DataRobot

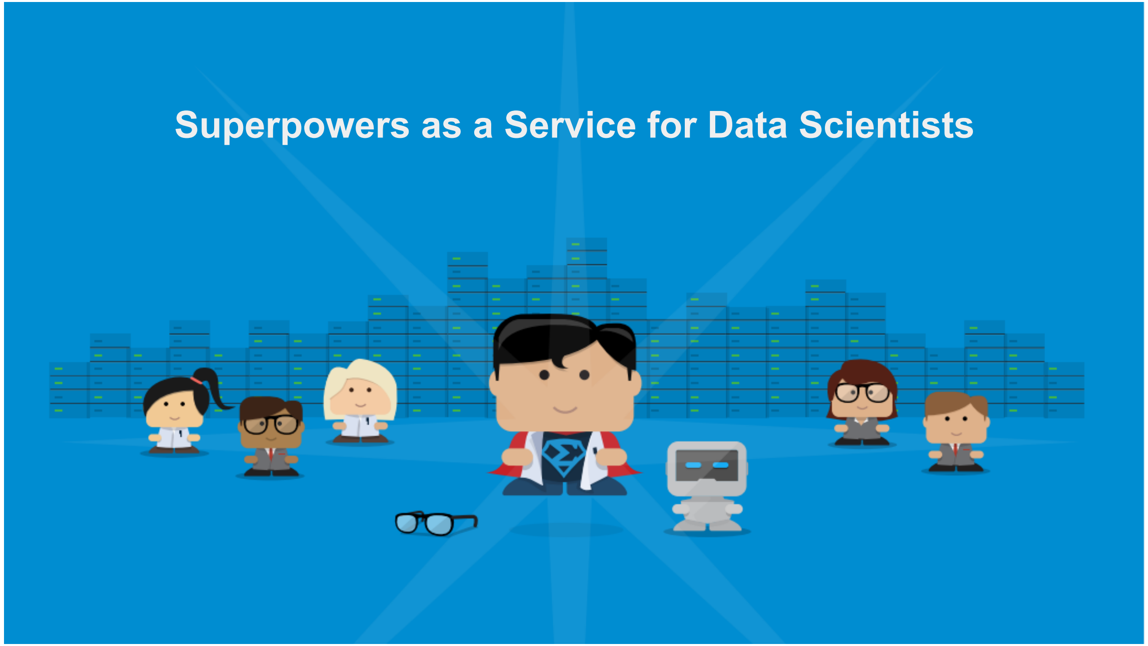
Winning the Liberty Mutual Kaggle Competition

Disclaimer:

- We do NOT state or imply Liberty Mutual endorses me, our team, our submission, Data Robot, Inc. or other companies I am involved with, or any of its or their products
- We thank Liberty Mutual for hosting the competition and identifying us as the winner of this competition.

