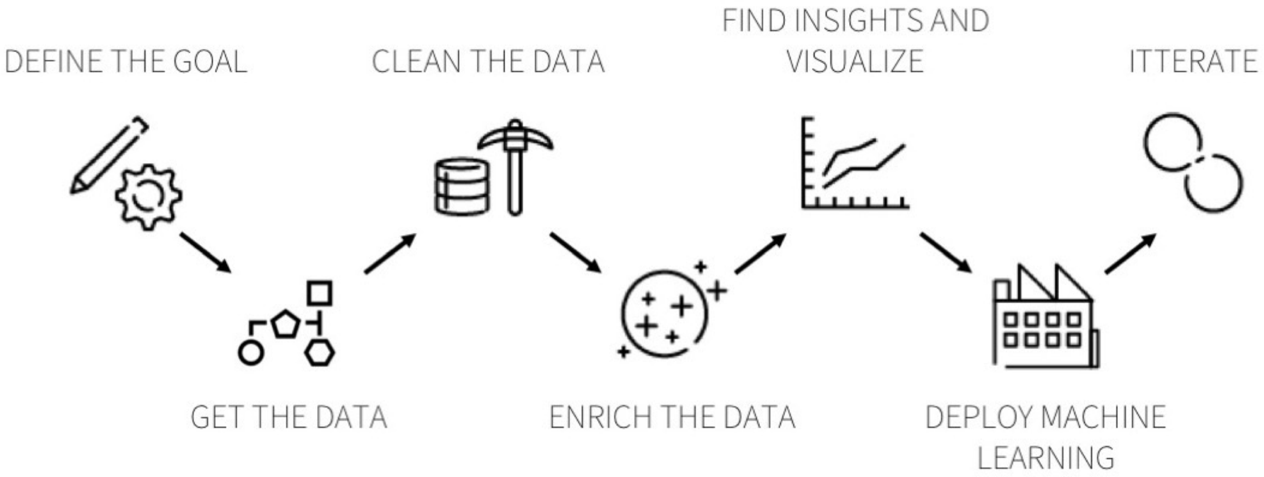
Class 3: Data Preparation

January 27nd, 2020

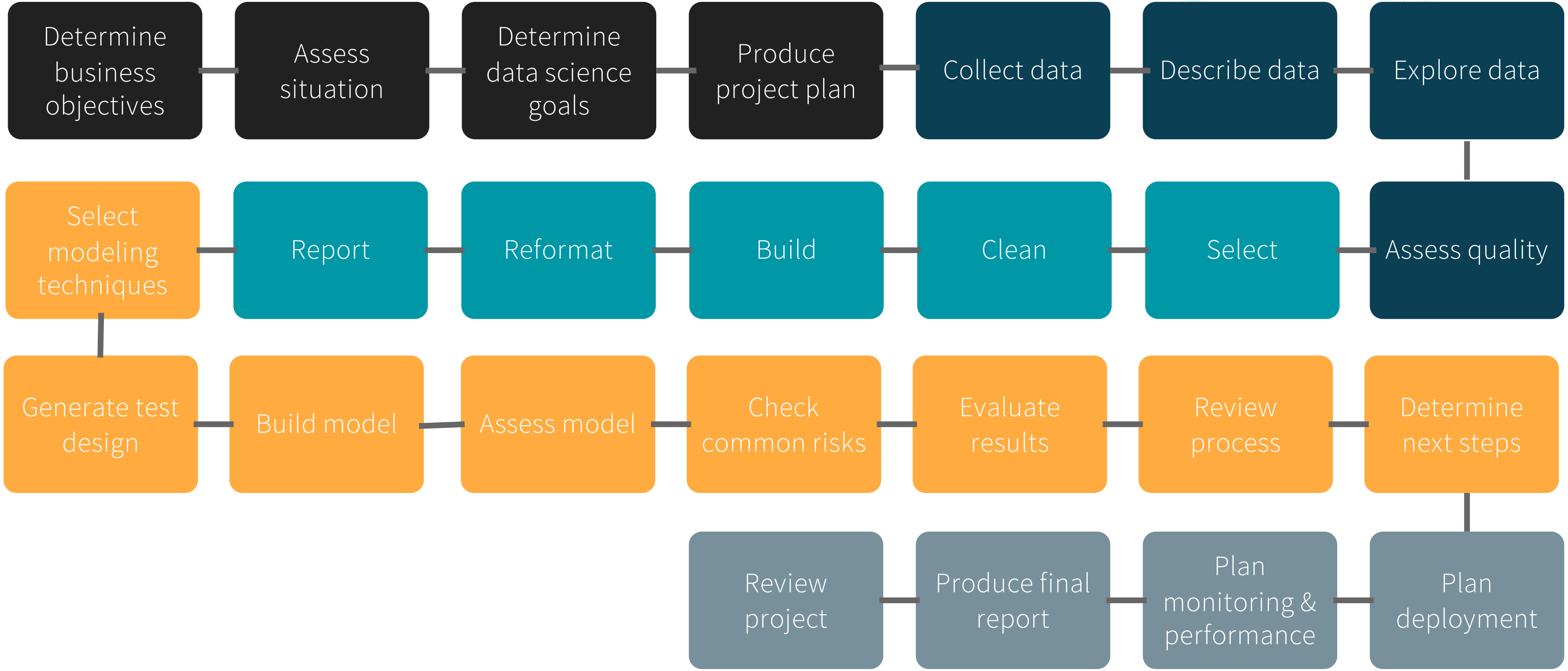
7 steps of a data projects

The Data Science Workflow



Advanced version of the workflow

The Data Science Workflow



Data preparation

70% of the work

Different Types of Data

Different types of Data

Definitions

Structured

Data stored with clearly defined data types whose pattern makes them easily searchable and linkable - most often tabular format

Examples:

- Database with columns for name, phone number etc
- XLS, SQL, CSV

Unstructured

Data is not structured via pre-defined data models or schema.

Examples:

- Text files
- Websites and social media
- Audio, pictures

Different types of Data

Structured data

A Categorical

Data can be one of several categories

Examples:
- Gender
- Nationality
- Hair color

Numerical

Data is a number

Examples:
- Age
- Weight
- Salary

I Text

Data is free-form text

Examples:
- Tweets
- Documents
- Business name

+ semi-structured data: json

Different type of structured data

Examples

	A Categorical	# Numerical	*I* Text
Eye Color (e.g. Blue)	✔		
Height (e.g. 170 cm)		✔	
Country of Birth (e.g. France)	✔		
Postal Code (e.g. 75001)	✔		
Date (e.g. Wednesday, 15 Jan 1976)	✔	✔	
Address (e.g. 10 Rue Saint Martin, Paris)	✔		✔
Curriculum Vitae			✔

