CSCI E-96

Data Mining Workflow

Data Types & EDA

Agenda

Start	End	Item
		Week 2 Review

- "C_mapping_inR.R"
- ~~"D_R objects.R"~~
- "E_EDA work.R"

Agenda

Start	End	Item
		Data For Modeling
		Data Mining Workflows
		Data Modification & Preprocessing
		Business Case: Donor Bureau
		Housekeeping, Reading & Homework

Data Structure for Analysis & Modeling

Often the 1st Column is a unique identifier but the identifier could also be a row attribute (not actually a vector)

name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
100%_Bran	N	C	70	4	1	130	10	5	6	280	25	3	1	0.33	68.40297
100%_Natural_Bran	Q	C	120	3	5	15	2	8	8	135	0	3	1	1	33.98368
All-Bran	K	C	70	4	1	260	9	7	5	320	25	3	1	0.33	59.42551
All-Bran_with_Extra_Fiber	K	C	50	4	0	140	14	8	0	330	25	3	1	0.5	93.70491
Almond_Delight	R	C	110	2	2	200	1	14	8		25	3	1	0.75	34.38484
Apple_Cinnamon_Cheerios	G	C	110	2	2	180	1.5	10.5	10	70	25	1	1	0.75	29.50954
Apple_Jacks	K	C	110	2	0	125	1	11	14	30	25	2	1	1	33.17409
Basic_4	G	C	130	3	2	210	2	18	8	100	25	3	1.33	0.75	37.03856
Bran_Chex	R	C	90	2	1	200	4	15	6	125	25	1	1	0.67	49.12025
Bran_Flakes	P	C	90	3	0	210	5	13	5	190	25	3	1	0.67	53.31381
Cap'n'Crunch	Q	C	120	1	2	220	0	12	12	35	25	2	1	0.75	18.04285

Generally we will use data frames to avoid complexity but you will be exposed to other data types.

Data Structure for Analysis & Modeling

Informative features are usually independent & do not lend information to other rows (auto-correlation). Can be called informative columns, independent variables, or features.
Remember in a DF, these can be mixed with decimals, integers, factors, strings, T/F.

name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
100%_Bran	N	C	70	4	1	130	10	5	6	280	25	3	1	0.33	68.40297
100%_Natural_Bran	Q	C	120	3	5	15	2	8	8	135	0	3	1	1	33.98368
All-Bran	K	C	70	4	1	260	9	7	5	320	25	3	1	0.33	59.42551
All-Bran_with_Extra_Fiber	K	C	50	4	0	140	14	8	0	330	25	3	1	0.5	93.70491
Almond_Delight	R	C	110	2	2	200	1	14	8		25	3	1	0.75	34.38484
Apple_Cinnamon_Cheerios	G	C	110	2	2	180	1.5	10.5	10	70	25	1	1	0.75	29.50954
Apple_Jacks	K	C	110	2	0	125	1	11	14	30	25	2	1	1	33.17409
Basic_4	G	C	130	3	2	210	2	18	8	100	25	3	1.33	0.75	37.03856
Bran_Chex	R	C	90	2	1	200	4	15	6	125	25	1	1	0.67	49.12025
Bran_Flakes	P	C	90	3	0	210	5	13	5	190	25	3	1	0.67	53.31381
Cap'n'Crunch	Q	C	120	1	2	220	0	12	12	35	25	2	1	0.75	18.04285

Generally we will use data frames to avoid complexity but you will be exposed to other data types.

Data Structure for Analysis & Modeling

If we are doing supervised learning, there is a dependent variable.

This is the outcome and is “dependent” on the informative columns. An analysis with this vector can be binary, classification, or predictive.

name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
100%_Bran	N	C	70	4	1	130	10	5	6	280	25	3	1	0.33	68.40297
100%_Natural_Bran	Q	C	120	3	5	15	2	8	8	135	0	3	1	1	33.98368
All-Bran	K	C	70	4	1	260	9	7	5	320	25	3	1	0.33	59.42551
All-Bran_with_Extra_Fiber	K	C	50	4	0	140	14	8	0	330	25	3	1	0.5	93.70491
Almond_Delight	R	C	110	2	2	200	1	14	8		25	3	1	0.75	34.38484
Apple_Cinnamon_Cheerios	G	C	110	2	2	180	1.5	10.5	10	70	25	1	1	0.75	29.50954
Apple_Jacks	K	C	110	2	0	125	1	11	14	30	25	2	1	1	33.17409
Basic_4	G	C	130	3	2	210	2	18	8	100	25	3	1.33	0.75	37.03856
Bran_Chex	R	C	90	2	1	200	4	15	6	125	25	1	1	0.67	49.12025
Bran_Flakes	P	C	90	3	0	210	5	13	5	190	25	3	1	0.67	53.31381
Cap'n'Crunch	Q	C	120	1	2	220	0	12	12	35	25	2	1	0.75	18.04285

Generally we will use data frames to avoid complexity but you will be exposed to other data types.

Agenda

Start	End	Item
		Data For Modeling
		Data Mining Workflows
		Data Modification & Preprocessing
		Business Case: Donor Bureau
		Housekeeping, Reading & Homework

Modeling Process

SEMMA (from SAS)

1. Sample
2. Explore
3. Modify
4. Model
5. Assess

In this course…

In this course, most data sets are curated for you.

As part of your course analysis you perform basic exploratory data analysis (EDA)

Data cleanup, Dimension Reduction, Feature Engineering & Feature Enrichment

Regression, Logistic Regression, KNN, Decision Trees, Random Forest etc.

There are many ways to evaluate a model. We will cover specific KPI and business implications.

The first day we discussed data mining and structure. Week 2 was devoted to basics of R then how to sample, explore and visualize data. Now we will modify data for modeling!

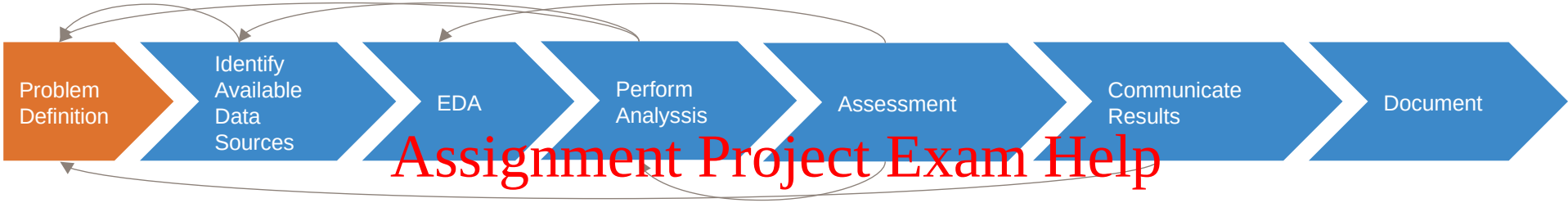
In addition to SEMMA, Data Mining (from the book)

1. Define/understand purpose
2. Obtain data (may involve random sampling)
3. Explore, clean, pre-process data
4. Reduce the data; if supervised DM, partition it
5. Specify task (classification, clustering, etc.)
6. Choose the techniques (regression, CART, neural networks, etc.)
7. Iterative implementation and “tuning”
8. Assess results – compare models
9. Deploy best model

The book’s process is excellent but is focused largely on the modeling process not how the process is part of a business context or if the effort doesn’t require a model.

