TFM: Predicción de la duración del tiempo de viaje en un taxi en la ciudad de New York

Master-Data-Science

Autor: Michael Quintana

La idea principal del proyecto es determinar si se puede estimar la duración del tiempo de viaje en un taxi en la ciudad de New York a partir de ciertas características del viaje como son: distancia del trayecto, zona inicial del trayecto, zona final del trayecto, hora en la que se tomo el viaje, etc. Con la finalidad de conocer anticipadamente el costo que implicara realizar dicho viaje, asimismo lograr mayor eficiencia en los viajes que se realizan diariamente.

TLC – Taxi and Limousine Commission

La comision de taxis y limosinas en New York, es una agencia encomendada por la Carta, cuyo objetivo es la continuación, el desarrollo y la mejora del servicio de taxis y de alquiler en la Ciudad de New York. Dicha entidad tiene diferentes reglas, leyes, órdenes, decisiones u otras fuentes legales que se debe cumplir dentro de la ciudad.

TLC otorga licencias a más de 130,000 vehículos en la ciudad de Nueva York. Cada vehículo recibe inspecciones exhaustivas de seguridad y emisiones por TLC y debe ser conducido por conductores con licencia de TLC que se hayan sometido a una verificación de antecedentes y hayan aprobado los requisitos de educación de TLC.

- Los taxis verdes brindan servicio de granizo en la calle y servicio preestablecido en el norte de Manhattan (sobre E 96th St y W 110th St) y en los distritos exteriores.
- Los taxis verdes cobran tarifas medidas estándar para todos los viajes de granizo callejero. El precio de los viajes preestablecidos se establece por la base o aplicación de teléfono inteligente utilizada para reservar el viaje.
- Los taxis verdes se identifican fácilmente por su color verde, las marcas de taxi en “T” y los números de licencia en el techo y los lados del vehículo.

La web de dicha entidad para mayor información: https://www1.nyc.gov/site/tlc/about/about-tlc.page

Los datos con los que se trabajo en dicho proyecto se descargaron de dicha web: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page, estos están agrupados por año, por mes y por tipo de vehiculo, Especificamente para este proyecto se esta considerando el año 2018 y viajes en taxi verde.

Descripcion de los datos

Se ha descargado los datos del año 2018, entre las variables que hay son:

VendorID: Un código que indica el proveedor de LPEP que proporcionó el registro.
1 = Creative Mobile Technologies, LLC;
2 = VeriFone Inc.

lpep_pickup_datetime: La fecha y hora en que inicio el viaje.
lpep_dropoff_datetime: La fecha y hora en que finalizo el viaje.

Passenger_count: El número de pasajeros en el vehículo.

Trip_distance: La distancia de viaje transcurrida en millas reportada por el taxímetro.

PULocationID: TLC Zona de Taxi en la que inicio el taxímetro.

DOLocationID: TLC Zona de Taxi en la que finalizo el taxímetro.

RateCodeID: El código de tarifa final vigente al final del viaje:
1 = tasa estándar
2 = JFK
3 = Newark
4 = Nassau o Westchester
5 = tarifa negociada
6 = Paseo en grupo

Store_and_fwd_flag: Esta bandera indica si el registro de viaje se llevó a cabo en el vehículo memoria antes de enviar al proveedor, también conocido como "almacenar y reenviar", porque el vehículo no tenía conexión con el servidor.
Y = tienda y viaje de ida
N = no es una tienda y viaje de ida

Payment_type: Un código numérico que indica cómo pagó el pasajero por el viaje.
1 = tarjeta de crédito
2 = efectivo
3 = Sin cargo
4 = Disputa
5 = Desconocido
6 = viaje vacío

Fare_amount: La tarifa de tiempo y distancia calculada por el medidor.

Extra: Extras y recargos varios. Actualmente, esto solo incluye los cargos de $ 0.50 y $ 1 por hora pico y por la noche.

MTA_tax: $ 0.50 de impuestos MTA que se activan automáticamente en función del medidor tarifa en uso.

Improvement_surcharge: Recargo por mejora de $ 0.30 en viajes aclamados en la bandera soltar. El recargo por mejora comenzó a percibirse en 2015.

Tip_amount: Cantidad de propina: este campo se rellena automáticamente para tarjeta de crédito

Tolls_amount: Importe total de todos los peajes pagados en viaje.

Total_amount: El importe total cobrado a los pasajeros. No incluye propinas en efectivo.

