



DataRobot

Essentials

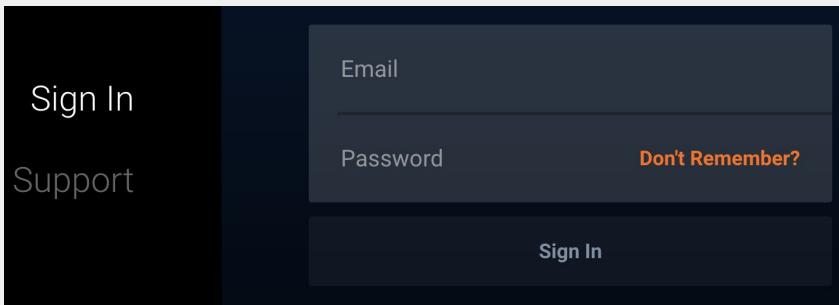




Logins

Using google chrome

<https://app.datarobot.com>



The screenshot shows a dark-themed login interface. On the left, there are links for "Sign In" and "Support". The main area has fields for "Email" and "Password", with a "Don't Remember?" link next to the password field. A "Sign In" button is at the bottom.

Username: {____}+dru{DMMMYYYY}@{____}.com

Password: DataRobot123!

<https://datarobot.litmos.com>



Please enter your username & password to access your online learning

Remember me on this computer

Login

[I've forgotten my username/password](#)

Username: {____}@{____}.com

Password: [check your email]



Who We Are

Founders



Jeremy Achin
CEO & Co-Founder



Tom de Godoy
CTO & Co-Founder

Top Data Scientists



Xavier Conort
Chief Data Scientist



Owen Zhang
Data Science Advisor



Sergey Yurgenson
Data Scientist



hosts worldwide data science competitions with over 1 million registered competitors as of June 2017. 12 members of the DataRobot team have been ranked in Kaggle's top 100 data scientists. Six are Kaggle Grandmasters.



Objectives



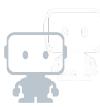
You Will Be Able To

- Follow a workflow giving you the best chance of success in machine learning projects
- Use DataRobot to approach and solve typical machine learning problems
- Discuss basic machine learning terminology with peers
- Tailor modeling decisions to meet business objectives



Not Covered

- Programming
- Advanced Extract-Transform-Load (ETL)
- Mathematical Theory / Machine Learning Algorithms



Structure



Explanation



Hands-on Exercises



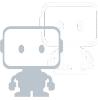
Hands-on Project



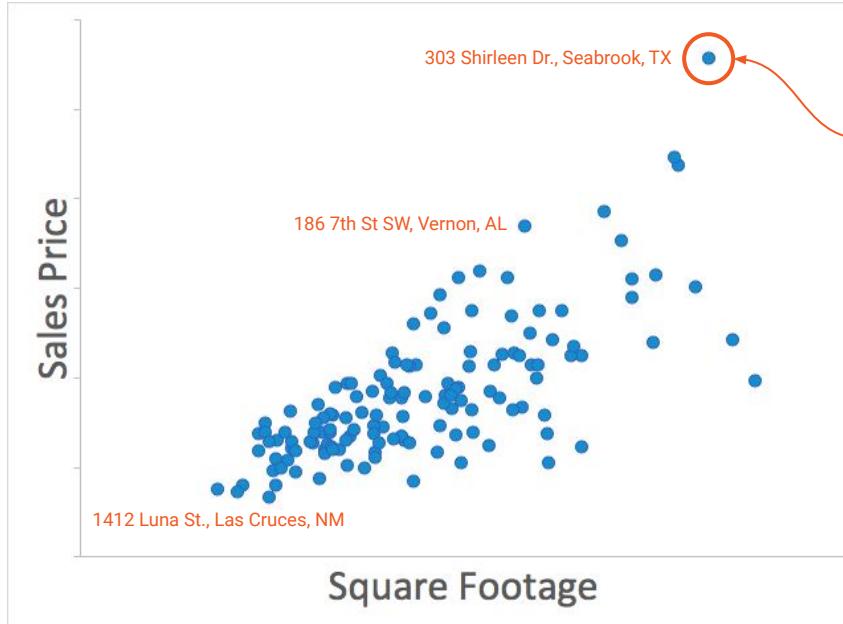
Surveys/Exam



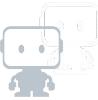
What is Machine Learning? AI?



Why do they call it “Machine Learning”?



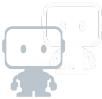
Every point is **an example** that the machine learns from



Why do they call it “Machine Learning”?

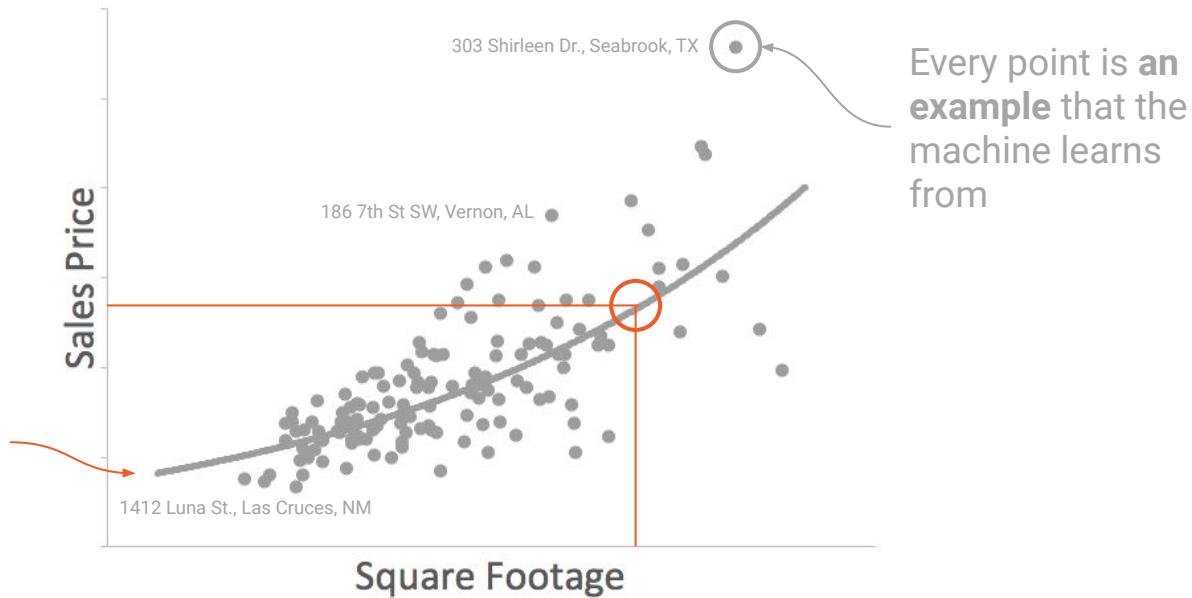
The line is a model. If you tell the model a square footage, it will make a prediction.





Why do they call it “Machine Learning”?

The line is a model. If you tell the model a square footage, it will make a prediction.



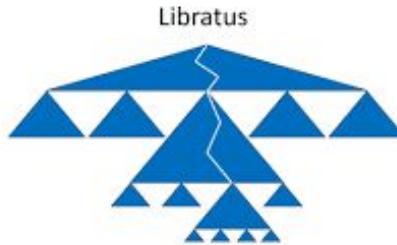
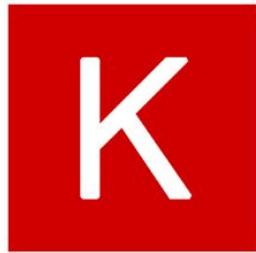
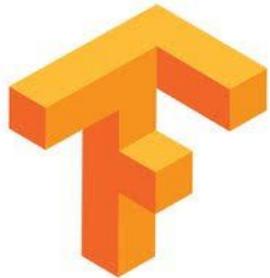
ML Working Definition



Using computer algorithms to uncover insights, determine relationships, and make predictions about future trends.



What About Deep Learning?



AI Working Definition



Enabling computer systems to perform tasks that ordinarily require human intelligence.

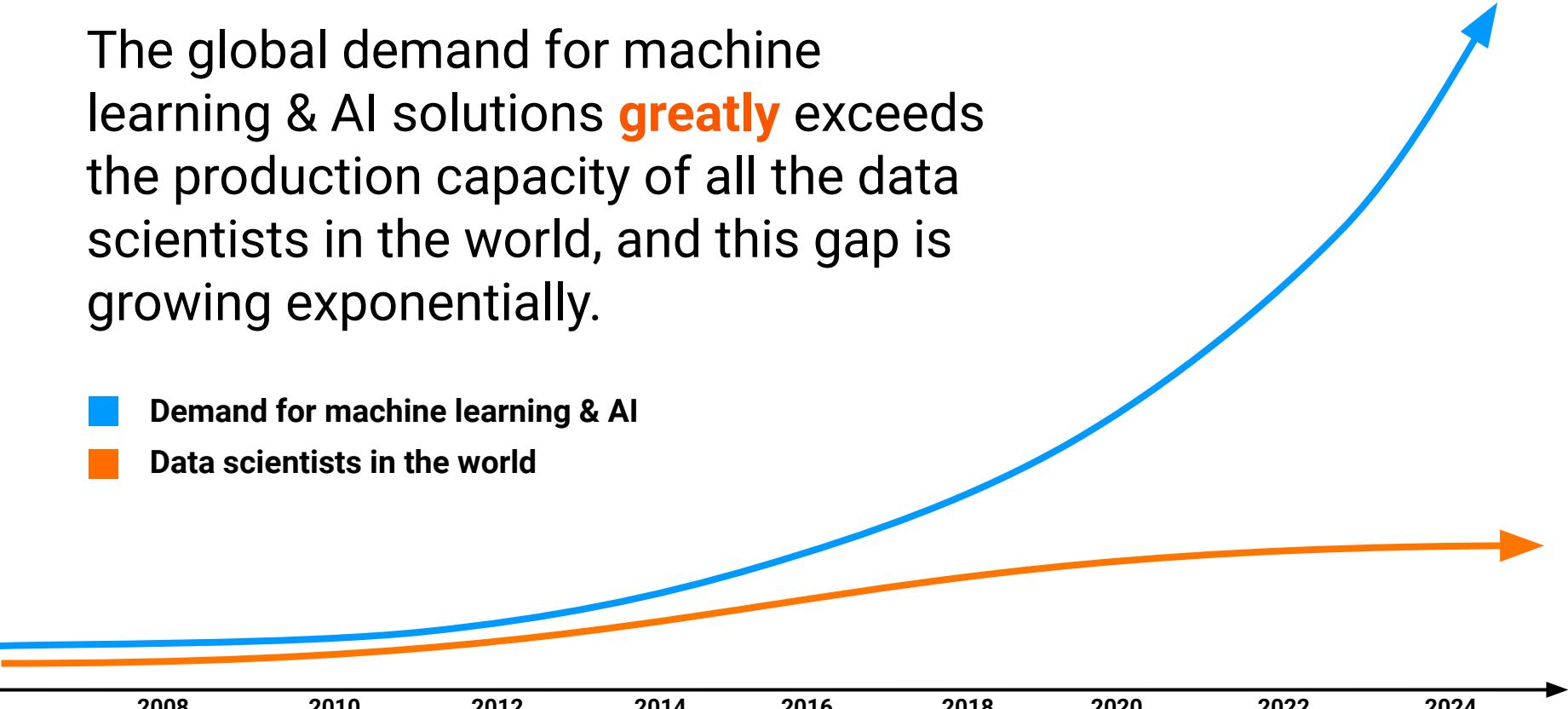
We use machine learning methods to create AI systems.



The Demand for AI

The global demand for machine learning & AI solutions **greatly** exceeds the production capacity of all the data scientists in the world, and this gap is growing exponentially.

- Demand for machine learning & AI
- Data scientists in the world

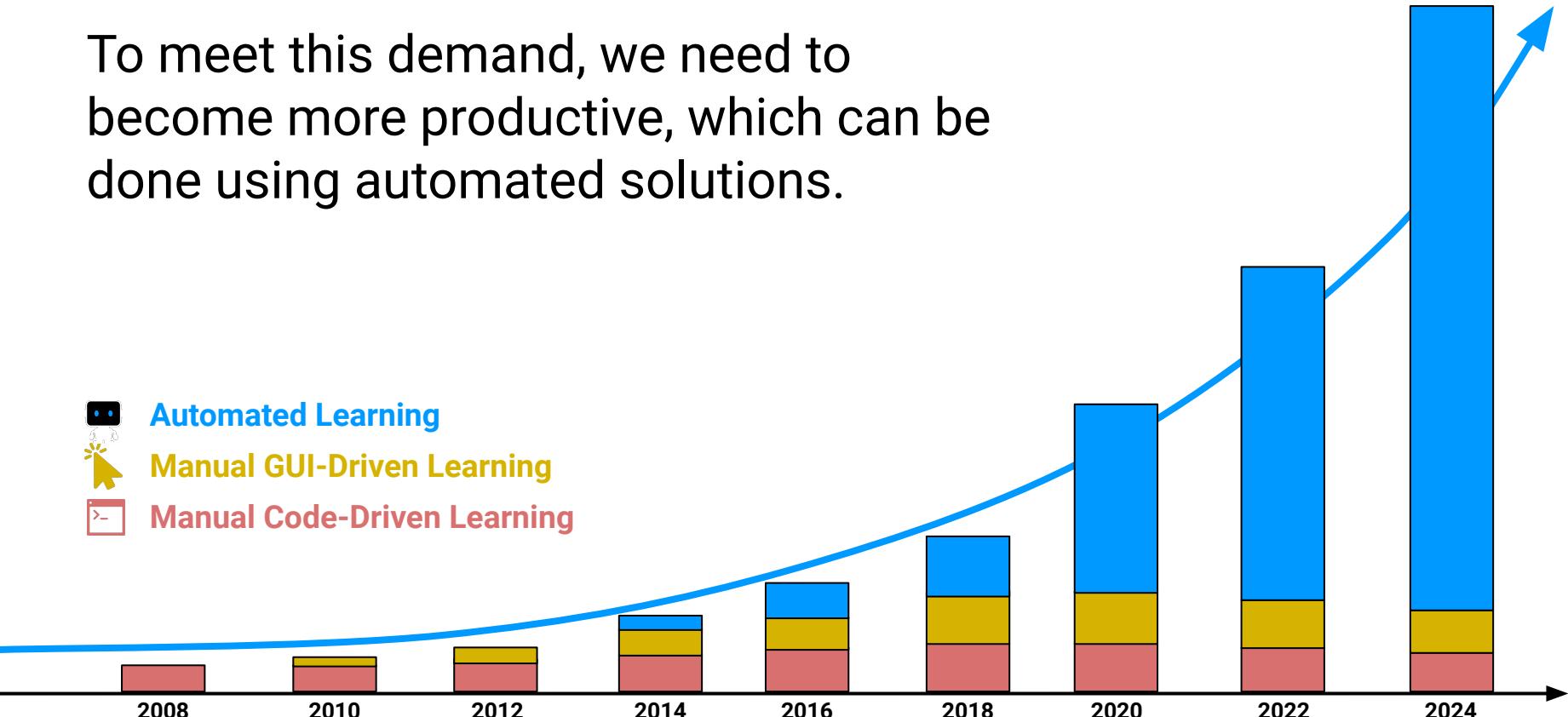


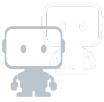


Bridging the Gap

To meet this demand, we need to become more productive, which can be done using automated solutions.

-  Automated Learning
-  Manual GUI-Driven Learning
-  Manual Code-Driven Learning





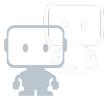
The DataRobot Interface

The screenshot shows the DataRobot interface homepage. At the top, there is a navigation bar with links for Data, Models, Deployments, Insights, Jupyter, and Repository. To the right of the navigation bar are several icons: a gear, a play button, a folder, and a user profile. Below the navigation bar, there is a large central area with a dark background featuring a white cloud-like shape. Inside this shape, the text "Begin a project by dragging a dataset here" is displayed above the word "or". Below "or", the text "simply import from:" is shown. Underneath this text are four orange buttons with white text: "Data Source", "URL", "HDFS", and "Local File". At the bottom of the central area, there is a note: "We currently accept .csv, .tsv, .dsv, .xls, .xlsx, .sas7bdat, .bz2, .gz, .zip, .tar, .tgz".

Begin a project by dragging a dataset here
or
simply import from:

Data Source URL HDFS Local File

We currently accept .csv, .tsv, .dsv, .xls, .xlsx, .sas7bdat, .bz2, .gz, .zip, .tar, .tgz



The Machine Learning Life Cycle



1. Define Project Objectives

- Specify problem
- Acquire subject matter expertise
- Define target and unit of analysis
- Prioritize modeling criteria
- Consider success criteria and risks
- Decide whether to continue

2. Acquire & Explore Data

- Find appropriate data
- Format data into single table
- Conduct exploratory data analysis
- Find and remove any target leakage
- Perform feature engineering

3. Model Data

- Select features
- Build candidate models
- Validate models

4. Interpret & Communicate

- Assess model quality
- Determine important features
- Identify relationships
- Explain predictions

5. Implement, Document & Maintain

- Select a model for deployment
- Document modeling process
- Create model monitoring and maintenance plan



Specify the Business Problem

Criteria

- State problem in language of business (no technical jargon)
- Specify actions that might result
- Include specifics (number of customers affected, costs etc.)
- Explain impact to the bottom line



Critique These Project Statements

Readmissions cost our hospital \$65m last year, and we don't have a way to determine which patients are at risk of readmission.

