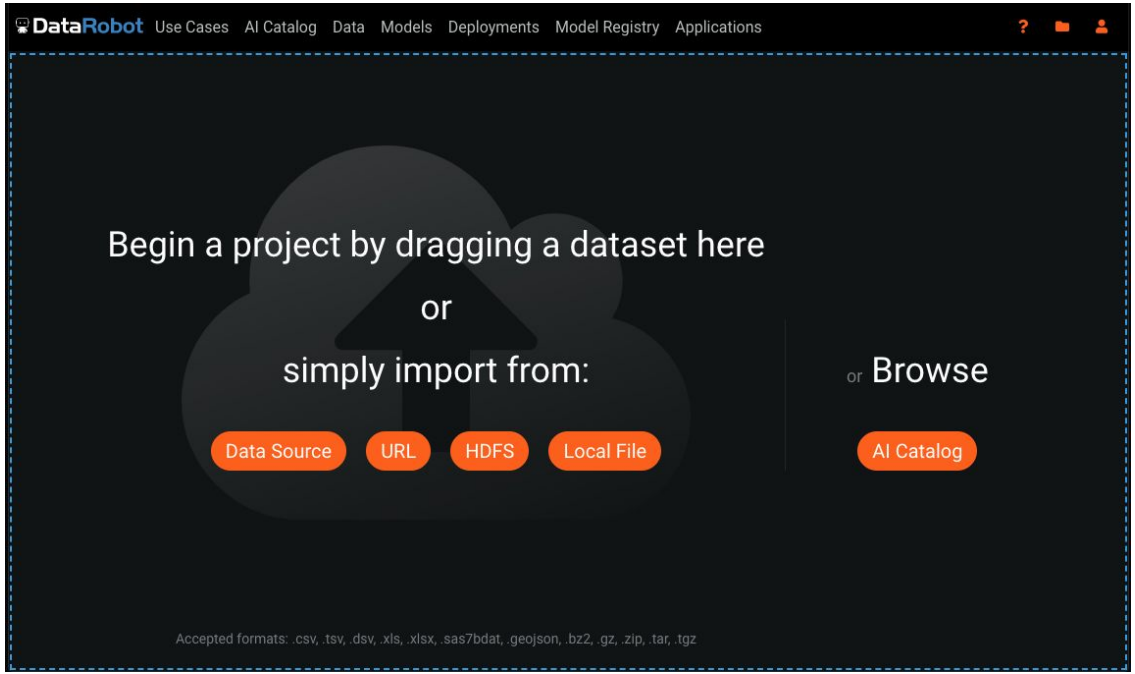
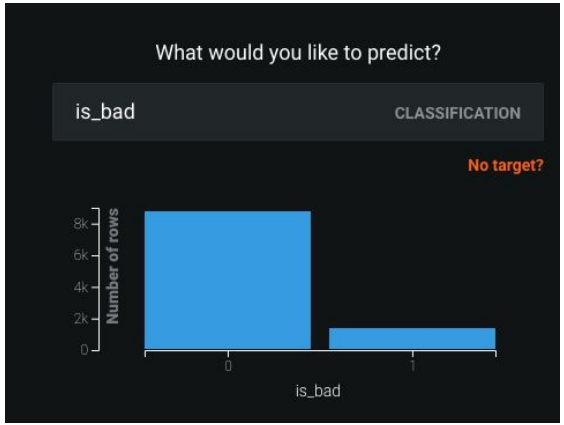
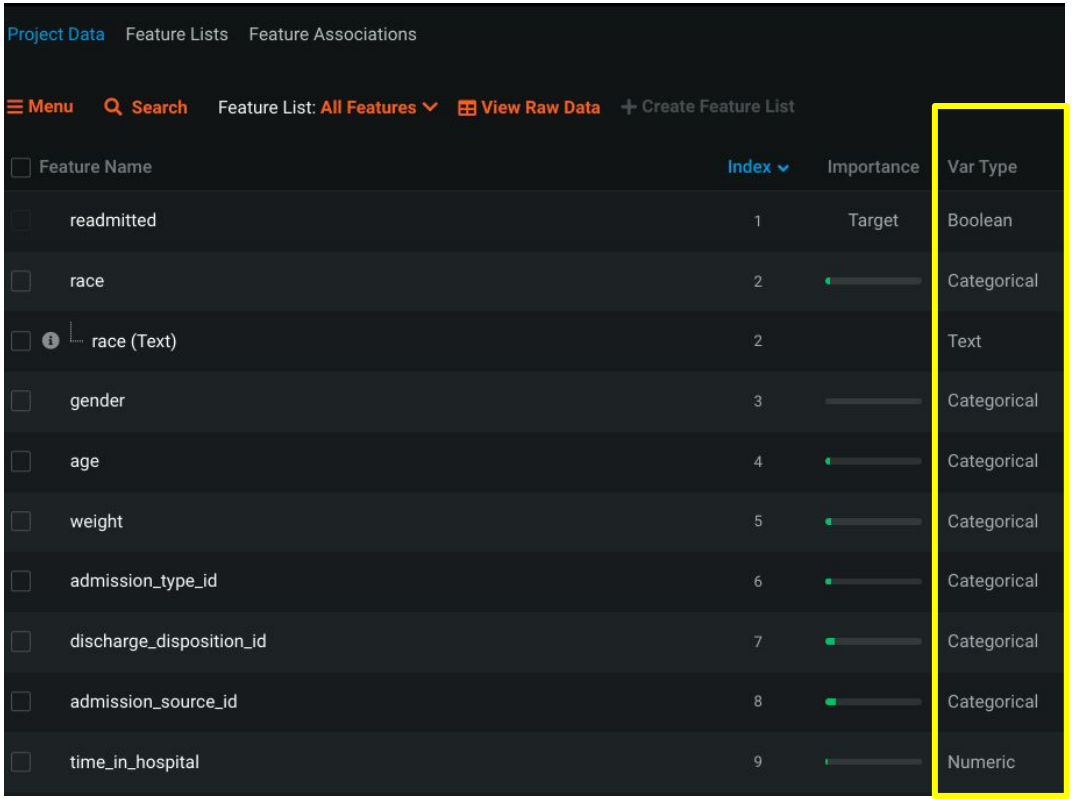
Load Your Data & Select Your Target Column





- No duplicate column names
- Most delimited formats accepted

Before you hit start...

