

eDream Odigeo Baggage Likelihood Model

Stéphane Couvreur

16/9/2018

“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

Exploratory Data Analysis

Plotting and visualising the distributions of different variables

Overall proportion of people having booked extra baggage:

```
##  
## No Extra Baggage    Extra Baggage  
##           80.4           19.6
```

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 multiple logistic 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

Severe class imbalance with the two variables TRAIN and PRODUCT

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

```
##  
## False  True  
##  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, comparing baggage selection rates among different variables

Feature engineering

Mapping of the booking company and encoding

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.. Also, extracting different countries would just lead to a categorical factor variable with potentially many levels - which is not so good for a machine learning algorithm.

There is interesting subtle variability between different booking websites, not all the same

```
##
##           No Extra Baggage Extra 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 is therefore likely a strong predictor of our outcome variable

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

Synthetic variable family size

We investigate a potential relation between infants and baggage

After creating a synthetic variable combining ADULTS + CHILDREN + INFANTS called FAMILY_SIZE

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

Not much going on there actually

It seems that there is not much interesting with SMS confirmation

```
##
##           No Extra Baggage Extra Baggage
##  False           80.3           19.7
##  True            80.5           19.5
```

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

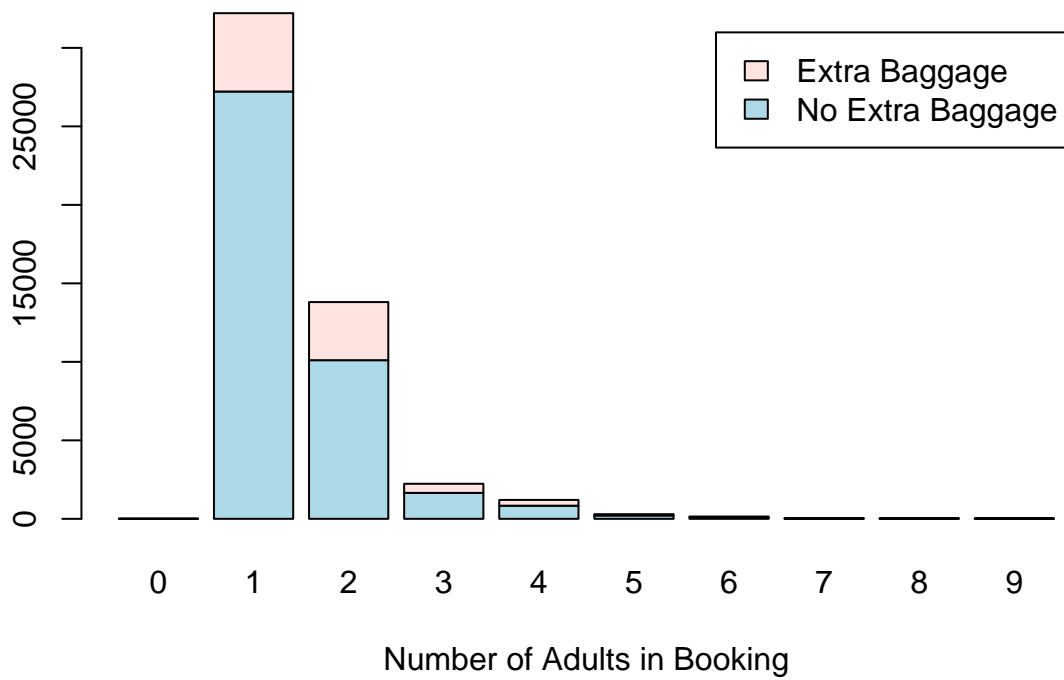
```
##
##           No Extra Baggage Extra Baggage
##  COMPUTER           79.5           20.5
##  OTHER              76.6           23.4
##  SMARTPHONE         83.2           16.8
##  TABLET            79.6           20.4
```

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:

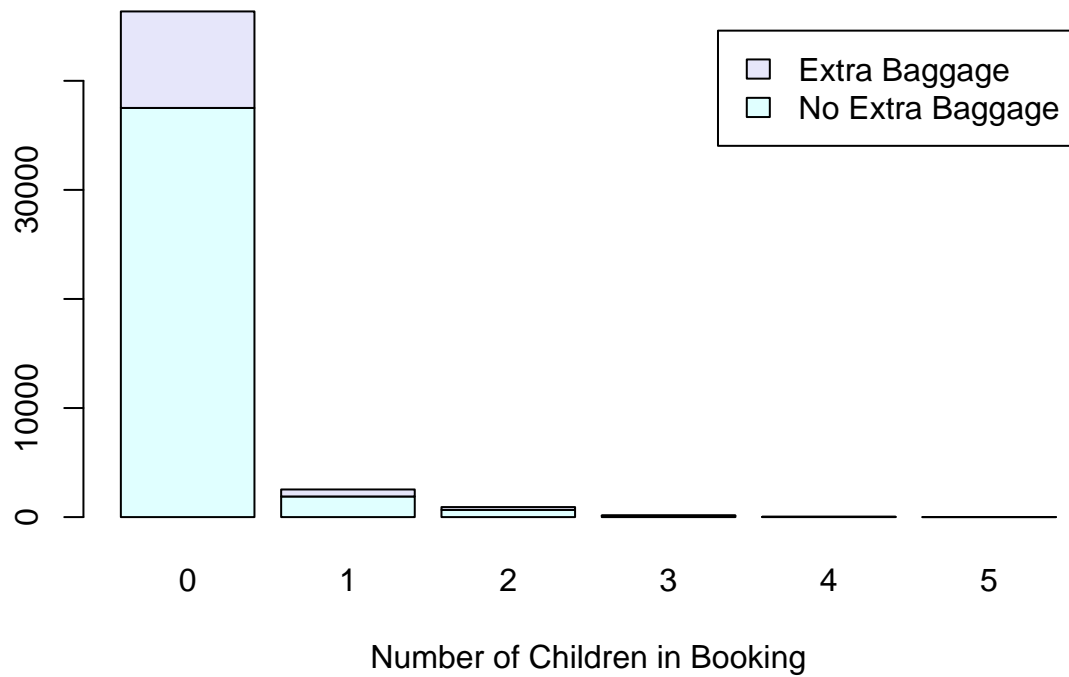
Adult Booking Distribution



##			
##		No Extra Baggage	Extra Baggage
##	0	100.0	0.0
##	1	84.5	15.5
##	2	73.2	26.8
##	3	74.1	25.9
##	4	69.5	30.5
##	5	71.7	28.3
##	6	63.3	36.7
##	7	69.4	30.6
##	8	80.0	20.0
##	9	73.9	26.1

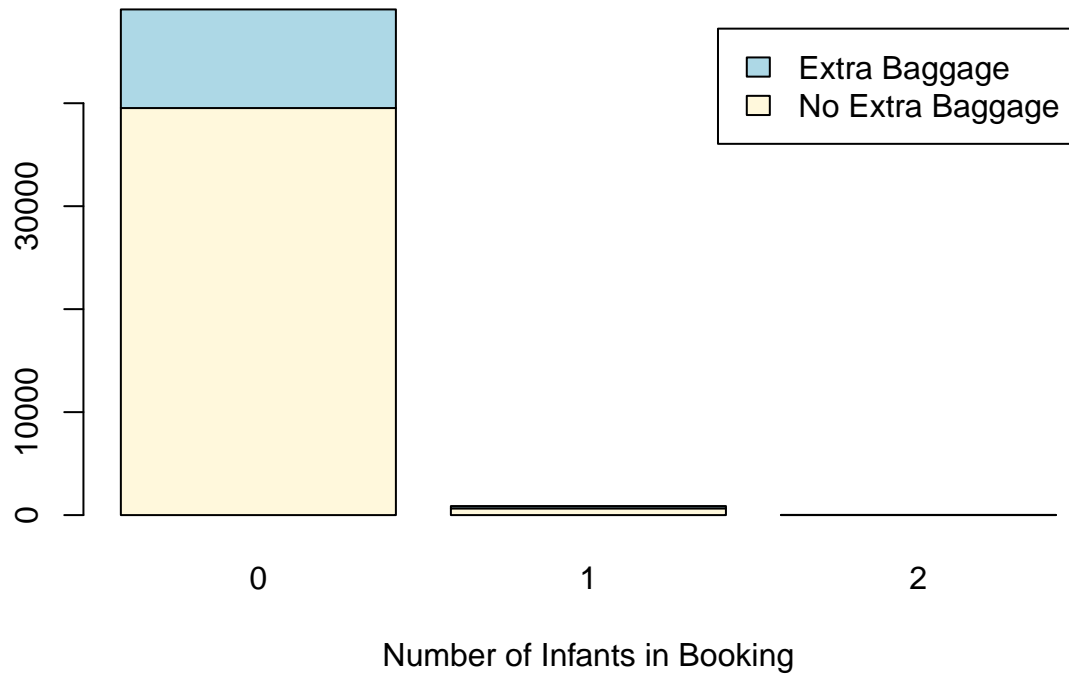
It seems that the more adults are travelling, the more likely they are to book luggage

