Estimación de la duración promedio de un viaje en taxi amarillo en la ciudad de Nueva York

Minería de datos para grandes volúmenes de información

David Betancur Londoño
Diego Andrés Jaramillo Zapata
Susana Londoño Muñoz
Catalina Piedrahita Jaramillo
Diego Andrés Valderrama Laverde

Pregunta de Investigación

Nuestro cliente es la *Comisión de Taxis y Limosinas de la Ciudad de Nueva York* (TLC por sus siglas en inglés), que tiene la necesidad de asignar, de manera eficiente, los taxis a los pasajeros para que el servicio sea fluido y sin complicaciones. Para esto, es clave conocer la duración de viaje de cada taxi, con el fin de predecir cuándo estará libre y, así, asignar el próximo viaje.

Por lo tanto, nuestra pregunta de investigación fue:

¿Cuál es la duración promedio estimada de un viaje en taxi amarillo en la ciudad de Nueva York?

Objetivos

General:
Desarrollar un modelo de aprendizaje de máquina en el contexto de Big Data para predecir la duración promedio de un viaje en taxi amarillo en la ciudad de Nueva York.

Específicos:

- Diseñar una arquitectura en AWS para el procesamiento y análisis de datos.
- Realizar ingeniería de características para determinar las principales variables que se considerarán en el modelo.
- Evaluar algunos modelos de aprendizaje de máquina para llevar a cabo una estimación de la duración promedio de viaje en taxi amarillo en la ciudad de Nueva York.

Estado del arte

La exploración inicial que se realizó fue buscar retos o documentación, donde se hubiese utilizado total o parcialmente el dataset seleccionado. En este proceso, se encontró un reto de Kaggle titulado: "*New York City Taxi Trip Duration. Share code and data to improve ride time predictions"*. Donde se pretendía estimar la variable respuesta del tiempo de viaje, aquí se observaron algunas estrategias en relación con la exploración inicial de los datos y se encontró que en gran medida los participantes utilizaron regresiones lineales para predecir la variable.
Asimismo, se encontró en Medium un documento titulado "*Linear Regression Model on the New York Taxi Trip Duration Dataset using Python*" en el cual se realizó un análisis univariante de algunas de las variables y bivariante de algunas otras con respecto a la variable objetivo -duración del viaje-. Se utilizaron algunos métodos allí presentados para identificar valores atípicos, además se imputaron valores que estuvieran por fuera de los valores promedio.

Otro punto de referencia fue el artículo "*Building A Linear Regression with PySpark and MLlib*" también del portal *Towards Data Science* donde se encontraron recomendaciones para construir modelos con PySpark de regresión lineal, gradient-boosted trees regression y decision tree regresion que motivaron la exploración del random forest.

En general es de destacar que diversos puntos de referencia para determinar las estrategias de exploración del conjunto de datos, creación de variables y código de modelos se obtuvieron mediante la exploración de artículos de *Medium* y *Towards Data Science,* blogs que se han posicionado en estos años como punto de referencia para el intercambio de conocimiento e ideas en el área de ciencia de datos.

En los diversos artículos leídos se destaca el uso de librería de Pandas de Python para realizar la exploración de datos, pero dado que esta no soporta grandes volúmenes, se definió tomar una muestra para realizar este proceso con las diversas opciones que ofrece esta librería. En ese sentido, detectada la limitación de Pandas, se utilizó PySpark para la creación de modelos, con su librería MLlib, que de acuerdo con los diversos artículos es la de mayor uso actual.

Metodología de Investigación

Fase I. Definición de necesidades del cliente (comprensión del negocio):

Para una empresa de taxis, la movilización de sus vehículos en una ciudad grande y compleja como Nueva York implica grandes retos, en especial, el tomar la decisión más acertada para lograr el mayor número de viajes, o mejorar la eficiencia de los

sistemas de despacho de taxi, para definir la presencia en el lugar y momento correctos.

El análisis de los tiempos de viaje puede ayudar al sector de los taxis a determinar planes de acción para programar de manera más eficiente los viajes, optimizando la cantidad de viajes posibles por día y la asignación de despachos, garantizando unos mejores niveles de satisfacción por parte de los usuarios.

Fase II. Estudio y comprensión de los datos:
En este trabajo se usaron los datos recolectados por TLC. Este proyecto se centró solo en los taxis tipo amarillo, además, se estableció como rango de años, 2018 y 2019, considerando un tamaño máximo de los datos de 20 GB, equivalente a más de 180 millones de registros.