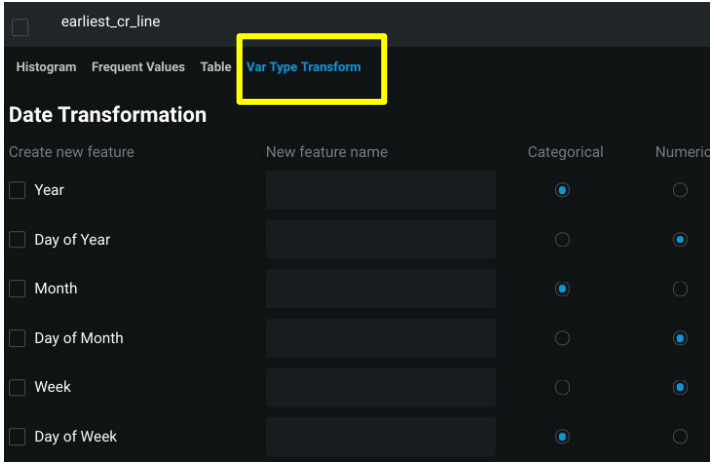


- Check out your feature types. Do they make sense?
- Did DataRobot denote any features as non-informative?
- Are dates read in as Type = Date and is DataRobot deriving date-based fields?
- Click on any field to expand and see its distribution...

How Should I Handle Dates?

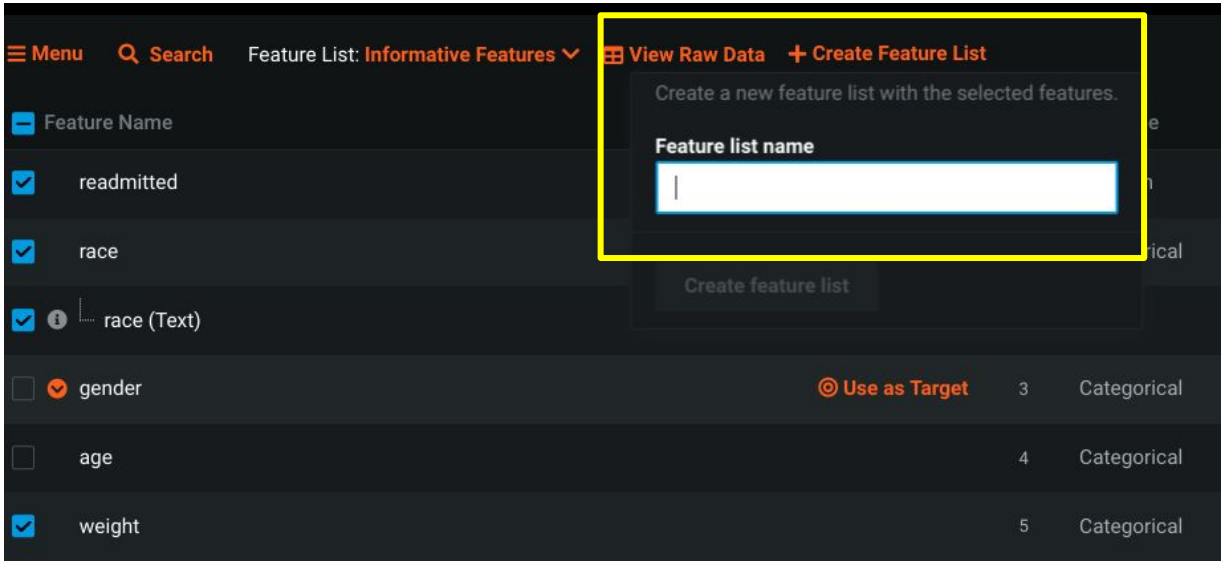
A: Raw Dates are generally not useful for ML - they should be transformed...

If your date range is long enough, DataRobot will automatically extract features from your date fields...



If a feature wasn't extracted automatically, you can do it manually using "Var Type Transform".

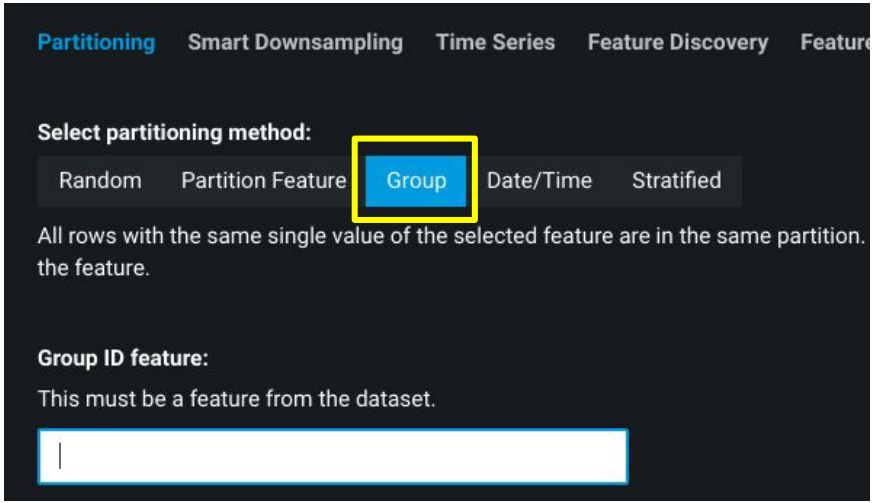
If you need to remove a feature...



