How We Use R in Production

Andy Pryke
Chief Data Scientist (Sales & Marketing Solutions)
Sidetrade

What Sidetrade Sales & Marketing Do

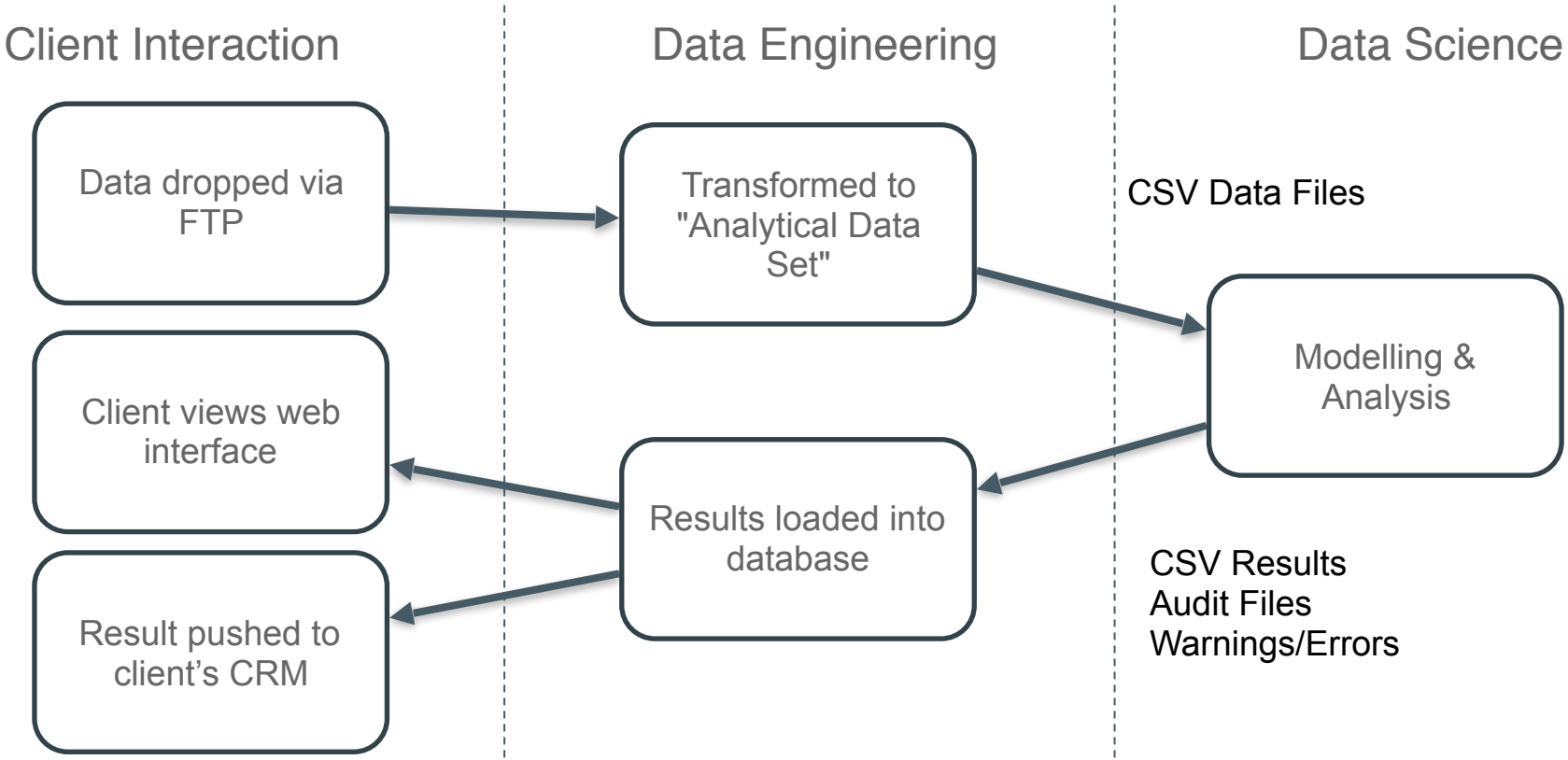
Take data from businesses who sell to businesses (B2B)
Do useful "Data Science" stuff with it

Churn Prediction - Will this customer leave or stay in next year?
Lead Prioritisation - Which leads are most likely to become customers?
Product Recommendations - Based on what they buy already, what might they buy in the future
Value - How much will a customer/lead spend?
Topic Modelling – What topic do customers websites cover
Segmentation - Customers by behaviour & background?

