

Anomaly Detection

Detecting Outliers using H2O

Difficulty: **Intermediate**

Outlier ● 0 ● 1



Matt Dancho & David Curry
Business Science Learning Lab



Learning Lab Structure

- **Presentation**
(30 min)

- **Demo's**
(30 min)

- **3 Pro-Tips**
(15 mins)

Business Science
Learning Lab 17
ANOMALY DETECTION WITH H2O

WEDNESDAY, AUG 28 @ 2PM EST

SPECIAL GUEST

Dr. Erin LeDell

Chief Machine Learning Scientist at H2O.ai

Dr. Erin LeDell is the Chief Machine Learning Scientist at H2O.ai the company that produces the open source, distributed machine learning platform, H2O.

Before joining H2O.ai, she was the Principal Data Scientist at two AI startups (both acquired), the founder of DataScientific, Inc. and a software engineer at a large consulting firm. She received her Ph.D. from UC Berkeley where her research focused on machine learning and computational statistics. She also holds a B.S. and M.A. in Mathematics.

H₂O.ai

Matt Dancho
Founder of Business Science, Matt designs and executes educational courses and workshops that deliver immediate value to organizations. His passion is up-leveling future data scientists coming from untraditional backgrounds.

David Curry
Founder of Sure Optimizze, David works with businesses to help improve website performance and SEO using data science. His passion is ethical Machine Learning initiatives.

Success Story

Robert M. Davis

- Works in EMS analytics
- Has struggled learning R & Python for **3 years!**
- Now, generating customized distributions & box plots
- Robert's **transformation** is **"eye-opening!"**



“Spent last 3 years trying to learn Python & R... This course [DS4B 101-R] is the real deal.”



Robert M. Davis, MPA • 2nd
The Analytics "Wizard" | Data Intelligent | Data Science in EMS | The Guy In GI...
2d

Learning `#ggplot` in week 4 has been a lot of fun...
After a rough week learning `#forcats` and categorical `#datawrangling` `ggplot` has been a nice transition

Being able to generate customized distribution, line, column, and box plots with only a few lines of code has been eye opening...

As someone that grew up with `#excel` as the go to data analysis and visualization platform for a decade, being able to replicate everything one would normally do in excel via `#rstudio` has been a trip...

Props to [Matt Dancho](#) on creating this course content via his [#BusinessScienceUniversity](#) platform...

I have spent the last 3 years trying to piece together learning `#python` and `#r`, but always coming up short and ultimately giving up...

What Matt has created here with his course DS4B-101 has been night and day in comparison to other online learning courses and resources...

If you are looking to get it wet in R, but you find `#programming` intimidating, or have struggled with courses in the past...

From someone that constantly misspells "price" as "proce" as he uses `select()` statements during the examples...

This course is the real deal 🎉🔥

DS4B 101 Business Analysis with R Information:
<https://lnkd.in/dx3qfA>

#datascience #continuedlearning

61 • 6 Comments

Reactions



Learning Labs PRO

Every 2-Weeks

1-Hour Course

Recordings + Code + Slack

\$19/month

university.business-science.io

Lab 16

**R's Optimization Toolchain, Part 2
- Nonlinear Programming**

Lab 15

**R's Optimization Toolchain, Part 1
- Linear Programming**

Lab 14

Customer Churn Survival Analysis

Lab 13

**Wrangling 4.6M Rows of Financial
Data w/ data.table**

Lab 12

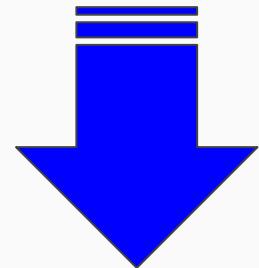
How I built anomalize

Lab 11

**Market Basket Analysis w/
recommenderLab**



Continuous Learning
Jet Fuel for your Brain



Learning Labs Pro

Community-Driven Data Science Courses

 Matt Dancho

\$19/m

H2O Company & Technology Updates



A Growth Company

- \$146M Raised
- **Series D \$72.5M**
- 18 Investors
- **Finance & AI**



Overview ?

Total Funding Amount	\$146.1M	CB Rank (Company)	183
----------------------	----------	-------------------	-----

H2O.ai
H2O.ai is an open source machine learning platform that makes it easy to build smart applications.
Mountain View, California, United States

Transaction Name	Organization Name	Funding Type	Money Raised
1. Series D - H2O.ai	H2O.ai	Series D	\$72,500,000
2. Series C - H2O.ai	H2O.ai	Series C	\$40,000,000
3. Series B - H2O.ai	H2O.ai	Series B	\$20,000,000
4. Series A - H2O.ai	H2O.ai	Series A	\$8,900,000
5. Seed Round - H2O.ai	H2O.ai	Seed	\$3,000,000

Investor Name	Lead Investor	Funding Round
Goldman Sachs	Yes	Series D - H2O.ai
Wells Fargo	—	Series D - H2O.ai
Nvidia GPU Ventures	—	Series D - H2O.ai
Nexus Venture Partners	—	Series D - H2O.ai
Ping An Global Voyager Fund	Yes	Series D - H2O.ai
Barclays Investment Bank	—	Series C - H2O.ai
CreditEase Fintech Investment Fund	—	Series C - H2O.ai
Nvidia GPU Ventures	Yes	Series C - H2O.ai
New York Life Insurance Co	—	Series C - H2O.ai
Nexus Venture Partners	—	Series C - H2O.ai

Source: <https://www.crunchbase.com/organization/h2o-2>



Open Source Technology Updates

The image shows a tablet displaying the H2O AutoML interface. On the left side of the screen, there is a code editor window showing R code for generating model metrics plots. On the right side, there are four plots: ROC, Precision vs Recall, Gain, and Lift. Below the plots, a terminal window shows the command used to generate the plots. A large yellow box at the bottom right contains the H2O.ai logo.

H2O AutoML

```
# Continue using ggplot
  p_legend <- get_legend(g)
  p_d <- gg + theme(legend.position = "none")
  p <- complete(plot_grid(p, p_d, n, n, ncol = 2))
  p$title <- ggplot()
  draw_labels(p$title, "Model Metrics", size = 15, fontface = "bold",
              colour = palette_light[1][1][1])
  p$subtitle <- ggplot()
  draw_labels(p$subtitle, "Ordered by (logloss)", size = 10,
              colour = palette_light[1][1][1])
  ret <- plot_grid(title, p_subtitle, p, p_legend,
                    ncol = 1, rel_heights = c(0.05, 0.05, 1, 0.05 * max_models))
  h2o.show_progress()
  return(ret)
}

actual_models_h2o_leaderboard %>
  plot.h2o.performance(metadata = test_tbl, order_by = "logloss") %>
  geom_vline(xintercept = 0.5, color = "black", size = 1)
```

H2O Model Metrics
Ordered by LOGLOSS

ROC

Precision vs Recall

Gain

Lift

H2O.ai



Psst... We teach H2O AutoML in 201

DS4B 201-R: Data Science For Business With R

Solve a real-world churn problem with H2O AutoML (automated machine learning) & LIME black-box model explanations using R

Matt Dancho

H2O AutoML

H2O.ai

Anomaly Detection

Business Case



Detecting Fraudulent Transactions

Credit Cards

Customer account becomes compromised

Bank is able to detect fraudulent transactions within minutes

Account is placed on hold

Saves both parties **billions** each year





Detecting Fraudulent Transactions

Key Issues

What does **abnormal behavior** mean?

