

## **Business case : Jamboree Education<sup>1</sup>**

### **Table of Contents**

#### **1. Problem Statement**

#### **2. Data Profiling**

#### **3. Exploratory Data Analysis**

- a. Null value Check
- b. Outlier Handling
- c. Assumptions of Linear Regression
  - i. Check for Linearity using Pair Plots and Correlation
  - ii. Check for Normality of residuals
  - iii. Check for Heteroskedasticity
  - iv. Check for Multicollinearity

#### **4. Training Models**


- a. Linear Regression Model
- b. Polynomial Regression Model
- c. Ridge Regression Model (**L2 Regularization**)
- d. Lasso Regression Model (**L1 Regularization**)

#### **5. Observations and Conclusions**

#### **6. Business Insights**

**\*\*** [Link to google colab provided as a footnote.](#)

---

<sup>1</sup>  Business\_Case : Jamboree .ipynb

## 1. Problem Statement

Jamboree, a leading test prep company, recently launched a feature on its website that predicts a student's probability of admission to Ivy League colleges from an Indian applicant's perspective. The goal is to understand the key factors influencing graduate admissions and develop a predictive model that estimates the likelihood of acceptance based on applicant attributes.

Our analysis will help Jamboree:

- Identify the most significant factors affecting admissions.
- Understand the relationships among these factors.
- Build a reliable predictive model using linear regression techniques.

The dataset consists of various features such as GRE and TOEFL scores, university ratings, undergraduate GPA, research experience, and statement of purpose strength. Using exploratory data analysis and regression modeling, we will analyze patterns, test assumptions, and derive actionable insights to improve admission predictions.

## **2. Data Profiling**

### **Dataset Overview**

- The dataset consists of 500 rows and 9 columns.
- Each row represents an applicant's data, containing scores, ratings, and other factors influencing admission.

### **Column Descriptions**

- GRE Score (Integer): GRE scores range from 290 to 340.
- TOEFL Score (Integer): TOEFL scores range from 92 to 120.
- University Rating (Integer): Ratings range from 1 to 5.
- SOP (Statement of Purpose Strength) (Float): Rated between 1.0 and 5.0.
- LOR (Letter of Recommendation Strength) (Float): Rated between 1.0 and 5.0.
- CGPA (Undergraduate GPA) (Float): Ranges from 6.8 to 9.92.
- Research (Binary: 0 or 1): Indicates whether an applicant has research experience.
- Chance of Admit (Float): Probability value between 0.34 and 0.97.

### **Missing Values & Data Quality**

- No missing values were detected (all columns have 500 non-null values).
- Data types are appropriate for analysis (integers and floats).

### **Statistical Summary**

- Mean GRE Score: 316.47, with a standard deviation of 11.30.
- Mean TOEFL Score: 107.19, with a standard deviation of 6.08.
- Mean CGPA: 8.58, indicating a generally high GPA distribution.
- Research Experience Distribution: 56% of applicants have research experience (1), while 44% do not (0).

### 3. Exploratory Data Analysis

#### 3.a) Null Value Check

No missing values in any column, ensuring completeness of data.

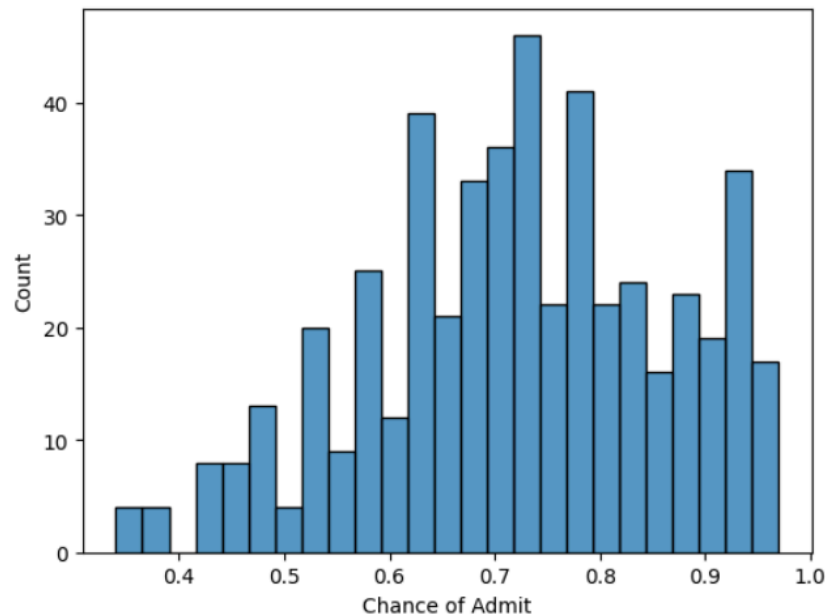
```
#Checking for NULL values  
print(df.isnull().sum())
```

```
Serial No.      0  
GRE Score       0  
TOEFL Score     0  
University Rating 0  
SOP            0  
LOR            0  
CGPA           0  
Research       0  
Chance of Admit 0  
dtype: int64
```