Data Mining in a Complete Business Workflow

Iterative Business Data Mining Project Life Cycle



1. **Problem Formulation**
2. **Define data requirements**
3. **Explore the data**

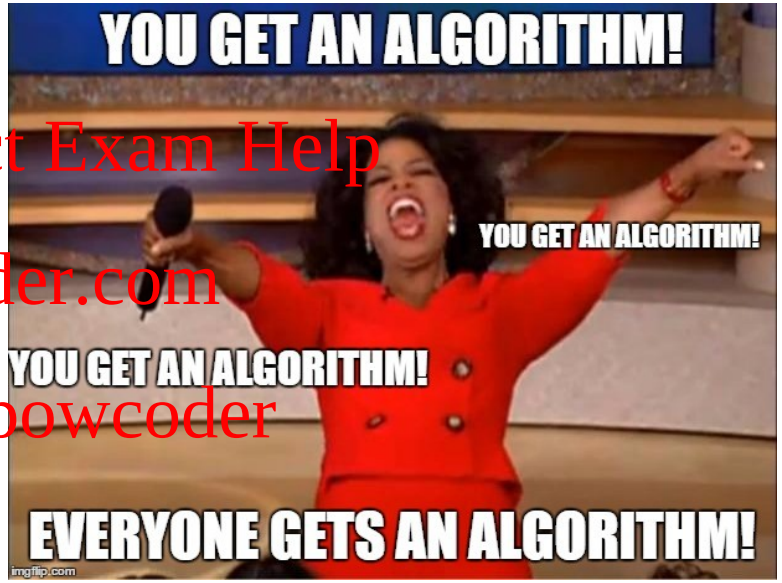
4. **Perform Analysis & Create Project Artifacts**
5. **Asses/Adjust the Project Artifacts**
6. **Communicate Results**
7. **Document to make it repeatable**
8. **Deploy & Monitor**

In this view the steps of a project are not solely for modeling, more iterative and not in isolation because the results are communicated to stakeholders.

What is a model?

An model is a set of rules governing actions or phenomena.

- **Empirical Support without math**
  - Your brain:
    - **Algo#1**: “fur” =Y, “tail”=Y, “claws”=Y, “meow”=Y therefore high probability that’s a cat
    - **Algo#2**: “another meme”=Y, “short & bald professor”=Y therefore “professor trying to hard to be cool”

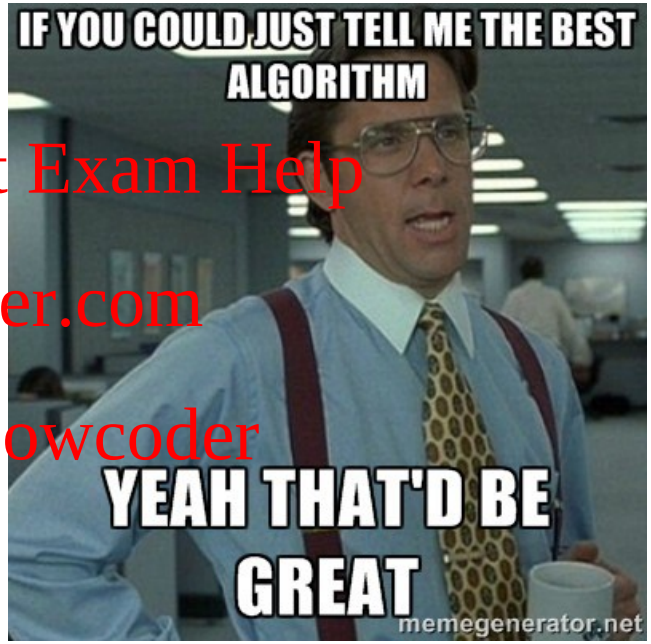


Algorithms or mental models of reality can be correct or lead you to incorrect assumptions.

What is a model?

A model is a set of rules governing actions or phenomena.

- **Empirical Support w/math**
  - *The way you instruct the rules to be constructed is the algorithm (KNN, RF, LogReg etc)*
  - Computers can learn complex representations of phenomena



Just like mental algorithms, the observations we give a mathematical algorithm which result in a final model (set of rules) can lead to correct or incorrect assumptions. "Garbage in…Garbage out."

Vocabulary

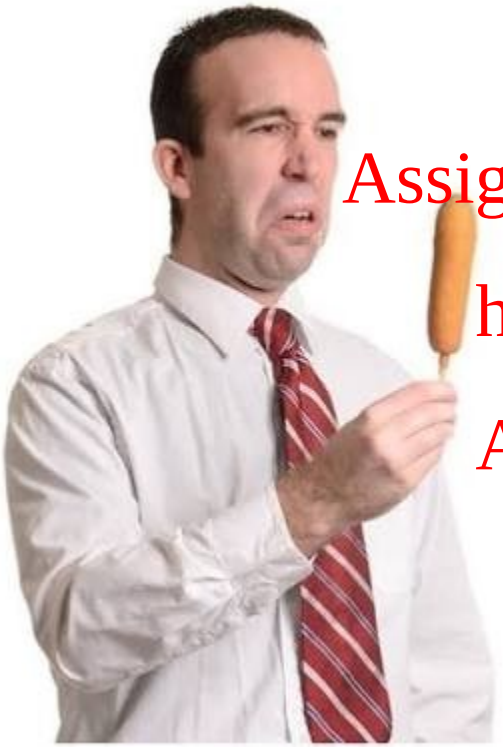
What is a model?	What is an algorithm?
A set of rules governing actions, or describing phenomena. Some representation of reality.	The method you choose to construct the model rules. The way you expect to learn about reality.

The algorithm produces the model.

Don’t be Dale!

Force stakeholders to agree on a problem, plan and output.
Put together a plan, adjust the plan as needed…but have a plan!!

