Bayesian Analysis of Data for Admission in the University

Contents

1. Introduction

Preparation is paramount in shaping one's academic journey. Students embarking on educational pursuits often navigate a labyrinth of inquiries regarding potential universities, admission processes, scholarships, and accommodations. Foremost among their concerns is securing a spot in their desired institution a gateway to realizing their aspirations.

For many, the allure of internationally recognized universities remains irresistible, with the United States emerging as a favored destination. Renowned for its prestigious colleges, a diverse array of academic offerings, and robust support systems including scholarships tailored for international students, the U.S. beckons aspirants from every corner of the globe. The country hosts a staggering population of over 10 million international students, hailing predominantly from Asian nations such as India, Pakistan, Sri Lanka, Japan, and China. Yet, it's not just America that captures the academic imagination; Canada, the United Kingdom, Germany, Italy, Australia, and Canada also witness a surge in enrollment from overseas scholars.

This surge in international pursuit of higher education is fueled by various factors, notably the scarcity of job opportunities full of fierce competition in domestic job markets. The quest for postgraduate studies thus emerges as a strategic move, equipping individuals with specialized skills and expertise to navigate an increasingly competitive professional landscape. Notably, within the realm of postgraduate education, the field of computer science stands out as a magnet for students seeking to harness the latest advancements in technology.

While the admission criteria may vary across institutions, many colleges in the U.S. adhere to standardized protocols, considering factors such as aptitude assessments(GRE) and academic track records(CGPAA etc,.). Proficiency in the English language, crucial for academic success, is often evaluated through standardized tests like TOEFL. Ultimately, the decision to admit or deny a candidate hinges on a holistic assessment of their application, encompassing academic achievements, extracurricular engagements, and language proficiency.

2. Literature review

Extensive research and development efforts have been dedicated to refining the university admission process, leveraging the power of machine learning models to aid prospective students in securing placements at their desired institutions. Past studies have explored the efficacy of various algorithms, including the Naive Bayes model, in assessing the likelihood of student acceptance into universities. However, these endeavors have encountered limitations, particularly in failing to account for all pertinent factors influencing the admission process, such as TOEFL/IELTS scores, statements of purpose (SOP), letters of recommendation (LOR), and undergraduate performance.

In response to these shortcomings, Bayesian Networks Algorithm has emerged as a promising tool for creating decision support networks tailored to evaluate the applications of foreign students. This innovative model seeks to predict the potential success of prospective students by drawing comparisons with the academic performance of current university attendees. By analyzing a range of metrics, the algorithm aims to forecast whether an applicant is likely to be admitted to the institution. Nevertheless, it is important to note a significant limitation of this approach its reliance solely on data from admitted students, without considering those whose applications were rejected. Consequently, the accuracy of predictions may be compromised, highlighting the ongoing challenges in developing comprehensive and reliable admission evaluation systems.

3. Dataset

The dataset has been sourced from Kaggle and includes various attributes related to university admissions. The dataset can be found on this site. The total number of observations is 400 and seven variables. However, we are only considering performing the analysis on GRE score, TOEFL score, and CGPA.

Table 1: Description of the Dataset

Variable	Type
GRE Scores (out of 340)	Continuous
TOEFL Scores (out of 120)	Continuous
University Rating (out of 5)	Categorical
Statement of Purpose (SOP) and Letter of Recommendation (LOR) Strength (out of 5)	Categorical
Undergraduate GPA (out of 10)	Continuous
Research Experience (either 0 or 1)	Categorical
Chance of Admit (ranging from 0 to 1)	Categorical

4. Objective

The primary aim of this research endeavor is to meticulously discern the most fitting data-generating model tailored to the multifaceted features encapsulated within this dataset. With a comprehensive array of attributes spanning student demographics, academic behaviors, and performance metrics, the task at hand necessitates a thorough exploration of various modeling approaches to accurately capture the intricate interplay between these factors. Through

systematic analysis and comparison of candidate models, this study endeavors to unravel the underlying patterns and relationships inherent within the dataset, shedding light on the optimal framework for generating data that aligns most closely with observed empirical phenomena. By identifying the most suitable data-generating model, this research not only enhances our understanding of the complex dynamics governing student outcomes.