How do we handle **Big Data Sets**?

Which techniques work gracefully in the presence of **High Imbalance**?

```
> # 1.1 CLASS IMBALANCE ----  
> credit_card_tbl %>%  
+   count(Class) %>%  
+   mutate(prop = n / sum(n))  
# A tibble: 2 x 3  
  Class     n     prop  
  <dbl> <int>    <dbl>  
1     0 284315  0.998  
2     1    492 0.00173
```



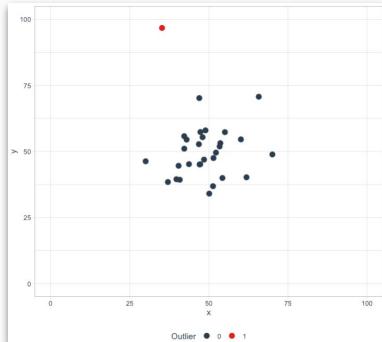
Other Business Reasons to Detect Anomalies



Exploratory Analysis

Understanding data

Identifying data issues

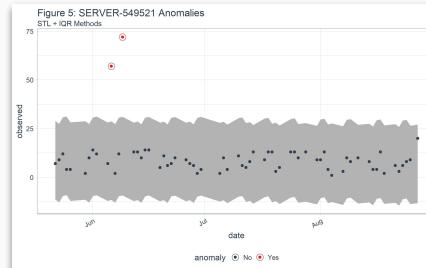


Key Business Cases

Spikes in Sales Demand

Detecting Machinery Malfunction

Identifying Malicious Behavior



What are Anomalies?

Key Terminology



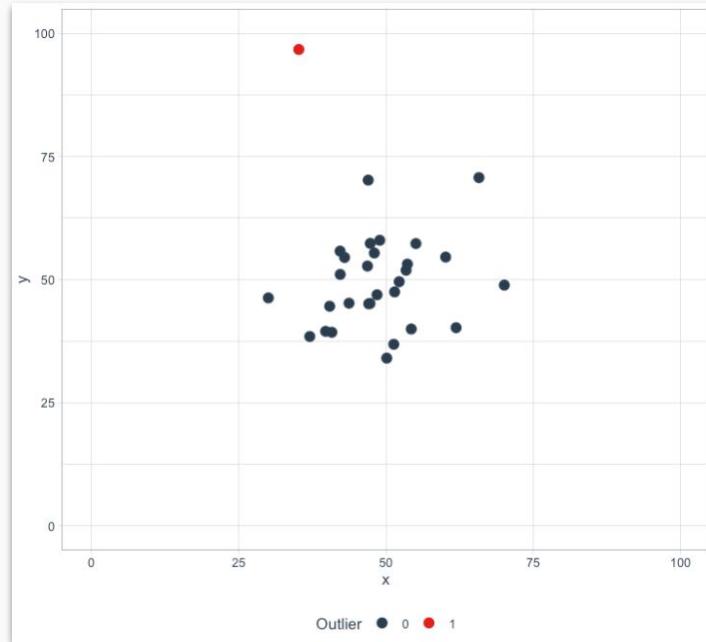
What are Anomalies?

Outliers

Must be understood

Common Causes:

