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 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).

```
##
## 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.

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

```
##
##             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 can therefore be ommitted

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

Synthetic variable family size

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

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

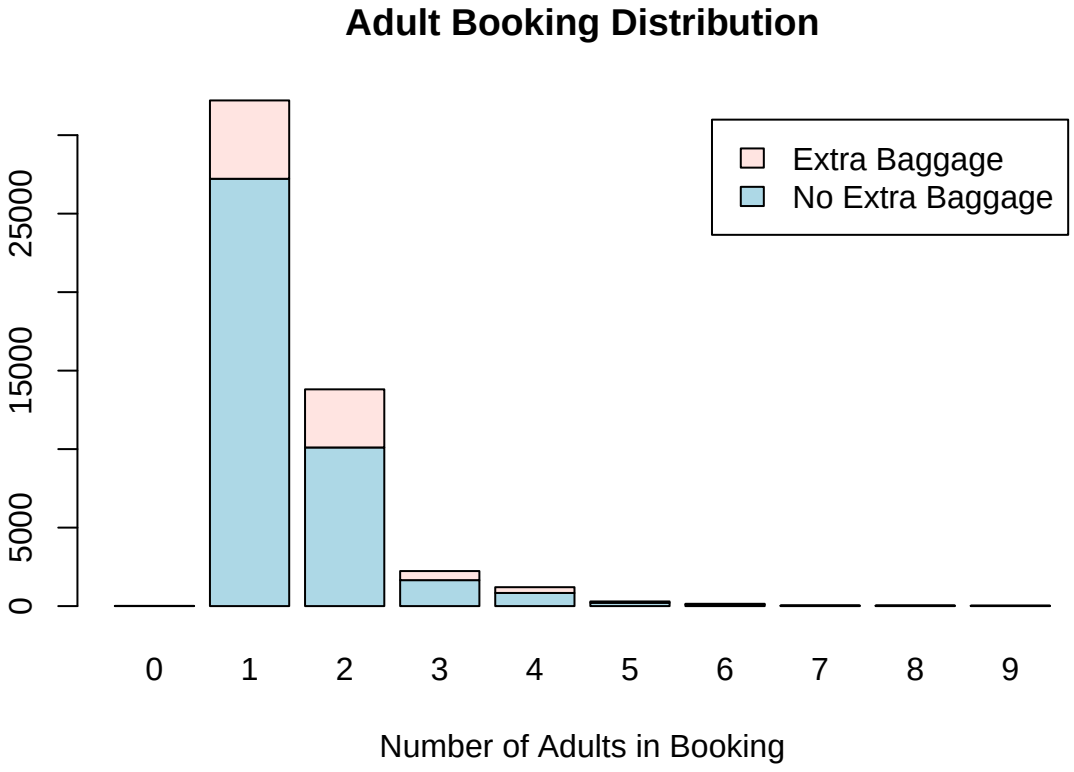
Could it be that with certain devices more customers book devices ?

```
##
##              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

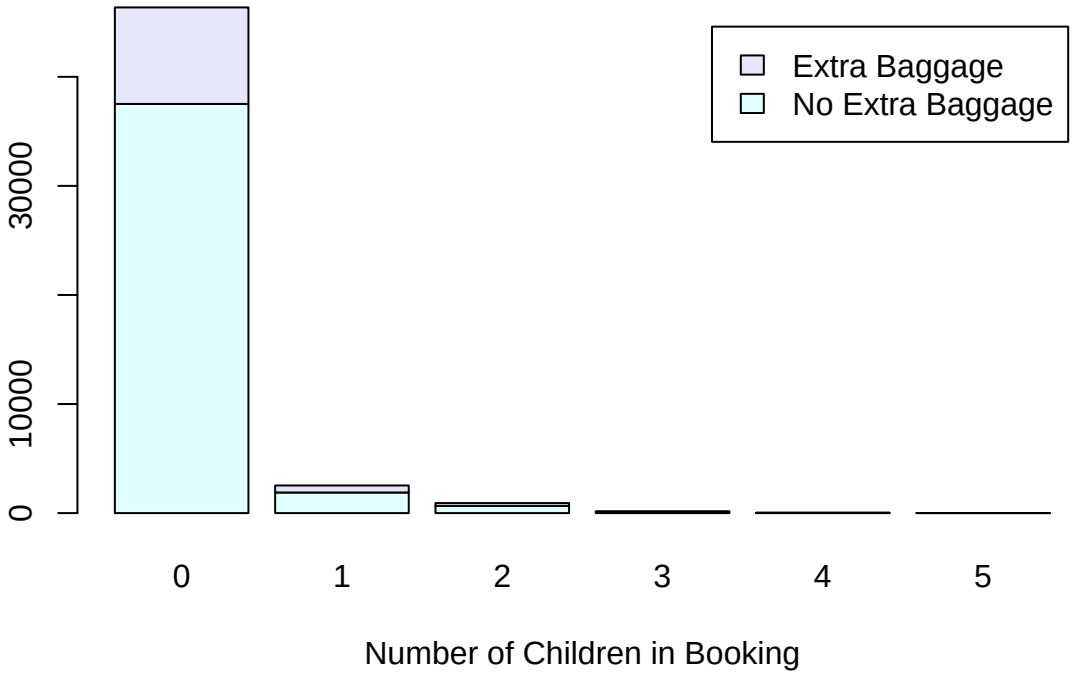
Adults travelling alone I would assume would be less likely to book luggage, but with one or more children much more likely to get luggage, especially with infants

Indeed from a small table you can see that:

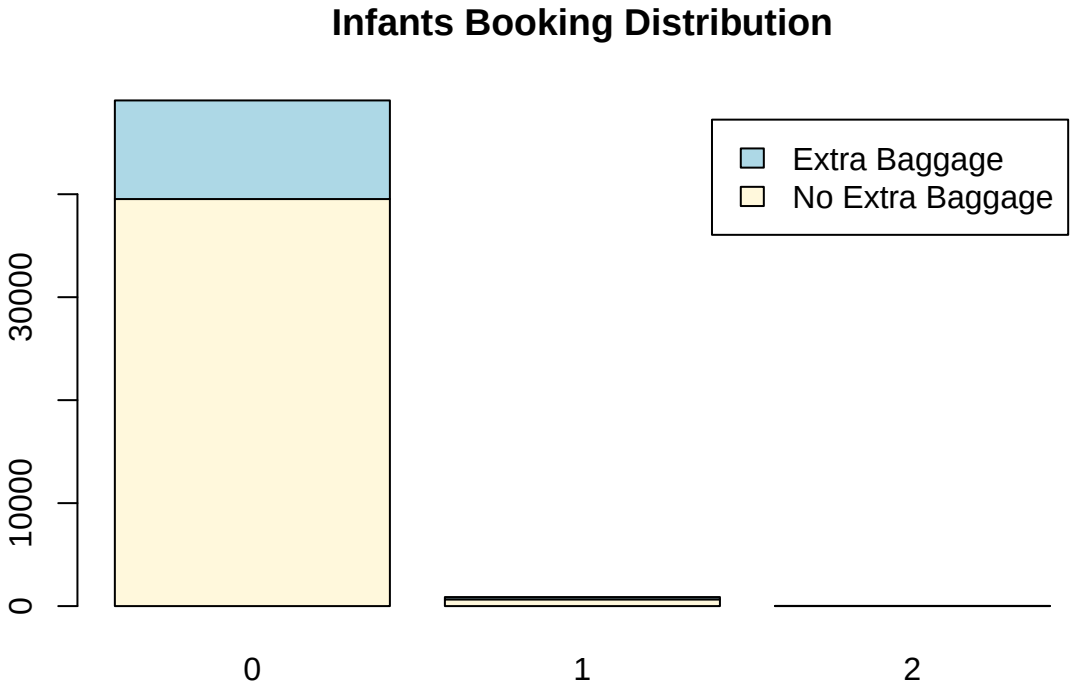



