Space Titanic: Análisis de datos

Presentación de hallazgos y desarrollo

Ramo: Aplicaciones de la matemática en la ingeniería

Universidad Técnica Federico Santa MarÃa

Noviembre 2023

Tabla de Contenidos

1. Presentación del problema
2. Estadística descriptiva
3. Visualización
4. Preprocesamiento
5. Selección de modelo
6. Métricas y análisis de resultados
7. Visualizaciones del modelo

Tabla de Contenidos

1. Presentación del problema
2. Estadística descriptiva
3. Visualización
4. Preprocesamiento
5. Selección de modelo
6. Métricas y análisis de resultados
7. Visualizaciones del modelo

Descripción del problema:

Bienvenido al año 2912, donde se necesitan tus habilidades de ciencia de datos para resolver un misterio cósmico. Hemos recibido una transmisión desde cuatro años luz de distancia y las cosas no pintan bien.
La nave espacial Titanic fue un transatlántico de pasajeros interestelar lanzado hace un mes. Con casi 13.000 pasajeros a bordo, la nave emprendió su viaje inaugural transportando emigrantes de nuestro sistema solar a tres exoplanetas recientemente habitables que orbitan estrellas cercanas.
Mientras rodeaba Alpha Centauri en ruta hacia su primer destino, el tórrido 55 Cancri E, la desprevenida nave espacial Titanic chocó con una anomalÃa del espacio-tiempo escondida dentro de una nube de polvo.
Lamentablemente, tuvo un destino similar al de su homónimo de 1000 años antes. Aunque la nave permaneció intacta, ¡casi la mitad de los pasajeros fueron transportados a una dimensión alternativa!

Tabla de Contenidos

1. Presentación del problema
2. Estadística descriptiva
3. Visualización
4. Preprocesamiento
5. Selección de modelo
6. Métricas y análisis de resultados
7. Visualizaciones del modelo

Data disponible

train

	PassengerId	HomePlanet	CryoSleep	Cabin	Destination	Age	VIP	RoomService	FoodCourt	ShoppingMall	Spa	VRDeck	Name	Transpo
0	0001_01	Europa	False	B/0/P	TRAPPIST-1e	39.0	False	0.0	0.0	0.0	0.0	0.0	Maham Ofracculy	F
1	0002_01	Earth	False	F/0/S	TRAPPIST-1e	24.0	False	109.0	9.0	25.0	549.0	44.0	Juanna Vines	
2	0003_01	Europa	False	A/0/S	TRAPPIST-1e	58.0	True	43.0	3576.0	0.0	6715.0	49.0	Altark Susent	F
3	0003_02	Europa	False	A/0/S	TRAPPIST-1e	33.0	False	0.0	1283.0	371.0	3329.0	193.0	Solam Susent	F
4	0004_01	Earth	False	F/1/S	TRAPPIST-1e	16.0	False	303.0	70.0	151.0	565.0	2.0	Willy Santantines	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
8688	9276_01	Europa	False	A/98/P	55 Cancri e	41.0	True	0.0	6819.0	0.0	1643.0	74.0	Gravior Noxnuther	F
8689	9278_01	Earth	True	G/1499/S	PSO J318.5-22	18.0	False	0.0	0.0	0.0	0.0	0.0	Kurta Mondalley	F
8690	9279_01	Earth	False	G/1500/S	TRAPPIST-1e	26.0	False	0.0	0.0	1872.0	1.0	0.0	Fayey Connon	
8691	9280_01	Europa	False	E/608/S	55 Cancri e	32.0	False	0.0	1049.0	0.0	353.0	3235.0	Celeon Hontichre	F
8692	9280_02	Europa	False	E/608/S	TRAPPIST-1e	44.0	False	126.0	4688.0	0.0	0.0	12.0	Propsh Hontichre	