Criteria

- State problem in language of business (no technical jargon)
- Specify actions that might result
- Include specifics (number of customers affected, costs etc.)
- Explain impact to the bottom line



Critique These Project Statements

Warehouse operators want to know how long each truck will take to unload, so they can tell drivers of subsequent deliveries when the warehouse bay will be ready.

Criteria

- State problem in language of business (no technical jargon)
- Specify actions that might result
- Include specifics (number of customers affected, costs etc.)
- Explain impact to the bottom line



Critique These Project Statements

Our hedge fund is considering investing \$40 Million/yr in loans on LendingClub.com. These loans are an appealing investment with interest rates averaging 17%. However, 5% of invested money on the site is lost due to borrower defaults. You would like to be able to screen out the riskiest borrowers. If successful, your organization will use your approach to justify investing on the site. They will then apply the method in an automated fashion to choose which loans to fund.

Criteria

- State problem in language of business (no technical jargon)
- Specify actions that might result
- Include specifics (number of customers affected, costs etc.)
- Explain impact to the bottom line



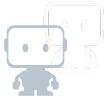
Acquire Subject Matter Expertise

Why

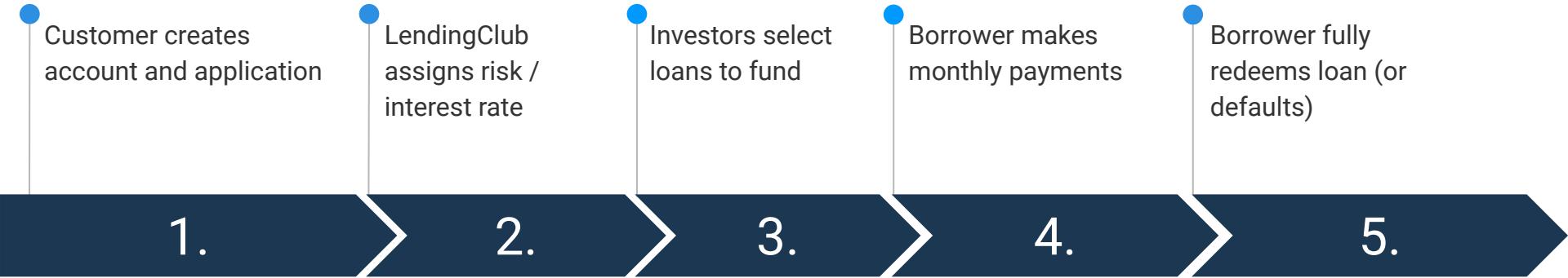
- Indicates obstacles and opportunities
- Suggests data collection and modeling ideas
- Finds data quality problems and improvement opportunities
- Sets expectations for model performance
- Clarifies alternatives to building model

How

- Talk to colleagues or subject matter experts (SMEs)
- Read (trade journals, Google, etc)

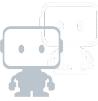


Subject Matter Expertise on Lending Club



Has disbursed \$42B in loans (as of 2018)

Mix of individual and institutional investors



Define Target

Retail Sales Data

Date	Customer ID	Store ID	Purchase Amount	Number of Items
10/2/2015	1037	17	\$107.23	3
10/2/2015	1038	17	\$99.50	2
10/5/2015	1037	17	\$212.49	5
10/5/2015	1091	19	\$37.04	2
10/5/2015	1302	19	\$18.02	1

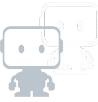
What do you want to predict?



Classification



Regression

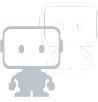


Define Unit of Analysis

Retail Sales Data

Date	Customer ID	Store ID	Purchase Amount	Number of Items
10/2/2015	1037	17	\$107.23	3
10/2/2015	1038	17	\$99.50	2
10/5/2015	1037	17	\$212.49	5
10/5/2015	1091	19	\$37.04	2
10/5/2015	1302	19	\$18.02	1

What does each row represent?



Define Unit of Analysis

Retail Sales Data

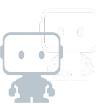
Date	Store ID	Purchase Amount	Number of Items
10/2/2015	17	\$206.73	5
10/5/2015	17	\$212.49	5
10/5/2015	19	\$55.06	3

What would be a good target and unit of analysis for Lending Club?



Prioritize Modeling Criteria

- Predictive performance
- Familiarity
- Prediction speed
- Speed to build model
- Interpretability?



Success Criteria



Who uses the model?



How much value can the model drive?



What modeling criteria will help get you there?

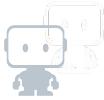


Risk Scenarios

Play devil's advocate. Be creative.

Examples:

- Change in underlying patterns so past is no longer predictive
- Business loses interest in outcome being modeled
- Model insufficiently predictive



Decide Whether To Continue



Estimate resources required



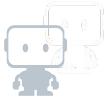
**Understand alternatives to
creating model**



Consider organizational risks



Estimate model's business value



The Machine Learning Life Cycle



1. Define Project Objectives

- Specify problem
- Acquire subject matter expertise
- Define target and unit of analysis
- Prioritize modeling criteria
- Consider success criteria and risks
- Decide whether to continue

2. Acquire & Explore Data

- Find appropriate data
- Format data into single table
- Conduct exploratory data analysis
- Find and remove any target leakage
- Perform feature engineering

3. Model Data

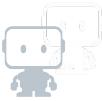
- Select features
- Build candidate models
- Validate models

4. Interpret & Communicate

- Assess model quality
- Determine important features
- Identify relationships
- Explain predictions

5. Implement, Document & Maintain

- Select a model for deployment
- Document modeling process
- Create model monitoring and maintenance plan



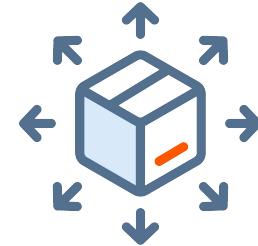
Find Appropriate Data



Internal



External



Public

What about for LendingClub?



Format Data for Modeling



**Modeling requires data
be in one “table”**



**Result should match
desired unit of analysis**



First Steps in Exploratory Data Analysis



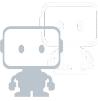
View raw table contents



Descriptive statistics



Verify variable/column types match expectation



Defining our Target *loan_is_bad*

FALSE

Fully Paid, Current

TRUE

In Grace Period, Late (16-30 days), Late (31-120 days), Default, Charged Off

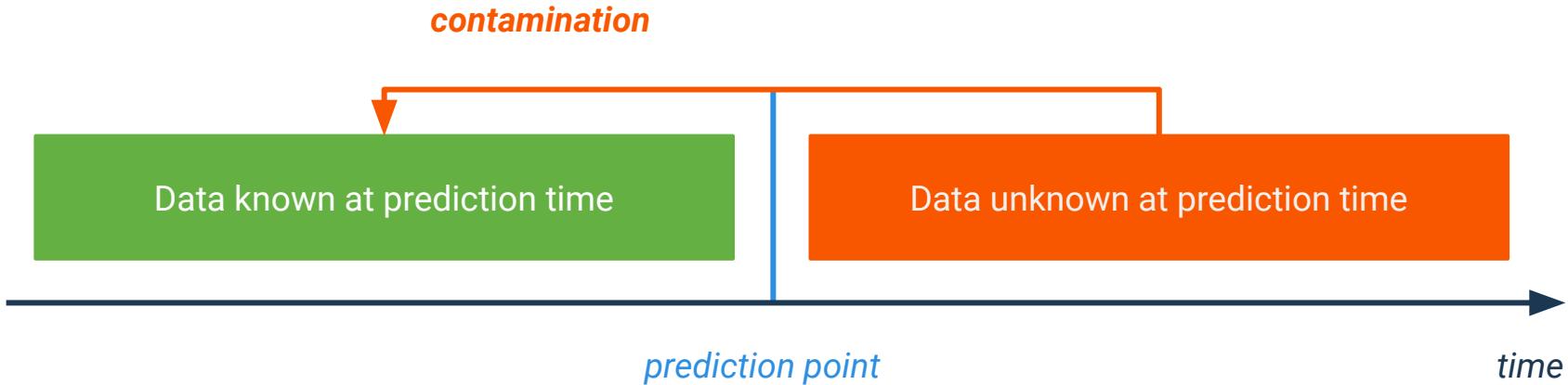


Load Data Into DataRobot

- Start a new project by clicking on the DataRobot icon.
- Load *early_2012_2013_loan_sample_with_outcome.csv.zip* into DataRobot.
- Explore the raw data in DataRobot and view some of the features.
- Set the target to `loan_is_bad` when prompted and run manual mode.
- Which features are most important according to the green bars?



Visualizing Target Leakage





Target Leakage Examples

Employee ID	Title	Experience Years	Monthly Salary GBP	Annual Income USD
315981	Data Scientist	3	5,000.00	78,895.44
4691	Data Scientist	4	5,500.00	86,784.98
23598	Data Scientist	5	6,200.00	97,830.35



Target Leakage Examples

Employee ID	Title	Experience Years	Monthly Salary GBP	Annual Income USD
315981	Data Scientist	3	5,000.00	78,895.44
4691	Data Scientist	4	5,500.00	86,784.98
23598	Data Scientist	5	6,200.00	97,830.35



Target Leakage Examples

Subscriber ID	Group	Daily Voice Usage	Daily SMS Usage	Daily Data Usage	Gender
24092091	M18-25	15.31	25	135.10	0
4092034091	F40-60	35.81	3	5.01	1
329815	F25-40	13.09	32	128.52	1
94721835	M25-40	18.52	21	259.34	0



Target Leakage Examples

Subscriber ID	Group	Daily Voice Usage	Daily SMS Usage	Daily Data Usage	Gender
24092091	M18-25	15.31	25	135.10	0
4092034091	F40-60	35.81	3	5.01	1
329815	F25-40	13.09	32	128.52	1
94721835	M25-40	18.52	21	259.34	0



Target Leakage Examples

Education	Married	Annual Income	Purpose	Late Payment Reminders	Is Bad Loan
1	Y	80k	Car Purchase	0	0
3	N	120k	Small Business	3	1
1	Y	85k	House Purchase	5	1
2	N	72k	Marriage	1	0



Target Leakage Examples

Education	Married	Annual Income	Purpose	Late Payment Reminders	Is Bad Loan
1	Y	80k	Car Purchase	0	0
3	N	120k	Small Business	3	1
1	Y	85k	House Purchase	5	1
2	N	72k	Marriage	1	0



How Can DataRobot Help?

DataRobot detected target leakage

As part of the feature analysis process, DataRobot has generated and run on a new feature list (Informative Features - Leakage Removed). This list excludes feature(s) that are at risk of causing target leakage and any features providing little or no information useful for modeling. To determine what was removed, you can see these features labeled on the Data table, All Features.

[READ MORE](#) [DISMISS](#)

Project Data Feature Lists

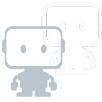
☰ Menu Search Feature List All Features View Raw Data Create Feature List 1-13 of 13

Feature Name	Index	Importance	Var Type	Unique	Missing	Mean	Std Dev	Median	Min	Max
wine_is_good	13	Target	Categorical	2	0					
[Target Leakage] quality	12	High	Numeric	7	0	5.87	0.89	6	3	10
alcohol	11	Medium	Numeric	99	0	10.52	1.23	10.40	8	14.9
density	8	Medium	Numeric	817	0	0.99	3.03e-3	0.99	0.99	1
chlorides	5	Medium	Numeric	141	0	0.05	0.02	0.04	9.00e-3	0
total_sulfur_dioxide	7	Low	Numeric	246	0	138	43.05	133	9	4
residual_sugar	4	Low	Numeric	291	0	6.37	5.13	5.10	0.60	65
pH	9	Low	Numeric	101	0	3.19	0.15	3.17	2.74	3
volatile_acidity	2	Low	Numeric	120	0	0.28	0.10	0.26	0.08	1

WORKERS
Using 0 of 20 total workers across all projects 04

STATUS
Autopilot has finished

ACTIONS
Rerun Autopilot
Unlock Holdout for all models



Load Data Into DataRobot

- Start a new project again by clicking on the DataRobot icon.
- Load `early_2012_2013_train.csv.zip` into DataRobot.
- Set the target to `loan_is_bad` and run autopilot.



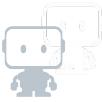
Lunch



Feature Engineering is the Art Part of Data Science

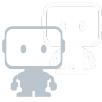
Sergey Yurgenson

Former #1 ranked modeler on Kaggle.com



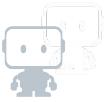
Feature Engineering - Dates

Manufacture Date	Is Defective
10/5/15	True
10/6/15	False
10/7/15	False
10/8/15	False
10/9/15	True



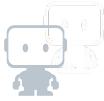
Feature Engineering - Dates

Manufacture Date	Day of the Week	Is Defective
10/5/15	Monday	True
10/6/15	Tuesday	False
10/7/15	Wednesday	False
10/8/15	Thursday	False
10/9/15	Friday	True



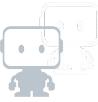
Feature Engineering - Generalizing

Birth Year	Is Bad
10 Apr 1982	True
3 Dec 1978	False
5 Jan 1993	False
15 Jul 1994	False
1 Jan 1980	True



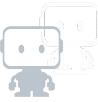
Feature Engineering - Generalizing

Birth Year	Age at Application Time	Is Bad
10 Apr 1982	30	True
3 Dec 1978	36	False
5 Jan 1993	24	False
15 Jul 1994	22	False
1 Jan 1980	31	True



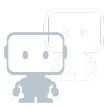
Feature Engineering - Aggregation

Date	Customer ID	Store ID	Purchase Amount	Number of Items
10/2/2015	1037	17	\$107.23	3
10/2/2015	1038	17	\$99.50	2
10/5/2015	1037	17	\$212.49	5
10/5/2015	1091	19	\$37.04	2
10/5/2015	1302	19	\$18.02	1



Feature Engineering - Aggregation

Date	Store ID	Purchase Amount	Number of Items	Number of Customers
10/2/2015	17	\$206.73	5	2
10/5/2015	17	\$212.49	5	1
10/5/2015	19	\$55.06	3	2



Feature Types

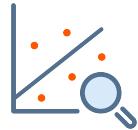
- Numerical
- Categorical
- Text



Numeric Features



Impute missing value and create flag for what was imputed



**Scaling transformations
(Ridit, Standardize, Squared, Log)**



Differences of features



Ratios of features



Categorical Features



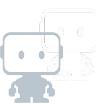
One-hot encoding



Count / ordinal encoding



Credibility estimates

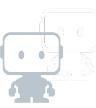


One-hot Encoding

Group
A
B
C
A
A



	Group_A	Group_B	Group_C
A	1	0	0
B	0	1	0
C	0	0	1
A	1	0	0
A	1	0	0

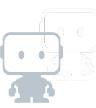


Count Encoding

Group
A
B
C
A
A



Count Encoding
3
1
1
3
3

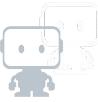


Ordinal Encoding

Group
A
B
C
A
A



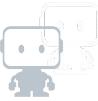
Ordinal Encoding
0
1
2
0
0



Credibility Estimates

$$count_k \times (\bar{y}_k - \bar{y}_{\cdot})$$

**The more we see a group, the more we “trust”
the group’s deviation from the overall mean**

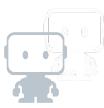


Credibility Estimates

Defaulted	Group
0	A
0	B
1	C
1	A
0	A

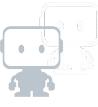


Credibility Estimates
$3 * (0.33 - 0.4)$
$1 * (0 - 0.4)$
$1 * (1 - 0.4)$
$3 * (0.33 - 0.4)$
$3 * (0.33 - 0.4)$



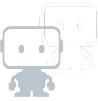
Feature Types

- ✓ **Numerical**
- ✓ **Categorical**
- **Text**



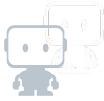
Text Features

Churn	Comments
0	I like AAA coffee
1	I dislike AAA coffee
1	I like AAA coffee, but prefer BBB coffee



Text Features (Tokenizing)

Churn	Comments	I	like	dislike	coffee	AAA	BBB	but	prefer
0	I like AAA coffee	1	1	0	1	1	0	0	0
1	I dislike AAA coffee	1	0	1	1	1	0	0	0
1	I like AAA coffee, but prefer BBB coffee	1	1	0	2	1	1	1	1



Text Features (N-grams)

Churn	Comments	Like AAA	Dislike AAA	Prefer BBB	...
0	I like AAA coffee	1	0	0	...
1	I dislike AAA coffee	0	1	0	...
1	I like AAA coffee, but prefer BBB coffee	1	0	1	...



