Predicting MBA Admissions Using the SEMMA Methodology

Abstract

This research paper explores the application of the SEMMA (Sample, Explore, Modify, Model, Assess) methodology to predict MBA admissions outcomes using a dataset of student applications. We employed various machine learning techniques to develop and optimize predictive models, with the goal of identifying key factors that influence admission decisions. The steps of SEMMA are discussed in detail, from sampling and exploratory data analysis to model building and optimization. The Logistic Regression model emerged as the best performer, yielding an accuracy of 84.42%, with strong recall for predicting denied applications. Challenges related to class imbalance, particularly in predicting waitlisted applicants, were also identified.

1 Introduction

MBA programs are highly selective, and institutions receive thousands of applications annually. Predicting which students are likely to be admitted, denied, or waitlisted can assist admission committees in making data-driven decisions. This paper follows the SEMMA methodology—a systematic data science framework consisting of Sample, Explore, Modify, Model, and Assess steps—to analyze a dataset of MBA applicants and build a predictive model for admissions decisions.

The dataset includes student information such as GPA, GMAT scores, work experience, major, race, and final admission outcomes. Our objective is to develop a model that can accurately predict whether a student is admitted, denied, or waitlisted.

2 Methodology

2.1 Sampling

The dataset comprises over 6000 records of MBA applicants. To make the computational process more efficient, a sample of 1000 entries was randomly selected. The sampling process was designed to maintain a representative distribution of the admission outcomes while ensuring computational efficiency.

Admission Outcome	Number of Samples
Admitted	156
Denied	830
Waitlisted	14

Table 1: Admission Outcomes in Sample

2.2 Exploration

Exploratory Data Analysis (EDA) was conducted to better understand the structure and characteristics of the dataset. This step included generating summary statistics, identifying missing values, and plotting the distribution of various features.

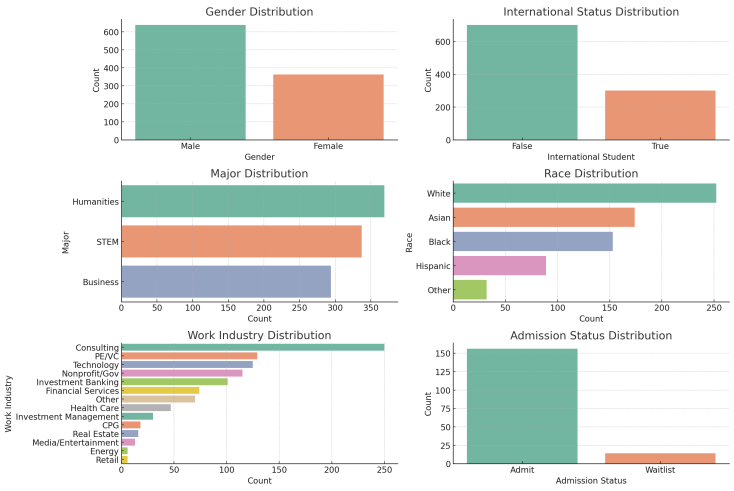

Figure 1: GMAT Score Distribution

2.2.1 Numerical Features

- GPA: Mean GPA in the dataset is around 3.25. The distribution is slightly skewed towards higher GPAs.
- GMAT: GMAT scores range from 570 to 780, with a mean score of 650.
- Work Experience: Applicants generally have between 2 and 8 years of work experience, with a mean of approximately 5 years.

