

Case: 1
Label: 1
Probability: 0.94
Explanation Fit: 0.00094

14.50 < path_id <= 16.75

8.67 < display <= 10.83

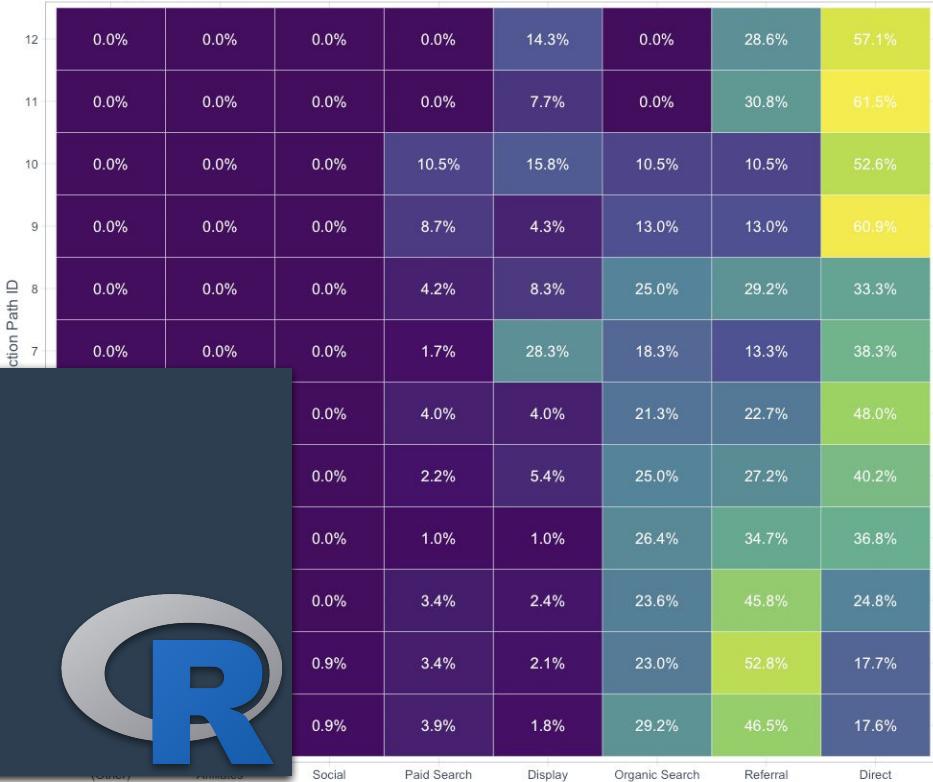
referral <= 4

social <= 9.92

Machine Learning for Customer Journey

Using Google Analytics Data

Components of Customer Journey By Purchase Number
As customers return, channel shifts from Referral and Organic to Direct, Display, and Paid Search



Matt Dancho & David Curry
Business Science Learning Lab





Learning Lab Structure

- **Presentation**
(20 min)
- **Demo's**
(30 min)
- **Pro-Tips**
(15 mins)



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 Optimize, David works with businesses to help improve website performance and SEO using data science. His passion is ethical Machine Learning initiatives.

Marketing Series

- **Lab 24 - A/B Testing**
 - Business Science's Website
 - Infer - Bootstrap & Permutation
- **Lab 25 - Multi-Channel Attribution (Part 1)**
 - Google Analytics Data
 - ChannelAttribution
- **Lab 26 - ML for Customer Journey (Part 2)**
 - Path Splitting
 - Applied ML for Conversion Probabilities
- **Lab 27 - Automated Prediction & Tracking Google Trends**
 - Google Trend Forecasting
 - gtrendsR, forecast
 - chronR, taskscheduleR



Learning Labs Pro
Community-Driven Data Science Courses
 Matt Dancho \$19/m



Learning Labs PRO

Every 2-Weeks

1-Hour Course

Recordings + Code + Slack

\$19/month

university.business-science.io

Lab 25 - Marketing Series
Attribution with ChannelAttribution

Lab 24 - Marketing Series
A/B Testing with Infer

Lab 23 - SQL Series
SQL with BigQuery & Conversion Funnel

Lab 22 - SQL Series
SQL for Time Series

Lab 21 - SQL Series
SQL for Data Science

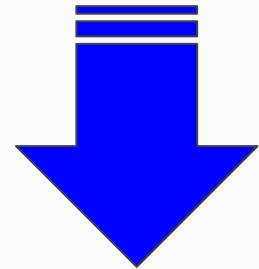
Lab 20 - Machine Learning
Explainable Machine Learning

Lab 19 - Network Analysis
Using Customer Credit Card History for
Networks Analysis

