Reporte prueba técnica Data Science

Autor:
Wolph Ronald Shwagger
Paul Fils

7 de febrero de 2023

Índice

1. Descripción de los datos

En Los Ángeles existe un sistema compartido de bicicletas que brinda datos anónimos acerca del uso del servicio. La tabla que se proporciona contiene el histórico de viajes que se han realizado desde 2016 y contiene una columna que es de particular interés y que se buscará analizar a más profundidad: *Passholder_type*. A continuación se presentan las columnas que contiene la tabla:

- *trip_id*: identificador único para el viaje.
- *duration*: duración del viaje en minutos.
- *start_time*: día/hora donde en viaje inicia en formato ISO 8601 tiempo local.
- *end_time*: día/hora donde el viaje termina en formato ISO 8601 tiempo local.
- *start_lat*: la latitud de la estación donde el viaje se originó.
- *end_lat*: la latitud de la estación donde terminó el viaje.
- *start_lon*: la longitud de la estación donde el viaje se originó.
- *end_lon*: la longitud de la estación donde terminó el viaje.
- *bike_id*: un entero único que identifica la bicicleta.
- *plan_duration*: número de días que el usuario tendrá el paso. 0 significa un viaje único (Walk-up plan).
- *trip_route_category*: "Round trip" son viajes que empiezan y terminan en la misma estación.
- *passholder_type*: El nombre del plan de passholder.
- *start_station*: la estación donde el viaje inició.
- *end_station*: la estación donde el viaje terminó.

2. Análisis exploratorio de los datos

Un dato es una representación simbólica (numérica, alfabética, algorítmica, espacial, etc.) de un atributo o variable cuantitativa o cualitativa. Al agregarle contexto a esos datos (su origen, cuando se obtuvieron, etc.) se obtiene información. Un conjunto de datos representa la información concreta sobre hechos o elementos, que al estudiar y analizarlo permite obtener conocimientos o realizar deducciones. La información recompilada puede estar alterada de alguna manera, eso significa que puede contener imperfecciones. Para una tarea de análisis de datos, es de suma importancia

identificar y deshacerse de las imperfecciones dentro del conjunto de datos ya que inferirse de manera negativa en los resultados obtenidos.

2.1. Identificación de tipos de datos

La primera tarea llevada a cabo al cargar los datos fue identificar y asignar los tipos de datos correspondientes a cada columna. En la Tabla 1 se puede ver a que tipo de datos pertenece cada una de las columnas.

Cuadro 1: Tipo de datos de cada columna.

Variable	Tipo de datos
trip_id	Categórico-nominal
duration	Discreto
start_time	Fecha
end_time	Fecha
start_lat	Continuo
end_lat	Continuo
start_lon	Continuo
end_lon	Continuo
bike_id	Categórico
plan_duration	Discreto
trip_route_category	Categórico
passholder_type	Categórico
start_station	Categórico-nominal
end_station	Categórico-nominal

