Semester Project

Arinah Karim[1]*

Abstract

While the human eyes may not be able to perceive patterns in data, machine learning algorithms most definitely can. By mining data, hidden patterns can be observed and relationships among variables can be established. In this Kaggle project, I will tackle the problem at hand using Logistic Regression and Random Forest, two types of classification methods, to solve the Space Titanic problem and explore all the possible variables and relationships from the given dataset. The datasets used in this problem were from Kaggle's Website

Keywords

Logistic Regression — Random Forest — Kaggle

[1] *Computer Science, School of Informatics , Computing and Engineering, Indiana University, Bloomington, IN, USA*

Contents

1. Problem and Data Description

The Spaceship Titanic collided with some sort of space anomaly and the task is to try to determine which passengers have been transported to another dimension given training data. The training data is composed of various features for 8693 people. These features are: the person's ID, their home planet, whether or not the person had entered cryosleep, the cabin number the person was staying in during their voyage, the planet destination, their age, whether or not they paid for VIP, various fees for the ship's amenities, the name of the passenger, and whether or not they had been transported to another dimension.

2. Data Preprocessing & Exploratory Data Analysis

2.1 Handling Missing Values

There were several categories of data that had N/A column values. These were: age and various fees for the ship's amenities. Additionally, there were features with blank cells, which were: home planet, cyrosleep, cabin number, destination planet, and VIP.

First, I will tackle the issue of empty cells in a very naive implementation for this milestone. For home planet, there are 3 possibilities where one can come from: Earth, Mars, and Europa. Around 53% of people came from Earth, roughly 25% from Mars, and 20% from Europa. The last 2% were unidentified, and thus the values to fill in. Based on the probabilities above, I created a random number generator that would pick a number from 1 through 100 and assign a planet to someone based on what number was chosen. This tactic was used for the other features containing empty cells.

Age was missing data, so I wanted to look at the distribution of ages visually. The histogram and QQ-plot show that the age distribution is skewed to the right, so rather than looking at the mean for a filler, I looked at the median and made that the value to be filled in for missing values.

For all of the amenities (food, mall, VR deck, spa) spendings, I filled them with 0s because that was what the mean was for all of them.

199 cabin assignment values were missing in the training data. Because all passengers had a passenger ID that indicated if they were a part of a group, I reduced some of the missing values by assigning someone of the same cabin to one of the people they traveled with randomly. This took care 64 missing values, so the 135 remaining values I assigned based on the previous value before.

2.2 Exploratory Data Analysis

First, I took a look at the number of people who were transported in the training data. Looking at the bar graph in Figure 1 of how many people were and were not transported, it looks almost evenly split.

The next step I took was to find patterns in the amount of people who were transported. To do this, I created various bar graphs of the transported data with things such as home planet, destination planet, various amenities etc. The graphs below show the results of these.

Examining the results of the number of people transported

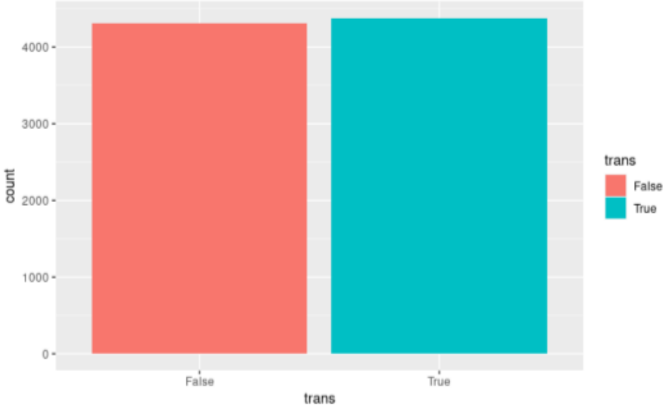


Figure 1. Bar graph of distribution of transported

and what planet they hail from, most people from Earth were not transported, while it was more likely that people from Mars and Europa were transported than not, but only slightly as can be seen in Figure 2.