Term Frequency

{number of times the word appears in the document}

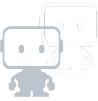
{total number of words in the document}

Inverse Document Frequency

{number of documents}

{number of documents that contain word}

TF * IDF



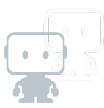
Text Features (TF-IDF)

Churn	Comments	I	like	dislike	coffee	AAA	BBB	but	prefer
0	I like AAA coffee	$\begin{aligned} &(1/4) * (3/3) \\ &= 0.25 \end{aligned}$	0.375	0	0.25	0.25	0	0	0
1	I dislike AAA coffee	0.25	0	0.75	0.25	0.25	0	0	0
1	I like AAA coffee, but prefer BBB coffee	0.125	0.188	0	$\begin{aligned} &(2/8) * (3/3) \\ &= 0.25 \end{aligned}$	0.125	0.375	0.375	0.375



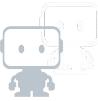
Other Text Pre-processing

- Word embeddings (word2vec, fastText)
- Deep learning methods (such as autoencoders)
- Cosine similarity between pairs of text columns (on datasets with 2+ text columns)
- Support for multiple languages, including English, Japanese, French, Spanish, Chinese, Portuguese, etc.



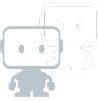
Feature Types

- ✓ Numerical
- ✓ Categorical
- ✓ Text



Domain-Specific Feature Engineering

Loan Amount	Term Length	Annual Income
\$11,200	36	\$108,000
\$10,000	60	\$65,000
\$8,000	36	\$35,000
\$16,000	36	\$110,000
\$4,000	60	\$155,000

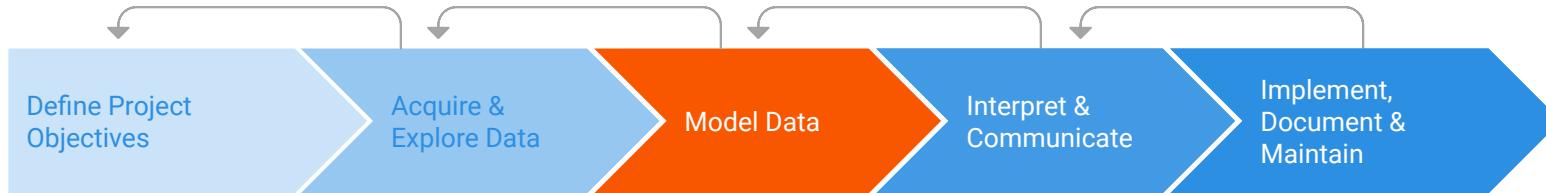


Domain-Specific Feature Engineering

Loan Amount	Term Length	Annual Income	Loan To Income Ratio
\$11,200	36	\$108,000	$(11200/36) / (108,000/12) = 0.035$
\$10,000	60	\$65,000	0.031
\$8,000	36	\$35,000	0.076
\$16,000	36	\$110,000	0.048
\$4,000	60	\$155,000	0.005



The Machine Learning Life Cycle



1. Define Project Objectives

- Specify problem
- Acquire subject matter expertise
- Define target and unit of analysis
- Prioritize modeling criteria
- Consider success criteria and risks
- Decide whether to continue

2. Acquire & Explore Data

- Find appropriate data
- Format data into single table
- Conduct exploratory data analysis
- Find and remove any target leakage
- Perform feature engineering

3. Model Data

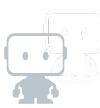
- Select features
- Build candidate models
- Validate models

4. Interpret & Communicate

- Assess model quality
- Determine important features
- Identify relationships
- Explain predictions

5. Implement, Document & Maintain

- Select a model for deployment
- Document modeling process
- Create model monitoring and maintenance plan



Selecting Features

DataRobot Data Models 52 Deployments Insights Jupyter Repository LendingClub

Project Data Feature Lists

☰ Menu Q Search Feature List All Features ▾ View Raw Data + Create Feature List 1-34 of 34

Feature Name	Index	Importance	Type	Count	Mean	Std Dev
loan_is_bad	34	Targ		16	0.36	
sub_grade	7			16	0.36	
<input checked="" type="checkbox"/> grade	6			249	62,497	
<input checked="" type="checkbox"/> annual_inc	11			249	62,497	
dti	16		Numeric	3,468	0	17.38
inq_last_6mths	19		Numeric	9	0	0.84
revol_util	25		Percentage	1,011	23	0.59
<input checked="" type="checkbox"/> purpose	12		Categorical	13	0	0.23
installment	5		Numeric	9,627	0	437
funded_amnt	4		Numeric	1,046	0	13,903

Create a new feature list with the selected features.

Feature list name: my list

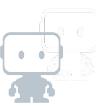
Create feature list

WORKERS
Using 3 of 20 total workers across all projects 20

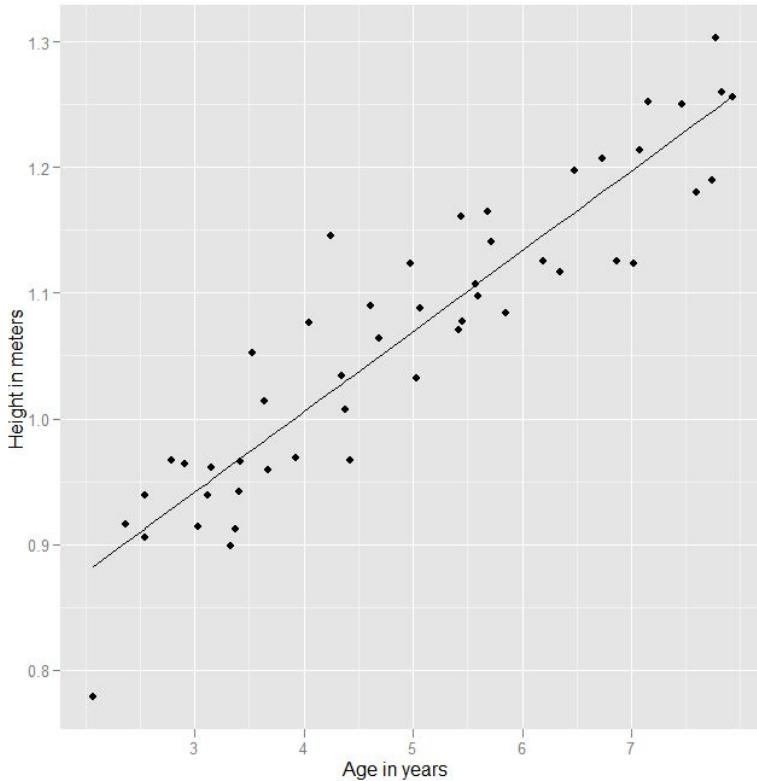
ACTIONS

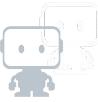
PROCESSING 3

- Gradient Boosted Greedy Trees ...
32.00% sample, CV #1 0% 0.2 GB RAM 4.2 CPUs
- RandomForest Classifier (Gini) ...
32.00% sample, CV #1 0% 0.3 GB RAM 0.3 CPUs
- Balanced ExtraTrees Classifier (...)
32.00% sample, CV #1 0% 0.3 GB RAM 0.5 CPUs

