Datarobot, Automated ML

Workflow and Benefits therein

홍운표 DataRobot

DataRobot

The world's most advanced Enterprise Machine Learning Automation platform

2012

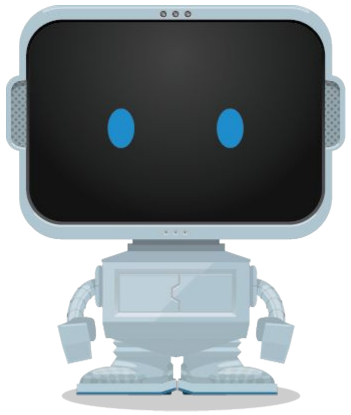
Founded, HQ in Boston, MA

$224M

In funding

1,000,000,000+

Models built on DataRobot Cloud



250+

Data Scientists & Engineers (of 600+)

4

#1 ranked Data Scientists kaggle

50+

Top 3 finishes kaggle

INSURANCE FINTECH HEALTHCARE MARKETING BANKING MANY MORE

Best Practices and Technology

The top ranked Data Scientists in the world

Owen Zhang
Product Advisor
Highest: 1st
MASTER

Xavier Conort
Chief Data Scientist
Highest: 1st
MASTER

Sergey Yurgenson
Data Scientist
Highest: 1st
MASTER

Amanda Schierz
Data Scientist
Current: 1st Female, 1st in UK
MASTER

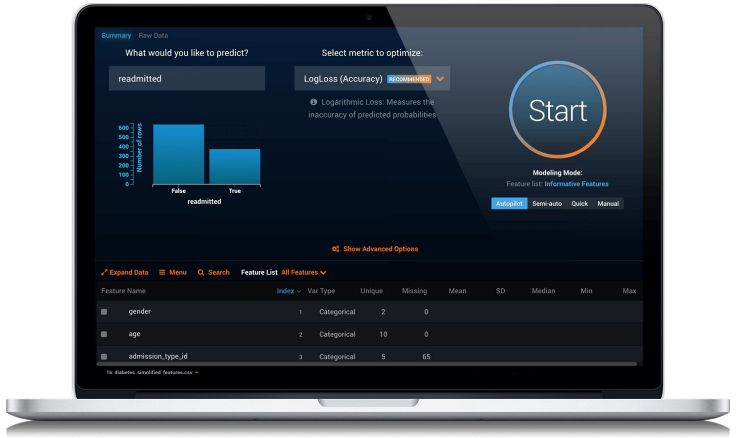
Jeremy Achin
CEO & Co-Founder
Highest: 20th
MASTER

Tom de Godoy
CTO & Co-Founder
Highest: 20th
MASTER

The best technologies in the world

XG Boost

H_2O



What would you like to predict?

readmitted

Select metric to optimize:

LogLoss (Accuracy)

Logarithmic Loss: Measures the inaccuracy of predicted probabilities

Modeling Mode:

Feature list: Informative Features

Autopilot | Semi-auto | Quick | Manual

Feature List All Features

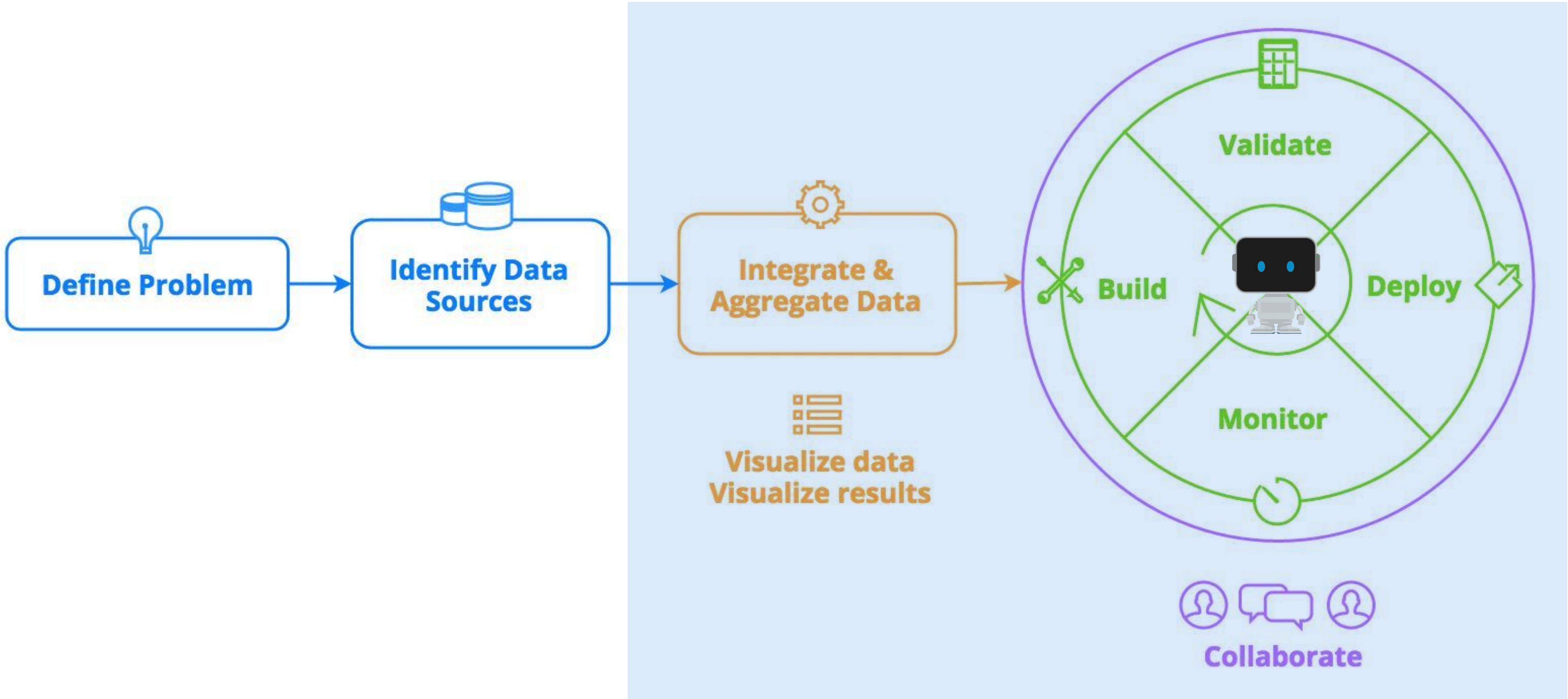
Feature Name	Index	Var Type	Unique	Missing	Mean	SD	Median	Min	Max
gender	1	Categorical	2	0					
age	2	Categorical	10	0					
admission_type_id	3	Categorical	5	65					

