INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY, CAMPUS ESTADO DE MÉXICO.

Escuela de Ingenierías

Reporte final del Reto de Kaggle 'Spaceship Titanic'

Inteligencia artificial avanzada para la ciencia de datos I (Gpo 101)

Equipo Ciete:

Diana Cañibe Valle	A01749422
Andrea Viany Díaz Álvarez	A01750147
José Benjamín Ruiz García	A01750246
Edna Jacqueline Zavala Ortega	A01750480
Cristian Aldo Sandoval Suárez	A01751137

Fecha de entrega: 15 de septiembre del 2022

Índice

Introducción

En este documento se reporta el proceso y los resultados obtenidos durante el desarrollo del reto ‘Spaceship Titanic’. Para dicho desarrollo se utilizó la metodología CRISP-DM, por lo tanto se plantea a continuación lo que se realizó en cada etapa (entendimiento de los datos y del negocio, procesamiento de datos, modelado y evaluación, despliegue).

Etapa 1 - Entendimiento del negocio

En el año 2912 el transatlántico interestelar 'Spaceship Titanic' emprendió su travesía con 13,000 pasajeros a bordo, desde nuestro sistema solar a tres exoplanetas (55 Cancri E,PSO J318.5-22,TRAPPIST-1e).

Lamentablemente mientras rodeaba Alpha Centauri en ruta hacia su primer destino,la nave chocó con una anomalía del espacio-tiempo escondida dentro de una nube de polvo, la cual causó la teletransportación de casi la mitad de sus pasajeros a una dimensión alternativa. (Kaggle,s.f.)

La misión es predecir si un pasajero fue transportado a una dimensión alternativa durante la colisión de la nave espacial Titanic con la anomalía del espacio-tiempo. Para lograrlo es necesario generar un modelo de predicción utilizando los registros recuperados de la nave espacial para determinar qué pasajeros fueron transportados, y habilitar una plataforma para realizar consultas al modelo.

Se tiene 3 tareas principales:

1. Análisis y Procesamiento de Datos: Realizar el estudio inicial de los datos proporcionados, y la limpieza, imputación y cambios considerados adecuados
2. Modelado: Construcción, pruebas y ajustes de diferentes modelos para la selección del mejor modelo
3. Despliegue del modelo: Desarrollo de aplicación para demostración práctica del modelo desarrollado, proporcionando los resultados de clasificación

Etapa 2 - Entendimiento de los datos

Descripciones de archivos y campos de datos

A) train.csv : registros personales de aproximadamente dos tercios (~ 8700) de los pasajeros, que se utilizarán como datos de capacitación.

- PassengerId- Un Id único para cada pasajero. Cada Id toma la forma gggg_p pdonde gggg indica un grupo con el que viaja el pasajero y ppes su número dentro del grupo. Las personas en un grupo a menudo son miembros de la familia, pero no siempre.
- HomePlanet- El planeta del que partió el pasajero, normalmente su planeta de residencia permanente.
- CryoSleep- Indica si el pasajero eligió ser puesto en animación suspendida durante la duración del viaje. Los pasajeros en criosueño están confinados en sus cabinas.
- Cabin- El número de cabina donde se hospeda el pasajero. Toma la forma deck/num/side, donde Side puede ser ‘P’ por Babor, o ‘S’ por Estribor .
- Destination- El planeta donde desembarcará el pasajero.
- Age- La edad del pasajero.

- VIP- Si el pasajero ha pagado por servicio VIP especial durante el viaje. RoomService, FoodCourt, ShoppingMall, Spa, VRDeck- Monto que el pasajero ha facturado en cada uno de los muchos servicios de lujo del Spaceship Titanic .
- Name- El nombre y apellido del pasajero.
- Transported- Si el pasajero fue transportado a otra dimensión. Este es el objetivo, la columna que está tratando de predecir.

B) test.csv : registros personales del tercio restante (~4300) de los pasajeros, que se utilizarán como datos de prueba. Su tarea es predecir el valor de Transported para los pasajeros en este conjunto.

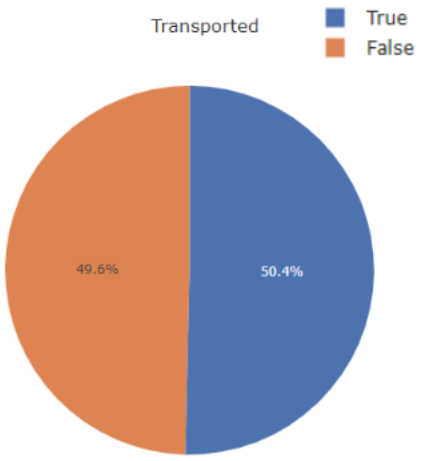
C) sample_submission.csv : un archivo de envío en el formato correcto.

- PassengerId- Id para cada pasajero en el conjunto de prueba.
- Transported- El objetivo. Para cada pasajero, prediga True o False.

Posterior a la descarga de los datos, podemos proceder a hacer una exploración inicial de los datos previa a su preprocesamiento, con el fin de conocer más información sobre los mismos; como tipo de dato, cantidad, etc.