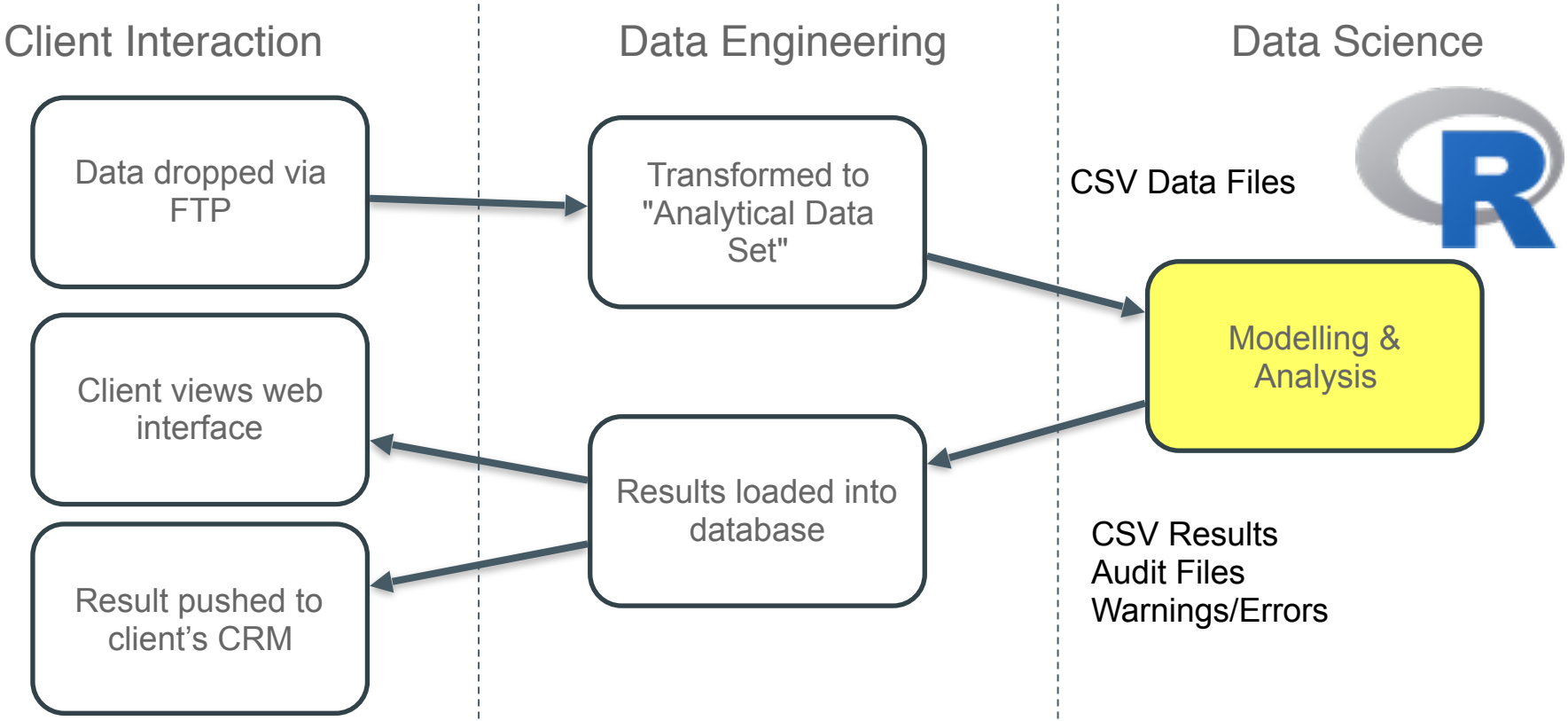
What is the Data?