Las principales actividades desarrolladas fueron:

- Análisis exploratorio de los datos, para determinar y comprender las variables disponibles, la relación entre ellas y con la pregunta de investigación.
- Limpieza de datos nulos.
- Validación de la coherencia de los datos, por ejemplo, que no se presentaran tiempos o cobros en cero, negativos o números muy grandes con respecto a los demás.

Fase III. Análisis de los datos y selección de características:
Se procedió de la siguiente manera:

- Se realizaron histogramas para identificar puntos de corte y eliminar datos atípicos
- Codificación de las variables categóricas.
- Generación de características.
- Implementación de técnicas para la selección de características: análisis de correlación, regresión lineal múltiple de todas las variables y porcentaje de influencia de las variables con *random forest*.

Fase IV. Modelado
A partir del estado del arte, se implementaron los siguientes modelos:

Figura 1: modelos entrenados para el proyecto.

Los cuales, en su modelación inicial, permitieron seleccionar las variables de mayor incidencia y representatividad en términos de explicación e influencia.

Fase V. Evaluación:
Realizamos la evaluación de los modelos con las siguientes métricas:

- Coeficiente de determinación- R2.
- RMSE (Raíz del Error Cuadrático Medio).
- MAE (Error Porcentual Absoluto Medio).

Estudio y comprensión de los datos

Descripción de los campos:

Nombre del campo	Descripción	Tipo
VendorID	Cod_proveedor. Un código que indica el proveedor de TPEP (Passenger Enhancement Programs) que proporcionó el registro. 1= Creative Mobile Technologies, LLC 2= VeriFone Inc.	IntegerType
tpep_pickup_datetime	Fecha_hora_inicio_tpep. La fecha y la hora en la que el medidor (del tpep) fue iniciado.	TimestampType
tpep_dropoff_datetime	Fecha_hora_fin_tpep. La fecha y la hora en que el medidor (del tpep) fue parado.	TimestampType
Passenger_count	Cont_pasajero. El número de pasajeros del vehículo. Este es un valor introducido por el conductor.	IntegerType
Trip_distance	Distancia_viaje. La distancia de viaje transcurrida en millas reportada por el taxímetro.	DoubleType
PULocationID	Cod_ubicacion_recogida. Zona de taxis TLC (Comisión de Taxis y Limosinas) en la que se contrató el taxímetro.	StringType
DOLocationID	Cod_ubicacion_descarga. Zona de taxis TLC en la que se desconectó el taxímetro	StringType
RateCodeID	Cod_tarifa. El código de tarifa final en uso al final del viaje. 1 = Tarifa estándar 2 = JFK 3 = Newark 4 = Nassau o Westchester 5 = Tarifa negociada 6 = Viaje en grupo	StringType
Store_and_fwd_flag	Bandera_almac_reenv. Esta bandera indica si el registro de viaje se mantuvo (local) en la memoria del vehículo antes de enviarlo al proveedor, también conocido como "store and forward", porque el vehículo no tenía conexión con el servidor. Y= viaje almacenamiento y reenvío N= no es un viaje de almacenamiento y reenvío	StringType
Payment_type	Tipo_pago. Un código numérico que indica cómo pagó el viaje el pasajero. 1 = Tarjeta de crédito 2 = Dinero en efectivo 3 = Sin cargo 4 = Disputa 5 = Desconocido 6 = Viaje anulado	StringType
Fare_amount	Cantidad_tarifa. La tarifa de tiempo y distancia calculada por el contador.	DoubleType
Extra	Extra. Extras y recargos diversos. Actualmente, esto sólo incluye los cargos de $0,50 dólares y 1 dólar por hora pico y recargo nocturno	DoubleType

Nombre del campo	Descripción	Tipo
MTA_tax	Impuesto_MTA. Tasa de $0,50 dólares de la MTA (Metropolitan commuter transportation mobility tax) que se activa automáticamente en función de la tarifa del contador en uso.	DoubleType
Improvement_surcharge	Banderazo. Recargo de mejora de $0,30 dólares aplicado a los viajes en la bajada de bandera. El recargo de mejora comenzó a cobrarse en 2015.	DoubleType
Tip_amount	Propina. Importe de la propina - Este campo se rellena automáticamente para las propinas con tarjeta de crédito. No se incluyen las propinas en efectivo.	DoubleType
Tolls_amount	Peajes. Importe total de todos los peajes pagados en el viaje.	DoubleType
Total_amount	Cobro_total. El importe total cobrado a los pasajeros. No incluye las propinas en efectivo.	DoubleType

Tabla 1: tabla de descripción de los campos del conjunto de datos.