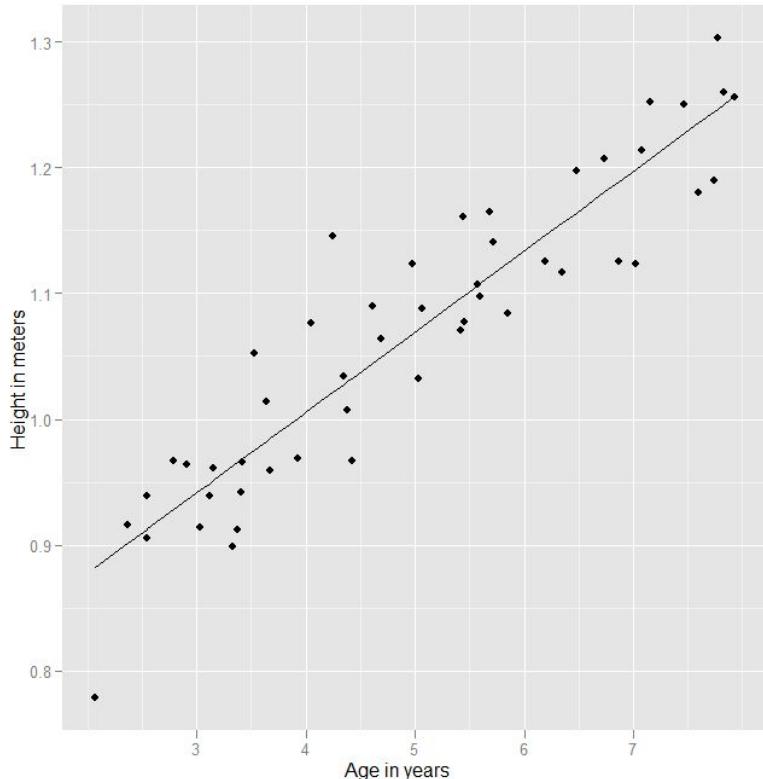


Linear Regression

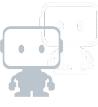




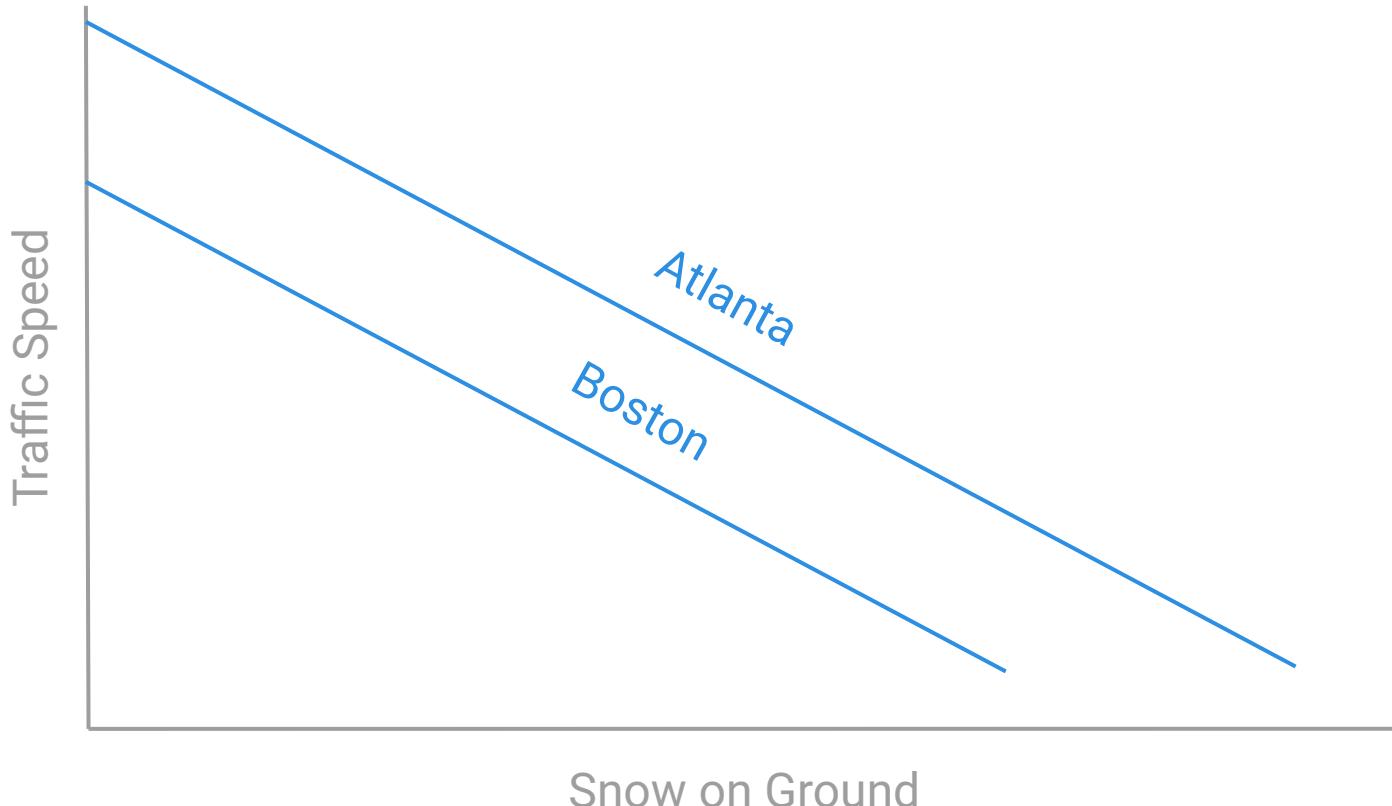
Linear Regression

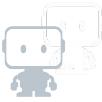


$$= \underline{\hspace{2cm}} + \underline{\hspace{2cm}} * \text{age}$$

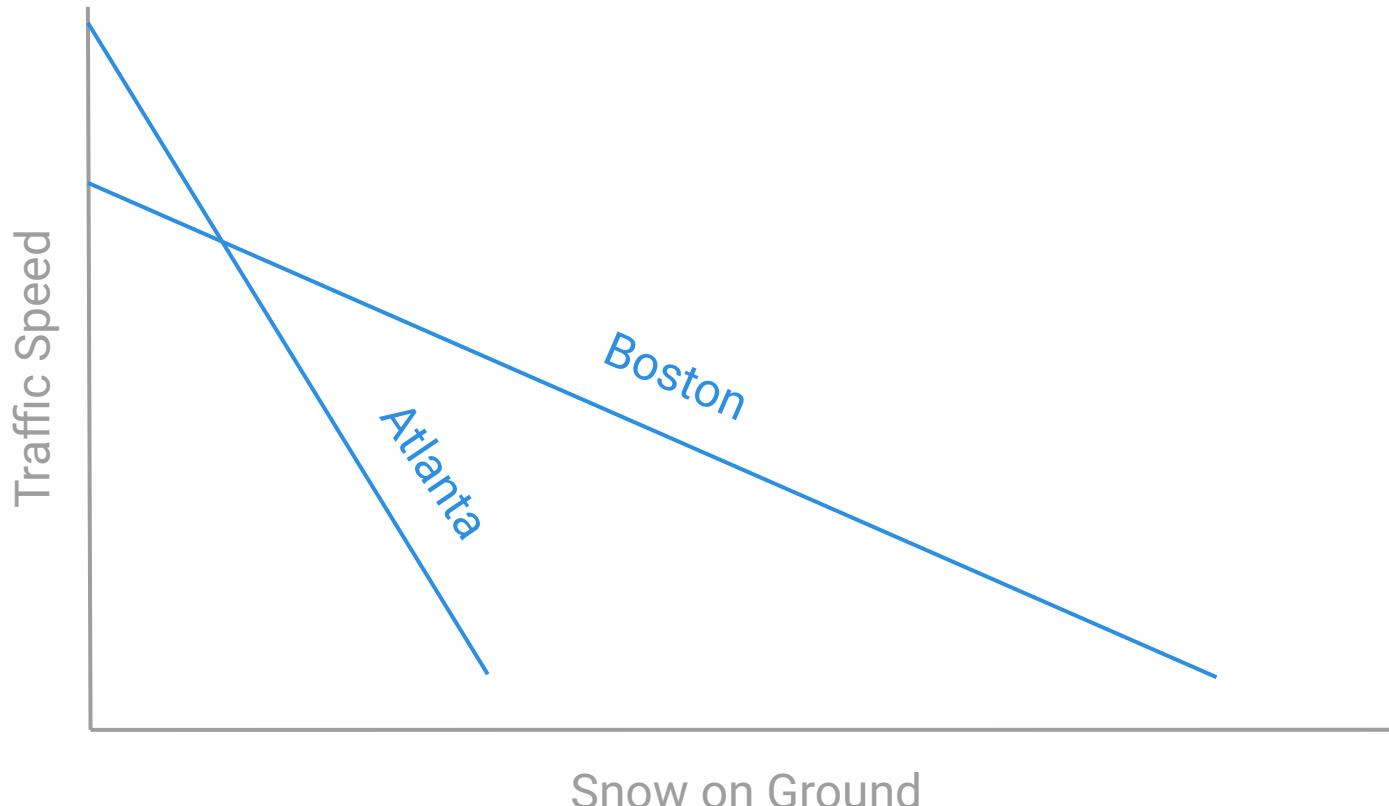


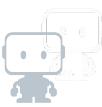
What Linear Models Miss



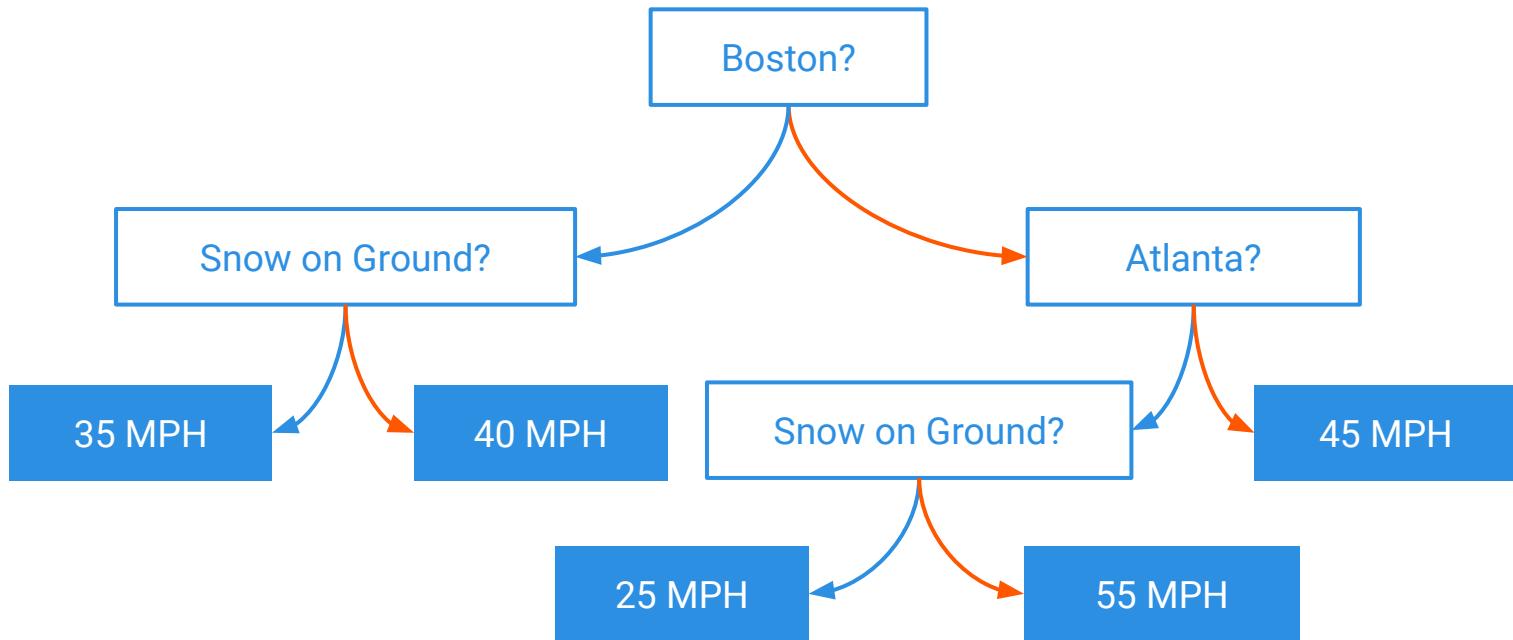


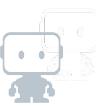
What Linear Models Miss



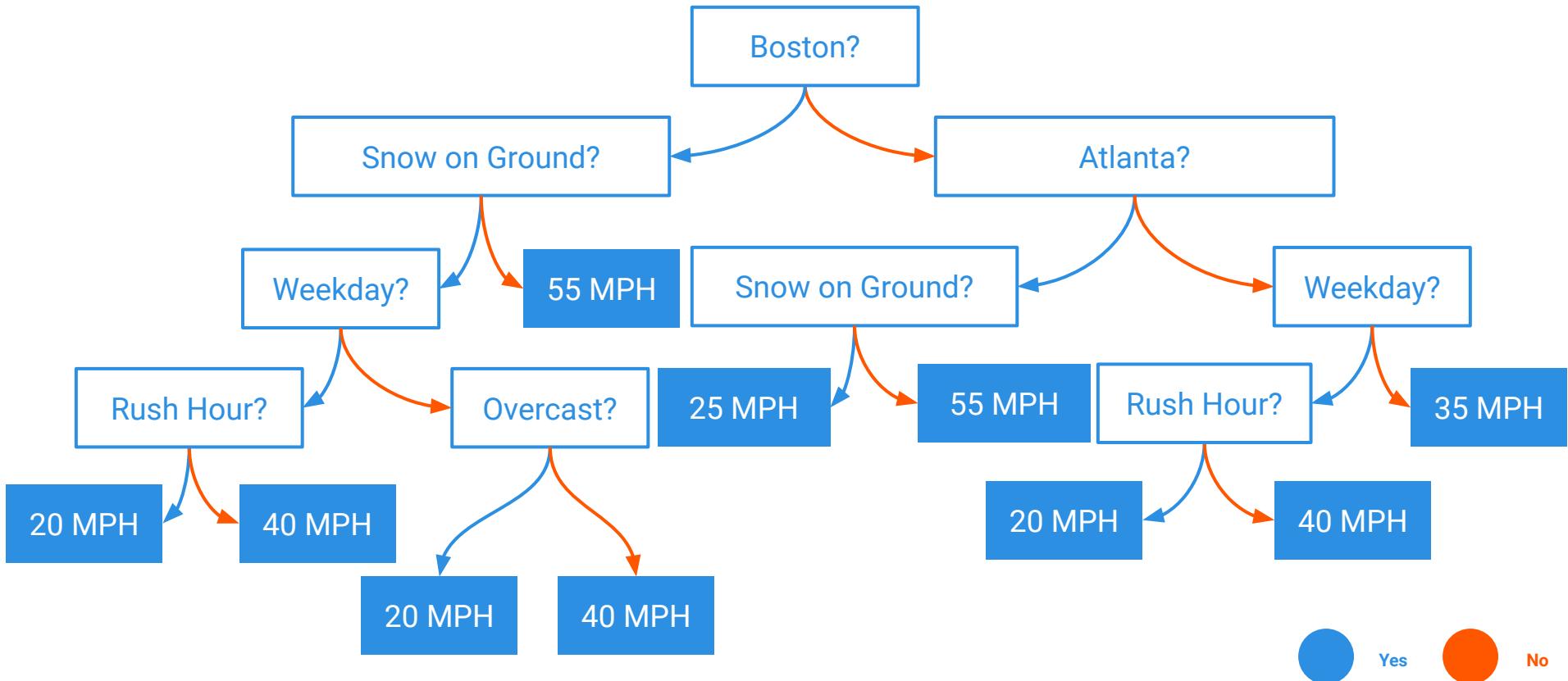


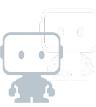
Decision Trees



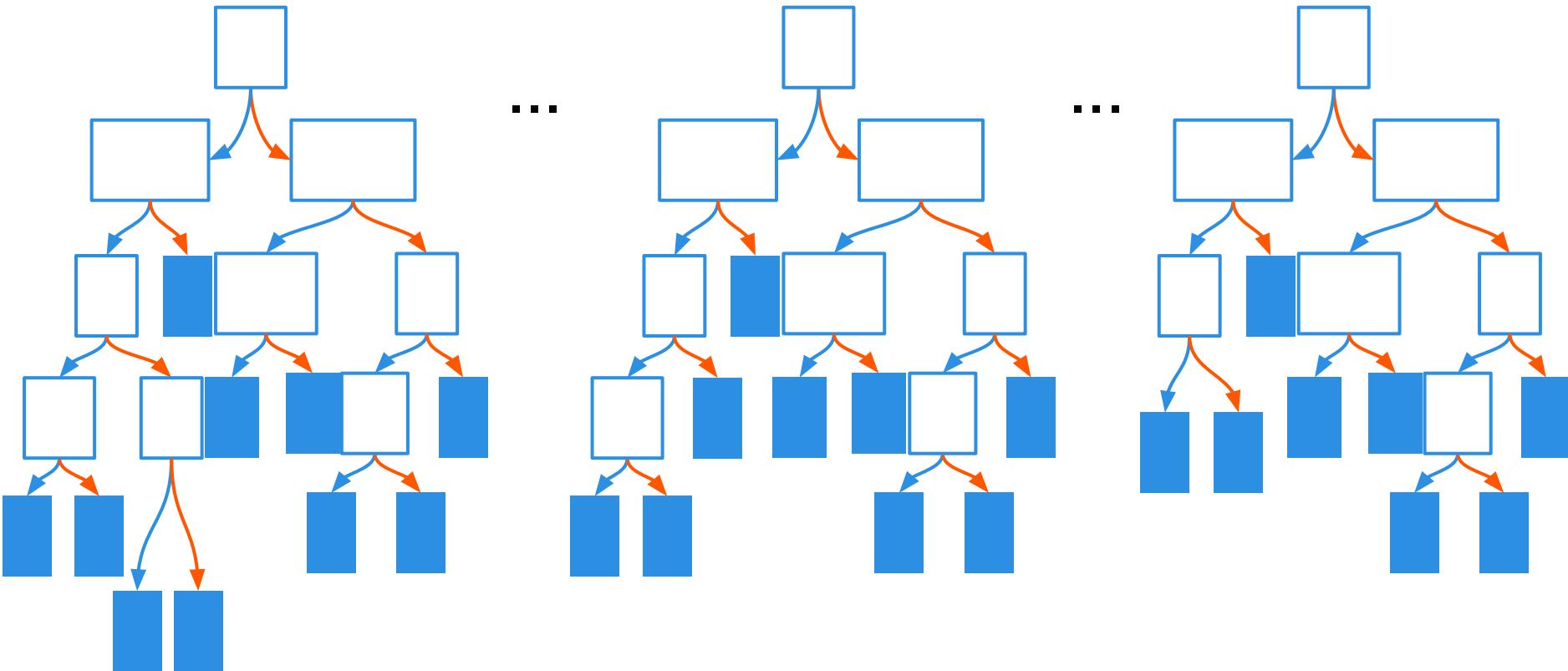


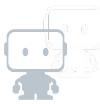
Decision Trees



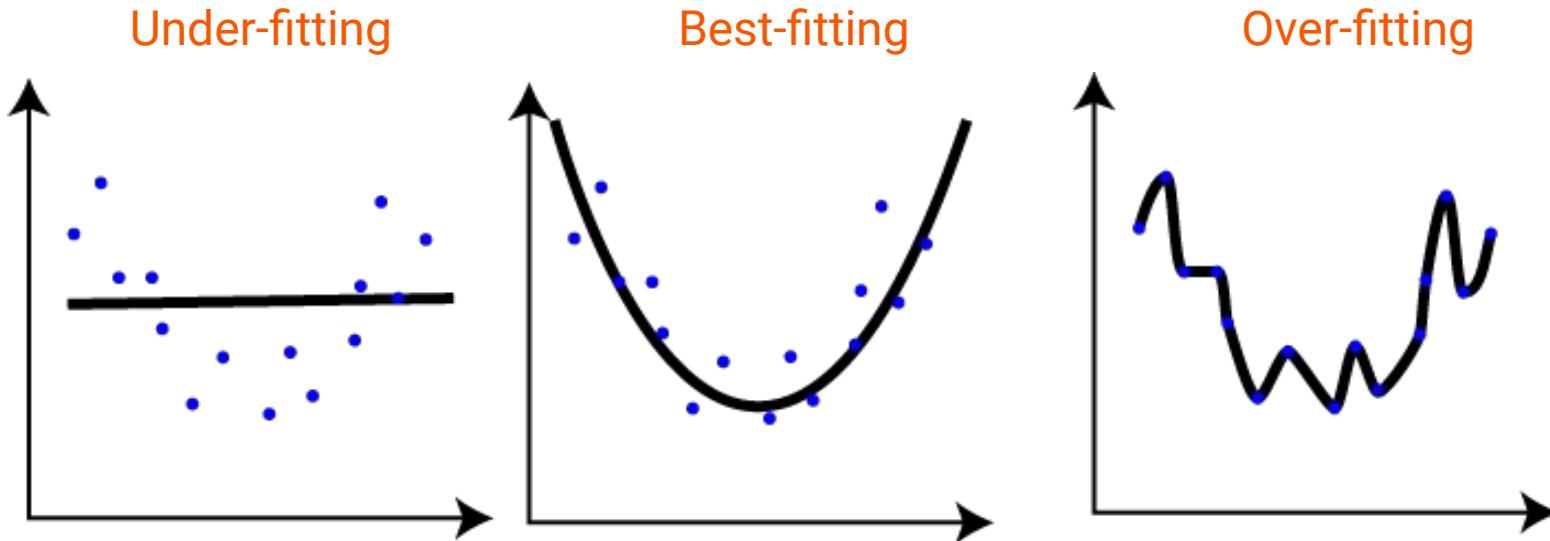


Tree-based Ensembles



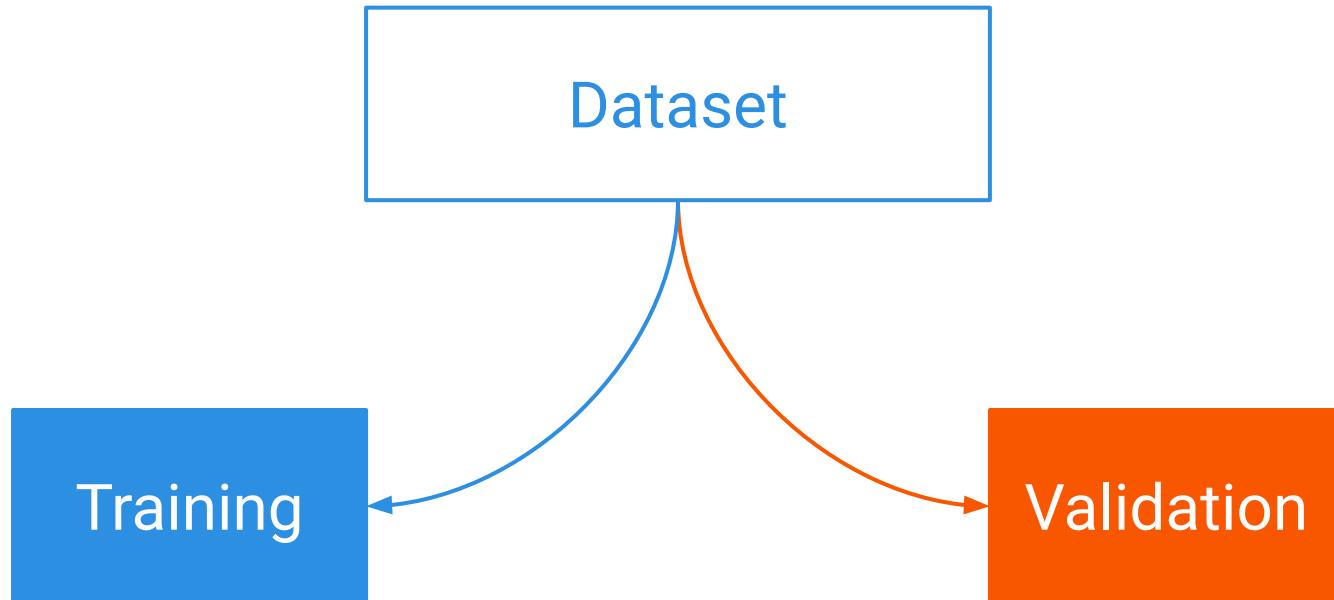
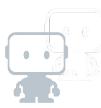


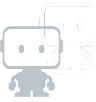
Introduction to Underfitting and Overfitting



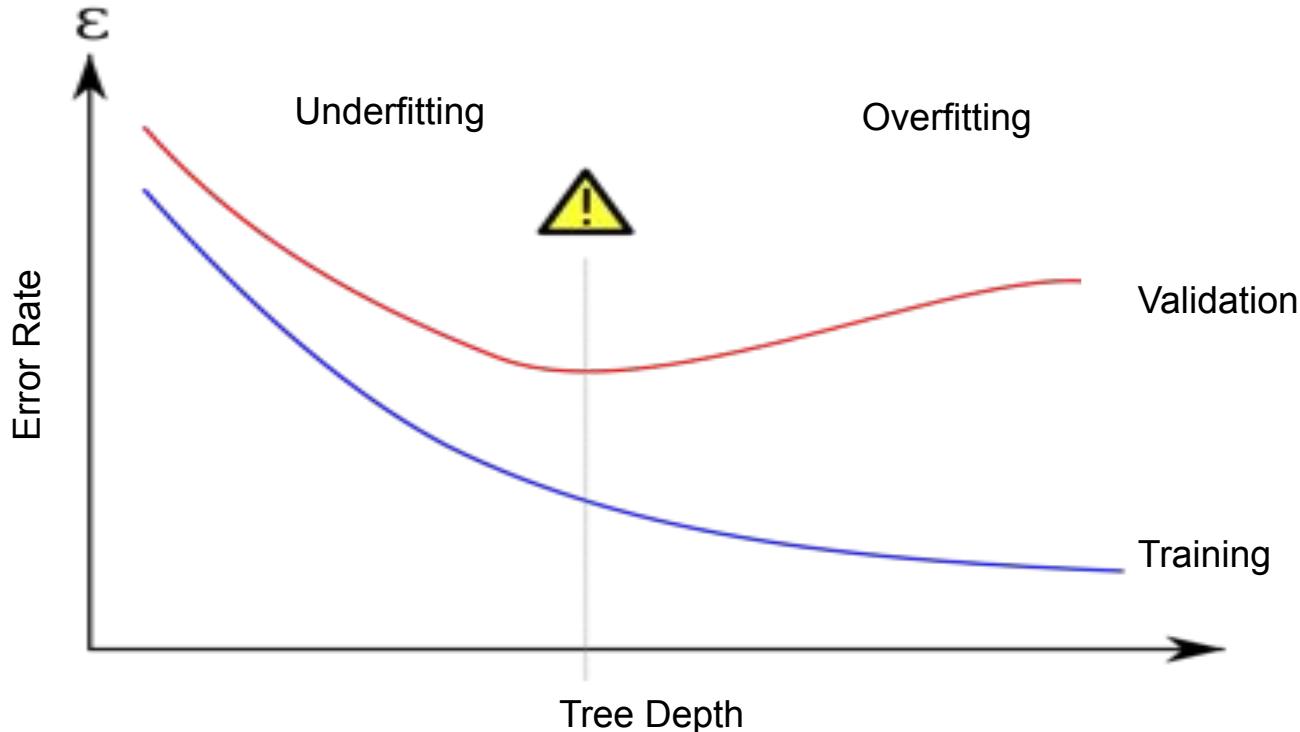
But how do we find the best fit?