Transactions - Who bought what, when, for how much
Customer - Date of 1st purchase, how acquired, any background
Leads - Any background data
Products - names, price, profit, product hierarchy
External Data – We add business sector, turnover, segmentation based on customer website etc... if available.
Data Volumes - from 100s to 100,000s of customers, sometimes very unbalanced data

Overall Architecture



Overall Architecture



How We call R

Shell calls to a R file which loads a text file (YML file) specifying what “function” to run and what parameters to use

```
Rscript.exe runDataScience.R "parameterFile=testBinaryModelAndScore.YML“
```

This calls multiple R Markdown Files with parameters via shell:

```
Rscript.exe -e "renderWithReportOnKnitrError('10.2_-_Preprocess_data.Rmd', output_file='Preprocess_HealthInsurance.html',
output_dir='x:/Test', quiet = TRUE)"
"baseYearFile=HealthInsurance_Train.txt"
"validationYearFile=HealthInsurance_Validation.txt"
"outputDataFilename=HealthInsurance.Rdata"
"randomSeed=0"
"numericDataChangeThreshold=0.4"
:
```

R Markdown is Key

Each “function” in our data science API is a RMD file. For example

- “Create new binary prediction model”
- “Score new data against binary prediction model”
- “Compare model performance with prior performance”

Reports are generated as we process data & build model.
If something goes wrong we can look at a partial report
Warnings like “Dataset is very small” are logged to file.
Warnings include machine readable recommendations like “set option to retrain on all data”, which can be automated.
Lots of reporting & charts

Challenges

2 Data scientists(!)

100s of models in production.

- Making sure they are "good"
- Justifying decisions
- Understanding what models are doing
- Detecting changes in data and model performance
- Detecting data errors

Challenges (continued)

- Large & small data volumes: from < 1k rows to 500k rows
- Balanced & very unbalanced (1:10,000) data
  (ratio of positive to negative examples)
- Making configuration easy
- Automating defaults

Detecting Dodgy Data

- Check for too many missing values & warn
- Data with no or low variance caret::nearZeroVar
- Detect ID like fields (too many different discrete values)
- Data distributions are different in train & score
  stats::ks.test & Cramer's V test
- Are some fields too predictive? Use caret::filterVarImp() to check

Preventing Perilous Predictions

- Are model predictions different on test & score data (KS & Cramer's V)
- How does our complex model compare with a simple one? Compare with a decision tree model.
- Track various metrics over time (e.g. lift in top 10%, false positive rate etc…)
- Warn if metrics change "too much" between runs
- Lots of diagnostics: Variable importance, show decision trees for most influential variables, what values of variables have the highest impact, examples of high & low scoring customers (rows)

Example Reports

Coding Principles

Pragmatic programmer book – old but lots of good tips

- Write for your future self
- Clear code rather than conciseness or comments
- Intermediate variables, rather than excessively chained code ("%>%")
- Few abbreviations in names - so you don't have to remember what /when you abbreviated
- Keeping things generic - no hard coding!
- Literate (like) Programming / Processing – Using R Markdown

Key packages

Key packages - Reporting

rmarkdown, knitr - All code runs as R Markdown reports

ggplot - plots in reports (use base R plots & other too)

DT - Print tables nicely in markdown

Key packages - Data manipulation

data.table - Fast manipulation of larger datasets

dplyr – data transformation (but not used much as can be slower than data.table on larger datasets)

fasttime – Quickly turn text into a date object

Key packages - Modelling

caret - Standardised interface to many many classification & regression methods

randomForest - Prediction

party, rpart, arules, discretization, infotheo - Explaining relationship of data & model

arules, recommenderlab - Recommendations

cluster - K-medians segmentation (more robust than k-means)

topicmodels - modelling topics of texts

Key Packages - Misc

RPostgres, DBI – Database access

RColorBrewer – Nice, principled, colours

xml2, yaml – Read XML & YML files

tm – Text Mining

lubridate – text to date, timespans

testthat – Test functions to check they do what they should

Thanks!