Tamaño en memoria de los datos: 15,9 GB. Los datos se encontraban en formato CSV, donde cada archivo era un mes de un año determinado. En total, se tuvieron 24 archivos y 187,2 millones de registros.

Para el análisis exploratorio se consideró una muestra aleatoria correspondiente a 1,87 millones de registros. A continuación, se presenta el análisis realizado:

- Análisis gráfico de las variables. Esto permitió identificar puntos atípicos.
- Cálculo de la variable respuesta (duración del viaje), a partir de la variable fecha hora de recogida y fecha hora de descarga del pasajero.
- Eliminación de algunos registros a partir de la distribución de la variable respuesta.

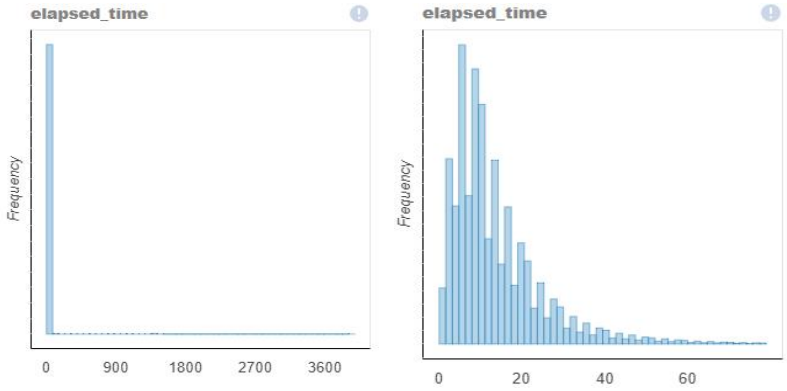


Figura 2: histogramas antes y después de eliminar valores atípicos en la variable respuesta.

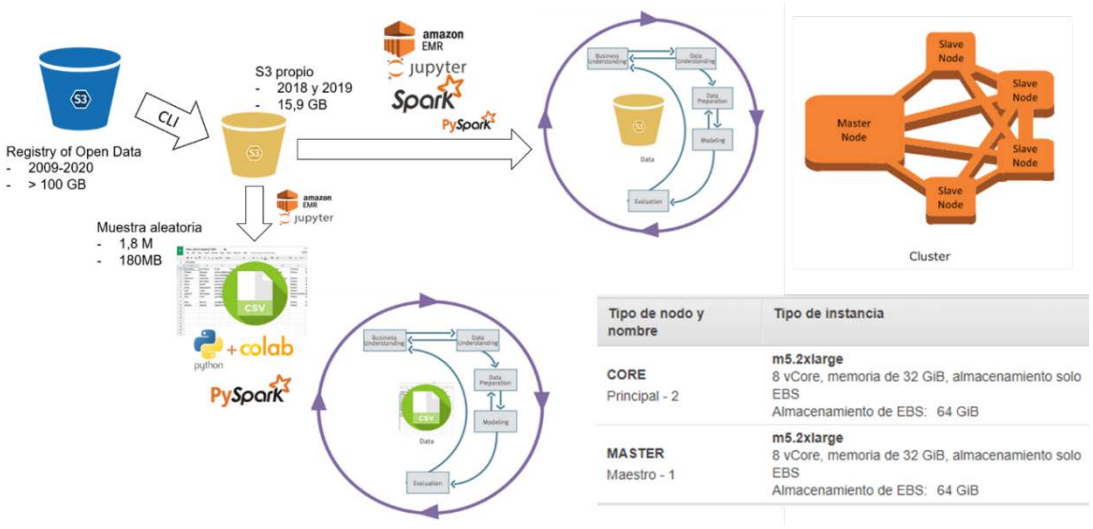
- A partir de la fecha, se codificaron las variables *hora* y *día* de la semana.

Arquitectura

Los datos utilizados están disponibles para uso público, y dispuestos en *Registry of Open Data on AWS* (https://registry.opendata.aws/nyc-tlc-trip-records-pds/). Fueron ingeridos por medio de la interfaz de línea de comandos a un *bucket* en S3 propio (s3://taxidatamineria/yellow_trip_data/) solo con los datos que nos interesaban (años 2018 y 2019), para un total de 15,9 GB.
Teniendo acceso a los datos, se creó un *clúster* en EMR, y se inició un *notebook* con el que se obtuvo una muestra del 1% de los datos (1,87 millones de registros, 180MB). Este conjunto de datos pudo ser manejado en una carpeta compartida de *Google drive,* que permitió realizar todo el proceso de análisis exploratorio de datos, preparación de datos, modelación y evaluación en Google *Colaboratory* de manera colaborativa y con el uso de librerías como Pandas y PySpark.

