Spaceship Titanic Analysis

Liam Glennie

March 17, 2023

Abstract

This paper presents an analysis of the Spaceship Titanic dataset using the data mining tool RapidMiner. The dataset was part of a Kaggle competition, and the analysis includes preprocessing, descriptive analysis, and modelling techniques. The conclusion highlights the strengths and limitations of RapidMiner as a tool for data mining and presents findings regarding the importance of exploring variables to improve the accuracy and reliability of predictive models.

1 Introduction

This paper analyses the Spaceship Titanic dataset [1]. The dataset is part of one of many competitions hosted by Kaggle. The analysis will be carried out using the data mining tool RapidMiner [2].

This paper contains a brief description of the problem and dataset at hand, the preprocessing steps applied, descriptive analysis in the form of univariate and bivariate plots, and modelling techniques. Finally, a conclusion is drawn based on RapidMiner and the findings discovered throughout the analysis.

2 Problem Context

The dataset contains information on passengers of the fictitious 2912 *Spaceship Titanic*. The *Spaceship Titanic* is an interstellar vessel that transports people from our galaxy to three habitable exoplanets orbiting nearby stars.

However, tragedy struck when rounding Alpha Centauri on its way to the first destination, 55 Cancri E. The *Spaceship Titanic* collided with a spacetime anomaly, causing almost half of its passengers to be transported to an alternate dimension.

Therefore, the problem at hand is to classify which passengers were transported to another dimension, given a set of attributes.

3 Dataset

The complete dataset contains 17386 personal records of passengers of the *Spaceship Titanic*. The attributes are:

- **PassengerID:** A unique ID for each passenger.
- **HomePlanet:** The planet the passenger departed from.
- **CryoSleep:** Indicates if the passenger chose to be put to sleep for the duration of the voyage. Passengers in cryo-sleep are confined to their cabins.

- **Cabin:** Cabin number where the passenger is staying. Follows the pattern *deck/num/side*, where side can be either P for Port or S for Starboard.
- **Destination:** The planet the passenger is travelling to.
- **Age:** The age of the passenger.
- **VIP:** Whether the passenger has paid for special VIP service.
- **RoomService, FoodCourt, ShoppingMall, Spa, VRDeck:** Amount the passenger was billed at each of the spaceship's luxury amenities.
- **Name:** First and last names of the passenger.
- **Transported:** Whether the passenger was transported or not. **The target variable**.

4 RapidMiner

RapidMiner is a popular data mining tool that allows users to perform data analysis tasks, such as data cleaning, preprocessing, visualisation and modelling without writing any code.

A secondary goal of this paper is to assess the ease of use, diversity and general performance compared to the popular programming language amongst data scientists, Python and R.

5 Preprocessing

RapidMiner offers an interesting feature called *TurboPrep* which allows you to quickly perform typical preprocessing techniques such as removing duplicates and missing values, transforming variables and creating new ones. Once you select all the operations you want to perform, RapidMiner translates them into a process map.

5.1 Duplicates

No duplicates were detected by checking the **PassengerID** column for repeated values. Obviously, no rows were removed.

5.2 Outliers

Only numerical variables were analysed for outliers. First of all, age seemed to show an unusual number of passengers at age zero, as shown in Figure 1. These cases were analysed, and it was assumed that zero was used to indicate that the age of that passenger was missing because it was unlikely that there would be so many recently born children boarding the spaceship. Additionally, all the cases with age zero also had a zero value for their expenses on luxury amenities. Thus, it is likely that these rows were missing some data.

Interestingly, most cases with missing information seemed to be transported more often than not. This can be seen in Figure 2. Having obtained this insight, it is likely that **Age** be discretised in the section of feature engineering 6.

The rest of the numerical variables, those related to passengers' expenses on luxury amenities, seemed to show a presence of outliers. Most of the passengers did not spend much money on luxury amenities. On the other hand, some specific passengers spent over one thousand monetary units. Figure 3 shows the distribution that all the luxury amenity expenses follow.