1k_diabetes_simplified_features.csv

DataRobot

Data Science Methodology

Due to limited resource, call for amelioration



No target goal

A few algorithms & prone to overfit

Aging of model

Not sufficient Explanations

Motivations for AutoML

Value of diverse set of algorithms

Methodology driven

↓

Problem driven

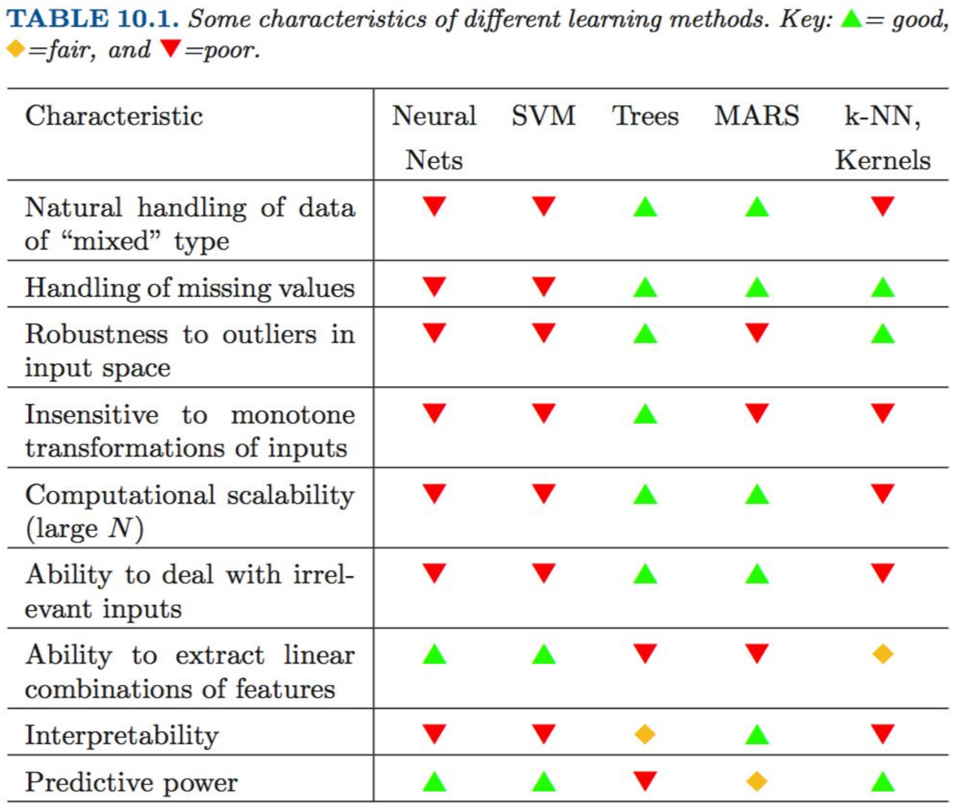


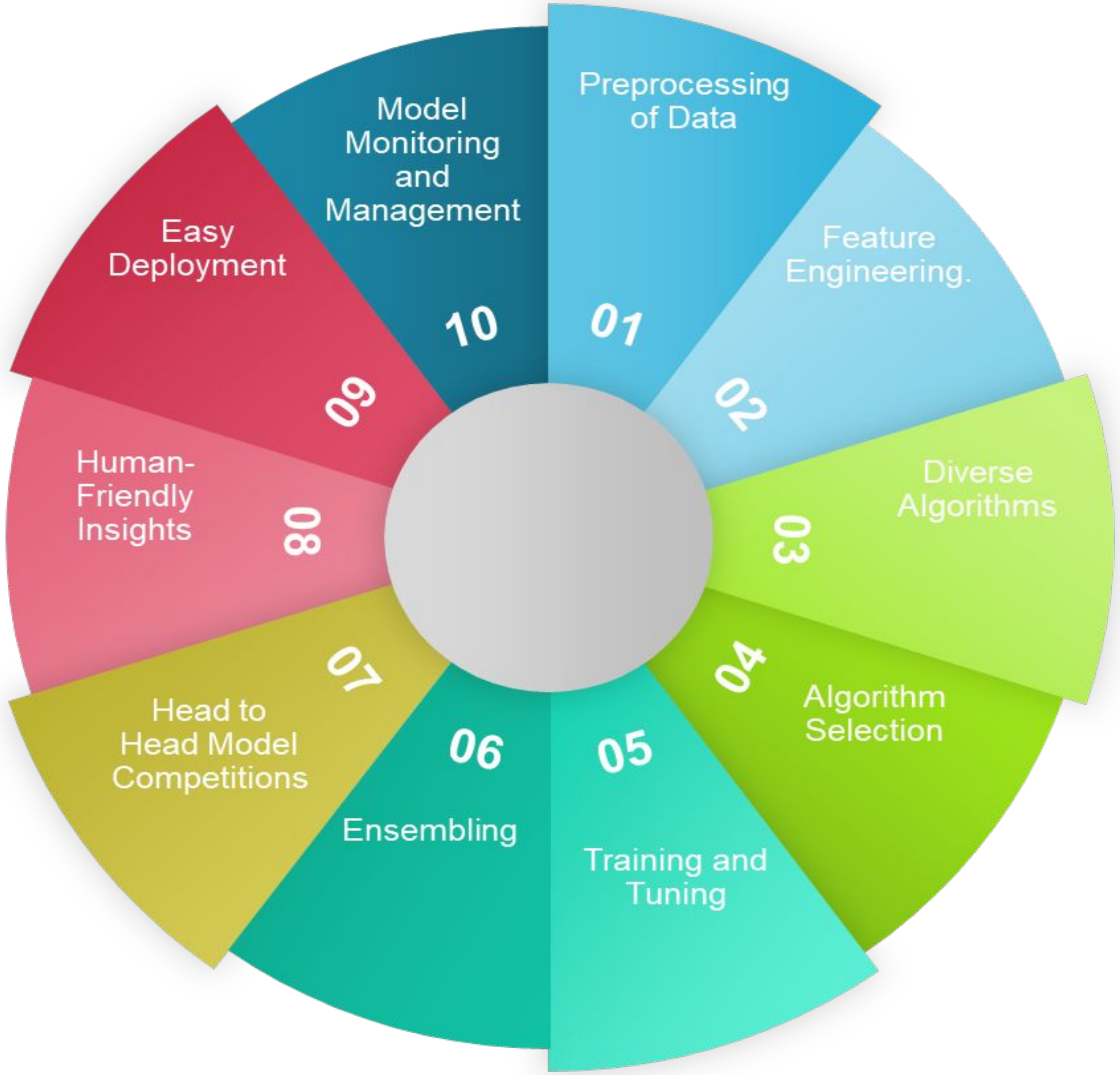
TABLE 10.1. *Some characteristics of different learning methods. Key: ▲= good, ◆=fair, and ▼=poor.*

Characteristic	Neural Nets	SVM	Trees	MARS	k-NN, Kernels
Natural handling of data of "mixed" type	▼	▼	▲	▲	▼
Handling of missing values	▼	▼	▲	▲	▲
Robustness to outliers in input space	▼	▼	▲	▼	▲
Insensitive to monotone transformations of inputs	▼	▼	▲	▼	▼
Computational scalability (large N)	▼	▼	▲	▲	▼
Ability to deal with irrelevant inputs	▼	▼	▲	▲	▼
Ability to extract linear combinations of features	▲	▲	▼	▼	◆
Interpretability	▼	▼	◆	▲	▼
Predictive power	▲	▲	▼	◆	▲

