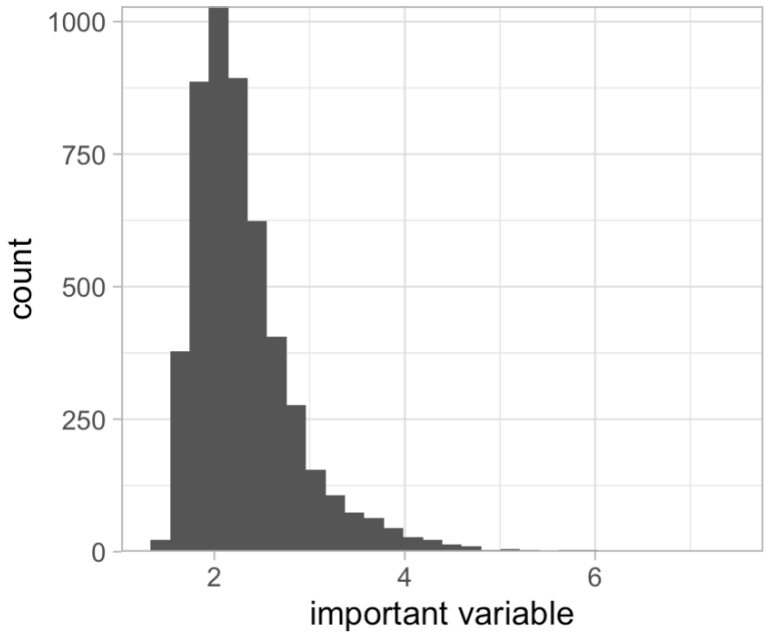
Data processing and feature engineering

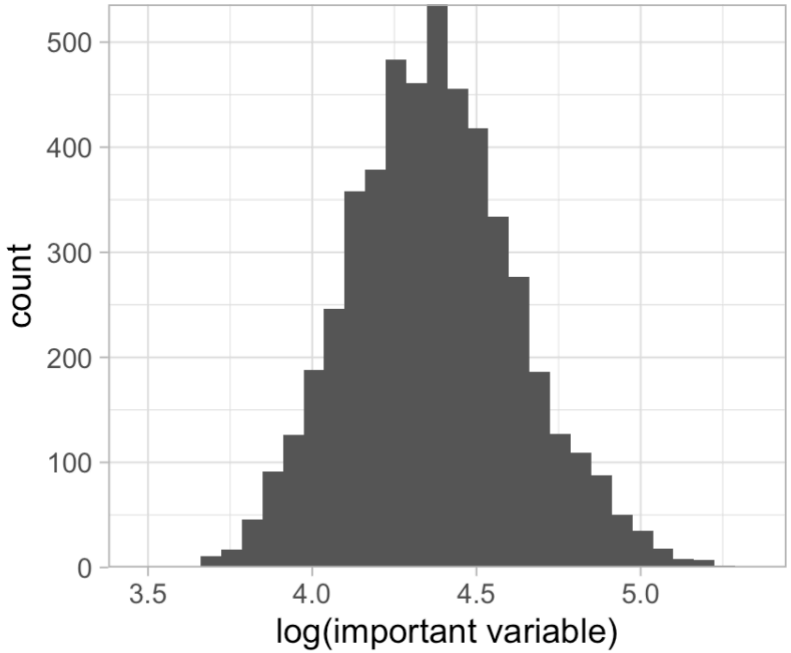
Dr Rebecca Barter

Transformations

A **logarithmic transformation** of a **skewed** important predictive variable can sometimes improve predictive performance

A logarithmic transformation of a skewed **response** variable can sometimes improve predictive performance, but you have to **un-transform** it when generating predictions





Numeric conversions for ordered categorical features

health_condition_self_reported
good
very good
poor
fair
excellent
good
good
poor
excellent
very good

Numeric conversions for ordered categorical features

health_condition_self_reported	health_condition_self_reported
good	
very good	
poor	0
fair	
excellent	
good	
good	
poor	0
excellent	
very good	

If the levels are **ordered**, they should be converted to a numeric variable

Numeric conversions for ordered categorical features

health_condition_self_reported	health_condition_self_reported
good	
very good	
poor	0
fair	**1**
excellent	
good	
good	
poor	0
excellent	
very good	

If the levels are **ordered**, they should be converted to a numeric variable

Numeric conversions for ordered categorical features

health_condition_self_reported	health_condition_self_reported
good	**2**
very good	
poor	0
fair	1
excellent	
good	**2**
good	**2**
poor	0
excellent	
very good	

If the levels are **ordered**, they should be converted to a numeric variable

Numeric conversions for ordered categorical features

health_condition_self_reported	health_condition_self_reported
good	2
very good	**3**
poor	0
fair	1
excellent	
good	2
good	2
poor	0
excellent	
very good	**3**

