ASDS – 5302

Final Project Report

MBA Admission Classification Using Logistic Regression with Backward Elimination

Hareen Maddipatla	**(1002233075)**
Rushika Badri Prasad	**(1002235923)**
Karunakar Kalvala	**(1002198623)**

UTA

THE UNIVERSITY OF TEXAS
AT ARLINGTON

1. INTRODUCTION:

1.1. Problem Statement:

The MBA admissions process involves evaluating numerous applications based on diverse factors such as academic performance, test scores, work experience, and demographics. However, inconsistencies, biases, and inefficiencies in traditional evaluation methods can hinder accurate predictions of admission outcomes. A systematic and data-driven approach is needed to improve the accuracy and fairness of admissions decisions while identifying the key factors influencing these outcomes.

1.2. Objective:

The objective of this project is to develop a robust machine learning model using logistic regression to predict MBA admissions outcomes and analyze the influence of critical factors like GPA, GMAT scores, and work experience. This project aims to enhance decision-making for institutions by providing actionable insights and offering a framework for addressing key challenges like data imbalance and complex feature relationships.

1.3. Motivation:

The growing competitiveness in MBA admissions, coupled with the need for fair and efficient evaluation methods, underscores the importance of leveraging data analytics. By harnessing the potential of machine learning, this project seeks to provide an evidence-based solution that benefits students, institutions, and recruiters alike. The project aspires to contribute to a transparent and merit-based admissions process while exploring the role of academic, professional, and demographic variables in predicting success.

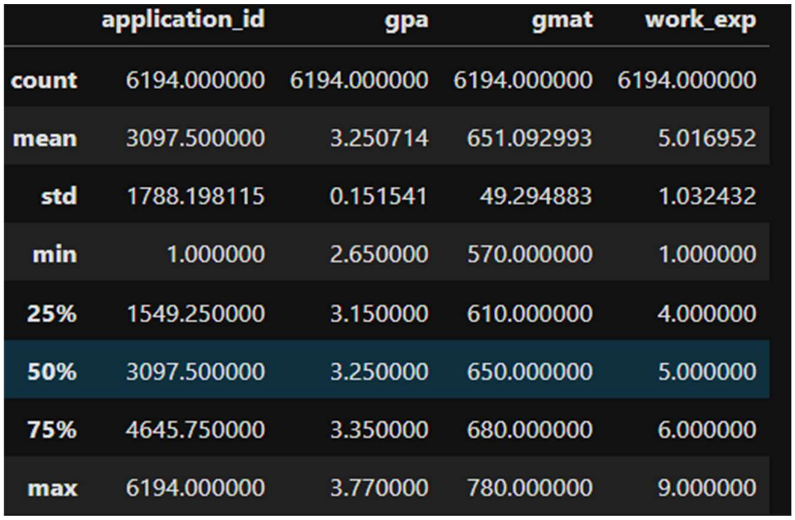
1.4. Scope:

This project focuses on analyzing a comprehensive dataset of MBA applicants to identify trends, predict admissions outcomes, and highlight critical success factors. Key activities include preprocessing data, addressing missing values, balancing imbalanced classes using SMOTE, and applying logistic regression with backward elimination for feature selection. The project's outcomes are designed to inform admissions strategies, improve predictive accuracy, and lay the groundwork for future exploration of advanced machine learning models like Random Forest or XGBoosting.