Luego de tener claridad de cómo se iba a realizar la preparación de los datos (eliminación de valores nulos y datos atípicos), validar su funcionamiento en el conjunto de datos y la certeza de que los modelos entrenados con la muestra tenían un buen rendimiento, se procedió a realizar los mismos procedimientos con todo el conjunto de datos (>180 millones de registros, 15,9 GB).

Para lograr ejecutar los procesos de preparación de datos, modelación y evaluación en un conjunto de datos de grandes dimensiones, se exploraron diferentes tipos de instancias en AWS EC2, tras leer los tipos de instancias (referencia) y cuáles están disponibles para las cuentas AWS *Educate*, la que más se ajustó a los requerimientos fue la m5.2xlarge, con 32GiB de memoria y 8 núcleos, permitiendo procesar todos los datos en un tiempo sensato. El clúster se creó con 1 nodo maestro y 4 nodos esclavos.



Tipo de nodo y nombre	Tipo de instancia
CORE Principal - 2	m5.2xlarge 8 vCore, memoria de 32 GiB, almacenamiento solo EBS Almacenamiento de EBS: 64 GiB
MASTER Maestro - 1	m5.2xlarge 8 vCore, memoria de 32 GiB, almacenamiento solo EBS Almacenamiento de EBS: 64 GiB

Figura 3: diagrama de la arquitectura utilizada para la realización del proyecto. Se incluyen los diferentes ambientes (Colaboratory y AWS EMR).

Finalmente se creó un repositorio en GitHub en el cual se encuentran los códigos utilizados y una muestra pequeña del conjunto de datos utilizado.

Enlace al repositorio: https://github.com/SusanaLondono/NYC_TLC_CDDDS

Modelos y evaluación

Al realizar la regresión lineal múltiple, donde se modeló la variable dependiente a partir de las 26 variables disponibles, se identificaron solo aquellas que fueron significativas en la modelación. Seguidamente, se corrió el modelo solo con estas variables significativas.

De igual forma, utilizando estas variables que fueron significativas, se corrió el modelo de *Gradient Boosted Trees*.

En el caso de la modelación de la regresión con *random forest*, se obtuvo el siguiente resultado en términos de la importancia de cada variable:

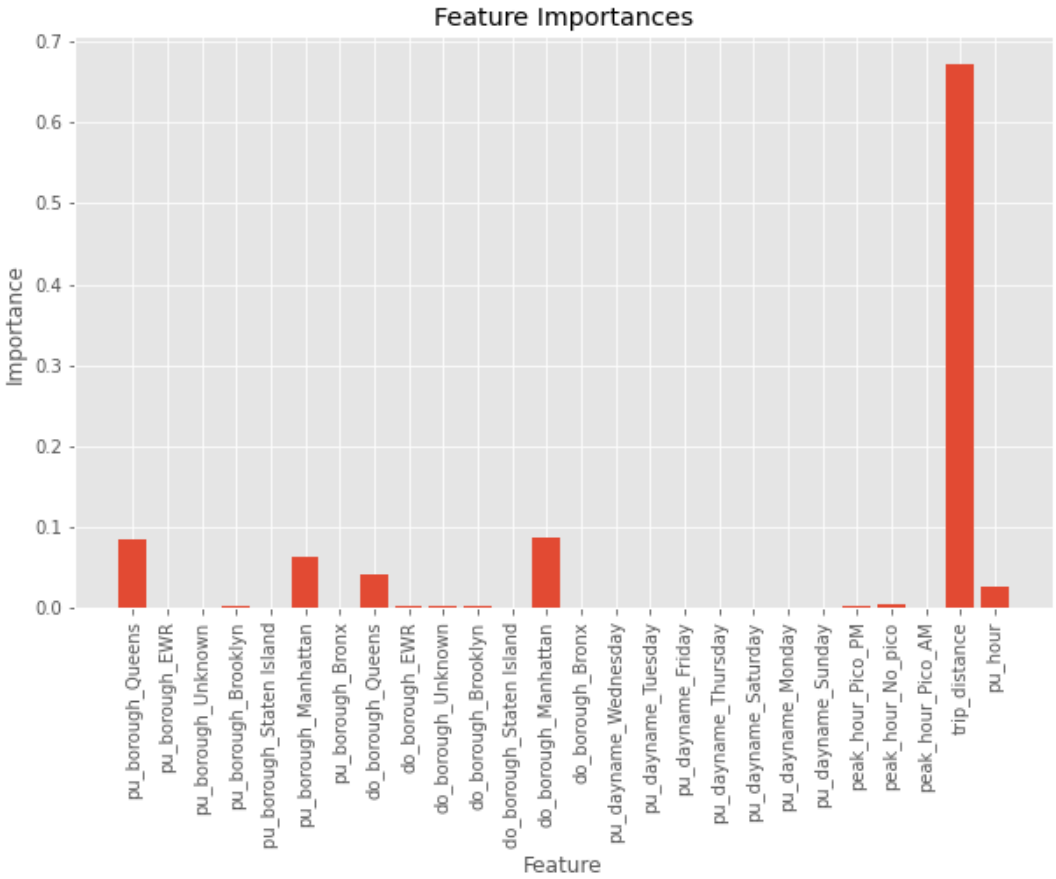


Figura 4: importancia de las variables según el modelo de random forest.

De esta manera, con una modelación con 20 árboles y 26 variables, solo 11 presentaron una influencia en la explicación de la variable dependiente.

Posteriormente, con estas variables, nuevamente se corrió el modelo, buscando tener una mayor eficiencia computacional y una leve afectación sobre los criterios de evaluación.

Los resultados obtenidos con la muestra son los que se presentan a continuación:

Modelo	Cant. de variables	R2	RMSE	MAE
Regresión lineal	9	0.64	6.51	4.63
Gradient boosted tree regression	9	0.69	5.97	4.04
Random Forest	11	0.69	6.00	4.20

Tabla 2: resultados obtenidos en los modelos con la muestra (1.8 millones de registros)

Finalmente, estos modelos fueron aplicados al conjunto total de los datos, los 180 millones de registros, donde se observó que los resultados fueron similares a los obtenidos con la muestra, con una leve disminución en el ajuste para el caso de la regresión lineal múltiple. Es de destacar que los siguientes resultados son las métricas obtenidas sobre el conjunto de prueba (test).

Modelo	Variables	R2	RMSE	MAE
Regresión Lineal	'pu_borough_Queens', 'do_borough_EWR', 'do_borough_Unknown', 'do_borough_Brooklyn', 'pu_dayname_Thursday', 'pu_dayname_Monday', 'pu_dayname_Sunday', 'peak_hour_No_pico', 'trip_distance'	0.63	6.56	4.67
Gradient Boosted Tree Regression		0.698	5.97	4.04
Random Forest	'pu_borough_Queens', "pu_borough_Manhattan", "do_borough_Queens", 'do_borough_EWR', 'do_borough_Brooklyn', "do_borough_Manhattan", 'pu_dayname_Sunday', "peak_hour_Pico_PM", 'peak_hour_No_pico', "pu_hour", 'trip_distance'	0.698	5.94	4.12

Tabla 3: resultados obtenidos en los modelos el total de datos (>180 millones de registros).

En términos generales, los modelos de *Gradient Boosted Tree* y *Random Forest*, fueron los de mejor ajuste, siendo este último el más sobresaliente. Es importante considerar que dentro de las bondades de estos modelos se encuentran su flexibilidad y que están pensados para grandes volúmenes de datos. Sin embargo, su interpretación puede ser compleja ante la presencia de una cantidad importante de árboles, que puede limitar su interpretación.

Conclusiones

- La selección de características vía muestra y modelación, permite detectar las variables de mayor influencia o poder explicativo, permitiendo ganar eficiencia computacional para la modelación del total de los datos.
- Tener claridad sobre la configuración del clúster en AWS es esencial para poder realizar de forma exitosa las ejecuciones con el conjunto de datos.
- Las instancias M5 de Amazon EC2 son las que presentaron mejor adaptación a nuestro proyecto, dado que son instancias de propósito general, además, ofrecen equilibrio entre recursos computacionales y económicos.
- Si bien, en nuestras prácticas de la maestría se utiliza con cierta frecuencia la librería pandas de Python, se pudo evidenciar claramente que para el manejo de grandes volúmenes de datos es insuficiente. La teoría vista en clase sobre el funcionamiento de una técnica como Map-Reduce y posteriormente la actualización a lenguajes y técnicas como Spark nos hacen tener un punto de vista diferente a la hora de abordar este tipo de retos.