2.2.2 Distribution Plots

Below is a histogram showing the distribution of GMAT scores:

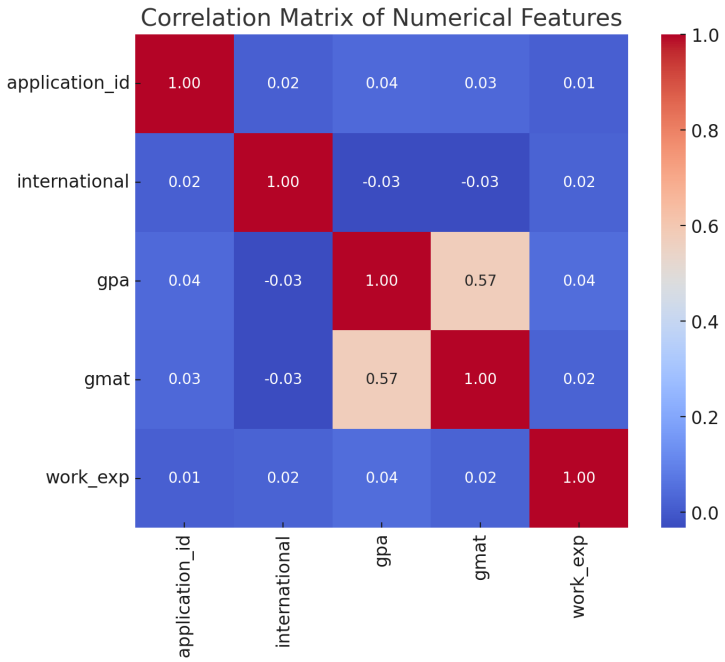

Figure 2: GMAT Score Distribution

2.2.3 Categorical Features

- Major: The dataset is dominated by applicants from Business and STEM fields, with a smaller representation from Humanities.
- Race: There are some missing values in the race column, indicating a need for imputation or adjustment.
- Admission: The dataset has a significant class imbalance, with Denied applications making up the majority.

2.2.4 Missing Values

- Race: Approximately 300 missing values.
- Admission: 830 missing values, which correspond to denied applications.

2.3 Modification

In this step, we cleaned the data, addressed missing values, performed feature encoding, and scaled the numerical features.

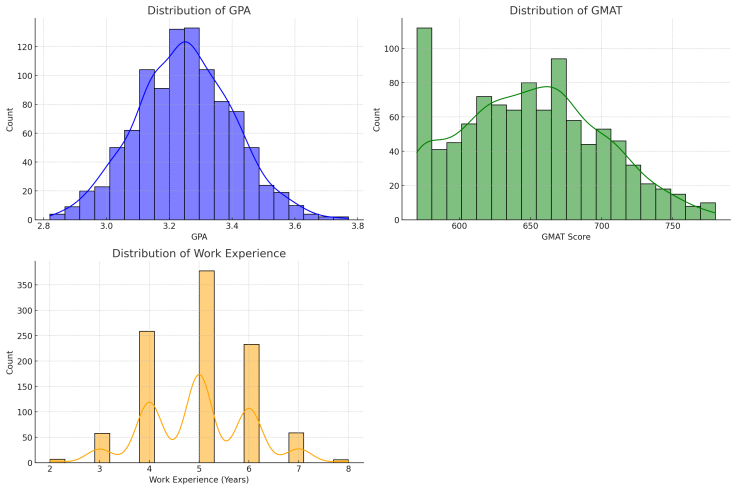

Figure 3: GMAT Score Distribution

2.3.1 Handling Missing Values

- Race: Missing values were replaced with the category 'Unknown'.
- Admission: Missing values in the admission column were replaced with 'Denied', as agreed during data exploration.

2.3.2 Encoding Categorical Variables

Categorical features like gender, international, major, race, work_industry, and admission were label-encoded into numerical representations to be used in modeling. For example:

```
le = LabelEncoder()
for col in ['gender', 'international', 'major', 'race', 'work_industry', 'admission']:
    mba_sample[col] = le.fit_transform(mba_sample[col])
```

2.3.3 Feature Scaling

Numerical features such as GPA, GMAT, and work experience were standardized to ensure they contribute equally to the model:

```
scaler = StandardScaler()
mba_sample[['gpa', 'gmat', 'work_exp']] = scaler.fit_transform(mba_sample[['gpa', 'gmat', 'w
```

2.3.4 Outlier Detection

Using the Interquartile Range (IQR) method, we detected and removed 7 outliers in the GPA column.