- Any Questions

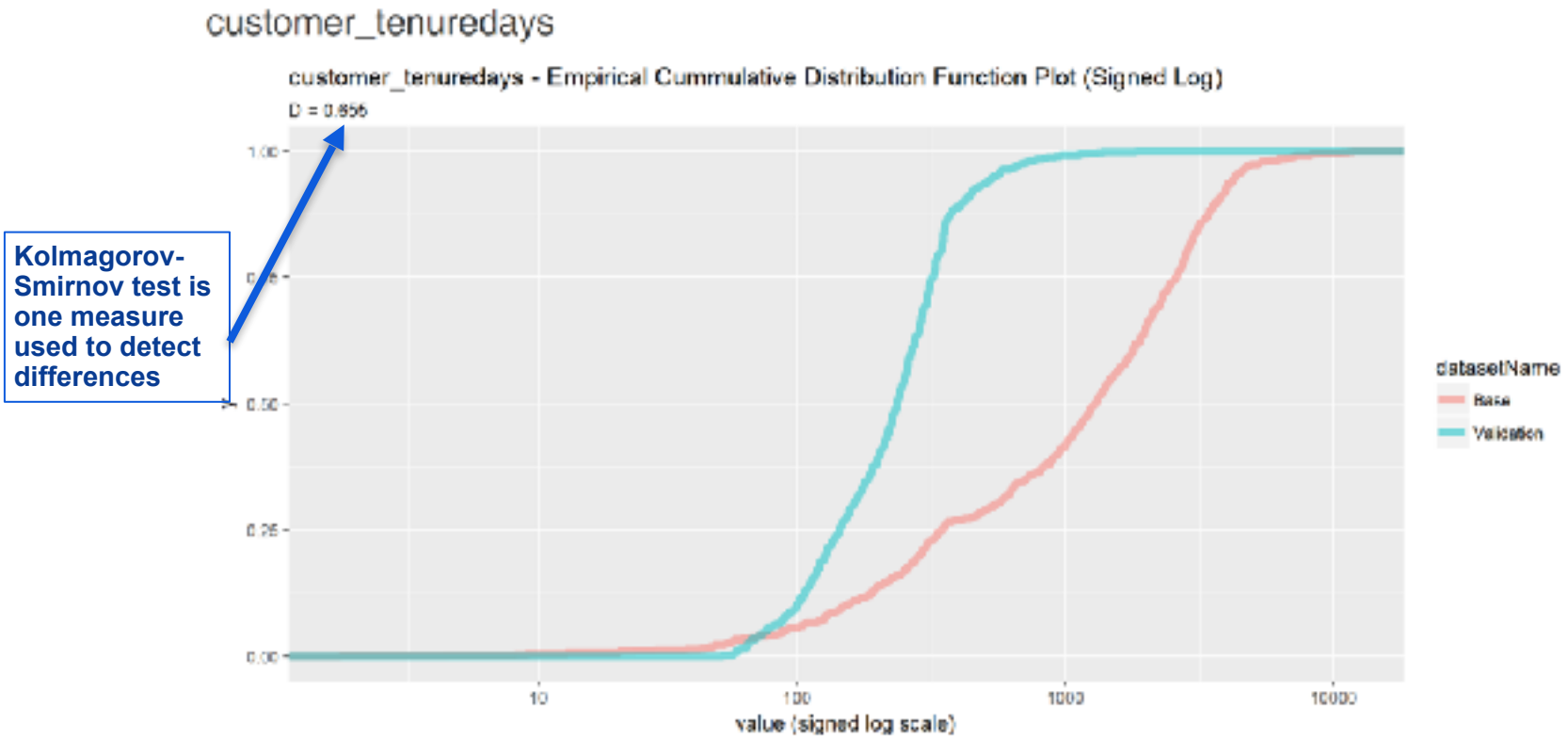
Contact Details

Andy Pryke on LinkedIn
https://www.linkedin.com/in/andypryke/ (include a message please)