Handling Missing value

What to do with missing values?

id	Amount_Requested	Loan_Purpose	Loan_Length	Status	Debt_To_Income_Ratio	Home_Ownership	FICO_Range	City
string Integer	string Decimal	string Text	string Text	string Text	string Text	string Text	string Text	string Text
81174	20.000	debt_consolidation	36 months	accepted	14.90%	MORTAGE	735-739	Paris
99592	19200	debt_consolidation		accepted	28.36%	ORTGAGE	715-719	Paris
80059	35000	debt_consolidation	60 months	accepted	0.2381	ORTGAGE	690-694	Lyon
15825	10000	debt_consolidation	36 months	accepted	0.1430	MORTGAGE	695-699	Londres
33182	12000	credit_card	36 months	accepted	18.78%	RENT	695-699	Marseille
62403	6000	other		accepted	20.05%	OWN	670-674	Nice
48808	10000	debt_consolidation	36 month	accepted	26.09%	RENT	720-724	London
22090	33500	credit_card	60 months	accepted	14.70%	MOTGAGE	705-709	Aix-en-Provence
76404	14675	credit_card	36 months	accepted	26.92%	RENT	685-689	Marseille
15867	7000	credit_card		accepted	7.10%	RENT	715-719	Marseille
94971	2000	moving	36 months	accepted	10.29%	RENT	670-674	aix-en-provence
36911	10625	debt_consolidation	36 months	accepted	12.54%	MORTGAGE	665-669	Paris
41200	28000	debt_consolidation	60 months	accepted	13.07%	MORTGAGE	670-674	Aix en Provence
83869	35000	debt_consolidation	36 months	pending	20.46pct	RENT	735-739	Paris