```
##
##     No Extra Baggage Extra Baggage
##   0            100.0           0.0
##   1             84.5          15.5
##   2             73.2          26.8
##   3             74.1          25.9
##   4             69.5          30.5
##   5             71.7          28.3
##   6             63.3    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

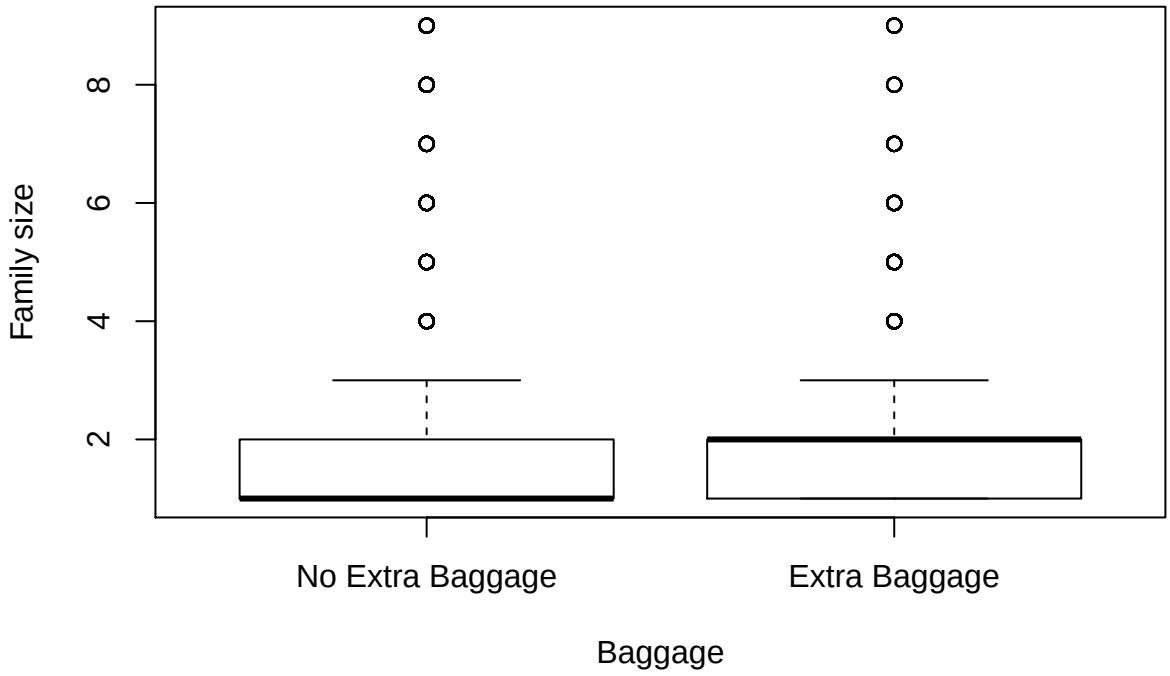
It seems that the more adults are travelling, the more likely they are to book luggage




```
##
##     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
```




```
##
##     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
```

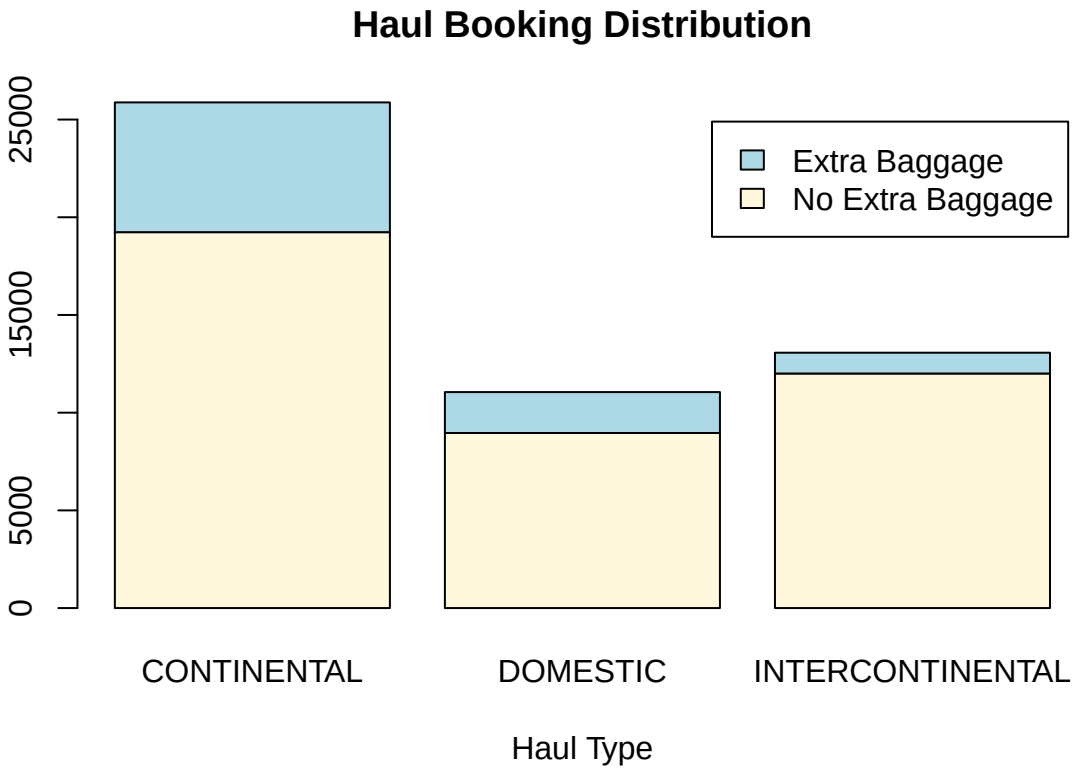
Increased family size also seems to bring with it increased probability of extra baggage selection.