1. Data Entry Errors
2. Unusual Events
3. Unusual Patterns



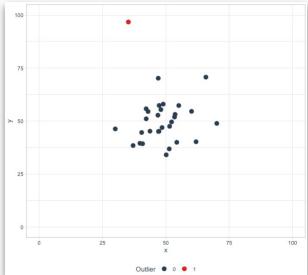


Types of Anomalies

1

Point Anomalies

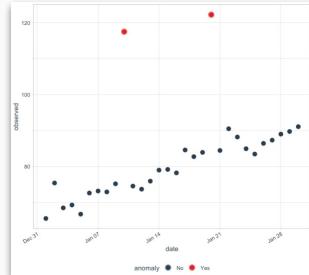
Single Point



2

Contextual

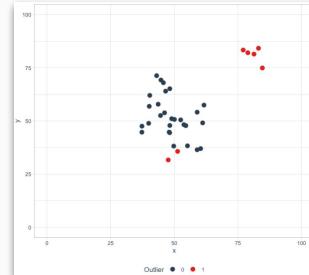
Time Series



3

Collective

Cluster of Points



Anomaly Detection Algorithms

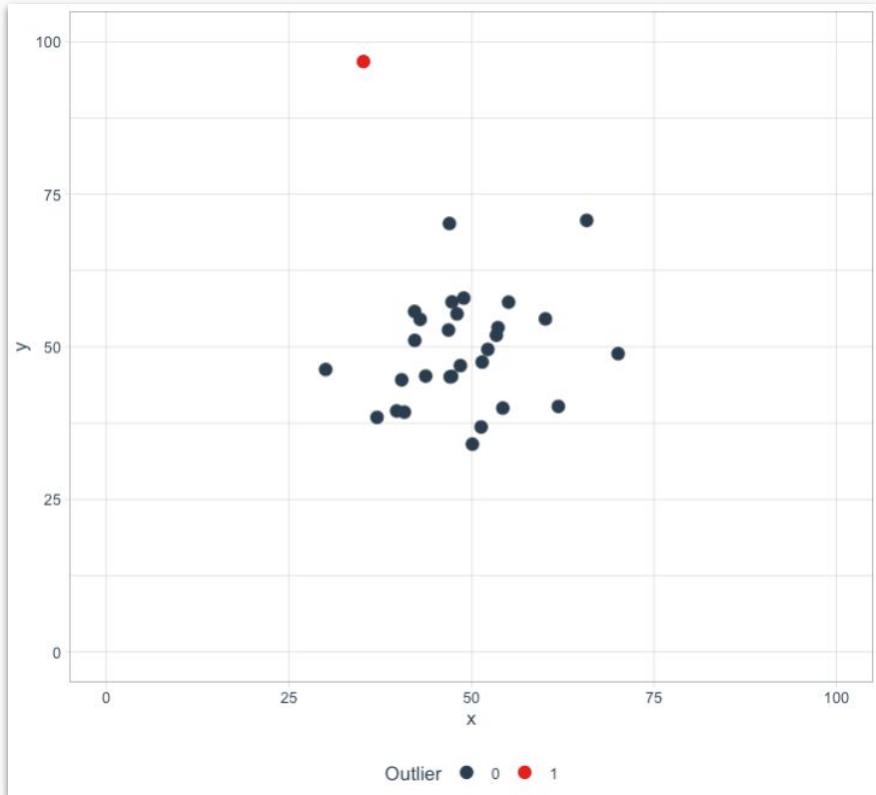


Unsupervised

- kNN
- K-Means
- [NEW] Isolation Forest

Supervised

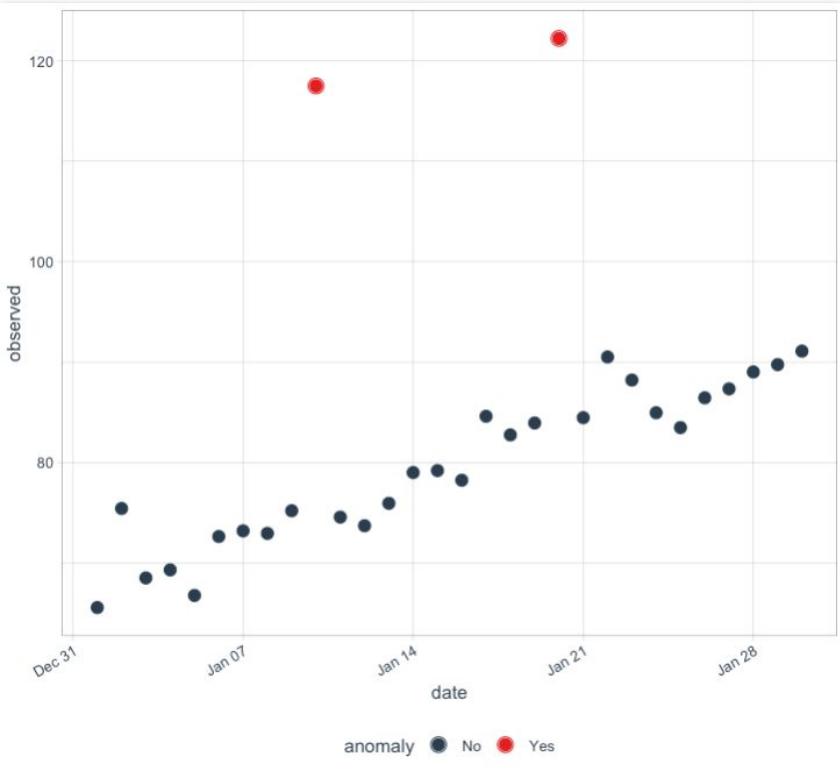
- SVM & XGBoost (great if data is labeled)





Unsupervised

- Anomalize (Lab Coming Soon)





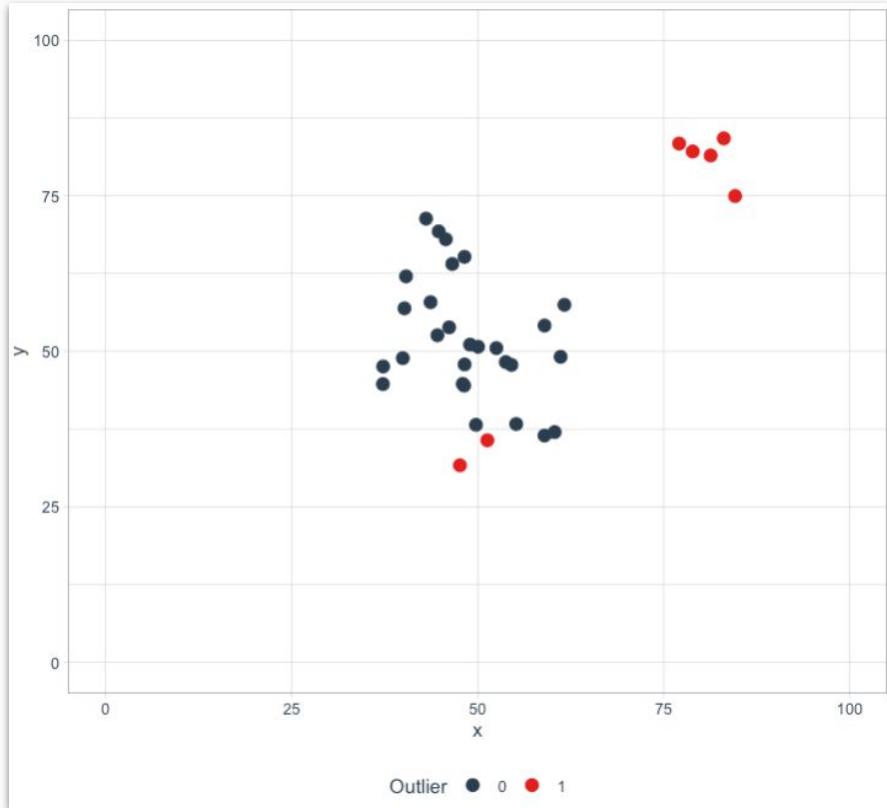
Groups (Collective)

Unsupervised

- kNN
- K-Means
- [NEW] Isolation Forest

Supervised

- SVM & XGBoost (great if data is labeled)





Point & Cluster

Not Time Series / Contextual

Algorithm	Unsupervised	Scalable
kNN	✓	
K-Means	✓	
Isolation Forest (H2O)	✓	✓
SVM		✓
XGBoost		✓