Lab 18 - Anomaly Detection
Time Series Anomaly Detection with
anomalize



Continuous Learning
Advanced Topics



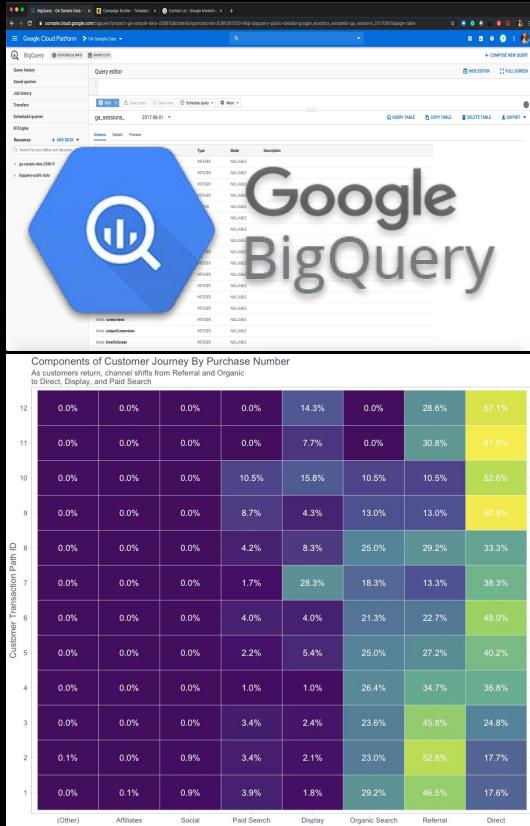
Learning Labs Pro

Community-Driven Data Science Courses

 Matt Dancho

\$19/m

Agenda



The screenshot shows the Google BigQuery web interface. At the top, there's a navigation bar with tabs like 'BigQuery', 'Data Catalog', 'Machine Learning', and 'Google Sheets'. Below the navigation is a search bar and a 'Query editor' section. The main area displays a table titled 'Components of Customer Journey By Purchase Number'. The table has columns for 'Customer Transaction Path ID' (1 to 12), 'Affiliates', 'Social', 'Paid Search', 'Display', 'Organic Search', 'Referral', and 'Direct'. The data shows various percentages across these categories for each path ID.

Customer Transaction Path ID	Affiliates	Social	Paid Search	Display	Organic Search	Referral	Direct	
12	0.0%	0.0%	0.0%	0.0%	14.3%	0.0%	28.6%	57.1%
11	0.0%	0.0%	0.0%	0.0%	7.7%	0.0%	30.8%	61.5%
10	0.0%	0.0%	0.0%	10.5%	15.8%	10.5%	10.5%	52.6%
9	0.0%	0.0%	0.0%	8.7%	4.3%	13.0%	13.0%	60.9%
8	0.0%	0.0%	0.0%	4.2%	8.3%	25.0%	29.2%	33.3%
7	0.0%	0.0%	0.0%	1.7%	28.3%	18.3%	13.3%	38.3%
6	0.0%	0.0%	0.0%	4.0%	4.0%	21.3%	22.7%	48.0%
5	0.0%	0.0%	0.0%	2.2%	5.4%	26.0%	27.2%	40.2%
4	0.0%	0.0%	0.0%	1.0%	1.0%	26.4%	34.7%	36.8%
3	0.0%	0.0%	0.0%	3.4%	2.4%	23.6%	45.8%	24.8%
2	0.1%	0.0%	0.9%	3.4%	2.1%	23.0%	52.8%	17.7%
1	0.0%	0.1%	0.9%	3.9%	1.8%	29.2%	46.5%	17.6%
(Other)								

- **Business Case Study**
 - Google Merchandise Store
 - Understand how customers buy
 - New Customer
 - Repeat Customer
- **Tools & Process**
 - Path Splitting
 - Large Data
 - Predicting & Explaining
- **30-Min Demo**
 - dplyr
 - dtplyr
 - parsnip
 - lime
- **Pro-Tips & Learning Guide**



Google Merchandise Store

Business Case



Business Case

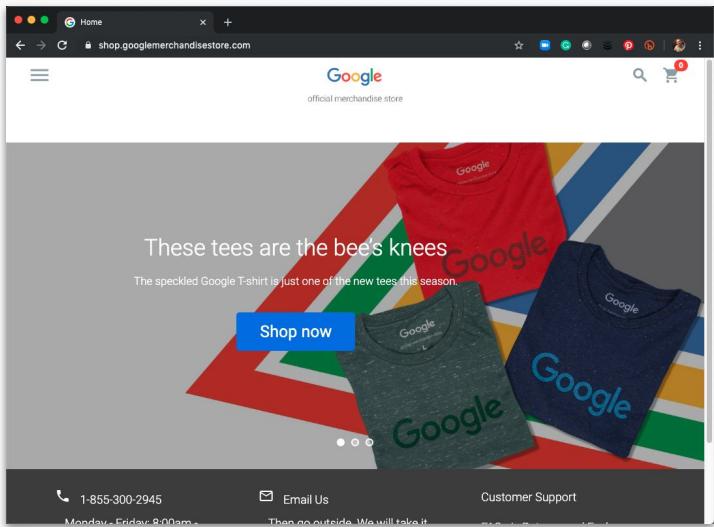
Google Merchandise Store

Customers can purchase t-shirts, gear, etc

Google Analytics tracks **channel sources** leading into the website.

We are interested in **Channels that lead to Transactions.**

Often called **Customer Journey**. We seek to understand this. Then make actions to increase conversion.



<https://shop.googlemerchandise.com/>

Business Case



```
> query_tbl
# A tibble: 847,224 x 6
  fullVisitorId      date channelGrouping traffic_source total_transactions total_transaction_revenue
  <chr>            <dbl> <chr>           <chr>           <dbl>                  <dbl>
1 4702386946621457676 20170103 Organic Search (direct)          NA                   NA
2 2087993472864421231 20170103 Organic Search (direct)          NA                   NA
3 0577469839995590230 20170103 Organic Search (direct)          NA                   NA
4 7618446014168949772 20170103 Paid Search   (direct)          NA                   NA
5 4837057017588527755 20170103 Paid Search   (direct)          NA                   NA
6 1555778933169102070 20170103 Direct        (direct)          NA                   NA
7 1116770359875973709 20170103 Organic Search (direct)          NA                   NA
8 5465941681642086837 20170103 Direct        (direct)          NA                   NA
9 2383851094930392840 20170103 Direct        (direct)          NA                   NA
10 6589579712030812313 20170103 Referral     sites.google.com    NA                   NA
# ... with 847,214 more rows
```

