HR Analytics: Using ML To Predict ***Employee Turnover***

EARL BOSTON, 2017

Matt Dancho
Founder
Business Science
mdancho@business-science.io

Business Science
570.419.4337
@bizScienc
www.business-science.io

EARL
ENTERPRISE APPLICATIONS OF THE R LANGUAGE

Business Science
Applying Data Science to Business and Financial Analysis

What We Do

- Consulting
  - Executive Leadership
  - Bolt-on data science team
  - **ML + Leadership = Good Decision Making**

- Community-driven
  - Educate data scientists
  - Open Source Software
  - Courses coming in 2018!

How We Help The Business

Executive Leadership

Education

- Focus on significant problems
- How data fits into the picture
- How ML can help
- Risk mitigation

Data Management

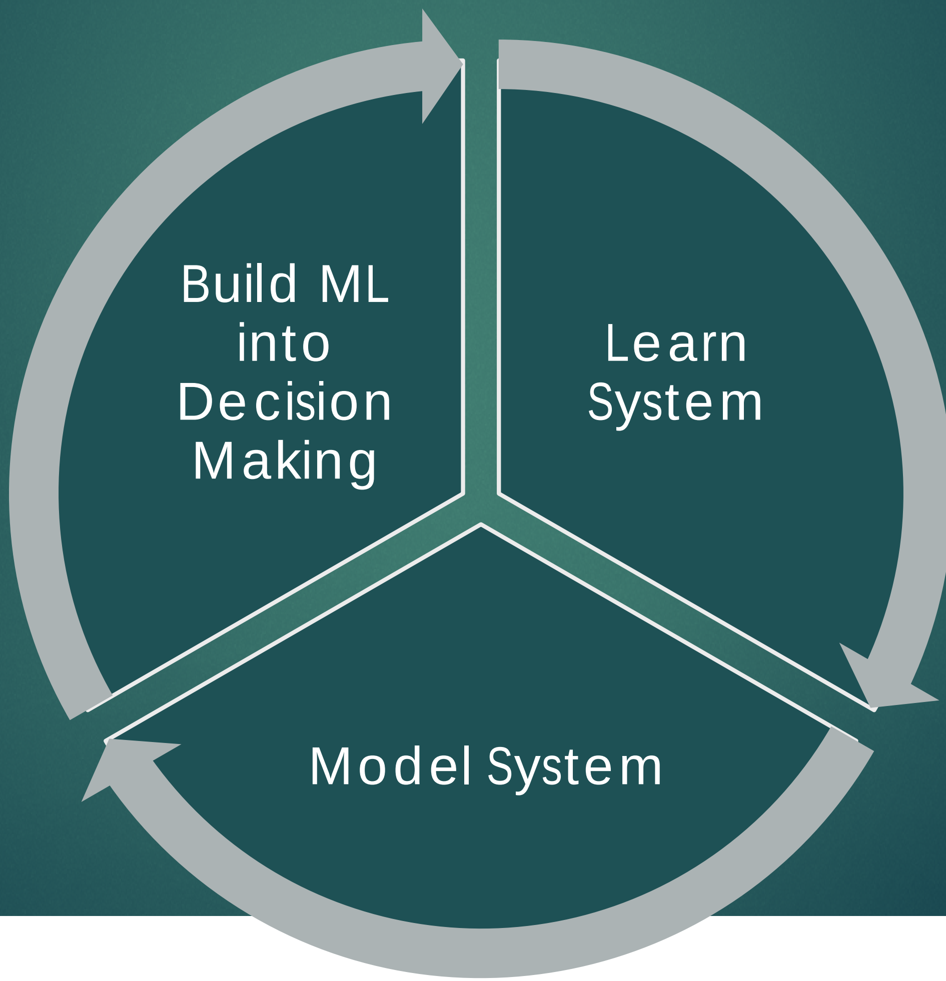
- Collecting data that yields results
- Building a data management process

Data Science

- Benefit from machine learning
- Distributing analytics
- Making decisions with ML insights

