What's Data Science Reporting?

Amy Tzu-Yu Chen

@amy17519 on LinkedIn, Twitter, Github

LA East R Users Group

Dec 9, 2019

About Me

Data Scientist

SYSTEM

Organizer







(My) Definition

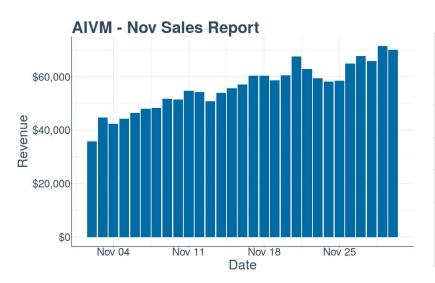
Definition

- Data science reports inform data scientists and stakeholders
 - whether a data science solution worked
 - how well it worked
 - whether it continues to work

Vending Machine

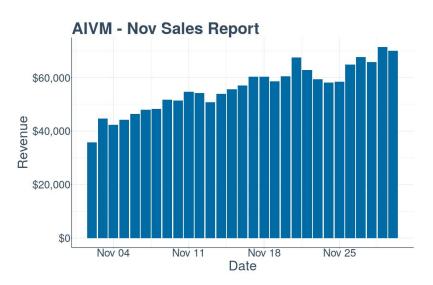




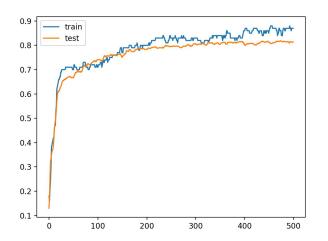


Date	Passerby	Sold	Revenue	Conversion	\$\$/Passerb y
Nov 2	253,445	47,418	\$35771.66	18.71%	\$0.14
Nov 3	355,658	62,896	\$44658.3	17.68%	\$0.13
Nov 4	316,322	54,501	\$42307.98	17.23%	\$0.13
Nov 29	390,060	67,774	\$71425.41	17.38%	\$0.18
Nov 30	429,267	80,620	\$70074.33	18.78%	\$0.16

DS Reports are not BI, DA, Business Report



Date	Passerby	Sold	Revenue	Conversion	\$\$/Passerb y
Nov 2	253,445	47,418	\$35771.66	18.71%	\$0.14
Nov 3	355,658	62,896	\$44658.3	17.68%	\$0.13
Nov 4	316,322	54,501	\$42307.98	17.23%	\$0.13
Nov 29	390,060	67,774	\$71425.41	17.38%	\$0.18
Nov 30	429,267	80,620	\$70074.33	18.78%	\$0.16



Decision Tree Model	Classification Rate (%)	Ensemble Output (%)
M1	92.6	94.7
M2	87.9	91.8
М3	95.2	97.3

Predicted/Classified

Actual

	Positive	
Negative	998	0
Positive	1	1

Date	Passerby	Sold	Revenue	Conversion	\$\$/Passerby	Accuracy
Nov 2	253,445	47,418	\$35771.66	18.71%	\$0.14	***
Nov 3	355,658	62,896	\$44658.3	17.68%	\$0.13	***
Nov 4	316,322	54,501	\$42307.98	17.23%	\$0.13	***
Nov 29	390,060	67,774	\$71425.41	17.38%	\$0.18	***
Nov 30	429,267	80,620	\$70074.33	18.78%	\$0.16	***

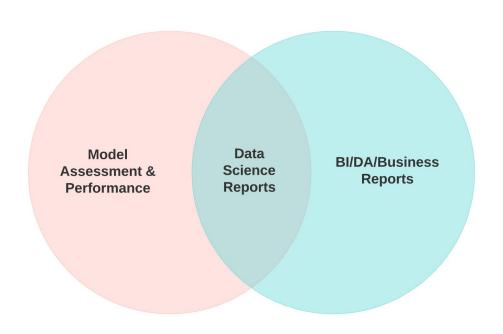
DS Reports are not model assessment

Date	Passerby	Sold	Revenue	Conversion	\$\$/Passerby	Accuracy
Nov 2	253,445	47,418	\$35771.66	18.71%	\$0.14	***
Nov 3	355,658	62,896	\$44658.3	17.68%	\$0.13	***
Nov 4	316,322	54,501	\$42307.98	17.23%	\$0.13	***
Nov 29	390,060	67,774	\$71425.41	17.38%	\$0.18	***
Nov 30	429,267	80,620	\$70074.33	18.78%	\$0.16	***

DS Reports are not model assessment

Date	Passerby	Sold	Revenue	Conversion	\$\$/Passerby	Accuracy
Nov 2	253,445	47,418	\$35771.66	18.71%	\$0.14	****
Nov 3	355,658	62,896	\$44658.3	17.68%	\$0.13	$\overline{}$
Nov 4	316,322	54,501	\$42307.98	17.23%	\$0.13	
Nov 29	390,060	67,774	\$71425.41	17.38%	\$0.18	
Nov 30	429,267	80,620	\$70074.33	18.78%	\$0.16	/**