When the destination planet is plotted with the people who were transported or not, it looks much more visually different than the results of the home planet as seen in Figure 3. PSO J318.5-22 has a nearly even split as to who was transported or not. It was more likely for people going to 55 Cancri e to be transported, while the opposite was true for people traveling to TRAPPIST-1e.

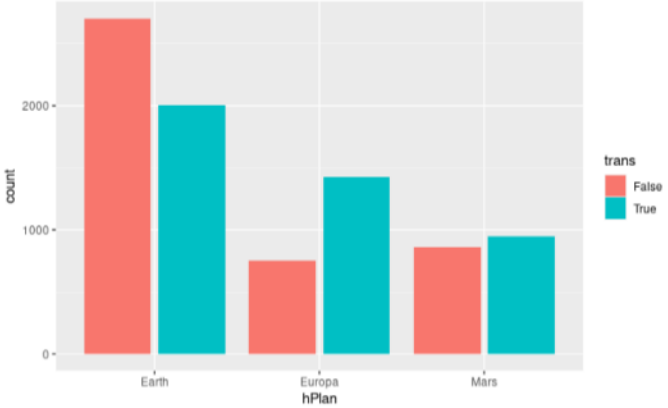


Figure 2. Transported passengers based on home planet

Seeing that both destination planet and home planet seemed to be a factor in who was transported, I decided to combine the two data and create a new column in both the training and testing data. I used the training data to see patterns in who was transported, as shown in Figure 4. Looking at the data, there are 9 possible different combinations, which makes sense given that there are 3 home planets and 3 destination planets. One of the most interesting pairs of home planet and destination planet is the home planet Europa paired with PSO J318.5-22 where everyone who came from that home planet and destination planet were teleported. However, compared to the amount of people who fell into this category (26), was a much smaller amount than the other possibilities. For instance, people hailing from Earth and went to TRAPPIST-1e were mostly not teleported, however, this data set contains 3252 people. Of the people hailing from Earth and traveling to the other planets, it was a pretty close split on who was transported or not, except for the case of Earth travelers going to TRAPPIST-1e. Looking at Europa travelers, it was more likely that the passenger would be transported across the 3 possible destination planets. And finally, Mars. Mars had a very odd distribution with destination planets, as most people who left from Mars went to TRAPPIST-1e as well. It looked more likely for people to be teleported if they left from Mars, but Mars passengers going to PSO J318.5-22 was a pretty even split as well.

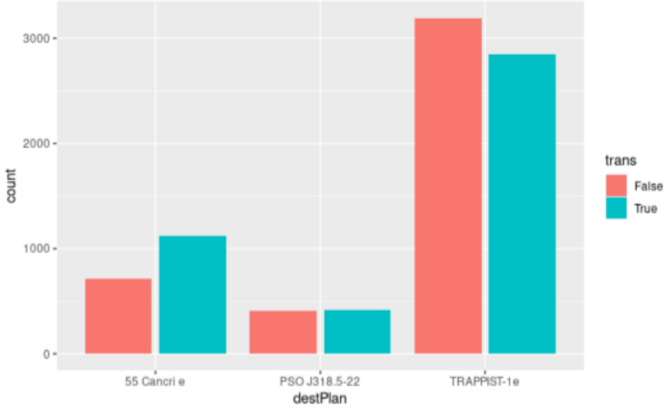


Figure 3. Transported passengers based on destination planet

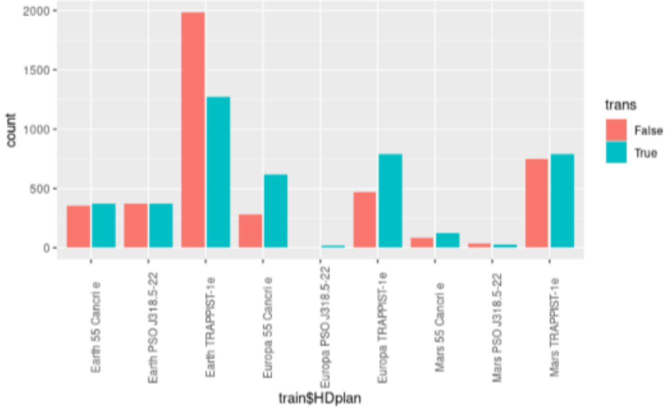


Figure 4. Bar graph of distribution of transported

Next, I examined how cryosleep affected transportation. Looking at Figure 5, one can see that when a passenger did cryosleep, it was more likely they were transported. The opposite of that could be said for people who did not partake in cryosleep.