Common Pitfalls



- State the problem in business terms. Avoid open ended questions or just curiosity analysis.
- When modeling avoid multi-collinearity (more on that later)
  - Don't measure things twice, like Fahrenheit and Celsius in the same data set
  - Understand data integrity & human behavior
  - Customers filling out surveys are already self selecting & therefore biased compared to *all* customers
  - Employees that are on an improvement plan may leave voluntarily rather than get fired. Data may record them as voluntarily leaving but that may mask the real issue
- Beware of perfection
  - Perfectly aligned analyses and SME expectations seldom occur
  - Model perfection likely means you have an error (data integrity)
- Beware of data leak & misunderstanding of the project!
  - Predicting hospital readmissions: use historical record of whether a person dies in the hospital…so obviously they wouldn't readmit.

Agenda

Start	End	Item
		Data For Modeling
		Data Mining Workflows
		Data Modification & Preprocessing
		Business Case: Donor Bureau
		Housekeeping, Reading & Homework

Modeling Process

SEMMA (from SAS)	In this course…
1. Sample	In this course, most data sets are curated for you.
2. Explore	As part of your course analysis you perform basic exploratory data analysis (EDA)
3. Modify	Data cleanup, Dimension Reduction, Feature Engineering & Feature Enrichment
4. Model	Regression, Logistic Regression, KNN, Decision Trees, Random Forest etc.
5. Assess	There are many ways to evaluate a model. We will cover specific KPI and business implications.

Often data is sampled from a large database so you can more quickly explore, apply methods and prototype before reassessing on full data.

PreProcessing

Many algorithms cannot accept the data directly.
Thus you must preprocess your data before training.

Common Pre-Processing Categorical Variables	Common Pre-Processing Numeric Variables
• Dummy Variables • Binning Low Frequency Levels • Changing to Numeric for Ordinal • Deal with Missing Levels	• Deal with Missing • Outlier Detection • Binning

The book does these steps in a traditional manner but we will use an easier method called variable treatment (vtreat). Review the book if you want to see the manual methods.

Dummy Variables

Always "1 Less than the State of Nature"

2 States of Nature = 1 Light Switch or "dummy variable"

- Dog or Cat needs one column where dog =1, cat=0

If the column has a 0 then its definitely a cat.

Represent non-numeric information as light switches.

Dummy Variables

Always "1 Less than the State of Nature"

3 States of Nature = 2 Light Switches or "dummy variables"

- Dog, Cat, Fish: Column A dog =1, cat=0, fish=0 & Column B dog=0, cat=1, fish=0.

If both columns are 0 then by default, this is the same information as the row having a fish.

Represent non-numeric information as light switches.

Dummy Variables

Dummy Variables represent category levels as 1/0 within new vectors. For some approaches, this lets the algorithm understand the information.

Consider this data:

CuID	Gender	Profession	Affiliation
CU43E439	Male	Nurse	Democrat
CU36E506	Male	Teacher	Independent
CU91E65	Female	Manager	Republican
CU17E255	Male	Student	Democrat
CU27E792	Female	Executive	
CU74E430	Female	Manager	Republican
CU25E466	Female	Teacher	Democrat

Dummy Variables

Dummy Variables represent category levels as 1/0 within new vectors. For some approaches, this lets the algorithm understand the information.

Consider this data:

CuID	Gender	Profession	Affiliation
CU43E439	Male	Nurse	Democrat
CU36E506	Male	Teacher	Independent
CU91E65	Female	Manager	Republican
CU17E255	Male	Student	Democrat
CU27E792	Female	Executive	
CU74E430	Female	Manager	Republican
CU25E466	Female	Teacher	Democrat

Through EDA you realize:

- You wouldn't use the ID var for training
- Gender has 2 levels
- Profession has 5 levels
- Affiliation has 3 levels & missing

Dummy Variables

Always make dummy variables “1 less than the state of the data nature”

Consider this data:

CuID	Gender	Profession	Affiliation
CU43E439	Male	Nurse	Democrat
CU36E506	Male	Teacher	Independent
CU91E65	Female	Manager	Republican
CU17E255	Male	Student	Democrat
CU27E792	Female	Executive	
CU74E430	Female	Manager	Republican
CU25E466	Female	Teacher	Democrat

CuID	Gender__Male	Profession_Nurse	Profession_Teacher	Profession_Manager	Profession_Student	Affiliation_D	Affiliation_I	Affiliation_Missing
CU43E439	1	1	0	0	0	1	0	0
CU36E506	1	0	1	0	0	0	1	0
CU91E65	0	0	0	1	0	0	0	0
CU17E255	1	0	0	0	1	1	0	0
CU27E792	0	0	0	0	0	0	0	1
CU74E430	0	0	0	1	0	0	0	0
CU25E466	0	0	1	0	0	1	0	0

A light switch has 2 states, on/off, yet you only need 1 switch. The same is true as more levels are added, you don’t need one for each level.

Dummy Variables

Always make dummy variables “1 less than the state of the data nature”

Dummy Data:

CuID	Gender__Male	Profession_Nurse	Profession_Teacher	Profession_Manager	Profession_Student	Affiliation_D	Affiliation_I	Affiliation_Missing
CU43E439	1	1	0	0	0	1	0	0
CU36E506	1	0	1	0	0	0	1	0
CU91E65	0	0	0	1	0	0	0	0
CU17E255	1	0	0	0	1	1	0	0
CU27E792	0	0	0	0	0	0	0	1
CU74E430	0	0	0	1	0	0	0	0
CU25E466	0	0	1	0	0	1	0	0

Applying Judgment:

CuID	Gender__Male	Profession_LowCount	Profession_Teacher	Profession_Manager	Affiliation_D	Affiliation_MissingOther
CU43E439	1	1	0	0	1	0
CU36E506	1	0	1	0	0	1
CU91E65	0	0	0	1	0	0
CU17E255	1	1	0	0	1	0
CU27E792	0	0	0	0	0	1
CU74E430	0	0	0	1	0	0
CU25E466	0	0	1	0	1	0

Never throw the “kitchen sink” at an algo, exercise your problem knowledge to reduce the number of vectors.

Numeric Variables

Numeric variables need to be examined, corrected and missing flags need to be created.

Consider this data:

Model	mpg	cyl	disp	hp	drat	wt	qsec
Mazda RX4	21	6	160	110	3.9	2.62	
Mazda RX4 Wag	21	6	160	-110	3.9	2.875	17.02
Datsun 710	122.8	NA	108	93	3.85	2.32	18.61
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44
Hornet Sportabout	18.7	8	?	175	3.15	3.44	17.02
Valiant	18.1	6	225	105	2.76	NA	20.22
Duster 360	14.3	8	360	245	3.21	3.57	15.84
Merc 240D	24.4	4	146.7	62	3.69	3.19	
Merc 230	22.8	4	140.8	95	3.92	3.15	22.9
Merc 280	19.2	6	167.6	123	3.92	3.44	18.3

With domain expertise you realize:

- Its unlikely a car has 122mpg
- No way -110 horsepower!