Process & Tools

What we need to know



Fraud Anomaly Detection

Step-By-Step



Isolation Forest

80/20 Concepts & Important Operations



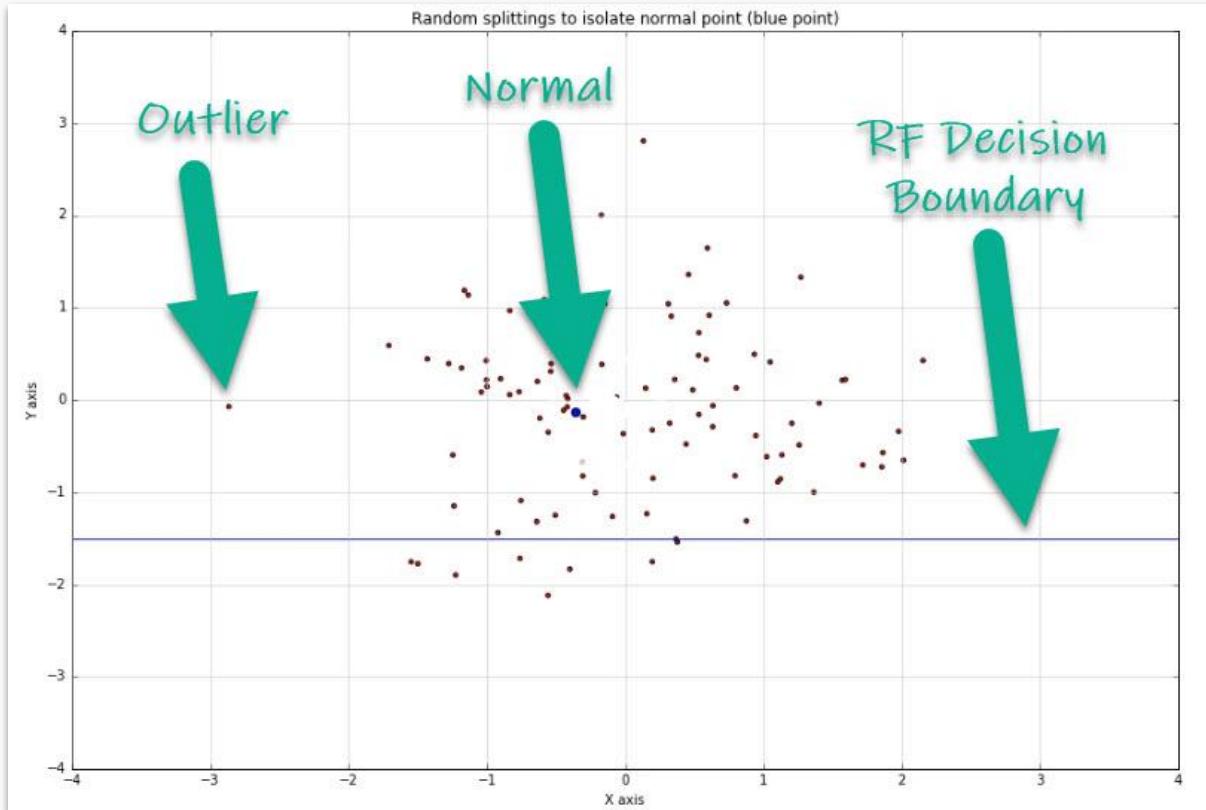
How Isolation Forest Works

Algorithm Internal Process

- Uses Random Forest Algorithm
- Randomly Selects One Feature (Target)
- Random Splits, Separating & Classifying Data

Key Concept

Outliers have **fewer splits** to isolate in the decision tree

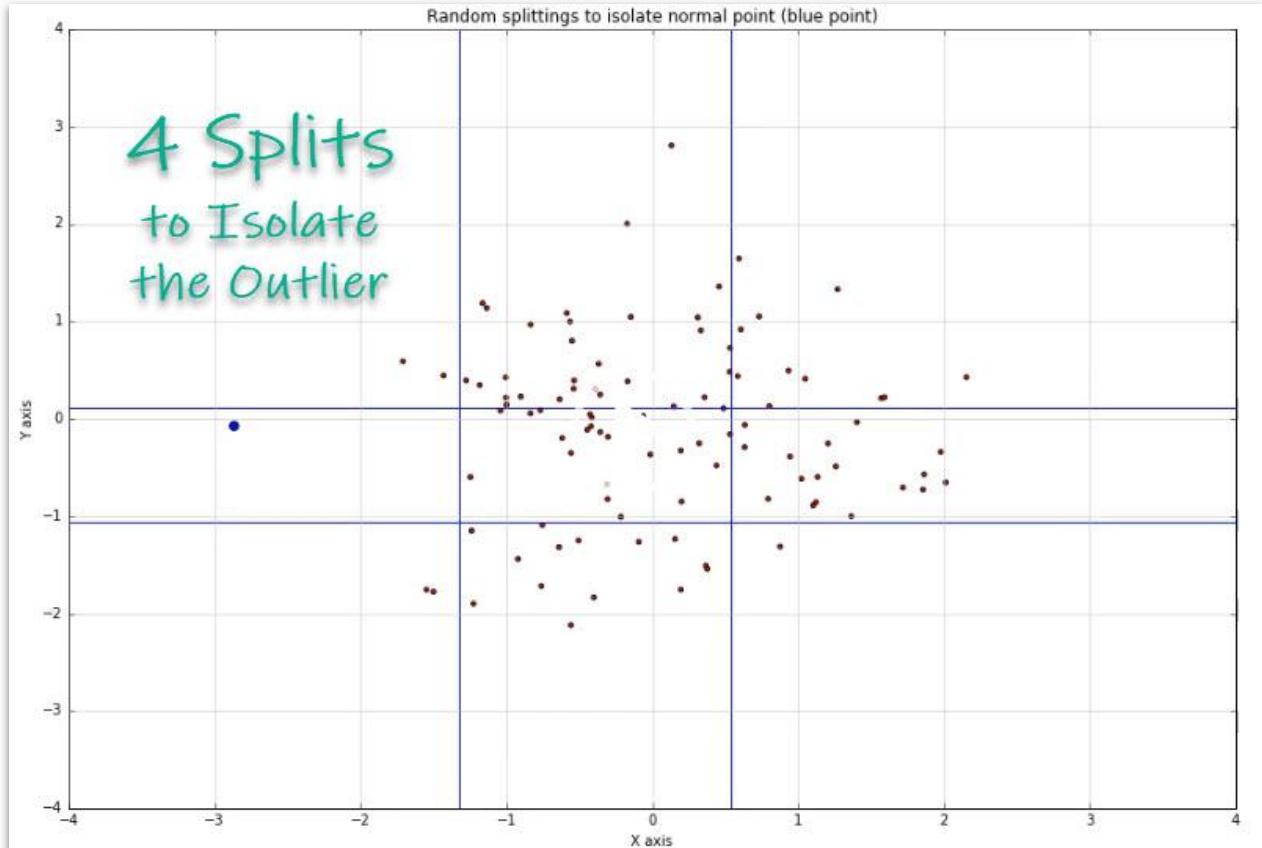


How Isolation Forest Works



Anomalies (Outliers)