Children Booking Distribution

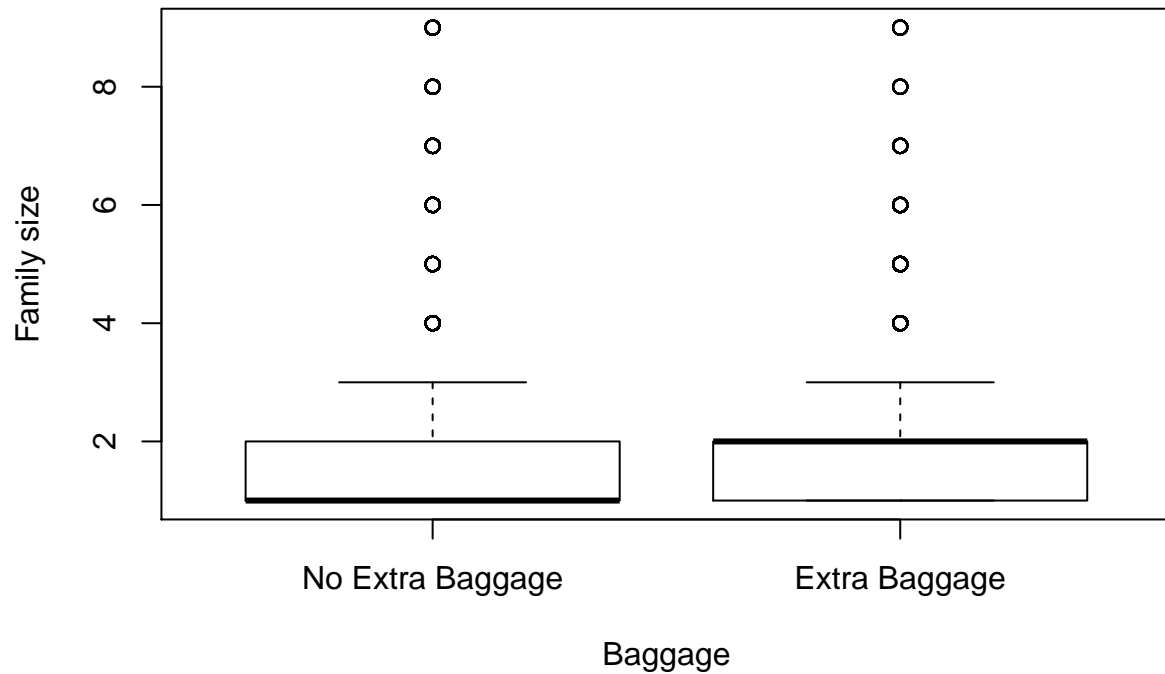


##			
##	No Extra Baggage	Extra Baggage	
##	0	80.9	19.1
##	1	74.6	25.4
##	2	71.8	28.2
##	3	72.6	27.4
##	4	88.5	11.5
##	5	0.0	100.0

Infants Booking Distribution



```
##
##      No Extra Baggage Extra Baggage
## 0          80.5         19.5
## 1          74.6         25.4
## 2         100.0          0.0
```



```
##
##      No Extra Baggage Extra Baggage
```

##	1	84.7	15.3
##	2	74.3	25.7
##	3	74.8	25.2
##	4	69.1	30.9
##	5	71.2	28.8
##	6	69.9	30.1
##	7	63.5	36.5
##	8	77.8	22.2
##	9	62.9	37.1

Increased family size also seems to bring with it increased probability of extra baggage selection.