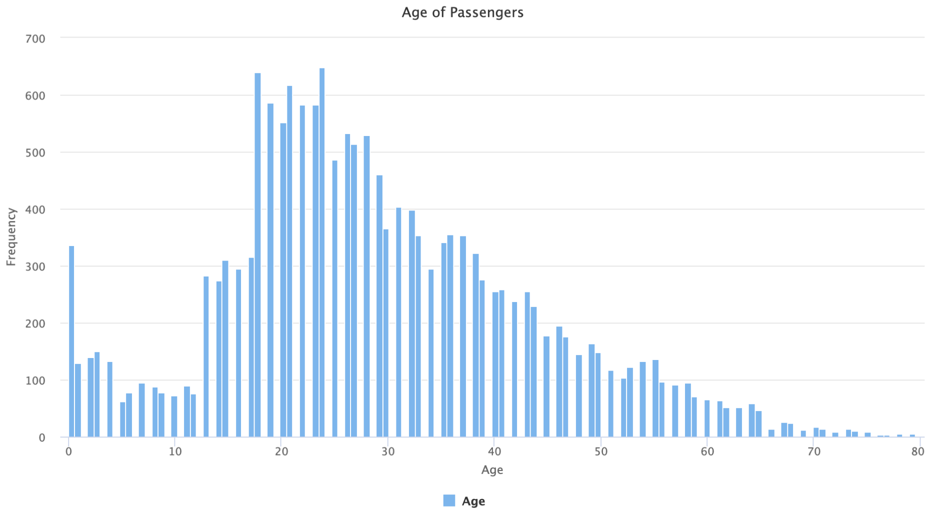


Figure 1: Age Histogram

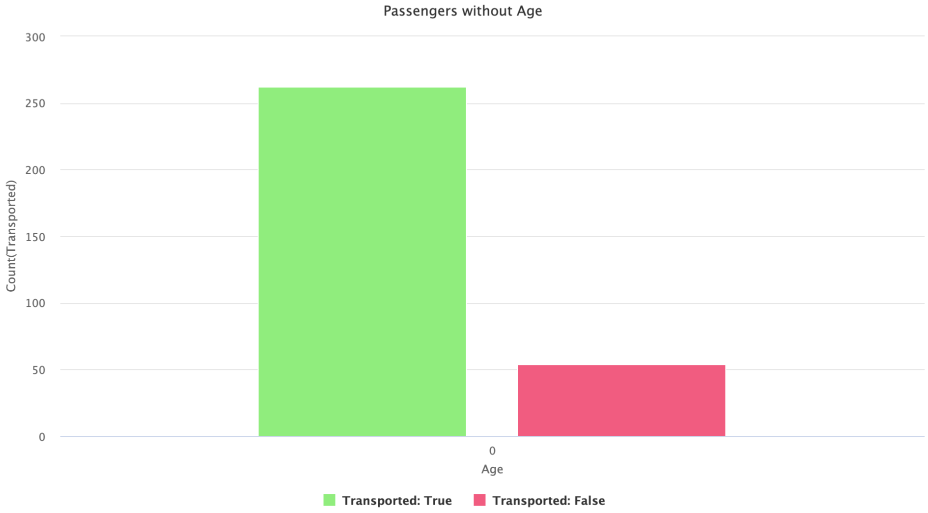


Figure 2: Missing Age Transported Barplot

Despite these outlier values being outside of 1.5*IQR and representing less than 10% of the population, these rows have not been removed. The Feature Engineering section 6 explains how these outliers were dealt with.

Finally, no multivariate outliers were able to be detected as RapidMiner requested more memory than the amount the machine had available.

5.3 Missing Values

An essential step in every data analysis is detecting how clean or complete the dataset at hand really is. Ideally, a dataset would not have any missing values or NAs. Unfortunately, that is not always the case.

RapidMiner offers a quick insight into the percentage of NAs per column, and as can be observed in Figure 4, there were around 2% of missing values per attribute. Of course, if all the missing values for attributes coincided in the same rows, it would mean that only 2% of the dataset would be missing,

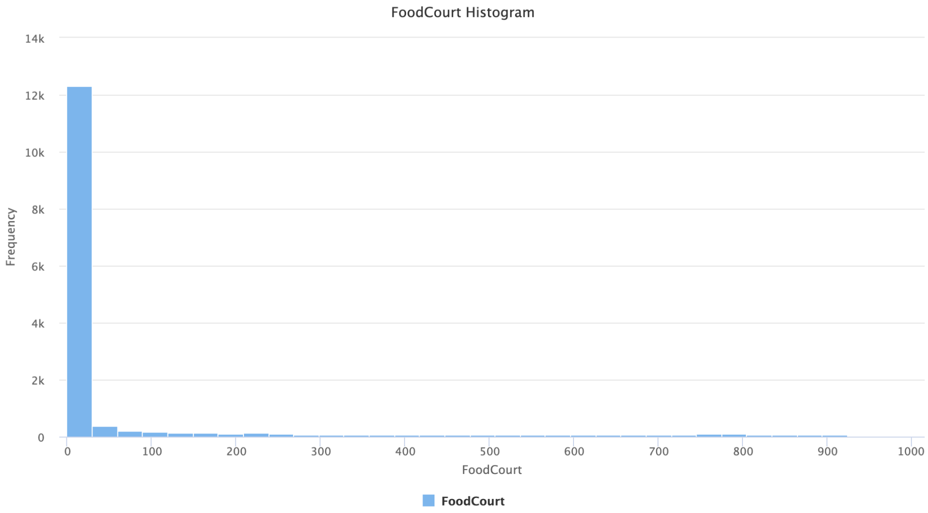
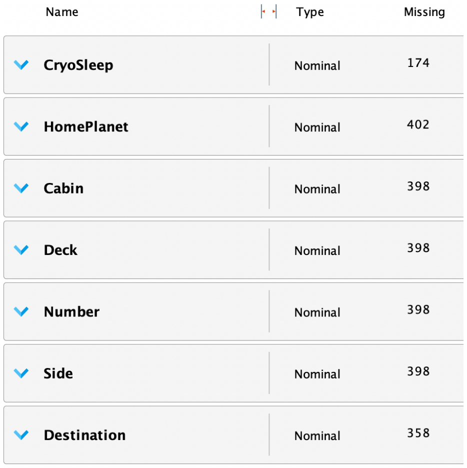


