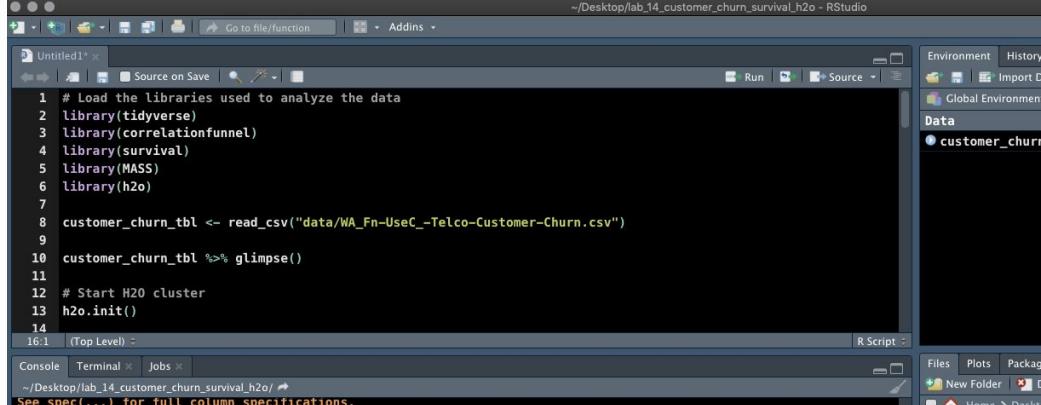


Cox Proportional Hazards Model Algorithm

H2O uses the Newton-Raphson algorithm to maximize the partial log-likelihood, an iterative procedure defined by the steps:

To add numeric stability to the model fitting calculations, the numeric predictors and offsets are demeaned during the model fitting process.

1. Set an initial value, $\beta^{(0)}$, for the coefficient vector and assume an initial log partial likelihood of $-\infty$.
2. Increment iteration counter, n , by 1.
3. Calculate the log partial likelihood, $pl(\beta^{(n)})$, at the current coefficient vector estimate.
4. Compare $pl(\beta^{(n)})$ to $pl(\beta^{(n-1)})$.



The screenshot shows an RStudio interface with an R script file open. The code loads libraries (tidyverse, correlationfunnel, survival, MASS, h2o), reads a CSV file named 'customer_churn_tbl' into memory, starts an H2O cluster, and initializes it. The R console shows the command 'h2o.init()' and the message '(Top Level) -'. The status bar indicates the script is located at 'Desktop/lab_14_customer_churn_survival_h2o.R'.

```
Untitled1* x
1 # Load the libraries used to analyze the data
2 library(tidyverse)
3 library(correlationfunnel)
4 library(survival)
5 library(MASS)
6 library(h2o)
7
8 customer_churn_tbl <- read_csv("data/WA_Fn-UseC-Telco-Customer-Churn.csv")
9
10 customer_churn_tbl %>% glimpse()
11
12 # Start H2O cluster
13 h2o.init()
14
16.1 (Top Level) -
```

Console Terminal Jobs

~/Desktop/lab_14_customer_churn_survival_h2o.R

See spec(...) for full column specifications.

Customer Churn Survival Analysis

With **correlationfunnel**, **parsnip** & **H2O**

Difficulty: **Intermediate**



$$LRE(x, y) = -\log_{10}(|x|), \text{ if } y = 0$$

```
$ Churn      <chr> "No", "No", "Yes", "No", "Yes", "Yes", "No", "No", "Yes", "No", "No", "No", "No", "No", "Yes"
> |
```

Matt Dancho & David Curry
Business Science Learning Lab #14





Learning Lab Structure

- **Presentation**

(20 min)

- **Demo's**

(20 min)

- **Presentation**

(20 mins)

Your Hosts!



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**.



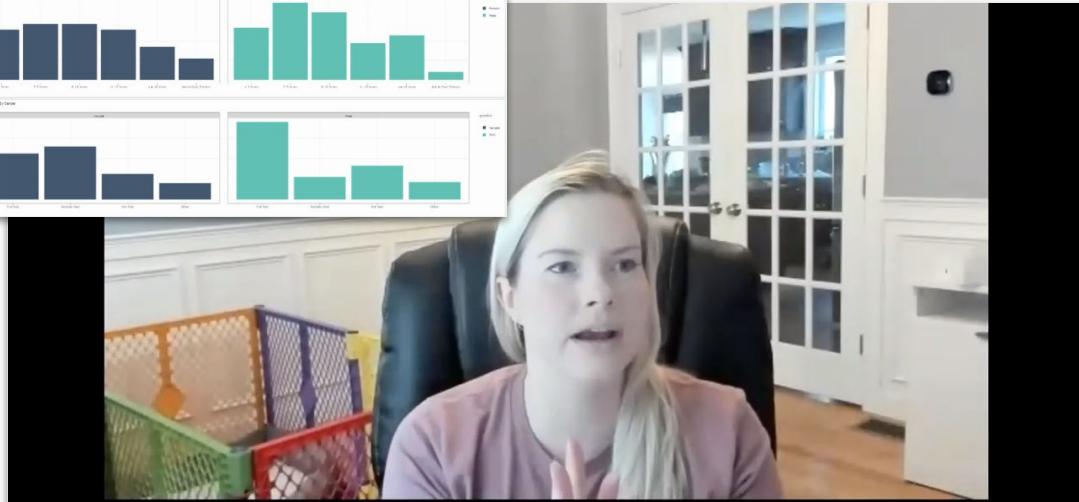
Success Story

Kristen Kehrer

- Founder, DataMovesMe
- Consultant & Keynote Speaker
- Built Shiny App **3-Days** after STARTING our 102 Course



"Now, all my clients are getting SHINY APPS!"



www.business-science.io

#BusinessScienceSuccess

A screenshot of the RStudio interface. On the left, a file browser shows a directory structure under 'mortege_loans_datatable'. It contains a folder named 'data' which holds four files: '.DS_Store', '2018Q1.zip', 'Acquisition_2018Q1.txt', and 'Performance_2018Q1.txt'. The 'Acquisition_2018Q1.txt' file is selected. A purple arrow points from the 'Values' section of the code editor down to the file in the file browser. The code editor on the right displays the following R code:

```
mortege_loans_datatable
Environment History Connections
Import Dataset > List >
Global Environment Data
Data_A 426207 obs. of 25 variables
Data_P 4645448 obs. of 31 variables
Values
  Acquisitions "data://Acquisition_2018Q1.txt"
  filelocation "data://"
  k 1
  Performance "data://Performance_2018Q1.txt"
  Performance_ColClass chr [1:31] "character" "character" "character" "numeric"
  Performance_Variables chr [1:31] "LOAN_ID" "Monthly.Rpt.Prd" "Servicer.Name" ...
Functions
na.lomf function (x)
Files Plots Packages Help Viewer
New Folder Delete Rename More
Home > Desktop > mortege_loans_datatable > data
Name Size Modified
.. 6 KB Jun 27, 2019, 11:13 AM
.DS_Store 56 MB Jun 27, 2019, 11:10 AM
2018Q1.zip 44.2 MB Jun 27, 2019, 11:15 AM
Acquisition_2018Q1.txt 375.5 MB Jun 27, 2019, 11:15 AM
```

● Business Case Study

- Customers Leaving
Telecommunications
Company

● 30-Min Demo

- correlationfunnel
- Survival Analysis
- H2O + LIME

● Process & Tools

● Survival Analysis

- 80/20 Concepts
- Methods & Pros/Cons
- Game Plan

● Tactics for Customer Retention

- What to look for
- What to learn to be able to implement at scale



Learning Labs PRO

Every 2 Weeks

Get Code

Recordings

Slack Community

\$19/month

university.business-science.io

Lab 13
**Wrangling 4.6M Rows w/
data.table**



Lab 12
How I built anomalize

Lab 11
**Market Basket Analysis w/
recommenderLab**

Lab 10
**Building API's with
plumber & postman**



Lab 9
**Finance in R with
tidyquant**

Learning Labs Pro

Community-Driven Data Science Courses



Matt Dancho

\$19/m

Business Case Study

Telecom Company that is losing
customers



Customers Leaving Telecommunications Company

Business Objectives

Customers are the lifeblood of subscription business.