Andy@AndyPryke.com

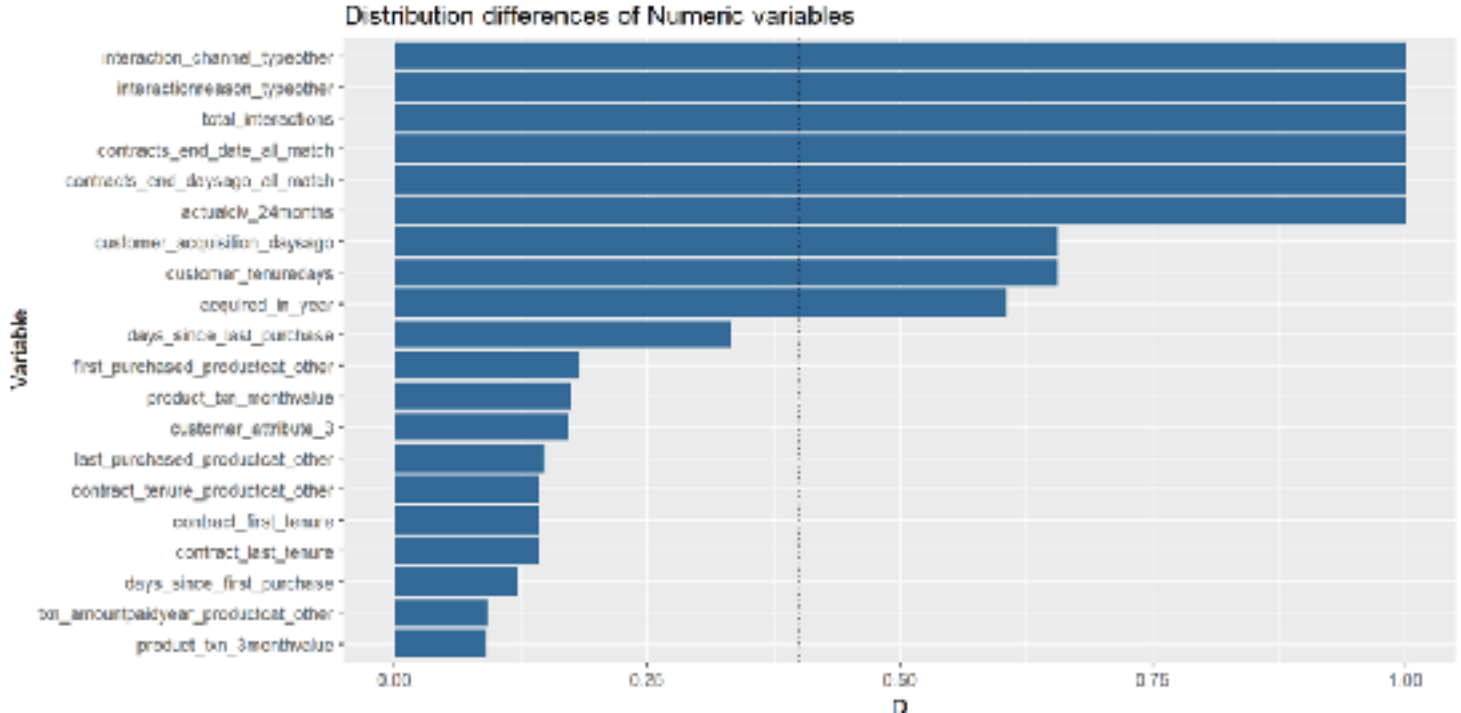
Example Diagnostic Charts

Detecting Differences between Train & Score Data



Overview of Numeric Columns which Differ (using KS-Test)

