



Predicción de tarifas de taxi en NY

Ismael Orihuela

01-Abril-2024

Resumen presentación

Se analizaron más de 55 millones de registros con información de viajes realizados en taxi durante el periodo de 2009 al primer semestre de 2015 en la ciudad de Nueva York.

Con base es este análisis se encontraron algunos hallazgos, como las zonas de mayor actividad de viajes, la tarifa promedio, la distancia promedio recorrida, siendo ésta el factor principal en el costo, o la relación entre las horas día y las tarifas.

También se desarrollo un modelo de Machine Learninig, en concreto un modelo de arboles de decisión, con la técnica de "Gradiet boosting machines" con el objetivo de predecir el costo de un nuevo viaje en la ciudad de NY.

Entendimiento del negocio:

De acuerdo con el **informe anual del 2015** de The Taxi & Limousine Commission (encargada de regular y supervisar la industria de taxis y limusinas), existen **13,587 taxis** autorizados en NY.

La costos establecidas en 2015 dicen que la tarifa básica del viaje se estima de acuerdo las siguientes condiciones

- Cargo inicial \$2.5.
- + 50 Centavos por 1/5 de milla o 60 segundos en trafico lento.
- + 50 Centavos por viajes que el destino sea New York, Nassau, Suffolk, Westchester, Rockland, Dutchess, Orange o Putnam Counties.
- + \$1 en fines de semana de 4pm a 8pm.
- + 50 Centavos de 8PM a 6PM.

Fuentes:

<https://pix11.com/news/nyc-cab-software-could-have-you-tipping-too-much/>
https://www.nyc.gov/assets/tlc/downloads/pdf/annual_report_2015.pdf

Entendimiento del negocio:

A pesar de que existe una regulación de la tarifa que cobra un taxi, podemos hacer uso de modelos de ML para inferir la tarifa de un viaje y generalizar su comportamiento.

Un modelo predictivo puede ser de utilidad para los usuarios para:

- Obtener una estimación de la tarifa de un viaje y planear gastos.
- Comparar costos entre alternativas de transporte (como el tren o metro).

De igual forma para las compañías de taxis y órganos reguladores:

- Tener un punto de partida generalizado para establecer tarifas competitivas.

También es necesario mencionar factores que pueden influir en las tarifas como el tráfico, las condiciones climáticas, la trayectoria real del viaje, zonas de partida o destino, la demanda del servicio y condiciones sociales que escapan a este análisis.

Comprensión de los datos-Información básica

Ejemplo de un registro de la información disponible:

key	2009-06-15 17:26:21.0000001
fare_amount	4.5
pickup_datetime	"2009-06-15 17:..."
pickup_longitude	-73.8
pickup_latitude	40.7
dropoff_longitude	-73.8
dropoff_latitude	40.7
passenger_count	1

- **key:** Identificar único de cada viaje.
- **fare_amount:** costo del viaje.
- **pickup_datetime:** Fecha y hora de inicio del viaje.
- **pickup_longitude, pickup_latitude:** Coordenas del punto de origen del viaje
- **dropoff_longitude, dropoff_latitude:** Coordenas del punto de destino del viaje
- **passenger_count:** Cantida de pasajeros.

Preparación de los datos – limpieza

Dentro del análisis se detectaron algunas inconsistencias en la información como:

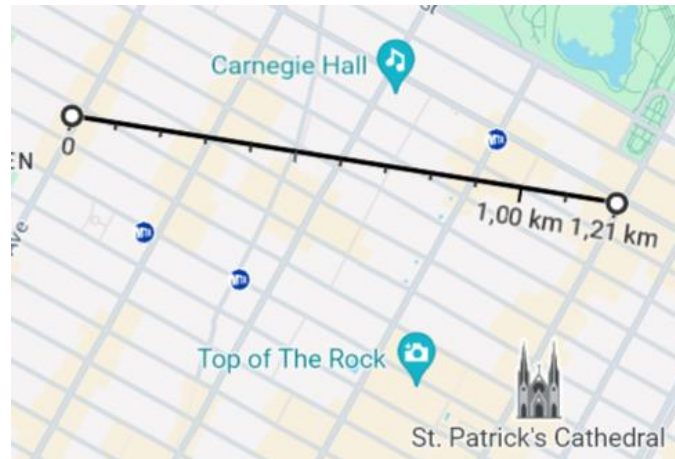
- Datos faltantes en coordenadas de destinos.
- Tarifas con costo negativo, menores a \$2.5 dls o poco comunes de más de \$100 dls.
- Viejes con orígenes y destinos fuera de la ciudad de NY.
- Viajes con cero KM recorridos.
- Viajes donde el costo por KM es mayor a \$50, viajes demasiado costosos poco comunes


