Libraries

```r
library(tidyverse)
library(tidymodels)
library(reticulate)
use_condaenv("r-keras2")
library(keras)
library(embed)

set.seed(1818)
```

Data

```r
data(sales, package = "DMwR2")
glimpse(sales)
```

```
## Rows: 401,146
## Columns: 5
## $ ID    <fct> v1, v2, v3, v4, v3, v5, v6, v7, v8, v9, v10, v12, v13, v14, v13…
## $ Prod  <fct> p1, p1, p1, p1, p1, p2, p2, p2, p2, p2, p2, p3, p3, p3, p3, p4,…
## $ Quant <int> 182, 3072, 20393, 112, 6164, 104, 350, 200, 233, 118, 233, 108,…
## $ Val   <dbl> 1665, 8780, 76990, 1100, 20260, 1155, 5680, 4010, 2855, 1175, 1…
## $ Insp  <fct> unkn, unkn, unkn, unkn, unkn, unkn, unkn, unkn, unkn, unkn, unk…
```

```r
n_distinct(sales$Prod)
```

```
## [1] 4548
```

```r
sales_split <- initial_split(sales, strata = Insp)
sales_train <- training(sales_split)
```

Recipe

```
sales_recipe <- recipe(sales_train) %>%
  update_role(Quant, Val, Prod, new_role = "predictor") %>%
  update_role(Insp, new_role = "outcome") %>%
  step_rm(ID) %>%
  step_naomit(Quant, Val) %>%
  step_filter(Insp != "unkn") %>%
  step_center(Quant, Val) %>%
  step_scale(Quant, Val) %>%
  step_embed(Prod, num_terms = 4, hidden_units = 16, outcome = vars(Insp),
             options = embed_control(loss = "binary_crossentropy", epochs = 10))
```

```
trained_sales_recipe <- prep(sales_recipe)
```

Embeddings

```
trained_sales_recipe$steps[[6]]$mapping$Prod %>%
  relocate(..level) %>%
  head(5)
```

..level	Prod_embed_1	Prod_embed_2	Prod_embed_3	Prod_embed_4
..new	0.0491049	-0.0320730	0.0072949	-0.0053174
p1	-0.0034932	-0.0110625	0.0385611	0.0228304
p2	-0.0304066	0.0419249	0.0021925	0.0343973
p3	0.0148002	-0.0117648	-0.0215916	0.0337168
p4	0.0095341	0.0453324	0.0109225	0.0375945

Thank you!

- slides & code: https://github.com/rstudio/rstudio-conf/tree/master/2021/alanfeder
- contact: AlanFeder@gmail.com
- Twitter: @AlanFeder

Acknowledgements

- I first learned about categorical embeddings from **fastai: Practical Deep Learning for Coders**, by Jeremy Howard
- Data and analysis steps taken from **TensorFlow training** at **RStudio::conf(2019)**, by Sigrid Keydana, Kevin Kuo, and Rick Scavetta