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 traevellers 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
```

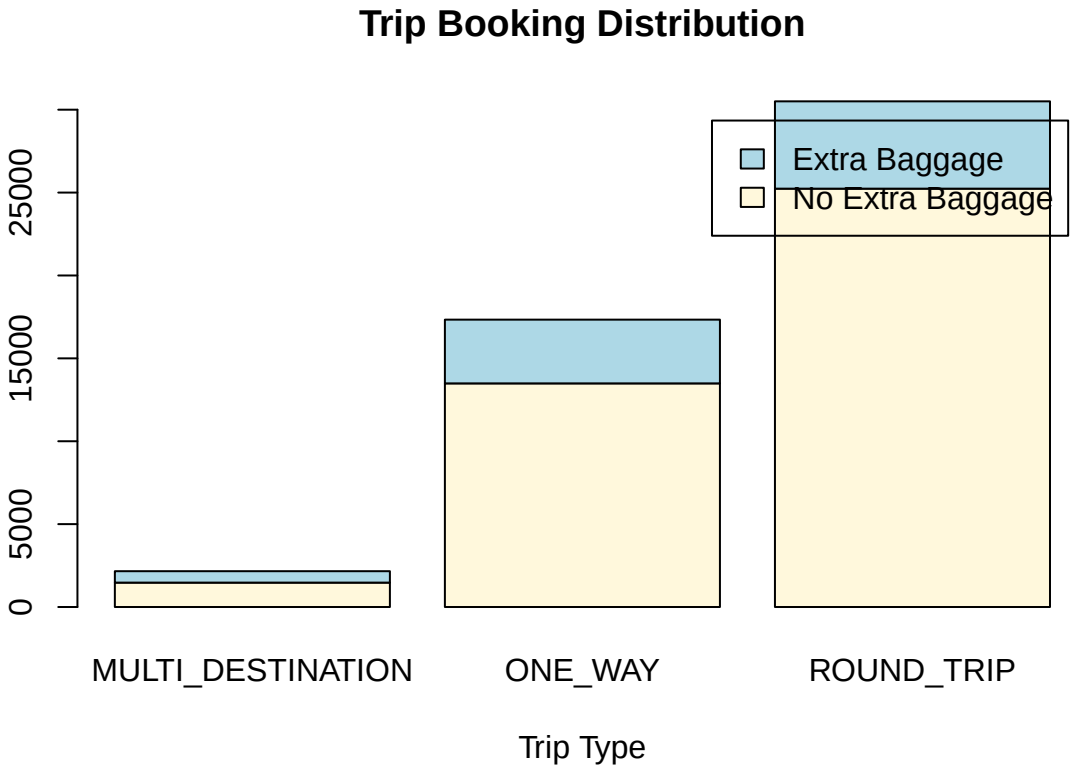
I would imagine that flight distance would account for a lot of the probability of luggage selection (high R2), as people who travel further I would assume need to carry more than if they are doing a short weekend trip within Europe




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