The cabin information consists of 3 parts: the deck level, the room number, and what side of the ship the cabin was on (either Starboard or Port). Each of these might be an important aspect to who was transported, so I created a column for

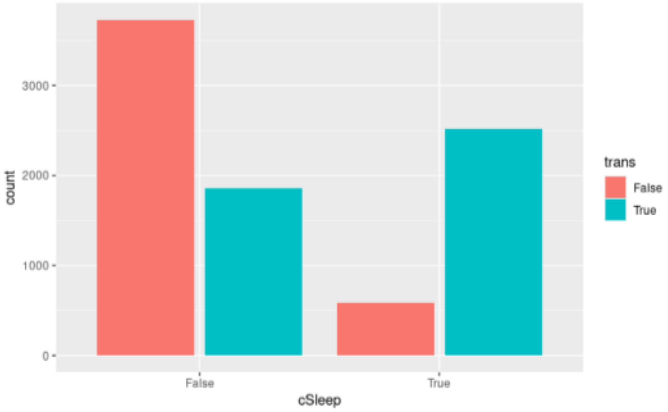


Figure 5. Cryosleep affects transportation rate

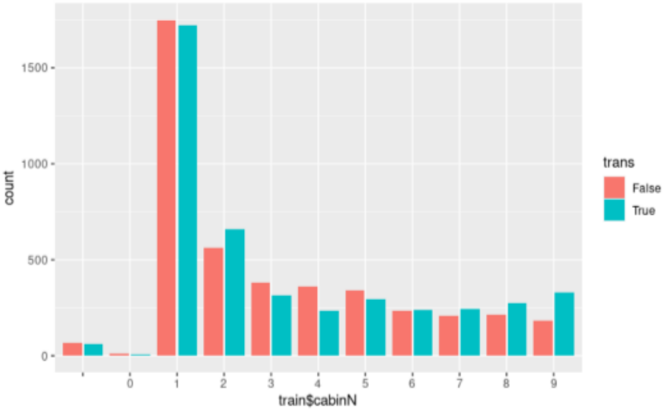


Figure 7. Cabin room number and transportation

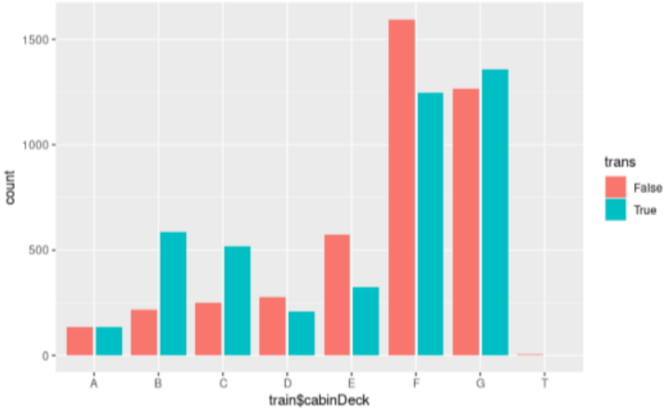


Figure 6. Cabin Deck Level and Transportation

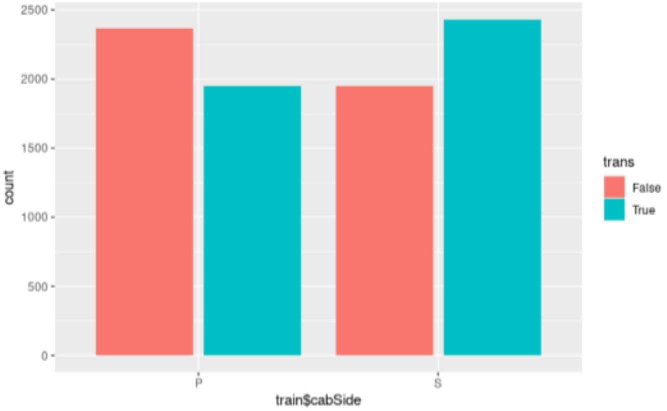


Figure 8. Side of ship and transportation

each of these 3 factors in both the testing and training datasets. These results can be seen in Figure 6. Looking at the deck level, 4 out of the 8 had results of being more likely than not of the passenger being transported. However, decks A and D were pretty close in numbers for who was transported or not. Decks F and G had the most amounts of passengers. F had significantly more people not transported, as well as Deck E. Decks B and C were the opposite of Decks F and E where more people were transported than not.
Further examining the cabin information, the number with the highest distribution was 1. Visually, it looks like whether the person was transported or not was a 50-50 split. To examine this information more, look at Figure 7.
Based on the distributions seen in Figure 8 of the passengers who were transported based off of what side of the ship they were on, passengers on the Starboard side of the ship were transported more than if they were on the Port side. Additionally, it was more likely for someone on the Port side of the ship to not be transported.
Looking at the various amenities, I did not see a correlation.

While looking at Figure 9 for the distribution of VIP, it looks like the status did not affect the rate of transportation.

Because age is all over the place, looking at the bar graph alone was not so helpful, so I also created a density graph to take a look at the distribution, as seen in Figure 10. Examining the figure, we see that younger children were transported more often than not, while 20+ were not transported as much. But looking at ages 60+, the number of people who were teleported was nearly the same as people who were not teleported. Therefore, I decided to partition the data so that the ages would be in bins that depended on what stage of life the passenger was in. Therefore, pre-teens would be 0-12, young adult would be 13-19, adult would be 20-60, and seniors would be 60 onwards. However, R did not like when I set the bound to 0, so I set it to -infinity instead. The results can be seen in Figure 11. For seniors, it looks like there really is no correlation. Looking at the pre-teen distribution, it looks like it was more likely they were transported than not. Adults were more likely to not be transported.

Amenities were quite messy to look at. The mean was 0 for all of them, so I decided to approach the amenities like I had done for the age stages. I separated the data based on whether the spending was equal to 0 or more than 0. The results showed significance across them. For those who did not spend money on the amenities, more of those people were