Metodologia

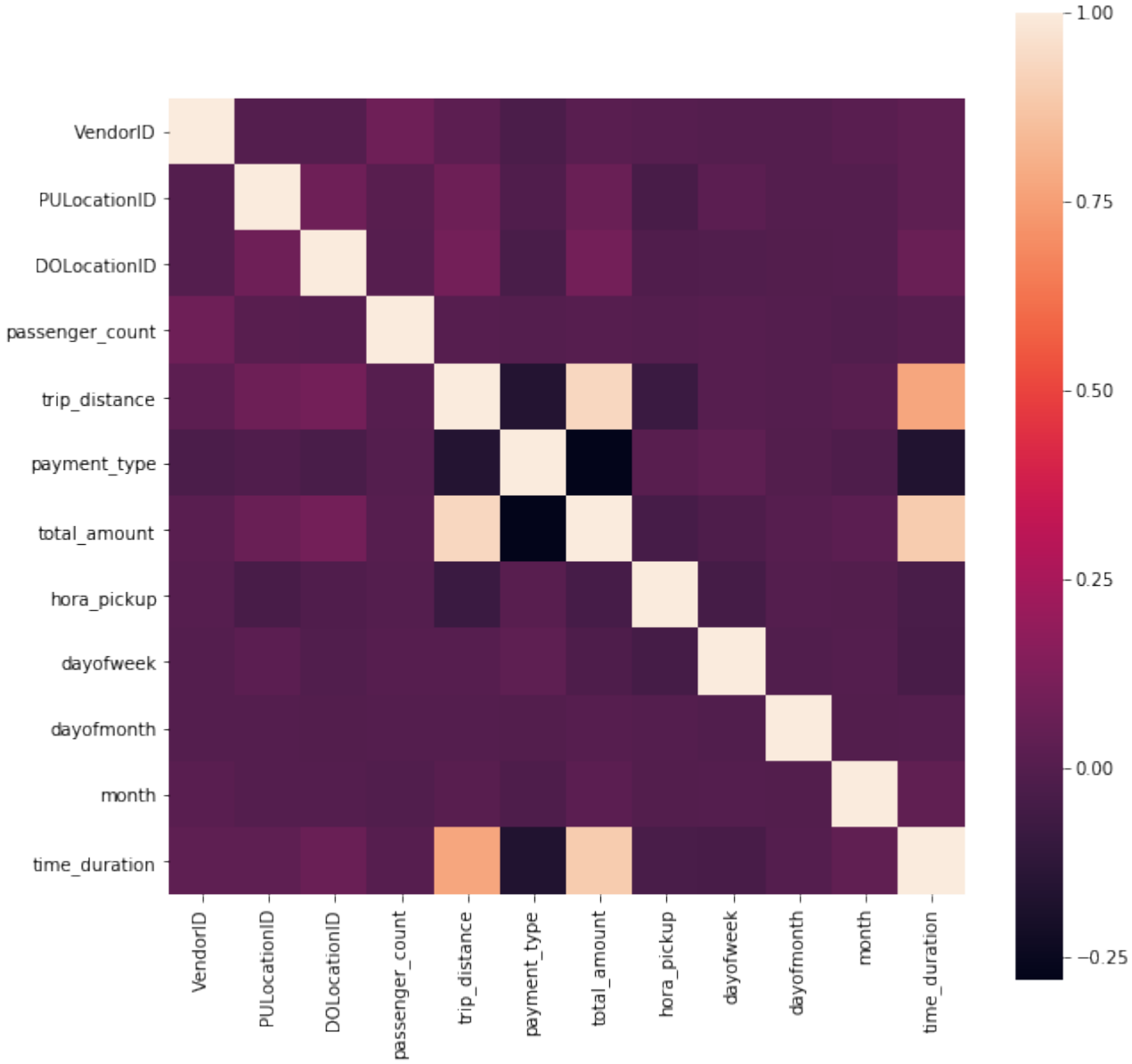
1 **Carga de datos:** Se ha cargado los datos desde un archivo .csv (DATOS_TFM.csv)**.**

2 **Tratamiento de datos:** Esta parte se ha dividido en subpartes:

Creación del target: A partir de los campos fecha inicial de viaje y fecha final del viaje se ha calculado el tiempo de duración del viaje, asimismo se considerara para el proyecto el intervalo desde 4 minutos hasta 3 horas de duracion de viaje.

Limpieza de variables: se han revisado las variables store_and_fwd_flag, RateCodeID, passenger_count, payment_type y filtrado registros con poca información de cada una de ellas, para el resto de variables se puede observar su distribución. En el caso de la variable trip_distance se ha considerado el intervalo entre 500 metros hasta 12 kilometros de distancia de viaje.