```
##   INTERCONTINENTAL             91.9           8.1
```

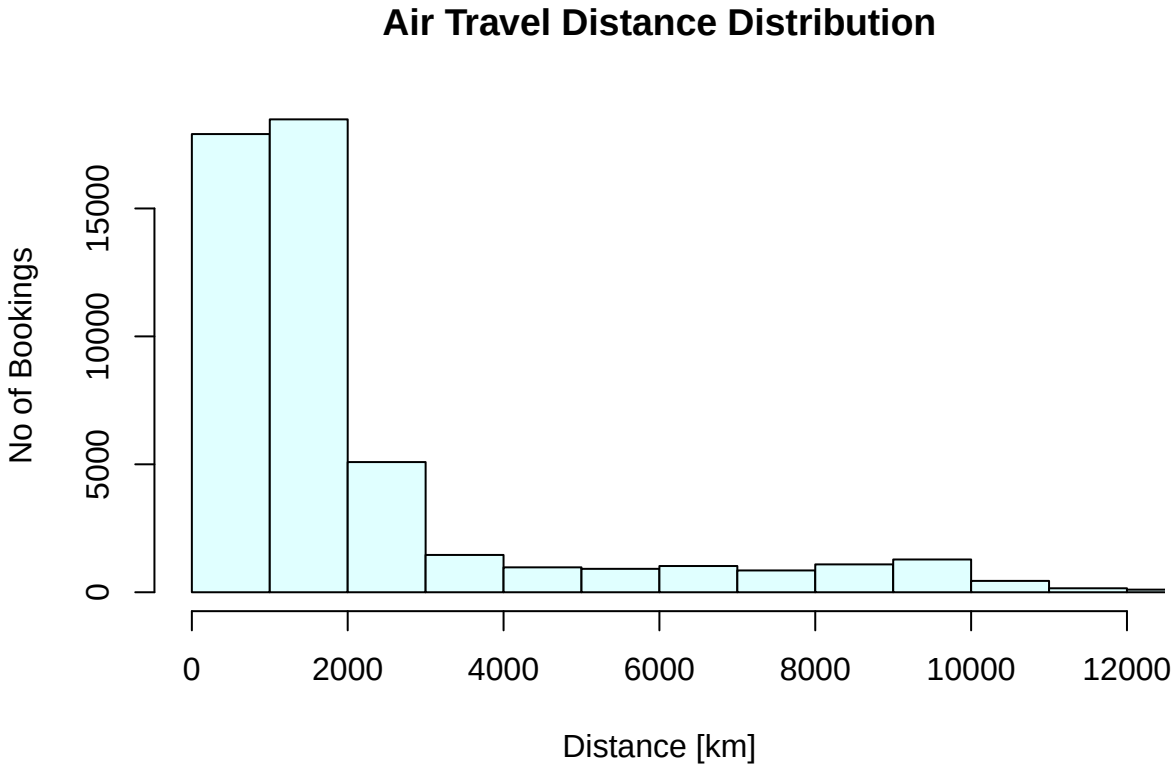
There are quite significan differences here between groups. One can imagine that in intercontinental flights, the luggage from more premium companies will be complimentary so no extra is needed. And for domestic flights it makes sense - travelling at home you might need less luggage.

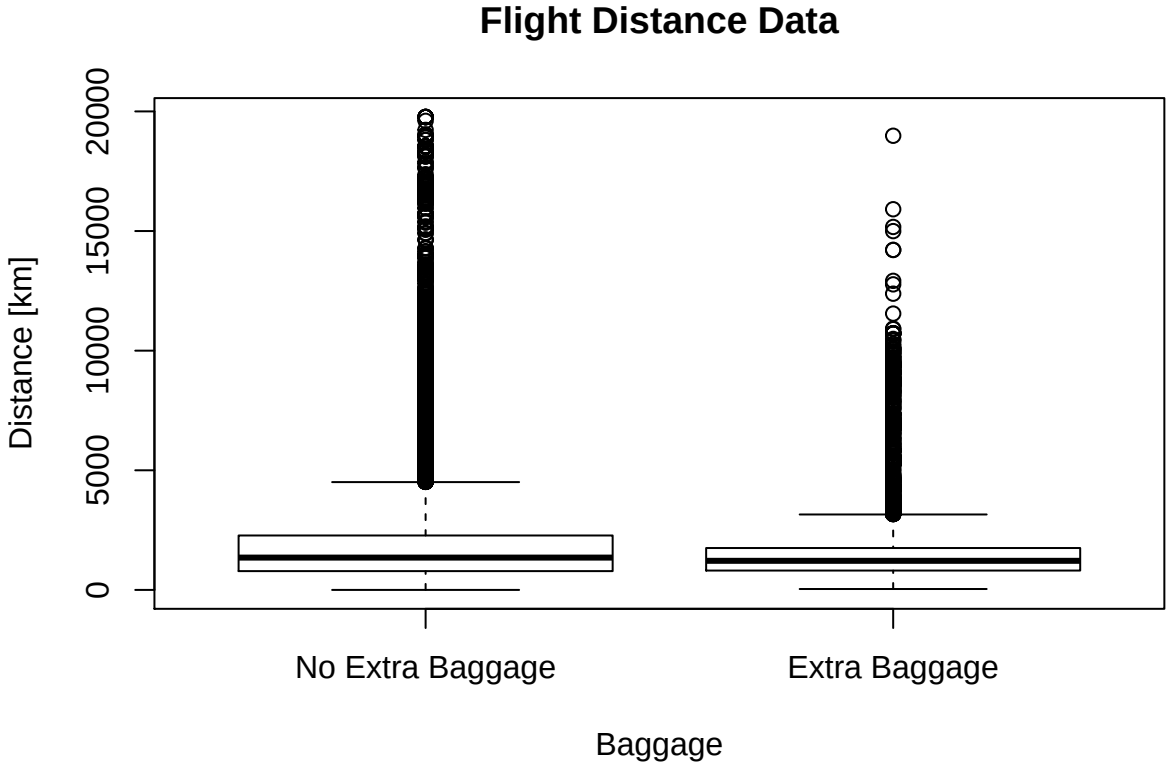



```
##
##                      No Extra Baggage Extra Baggage
##   MULTI_DESTINATION             68.1          31.9
##   ONE_WAY                       77.8          22.2
##   ROUND_TRIP                    82.7          17.3
```

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 skeweved normal distribution with alot of short flights between 0-3000km and then drastic reductions from then onwards.


Air Travel Distance Distribution
No of Bookings
0
5000
10000
15000
0
2000
4000
6000
8000
10000
12000
Distance [km]


Flight Distance Data
Distance [km]
0
5000
10000
15000
20000
No Extra Baggage
Extra Baggage
Baggage

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 hightly significant in terms of p-value:

```
##
## Call:
## glm(formula = EXTRA_BAGGAGE ~ DISTANCE + factor(HAUL_TYPE) +
##     factor(TRIP_TYPE) + ADULTS + CHILDREN + INFANTS, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6971  -0.7341  -0.6139  -0.3449   2.6021
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -8.539e-01  5.373e-02 -15.892  < 2e-16
## DISTANCE                        -5.518e-05  8.476e-06  -6.510 7.52e-11
## factor(HAUL_TYPE)DOMESTIC       -4.215e-01  2.887e-02 -14.602  < 2e-16
## factor(HAUL_TYPE)INTERCONTINENTAL -1.139e+00  4.366e-02 -26.092  < 2e-16
## factor(TRIP_TYPE)ONE_WAY        -4.411e-01  5.141e-02  -8.580  < 2e-16
## factor(TRIP_TYPE)ROUND_TRIP     -7.569e-01  5.022e-02 -15.073  < 2e-16
## ADULTS                           2.770e-01  1.266e-02  21.873  < 2e-16
## CHILDREN                         2.368e-01  2.774e-02   8.535  < 2e-16
## INFANTS                          2.438e-01  8.133e-02   2.998  0.00272
##
## (Intercept)                     ***
## DISTANCE                        ***
## factor(HAUL_TYPE)DOMESTIC       ***
## factor(HAUL_TYPE)INTERCONTINENTAL ***
## factor(TRIP_TYPE)ONE_WAY        ***
## factor(TRIP_TYPE)ROUND_TRIP     ***
## ADULTS                          ***
## CHILDREN                        ***
## INFANTS                         **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 49478  on 49999  degrees of freedom
## Residual deviance: 46638  on 49991  degrees of freedom
## AIC: 46656
##
## Number of Fisher Scoring iterations: 5
```

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.

```
##
## Call:
```

```
## glm(formula = EXTRA_BAGGAGE ~ DISTANCE + factor(HAUL_TYPE) +
##     factor(TRIP_TYPE) + factor(ADULTS) + factor(CHILDREN) + factor(INFANTS),
##     family = binomial(link = "logit"), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3694  -0.7246  -0.5959  -0.3293   2.6493
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -1.187e+01  1.008e+02  -0.118  0.90624
## DISTANCE                        -5.727e-05  8.502e-06  -6.736 1.62e-11
## factor(HAUL_TYPE)DOMESTIC       -4.314e-01  2.897e-02 -14.891  < 2e-16
## factor(HAUL_TYPE)INTERCONTINENTAL -1.113e+00  4.385e-02 -25.375  < 2e-16
## factor(TRIP_TYPE)ONE_WAY        -3.974e-01  5.177e-02  -7.676 1.64e-14
## factor(TRIP_TYPE)ROUND_TRIP     -7.448e-01  5.051e-02 -14.747  < 2e-16
## factor(ADULTS)1                  1.116e+01  1.008e+02   0.111  0.91183
## factor(ADULTS)2                  1.180e+01  1.008e+02   0.117  0.90676
## factor(ADULTS)3                  1.172e+01  1.008e+02   0.116  0.90744
## factor(ADULTS)4                  1.197e+01  1.008e+02   0.119  0.90544
## factor(ADULTS)5                  1.184e+01  1.008e+02   0.117  0.90651
## factor(ADULTS)6                  1.225e+01  1.008e+02   0.122  0.90324
## factor(ADULTS)7                  1.186e+01  1.008e+02   0.118  0.90632
## factor(ADULTS)8                  1.138e+01  1.008e+02   0.113  0.91010
## factor(ADULTS)9                  1.168e+01  1.008e+02   0.116  0.90776
## factor(CHILDREN)1                2.991e-01  5.037e-02   5.938 2.89e-09
## factor(CHILDREN)2                4.770e-01  7.818e-02   6.102 1.05e-09
## factor(CHILDREN)3                3.966e-01  1.890e-01   2.099  0.03584
## factor(CHILDREN)4               -7.105e-01  6.272e-01  -1.133  0.25734
## factor(CHILDREN)5                1.354e+01  3.247e+02   0.042  0.96673
## factor(INFANTS)1                 2.346e-01  8.371e-02   2.803  0.00507
## factor(INFANTS)2                -1.169e+01  8.089e+01  -0.145  0.88504
##
## (Intercept)
## DISTANCE                        ***
## factor(HAUL_TYPE)DOMESTIC       ***
## factor(HAUL_TYPE)INTERCONTINENTAL ***
## factor(TRIP_TYPE)ONE_WAY        ***
## factor(TRIP_TYPE)ROUND_TRIP     ***
## factor(ADULTS)1
## factor(ADULTS)2
## factor(ADULTS)3
## factor(ADULTS)4
## factor(ADULTS)5
## factor(ADULTS)6
## factor(ADULTS)7
## factor(ADULTS)8
## factor(ADULTS)9
## factor(CHILDREN)1               ***
## factor(CHILDREN)2               ***
## factor(CHILDREN)3               *
## factor(CHILDREN)4
## factor(CHILDREN)5
## factor(INFANTS)1                **
```