850K Rows

700K Visitor ID's

8 Channel Groups

Time Series

Customer Journey

Concepts & Business Goals

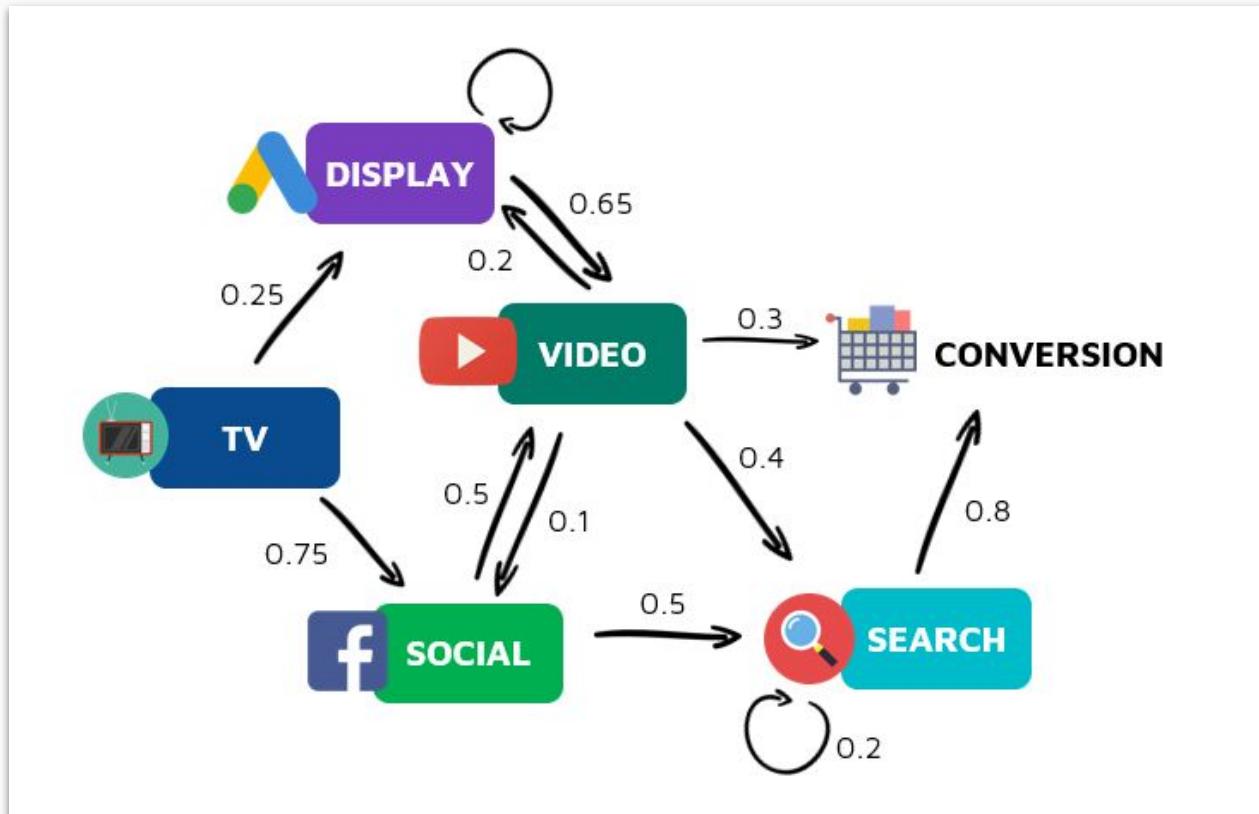


Touch Points

User interacts with media, website, referral, social, and search.

Tracked in **Google Analytics**:

- User ID
- Session ID
- Channel Group



Customer Journey



Components of Customer Journey By Purchase Number

As customers return, channel shifts from Referral and Organic to Direct, Display, and Paid Search

Customer Transaction Path ID	(Other)	Affiliates	Social	Paid Search	Display	Organic Search	Referral	Direct
12	0.0%	0.0%	0.0%	0.0%	14.3%	0.0%	28.6%	57.1%
11	0.0%	0.0%	0.0%	0.0%	7.7%	0.0%	30.8%	61.5%
10	0.0%	0.0%	0.0%	10.5%	15.8%	10.5%	10.5%	52.6%
9	0.0%	0.0%	0.0%	8.7%	4.3%	13.0%	13.0%	60.9%
8	0.0%	0.0%	0.0%	4.2%	8.3%	25.0%	29.2%	33.3%
7	0.0%	0.0%	0.0%	1.7%	28.3%	18.3%	13.3%	38.3%
6	0.0%	0.0%	0.0%	4.0%	4.0%	21.3%	22.7%	48.0%
5	0.0%	0.0%	0.0%	2.2%	5.4%	25.0%	27.2%	40.2%
4	0.0%	0.0%	0.0%	1.0%	1.0%	26.4%	34.7%	36.8%
3	0.0%	0.0%	0.0%	3.4%	2.4%	23.6%	45.8%	24.8%
2	0.1%	0.0%	0.9%	3.4%	2.1%	23.0%	52.8%	17.7%
1	0.0%	0.1%	0.9%	3.9%	1.8%	29.2%	46.5%	17.6%

Transaction Path Patterns

Each transaction is part of a path.