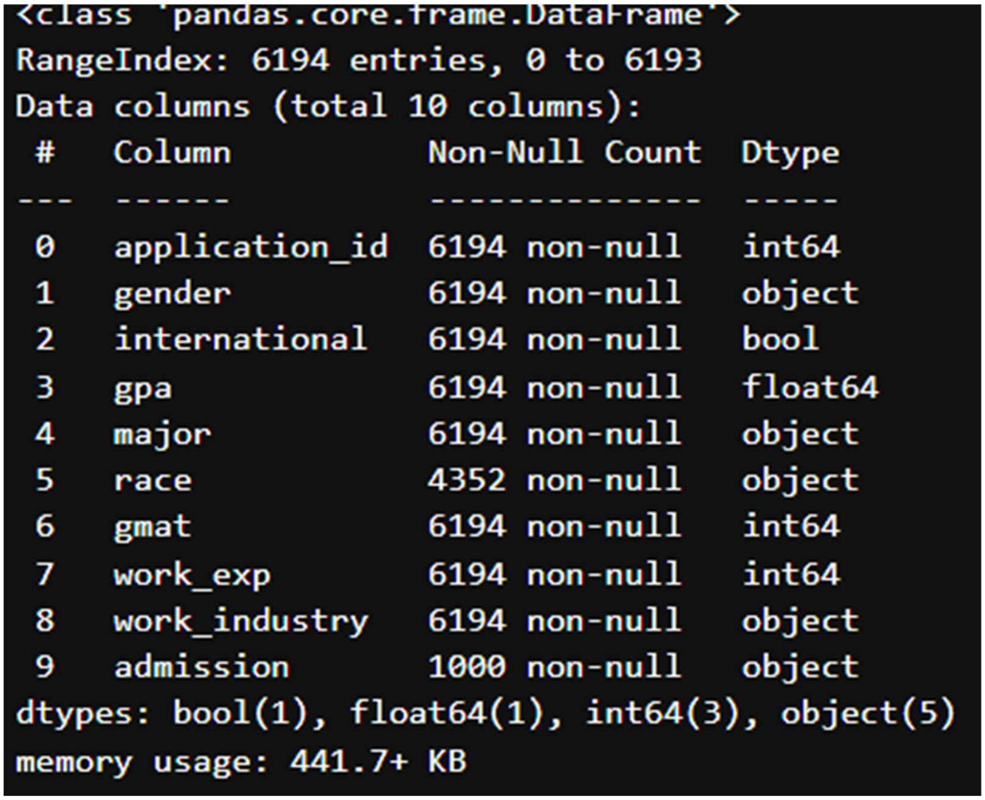
2. DESCRIPTIVE ANALYSIS:

The dataset comprises 6,194 MBA application records, providing insights into candidates' academic and professional profiles. The key variables include **GPA**, **GMAT scores**, and **work experience**. The GPA scores range from **2.65 to 3.77**, with an average of **3.25** and a standard deviation of **0.15**, indicating that the majority of candidates have a strong academic background with slight variability. GMAT scores span from **570 to 780**, with a mean of **651.09** and a standard deviation of **49.29**, reflecting high competitiveness in test performance. For work experience, the range is **1 to 9 years**, averaging **5 years**, with a standard deviation of **1.03**, highlighting that candidates generally possess moderate professional experience.

The 25th percentile values for GPA, GMAT, and work experience are **3.15**, **610**, and **4 years**, respectively, while the 75th percentiles are **3.35**, **680**, and **6 years**, showcasing well-distributed characteristics across the dataset. These distributions suggest a well-rounded applicant pool, with most candidates demonstrating solid academic performance, competitive test scores, and relevant work experience, which are likely critical factors in MBA admissions decisions.