Have **fewer splits** to isolate
in the decision tree

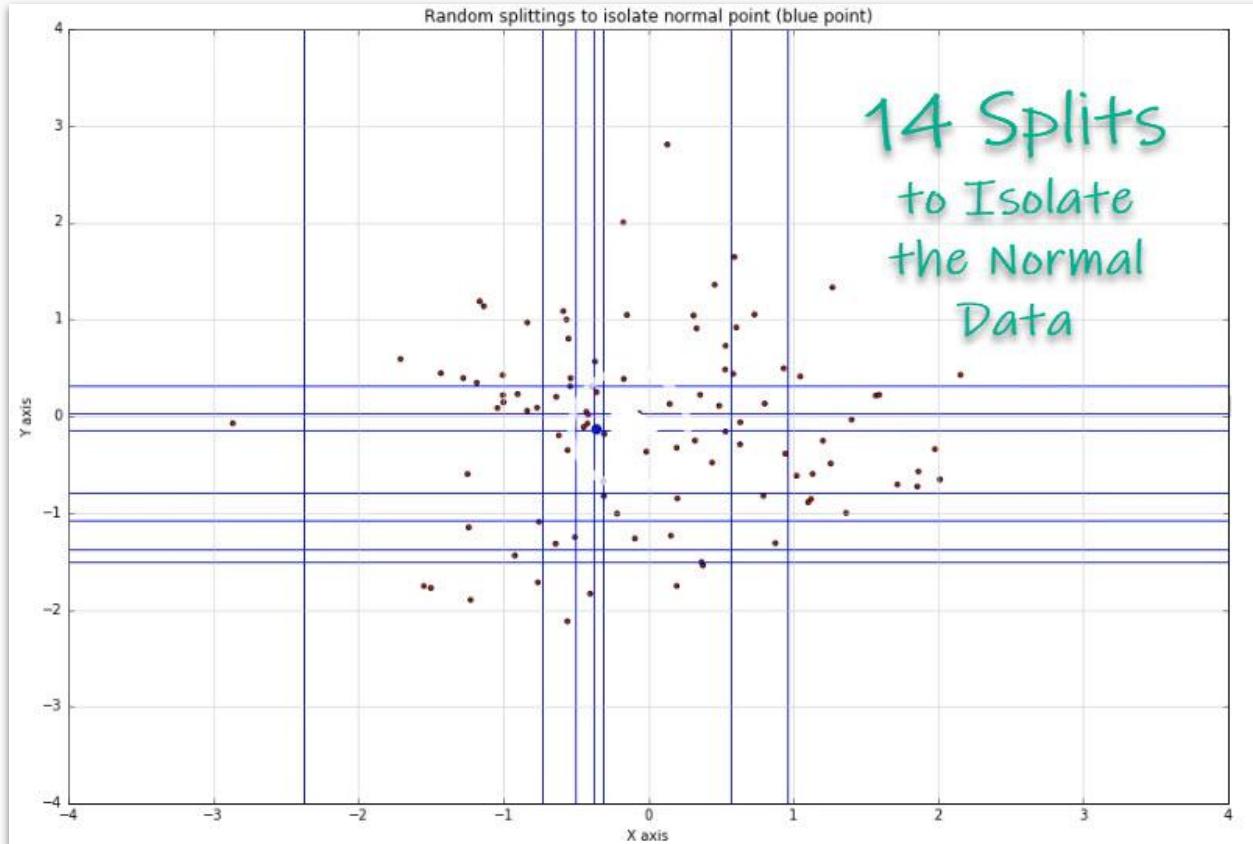


How Isolation Forest Works



Normal Data

Have many splits to isolate
in the decision tree



30-Min Demo

Detecting Fraud with Isolation
Forest

Isolation Forest

Secret Tactics & Pro Tips

Secret Tactics for

Getting Great Results

Use these tips to
increase your anomaly detection performance

Pro Tip #1

Run Algorithm Multiple Times & Average to Stabilize



Isolation Forest Algorithm

- Randomly assigns a single target
- If selects bad target, will get bad results

Prevent Bad Results

- Run multiple times
- Change Seed Parameter
- Average Results

Single Run

```
> predictions_tbl %>% pr_auc(class, predict)
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>        <dbl>
1 pr_auc  binary      0.933
```

Stabilized Performance after Averaging Multiple Runs

```
> stabilized_predictions_tbl %>% pr_auc(Class, mean_predict)
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>        <dbl>
1 pr_auc  binary      0.987
```

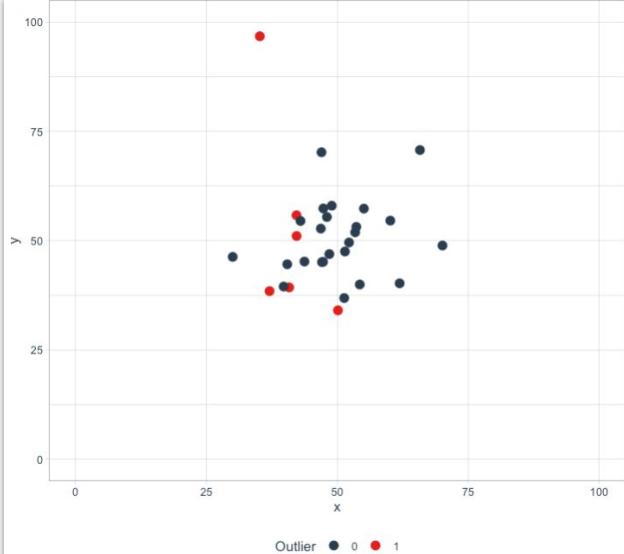
Pro Tip #2

Adjust Quantile / Threshold Based On Visualizing Outliers



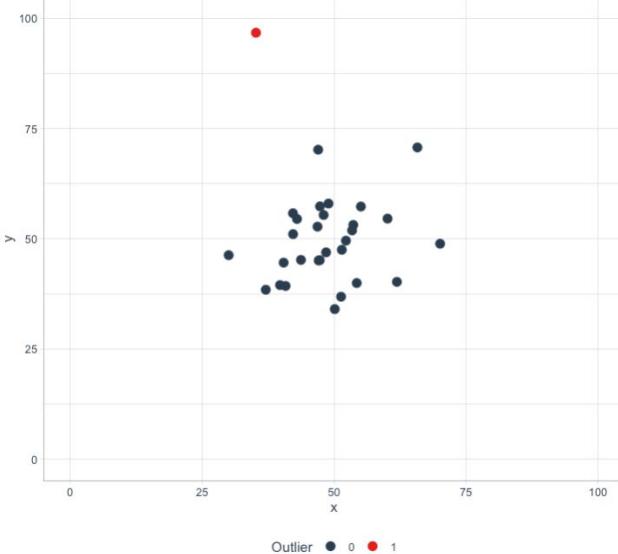
Too Low

```
50 quantile <- h2o.quantile(predictions, probs = 0.80)
51 quantile
```



Just Right

```
50 quantile <- h2o.quantile(predictions, probs = 0.99)
51 quantile
```