The good old Venn Diagram



My AIVM Report

Date	Booorby	Passerby Sold	Sold Revenue		\$\$/Pas	sserby	DS Lift %
Date	Passerby	Solu	Revenue	n	No Model	DS Model	DS LIII %
Nov 2	253,445	47,418	\$35771.66	18.71%	\$0.14	\$0.16	14.28%
Nov 3	355,658	62,896	\$44658.3	17.68%	\$0.13	\$0.14	7.14%
Nov 4	316,322	54,501	\$42307.98	17.23%	\$0.12	\$0.14	16.67%
Nov 29	390,060	67,774	\$71425.41	17.38%	\$0.18	\$0.19	5.56%
Nov 30	429,267	80,620	\$70074.33	18.78%	\$0.16	\$0.18	12.5%

Before We Begin



The Metric

- Decide on **one** metric
- Communicate and agree on this in advance
- It's often but not necessarily the outcome variable



DS model recommends drinks to maximize \$\$ per passerby

Date	Bassarby	Sold	Revenue	Conve	ersion	\$\$/Pas	sserby	DS Lift %
Date	Passerby	Solu	Revenue	No Model	DS Model	No Model	DS Model	D3 LIII %
Nov 2	253,445	47,418	\$35771.66	20%	18%	\$0.14	\$0.16	14.28%
Nov 3	355,658	62,896	\$44658.3	24%	17%	\$0.13	\$0.14	7.14%
Nov 4	316,322	54,501	\$42307.98	26%	19%	\$0.12	\$0.14	16.67%
Nov 29	390,060	67,774	\$71425.41	20%	16%	\$0.18	\$0.19	5.56%
Nov 30	429,267	80,620	\$70074.33	21%	19%	\$0.16	\$0.18	12.5%

DS model recommends drinks to maximize a

\$\$ per passerby

Doto	Doggarby	Cald	Revenue	Conve	ersion	\$\$/Pas	sserby	DC 1 :f4 0/
Date	Passerby	Sold	Revenue	No Model	DS Model	No Model	DS Model	DS Lift %
Nov 2	253,445	47,418	\$35771.66	20%	18%	\$0.14	\$0.16	14.28%
Nov 3	355,658	62,896	\$44658.3	24%	17%	\$0.13	\$0.14	7.14%
Nov 4	316,322	54,501	\$42307.98	26%	19%	\$0.12	\$0.14	16.67%
Nov 29	390,060	67,774	\$71425.41	20%	16%	\$0.18	\$0.19	5.56%
Nov 30	429,267	80,620	\$70074.33	21%	19%	\$0.16	\$0.18	12.5%



DS model recommends drinks to maximize \$\$ per passerby

Date	Посоливу	Sold	Revenue	Conve	ersion	\$\$/Pas	sserby	DC 1 :#4 0/
Date	Passerby	Solu	Revenue	No Model	DS Model	No Model	DS Model	DS Lift %
Nov 2	253,445	47,418	\$35771.66	20%	18%	\$0.14	\$0.16	14.28%
Nov 3	355,658	62,896	\$44658.3	24%	17%	\$0.13	\$0.14	7.14%
Nov 4	316,322	54,501	\$42307.98	26%	19%	\$0.12	\$0.14	16.67%
Nov 29	390,060	67,774	\$71425.41	20%	16%	\$0.18	\$0.19	5.56%
Nov 30	429,267	80,620	\$70074.33	21%	19%	\$0.16	\$0.18	12.5%

DS model recommends drinks to maximize a

\$\$ per passerby

- Decide on **one** metric
- Communicate and agree on this in advance
- It's often but not necessarily the outcome variable

Design Components

Design Components

- Measurability
- Format
- Channel
- Alerts
- Audience

- Can you regularly and automatically build the DS report
- Questions to ask:
 - Do we have data daily? Weekly? Monthly?
 - Data quality
 - Data processing

- Measurability
- Format
- Channel
- Alerts
- Audience

- Can you regularly and automatically build the DS report
- Questions to ask:
 - Do we have data daily? Weekly? Monthly?
 - Data quality
 - Data processing

- Measurability
- Format
- Channel
- Alerts
- Audience

- Can you **regularly and automatically** build the DS report
- Questions to ask:
 - Do we have data daily? Weekly? Monthly?
 - Data quality is bad
 - Data processing

- Measurability
- Format
- Channel
- Alerts
- Audience

- Can you regularly and automatically build the DS report
- Questions to ask:
 - Do we have data daily? Weekly? Monthly?
 - Data quality is bad
 - Data processing is not automated