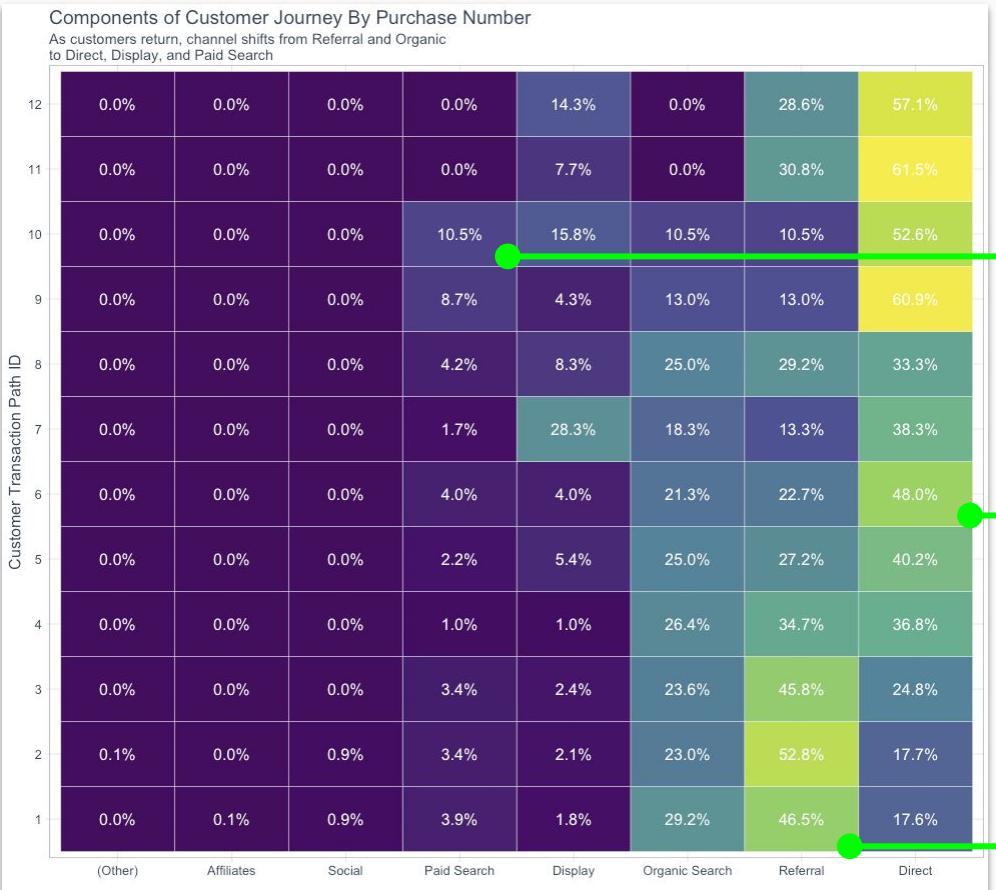
New Customers have only one transaction path.

Returning Customers have multiple transaction paths.

Patterns for New vs Returning are different.



Customer Journey Composition



Path 10

Customers making their **10th purchase** come composed of **Direct** and **Paid Search / Display (Adwords)**

Path 6

Customers making their **6th purchase** comes through **Direct (Email)**

Path 1

Customers that are making their **first purchase** come through **Referral (Hyperlinks on websites)** & **Organic Search (Googling)**.



Components of Customer Journey By Purchase Number

As customers return, channel shifts from Referral and Organic to Direct, Display, and Paid Search



Create an actionable marketing plan

Path 10

Customers making their **10th purchase** come composed of Direct and Paid Search / Display

Path 6

Customers making their **6th purchase** comes through Direct (Email)

Customers that are making their **first purchase** come through Referral (Hyperlinks on websites) & Organic Search (Googling).

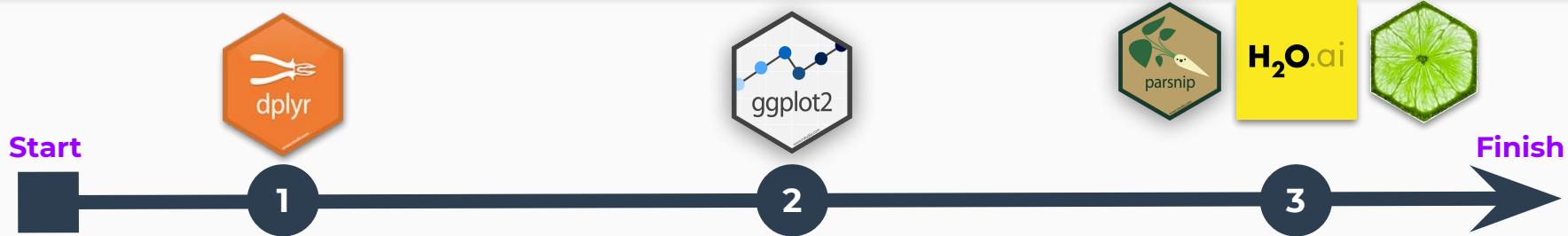
Tools & Process

Cutting-edge



Customer Journey Workflow

Step-By-Step



Large Data ETL

Biggest Challenge

Transformation to split paths
Time Series Data
Grouped Lag & Cumulative Sums
using **dtplyr**

Channel Path Visualization

Most important step
Can make strategies.

Problem

Strategies don't incorporate probability.

Machine Learning

Model all conversion paths with tools
like **Parsnip (XGBoost) & H2O**.
Paths scored using probability.
Explain a collection of paths using **LIME**.

Path Splitting



Customer Channel Event History

Path Splitting

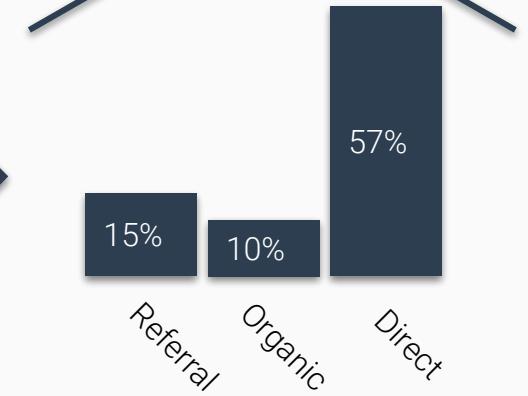
Transaction 1



Transaction 2



Trans.
10





Path Splitting

Path splitting with dplyr

800K+ unique sessions

700K+ unique users (groups)

Tools (learn these)

- Big Data (dtplyr)
- ETL (dplyr)



