**Business and Data Understanding**

The university admissions office in the US and wants to assess the quality of its admission program. It has pledged to nurture an appreciation for diversity and inclusion (gender, education) and select highly skilled students who could be placed well. The management wants to assess if needs to revise the school’s admission criteria to meet its strategic objectives.

The data pertains to a single student who applied to the course. The data structure is shown below:

|  |  |  |  |
| --- | --- | --- | --- |
| Variable / Feature / Attribute | Usage | Type | Values |
| admit | Target | Categorical | [0, 1] |
| gre | Input | Continuous | 0-800 |
| gpa | Input | Continuous | 0-4 |
| rank | Input | Categorical | [1, 2,3,4 ] |
| gender | Input | Categorical | [Male,Female] |

**Data Preparation**

Missing values, outliers, binning, multicollinearity

|  |  |  |
| --- | --- | --- |
| **Variable** | **Data Quality Issue** | **Potential Handling Strategies** |
| admit | No Issues | - |
| gre | No Issues | - |
| gpa | No Issues | - |
| rank | No Issues | - |

**Modeling**

**Diagnostic model**:

Logistic Regression is used to identify the students who will be admitted. We will model the probability of admission based on the input variables and identify their influence on admission.

**Predictive model**:

Logistic Regression is used to identify the students who will be admitted. We will model the probability of admission based on the data available. The model will be used to rank the probabilities. The priority will be to discriminate the top

The model does not involve any costs as of now. It is used as a diagnostic model. Generalizability is assessed using hold-out data.