AGENDA

1. Team
2. Competition Summary
3. Challenges
4. Our Approach
5. Key Takeaways



DATAROBOT TEAM

Xavier Conort

Chief Data Scientist

Kaggle Rank: #1   2012-2013

Experience:
Principal Research Engineer, I2R

Actuary, Risk Manager and CFO at CNP Assurances and AXA

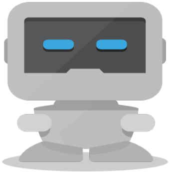
Owen Zhang

Chief Product Officer

Kaggle Rank: #1   2013-present

Experience:
VP, Science, AIG

Sr. Director, Analytics at Travelers Insurance



DataRobot

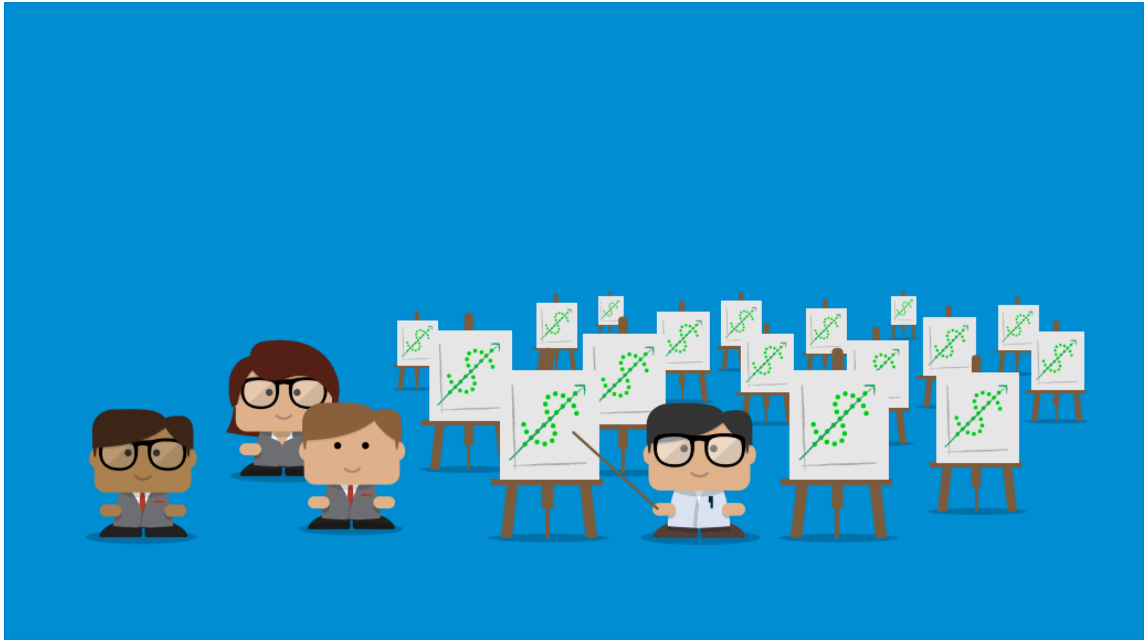
Data Science Master

Kaggle Rank: Master (top 1%)

Experience:
10+ million predictive models and counting ...

AGENDA

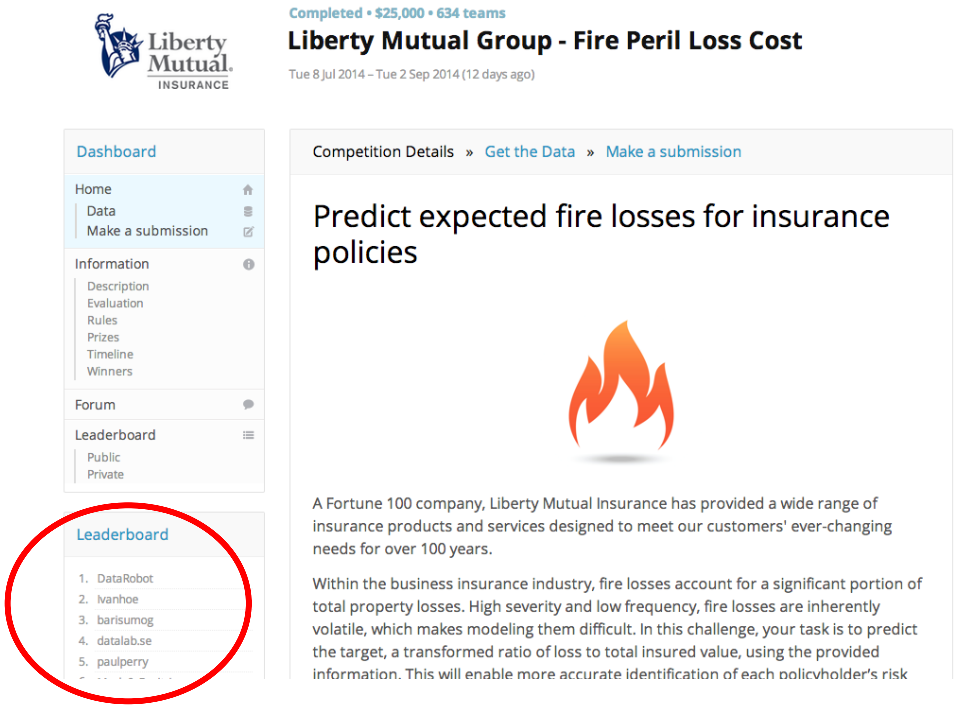
1. Team
2. Competition Summary
3. Challenges
4. Our Approach
5. Key Takeaways