Through EDA you realize:

- Missing values are blank, NA, and “?”

Outlier Numeric Variables

Outliers can be removed or the values can be replaced with imputation.

Consider this data:

Model	mpg	cyl	disp	hp	drat	wt	qsec
Mazda RX4	21	6	160	110	3.9	2.62	
Mazda RX4 Wag	21	6	160	-110	3.9	2.875	17.02
Datsun 710	122.8	NA	108	93	3.85	2.32	18.61
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44
Hornet Sportabout	18.7	8	?	175	3.15	3.44	17.02
Valiant	18.1	6	225	105	2.76	NA	20.22
Duster 360	14.3	8	360	245	3.21	3.57	15.84
Merc 240D	24.4	4	146.7	62	3.69	3.19	
Merc 230	22.8	4	140.8	95	3.92	3.15	22.9
Merc 280	19.2	6	167.6	123	3.92	3.44	18.3

Drop a row if:

- You have many other records for training
- The record contains multiple integrity issues

Impute (use a method to change the value):

- **Hotdeck** – choose a random value in the vector say 19.2
- **Mean** Imputation – mean avg of vector
- **Median** Imputation – median avg
- Train an **algorithm** to fill in the values (KNN)

Detecting Outliers

- An outlier is an observation that is “extreme”, being distant from the rest of the data (definition of “distant” is deliberately vague)
- Outliers can have disproportionate influence on models
- An important step in data pre-processing/EDA is detecting outliers
- Once detected, domain knowledge is required to determine if it is an error, or truly extreme.
  - Correct them to a more normal (avg) value?
  - Remove the record altogether?

Detecting Outliers

- In some contexts, finding outliers is the purpose of the DM exercise (airport security screening). This is called “anomaly detection”

Missing Numeric Variables

Imputation is a best practice over to fill in for missing numeric values. Usually start with an easy method like median or hotdeck.

Model	mpg	FixMPG
Mazda RX4	21	0
Mazda RX4 Wag	21	0
Datsun 710	20.1	1
Hornet 4 Drive	21.4	0
Hornet Sportabout	18.7	0
Valiant	18.1	0
Duster 360	14.3	0
Merc 240D	24.4	0
Merc 230	22.8	0
Merc 280	19.2	0

Drop a row if:

- You have many other records for training
- The record contains multiple integrity issues

Impute (use a method to change the value):

- **Hotdeck**
- **Mean**
- **Median**
- **Algorithm** (KNN)

Original Data

Imputation through domain expertise can be VERY time consuming but is sometimes worth it though not often.

Model	mpg	cyl	disp	hp	drat	wt	qsec
Mazda RX4	21	6	160	110	3.9	2.62	
Mazda RX4 Wag	21	6	160	-110	3.9	2.875	17.02
Datsun 710	122.8	NA	108	93	3.85	2.32	18.61
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44
Hornet Sportabout	18.7	8	?	175	3.15	3.44	17.02
Valiant	18.1	6	225	105	2.76	NA	20.22
Duster 360	14.3	8	360	245	3.21	3.57	15.84
Merc 240D	24.4	4	146.7	62	3.69	3.19	
Merc 230	22.8	4	140.8	95	3.92	3.15	22.9
Merc 280	19.2	6	167.6	123	3.92	3.44	18.3

Drop a row if:

- You have many other records for training
- The record contains multiple integrity issues

Impute (use a method to change the value):

- **Hotdeck**
- **Mean**
- **Median**
- **Algorithm** (KNN)

Missing Numeric Variables

Imputation through domain expertise can be VERY time consuming but is sometimes worth it though not often.

Consider this data:

Model	mpg	cyl	disp	hp	drat	wt	qsec
Mazda RX4	21	6	160	110	3.9	2.62	17.02
Mazda RX4 Wag	21	6	160	110	3.9	2.875	17.02
Datsun 710	20.1	8	108	93	3.85	2.32	18.61
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44
Hornet Sportabout	18.7	8	146.7	175	3.15	3.44	17.02
Valiant	18.1	6	225	105	2.76	3.46	20.22
Duster 360	14.3	8	360	245	3.21	3.57	15.84
Merc 240D	24.4	4	146.7	62	3.69	3.19	22.9
Merc 230	22.8	4	140.8	95	3.92	3.15	22.9
Merc 280	19.2	6	167.6	123	3.92	3.44	18.3

Impute (use a method to change the value):

- -110 looks like a data entry issue & other Mazda has 110.
- 146.7 is the average etc

Missing Flags

Add missing indicator dummy variables similar to the categorical exercise.

Consider this data:

Model	mpg	cyl	disp	hp	drat	wt	qsec	missing MPG	missing Cyl	missing Disp	Adjust_ HP	missing WT	missing QSec
Mazda RX4	21	6	160	110	3.9	2.62	17.02	0	0	0	0	0	1
Mazda RX4 Wag	21	6	160	110	3.9	2.875	17.02	0	0	0	1	0	0
Datsun 710	20.1	8	108	93	3.85	2.32	18.61	1	1	0	0	0	0
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	0	0	0	0	0	0
Hornet Sportabout	18.7	8	146.7	175	3.15	3.44	17.02	0	0	1	0	0	0
Valiant	18.1	6	225	105	2.76	3.15	20.22	0	0	0	0	1	0
Duster 360	14.3	8	360	245	3.21	3.57	15.84	0	0	0	0	0	0
Merc 240D	24.4	4	146.7	62	3.69	3.19	22.9	0	0	0	0	0	1
Merc 230	22.8	4	140.8	95	3.92	3.15	22.9	0	0	0	0	0	0
Merc 280	19.2	6	167.6	123	3.92	3.44	18.3	0	0	0	0	0	0

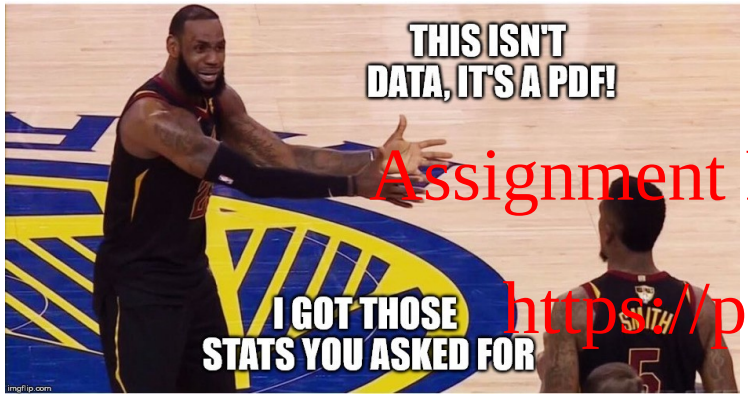
Feature Engineering- Still Pre-Processing!!

