

tidymodels Discussion

Max Kuhn (RStudio)

On the Horizon

There is a project list in the
be tackling.

org that has a list of activities and potential projects that we will

Pipelines

As previously mentioned, the modeling includes pre-modeling activities (e.g. feature engineering) as well as post-processing actions such as

- choosing an appropriate probability threshold
- calibrating probabilities
- applying equivocal zones and model applicability domain analyses

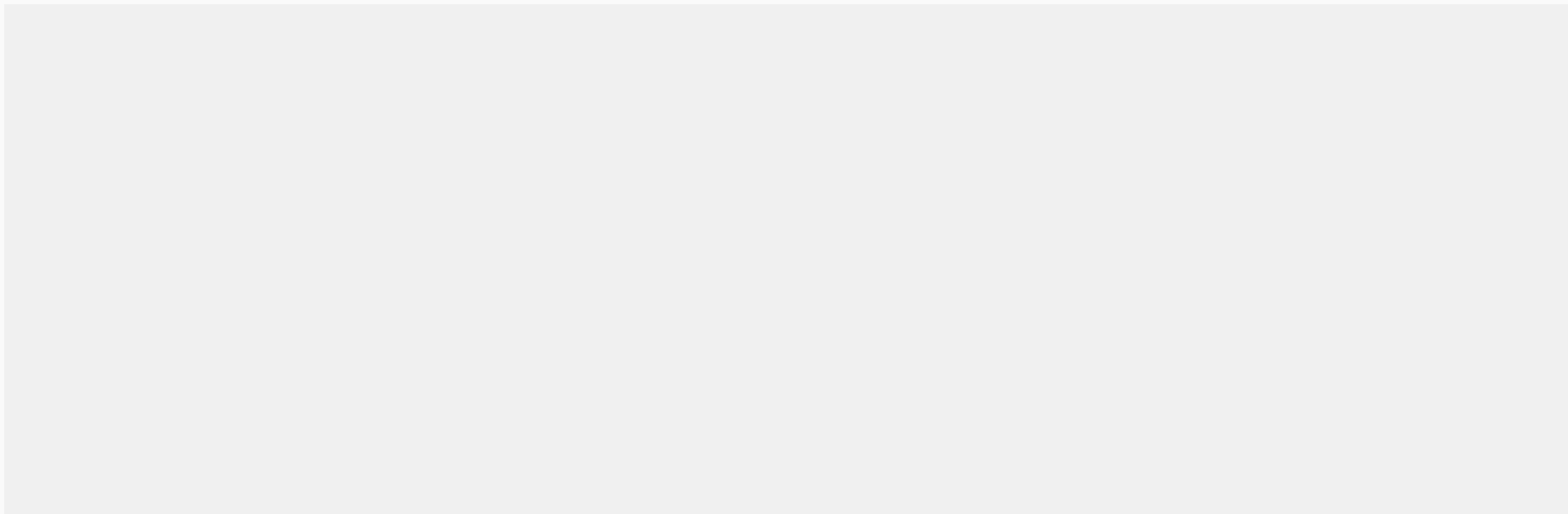
Modeling pipelines exist in `python` and `spark`.

Our implementation will be tidy and allow users to quickly try different combinations of techniques.