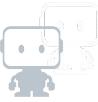
Data Partitioning



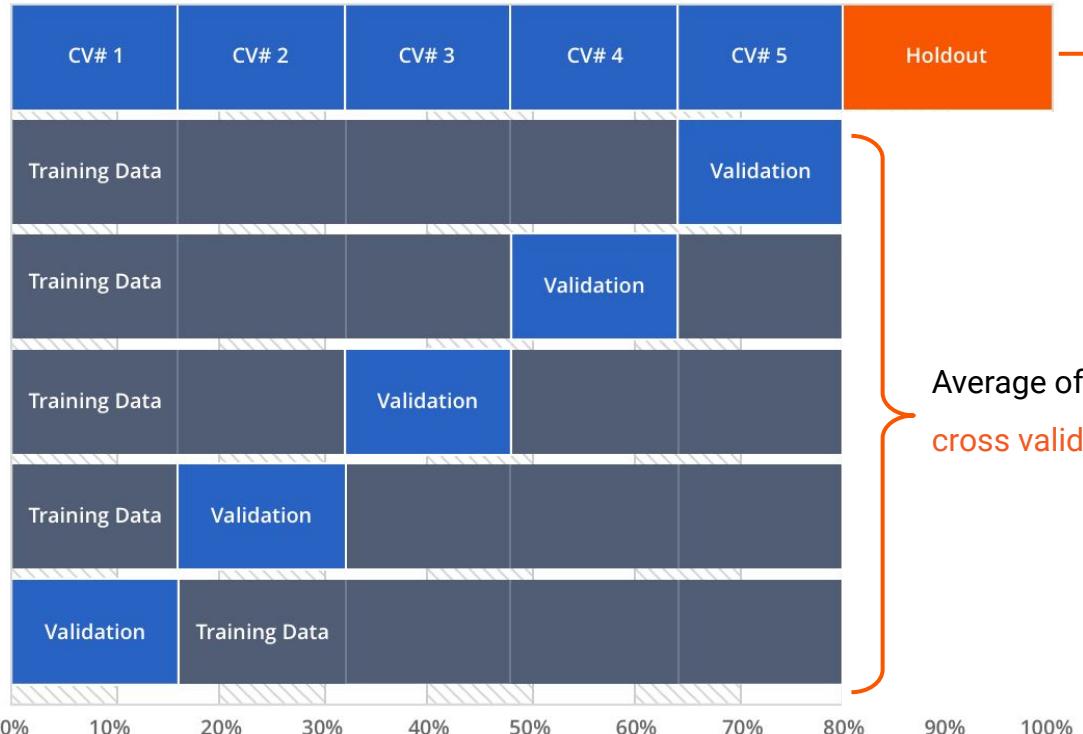


Finding the Best Fit





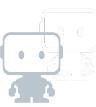
Cross Validation in DataRobot



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**.

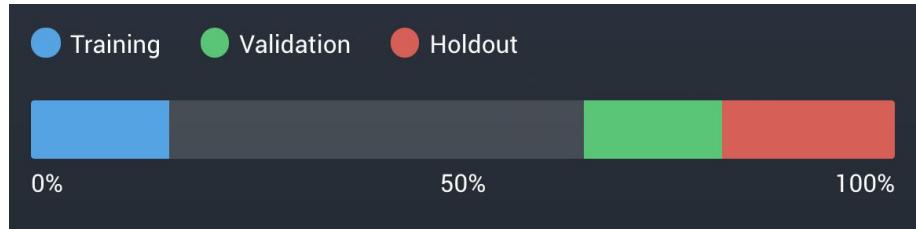
Average of these 5 validation scores is the **cross validation score**



Modeling Modes

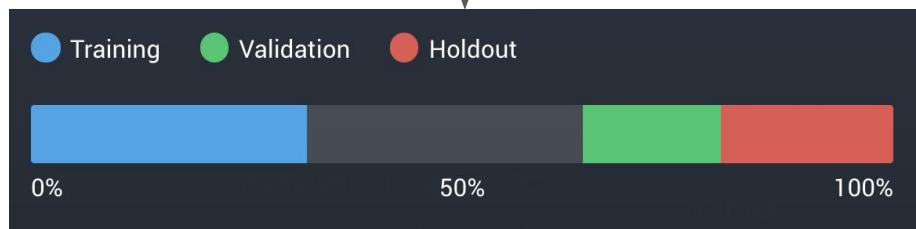
Autopilot

- Default setting
- Multiple elimination rounds



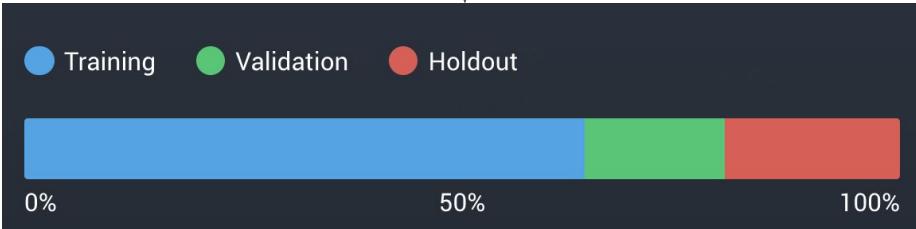
Quick

- Fewer blueprints
- One elimination round



Manual Mode

- User chooses blueprints
- Can customize elimination

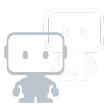


Benefits of Automation



Iteration is easy and safe

Leaderboard



DataRobot Data Models 67 Deployments Insights Jupyter Repository LendingClub

Select some or all feature sets Speed vs Accuracy Model Comparison Prediction Apps

☰ Menu Search Add New Model Filter Models Metric LogLoss ▾

Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
Nystroem Kernel SVM Classifier One-Hot Encoding Missing Values Imputed Standardize Smooth Ridit Transform Converter for Text Mining Auto-Tuned Word N-Gram Text Modeler using token occurrences Transform on the link function scale Nystroem Kernel SVM Classifier M73 BP62	DR Reduced Features M60 80.0 %	0.4029 *	0.4080 *	0.4077
AVG Blender M75 M60+71+61	Multiple Feature Lists 64.0 %	0.4028	Run	
ENET Blender M78 M60+71+61	Multiple Feature Lists 64.0 %	0.4028	Run	
Advanced AVG Blender M76 M66+64+60+71+65+...	Multiple Feature Lists 64.0 %	0.4029	Run	
Nystroem Kernel SVM Classifier One-Hot Encoding Missing Values Imputed Standardize Smooth Ridit Transform Converter for Text Mining Auto-Tuned Word N-Gram Text Modeler using token occurrences Transform on the link function scale Nystroem Kernel SVM Classifier M71 BP62	DR Reduced Features M60 64.0 %	0.4029	Run	

WORKERS
Using 0 of 20 total workers across all projects

STATUS
 Autopilot has finished

ACTIONS
 Rerun Autopilot
 Unlock Holdout for all models

What about Fast Iron?



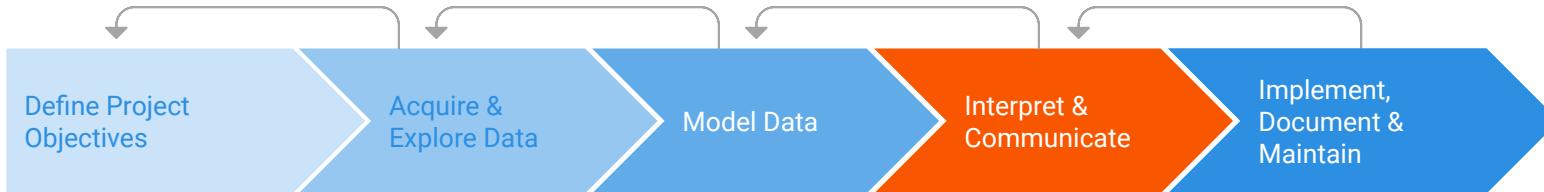
Go back to the fast_iron_100k project. Run any model(s) from the repository and explore the dataset further using DataRobot.

Discussion Questions

1. Which models did you choose and why? Which performed the best?
2. Are there any other data quality issues in the dataset?
3. What other data sources might be useful to predict the sales price?



The Machine Learning Life Cycle



1. Define Project Objectives

- Specify problem
- Acquire subject matter expertise
- Define target and unit of analysis
- Prioritize modeling criteria
- Consider success criteria and risks
- Decide whether to continue

2. Acquire & Explore Data

- Find appropriate data
- Format data into single table
- Conduct exploratory data analysis
- Find and remove any target leakage
- Perform feature engineering

3. Model Data

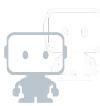
- Select features
- Build candidate models
- Validate models

4. Interpret & Communicate

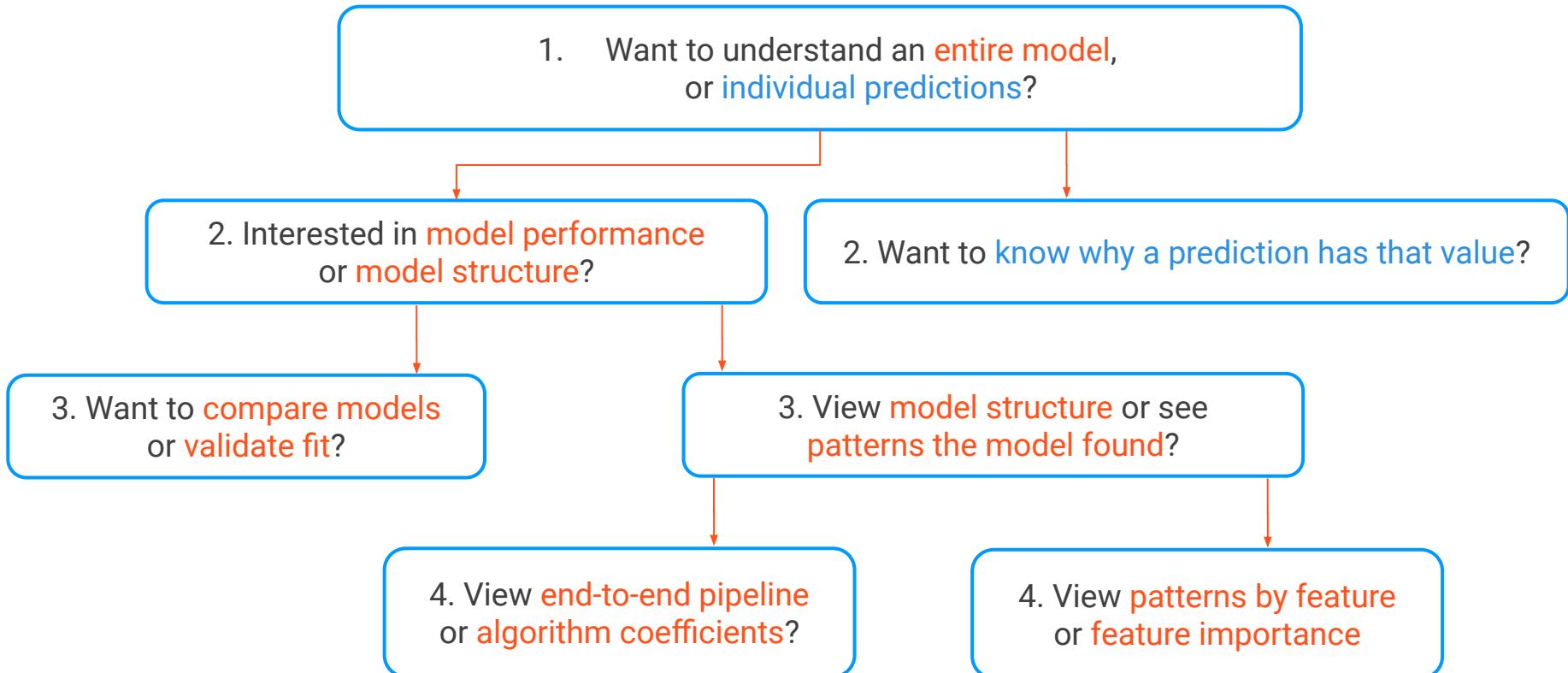
- Assess model quality
- Determine important features
- Identify relationships
- Explain predictions

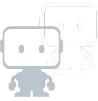
5. Implement, Document & Maintain

- Select a model for deployment
- Document modeling process
- Create model monitoring and maintenance plan



Interpreting a Model





Confusion Matrix

Strengths

- Straightforward Interpretation

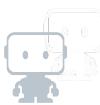
Weaknesses

- Imprecise

Use Cases

- Quick view of model
- Easy to equate to real business value

		Predicted		
		-	+	
Actual	-	4603 (TN)	2146 (FP)	6749
	+	488 (FN)	763 (TP)	1251
		5091	2909	8000



Calculating Payoffs from a Confusion Matrix

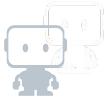
Payoff Matrix

\$3,000 (TN)	\$0 (FP)
-\$8,000 (FN)	\$0 (TP)

Payoff = \$9,905,000

$$= (\$3K \times 4603) + (\$0 \times 2146) + \\ (-\$8K \times 488) + (\$0 \times 763)$$

		Predicted		
		-	+	
Actual	-	4603 (TN)	2146 (FP)	6749
	+	488 (FN)	763 (TP)	1251
		5091	2909	8000



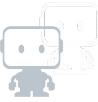
What a Confusion Matrix Doesn't Show

Log Loss = 0.31

Log Loss = 1.11

Actual Value	Model 1	Model 2	Are Models Correct?
True	0.9	0.6	Yes
False	0.51	0.9	No
False	0.1	0.4	Yes

Assume classification threshold = 0.5



Lift Chart

Strengths

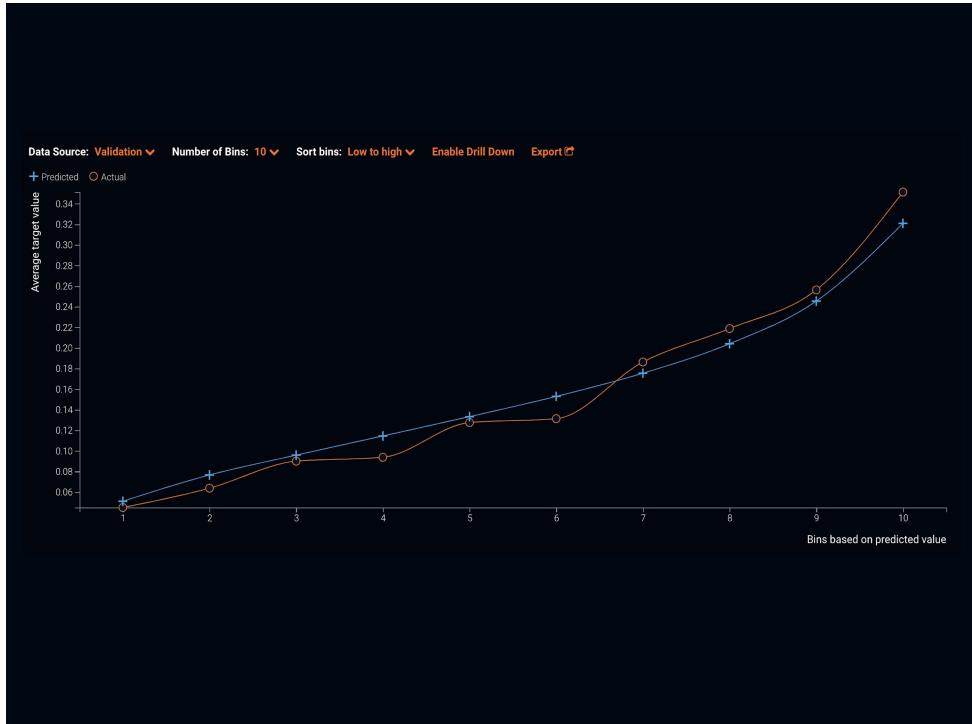
- Straightforward interpretation
- Visualizes the model's error and its ability to discriminate

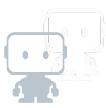
Weakness

- Does not display the variation within the bins

Use Cases

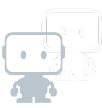
- Many





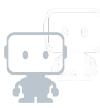
End of Day 1

Please complete Survey - *Day One* in Litmos



Lift Chart Data

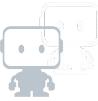
Predicted	Actual
4	7
5	3
7	9
12	7
13	8
15	10
17	12
20	12
21	30
23	16