Figure 3: FoodCourt Barplot X-axis limited to 1000



Name	Type	Missing
CryoSleep	Nominal	174
HomePlanet	Nominal	402
Cabin	Nominal	398
Deck	Nominal	398
Number	Nominal	398
Side	Nominal	398
Destination	Nominal	358

Figure 4: Missing Values in RapidMiner

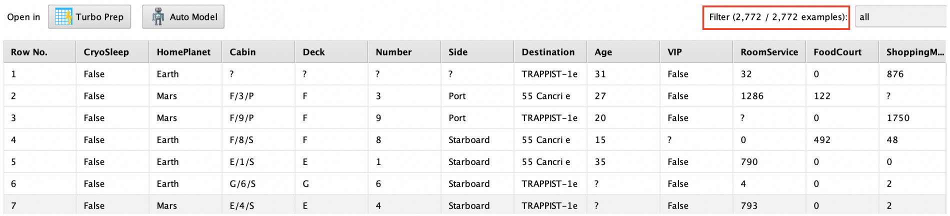
and these instances could have been removed without harming the analysis. However, this was not the case. Figure 5 shows that over 2700 rows contained NA values. This is a relatively large number of rows to remove without attempting to impute some data.

Crucially, there were no NAs for the target variable, **Transported**, meaning there may still be some hope of saving the rows with missing data.

5.3.1 Imputation

Imputation is a common data preprocessing technique in statistical analysis and machine learning. The goal of imputation is to fill in missing values in a dataset with plausible estimates to extract the most information as possible from the dataset.

Luxury Amenities Passengers in **CryoSleep** were confined to sleep in their cabin for the duration of their trip. Thus, any passenger that is in **CryoSleep** cannot spend any money on luxury amenities. So, the values of the attributes **RoomService, FoodCourt, ShoppingMall, Spa and VRDeck** were all set to zero for passengers with **CryoSleep** set to True. The opposite has also been considered.



Filter (2,772 / 2,772 examples): all

Row No.	CryoSleep	HomePlanet	Cabin	Deck	Number	Side	Destination	Age	VIP	RoomService	FoodCourt	ShoppingM...
1	False	Earth	?	?	?	?	TRAPPIST-1e	31	False	32	0	876
2	False	Mars	F/3/P	F	3	Port	55 Cancri e	27	False	1286	122	?
3	False	Mars	F/9/P	F	9	Port	TRAPPIST-1e	20	False	?	0	1750
4	False	Earth	F/8/S	F	8	Starboard	55 Cancri e	15	?	0	492	48
5	False	Earth	E/1/S	E	1	Starboard	55 Cancri e	35	False	790	0	0
6	False	Earth	G/6/S	G	6	Starboard	TRAPPIST-1e	?	False	4	0	2
7	False	Mars	E/4/S	E	4	Starboard	TRAPPIST-1e	?	False	793	0	2

Figure 5: Total number of rows with at least one NA

If the passenger had some expense, they must not be in **CryoSleep** and this value was set to False.

VIP Figure 6 shows that not a single passenger coming from Earth paid for the VIP experience on the *Spaceship Titanic*. To follow the pattern, all passengers from Earth with a missing value for the attribute **VIP**, were imputed to False.

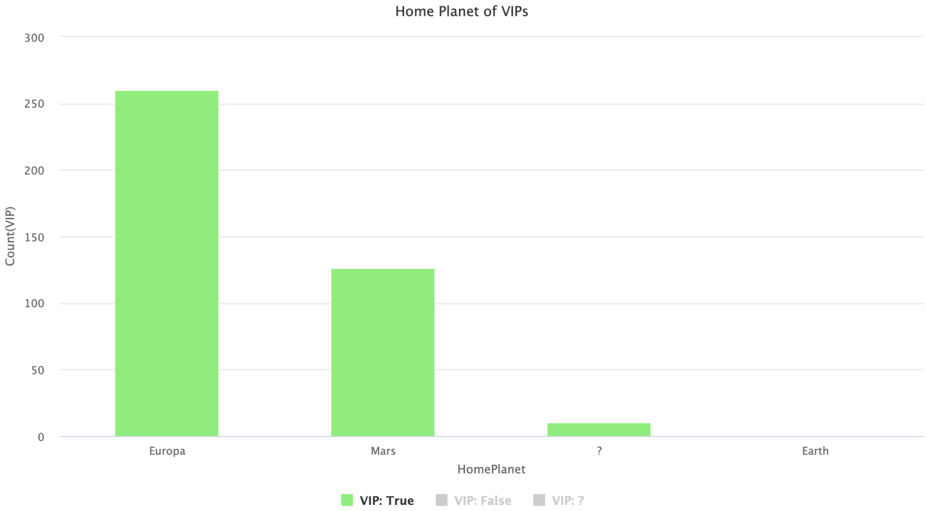


Figure 6: Home Planet of VIP passengers

There were no VIP passengers under the age of 18 as is portrayed by Figure 7. Therefore, the **VIP** attribute of these passengers was also imputed to False.

