583 Final Report

Henry Luo, Renghe Tang, Weijia Lyu

2023-03-24

Introduction

which encountered a spacetime anomaly recently. Half of the passengers were lost to an alternative dimension. An exploratory analysis investigating the relationship between passengers' features and their probability of being transported was launched and yielded troubling results. The transportation does not seem to be random. Passengers' chance of being transported appears to differ by their features. Cryosleep status, age, and even deck location seem to affect passengers' chance of being transported to the alternative dimension. The current project further investigates the situation with logistic regression analysis, hoping to gain further understanding of how and by what magnitude different features affect the chance of being transported. A better understanding of the relationship between passenger features and the chance of transportation could help other spaceships and passengers prepare for potential encounters with the spacetime anomaly.

Further processing of the data set

Additional cleaning and processing of the data set are conducted. 24% of the total passenger records contain missing values. Since the function in R for logistics regression ignores rows containing missing values by default, it is necessary to fill in the missing values to retain the information from 24% of the data in the analysis. Missing values in categorical variables are filled with "NR," representing not reported. Missing `RoomNum` is filled with -1 to distinguish it from the other room numbers. Missing values in numeric variables outlining passenger's spending are filled with mode 0 since those variables are extremely skewed and heavy on zero. Filing with mean or median does not seem appropriate. Passengers `Name` was found to be an important variable by the random forest model in the exploratory analysis. It seems to suggest that family status is a playing a role, as there are many passengers sharing the same last name. Therefore a new binary categorical variable family is added to denote if family members are on board. If a passenger shares the same last name and is in the same `Group` with other passengers, their `Family` variable is denoted as "True."

Logistic regression models

Model 1

allegedly fictional records of passengers aboard the spaceship Titanic. A logistic regression model, model 1, is fitted on the data. Backward selection method is used for variable selection. Variables are removed one by one from the full model until the simplest model with the lowest AIC is obtained. The resulting model contains 13 variables. Categorical variables `HomePlanet, CryoSleep, Deck, Side, Destination` and numerical variables `Age, RoomService, FoodCourt, ShoppingMall, Spa,` and `VRDeck` are included. `VIP`, `RoomNum,` and `Famly` are excluded The result is largely consistent with the finding in the exploratory analysis. The model obtained an in-sample log-loss of 0.431, a misclassification rate of 0.205, and a recall of 0.817. The relatively small in-sample log-loss and misclassification rate suggest that the model is a decent fit

for the data for the purpose of the current investigation. Since false negatives, falsely predicting a passenger to survive the anomaly, is particularly harmful, recall is selected to be an important metric to examine as it can be considered a measure of the model's ability to avoid false negatives. A recall of 0.817 indicates that the model can detect 81.7% of the true positive cases, leaving about 18.3% to potentially be false negatives. Model 1 is summarized in Table 1.

Table 1: Summary of Model 1

term	estimate	std.error	statistic	p.value
(Intercept)	-0.9259624	0.3064672	-3.0214078	0.0025160
Group	0.0000355	0.0000107	3.3233457	0.0008894
InGroupId02	0.1271100	0.0841635	1.5102748	0.1309733
InGroupId03	0.3855666	0.1233681	3.1253355	0.0017760
InGroupId04	0.2328078	0.1884792	1.2351911	0.2167594
InGroupId05	-0.2027269	0.2385149	-0.8499547	0.3953503
InGroupId06	0.0738403	0.2902590	0.2543945	0.7991908
InGroupId07	-0.3052866	0.3577581	-0.8533324	0.3934750
InGroupId08	-0.2233028	0.6474211	-0.3449112	0.7301611
HomePlanetEuropa	1.6481006	0.2175044	7.5773224	0.0000000
HomePlanetMars	0.5369988	0.0937765	5.7263698	0.0000000
HomePlanetNR	0.4600720	0.1873037	2.4562892	0.0140380
CryoSleepNR	0.3187379	0.1749380	1.8220045	0.0684543
CryoSleepTrue	1.3895050	0.0803738	17.2880256	0.0000000
DeckB	1.3185594	0.2595632	5.0799168	0.0000004
DeckC	2.4442840	0.2896543	8.4386255	0.0000000
DeckD	0.7683669	0.2836545	2.7088128	0.0067524
DeckE	0.3276287	0.2845483	1.1513992	0.2495680
DeckF	0.9582887	0.2868968	3.3401855	0.0008372
DeckG	0.5120737	0.2942372	1.7403430	0.0817988
DeckNR	0.9011037	0.3353507	2.6870484	0.0072087
DeckT	-0.2244331	1.7644864	-0.1271946	0.8987864
SideNR	NA	NA	NA	NA
SideS	0.5865482	0.0582112	10.0762121	0.0000000
DestinationNR	-0.0428577	0.2086582	-0.2053967	0.8372622
DestinationPSO J318.5-22	-0.4220829	0.1123177	-3.7579371	0.0001713
DestinationTRAPPIST-1e	-0.4633489	0.0811116	-5.7124828	0.0000000
Age	-0.0077990	0.0021066	-3.7021648	0.0002138
RoomService	-0.0015092	0.0000915	-16.4861683	0.0000000
FoodCourt	0.0005083	0.0000405	12.5469052	0.0000000
ShoppingMall	0.0005371	0.0000665	8.0729857	0.0000000
Spa	-0.0020393	0.0001053	-19.3676005	0.0000000
VRDeck	-0.0019153	0.0001030	-18.6022239	0.0000000