Lift Chart Data

Predicted	Actual
4	7
5	3
7	9
12	7
13	8
15	10
17	12
20	12
21	30
23	16

Predicted	Actual
4.5	5

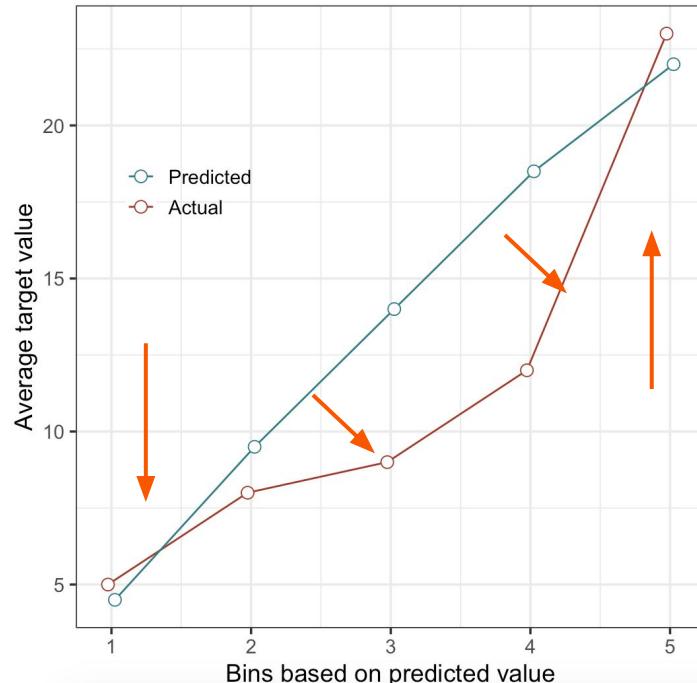


Lift Chart Data

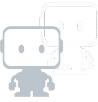
Predicted	Actual
4	7
5	3
7	9
12	7
13	8
15	10
17	12
20	12
21	30
23	16

Predicted	Actual
4.5	5
9.5	8
14	9
18.5	12
22	23

1. How close are the lines?

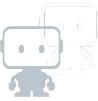


2. How steep is the actual line?



Feature Fit





Prediction Distribution

Strengths

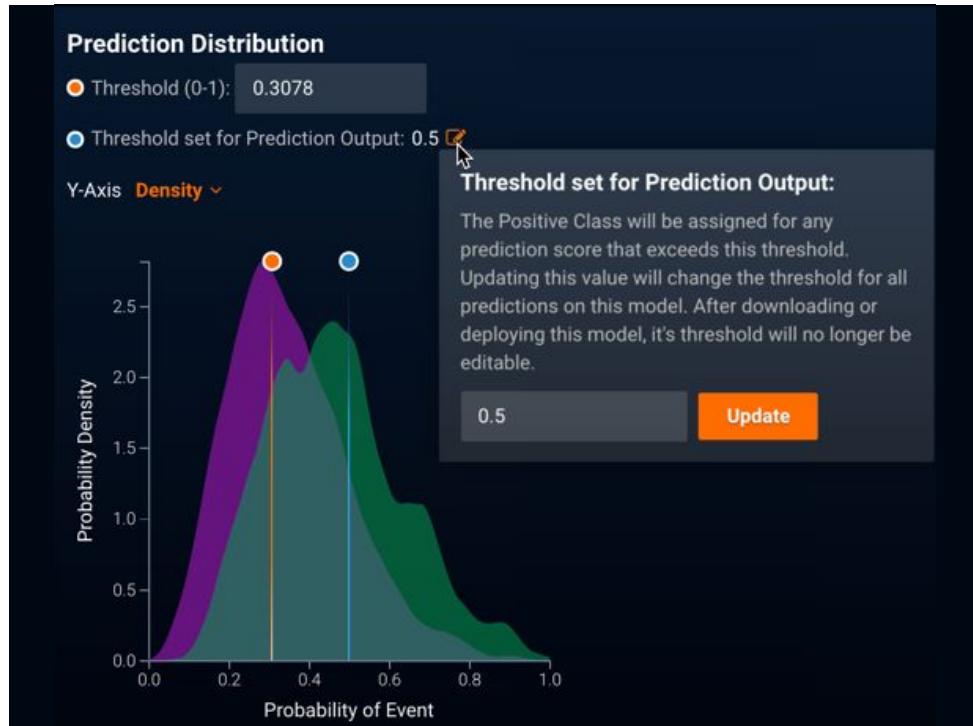
- Visualizes a model's performance by showing how well it separates the two groups in the target

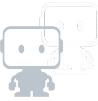
Weakness

- Less "executive friendly" than a simple accuracy score

Use Case

- Set a threshold for model output
- Model comparison





ROC Curve

Strengths

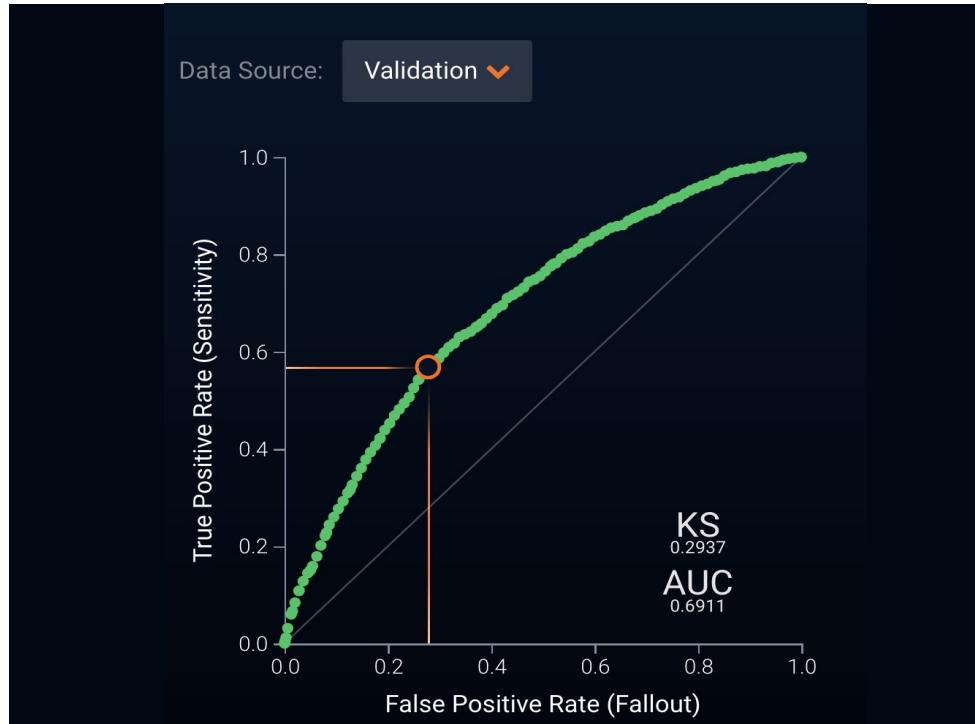
- Precise measure of model's ability to discriminate risk
- Interpretable metrics

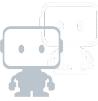
Weakness

- Not “executive friendly”
- Only captures rank ordering of probabilities

Use Case

- Model comparison





Cumulative Gain and Lift

Strengths

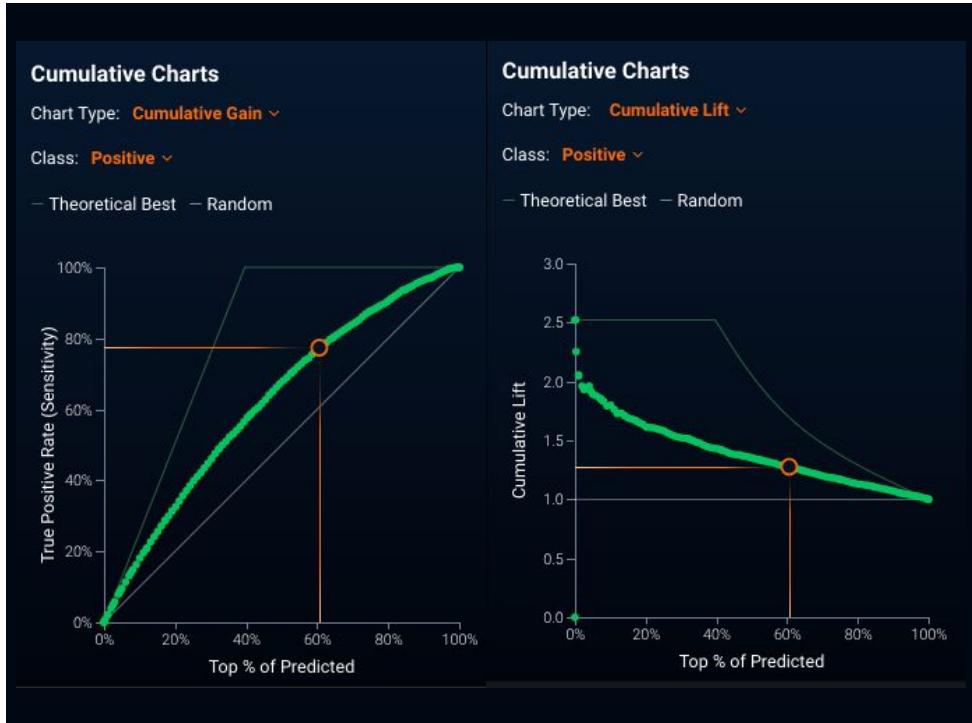
- Widely used in marketing
- Shows how many predictions of a model are expected to be needed to achieve a certain true positive (or true negative) rate

