**Jamboree Education - Linear Regression**

Problem Statement :

Jamboree is an organization that helps students to secure admission to various top tier colleges via various competitive examinations. The streamlined approach of the organization have helped students to deliver maximum score with minimal effort.

A new feature of checking the probability of selection into the IVY league colleges has been introduced. We need to understand what are the factors that are important for the selection procedures and how are they inter-related.

Exploratory Data Analysis

Q) Observation of data and its column names

A screen shot of a computer

AI-generated content may be incorrect.

Q) Shape of the data set

Ans)

A black and green text

AI-generated content may be incorrect.

Insights : There are 500 rows and 9 columns present in the data set.

Q) Describe the statistical summary of the data set ?

Ans)

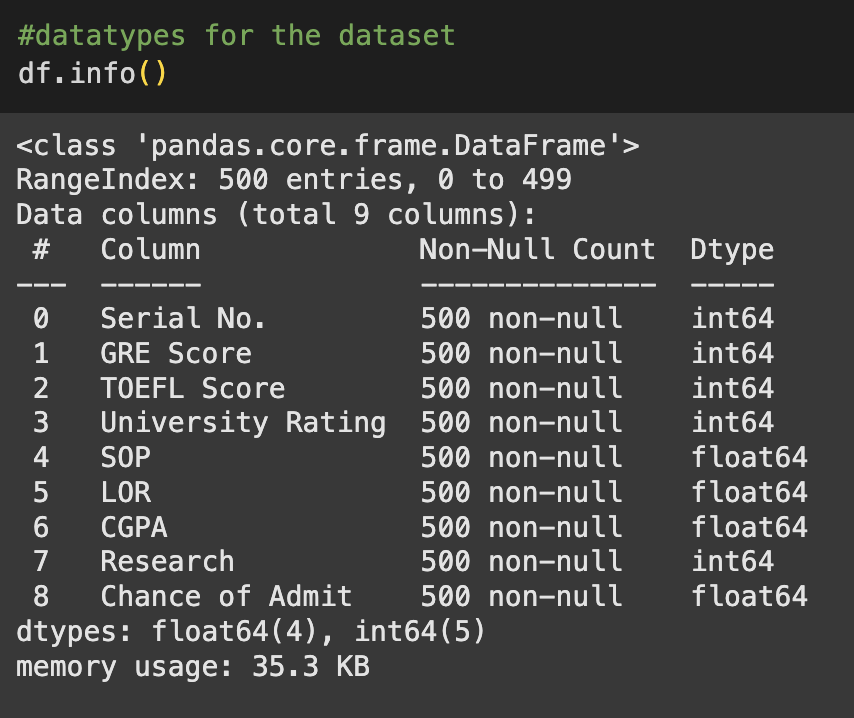
A screen shot of a computer

AI-generated content may be incorrect.

Insights : Serial No. is a unique identifier for the data set hence avoiding that the statistical summary of the features are listed out.

Q) What are the data types for the various columns of the data set ?

Ans)



Insights : there are total of 9 columns of which 5 of them are integer data type and rest 4 of them are float datatype.

Q) List the details of the missing values in the data set ?

Ans )

Insight : There are no missing values present in the given data set. This means that the data set is complete and the information can be successfully extracted without any assumptions to be done at record level.

A screenshot of a black screen

AI-generated content may be incorrect.

Q) List out the Duplicate values in the data set ?

Ans)

A screen shot of a computer code

AI-generated content may be incorrect.

Insights : There are no duplicate rows in the given data set. This means that the data that is under consideration is all unique and there is no error that can happen due to duplicated values

Q) Analyse the variables and do the univariate analysis for the data set?

Ans)

A screenshot of a computer screen

AI-generated content may be incorrect.

A group of graphs showing different types of data

AI-generated content may be incorrect.

Insights : GRE Score : The values are spread across the range 290 - 340 with the max value near 320 with a count of around 70 and comes as low as near to 5

TOEFL Score : The values are spread across the range of 0-120 with the max. Value near 105 with a count. Near to 100 and lowest value near to 2.