Rest of Attributes For the attributes like **Age, Cabin, HomePlanet and Destination** there did not seem to be an obvious pattern that could have been used to impute the missing values. Therefore, rows containing missing values for these attributes were been removed.

6 Feature Engineering

Feature engineering refers to the process of selecting, extracting, transforming and creating new features from raw data to improve the performance of machine learning models.

To begin with, three new variables were extracted from the attribute **Cabin**. As the values of **Cabin** are structured in the following manner: *Deck/Number/Side*, the new variables were created from splitting by "/". This operation was done to see if the location of a passenger on the spaceship had a relationship with them being transported or not.

New variables:

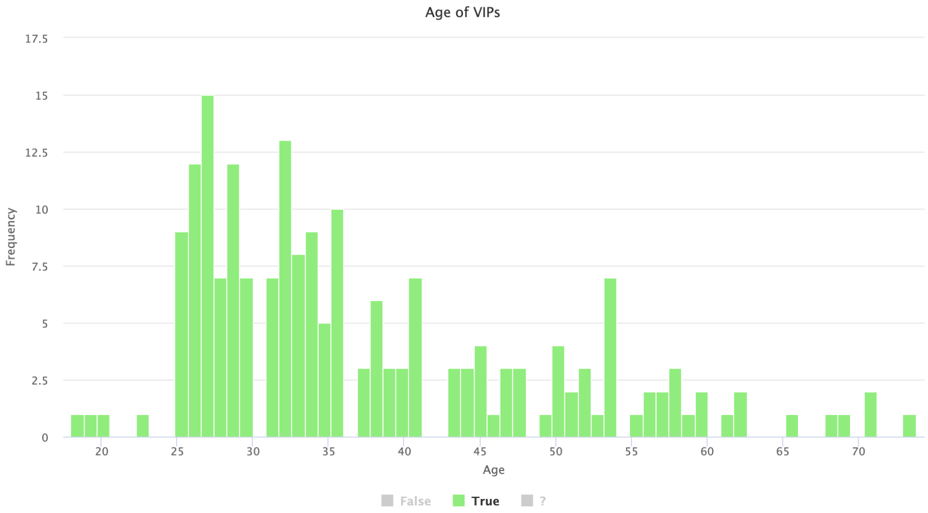


Figure 7: Age of VIP passengers

- **Deck**: The deck of the passenger's cabin was on the spaceship.
- **Number**: The number of the passenger's cabin on the spaceship.
- **Side**: The side of the passenger's cabin was on the spaceship.

Figure 8 showcases the relationship between the passenger being transported or not and the deck and side of their cabin on the spaceship. For example, passengers on the starboard side were slightly more likely to be transported than those on the port side. Or if a passenger's cabin was on deck B or C, it was more likely that they were transported than not.

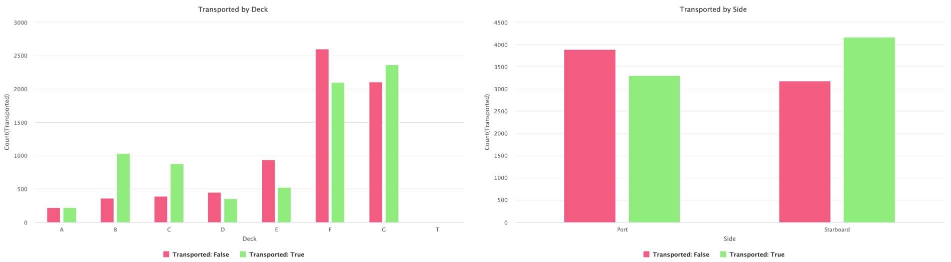


Figure 8: Transported by Deck and Side

Next, as mentioned previously, the luxury amenity expenses variables caused some problems due to having many outliers and heavily skewed data. So, to mitigate these problems new variables were created. Essentially, these new variables represent if a passenger spent any money on each of the luxury amenities. These variables were also created to better understand the data and the effect they may have on the models created in a future section, 8.

New variables:

- **hasRoomService:** Has the passenger spent any money on room service?
- **hasFoodCourt:** Has the passenger spent any money in the food court?
- **hasShoppingMall:** Has the passenger spent any money in the shopping mall?
- **hasSpa:** Has the passenger spent any money in the Spa?

- **hasVRDeck:** Has the passenger spent any money on the VR deck?
- **hasExpenses:** Has the passenger spent any money onboard?

