Q1. Jamboree Education - Linear Regression

Mindset

1. Evaluation will be kept lenient, so make sure you attempt this case study.
2. Read the question carefully and try to understand what exactly is being asked.
3. Brainstorm a little. If you’re getting an error, remember that Google is your best friend.
4. You can watch the lecture recordings or go through your lecture notes once again if you feel like you’re getting confused over some specific topics.
5. Discuss your problems with your peers. Make use of the Slack channel and WhatsApp group.
6. Only if you think that there’s a major issue, you can reach out to your Instructor via Slack or Email.
7. There is no right or wrong answer. We have to get used to dealing with uncertainty in business. This is exactly the skill we want to develop.

Context

Jamboree has helped thousands of students like you make it to top colleges abroad. Be it GMAT, GRE or SAT, their unique problem-solving methods ensure maximum scores with minimum effort.

They recently launched a feature where students/learners can come to their website and check their probability of getting into the IVY league college. This feature estimates the chances of graduate admission from an Indian perspective.

How can you help here?

Your analysis will help Jamboree in understanding what factors are important in graduate admissions and how these factors are interrelated among themselves. It will also help predict one's chances of admission given the rest of the variables.

Dataset:

Dataset Link: jamboree_admission.csv

Column Profiling:

- Serial No. (Unique row ID)
- GRE Scores (out of 340)
- TOEFL Scores (out of 120)
- University Rating (out of 5)
- Statement of Purpose and Letter of Recommendation Strength (out of 5)
- Undergraduate GPA (out of 10)

- Research Experience (either 0 or 1)
- Chance of Admit (ranging from 0 to 1)

Concept Used:

- Exploratory Data Analysis
- Linear Regression

What does good looks like?

- Import the dataset and do usual exploratory data analysis steps like checking the structure & characteristics of the dataset.
- Drop the unique row Identifier if you see any. This step is important as you don't want your model to build some understanding based on row numbers.
- Use Non-graphical and graphical analysis for getting inferences about variables.
    - This can be done by checking the distribution of variables of graduate applicants.
- Once you've ensured that students with varied merit apply for the university, you can start understanding the relationship between different factors responsible for graduate admissions.
- Check correlation among independent variables and how they interact with each other.
- Use Linear Regression from (Statsmodel library) and explain the results.
- Test the assumptions of linear regression:
    - Multicollinearity check by VIF score
    - Mean of residuals
    - Linearity of variables (no pattern in residual plot)
    - Test for Homoscedasticity
    - Normality of residuals
- Do model evaluation- MAE, RMSE, R2 score, Adjusted R2.
- Provide actionable Insights & Recommendations
- Try out different Linear Regressions

Evaluation Criteria (100 Points):

1. Define Problem Statement and perform Exploratory Data Analysis **(10 points)**
    - Definition of problem (as per given problem statement with additional views)
    - Observations on shape of data, data types of all the attributes, conversion of categorical attributes to 'category' (If required) , missing value detection, statistical summary.
    - Univariate Analysis (distribution plots of all the continuous variable(s) barplots/countplots of all the categorical variables)
    - Bivariate Analysis (Relationships between important variables such as workday and count, season and count, weather and count.
    - Illustrate the insights based on EDA

- Comments on range of attributes, outliers of various attributes
- Comments on the distribution of the variables and relationship between them
- Comments for each univariate and bivariate plots

2. Data Preprocessing **(10 Points)**
   - Duplicate value check
   - Missing value treatment
   - Outlier treatment
   - Feature engineering
   - Data preparation for modeling
3. Model building **(10 Points)**
   - Build the Linear Regression model and comment on the model statistics
   - Display model coefficients with column names
   - Try out Ridge and Lasso regression
4. Testing the assumptions of the linear regression model **(50 Points)**
   1. Multicollinearity check by VIF score (variables are dropped one-by-one till none has VIF>5) (10 Points)
   2. The mean of residuals is nearly zero (10 Points)
   3. Linearity of variables (no pattern in the residual plot) (10 Points)
   4. Test for Homoscedasticity (10 Points)
   5. Normality of residuals (almost bell-shaped curve in residuals distribution, points in QQ plot are almost all on the line) (10 Points)
5. Model performance evaluation **(10 Points)**
   - Metrics checked - MAE, RMSE, R2, Adj R2
   - Train and test performances are checked
   - Comments on the performance measures and if there is any need to improve the model or not
6. Actionable Insights & Recommendations **(10 Points)**
   - Comments on significance of predictor variables
   - Comments on additional data sources for model improvement, model implementation in real world, potential business benefits from improving the model (These are key to differentiating a good and an excellent solution)

Submission Process:

- Type your insights and recommendations in the text editor.
- Convert your jupyter notebook into PDF (Save as PDF using Chrome browser's Print command), upload it on our platform
- Optionally, you may add images/graphs in the text editor by taking screenshots or saving matplotlib graphs using plt.savefig(...).
- After submitting, you will not be allowed to edit your submission.