Once you have your data identified, collected and organized, you may want to create new vectors using existing data to aid the analysis.

Feature engineering or “feature crossing” is the act of using existing data to form new data inputs for analysis. For example, dividing one data point by another to derive a new data point.

What is Feature Engineering?

Using qualitative or technical expertise to derive new features for machine learning.



Example

Predict basketball team wins:

Raw Data:

Total Rebounds (2017-18): 2749

Total Games: 56

LBJ triple doubles: 10

Engineered

Name	Value	Type	Reasoning
Rebounds per Game	41	Technical	Simple ratio of two team level stats
Percent of Games that Lebron James has a "triple double"	10/56 = 17.8%	Qualitative	Capturing information about the best player's performance e.g. getting double digit stats in 3 of assists, blocks, points, rebounds, or steals

Why is Feature Engineering Effective?

Using qualitative or technical expertise to derive new features for machine learning.

$$f\left(Modeling\ Results\right) = the\ Algo + Parameters + Data\ provided$$

- Better features means **flexibility**.
  - less than optimal models can still yield good results
- Betters features means **simpler models**.
  - less than optimal parameters can still yield good results
- Better features means **better results.**
  - *"The algorithms we used are very standard for Kagglers. We spent most of our efforts in feature engineering." Xavier Conort describing his winning "Flight Quest" submission*
  - *Way to differentiate & squeeze out more accuracy*

Example Feature Engineering Methods

- `library(vtreat)`
  - Automatic variable treatment functions
- `library(acepack)`
  - Alternating conditional expectations for feature importance
- Hand-Coded Variables
  - Subject Matter Experts tell you which variables to interact
- Dimension Reduction - PCA
  - Principle component analysis

Grey not covered
but if interested can share code

Variable Treatment: `library(vtreat)`

Vtreat automates some data cleaning, imputation and engineers specific response encoded variables.


```
> dTrainNTreated <- prepare(treatmentsN,dTrainN,pruneSig=0.99)
> dTrainNTreated
  x_lev_x.a x_lev_x.b  z_clean z_isBAD y
1         1         0 1.000000       0 0
2         1         0 2.000000       0 0
3         1         0 3.000000       0 0
4         1         0 4.000000       0 1
5         0         1 5.000000       0 0
6         0         1 3.666667       1 1
7         0         1 7.000000       0 1
```

```
> dTrainN
  x  z y
1 a  1 0
2 a  2 0
3 a  3 0
4 a  4 1
5 b  5 0
6 b NA 1
7 b  7 1
```

Step 1

Organize a Informative Data

Step 2

```
treatmentsN <-
designTreatmentsN(...)
```

"***Design Treatment***" - Categorical or Numeric

Step 3

"***Prepare***" Data into a treated Modeling Matrix

Vtreat adjusts data in many ways.

Action	Type	Description
Imputation	Cleaning	Na, Nan, Inf replacement
Imputation Indicator	Cleaning	Append a binary column as imputation flag
Dummy Variables	Cleaning	Create dummy variables for categorical variables
Constant Attribute Suppression	Cleaning	Remove variables with a single value.
Level Deviation –"cat_D"	Engineering	A "deviation fact" about a categorical level. Tells us if 'y' is concentrated or diffuse when conditioned on the observed level of the original categorical variable.
Level Prevalence- "cat_P"	Engineering	A "prevalence fact" about a categorical level. Tells us if the original level was rare or common.
Estimated Single Variable Effects – • **"cat_B" = categorical outcome w/Bayesian** • **"cat_N" = numeric outcome w/Regression**	Engineering	A single variable Bayesian model of the change in logit-odds in outcome from mean distribution conditioned on the observed value of the original variable. A single variable regression model of the difference in outcome expectation conditioned on the observed value of the original variable.
Rare Cats*	Engineering	For categorical levels below a frequency threshold, pool different levels into a common "rare-level" variable

Mean Imputation - PreProcessing

Most of the time we will use vtreat to clean data…its faster and easier.

Vtreat Actions	Common Name
NA, NAN, and Infinity replaced with mean	Mean imputation

- When a numeric value is missing
  - Replace that value with the mean average
- Not applicable for categorical variables


```
> dTrainN
   x  z y
1  a  1 0
2  a  2 0
3  a  3 0
4  a  4 1
5  b NA 0
6  b  6 1
7  b  7 1
```

Design a treatment plan and apply it.

```
> dTrainNTreated
  z_clean z_isBA
1 1.000000
2 2.000000
3 3.000000
4 4.000000
5 3.833333
6 6.000000
7 7.000000
```

Missing Flags- PreProcessing

Vtreat Actions	Common Name
Indicator	Missing indicator

- When a numeric value is missing
  - Add a missing flag variable
- When a factor level is missing
  - Add a missing flag variable

Dummy Variables - PreProcessing

Vtreat Actions	Common Name
Indicator variables	Dummy variable

- Not applicable for numeric values
- For each factor level, will create dummy variables


```
> dTrainN
  x  z y
1 a  1 0
2 a  2 0
3 a  3 0
4 a  4 1
5 b NA 0
6 b  6 1
7 b  7 1
```

Design a treatment plan and apply it.

x_lev_x_a	x_lev_x_b	y
1	0	0
0	0	0
1	0	0
1	0	1
0	1	0
0	1	1
0	1	1

Beware! Since some algorithms don't care, vtreat will return ALL dummy variables and not drop one to represent all 0s.

Vtreat Engineered CAT Variables Example

Simple Calculations

CatN (Numeric):
Level A value =1: (1/3) minus Overall Y value = 1: (3/7)
= -0.0952381

CatD (Deviation):
Level B sd(c(0,1)) = 0.7071068

Cat P (Prevalence):
Level C Occurred (2/7) = 0.2857143

```
> df
  x y
1 a 0
2 a 0
3 c 0
4 a 1
5 b 0
6 b 1
7 c 1
```

Step 1

Organize a Modeling Matrix

Step 2

"***Design Treatment***" - Categorical or Numeric

```
treatmentsN <-
designTreatmentsN(...)
```

Step 3*

"***Prepare***" Data into a treated Modeling Matrix

**abridged data shown*

```
     x_catP      x_catN    x_catD y
0.4285714 -0.09523810 0.5773503 0
0.4285714 -0.09523810 0.5773503 0
0.2857143  0.07142857 0.7071068 0
0.4285714 -0.09523810 0.5773503 1
0.2857143  0.07142857 0.7071068 0
0.2857143  0.07142857 0.7071068 1
0.2857143  0.07142857 0.7071068 1
```

Non-Informative Check

Vtreat Actions	Common Name
Constant/Near Constant	Suppress uninformative variables

```
> dTrainN
  x   z y
1 a 2.4 0
2 a 2.4 0
3 c 2.4 0
4 a 2.4 1
5 b 2.4 0
6 b 2.4 1
7 c 2.4 1
```

In this fake example, since z contains a constant value an algorithm can't learn from it.