Drop values

Delete all data from any participant with missing values

Loan_Purpose	Loan_Length	Status	Debt_To_Income_Ratio
Text	Text	Text	Text
debt_consolidation	36 months	accepted	14.90%
~~debt_consolidation~~		~~accepted~~	~~28.36%~~
debt_consolidation	60 months	accepted	0.2381

Few Warnings:

- Be sure your sample is **large enough**, then you likely can drop data without substantial loss of statistical power.
- Be sure **data is not missing at Random**: There is a pattern in the missing data that affect your primary dependent variables.

  For example, lower-income participants are less likely to respond income column.

Imputation

Replacing missing values with substitute values.

Method #1: Common Value

For Number:
- Average
- Median
- Constant Value

For Category:
- Treat like the category « Empty »
- Most frequent value
- A constant value

Method #2: Educated Guess

Infer a missing value:
- If Age is lower than 20, Income is likely to be 0
- If living in a house in a rich city: income is likely to be higher than average
- Nb. of child is likely to not be 0 if age is high and situation not married

Method #3: Sub-Model

- Create a specific model of machine learning to predict the missing (Regression, Classification)

Example

How to handle the missing data in this doc?

Adress	SPCategory	City	Birth_date	Income
Text	Natural lang.	Text	Date (unparsed)	Integer
48238 Ella Manor	Intermediate occupations	Vinniestad	4/20/84	64014
1377 Bahringer Street	Lower supervisory and technical occupations	Gladysport	7/1/83	33274
739 Bashirian Burg	Lower managerial and professional occupations	West Maximo	11/26/94	
969 Sandy Mount St.	Intermediate occupations	North Porterbury	8/24/60	27232
454 Walter Stream St.	Small employers and own account workers	North Angelica	2/18/90	64851
2311 Connor Views	Intermediate occupations	North Nelshaven	12/21/38	48970
689 Schmitt Rapids	Small employers and own account workers	North Esteban	2/4/83	40529
9147 Bernier Common Ave.	Higher managerial and professional occupations	Carrollton	11/29/95	
985 Hodkiewicz Courts Suite 951	Lower managerial and professional occupations	East Howardberg	12/9/54	
05004 Arlo Oval	Small employers and own account workers	South Lempiland	3/24/44	

Not OK

- Drop Rows
- Average

Better :

- Educated guess from SPCategory
- Sub-prediction Model

Data

Group by operations

Definition

"CLOUD" table

ID	NeighborhoodID	power	tag	ip
1	1	5	funny	
2	1	2	likeable	
3	4	1	funny	
4	1	9	funny	
5	3	2	smart	
6	3	8	nice	
7	3	1	smart	
8	8	3	okay	
9	4	5	alright	

t1_temp current outcome

id	NeighborhoodID	power	tag	ip
1	1	15	funny	
2	1	2	likeable	
5	3	3	smart	
6	3	8	nice	
9	4	5	alright	
8	8	3	okay	

Aggregate data from an entity

How can you aggregate this dataset?

Client_ID	QUANTITY	SALES_VALUE	STORE_ID	RETAIL_DISC	Product_Category
Integer	Integer	Decimal	Integer	Decimal	Text
40	1	0.99	406	0.0	CANDY - CHECKLANE
40	1	2.49	406	0.0	DISPOSIBLE FOILWARE
40	1	0.99	406	0.0	CANDY - CHECKLANE
41	1	3.5	330	-0.49	DEODORANTS
41	1	0.99	295	0.0	CANDY - CHECKLANE
42	2	7.38	380	0.0	CIGARETTES
42	1	1.99	380	0.0	HARDWARE SUPPLIES
43	1	0.6	345	0.0	CANDY - CHECKLANE
43	1	0.42	345	0.0	BABY FOODS
44	1	1.99	297	0.0	CANDY - PACKAGED
45	1	4.99	300	0.0	SINUS AND ALLERGY
45	1	4.99	300	0.0	SINUS AND ALLERGY
45	1	4.67	300	-0.52	SINUS AND ALLERGY
45	1	1.99	300	0.0	DEODORANTS

Aggregate data from an entity

How can you aggregate this dataset?