Data Skills Needed

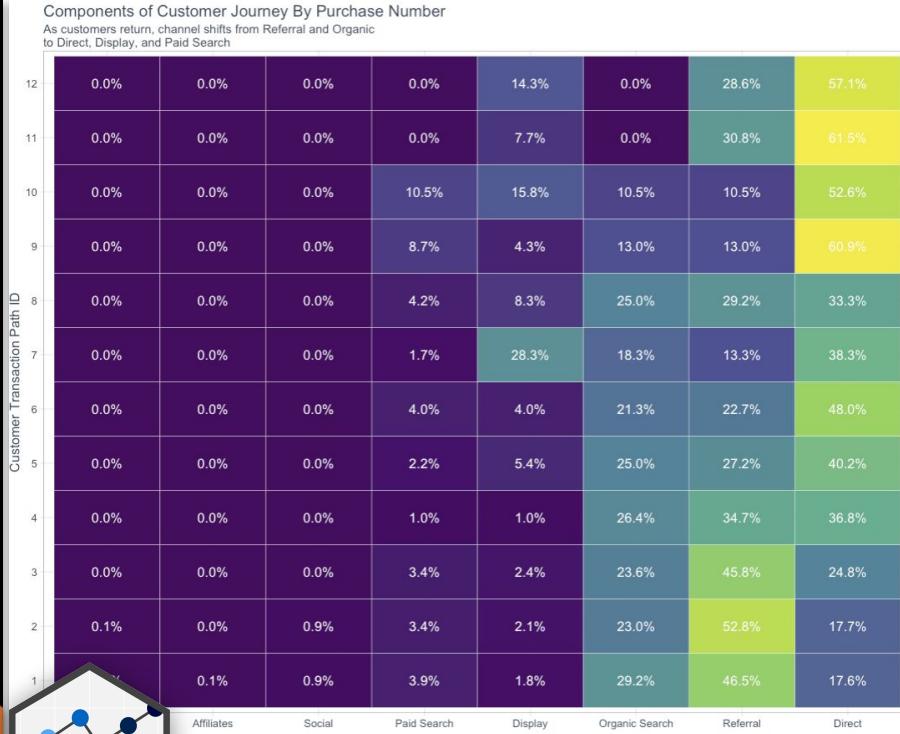
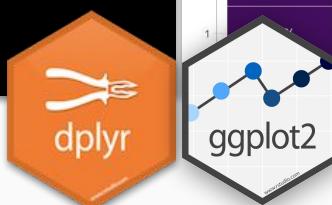
- Grouped calculations
- Time series - lag(), cumsum(), last()
- Started at 2-min per calc, **reduced to 17-sec**

```
106 query_paths_split_dt <- query_dt %>%
107
108     # Format data
109     mutate(date = ymd(date)) %>%
110     select(fullVisitorId, date, channelGrouping, total_transactions) %>%
111
112     # Create flag if transaction > 0
113     mutate(flag = case_when(
114         total_transactions > 0 ~ 1,
115         TRUE ~ 0
116     )) %>%
117
118     # Group by full Visitor Id
119     group_by(fullVisitorId) %>%
120
121     # Order by date
122     arrange(date) %>%
123
124     # Add paths by group
125     mutate(paths = cumsum(flag)) %>%
126     mutate(path_id = lag(paths, n = 1)) %>%
127
128     ungroup() %>% # End Group by Full Visitor ID
129
130     # Fill path_id NA's
131     mutate(path_id = ifelse(is.na(path_id), 0, path_id)) %>%
132
133     # Move from 0-base to 1-base index
134     mutate(path_id = path_id + 1) %>%
135
136     # Create visitor-path ids
137     mutate(visitor_path_id = str_c(fullVisitorId, "_", path_id)) %>%
138
139     # Group By All Unique Visitor-Paths - Flag if full path has a transaction
140     group_by(visitor_path_id) %>%
141     mutate(transaction_flag = ifelse(last(flag) == 1, 1, 0)) %>%
142     ungroup()
```



Visualize Customer Journey Composition

```
156 customer_journey_tbl <- query_paths_split_tbl %>%
157   select(visitor_path_id, transaction_flag, path_id, channelGrouping) %>%
158
159   # Only evaluate successful paths
160   filter(transaction_flag == 1) %>%
161
162   # Count successful channels by path id
163   group_by(channelGrouping, path_id) %>%
164   summarize(count = n()) %>%
165   ungroup() %>%
166
167   # Calculate successful channel composition by path id
168   group_by(path_id) %>%
169   mutate(prop = count / sum(count)) %>%
170   ungroup() %>%
171
172   mutate(channelGrouping = as_factor(channelGrouping)) %>% fct_reorder(prop)) %>%
173   ungroup() %>%
174
175   # Filter out paths > 12 (few people have more than 12 transactions in 1 year)
176   select(-count) %>%
177   filter(path_id <= 12) %>%
178   mutate(path_id = as.factor(path_id)) %>%
179
180   # Trick to fill in missing observations with zeros
181   pivot_wider(names_from = channelGrouping,
182             values_from = prop,
183             values_fill = list(prop = 0)) %>%
184   pivot_longer(cols = `Other`:`Social`, names_to = "channelGrouping", values_to = "prop") %>%
185
186   mutate(channelGrouping = as_factor(channelGrouping)) %>% fct_reorder(prop))
187
188 customer_journey_tbl
189
```