#### 3.b) Univariate Analysis

Distribution of Chance of Admit:

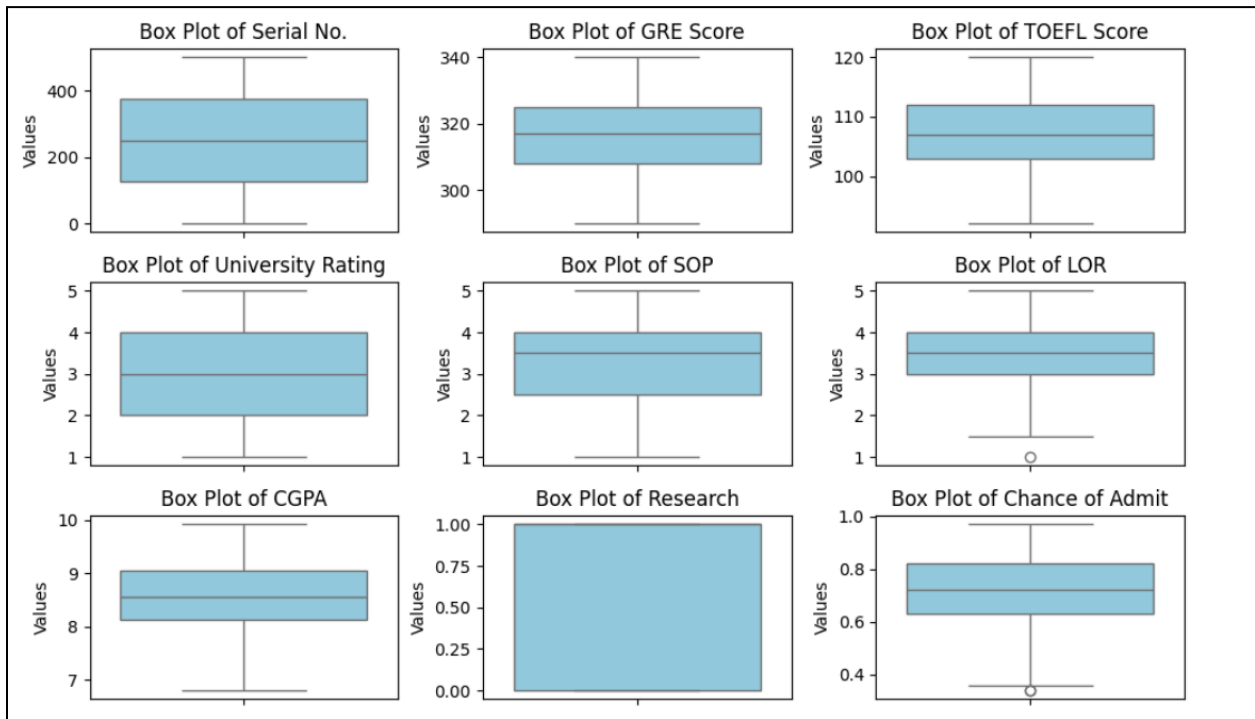
- A histogram with 25 bins shows the distribution of **Chance of Admit**.
- Helps understand the probability distribution of admission chances.