Todos estos registros fueron omitidos, dado que no se conoce el comportamiento que da origen a estos casos particulares, además se cuenta con una gran cantidad de información.

Preparación de los datos – limpieza

Se generaron variables auxiliares referente a la temporalidad de viaje como **año, mes, día de la semana** y **hora del viaje**.

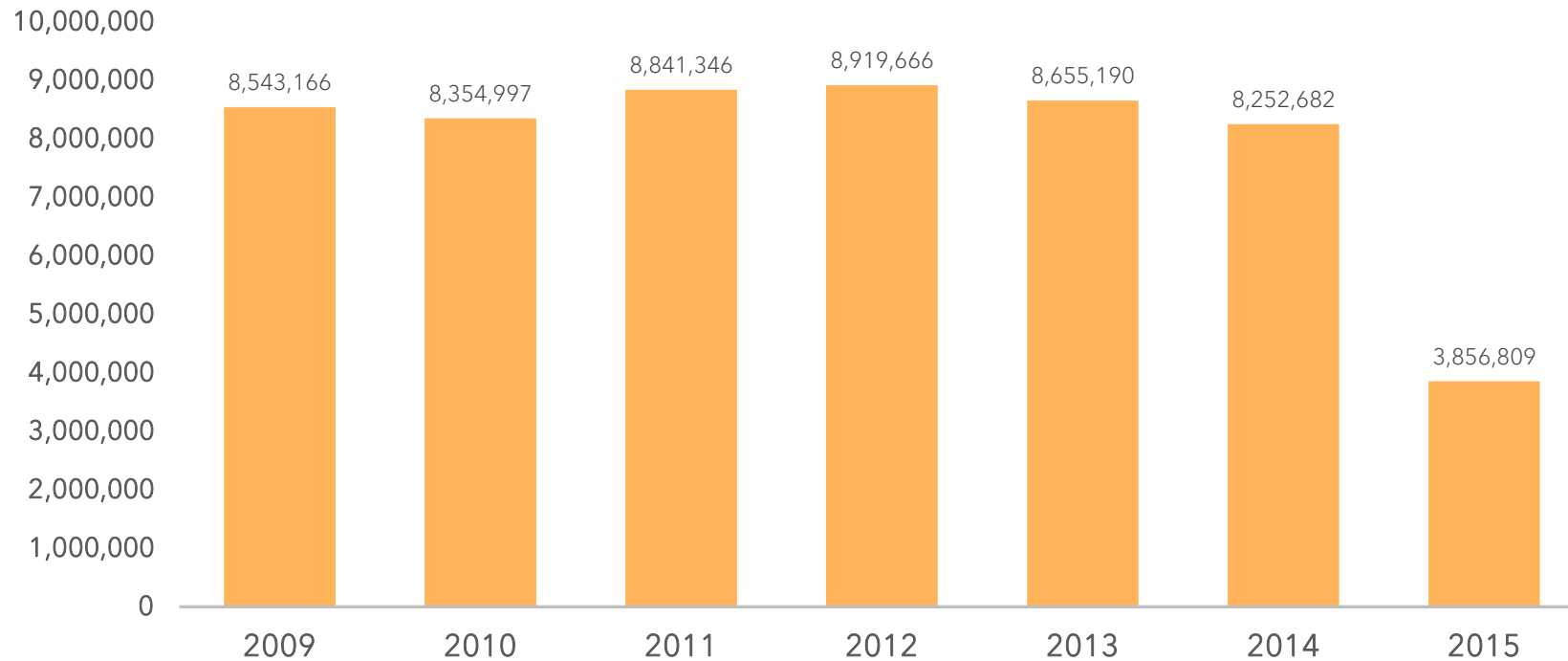
Adicional se **cálculo la distancia** entre puntos de origen y destino mediante una aproximación, dado que no se cuenta con las rutas reales de cada viaje.