University Rating - the values are spread across the range of 1-5 with the max. Rating of the universities between 3-3.5 while the lowest rating is across the band of 1-1.5 The highest rating has a count of near 160 while the lowest rating band has a count near to 40.

SOP strength- the values are spread across 1-5 with the max values in the range of 4-4.5 and 3.25 - 3.5 while the lowest chances are in the range of 1-1.5 with a count of around 5. The highest range has the count near to 100.

University rating and SOP strength show varied distributions, suggesting that applicants apply to universities of different ratings and have varied SOP strengths.

Distribution of CGPA - the values are spread across 6.5 - 10 with the max values between 8.5 - 9 and the lowest near to 7.0. The max count coms around 80 while the lowest is near to 1.This has an almost bell shaped curve subscribing to a normal distribution pattern for the feature.

Chance of Admit - the values are spread across 0.2 - 1.0 with the max values near 0.7 and the lowest near 0.4 i.e. there is almost 70% probability on average for the chance of admission and a 40% chance for rejection on average. This distribution is also left-skewed, suggesting a generally high likelihood of admission among the applicants in the dataset.

Q) Count plot for Research experience ?

Ans )

Insights : The Research experience count says that there are over 250+ candidates who have research experience while a little of 200 do no have the research experience.

A screenshot of a computer screen

AI-generated content may be incorrect.

Bivariate Analysis

Q) Compute the correlation matrix

Ans)

A screenshot of a computer screen

AI-generated content may be incorrect.

Insights : This graph shows the correlation of various features on to the admission into the IVY league colleges. The correlation is represented via heatmap with color grading ranging from blue to red where red shows highest degree of correlation.

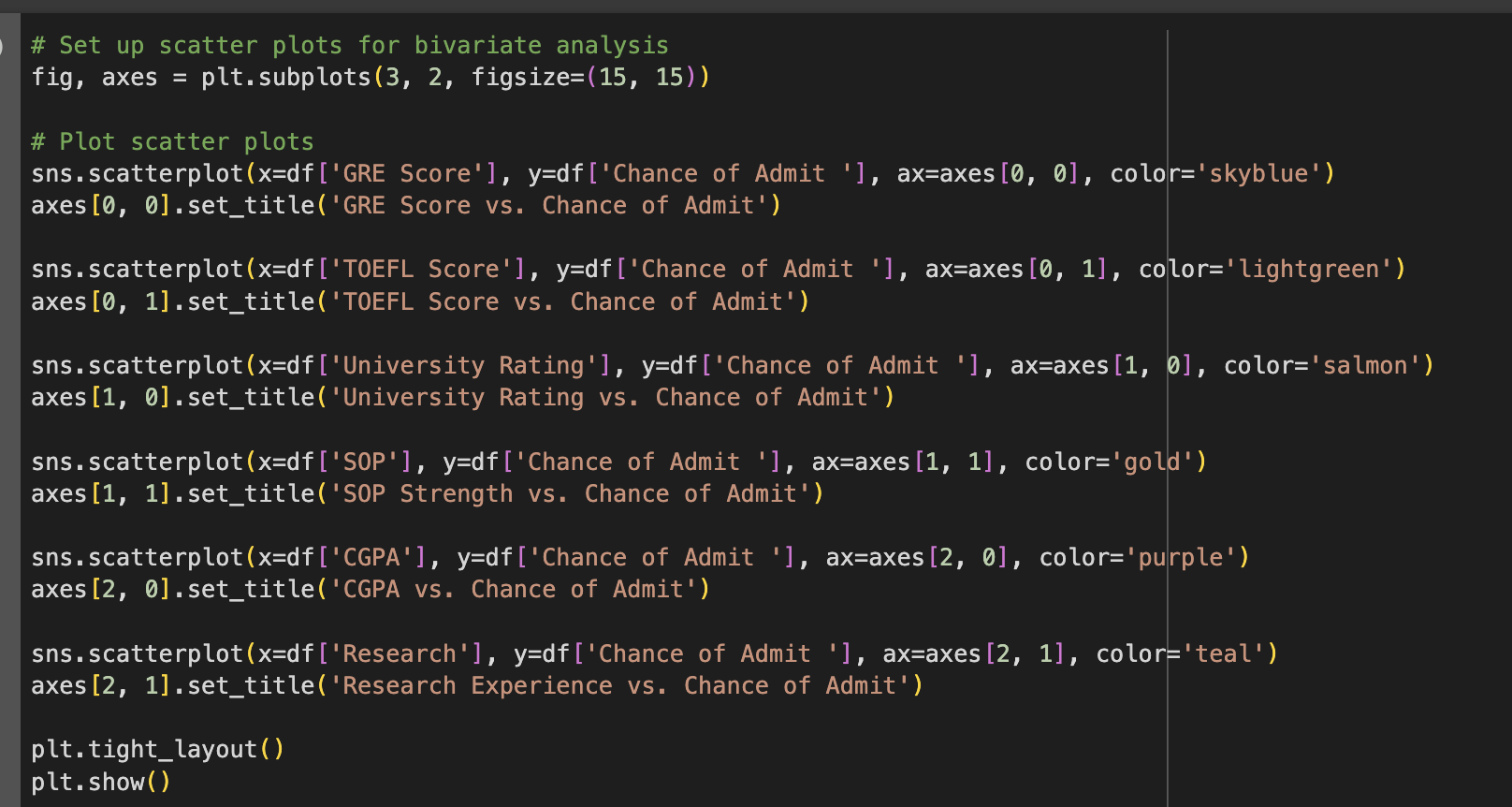
CGPA, GRE Score, and TOEFL Score show a high positive correlation with the Chance of Admit, which is expected as per the real life scenario.