```
## factor(INFANTS)2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 49478  on 49999  degrees of freedom
## Residual deviance: 46313  on 49978  degrees of freedom
## AIC: 46357
##
## Number of Fisher Scoring iterations: 11
```

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:

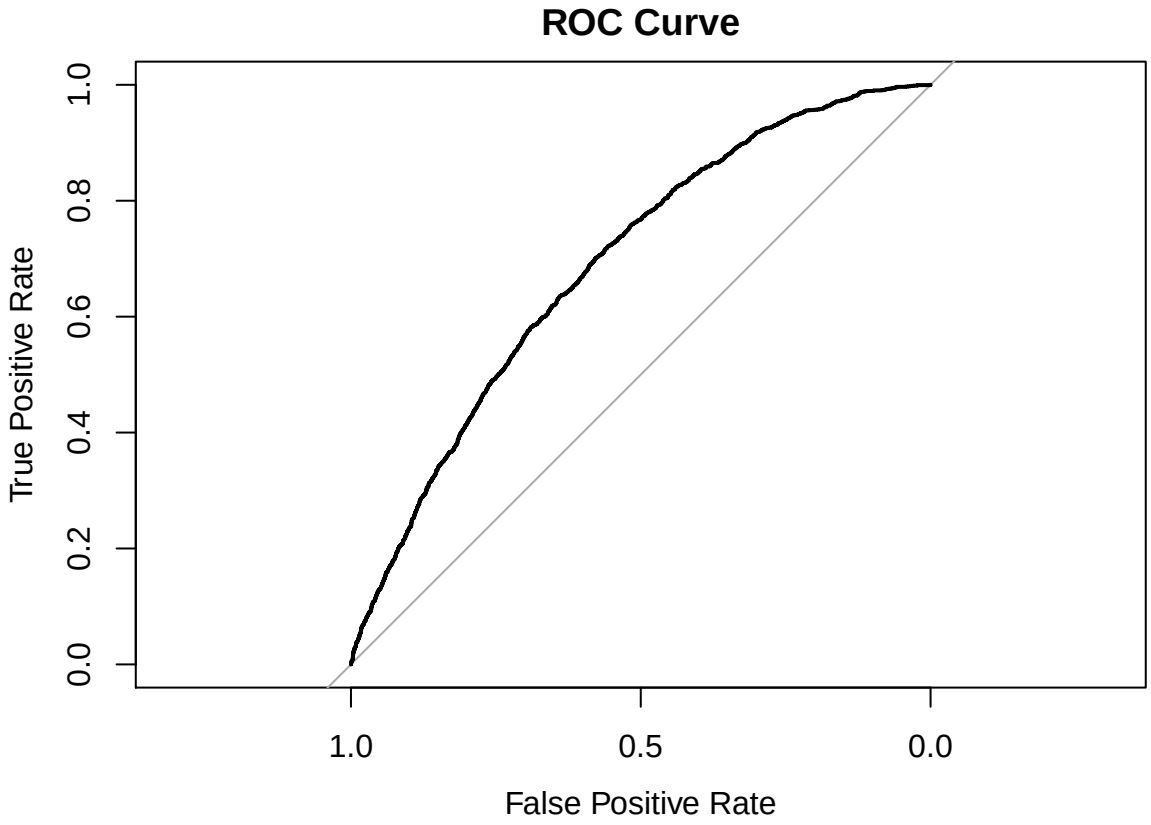
```
##
## Call:
## glm(formula = EXTRA_BAGGAGE ~ DISTANCE + factor(HAUL_TYPE) +
##     factor(TRIP_TYPE) + factor(DEVICE) + factor(COMPANY) + factor(FAMILY_SIZE),
##     family = binomial(link = "logit"), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3171  -0.7225  -0.5624  -0.3321   2.5300
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -8.204e-01  5.954e-02 -13.777  < 2e-16
## DISTANCE                        -4.363e-05  9.452e-06  -4.616 3.92e-06
## factor(HAUL_TYPE)DOMESTIC       -4.007e-01  3.271e-02 -12.249  < 2e-16
## factor(HAUL_TYPE)INTERCONTINENTAL -1.205e+00  5.070e-02 -23.768  < 2e-16
## factor(TRIP_TYPE)ONE_WAY        -3.155e-01  5.853e-02  -5.390 7.04e-08
## factor(TRIP_TYPE)ROUND_TRIP     -7.060e-01  5.709e-02 -12.367  < 2e-16
## factor(DEVICE)OTHER             -1.215e+01  8.917e+01  -0.136  0.89159
## factor(DEVICE)SMARTPHONE        -2.078e-01  3.296e-02  -6.303 2.91e-10
## factor(DEVICE)TABLET            -5.104e-02  5.501e-02  -0.928  0.35350
## factor(COMPANY)GO VOYAGE         3.616e-01  4.402e-02   8.214  < 2e-16
## factor(COMPANY)OPODO             1.871e-01  3.022e-02   6.193 5.90e-10
## factor(COMPANY)OTHER             1.239e+01  8.917e+01   0.139  0.88948
## factor(FAMILY_SIZE)2             6.127e-01  2.936e-02  20.865  < 2e-16
## factor(FAMILY_SIZE)3             6.410e-01  5.105e-02  12.556  < 2e-16
## factor(FAMILY_SIZE)4             8.750e-01  6.096e-02  14.353  < 2e-16
## factor(FAMILY_SIZE)5             8.928e-01  1.035e-01   8.625  < 2e-16
## factor(FAMILY_SIZE)6             9.211e-01  1.558e-01   5.912 3.38e-09
## factor(FAMILY_SIZE)7             1.236e+00  2.609e-01   4.738 2.16e-06
## factor(FAMILY_SIZE)8             6.004e-01  3.555e-01   1.689  0.09127
## factor(FAMILY_SIZE)9             1.047e+00  4.039e-01   2.592  0.00955
##
## (Intercept)                     ***
## DISTANCE                        ***
## factor(HAUL_TYPE)DOMESTIC       ***
## factor(HAUL_TYPE)INTERCONTINENTAL ***
## factor(TRIP_TYPE)ONE_WAY        ***
## factor(TRIP_TYPE)ROUND_TRIP     ***
```