5. Methodology

5.1 Preprocessing

Before delving into the analysis of the dataset, a meticulous examination was conducted to ascertain the presence of missing values, ensuring the integrity and completeness of the data. This meticulous scrutiny revealed no instances of missing data across the entire dataset, affirming its robustness and reliability for subsequent analyses. Following this preliminary quality assessment, the dataset was partitioned into distinct training and testing subsets, adhering to a stratified sampling approach to preserve the representative nature of the data distribution. Seventy-five percent of the dataset was allocated to the training subset, serving as the foundation for model development and parameter estimation, while the remaining twenty-five percent constituted the testing subset, facilitating the evaluation of model performance and generalization capabilities on unseen data. By partitioning the dataset in this manner, the integrity of the modeling process is upheld, enabling rigorous validation and assessment of predictive models against real-world data scenarios.

5.2 Descriptive Analysis

A comprehensive summary encompassing GRE score, TOEFL score, and CGPA features within the dataset was meticulously computed, offering a detailed overview of their central tendencies, dispersions, and distributions. This encompassed statistical measures such as mean, median, standard deviation, minimum, and maximum values, providing valuable insights into the inherent variability and characteristics of each feature. Subsequently, density curves were employed as a powerful visualization tool to delve deeper into the distributional properties of the features. These density curves facilitated a nuanced exploration of the shape, skewness, and modality of each feature's distribution, enabling a deeper understanding of their underlying patterns and structures.

5.3 Analysis – Approach 1

Given the apparent symmetry observed in the distribution of features within the dataset, the initial exploration focused on fitting a normal distribution as the data-generating distribution, with a fixed standard deviation. This approach aimed to leverage the assumption of normality to model the underlying structure of the data, thereby facilitating subsequent inference and analysis. In tandem with this endeavor, two distinct prior distributions were considered: one characterized by a normal distribution(Tried with different combinations of mean and standard deviation) and the other adopting a flat prior proportional to 1 and 100. The utilization of these differing prior distributions allowed for a comprehensive exploration of the impact of prior beliefs and assumptions on the resulting posterior inference. By systematically evaluating the performance of each prior in conjunction with the normal data-generating distribution, a nuanced understanding of the interplay between prior specification and model outcomes was achieved, paving the way for robust Bayesian inference and decision-making processes.

Data-generating distribution is as follows,

$$P(X_{1:n}|\mu) \sim N(\mu,\ \sigma^2)$$

Two priors as follows,

1. $P(\mu) \sim N(\mu_1, \sigma_1^2\ )$
2. $P(\mu) \propto 1$

The methodology adopted in this study entails a meticulous computation of the posterior distribution based on the training data, encapsulating the updated beliefs about model parameters following the incorporation of observed data. This posterior distribution serves as a foundational component for subsequent analyses. Leveraging this posterior distribution, the posterior predictive distribution for the testing data was systematically derived, enabling the generation of probabilistic forecasts for unseen data instances. Furthermore, to quantify the uncertainty associated with these predictions, a 95% credible interval was computed for the posterior predictive distribution, providing a robust measure of prediction uncertainty. This credible interval serves as a probabilistic range within which the true values of the target variable are expected to lie with a specified level of confidence. Finally, to gauge the efficacy of the predictive model, the number of testing samples falling within the calculated credible interval was meticulously tallied, offering valuable insights into the model's calibration and accuracy in capturing the underlying data distribution.

Posterior distribution is as follows,

$$P(\mu|X_{1:n}) \propto P(X_{1:n}|\mu) * P(\mu)$$

Posterior predictive distribution is as follows,

$$P(X_{New}|X_{1:n}) = \int P(X_{\text{new}}|\mu) * P(\mu|X_{1:n})$$

An exhaustive examination was conducted, systematically testing diverse combinations of mean and standard deviation parameters to prior distribution and flat prior the intricate distributional nuances inherent within the dataset. This rigorous exploration encompassed a broad spectrum of parameter values, spanning a range of locational and scale properties, in order to comprehensively assess their impact on the resultant distributional shapes and characteristics