	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
```

3. DATA PREPROCESSING:

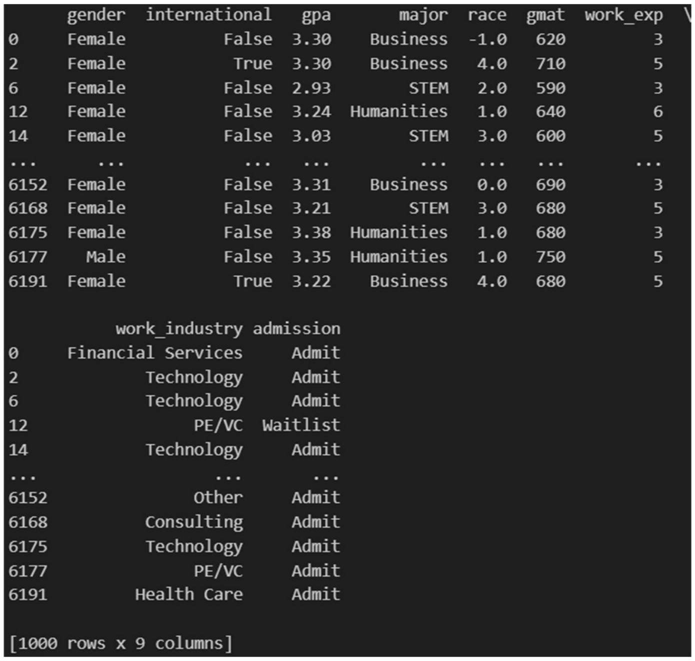
Data cleaning and preprocessing are critical steps in ensuring that the dataset is reliable and suitable for analysis. For this project, significant efforts were made to clean the data, address quality issues, and preprocess the variables to extract meaningful insights.

3.1. Handling Missing Values:

Missing data in columns like "race" and "admission" was addressed through appropriate techniques. We have implemented KNN imputation for “race” feature due to missing values and dropped the blank rows data in “admission” feature. Later performed label-encoding and one-hot encoding for training the model and classifying the results. This ensured that the model had a complete dataset without introducing biases or inconsistencies.

3.2. Data Cleaning:

Irrelevant or redundant columns were removed, and this step streamlined the dataset, ensuring only the most relevant information was retained for analysis.



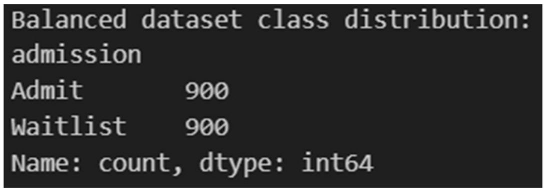
```
      gender  international   gpa       major  race  gmat  work_exp  \
0     Female          False  3.30    Business  -1.0   620         3
2     Female           True  3.30    Business   4.0   710         5
6     Female          False  2.93        STEM   2.0   590         3
12    Female          False  3.24  Humanities   1.0   640         6
14    Female          False  3.03        STEM   3.0   600         5
...      ...            ...   ...         ...   ...   ...       ...
6152  Female          False  3.31    Business   0.0   690         3
6168  Female          False  3.21        STEM   3.0   680         5
6175  Female          False  3.38  Humanities   1.0   680         3
6177    Male          False  3.35  Humanities   1.0   750         5
6191  Female           True  3.22    Business   4.0   680         5

           work_industry admission
0     Financial Services     Admit
2             Technology     Admit
6             Technology     Admit
12                 PE/VC  Waitlist
14            Technology     Admit
...                  ...       ...
6152               Other     Admit
6168          Consulting     Admit
6175          Technology     Admit
6177               PE/VC     Admit
6191         Health Care     Admit