Selección y correlación de variables:



Escalado de variables: Como tenemos diferentes variables numericas y en diferente escala, se ha estandarizado la escala para todas las variables, esto con la finalidad de obtener mejores resultados.

3 Particion de los datos: Se ha utilizado la librería model_selection – train_test_split que particiona los datos de manera aleatoria, esto con la finalidad de entrenar el modelo con la parte train y validar los resultados del modelo con la parte test.

4 Modelamiento de los datos: Esta etapa se ha realizado en 3 partes:

Modelos Básicos: Se probo los modelos de regresión lineal y sus variantes Lasso y Ridge, los resultados de estos se muestran a continuación:

Modelo	mean_squared_error	mean_absolute_error	mean_absolute_percentaje_error	correlation_target_prediccion
Regresion Lineal	111152.1107	228.8491	30.2038	0.7771
Lasso	111175.5925	228.8782	30.2316	0.7770
Ridge	111153.0357	229.0062	30.2713	0.7771

De acuerdo a los resultados observamos que el que tiene mejores métricas es la regresion lineal ya que el error es el menor de los tres.

Modelos Ensamblados: Se probo los modelos arboles ensamblados GBM y RANDOMFOREST, los resultados de estos se muestran a continuación:

Modelo	mean_squared_error	mean_absolute_error	mean_absolute_percentaje_error	correlation_target_prediccion
GBM	78385.0978	186.3683	23.2331	0.8493
RandomForest	74708.8962	182.7290	22.8570	0.8574

De acuerdo a los resultados observamos estos modelos obtienen mejores resultados que los modelos basicos probados antes, entre los dos el que tiene mejores métricas es Randomforest ya que el error es menos que GBM.

LightGBM: Se prueba este modelo con la finalidad de tunear los parametros mas importantes ya que el procesamiento de este es mucho mas rapido que los anteriores, y los resultados se muestran a continuacion:

Modelo	mean_squared_error	mean_absolute_error	mean_absolute_percentaje_error	correlation_target_prediccion
LightGBM	60677.7650	163.2641	20.3337	0.8854

Observamos finalmente que las métricas de este modelo son incluso mejores que GBM y RANDOMFOREST, por lo que nos quedaremos con este ultimo modelo con sus parámetros tuneados.

5 Importancia de variables:

Observamos que las principales variables son la zona inicial de trayecto, la zona final de trayecto, la distancia del trayecto, la hora en la que inicia el trayecto.

6 Evaluación de las predicciones: Aqui evaluamos el error que tienen las predicciones:

	range_dif
0.00	0.000527
0.10	18.873522
0.20	37.967869
0.30	57.834485
0.40	79.654525
0.50	105.001794
0.60	136.756686
0.70	179.977847
0.80	247.299757
0.95	517.023660
1.00	3220.862271

Esta tabla muestra segun el porcentaje de los datos cuantos segundos falla el modelo, es decir en el 70% de los datos el error aproximado es de +-3 minutos (179.97 segundos).

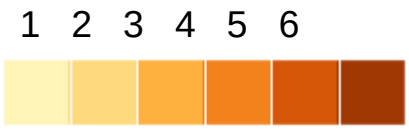
	range_dif_real
0.00	-3220.862271
0.10	-259.471695
0.15	-175.488691
0.30	-55.531575
0.40	-10.145611
0.50	26.069725

	range_dif_real
0.60	59.329183
0.70	95.715358
0.85	182.873865
0.90	238.821231
1.00	1855.208280