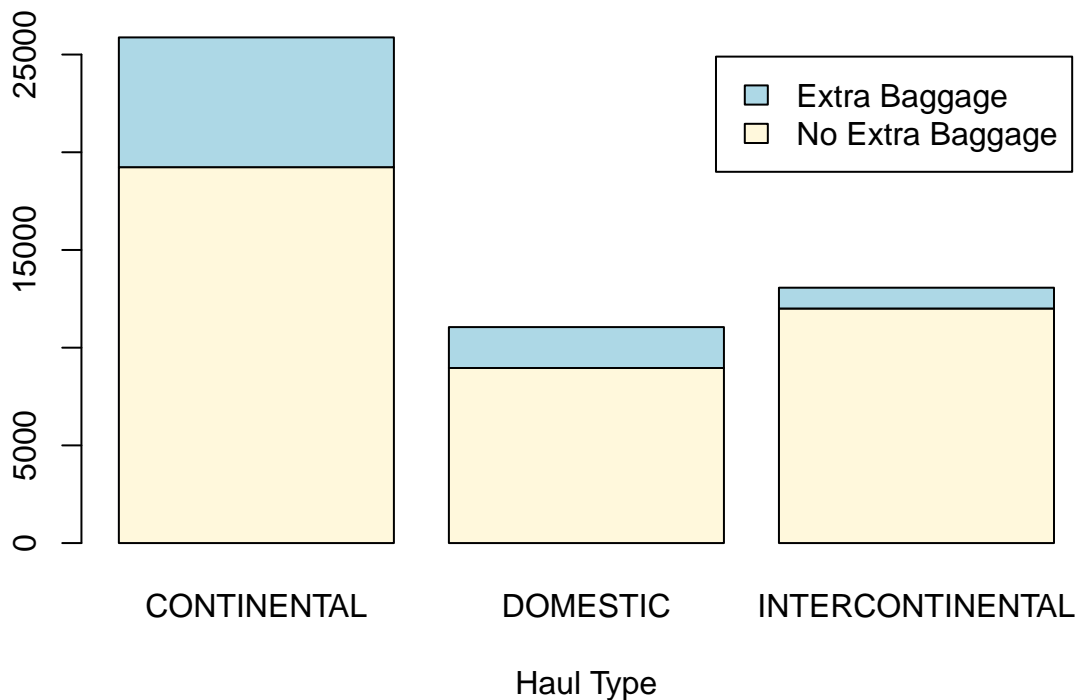
Synthetic variable adult alone

It would be interesting to see if the adults travelling alone tend to not book luggage as would be my initial assumption - we could create a binary variable IS_ALONE. Indeed from extracting this information it seems that we can improve our model as travellers not alone have much more probability of booking luggage.

##			
##		No Extra Baggage	Extra Baggage
##	Not alone	73.6	26.4
##	Alone	84.7	15.3

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

Haul Booking Distribution

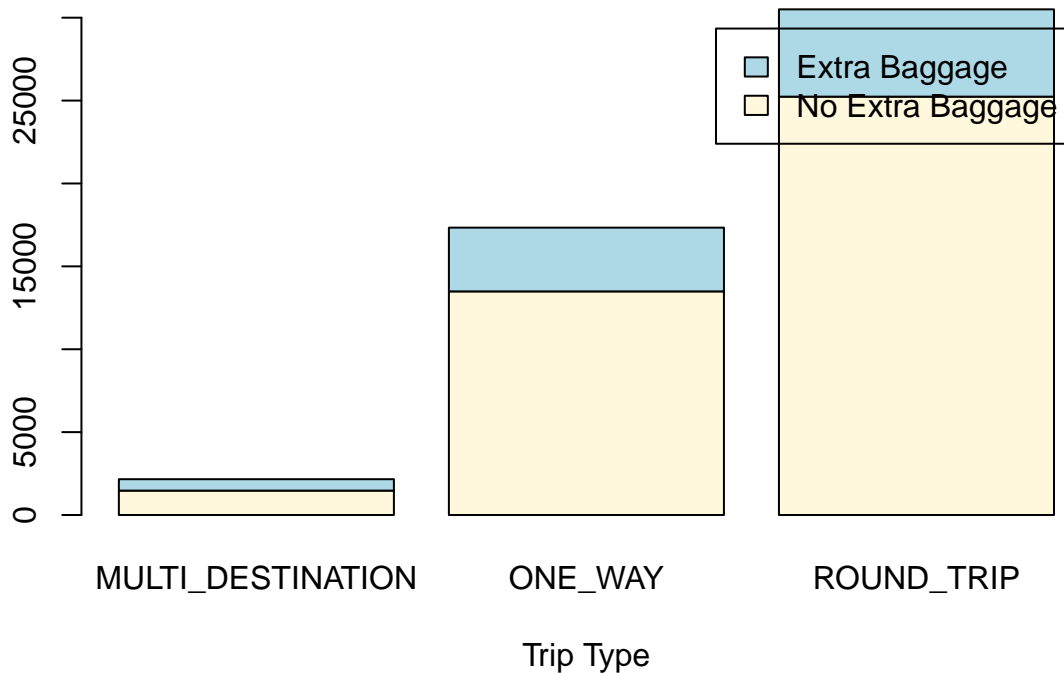


##			
##		No Extra Baggage	Extra Baggage
##	CONTINENTAL	74.3	25.7
##	DOMESTIC	81.1	18.9

```
## INTERCONTINENTAL          91.9          8.1
```

There are quite significant differences here between groups. One can imagine that in intercontinental flights, the luggage from more premium companies will be complimentary so no extra is needed. And for domestic flights it makes sense - travelling at home you might need less luggage.