- Select all & remove (uncheck) features
- Click “create feature list”
- Name your new list (e.g. “no gender”)

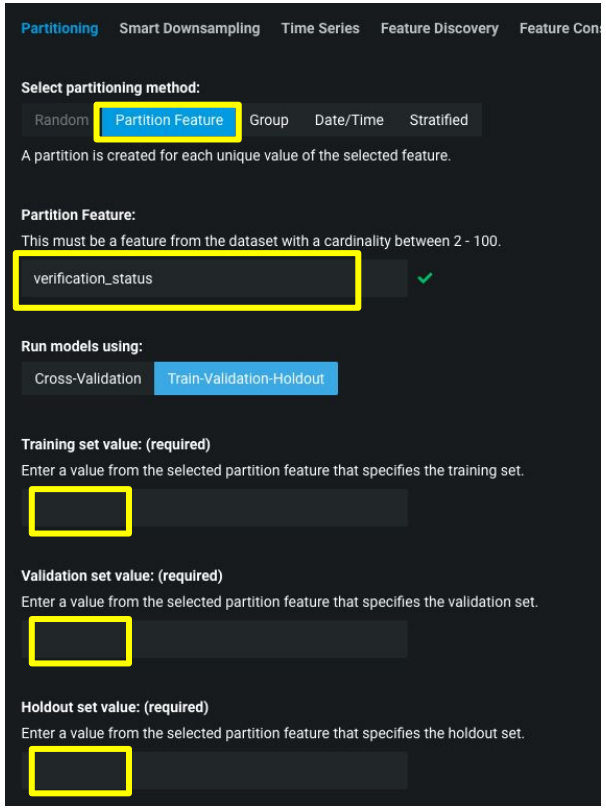
What if Rows (i.e. entities, people) Aren't Unique?

A: If you don't adjust the sampling method, you might end up with a patient/user records in **both** training and validation sets. This is a very *subtle* form of target leakage!



To avoid this, you can use the **Group** partitioning method to supply a column with the ID (i.e. PatientID). DataRobot will make sure not to split patients across training vs. validation sets

What if my Data is Already Split into Training/Validation?

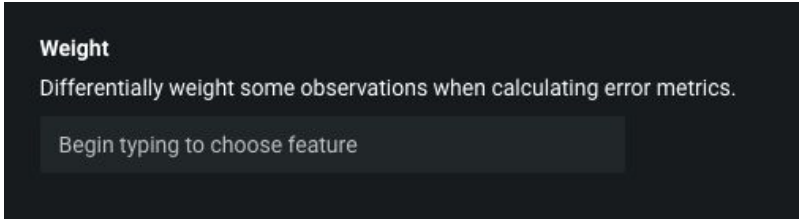


A: You can provide DataRobot this information in the form of a column in your raw data...

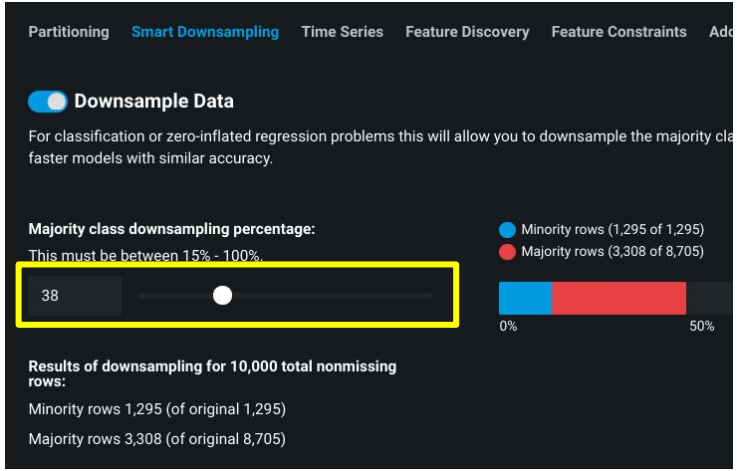
- Under **Partitioning** options, choose "**Partition Feature**"
- Supply column name which denotes the partitions
- Provide the values that represent the different samples (these are just examples…):
  - T or 0 for training
  - V or 1 for validation
  - H or 2 for holdout

Should I Be Sampling?

A: It may be counterintuitive, but usually you can build a model from a **sample** of your data that is just as good as using 100%. The larger your data is (or more imbalanced), the more likely you should be sampling. Two options:

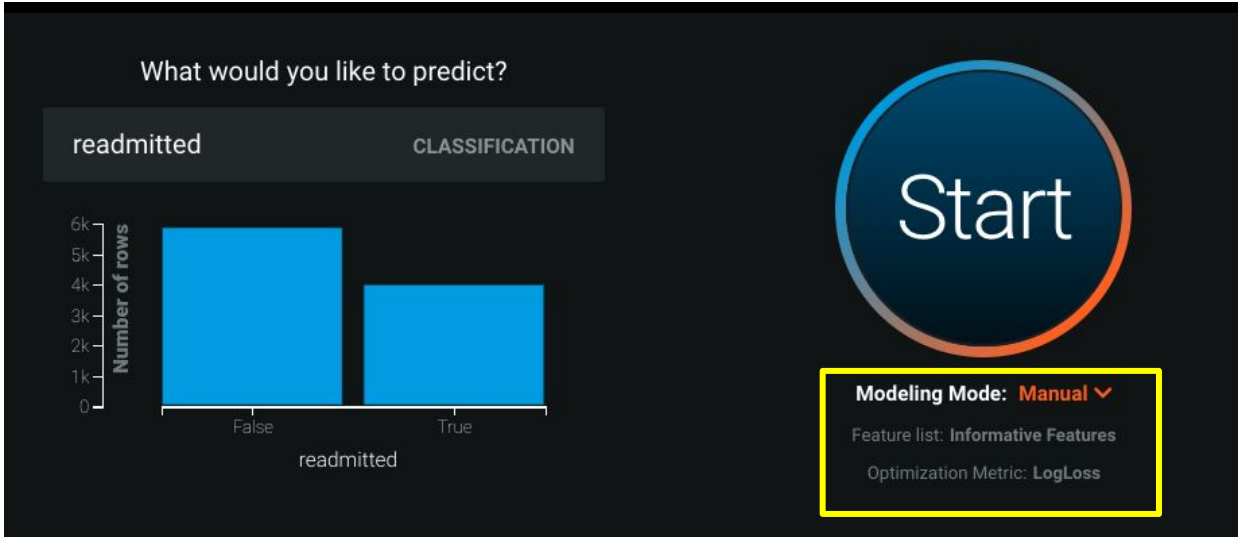


1. Sample the data BEFORE you create the project in DataRobot. Specify the column in your data that reflects sampling weight of records - available under “Additional” Advanced Options.



2. Adjust the sampling AFTER you start the project in DataRobot. Accuracy metrics are automatically adjusted.

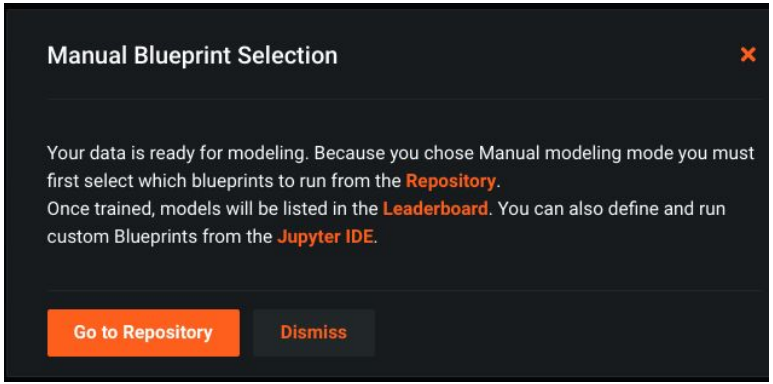
Start in Manual Mode

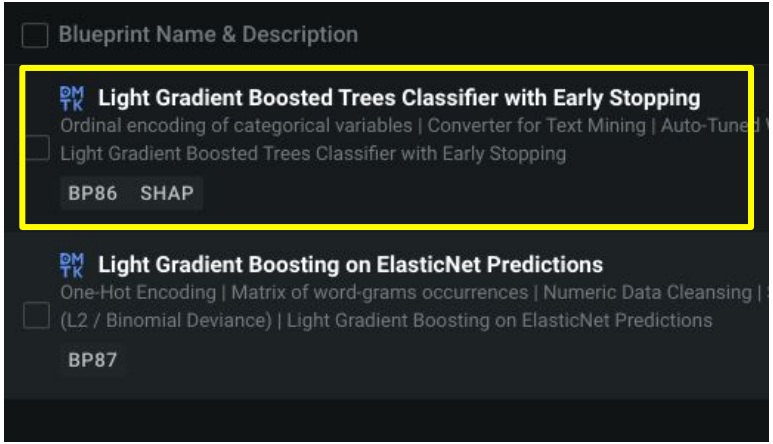


- For first run, try Manual
- Select your feature list below at the top of your features
- Default is “informative features”

Run one model - sanity check

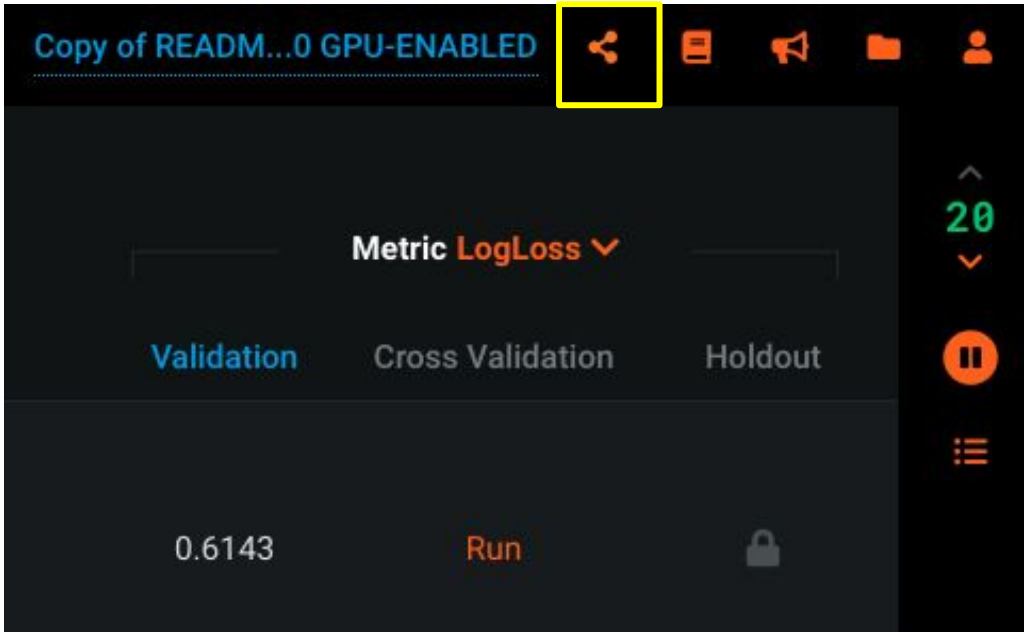
Go to Repository





- Use search field to find a blueprint - e.g. “Light Gradient Boosting”
- Select “Add” from blueprint menu on right
- Set sample size to 64% - Run Task

Don’t forget to share with your team-mates



This will save time, take the load off workers and you can collaborate on the same project...

Model Review - Step 1. Are My Results *"Too Good To Be True"*?