[1000 rows x 9 columns]
```

3.3. Addressing Class Imbalance:

The target variable "admission" showed an imbalanced distribution, which was addressed using Synthetic Minority Oversampling Technique (SMOTE). This created a balanced dataset by generating synthetic samples for underrepresented classes, improving model performance and prediction reliability.



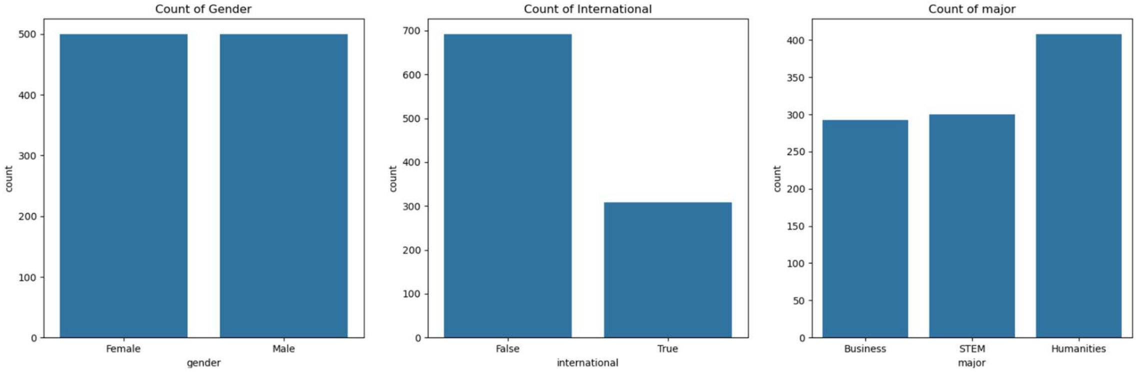
```
Balanced dataset class distribution:
admission
Admit       900
Waitlist    900
Name: count, dtype: int64
```

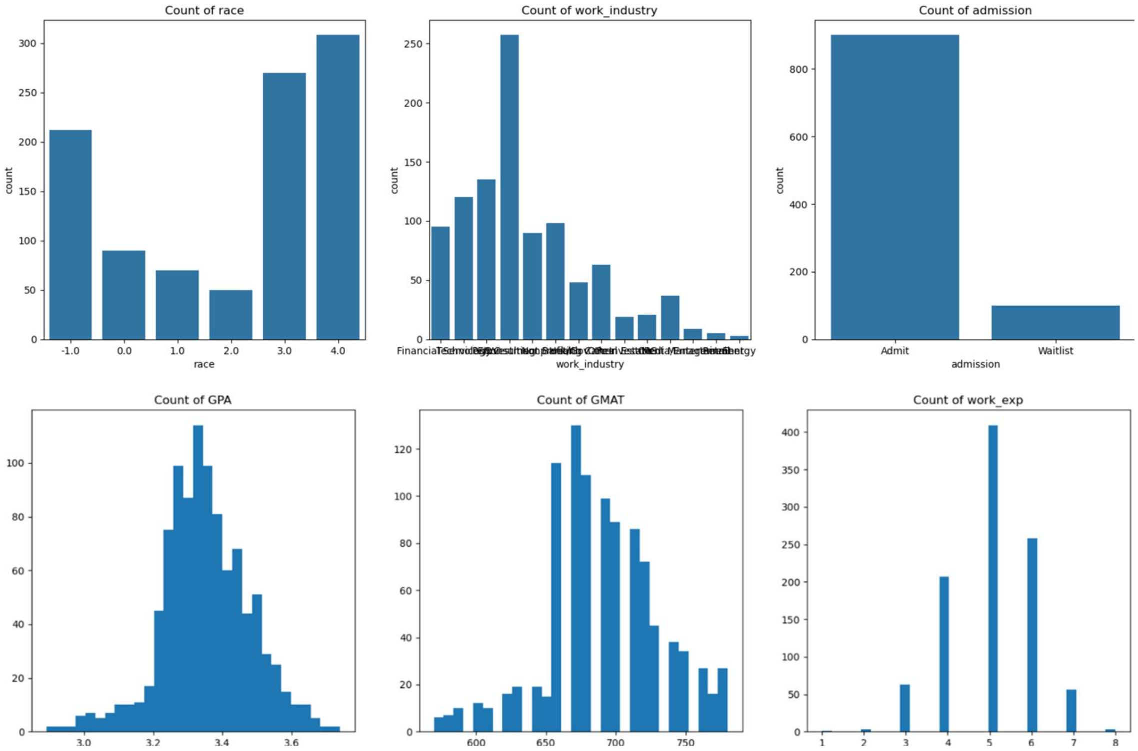
4. EXPLORATORY DATA ANALYSIS:

The exploratory data analysis (EDA) phase focused on uncovering patterns, relationships, and insights within the dataset to better understand the factors influencing garment worker productivity. A combination of univariate and bivariate analyses was performed, supported by visualizations, to comprehensively explore the data.

4.1. Univariate Analysis:

The univariate analysis provides valuable insights into the characteristics of the MBA applicant pool. GPA scores show a concentrated range, with most applicants scoring between 3.15 and 3.35 and a mean of 3.25, indicating consistently strong academic performance. GMAT scores also reflect a competitive applicant base, with an average score of 651 and scores clustering around this value, demonstrating a high level of preparation among candidates. Work experience ranges from 1 to 9 years, with an average of 5 years, suggesting that most applicants are mid-career professionals with substantial experience. Overall, the analysis reveals a relatively homogeneous pool of high-performing candidates with strong academic and professional profiles, highlighting the competitive nature of the MBA admissions process.



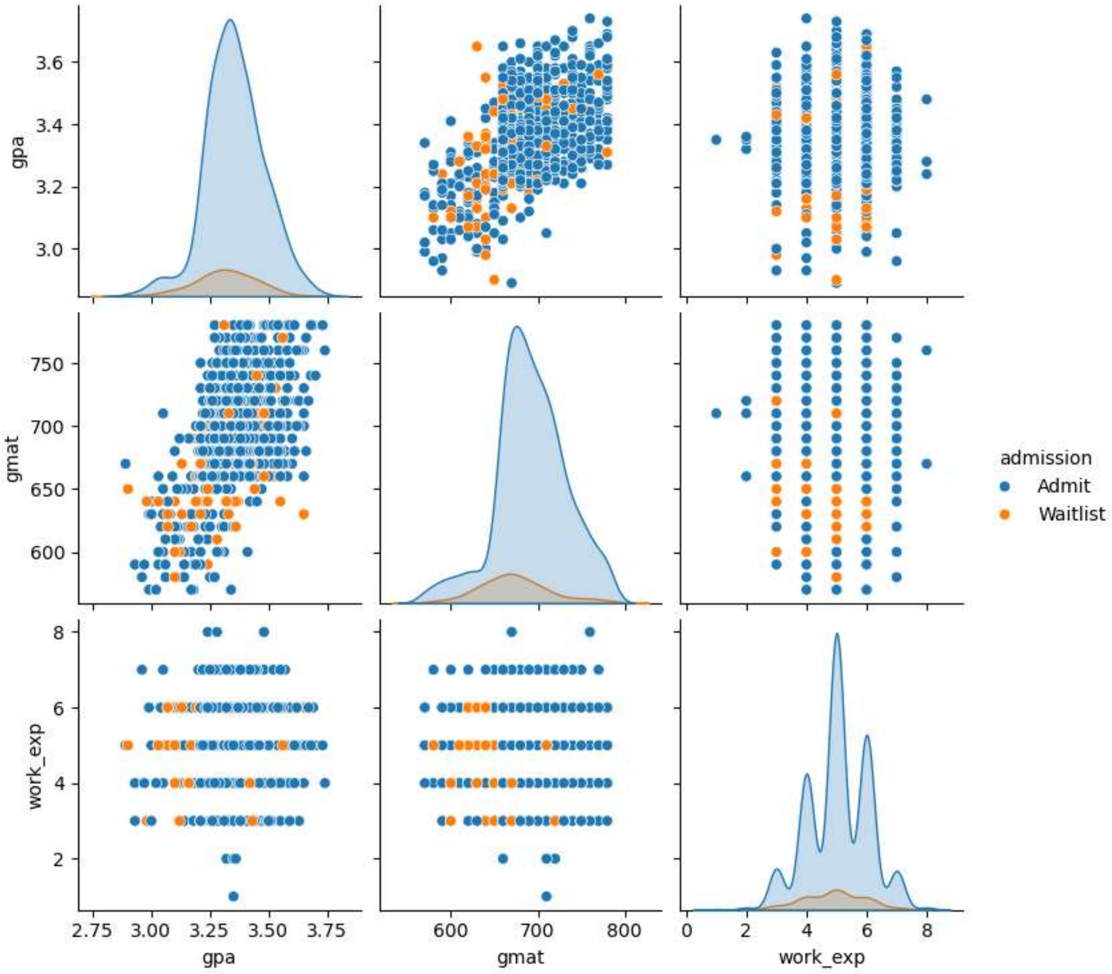


4.2. Bivariate Analysis:

The bivariate analysis uncovers several key relationships between the features in the dataset and their influence on MBA admissions outcomes. A clear positive correlation is observed between GPA and GMAT scores, suggesting that candidates who perform well academically also tend to excel in standardized tests. This alignment highlights the interconnected nature of