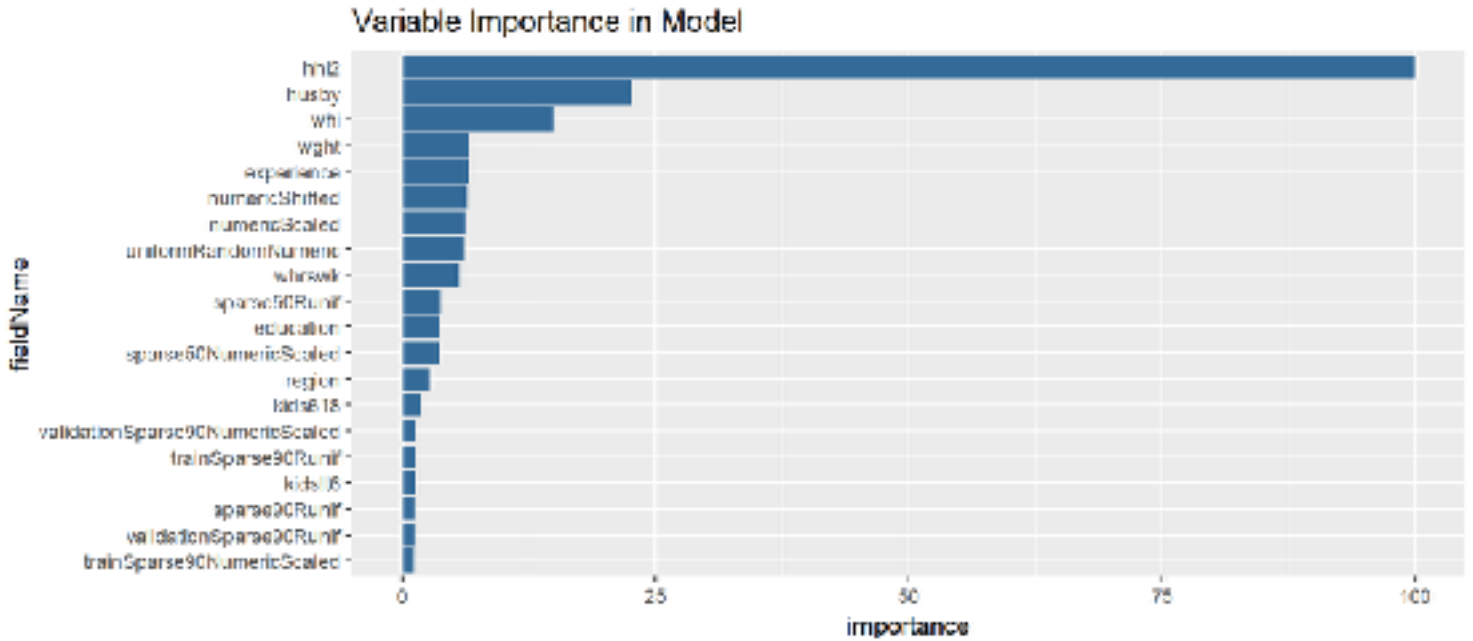
DataScienceWarning: The following Numeric column(s) differ significantly between train and score sets that haven't already been excluded: interaction_channel_typeother, interactionreason_typeother, total_interactions, contracts_end_date_all_match, contracts_end_daysago_all_match, customer_acquisition_daysago, customer_tenuredays, acquired_in_year



What Variables Influence Model?

Variable Importance Based on Model

This importance measure is calculated from the model. Note that it is not directly comparable with the simple importance measure, as it is scaled differently. For this measure, the variable which is most useful in predictions has a score of 100.



What Values of Variable Matter?

	Importance	Multiplier	AbsImportance
	All	All	All
hh?=1	57.98	1.65	57.98
husby=20.8 <= x <= 168.8	12.27	1.37	12.27
whi=0	6.99	1.24	6.99
whrswk=0 <= x < 33	1.83	1.16	1.83
hispanic=0	0.54	1.09	0.54
experience=7.8 <= x < 37.8	0.48	1.05	0.48
region=other	0.34	1.13	0.34
race=white	0.18	1.01	0.18
wght=113644 <= x <= 668500	0.18	1.04	0.18
region=northcentral	0.17	1.09	0.17
kida618=2 <= x < 3	0.15	1.11	0.15
kida618=1 <= x < 2	0.12	1.08	0.12
education=13-15years	0.1	1.06	0.1

What Influences an Individual Prediction?

	rowId	rank	driverDetailId	driverDetailsText	importance
12	12264	1	23	hhi2=0	-57.98
6	12264	2	21	whi=0	6.99
5	12264	3	6	whrswk=10 <= x < 11	-0.73
27	12264	4	34	hispanic=0	0.54
31	12264	5	55	experience=7.0 <= x < 37.0	0.40
26	12264	6	33	race=white	0.18
57	12264	7	89	wght=113641 <= x <= 958500	0.18
52	12264	8	85	region=northcentral	0.17
47	12264	9	73	husby=15.1 <= x < 20.0	-0.14

Comparing Distributions of Predictions