Titanic Spaceship

Javier Corral
Damian Perez
Ricardo Meadowcroft

El suceso

Es el año 2912 y la nave interestelar “Titanic” con 13,000 pasajeros mientras rodeaba Alfa Centauri colisionó contra una anomalía espacio-temporal escondida en una nube de polvo en la cual casi la mitad de los pasajeros fue transportada a una dimensión alternativa

Nuestro papel

Para ayudar a rescatar a estos pasajeros intentaremos predecir quienes serán transportados basándonos en la información registrada por la nave acerca de los pasajeros

Los datos que tenemos por pasajero son:

- `PassengerId (separación)`
- `HomePlanet (dummies)`
- `CryoSleep`
- `Cabin deck/num/lado (separación y dummies)`
- `Destination (dummies)`
- `VIP`
- `RoomService`
- `FoodCourt`
- `ShoppingMall`
- `Spa`
- `VRDeck`
- `Name (descartar)`
- `Transported`

Consiguiendo un criterio

Seguimos la metodología CRISP-DM:

- Utilizamos R para limpiar los datos y crear variables derivadas
- Utilizamos R para un análisis exploratorio de los datos (EDA)
- Utilizamos Python para entrenar modelos a la vez que optimizamos los hiperparametros con Grid Search
- Comparamos los scores de cada modelo y elegimos el mejor con la matriz de confusión
- Utilizamos el mejor modelo para predecir si los pasajeros de los datos de prueba serian transportados y lo subimos a Kaggle el cual nos dio un score de 73%

Limpieza y análisis de datos

- Se modificaron las variables para representar características más fundamentales y fáciles de relacionar
  - La cabina -> Nivel, lado y Número
  - El id -> Número de grupo, y número dentro del grupo
- Se imputaron valores faltantes usando k-Nearest Neighbors
- En el modelado, las variables categóricas (Exceptuando nombre) se convirtieron en binarias con variables dummy
- El análisis uni-, bi- y multivariado se efectuó con la ayuda de Shiny en R, y en Python TSNE permite también visualizar su agrupamiento en altas dimensiones

C:/Users/HP/Downloads/MineriaDeDatos-main/Final/spaceship-titanic - Shiny

http://127.0.0.1:4984 | Open in Browser | Publish

Auto Information

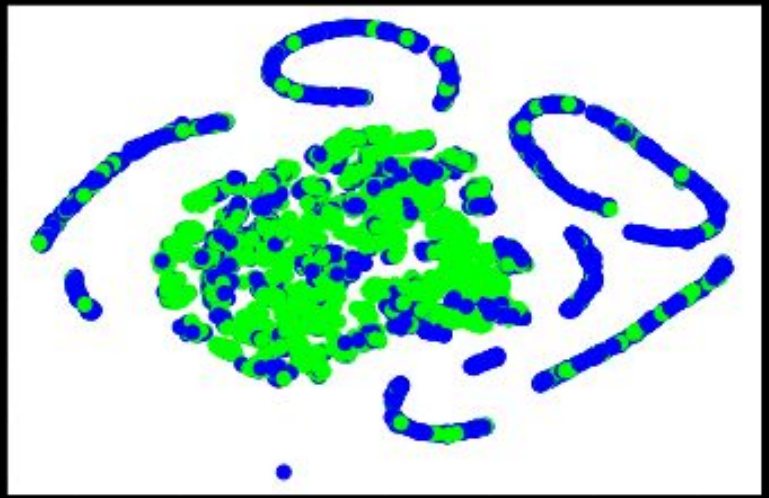
Caracteristica

Destination

Segunda Caracteristica (Bivariado)