Moderate Correlation can be seen for the features of University Rating and SOP with the Chance of Admit.

The lowest of the correlation is shown by the feature – research experience.

Q) Scatter plot for bivariate analysis

Ans)



A screenshot of a graph

AI-generated content may be incorrect.

A comparison of a graph

AI-generated content may be incorrect.

Insights : We have created the scatter plot for Features GRE Score, TOEFL Score, University Ranking and SOP strength against Chance of Admit

The max. chance of admit is more scattered in the area of the range 310 – 330. Once the value is higher or lower, then the chance of admit is either very high or very low. Though the chance is extreme values the number of candidates are low in this region.

Similar is the case for the features – TOEFL score with the range 100-115 that has max. candidature and above average change of admission . In case of university ranking, the chane of admit is more for the middle values of the rating – i.e. 3-4. When the rating is lower than this or higher than this cases becomes extreme. In the former the students those who are applying would be less, while in the latter case, the universtiy acceptance is low but Chances are more if selected. So is the case for SOP strength too. If the SOP is good enough then the chances of getting into the college increases.

In case of CGPA, when the value is more the chance of admit is also very high. While the research experience is showing a balanced output. If there is experience then we are to get it. But not having is never the deal breaker for admission as per the scatter plot.

In totality we can say that Higher GRE scores, TOEFL scores, and CGPA generally correspond to a higher chance of admission to the colleges.

Q) Outlier Treatment

Ans)

A screen shot of a computer program

AI-generated content may be incorrect.

Insights : As per the above code results, there are no outliers present in the given data set.

Suppose if there were any outliers present , then we could have implemented the below code within the for loop for clipping the oultier values from the data set .

Q1 = df[col].quantile(0.25)

Q3 = df[col].quantile(0.75)

IQR = Q3 - Q1

lower\_bound = Q1 - 1.5 \* IQR

upper\_bound = Q3 + 1.5 \* IQR

df[col] = df[col].clip(lower=lower\_bound, upper=upper\_bound)

Q) Feature Engineering

Ans) There is nothing to be created as to get the data manipulated.However we can delete the column Serial No. since its just an identifier column and does not have any impact on the creation of the inter connection or correlation for the features.