CLASSIFICATION

- Select AUC (area under the curve) as Metric
- Measures ability to rank-order your classes
- If AUC is in high 90's - probably too good.
- Go to step 2. May need to rethink features

Metric AUC

Validation	Cross Validation	Holdout
0.7095 *	0.6972 *	0.6922 *
0.7138	0.6987	0.6928
0.7125	0.6985	0.6922
0.7124	0.6983	0.6915

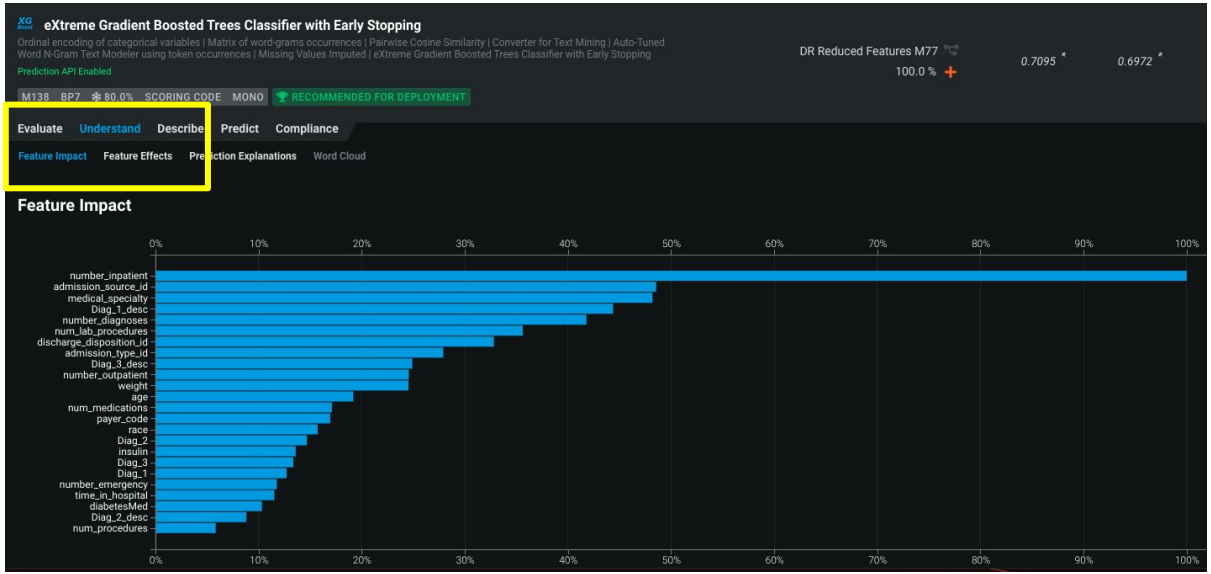
REGRESSION

- Select MAPE (mean abs percent error) as Metric
- A MAPE of 0.0 means that model estimated every target value perfectly.
- A very low MAPE (<1%) might be too good.
- Check for leakage in step 2.

Metric MAPE

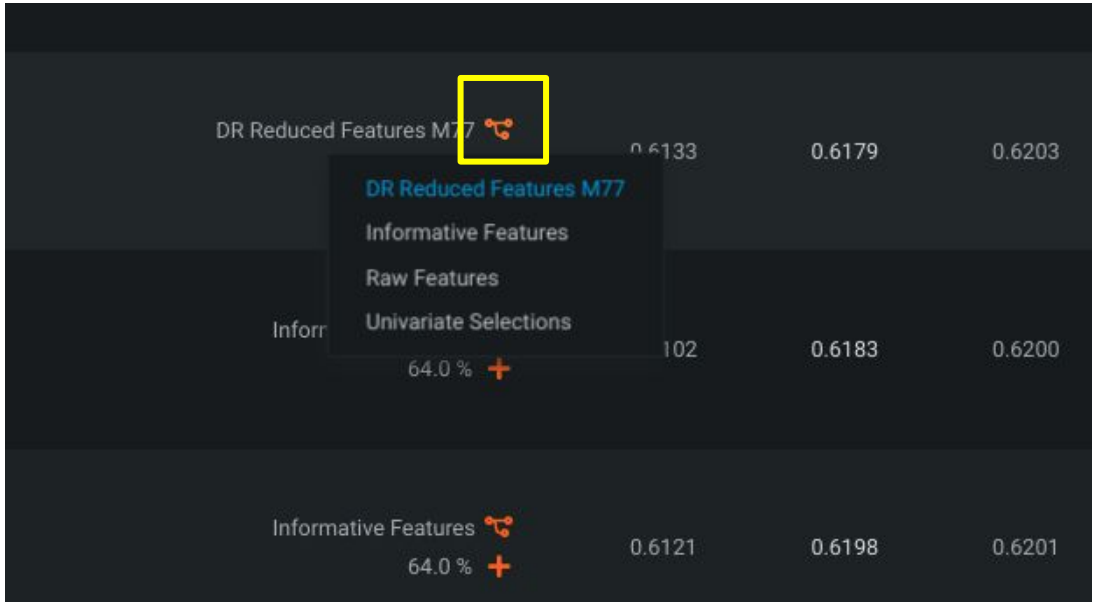
Validation	Cross Validation	Holdout
23.4761 *	24.6858 *	29.8133 *
22.5255 *	25.3280 *	28.1640
22.5954	26.3792	28.4790

Model Review - Step 2. “Which Features are Driving Prediction?”