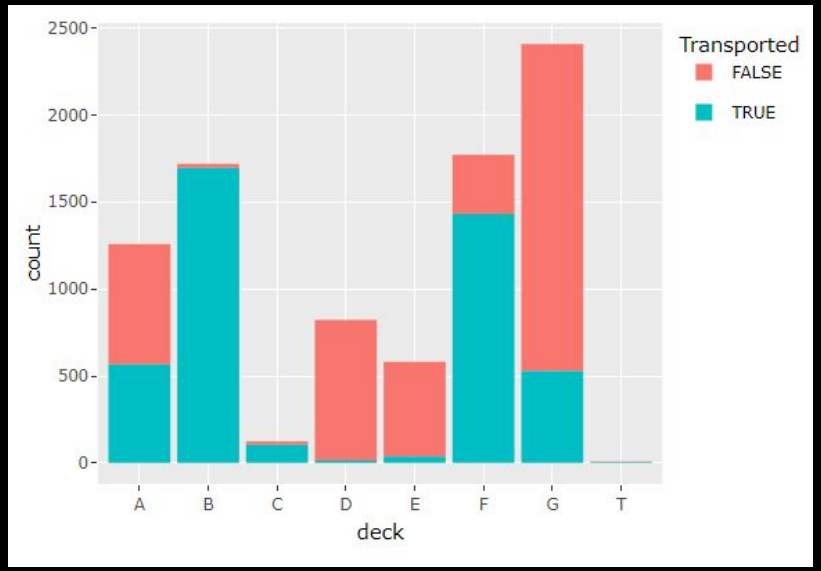
HomePlanet

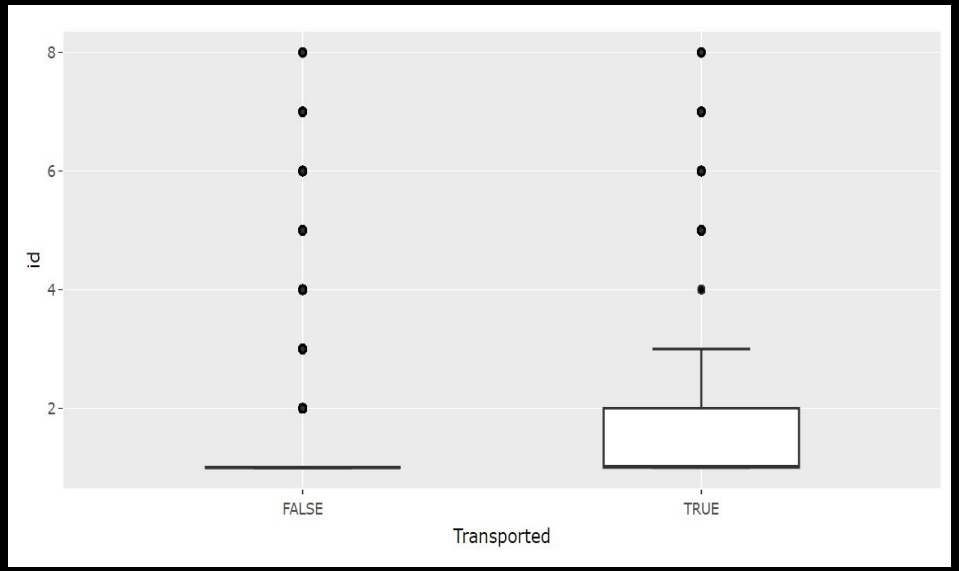
Univariado | Bivariado | Multivariado



Entendiendo el criterio

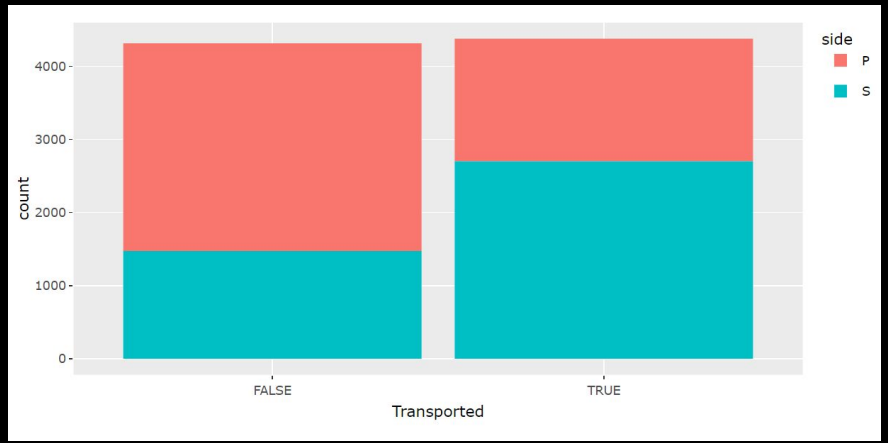
Para entender el criterio que tomó el modelo podemos ver cómo se relaciona la información:

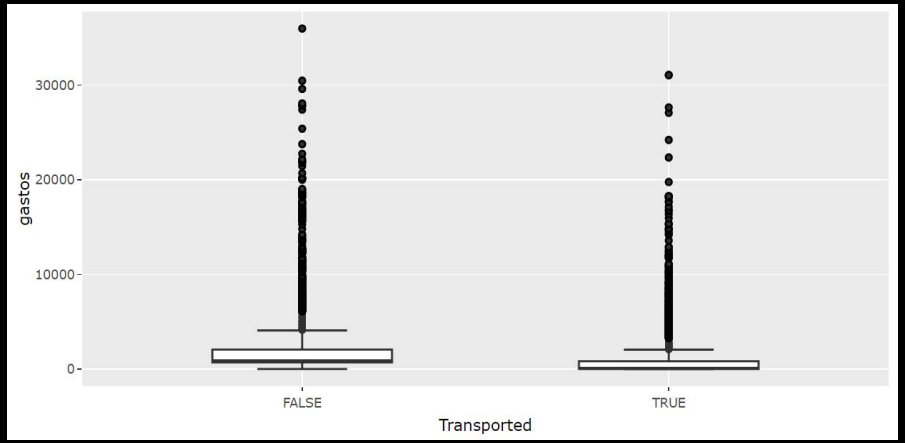





HomePlanet
Earth
Europa
Mars
count
4000
3000
2000
1000
0
FALSE
TRUE
Transported

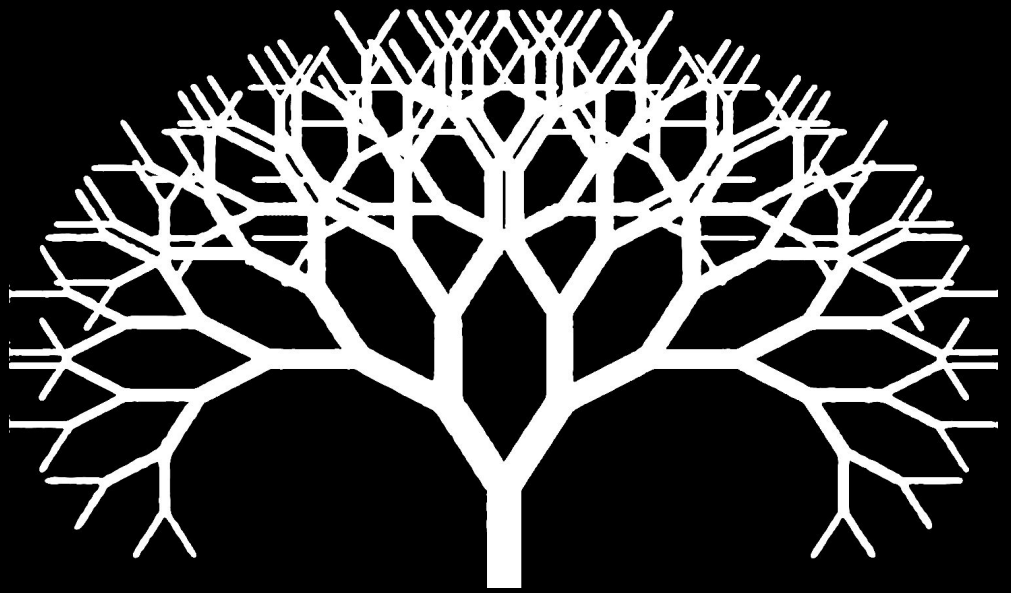

CryoSleep
FALSE
TRUE
count
4000
3000
2000
1000
0
FALSE
TRUE
Transported


side
P
S
count
4000
3000
2000
1000
0
FALSE
TRUE
Transported

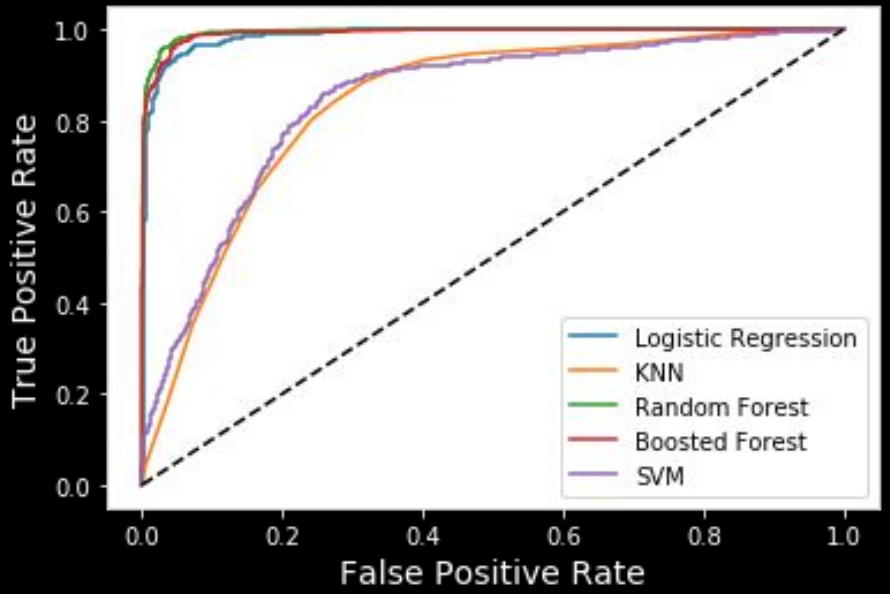

gastos
30000
20000
10000
0
FALSE
TRUE
Transported

Modelos probados

1. Logistic Regression
2. KNN Classifier
3. Random Forest
4. Boosted Forest
5. Support Vector Machine



Resultados



Vimos que Boosted Forest, Random Forest, y Logistic Regression a través del uso de dummies fueron mucho más adecuados para este desafío; de estos, ***Random Forest*** fue el elegido.

Entregable

Hicimos una página web con el modelo embedded usando el framework de django, al cual si le pasas al formulario los datos de un nuevo pasajero predice si será transportado o no.

Predictor de catástrofes interdimensionales

¿Serás enviado a otra dimensión?

Campo	Valor
Grupo:	
Miembro:	
CryoSleep:	True
Deck:	A
Número de habitación:	
Side:	P
Edad:	
VIP:	True
Home Planet:	Europa
Destination:	TRAPPIST-1e
RoomService:	
FoodCourt:	
ShoppingMall:	
SPA:	
VRDeck:	
Limpiar	Enviar

GRACIAS