### 3.b) Outlier Handling

Box Plot Analysis for Numerical Features:

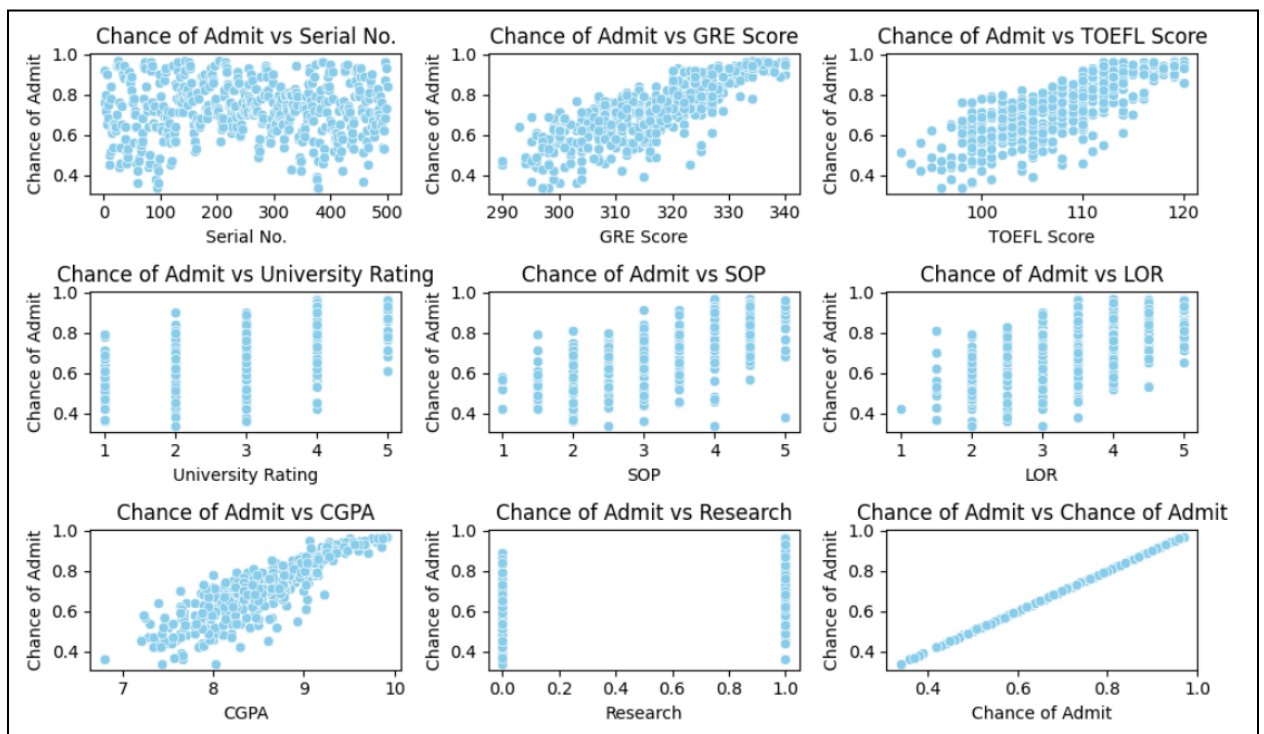
- Box plots were generated for all numerical variables.
- No significant outliers were observed, indicating clean data.
- No extreme values that require immediate treatment.



### 3.c) Assumptions of Linear Regression

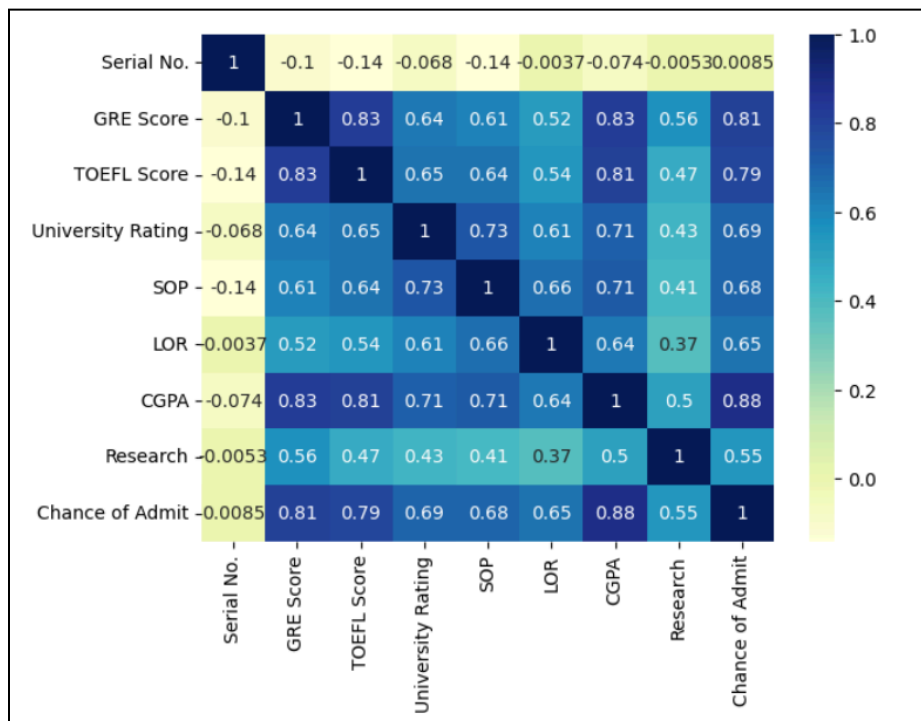
#### 3.c.i) Linearity Check using Scatter Plots