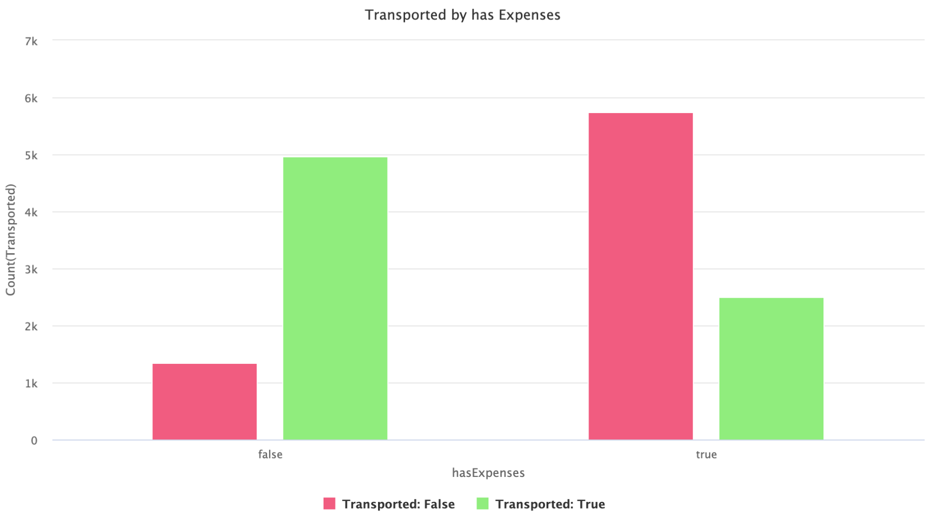


Figure 9: Transported by hasExpenses

Figure 9 displays that passengers without any expenses were more likely to be transported than those that spent at least one monetary unit onboard.

Finally, the attribute **Age** was discretised into seven different categories. Figure 10 shows the distribution of the age groups chosen. The main idea behind the discretisation of **Age** was to differentiate between passengers with a known age from those without one and to identify if there were age groups that were more likely to be transported. Next, the rest of the bins (1-10, 11-18, 19-25, 26-35, 46-55, 55+) were chosen based on common age divisions while attempting to keep the categories somewhat balanced.

It is interesting to observe that passengers unknown or under 18 years old seemed to be more likely to be transported than the rest of the categories.

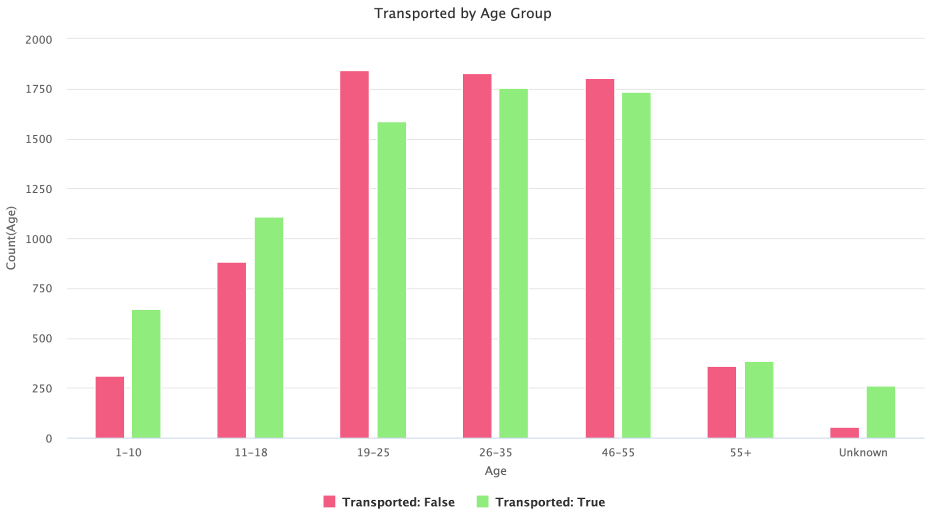


Figure 10: Transported by Age Group

7 Bivariate Analysis

Throughout the paper, some plots have already hinted at possible relationships between a passenger being transported or not based on if they had any expenses onboard the spaceship, their age group, the deck and side their cabin was located on the space shuttle. Thus, this section focuses on further identifying possible relationships between the explanatory attributes and the target, **Transported**.

7.1 Transported and HomePlanet

As can be seen in Figure 11, there was very little difference between the odds of being transported for passengers from Mars. On the other hand, Europa passengers had higher odds of being transported than not. Finally, Earth passengers were slightly less likely to be transported.

