Building Data Fluency through Analogies from the Real World and Movies

Chester Ismay, Ph.D.
Data Science Evangelist
DataRobot

How can we explain the following concepts via analogy?

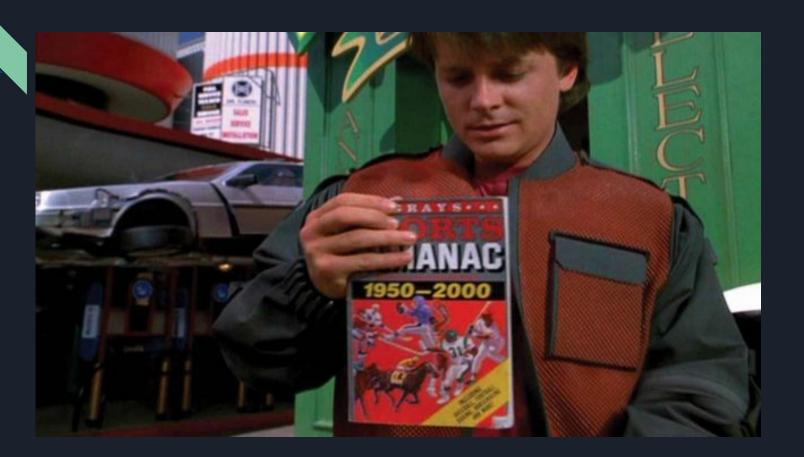
- R, RStudio, and R packages
- Cross-validation
- Machine Learning Life Cycle
- Target Leakage

Back to the Future Part II

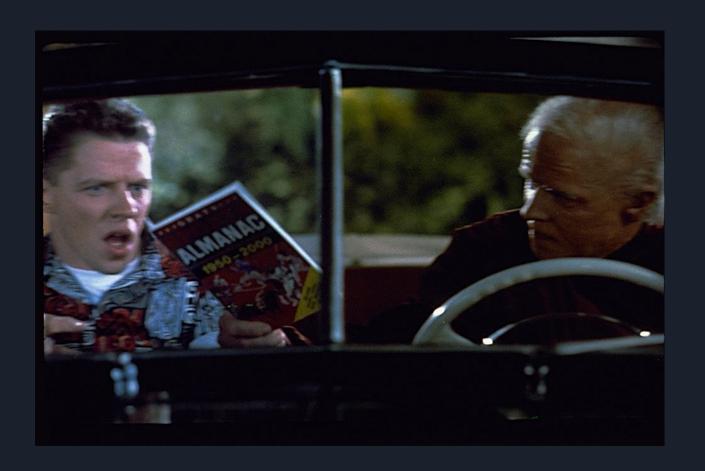












The Weather

Tackey - Month season high wear 50 kee year 60. The chance of precipitation is seen seen through knowled beautiful Depretor - Versiche cloudings with

Hill Halley Telegraph

Vol. XVII. No. 32

COMPLETE NEWS SERVICE

THURSDAY, APRIL 16, 1959

BIFF TANNEN

Luckiest Man On Earth

PAUL CRUMRINE

A suggestion that public hearings on applications be limited to one every six months was taken under ad-

visement by the commission. Many persons feel at this stage that some logal action is forthcoming but it now becomes common knowledge that there is pressure from the inside which will materially change the aspect of the case.

Of no less importance was the common recognition shown of the fact that any menace from without to the peace of our continents concerns all of us and therefore properly is a subject for consultation and cooperation. This was reflected in the instruments adopted by the con-

Thus at this conference all our governments found themselves in unanimous agreement regarding this undertaking. Arrangements for doubing with questions and disputes between the republics were further improved

A suggestion that public hearings on applications be limited to one every aix months was taken under adviscount by the commission.