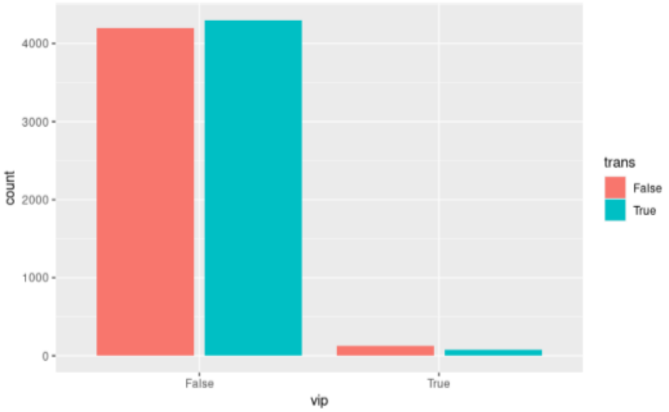


Figure 9. VIP status and Transportation

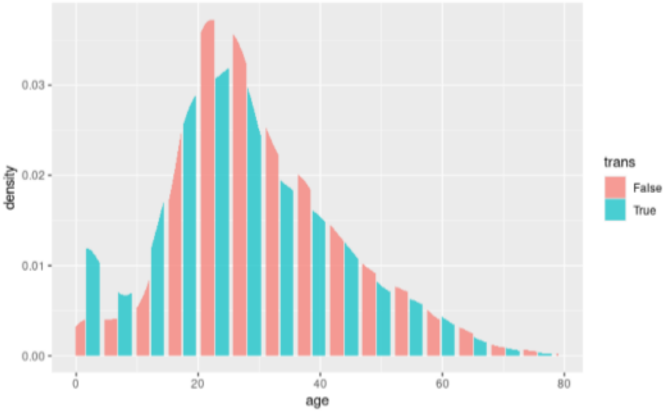


Figure 10. Age and Transportation

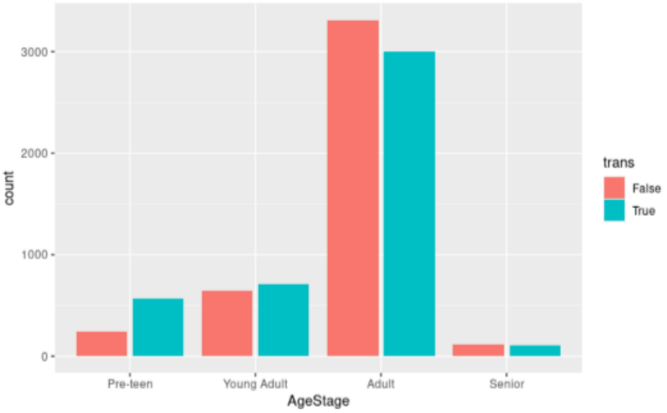


Figure 11. Stage of Life and Transportation

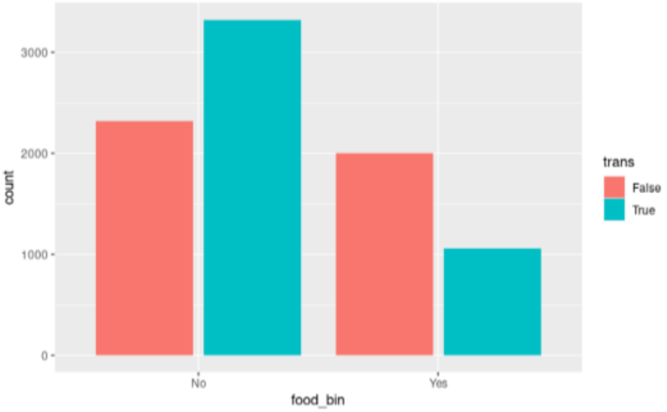


Figure 12. Food and Transportation

transported than those who had spent money.

Looking at the relationships that exist, some columns, the passenger name and passenger ID, can be removed from the training and testing files. Additionally, transported and age stage need to be factor variables for this problem.

3. Algorithm and Methodology

The algorithm used was logistic regression. Logistic regression is often used for binary classification problems, i.e., classification problems where the label only has two possible values. The algorithm is a supervised type of machine learning, meaning that it contains labels for the data (true or false for transportation for this problem). Whereas linear regression uses a fitted straight line to make predictions, logistic regression has a curved 'S' shape curve to it. Logistic regression will calculate the probability of someone being transported or not based on the remaining data in the training file.

A model can visually look great, but to ensure that the model is actually great, I will calculate the R^2 value using one of several R^2 methods for logistic regression. I will then calculate the p-value to ensure the R^2 value is not due to chance.

In order to create the logistic regression model, the data needs to be preprocessed and contain no NA values, as the glm function will throw out any NA values and will result in an error in the build. Therefore, I checked that all the values in the training set had values and all the columns that were removed from the training set were also removed from the testing set.

Another model I wanted to explore was Random Forest. Random Forest uses decision trees to make predictions. The nature of decision trees is not very flexible as they work very well on the data they train on, but not for new data. But Random Forest improves decision trees by making them more flexible, and that is why I decided to look into it more. To test how placing data into bins improves accuracy, I created two Random Forest algorithms: one which uses the data provided and the other one using the binned data I created from the preprocessing step. The Random Forest library in R helped create these models quite easily. Random forest uses splits and the variables used per split varies on how many variables you are using in the model, usually it is the square root of how many variables you are using in the model.

