Coursework: EDA & Regression

Statistics and Machine Learning 1 - MSc Data Science

11549067

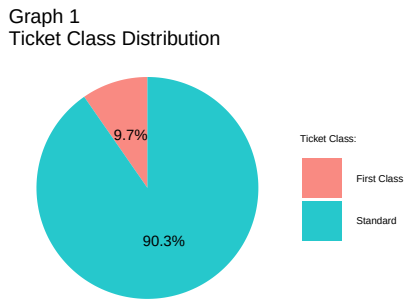
1. Brief description of the Data

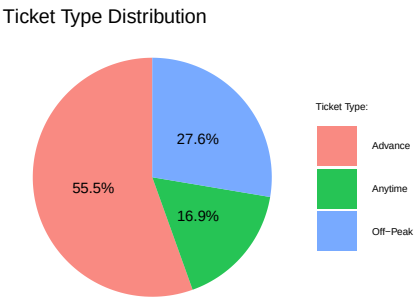
The dataset used in this report contains information on rail journeys made by passengers in the UK between 1 January and 30 April 2024. In total, the dataset includes records of 31,645 journeys with attributes like the payment method, departure and arrival stations, type of tickets, prices and dates and hours of departure and arrival.

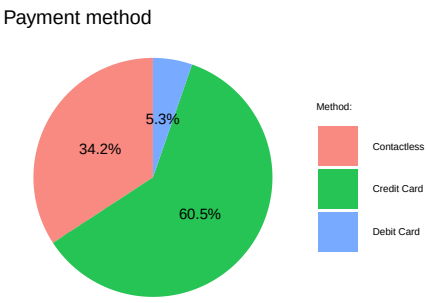
2. Exploratory data analysis

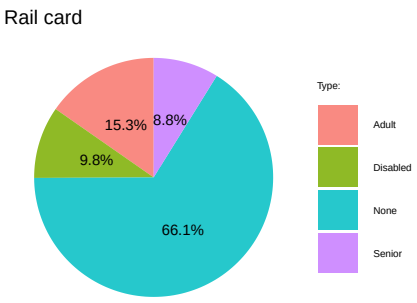
2.1. Ticket class, ticket types, rail card usage and payment methods

Roughly nine out of ten tickets were purchased in the "Standard" class. Slightly more than half of the passengers bought their tickets in advance. Six out of ten tickets were purchased using credit cards. Over 60% of passengers did not use a rail card.



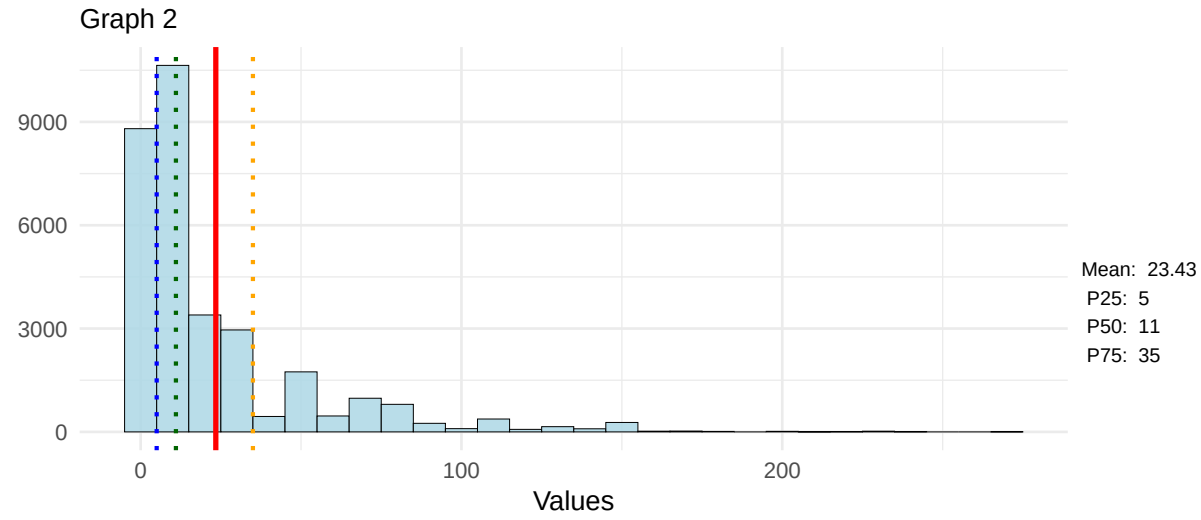






2.2. Ticket prices distribution

The average ticket price was £23. Graph 2 shows the distribution of ticket prices, revealing a concentration of tickets at the lower end of the price range. 25% of tickets cost £5 or less and at least half of the tickets were priced at £11 or less.