Prior	Posterior	Posterior Predictive
$P(\mu) \sim N(\mu_1, \sigma_1^2)$	$P(\mu\|X_{1:n}) \sim N(\mu_2, \sigma_2^2)$ $\sigma_2^2 = \frac{\sigma_1^2\sigma^2}{n\sigma_1^2 + \sigma^2}, \mu_2 = \frac{n\sigma_1^2\bar{X} + \mu_1\sigma^2}{n\sigma_1^2 + \sigma^2}$	$P(X_{new}\|X_{1:n}) \sim N(\mu_3, \sigma_3^2)$ $\mu_3 = \mu_2$ $\sigma_3^2 = \sigma_2^2 + \sigma^2$
$P(\mu) \propto 1$	$\mu_2 = \mu\, P(\mu\|X_{1:n}) \sim N(\mu_2, \sigma_2^2)$ $\sigma_2^2 = \sigma^2$	$P(X_{new}\|X_{1:n}) \sim N(\mu_3, \sigma_3^2)$ $\mu_3 = \mu_2$ $\sigma_3^2 = \sigma_2^2 + \sigma^2$

6. Results

6.1 Results - Descriptive Analysis

Table 2 Descriptive Statistics

Feature	Median	Mean	Standard Deviation
CGPA	8.640	8.625	0.584
GRE score	317.5	317.2	11.408
TOEFL score	108.0	107.7	6.014

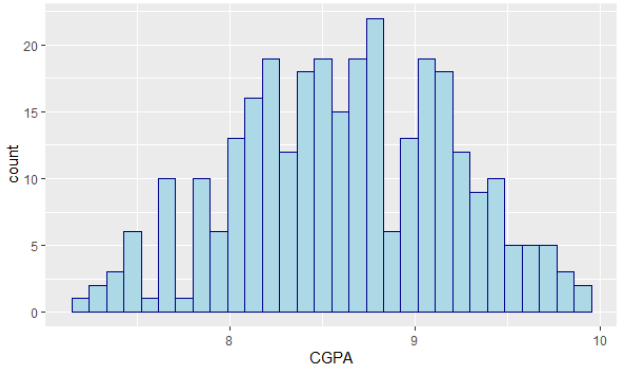

Figure 1: Distribution of CGPA

Figure 2: Distribution of GRE Score

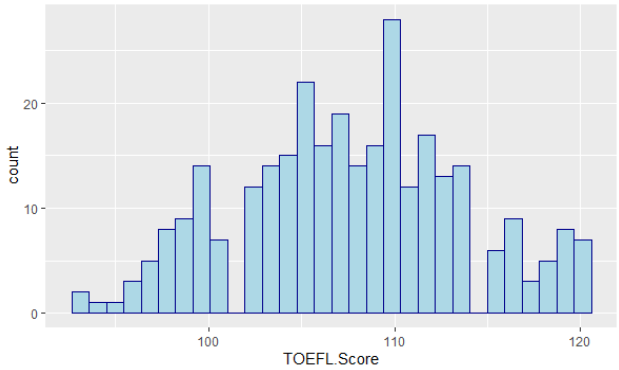

Figure 3: Distribution of TOEFL Score

All three features seem to have an approximate symmetric distribution.

6.1 Results - Analysis – Approach 1

A comprehensive array of graphical summaries was generated to elucidate the intricacies of the posterior predictive distribution and the number of test samples inside the 95% credible interval.

- CGPA

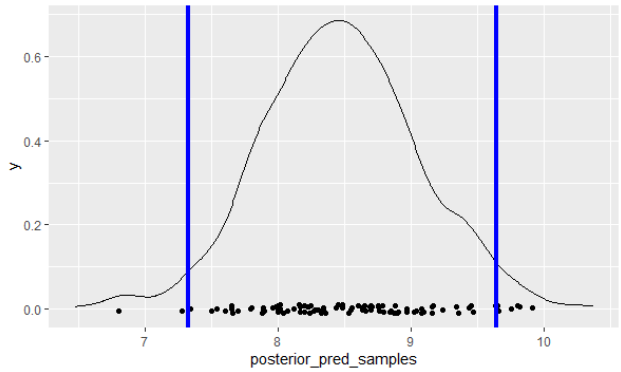

Figure 5: Posterior Predictive Distribution for CGPA $P(X\_(1:n)|\mu)\sim N(\mu, 0.35)$ & $P(\mu)\sim N(2,1)$