Losing customers (churn) requires gaining new customers to replace a **10X more expensive** alternative than retaining existing.

Solution: Understand why & implement retention strategies.

Prediction + Explanation





Telecom Customer Data

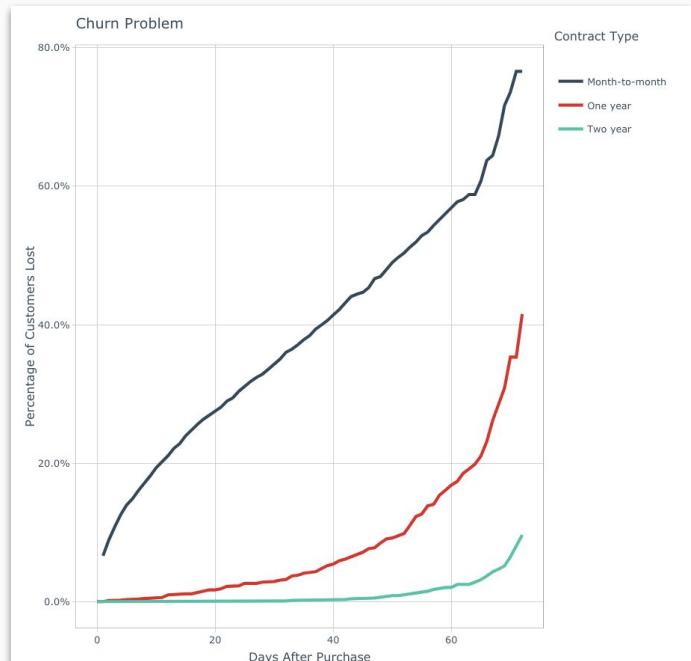
Customer Product History

Tenure (numeric)

- Number of months that the customer has been with the company
- Time series converted to duration

Churn (yes/no)

- Whether or not customer has cancelled their service





Feature List

7043 Customers

21 Features

- Time varying - Tenure
- Whether Canceled - Churn
- What services they had
- Customer Information
- Services Purchased
- Contract Information
 - Payment Method
 - Contract Type

Details

Telecom Customer Data:

- customerID (chr): CUSTOMER ID
- gender (chr): Customer's gender ("Female", "Male")
- SeniorCitizen (dbl): 1 = Senior Citizen, 0 = Not Senior Citizen
- Partner (chr): Whether the customer has a partner or not (Yes, No)
- Dependents (chr): Whether the customer has dependents or not (Yes, No)
- tenure (dbl): Number of months the customer has stayed with the company
- PhoneService (chr): Whether the customer has a phone service or not (Yes, No)
- MultipleLines (chr): Whether the customer has multiple lines or not (Yes, No, No phone service)
- InternetService (chr): Customer's internet service provider (DSL, Fiber optic, No)
- OnlineSecurity (chr): Whether the customer has online security or not (Yes, No, No internet service)
- OnlineBackup (chr): Whether the customer has online backup or not (Yes, No, No internet service)
- DeviceProtection (chr): Whether the customer has device protection or not (Yes, No, No internet service)
- TechSupport (chr): Whether the customer has tech support or not (Yes, No, No internet service)
- StreamingTV (chr): Whether the customer has streaming TV or not (Yes, No, No internet service)
- StreamingMovies (chr): Whether the customer has streaming movies or not (Yes, No, No internet service)
- Contract (chr): The contract term of the customer (Month-to-month, One year, Two year)
- PaperlessBilling (chr): Whether the customer has paperless billing or not (Yes, No)
- PaymentMethod (chr): The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
- MonthlyCharges (dbl): The amount charged to the customer monthly
- TotalCharges (dbl): The total amount charged to the customer
- Churn (chr): Outcome. Whether the customer churned or not (Yes or No)

Process & Tools

Churn Modeling & Machine Learning Tools

Churn Modeling Process

Step-By-Step



dplyr

Format Data

For EDA with **correlationfunnel**

parsnip

Modeling

Also need **survival** (Survival Curves)

h2o & lime

Automated Machine Learning

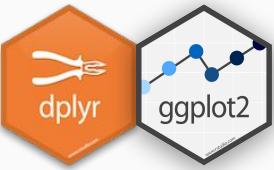
Local Feature Explanation

(50+ Models in seconds)



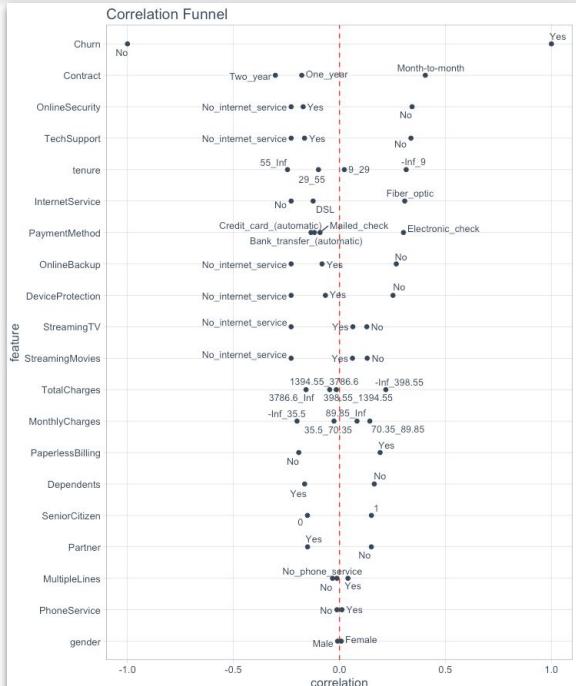
Tools Needed

dplyr & ggplot2



Data preparation is critical for correlation funnel

- Missing values
- Data imbalance





Tools Needed

correlationfunnel

- Speeds Up Exploratory Data Analysis
- Improves Feature Selection
- Gets You To Business Insights Faster

The screenshot shows the GitHub page for the `correlationfunnel` package. The page has a header with the package name and version (0.0.3), navigation links for Home, Function Reference, Articles, and News. Below the header is a section titled "correlationfunnel" with a subtitle "Speed Up Exploratory Data Analysis (EDA)". It includes a brief description of the package's goal: "The goal of `correlationfunnel` is to speed up Exploratory Data Analysis (EDA). Here's how to use it." A code snippet shows how to install it from GitHub: `devtools::install_github("business-science/correlationfunnel")`. The page then discusses the "Correlation Funnel in 2-Minutes", explaining the problem of time-consuming feature-target relationship analysis and the solution provided by the package. It lists "Main Benefits" and provides an example from a "Bank Marketing Campaign". On the right side, there are sections for "Links" (source code at <https://github.com/business-science/correlationfunnel>, report a bug at <https://github.com/business-science/correlationfunnel/issues>), "License" (MIT), and "Developers" (Matt Dancho).