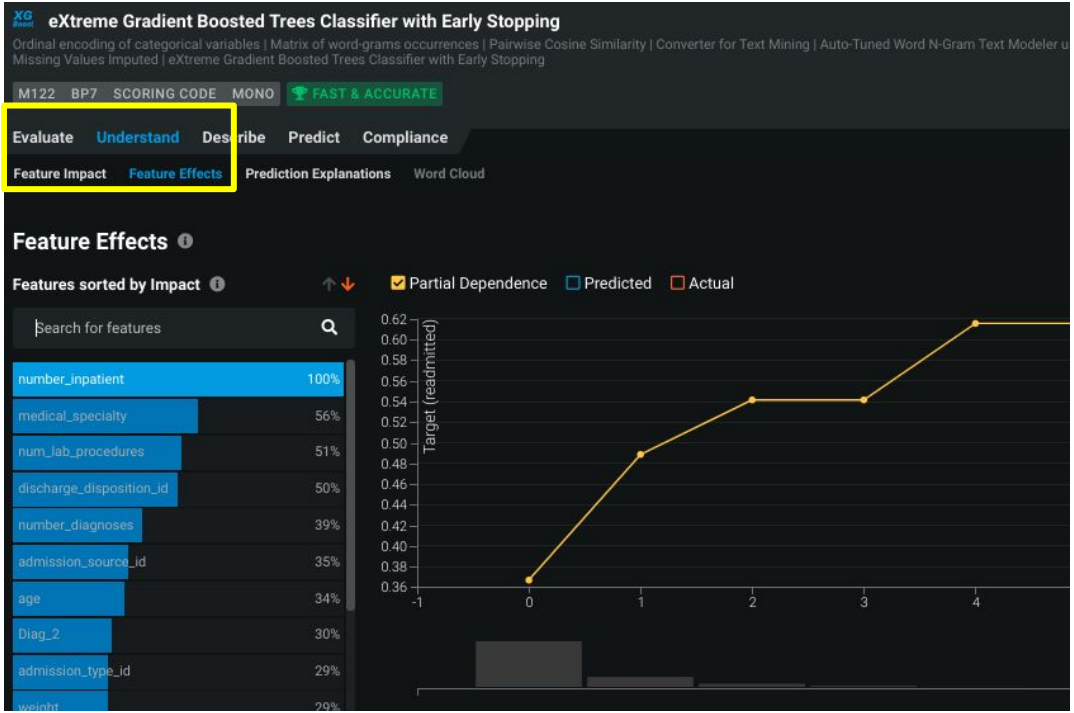
- **Feature Impact** measure perf. decrease when that feature is scrambled
- Top feature always set to 100%.
- Is the model depending heavily on **one** feature (i.e. leakage)?
- If found, remove leakage feature by creating new feature list

If you need to re-run a model on a new feature list...



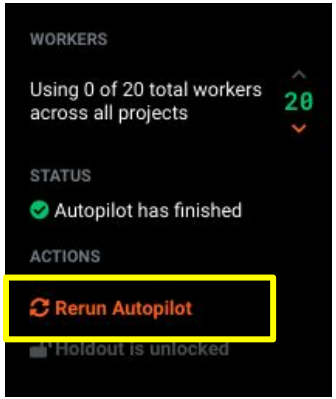
Click on the icon next to any model's feature list to re-run on a different feature list

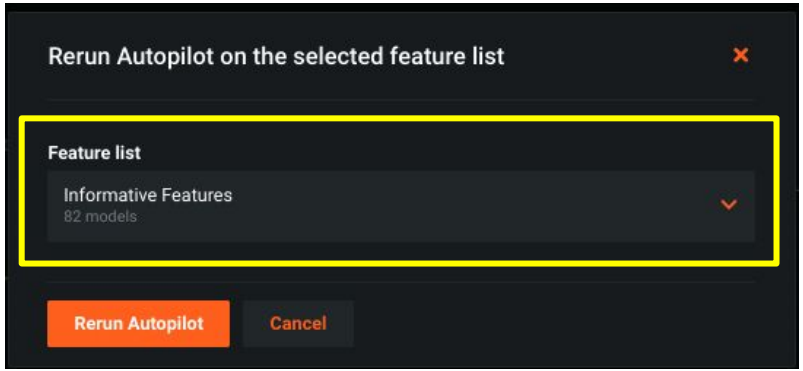
Model Review - Step 3. “How do Features Affect Predictions?”



- **Feature Effects** show average *prediction* over validation/holdout set as the feature value varies (i.e partial dependence).
- Does this shape match your common sense / pre-conceptions about the relationship?
- Red flags may indicate a badly-formed feature or interaction.

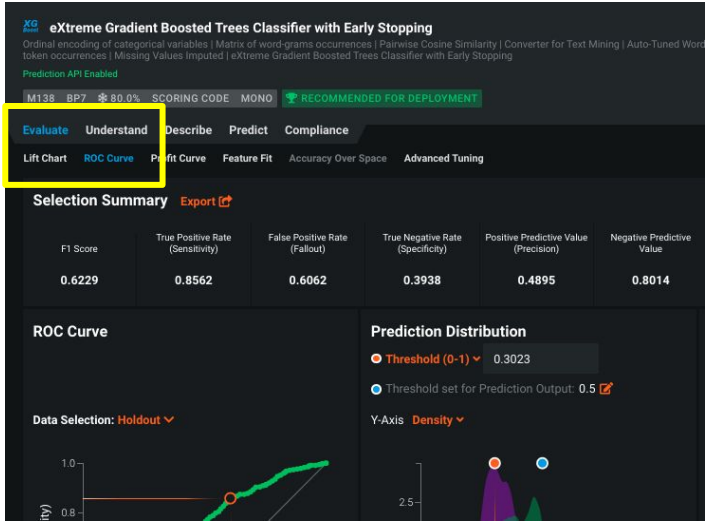
If you don’t see leakage and your accuracy is reasonable, run **Autopilot** (make sure you pick the right feature list) & max out your workers