- Measurability
- Format
- Channel
- Alerts
- Audience

- Can you regularly and automatically build the DS report
- Questions to ask:
 - Do we have data daily? Weekly? Monthly?
 - Data quality is bad
 - Data processing is not automated

Measurability V

- Format
- Channel
- Alerts
- Audience

Sorry, You are not ready to automate this report yet

•	Measurahility

- Format
- Channel
- Alerts
- Audience

Date	Passerby	Sold	Revenue	Conversio n	\$\$/Passerby		DS Lift %
					No Model	DS Model	D3 Liit %
Nov 2	253,445	47,418	\$35771.66	18.71%	\$0.14	\$0.16	14.28%
Nov 3	355,658	200	\$44658.3	17.68%	\$0.13	\$0.14	7.14%
Nov 4	316,322	54,501	\$42307.98	17.23%	\$0.12	\$0.14	16.67%
Nov 29	390,060	67,774	\$71425.41	17.38%	\$0.18	\$0.19	5.56%
Nov 30	429,267	80,620	\$0	18.78%	\$0	\$0	NA

Should I trust this?

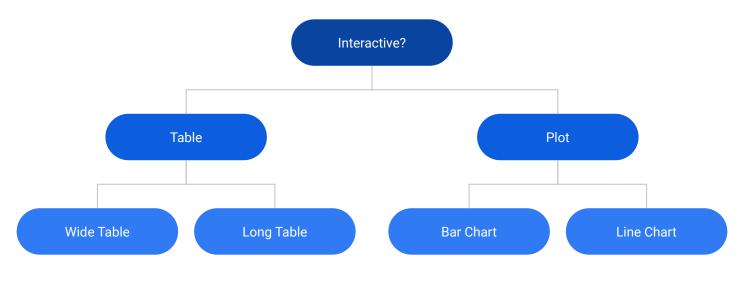


Date	Passerby	Sold	Revenue	Conversio n	\$\$/Passerby		DS Lift %
					No Model	DS Model	DS LIII %
Nov 2	253,445	47,418	\$35771.66	18.71%	\$0.14	\$0.16	14.28%
Nov 3	355,658	200	\$44658.3	17.68%	\$0.13	\$0.14	7.14%
Nov 4	316,322	54,501	\$42307.98	17.23%	\$0.12	\$0.14	16.67%
Nov 29	390,060	67,774	\$71425.41	17.38%	\$0.18	\$0.19	5.56%
Nov 30	429,267	80,620	\$0	18.78%	\$0	\$0	NA

- Measurability
- Format
- Channel
- Alerts
- Audience

• • • • • •

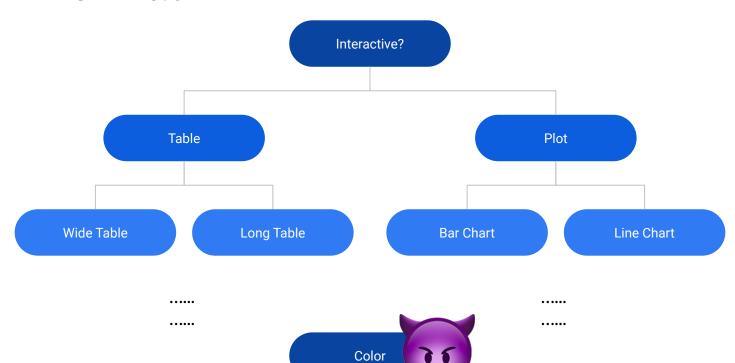
• • • • • •



• • • • • •

• • • • • •

- Measurability
- Format
- Channel
- Alerts
- Audience



- Measurability
- Format
- Channel
- Alerts
- Audience

- Tips
 - Transparency
 - Clarity
 - Moderate amount of info
 - Table: Format numbers (Add % and \$)
 - Plot: Avoid using too many geometric shapes (lines, bars, and dots should be sufficient)
 - 0

- Measurability
- Format
- Channel
- Alerts
- Audience

- Tips
 - Transparency
 - Clarity
 - Moderate amount of info
 - \circ Table: Format numbers (Add % and \$)
 - Plot: Avoid using too many geometric shapes (lines, bars, and dots should be sufficient)
 - 0

Bottomline:

- You enjoy reading the report
- Anyone can make a call whether both business and DS models are healthy in 5 sec

- Measurability
- Format
- Channel
- Alerts
- Audience

Channel

 Choose the right place to send your report & be your own advocate

- Measurability
- Format
- Channel
- Alerts
- Audience

Channel

	R	python		
Slack	slackr	python-slackclient		
Email	mailR, gmailr, blastula, blatr, mail, sendmailR	smtplib+email		
Google Sheet	googlesheets	gsheets		
Dashboards	shiny	plotly dash, flask		

- Measurability
- Format
- Channel
- Alerts
- Audience

Alerts

WARNING: Lift is dropping

- Make your report work for you!