<https://business-science.github.io/correlationfunnel/>



Tools Needed

parsnip



Like scikit-learn for R

Included:

- **survival regression (surv_reg)**

Not Included:

- **Survival Curves**
- **CoxPH**

https://tidymodels.github.io/parsnip/reference/surv_reg.html

A screenshot of a web browser showing a table titled "List of Models" from the parsnip package. The table compares various modeling functions across different model types: nnet, glmnet, multinom_reg(), keras, spark, nearest_neighbor(), null_model(), randomForest, rand_forest(), ranger, spark, flexsurv, survreg, svm_poly(), and svm_rbf(). The columns represent model types: Basic Usage (✓), Model List (✓), Articles (✓), News (✗), and Reference (✗). A note at the bottom states: "Models can be added by the user too. See the 'Making a parsnip model from scratch' vignette."

	part of tidymodels	Basic Usage	Model List	Articles	News	Reference
nnet	✓ ✓	✗	✓	✗		
glmnet	✓ ✓	✗	✗	✗		
multinom_reg()	keras	✓ ✓	✗	✗		
spark	✓ ✓	✗	✗	✗		
nearest_neighbor()	kknn	✓ ✓	✗	✓	✗	
null_model()	parsnip	✓ ✓	✗	✓	✗	
	randomForest	✓ ✓	✗	✓	✗	
rand_forest()	ranger	✓ ✓	✓	✓	✗	
	spark	✓ ✓	✗	✓	✗	
	flexsurv	✗ ✗	✗	✓	✗	
surv_reg()	survreg	✗ ✗	✗	✓	✗	
	svm_poly()	kernlab	✓ ✓	✗	✓	✗
	svm_rbf()	kernlab	✓ ✓	✗	✓	✗



Tools Needed

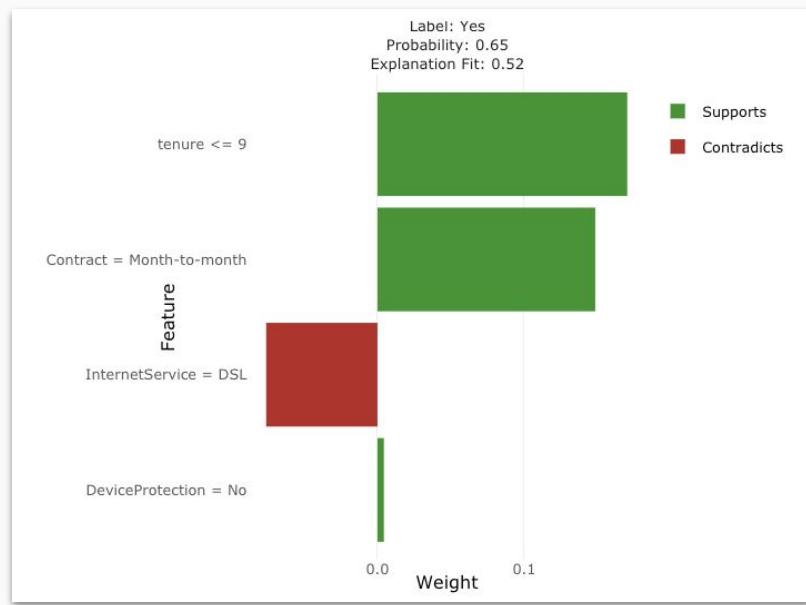
H2O & LIME



Automatic Machine Learning

Predicts Churn Risk (%)

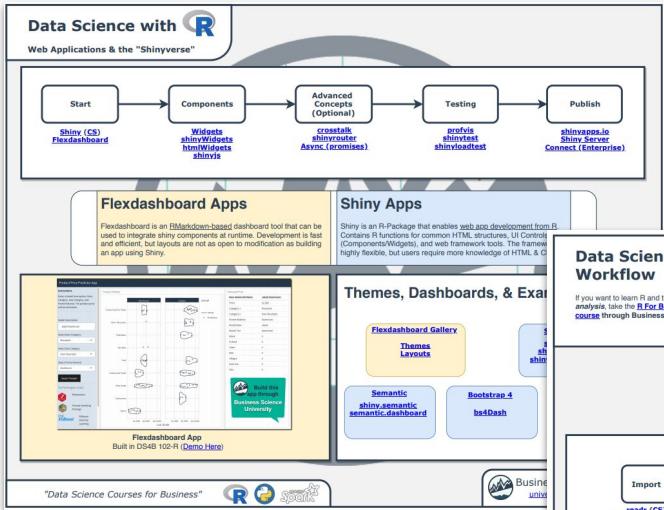
Tells what features contribute to **the person of interest** leaving