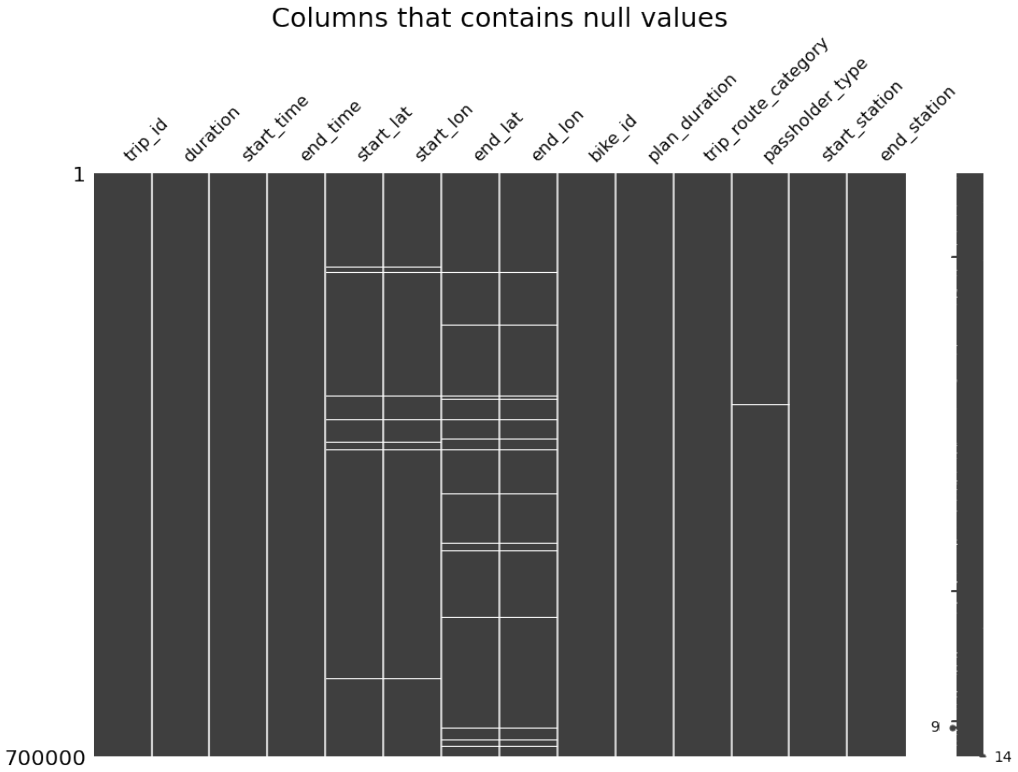

Figura 1: Mapa de valores nulos.

2.2. Identificación de valores nulos

No todos los campos del conjunto de datos contienen datos validos, existen valores nulos representados por los huecos blancos de la Figura 1. Las columnas con valores nulos son: *start_lat*, *start_lon*, *end_lat*, *end_lon*, *plan_duration*, *passholder_type*.

En la Figura 2 se logra ver la cantidad de instancias válidas con una gráfica de barras, las filas con valores nulos son muy pequeñas a comparación del conjuntos de datos en general que contiene $700,000$ instancias. Por lo tanto, eliminar algunas filas no tendría efecto negativo sobre los próximos análisis tanto exploratorio como predictivo, sobre todo si esas instancias contienen valores nulos. En estadística, se puede hacer deducciones de una muestra poblacional, lo que se conoce como inferencia estadística. Al eliminar las filas con valores nulos, los datos pasan de $700,000$ instancias a $675,626$ instancias válidas por así decirlo.