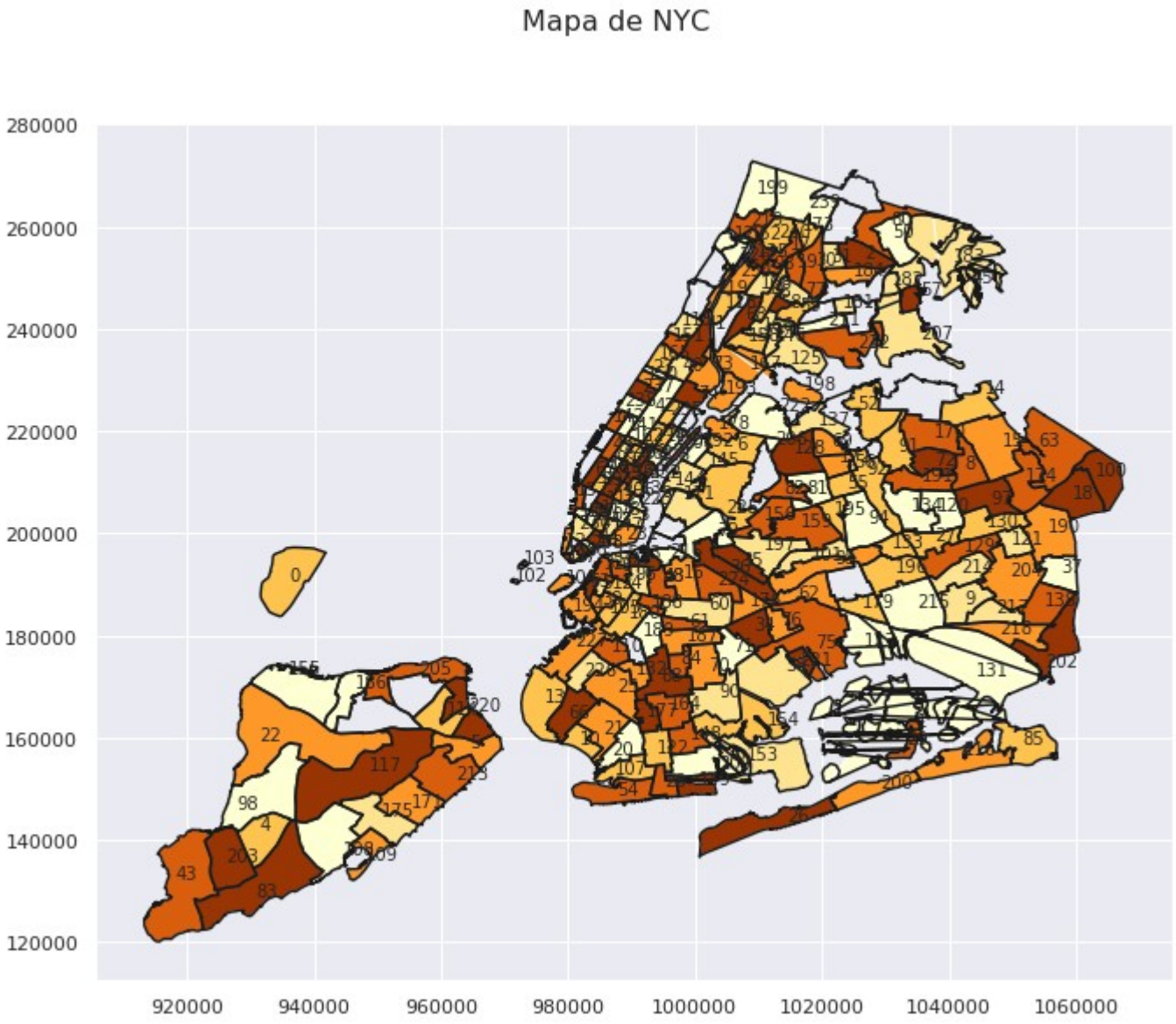
En esta tabla observamos que el error subestimado en 15% de los datos es menor a 3 minutos y el error sobreestimado 15% de los datos es mayor a 3 minutos.

Visualizacion de las zonas

Se tiene el archivo POLYGON_ZONE.shp que contiene las coordenadas de las zonas que realizan viajes.




```
<Figure size 792x648 with 0 Axes>
```



Podemos observar que las zonas con el color mas fuerte indican aquellas donde hay mayor recaudación de dinero.

Conclusiones Finales:

Finalmente podemos concluir con los resultados del proyecto, que el tiempo de duración de los viajes que realizan los taxis verdes en la ciudad de New York si dependen de las características planteadas al inicio del proyecto. Es decir que el tiempo de duración de los viajes en New York depende de la zona inicial del viaje, la zona final del viaje, la distancia del viaje, la hora en la que se tomo el viaje. Esto se evidencia en el grafico de importancia de variables del modelo final. Asimismo podemos plantear con estos resultados mejoras en la eficiencia de los viajes tales como:

- Plantear costos de viaje acorde al tiempo de duración a priori.

- Plantear una mejora en los costos de viaje en las zonas iniciales que tienen mayor o menor demanda de viajes y que los tiempos de viaje sean cortos.

- Plantear mejoras en la ruta de viaje, esto con la finalidad de disminuir el tiempo de viaje.
- Aumentar la distribución de taxis en las rutas de viaje con mayor tiempo de duración.
- Mejorar la calidad de atención en las rutas con mayor tiempo de viaje.