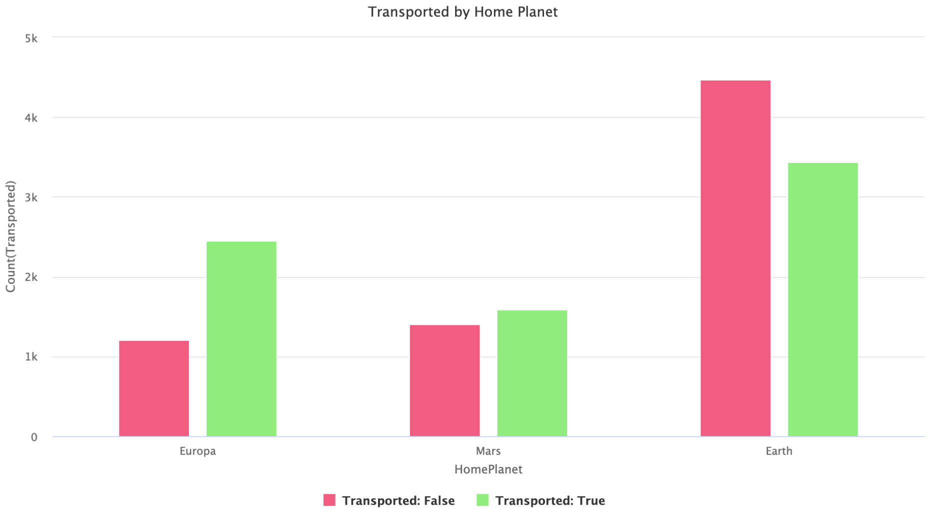


Figure 11: Transported by HomePlanet

7.2 Transported and Destination

When it comes to being transported based on the destination of a passenger, Figure 12 shows that *55 Cancri E* was the only destination planet where passengers were more likely to be transported than not. This could have occurred because the *Spaceship Titanic* collided with the spacetime anomaly when it was approaching its first destination, *55 Cancri E*. Therefore, passengers may have been getting ready to disembark in a specific location of the spaceship, causing them to be transported.

7.3 Transported, Cabin and CryoSleep

First of all, Figure 13 displays, on the left-hand side, cabins associated with passengers that were not transported. And as can be seen, there is a group of specific cabins toward the left-center of the plot that are associated with passengers that were transported.

Now, the right-hand side plot shows that the same group of cabins are associated with passengers that have been put into cryo-sleep. Therefore, the passengers of that group of cabins were passengers that were in cryo-sleep and were transported. Figure 14 strengthens the idea that **Transported** and **CryoSleep** seem to have some sort of relationship because it showcases that passengers in cryo-sleep were more likely to be transported than not.

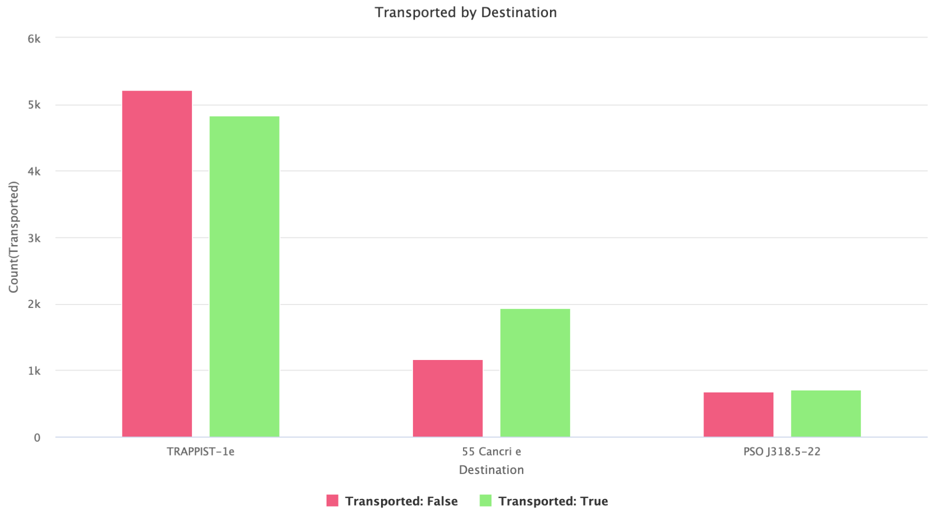


Figure 12: Transported by Destination

Consequently, it is likely that **Cabin** and **CryoSleep** will be strong predictors when it comes to modelling.

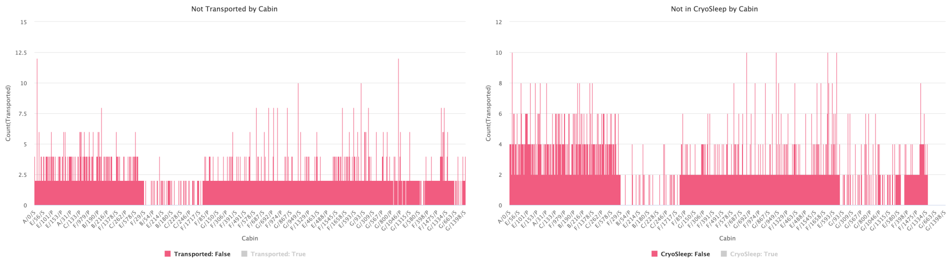


Figure 13: Not Transported Passenger cabins (Left) Not in CryoSleep passenger cabins (Right)

8 Modelling

This section describes the iterative process of finding the logistic regression model that best fits the data.

Logistic regression [3] is a statistical model used to predict binary outcomes based on one or more predictor variables. It is a type of generalized linear model that uses the logistic function to estimate the probability of the binary outcome.

