ML & Statistics in Industry - Best of both worlds

Background

Many projects involve observational data. One example is *logged data* from complicated machines. For this internship, we focus on studying relations y=f(several predictors x). A typical goal in CQM's consultancy projects is *explaining* the relation in order to better understand the underlying domain problem. This is similar to studying regression coefficients in a small linear regression application from classical statistics: effects of predictors given the others are constant, and the uncertainty of these coefficients. This is in contrast to making a *predictive model,* which is often the focus in supervised learning in Machine Learning. The latter is considered much easier. Explaining goes towards cause-and-effect.

In CQM projects, getting a sense of the input space (the predictors) can be time consuming: visualization, tabulation, summary, missing values, etc. To go further to make a regression-type model includes more detailed checks (collinearity, etc.). In contrast, in ML there are techniques for *prediction* that are more automatic (e.g. some accept missing values as input, robust for outliers), but do not readily offer the convenient and useful regression-type inference.

The question:

Can we set up a fairly automatic and robust workflow for building an explorative model for input, which may contribute to the exploratory phases of a project. The output of such a workflow has e.g.

- List of strange aspects of the input & output data
- A rough model that is not primarily used for prediction but more for the effect of each predictor.
- Names of the most important predictors
- Effect plots of the most important predictors / insight in the nature of the relation.
- Uncertainty (i.e. the standard errors from linear regression, bootstrap?)

Aspects we think of here: dependency in the data (groups as known from linear mixed models, time dependency); proper treatment of such groups in cross-validation; partial dependence plots; checking whether new input is not a multivariate outlier; behavior with few observations (when does it break down); bootstrap; regularization.

Possible approach:

1) Literature search
2) Simulation-based assessment: simulate 100x a true situation, analyse using candidate methods, collect performance. Around this: vary the true situation by some dimensions (i.e. number of observations, predictors, noise size, number of missing values, outliers,…)
3) Apply to example datasets.