Machine Learning Model



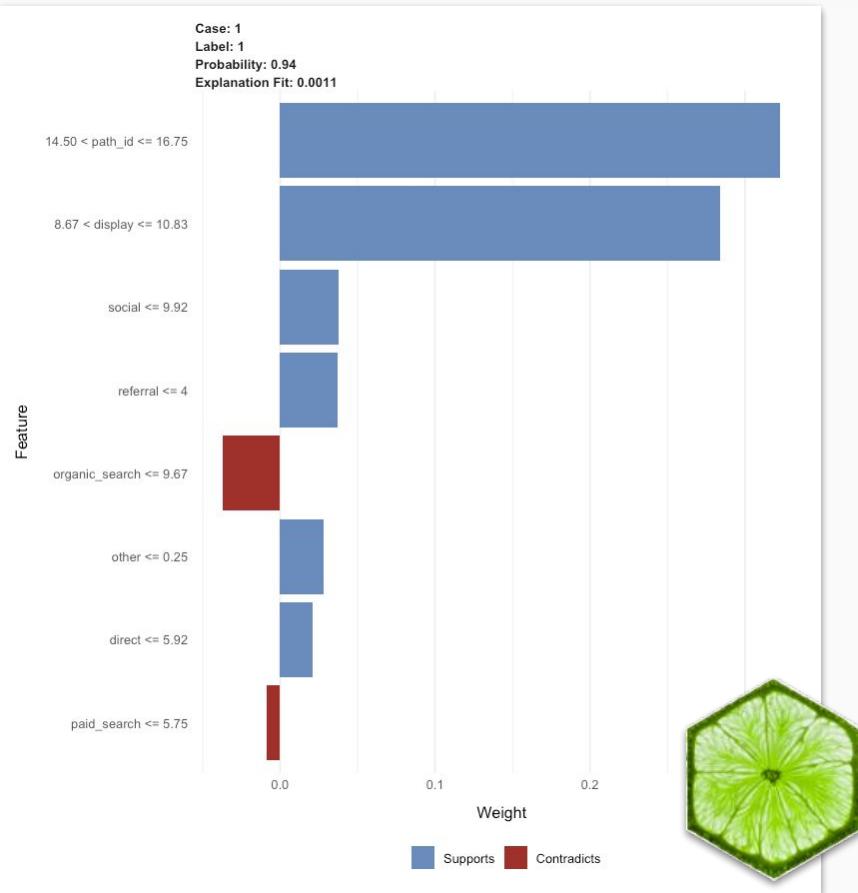
```
250# 4.2 Modeling ----  
251# WARNING - LONG RUNNING SCRIPT - TAKES 7 MINUTES TO RUN ----  
252# > toc()  
253# 428.923 sec elapsed  
254  
255tic()  
256model_xgboost <- boost_tree(mode = "classification",  
257  mtry = 12,  
258  trees = 1000,  
259  min_n = 3,  
260  tree_depth = 5,  
261  learn_rate = 0.01,  
262  loss_reduction = 0.01) %>%  
263  set_engine("xgboost") %>%  
264  fit.model_spec(transaction_flag ~ ., data = data_aggregated_tbl %>% select(-visitor_path_id))  
265toc()
```

```
> predictions_tbl %>% arrange(desc(.pred_1))  
# A tibble: 719,113 x 13  
  .pred_0 .pred_1 visitor_path_id transaction_flag path_id other affiliates direct display organic_search paid_search referral social  
    <dbl>   <dbl> <chr>           <fct>        <dbl>   <dbl>     <dbl>   <dbl>     <dbl>        <dbl>      <dbl>   <dbl>     <dbl>  
1 0.0624  0.938 1957458976293878100... 0             15     0     0     0     9     0     0     0     0     0     0  
2 0.0701  0.930 1957458976293878100... 1             12     0     0     0     2     0     0     0     0     0     0  
3 0.0713  0.929 1957458976293878100... 1             13     0     0     0     5     0     0     0     0     0     0  
4 0.0750  0.925 1957458976293878100... 1             10     0     0     0     3     0     0     0     0     0     0  
5 0.0808  0.919 7311242886083854158... 1             12     0     0     5     0     0     0     0     0     0     0  
6 0.0811  0.919 2402527199731150932... 1             15     0     0     3     0     0     0     0     0     0     0  
7 0.0811  0.919 7813149961404844386... 1             13     0     0     3     0     0     0     0     0     0     0  
8 0.0811  0.919 7813149961404844386... 1             16     0     0     3     0     0     0     0     0     0     0  
9 0.0811  0.919 7813149961404844386... 1             17     0     0     3     0     0     0     0     0     0     0  
10 0.0811 0.919 7813149961404844386... 1            25     0     0     3     0     0     0     0     0     0     0  
# ... with 719,103 more rows
```



Gives us conversion probabilities

Explainable Machine Learning



Describe why ML Model concludes 94% Probability

Use **LIME Explainable Machine Learning** to interpret prediction for single observation.

- **94% Probability of Conversion**
- Customer has **purchase path count** is between 14.5 & 16.75
- **Display** links clicked between 8.7 and 10.8

Action: Send Email (Direct) to target person and gain conversion

30-Min Demo

Let's do this!

PRO-TIPS

Yeahhhhhh!

Pro-Tip #1: We need to cross-validate model



We did not do Cross Validation

This is where H2O comes in.

Run H2O AutoML overnight, come in the next day with **100+ models** that have been **5-Fold Cross Validated**.

Learn H2O AutoML in

- **Advanced ML & Business Consulting**
DS4B 201-R

```
> h2o.predict(h2o_model, newdata = as.h2o(credit_card_group_tbl)) %>%  
+   as_tibble()  
#=====  
# A tibble: 1,125 x 7  
#>   predict    p1     p2     p3     p4     p5 Other  
#>   <fct>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  
#> 1 other    0.0009704 0.0228  0      0      0.977  
#> 2 other    0.0232   0.000717 0.000553 0      0.00376 0.973  
#> 3 other    0       0.0009737 0.0238  0      0.000107 0.976  
#> 4 other    0.00643  0.0009724 0.000558 0.00343 0.000105 0.990  
#> 5 Other    0       0.0009720 0.000555 0      0.000104 1.000  
#> 6 3       0       0.0009704 0.909   0      0.000102 0.0909  
#> 7 3       0       0.0009761 0.995   0      0.000110 0.00491  
#> 8 1       0.984   0.0009735 0.000567 0.00349 0.000106 0.0127  
#> 9 Other    0.195   0.0009602 0.000464 0.00285 0.0009870 0.802  
#> 10 Other   0       0.0009737 0.000568 0      0      1.000  
# ... with 1,115 more rows
```