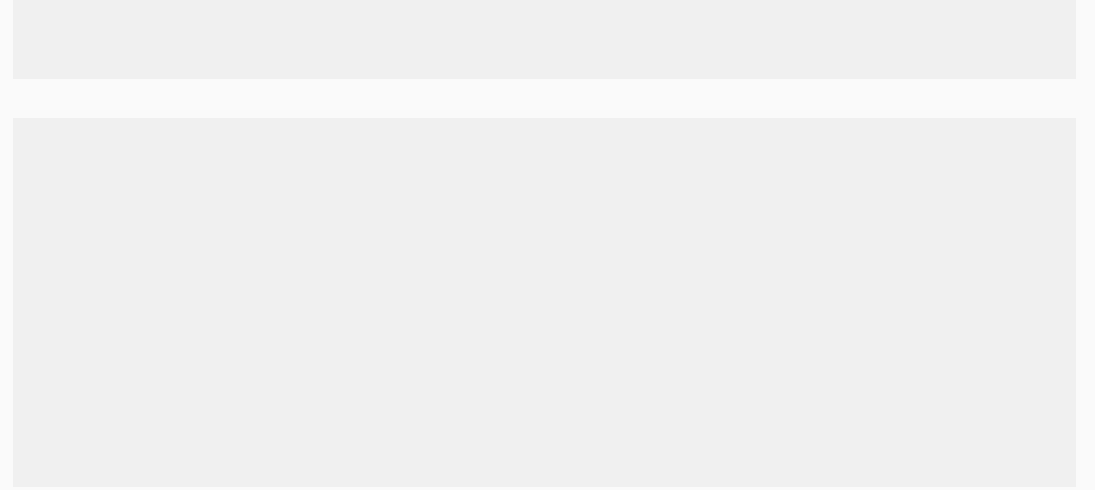
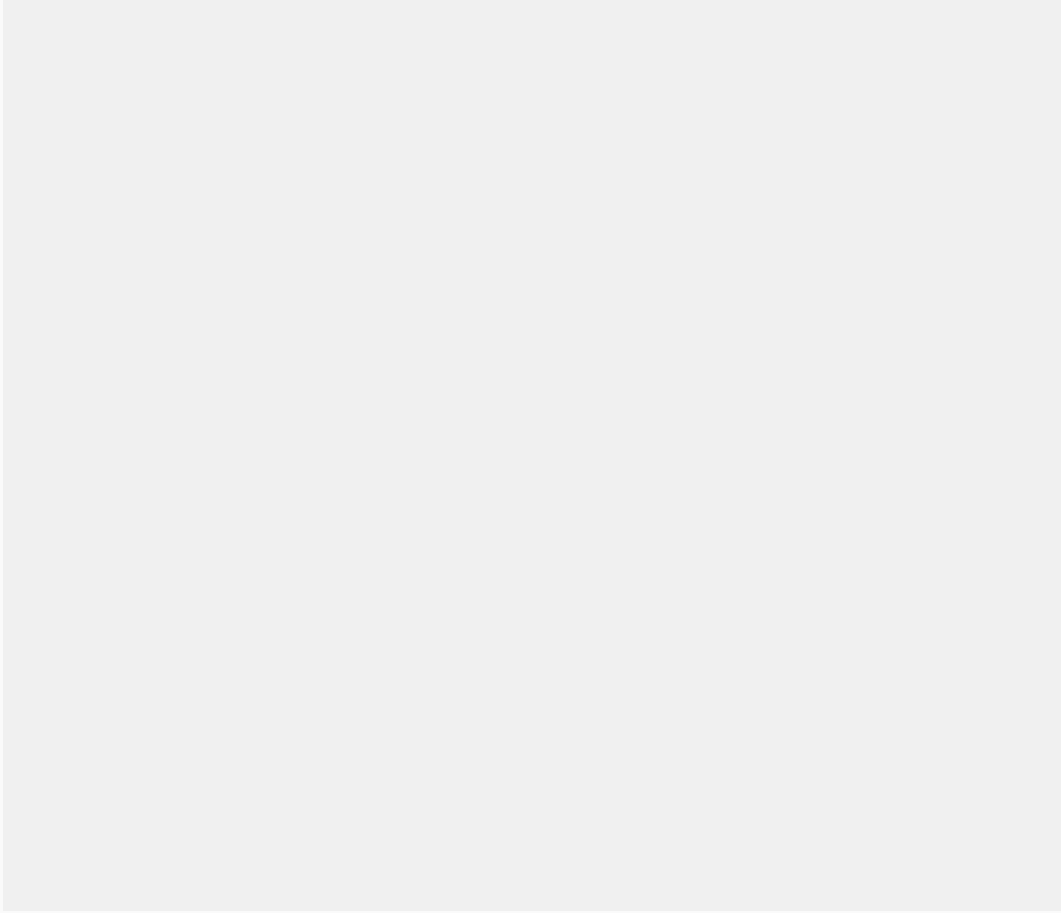
Pipelines Syntax

Suppose we need to impute some data, fit a logistic regression, then choose an appropriate probability threshold.