```
## factor(DEVICE)OTHER
## factor(DEVICE)SMARTPHONE          ***
## factor(DEVICE)TABLET
## factor(COMPANY)GO VOYAGE          ***
## factor(COMPANY)OPODO              ***
## factor(COMPANY)OTHER
## factor(FAMILY_SIZE)2              ***
## factor(FAMILY_SIZE)3              ***
## factor(FAMILY_SIZE)4              ***
## factor(FAMILY_SIZE)5              ***
## factor(FAMILY_SIZE)6              ***
## factor(FAMILY_SIZE)7              ***
## factor(FAMILY_SIZE)8              .
## factor(FAMILY_SIZE)9              **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 39500  on 39889  degrees of freedom
## Residual deviance: 36964  on 39870  degrees of freedom
##   (110 observations deleted due to missingness)
## AIC: 37004
##
## 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
```

	odds	2.5 %	97.5 %
(Intercept)	4.402769e-01	3.915665e-01	4.945276e-01
DISTANCE	9.999564e-01	9.999377e-01	9.999748e-01
factor(HAUL_TYPE)DOMESTIC	6.698647e-01	6.281477e-01	7.140877e-01
factor(HAUL_TYPE)INTERCONTINENTAL	2.996791e-01	2.711397e-01	3.307619e-01
factor(TRIP_TYPE)ONE_WAY	7.294261e-01	6.506870e-01	8.185287e-01
factor(TRIP_TYPE)ROUND_TRIP	4.936250e-01	4.415957e-01	5.523684e-01
factor(DEVICE)OTHER	5.272689e-06	1.104940e-17	3.747769e-04
factor(DEVICE)SMARTPHONE	8.123831e-01	7.613864e-01	8.664154e-01
factor(DEVICE)TABLET	9.502433e-01	8.523682e-01	1.057529e+00
factor(COMPANY)GO VOYAGE	1.435614e+00	1.316499e+00	1.564477e+00
factor(COMPANY)OPODO	1.205794e+00	1.136366e+00	1.279273e+00
factor(COMPANY)OTHER	2.406503e+05	3.403378e+03	1.181690e+17
factor(FAMILY_SIZE)2	1.845350e+00	1.742077e+00	1.954606e+00
factor(FAMILY_SIZE)3	1.898366e+00	1.716672e+00	2.097058e+00
factor(FAMILY_SIZE)4	2.398817e+00	2.127210e+00	2.701559e+00
factor(FAMILY_SIZE)5	2.441988e+00	1.988672e+00	2.984773e+00
factor(FAMILY_SIZE)6	2.511944e+00	1.840634e+00	3.393657e+00
factor(FAMILY_SIZE)7	3.442593e+00	2.039806e+00	5.699803e+00
factor(FAMILY_SIZE)8	1.822775e+00	8.689566e-01	3.552327e+00
factor(FAMILY_SIZE)9	2.848256e+00	1.244662e+00	6.178388e+00

```
## Area under the curve: 0.6895
```



We notice that looking at the odds ratio table that 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: 0.6895
```

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