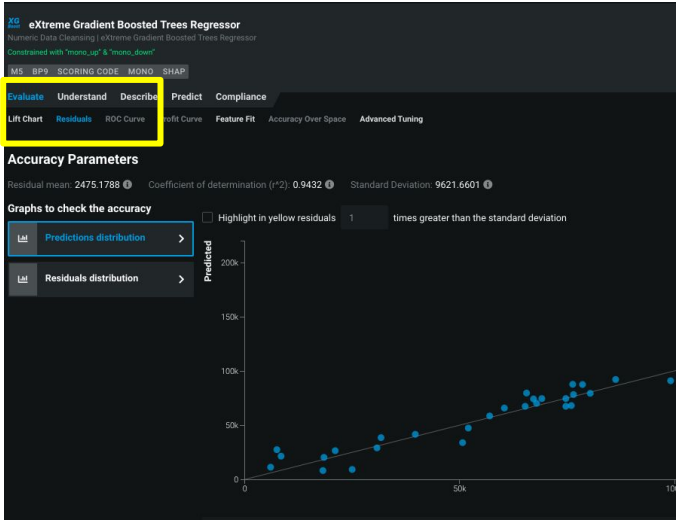
Model Review - Step 4a. “Let’s really dig into accuracy...”

CLASSIFICATION



Check out Lift Chart, ROC, Confusion Matrix

REGRESSION



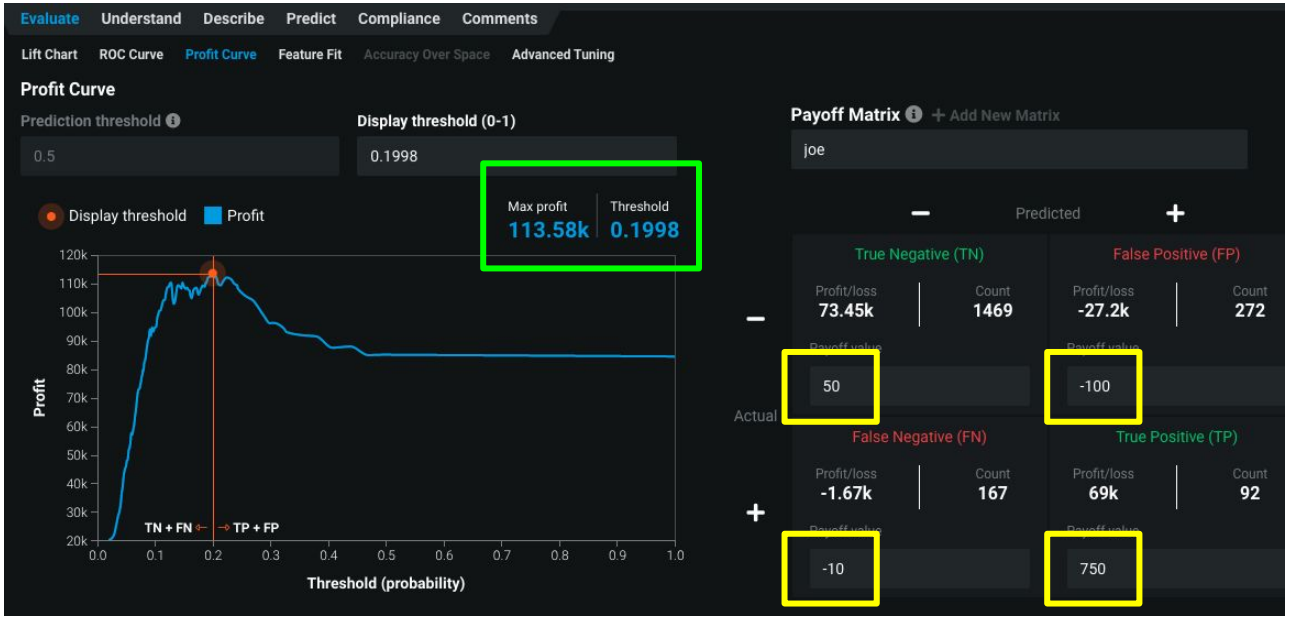
Check out Lift Chart, Residuals

Model Review - Step 4b. “How to adjust the threshold?”

1) Changing the decision boundary by dragging the orange line will affect the numbers in the confusion matrix.

2) The confusion matrix allows you to consider the value of a correct decision vs. the cost of a poor decision to get the most accurate assessment of value.

Model Review - Step 4c. “What is the optimal threshold?”

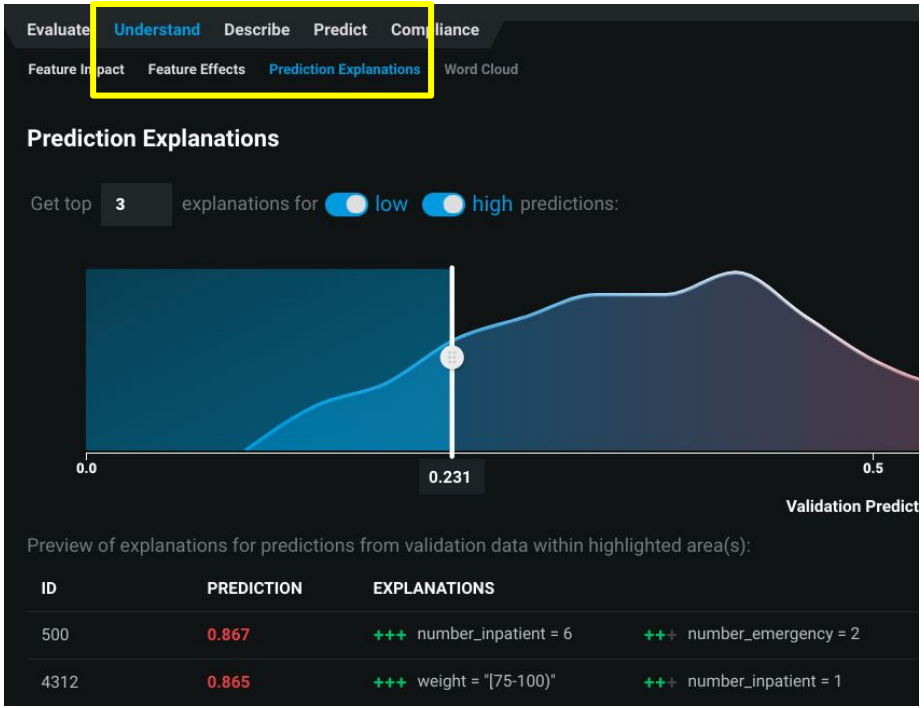


1) Fill in the 4 values corresponding to the cells of the confusion matrix. Adjust as needed

2) As you change the cost/payoff values you will get the optimal threshold and the maximum profit for the set you’re evaluating*.