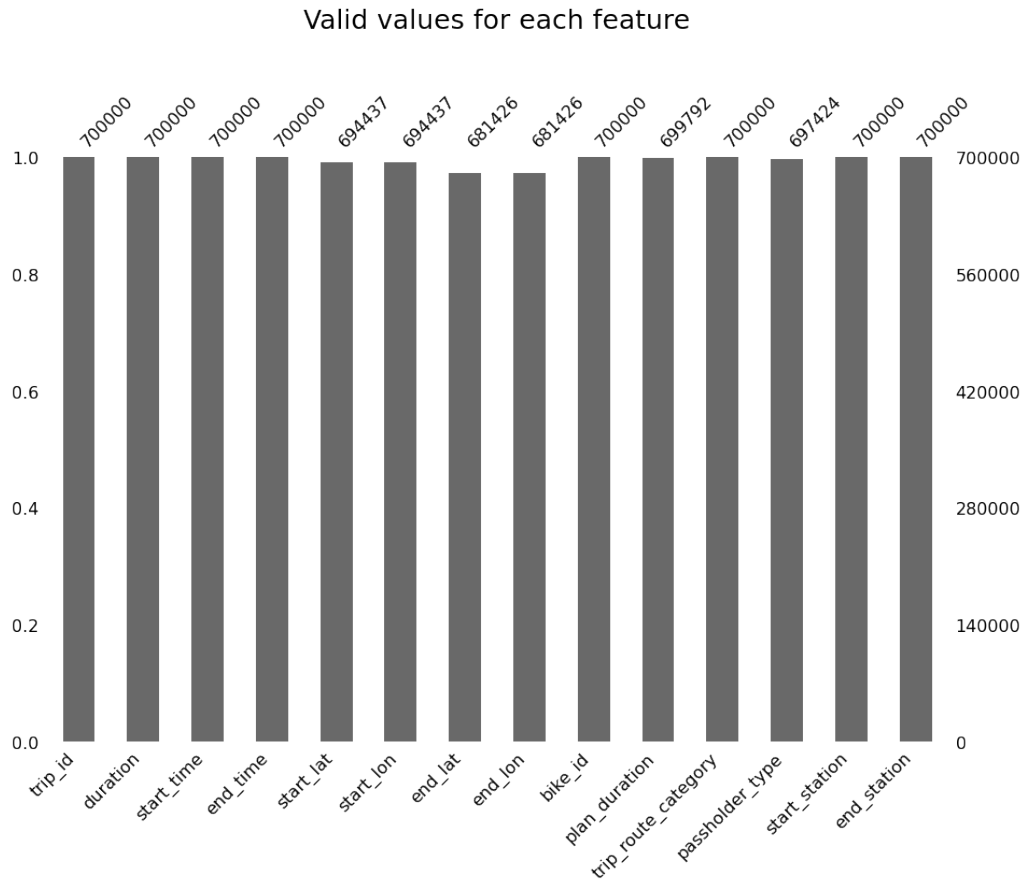

Figura 2: Cantidad de valores que no son nulos.

2.3. Detección de inconsistencia en el conjunto de datos

La inconsistencia se refiere a que existe información contradictoria o incongruente en el conjunto de datos. Inicialmente se comentó que los datos son de un sistema de bicicletas compartidas en Los Ángeles, por lo tanto se espera que todas las estaciones sean de Los Ángeles. En el dataset cada estación tiene sus coordenadas (longitud y latitud) y existen algunas que no están dentro de las coordenadas de los Ángeles, sabiendo que las coordenadas de los ángeles son 34,05223 para latitud, $-118{,}24368$ para longitud. En este caso, las inconsistencias que hay en las coordenadas se presentan como *Outliers* y se pueden ver gráficalmente con los boxplots de la Figura 3.

Se puede lidiar con las inconsistencias de las coordenadas de dos formas: la primera es identificar la estación en cuestión, encontrar sus coordenadas correctas y añadirlas al conjunto de datos; la segunda es simplemente eliminar las filas que contienen

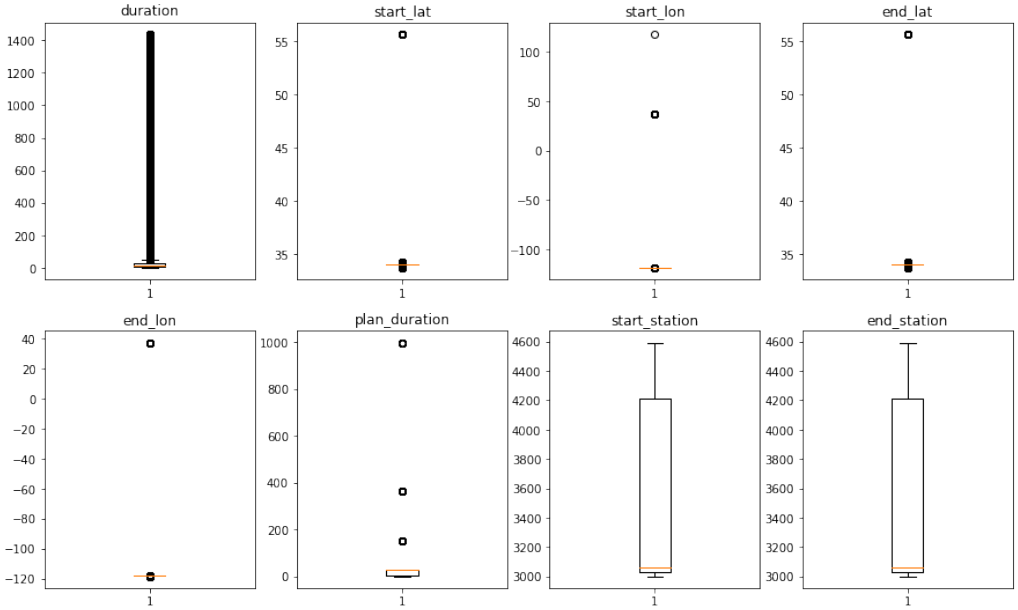

Figura 3: Visualización de outliers.