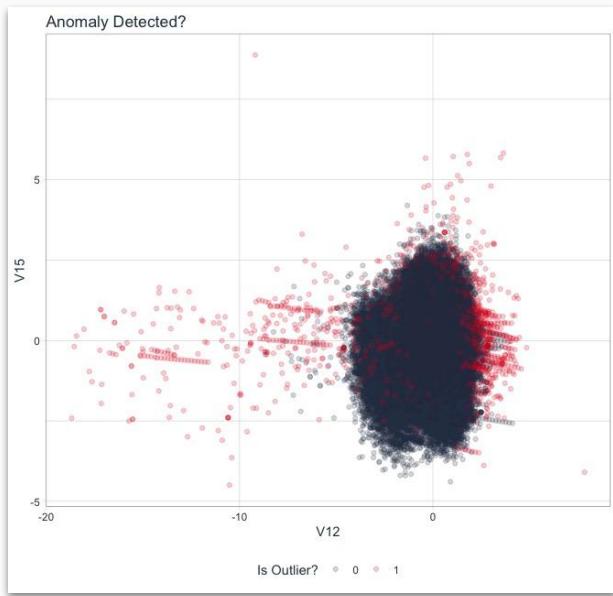




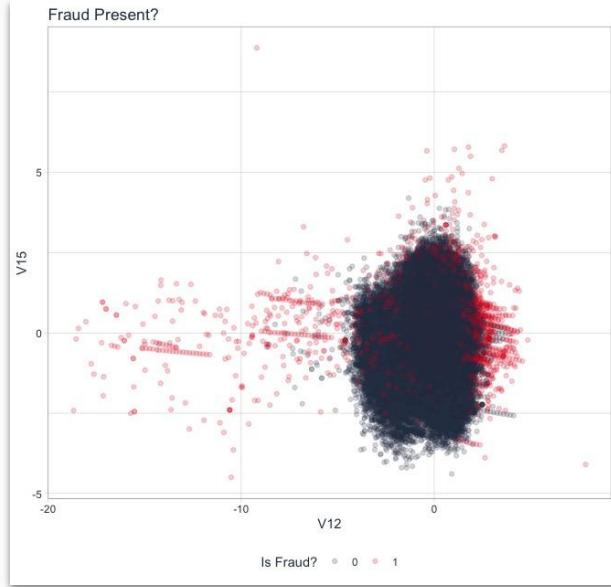
Pro Tip #3

Visualize to See What's Going On

Anomalies



Fraud

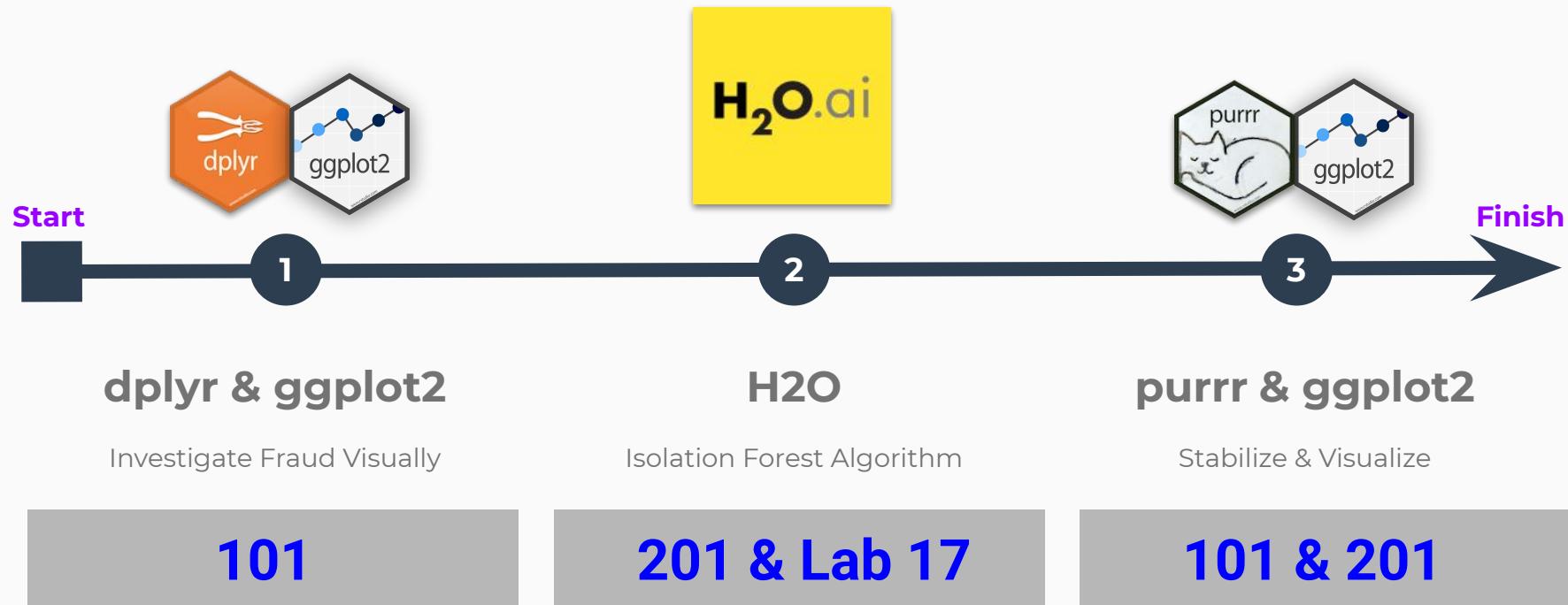


Data Science Transformational Learning Plan



Fraud Anomaly Detection

Step-By-Step





101 & 201

```
157 # 6.3 CALCULATE AVERAGE PREDICTIONS ----
158 stabilized_predictions_tbl <- multiple_predictions_tbl %>%
159   unnest(predictions) %>%
160   select(row, seed, predict) %>%
161
162   # Calculate stabilized predictions
163   group_by(row) %>%
164     summarize(mean_predict = mean(predict)) %>%
165   ungroup() %>%
166
167   # Combine with original data & important columns
168   bind_cols(
169     credit_card_tbl
170   ) %>%
171   select(row, mean_predict, Time, V12, V15, Amount, Class) %>%
172
173   # Detect Outliers
174   mutate(outlier = ifelse(mean_predict > quantile(mean_predict, probs = 0.99), 1, 0)) %>%
175   mutate(Class = as.factor(Class))
176
177 # 6.4 MEASURE ----
178 stabilized_predictions_tbl %>% pr_auc(Class, mean_predict)
179
```

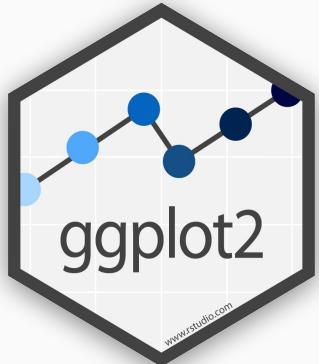


H₂O.ai

```
39 # 2.0 H2O ISOLATION FOREST ----
40 h2o.init()
41
42 credit_card_h2o <- as.h2o(credit_card_tbl)
43 credit_card_h2o
44
45 target <- "Class"
46 predictors <- setdiff(names(credit_card_h2o), target)
47
48 isoforest <- h2o.isolationForest(
49   training_frame = credit_card_h2o,
50   x      = predictors,
51   ntrees = 100,
52   seed    = 1234
53 )
54
55 isoforest
```