8693 rows × 14 columns

Figura: Data de entrenamiento

Out[3]:	erld	HomePlanet	CryoSleep	Cabin	Destination	Age	VIP	RoomService	FoodCourt	ShoppingMall	Spa	VRDeck	Name
	_01	Earth	True	G/3/S	TRAPPIST-1e	27.0	False	0.0	0.0	0.0	0.0	0.0	Nelly Carsoning
	_01	Earth	False	F/4/S	TRAPPIST-1e	19.0	False	0.0	9.0	0.0	2823.0	0.0	Lerome Peckers
	_01	Europa	True	C/0/S	55 Cancri e	31.0	False	0.0	0.0	0.0	0.0	0.0	Sabih Unhearfus
	_01	Europa	False	C/1/S	TRAPPIST-1e	38.0	False	0.0	6652.0	0.0	181.0	585.0	Meratz Caltilter
	_01	Earth	False	F/5/S	TRAPPIST-1e	20.0	False	10.0	0.0	635.0	0.0	0.0	Brence Harperez
	...	...	...	...	...	...	...	...	...	...	...	...	...
	_02	Earth	True	G/1496/S	TRAPPIST-1e	34.0	False	0.0	0.0	0.0	0.0	0.0	Jeron Peter
	_01	Earth	False	NaN	TRAPPIST-1e	42.0	False	0.0	847.0	17.0	10.0	144.0	Matty Scheron
	_01	Mars	True	D/296/P	55 Cancri e	NaN	False	0.0	0.0	0.0	0.0	0.0	Jayrin Pore
	_01	Europa	False	D/297/P	NaN	NaN	False	0.0	2680.0	0.0	0.0	523.0	Kitakan Conale
	_01	Earth	True	G/1498/S	PSO J318.5-22	43.0	False	0.0	0.0	0.0	0.0	0.0	Lilace Leonzaley

Figura: Data de testeo (Variables independientes)

transports

	PassengerId	Transported
0	0013_01	False
1	0018_01	False
2	0019_01	False
3	0021_01	False
4	0023_01	False
...	...	...
4272	9266_02	False
4273	9269_01	False
4274	9271_01	False
4275	9273_01	False
4276	9277_01	False

4277 rows × 2 columns

Figura: Registro de pasajeros transportados

Estadística descriptiva

```
In [9]: train.nunique()

Out[9]: PassengerId     8693
        HomePlanet         3
        CryoSleep          2
        Cabin           6560
        Destination        3
        Age               80
        VIP                2
        RoomService     1273
        FoodCourt       1507
        ShoppingMall    1115
        Spa             1327
        VRDeck          1306
        Name            8473
        Transported        2
        dtype: int64
```

Figura: train.nunique()

Estadística descriptiva

	Age	RoomService	FoodCourt	ShoppingMall	Spa	VRDeck
count	8514.000000	8512.000000	8510.000000	8485.000000	8510.000000	8505.000000
mean	28.827930	224.687617	458.077203	173.729169	311.138778	304.854791
std	14.489021	666.717663	1611.489240	604.696458	1136.705535	1145.717189
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	19.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	27.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	38.000000	47.000000	76.000000	27.000000	59.000000	46.000000
max	79.000000	14327.000000	29813.000000	23492.000000	22408.000000	24133.000000

Figura: train.describe()

