PS1 Data Analysis

Basics

- We have to ensure we don’t have too many parameters or else too much time to train
- We have to ensure to chose all required parameters
- We have to ensure to change the form of certain parameters (Hot-key-encoding)

Being PersonId 3 or 4 significantly increases chance of being Transported =TRUE
Since the %Diff is approximately same, let us club it into one parameter to be fed into our model.
Therefore two parameters we get are:
1) PID – PersonId except 3 or 4
2) PID3/4 – PersonId is 3 or 4

PersonId	TRUE	FALSE	% DIFF
1	2959	3258	5
2	789	623	12
3	353	218	24
4	137	94	19
5	70	58	9
6	41	34	9
7	23	23	0
8	6	7	8

Since we have only 3 home planets, we can add a parameter for each of them.
Hot-Key Encoding.
Being in Home Planet 1 or 3 is having significant effect on probability

Home Planet	TRUE	FALSE	% DIFF
1	1951	2651	15
2	920	839	5
3	1404	727	32

Whether having CryoSleep or not having it is not only having “% DIFF” significantly different, but also the effects on the probability of being transported is completely opposite.

Y – Higher True
N – Higher False

To Do: Y = 1, N = -1

We could also have done as 1/0 but 1/-1 won’t cause any harm since it is just the bias term would be different in both cases

CryoSleep	TRUE	FALSE	% DIFF
Y	2483	554	64
N	1895	3761	33

Cabin

- This is a parameter we would suspect logically also to hold significant importance since it is possible that people from specific cabins or specific part of the ship were transported
- Now, one thing we could do is to make a huge list of parameters, one for each cabin, but this would increase the parameters by a enormous amount
- Lets see how Data Analysis helps us to safely remove our way through this at the loss of some accuracy

- Even though T has the maximum effect but since the number of data points we have is not large enough to make any comments and also the number of such data points in testing case is also negligible (<10) so we can safely ignore it.
- As per the data with us, we can see that being in Deck B or Deck C is having the maximum effect on the probability of a person being transported
- D,E,F also have significant effect but if we keep on taking parameters we will have too many so approximate by dropping them
- A,G no significant effect on probability
- Let's analyse B,C if there are specific cabins which are causing this or any variations

	TRUE	FALSE	% Diff
A	127	129	1
B	572	207	47
C	508	239	36
D	207	271	13
E	313	563	29
F	1229	1565	12
G	1321	1238	3
T	1	4	60

- B0 – Negligible Data
- Rest all cause increase
- Increase we take it to be approximately same and consider Deck B as a single parameter

B BLOCK	TRUE	FALSE	% DIFF
0	0	1	100
1	207	70	49
2	191	71	46
3	66	22	50
4	20	7	48
5	18	7	44
6	17	9	31
7	13	4	53
8	22	9	42
9	18	7	44

- C0 – Negligible Data
- C4,C5 – Approximately No effect
- Rest all cause increase
- Increase for these we take it to be approximately same and consider Deck C as a two parameters 1) DC – Without C4,C5 2) DC4/5 – C4,C5

C BLOCK	TRUE	FALSE	% DIFF
0	0	1	100
1	185	83	38
2	177	78	39
3	56	30	30
4	9	9	0
5	13	12	4
6	17	5	55
7	14	6	40
8	19	8	41
9	18	7	44

Side

- The side on which the cabin is located is also having effect on the probability as we can observe below.
- We use S = 1 and P = -1 to make “Side” as a parameter which we can use to train our model