academic capabilities and test preparation, which likely plays a crucial role in the admissions process. On the other hand, work experience shows no significant correlation with either GPA or GMAT scores, indicating that professional tenure is an independent factor that does not necessarily align with academic or test performance.

When examining admission outcomes, a strong trend emerges, candidates with higher GPA and GMAT scores are more likely to be admitted. Admissions decisions are clustered around applicants with superior performance in these metrics, emphasizing their importance as primary predictors. Meanwhile, waitlisted candidates exhibit greater variability in their GPA and GMAT scores, which could imply that these applicants are on the borderline of meeting the academic and test performance benchmarks but lack other differentiating qualities. Work experience, while not directly correlated with academic metrics, may still act as a complementary factor influencing decisions, particularly for candidates with mid-range GPAs or GMAT scores.

This analysis underscores the centrality of academic performance and GMAT scores in driving MBA admissions outcomes, while work experience may contribute more subtly, depending on individual applicant profiles. The insights suggest that admissions committees heavily weigh academic and test performance, possibly considering professional experience as a secondary but important factor for a well-rounded evaluation. This highlights the importance of balancing academic excellence with relevant work experience to maximize admission success.



5. MODEL

5.1. Model Selection:

The use of logistic regression in this project was both strategic and practical, as it aligns well with the goals of predicting MBA admissions outcomes and identifying the key factors that influence decisions. Logistic regression is particularly effective for binary classification problems like this, where the target variable (admission) has two possible outcomes (admitted or waitlisted). The model operates by estimating the probability of an outcome based on the given predictor variables, such as GPA, GMAT scores, and work experience. One of the main