Tabla de Contenidos

1 Presentación del problema

2 Estadística descriptiva

3 Visualización

4 Preprocesamiento

5 Selección de modelo

6 Métricas y análisis de resultados

7 Visualizaciones del modelo

Variables cuantitativas

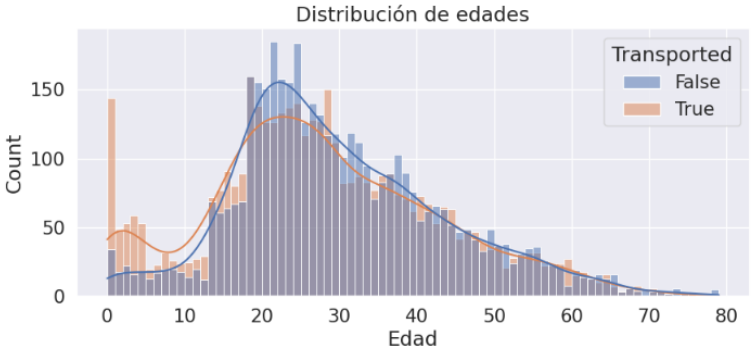

Figura: Distribución de edades

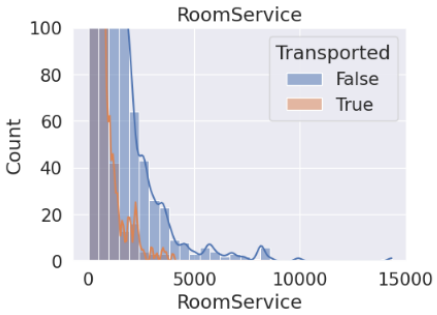

Figura: Gasto en servicio a la habitación

Variables cuantitativas

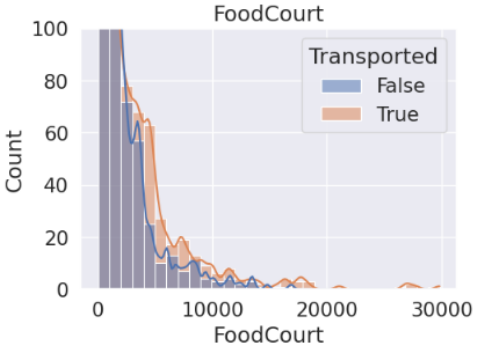

Figura: Gasto en patio de comidas

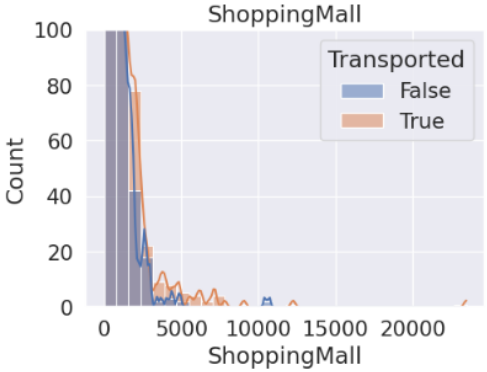

Figura: Gasto en centro comercial

Variables cualitativas

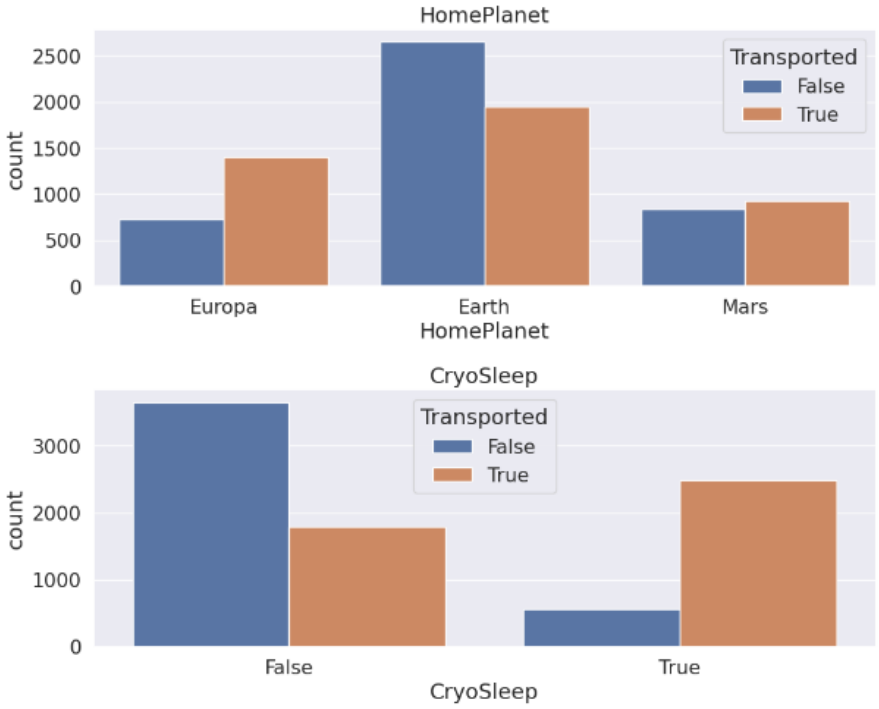

Figura: Visualización de las columnas HomePlanet y CryoSleep

Variables cualitativas

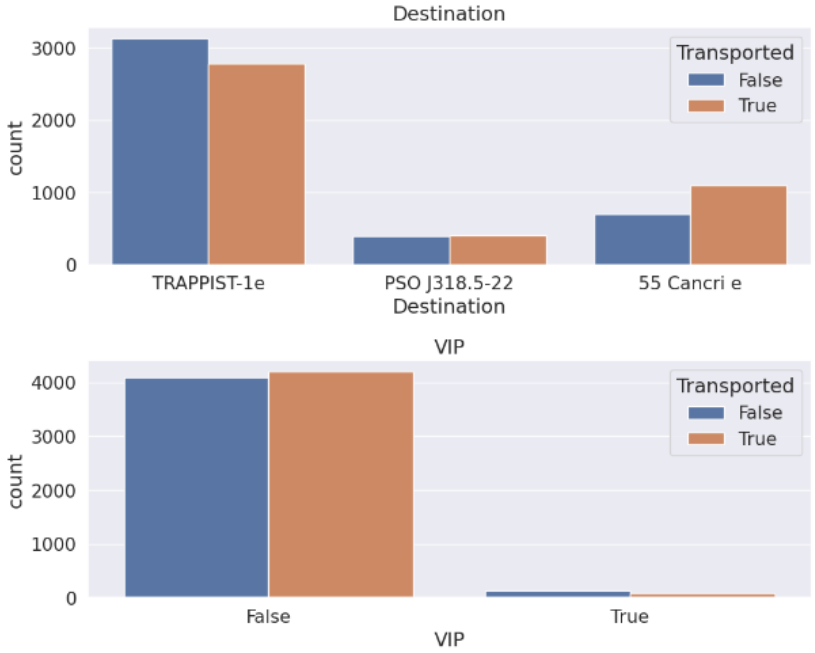

Figura: Visualización de las columnas Destination y Vip

Tabla de Contenidos

1. Presentación del problema
2. Estadística descriptiva
3. Visualización
4. Preprocesamiento
5. Selección de modelo
6. Métricas y análisis de resultados
7. Visualizaciones del modelo

Transformación de los datos de entrenamiento

	PassengerId	CryoSleep	Age	VIP	RoomService	FoodCourt	ShoppingMall	Spa	VRDeck	Transported	...
0	0001_01	0	39.0	0	0.0	0.0	0.0	0.0	0.0	0	...
1	0002_01	0	24.0	0	109.0	9.0	25.0	549.0	44.0	1	...
2	0003_01	0	58.0	1	43.0	3576.0	0.0	6715.0	49.0	0	...
3	0003_02	0	33.0	0	0.0	1283.0	371.0	3329.0	193.0	0	...
4	0004_01	0	16.0	0	303.0	70.0	151.0	565.0	2.0	1	...
...	...	...	...	...	...	...	...	...	...	...	...
8688	9276_01	0	41.0	1	0.0	6819.0	0.0	1643.0	74.0	0	...