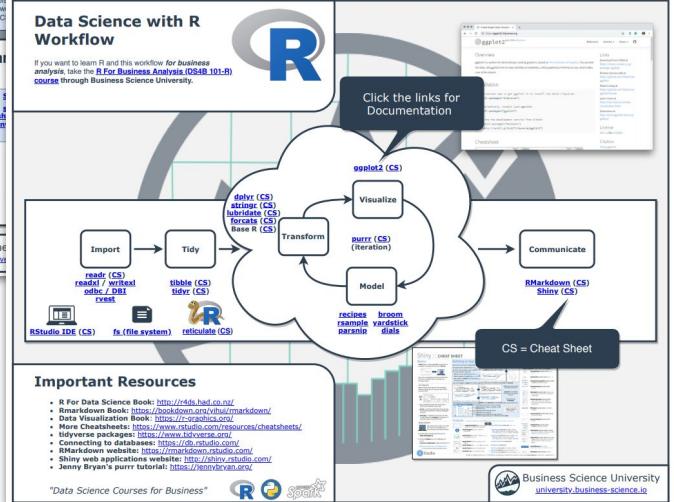
Links To The Resources



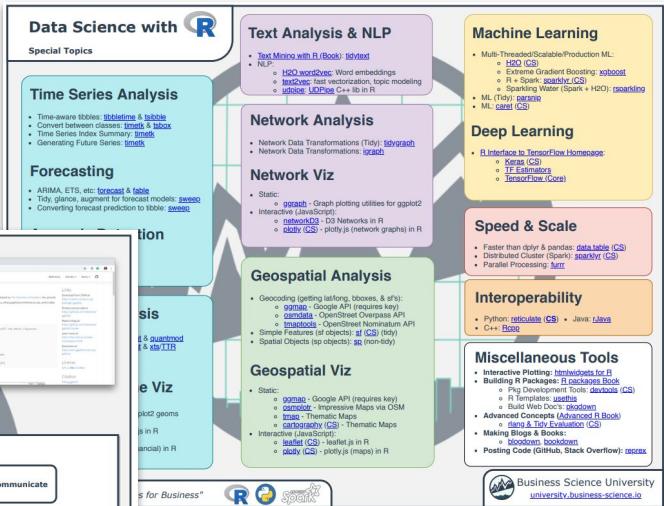
Page 2



Page 1



Page 3



Survival Analysis

80/20 Concepts

Mortality Table

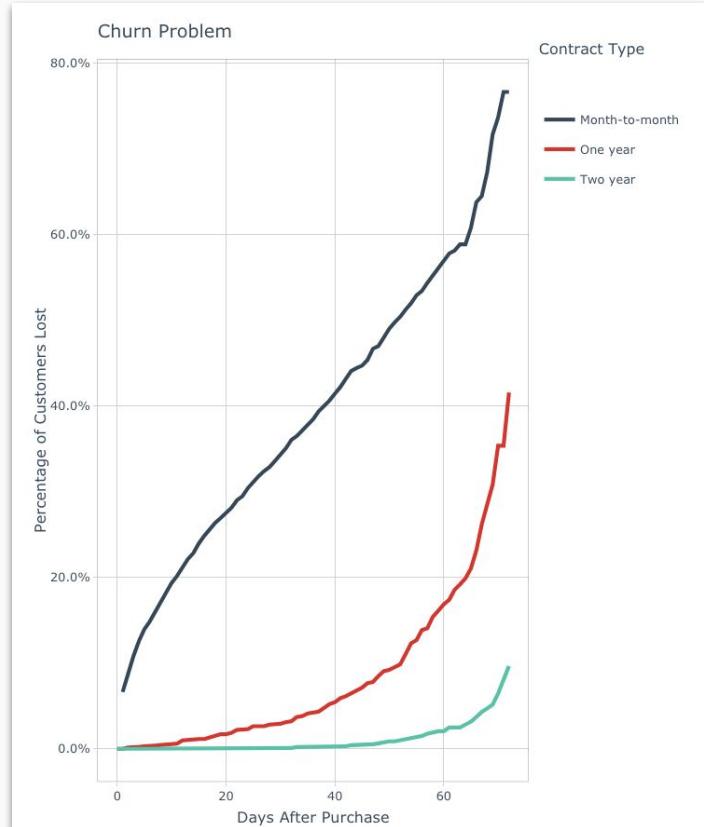
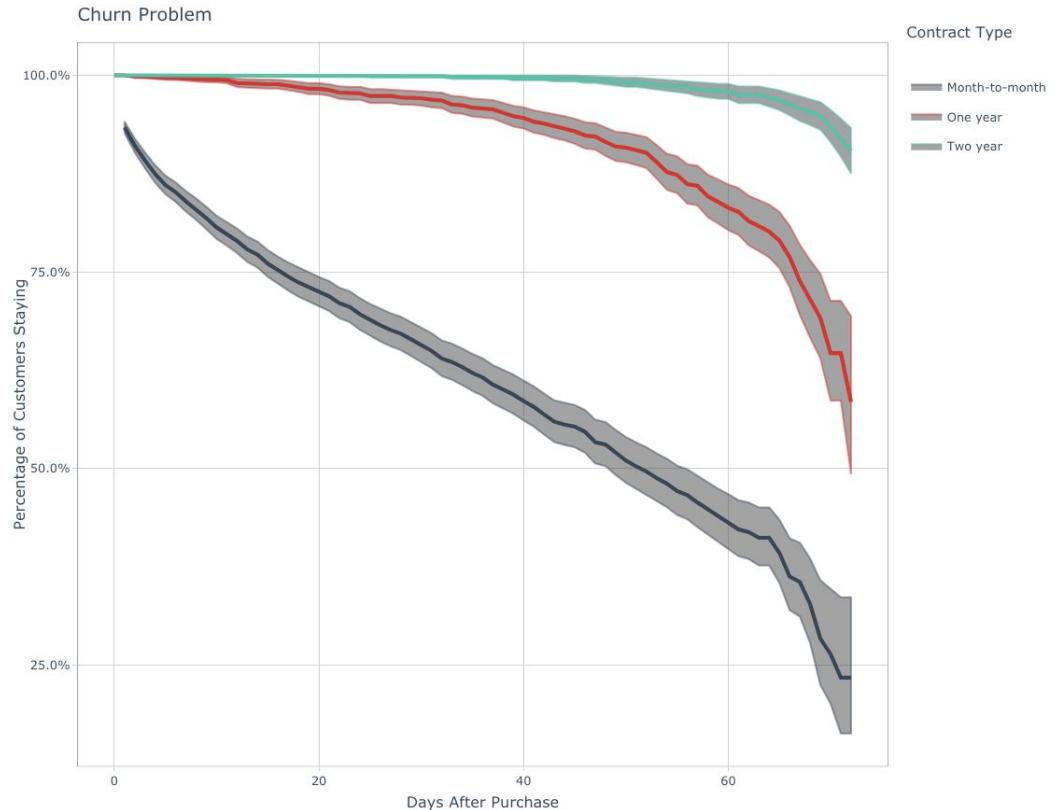


- Based on Time
- How long until churn

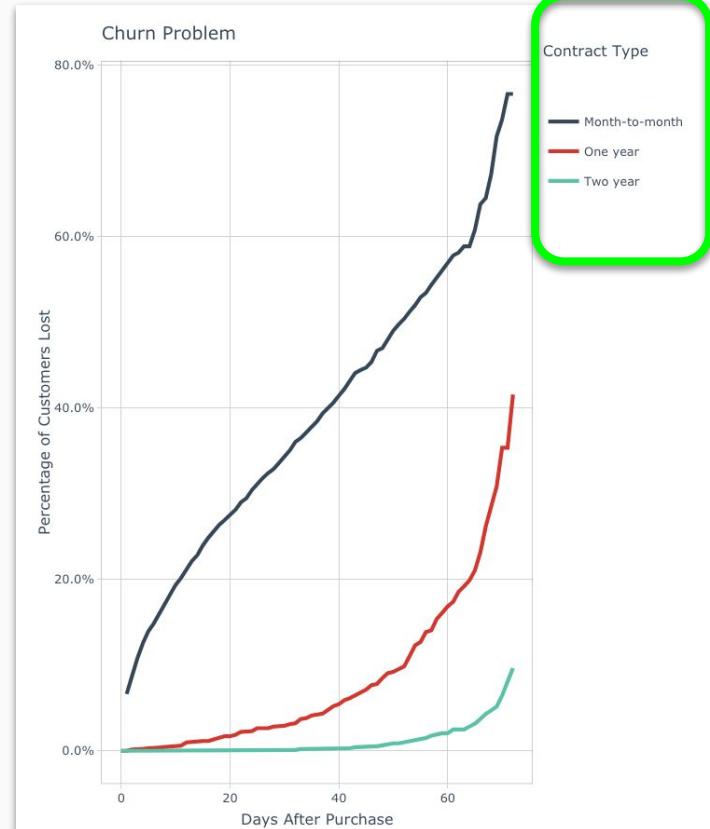
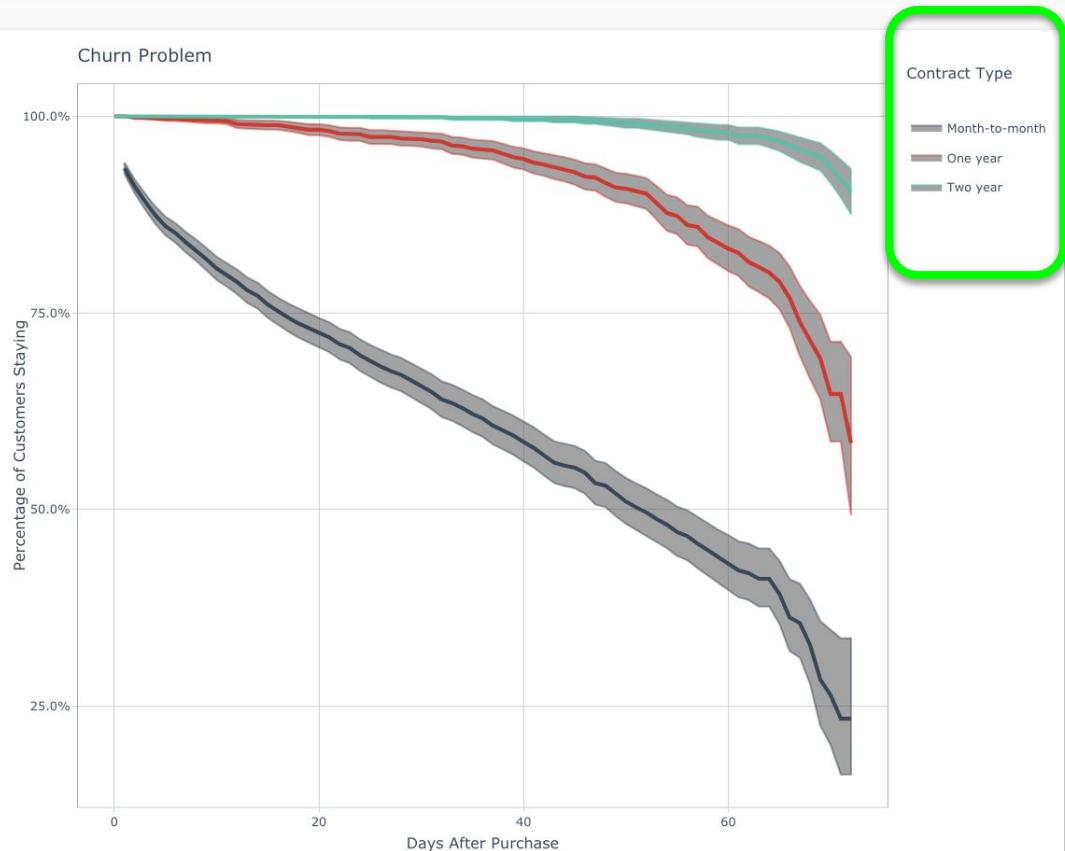
```
# A tibble: 218 x 9
  time n.risk n.event n.censor estimate std.error conf.high conf.low strata
  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
1     1   3875    380     224   0.934  0.00405   0.941   0.926 Month-to-month
2     2   3271    121     109   0.911  0.00500   0.920   0.902 Month-to-month
3     3   3041     94      97   0.892  0.00578   0.902   0.882 Month-to-month
4     4   2850     82      83   0.874  0.00649   0.886   0.863 Month-to-month
5     5   2685     63      65   0.860  0.00706   0.872   0.849 Month-to-month
6     6   2557     40      55   0.851  0.00744   0.864   0.839 Month-to-month
7     7   2462     50      63   0.840  0.00793   0.853   0.827 Month-to-month
8     8   2349     41      60   0.830  0.00836   0.843   0.816 Month-to-month
9     9   2248     45      60   0.818  0.00885   0.833   0.804 Month-to-month
10   10   2143     45      51   0.807  0.00936   0.821   0.792 Month-to-month
# ... with 208 more rows
```