Trip Booking Distribution

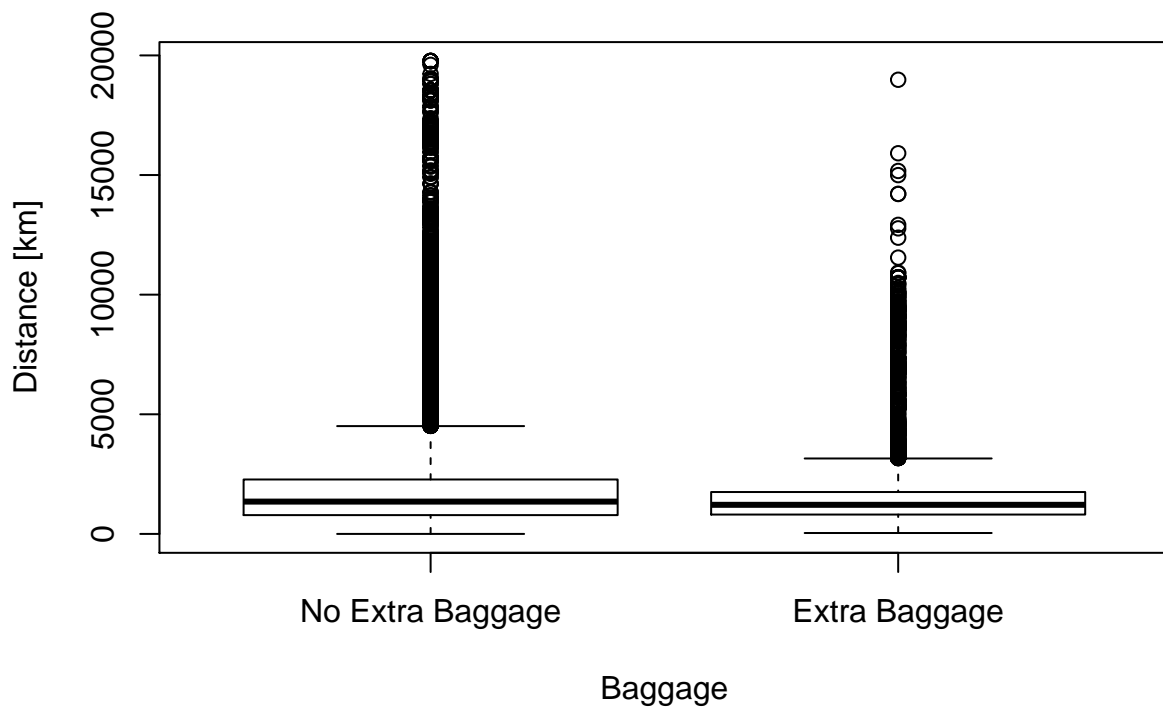
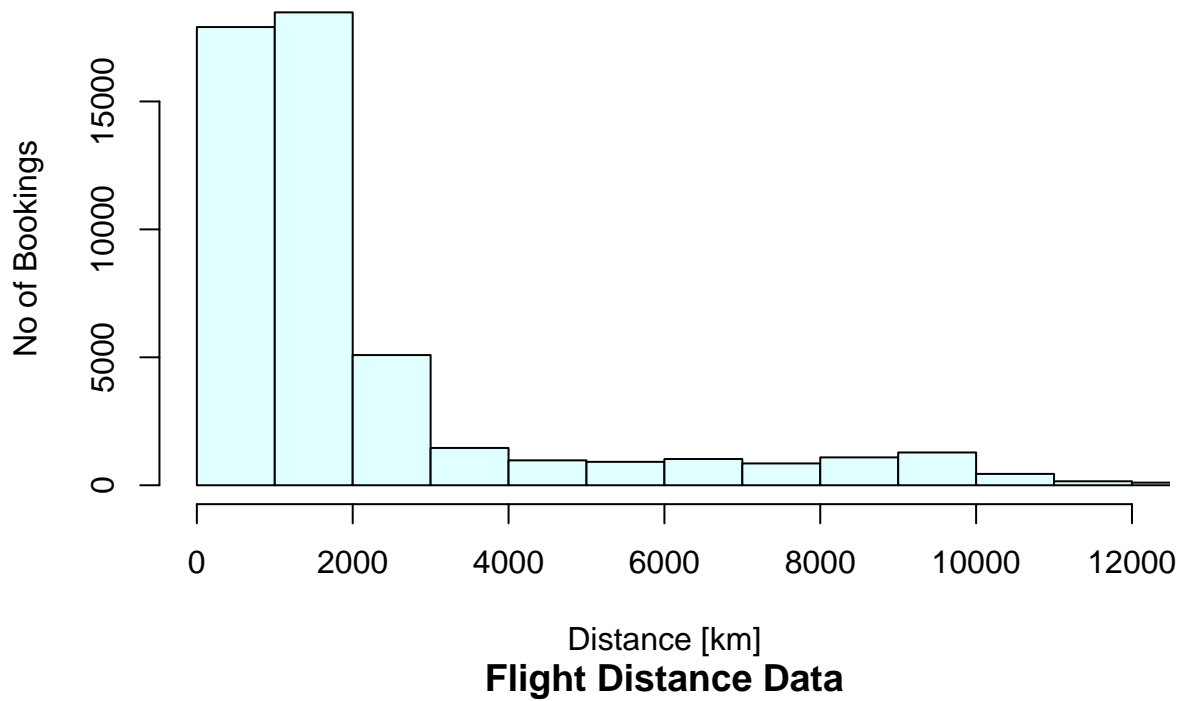


```
##
##           No Extra Baggage Extra Baggage
## MULTI_DESTINATION          68.1          31.9
## ONE_WAY                  77.8          22.2
## ROUND_TRIP                82.7          17.3
```