COMPETITION SUMMARY

Fire Peril Loss Cost:

- Organized by Liberty Mutual
- Business problem: Predict expected fire losses for insurance policies
  - Significant portion of property losses
  - Volatile and hard to model correctly
- Started **July 8, 2014**
- Finished **Sept 2, 2014**



Team DataRobot finished 1st out of 634 teams competing globally!

DATA OVERVIEW

- ~1 million insurance records
- 300 variables:

target : The transformed ratio of loss to total insured value

id : A unique identifier of the data set

dummy : Nuisance variable used to control the model, but not a predictor

var1 – var17 : A set of normalized variables representing policy characteristics

crimeVar1 – crimeVar9 : Normalized Crime Rate variables

geodemVar1 – geodemVar37 : Normalized geodemographic variables

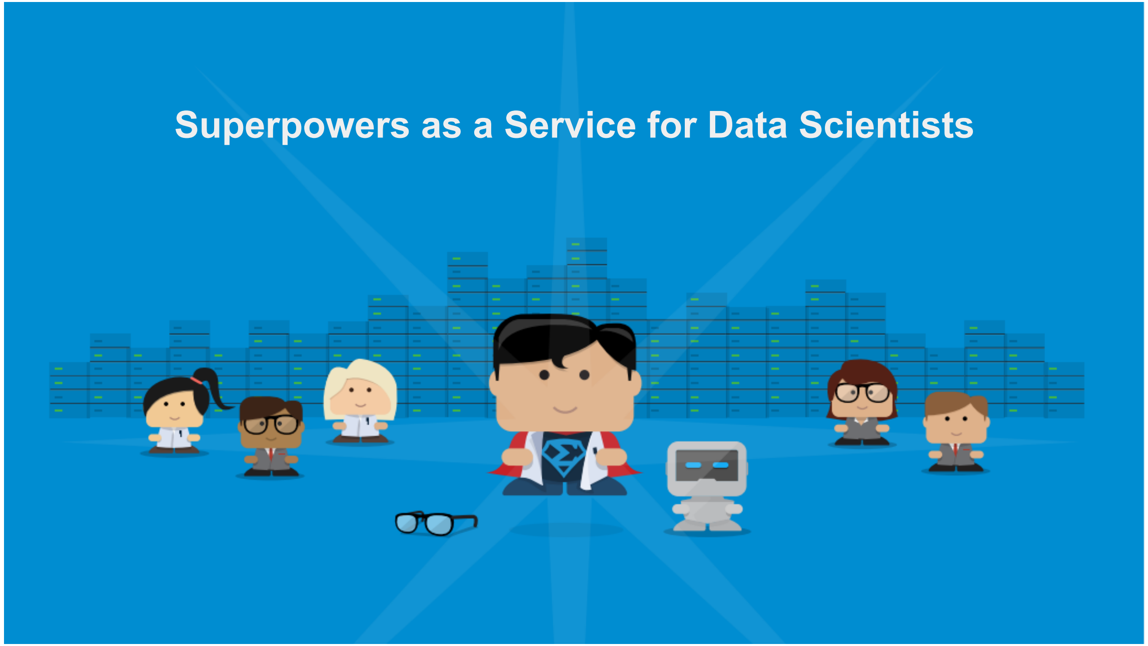
weatherVar1 – weatherVar236 : Normalized weather station variables