- Scatter plots were generated for all numerical features against **Chance of Admit**.
- Observations:
  - **GRE Score**, **TOEFL Score**, and **CGPA** show a strong linear relationship.
  - **SOP** and **LOR** exhibit moderate linearity.
  - **Research** and **University Rating** do not show clear linear trends.
  - **Serial No.** is irrelevant and can be dropped.



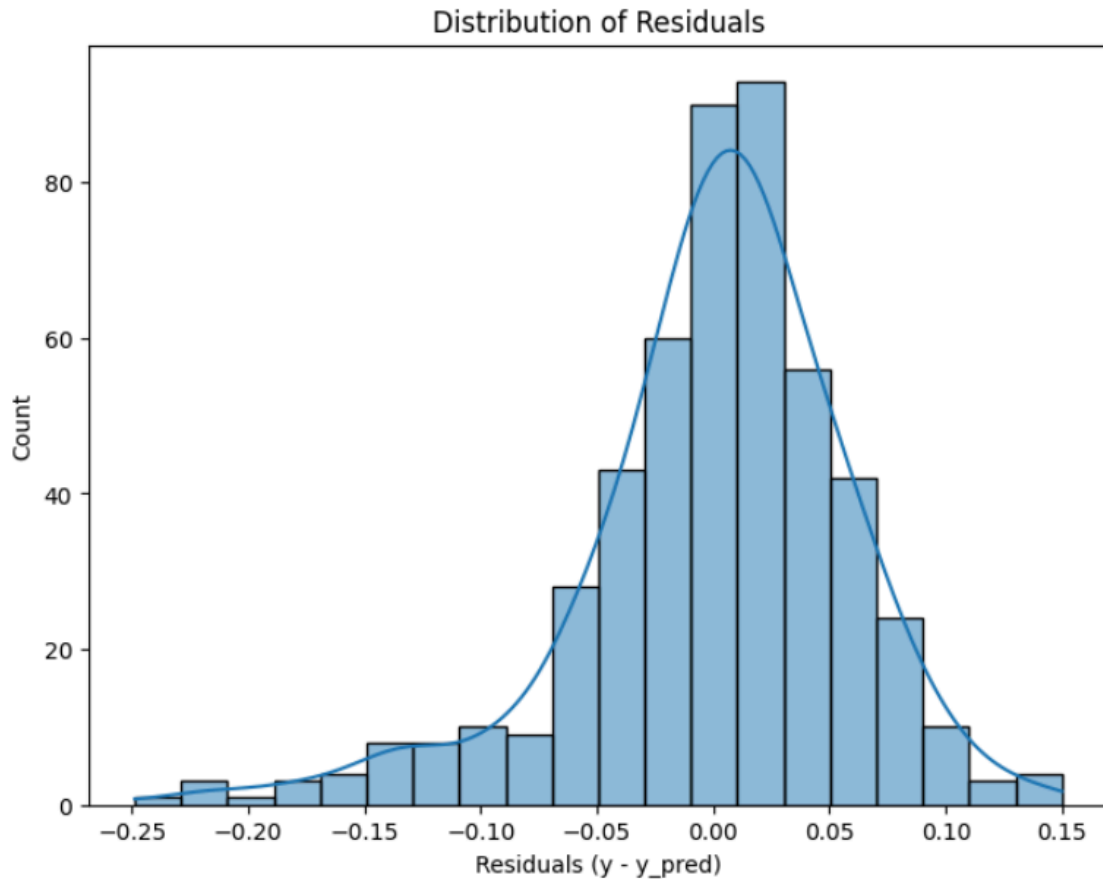
## Correlation Analysis

- A correlation heatmap was generated to quantify relationships.
- **CGPA**, **GRE Score**, and **TOEFL Score** show high correlation with **Chance of Admit**.
- **SOP**, **LOR**, and **University Rating** exhibit moderate correlation.
- No strong negative correlations observed.