	PassengerId	HomePlanet	CryoSleep	Cabin	Destination	Age	VIP	RoomService	FoodCourt	ShoppingMall	Spa	VRDeck	Name	Transported
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	False
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	True
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	False
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	False
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	True

Tenemos que validar el balance de los datos, específicamente la variable objetivo que es 'Transported':



Este proceso se puede repetir para validarlo por cada atributo del set de datos, para visualizar las gráficas ver la sección '*2. Entendimiento de los datos*' en el siguiente google colab: Spaceship Titanic.ipynb

Etapa 3 - Preparación de los datos

Una vez que podemos afirmar conocer los datos, empezamos el proceso de limpieza e imputación para poder contar con un set de datos íntegro y funcional para poder utilizar en los modelos de ML.

La lista de cambios totales durante las distintas fases de cambios, es decir la instancia original, los cambios tras el primer modelo, y los cambios para el modelo final son los siguientes:

- Fase 1
    - Separación de cabina en: puerto (Deck) , número de puerto (deck_number) y lado (Side)
    - Cambio de tipo de dato bool a int para Cryosleep, VIP
    - One hot encoding : HomePlanet, Destination, Side
    - Label encoding: Deck
    - Llenado de variables categóricas con la moda (train)
    - Llenado de variables numéricas con la mediana (train)
- Fase 2
    - Nueva columnas:
        - Separación por grupos de edad (age_group)
        - Suma de gastos (total_expenses)
        - Cualidad de haber gastado (no_expenses)
        - Separación por grupos en base a Id del pasajero (group)
    - Llenado de variables categóricas con la moda (train+test)
    - Llenado de variables numéricas con la mediana (train+test)
- Fase 3
    - Estandarización de los datos

Para visualizar los cambios directos ver la sección ‘Preprocesamiento de datos’ en el siguiente google colab: Reto Titanic Espacial.ipynb

Etapa 4 y 5 - Modelado y Evaluación

En esta etapa realizamos distintas pruebas y validaciones para seleccionar el modelo final, podemos considerar 5 fases:

Fase 1

Intento inicial con datos llenos y separados, decidimos probar con regresión logística ya que la predicción requerida se divide en 2 (verdadero,falso), además queríamos obtener un indicador inicial para realizar modificaciones a los datos antes de probar cambios en los hiper parámetros. El puntaje obtenido en kaggle con este modelo fue de: 0.72340.

Fase 2

Decidimos agregar más columnas calculadas al dataset y comprobar si tenían un impacto en la predicción, de igual manera probamos diferentes algoritmos de machine learning para ver cual obtenía un mejor resultado en el entrenamiento y después hacer la prueba.

Modelo	Logistic Regression	KNN	SVC	Random Forest	Decision Tree	XGBoost	LGMB	CatBoost	Naive Bayes
Score	0.7853	0.7842	0.7997	0.7741	0.7534	0.8141	0.8037	0.8037	0.7359

Fase 3

En paralelo realizamos un intento de experimentar la funcionalidad de AutoML, aspirando a tener mejores resultados. Sin embargo, aunque el modelo proporcionado validaba un 81% de efectividad con los datos de prueba, al hacer el test y subirlo a Kaggle, obtuvo un puntaje de tan solo 0.58569, siendo incluso menor a nuestro primer intento, por lo tanto, descartamos su uso.

Fase 4

Aunque los resultados de las pruebas eran ya considerablemente buenos, aspiramos a más y realizamos una estandarización de los datos y combinamos el conjunto de train y test proporcionados para obtener la mediana y moda de los datos utilizados para la imputación de los valores faltantes. Probamos nuevamente con diferentes modelos, pero realizamos una validación cruzada para el entrenamiento, obteniendo así el modelo que obtuvo el puntaje más alto hasta el momento (0.80009) siendo una regresión logística.

Fase 5

Siendo ambiciosos y buscando mejorar realizamos un grid search y un k best selection, para ajustar el modelo y los hiper parámetros con los que esperábamos obtener un score más alto, sin embargo los resultados no fueron los esperados ya que el puntaje se mantuvo constante oscilando entre 0.78 y 0.80. Finalmente mantuvimos como modelo final el obtenido previamente en la fase 4.

Para visualizar los modelos ver la sección ‘Modelado’ y ‘4.Modelo’ en los siguientes google colab: Reto Titanic Espacial.ipynb , Spaceship Titanic.ipynb

Etapa 6 - Despliegue

Una vez que determinamos el modelo final, procedimos a desarrollar la aplicación con react native, tras algunas complicaciones y considerando el tiempo restante se tomó la decisión de mejor trabajar con react. Al mismo tiempo se realizó el pipeline para el procesamiento de los datos que se obtendrían del formulario realizado para las pruebas. Se realizó la API necesaria y el servidor con la EC2 de AWS. Así también se desarrolló la propuesta de negocio y la creación de una empresa para desplegar el producto final y ofrecer un contexto más completo al reto.