**OUTLINE**

[**1. Executive Summary 2**](#_kppoqkktuwqb)

[**2. Background/ Context 2**](#_mcj1jov1d31p)

[Domain 2](#_1qjucgs2t7qn)

[Brief description of the scenario 2](#_pwayhm9xebxm)

[Decisions of interest 2](#_yll1z9lqj4dg)

[Decision makers 2](#_atwv6i8dpzdo)

[**3. Business Understanding 3**](#_3iweu71xmps)

[Business Objective 3](#_a0uyfkw37thh)

[Situation Assessment 3](#_herq4z2iymv9)

[Data Mining Goals 3](#_22s8j25egj8i)

[**4. Data Understanding 4**](#_agv5con5t6ey)

[Data Requirements 4](#_gzcezfmwxnmf)

[Data Description 4](#_6yere5j6cg4g)

[Data Sources 4](#_rmnotqkec93f)

[Data Quality 4](#_a72rxbs1dli7)

[**5. Data Preparation 4**](#_kwas30ba27f5)

[Data selection 4](#_az2y20p7xvqb)

[Data cleaning 5](#_t7irgngkzej4)

[Data preparation 5](#_he3pzzeq2krt)

[**6. Modeling ‐ Building Models 5**](#_rj9b2xglv8vy)

[Describe data in detail 5](#_ugqsyktorrdk)

[What type of decision making model(s) is appropriate for the decision‐making tasks? 5](#_7ynlo3rp9ge3)

[Provide rationale for choice of model(s) 5](#_5cbcqnn2zr4x)

[Detail model development and output 6](#_p4zbchhdckpn)

[**7. Model Evaluation 7**](#_6oa850dyz44g)

[**8. Discussion 8**](#_hnhlwdceodbk)

[**9. Conclusion 10**](#_9sr8utsrhwrb)

[**Appendix 11**](#_s1ol0kwmw8uq)

# 

# **Executive Summary**

This project is centered on optimizing the college admissions process using data-driven techniques. Our objective was to assist higher education institutions in making more efficient and informed decisions regarding the admission and enrollment of prospective students.

# **Background/ Context**

## **Domain**

The project falls within the domain of the college admissions application review process and feedback mechanisms for decisions made during this process.

## **Brief description of the scenario**

This project aims to support higher education institutions in refining their admissions and enrollment procedures, ultimately leading to saving their valuable time and resources. Higher education institutions regularly receive a substantial volume of applications from diverse backgrounds, including applicants with varying parental backgrounds, high-school performance, standardized test scores, and ethnic backgrounds. Our objective is to assist admission committees in efficiently filtering this extensive pool of applications based on historical admission decisions. Additionally, we aim to identify applications that are less likely to result in enrollment after an admission offer, allowing committees to save valuable effort on selecting students that are less likely to enroll. These data-driven decisions not only enhance the efficiency of the admissions process but also optimize enrollment numbers while upholding academic standards.

## **Decisions of interest**

The primary decision of interest in this project concerns mainly two things:

1. The admissions and enrollment process of higher education institutions. This enables institutions to streamline their human resources and ensure a smooth and efficient process.
2. The project also encompasses decisions related to targeted marketing efforts and enhancing student support services to improve enrollment and retention rates.

## **Decision makers**

The major decision factors in this project will be parent education, high school and standard test scores, gender, and ethnicity.

Multiple stakeholders within the institution will be involved, including:

*Admissions Committee*- They will be responsible for the final evaluation of applications and issuing admissions offers. The project will use data predictive models that will assist the Admissions Committee in identifying the most promising applicants.

*Enrollment management*- They will be responsible for enhancing the retention rates of the accepted applicants.

*Marketing and Outreach team*- They will be responsible for improving the market strategy of the institution by targeting high-quality applicants.

# **Business Understanding**

## **Business Objective**

To provide a data-driven admission and enrollment optimization strategy to colleges to make their application process more efficient, less time consuming and assisting them in selecting bright students that will help these institutes in improving academic quality for better retention rates in future.