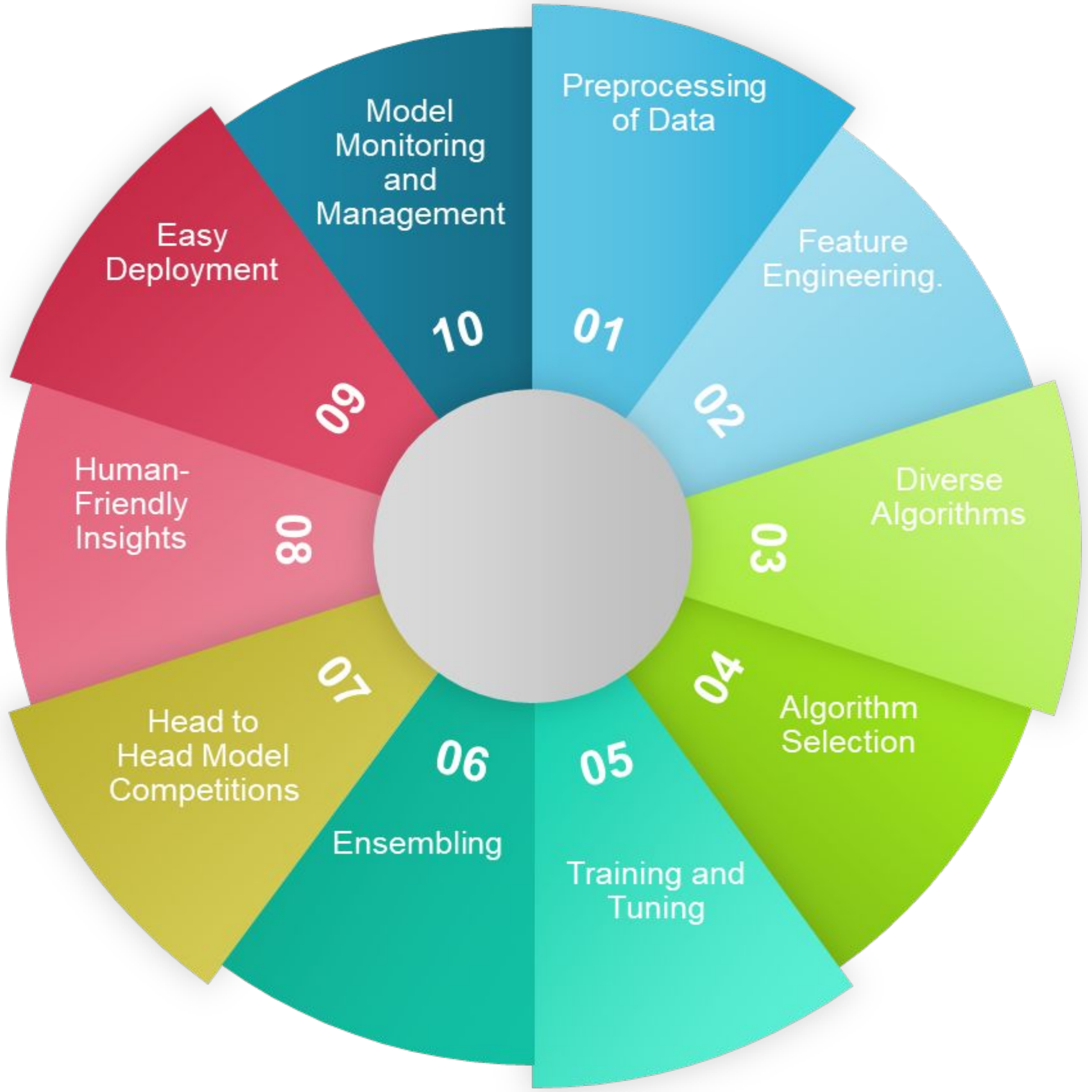
Source: *http://statweb.stanford.edu/~tibs/ElemStatLearn/*

What is Automated Machine Learning

- 10 steps to building models
- An expert system that knows how to do each of these 10 steps, without human instructions
- Human friendly – not a black box
- Fast and accurate
- Replicable data science



What about DataRobot?

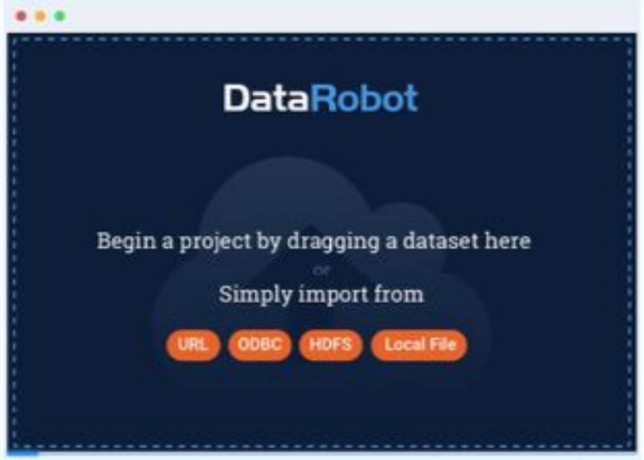


Key Points

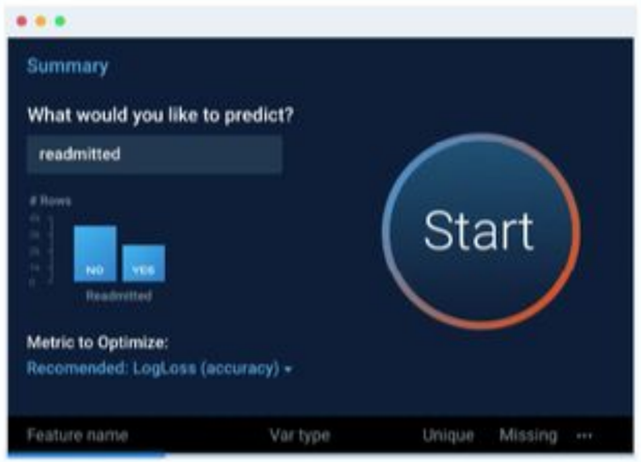
- End to end automated machine learning – all 10 steps are automated
- Hundreds of algorithms in the repository with new algorithms being added regularly
- Chooses the best algorithms for your data
- Best-in-class human-friendly insights
- Widest range of deployment options
- Enterprise ready
- Automatic model reports
- Large support team around the world

DataRobot Workflow

1. Ingest data



2. Select your prediction variable


Summary

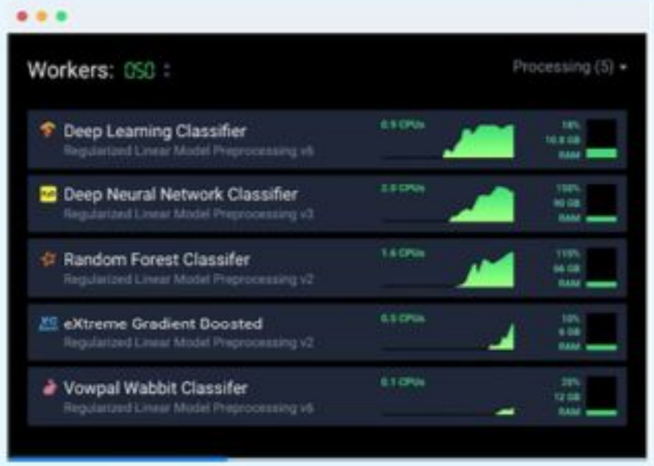
What would you like to predict?

readmitted

Metric to Optimize:

Recomended: LogLoss (accuracy) ▾

3. DataRobot builds and tests hundreds of models to find the best fit


Workers: 050 ⁝

Processing (5) ▾

Deep Learning Classifier
Regularized Linear Model Preprocessing v6

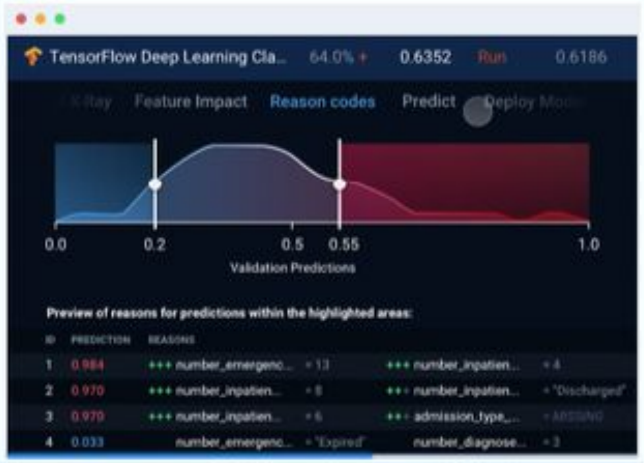
Deep Neural Network Classifier
Regularized Linear Model Preprocessing v3

Random Forest Classifer
Regularized Linear Model Preprocessing v2

eXtreme Gradient Doosted
Regularized Linear Model Preprocessing v2

Vowpal Wabbit Classifer
Regularized Linear Model Preprocessing v6

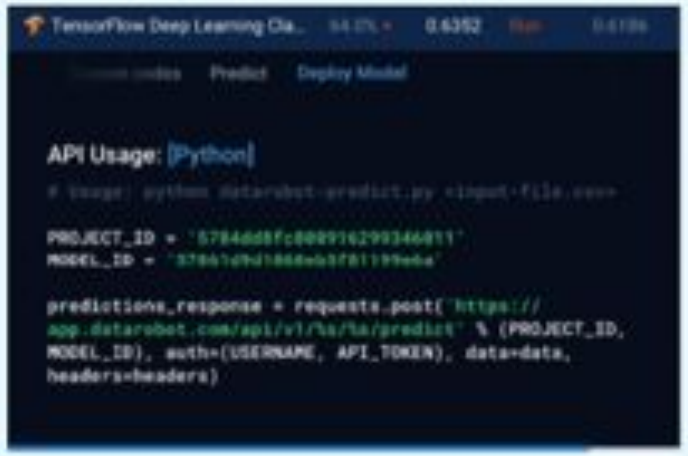
4. Explore the best models



Preview of reasons for predictions within the highlighted areas:

ID	PREDICTION	REASONS			
1	0.984	+++ number_emergenc...	= 13	+++ number_inpatien...	= 4
2	0.970	+++ number_inpatien...	= 8	++ number_inpatien...	= "Discharged"
3	0.970	+++ number_inpatien...	= 6	++ admission_type_...	= MISSING
4	0.033	number_emergenc...	= "Expired"	number_diagnose...	= 3

5. Deploy the model and make predictions in your application



API Usage: [Python]

```
# Usage: python datarobot-predict.py <input-file.csv>

PROJECT_ID = '5784dd8fc800916299346011'
MODEL_ID = '57861d9d1868eb3f81199eba'

predictions_response = requests.post('https://
app.datarobot.com/api/v1/%s/%s/predict' % (PROJECT_ID,
MODEL_ID), auth=(USERNAME, API_TOKEN), data=data,
headers=headers)
```

Different but powerful way of analysis

A few perspectives (many more)

Single model

No need of **Hold-out** partition : just train/test or k-fold CV

Only **interpretable** algorithm is chosen □ Linear model is preferable

Blending starts from existing model

Interaction should be considered for model performance (linear model)

parameter tuning is limited for a model and time-consuming

Multiple models

Hold-out partition for evaluation of several models

Interpretability is model-agnostic

Blending is fair-basis reflecting multiple models performance, with speed vs accuracy data

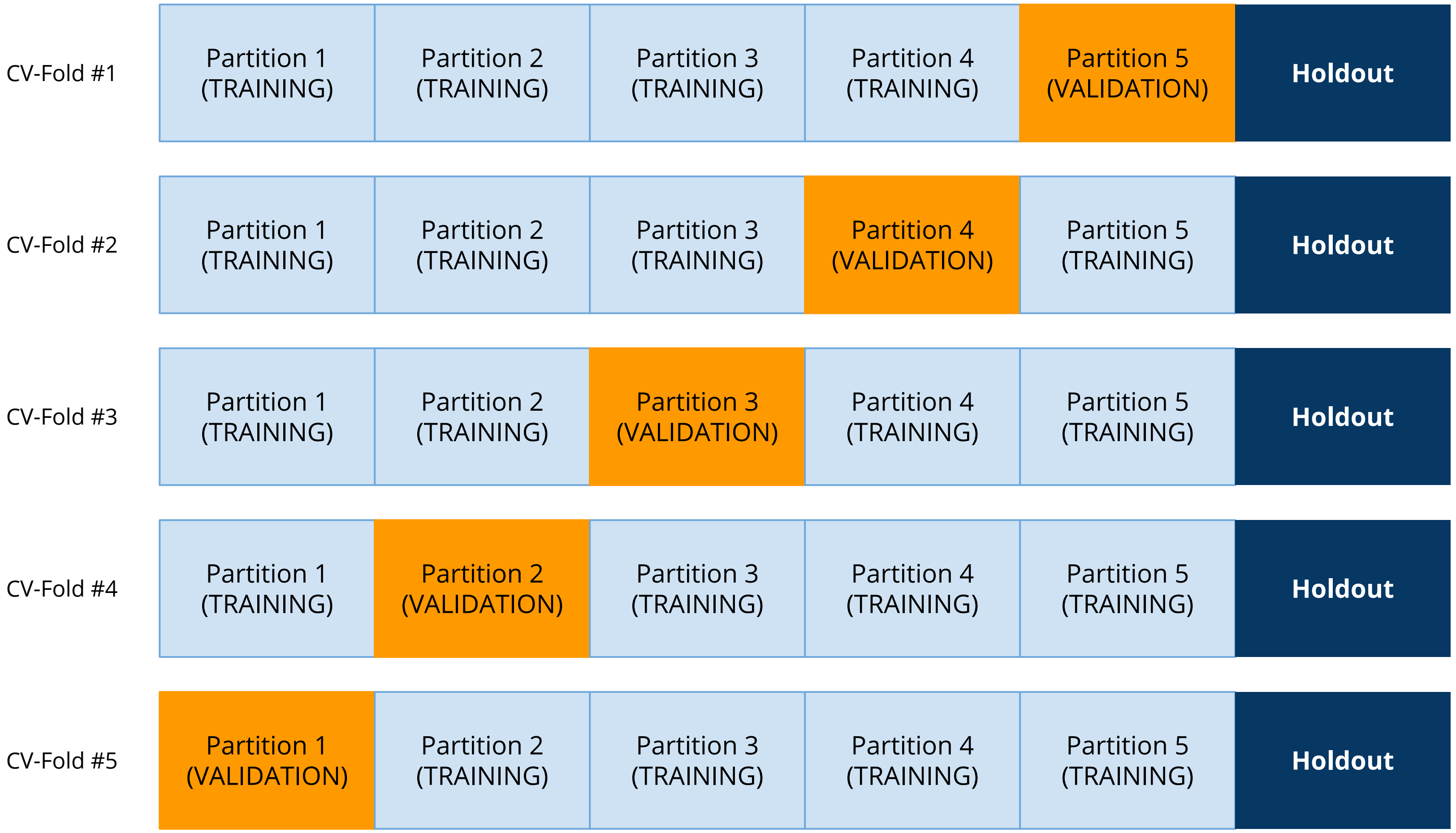
Interaction automatically reflected in tree-based algorithms. If interaction should be of importance, DR has GA2M model and R/Python api support for that

Parameter tuning is exhaustive for all candidate models.