### 3.c.ii) Normality of Residuals

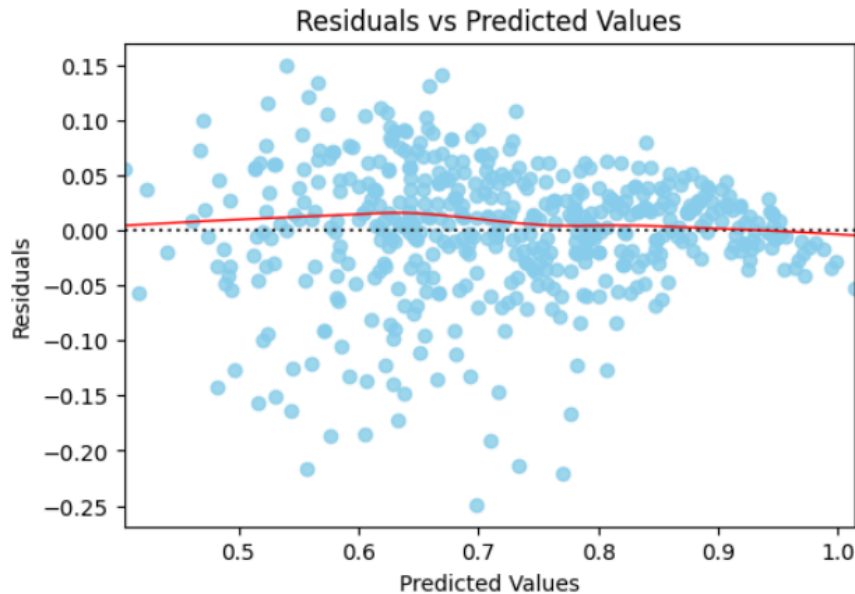
- Histogram of residuals confirms a fairly normal distribution.
- Indicates that the model's errors are symmetrically distributed.
- Supports assumptions for linear regression and gradient descent convergence.





### 3.c.iii) Homoscedasticity Check (Residuals vs Predicted Values)

- Residual plot does not show a pattern, confirming homoscedasticity.
- No clear funnel shape, meaning variance is constant across predictions.



### 3.c.iv) Multicollinearity Check (Variance Inflation Factor )

- VIF values for all predictors are below 5.
- No evidence of strong multicollinearity.
- Ensures stable and interpretable regression coefficients.

	Feature	VIF
1	GRE Score	4.464249
2	TOEFL Score	3.904213
3	CGPA	4.777992
4	University Rating	2.621036
5	SOP	2.835210
6	LOR	2.033555
7	Research	1.494008

## 4. Training Models

### a. Linear Regression Model

- **Feature Selection & Data Splitting**

- `Serial No.` dropped as it is irrelevant.
- `Chance of Admit` set as the target variable (`y`).
- Data split into 80% training and 20% testing using `train_test_split()`.

- **Model Training**

- Standardized input features using `StandardScaler()` to ensure proper scaling.
- Used `LinearRegression()` within a pipeline to streamline preprocessing and modeling.
- Model trained on the training dataset.

- **Model Evaluation ( $R^2$  Score)**

- Achieved a  $R^2$  score of 0.8188 on the test set.
- Indicates that 81.88% of the variance in `Chance of Admit` is explained by the model.
- Suggests a strong linear relationship, but there is room for improvement.

- **Next Steps**

- Try Polynomial Regression to capture non-linear relationships.
- Feature Engineering: Create interaction terms or higher-order transformations.
- Consider Regularization: Use Ridge/Lasso regression to handle multicollinearity if needed.

## **b. Polynomial Regression Model**

- **Data Splitting Strategy**

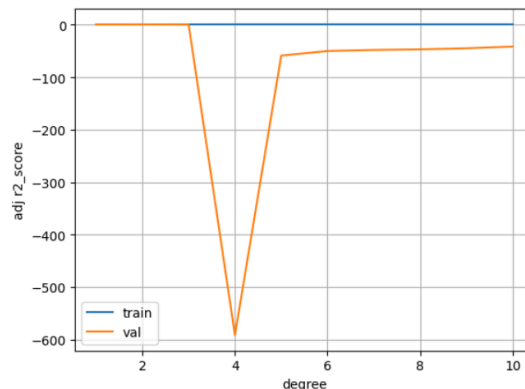
- Data split into training (60%), validation (20%), and test (20%) sets.
- Ensures proper hyperparameter tuning without overfitting on the test set.