Interestingly, in round trips customers select extra baggage the least - perhaps they travel lighter as they know their belongings are at home. However much more take luggage on one ways (moving, expatriation or immigration perhaps ?) and even more on multi-destination trips.

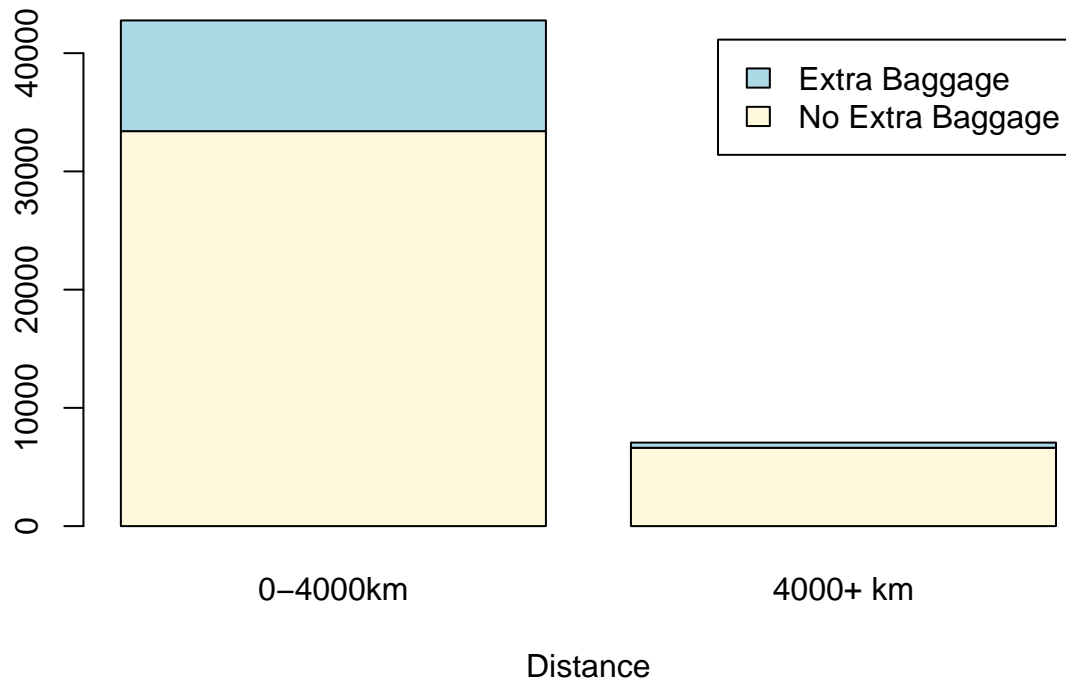
As one would imagine, flight DISTANCE seems to follow a skewed normal distribution with a lot of short flights between 0-3000km and then drastic reductions from then onwards.

Air Travel Distance Distribution



Using distance cut into categories for ease of use in our model. We group together values between 0-4000 and 4000+ km to make things even easier for our classifier.

