possibility of using models like Random Forest and XGBoosting.

Q & A

THANK YOU

By NEXUS