If the levels are **ordered**, they should be converted to a numeric variable

Numeric conversions for ordered categorical features

health_condition_self_reported	health_condition_self_reported
good	2
very good	3
poor	0
fair	1
excellent	**4**
good	2
good	2
poor	0
excellent	**4**
very good	3

If the levels are **ordered**, they should be converted to a numeric variable

One-hot encoding (dummy variables) for unordered categorical features

health_condition_self_reported
good
very good
poor
fair
excellent
good
good
poor
excellent
very good

If the levels are **ordered**, they should be converted to a numeric variable

If the levels are **unordered**, they should be converted to a collection of binary variables (one-hot encoding)

One-hot encoding (dummy variables) for unordered categorical features

health_condition_self_reported	health_condition_poor	health_condition_fair	health_condition_good	health_condition_very_good	health_condition_excellent
good					
very good					
poor					
fair					
excellent					
good					
good					
poor					
excellent					
very good					

If the levels are **ordered**, they should be converted to a numeric variable

If the levels are **unordered**, they should be converted to a collection of binary variables (one-hot encoding)

One-hot encoding (dummy variables) for unordered categorical features

health_condition_self_reported	health_condition_poor	health_condition_fair	health_condition_good	health_condition_very_good	health_condition_excellent
good	0				
very good	0				
poor	**1**				
fair	0				
excellent	0				
good	0				
good	0				
poor	**1**				
excellent	0				
very good	0				

If the levels are **ordered**, they should be converted to a numeric variable

If the levels are **unordered**, they should be converted to a collection of binary variables (one-hot encoding)

One-hot encoding (dummy variables) for unordered categorical features

health_condition_self_reported	health_condition_poor	health_condition_fair	health_condition_good	health_condition_very_good	health_condition_excellent
good	0	0			
very good	0	0			
poor	1	0			
fair	0	**1**			
excellent	0	0			
good	0	0			
good	0	0			
poor	1	0			
excellent	0	0			
very good	0	0			

If the levels are **ordered**, they should be converted to a numeric variable

If the levels are **unordered**, they should be converted to a collection of binary variables (one-hot encoding)

One-hot encoding (dummy variables) for unordered categorical features

health_condition_self_reported	health_condition_poor	health_condition_fair	health_condition_good	health_condition_very_good	health_condition_excellent
good	0	0	**1**		
very good	0	0	0		
poor	1	0	0		
fair	0	1	0		
excellent	0	0	0		
good	0	0	**1**		
good	0	0	**1**		
poor	1	0	0		
excellent	0	0	0		
very good	0	0	0		

If the levels are **ordered**, they should be converted to a numeric variable

If the levels are **unordered**, they should be converted to a collection of binary variables (one-hot encoding)

One-hot encoding (dummy variables) for unordered categorical features

health_condition_self_reported	health_condition_poor	health_condition_fair	health_condition_good	health_condition_very_good	health_condition_excellent
good	0	0	1	0	
very good	0	0	0	**1**	
poor	1	0	0	0	
fair	0	1	0	0	
excellent	0	0	0	0	
good	0	0	1	0	
good	0	0	1	0	
poor	1	0	0	0	
excellent	0	0	0	0	
very good	0	0	0	**1**	

If the levels are **ordered**, they should be converted to a numeric variable

If the levels are **unordered**, they should be converted to a collection of binary variables (one-hot encoding)

One-hot encoding (dummy variables) for unordered categorical features

health_condition_self_reported	health_condition_poor	health_condition_fair	health_condition_good	health_condition_very_good	health_condition_excellent
good	0	0	1	0	0
very good	0	0	0	1	0
poor	1	0	0	0	0
fair	0	1	0	0	0
excellent	0	0	0	0	**1**
good	0	0	1	0	0
good	0	0	1	0	0
poor	1	0	0	0	0
excellent	0	0	0	0	**1**
very good	0	0	0	1	0

If the levels are **ordered**, they should be converted to a numeric variable

If the levels are **unordered**, they should be converted to a collection of binary variables (one-hot encoding)

One-hot encoding (dummy variables) for unordered categorical features

health_condition_poor	health_condition_fair	health_condition_good	health_condition_very_good	health_condition_excellent
0	0	1	0	0
0	0	0	1	0
1	0	0	0	0
0	1	0	0	0
0	0	0	0	1
0	0	1	0	0
0	0	1	0	0
1	0	0	0	0
0	0	0	0	1
0	0	0	1	0

If the levels are **ordered**, they should be converted to a numeric variable

If the levels are **unordered**, they should be converted to a collection of binary variables (one-hot encoding)