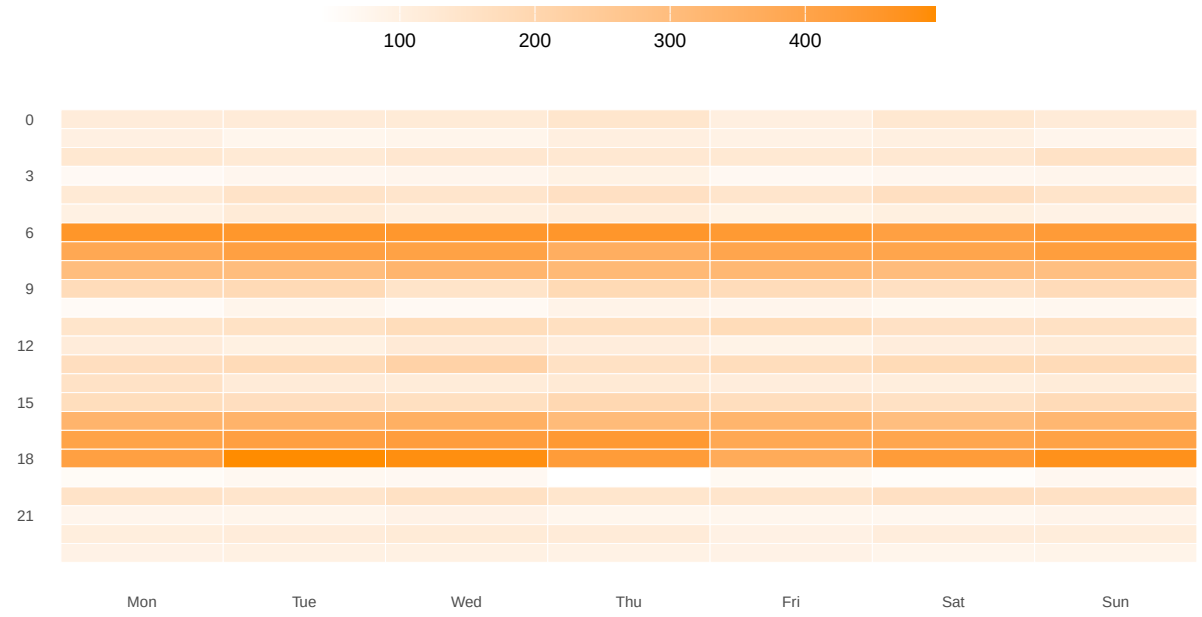
2.3. Origin and destination

10 station combinations accounted for 80% of all journeys, equivalent to 26,094 trips. Graph 3 illustrates the most common departure and arrival stations and the connections between them. For clarity, the graph only includes station combinations with 500 or more journeys.

2.4. Peak travel hours

The data also highlights clear peak travel hours. As shown in Graph 4, there is a marked increase in the number of journeys between 6:00 am and 9:00 am, and between 4:00 pm and 7:00 pm.

