FERRERO

Business Case

Welcome to the business case study. Below are a set of questions that you would need to attempt and present the results in powerpoint on the day of the interview. There are two datasets that are available to you for this business case. Please be ready to share the underlying codes and outputs and explain the reasoning of your answers.

Exercise 1:

1. Read 'loan_sanction_train.csv' and print the first 15 rows
2. Check the following characteristics of the dataset:
   - Number of rows and columns
   - Data type
   - Print the names of columns
3. Print the names of the columns that have some null values
4. For each of those columns please provide a few examples on how you would treat them and implement one
5. Create a distribution plots for each variable by Loan_Status value
6. Split loan_sanction_train into training and validation set
   Set seed : 42 and validation size 10%
   What does the "seed" part do here?
7. Build a Random Forest model on the training set to predict: 'Loan_Status'.
   a) Evaluate the performance both on training and validation set. Display the confusion matrix.
   b) Explain the reasons for the performance criteria you chose.
   c) What are the conclusions that you can draw from this analysis to help the loan approval process?
   d) Any enhancements you would suggest in this exercise?
8. Retrain the model and this time, use cross validation using k-fold (k:5).
   What are the advantages of using cross validation?
9. Average the coefficients of the folds and predict using the test dataset
10. Save your prediction dataframe and save your trained model

Exercise 2:

1. Use the data from loan_sanction_train.csv to conduct a linear regression analysis, aiming to discern the influence of each variable on ApplicantIncome.
2. Present a comprehensive diagnostic assessment to evaluate the model quality.
3. How does Bayesian linear regression differ from traditional linear regression?

4. Apply Bayesian linear regression on the loan_sanction_test.csv dataset, examining the influence of each variable on ApplicantIncome by incorporating the coefficients obtained from the linear regression as priors.
5. Present a comprehensive diagnostic assessment to evaluate the Bayesian's model quality.