Numeric Variable Name	Variable Type
target	Continuous
id	Discrete
var10	Continuous
var11	Continuous
var12	Continuous
var13	Continuous
var14	Continuous
var15	Continuous
var16	Continuous
var17	Continuous
crimeVar1 – crimeVar9	Continuous
geoDemVar1 – geoDemVar37	Continuous
weatherVar1 – weath	

Categorical Variable Name	Variable Type	Possible Values
var1	Ordinal	1, 2, 3, 4, 5, Z*
var2	Nominal	A, B, C, Z*
var3	Ordinal	1, 2, 3, 4, 5, 6, Z*
var4+	Nominal	A1, B1, C1, D1, D2, D3, D4, E1, E2, E3, E4, E5, E6, F1, G1, G2, H1, H2, H3, I1, J1, J2, J3, J4, J5, J6, K1, L1, M1, N1, O1, O2, P1, R1, R2, R3, R4, R5, R6, R7, R8, S1, Z*
var5	Nominal	A, B, C, D, E, F, Z*
var6	Nominal	A, B, C, Z*
var7	Ordinal	1, 2, 3, 4, 5, 6, 7, 8, Z*
var8	Ordinal	1, 2, 3, 4, 5, 6, Z*
var9	Nominal	A, B, Z*
dummy	Nominal	A, B

AGENDA

1. Team
2. Competition Summary
3. Challenges
4. Our Approach
5. Key Takeaways



CHALLENGES

Low frequency
Very few homes get burned down!

High severity
A burnt down home may cost a lot of $$$

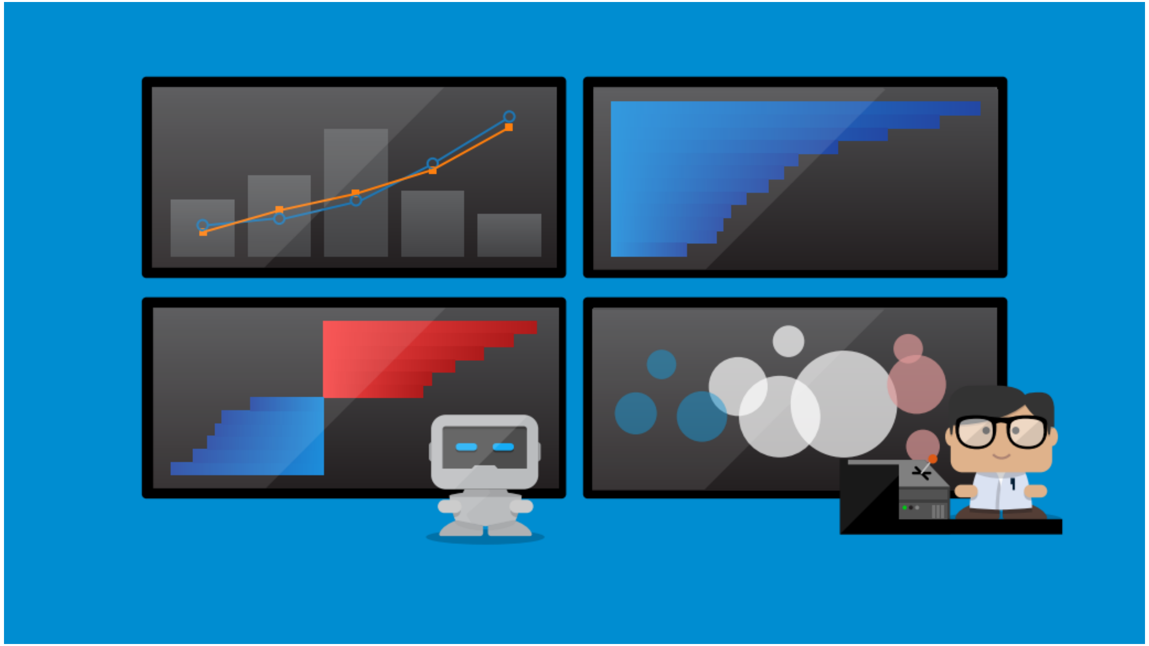
Many features

300 different variables!

