eDream Odigeo Baggage Likelihood Model

Stéphane Couvreur

5/9/2018

"The simulacrum is never that which conceals the truth — it is the truth which conceals that there is none. The simulacrum is true."
Jean Baudrillard, *Simulacra and Simulation*, 1988

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

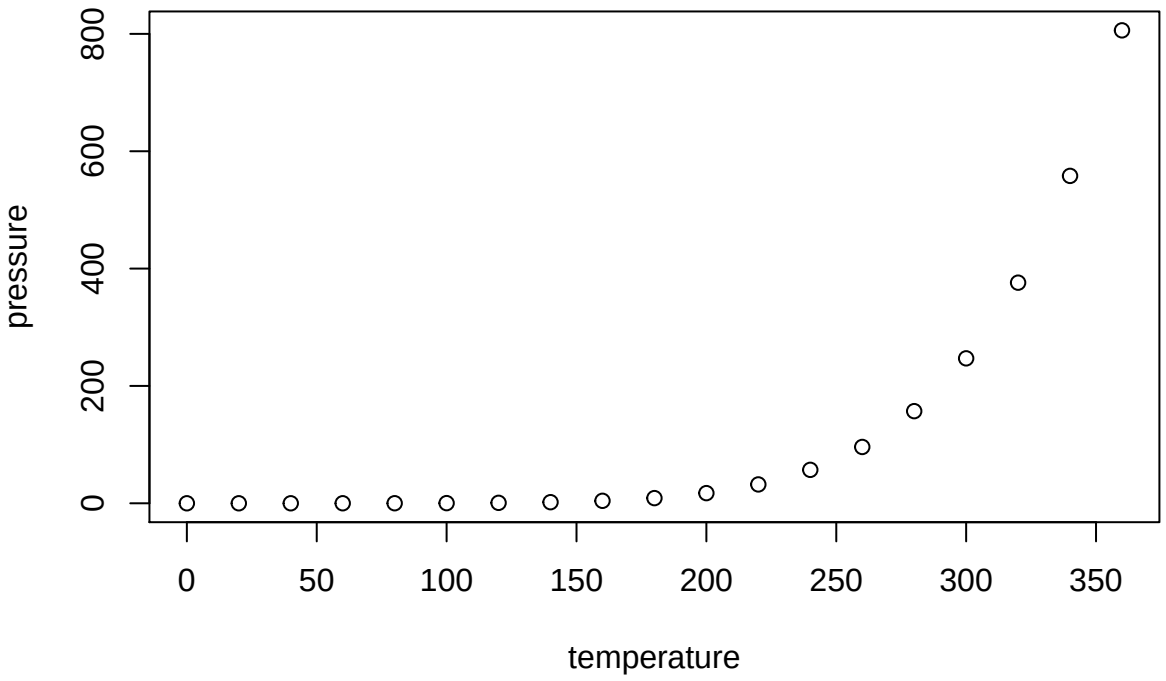
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Appendix - Logged thoughts

Day 1 - Wednesday 9th of May 2018

First thoughts looking at the data

How to make the model as parsimonious as possible ? Let's see which values do not have such relevance

I think the best would be to try a good old mulitple logisitic regression model with several dummy variables, but omit data which at first you find useless.

Data which seems irrelevant at first sight:

- TIMESTAMP
- DEPARTURE
- ARRIVAL

As there is very strong class imbalance within the TRAIN booking binary variable (99.5% in the training set did not book a train).

No train booking [%]	Train booking [%]
99.5	0.5

Similarly within the PRODUCT variable (98.1% booked a Trip compared to a Dynpack) - both these variables were not considered.

DYNPACK [%]	TRIP [%]
1.9	98.1

Investigating further

It would be interesting to see if there are significant variations in baggage booking between eDreams (ED), Opodo (OP) or Go Voyage (GO) - should use a string operation on this

To simplify I assume that there is no local variability between bookings in UK, Italy, Spain, France etc..

It seems that however there is nothing interesting there - the proportions are virtually exactly the same everywhere

	No baggage [%]	Baggage [%]
EDREAMS	80.6	19.4
GO VOYAGE	81.2	18.8
OPODO	79.9	20.1
Other	75.7	24.3

The website variable can therefore be ommitted

Don't know what to do with GDS variables, I remove them for now and come back later

Maybe those who pick SMS as an extra are more likely to pick other extras ? To investigate

Not much going on there actually

	No baggage [%]	Baggage [%]
No SMS confirmation	80.3	19.7
SMS confirmation	80.5	19.5

Could it be that with certain devices more customers book devices ?

