Predictive Analytics Lecture 1

Adam Kapelner

Stat 422/722
at The Wharton School of the University of Pennsylvania

January 17 & 18, 2017

Define: Prediction and Forecast

“statement about an uncertain event”,

Define: Prediction and Forecast

“statement about an uncertain event”, “informed guess or opinion”

predict (v.) 1620s (implied in predicted), "*foretell, prophesy,*" a back formation from prediction or else from Latin praedicatus, past participle of praedicere "foretell, advise, give notice,"

forecast (n.) early 15c., "*forethought, prudence,*" probably from forecast (v.). Meaning "conjectured estimate of a future course" is from 1670s.

I will be using predict and forecast interchangeably.

Examples

We make predictions all the time, saying things like:

- "Apple stock will go up tomorrow",
- "This condo will sell for $500K"

and sometimes unknowingly

- "Going skiing this weekend will make me happy",

How do we make predictions?

Examples

We make predictions all the time, saying things like:

- "Apple stock will go up tomorrow",
- "This condo will sell for $500K"

and sometimes unknowingly

- "Going skiing this weekend will make me happy",

How do we make predictions? We use a *model*.

Define: model

"a functional decription of a system"

Define: model

"a functional decription of a system"

An example model is:

> *Early to bed and early to rise makes a man healthy, wealthy and wise.*

All models have **input(s)** and **output(s)**. In the model above, what are the...

Inputs?

Define: model

"a functional decription of a system"

An example model is:

> *Early to bed and early to rise makes a man healthy, wealthy and wise.*

All models have **input(s)** and **output(s)**. In the model above, what are the...

Inputs? bedtime schedule, waking schedule, ...
Outputs?

Define: model

"a functional decription of a system"

An example model is:

> *Early to bed and early to rise makes a man healthy, wealthy and wise.*

All models have **input(s)** and **output(s)**. In the model above, what are the...

Inputs? bedtime schedule, waking schedule, ...
Outputs? health, wealth and wisdom

Synonyms for Inputs and Outputs

Here, the inputs and outputs are

- *features*
- *attributes*
- *characteristics*
- *variables* / *variates*

of a person. A person features health, a person has the characteristic of going to bed early.

What are "observations"?

Here, we have features of a person. Generally, inputs and outputs are features of the

- *observation* or
- *unit* or
- *record* or
- *subject*.

Thus the model relates some *feature(s) of the observation* to other *feature(s) of the observation*. Here, we are relating specific people's bedtime schedule and waking schedule to their health, wealth and wisdom.

Ambiguity of Models Defined by Words

Models phrased in language such as:

> *Early to bed and early to rise makes a man healthy, wealthy and wise.*

are usually ambiguous, imprecise, vague and ill-defined. Why?

Ambiguity of Models Defined by Words

Models phrased in language such as:

> *Early to bed and early to rise makes a man healthy, wealthy and wise.*

are usually ambiguous, imprecise, vague and ill-defined. Why?

- What does "early to bed" mean?
- What does "early to rise" mean?
- What does "healthy" mean?
- What does "wealthy" mean?
- What does "wise" mean?

Without resolving these ambiguities,

Ambiguity of Models Defined by Words

Models phrased in language such as:

> *Early to bed and early to rise makes a man healthy, wealthy and wise.*

are usually ambiguous, imprecise, vague and ill-defined. Why?

- What does "early to bed" mean?
- What does "early to rise" mean?
- What does "healthy" mean?
- What does "wealthy" mean?
- What does "wise" mean?

Without resolving these ambiguities, the model is unusable and of course, untestable.

Ambiguity of Models Defined by Words

Models phrased in language such as:

> *Early to bed and early to rise makes a man healthy, wealthy and wise.*

are usually ambiguous, imprecise, vague and ill-defined. Why?

- What does "early to bed" mean?
- What does "early to rise" mean?
- What does "healthy" mean?
- What does "wealthy" mean?
- What does "wise" mean?

Without resolving these ambiguities, the model is unusable and of course, untestable.

In order to make this precise and defined, there is a necessity to use numbers.

Ambiguity of Models Defined by Words

Models phrased in language such as:

> *Early to bed and early to rise makes a man healthy, wealthy and wise.*

are usually ambiguous, imprecise, vague and ill-defined. Why?

- What does "early to bed" mean?
- What does "early to rise" mean?
- What does "healthy" mean?
- What does "wealthy" mean?
- What does "wise" mean?

Without resolving these ambiguities, the model is unusable and of course, untestable.

In order to make this precise and defined, there is a necessity to use numbers. Thus, features of the observation must be *measured* (to be defined later).

Ambiguity of Models Defined by Words

Models phrased in language such as:

> *Early to bed and early to rise makes a man healthy, wealthy and wise.*

are usually ambiguous, imprecise, vague and ill-defined. Why?