Client_ID Integer	QUANTITY Integer	SALES_VALUE Decimal	STORE_ID Integer	RETAIL_DISC Decimal	Product_Category Text
40	1	0.99	406	0.0	CANDY - CHECKLANE
40	1	2.49	406	0.0	DISPOSIBLE FOILWARE
40	1	0.99	406	0.0	CANDY - CHECKLANE
41	1	3.5	330	-0.49	DEODORANTS
41	1	0.99	295	0.0	CANDY - CHECKLANE
42	2	7.38	380	0.0	CIGARETTES
42	1	1.99	380	0.0	HARDWARE SUPPLIES
43	1	0.6	345	0.0	CANDY - CHECKLANE
43	1	0.42	345	0.0	BABY FOODS
44	1	1.99	297	0.0	CANDY - PACKAGED
45	1	4.99	300	0.0	SINUS AND ALLERGY
45	1	4.99	300	0.0	SINUS AND ALLERGY
45	1	4.67	300	-0.52	SINUS AND ALLERGY
45	1	1.99	300	0.0	DEODORANTS

Group By Options:

For Number:

- Average
- Sum
- Minimum and Maximum
- Standard Deviation

For Category and Number:

- Count of Value
- Count of Distinct Value
- First and Last Value
- Most Frequent

Result

Client_ID	QUANTITY	SALES_VALUE	STORE_ID	RETAIL_DISC	Product_Category
Integer	Integer	Decimal	Integer	Decimal	Text
40	1	0.99	406	0.0	CANDY - CHECKLANE
40	1	2.49	406	0.0	DISPOSIBLE FOILWARE
40	1	0.99	406	0.0	CANDY - CHECKLANE
41	1	3.5	330	-0.49	DEODORANTS
41	1	0.99	295	0.0	CANDY - CHECKLANE
42	2	7.38	380	0.0	CIGARETTES
42	1	1.99	380	0.0	HARDWARE SUPPLIES
43	1	0.6	345	0.0	CANDY - CHECKLANE
43	1	0.42	345	0.0	BABY FOODS
44	1	1.99	297	0.0	CANDY - PACKAGED
45	1	4.99	300	0.0	SINUS AND ALLERGY

Client_ID	QUANTITY_s...	SALES_VALUE_max	SALES_VALUE_avg	STORE_ID_distinct	RETAIL_DISC_count	RETAIL_DISC_sum	Product_Category_distinct
bigint	bigint	double	double	bigint	bigint	double	bigint
Integer	Integer	Decimal	Decimal	Integer	Integer	Decimal	Integer
40	710	39.99	5.025763546797998	6	609	-227.90000000000003	48
41	62	7.99	3.3271186440677973	2	59	-27.549999999999997	13
42	178	10.58	2.329363636363637	1	110	-41.64999999999999	16
43	190	67.47	2.688385093167706	1	161	-22.069999999999997	18
44	14	9.88	3.8776923076923087	4	13	-2.16	8
45	70	9.99	4.443428571428575	2	70	-8.799999999999999	17

Reminder: Dummification & Rescaling

Dummification & Rescaling: the issue

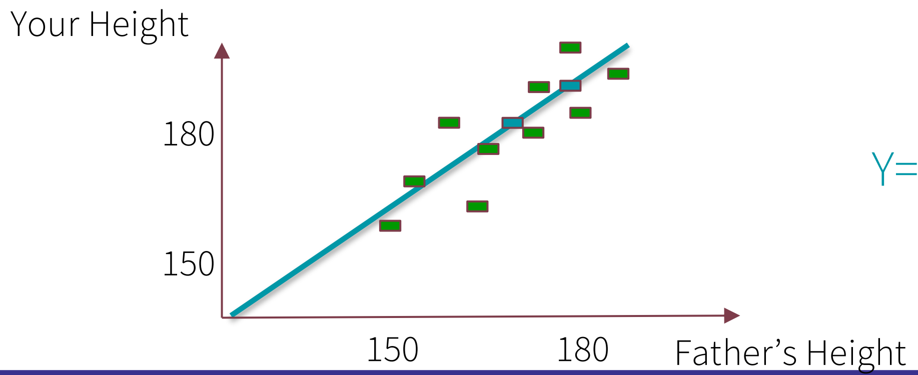
The specificity of your data can create bias in your model

