Jamboree Business Case

From company's perspective:

- Jamboree is a renowned educational institution that has successfully assisted numerous students in gaining admission to top colleges abroad. With their proven problem-solving methods, they have helped students achieve exceptional scores on exams like GMAT, GRE, and SAT with minimal effort.
- To further support students, Jamboree has recently introduced a new feature on their website. This feature enables students to assess their probability of admission to Ivy League colleges, considering the unique perspective of Indian applicants.
- By conducting a thorough analysis, we can assist Jamboree in understanding the crucial factors impacting graduate admissions and their interrelationships. Additionally, we can provide predictive insights to determine an individual's admission chances based on various variables.

From learner's perspective:

- Solving this business case holds immense importance for aspiring data scientists and ML engineers.
- Building predictive models using machine learning is widely popular among the data scientists/ML engineers. By working through this case study, individuals gain hands-on experience and practical skills in the field.
- Additionally, it will enhance one's ability to communicate with the stakeholders involved in data-related projects and help the organization take better, data-driven decisions.

Dataset: link

- **Serial No.:** This column represents the unique row identifier for each applicant in the dataset.
- **GRE Scores:** This column contains the GRE (Graduate Record Examination) scores of the applicants, which are measured on a scale of 0 to 340.
- **TOEFL Scores:** This column includes the TOEFL (Test of English as a Foreign Language)

scores of the applicants, which are measured on a scale of 0 to 120.

- **University Rating:** This column indicates the rating or reputation of the university that the applicants are associated with.
  - The rating is based on a scale of 0 to 5, with 5 representing the highest rating. ● **SOP:** This column represents the strength of the applicant's statement of purpose, rated on a scale of 0 to 5, with 5 indicating a strong and compelling SOP. ● **LOR:** This column represents the strength of the applicant's letter of recommendation, rated on a scale of 0 to 5, with 5 indicating a strong and compelling LOR. ● **CGPA:** This column contains the undergraduate Grade Point Average (GPA) of the applicants, which is measured on a scale of 0 to 10.
- **Research:** This column indicates whether the applicant has research experience (1) or not (0).
- **Chance of Admit:** This column represents the estimated probability or chance of admission for each applicant, ranging from 0 to 1.

These columns provide relevant information about the applicants' academic qualifications, test scores, university ratings, and other factors that may influence their chances of admission.

How to get started?

To complete the case study, begin by downloading the CSV files from the provided link. Afterward, proceed to upload them onto Google Colab / Jupyter Notebook for further analysis.

If you are using Google Colab, you can directly start working on the notebook on Colab.

If you wish to use Jupyter Notebook, Install Anaconda using the link. Once Anaconda has been installed on your system, open Jupyter Notebook. Refer link.

Now, the CSV data file needs to be uploaded/imported in the Colab/Jupyter notebook respectively.
Once the files have been successfully uploaded/imported, you can conveniently access them within the notebook using the Pandas read_csv( ) method.

What is expected?

Assuming you’re a data scientist/ML engineer hired by Jamboree, your primary objective is to analyze the given dataset and derive valuable insights from it. Additionally, utilize the dataset to construct a predictive model capable of estimating an applicant's likelihood of admission based on the available features.

---

1. Define the problem statement.

   Hint:
   We want you to discuss the given problem statement with additional views.

2. Import the dataset and do usual exploratory data analysis steps like checking the structure & characteristics of the dataset.

   Hint:
   We want you to...
   - Drop any irrelevant column present in the dataset.
   - Check the shape of the dataset provided.
   - Check the data type of each column.
   - Comment on the range of attributes.
   - Display the statistical summary of the entire dataset.

3. Use Non-graphical and graphical analysis for getting insights about variables.

   Hint:
   We want you to...
   - Perform a Univariate Analysis - Check the distribution of different continuous/categorical variables.
     i. You could use a histplot, displot or kdeplot.
   - Perform a Bivariate Analysis - Check the relationship between different variables.
     i. You could use a scatter plot, regplot or pairplot.

4. Perform data preprocessing.

   Hint:
   We want you to...
   - Check for duplicate records and treat them accordingly if found.
   - Check for missing values and treat them accordingly if found.
   - Check for outlier values and treat them accordingly if found.

5. Check the correlation among independent variables and how they interact with each other.

   Hint:
   We want you to create a correlation matrix/heatmap and drop any feature that has a high correlation (>0.90) with some other feature.

6. Prepare the data for modeling.

   Hint:
   - Encode categorical variables (if any) using a suitable method
   - Perform the train-test split
   - Perform data normalization/standardization
     NOTE: Feature scaling should be performed after the train-test

split. 7. Build the Linear Regression model.

   Hint:
   - Use Linear Regression from Statsmodel library to create a model and comment on the model statistics.
   - Also, display model coefficients with column names.
   - Drop columns with p-value > 0.05 (if any) and re-train the model.

8. Test the assumptions of linear regression.

   a. Multicollinearity check by VIF score
      Hint:
      - Variables are dropped one-by-one till none has a **VIF>5**.

   b. Mean of residuals should be close to **zero**.

   c. Linear relationship between independent & dependent variables.
      Hint:
      This can be checked using the following methods:
      - Scatter plots
      - Regression plots
      - Pearson Correlation

   d. Test for Homoscedasticity
      Hint:

- Create a scatterplot of residuals against predicted values.
- Perform a **Goldfeld-Quandt** test to check the presence of Heteroscedasticity in the data.
- If the obtained **p-value>0.05**, there is no strong evidence of heteroscedasticity and hence, the presence of Homoscedasticity is validated.

e. Normality of residuals

Hint:

- Almost **bell-shaped** curve in residuals distribution.
- Points in the **Q-Q plot** are almost all on the line.

9. Evaluate the model's performance.

Hint:

- ○ MAE (Mean Absolute Error)
- ○ RMSE (Root Mean Square Error)
- ○ R2 score / R-squared value
- ○ Adjusted R-squared value

10. Provide Actionable Insights & Recommendations.

FAQs

Q. Which platform am I supposed to use?

You may use either Google Colab or Jupyter notebook.

Q. I am having issues setting up a Jupyter notebook.

Install Anaconda using the link. Once Anaconda has been installed on your system, open Jupyter Notebook. Refer link.