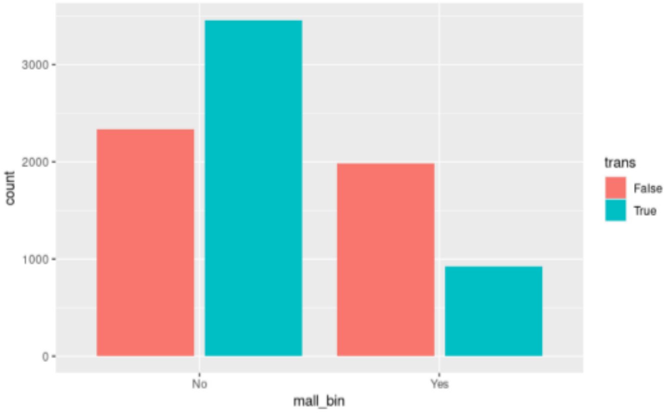


Figure 13. Shopping Mall and Transportation

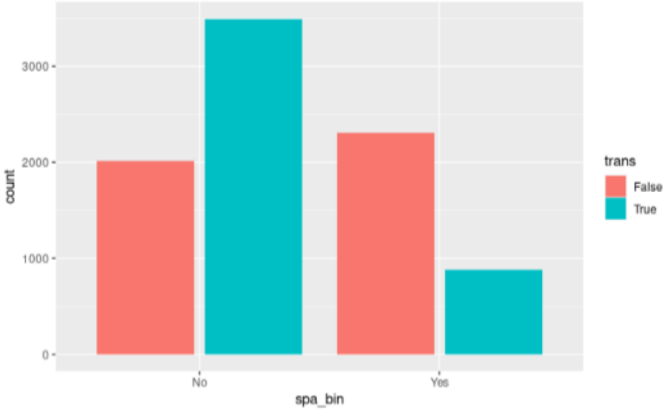


Figure 14. Spa and Transportation

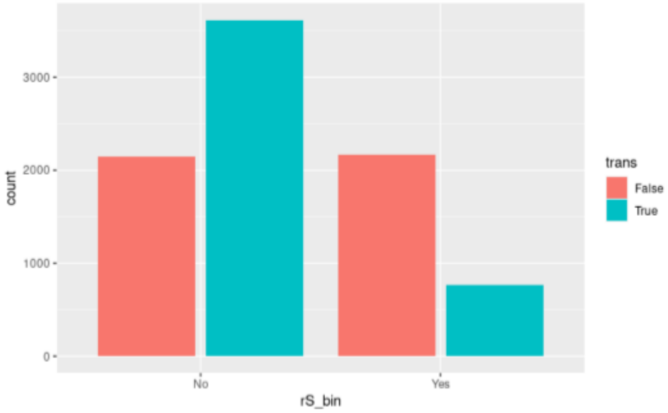


Figure 15. Room service and Transportation

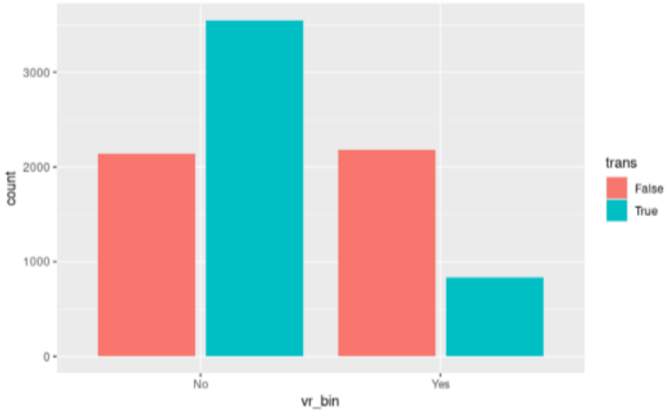


Figure 16. VR Deck and Transportation

4. Experiments and Results

At first, I kept age as a variable to influence the model, however, that resulted in many errors as it was the only continuous variable, so I removed it.

I also had to experiment to realize that all of the variables needed to be converted to factors, rather than staying as ints, chars, etc. This is because the data I kept in the end was all categorical and needed to be reflected as such in the model which I will now elaborate on.

R has a built-in function called glm that allows a user to create a logistic regression model. After cleaning up the data, I used all of the variables together to create a model, which can be seen in the figure below. Upon creating the model, p-values were created and the p-values with scores less than 0.05 were marked as significant. In the model, the following were marked as significant: the passenger was in cryosleep, the passenger was a VIP, the passenger was residing in one of cabins A-G, the cabin number the person was staying in where 1-6 were more significant than the others, the passenger was on the Port side of the ship, the age of the passenger, and whether the person had spent money or not on amenities.