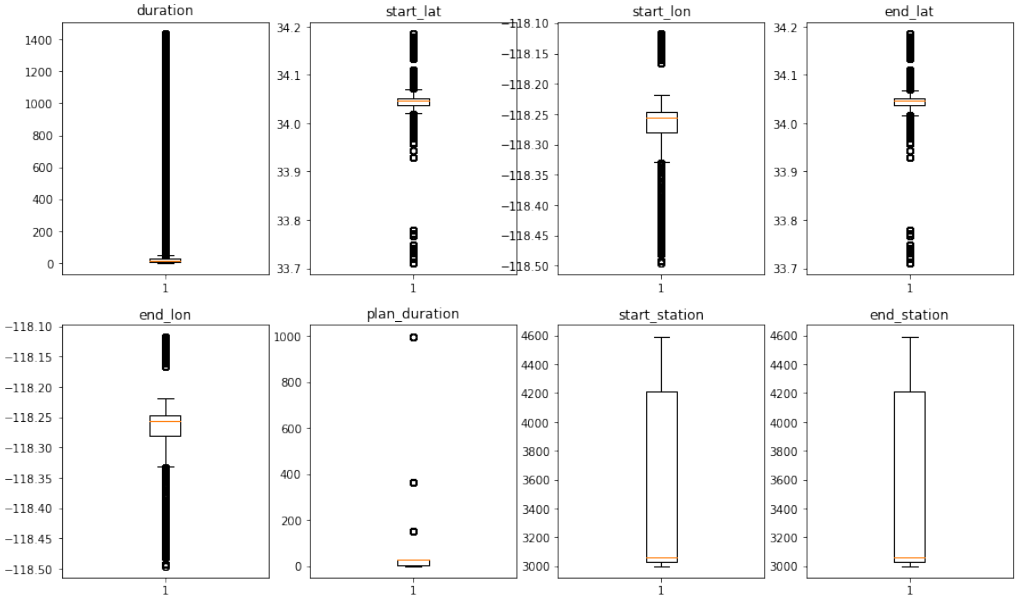

Figura 4: Visualización de outliers.

inconsistencias siendo consiente que eso no afecte los análisis que se realizarán. Son un total de 182 filas de 675, 626, eso no tendrá impacto negativo sobre los análisis, por lo tanto se procedió a eliminar del conjunto de datos los outliers. En la Figura 4 se puede ver la nueva gráfica con los boxplots después de eliminar los outliers de las columnas *start_lat*, *start_lon*, *end_lat*, *end_lon*.

2.4. Saturación del servicio

La empresa busca contar con la disponibilidad más alta de servicio en el mercado, por lo que quiere entender cómo se comporta la demanda para cada plan. Se agregaron 2 columnas más al conjunto de datos: *schedule* que tiene como valores *morning*, *afternoon*, *evening*, *night* y *year* que tiene como valores los años 2016, 2017, 2018, 2019, 2020, 2021 según corresponde a cada registro.