Distance Category Booking Distribution



```
##
##           No Extra Baggage Extra Baggage
## 0-4000km           78.1           21.9
## 4000+ km           93.8           6.2
```

Building the model

Logisitic Regression

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 highly significant in terms of p-value:

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.

From the likelihood ratio test, it seems that there is strong evidence that

If we remove the categories using the family size feature, we get:

```
##
## Call:
## glm(formula = EXTRA_BAGGAGE ~ factor(HAUL_TYPE) + factor(TRIP_TYPE) +
##      factor(DISTANCE_CAT) + factor(DEVICE) + factor(COMPANY) +
##      FAMILY_SIZE + factor(IS_ALONE) + factor(SMS), family = binomial(link = "logit"),
##      data = train)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.2870  -0.7287  -0.5599  -0.3067   2.5758
```

```

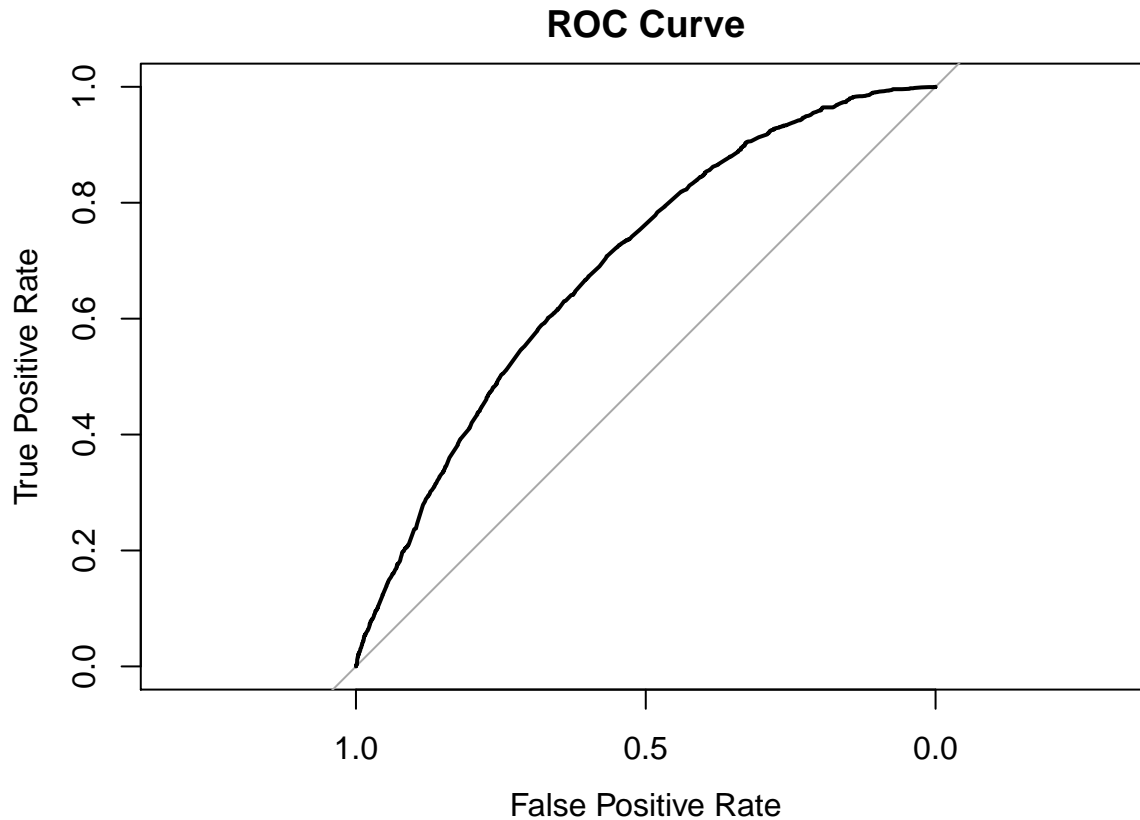
##
## Coefficients:
##
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.43223    0.07434  -5.814 6.09e-09
## factor(HAUL_TYPE)DOMESTIC      -0.37785    0.03229 -11.702 < 2e-16
## factor(HAUL_TYPE)INTERCONTINENTAL -1.09190    0.04890 -22.327 < 2e-16
## factor(TRIP_TYPE)ONE_WAY      -0.30735    0.05849  -5.254 1.48e-07
## factor(TRIP_TYPE)ROUND_TRIP    -0.69332    0.05712 -12.139 < 2e-16
## factor(DISTANCE_CAT)4000+ km   -0.55875    0.06862  -8.143 3.86e-16
## factor(DEVICE)OTHER      -12.14857   88.97834  -0.137  0.891
## factor(DEVICE)SMARTPHONE    -0.21953    0.03298  -6.656 2.82e-11
## factor(DEVICE)TABLET      -0.05320    0.05502  -0.967  0.334
## factor(COMPANY)GO VOYAGE       0.34282    0.04401   7.789 6.76e-15
## factor(COMPANY)OPODO         0.18077    0.03022   5.981 2.21e-09
## factor(COMPANY)OTHER        12.37199   88.97838   0.139  0.889
## FAMILY_SIZE         0.08585    0.01777   4.830 1.37e-06
## factor(IS_ALONE)Alone      -0.52364    0.03858 -13.573 < 2e-16
## factor(SMS)True          -0.01318    0.02605  -0.506  0.613
##
## (Intercept) ***
## factor(HAUL_TYPE)DOMESTIC ***
## factor(HAUL_TYPE)INTERCONTINENTAL ***
## factor(TRIP_TYPE)ONE_WAY ***
## factor(TRIP_TYPE)ROUND_TRIP ***
## factor(DISTANCE_CAT)4000+ km ***
## factor(DEVICE)OTHER
## factor(DEVICE)SMARTPHONE ***
## factor(DEVICE)TABLET
## factor(COMPANY)GO VOYAGE ***
## factor(COMPANY)OPODO ***
## factor(COMPANY)OTHER
## FAMILY_SIZE ***
## factor(IS_ALONE)Alone ***
## factor(SMS)True
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 39439  on 39751  degrees of freedom
## Residual deviance: 36893  on 39737  degrees of freedom
## (248 observations deleted due to missingness)
## AIC: 36923
##
## Number of Fisher Scoring iterations: 12
##
## Waiting for profiling to be done...
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
##
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
##
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
##
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
##
```