201
Lab 17

ggplot2 & purrr



```
24 * # 1.2 AMOUNT SPENT VS FRAUD ----
25   g <- credit_card_tbl %>%
26     select(Amount, Class) %>%
27     ggplot(aes(Amount, fill = as.factor(Class))) +
28     geom_histogram() +
29     # geom_density(alpha = 0.3) +
30     facet_wrap(~ Class, scales = "free_y", ncol = 1) +
31     scale_x_log10(label = scales::dollar_format()) +
32     scale_fill_tq() +
33     theme_tq() +
34     labs(title = "Fraud by Amount Spent",
35          fill = "Fraud")
36
37 ggplotly(g)
--
```

101 + 201



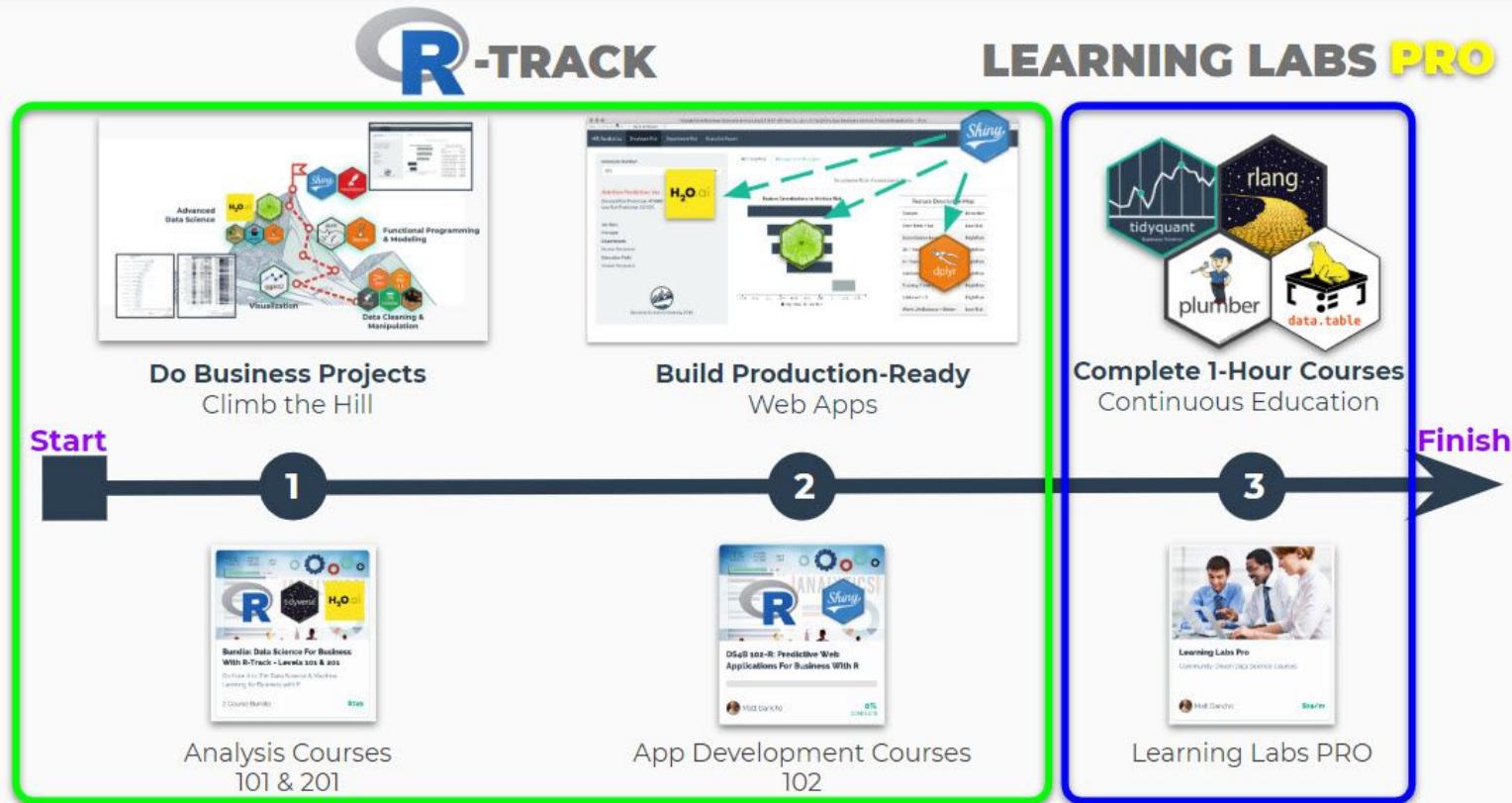
```
150
151 * # 6.2 MAP TO MULTIPLE SEEDS ----
152   multiple_predictions_tbl <- tibble(seed = c(158, 8546, 4593)) %>%
153     mutate(predictions = map(seed, iso_forest))
154
155   multiple_predictions_tbl
156
```

101 + 201

Business Science University

Our program that will TRANSFORM YOU in weeks, not years.

The program that will deliver YOUR Transformation



Everything is **Taken Care of** For You in Our Platform



3-Course R-Track System



Business Analysis with R (DS4B 101-R)

Data Science For Business with R (DS4B 201-R)

R Shiny Web Apps For Business (DS4B 102-R)

Project-Based Courses with Business Application

Data Science Foundations
7 Weeks



DS4B 101-R: Business Analysis With R

Your Data Science Journey Starts Now! Learn the fundamentals of data science for business with the tidyverse.



Machine Learning & Business Consulting
10 Weeks



DS4B 201-R: Data Science For Business With R

Solve a real-world churn problem with H2O AutoML (automated machine learning) & LIME black-box model explanations using R



Web Application Development
4 Weeks



DS4B 102-R: Shiny Web Applications For Business (Level 1)

Build a predictive web application using Shiny, Flexdashboard, and XGBoost

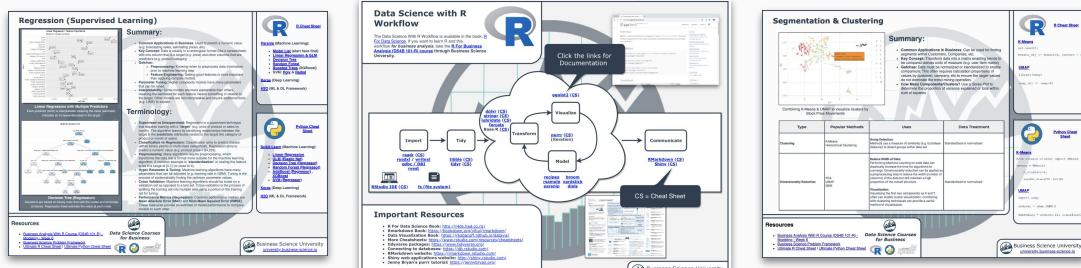


Key Benefits

- Fundamentals - Weeks 1-5 (25 hours of Video Lessons)
 - Data Manipulation (dplyr)
 - Time series (lubridate)
 - Text (stringr)
 - Categorical (forcats)
 - Visualization (ggplot2)
 - Programming & Iteration (purrr)
 - 3 Challenges
- **Machine Learning - Week 6 (8 hours of Video Lessons)**
 - Clustering (3 hours)
 - Regression (5 hours)
 - 2 Challenges
- Learn Business Reporting - Week 7
 - RMarkdown & plotly
 - 2 Project Reports:
 1. Product Pricing Algo
 2. Customer Segmentation