## **Situation Assessment**

Dataset contains data from 17,339 applicants of three colleges namely Arts & Letters, Business & Economics, and Math & Science of an undergraduate university in North America. Variables included in the data are demographics of the applicants, the education of both parents of the applicants, performance in the high school of the candidates, and finally whether the student opted to enroll in the colleges after getting an admission offer from the colleges. In the case of applicants enrolling in the college, performance of the students in these colleges after enrollment.

If we look at the demographics of the students who applied to these applicants, we have 42% of males and 58% of females. Ethnicities wise, we have 53% White applicants, 17% Asians and ethnicity of 30% applicants is not provided in the population data set. [[1]](#footnote-1)Moreover, 40% of applications were from the Arts & Letters college, 36% from Math & Science and only 24% of the entire population applied to Business & Economics college. The majority of students’ both parents have 4 years of college degree and postgraduate.

## **Data Mining Goals**

We have a supervised dataset of 5,323 applicants who successfully cleared the application process and received offers from these three colleges but only 1,528 applicants decided to enroll. As we do not know about the other factors that most applicants were considering while making their college decision, therefore we cannot predict the GPA of these students based on their demographics and other available variables. Our project will be focused on the sample of 1,528 students who are enrolled in the three colleges.

Our data mining goals are.

* Clustering the data based on the demographics of applicants.
* Classification of parents in seven education level categories
* Predicting the impact of the education level of 1 parent on the decision of enrollment in a college by the student.
* Predicting the impact of education level of both parents on the decision of enrollment in a college by the students.
* Predicting impact of all predictor variables on the admission and enrollment of the students.

The following Models /Methods will be used in the project.

· Logistic Regression

· Classification And Regression Tree

# **Data Understanding**

## **Data Requirements**

The focus of the project is to streamline the admissions and enrollment process of a university. For this the data related to their high school GPA, SAT scores, gender, Ethnicity, and the College that they have applied to would be needed.

## **Data Description**

The college admissions dataset consists of data from 17,339 students. The dataset has 12 attributes and it is split between the 3 colleges where the students have applied to the university.

## **Data Sources**

The college admissions dataset was obtained from McGraw Hill- Business Analytics Communication with numbers by Sanjiv Jaggia**.**

## **Data Quality**

The dataset overall is complete except for the College\_GPA attribute where 15,811 applicants do not have a GPA. This is because the dataset on a whole has the data of both enrolled and not enrolled students, and the college GPA is only available for those students who have completed 4 years of enrollment. Along with the college GPA the parents’ education field does not have the description of the value 5. 2,151 rows have value 5 in their parent’s education field. So, we cannot predict what it represents for the applicants.

# **Data Preparation**

## **Data selection**

* The selection of data was driven by relevance to the project’s objectives. We focused on quality and number of applicants encompassed in data. This dataset had good performance indicators such as (High School GPA, SAT/ACT scores), demographic information (gender, ethnicity), and parental educational background.
* The dataset encompassed 17,339 applicants across three colleges: Arts & Letters, Business & Economics, and Math & Science. It included a good ratio of successful and unsuccessful admission applications.

## **Data cleaning**

* The dataset was thoroughly examined for missing values using different functions in Excel and R. Special attention was given to the ‘College\_GPA’ attribute, where a significant number of entries were missing.
* The parents’ education field had an unexplained category labeled ‘5’. We treated this category with caution.

## **Data preparation**

Targeted variables:

* **Admitted:** Binary variable indicating whether an applicant was admitted (‘Yes’) or not (‘Not’).
* **Enrolled:** Binary variable indicating whether an admitted applicant enrolled (‘Yes’) or not (‘Not’).

Attribute Data Type and handling:

* The attributes were categorized as nominal, categorical and ratio.
* The categorical variables were converted into ordered categories, while sex and ethnicity were handled as nominal categories.

Splitting Data for Model Training and Testing:

* The dataset is divided into training and testing sets, ensuring a balanced representation of the various categories in both sets. This helped in unbiased model training and evaluation.