Looking at the deviance residuals, which tells us if each of the observations fits the model well, they are pretty good because they are symmetrical around the median. The Akaike Information Criterion(AIC) is 9114, and the lower the number the better. I don't have another number to compare it to at the time, but I think this number is at least decent. The null deviance and residual deviance have a bit of a difference between them which is a good indication.

The method I chose to use to calculate the R^2 value was McFadden's Pseudo R^2. The value I calculated was 0.2497411. While this number is low, it does indicate that there is some sort of pattern in this huge dataset. The p-value was so tiny that we can conclude that the R^2 value was due to chance.

After making the model, I created a confusion matrix to see how the model performed on the training data. The resulting accuracy rate is 75.03%.

Moving on to the Random Forest models, I will first list the variables I used in the first model: home planet, cryosleep, home-destination planet, cabin number, cabin deck level, the side of the ship the cabin was on, age, and the five amenities. The first model that contains mostly non-binned data and it took me a lot of experimenting to find which variables, excluding binned data, created a low error rate. This model produced an OOB (out-of-bag estimate) mis-classification rate of 20.12%Ṫhis model also produced a confusion matrix, which determines the proportions of predicted labels given the labels you have. The false positive rate was 21.25%,

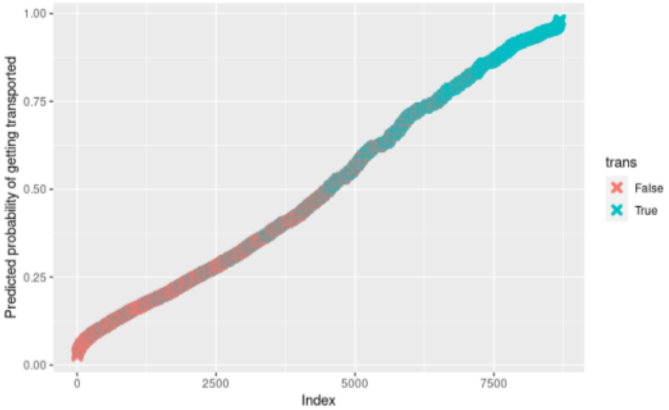
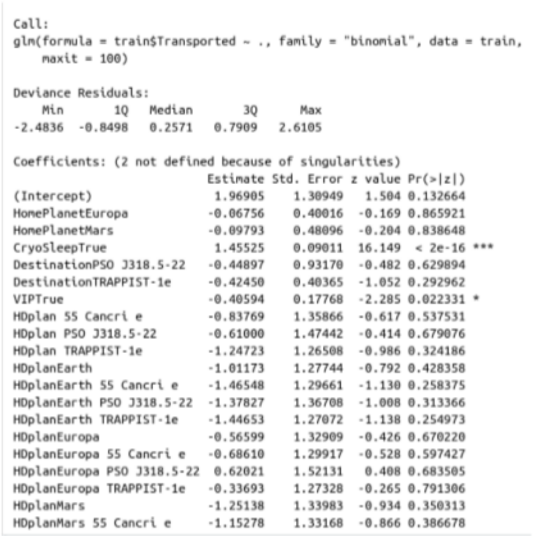