Survival Curves



Strata = Group of Interest



Survival Analysis

Methods + Pros & Cons



Modeling Churn

Strengths

Survival Curves - Communication Tool
that **Business Leaders Understand**

CoxPH & Survival Regression - Used to
incorporate **multivariate** analysis

Weaknesses

CoxPH & Survival Regression - Not as
high performance as **Machine Learning**

Use **Machine Learning** to model true risk



Kaplan-Meier Method

Pros

Simple Method

Cons

Simple Method (**univariate**)

Does not account for other variables
in the data

Only time, churn, and strata

The survival probability at time t_i , $S(t_i)$, is calculated as follow:

$$S(t_i) = S(t_{i-1})(1 - \frac{d_i}{n_i})$$

Where,

- $S(t_{i-1})$ = the probability of being alive at t_{i-1}
- n_i = the number of patients alive just before t_i
- d_i = the number of events at t_i
- $t_0 = 0, S(0) = 1$



Cox Proportional Hazard

Pros

Multivariate!

Super easy to get Survival Curves

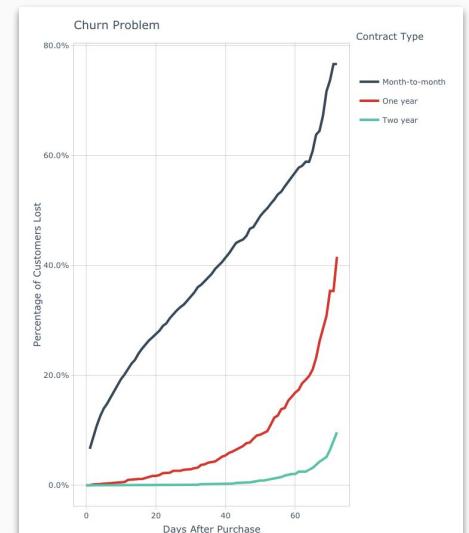
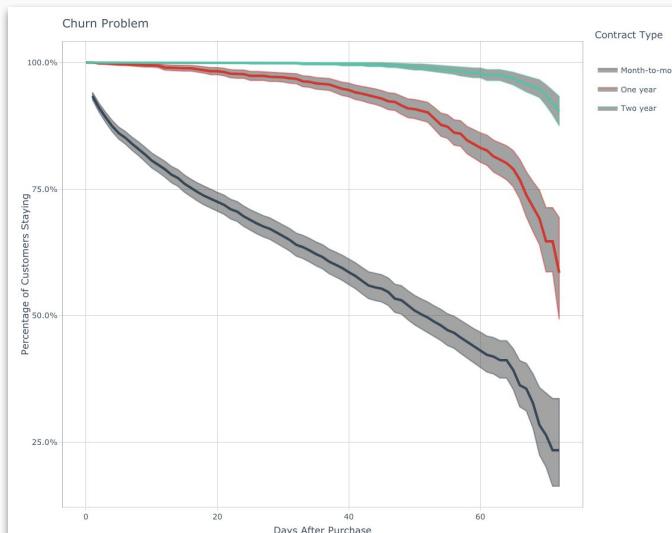
Can predict churn

Cons

Predictive Accuracy

Underlying model assumes covariates are not time dependent

term	estimate	std.error	statistic	p.value	conf.low	conf.high
1 OnlineSecurity_NoTRUE	0.583	0.0609	9.71	2.75e-22	0.465	0.700
2 TechSupport_NoTRUE	0.316	0.0598	5.29	1.25e-7	0.199	0.434
3 InternetService_FiberOpticTRUE	0.0440	0.0537	0.820	4.12e-1	-0.0612	0.149
4 PaymentMethod_ElectronicCheckTRUE	0.379	0.0488	7.76	8.46e-15	0.283	0.475
5 DeviceProtection_NoTRUE	0.352	0.0506	6.94	3.82e-12	0.252	0.451



Survival Regression



Pros

Multivariate!

Parsnip provides convenient API

```
158 # 6.3 Survival Regression w/ ParSNIP
159 # 4.3 Using ParSNIP ----
160 model_survreg <- parsnip::surv_reg(mode = "regression", dist = "t") %>%
161   set_engine("survreg") %>%
162   fit.model_spec(Surv(I(tenure + 1), Churn_Yes) ~ . - Contract + strata(Contract), data = train_tbl)
163
164 model_survreg$fit %>% survfit()
165
166 predict(model_survreg, new_data = train_tbl) %>%
167   bind_cols(train_tbl %>% select(Churn_Yes, everything()))
168
169
```

Cons

Predictive Accuracy

Underlying model assumes covariates are not time dependent

More difficult to obtain survival curves

```
> model_survreg$fit %>% tidy()
# A tibble: 9 x 7
  term                estimate std.error statistic p.value conf.low conf.high
  <chr>              <dbl>     <dbl>      <dbl>    <dbl>     <dbl>     <dbl>
1 (Intercept)        5.01      0.0653     76.7     0.        4.88      5.14
2 OnlineSecurity_NoTRUE -0.266    0.0367    -7.25    4.12e-13  -0.338    -0.194
3 TechSupport_NoTRUE -0.176    0.0331     -5.32    1.04e-7   -0.241    -0.111
4 InternetService_FiberOpticTRUE -0.0539   0.0309     -1.75    8.07e-2   -0.114    0.00658
5 PaymentMethod_ElectronicCheckTRUE -0.230    0.0331     -6.97    3.28e-12  -0.295    -0.165
6 DeviceProtection_NoTRUE -0.217    0.0329     -6.58    4.86e-11  -0.281    -0.152
7 Two year          -1.79      0.0867    -20.6     3.02e-94  NA        NA
8 Month-to-month      0.286    0.0189     15.2     5.45e-52  NA        NA
9 One year           -1.07      0.0523    -20.5     8.96e-94  NA        NA
> |
```

