AB InBev - Technical Test
Data Scientist

1. Credit Scoring Model Challenge (Supervised Learning)

Description

Banks and entities that give out credit play a crucial role in market economies. They decide who can be financed, on what terms and this can make or break investment decisions. For markets and society to function, individuals and companies need access to credit.

Credit scoring algorithms, which make a guess at the probability of default, are the method banks and entities use to determine whether or not a loan should be granted. This challenge requires participants to improve on the state of the art in credit scoring, by predicting the probability that somebody will be credit worthy by first predicting the probability that a client will experience financial distress in the next quarter. Historical data are provided on 250,000 borrowers.

The **goal** of this challenge is to build a model that lenders can use to help make the best financial decisions from a credit risk perspective.

Given Inputs

- ***data_dictionary.xlsx***: Brief description of the variables that are available for the challenge.
- ***cs-training.csv and cs-test.csv***: Database with all the available business variables.
    - Use the target variable ***SeriousDlqin2yrs***, already included in the dataset.
    - Proceed with any feature engineering to build any variables you consider important for the challenge.
- ***sample_results.csv:*** Database with a sample of the credit scoring model results for clients in the training database.

Questionnaire:

1. Briefly describe how you would build the target variable if it were not available in the dataset.
2. What metrics do you consider relevant to consider in the decision of choosing the best model? Discuss why you might give priority to one metric over another.
3. In the metric optimization process, how may the cutoff point or threshold of the prediction of the probability of default affect the result?
4. What additional variables do you think should be considered given the context of our clients at AB InBev? (Bars, Restaurants, Convenience Stores, etc.)

Expected outputs

- Python scripts used to develop the model.
- Executive summary including any visuals or plots used to explain the model building process and results.
- Insights gathered on any of the data exploring process and model building.

2. Client Segmentation Challenge (Unsupervised Learning)

Description

Understanding a company’s client basis is key to proceed with proper commercial strategy, logistics and technology project development.

Clustering algorithms, which associate clients based on their characteristics, are one of the methods banks and commercial entities use to understand common characteristics, find look-alikes and better establish what are the key differentiating factors amongst them.

The **goal** of this challenge is to build a segmentation of clients that decision makers can use to help understand customer basis and further develop any strategies that are aligned with the company’s interests.

Given Inputs

- Same as the ones used in the last challenge.
- Proceed with any feature engineering to build any variables you consider important for the challenge.

Questionnaire:

1. Briefly explain the algorithms tested.
2. Write a brief discussion on the process of selecting the proper number of groups.
3. Provide a brief interpretation of the clusters formed.

Expected outputs

- Python scripts used to develop the model.
- Executive summary including any visuals or plots used to explain the model building process and results.
- Insights gathered on any of the data exploring process and model building.

3. Concepts and understanding

Analyze and respond the following questions to the best of your understanding:

1. Briefly discuss any ML Ops standards you’ve been exposed to or are familiar with based on your understanding and experience.

2. In the development of a Machine Learning project with a business objective such as predicting sales, calculating propensity scores or recommending a product or service, what should the script structure of the project contain?
3. Why is it of importance to think about the implementation and integration of an ML Project to a company's system framework?
4. Do you have experience using cloud architectures? If so, please discuss which ones you've used. Are you familiar with the AzureML framework?
5. Are you familiar with the Spark framework and have you developed any models using languages such as PySpark or Scala?

Concept	Description	Example of Applicable Business Scenario
Recommender Systems		
XGBoost		
Cross-Validation		
Silhouette Test		
ROC Curve		
AUC Metric		