Business Science Approach

Systematic Process, Adaptive Approach

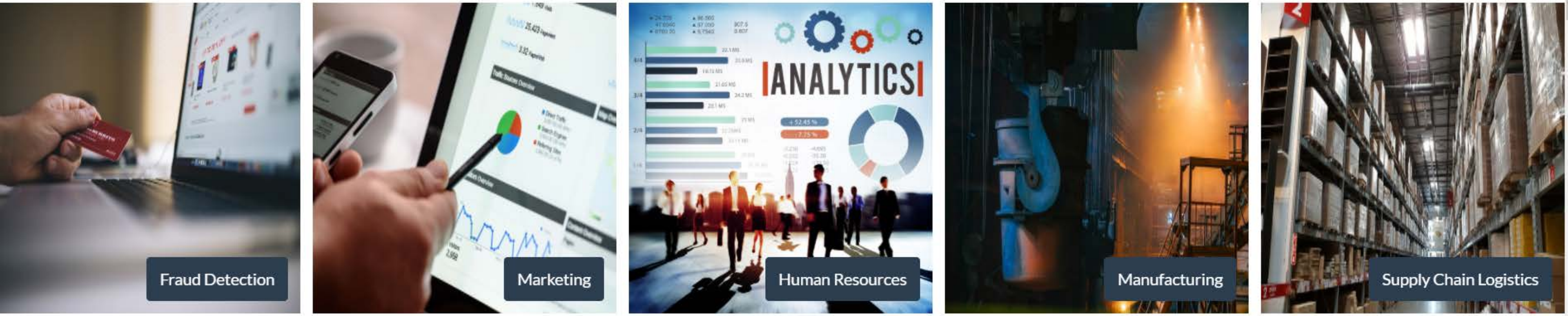


EARL
ENTERPRISE APPLICATIONS OF THE R LANGUAGE

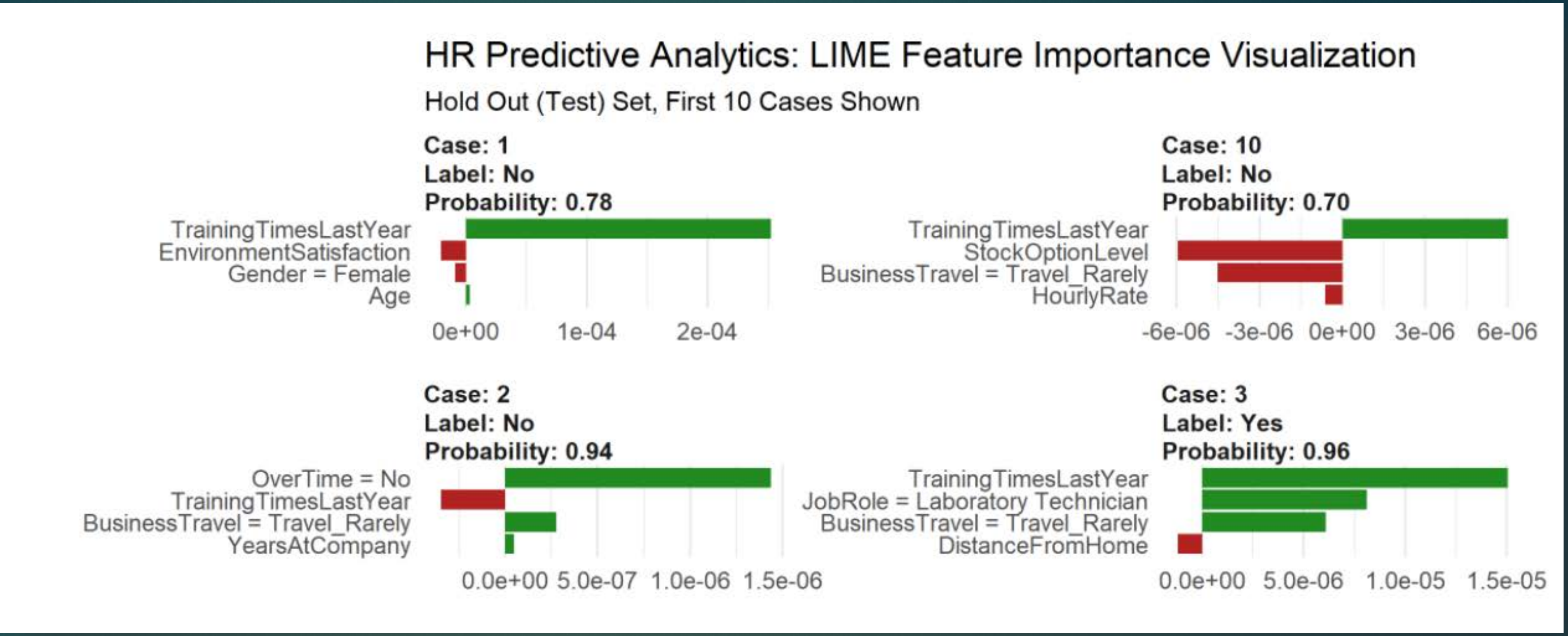
Business Science Expertise

Apply Systematic Approach To Any Problem

OUR EXPERTISE



EARL
ENTERPRISE APPLICATIONS OF THE R LANGUAGE



Case Study: HR Analytics

USING MACHINE LEARNING TO *PREDICT* & *EXPLAIN*

EMPLOYEE TURNOVER

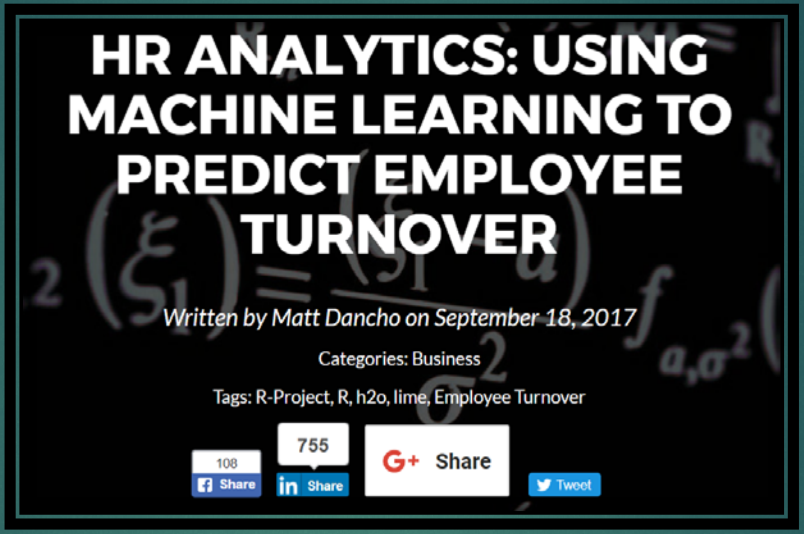
3 Reasons You Should Listen

1. Employee attrition: **A HUGE PROBLEM**

2. New techniques to **predict** & **explain** turnover

3. **Framework** for ML in business applications

A 4th Reason: It's Popular



Also featured on:

R-Bloggers ● KDNuggets ● LinkedIn

Code available in article:

http://www.business-science.io/business/2017/09/18/hr_employee_attrition.html

Just google:

"Predict Employee Turnover"

Employee Turnover

"You take away our top 20 employees and overnight we [Microsoft] become a mediocre company."

-Bill Gates

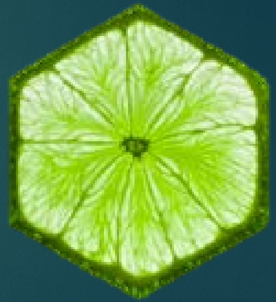
Cost Of Turnover

Organizations face huge costs resulting from employee turnover

- **Most important costs are intangible**:
  - When productive employee quits
  - **Lost**: New product ideas, great project management, or customer relationships

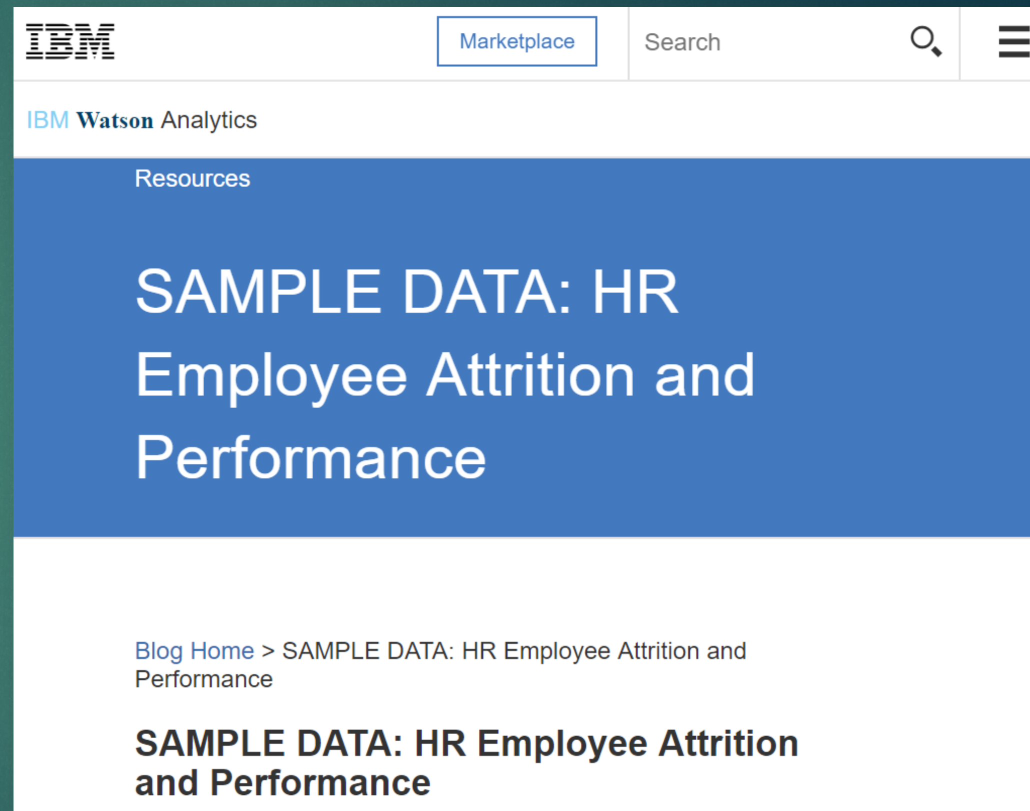
ML Tools Are Evolving

- **H2O**
  - **Automated Machine Learning**
  - Predict at very high accuracy
  - Complex models can’t be explained
- **LIME**
  - Used to explain ML classifiers
  - **Deep learning, stacked ensembles now explainable**