Solution to Accuracy: Machine Learning

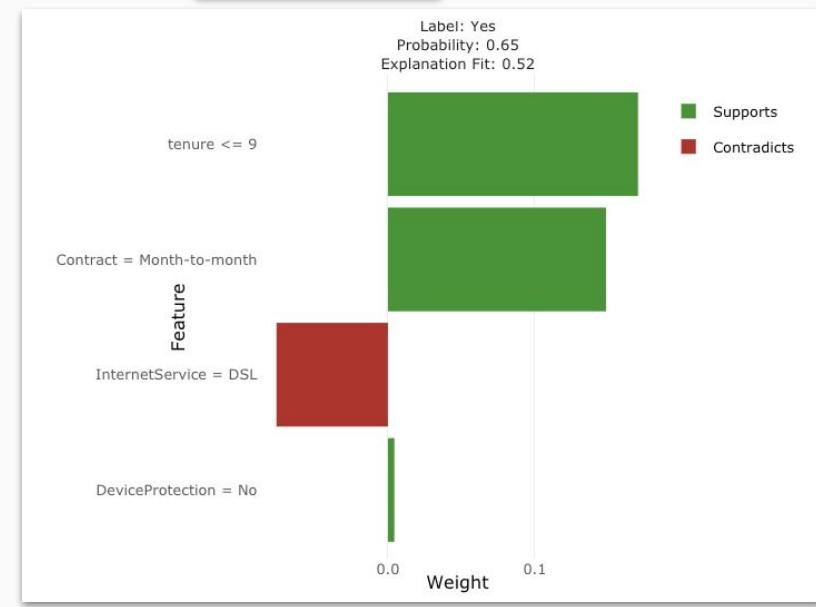


Pros

Multivariate!
High Accuracy
Explainability
By Person/Observation

Cons

No time-varying survival curves



Game Plan

Using Survival Analysis & ML

Feature Selection



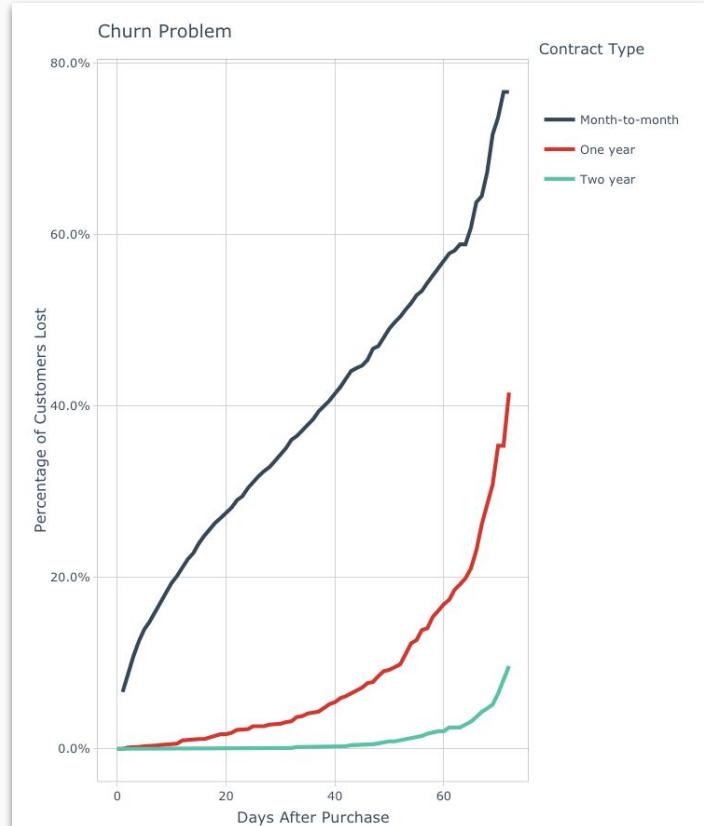
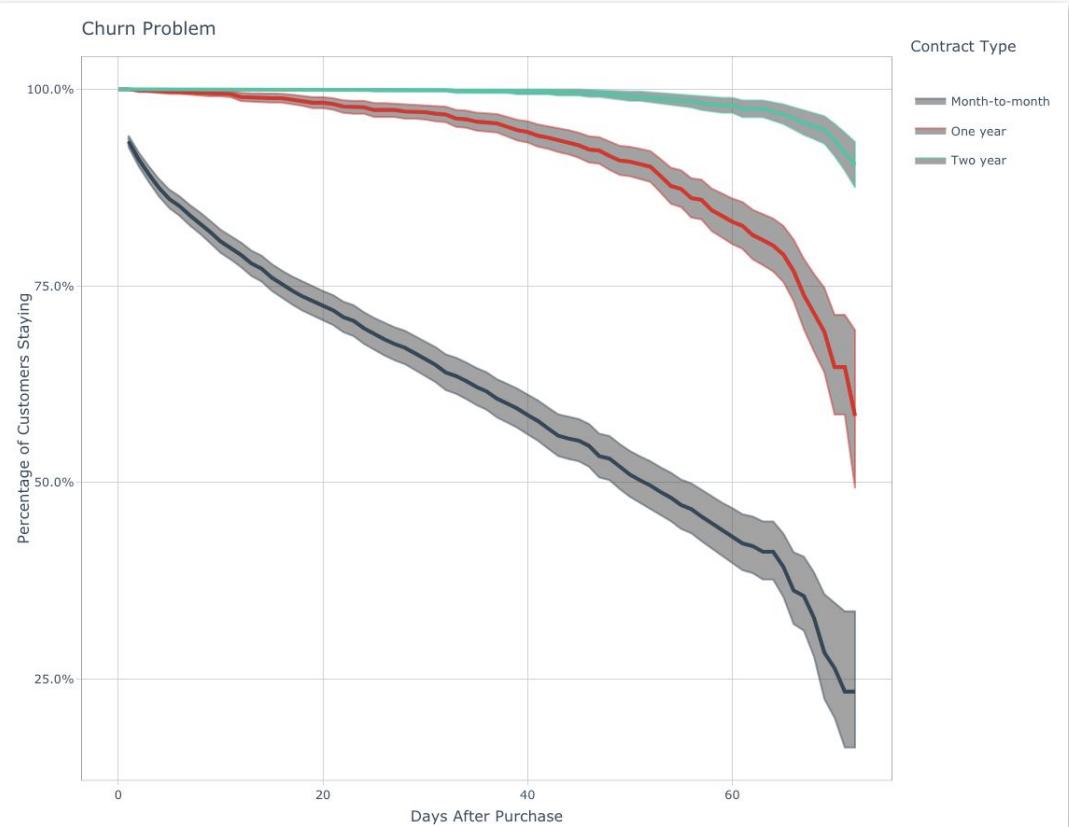
correlationfunnel

- Speeds Up Exploratory Data Analysis
- Improves Feature Selection
- Gets You To Business Insights Faster

The screenshot shows the GitHub project page for `correlationfunnel`. The page includes the following sections:

- Installation:** Instructions for installing the package from GitHub using devtools.
- Correlation Funnel in 2-Minutes:** A brief explanation of the problem of exploratory data analysis (EDA) and how the package provides a workflow and visualization tools.
- Main Benefits:** A list of three benefits: Speeds Up Exploratory Data Analysis, Improves Feature Selection, and Gets You To Business Insights Faster.
- Example - Bank Marketing Campaign:** A section showing a correlation matrix heatmap for a bank marketing campaign dataset.