H2O AutoML

H₂O.ai

Pro-Tip #2: Explain Results Locally



Executives need strategies to target a single customer

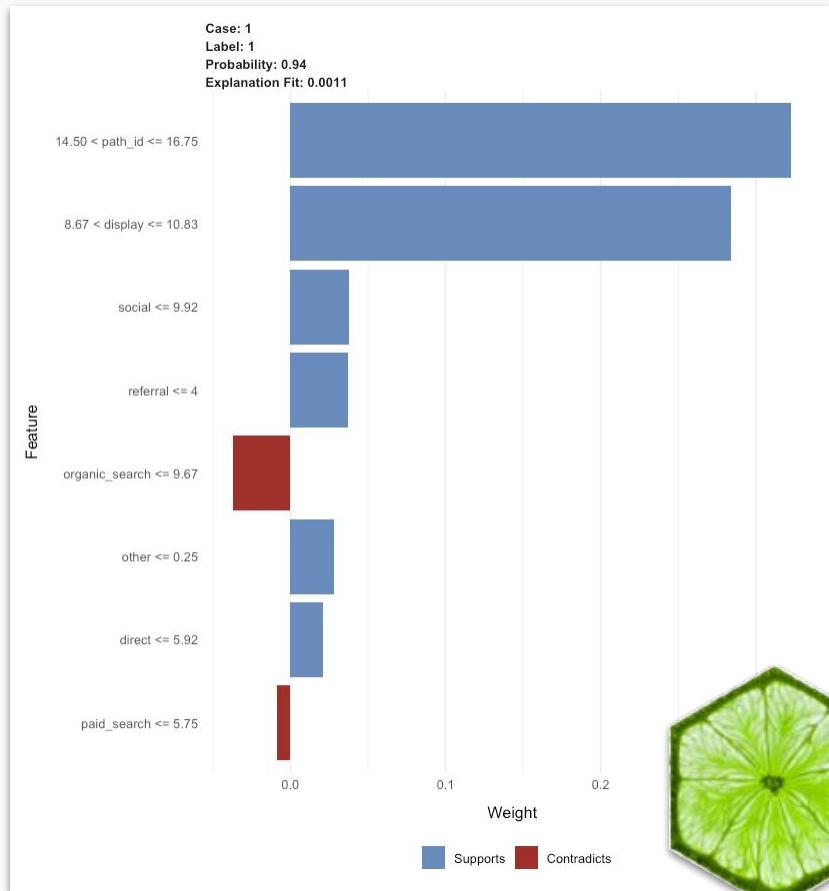
This is where **LIME** comes in.

Explain why an individual is high likelihood of converting.

Develop strategies your organization can use to target high probability customers and convert them.

Learn LIME in

- **Advanced ML & Business Consulting**
DS4B 201-R



Pro-Tip #3: Productionalize the Results



Businesses need apps

This is where Shiny comes in.

Make a shiny app to enable others to use your work.

Learn SHINY in:

- **Shiny Dashboards**
DS4B 102-R
- **Shiny Developer w/ AWS**
DS4B 202-R

The dashboard features a sidebar with icons for user management, dashboard creation, and help. The main content area has a header 'Employee Attrition Prevention' and a sub-header 'Select an employee, and learn how best to prevent their attrition.' A dropdown menu shows 'Denise Mooney, 71.8%'. Below this are two red cards: 'Attrition Risk 71.8%' and 'Prediction Leave'. To the right is a large image of two smiling business people. A yellow hexagon labeled 'H2O.ai' is overlaid on the image. Three blue arrows point from the bottom right towards the dashboard: one from a blue hexagon labeled 'dplyr', one from a white hexagon labeled 'ggplot2', and one from a red hexagon labeled 'Personal Development Recommendation Seek Mentorship Role'.

Employee Attrition Prevention

Select an employee, and learn how best to prevent their attrition.

Denise Mooney, 71.8%

Attrition Risk 71.8% Prediction Leave

Employee Card

Employee Name
Denise Mooney

Department
Sales

Job Role
Sales Executive

Cause & Prevention Recommendations

Overtime = Yes
StockOptionLevel = 0
4 < NumCompaniesWorked
JobLevel = 2
RelationshipSatisfaction = Low

Label: Yes
Probability: 0.72
Explanation Fit: 0.3

Supports Contracts

ggplot2

Personal Development Recommendation Seek Mentorship Role

Professional development Recommendation

4-Course R-Track System



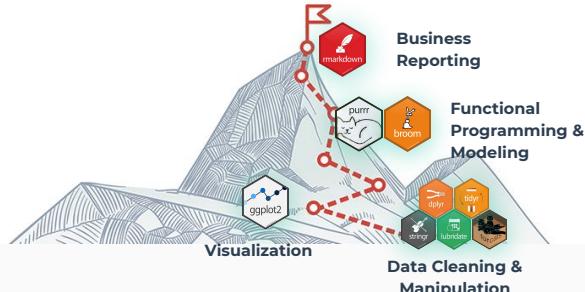
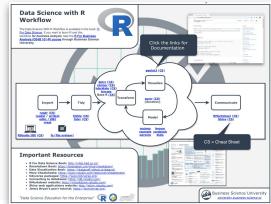
Business Analysis with R (DS4B 101-R)

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

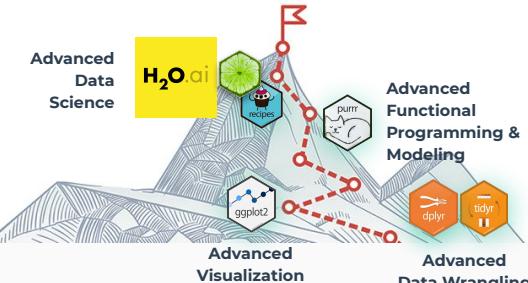
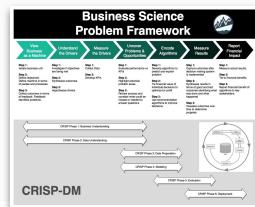
Web Apps & Shiny Developer (DS4B 102-R + DS4B 202A-R)