- **Adjusted R<sup>2</sup> Calculation**

- Adjusted R<sup>2</sup> used instead of R<sup>2</sup> to account for feature complexity.
- Formula:
- $\text{Adjusted } R^2 = 1 - \frac{(n - p - 1)(1 - R^2)}{n - 1}$  where  $n$  = number of samples,  $p$  = number of predictors.

- **Hyperparameter Tuning for Polynomial Degree**

- Tested polynomial degrees from 1 to 10.
- Train & validation scores plotted to find the optimal degree.
- Best degree = 1, meaning higher-degree polynomial features do not improve performance.



- **Model Training & Performance**

- Polynomial Regression with degree = 1 is trained.
- Achieved R<sup>2</sup> score of 0.8188 on test data.
- Same performance as Simple Linear Regression, indicating that non-linearity is not significantly contributing to prediction improvement.

- **Conclusion & Next Steps**

- Polynomial features >1 do not help much in this case.
- Explore Regularization (Ridge/Lasso) to see if small improvements can be made.

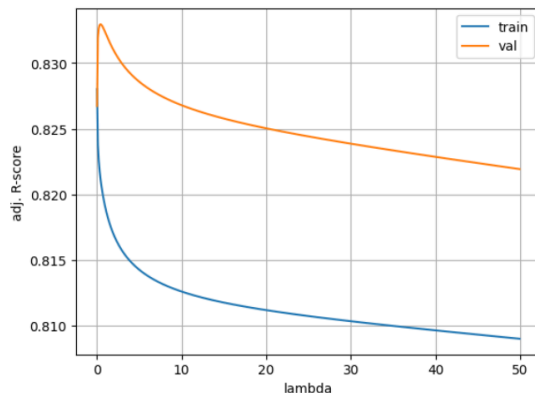
### c. Ridge Regression Model (L2 Regularization)

- **Data Splitting Strategy**

- Data split into training (60%), validation (20%), and test (20%) sets.
- Ensures proper model selection without overfitting.

- **Hyperparameter Tuning for Alpha ( $\lambda$  - Regularization Strength)**

- Lambda values tested: 0.01 to 50 with a step of 0.1.
- Train & validation scores plotted to find the optimal alpha value.
- Best lambda ( $\alpha$ ) = 0.41, where validation performance is maximized.



- **Model Training & Performance**

- Ridge Regression with  $\alpha = 0.41$  is trained.
- Achieved  $R^2$  score of 0.8188, slightly lower than Simple Linear Regression (0.81884).
- Indicates regularization does not improve performance for this dataset.

- **Intuition Behind Regularization's Impact**

- Regularization penalizes large coefficients, reducing model complexity.
- Since polynomial degree = 1, the model is already simple, and regularization unnecessarily restricts it.
- Regularization is more beneficial when dealing with high-degree polynomial features or multicollinearity issues.

- **Conclusion & Next Steps**

- L2 Regularization is unnecessary for this problem.
- Explore Lasso Regression (L1 Regularization) to check if feature selection helps.

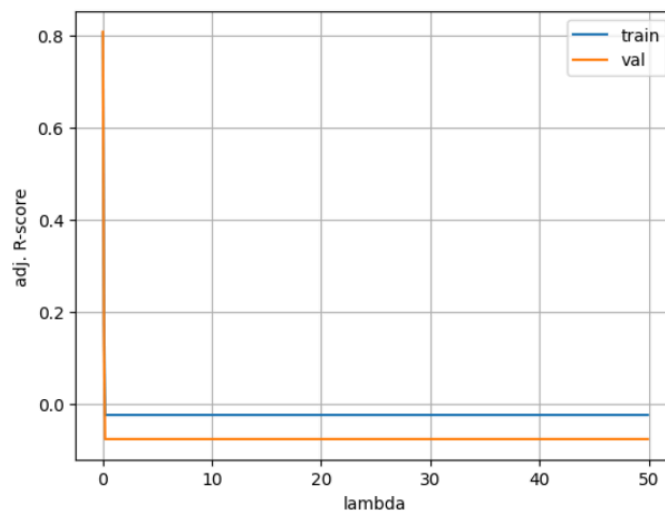
#### d. Lasso Regression Model (L1 Regularization)

- **Data Splitting Strategy**

- Training (60%), Validation (20%), and Test (20%) data split.
- Ensures model tuning without data leakage.