	No baggage [%]	Baggage [%]
COMPUTER	79.5	20.5
OTHER	76.6	23.4
SMARTPHONE	83.2	16.8
TABLET	79.6	20.4

Not so much actually really, a bit of a face value judgement but let's omit the DEVICE variable for now and investigate later if we have time

Adults travelling alone I would assume would be less likely to book luggage, but with one or more children much more likely to get luggage, especially with infants

Indeed from a small table you can see that:

I would imagine that flight distance would account for a lot of the probability of luggage selection (high R2), as people who travel further I would assume need to carry more than if they are doing a short weekend trip within Europe

As one would imagine, flight DISTANCE seems to follow a skeweved normal distribution with alot of short flights between 0-3000km and then drastic reductions from then onwards.

Although our first assumption that number of adults was a predictor of baggage selection - indeed fitting it to our general linear model it would seem so as it's hightly significant in terms of p-value:

	Estimate	Std. Error	z value	Pr(>z)	
(Intercept)	-9.070e-01	6.024e-02	-15.058	< 2e-16	***
DISTANCE	-5.051e-05	9.392e-06	-5.379	7.51e-08	***
factor(HAUL_TYPE)DOMESTIC	-4.285e-01	3.223e-02	-13.292	< 2e-16	***
factor(HAUL_TYPE)INTERCONTINENTAL	-1.139e+00	4.913e-02	-23.185	< 2e-16	***
factor(TRIP_TYPE)ONE_WAY	-3.732e-01	5.784e-02	-6.452	1.10e-10	***
factor(TRIP_TYPE)ROUND_TRIP	-6.986e-01	5.651e-02	-12.362	< 2e-16	***
ADULTS	2.680e-01	1.408e-02	19.031	< 2e-16	***
CHILDREN	2.344e-01	3.133e-02	7.483	7.26e-14	***
INFANTS	2.451e-01	9.171e-02	2.673	0.00752	**

However, considering adults childrens and infants as levels, it seems that having two children or one infant highly increases the change of selecting luggage. It might me interesting for the the sake of parsimony to remove the adult category.

	Estimate	Std. Error	z value	Pr(>z)	
(Intercept)	-1.299e+01	1.754e+02	-0.074	0.9410	
DISTANCE	-5.214e-05	9.418e-06	-5.536	3.09e-08	***

	Estimate	Std. Error	z value	Pr(>z)	
factor(HAUL_TYPE)DOMESTIC	-4.376e-01	3.235e-02	-13.527	< 2e-16	***
factor(HAUL_TYPE)INTERCONTINENTAL	-1.115e+00	4.934e-02	-22.588	< 2e-16	***
factor(TRIP_TYPE)ONE_WAY	-3.324e-01	5.823e-02	-5.708	1.14e-08	***
factor(TRIP_TYPE)ROUND_TRIP	-6.891e-01	5.683e-02	-12.126	< 2e-16	***
factor(ADULTS)1	1.222e+01	1.754e+02	0.070	0.9444	
factor(ADULTS)2	1.285e+01	1.754e+02	0.073	0.9416	
factor(ADULTS)3	1.277e+01	1.754e+02	0.073	0.9420	
factor(ADULTS)4	1.293e+01	1.754e+02	0.074	0.9412	
factor(ADULTS)5	1.294e+01	1.754e+02	0.074	0.9412	
factor(ADULTS)6	1.335e+01	1.754e+02	0.076	0.9393	
factor(ADULTS)7	1.327e+01	1.754e+02	0.076	0.9397	
factor(ADULTS)8	1.263e+01	1.754e+02	0.072	0.9426	
factor(ADULTS)9	1.265e+01	1.754e+02	0.072	0.9425	
factor(CHILDREN)1	3.014e-01	5.675e-02	5.311	1.09e-07	***
factor(CHILDREN)2	4.948e-01	8.751e-02	5.654	1.57e-08	***
factor(CHILDREN)3	3.535e-01	2.107e-01	1.677	0.0935	.
factor(CHILDREN)4	-1.226e+01	1.320e+02	-0.093	0.9260	
factor(CHILDREN)5	1.455e+01	5.354e+02	0.027	0.9783	
factor(INFANTS)1	2.496e-01	9.470e-02	2.635	0.0084	**
factor(INFANTS)2	-1.265e+01	1.382e+02	-0.091	0.9271	

First iteration of the logistic regression model gives an AUC of 0.6828

Day 2 - Thursday 10th of May 2018

Building the model

Preliminary 80/20 train/validation split to have internal validation mechanism