Time-Based Explanation: CoxPh Survival Curves



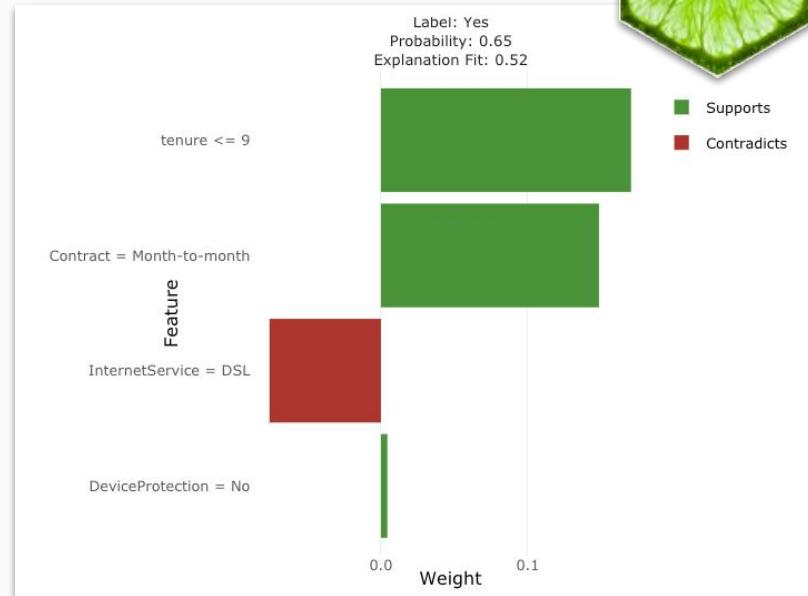
Individual Explanations & Churn Risk



```
H2OBinomialMetrics: stackedensemble
** Reported on cross-validation data. **
** 5-fold cross-validation on training data (Metrics computed for combined holdout predictions) **

MSE:  0.1348264
RMSE: 0.3671871
LogLoss: 0.4186951
Mean Per-Class Error: 0.235612
AUC: 0.848876
pr_auc: 0.6656359
Gini: 0.697752

Confusion Matrix (vertical: actual; across: predicted) for F1-optimal threshold:
      No  Yes  Error    Rate
No    4037 1137 0.219753  =1137/5174
Yes     470 1399 0.251471   =470/1869
Totals 4507 2536 0.228170  =1607/7043
```



30 Min Demo

Customer Churn Survival Analysis

Churn

Tactics for Retention

Churn

Impacts Every Company

Losing Customers Costs Millions.

We can reduce it.



Trick to Solving Churn Problems:

Shift users/people into lower probability cohorts.

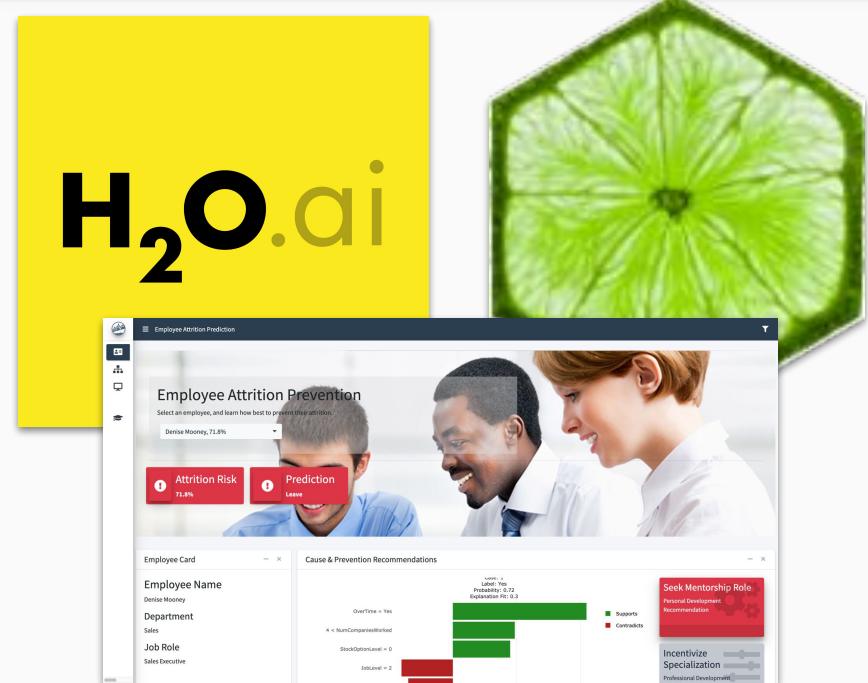
Churn

Key = Shift users

Develop insights with ML

Incentivize high risk users/people at individual level

Shift to low risk cohorts



Shift Cohorts

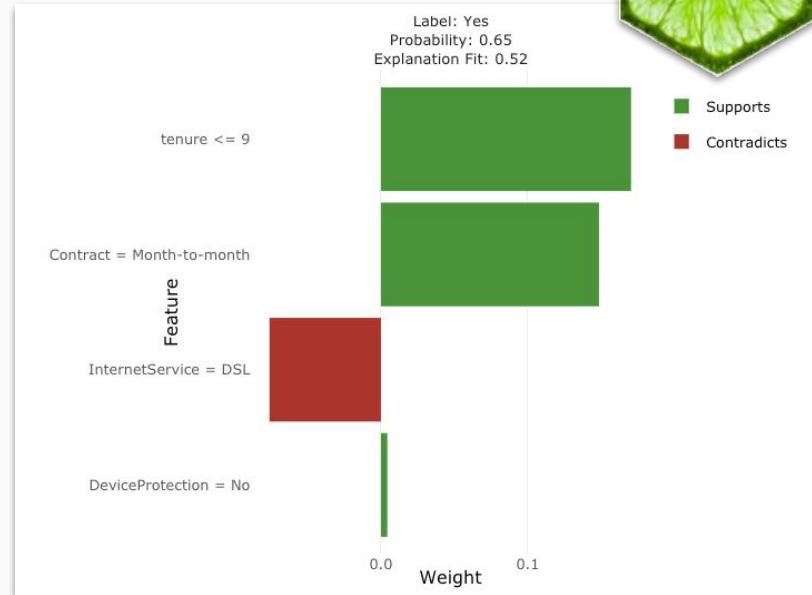


H₂O.ai

```
H2OBinomialMetrics: stackedensemble
** Reported on cross-validation data. **
** 5-fold cross-validation on training data (Metrics computed for combined holdout predictions) **

MSE:  0.1348264
RMSE: 0.3671871
LogLoss: 0.4186951
Mean Per-Class Error: 0.235612
AUC: 0.848876
pr_auc: 0.6656359
Gini: 0.697752

Confusion Matrix (vertical: actual; across: predicted) for F1-optimal threshold:
      No  Yes  Error    Rate
No    4037 1137 0.219753 =1137/5174
Yes     470 1399 0.251471 =470/1869
Totals 4507 2536 0.228170 =1607/7043
```



Learning Plan

Recap



Churn & Attrition Learning Plan



Learn **dplyr & ggplot2**

Gets data wrangling & visualization

Learn **parsnip**

Machine Learning

Learn **h2o & lime**

Automated ML & Local Feature
Explanation

(50+ Models in seconds)