IBM Watson Data

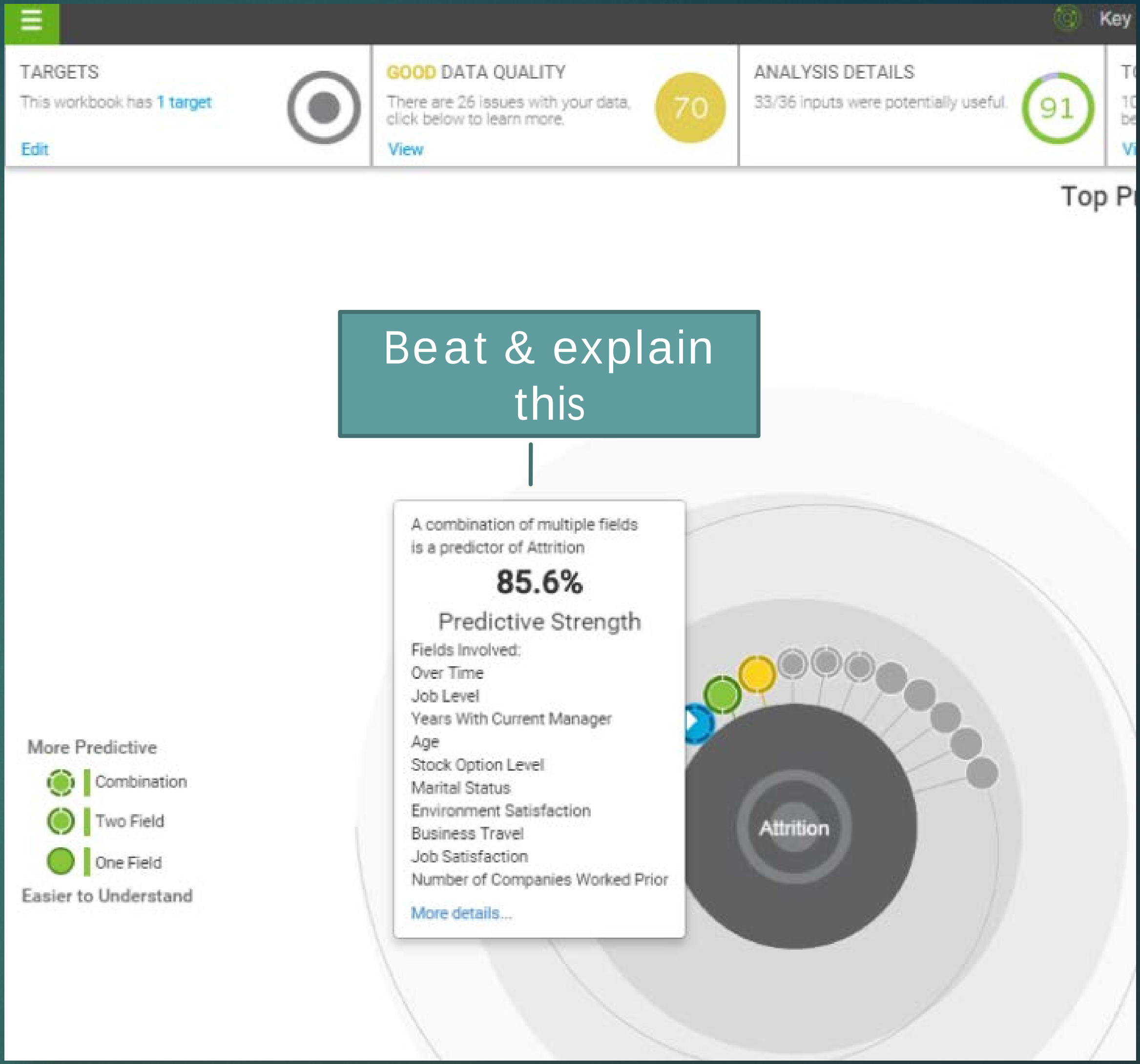
- Simulated HR Database
- Representative of real-world data
- Used for IBM Watson Case Study



Source: https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employee-attrition/

Goals

- Improve predictive accuracy
- Explain features that drive model



Source: https://www.ibm.com/communities/analytics/watson-analytics-blog/watson-analytics-use-case-for-hr-retaining-valuable-employees/

