Excellence Academy for Professionals

DATA SCIENCE COMPETITION 2024

16 June to 20 July 2024

Data Science Competition: Predicting Probability of Default

Problem Statement

Financial institutions face significant risks due to loan defaults. Accurately predicting the probability of default (PD) on loans is critical for risk management and strategic planning. In this competition, participants are tasked with developing a predictive model that estimates the probability of default on loans using historical loan data.

Dataset Description

The provided dataset contains historical information about borrowers, including various features that may impact the probability of default.

Competition Tasks

The competition runs from June 16, 2024 - July 20, 2024**.** Your submission should cover the following aspects:

1. **Data Cleaning**:
    - Clean the dataset and explain your decisions regarding techniques used.
2. **Basic EDA (Exploratory Data Analysis)**:
    - Explore the dataset to gain insights into feature distributions, correlations, and potential patterns.
    - Visualize key relationships and summarize your findings.
    - Discuss any interesting observations.
3. **Feature Selection**:
    - Select relevant features for model training.
    - Justify your feature selection methods (e.g., statistical tests, domain knowledge).
4. **Hyperparameter Tuning**:
    - Optimize hyperparameters for your chosen machine learning model(s).
    - Explain the rationale behind your choices.

5. **Cross Validation**:
    - Implement cross-validation to assess model performance.
    - Describe the cross-validation strategy.
    - Report evaluation metrics.
6. **Feature Scaling and Transformation**:
    - Apply appropriate scaling and transformation techniques.
    - Discuss why you chose specific transformations.
7. **Model Building**:
    - Train at least 5 models.
    - Explain your choice of algorithm.
    - Discuss any model assumptions and limitations.
8. **Model Evaluation**:
    - Evaluate your model(s) on a separate validation set.
    - Interpret performance metrics.
9. **Endpoint Development for Inference**:
    - Create API endpoints using Fast API.
    - Implement endpoints for model training and inference.
    - Provide clear documentation on how to use these endpoints.
10. **Data Drift Detection**:
    - Implement a mechanism for detecting data drift in the deployed model.
    - Explain why monitoring data drift is crucial for model maintenance.
11. **Model Analysis**:
    - Interpret model coefficients or feature importances.
    - Investigate instances where the model performs poorly.
    - Analyze the model for biases in predictions.

- Explain how the model is making predictions.
- Clearly communicate the limitations of the model, acknowledging situations where it might not perform well.
- Propose potential enhancements or future directions for model improvement.
- Discuss potential business implications of your findings.

Submission Requirements

Your submission should be sent no later than the **20th of July 2024, 2359hrs** and should include:

1. Jupyter notebook(s) containing your code for each step and notes explaining why each decision was made.
2. Python Script(s) with your FastAPI endpoints.
3. A detailed README explaining your approach, assumptions, and reasoning.
4. A txt or csv file showing your **GIT** commit history.

The submissions should be sent to mailto:competitions@claxonactuaries.com and CC mailto:tinaye.m@claxonactuaries.com with the subject **"Submission: Claxon Data Science Competition 2024"**

The submissions should be zipped in a folder with the name of the participant on each file and the folder name.

Evaluation Criteria

Participants will be evaluated based on the following criteria:

- **Technical Excellence**: Quality of code, model performance (**ROC-AUC**), and adherence to best practices.
- **Clear Explanations**: Well-documented decisions and explanations throughout the pipeline.
- **Deployment**: Successful implementation of API endpoints.
- **Creativity**: Innovative approaches and thoughtful problem-solving.

Those who would have completed all the competition tasks successfully will be called for an in-person presentation of their submissions at Claxon Actuaries which will also aid in selecting the winners.

Rules and Regulations for the Competition:

1. **Eligibility Criteria:**
    - Participants aged 30 years old or younger at the time of submission are eligible.
    - The competition is open to individuals only, and collaboration is not allowed**.**
    - Current employees of Claxon Actuaries are ineligible.
2. **Competition Duration:**
    - The competition will run from 15 June to 20 July 2024
    - The latest submission deadline is **July 20, 2024, 2359hrs**.
3. **Coding Language:**
    - All code submissions must be written in Python.
4. **Prizes and Opportunities:**

    🥇 **First Place:** USD $500 + Opportunity to work with Claxon Actuaries

    🥈 **Second Place:** USD $300

    🥉 **Third Place:** USD $150

    All winners will be publicized on our social media platforms and receive free career mentorship with Claxon Actuaries leadership.
5. **Questions:**
    - Any questions regarding the challenge should be directed to mailto:competitions@claxonactuaries.com

Good luck and may the best data scientists prevail! 🚀