The logistic function (also called the Sigmoid function) is defined as:

$$p(x) = \frac{1}{1 + e^{-z}}$$

where:

- $p(x)$ is the probability of the binary outcome for a given set of predictor variables x
- z is a linear combination of the predictor variables and their corresponding coefficients:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

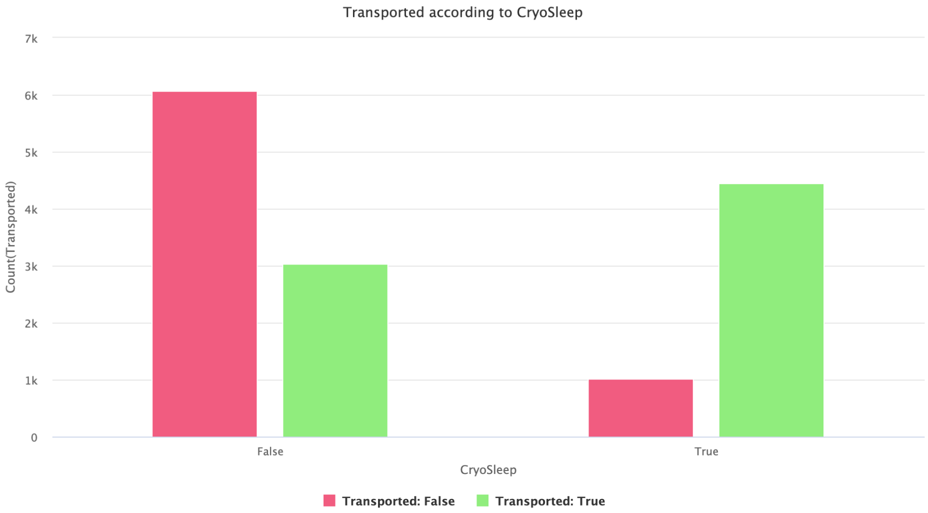


Figure 14: Transported according to CryoSleep

- $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the coefficients (also called weights or parameters) of the logistic regression model.

The logistic regression model estimates the coefficients of the logistic function using maximum likelihood estimation, which involves finding the values of the coefficients that maximize the likelihood of the observed binary outcomes given the predictor variables. The model can then be used to predict the probability of the binary outcome for new observations based on their predictor variables.

8.1 First Models

First, two simple models were developed. The first has as explanatory variables the **Deck, Side and Number** of passenger cabins. The second model only uses the passenger's **Cabin** as a predictor. These first two models were used to assess what would be the better predictor. The exact cabin location on the ship or this location decomposed into deck, side and number.

The model using the **Deck, Side and Number** obtained an accuracy of 65.14% without bias towards any class, Table 1. On the other hand, the **Cabin** model obtained an accuracy of 78.07%, however, it did show some bias toward the false class. Table 2. It must be noted that the metrics were extracted using test data, that is, data not used to train the model.

	true False	**true True**	**class precision**
pred. False	1492	829	64.28%
pred. True	915	1712	65.17%
class recall	61.99%	67.38%	

Table 1: Performance Deck, Side and Number model

	true False	**true True**	**class precision**
pred. False	2113	715	74.72%
pred. True	294	1826	86.13%
class recall	87.79%	71.86%	

Table 2: Performance Cabin model

Despite the bias, the **Cabin** variable was kept for future models.

8.2 Numeric Age vs Binned Age Model

Next, two additional models were developed to distinguish whether there would be an increase in performance when it came to the format of **Age**. Therefore, the first model used the numeric version of **Age** as a predictor, whereas the second model used the binned version of **Age**. Both models used **Cabin**.

There was no significant difference in the performance of the two models with both achieving an accuracy of around 78%. The numeric version of **Age** was kept due to it having less class bias as can be seen in Table 3 and 4

	true False	**true True**	**class precision**
pred. False	1914	571	77.02%
pred. True	493	1970	79.98%
class recall	79.52%	77.53%	

Table 3: Performance Binned Age model

	true False	**true True**	**class precision**
pred. False	1880	555	77.21%
pred. True	527	1986	79.03%
class recall	78.11%	78.16%	

Table 4: Performance Numeric Age model

8.3 Adding CryoSleep

As mentioned earlier, in Figure 13 it seemed that **CryoSleep** shared a slight relationship with **Transported**, that is, passengers in cryo-sleep also happened to be passengers that were transported. Therefore, **CryoSleep** was added. The performance of the model improved considerably to 86.4% while the slight bias to the false class still remains present, Table 5.

	true False	**true True**	**class precision**
pred. False	1880	555	77.21%
pred. True	527	1986	79.03%
class recall	78.11%	78.16%	

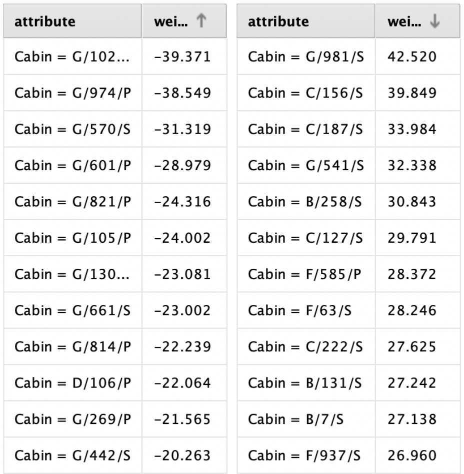
Table 5: Performance Cabin, Numeric Age and CryoSleep model

Additional models were created using attributes like **HomePlanet, Destination** and **VIP**, however, no improvement was shown compared to the model using **Age, Cabin** and **CryoSleep**.