Business Analysis with R (DS4B 101-R)

Data Science Foundations
7 Weeks



Key Benefits

End-to-End Churn Project

Understanding the Problem & Preparing Data - Weeks 1-4

- Project Setup & Framework
- Business Understanding / Sizing Problem
- Tidy Evaluation - rlang
- EDA - Exploring Data -GGally, skimr
- Data Preparation - recipes
- Correlation Analysis
- 3 Challenges

Machine Learning - Weeks 5, 6, 7

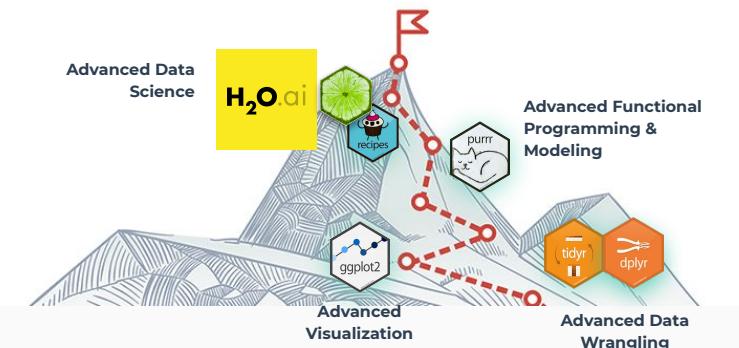
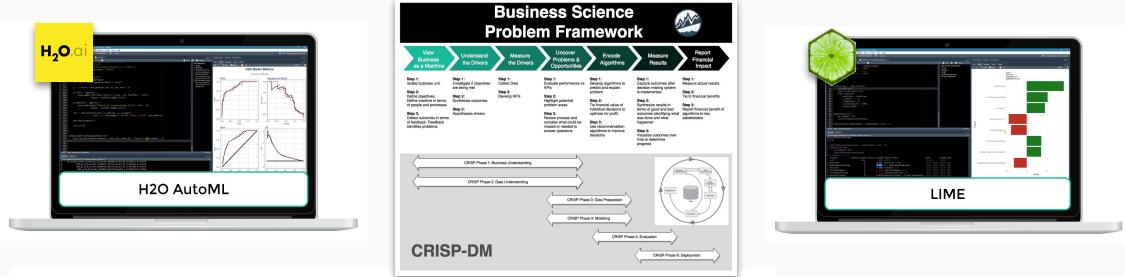
- H2O AutoML - Modeling Churn
- ML Performance
- LIME Feature Explanation

Return-On-Investment - Weeks 7, 8, 9

- Expected Value Framework
- Threshold Optimization
- Sensitivity Analysis
- Recommendation Algorithm

Data Science For Business (DS4B 201-R)

Machine Learning & Business Consulting
10 Weeks



Key Benefits

Learn Shiny & Flexdashboard

- Build Applications
- Learn Reactive Programming
- Integrate Machine Learning

App #1: Predictive Pricing App

- Model Product Portfolio
- XGBoost Pricing Prediction
- Generate new products instantly

App #2: Sales Dashboard with Demand Forecasting

- Model Demand History
- Segment Forecasts by Product & Customer
- XGBoost Time Series Forecast
- Generate new forecasts instantly

Shiny Apps for Business (DS4B 102-R)



Web Application Development
4 Weeks

The collage includes:

- A "Data Science with R" course page featuring a "Predictive Pricing App" dashboard.
- A "Flexdashboard Apps" section showing a dashboard with a map of the US and time series plots.
- A "Shiny Apps" section showing a dashboard with a scatter plot and a histogram.
- A "Themes, Dashboards, & Examples" section showing a dashboard with multiple panels and a sidebar.
- A "Business Analytics" section showing a dashboard with a map and a bar chart.
- A "Machine Learning" section showing a dashboard with a scatter plot and a sidebar.
- A "Data Science with R" course page featuring a "Sales Dashboard with Demand Forecasting" dashboard.



The collage includes:

- A "Shiny" logo and a bar chart.
- A "DATA ANALYTICS" section with a large blue "R" icon.
- A "Machine Learning" section with a green gear icon.
- A "Shiny" logo and a bar chart.
- A "DS4B 102-R: Shiny Web Applications for Business (Level 1)" course page.
- A "Build a predictive web application using Shiny, Flexdashboard, and XGBoost" section.
- A photo of Matt Dancho.



Testimonials



*“Your program allowed me to cut down to **50% of the time** to deliver solutions to my clients.”*

-Rodrigo Prado, Managing Partner Big Data Analytics & Strategy at Genesis Partners



*“I can already **apply** a lot of the early gains from the course to current working projects.”*

-Adam Mitchell, Data Analyst with Eurostar



*“My work became **10X easier**. I can spend quality time asking questions rather than wasting time trying to figure out syntax.”*

-Mohana Chittor, Data Scientist with Kabbage, Inc

Achieve
Results that
Matter to
the
Business

15% OFF PROMO Code: learninglabs



R-TRACK BUNDLE

R-TRACK BUNDLE

DS4B 101-R: Business Analysis With R
Your Data Science Journey Starts Now! Learn the fundamentals of data science for business with the tidyverse.

DS4B 201-R: Data Science For Business With R
Solve a real-world churn problem with H2O AutoML (automated machine learning) & LIME black-box model explanations using R

DS4B 102-R: Shiny Web Applications For Business (Level 1)
Build a predictive web application using Shiny, Flexdashboard, and XGBoost

Bundle - DS For Business + Web Apps (Level 1): R-Track - Courses 101, 102,

3 Course Bundle

0% COMPLETE

Get started now!

Payment Option	Description	Price	Action
<input checked="" type="radio"/>	Paid Course 15% COUPON DISCOUNT	\$149 \$976.65	Enroll
<input type="radio"/>	3 Monthly Payments 15% COUPON DISCOUNT 3X Monthly	3 payments of \$49/m	3 payments of \$381.65/m
<input type="radio"/>	6 Low Monthly Payments 15% COUPON DISCOUNT 6X Payment Plan	6 payments of \$24.99/m	6 payments of \$198.90/m
<input type="radio"/>	12 Low Monthly Payments 15% COUPON DISCOUNT 12X Plan	12 payments of \$12.50/m	12 payments of \$106.25/m

Begin Learning Today

university.business-science.io