advantages of logistic regression is its simplicity and interpretability—it provides a clear understanding of how each predictor contributes to the likelihood of admission. This transparency makes it an ideal choice for this project, where actionable insights are as important as predictive accuracy. For admissions committees or prospective students, understanding why certain factors influence decisions is just as crucial as the decisions themselves.

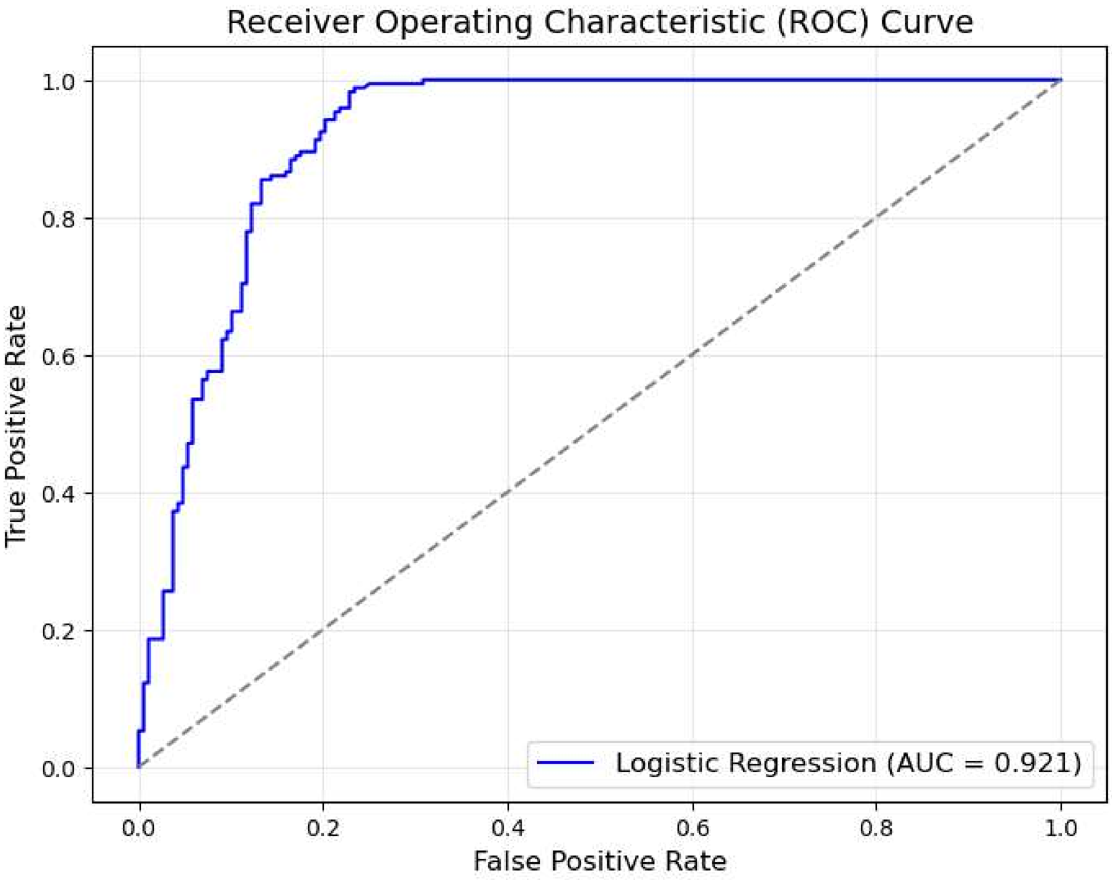
To refine the logistic regression model, backward elimination was employed as the feature selection technique. Backward elimination is a robust and systematic process that begins with all available features in the model and iteratively removes those that are statistically insignificant or contribute little to the predictive power. This is determined by assessing the p-values of each feature; those exceeding a predefined threshold are removed in successive iterations. This approach ensures that the final model retains only the most impactful predictors, enhancing its interpretability and performance. By eliminating redundant or irrelevant features, backward elimination reduces the complexity of the model and addresses potential issues like multicollinearity, where strong correlations between predictors can distort the results.

The choice of backward elimination was deliberate, as it aligns with the project's broader objective of identifying the most critical factors influencing MBA admissions. This method helped focus the analysis on features with the strongest relationships to admissions outcomes, such as, GMAT scores, and work experience, while discarding less relevant variables. By doing so, the model avoids overfitting and remains generalizable to new data, which is essential for practical applications. Furthermore, backward elimination complements the interpretability of logistic regression by narrowing the scope of the analysis to a manageable and meaningful set of predictors. This combination of simplicity, clarity, and statistical rigor makes the model highly effective for both predictive and explanatory purposes.

5.2. Results:

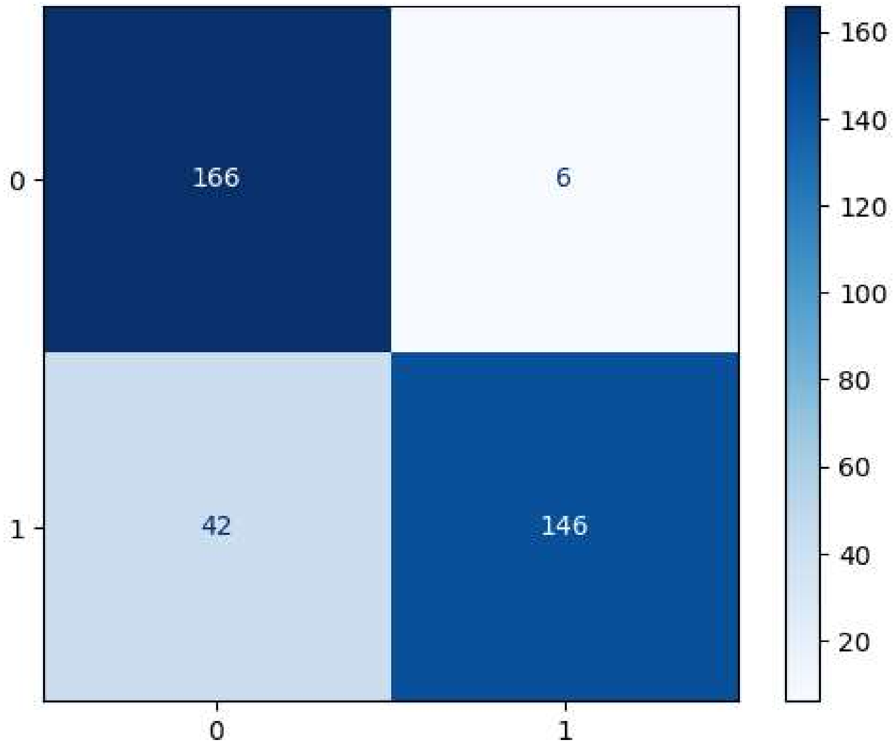