This model could have been considered as the final model, however, Figure 15 clearly shows that the value of **Cabin** is the biggest indicator of the prediction this final model will produce. This could be a concerning problem. Imagine that new records are found from the *Spaceship Titanic* with passengers from a new section of the space shuttle that was not included in the original data. That is, this new data could contain cabins that the model has never seen before, meaning that the **Cabin** variable would likely be rendered useless as a predictor. Thus this model could not be considered the final model.

8.4 Final Model

A similar iterative process was followed to obtain the final model. By using insight gained throughout the analysis, several combinations of predictors were used to obtain a definitive product. The final model used **AgeBinned, CryoSleep** and **Destination** as their predictors. Figure 16 illustrates that **CryoSleep:True** variable has the most influence on the outcome of the model along with

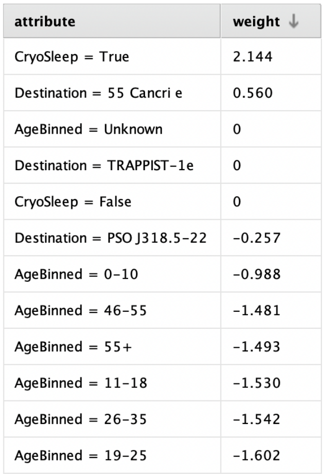


attribute	wei… ↑
Cabin = G/102…	-39.371
Cabin = G/974/P	-38.549
Cabin = G/570/S	-31.319
Cabin = G/601/P	-28.979
Cabin = G/821/P	-24.316
Cabin = G/105/P	-24.002
Cabin = G/130…	-23.081
Cabin = G/661/S	-23.002
Cabin = G/814/P	-22.239
Cabin = D/106/P	-22.064
Cabin = G/269/P	-21.565
Cabin = G/442/S	-20.263

attribute	wei… ↓
Cabin = G/981/S	42.520
Cabin = C/156/S	39.849
Cabin = C/187/S	33.984
Cabin = G/541/S	32.338
Cabin = B/258/S	30.843
Cabin = C/127/S	29.791
Cabin = F/585/P	28.372
Cabin = F/63/S	28.246
Cabin = C/222/S	27.625
Cabin = B/131/S	27.242
Cabin = B/7/S	27.138
Cabin = F/937/S	26.960

Figure 15: Cabin, Numeric Age and CryoSleep weights

AgeBinned:19-25 and **AgeBinned:26-35**. Essentially, this means that if the passenger was in cryo-sleep, their odds of being transported increased but if they belonged to the age groups 19-25 or 26-35, their odds of being transported decreased.



attribute	weight ↓
CryoSleep = True	2.144
Destination = 55 Cancri e	0.560
AgeBinned = Unknown	0
Destination = TRAPPIST-1e	0
CryoSleep = False	0
Destination = PSO J318.5-22	-0.257
AgeBinned = 0-10	-0.988
AgeBinned = 46-55	-1.481
AgeBinned = 55+	-1.493
AgeBinned = 11-18	-1.530
AgeBinned = 26-35	-1.542
AgeBinned = 19-25	-1.602

Figure 16: Final Model weights

Finally, the performance of the final model dropped considerably when comparing it to the previous model including the **Cabin** variable, achieving an accuracy of 73.95%. Also, this model shows a considerable amount of bias toward the false class, Table 6. While it is not an ideal solution, it is better than producing a model that is overfitted or biased to the training set.

	true False	**true True**	**class precision**
pred. False	2059	941	68.63%
pred. True	348	1600	82.14%
class recall	85.54%	62.97%	

Table 6: Performance Final model

9 Conclusion

9.1 RapidMiner

RapidMiner is a valuable tool for quickly gaining insights into data and is suitable for beginners or non-specialised individuals. However, its limitations in terms of visualisations can be a drawback, as the plot section is not persistent and requires exporting after every plot is created. RapidMiner's pre-optimised code allows for fast execution of processes, though development time may be prolonged compared to programming languages such as Python and R. Overall, RapidMiner is a useful option for those looking to perform basic data mining tasks efficiently and effectively.

9.2 Analysis Conclusion

In conclusion, the analysis of the *Spaceship Titanic* data showed that the **Cabin** variable was strongly associated with the target variable, **Transported**, indicating that bias may have been introduced in the data during its artificial creation. Although a highly accurate model was initially developed using the **Cabin** variable, it was found to be overfitted, and a less-performing model was ultimately selected. This model used **AgeBinned**, **CryoSleep** and **Destination** as its predictors, which were likely to be more robust in predicting new unseen data and did not rely on the **Cabin** variable. These findings emphasize the importance of exploring variables to improve the accuracy and reliability of predictive models, especially in datasets where bias may be present.

References

[1] Kaggle. Spaceship titanic. https://www.kaggle.com/competitions/spaceship-titanic. Accessed: March 16, 2023.

[2] RapidMiner. Home. https://rapidminer.com/. Accessed: March 16, 2023.

[3] Sandro Sperandei. Understanding logistic regression analysis. *Biochemia medica*, 24(1):12–18, 2014.