¿Cómo son los
viajes en taxi
en NY?

Viajes en taxi NY



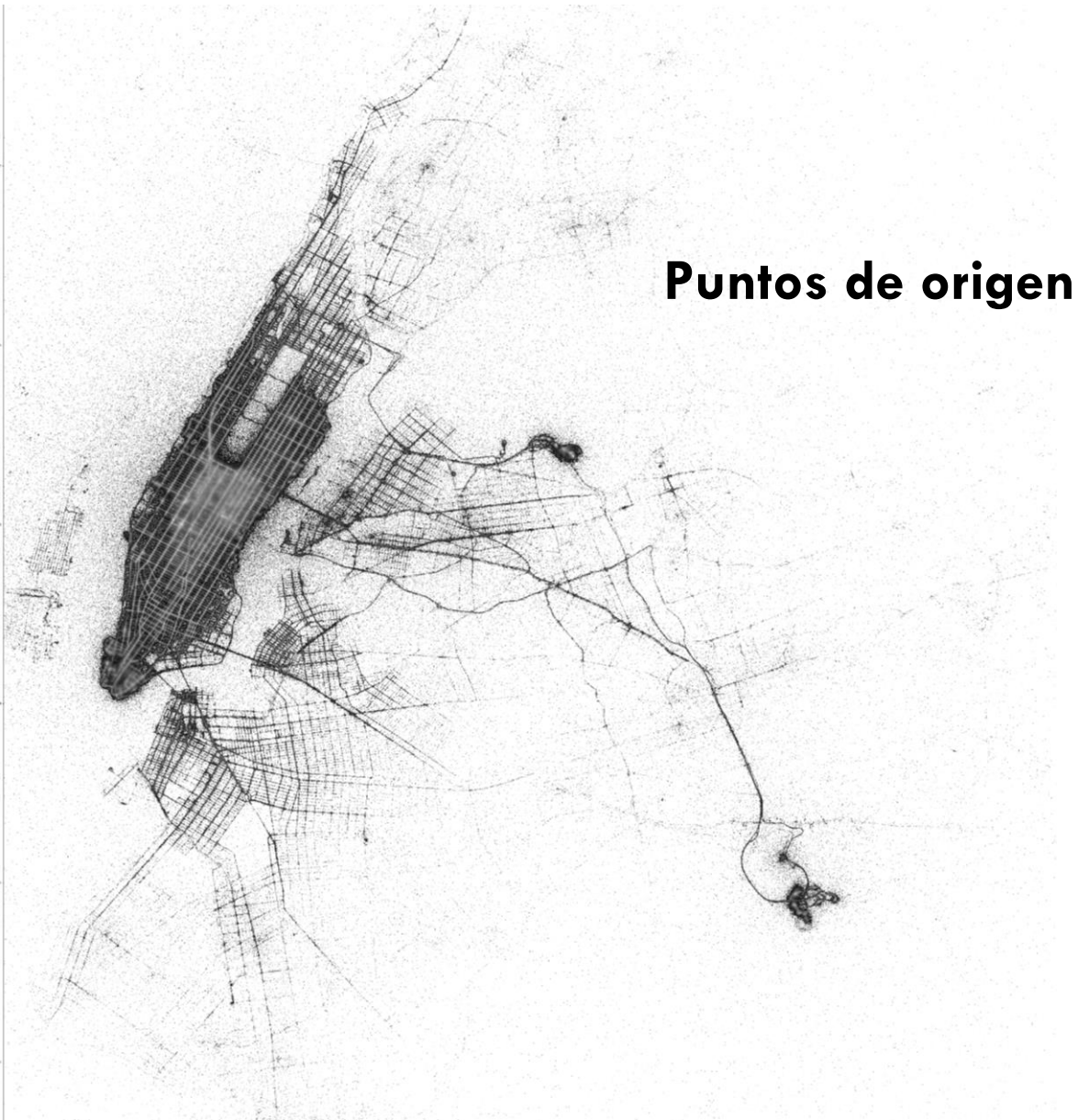
Desde 2009 al primer semestre de 2015 se tienen más de **55 millones**