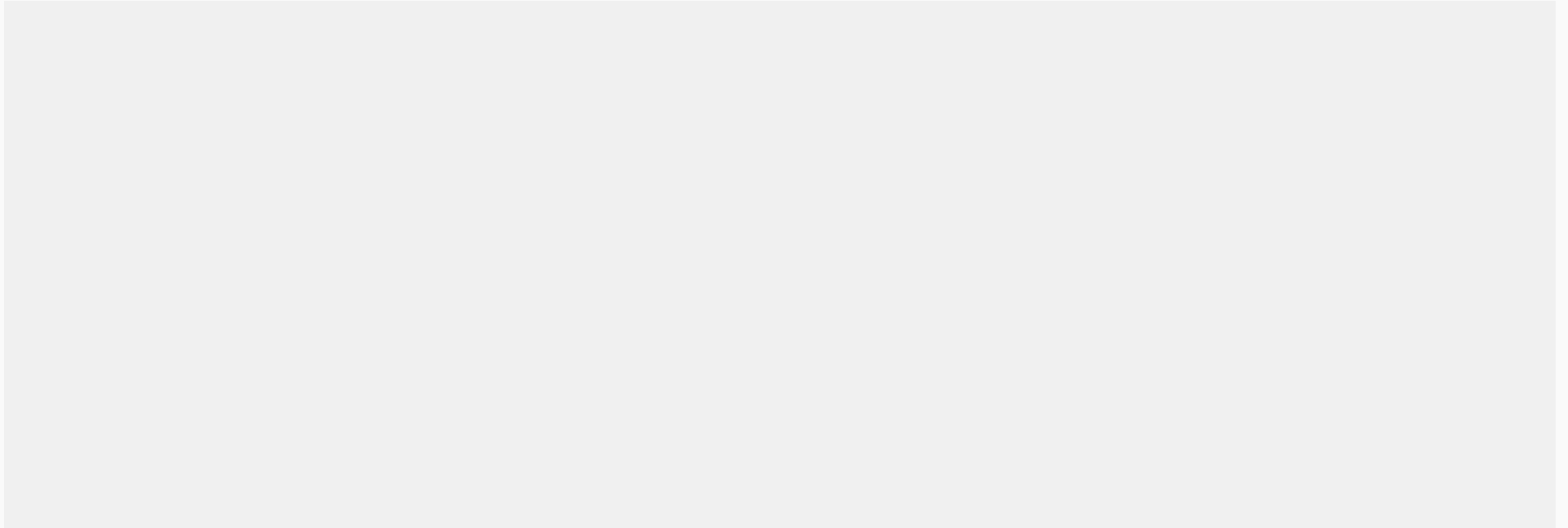
Although it isn't finalized, the syntax will look something like:



Automatically Identify Tunable Parameters



Model Tuning Syntax Prototype



Principles of Modeling Packages and Templates

We are in the process of developing a set of _____ for making good modeling packages. For example:

- Separate the interface that the _____ uses from the code to do the computations. They serve two very different purposes.
- Have multiple interfaces (e.g. formula, x/y, etc).
- The _____ should use the most appropriate data structures for the data (as opposed to the computations). For example, factor outcomes versus 0/1 indicators and data frames versus matrices.
- _____ for class probabilities 😊.
- Use S3 methods.
- The _____ method should give standardized, predictable results.

Rather than try to make methodologists into software developers, we will provide _____ with template packages that can be used to meet these guidelines (along with documentation and examples on _____).