Z is automatically dropped.

```
> dTrainNTreated
     x_catP      x_catN    x_catD x_lev_x_a x_lev_x_b x_lev_x_c y
1 0.4285714 -0.09523810 0.5773503         1         0         0 0
2 0.4285714 -0.09523810 0.5773503         1         0         0 0
3 0.2857143  0.07142857 0.7071068         0         0         1 0
4 0.4285714 -0.09523810 0.5773503         1         0         0 1
5 0.2857143  0.07142857 0.7071068         0         1         0 0
6 0.2857143  0.07142857 0.7071068         0         1         0 1
7 0.2857143  0.07142857 0.7071068         0         0         1 1
```

Vtreat summary

When specifying a plan you need to pass in :

1. The Data
2. The column names of informative variables.
3. The name of the Y or Response Variable

Best Practices:

Almost always drop cat_D variables
When possible design a treatment plan on separate data than the training & test sets
Review the output to ensure coherence, its not a free automated lunch!

REVIEW: Informative Variables

Vtreat needs these names

name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
100%_Bran	N	C	70	4	1	130	10	5	6	280	25	3	1	0.33	68.40297
100%_Natural_Bran	Q	C	120	3	5	15	2	8	8	135	0	3	1	1	33.98368
All-Bran	K	C	70	4	1	260	9	7	5	320	25	3	1	0.33	59.42551
All-Bran_with_Extra_Fiber	K	C	50	4	0	140	14	8	0	330	25	3	1	0.5	93.70491
Almond_Delight	R	C	110	2	2	200	1	14	8		25	3	1	0.75	34.38484
Apple_Cinnamon_Cheerios	G	C	110	2	2	180	1.5	10.5	10	70	25	1	1	0.75	29.50954
Apple_Jacks	K	C	110	2	0	125	1	11	14	30	25	2	1	1	33.17409
Basic_4	G	C	130	3	2	210	2	18	8	100	25	3	1.33	0.75	37.03856
Bran_Chex	R	C	90	2	1	200	4	15	6	125	25	1	1	0.67	49.12025
Bran_Flakes	P	C	90	3	0	210	5	13	5	190	25	3	1	0.67	53.31381
Cap'n'Crunch	Q	C	120	1	2	220	0	12	12	35	25	2	1	0.75	18.04285

REVIEW: Outcome/Target Variable

Vtreat needs this name as the outcome.

name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
100%_Bran	N	C	70	4	1	130	10	5	6	280	25	3	1	0.33	68.40297
100%_Natural_Bran	Q	C	120	3	5	15	2	8	8	135	0	3	1	1	33.98368
All-Bran	K	C	70	4	1	260	9	7	5	320	25	3	1	0.33	59.42551
All-Bran_with_Extra_Fiber	K	C	50	4	0	140	14	8	0	330	25	3	1	0.5	93.70491
Almond_Delight	R	C	110	2	2	200	1	14	8		25	3	1	0.75	34.38484
Apple_Cinnamon_Cheerios	G	C	110	2	2	180	1.5	10.5	10	70	25	1	1	0.75	29.50954
Apple_Jacks	K	C	110	2	0	125	1	11	14	30	25	2	1	1	33.17409
Basic_4	G	C	130	3	2	210	2	18	8	100	25	3	1.33	0.75	37.03856
Bran_Chex	R	C	90	2	1	200	4	15	6	125	25	1	1	0.67	49.12025
Bran_Flakes	P	C	90	3	0	210	5	13	5	190	25	3	1	0.67	53.31381
Cap'n'Crunch	Q	C	120	1	2	220	0	12	12	35	25	2	1	0.75	18.04285

SME – Factor Level Interactions

Domain Specific Knowledge can be applied to factors to create new variables.

Example Data w/Factors → *If we believe x2 & x3 have a qualitative interaction* → *Then concatenating the variables could help the algo learn faster*

```
> df
   id    x1 x2       x3
1   1  Tall  A  January
2   2  Tall  B February
3   3 Short  C    March
4   4  Tall  D  January
5   5  Tall  E February
6   6 Short  A    March
7   7  Tall  B  January
8   8  Tall  C February
9   9 Short  D    March
10 10  Tall  E  January
11 11  Tall  A February
12 12 Short  B    March
13 13  Tall  C  January
14 14  Tall  D February
15 15 Short  E    March
```



```
> df
   id    x1 x2       x3    NewVar
1   1  Tall  A  January  A_January
2   2  Tall  B February  B_February
3   3 Short  C    March    C_March
4   4  Tall  D  January  D_January
5   5  Tall  E February  E_February
6   6 Short  A    March    A_March
7   7  Tall  B  January  B_January
8   8  Tall  C February  C_February
9   9 Short  D    March    D_March
10 10  Tall  E  January  E_January
11 11  Tall  A February  A_February
12 12 Short  B    March    B_March
13 13  Tall  C  January  C_January
14 14  Tall  D February  D_February
15 15 Short  E    March    E_March
```

With factors, you combine by concatenating the levels, capturing the information contained in both levels such as “bald” “male” to “bald_male”

SME – Numeric Interactions

Domain Specific Knowledge can be applied to numeric variables to create new variables.

Example Data w/Numeric Vectors → *If we believe x2 & x3 have a qualitative interaction* → *Then interacting them with simple operators Such as subtraction or division may help the Algorithm learn faster*


```
> df
   id         x1         x2        x3
1   1 -0.88136829  0.2501651 -13.14273
2   2  0.02733313 -1.4333039 -11.74314
3   3 -0.33639521 -1.6865764 -11.68639
4   4 -0.23421101  0.5190674 -12.01037
5   5  0.27393831  0.5816568 -11.84066
6   6  1.41036156 -0.9008710 -12.34022
7   7  1.20575282 -1.1239448 -11.24841
8   8  0.68555598  0.3259345 -13.60888
9   9  2.07185790 -0.8585351 -11.31116
10 10 -0.05889832  0.7912153 -13.02964
```

```
> df
   id         x1         x2        x3 SubtractedNewVar DivisionNewVar
1   1 -0.88136829  0.2501651 -13.14273        13.392900    0.067061255
2   2  0.02733313 -1.4333039 -11.74314        10.309840   -0.002327582
3   3 -0.33639521 -1.6865764 -11.68639         9.999812    0.028785217
4   4 -0.23421101  0.5190674 -12.01037        12.529433    0.019500739
5   5  0.27393831  0.5816568 -11.84066        12.422314   -0.023135397
6   6  1.41036156 -0.9008710 -12.34022        11.439347   -0.114289842
7   7  1.20575282 -1.1239448 -11.24841        10.124465   -0.107193180
8   8  0.68555598  0.3259345 -13.60888        13.934819   -0.050375619
9   9  2.07185790 -0.8585351 -11.31116        10.452620   -0.183169438
10 10 -0.05889832  0.7912153 -13.02964        13.820857    0.004520334
```

X1 X3 Div. New Var

In sports, variables like “turnovers per game” are expert led engineered variables. “per” is interaction through division of two seasonal level statistics.