Trabajo futuro

- Probar otros modelos de ensamble que puedan incrementar la precisión del resultado
- Realizar un análisis en más detalle de los resultados de los parámetros (betas) del modelo para el caso de la regresión lineal, que ayuden a comprender cuales son las comunas o días de la semana que representan un incremento o disminución en el tiempo promedio de viaje

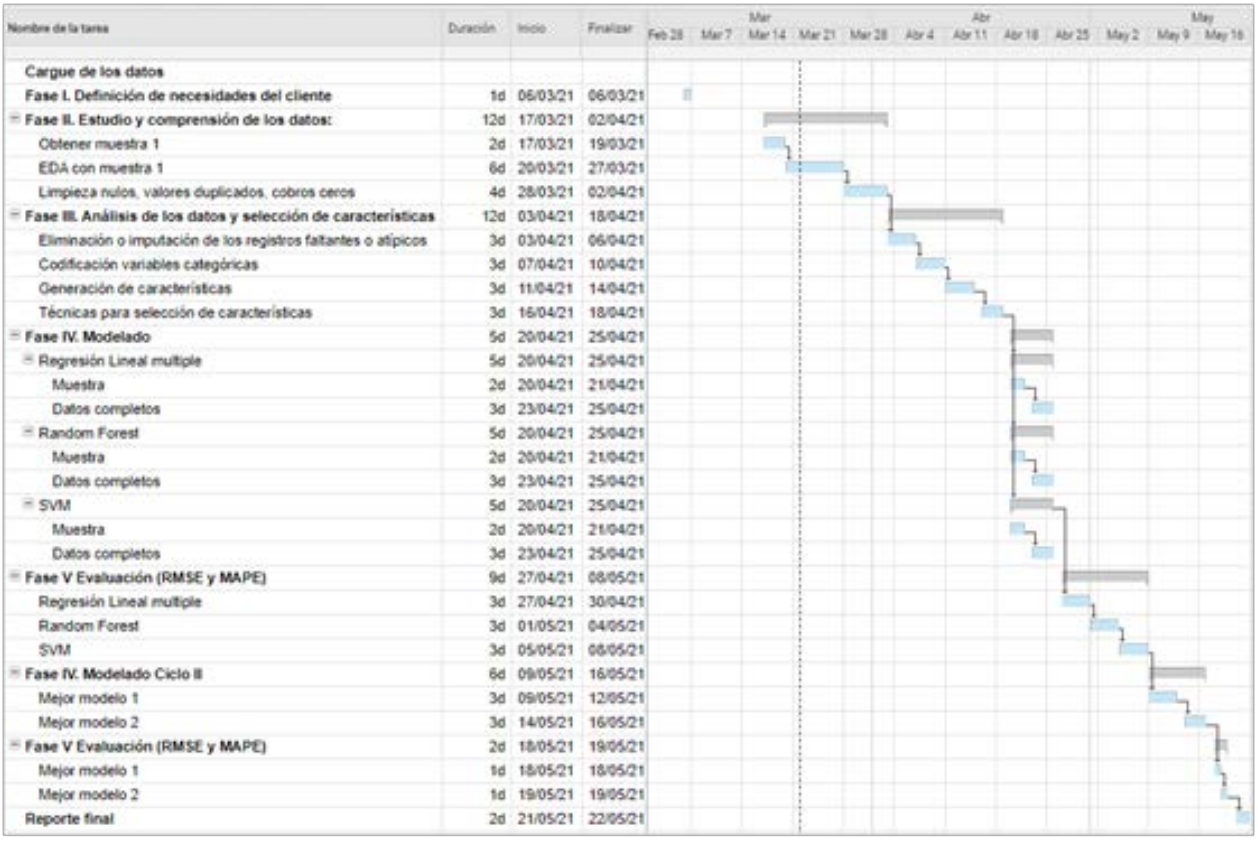
Ejecución del Plan

Con respecto al plan inicialmente propuesto, se destacan los siguientes elementos, que llevaron a una inversión de mayor tiempo o esfuerzo a lo inicialmente concebido:

- Al entender a mayor profundidad los datos, y los requerimientos de limpieza y tratamiento, fue necesario adicionar nuevas actividades que requirieron de un mayor tiempo.
- Derivado de una mayor comprensión de los datos, sumado a una nueva revisión del estado del arte, se definió cambiar una de las técnicas de modelación inicialmente planteada, y ajustar las métricas de evaluación, lo que requirió de nuevos entendimientos y análisis.
- La configuración del clúster de AWS, tomó más tiempo de lo esperado, dado que es un proceso de ensayo – error.