# **Modeling ‐ Building Models**

## **Describe data in detail**

The college admissions dataset consists of data from 17,339 students who applied to three colleges namely Arts & Letters, Business & Economics, and Math & Science of an undergraduate university in North America. The dataset has seven nominal, two categorical and three ratio variables. Nominal variables are Applicant ID, Gender, ethnicity, college name, status of admission, status of enrollment. Categorical variables are education of parents. And ratio variables are College GPA of Students, High School GPA of Students, SAT/ACT score of students.

## **What type of decision-making model(s) is appropriate for the decision‐making tasks?**

* Logistics Regression Model
* Classification And Regression Tree Mode**l**

## **Provide rationale for choice of model(s)**

*Reasons to select Regression Model:*

The First and foremost reason is its capability to incorporate many predictor variables and evaluate the relationship of all variables with respect to the response variable. Moreover, the logistic regression model doesn’t assume a linear relationship between predictors and response variable hence it gives us a chance to find the strength of relationships and eliminate the variable that has negligible influence on the target variable.

Considering that we had numerical, ratio and categorical types of variables present in our data and the Logistic Regression Model can work with these types of data, that’s why we have selected this model for our data.

*Reasons to select CART Model:*

We are using Classification Tree for our project because it is a more convenient model to interpret data without getting into more complex Machine Learning Models. Secondly, Data visualization is more structured in this model. Just looking at the Decision tree we can reveal about the relationship and impact of other variables on the target variable. Thirdly, this model also accommodates non-linear relations present in the data as well. We can also check their influence on target variables without running separate models for these variables. Lastly, Classification trees are also not disrupted by the outliers present in the data. Therefore, it will not impact our output at large and our observations will be relevant to similar kinds of dataset.

## **Detail model development and output**

*Logistic Regression Model*

We first developed a model for admission prediction based on exploratory data analysis, using all variables available in the dataset. Based on the summary of the model we decided the best variables to consider and the variables that are insignificant like White, Asian and sex to eliminate.

We developed another model with the selected variables and validated the model with a randomly split dataset into a training set and validation set. For logistic regression predicting admitted decision, model 4 is best with accuracy 81.75%, sensitivity 89.04%. The equation of the model is AdmittedNumeric ~ Edu\_Parent1 + Edu\_Parent2 + HSGPA + SAT\_ACT. The data is splitted into a 70-30 ratio for training and validating. For logistic regression predicting enrolled decision, model 4 gives an accuracy of 80.90%, sensitivity of 20.27% and specificity of 94.69%, performing the best in comparison to other models. The equation of the model is EnrolledNumeric ~ Edu\_Parent1 + Edu\_Parent2 +White+ HSGPA + SAT\_ACT+College. The cutoff for classifying predicted probabilities in logistic regression is 0.5. We decided on this value after testing the model with cutoff values in range [0.3,0.8]. We got the best accuracy at 0.5.

*Classification Tree Model*

Similarly, we have two classification models for Admitted and Enrolled Students. Details of both models are:

*CART for Admitted*

We have run a Classification Model for response variable Admitted and predictor variables are Education of Parent 1, Education of Parent 2, Sex of student, Ethnicity of student, High School GPA of student, SAT Score and College. We have initially created a Full Classification tree, but working on Full tree not only creates the issue of overfitting, it also requires more resources and scalability. Therefore, we have worked on pruned trees. That eliminated all the variables that were insignificant to our analyses. The root node is High School GPA, the Internal node is SAT\_ACT and branches are education of parents and ethnicity.

In the second step, we have run a Confusion Matrix to look at the Accuracy, Sensitivity, Specificity and Prevalence of our model to analyze how accurate our model is regarding the predictability of validation data. At the default cutoff value of 0.5, accuracy of the model is 81.46% and Sensitivity is 59.09% whereas, with the new cut off value of 0.307 we have increased the sensitivity to 77.13% but the accuracy of the model has decreased to 76.85%. We will still go with our model with a cutoff value of 0.307 as it corresponds to our data more than the first model.