Our ROC (Receiver Operating Characteristic) curve results highlight the excellent performance of our logistic regression model in predicting MBA admissions. With an AUC (Area Under the Curve) of **0.921**, our model demonstrates outstanding discriminatory power, effectively distinguishing between applicants who are admitted and those who are not. This high AUC score indicates that our model can reliably rank applicants, assigning higher probabilities to those more likely to be admitted.

The curve reflects that our model strikes a good balance between sensitivity (true positive rate) and specificity (true negative rate) across different thresholds. However, while the overall performance is strong, we acknowledge that there may still be room for fine-tuning the balance between these metrics, especially if we prioritize minimizing missed admissions (false negatives) over other considerations.



The confusion matrix provides a detailed view of how our model performs in predicting admissions outcomes. Here's how we interpret the results:

1. **True Positives (TP)**: These are cases where our model correctly predicted admission for applicants who were actually admitted. With a high precision of **96%**, we can confidently state that the majority of our positive predictions are accurate, making our model reliable in identifying admitted applicants.
2. **True Negatives (TN)**: These are instances where our model correctly identified non-admitted applicants. The overall accuracy of ~91% demonstrates that our model is effective at distinguishing between admitted and non-admitted candidates.
3. **False Positives (FP)**: These occur when our model incorrectly predicts admission for applicants who were not actually admitted. While the number of false positives is relatively low, these errors might lead to inefficiencies, such as allocating resources to applicants who are unlikely to be admitted.
4. **False Negatives (FN)**: These are applicants who should have been admitted but were misclassified as not admitted. With a recall of **78%**, we recognize that some strong candidates are being overlooked. Reducing false negatives is a priority for us, as it would ensure that fewer deserving applicants are missed, improving both fairness and inclusivity.