A black background with white text and green text

AI-generated content may be incorrect.

Q) Data Preparation for Modelling.

Ans) A screenshot of a computer program

AI-generated content may be incorrect.

Insights : Data preparation for modelling mainly deals with the scaling / standardisation of the data so that ML model / algorithm can understand it on a regular scale and not upon a skewed value where the baseline is different for different columns. Once the data is standardised, we need to split the data into train data and test data generally as 80 : 20 split. Looking at the shape of the X matrix, there are 7 columns available ( devoid of output column chance of admit and the identifier column Serial No. ) y represents the output variable or the feature.

Q) Linear Regression model

Ans)

A screenshot of a computer program

AI-generated content may be incorrect.

Insights : Linear regression model is built using the train data. The model coefficients indicate the impact of each feature on the chance of admission.We can clearly see that among the features, CGPA has the highest coefficient followed by GRE Score and Then TOEFL Score.This says that the impact of these features are very high on the outcome of Chance of admitting into the IVY league colleges.

Recommendation : A very good CGPA Score along with good GRE Score and TOEFL score can have positive imapct on Chance of Admit.Students could be adviced to achieve good marks during their college as well as for the competitive exam. These score will elevate the candidature to the IVY league.

Q) Model Statistics

Ans)

A screenshot of a computer program

AI-generated content may be incorrect.

Insights : The variables x1 through x7 are actually : GRE Score(x1), TOEFL Score(x2), University Rating(x3), SOP(x4), LOR(x5),CGPA(x6), Research(x7)

R-squared (0.821): This means that approximately 82.1% of the variance in "Chance of Admit"

Adjusted R-squared (0.818): Slightly lower than R-squared, as expected. This is a more conservative estimate of the explained variance, adjusting for the number of predictors in the model. Since it's very close to the R-squared, it suggests that adding the 7th variable didn't hurt the model but didn't help it much either.

F-statistic (257.0) and Prob (F-statistic) (3.41e-142): This tests the overall significance of the model. The extremely low p-value (Prob) indicates that the model is highly statistically significant.

**University Rating and SOP**  These variables have high p-values (0.541 and 0.721, respectively). This means that they are not statistically significant. The rest of the variables have a statistically significant relationship with "Chance of Admit."

Recommendations : This again re iterates the value that CGPA GRE Score and TOEFL Score are crucial for the chance of admit into the IVY league colleges. Hence the students need to take extra precaution as to get the best of the marks in these sections.

Q) Ridge and Lasso regression

Ans)