8689	9278_01	1	18.0	0	0.0	0.0	0.0	0.0	0.0	0	...
8690	9279_01	0	26.0	0	0.0	0.0	1872.0	1.0	0.0	1	...
8691	9280_01	0	32.0	0	0.0	1049.0	0.0	353.0	3235.0	0	...
8692	9280_02	0	44.0	0	126.0	4688.0	0.0	0.0	12.0	1	...

8493 rows × 27 columns

Figura: Data modificada

Total_spending	HomePlanet_Earth	HomePlanet_Europa	HomePlanet_Mars	Destination_55 Cancri e	Destination_PSO J318.5-22	Destination_TRAPPIST-1e	Cabin_Deck_A	Cabin_Deck_B	Cabin_Deck_C
0.0	0	1	0	0	0	1	0	1	0
736.0	1	0	0	0	0	1	0	0	0
10383.0	0	1	0	0	0	1	1	0	0
5176.0	0	1	0	0	0	1	1	0	0
1091.0	1	0	0	0	0	1	0	0	0
774.0	1	0	0	0	1	0	0	0	0
1584.0	1	0	0	0	0	1	0	0	0
0.0	1	0	0	0	0	1	0	0	0
1018.0	1	0	0	0	0	1	0	0	0
0.0	0	1	0	1	0	0	0	1	0
0.0	0	1	0	0	0	1	0	1	0
8157.0	0	1	0	1	0	0	0	1	0
1309.0	0	0	1	0	0	1	0	0	0
809.0	1	0	0	0	0	1	0	0	0
1003.0	1	0	0	0	0	1	0	0	0

Figura: Data modificada

Transformacion de los datos de prueba

index	PassengerId	CryoSleep	Age	VIP	RoomService	FoodCourt	ShoppingMall	Spa	VRDeck	Total_spending
0	0013_01	1	27.0	0	0.0	0.0	0.0	0.0	0.0	0.0
1	0018_01	0	19.0	0	0.0	9.0	0.0	2823.0	0.0	2832.0
2	0019_01	1	31.0	0	0.0	0.0	0.0	0.0	0.0	0.0
3	0021_01	0	38.0	0	0.0	6652.0	0.0	181.0	585.0	7418.0
4	0023_01	0	20.0	0	10.0	0.0	635.0	0.0	0.0	645.0
5	0027_01	0	31.0	0	0.0	1615.0	263.0	113.0	60.0	2051.0
6	0029_01	1	21.0	0	0.0	0.0	0.0	0.0	0.0	0.0
7	0032_01	1	20.0	0	0.0	0.0	0.0	0.0	0.0	0.0
8	0032_02	1	23.0	0	0.0	0.0	0.0	0.0	0.0	0.0
9	0033_01	0	24.0	0	0.0	639.0	0.0	0.0	0.0	639.0
10	0037_01	0	19.0	0	339.0	3.0	136.0	237.0	0.0	715.0