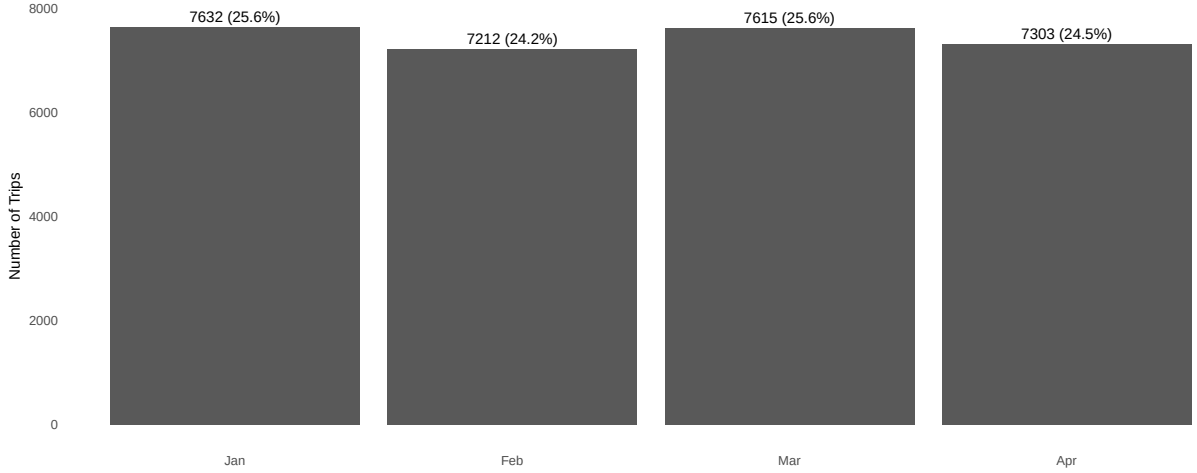
Graph 4



2.5. Journeys per month

As shown in Graph 5, the distribution of journeys is relatively consistent across the four months of data collection, with an average of 7,440 journeys per month.

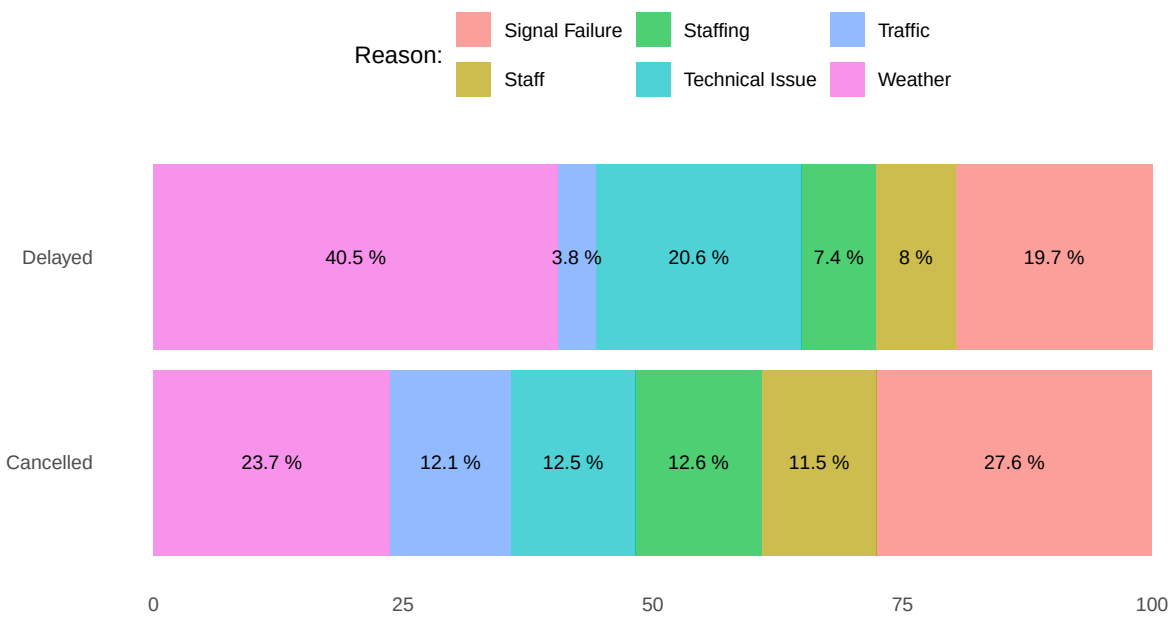
Graph 5



2.6. Delayed and cancelled journeys

Out of all recorded journeys, 87% (27,479) were on time, while 7% (2,289) were delayed, and 6% (1,877) were cancelled. Graph 6 presents the main reasons for delayed and cancelled journeys.

Graph 6



3. Column *DelayInMinutes*

We created a new variable named *DelayInMinutes*, which calculates the delay duration in minutes for all delayed journeys. If a journey was on time or cancelled, *DelayInMinutes* is assigned a value of NA.