Reduced model 1

The random forest model in the exploratory analysis suggested that `InGroupId` has the second lowest variable importance, with the lowest being `VIP.` In the logistic regression model, only one level of `InGroupId` is found to have a significant difference from the reference level. It seems possible that `InGroupId` could be potentially excluded from the model without significantly affecting model fit to further simplify the model. A likelihood ratio test (LRT) is therefore conducted to see if `InGroupId` should be included in the model. A reduced model without `InGroupId` is fitted and compared with the full model in LRT. The resulting p-value is 0.000163042. The p-value is smaller than 0.05, and the null hypothesis that `InGroupId` should not be included in the

model is rejected. LRT test suggests that no additional variable can be removed. The model appears to be the simplest in terms of the number of variables

Model 2: treatment of complex categorical variables

Model 1 is not optimal in terms of its interpretability. Some categorical variables, such as `InGroupId` and `Deck` have too many levels. Since the levels of the categorical variables are only compared to the reference level, it is inefficient and difficult to interpret categorical variables with many levels. To cope with the problem, model 2 is fitted with the variable `Deck` re-coded according to findings from the exploratory analysis. Deck A, T, and NR, which are found to have no difference in the probability of transportation are coded as 0. Deck B, C, and G, which have a higher probability of being transported are coded as 2. Deck D, E, and F, with lower probability of being transported, are coded as 1. `InGroupId` and `RoomNum`are treated as numeric. Backward selection with AIC is applied. Model 2 is summarized in Table 2.

The in-sample log-loss of model 2 is 0.437, which is slightly higher than that of model 1. The misclassification rate is 0.205, and the recall is 0.817, which is the same as that of model 1. It appears that the prediction accuracy of the model is not significantly affected by the treatment of the categorical variables. However, the interpretability of model 2 is much better than that of model 1. It is clear now that both deck category 1 and 2 actually have an increased chance of being transported, and `InGroupId` is excluded from the model when treated as a numeric variable.

Table 2: Summary of Model 2

term	estimate	std.error	statistic	p.value
(Intercept)	-1.5730123	0.2448063	-6.425540	0.0000000
Group	-0.0000329	0.0000193	-1.708889	0.0874715
HomePlanetEuropa	2.7022543	0.1460595	18.501051	0.0000000
HomePlanetMars	0.6882510	0.0899633	7.650349	0.0000000
HomePlanetNR	0.6272777	0.1814790	3.456476	0.0005473
CryoSleepNR	0.3140497	0.1731541	1.813701	0.0697237
CryoSleepTrue	1.3412688	0.0792398	16.926701	0.0000000
Deck1	1.4114958	0.2297372	6.143958	0.0000000
Deck2	1.3065909	0.2220593	5.883974	0.0000000
RoomNum	0.0004503	0.0001093	4.117794	0.0000383
SideNR	1.8088253	0.3006340	6.016703	0.0000000
SideS	0.5856648	0.0578344	10.126577	0.0000000
DestinationNR	-0.0665490	0.2064283	-0.322383	0.7471625
DestinationPSO J318.5-22	-0.4481952	0.1119037	-4.005186	0.0000620
DestinationTRAPPIST-1e	-0.4999876	0.0803122	-6.225551	0.0000000
Age	-0.0085944	0.0020383	-4.216454	0.0000248
RoomService	-0.0014714	0.0000902	-16.305991	0.0000000
FoodCourt	0.0005171	0.0000399	12.970457	0.0000000
ShoppingMall	0.0005454	0.0000669	8.154776	0.0000000
Spa	-0.0019673	0.0001042	-18.877177	0.0000000
VRDeck	-0.0018170	0.0000994	-18.283714	0.0000000

Model 3: Log-transformation

The numerical variables are highly skewed, with the majority of values being 0. To assess potential effects of extreme skewness on the model fit, model 3 is fitted on log-transformed numerical data. The in-sample log-loss of model 3 is 0.498, which is higher than model 1 and model 2. The misclassification rate and recall are both worse for model 3 than for the other models, which are 0.237 and 0.760, respectively. It seems that

log transformation may not have preserved the underlying structure of the numerical variables. Model 3 is summarized in Table 3.