An immediate investigation to assured and indications are that some new light will be ahed on the situation in the near future. Available facto seem vague but authorities



BIFF TANNEN

Youth Americal After Glitter Was Removed From Campaign Sign

From Vestinary Late Millions
BERGERS assets of the sparitie
P and giftine was taken one of a
Prince-for-Greener' sign on the
of malletine minded was filled in
summer year care! publishing
against Fred Ragner. For-raining
parity proceedings of mapping of
the filled by the proceeding of the proceeding of
the proceedings of the proceeding of
the proceedings of the proceedings of
the proceedings of the proceeding

The generatorial night-and was The general-oral negations was closeled with glass disast which year close accessing highes at eight giving the eigh sufficient for eight sufficient and that linguing removed the fitter between the eight sufficient for eight sufficient for

State Likely to Start Payroll Tax in 1960

Of no less importance was the common recognition shown of the fact that any menace from without to the peace of our continents conperns all of us and therefore properly is a subject for consultation and cooperation.

Nasser Accuses Reds of Plotting His Overthrow

are going forward toward what is hoped will be a friendly meeting of both sides at the conference. This debate has certainly stirred up the local media but things have been somewhat amouthed over by the decision of the committee to allow the local television station to cover the conference

The facts regarding the situation remain the same, state the authorities. Details concorning the action have been given a prolimonary investigation but it is felt that only by a more detailed study will the true facts become known.

It would appear that the preliminary inquiry into this matter has in fact not settled any of the minor obferences arining from the situation but rather has aggregated the mood of thuse petitioning for more local onvolvement by the

Many persons feel at this stage that some legal action is forthcoming but it now beounes common knowledge that there is pressure from the inaide which will motorially

The facts regarding the sitsistion remain the same, state the authorities. Details concerning the action have been given a preliminary investigation but it is nelt that only by a more detailed study will the true facts become known.

Thus at this conference all our governments found thempelves in ununimous agree ment regarding this undertaking Arrangements for dealing between the republics were

The Mayor, mountable has dislomatically heat a low preat City Hall confirm that usprivate meetings concerning the same, that the Mayor will have no public statement to

An immediate investigation is assured and indications are that some new light will be shed on the situation in the near future Available facts seem vague but authorities feel that time will disclose some means of arriving of a

Many persons feel at this change the aspect of the case. stage that some legal action is

Khruschev Offers **Dates for Summit**

An interesting telephone in the appropriate and introduced part. MARKS PROVIDE AND ADDRESS OF were married of courting at a and the same

Potent place rule, of born. 400, have given benefits on the sphanos and non-stands. Inseptember with facts to be made on the principal planeters of the principle will take manufaction Seep tool, A to Roll that Dieter Statute and right displayment.

Pinters want both of tensor select Review private horseling on their school or at 3, year marries. For

the tell and manufacture was The contracts surrage, time that more now legits mill par property of the fact, that are that is its pleasure in the contact from Advant to the many richery. Arranged State. Space of the transporting term. Series all of the last States Comthat that you this discuss , printed in a printed to trive Then was confincted in the coproduced without he directly Statute

> man Julius, Alexandra Saint. Now Man, Asian 11th Administra



SECTION AND PERSONS.

White the West arrivated of Discontinuities had not to be the minimal of an possible float more of the July name to

Strachage or applications by

IN ROOM access. Disc the problemate bearing the Mar. mediter had to duck but settlers. MAY OF THE SOUR STREET, person of the conference told principle and the Continue by artered for each work, Park . region floor appropriate the a limited with the state of a world of these partitioning for property spirity the made and . Spark half confirmmed by the southful.

> The lasts cognising the allterfiner contracts from relative, prome the builty blue, beliefly non-SPIRITE THE MIGHL SHITE THERE. gives a southerney terretion. tion but it is not that made and but a facts benefit shall will the Street Variety between the bright of

The Storm manuable has Title-At book in position Manhouse at this Red makes that we of the brister I may be added the private provings transcring ther series, birth Dat Mission with NAME AND ADDRESS OF OWNER, OR







Back to the Future writer: ... Biff was based on Donald Trump



Two-Stage Modeling Process



Stage 1: **Training / Building**

Historical Data

Algorithm (Blueprint)

+

Testing / Validation



Stage 2:

Predicting / Scoring

New Data



Model is BUILT

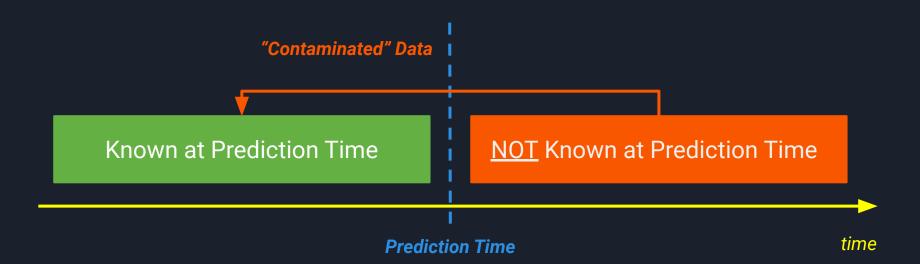
(usually not frequent)

Predictions and Decisions

Model is **USED**

(model is "static")

Target Leakage





Predicting Recidivism



Preventing Target Leakage



How to avoid the data leakage? blogpost

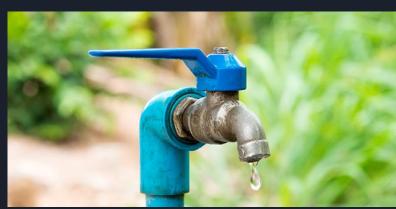
Preventing Target Leakage



Further References

- What is Target Leakage? DataRobot Wiki
- Leakage Data Skeptic 12 minute podcast

Prevent Target Leakage to Prevent Bad Parallel Universes





Prevent Target Leakage to Prevent Bad Parallel Universes



Time Travel from Back to the Future Part II

→ Target Leakage



- Define Project Objectives
 - Specify business problem
 - Acquire subject matter expertise
 - Define unit of analysis and prediction target
 - Prioritize modeling criteria
 - Consider risks and success criteria
 - Decide whether to continue
- 2. Acquire & Explore Data
 - Find appropriate data
 - Merge data into single table
 - Conduct exploratory data analysis
 - Find and remove any target leakage
 - Feature engineering

- Model Data
 - Variable selection
 - Build candidate models
 - Model validation and selection
- 4. Interpret & Communicate
 - Interpret model
 - Communicate model insights
- 5. Implement, Document & Maintain
 - Set up batch or API prediction system
 - Document modeling process for reproducibility
 - Create model monitoring and maintenance plan

DataRobot wiki



1. Define Project Objectives

- Specify business problem
- Acquire subject matter expertise
- Define unit of analysis and prediction target
- Prioritize modeling criteria
- Consider risks and success criteria
- Decide whether to continue

Acquire & Explore Data

- Find appropriate data
- Merge data into single table
- Conduct exploratory data analysis
- Find and remove any target leakage
- Feature engineering

Model Data

- Variable selection
- Build candidate models
- Model validation and selection

4. Interpret & Communicate

- Interpret model
- Communicate model insights

Implement, Document & Maintain

- Set up batch or API prediction system
- Document modeling process for reproducibility
- Create model monitoring and maintenance plan

Define Project Objectives Acquire & Explore Data

Model Data

Interpret & Communicate Implement, Document & Maintain

1. Define Project Objectives

- Specify business problem
- Acquire subject matter expertise
- Define unit of analysis and prediction target
- Prioritize modeling criteria
- Consider risks and success criteria
- Decide whether to continue

2. Acquire & Explore Data

- Find appropriate data
- Merge data into single table
- Conduct exploratory data analysis
- Find and remove any target leakage
- Feature engineering





- Define Project Objectives
 - Specify business problem
 - Acquire subject matter expertise
 - Define unit of analysis and prediction target
 - Prioritize modeling criteria
 - Consider risks and success criteria
 - Decide whether to continue
- Acquire & Explore Data
 - Find appropriate data
 - Merge data into single table
 - Conduct exploratory data analysis
 - Find and remove any target leakage
 - Feature engineering

- Model Data
 - Variable selection
 - Build candidate models
 - Model validation and selection
- 4. Interpret & Communicate
 - Interpret model
 - Communicate model insights
- Implement, Document & Maintain
 - Set up batch or API prediction system
 - Document modeling process for reproducibility
 - Create model monitoring and maintenance plan