*You may need to adjust total value for sampling. For example, if your evaluation set is a 20% sample of one month - estimated annual value = max profit * 5 * 12 mo.

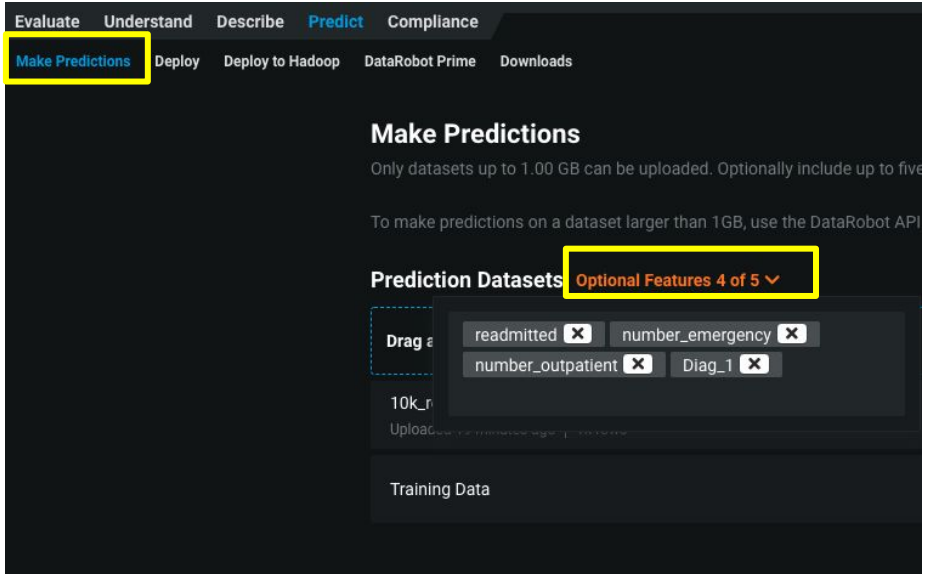
Model Review - Step 5. “Why did this record score high?”



- **Prediction Explanations** are the main drivers for each individual score are provided in decreasing order of impact
- **Strength** shows *direction* (impact is positive or negative) and *intensity* (small, medium, & large effect).
- If acting on your predictions are important, it’s crucial to have actionnable features.

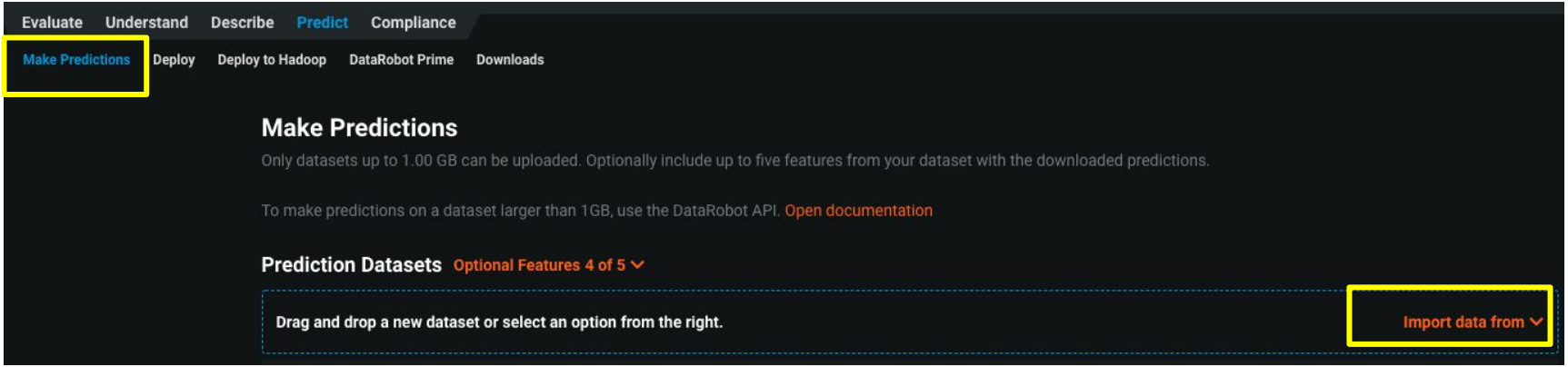
Predictions - Step 1. Get Ready

Update your ***decision boundary*** threshold (binary classification only)



DataRobot will just give you the score unless you specifically add other columns to “pass through” - it’s a good idea to add the ***TARGET*** and an ***ID*** field

Predictions - Step 2. Load Prediction File



- Import (or drag) the file into the Prediction Dataset area
- **Gotcha #1** - your file **MUST** have all raw columns used by the model; processing *<u>inside</u>* DataRobot during training will be replicated - features are *case sensitive*...
- **Gotcha #2** - if a column has changed since training (e.g. temp was ‘F, now in ‘C) it may significantly affect your predictions.

Predictions - Step 3. Prediction explanations

- Specify the number of explanations and range (this is time-consuming so focus on just the “interesting” score range)
- Select “**Update**”

- Next, under Compute & Download, select Calculator. Then, download the file.