Common Challenge for P&C Insurance dataset

AGENDA

1. Team
2. Competition Summary
3. Challenges
4. Our Approach
5. Key Takeaways



OUR APPROACH

High severity
(presence of large claims)

↓

Censor large claims

+

Low frequency
(few claims events)

↓

Downsample non-claims

+

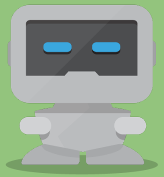
Many Features

↓

Reduce noise through feature selection

+

Explore best pre-processing and algorithms



+

Explore other models manually

CENSORING LARGE CLAIMS

Technique

- To censor large claims, we capped the target at the 80% quantile of non-zero losses

Impact

- More accurate and robust results than modeling with the raw target or its log transformation

DOWNSAMPLING MAJORITY EVENTS

Technique

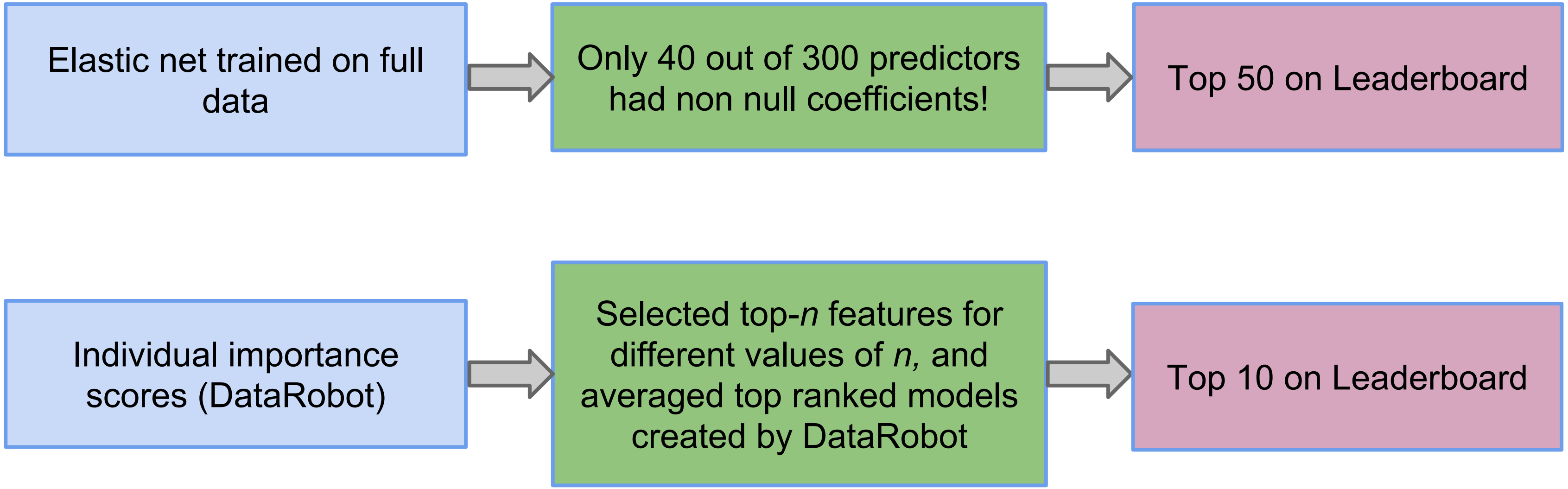
- Kept all non-0 records but only small % of 0 records from the training dataset

Impact

- Sped up model training significantly, and improved accuracy of certain ML algorithms (RandomForest, ExtraTrees, etc.)