One can easily confine the search space and quickly get the results

Benefits : safer model

Robust model free from the risk of overfitting



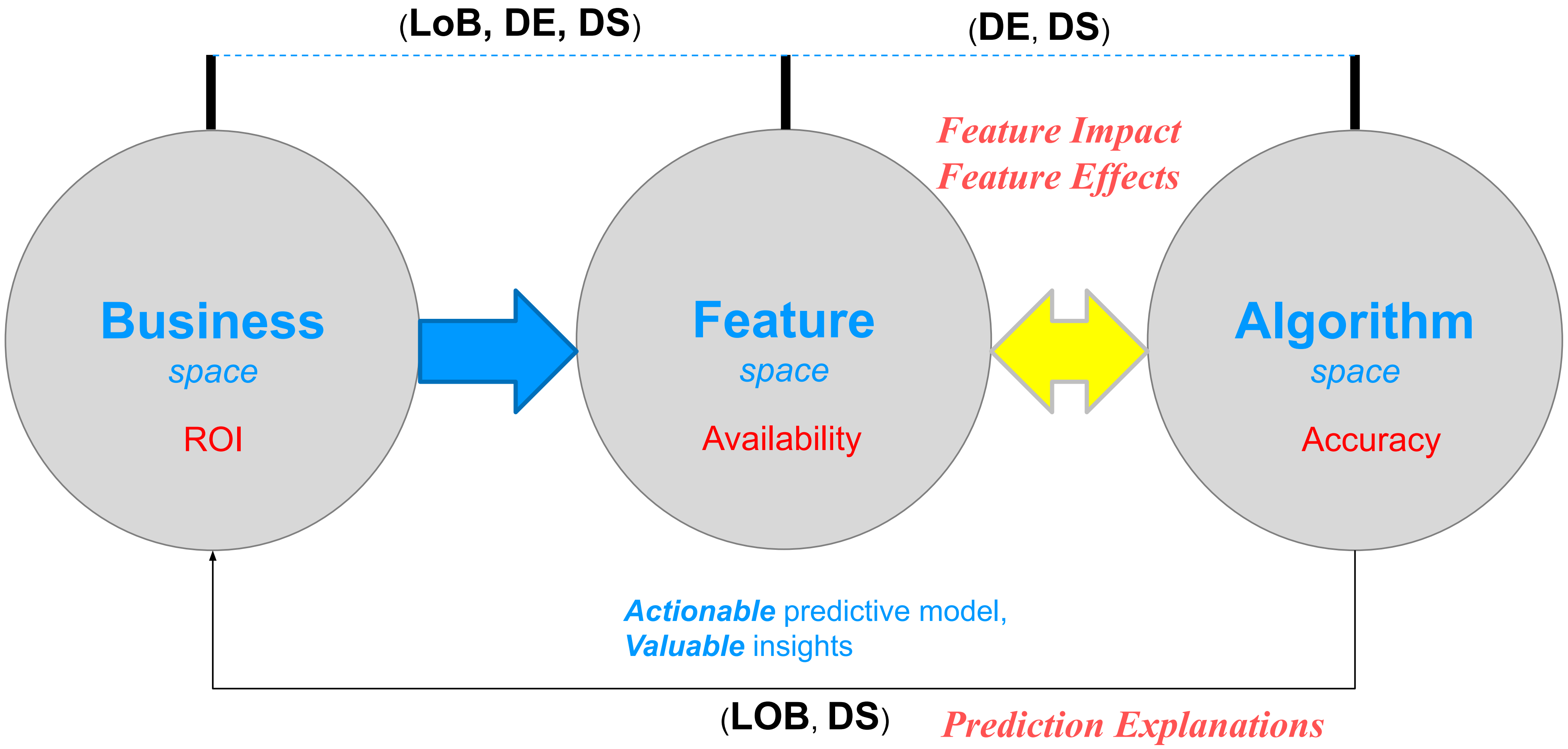
CV-Fold #1	Partition 1 (TRAINING)	Partition 2 (TRAINING)	Partition 3 (TRAINING)	Partition 4 (TRAINING)	Partition 5 (VALIDATION)	**Holdout**
CV-Fold #2	Partition 1 (TRAINING)	Partition 2 (TRAINING)	Partition 3 (TRAINING)	Partition 4 (VALIDATION)	Partition 5 (TRAINING)	**Holdout**
CV-Fold #3	Partition 1 (TRAINING)	Partition 2 (TRAINING)	Partition 3 (VALIDATION)	Partition 4 (TRAINING)	Partition 5 (TRAINING)	**Holdout**
CV-Fold #4	Partition 1 (TRAINING)	Partition 2 (VALIDATION)	Partition 3 (TRAINING)	Partition 4 (TRAINING)	Partition 5 (TRAINING)	**Holdout**
CV-Fold #5	Partition 1 (VALIDATION)	Partition 2 (TRAINING)	Partition 3 (TRAINING)	Partition 4 (TRAINING)	Partition 5 (TRAINING)	**Holdout**

The holdout is completely hidden from the models during the training process. After you have selected your optimal model, you can score your model on this to get your holdout score.

Average of these 5 validation scores is the cross validation score

Benefits : more effort on feature space

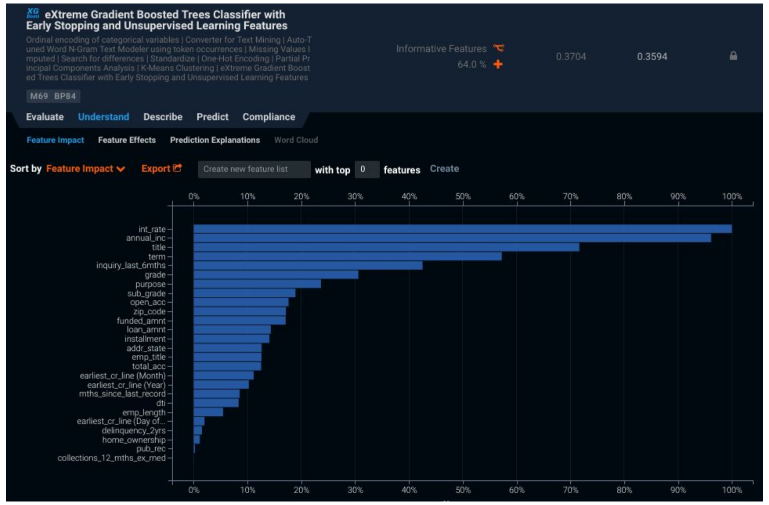
”Feature engineering is the art of data science” (Sergey Yurgenson)



DataRobot

Benefits : Explainability

Model-agnostic explanation



eXtreme Gradient Boosted Trees Classifier with Early Stopping and Unsupervised Learning Features