Table 3: Summary of Model 3

term	estimate	std.error	statistic	p.value
(Intercept)	-1.2976461	0.3586666	-3.6179729	0.0002969
HomePlanetEuropa	1.9898681	0.1250478	15.9128541	0.0000000
HomePlanetMars	0.3464210	0.0882728	3.9244391	0.0000869
HomePlanetNR	0.4472310	0.1807364	2.4744932	0.0133425
CryoSleepNR	0.2639506	0.1628999	1.6203237	0.1051628
CryoSleepTrue	1.2837934	0.0941617	13.6339280	0.0000000
Deck1	0.5871869	0.1935915	3.0331234	0.0024204
Deck2	0.2364313	0.1832843	1.2899699	0.1970611
SideS	0.5345194	0.0539560	9.9065834	0.0000000
DestinationNR	0.0362138	0.1965598	0.1842382	0.8538266
DestinationPSO J318.5-22	-0.3100037	0.1096797	-2.8264443	0.0047068
DestinationTRAPPIST-1e	-0.3789812	0.0714348	-5.3052725	0.0000001
Group	-0.1536153	0.0658174	-2.3339598	0.0195978
RoomNum	0.2289223	0.0572451	3.9989867	0.0000636
Age	-0.1416986	0.0239861	-5.9075243	0.0000000
RoomService	-0.0971865	0.0071103	-13.6684372	0.0000000
FoodCourt	0.0650429	0.0072245	9.0030408	0.0000000
ShoppingMall	0.0481858	0.0069472	6.9360329	0.0000000
Spa	-0.1203082	0.0071605	-16.8016827	0.0000000
VRDeck	-0.1219958	0.0073803	-16.5299395	0.0000000

The best model

The performance metrics of the three models are summarized in Table 4. It is apparent that model 1 has the best performance for prediction, but the interpretability of model 1 is not ideal due to the many levels in the categorical variables. Model 2 has a very similar prediction performance compared to model 1 and better interpretability due to the further processing of the categorical variables. Therefore, model 2 appears to be the most appropriate model for investigating the relationship between passenger features and their chance of being transported.

Table 4: Performance Metric of Fitted Models

Model	MisclassificationRate	Recall	LogLoss
Model 1	0.2053376	0.8174966	0.4310090
Model 2	0.2053376	0.8170397	0.4365781
Model 3	0.2372263	0.7599345	0.4983272

Diagnostics of model 2

Model 2 is diagnosed for potential issues.

Multicollinearity

Multicollinearity among variables is a problem that affects logistic models. To rule out the effect of multicollinearity on model 2, the variance inflation factor (VIF) is calculated. The VIF of all variables is below

5, which suggests that there is no significant multicollinearity in the data.

Residual plots

The Pearson residual plots of each variable are displayed in Figure 1 and Figure 2. Figure 1 contains nine variables, Figure 2 contains the remaining four variables, and the plot of the residuals against the predicted values. The x-axis represents the values of the variables. The y-axis represents the Pearson residual corresponding to each value of the variable. It seems that model 2 sufficiently captures the relationship between the predictor variables and the response, as the residuals are mostly randomly distributed above and below zero with no significant patterns. However, there seem to be some outliers in the data whose residuals deviate significantly from the rest of the data, especially in spending-related variables such as `ShoppingMall`, `VRDeck`, and `Spa`.