- Some data is in a textual or numerical format but should be understood as a category by your model
- This also allows you to use non numerical data in a linear model

  > Create a dummy variable that corresponds to these categories: 0-1 for linear model

- Your numerical data can be distributed in a way that will be misunderstood by your model

  > Change the values to rearrange them on a scale



$$Y = a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5$$

Dummification for linear models

For categorical variables

Client_ID	Product_Category
bigint Integer	string Text
1060	BATTERIES
1916	SOAP - LIQUID & BAR
718	CANDY - PACKAGED
718	CANDY - PACKAGED
293	EASTER
293	CANDY - PACKAGED

Client_ID	BATTERIES	SOAP - LIQUID & BAR	CANDY - PACKAGED	EASTER	HAIR CARE PRODUCTS
bigint Integer	bigint Integer	bigint Integer	bigint Integer	bigint Integer	bigint Integer
1060	1	0	0	0	0
1916	0	1	0	0	0
718	0	0	1	0	0
718	0	0	1	0	0
293	0	0	0	1	0
293	0	0	1	0	0

Then your formula will look like this: $Y = a_{batteries}X_1 + a_{Soap}X_2 + a_{Candy}X_3 \ldots$

And $X_1, X_2, X_3 \ldots$ are either 0 or 1

Rescaling for numerical variables

LastName	First_Name	nb_childs	Income
string	string	string	string
Text	Text	Integer	Integer
Rosenbaum	Rosalind	1	28114
Stamm	Brett	3	47901
White	Verla	1	25700
Lindgren	David	1	18305
Brakus	Darryl		37341
Robel	Lilyan	0	31194

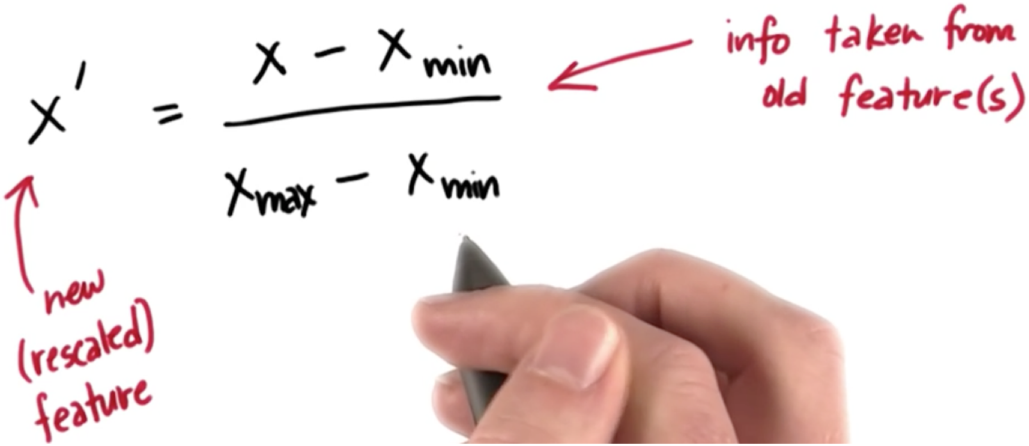
Without rescaling your formula will look like this:

$$Y_{rosalindrosenbaum} = a_{income} {}^{*}28114 + a_{Child} {}^{*}1 \ldots$$

$$Y_{BrettStamm} = a_{income} {}^{*}47901 + a_{Child} {}^{*}3 \ldots$$

Feature scaling is a method used to standardize the range of independent variables

Example of Rescaling – Min-Max Rescaling



Questions?

Hands-On