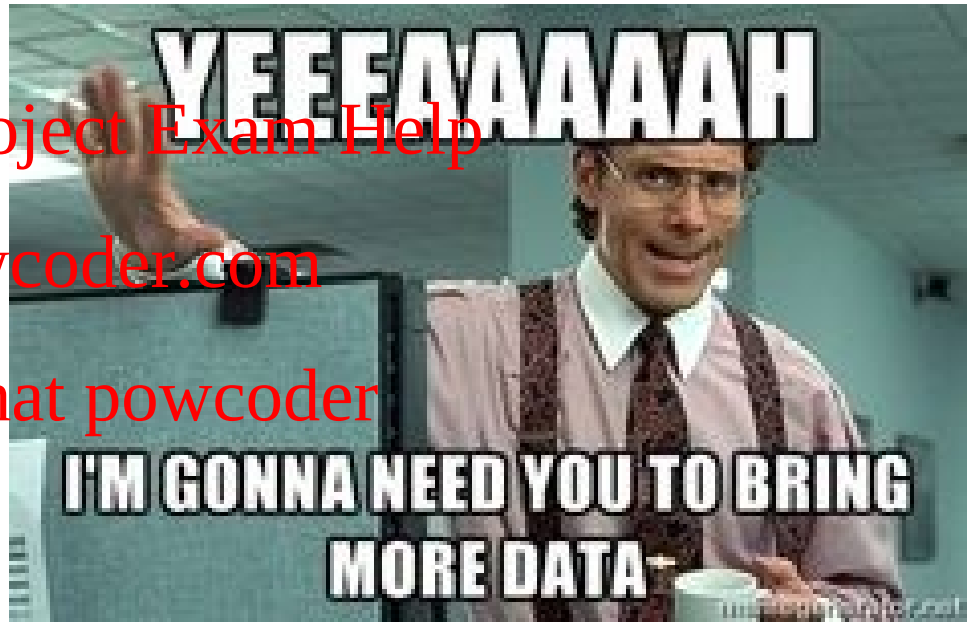
Feature Enrichment

When training a model you can

- Add more training records
  - Find more rows!

OR

- Add more information
  - Find more columns!



Why is Feature Enrichment Effective?

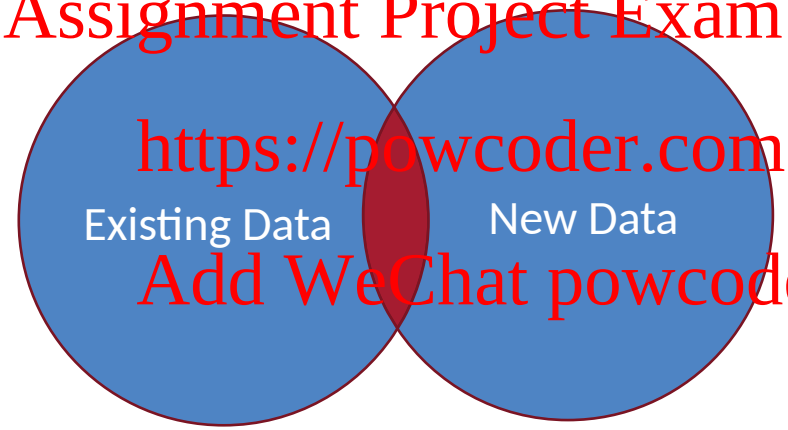
Using qualitative or technical expertise to derive new features for machine learning.

$$f(\textit{Modeling Results}) = \textit{the Algo} + \textit{Parameters} + \textit{Data provided}$$

- Better features means **flexibility**.
  - less than optimal models can still yield good results
- Betters features means **simpler models**.
  - less than optimal parameters can still yield good results
- Better features means **better results.**
  - *"The algorithms we used are very standard for Kagglers. We spent most of our efforts in feature engineering." Xavier Conort describing his winning "Flight Quest" submission*
  - *Way to differentiate & squeeze out more accuracy*

Data Enrichment aids Model Performance

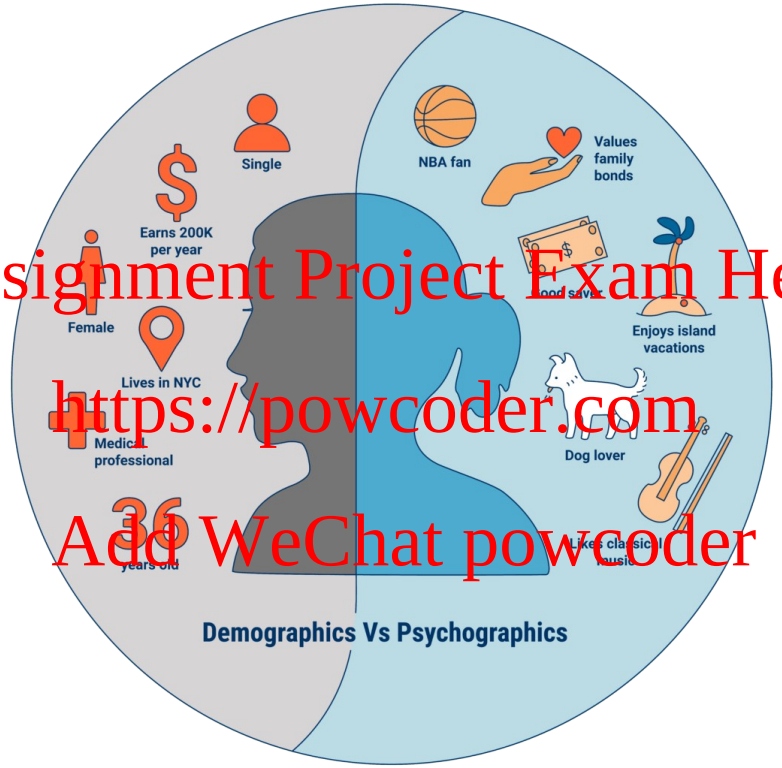
Feature Enrichment is the act of adding new information to your dataset. You are enriching your existing data, often with public or 3rd party data.



There are limits to what an organization has internally for data.
Companies exist solely to enrich data sources.

An example of Data Enrichment

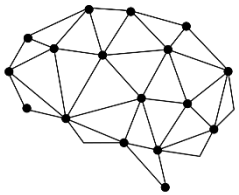
An organization may collect some information about employees during interactions.



