Byron Washington

CDS 302

March 18, 2025

Project Proposal

My database would be a repository for machine learning models. This would allow users to store, track, and compare different models. The database would include models, their versions, performance logs, their outputs, and the datasets they were trained on. The intended users would be professional or amateur data scientists/analysts.

A function that is expected by users is registering different models. Users would register a new model and include its name, type (regression, classification, etc.), its hyperparameter configuration, and a description. Users would also be able to store different versions of models, allowing them to keep track of each model's version and which dataset it was trained on.

Another expected function would be registering datasets that the models were trained on. Users could also test models on other datasets to improve their models across different sources. If an issue came up with a dataset, users could query all active models trained on that dataset to find out which models need to be retrained.

Another function would be allowing users to log the performance of each model with its hyperparameter configuration. It would also store scoring variables that are needed to determine how accurate and precise a model is. For example, a regression model could store R^2 and Mean Squared Error (MSE) while a classification model could score F1, precision, or Log Loss[1].

Additionally, users could store the predictions of the models, storing the input and corresponding output. This would let users compare output not only by inputs, but also by which dataset the model was trained on.

References:

1. "3.4. Metrics and Scoring: Quantifying the Quality of Predictions." *Scikit*, scikit-learn.org/stable/modules/model_evaluation.html. Accessed 18 Mar. 2025.