Feature Set

- HR Dataset
- 35 Features
- 1,470 Observations

Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	Education
41	Yes	Travel_Rarely	1102	Sales	1	2	Life Science
49	No	Travel_Frequently	279	Research & Development	8	1	Life Science
37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other
33	No	Travel_Frequently	1392	Research & Development	3	4	Life Science
27	No	Travel_Rarely	591	Research & Development	2	1	Medical
32	No	Travel_Frequently	1005	Research &	2	2	Life Science

Modeling With H2O

- Training the model

```
# Split data into Train/Validation/Test Sets
hr_data_h2o <- as.h2o(hr_data)

split_h2o <- h2o.splitFrame(hr_data_h2o, c(0.7, 0.15), seed = 1234 )

train_h2o <- h2o.assign(split_h2o[[1]], "train" ) # 70%
valid_h2o <- h2o.assign(split_h2o[[2]], "valid" ) # 15%
test_h2o  <- h2o.assign(split_h2o[[3]], "test" )  # 15%
```

```
# Run the automated machine learning
automl_models_h2o <- h2o.automl(
    x = x,
    y = y,
    training_frame    = train_h2o,
    leaderboard_frame = valid_h2o,
    max_runtime_secs  = 30
    )
```

Automated ML:
-Deep Learning
-Ensembles
-GBM

Modeling With H2O

- Prediction: Test Data (Unseen)

```
# Predict on hold-out set, test_h2o
pred_h2o <- h2o.predict(object = automl_leader, newdata = test_h2o)
```

- Performance: 88% Accuracy

```
## [[1]]
## [[1]]$accuracy
## [1] 0.8767773
##
## [[1]]$misclassification_rate
## [1] 0.1232227
##
## [[1]]$recall
## [1] 0.6206897
##
## [[1]]$precision
## [1] 0.5454545
##
## [[1]]$null_error_rate
## [1] 0.7914692
```

Important for Goal

Important for Business Case

Puts Accuracy Into Perspective

EAR

ENTERPRISE APPLICATIONS OF THE R LANGUAGE

HR Implications

- Recall = 62%
  - Will correctly classify those at risk of turnover 62 of 100 times
  - Critical to the business
    - 62% of at risk employees that can be targeted preemptively
- Precision = 54%
  - Will avoid incorrectly assigning “Yes” 54 of 100 times
  - Better to target incorrectly than miss
    - Should not sacrifice Recall

*Have a **great model**, but…*

*how do we **prevent turnover**?*

EARL
ENTERPRISE APPLICATIONS OF THE R LANGUAGE

LIME

- **L**ocal **I**nterpretable **M**odel-Agnostic **E**xplanation

- Theory
  - LIME approximates model locally as logistic or linear model
  - Repeats process 5000X
  - Outputs features that are important to local models

- Result: Data Scientists Understand Why Model Predicts What it Predicts

LIME

- Complex classification models can now be interpreted
  - Black Box Models
  - Neural Networks, Ensembles, Random Forests

- **H2O and LIME now integrated!**
  - https://github.com/thomasp85/lime

LIME

- Step 1: Create explainer using lime()

```r
# Run lime() on training set
explainer <- lime::lime(
    as.data.frame(train_h2o[,-1]),
    model          = automl_leader,
    bin_continuous = FALSE)
```

Create explainer object

LIME

- Step 2: Create explanation using explain()