Side	TRUE	FALSE	% Diff
P	1898	2308	10
S	2380	1908	11

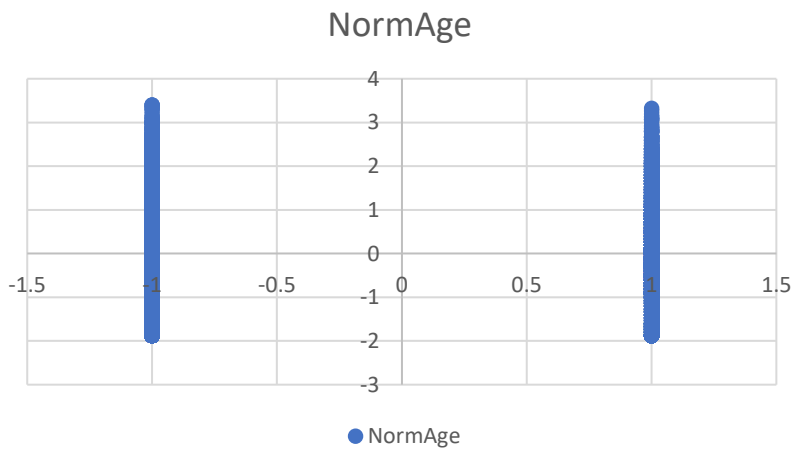
Destination Planet

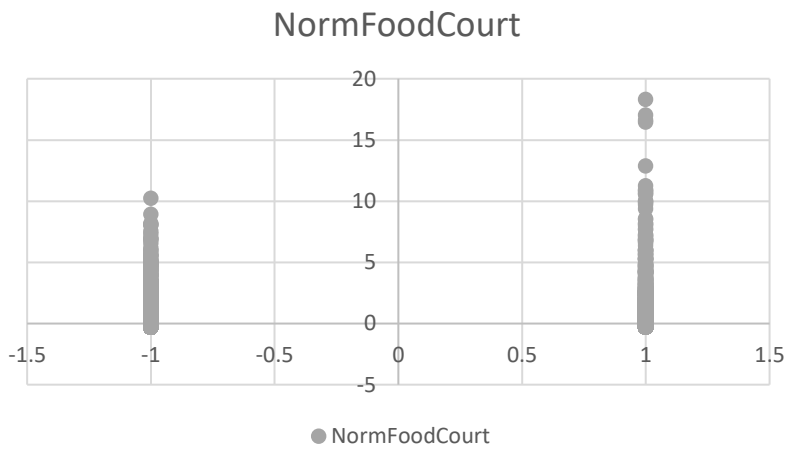
- We use an approach similar to Home Planet
- We create 3 parameters D1, D2 and D3

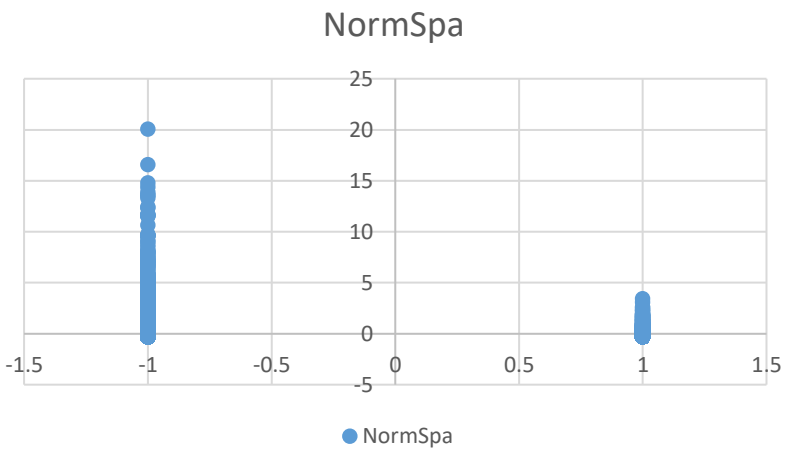
Normalization

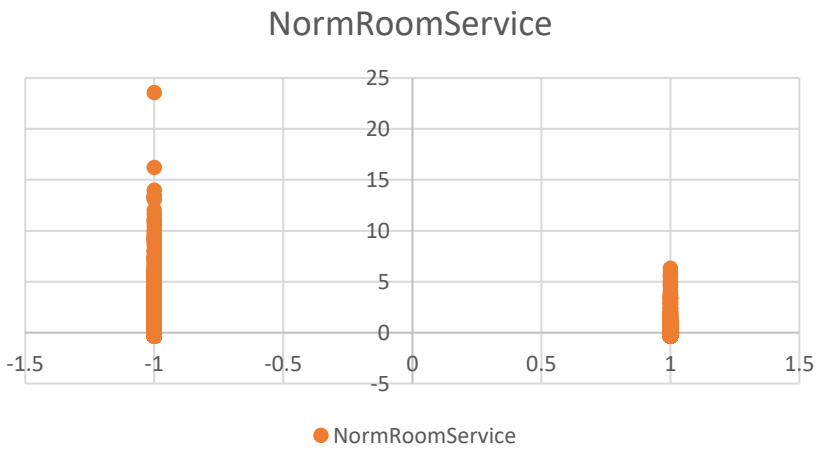
- This is a very important step in setting up our data
- Speeds up the learning rate of the model
- We process all our parameters such that their range has similar order
- Here we have subtracted the mean and then divided the result by the standard deviation
- This has brought a major fraction of all our data to approximately between -1 and 1
- Used for Age, Room Service, Food Court, Shopping Mall, Spa and VR Deck