Figure 17. Logistic Regression Model



```
Call:
glm(formula = train$Transported ~ ., family = "binomial", data = train,
    maxit = 100)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4836  -0.8498   0.2571   0.7909   2.6105

Coefficients: (2 not defined because of singularities)
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                  1.96905    1.30949   1.504 0.132664
HomePlanetEuropa            -0.06756    0.40016  -0.169 0.865921
HomePlanetMars              -0.09793    0.48096  -0.204 0.838648
CryoSleepTrue                1.45525    0.09011  16.149  < 2e-16 ***
DestinationPSO J318.5-22    -0.44897    0.93170  -0.482 0.629894
DestinationTRAPPIST-1e      -0.42450    0.40365  -1.052 0.292962
VIPTrue                     -0.40594    0.17768  -2.285 0.022331 *
HDplan 55 Cancri e          -0.83769    1.35866  -0.617 0.537531
HDplan PSO J318.5-22        -0.61000    1.47442  -0.414 0.679076
HDplan TRAPPIST-1e          -1.24723    1.26508  -0.986 0.324186
HDplanEarth                 -1.01173    1.27744  -0.792 0.428358
HDplanEarth 55 Cancri e     -1.46548    1.29661  -1.130 0.258375
HDplanEarth PSO J318.5-22   -1.37827    1.36708  -1.008 0.313366
HDplanEarth TRAPPIST-1e     -1.44653    1.27072  -1.138 0.254973
HDplanEuropa                -0.56599    1.32909  -0.426 0.670220
HDplanEuropa 55 Cancri e    -0.68610    1.29917  -0.528 0.597427
HDplanEuropa PSO J318.5-22   0.62021    1.52131   0.408 0.683505
HDplanEuropa TRAPPIST-1e    -0.33693    1.27328  -0.265 0.791306
HDplanMars                  -1.25138    1.33983  -0.934 0.350313
HDplanMars 55 Cancri e      -1.15278    1.33168  -0.866 0.386678
```

Figure 18. glm function results

which means that nearly 21.25% of the true negative cases were identified wrong. The false negative rate was 18.50% so 18.50% of the true positive cases were identified incorrectly. The variable importance, i.e., a variable that appears in more splits than others, indicates how significant it is to the desired output, which would be if someone was transported or not. The order of variable importance was: cryosleep, spa, room service, vr deck, food court, shopping mall, home-destination planet, cabin deck, age, cabin number, home planet, and the side of the ship one stayed on. However, we know age should be a bit more significant after it was put into bins so perhaps this model is not the greatest one to look at. But this output is something to take not as heavily as the confusion matrix because confounding variables and etc. This Random Forest model produced a 79.88% accuracy which is kind of bizarre. The second Random Forest model will use the bins for the amenities, as well as the age. This model did much worse than I thought it would, with an accuracy rate of 75.39%. While the false positive rate decreased by a little over 2%, the false negative rate nearly doubled.

I also experimented with various column values to see if I could beat the 18.50%. I replaced variables and also added them one at a time to see which ones may have been confounding and had not improved the model. For instance, looking at the binned mall with the binned food data reduced the error rate by 0.01%.

5. Summary and Conclusions

Based on the three different tests, the random forest models outperformed the logistic regression model. What was very interesting to me was that the bins that were created from the data did not really seem to improve the accuracy of the model. I thought that based off of the graphs that binning the data would be important and there were definite trends. Perhaps there was another key variable I was missing that caused the bins to work as well as I thought. Or perhaps the testing data is vastly different and the trends we saw in the training data is much different from the testing data. An afterthought I had shortly before turning this assignment in was that the group number may have an effect on the rate, but that is something I could test at a later point. I also hope to come back and experiment with different variables for the logistic regression model to improve its accuracy.

Acknowledgments

I would like to thank all the YouTube videos out there that exist on logistic regression. I would also like to thank Professor Hasan Kurban. I know I may not have been in class very often, but I always had a fun time learning from him. And thank you to the TAs and being by far the most consistent graders I've ever had. You've kept my stress levels low because I always knew what to expect.

References

https://v8doc.sas.com/sashtml/insight/chap39/sect55.htm
https://www.youtube.com/watch?v=C4N3$_{X}JJ - jU$
https://www.listendata.com/2015/06/r-keep-drop-columns-from-data-frame.html
https://r-coder.com/cut-r/
https://r-graph-gallery.com/21-distribution-plot-using-ggplot2.html
https://statisticsglobe.com/r-extract-first-or-last-n-characters-from-string
https://www.datanovia.com/en/blog/ggplot-axis-ticks-set-and-rotate-text-labels/
https://www.marsja.se/how-to-concatenate-two-columns-or-more-in-r-stringr-tidyr/
https://www.r-bloggers.com/2015/09/how-to-perform-a-logistic-regression-in-r/ https://www.youtube.com/watch?v=1iAf8AOw7wA