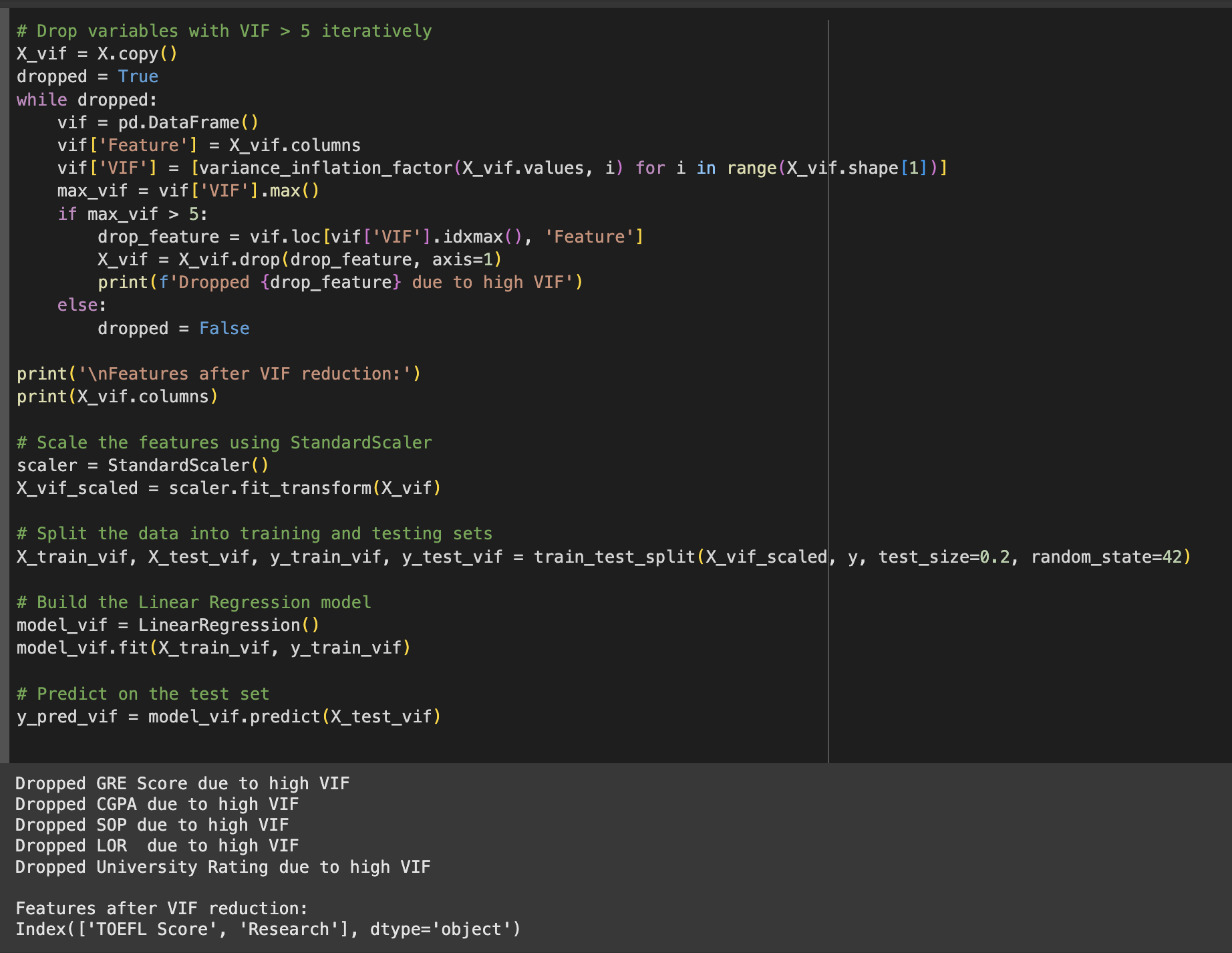
A computer screen shot of a program code

AI-generated content may be incorrect.

Q) Multicollinearity check by VIF score

Ans) A screenshot of a computer

AI-generated content may be incorrect.



Insights : Multicollinearity occurs when two or more independent variables in your model are highly correlated with each other. A common rule of thumb is that VIF scores above 5 or 10 indicate problematic multicollinearity. The VIF score for Research is relatively low, suggesting that Research experience is not strongly correlated with the other variables in the model. The High scores of GRE and TOEFL indicates that these are highly correlated (students who score well on one tend to score well on the other).

After removing and re calculating the VIF values the features that would remain are only TOEFL Score and Research Experience.

Q) Mean of Residuals

Ans )

A computer screen shot of a program

AI-generated content may be incorrect.

Insights : The mean of residuals is very close to zero. While it's slightly negative, the magnitude is so small that it likely doesn't indicate a practically significant bias in the model's predictions.

This mean that the prediction of the model is having high confidence since the bias is very minimal or practically none.

Recommendations : Any new data can be checked with the current model and the result of Chance of Admit can be as close as to the real world scenario. The percentage error for the same is virtually zero.

Q) Linearity of variables

Ans)

A screen shot of a graph

AI-generated content may be incorrect.

Insights :

The residuals appear to be mostly scattered around zero. This suggests that the assumption of linearity is likely met.The spread of the residuals might be a bit smaller in the middle range of fitted values and slightly larger towards the edges. However, it's not sufficiently large enough.