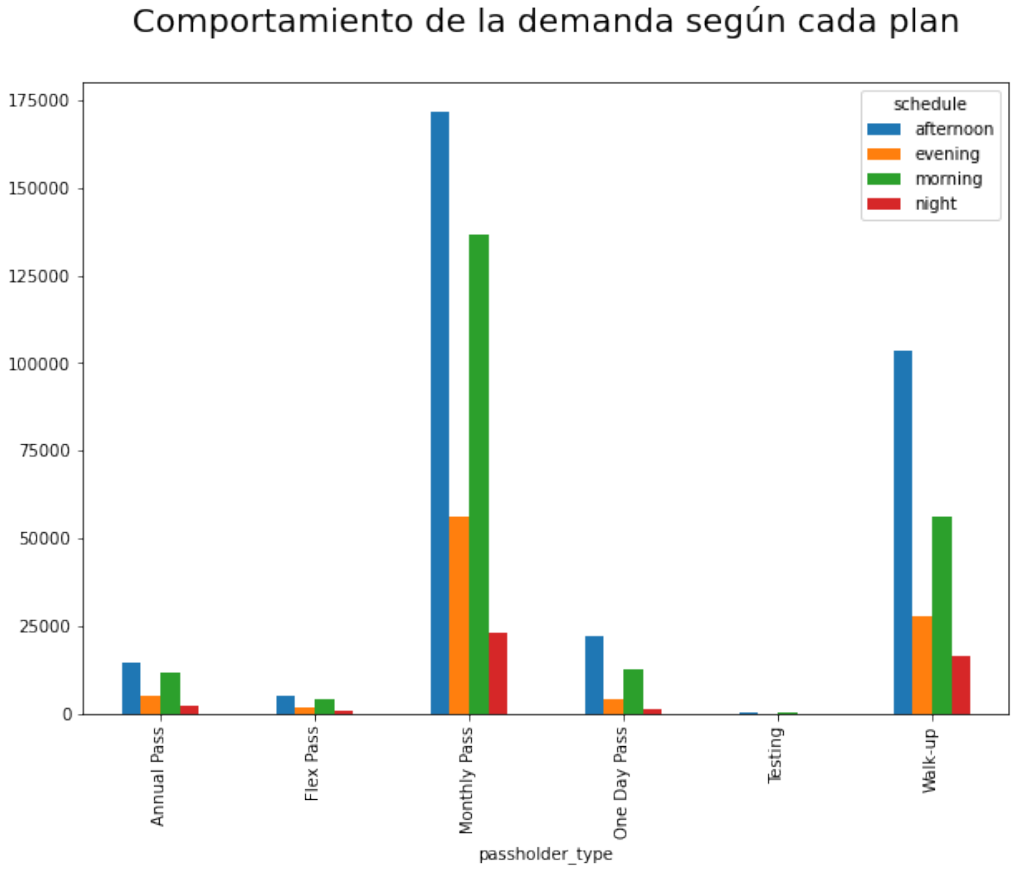

Figura 5: Uso del servicio por planes y horarios.

En la Figura 5 se puede ver que el plan con más demanda en el mercado es el de

abono mensual. Para cada plan, el horario que la gente más usa las bicicletas es después del medio día (de las 13h a las 18h).

2.5. Crecimiento de planes

Se tiene la intuición que la tendencia en uso de bicicletas compartidas entre estaciones va a la alta, por lo que se requiere realizar una correcta planificación de bicicletas que deben tener. Eso es parcialmente cierto, según la Figura 6 se puede concluir lo siguiente:

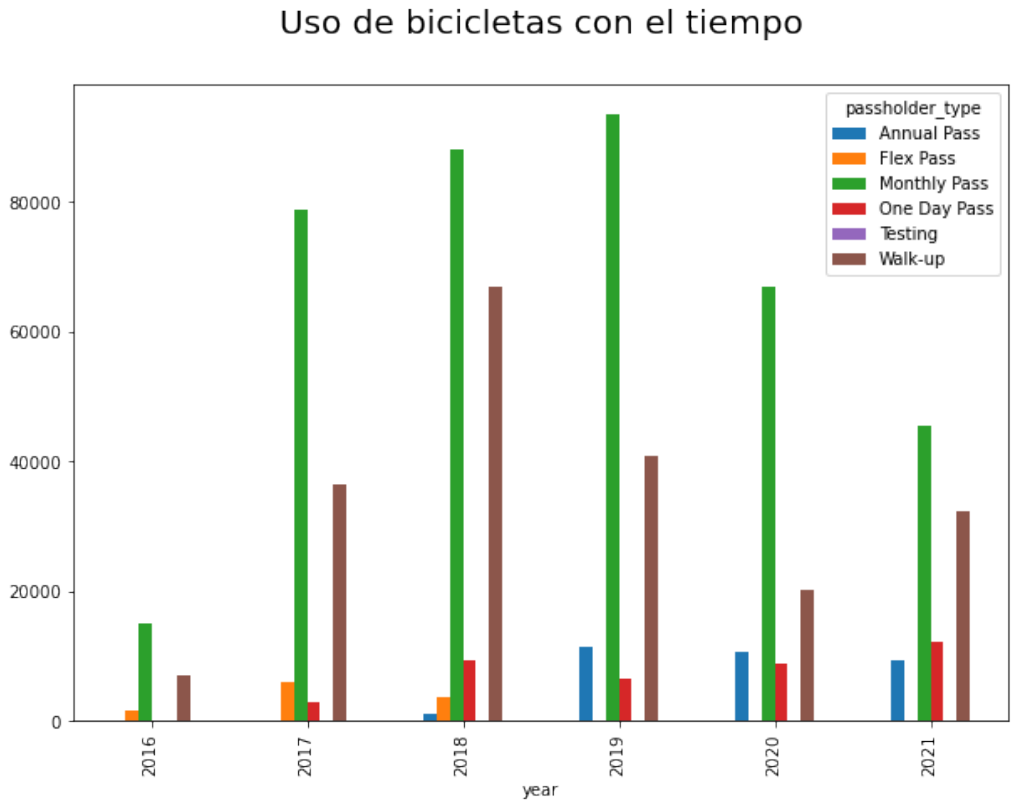

Figura 6: Abono y uso de los planes con el tiempo.