Seguidamente, se presenta el plan inicial versus el final.

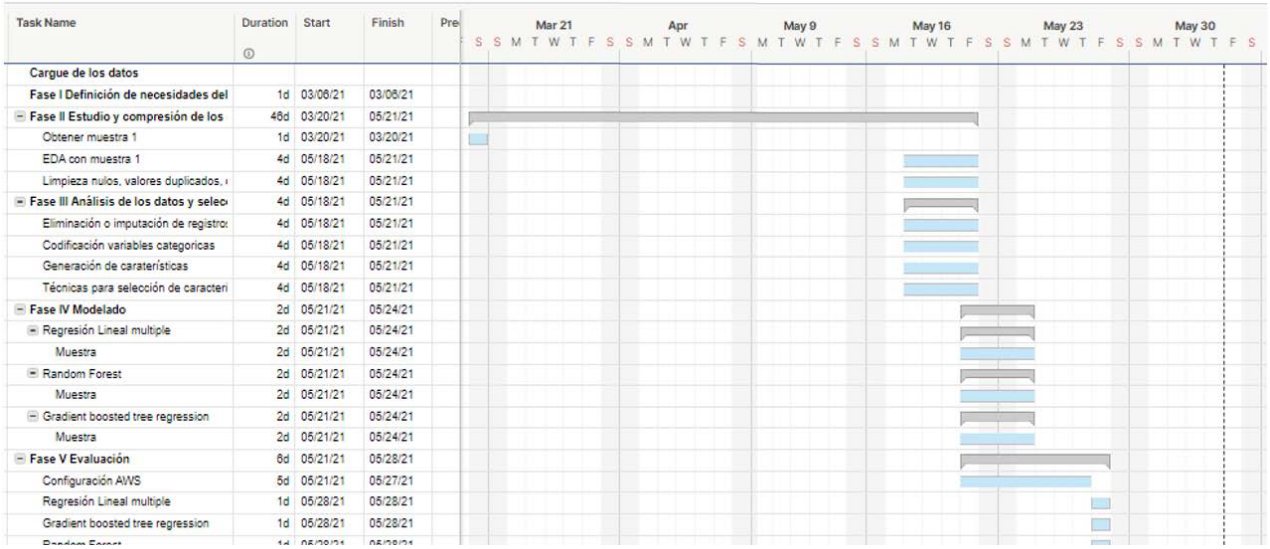
Plan Inicial



Nombre de la tarea	Duración	Inicio	Finalizar
Cargue de los datos			
Fase I. Definición de necesidades del cliente	1d	06/03/21	06/03/21
Fase II. Estudio y comprensión de los datos:	12d	17/03/21	02/04/21
Obtener muestra 1	2d	17/03/21	19/03/21
EDA con muestra 1	6d	20/03/21	27/03/21
Limpieza nulos, valores duplicados, cobros ceros	4d	28/03/21	02/04/21
Fase III. Análisis de los datos y selección de características	12d	03/04/21	18/04/21
Eliminación o imputación de los registros faltantes o atípicos	3d	03/04/21	06/04/21
Codificación variables categóricas	3d	07/04/21	10/04/21
Generación de características	3d	11/04/21	14/04/21
Técnicas para selección de características	3d	16/04/21	18/04/21
Fase IV. Modelado	5d	20/04/21	25/04/21
Regresión Lineal multiple	5d	20/04/21	25/04/21
Muestra	2d	20/04/21	21/04/21
Datos completos	3d	23/04/21	25/04/21
Random Forest	5d	20/04/21	25/04/21
Muestra	2d	20/04/21	21/04/21
Datos completos	3d	23/04/21	25/04/21
SVM	5d	20/04/21	25/04/21
Muestra	2d	20/04/21	21/04/21
Datos completos	3d	23/04/21	25/04/21
Fase V Evaluación (RMSE y MAPE)	9d	27/04/21	06/05/21
Regresión Lineal multiple	3d	27/04/21	30/04/21
Random Forest	3d	01/05/21	04/05/21
SVM	3d	05/05/21	06/05/21
Fase IV. Modelado Ciclo II	6d	09/05/21	16/05/21
Mejor modelo 1	3d	09/05/21	12/05/21
Mejor modelo 2	3d	14/05/21	16/05/21
Fase V Evaluación (RMSE y MAPE)	2d	18/05/21	19/05/21
Mejor modelo 1	1d	18/05/21	18/05/21
Mejor modelo 2	1d	19/05/21	19/05/21
Reporte final	2d	21/05/21	22/05/21

Figura 5: diagrama del plan inicialmente propuesto para el desarrollo del proyecto.