REDUCING NOISE

- Ran 2 automated feature selections to detect large presence of noise
- Produced our own feature list with significantly low noise



FEATURE ENGINEERING

- Broke feature set into 4 components
- Created surrogate ID based on identical crime, geodemographics and weather variables

Policy Characteristics

30 features:

- All policy characteristics features (17)
- Split V4 into 2 levels (8)
- Computed ratio of certain features
- Combined surrogate ID and subsets of policy vars

Geodemographics

1 feature:

- Derived from PCA trained on scaled vars

Weather

1 feature:

- Derived from elasticnet trained on scaled variables

Crime Rate

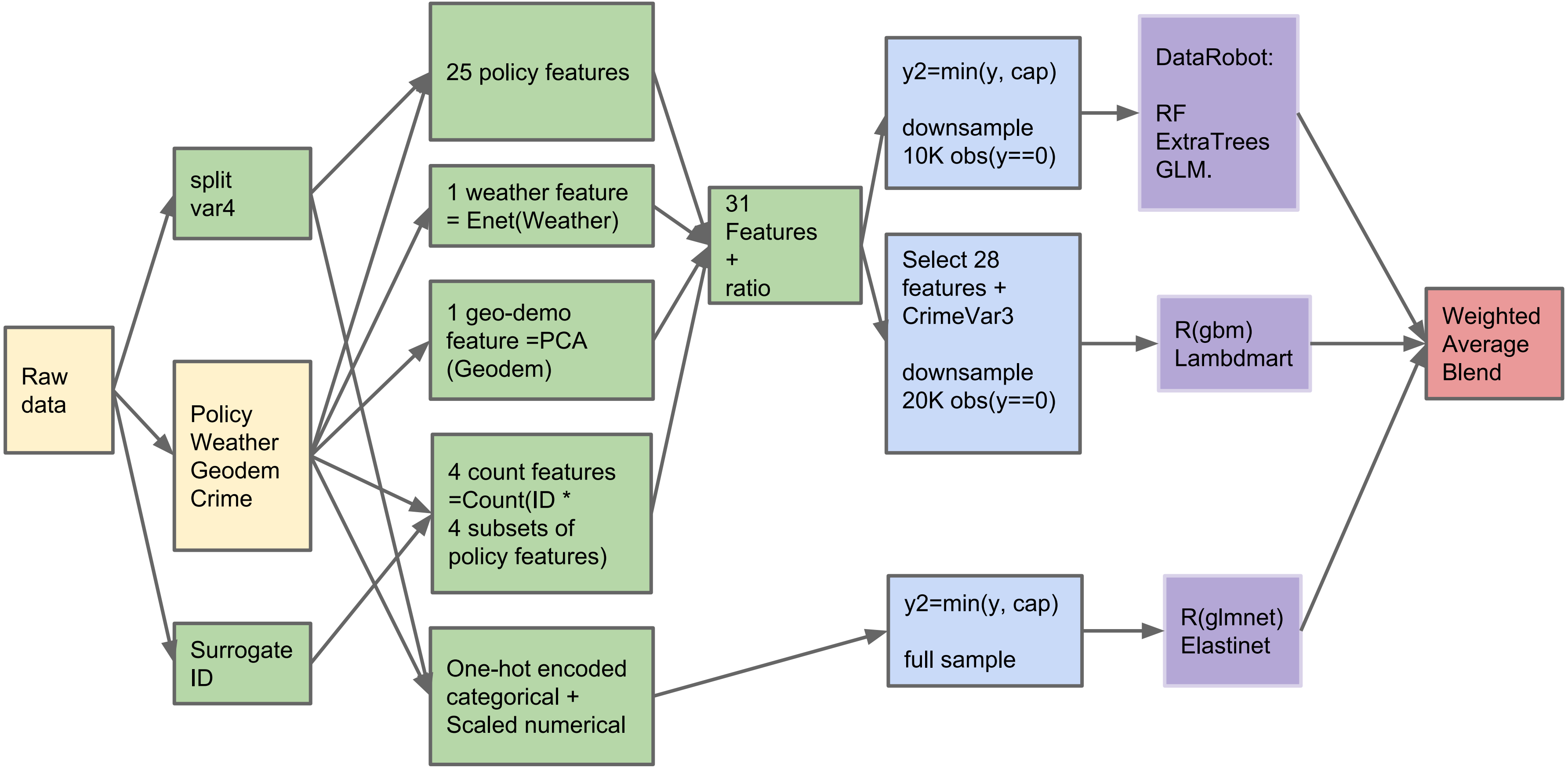
0 features

32 features

MODELING WITH DATAROBOT

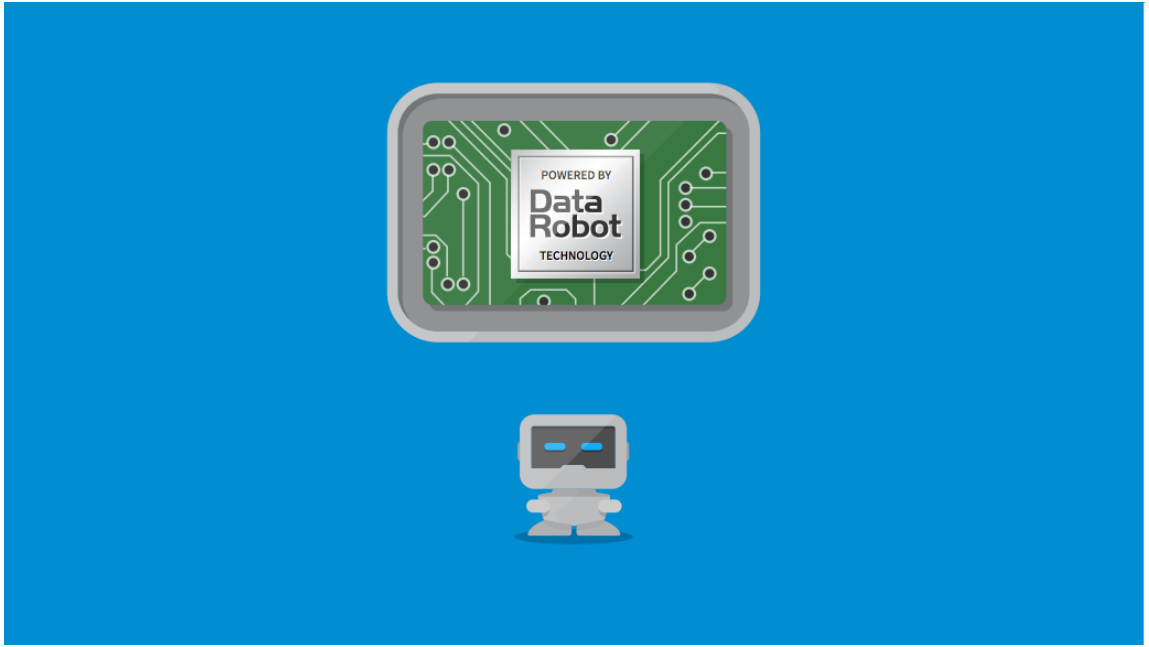
- Uploaded downsampled dataset with 32 features and capped target
- Ran multiple models in parallel with Normalized Gini metric to rank them
- Selected 3 promising models based on their 5-CV scores and trained them on 100% data
    - ExtraTrees Regressor
    - RandomForest Regressor
    - Ordinary Least Squares Regressor

FINAL SOLUTION SUMMARY



AGENDA

1. Team
2. Competition Summary
3. The Challenge
4. Our Approach
5. Key Takeaways



KEY TAKEAWAYS

1. Classic insurance problem set - keys to addressing this problem were:
    - Capping
    - Downsampling
    - Feature selection to reduce noise
    - Extracting value from blocks of features
2. Capabilities of DataRobot
    - Rapidly explore diverse combinations of feature transformations and ML algorithms
    - Build highly accurate models from this search space automatically