The average delay duration is 42 minutes. Half of the passengers (with delayed journeys) experienced a delay of 37 minutes or more, while 75% of passengers (with delayed journeys) faced a delay of at least 19 minutes.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   19.00   37.00   42.23   53.00  180.00   29357
```

4. Modeling probability of refund

Now, we will focus exclusively on the journeys that were either delayed or cancelled. We will fit an appropriate regression model to predict the probability that a passenger requests a refund for their ticket. Out of the 4,166 delayed or cancelled journeys, 27% of passengers (1,114) requested a refund.

First, a variable named *MediumPrice* will be created for tickets priced above £10 and up to £30:

```
dataModel = MavenRail %>%
  filter(Journey.Status != "On Time") %>% #Journey.Status is not On Time
  mutate(MediumPrice = ifelse(Price > 10 & Price <= 30, 1, 0))
```

Among the delayed and cancelled journeys, 749 cases meet this condition:

Table 1: Binomial regression with one predictor

	Model 1
(Intercept)	−1.076***
	[−1.153, −0.999]
MediumPrice	0.354***
	[0.182, 0.524]
Num.Obs.	4166
AIC	4826.0
BIC	4838.6
Log.Lik.	−2410.983
F	16.507
RMSE	0.44

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

```
##
##    0    1
## 3417  749
```

We fit a Generalized Linear Model (GLM) from the binomial family, as the response variable is categorical with two levels ("Yes" or "No"), using the newly created variable *MediumPrice* as the sole predictor. The model can be formally expressed with the following formula:

$$\log\left[\frac{P(\text{Refund. Request} = 1)}{1 - P(\text{Refund. Request} = 1)}\right] = \alpha + \beta_1(\text{MediumPrice}) \tag{1}$$

Table 1 shows that ***MediumPrice* has a positive and statistically significant effect on the probability of requesting a refund**, with a coefficient of 0.354.

4.1. Probability given that they paid £5

A £5 ticket price lies outside the range of *MediumPrice*, so the value of this variable in the formula is equal to zero. The right side of the equation, thus, simplifies to -1.076, giving us the odds expressed as follows:

$$\frac{\text{P(Refund.Request)}}{1 - \text{P(Refund.Request)}} = e^{-1.076} = 0.341$$

Solving the equation shows that the **probability of requesting a refund when the ticket price is £5 is 25%**.

$$P(\text{Refund.Request}) = \frac{0.341}{1.341} = 0.254$$

```
## [1] 0.2543168
```

4.2. Probability given that they paid £25

A ticket price of £25 falls within the range of *MediumPrice*, so the value on the right side of the formula is −0.722. This results in the odds being expressed as follows:

$$\frac{\text{P(Refund.Request)}}{1 - \text{P(Refund.Request)}} = e^{-0.722} = 0.486$$

Solving the equation shows that the **probability of requesting a refund when the ticket price is £25 is 33%**.

$$P(\text{Refund.Request}) = \frac{0.486}{1.486} = 0.327$$

