Titanic - Machine Learning

Samuel Peláez Alemán, Diego Velázquez Saldaña, María Fernanda Argueta, Alain Hurtado Escamilla, Diego Loyo Villagrán
Tecnológico de Monterrey
(Dated: 19 de Agosto, 2024)

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Keywords: first keyword, second keyword, third keyword

I. INTRODUCCIÓN

A continuación se describen todos los procedimientos y el análisis llevado a cabo para resolver el problema de Supervivencia Titanic aplicando ciencia de datos y generando un modelo de Machine Learning para llegar a la solución óptima. El objetivo principal de este reto es crear un modelo con el que se pueda predecir de forma efectiva la supervivencia de un pasajero del Titanic dadas ciertas características provenientes de la base de datos.

El reto a resolver se obtuvo de la plataforma de Kaggle, la cual proporciona una base de datos previamente dividida entre train y test. Dichas bases de datos cuentan con información detallada sobre los pasajeros indicando las siguientes features: survival, pclass, sex, Age, sibsp, parch, ticket, fare, cabin, y embarked. Se le realizó una limpieza a los datos utilizando herramientas de ETL.

Se hizo un exhaustivo análisis y exploración de los datos. En donde se procuró identificar el comportamiento de las distintas variables, su relación entre sí e incluso su influencia en la supervivencia del pasajero. Luego se realizó imputación, eliminación y creación de nuevos datos.

Con lo que finalmente se diseñó.....

II. LIMPIEZA DEL CONJUNTO DE DATOS

Durante el proceso de preparación de los datos de la base de supervivientes del Titanic, se realizaron varias acciones para garantizar la integridad y calidad del conjunto de datos; es importante considerar que los cambios que se les ajusta a una de las bases de datos, se le debía de hacer a la otra

1. Eliminación de la columna 'Cabin': Se decidió eliminar la columna 'Cabin' debido a que presentaba más del 77% de datos faltantes. La cantidad tan significativa de valores ausentes en esta columna podría haber introducido sesgos o ruido en los análisis posteriores, por lo que se optó por removerla completamente.

2. Remoción de filas con datos faltantes en la columna 'Embarked': Dentro del conjunto de datos de entrenamiento, se identificaron 2 pasajeros a los que les faltaba información en la columna 'Embarked'. Dado que esta información es crucial para el análisis, se decidió eliminar estas filas para evitar posibles inconsistencias en los resultados.

3. Remoción de fila con dato faltante en la columna 'Fare': En el conjunto de datos de prueba, se detectó que un pasajero carecía de información en la columna 'Fare'. Al igual que en el caso anterior, se decidió eliminar esta fila para mantener la coherencia del análisis.

4. Eliminación de la columna 'PassengerId': Tanto en el conjunto de datos de entrenamiento como en el de prueba, se removió la columna 'PassengerId'. Esta columna, que solo sirve como identificador, no aportaba valor al análisis predictivo, por lo que su eliminación ayudó a simplificar el modelo sin pérdida de información relevante.

III. EXPLORACIÓN DE DATOS

Para tener una mejor idea del comportamiento de los datos y de cuales variables eran de valor para conservar

o no se hizo realizó un análisis para saber que variables tienen valores nulos. En base a lo observado durante esta etapa se pudo tomar decisiones para poder eliminar ciertas columnas de los datasets que al tener valores faltantes no van a poder aportar al modelo. En la Fig 1 y Fig 2 podemos observar el porcentaje de datos faltantes por columnas en ambos datasets de entrenamiento y evaluación.

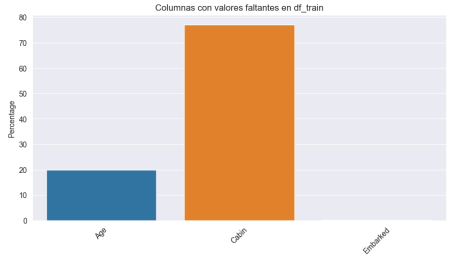

Figure 1. Features con valores faltantes en el set de entrenamiento

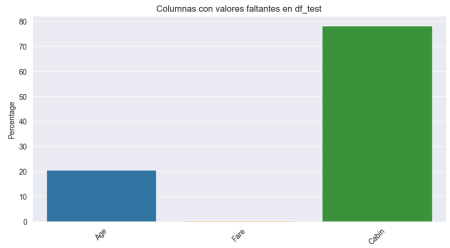

Figure 2. Features con valores faltantes en el set de evaluación

Otra variable que se exploro fue la probabilidad de supervivencia por grupo de edad y sexo. En la Fig 3 se nota como cada grupo tiene distintas probabilidades de supervivencia. esto nos permite generar esta nueva feature para posteriormente ingresarla al modelo.