*CART for Enrolled*

We have also implemented a classification model for the response variable "Enrolled” using the predictor variables that represent education level of parent 1, education level of parent 2, sex, white or not, Asian, or not, high school GPA, SAT/ACT scores and College. Note, that here we have filtered the dataset to include only those applicants who have gotten an admit (Admitted = “Yes”). Since the number of applicants who had enrolled was far less than the number of applicants who did not enroll, we oversampled the dataset so that the overall class balance is achieved to represent both cases and avoid bias. We created a full classification tree using this and pruned it to suit our needs. The pruned tree was used to model the data. We ran a confusion matrix with a default cutoff value of 0.50 on the validation set, the classification tree had an accuracy of 66.75%, sensitivity of 64.18% and specificity of 69.32%.

# **Model Evaluation**

For Admitted variables, we have higher accuracy rate, sensitivity that is 81.75%, 89.04%, in Logistic Regression as compared to CART model, therefore, we are selecting this model.

Similarly, for Enrolled we have accuracy as 76.77%, and sensitivity as 77.88% in the Logistics Model which is higher than the CART model. Same can be seen in our graph below. Therefore, we are selecting logistic regression instead of CART for both admitted and enrolled models.

# **Discussion**

Based on the Model, what would your decision/recommendation be? Why?

* The admissions team should deprioritize the sex and Asian demographic categories in predictive models, as they are not significant predictors for admission and enrollment.
* Consider admissions for applicants to Math & Science and Business & Economics, given their higher likelihood of enrollment compared to Arts & Letters College.
* Students whose parents have a college undergraduate or postgraduate degree can be given preference, as they are more likely to enroll. Parent’s educational background emerged as a significant predictor of student enrollment, indicating its utility in targeted recruitment strategies.

What are the limitations of the Model you have used?

* The model is built on a data set from a specific set of colleges and limited number admission and enrollment decisions. This might not represent the overall population of the institution. Therefore, it is difficult to generalize for all applications of the institution.
* The intake semester is also not defined which also might play a significant role in the number of applications and quality of applications.
* These models and findings might not be applicable to other institutions with different demographics or academic focuses.
* These models do not account for changing trends in education and student preferences over a period of time.
* Enrollment decisions could be affected by other unmeasured factors like financial aid provided, other offers available to students, or the overall quality of the domain.

What cognitive biases would you expect (most likely) to influence the decision‐making process? How does decision support mitigate some/all of these?

* If the decisions are made by the admission committee without any data driven insights, there may exist personal biases influenced by personal preferences, stereotypes, or anecdotal experiences. By using our logistic regression and CART models, decisions are based on analyzed data, reducing subjective judgements.
* Committees might tend to overvalue traditional test scores and GPA of students and overlook parents’ background and other factors. Our models will provide a holistic view of an applicant which has been modeled ensuring decisions are made on quantifiable and consistent criteria.
* The Admissions committee might favor applicants that have similar backgrounds to those of the institutions’ students or alumni. This will not only lead to lack of diversity but also discourage other students from applying. This might happen knowingly or unknowingly. Given a classification from our models, the deserving candidates and potential enrolling candidates will be filtered, avoiding chances of above biasness.

What enhancements would you aim for to enable better decision support for this task?

* We would suggest additional factors like financial aid requirements, past achievements, extracurricular activities, any notable referrals, session of application, alternative preferences etc. This will help to make decisions more holistically and can be generalized better.
* We would like to implement a feedback mechanism from the admissions team on the model’s performance. There might be a possibility that the model classifies a potential candidate as not qualified enough. This feedback can help to improve the model.
* We would also recommend developing a minimally intuitive user interface for the institution to feed in data and model to reflect the outcomes and decisions.
* The model should be updated regularly by considering current trends and patterns from student behavior perspective, patterns in higher education and institutions’ preferences.

# **Conclusion**

This project successfully demonstrates how data-driven models can revolutionize the college admissions process. By effectively predicting the likelihood of admission and enrollment, the models can guide institutions in making more informed decisions, thus optimizing both their resource allocation and the quality of their student body. Future enhancements and adaptations of these models promise even greater efficiency and accuracy in the ever-evolving landscape of higher education.