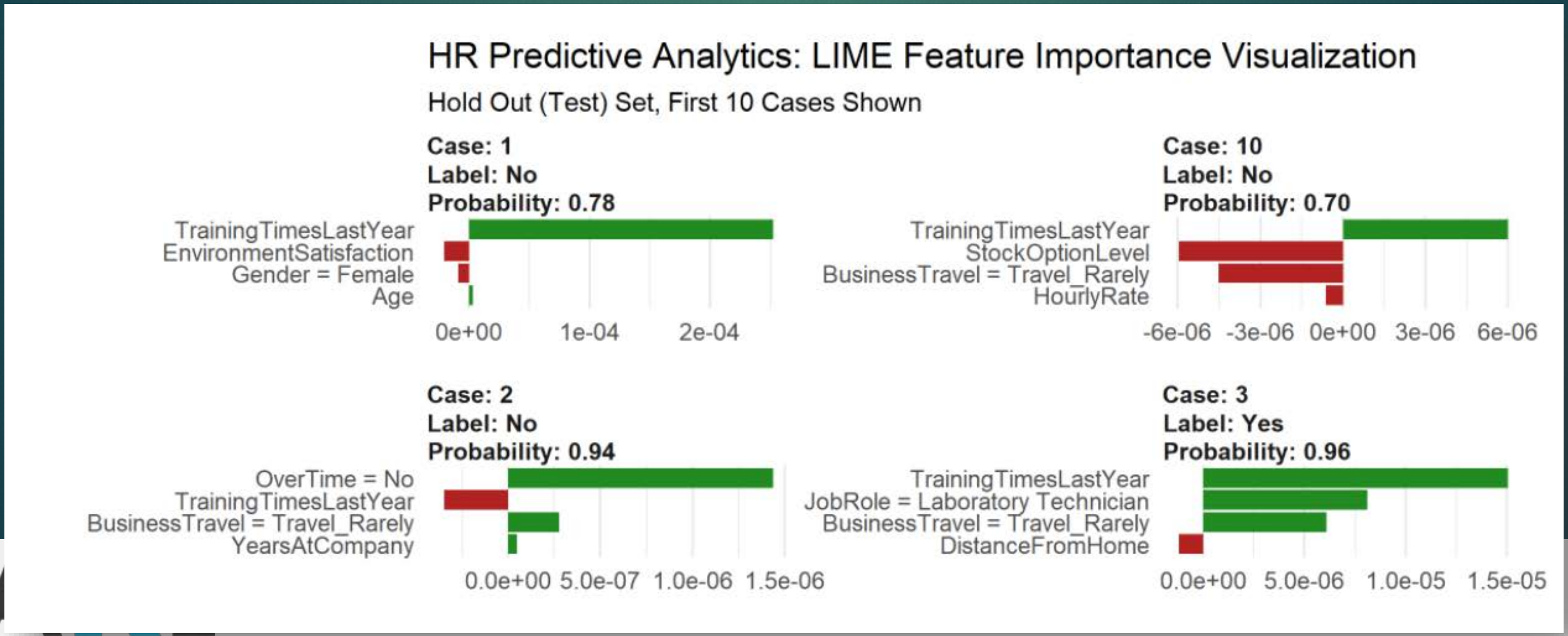
```r
# Run explain() on explainer
explanation <- lime::explain(
    as.data.frame(test_h2o[1:10,-1]),
    explainer    = explainer,
    n_labels     = 1,
    n_features   = 4,
    kernel_width = 0.5)
```

Explain new observations

LIME

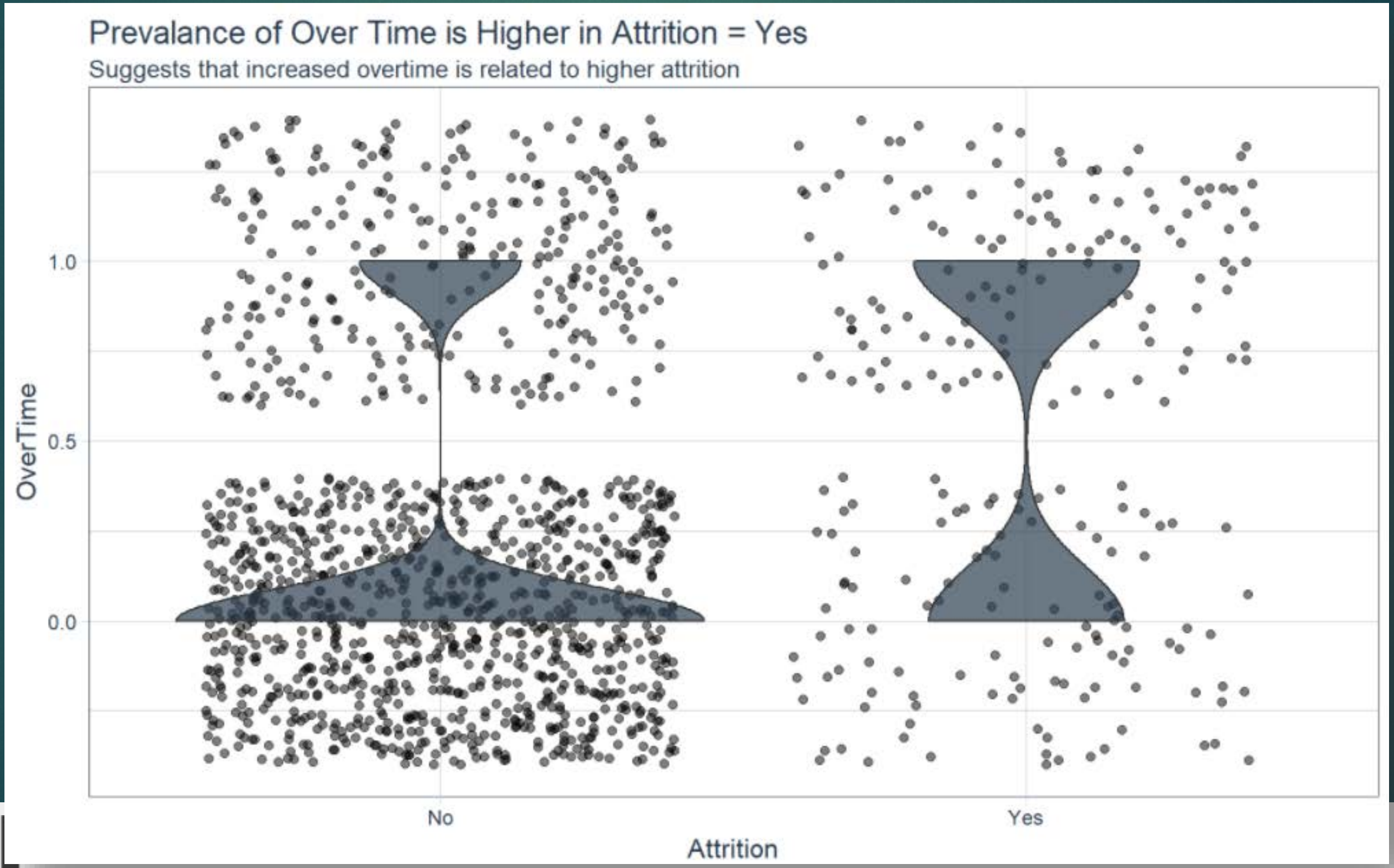
- Step 3: Plot Feature Importance

```r
plot_features(explanation) +
    labs(title = "HR Predictive Analytics: LIME Feature Importance Visualization",
         subtitle = "Hold Out (Test) Set, First 10 Cases Shown")
```