Ordinal encoding of categorical variables | Converter for Text Mining | Auto-Tuned Word N-Gram Text Modeler using token occurrences | Missing Values Imputed | Search for differences | Standardize | One-Hot Encoding | Partial Principal Components Analysis | K-Means Clustering | eXtreme Gradient Boosted Trees Classifier with Early Stopping and Unsupervised Learning Features

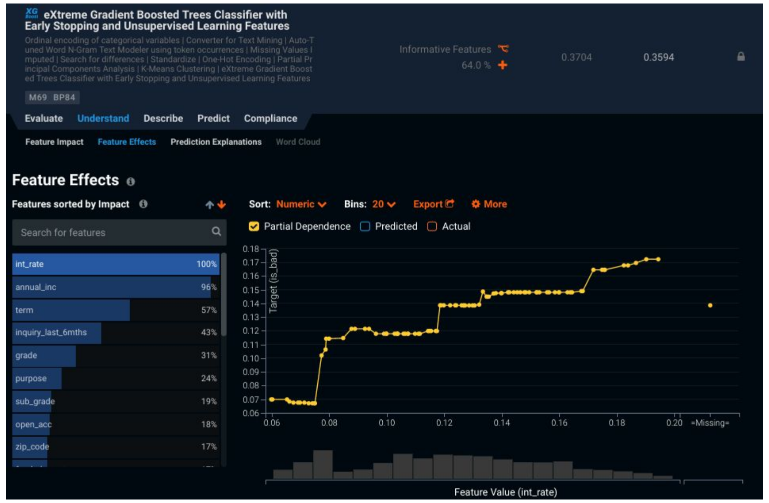
M69 BP84

Informative Features 64.0 % 0.3704 0.3594

Sort by Feature Impact Export Create new feature list with top 0 features Create

[Feature Impact]

- The importance of each feature
- Coincides with domain knowledge?
- Any new insights?



eXtreme Gradient Boosted Trees Classifier with Early Stopping and Unsupervised Learning Features

Ordinal encoding of categorical variables | Converter for Text Mining | Auto-Tuned Word N-Gram Text Modeler using token occurrences | Missing Values Imputed | Search for differences | Standardize | One-Hot Encoding | Partial Principal Components Analysis | K-Means Clustering | eXtreme Gradient Boosted Trees Classifier with Early Stopping and Unsupervised Learning Features

M69 BP84

Informative Features 64.0 % 0.3704 0.3594

Feature Effects

Features sorted by Impact

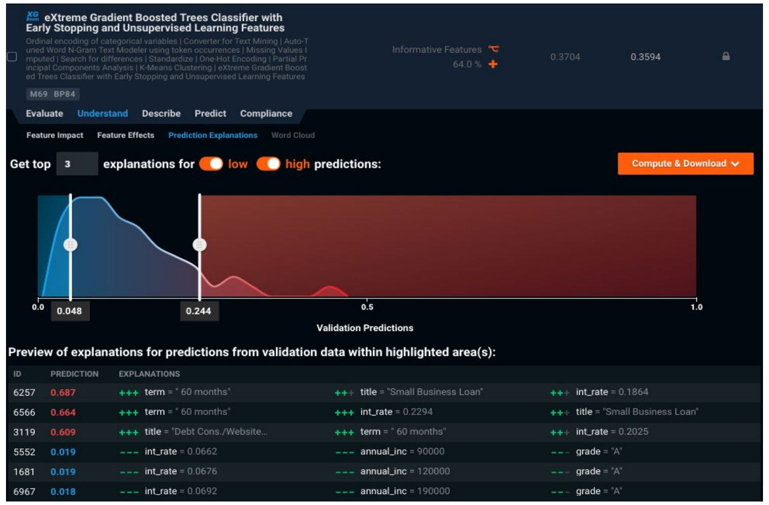
int_rate	100%
annual_inc	96%
term	57%
inquiry_last_6mths	43%
grade	31%
purpose	24%
sub_grade	19%
open_acc	18%
zip_code	17%

Sort: Numeric Bins: 20 Export More

☑ Partial Dependence □ Predicted □ Actual

[Feature Effect]

- Relationship among target and a feature
- Relationship reflects domain knowledge?
- Any new insights or feature transform?



eXtreme Gradient Boosted Trees Classifier with Early Stopping and Unsupervised Learning Features

Ordinal encoding of categorical variables | Converter for Text Mining | Auto-Tuned Word N-Gram Text Modeler using token occurrences | Missing Values Imputed | Search for differences | Standardize | One-Hot Encoding | Partial Principal Components Analysis | K-Means Clustering | eXtreme Gradient Boosted Trees Classifier with Early Stopping and Unsupervised Learning Features

M69 BP84

Informative Features 64.0 % 0.3704 0.3594

Get top 3 explanations for low high predictions:

Preview of explanations for predictions from validation data within highlighted area(s):

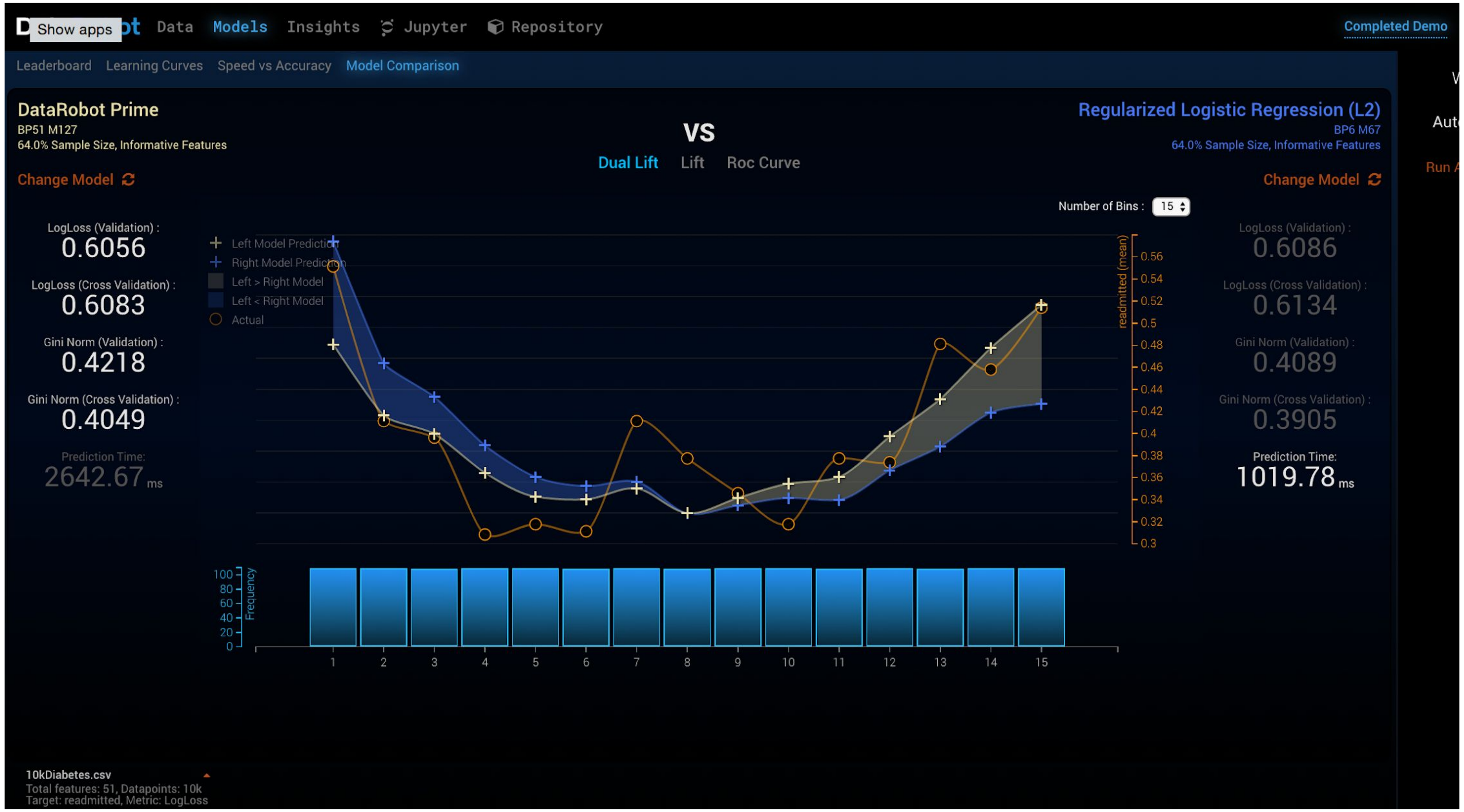
ID	PREDICTION	EXPLANATIONS		
6257	0.687	+++ term = " 60 months"	++ title = "Small Business Loan"	++ int_rate = 0.1864
6566	0.664	+++ term = " 60 months"	+++ int_rate = 0.2294	++ title = "Small Business Loan"
3119	0.609	+++ title = "Debt Cons./Website...	+++ term = " 60 months"	++ int_rate = 0.2025
5552	0.019	--- int_rate = 0.0662	--- annual_inc = 90000	-- grade = "A"
1681	0.019	--- int_rate = 0.0676	--- annual_inc = 120000	-- grade = "A"
6967	0.018	--- int_rate = 0.0692	--- annual_inc = 190000	-- grade = "A"