# **Appendix**

**Appendix-A**

A Summary of parents’ education of all applicants is as below.

|  |  |  |
| --- | --- | --- |
| **Education level of parents** | **Parent-1**  **% of Total population** | **Parent-2**  **% of Total population** |
| 1-No high school | 5% | 5% |
| 2-Some high school | 4% | 4% |
| 3-High school graduate | 12% | 13% |
| 4-Some College | 14% | 15% |
| 5-No Description | 5% | 8% |
| 6-4 years college | 32% | 34% |
| 7-Postgraduate | 27% | 20% |

**Appendix-B**

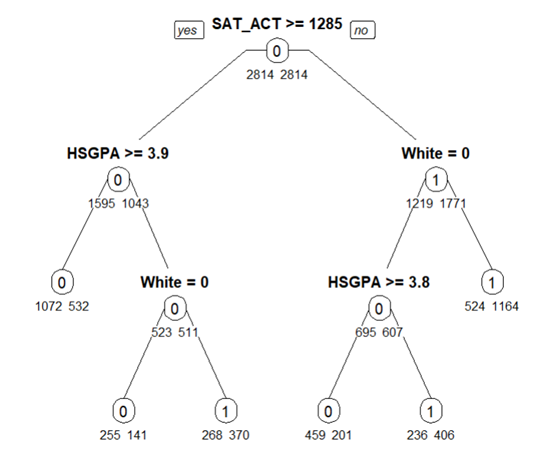
|  |  |  |
| --- | --- | --- |
| **Attributes** | **Description or Possible values** | **Variable Type** |
| Applicant | A modified ID for each Applicant | Nominal |
| Edu\_Parent 1 | 1- No high School  2- Some High School  3- High School Graduate | Categorical |
| Edu\_Parent 2 | 4- Some College  6- 4 Year College Graduate  7- Postgraduate | Categorical |
| Sex | M- Male  F-Female | Nominal |
| White | 1- White  0-Not White | Nominal |
| Asian | 1- Asian  0- Not Asian | Nominal |
| HSGPA | High School weighted GPA, ranging from 0-5 | Ratio |
| SAT/ACT | The highest of the SAT/ACT score is first converted into the equivalent SAT score for English & Math | Ratio |
| College | Arts & Letters  Business & Economics  Math & Science | Nominal |
| Admitted | 1- Admitted by college  0- Not admitted by college | Nominal |
| Enrolled | 1- Applicant enrolled  0- Applicant not enrolled | Nominal |
| College\_GPA | College GPA, ranging from 0-4, four years after enrollment; blanks for those who did not enroll | Ratio |

**Appendix -C**

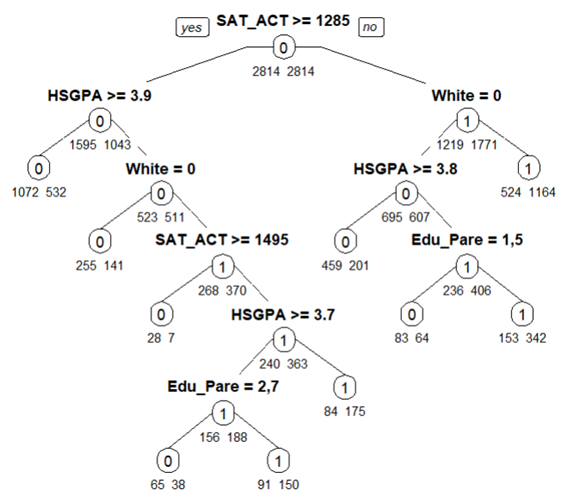
|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

|  |  |
| --- | --- |
| **Default Tree CART Admitted** | **CART Admitted Pruned Tree** |

**Default Tree CART Enrolled**



**CART Enrolled Pruned Tree**



1. A summary of dataset is available in Appendix A [↑](#footnote-ref-1)