Date	Passerby	Sold	Revenue	Conversio n	\$\$/Passerby		DS Lift %
Date					No Model	DS Model	D3 LIII %
Nov 2	253,445	47,418	\$35771.66	18.71%	\$0.14	\$0.16	14.28%
Nov 3	355,658	62,896	\$44658.3	17.68%	\$0.13	\$0.14	7.14%
Nov 4	316,322	54,501	\$42307.98	17.23%	\$0.12	\$0.14	16.67%
Nov 29	390,060	67,774	\$71425.41	17.38%	\$0.12	\$0.09	-25%
Nov 30	429,267	80,620	\$70074.33	18.78%	\$0.16	\$0.09	-43.75%

- Measurability
- Format
- Channel
- Alerts
- Audience

Audience

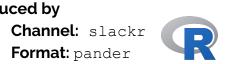
- Who?
 - Yourself
 - Stakeholders
- Use the report to
 - Measure + monitor DS performance
 - Discover missed opportunities
 - Educate your audience on DS philosophy

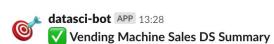
- Measurability
- Format
- Channel
- Alerts
- Audience

Showcase

Produced by





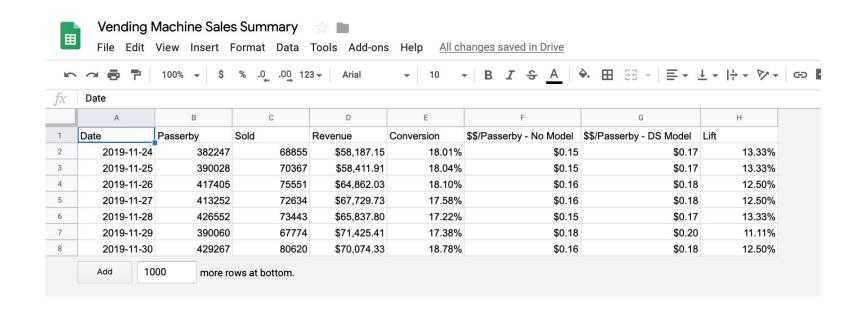


2019-11-24 382	32247 68855					
	1441 00000	\$58,187.15	18.01%	\$0.15	\$0.17	13.33%
2019-11-25 390	00028 70367	\$58,411.91	18.04%	\$0.15	\$0.17	13.33%
2019-11-26 41	.7405 75551	\$64,862.03	18.10%	\$0.16	\$0.18	12.50%
2019-11-27 413	.3252 72634	\$67,729.73	17.58%	\$0.16	\$0.18	12.50%
2019-11-28 420	6552 73443	\$65,837.80	17.22%	\$0.15	\$0.17	13.33%
2019-11-29 390	00060 67774	\$71,425.41	17.38%	\$0.18	\$0.20	11.11%
2019-11-30 429	9267 80620	\$70,074.33	18.78%	\$0.16	\$0.18	12.50%

Produced by

• Channel: googlesheets





Takeaway

Check your DS reports

- Data science reports inform data scientists and stakeholders
 - whether a data science solution worked
 - how well it worked
 - whether it continues to work

DS Report should be a part of DS workflow

- DS reports should be a part of data scientist's responsibility
- You build a DS report before someone else asks for it
- You do this for your
 - stakeholders
 - current self
 - future self

Backup Slides

A Note on Tidy Data

- Your report does not have to be in the format convenient for data processing.
 - Readability > <u>Tidy Data Principle</u>

A Note on Tidy Data

- Your report does not have to be in the format convenient for data processing.
 - Readability > <u>Tidy Data Principle</u>



Date	Passerby	Sold	Revenue	Conversion	\$\$/Passerby - No Model	\$\$/Passerby - DS Model	Lift
2019-11-24	382247	68855	\$58,187.15	18.01%	\$0.15	\$0.17	13.33%
2019-11-25	390028	70367	\$58,411.91	18.04%	\$0.15	\$0.17	13.33%
2019-11-26	417405	75551	\$64,862.03	18.10%	\$0.16	\$0.18	12.50%
2019-11-27	413252	72634	\$67,729.73	17.58%	\$0.16	\$0.18	12.50%
2019-11-28	426552	73443	\$65,837.80	17.22%	\$0.15	\$0.17	13.33%
2019-11-29	390060	67774	\$71,425.41	17.38%	\$0.18	\$0.20	11.11%
2019-11-30	429267	80620	\$70,074.33	18.78%	\$0.16	\$0.18	12.50%

What is Tidy Data Principle?

Tidy data is a standard way of mapping the meaning of a dataset to its structure. A dataset is messy or tidy depending on how rows, columns and tables are matched up with observations, variables and types. In **tidy data**:

- 1. Each variable forms a column.
- 2. Each observation forms a row.
- 3. Each type of observational unit forms a table.