DaQuiris

Abstract:

Data Quality is a very important, but also very difficult to automate. At some point, any DQ validation process will require manual effort. The goal of DaQuiri is to minimize the manual work as much as possible.

There is a currently operational implementation on e-commerce data warehouse, but it can’t scale to BigRed. The task will be to mimic the DaQuiri framework that is scalable to run production data after processing and, eventually, on the fly

The process will require various methods of regression analysis, a model evaluator, and applying a champion model to a new set of records to test whether the new record is within a reasonable prediction interval.

The result will be tremendous. More confidence in production data without a large manual effort to test multiple false positive flags in more simplistic processes.

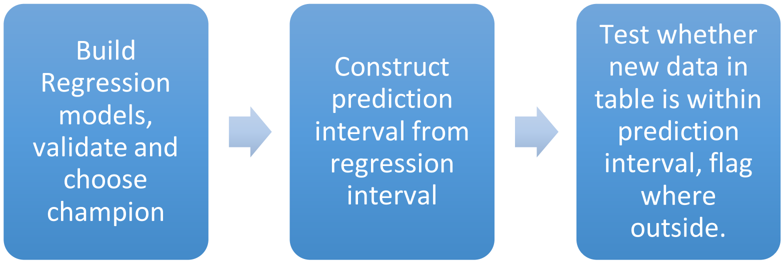
Data Source and Description:

Target BigRed

Description:

Every N minutes, build, train, and implement a regression model, and use the model to generate a prediction interval. Test new data based on prediction intervals, flag any that are outside.

Pipeline of Steps (in Block Diagram):



Build Regression models, validate and choose champion

Construct prediction interval from regression interval

Test whether new data in table is within prediction interval, flag where outside.

Metrics for Success:

Percent of data quality errors found
Timeliness of models (SLA)
Percent of flags that are false