Learn the tools

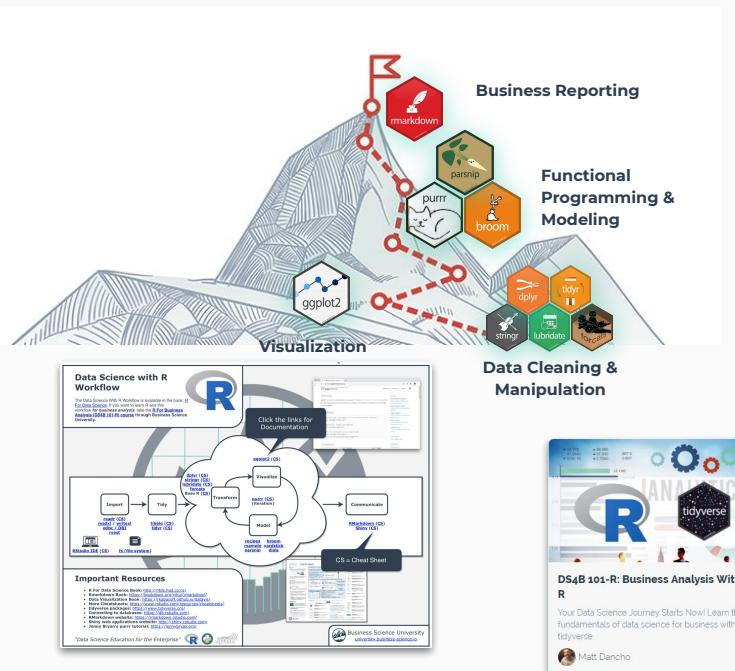


Step 1 - Learn the Foundations

Data Science Foundations

35 Hours of Video Lessons

- Machine Learning (parsnip)
- **Data Manipulation (dplyr)**
- Visualization (ggplot2)
- Reporting (rmarkdown)
- More packages



Matt Dancho

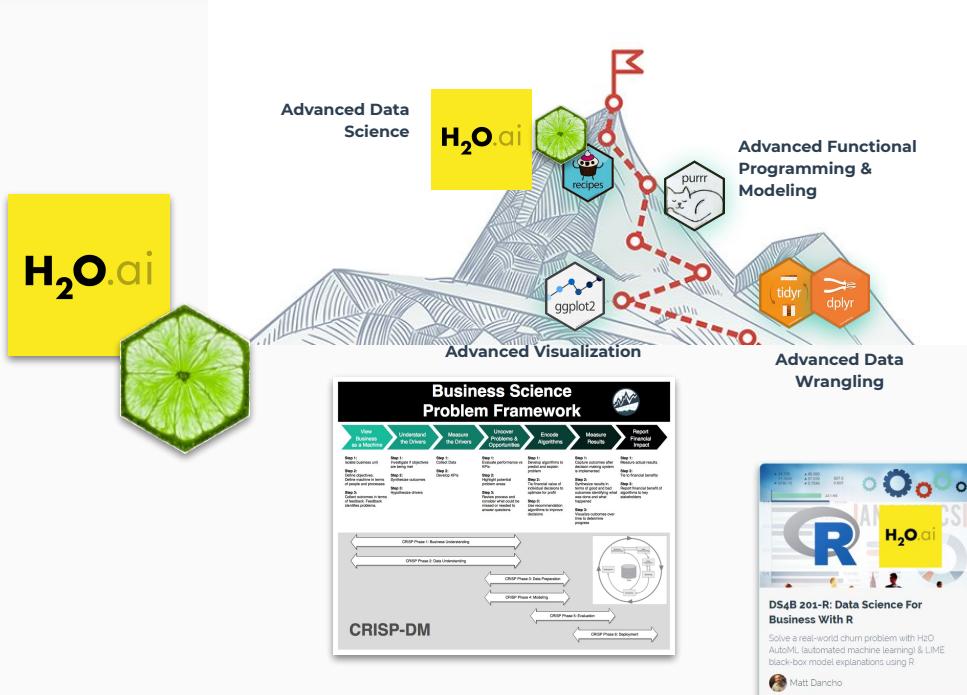


Step 2 - Learn Advanced ML

Advanced ML + Business Consulting

End-to-End Churn Project

- Machine Learning (H2O)
- Data Manipulation (lime)
- Repeatable Framework for Business Problems
- ROI Analysis for Project Benefit



Matt Dancho



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.



Project-Based Courses with Business Application

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



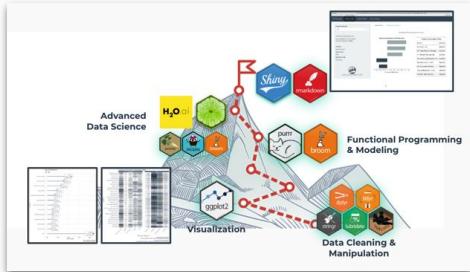
DS4B 102-R: Shiny Web Applications For Business (Level 1)

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

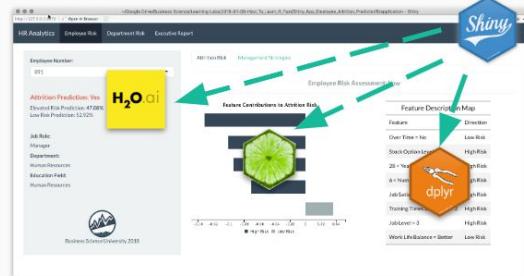


Program At A Glance

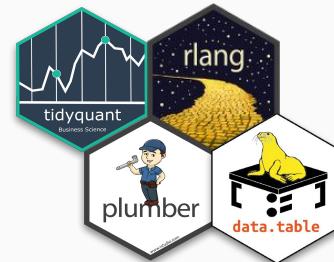
How It Works



Do Business Projects
Climb the Hill



Build Production-Ready
Web Apps



Complete 1-Hour Courses
Continuous Education

Start



Analysis Courses



App Development
Courses



Learning Labs PRO

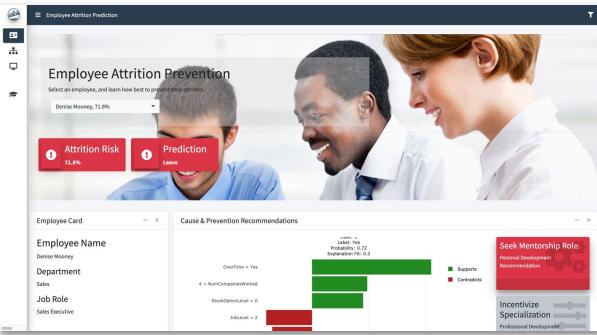
Finish

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

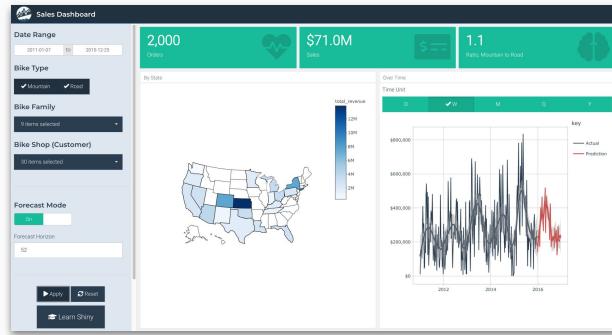
What You Build



Stock Portfolio Optimization



People Analytics Churn



Demand Forecasting

What You Learn



Advanced Machine Learning



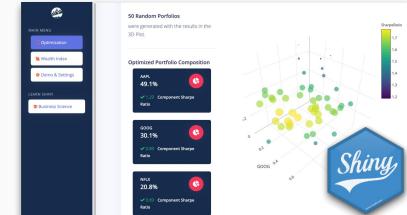
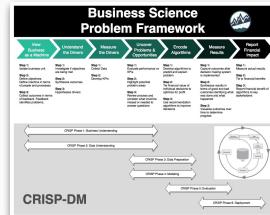
Applied to Business Problems



Following Repeatable Frameworks



Communicated with Data Products



Results



*“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



PROMO Code: **learninglabs**

R **H₂O.ai** **tidyverse** **Shiny**

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

3 Course Bundle **0%** COMPLETE

R **tidyverse**

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

R **H₂O.ai**

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

R **Shiny**

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

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

Matt Dancho

R-TRACK BUNDLE

MSRP: ~~\$234/mo~~

6 Low Monthly Payments
\$199/mo

Save: \$35/mo

Begin Learning Today

university.business-science.io