They could buy data about employees to have a more complete picture.

https://www.cbinsights.com/research/what-is-psychographics/

Modeling with Feature Enrichment is Widespread



Cambridge Analytica

Used social media information of individual and friends to model voting tendencies

evariant
MOVING HEALTHCARE AHEAD

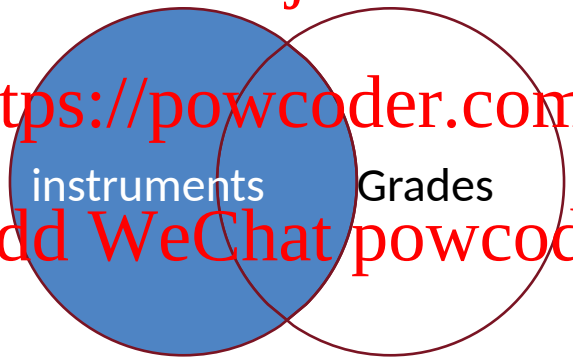
Uses 3rd party data to predict if a household member has diabetes for marketing.

Most consumers are not aware the type and amount of data that is available about them.

HARVARD UNIVERSITY

Feature Enrichment Requires a Join

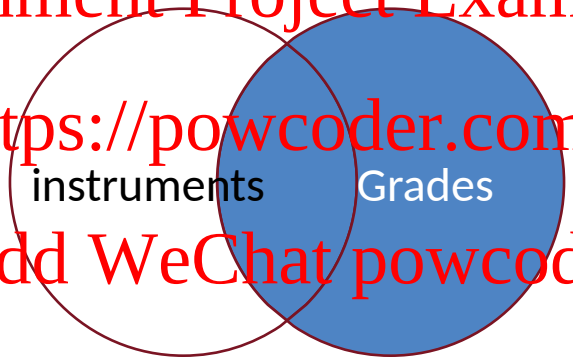
Left Join



Return all rows from x where there are matching values in y, and all columns from x and y. *If there are multiple matches between x and y, all combination of the matches are returned.*

Feature Enrichment Requires a Join

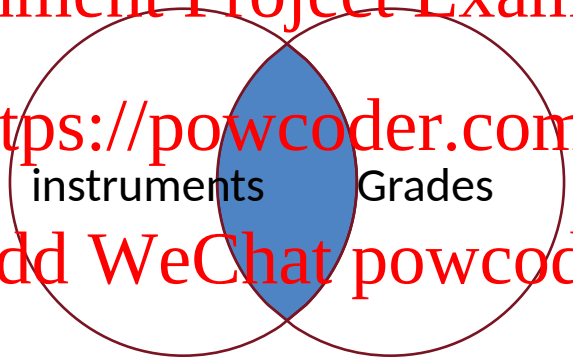
Right Join



Return all rows from y, and all columns from x and y. Rows in y with no match in x will have NA values in the new columns. If there are multiple matches between x and y, all combinations of the matches are returned.

Feature Enrichment Requires a Join

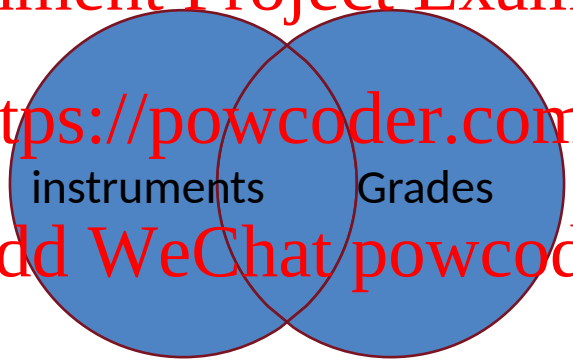
Inner Join



Return all rows from x where there are matching values in y, and all columns from x and y. If there are multiple matches between x and y, all combination of the matches are returned.

Feature Enrichment Requires a Join

Full Join



Return all rows and all columns from both x and y. Where there are not matching values, returns NA for the one missing.

Open A_Joins.R

Agenda

Start	End	Item
		Review in Context
		Data Mining Workflows
		Data Modification & Preprocessing
		Business Case: Donor Bureau
		Housekeeping, Reading & Homework

Meet Donor Bureau

Company Overview

- Founded in 2011
  - DonorBureau spun out of a direct mail agency

- Works with non-profits to optimize their direct mail fundraising

- Data Stats:
  - 900M past mail transactions
  - 140M past donations
  - 40M individuals

Modeling Problem

- Two modeling problems
  - What acquisition targets should NOT get mail?
  - What donor cultivation targets should get mail?
- Acquisition modeling
  - Drop bottom 20% of list to be mailed
  - Goal is to have the top 80% gross twice as much as the bottom 20%
- Cultivation modeling:
  - Add names to the monthly 10k piece cultivation mailings
  - Find 2k additional names per month to mail from a 40k donor file

Let's Practice

Open B_DataPreProcessingEngineering:

Variable	Description
rowID	Appended Row ID - noninformative
uniqueID	Unique Identifier – noninformative
Zip	Zip code group to anonymize
Homeowner	Y/N if they own a hom
NUMCHLD	Number of children in house
Income	Household Income
Gender	M/F
Wealth	Wealth rating uses median family income and population statistics from each area to index relative wealth within each state. The segments are denoted 0-9, with 9 being the highest wealth group and zero being the lowest. Each rating has a different meaning within each state.

Variable	Description
HV	Avg Home Value in potential donor's neighborhood in hundreds of dollars.
Icmed	Median Income in potential donor's neighborhood in hundreds of dollars.
Icavg	Avg Family Income in potential donor's neighborhood in hundreds of dollars.
IC15	% earning less than $15K in potential donor's neighborhood.
NUMPROM	Lifetime # of promotions received to date
RAMNTALL	Dollar amount of lifetime gifts to date.
MAXRAMNT	Dollar amount of largest gift to date.
LASTGIFT	Dollar amount of most recent gift.
TOTALMONTHS	Number of months from last donation
TIMELAG	Number of months between first and second gift.
AVGGIFT	Average dollar amount of gifts to date.
Y1_Donation	Y/N did they donate
Y2_DonatedAmt	Dollar Amt of donation

Agenda

Start	End	Item
		Review in Context
		Data Mining Workflows
		Data Modification & Preprocessing
		Business Case: Donor Bureau
		Housekeeping, Reading & Homework

Housekeeping , Reading & Homework

- Keep posting Piazza questions
- Make sure you can knit and submit correctly, better to solve it now than struggle all semester

- Chapter 6

- Week 2 HW script…**moved out, see Canvas for adjusted date**
- OKCupid Case **DUE (see syllabus)**
  - All files including slides, code (if applicable) & video