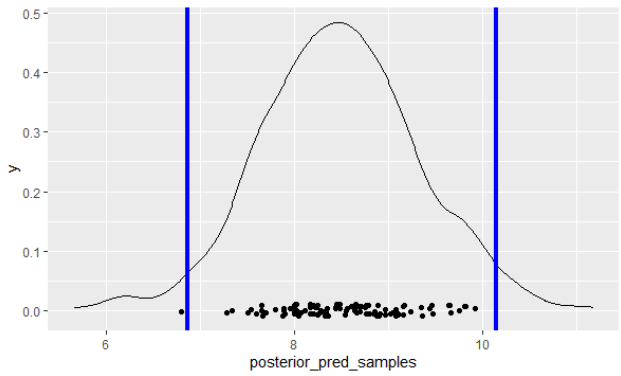

Figure 4: Posterior Predictive Distribution for CGPA $(X\_(1:n)|\mu)\sim N(\mu, 0.35)$ & $P(\mu)\sim 1$

- GRE Score

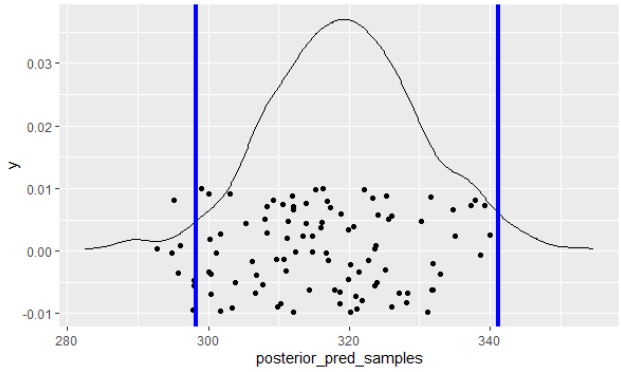

Figure 7: Posterior Predictive Distribution for GRE Score - $P(X\_(1:n)|\mu)\sim N(\mu, 120)$ & $P(\mu)\sim N(250,100)$

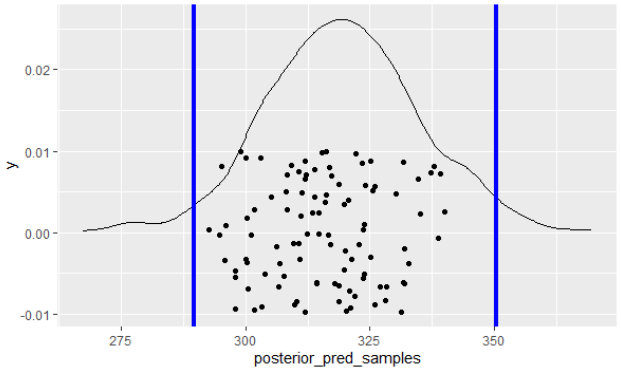

Figure 6: Posterior Predictive Distribution for GRE Score $(X\_(1:n)|\mu)\sim N(\mu, 120)$ & $P(\mu)\sim 1$

- TOEFL

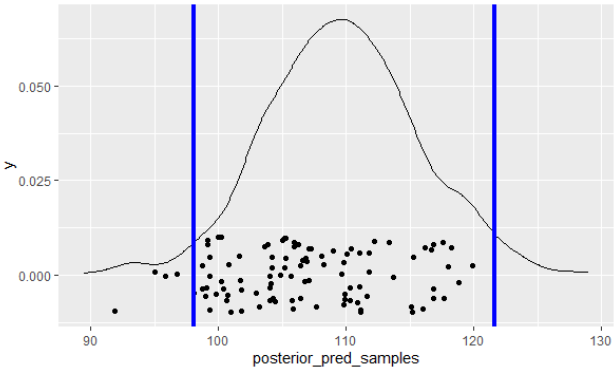

Figure 9: Posterior Predictive Distribution for TOEFL score- $P(X_{1:n}|\mu) \sim N(\mu, 36)$ & $P(\mu) \sim N(100,25)$