Project-Based Courses with Business Application

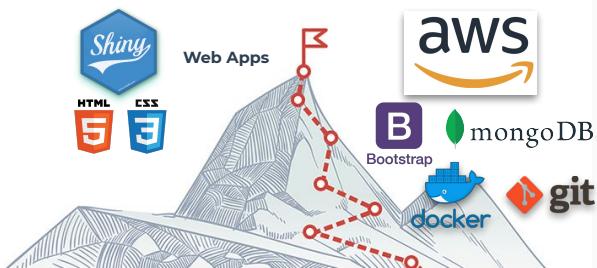
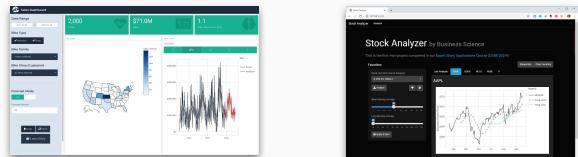
Data Science Foundations
7 Weeks



Machine Learning & Business Consulting
10 Weeks



Web Application Development
12 Weeks

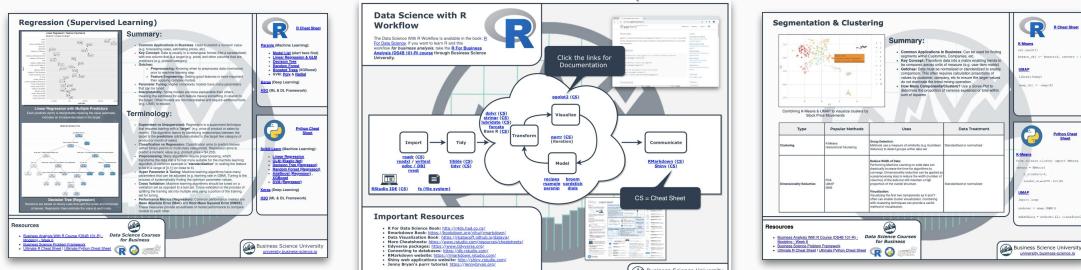


Key Benefits

- 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

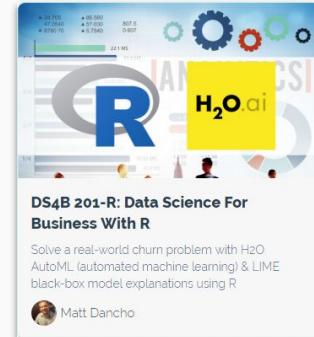
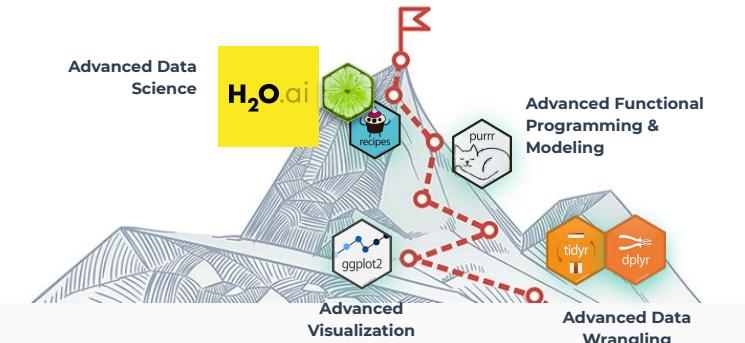
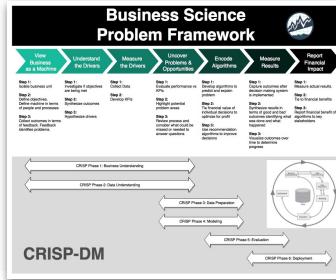
- H2O AutoML - Modeling Churn
- ML Performance
- LIME Feature Explanation

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 screenshot of a "Data Science with R" course page showing a flowchart from "Start" to "Publish" through "Components", "Advanced Forecasting", and "Testing". It highlights "Flexdashboard Apps" and "Shiny Apps".
- A "Sales Dashboard" showing a map of the US with data points, a bar chart for 2,000 units, and a line graph for \$71.0M.
- A "Demand Forecasting" dashboard with various charts and data tables.
- A "Shiny App" interface showing a histogram and other data visualizations.



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

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

Matt Dancho

Key Benefits

Frontend + Backend + Production Deployment

Frontend for Shiny

- Bootstrap

Backend for Shiny

- MongoDB
- Dynamic UI
- User Authentication
- Store & Write User Data

Production Deployment

- AWS
- EC2 Server
- VPC Connection
- URL Routing

Shiny Apps for Business (DS4B 202A-R)



Web Application Development
6 Weeks



15% OFF PROMO Code: **learninglabs**



R-TRACK BUNDLE

4-Course Bundle - Machine Learning + Expert Web Applications (R-Track)

Go from Beginner to Expert Data Scientist & Shiny Developer in Under 6-Months

4 Course Bundle ~~\$1,500~~

**\$127/mo
Limited Time**

DS4B 101-R: Business Analysis With R

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

Matt Dancho

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

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

Matt Dancho

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

DS4B 202A-R: Expert Shiny Developer with AWS

Learn how to build Scalable Data Science Applications using R, Shiny, and AWS Cloud Technology.

Matt Dancho

<input type="radio"/>	Paid Course 15% COUPON DISCOUNT	\$1,596 \$2,356.60
<input checked="" type="radio"/>	12 Low Monthly Payments 15% COUPON DISCOUNT	12 payments of \$149/m 12 payments of \$126.65/m

Begin Learning Today

university.business-science.io