- El uso de bicicletas compartidas se fue de alza de 2016 a 2019. Sin embargo, después de 2019 empezó a disminuirse de manera considerable. Muchos factores pueden influir en la disminución del uso especialmente el inicio de la COVID-19. La pandemia es sin duda un factor importante en la disminución del uso de bicicletas compartidas ya que la mayoría de la gente se quedaron en casa, trabajan desde casa y no salieron si no hay una emergencia.

- Según la gráfica desde 2016 el plan que más se usa es el de consumo mensual $Monthly\ Pass$ representado por la barra verde de la gráfica.
- el segundo plan con más uso es el $Walk - up$.
- El plan $Flex\ Pass$ por un lado solo se ha usado de 2016 a 2018, su mayor uso fue en 2017. Su uso se ha desaparecido de 2019 a 2021.
- El plan *Anual Pass* ha tenido sus primeros usos en 2018, y se ha incrementado en 2019 y 2020. Su uso en 2021 ha sido menor que en 2019 y 2020, pero la diferencia es mínima.

3. Análisis predictivo de los datos

Se desea saber si es posible inferir el tipo de pase tomando en cuenta las demás variables de viaje.

3.1. Conversión de variables categóricas a numéricas

La primera tarea que se realizó antes de aplicar el modelo analítico a los datos fue la conversión de los valores categóricos a numéricos aplicando la técnica de $One - HotEncoding$. Las columnas a las que se aplicaron la técnica fueron *trip_route_category*, *passholder_type*, *bike_id*.

La variable a predecir es *passholder_type*, es una variable categórica con 6 valores diferentes: $Annual\ Pass$, $Flex\ Pass$, $Monthly\ Pass$, $One\ Day\ Pass$, $Testing$, $Walk-up$. Esos valores fueron codificados en 0,1,2,3,4,5 con One-Hot Encoding.

3.2. Análisis de correlación de Pearson entre las variables

El análisis de correlación de Pearson permite obtener el grado de linealidad que existe entre dos variables. Una variable depende de otra si su índice de correlación se acerca a -1 o 1. El mapa de calor de la Figura 7 muestra que el tipo de plan está fuertemente relacionado con la duración del plan (*plan_duration*) y paracialmente relacionado con la categoría del viaje (*trip_route_category*).

Después del análisis, se consideran que las siguientes variables son las más impactadas para la predicción:

- *trip_route_category*.
- *duration*.
- *start_station*

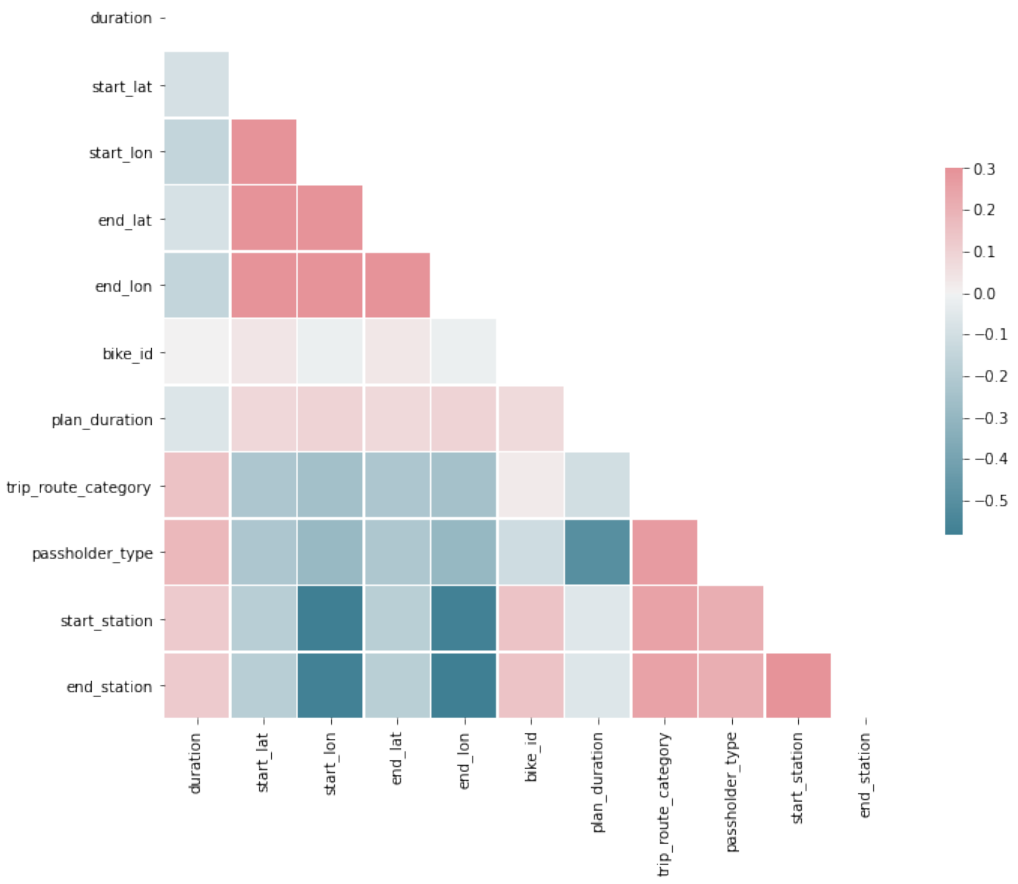

Figura 7: Correlación de Pearson entre las variables.

- *end_station*

3.3. División de los datos en entrenamiento y prueba

El conjunto de datos fue dividido entre entrenamiento y prueba, %80 para entrenamiento (540, 355 de los datos) y %20 para pruebas (135, 089 de los datos).

3.4. Aplicación de Random Forest

Se aplicó 10-Fold Cross Validation para evaluar el rendimiento de Random Forest. El accuracy del modelo es de % 68.

Para mejorar la predicción, se tomó en cuenta una nueva variable: la variable *year*. Esa variable fue agregada previamente al conjunto de datos para el análisis ex-

ploratorio. El modelo fue entrenado y probado nuevamente usando 10-Fold Cross Validation y el accuracy aumentó a % 70

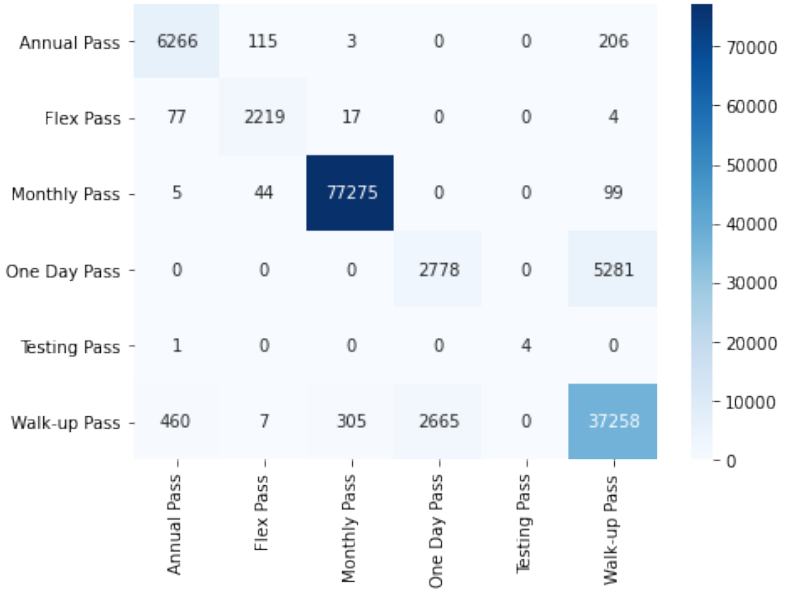

Figura 8: Matriz de confusión.

La nueva matriz de confusión obtenida al agregar la nueva columna se puede ver en la Figura 8.

4. CONCLUSIÓN

Para concluir, se puede decir que las variables utilizadas para predecir el plan de consumo del cliente no son suficientes. Para mejorar el rendimiento del modelo, hay que usar otras variables cómo la duración del plan de consumo, las promociones si es que hayan, el tipo de usuario etc.