Figura: Data modificada

Total_spending	HomePlanet_Earth	HomePlanet_Europa	HomePlanet_Mars	Destination_55 Cancri e	Destination_PSO J318.5-22	Destination_TRAPPIST-1e	Cabin_Deck_A	Cabin_Deck_B	Cabin_Deck_C
0.0	0	1	0	0	0	1	0	1	0
736.0	1	0	0	0	0	1	0	0	0
10383.0	0	1	0	0	0	1	1	0	0
5176.0	0	1	0	0	0	1	1	0	0
1091.0	1	0	0	0	0	1	0	0	0
774.0	1	0	0	0	1	0	0	0	0
1584.0	1	0	0	0	0	1	0	0	0
0.0	1	0	0	0	0	1	0	0	0
1018.0	1	0	0	0	0	1	0	0	0
0.0	0	1	0	1	0	0	0	1	0
0.0	0	1	0	0	0	1	0	1	0
8157.0	0	1	0	1	0	0	0	1	0
1309.0	0	0	1	0	0	1	0	0	0
809.0	1	0	0	0	0	1	0	0	0
1003.0	1	0	0	0	0	1	0	0	0

Figura: Data modificada

Tabla de Contenidos

1. Presentación del problema
2. Estadística descriptiva
3. Visualización
4. Preprocesamiento
5. Selección de modelo
6. Métricas y análisis de resultados
7. Visualizaciones del modelo

Métricas

```
Random forest accuracy: 0.804552590266876

Adaboost accuracy: 0.793171114599686

Gaussian NB accuracy: 0.6620879120879121

Multinomial NB accuracy: 0.6840659340659341
```

Figura: Accuracy score de 4 modelos

Tabla de Contenidos

1. Presentación del problema
2. Estadística descriptiva
3. Visualización
4. Preprocesamiento
5. Selección de modelo
6. Métricas y análisis de resultados
7. Visualizaciones del modelo

Accuracy score y matriz de confusión

```
[155] accuracy = accuracy_score(y_true['Transported'], y_pred)
      print(accuracy)

      0.5288070762610566
```

Figura: Accuracy score

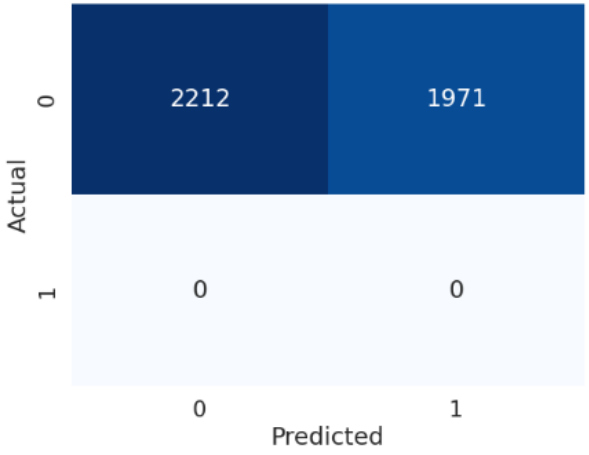

Figura: Matriz de confusión

Tabla de Contenidos

1 Presentación del problema

2 Estadística descriptiva

3 Visualización

4 Preprocesamiento

5 Selección de modelo

6 Métricas y análisis de resultados

7 Visualizaciones del modelo

Visualizaciones

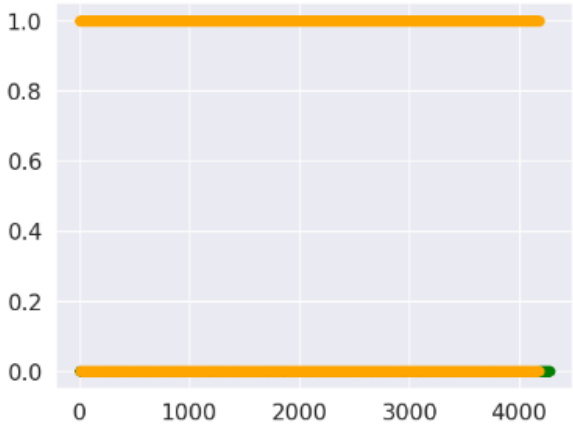

Figura: Visualizacion

Conclusiones

Verdaderos Positivos (VP): 2219
Falsos Positivos (FP): 1964
Verdaderos Negativos (VN): 0
Falsos Negativos (FN): 0
Esta matriz de confusión sugiere que nuestro modelo hizo predicciones positivas correctas (2219) pero realizó un gran número de predicciones falsas positivas (1964). El modelo no hizo ninguna predicción negativa correcta, y no hubo instancias en las que predijo correctamente la clase negativa.

Muchas Gracias