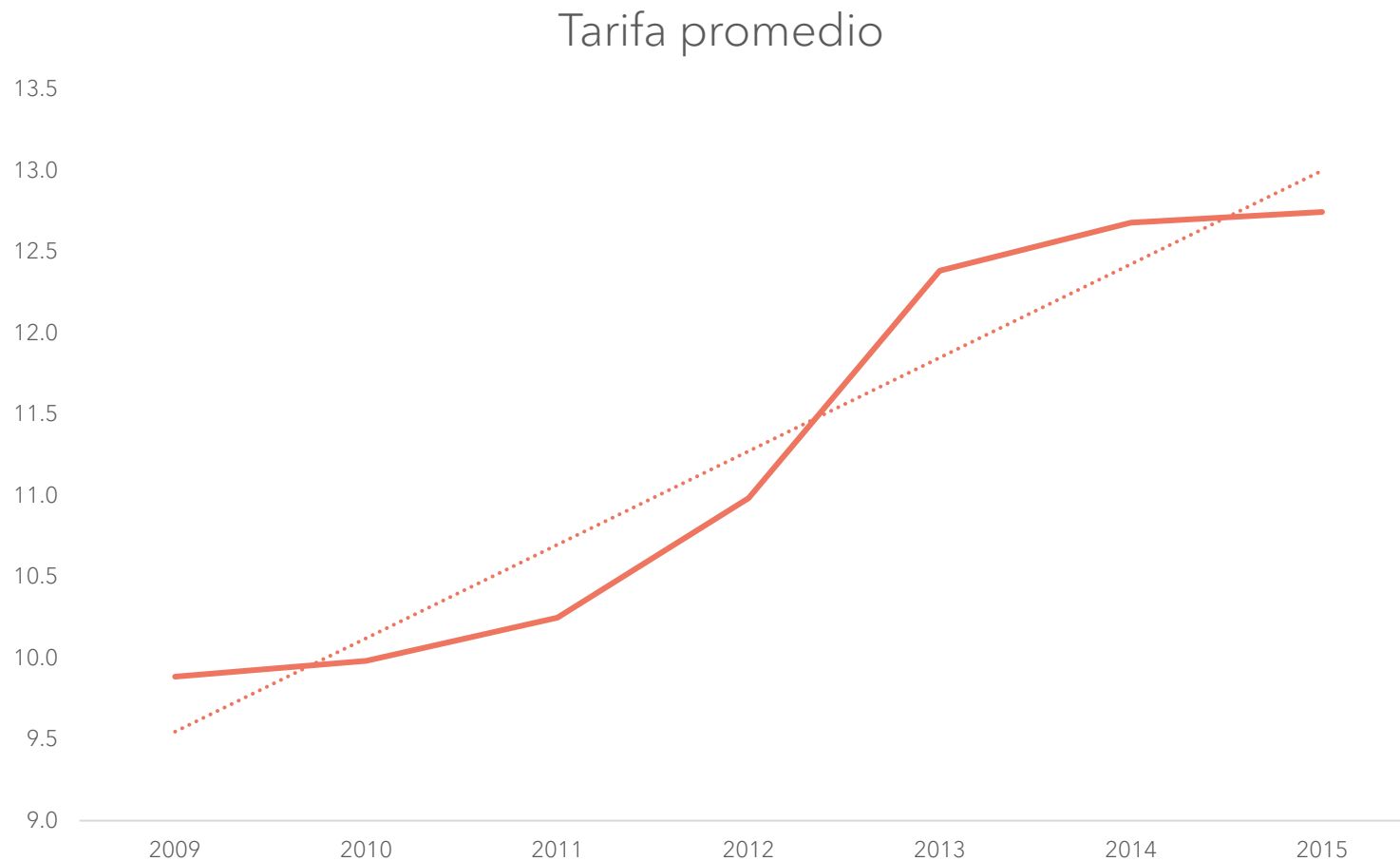
De 2009 a 2014 se realizan en **promedio más 8.5 millones de viajes al año.**