[Prediction Explanation]

- What is the basis of prediction?
- The predictions are reliablable to business people?

Benefits : effective blending

Search over candidates which promises tangible improvement



DataRobot Prime
BP51 M127
64.0% Sample Size, Informative Features

LogLoss (Validation) : 0.6056
LogLoss (Cross Validation) : 0.6083
Gini Norm (Validation) : 0.4218
Gini Norm (Cross Validation) : 0.4049
Prediction Time: 2642.67 ms

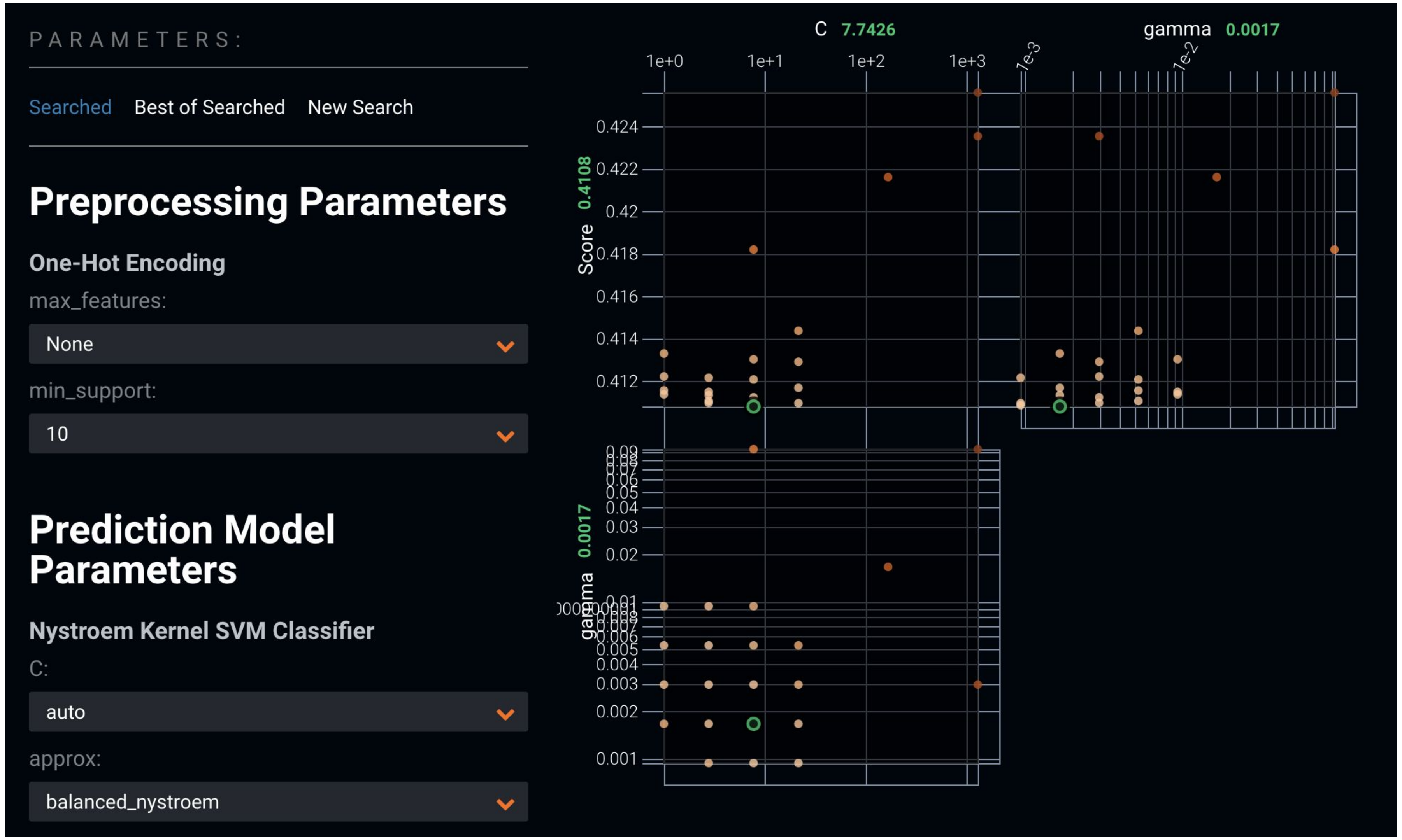
Regularized Logistic Regression (L2)
BP6 M67
64.0% Sample Size, Informative Features

LogLoss (Validation) : 0.6086
LogLoss (Cross Validation) : 0.6134
Gini Norm (Validation) : 0.4089
Gini Norm (Cross Validation) : 0.3905
Prediction Time: 1019.78 ms

10kDiabetes.csv
Total features: 51, Datapoints: 10k
Target: readmitted, Metric: LogLoss

Benefits : Hyper-param Tuning

Gradient-free and effective pattern search



PARAMETERS:

Searched | Best of Searched | New Search

Preprocessing Parameters

One-Hot Encoding

max_features: None

min_support: 10

Prediction Model Parameters

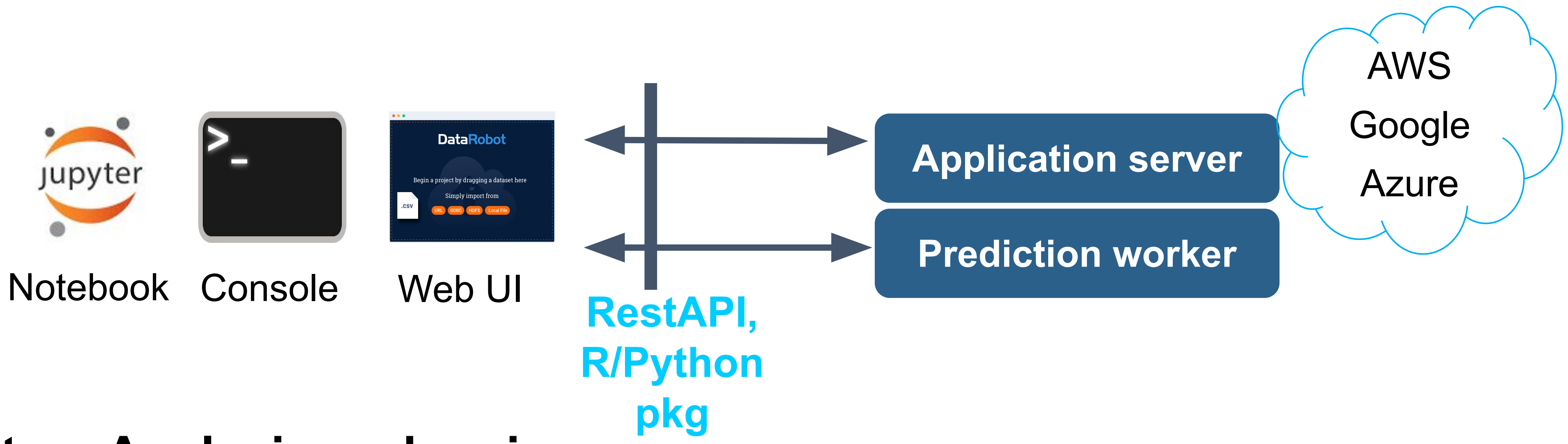
Nystroem Kernel SVM Classifier

C: auto

approx: balanced_nystroem

Benefits : API integration

data scientists and developers can use API



Custom Analysis and various Analysis

Model Factory

Automatic Model Refresh

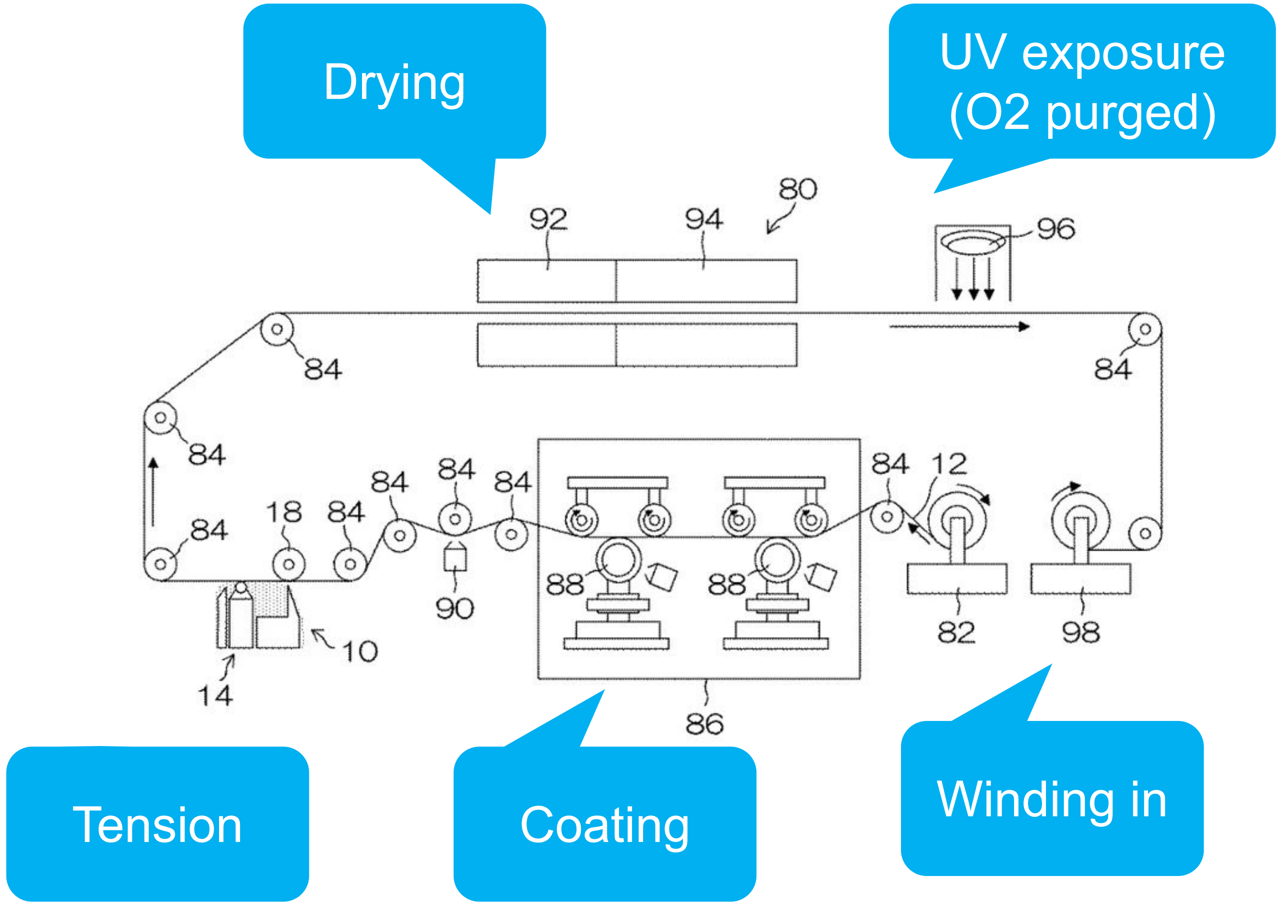
Model Diags & Viz

Feature Engineering

App. Integration

Demo : Bleedout prediction

Binary classification for QA



Drying

UV exposure (O2 purged)

Tension

Coating

Winding in

Process: Coating of thin film by covering the surface with coating solution and drying, followed by polymerizing with UV-light.

Problem: Unintended precipitation of powder such as unpolymerized monomer, antioxidant occurs causing “bleedout”. It spoils the product and contaminates the production line.

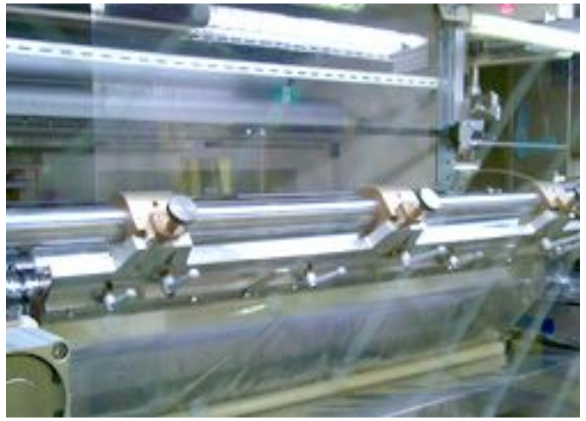
Data:

- Material: length of film roll
- Project type: production vs experiment
- Control: winding tension, UV-exposure duration, O2 concentration etc

Demo : Bleedout MFG process

1) Unwinding

2) Coating



3) Drying

4) UV exposure

5) Tension Control

6) Winding

DataRobot