Kuunda: Data Science Challenge

Kuunda is focused on enabling responsible digital financial services for underbanked populations. This technical assessment simulates a core component of Kuunda's business: the development and deployment of a **default prediction model or a rule-based scorecard** to assess credit risk for digital loan products.

You will work with a dataset that includes trading behavior features and loan performance outcomes aggregated over different time windows. The dataset is stored in AWS S3 and reflects a real-world, high-class-imbalance use case. Your goal is to develop a reproducible, explainable model pipeline from raw data ingestion to model saving for production use.

Output: Please provide your solution as a set of notebooks (one for each task) and model assets. You are welcome to use `git` for source control.

Tools: Python, AWS S3 (use `boto3` or `s3fs`), `pandas`, `scikit-learn`, `matplotlib/seaborn`, `joblib` or `pickle`, and optionally `xgboost`, `lightgbm`, or `catboost` and/or any other python modules you deem necessary.

Note: Take your time to think critically and creatively. The goal of this assessment is to gauge your creativity and problem-solving skills.

Task 1: Default risk

The goal of this exercise is to develop a machine learning pipeline or rule-based algorithm/scorecard to manage the default risk of customers using historical financial behavior. The notebooks should contain the end-to-end setup, including data loading, preprocessing, exploratory data analysis (EDA), model training and validation.

a) Access the dataset stored in the AWS S3 bucket `kuunda-datascience-challenge` using Python and the credentials provided in `datascience_trainees_accessKeys.csv`. Load the data into a Pandas DataFrame. Clearly document your method for accessing S3 securely i.e. How are the credentials stored.

b) Conduct an exploratory analysis to understand the structure and quality of the data. Provide descriptive statistics for all features, use appropriate visualizations. Explore correlations between features and the target variable. Investigate any class imbalance in the target, and summarize the initial insights and hypotheses you generate from this analysis.

c) Train a binary classifier to predict default or perform data analysis to create a rule-based score card. Please justify your choice. Please clearly explain the methodologies and techniques used. Describe any feature engineering that was done. What metric was used for training and why? Think of the implication for the business.

d) Describe and implement a (back) testing strategy to estimate model performance in production.

Hint: Think critically about the trade-off between *performance* and *explainability*. Kuunda may prioritize transparent models in some regions for regulatory or operational reasons.

Task 2: Prepare the Algorithm for Production

Deploying a scorecard at Kuunda requires not only high model performance but also robustness, traceability, and reusability. From a business perspective, being able to store and reuse scorecards ensures consistency in credit scoring decisions across environments, enables reproducibility for audits or regulatory reviews, and facilitates efficient updates when retraining is necessary. This task focuses on preparing the scorecard for production by saving both the algorithm and metadata. This provides a mechanism for preforming inference on unseen data.

a) After training the final model, save it to disk using a suitable serialization method such as `joblib` or `pickle`. Ensure that the model can later be reloaded exactly as it was at training time.

b) In addition to the model object, save relevant metadata required for production inference. This should include the list of feature names used during training, the full transformation pipeline (including scaling and encoding), model performance metrics computed on the test set, and a timestamp marking when the model was trained.

c) Implement a reusable function that loads the model and its associated metadata from disk. This function should return the python object representing the model pipeline.

d) Provide an example script that demonstrates inference on unseen dataset. This should include loading the data, applying the preprocessing steps and producing predictions using the saved model assets. Kuunda will use this script on a hidden dataset to evaluate your submission.

Task 3: Business and Ethical Questions

The following questions must be answered at the end of your notebook submission. Your responses will be assessed for clarity, business relevance, and your ability to balance technical insight with practical impact.

1. What is the most critical business consideration when evaluating model performance at Kuunda? Explain with examples.

2. If you had to explain this model to a bank or regulator, how would you communicate the risks and benefits?

3. How would you monitor this model in production? What data would you track and how would you trigger retraining?

4. What are the ethical implications of false positives and false negatives in this credit scoring model?