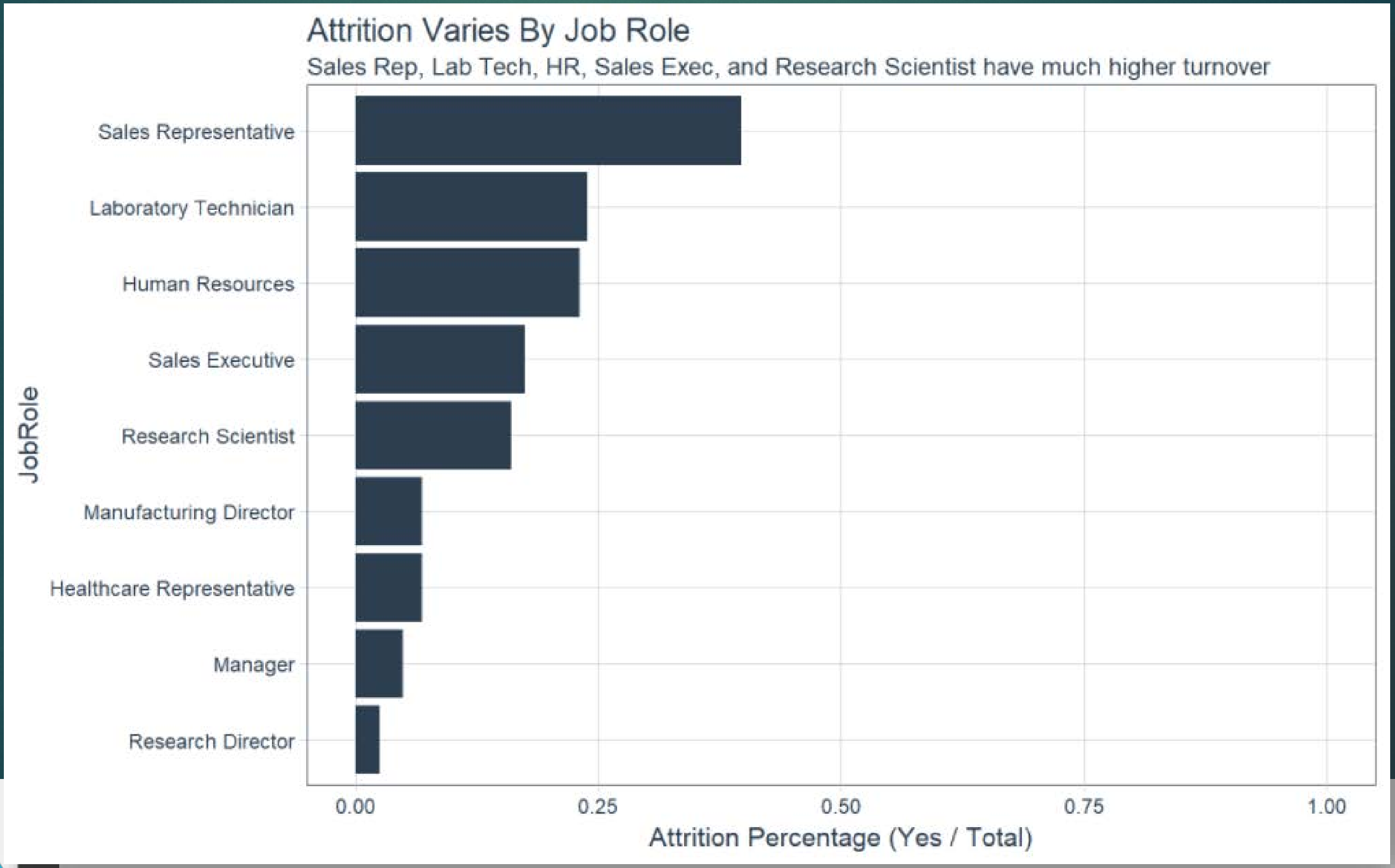
LIME

- Step 4: Investigate Important Features
  - Overtime



LIME

- Step 4: Investigate Important Features
  - Job Role



What About **Real World Applications**?