```
## [1] 0.3271024
```

5. Predicting another dataset

Using the *MavenRail* data, we will fit an appropriate regression model to determine the likelihood of passengers in the *ToPredict* file requesting a refund. As shown in Table 2, four binomial regression models with different combinations of predictors were built.

All variables maintain their original categories, except for the variable *Price*, that was converted from a numerical variable into a factor variable with five levels for easier interpretation.

Table 2: Regression Models and Predictors

Model	**Predictors**
First model	Price
Second model	Price, Railcard
Third model	Price, Railcard, Ticket class, Ticket type
Fourth model	Railcard, Ticket class, Ticket type, Price, Journey status, Reason for delay

Based on these results, we observe that all predictors have a statistically significant impact on the probability of requesting a refund. Additionally, **Model 4 demonstrates the best performance among the four models:** it has the lowest Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values, indicating a better fit.

We created confusion matrices for Model 2, Model 3 and Model 4 (Model 1 predicted all cases as zero). It can be observed that the last model has the highest number of correctly predicted refund requests.

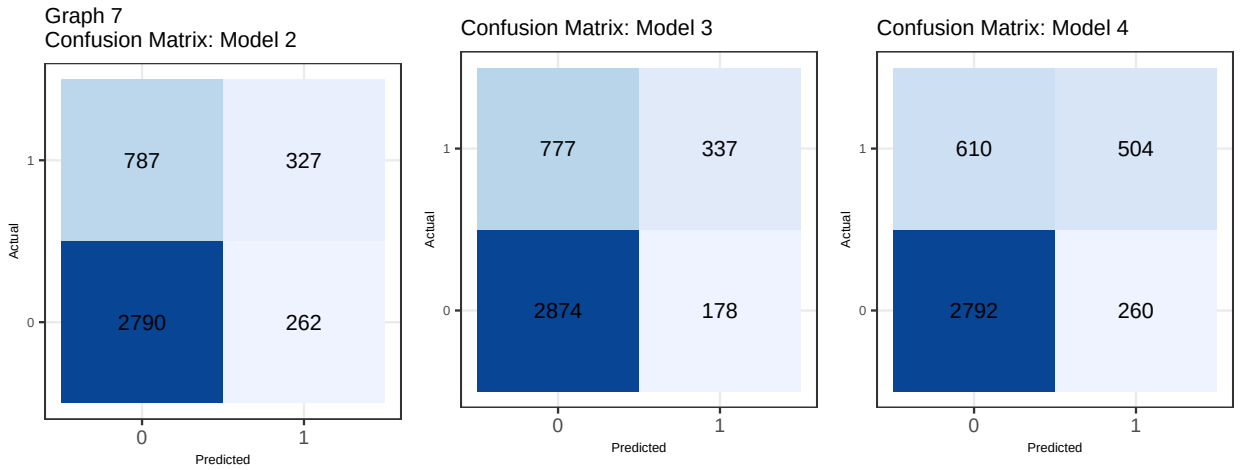
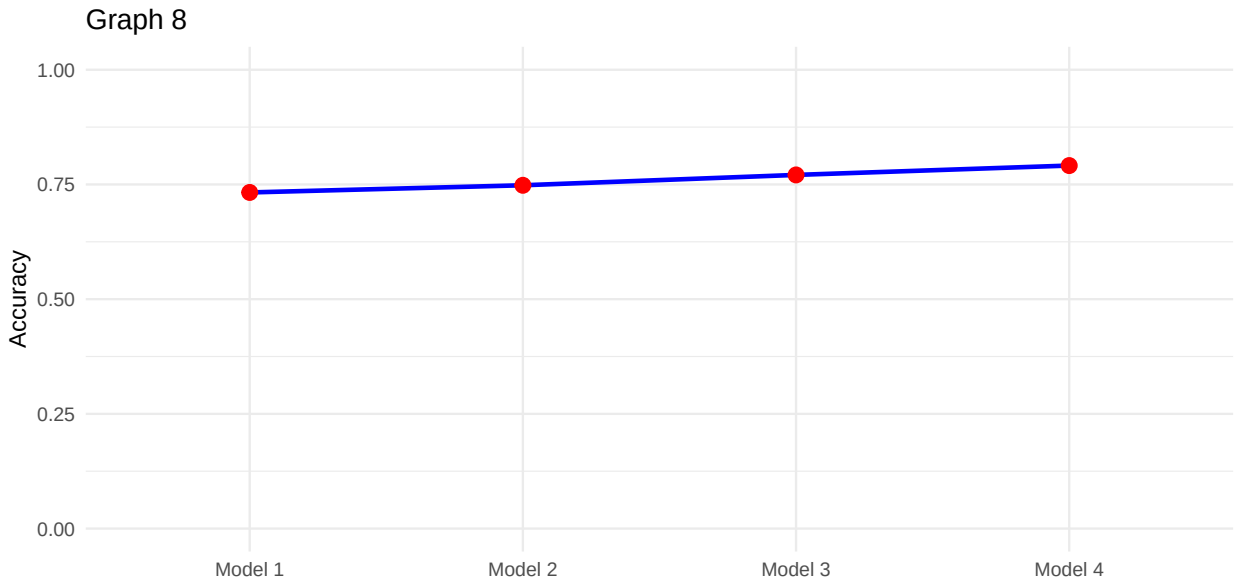


Table 3: Modeling probability of requesting refund

	Model 1	Model 2	Model 3	Model 4
(Intercept)	−1.278***	−0.582***	−0.869***	−0.856***
	[−1.426, −1.134]	[−0.762, −0.405]	[−1.173, −0.569]	[−1.215, −0.502]
PriceNEWBetween 6 and 16	0.418***	0.715***	0.768***	0.557***
	[0.222, 0.616]	[0.503, 0.928]	[0.551, 0.987]	[0.329, 0.786]
PriceNEWBetween 17 and 42	0.471***	1.022***	1.161***	0.891***
	[0.243, 0.698]	[0.774, 1.272]	[0.903, 1.419]	[0.624, 1.160]
PriceNEWBetween 43 and 76	0.476***	0.917***	1.014***	1.072***
	[0.265, 0.688]	[0.686, 1.149]	[0.778, 1.253]	[0.810, 1.336]
PriceNEWMore than 76	−0.008	0.565***	0.898***	1.405***
	[−0.241, 0.222]	[0.310, 0.819]	[0.617, 1.180]	[1.080, 1.732]
RailcardDisabled		−1.502***	−1.514***	−1.518***
		[−1.826, −1.193]	[−1.840, −1.202]	[−1.878, −1.169]
RailcardNone		−1.542***	−1.588***	−1.512***
		[−1.718, −1.367]	[−1.769, −1.408]	[−1.732, −1.293]
RailcardSenior		−0.103	−0.125	−0.178
		[−0.383, 0.176]	[−0.409, 0.157]	[−0.498, 0.140]
Ticket.ClassStandard			0.303*	0.409**
			[0.054, 0.558]	[0.144, 0.680]
Ticket.TypeAnytime			−0.765***	−0.991***
			[−0.998, −0.538]	[−1.247, −0.742]
Ticket.TypeOff-Peak			0.252**	0.146
			[0.080, 0.424]	[−0.036, 0.327]
Journey.StatusDelayed				−0.781***
				[−0.966, −0.597]
Reason.for.DelayStaff				0.168
				[−0.150, 0.479]
Reason.for.DelayStaffing				1.221***
				[0.948, 1.494]
Reason.for.DelayTechnical Issue				1.306***
				[1.076, 1.537]
Reason.for.DelayTraffic				0.748***
				[0.445, 1.049]
Reason.for.DelayWeather				−0.730***
				[−0.990, −0.472]
Num.Obs.	4166	4166	4166	4166
AIC	4808.1	4452.2	4378.1	3914.4
BIC	4839.7	4502.9	4447.8	4022.1
Log.Lik.	−2399.032	−2218.123	−2178.035	−1940.201
F	9.786	53.390	42.203	44.757
RMSE	0.44	0.42	0.41	0.39

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Examining the prediction accuracy of the four models, we also observe that **Model 4 has the highest performance, accurately predicting whether a passenger requested a refund (or not) in 80% of cases**.



Now, we show the calculations for the passengers with the highest and lowest probabilities of requesting a refund. For the passenger with the highest probability, Model 4 predicted an **88% chance of a refund request**. This passenger paid £76 for the ticket, used an Adult railcard, had a Standard ticket class, purchased the ticket during off-peak hours, experienced a cancelled journey, and the reason for the cancellation was "Staffing."

The "Adult" category in *Railcard* and "Cancelled" category in *Journey.Status* are reference categories, so their values are equal to zero:

$$\log(\text{Odds}) = -0.85625 + 1.07199 + 0.40870 + 0.14556 + 1.22060 = 1.991$$

Thus, the value of the odds is equal to 7.323:

$$\frac{\text{P(Refund.Request)}}{1 - \text{P(Refund.Request)}} = e^{1.991} = 7.323$$

Finally, the predicted probability of requesting a refund was 88%:

$$P(\text{Refund.Request}) = \frac{7.323}{1 + 7.323} = 0.88$$