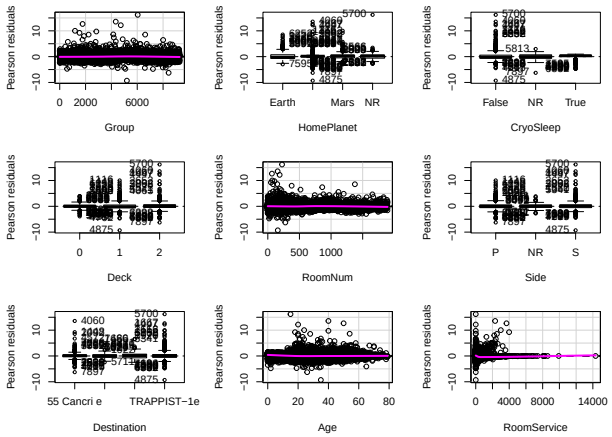

Figure 1: Pearson residual vs. variables of model 2

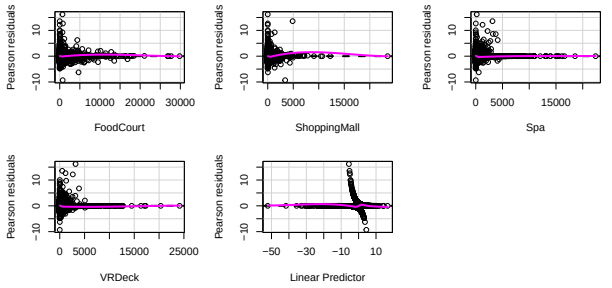

Figure 2: Pearson residual vs. variables of model 2

Outliers

Outliers of the spending-related numerical variables are removed by removing data that are two standard deviations away from the mean. 1294 entries are removed, which is 14.9% of the entire data set. The model

is fitted again without the outliers. The resulting model does not differ interms of the important variables. The in-sample log-loss increased to 0.458 from 0.437 and the misclassification rate increased to 0.222 from 0.205. Only recall is improved slightly from 0.817 to 0.824. It seems removing outliers does not improve model fit except for making the model less vulnerable to false negatives. It seems that the outliers' effect is not very significant for the current model.

Cross-validation

Cross-validation is conducted for model 2 to rule out potential overfitting. All of the logistic models fitted in the current analysis produced predictions of 0 and 1, which could be an indicator of overfitting. The 30% of the data set with categorical variables re-coded for model 2 is randomly sampled to be test set. The model is fitted on the remaining 70% of the data. The cross-validated log-loss is 0.439, which is very close to the in-sample log-loss of 0.437. The misclassification rate and recall are also very close to the in-sample measures, which are 0.203 and 0.811, respectively. The result indicates that the model is not likely to be overfitting.

Results

The most appropriate logistic regression model is determined for the analysis of the relationship between passenger features and the chance of being transported. 13 variables are found to affect the chance of being transported. Among the numerical variables, `Age`, `Group`, `RoomService`, `Spa`, and `VRDeck` negatively affect the probability of being transported. Passengers who are older, traveling with a group with a larger group number, using more room service, and spending more on entertainment on the ship are less likely to be transported. `FoodCourt`, `ShoppingMall`, and `RoomNum` are positively related to the probability of being transported. Passengers that eat more, shop more, and live in a room with a larger room number are more at risk of being transported.

For categorical variables, passengers whose home planet is Europa or Mars are more at risk of being transported than passengers who live on Earth, with passengers from Europa being more at risk than those from Mars. Passengers in CryoSleep or on the starboard side of the ship are more at risk of being transported. Passengers traveling to 55 Cancri e are more at risk than those traveling to PSO J318.5-22 and TRAPPIST-1e, with no difference between the latter two destinations. Deck area 1 (D, E, F) and 2 (B, C, G) are more vulnerable to transportation than deck area 0 (A, T, NR), with deck area 2 being relatively safer than deck area 1.

The magnitude of the change in probability of being transported for each passenger when the passenger's features change can also be investigated from the model. For example, for an infant below age one who was in cryosleep, their probability of being transported would decrease by 0.31 if they woke up from cryosleep. A 39 years-old passenger would only have a decreased probability of 0.07. A passenger who was in deck area 0 would become more at risk of being transported if they were in deck area 1 by having an increased transported probability of 0.2.

Conclusion

The current project identified an appropriate logistic model to investigate the relationship between the features of passengers on board the spaceship Titanic and their chance of being transported to an alternative dimension by the spacetime anomaly. Statistical evidence is obtained to determine that passengers were not transported at random. There are 13 distinct features affecting the probability of being transported. The recommendation from the current investigation to the spaceships in nearby areas and the passengers on board are the following. Prioritize evacuating children who are Europa residents in cryosleep on the starboard side of deck B, C, , or G, and are traveling to G 55 Cancri e. Direct passengers and crew unable

to evacuate to the port side of deck A and T. Passengers should be aware that spending in food courts and shopping malls appear to slightly increase the risk of being transported while spending on room service, spa, and VR deck seems to slightly decrease the risk of being transported.

The current logistic model has limitations. Its prediction accuracy is not as good as non-parametric models such as a random forest. The definition of outliers in the passengers' records may need to be further refined due to the complex and extremely skewed nature of the data. The insights gained from the logistic model can be used with non-parametric models and additional information about the situation to arrive at a more detailed understanding of the mechanism of the spacetime anomaly.