- Client Case Study
  - Fortune 500 firm
  - Modeled executive potential using more sophisticated process
  - Our algorithm identified 16 employees that predicted as executive potential but were not targeted by client

Conclusions

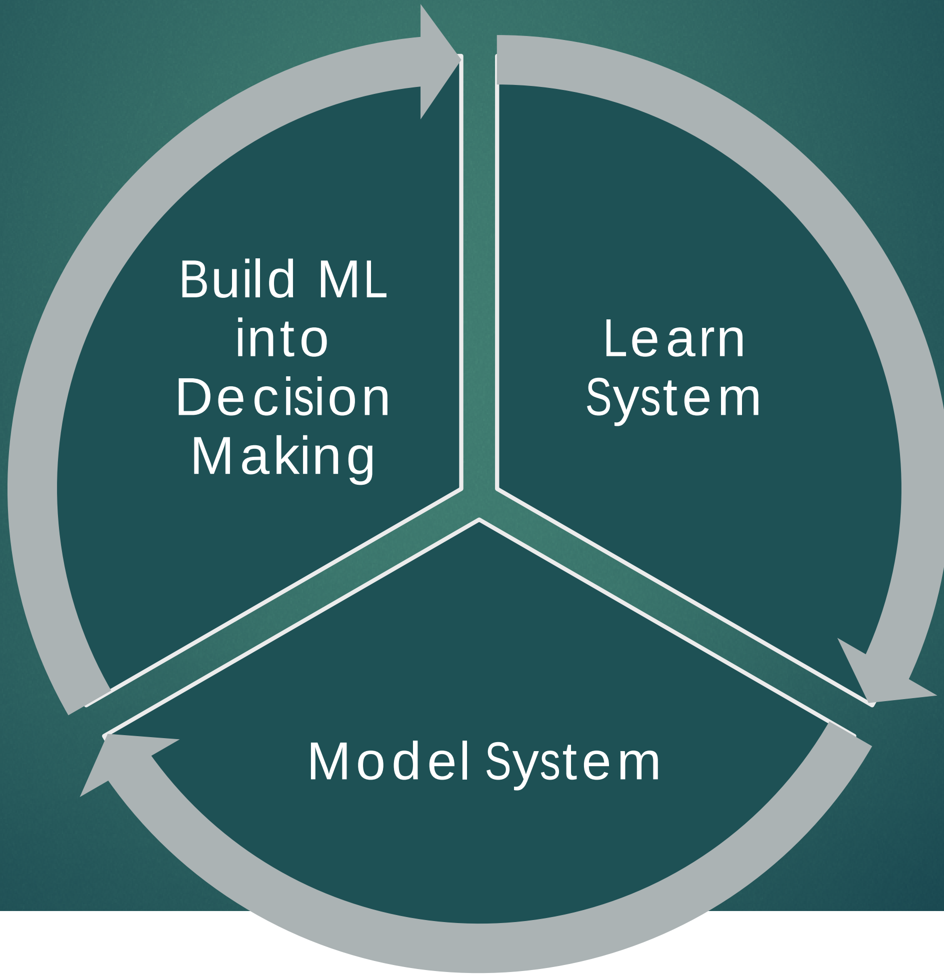
- Can use predictive analytics & ML for HR
  - Predicted turnover
  - 88% Accuracy
  - 62% Recall ← Important!
- Can explain black-box model
  - Turnover greater based on Job Role & Overtime
- Framework for high accuracy & explainability

We're done right? No!

- Risks
  - How do we know model is right? Model not back-tested
  - Time: Cross-sectional analysis, model not adaptive
  - What do we do when model breaks down?
  - Model: Your model will change, can't trust blindly
  - Only certainty: **CHANGE**

Business Science Approach

Systematic Process, Adaptive Approach



EARL
ENTERPRISE APPLICATIONS OF THE R LANGUAGE

Client Archetype

- Seeking predictive analytics to:
  - Understand business problem as a system
  - Increase profitability
  - Make better decisions
  - Mitigate data science risks
  - Convey insights to stakeholders

- No data science team

"Business Science is your Bolt-On Data Science Team"

Need Data Science for Business? Contact Business Science!

Business Science
www.business-science.io/contact
570.419.4337

Matt Dancho
Founder
mdancho@business-science.io

Try our software!

tidyquant
timetk
sweep
tibbletime

Business Science
Applying Data Science to Business and Financial Analysis

EARL
ENTERPRISE APPLICATIONS OF THE R LANGUAGE