One-hot encoding (dummy variables) for unordered categorical features

health_condition_poor	health_condition_fair	health_condition_good	health_condition_very_good	health_condition_excellent
0	0	1	0	0
0	0	0	1	0
1	0	0	0	0
0	1	0	0	0
0	0	0	0	1
0	0	1	0	0
0	0	1	0	0
1	0	0	0	0
0	0	0	0	1
0	0	0	1	0

If the levels are **ordered**, they should be converted to a numeric variable

If the levels are **unordered**, they should be converted to a collection of binary variables (one-hot encoding)

One level should be left out of the model to avoid **redundancy**

Missing values

Many models will **not accept** data with missing values or will **silently remove rows** with missing values

- Removing **rows** with missing values can bias your data & predictions and is *not* recommended
- If a **column** has more than >50% of its values missing, *and* it's not important for the prediction problem, you may want to remove it
- For columns with <50% of their values missing, you may want to **impute** missing values. Simplest imputation methods:
    - **Numeric** columns: impute with **mean**
    - **Categorical** columns: impute with **mode** (*before* creating dummy variables)

Avoiding “Data Leakage”

Any modifications that you want to do to your data should be based on information from **the training data** only

For example

- **Mean imputation** should use the **mean** of the **training data** for imputing both training and test sets
- When creating **dummy variables,** you should only include **levels** for values from the **training data**

Features included in ML models should be:

Response variable: healthcare expenses (for new people)

- **Relevant** to the outcome variable — use **domain knowledge**
  Demographics, comorbidities, health insurance are relevant to healthcare utilization
- Not be **ID** variables
- **Pre-outcome** — available before the outcome is observed
  Available: Patient history Unavailable: Dr visits in the *next (or current)* year
- **Non-redundant** — not highly correlated with other features
  BMI = weight / height^2
- Ethically **appropriate**
  Depends on how the predictions will be used

Which features are appropriate?

Response variable: healthcare expenses (for new people)

Variable Name	Appropriate?
SEQN	
age	
pregnant	
income_to_poverty_ratio	
n_overnight_hospital_stays_year	
seen_mental_health_professional_year	
health_condition_self_reported	
n_drinks_per_day	
smoker	
diabetes	
history_heart_disease	
history_stroke	
history_copd	
history_cancer	
health_insurance	
weight	
height	
bmi	
gender_m	

Which features are appropriate?

Response variable: healthcare expenses (for new people)

Variable Name	Appropriate?
SEQN	
age	yes
pregnant	yes
income_to_poverty_ratio	yes
n_overnight_hospital_stays_year	
seen_mental_health_professional_year	
health_condition_self_reported	yes
n_drinks_per_day	yes
smoker	yes
diabetes	yes
history_heart_disease	yes
history_stroke	yes
history_copd	yes
history_cancer	yes
health_insurance	yes
weight	
height	
bmi	
gender_m	yes

Which features are appropriate?

Response variable: healthcare expenses (for new people)

Variable Name	Appropriate?
SEQN	**ID Variable**
age	yes
pregnant	yes
income_to_poverty_ratio	yes
n_overnight_hospital_stays_year	**Proxy for outcome & not available pre-outcome**
seen_mental_health_professional_year	**Proxy for outcome & not available pre-outcome**
health_condition_self_reported	yes
n_drinks_per_day	yes
smoker	yes
diabetes	yes
history_heart_disease	yes
history_stroke	yes
history_copd	yes
history_cancer	yes
health_insurance	yes
weight	**Redundant**
height	**Redundant**
bmi	**Redundant**
gender_m	yes

These correspond to the same year as the outcome in the data

If they were measured for the previous year, they would be helpful!

$$\text{BMI} = \frac{\text{weight}}{\text{height}^2}$$

Which features are appropriate?

Response variable: healthcare expenses (for new people)

Variable Name	Appropriate?
~~SEQN~~	~~ID Variable~~
age	yes
pregnant	yes
income_to_poverty_ratio	yes
~~n_overnight_hospital_stays_year~~	~~Proxy for outcome & not available pre-outcome~~
~~seen_mental_health_professional_year~~	~~Proxy for outcome & not available pre-outcome~~
health_condition_self_reported	yes
n_drinks_per_day	yes
smoker	yes
diabetes	yes
history_heart_disease	yes
history_stroke	yes
history_copd	yes
history_cancer	yes
health_insurance	yes
weight	**Redundant**
height	**Redundant**
~~**bmi**~~	~~**Redundant**~~
gender_m	yes

These correspond to the same year as the outcome in the data

If they were measured for the previous year, they would be helpful!

$$\text{BMI} = \frac{\text{weight}}{\text{height}^2}$$

Choosing the “best” features summary

Domain knowledge is really important for building good models!

For fairly simple problems, using all *acceptable* variables will usually give the best predictive performance

You can compare the performance of several models using your validation set data