Define Project Objectives Acquire & Explore Data

Model D

Define Project Objectives

- Specify business problem
- Acquire subject matter expertise
- Define unit of analysis and prediction target
- Prioritize modeling criteria
- Consider risks and success criteria
- Decide whether to continue

Acquire & Explore Data

- Find appropriate data
- Merge data into single table
- Conduct exploratory data analysis
- Find and remove any target leakage
- Feature engineering



- Set up batch or API prediction system
- Document modeling process for reproducibility
- Create model monitoring and maintenance plan



1. Define Project Objectives

- Specify business problem
- Acquire subject matter expertise
- Define unit of analysis and prediction target
- Prioritize modeling criteria
- Consider risks and success criteria
- Decide whether to continue

2. Acquire & Explore Data

- Find appropriate data
- Merge data into single table
- Conduct exploratory data analysis
- Find and remove any target leakage
- Feature engineering

3. Model Data

- Variable selection
- Build candidate models
- Model validation and selection

4. Interpret & Communicate

- Interpret model
- Communicate model insights

Implement, Document & Maintain

- Set up batch or API prediction system
- Document modeling process for reproducibility
- Create model monitoring and maintenance plan



Acquire & Explore Data

- Find appropriate data
- Merge data into single table
- Conduct exploratory data analysis
- Find and remove any target leakage
- Feature engineering

Interpret & Data Communicate

Implement, Document & Maintain

Model Data 3.

- Variable selection
- Build candidate models
- Model validation and selection

Interpret & Communicate

- Interpret model
 - Communicate model insights

Implement, Document & Maintain 5.

- Set up batch or API prediction system
- Document modeling process for reproducibility
- Create model monitoring and maintenance plan



Define Project Objectives

- Specify business problem
- Acquire subject matter expertise
- Define unit of analysis and prediction target
- Prioritize modeling criteria
- Consider risks and success criteria
- Decide whether to continue

2. Acquire & Explore Data

- Find appropriate data
- Merge data into single table
- Conduct exploratory data analysis
- Find and remove any target leakage
- Feature engineering

Model Data

- Variable selection
- Build candidate models
- Model validation and selection

4. Interpret & Communicate

- Interpret model
 - Communicate model insights

Implement, Document & Maintain

- Set up batch or API prediction system
- Document modeling process for reproducibility
- Create model monitoring and maintenance plan



- Find appropriate data
- Merge data into single table
- Conduct exploratory data analysis
- Find and remove any target leakage
- Feature engineering

I Data

Interpret & Implement, Document & Maintain

Model Data

- Variable selection
- Build candidate models
- Model validation and selection

4. Interpret & Communicate

- Interpret model
 - Communicate model insights

Implement, Document & Maintain

- Set up batch or API prediction system
- Document modeling process for reproducibility
- Create model monitoring and maintenance plan



Define Project Objectives

- Specify business problem
- Acquire subject matter expertise
- Define unit of analysis and prediction target
- Prioritize modeling criteria
- Consider risks and success criteria
- Decide whether to continue

2. Acquire & Explore Data

- Find appropriate data
- Merge data into single table
- Conduct exploratory data analysis
- Find and remove any target leakage
- Feature engineering

Model Data

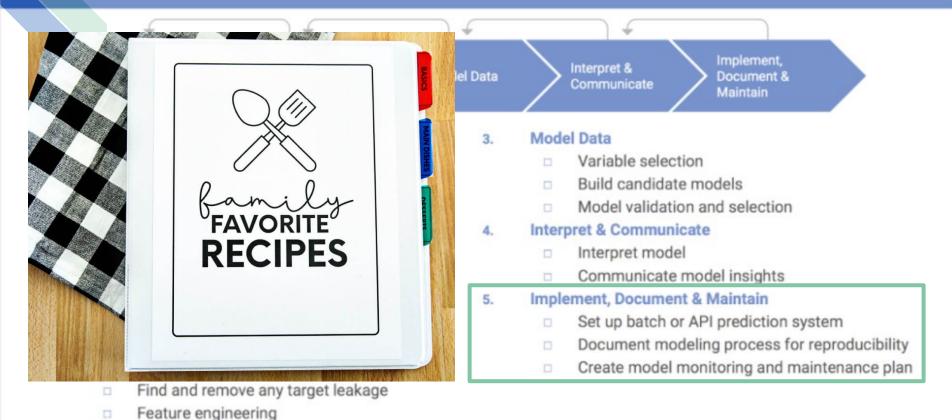
- Variable selection
- Build candidate models
- Model validation and selection