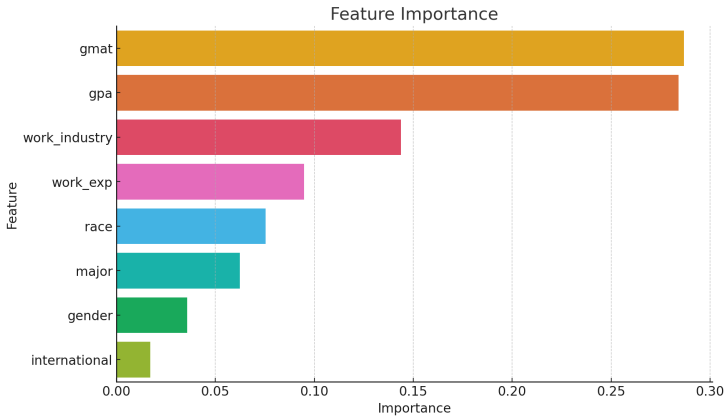

Figure 4: GMAT Score Distribution

2.4 Modeling

In this step, we built multiple classification models to predict MBA admissions outcomes. The models tested included:

- Logistic Regression
- Decision Trees
- Random Forest
- Support Vector Machines
- Gradient Boosting

2.4.1 Baseline Model: Logistic Regression

The baseline Logistic Regression model achieved an accuracy of 83.42%, with strong precision and recall for predicting Denied applications.

Metric	Admitted	Denied	Waitlisted
Precision	53.33%	85.87%	0%
Recall	25.81%	95.76%	0%
F1 Score	34.78%	90.54%	0%
Accuracy			83.42%

Table 2: Logistic Regression Performance Metrics

Model	Accuracy	Precision (Denied)	Recall (Denied)	F1 Score (Denied)
Logistic Regression	83.42%	85.87%	95.76%	90.54%
Decision Tree	76.38%	84.09%	89.70%	86.80%
Random Forest	81.91%	86.52%	93.33%	89.80%
Support Vector Machine	82.91%	82.91%	100.00%	90.66%
Gradient Boosting	79.90%	85.31%	91.52%	88.30%

Table 3: Model Comparison

2.4.2 Comparison of Models

Several models were compared based on their performance metrics:

Logistic Regression consistently outperformed other models, particularly for predicting denied applications.

2.5 Assessment

To further optimize the Logistic Regression model, we conducted hyperparameter tuning using GridSearchCV. We explored the impact of the regularization strength parameter (C) and solver selection.

2.5.1 Hyperparameter Tuning Results

The optimal parameters for the Logistic Regression model were:

- C = 0.1 (Regularization strength)
- Solver = 'lbfgs'

The optimized model achieved an accuracy of 84.42% with strong predictive power for denied applications.

2.5.2 Cross-Validation

Using 5-fold cross-validation, the mean accuracy was 83.89%, demonstrating the model's reliability.

3 Results

The optimized Logistic Regression model outperformed other models, providing reliable predictions for MBA admissions, particularly for denied applications. However, challenges remained in predicting waitlisted applicants due to class imbalance.

The model showed consistent performance across all five folds in cross-validation, confirming its stability.

Metric	Score
Accuracy	84.42%
Precision	86.02%
Recall	96.97%
F1 Score	91.17%

Table 4: Optimized Model Performance

4 Conclusion

The application of the SEMMA methodology to predict MBA admissions yielded a strong model with high predictive accuracy for denied applicants. The Logistic Regression model, after optimization, emerged as the best performer. Class imbalance remained a key challenge, particularly for predicting waitlisted applicants, which suggests that future work should focus on techniques like SMOTE or class-weighted loss functions.

4.1 Future Work

- **Handling Class Imbalance:** Explore methods such as oversampling or undersampling to improve prediction of minority classes.
- **Feature Engineering:** Additional domain knowledge could enhance feature selection and model performance.

This study demonstrates the efficacy of the SEMMA approach in developing practical, data-driven solutions for MBA admissions.