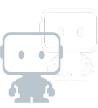
Weakness

- Doesn't show rate of wrong predictions

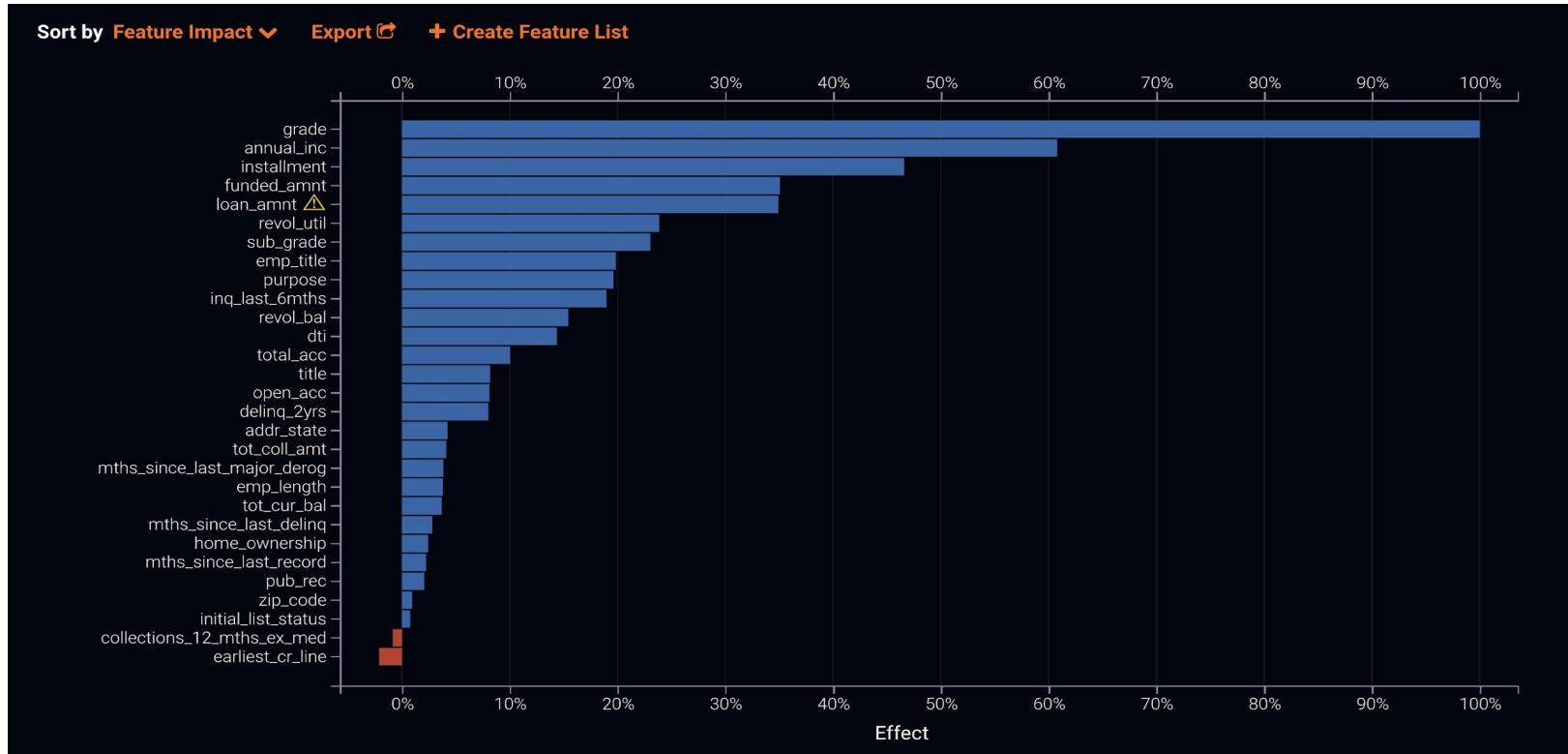
Use Case

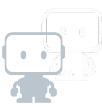
- Model comparison



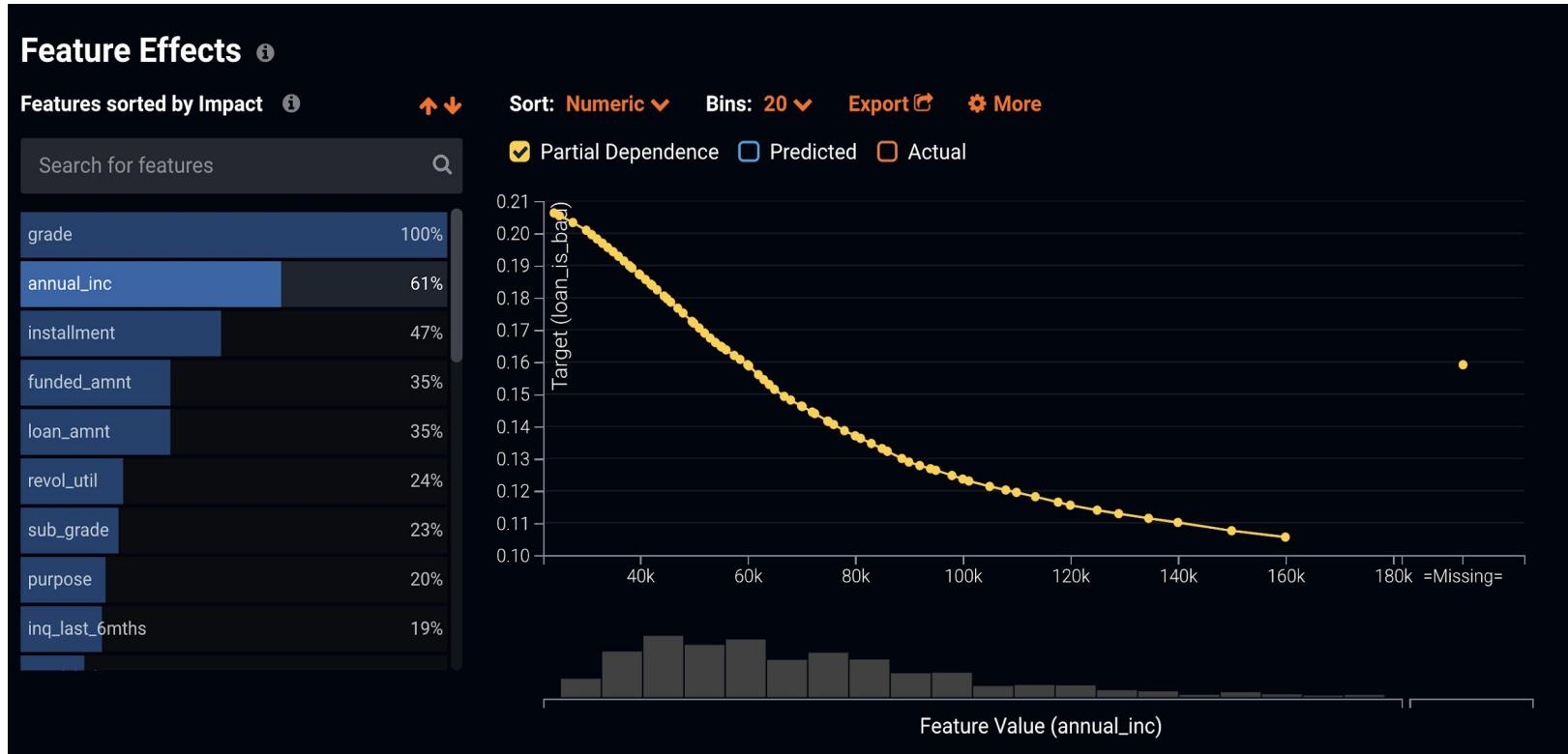


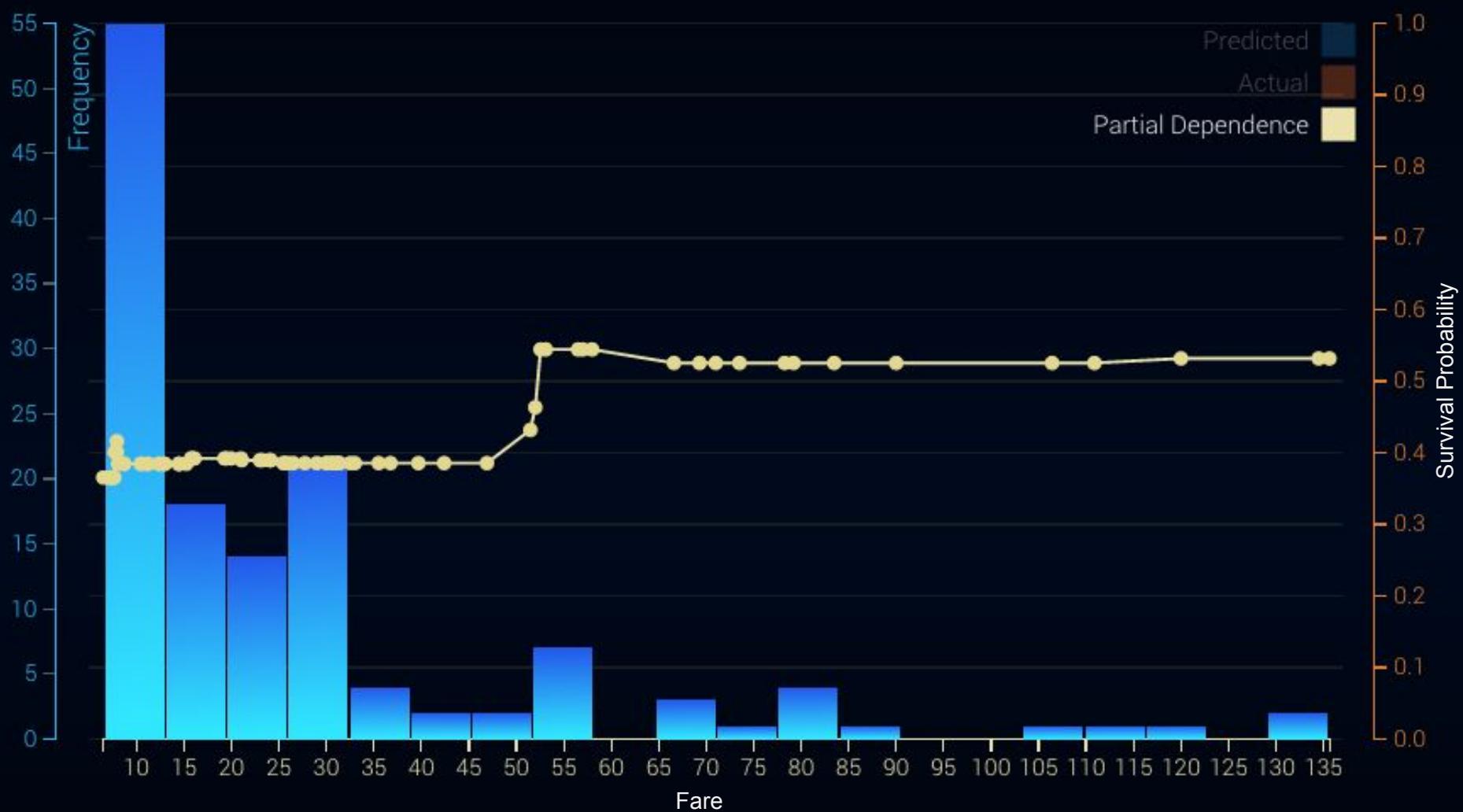
Feature Impact

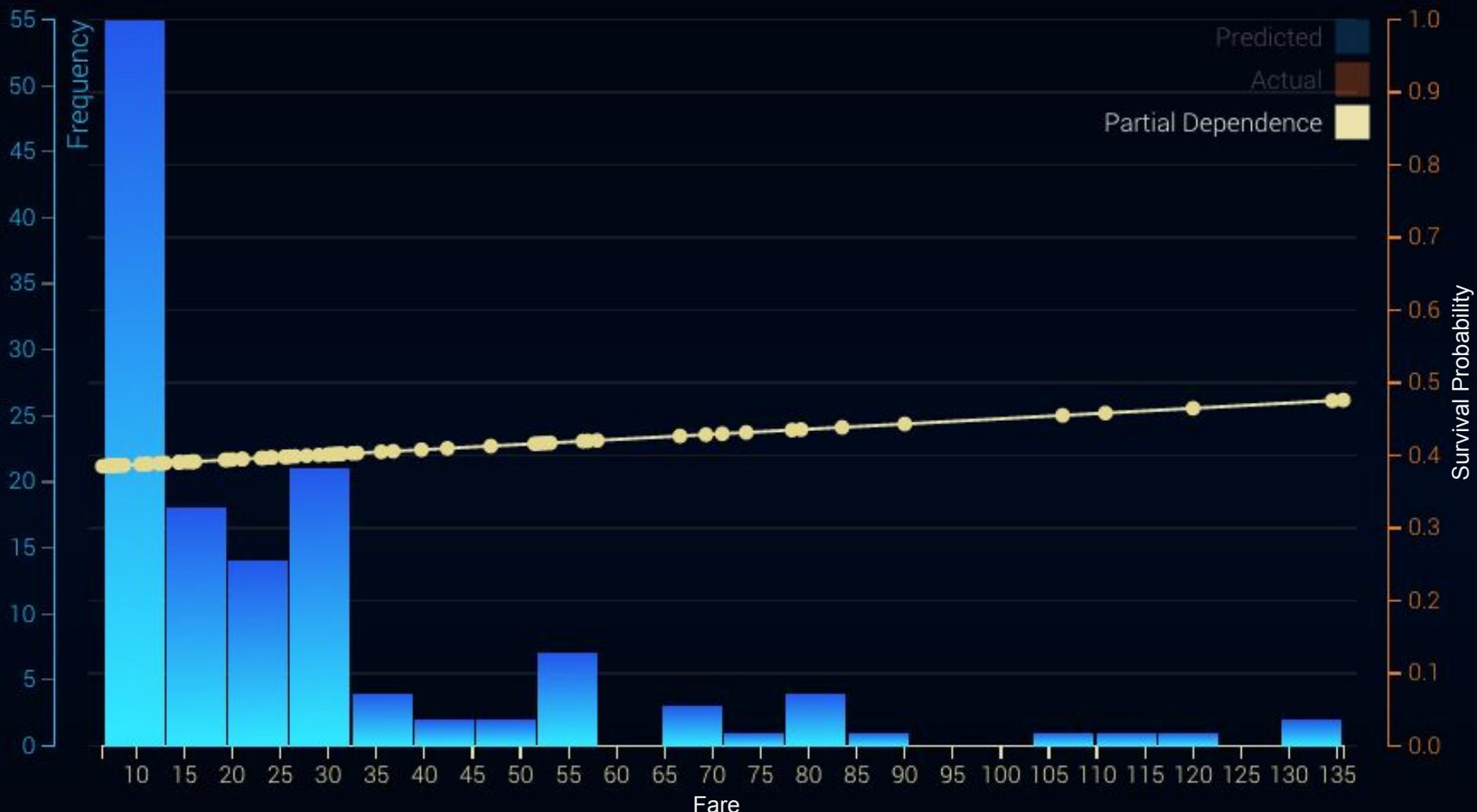


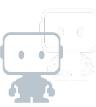


Feature Effects



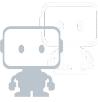




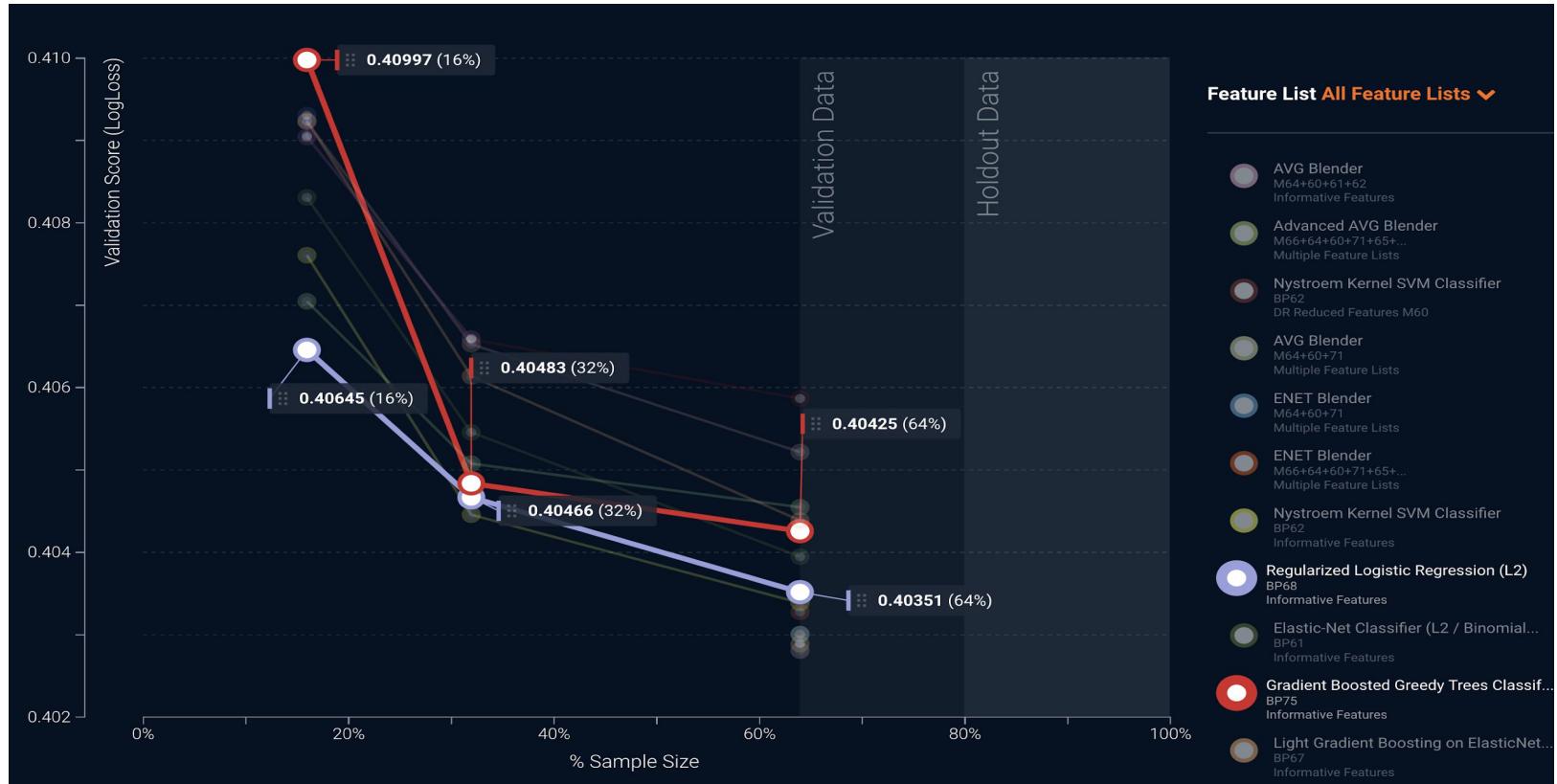


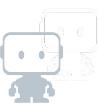
Prediction Explanations



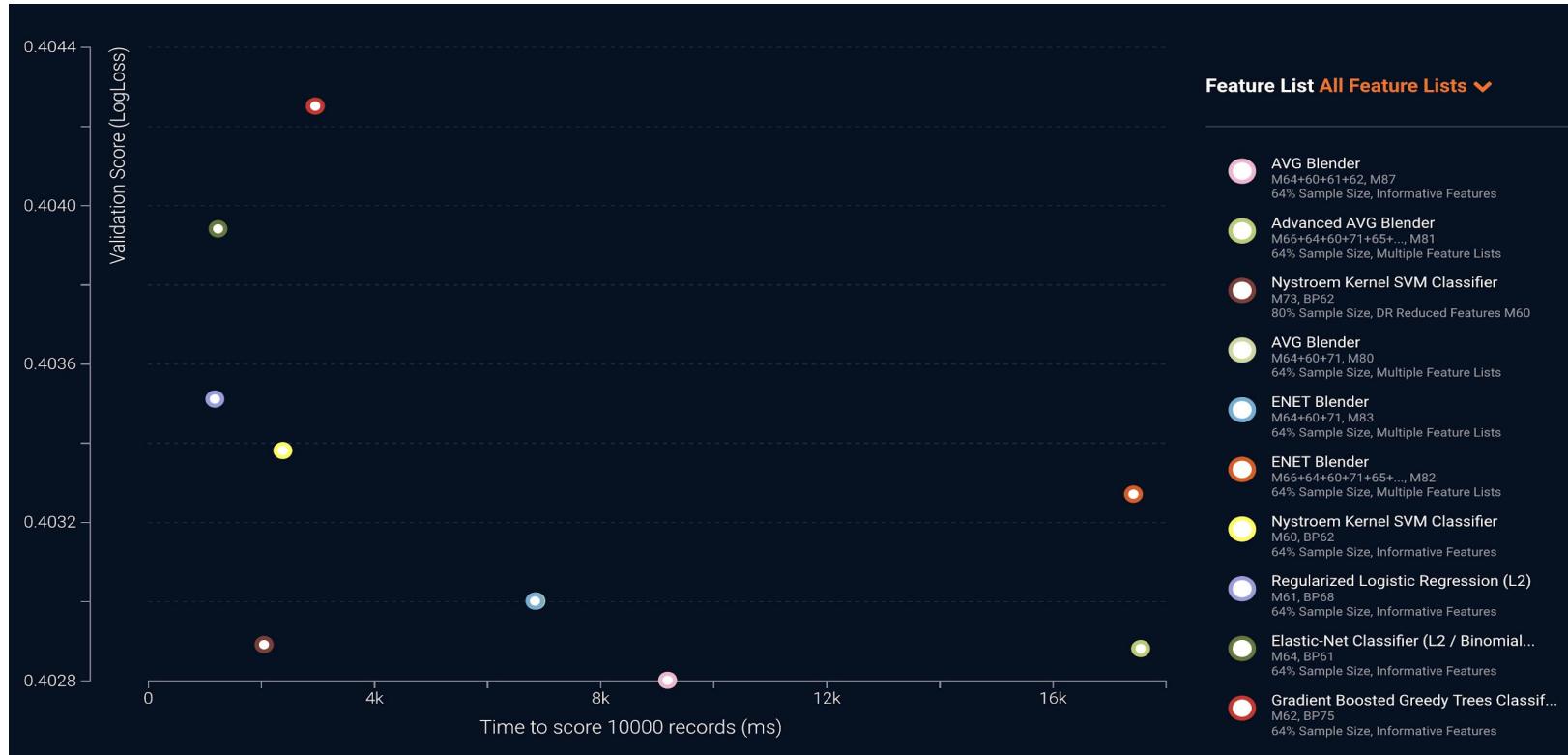


Learning Curves

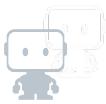




Speed vs. Accuracy



More Insights



Insights

Tree-based Variable Importance

For tree/forest models, illustrates the relevancy of different variables on the prediction.

Hotspots

Provides a simple, graphical interpretation of rules describing high predictive performance.



Variable Effects

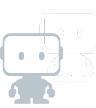
For some linear models, illustrates the strength and direction of the relationship between each feature in your model and the model prediction.