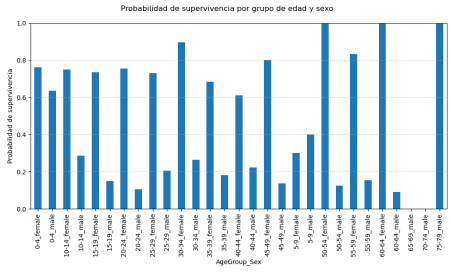

Figure 3. Probabilidad de supervivencia por grupo de edad y sexo

Por otra parte, se hizo el análisis de la influencia en la cantidad de acompañantes en la supervivencia de los sujetos. Tal como se muestra en la Fig 4 las personas que iban en grupos pequeños o solos tenían más probabilidad de sobrevivir que aquellos pertenecentes a un grupo mediano o grande.

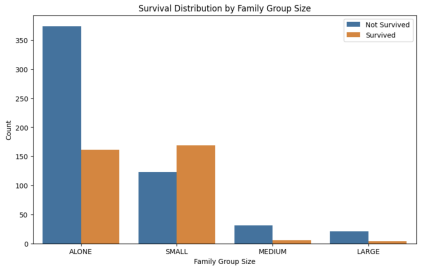

Figure 4. Probabilidad de supervivencia dependiendo acompañantes

IV. TRANSFORMACIÓN DE DATOS

Con el fin de agrupar elementos que tenían relación y para reducir la cantidad de features con los que se iba a trabajar se transformaron los datos,lo cual consistió en generar nuevas columnas en base a la unión de otras o tomando en cuenta información proveniente de otros features.

Pclass All	Ticket_FirstDigit	0	1	2	3	4	5	6	7	8	9
1	0	190	4	14	0	4	2	0	0	0	214
2	0	22	136	23	0	1	0	2	0	0	184
3	4	17	90	330	15	4	12	13	3	3	491
All	4	229	230	367	15	9	14	15	3	3	889

Table I. Distribución de los primeros dígitos de los boletos por clase.

1. Dado que la columna "Age" presentaba un 19% de datos nulos en el dataset de entrenamiento y al ser una variable que podría proporcionar información valiosa, utilizamos una técnica para completar los valores faltantes. Para ello, primero separamos el título de las personas de su nombre. Usamos los títulos: "Capt", "Col", "Countess", "Don", "Dr", "Jonkheer", "Lady", "Major", "Master", "Miss", "Mlle", "Mme", "Mr", "Mrs", "Ms", "Rev" y "Sir", para agrupar a los pasajeros y calcular la media de edad de cada grupo, excluyendo aquellos registros donde la edad era desconocida. Posteriormente, para aquellos pasajeros cuya edad no estaba registrada, asignamos un valor aleatorio dentro de un rango definido por la media de edad de

su grupo más o menos una desviación estándar. Esta técnica nos permitió imputar las edades de manera que se preservara la distribución original y se redujera el sesgo en la predicción del modelo.

2. Se comprobó la hipótesis que se tenía, que la cantidad de acompañantes era influyente en la supervivencia del pasajero tal como se puede observar en Fig 4. Por lo que se tomó la información del feature de sibsp que indicaba el número de hermanos o esposos. También se tomó la información de parch que indica el número de padres o hijos. Primero que nada se calculó la cantidad de personas en la familia de cada pasajero, para lo que se utilizó la siguiente fórmula.

$$FamilyMembers = sibsp + parch + 1 \quad (1)$$

A esto se le suma uno para tomar en cuenta al pasajero en la cuenta de miembros de familia. Después se separó por tamaño de familia, donde una familia pequeña tiene de 2-4 miembros, una mediana 5 o 6 y una grande 7-11. Los grupos que contienen un solo miembro se nombraron como solo.

3. Otra variable que se destacó como relevante durante la exploración de datos fue "Sex". La relación entre "Sex" y "Age" resultó especialmente interesante, ya que al agrupar estas dos variables, obtuvimos información valiosa. Esto se hizo evidente cuando calculamos la probabilidad de supervivencia, que variaba significativamente para los grupos formados por mujeres de 0 a 5 años, hombres de 0 a 5 años, y así sucesivamente en intervalos de 5 años. A partir de esta observación, generamos una nueva columna denominada "AgeGroupSex", donde combinamos las columnas "Age" y "Sex" para segmentar aún más a la población.

4. Para poder obtener más información de los pasajeros, se utilizo la variable "Ticket", según un proyecto en Kaggle [1] al descomponer en tres partes el "Ticket" se puede obtener informaci

[1] Oscar Takeshita, *Titanic Ticket-only study* (Kaggle, 2018).

Appendix: Appendix

Acceso al repositorio de Github con el código realizado. Aquí