6. CONCLUSION:

Our project successfully developed a logistic regression model to predict MBA admissions outcomes, achieving strong performance metrics, including an AUC of **0.921**, high precision (**96%**), and overall accuracy (~91%). These results highlight the model's ability to reliably distinguish between admitted and non-admitted applicants. The analysis emphasized the critical role of factors like, GMAT scores, work industry and work experience in influencing admissions decisions. While the model demonstrated excellent predictive power, areas for improvement, such as increasing recall to minimize false negatives, were identified. Overall, the project underscores the value of data-driven approaches in enhancing admissions processes, offering actionable insights for institutions, students, and recruiters.

7. RECOMMENDATIONS:

- **Enhance Data Balance**: Addressing class imbalance more comprehensively by gathering a more balanced dataset or exploring advanced resampling techniques beyond SMOTE can further improve model robustness.
- **Incorporate Additional Features**: Integrate new data points, such as interview scores, letters of recommendation, or essay evaluations, to provide a more holistic view of applicants and improve predictive accuracy.
- **Optimize Thresholds**: Experiment with different decision thresholds to strike an ideal balance between precision and recall, based on the institution's priorities, such as minimizing missed admissions or reducing resource allocation inefficiencies**.**

- **Explore Advanced Models**: Consider implementing more complex machine learning models, such as Random Forest, XGBoosting, or neural networks, to capture non-linear relationships and potentially improve performance further.
- **Longitudinal Analysis**: Analyze long-term outcomes, such as placement success or career progression, to validate and refine the model's predictive features for broader applicability.

8. REFERENCES:

- "Productivity Prediction of Garment Employees," UCI Machine Learning Repository, 2020. [Online]. Available: https://doi.org/10.24432/C51S6D