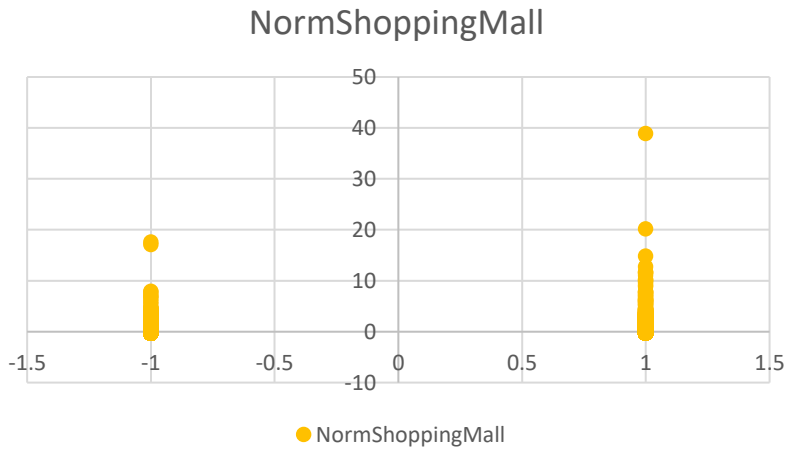
Scatter Plots of Parameters vs Transported to find Trends

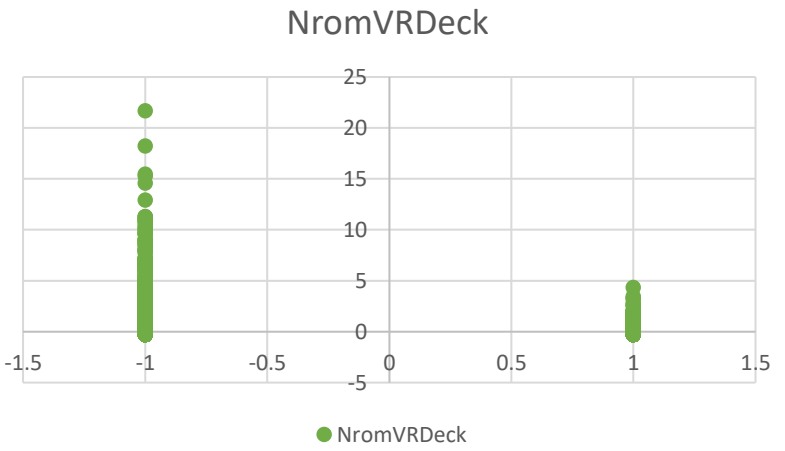












Conclusions from graphs

- Except for the Age, it is clearly observed that at higher values of the parameter, the probability of one of the consequences is higher.
- This is slightly counter-intuitive since logically the price we pay for a facility can’t affect our survival probability. At max, wether we have the facility or no can affect, but not price. This is regarding another dimension !
- NOTE: From this we made the first point as a conclusion but based on the previous slide we can’t surely conclude that close to zero the probability is equal for both. This is because they are scatter plots and since our data is not too small, we can make incorrect conclusions just by observation of scatter plots.

VIP

- Having availed the VIP service or has no significant role on the probability of being transported but not availing decreases the probability significantly
- Here if we take VIP: Y – 1 and N – (-1) then we can incorporate both the things in one parameter
- So let’s go with this

	FALSE	TRUE	% DIFF
VIP	4192	4302	1.30
NON-VIP	123	76	23.62

Conclusions

- By carefully analysing data we can wisely reduce our parameters which would help to speed up training our model
- It must be made a thumb rule to always analyse data before deciding which parameters to drop since relations can be found in places where it is intuitively never expected