- What does "early to bed" mean?
- What does "early to rise" mean?
- What does "healthy" mean?
- What does "wealthy" mean?
- What does "wise" mean?

Without resolving these ambiguities, the model is unusable and of course, untestable.

In order to make this precise and defined, there is a necessity to use numbers. Thus, features of the observation must be *measured* (to be defined later).

The Model as a Functional Relationship

Thus the model relates some *measured feature(s) of the observation* to other *measured feature(s) of the observation*. The relationship is a function taking in inputs (within the parentheses) and "returning" the outputs (the equal sign). For any observation,

$$\begin{array}{c}\text{the measured}\\ \text{outputs of an}\\ \text{observation}\end{array} = \text{model}\left(\begin{array}{c}\text{the measured}\\ \text{inputs of an}\\ \text{observation}\end{array}\right)$$

It is traditional to put the outputs on the left hand side. This is assumed that the outputs were measured. This type of observation is called

- old or
- historical or
- known

and predictions here are not needed (obviously). In our aphorism model, for the observation being a known person named Joe:

$$\left[\begin{array}{c}\text{a measured quantity of Joe's health}\\ \text{a measured quantity of Joe's wealth}\\ \text{a measured quantity of Joe's wisdom}\end{array}\right] = \text{model}\left(\left[\begin{array}{c}\text{a measured quantity of Joe's bedtime}\\ \text{a measured quantity of Joe's waketime}\\ \vdots\end{array}\right]\right)$$

Updated Definition of Prediction

Now we can hone our definition of prediction. For a

- new or
- heretofore unseen or
- future

observation, where the inputs have been measured / assessed but the output has not been measured / assessed,

$$\underbrace{\begin{array}{c}\text{the guessed}\\\text{output}\\\text{measurements}\end{array}}_{\text{prediction}} = \text{model}\left(\begin{array}{c}\text{the measured}\\\text{inputs of an}\\\text{observation}\end{array}\right)$$

$$\begin{bmatrix}\text{a guessed quantity of Bob's health}\\\text{a guessed quantity of Bob's wealth}\\\text{a guessed quantity Bob's wisdom}\end{bmatrix} = \text{model}\left(\begin{bmatrix}\text{a measured quantity of Bob's bedtime}\\\text{a measured quantity of Bob's waketime}\\\vdots\end{bmatrix}\right)$$

Measurements as Variables

Instead of "a measured quantity ..." we can use algebraic *variables* to denote the numerical quantities. It is traditional to use x's to represent inputs and y's to represent outputs. Here would be the relationship for Joe:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \text{model}\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \end{bmatrix}\right)$$

and for Bob:

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \end{bmatrix} = \text{model}\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \end{bmatrix}\right)$$

We will use the "hat" symbol (ˆ) to indicate a prediction of the response $\hat{y}$ to distinguish it from a known value of the response y.

Mathematical Model

Now that we have numbers for inputs / outputs and an equal sign relating them, we have created a *mathematical model*. The *model* now will be represented as a function, f. For an old observation,

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = f\left(x_1, x_2, \ldots\right)$$

and a new observation,

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \end{bmatrix} = f\left(x_1, x_2, \ldots\right)$$

Focus: models with univariate responses.

Although general models have any number of outputs, this semester we will only consider models with one output. Thus, we will be looking at models such as

Early to bed and early to rise makes a man healthy.

We picked the most interesting output. So, for an old observation,

$$y = f\left(x_1, x_2, \ldots\right)$$

and a new observation,

$$\hat{y} = f\left(x_1, x_2, \ldots\right)$$

More Vocabulary

Even though inputs and outputs are features of an observation, they each go by special names that emphasize their roles.

The output is called the

- *response* (the model "responds" to inputs)
- *outcome* (the result of inputs)
- ~~*endpoint*~~ (only used in clinical trial context)

and they are the target of prediction — what we want to ultimately predict.

Inputs then can go by the following terms of art:

- *covariates* (because the vary with the response, co-vary)
- *predictors* (since they will be the inputs used to make predictions)

and they are what we use to predict. I will try to use "response" and "predictors" in this course.

Define the measurements

Early to bed and early to rise makes a man healthy.

What is the response metric?

Define the measurements

Early to bed and early to rise makes a man healthy.

What is the response metric? What does "healthy" mean?

- Healthy for his whole life? Unlikely the model means this...
- Healthy for ages 25-65? Since we can expect health in infanthood and adolescence but not in elderly years

One also gets a feeling from the wording, there is either "healthy" or "not healthy". Thus the response metric will be *categorical*.

Categorical measurements consist of discrete, mutually exclusive *levels*. Here, {healthy, not healthy}. Generally, {a, b, c, . . .}. Metrics with a large number of levels are difficult to model — keep it low.

I there are two levels, it is called *binary* and the model would be called a "binary response model".

Define the response clearly