	odds	2.5 %	97.5 %
(Intercept)	6.490615e-01	5.609225e-01	7.507114e-01
factor(HAUL_TYPE)DOMESTIC	6.853354e-01	6.431853e-01	7.299722e-01
factor(HAUL_TYPE)INTERCONTINENTAL	3.355777e-01	3.046895e-01	3.690822e-01
factor(TRIP_TYPE)ONE_WAY	7.353896e-01	6.560535e-01	8.251605e-01
factor(TRIP_TYPE)ROUND_TRIP	4.999160e-01	4.471972e-01	5.594389e-01
factor(DISTANCE_CAT)4000+ km	5.719221e-01	4.995077e-01	6.537156e-01
factor(DEVICE)OTHER	5.295925e-06	1.181614e-17	3.734538e-04
factor(DEVICE)SMARTPHONE	8.028965e-01	7.524659e-01	8.563321e-01
factor(DEVICE)TABLET	9.481934e-01	8.505161e-01	1.055269e+00
factor(COMPANY)GO VOYAGE	1.408916e+00	1.292030e+00	1.535360e+00
factor(COMPANY)OPODO	1.198135e+00	1.129138e+00	1.271159e+00
factor(COMPANY)OTHER	2.360960e+05	3.365525e+03	1.088804e+17
FAMILY_SIZE	1.089638e+00	1.052177e+00	1.128123e+00
factor(IS_ALONE)Alone	5.923625e-01	5.492253e-01	6.388960e-01
factor(SMS)True	9.869090e-01	9.377823e-01	1.038604e+00

```
## Area under the curve: 0.6899
```



We can confirm from this that SMS is not significant in our model, we can therefore remove it.

We notice that looking at the odds ratio table that a unit increase in family size brings a 9.2% [95% CI 5.4 - 13.1] increase in probability of booking luggage as adjusting for several parameters. We can therefore say with 95% confidence that the true odds ratio of booking luggage after adjusting for flight distance, haul type, trip type, company and booking device in our population lies between the range [X - X] with mean

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

Area under the curve: 0.6899

Overall one of the challenges of building this model is that there is strong class imbalance - indeed it might be interesting to try an xgboost model with the data encoded as levels rather than as dummy variables

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

Linear model error estimation

To make sure that the AUC we get on the validation set we will also get on the test set (which is hidden from us), we should make a 5-fold cross validation where we can get a confidence interval on the AUC estimation

Graph Boosted Machine with XGBoost

Code done in Python here

Need to be just a little bit careful because the sklearn label encoder says 0 is baggage and 1 is no baggage, the opposite of what we define initially but that's ok