Real



Task Name	Duration	Start	Finish
Cargue de los datos			
Fase I Definición de necesidades del	1d	03/06/21	03/06/21
Fase II Estudio y compresión de los	46d	03/20/21	05/21/21
Obtener muestra 1	1d	03/20/21	03/20/21
EDA con muestra 1	4d	05/18/21	05/21/21
Limpieza nulos, valores duplicados,	4d	05/18/21	05/21/21
Fase III Análisis de los datos y selec	4d	05/18/21	05/21/21
Eliminación o imputación de registro	4d	05/18/21	05/21/21
Codificación variables categoricas	4d	05/18/21	05/21/21
Generación de caraterísticas	4d	05/18/21	05/21/21
Técnicas para selección de caracteri	4d	05/18/21	05/21/21
Fase IV Modelado	2d	05/21/21	05/24/21
Regresión Lineal multiple	2d	05/21/21	05/24/21
Muestra	2d	05/21/21	05/24/21
Random Forest	2d	05/21/21	05/24/21
Muestra	2d	05/21/21	05/24/21
Gradient boosted tree regression	2d	05/21/21	05/24/21
Muestra	2d	05/21/21	05/24/21
Fase V Evaluación	6d	05/21/21	05/28/21
Configuración AWS	5d	05/21/21	05/27/21
Regresión Lineal multiple	1d	05/28/21	05/28/21
Gradient boosted tree regression	1d	05/28/21	05/28/21

Figura 6: plan que se realmente se llevó a cabo para el desarrollo del proyecto.

Implicaciones éticas

- Los resultados pueden llevar a que algunos conductores de taxi eviten desplazarse hacia ciertas zonas, donde la duración promedio es mayor. Afectando la disponibilidad para algunos usuarios y zonas.
- Esto podrá llevar a un aumento de los niveles de supervisión y exigencias en términos de ingresos esperados hacia los conductores de taxi que no son los dueños de este.
- Se podrán determinar las zonas y horas de mayor afluencia de usuarios, y que podría llevar a: ataques terroristas; actos delictivos y campañas publicitarias invasivas no solicitadas.

Aspectos legales y comerciales

En términos de la administración del taxi, permitirá tener una mejor programación en la asignación de los vehículos a los usuarios, dado que se podrá hacer una mejor programación de los viajes con el objetivo de lograr el máximo de viajes posibles, lo que aumentará los ingresos y, a la vez, una mejor cobertura de la demanda.

Referencias

- Amazon EC2 Instance Types, https://aws.amazon.com/ec2/instance-types/?nc1=h_ls
- Das, Anuradha (2019). Exploratory Data Analysis of New York Taxi Trip Duration Dataset using Python https://medium.com/analytics-vidhya/exploratory-data-analysis-of-nyc-taxi-trip-duration-dataset-using-python-257fdef2749e
- Das, Anuradha (2019). Linear Regression Model on the New York Taxi Trip Duration Dataset using Python https://medium.com/analytics-vidhya/building-a-linear-regression-model-on-the-new-york-taxi-trip-duration-dataset-using-python-2857027c54f3
- Li, S. (2018). Towards Data Science. Building A Linear Regression with PySpark and MLlib https://towardsdatascience.com/building-a-linear-regression-with-pyspark-and-mllib-d065c3ba246a
- Research Prediction Competition, 2015. ECML/PKDD 15: Taxi Trajectory Prediction (I). Predict the destination of taxi trips based on initial partial trajectories. Disponible en: https://www.kaggle.com/c/pkdd-15-predict-taxi-service-trajectory-i
- Sarkar, M. 2020. ChaiEDA: NYC Taxi Trip Duration – análisis. Disponible en: https://www.kaggle.com/neomatrix369/chaieda-nyc-taxi-trip-duration-analysis

- Spark, 2021. Classification and regression - spark.ml https://spark.apache.org/docs/1.6.1/ml-classification-regression.html#random-forest-regression