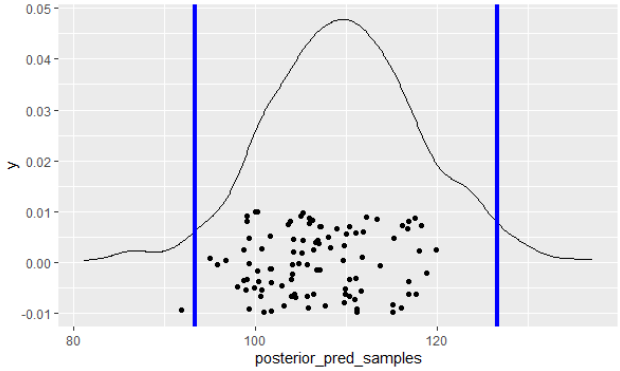

Figure 8: Posterior Predictive Distribution for TOEFL score- $(X_{1:n}|\mu) \sim N(\mu, 36)$ & $P(\mu) \sim 1$

Table 3: Summary results of posterior predictive distribution of CGPA

Data generating model	**Prior**	**Number of samples inside the credible interval**
$P(X_{1:n}\|\mu) \sim N(\mu, 0.35)$	$P(\mu) \sim N(2,1)$	92
$P(X_{1:n}\|\mu) \sim N(\mu, 0.35)$	$P(\mu) \propto 1$	99
$P(X_{1:n}\|\mu) \sim N(\mu, 0.35)$	$P(\mu) \sim N(20,10)$	92
$P(X_{1:n}\|\mu) \sim N(\mu, 0.35)$	$P(\mu) \sim N(200,100)$	92
$P(X_{1:n}\|\mu) \sim N(\mu, 0.35)$	$P(\mu) \propto 100$	99

Table 4: Summary results of posterior predictive distribution of GRE

Data generating model	Prior	Number of samples inside the credible interval
$P(X_{1:n}\|\mu)\sim N(\mu,\ 120)$	$P(\mu)\sim N(250,100)$	92
$P(X_{1:n}\|\mu)\sim N(\mu,\ 120)$	$P(\mu) \propto 1$	100
$P(X_{1:n}\|\mu)\sim N(\mu,\ 120)$	$P(\mu)\sim N(2.5,1)$	95
$P(X_{1:n}\|\mu)\sim N(\mu,\ 120)$	$P(\mu)\sim N(2500,1000)$	92
$P(X_{1:n}\|\mu)\sim N(\mu,\ 120)$	$P(\mu) \propto 100$	100

Table 5: Summary results of posterior predictive distribution of TOEFL

Data generating model	Prior	Number of samples inside the credible interval
$P(X_{1:n}\|\mu)\sim N(\mu,\ 36)$	$P(\mu)\sim N(100,25\ )$	95
$P(X_{1:n}\|\mu)\sim N(\mu,\ 36)$	$P(\mu) \propto 1$	99
$P(X_{1:n}\|\mu)\sim N(\mu,\ 36)$	$P(\mu)\sim N(1,0.25\ )$	96
$P(X_{1:n}\|\mu)\sim N(\mu,\ 36)$	$P(\mu)\sim N(1000,250\ )$	95
$P(X_{1:n}\|\mu)\sim N(\mu,\ 36)$	$P(\mu) \propto 100$	99

The meticulous exploration of various parameters for the prior distribution revealed a consistent pattern across all iterations: the 95% credible intervals of the posterior predictive distribution consistently encompassed the same number of test samples. This striking consistency serves as a compelling indicator suggesting that a normal distribution may indeed serve as a fitting model for the data generating process. The robustness of this finding underscores the potential suitability of the normal distribution in accurately capturing the underlying data dynamics and generating reliable predictions. Furthermore, the consistent containment of test samples within the credible intervals across different parameter configurations lends credence to the robustness and stability of the modeling approach. This convergence of results across diverse prior specifications bolsters confidence in the suitability of the normal distribution as a viable framework for modeling the

observed data, thereby providing valuable insights for subsequent analyses and decision-making processes.