The scatter is mostly random, and there are no strong patterns suggesting that there is reasonable fit for the data

Q) Test for Homoscedasticity

Ans)

A screen shot of a computer code

AI-generated content may be incorrect.

Insights : The Goldfeld-Quandt test is a statistical test for **heteroscedasticity** (non-constant variance) in a regression model.

Consider the alpha value to be 0.05. The p-value > alpha which indicates that the we fail to reject the null hypothesis. Here The null hypothesis is that the variances are equal

Q) Normality of residuals

Ans)

A screenshot of a computer program

AI-generated content may be incorrect.

A graph with a red line

AI-generated content may be incorrect.

A graph of a graph

AI-generated content may be incorrect.

Insights: The Q-Q plot suggests that the residuals from the model are approximately normally distributed. The deviations at the tails are minor and likely not a cause for concern.

The histogram generally supports the assumption of normality for residuals. The slight skewness is probably not a major concern based on the large sample size.

Q) Model performance evaluation

Ans)

A computer screen shot of a program

AI-generated content may be incorrect.

Insights : Mean Absolute Error (MAE), represents the average absolute difference between predictions and the actual values. Root Mean Squared Error (RMSE) square root of the average squared difference between predictions and actual values. The MAE and RMSE are relatively low, suggesting that model's predictions are fairly accurate. A lower RMSE indicates better performance. The R-squared and Adjusted R-squared are both above 0.8, indicating that model explains a substantial portion of the variance in the dependent variable.

A computer screen shot of a program

AI-generated content may be incorrect.

A computer screen shot of a program

AI-generated content may be incorrect.

A computer screen shot of a program

AI-generated content may be incorrect.

Overall Recommendations:

* Significance of Predictor Variables: Based on the model coefficients and p-values, the most significant predictors of a student's chance of admission are:

1. CGPA: Consistently shows a strong positive impact.
2. GRE Score and TOEFL Score: Higher scores significantly increase the chance of admission.
3. University Rating and SOP Strength: These factors also play a crucial role in admission decisions.

* Additional Data Sources for Model Improvement : To improve the model, consider incorporating the following data sources:
* Quality of Undergraduate Institution: The reputation and ranking of the applicant's undergraduate institution can provide valuable insights.
* Detailed Research Experience: Information on publications, conference presentations, and the impact of the research will also help in deciding the candidature at IVY league.
* Specific Programs Applied To: Different programs may have varying admission criteria and weigh different factors accordingly. Generalizing might lead to wrong data being deducted as the output.
* Demographic Information: While ethical considerations are paramount, understanding demographic trends can help in identifying disparities and ensuring fairness. Also the colleges tries to ensure a balanced count from all stratas. Hence this can be a useful feature.
* Model Implementation in the Real World
* Interactive Prediction Tool: Develop a user-friendly tool on Jamboree's website where prospective students can input their scores and other details to get an estimated chance of admission. This can be used to let know what all updates or improvements could be made which could help them increase their chances for getting admitted to IVY league.
* Personalized Counseling: Use the model to identify areas where students can improve their profiles. If possible one – on – one counselling for students to improve their GRE / TOEFL scores or better the SOP etc will be helpful during the time of admissions
* Refine Training Programs: Tailor GMAT, GRE, and TOEFL training programs to focus on areas that have the most impact on admission chances.
* Potential Business Benefits
* Improved Authenticity and Credibility: Providing realistic admission chances can enhance Jamboree’s credibility and build trust with prospective students. The more authentic the organization is, the better would be footfall of the students.
* Increased Engagement: An interactive prediction tool can attract more visitors to the website and increase engagement with Jamboree's services.
* Improved Student Outcomes: By focusing on the most impactful factors, Jamboree can help students optimize their profiles and increase their chances of getting into top universities.

By implementing these recommendations, Jamboree can leverage data-driven insights to improve its services, enhance student outcomes, and maintain its position as a leading test preparation and counseling institute.