Word Cloud

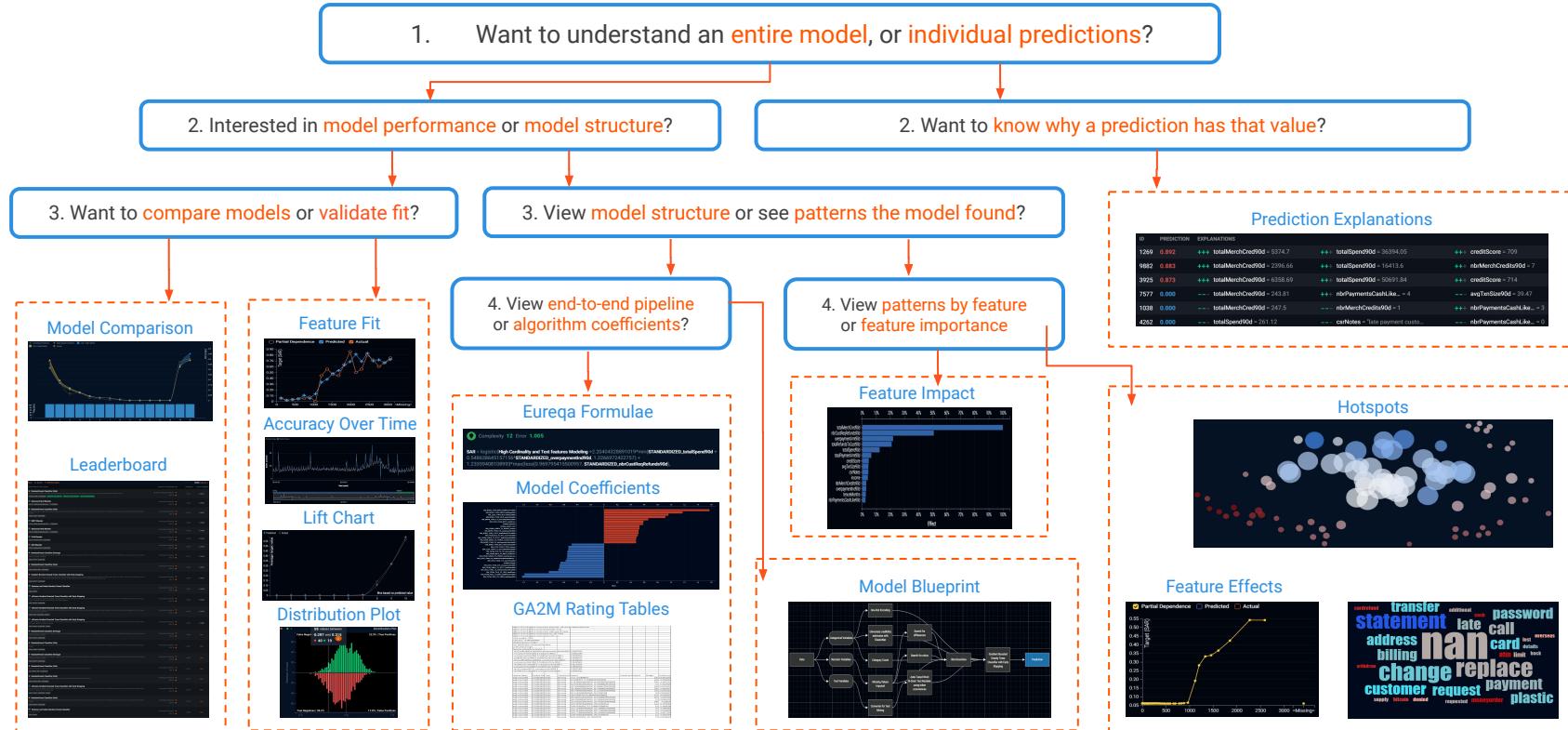


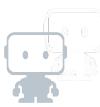
Text Mining

Displays the most relevant words and short phrases in variables detected as text, indicating the strength of correlation with the target.



How to Understand a DataRobot Model



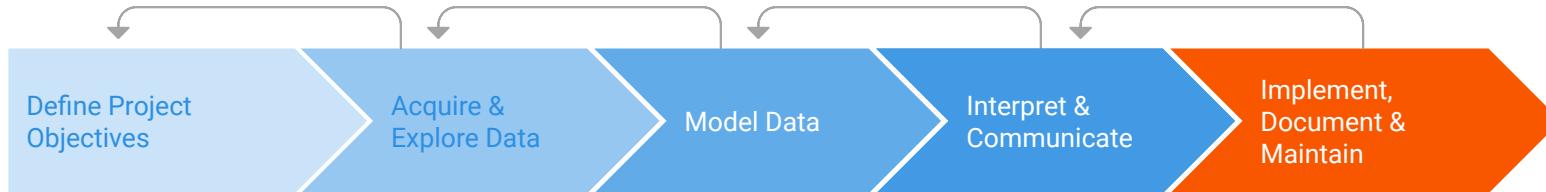


Model Tags

Leaderboard Learning Curves Speed vs Accuracy Model Comparison Prediction Apps					Metric LogLoss ▾
☰ Menu	🔍 Search	+ Add New Model	✖ Filter Models		
Model Name & Description	Feature List & Sample Size		Validation	Cross Validation	Holdout
Nystroem Kernel SVM Classifier One-Hot Encoding Missing Values Imputed Standardize Smooth Ridit Transform Converter for Text Mining Auto-Tuned Word N-Gram Text Modeler using token occurrences Transform on the link function scale Nystroem Kernel SVM Classifier	DR Reduced Features M60	80.0 %	0.4029 *	0.4080 *	0.4077
M73 BP62	🏆 RECOMMENDED FOR DEPLOYMENT				
Advanced AVG Blender M81 M66+64+60+71+65+...	Multiple Feature Lists	64.0 %	0.4029	Run	🔒
Nystroem Kernel SVM Classifier One-Hot Encoding Missing Values Imputed Standardize Smooth Ridit Transform Converter for Text Mining Auto-Tuned Word N-Gram Text Modeler using token occurrences Transform on the link function scale Nystroem Kernel SVM Classifier	DR Reduced Features M60	64.0 %	0.4029	Run	🔒
M71 BP62	🏆 FAST & ACCURATE				



The Machine Learning Life Cycle



1. Define Project Objectives

- Specify problem
- Acquire subject matter expertise
- Define target and unit of analysis
- Prioritize modeling criteria
- Consider success criteria and risks
- Decide whether to continue

2. Acquire & Explore Data

- Find appropriate data
- Format data into single table
- Conduct exploratory data analysis
- Find and remove any target leakage
- Perform feature engineering

3. Model Data

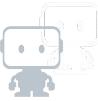
- Select features
- Build candidate models
- Validate models

4. Interpret & Communicate

- Assess model quality
- Determine important features
- Identify relationships
- Explain predictions

5. Implement, Document & Maintain

- Select a model for deployment
- Document modeling process
- Create model monitoring and maintenance plan



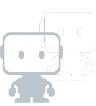
Deployment Notes

1. Choose model

- Which model satisfies my modeling criteria?
- Optionally use 100% of data to train model

2. Choose deployment options

- GUI batch predictions
- Prediction API
- DataRobot Prime (python, java)
- Scoring Code JAR (java)



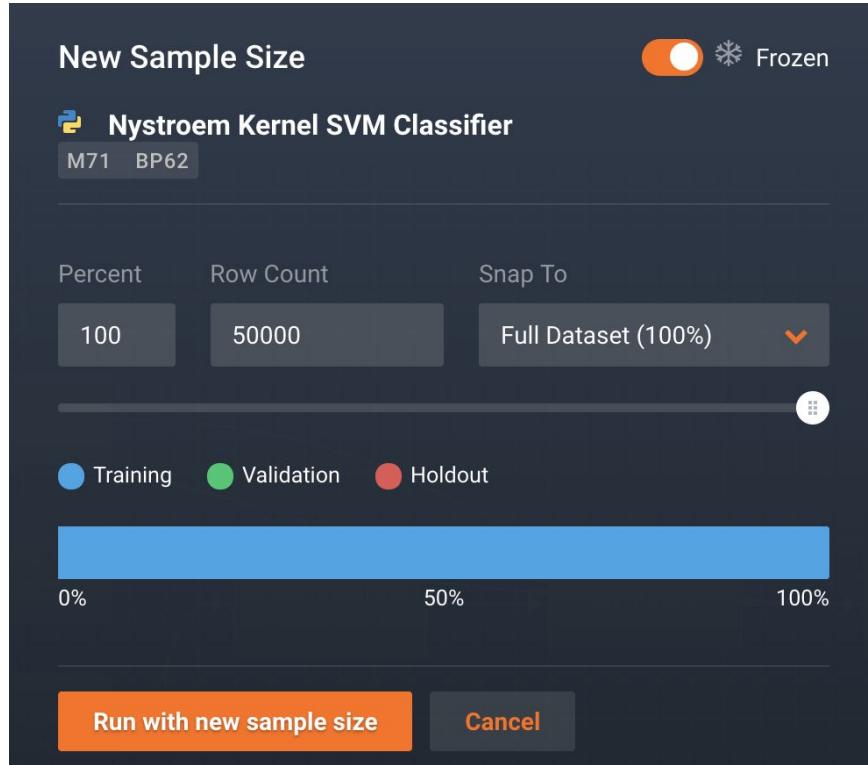
Training on 100%

Pros

- Using 100% of the data gives the model more examples to learn from and may significantly improve predictive performance

Cons

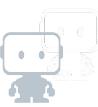
- Can't validate the model directly
- When monitoring the model after deployment, new predictions have no direct comparison





GUI Batch Predictions

- Load `2015_test.csv.zip` into the LendingClub project.
- Add optional features `id`, `member id`, `grade`, `annual_inc`, and `purpose`.
- Set threshold for prediction output to be 0.2 (under ROC curve tab)
- Make predictions with the chosen model.
- Download the csv and inspect the results.



Prediction API

DataRobot Data Models 68 Deployments Insights Jupyter Repository early_2012_2013_train.csv.zip

loan_is_bad Predictions DataRobot Prediction Server | early_2012_2013_train.csv.zip

Service	Drift	Accuracy	Activity	Avg. Requests/Day	Last Prediction
✓	?	?	Dec 3 now	0	-

Overview Service Health Data Drift Accuracy Integrations Settings

Integrations

Connect this deployment's predictions with downstream applications.

Language: Python

Copy to clipboard

```
# Usage: python datarobot-predict.py <input-file.csv>
#
# This example uses the requests library which you can install with:
# pip install requests
# We highly recommend that you update SSL certificates with:
# pip install -U urllib3[secure] certifi
import requests
import sys

API_TOKEN =
USERNAME = 'taylor.larkin@datarobot.com'

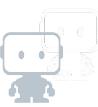
DEPLOYMENT_ID = '5be5b9ba79477102a0cee2a7'

# Set HTTP headers
# Note: The charset should match the contents of the file.
headers = {'Content-Type': 'text/plain; charset=UTF-8', 'datarobot-key': 'cc0e0c01-8463-3e63-4794-5bea42900997'}

data = open(sys.argv[1], 'rb').read()

# Make predictions on your data
# The URL has the following format:
# https://datarobot-predictions.orm.datarobot.com/predApi/v1.0/deployments/<DEPLOYMENT_ID>/predictions
# See docs for details:
# app.datarobot.com/docs/users-guide/deploy/api/new-prediction-api.html
predictions_response = requests.post('https://datarobot-predictions.orm.datarobot.com/predApi/v1.0/deployments/%s/predictions' % (DEPLOYMENT_ID),
                                    auth=(USERNAME, API_TOKEN), data=data, headers=headers)

predictions_response.raise_for_status()
print(predictions_response.json())
```



Model Documentation

DataRobot Data Models 69 Deployments Insights Jupyter Repository

early_2012_2013_train.csv.zip

Leaderboard Learning Curves Speed vs Accuracy Model Comparison Prediction Apps

☰ Menu Q Search + Add New Model Filter Models Metric LogLoss

Model Name & Description Feature List & Sample Size Validation Cross Validation Holdout

Nystroem Kernel SVM Classifier
One-Hot Encoding | Missing Values Imputed | Standardize | Smooth Ridit Transform | Converter for Text Mining |
Auto-Tuned Word N-Gram Text Modeler using token occurrences | Transform on the link function scale |
Nystroem Kernel SVM Classifier

M109 BP62

Evaluate Understand Describe Predict Compliance

Compliance Documentation

Model Compliance Documentation

Nystroem Kernel SVM Classifier M109 BP62

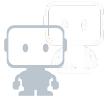
Model development overview and intended business use:
Provide a high-level summary of this model's business purpose, including an overview of the business problem that it is solving. DataRobot will populate the "Model Development Purpose and Intended Use" section of the model compliance documentation with this description.

Update Cancel

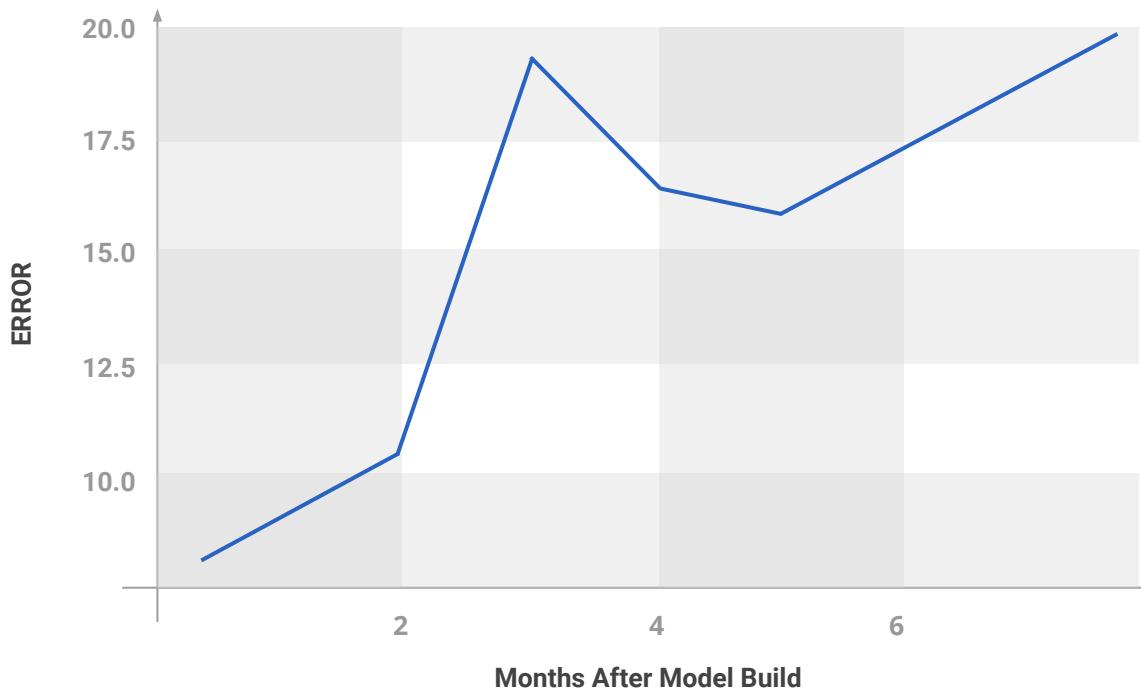
WORKERS
Using 0 of 20 total workers across all projects

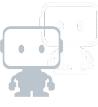
STATUS
Autopilot has finished

ACTIONS
Rerun Autopilot Holdout is Unlocked



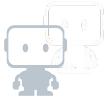
Deteriorating Model Performance



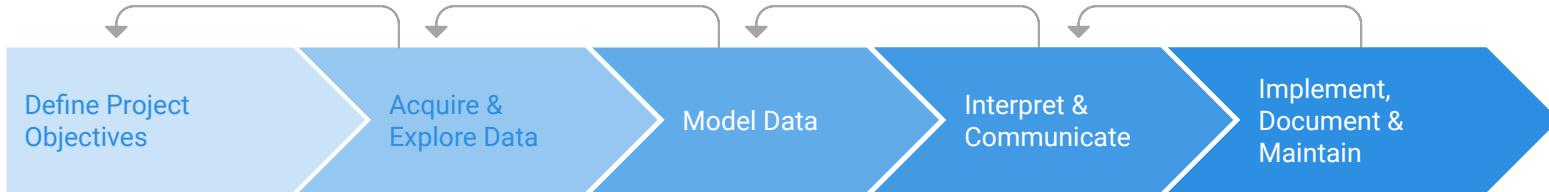


Data Drift





The Machine Learning Life Cycle



1. Define Project Objectives

- Specify problem
- Acquire subject matter expertise
- Define target and unit of analysis
- Prioritize modeling criteria
- Consider success criteria and risks
- Decide whether to continue

2. Acquire & Explore Data

- Find appropriate data
- Format data into single table
- Conduct exploratory data analysis
- Find and remove any target leakage
- Perform feature engineering

3. Model Data

- Select features
- Build candidate models
- Validate models

4. Interpret & Communicate

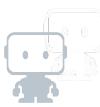
- Assess model quality
- Determine important features
- Identify relationships
- Explain predictions

5. Implement, Document & Maintain

- Select a model for deployment
- Document modeling process
- Create model monitoring and maintenance plan

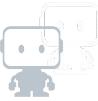


Lunch



Documentation and Support





Roles

Capability	Owner	User	Observer
View everything	✓	✓	✓
Launch IDEs	✓	✓	
Make predictions	✓	✓	
Create and edit feature lists	✓	✓	
Set target	✓	✓	
Delete jobs from queue	✓	✓	
Run Autopilot	✓	✓	
Share project with others	✓	✓	
Rename project	✓	✓	
Delete project	✓		
Unlock holdout	✓		
Clone the project	✓	✓	



DataRobot offers a growing variety of certifications, from introductory to advanced, with many specialties covered including:

- *DataRobot Deeper Dive Certification*
- *DataRobot Time Series Modeling Certification*
- *Advanced DataRobot with Python Certification*
- *Advanced DataRobot with R Certification*

Check out:

[datarobot.com/education](https://www.datarobot.com/education)

for more information





Hands-on Project



End of Day

Please complete *Course Exam* and
Course Survey in Litmos

DataRobot