EL modelo puede desplegarse en cualquier plataforma para su consumo por medio de API. En las Figuras **??** se puede ver un workflow de MLOps.

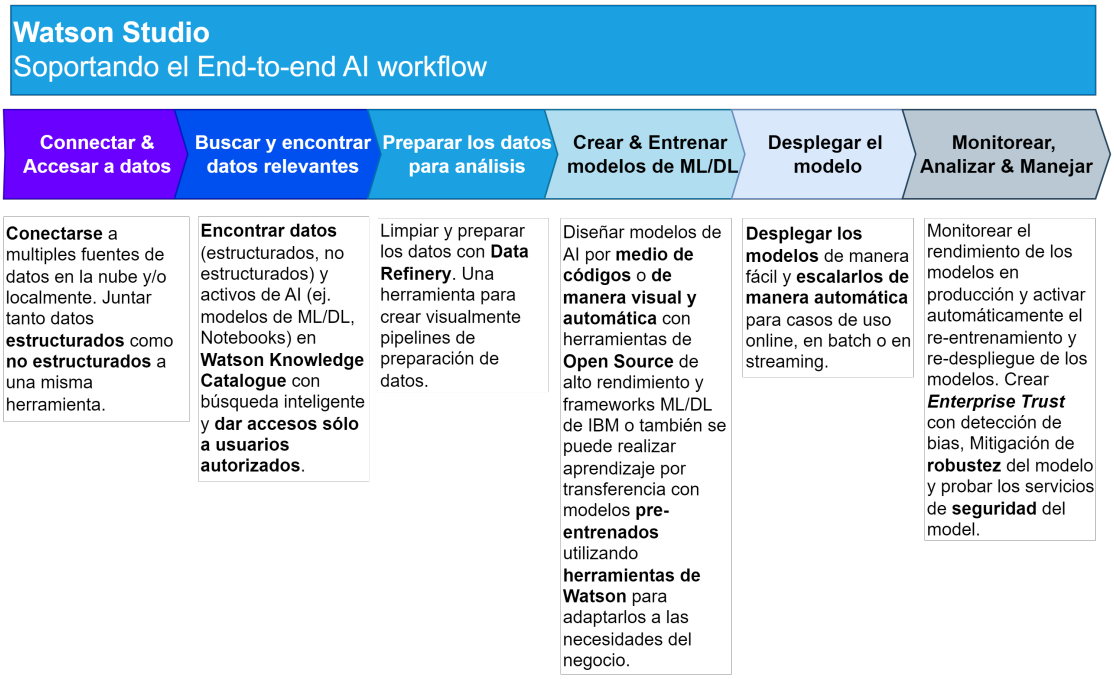

Figura 9: End-to-end AI Workflow.

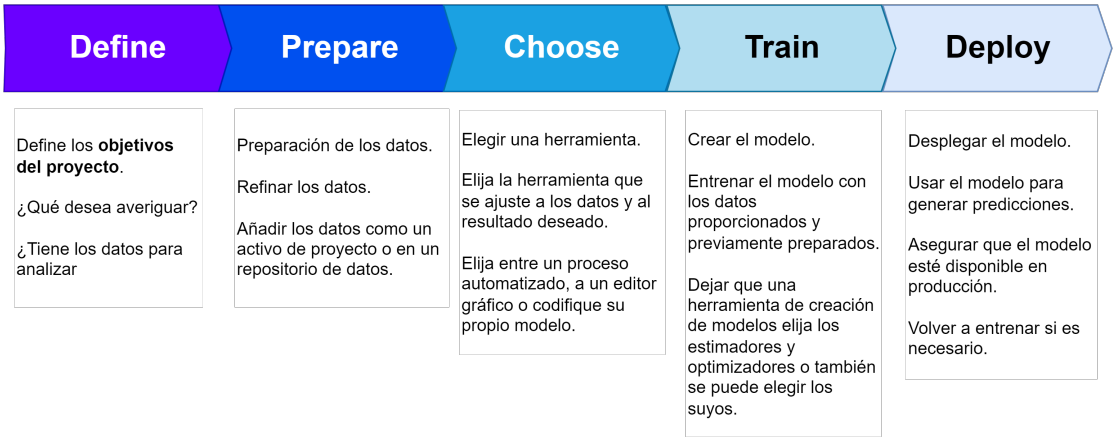

Figura 10: Workflow de MLOps.