- **Hyperparameter Tuning for Alpha ( $\lambda$  - Regularization Strength)**

- Lambda values tested from 0.01 to 50 with a step of 0.1.
- Train & validation scores plotted to identify optimal alpha value.
- Best alpha ( $\lambda$ ) = 0.01, where validation score is highest.



- **Model Training & Performance**

- Lasso Regression with  $\alpha = 0.01$  is trained.
- Achieved  $R^2$  score of 0.8139, which is lower than Ridge Regression (0.8188) and Simple Linear Regression (0.8188).
- Reason for performance drop:
  - L1 regularization shrinks some feature coefficients to zero.
  - This results in feature elimination, potentially removing important predictors.
  - Unlike Ridge, which only shrinks coefficients, Lasso forces some to be exactly zero.

- **Feature Importance (Sparse Feature Property of Lasso)**
  - Some coefficients are exactly 0, meaning those features were eliminated.
  - Lasso performs feature selection, which can be useful in high-dimensional datasets but may harm performance if essential features are removed.
  
- **Conclusion & Next Steps**
  - Lasso is not helpful here since the dataset does not have many irrelevant features.
  - Regularization is unnecessary when using only first-degree polynomial features.

## 5. Observations and Conclusions

	Model	HyperParameters	R2 Score	Observation
1	Simple Linear Regression	No hyperparameters (Baseline Model)	<b>0.8188</b>	<b>Baseline model, decent performance.</b>
2	Polynomial Regression	Polynomial Degree = 1	0.8188	Higher-degree polynomials didn't improve performance.
3	Ridge Regression (L2)	Polynomial Degree = 1, Alpha ( $\lambda$ ) = <b>0.41</b>	0.8188	Regularization had no impact, unnecessary for low-degree polynomials.
4	Lasso Regression (L1)	Polynomial Degree = 1, Alpha ( $\lambda$ ) = <b>0.01</b>	0.8139	Performed worst due to feature elimination (some coefficients became 0).

### Observations:

1. **Simple Linear Regression performed well** with an  $R^2$  score of **0.8188**, indicating a strong linear relationship between features and the target variable.
2. **Polynomial Regression with higher degrees didn't improve performance**, suggesting that the dataset does not benefit from additional polynomial features.
3. **Ridge Regression (L2 Regularization) had no impact**, showing that regularization is unnecessary when the model complexity is low (degree = 1).
4. **Lasso Regression (L1 Regularization) performed the worst** ( $R^2$  = **0.8139**) due to its feature elimination property, which likely removed useful predictors.

### Conclusions:

1. **Regularization (L1 & L2) is unnecessary** when using simple linear models with well-scaled features.
2. **Higher-degree polynomial features do not improve model performance**, indicating that the data follows a mostly linear pattern.
3. **Feature selection and transformation methods should be explored** instead of polynomial expansion or regularization.

## 6. Business Insights

### Identify Key Admission Factors to Optimize Student Preparation

- The regression models highlight the most influential factors affecting admission chances (e.g., GRE scores, SOP, and CGPA).
- **Business Impact:** Jamboree can prioritize coaching efforts on these key areas, refining course materials and training students for maximum admission success.

### Personalized Admission Counseling

- Since simple linear regression already predicts well, each student's admission probability can be estimated based on their current profile.
- **Business Impact:** Jamboree can offer data-driven personalized counseling, recommending students focus on specific areas (e.g., improving SOP or GRE scores) to boost their admission chances.

### Targeted Marketing for Course Enrollment

- By analyzing student data, Jamboree can identify high-potential applicants who need minimal improvement vs. those requiring extensive coaching.
- **Business Impact:** Adjust pricing models and marketing efforts—offer premium mentorship for borderline cases and standardized prep for high-potential students.

### Predictive Analytics for Student Success

- The performance of different models indicates that past student profiles can predict future admission trends.
- **Business Impact:** Jamboree can forecast admission rates for different universities and adjust course offerings accordingly, ensuring alignment with real-world trends.

### Competitive Differentiation via AI-Driven Insights

- Advanced modeling and insights can be used as a USP (Unique Selling Proposition) in the competitive test-prep industry.
- **Business Impact:** Jamboree can market AI-driven admission prediction tools to attract students looking for a data-backed approach to admissions, differentiating itself from competitors.