```
## [1] 0.8798062
```

Now, we present the calculations for the passenger with the lowest probability of requesting a refund. This passenger paid £5 for the ticket, had no railcard, chose a Standard ticket class, purchased an "Anytime" ticket, experienced a delayed journey, and the reason for the delay was the weather. **Model 4 predicts a 1% probability of a refund request for this passenger**.

$$\log(\text{Odds}) = -0.85625 - 1.51177 + 0.40870 - 0.99111 - 0.78077 - 0.72976 = -4.46096$$

Thus, the value of the odds is equal to .01:

$$\frac{\text{P(Refund.Request)}}{1 - \text{P(Refund.Request)}} = e^{-4.46} = 0.01$$

And the predicted probability of requesting a refund is 1%:

$$P(\text{Refund.Request}) = \frac{0.02}{1 + 0.02} = 0.01$$

```
## [1] 0.01141948
```

Finally, we determine the likelihood of passengers in the file ToPredict.csv requesting a refund. This analysis is restricted for passengers with delayed or cancelled journeys, as the model was fit on that specific group.

The results indicate that **2 out of the 6 passengers are likely to request a refund based on Model 4**. These two passengers paid more than £100 for a standard class ticket, purchased their tickets off-peak, did not use a railcard, and experienced delays due to staffing problems:

```
##   Journey.Status Reason.for.Delay predicted_probabilities
## 1        Delayed         Staffing               0.5077222
## 5        Delayed         Staffing               0.5077222
```