4. Interpret & Communicate

- Interpret model
 - Communicate model insights

5. Implement, Document & Maintain

- Set up batch or API prediction system
- Document modeling process for reproducibility
- Create model monitoring and maintenance plan



DataRobot, Inc. Confidential, All rights reserved.

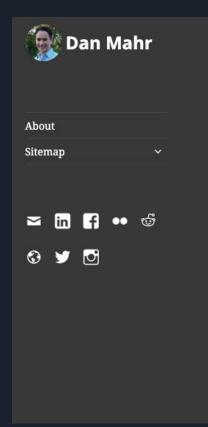
Cooking or Baking a Meal

→ Machine Learning Life Cycle

Studying and taking exams



One successful strategy



Seven-Week GRE Study Plan

After successfully completing the GRE last year, I posted my seven-week GRE study plan on the /r/GREhelp subreddit. Since that post is now locked, I've reproduced it here, along with a few other tips sent in direct messages.

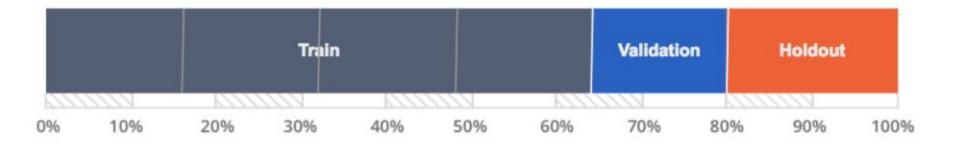
I started studying about seven weeks before my test date, and studied in two phases. In the first phase (weeks 1-4), I mostly followed the Magoosh 1-month study plan and completed all of the associated lesson videos and practice problems, as well as 13/20 vocab flashcards decks. In the second phase (weeks 5-7), I did ETS/Manhattan/Magoosh practice problems, reviewed vocab flash cards, and took five practice tests. You can follow along with in the studying tracking spreadsheet; just Make a Copy to edit, and then enter your test date in cell D52.

How can this relate to machine learning?

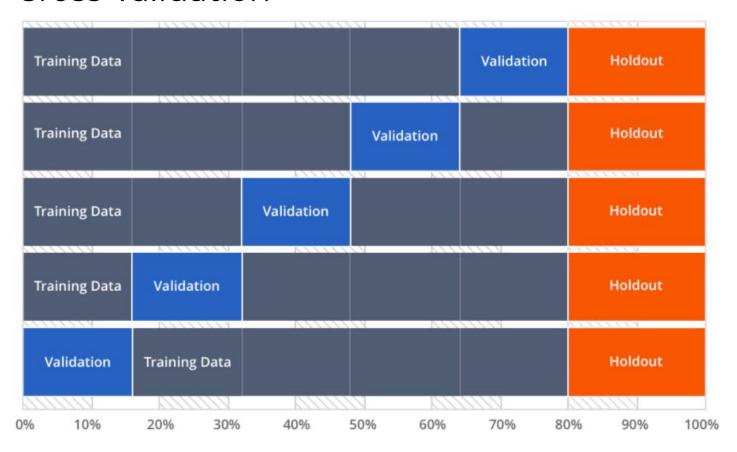




Training / Validation / Holdout



Cross-validation



Studying for an Exam

→ Cross-validation

R and RStudio

R: Engine



RStudio: Dashboard



R packages

R: A new phone



R Packages: Apps you can download



ModernDive.com

The R Series

Statistical Inference via Data Science

A ModernDive into R and the Tidyverse



Chester Ismay Albert Y. Kim



Thank you!



Dr. Chester Ismay

chester.ismay@datarobot.com

https://chester.rbind.io

Slides available at

http://bit.ly/ismay-analogies-psu