La tarifa promedio del viaje es de **\$11.1 dlls.**

La tarifa promedio por KM recorrido es de **\$4.4 dlls.**

Densidad de los viajes

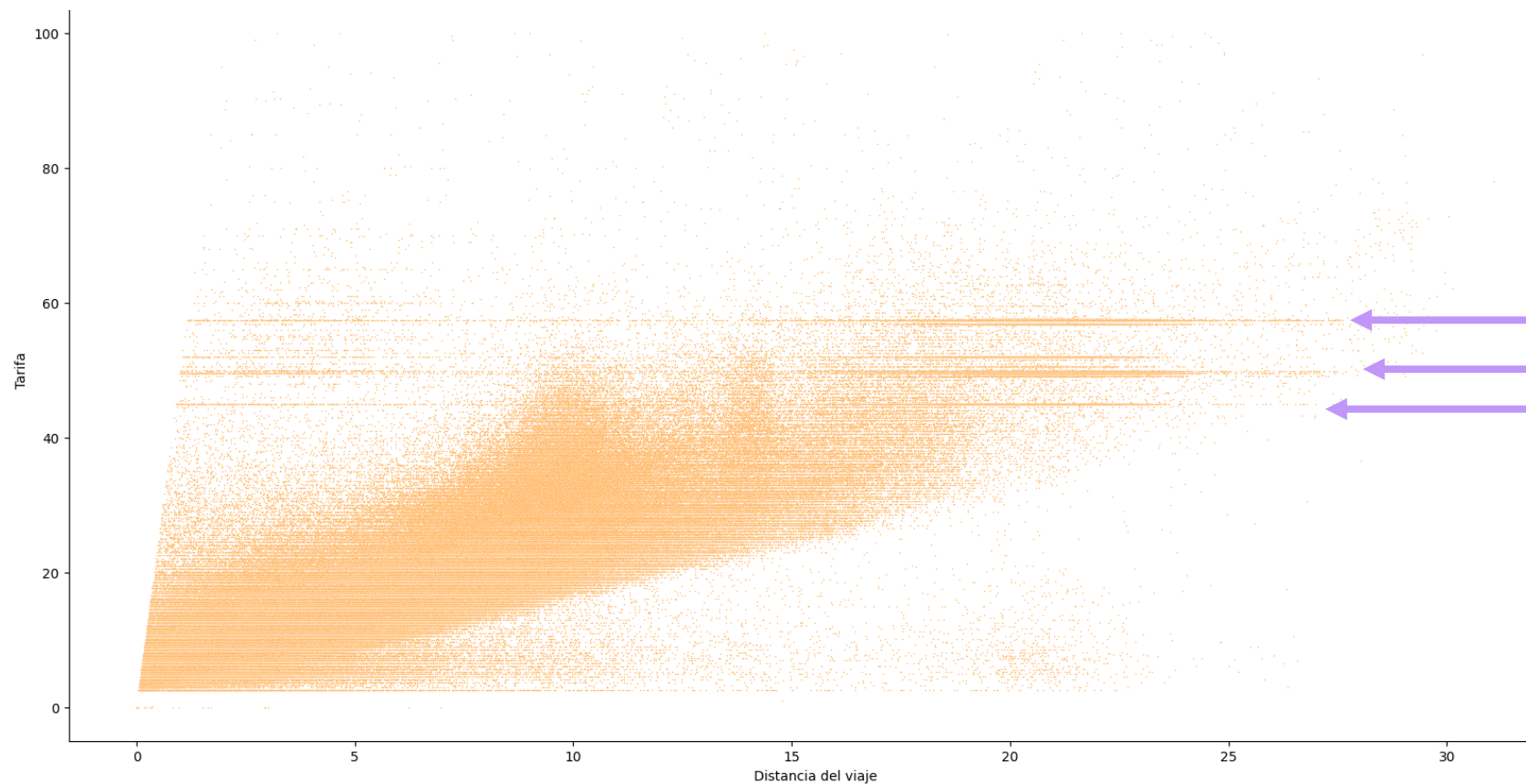




Desde el 2009 existe un **incremento** en la tarifa , particularmente de 2012 a 2013 el alza es más evidente.

Parece ser que existe una **relación creciente** el año el precio promedio del viaje.

Tarifa y distancia recorrida

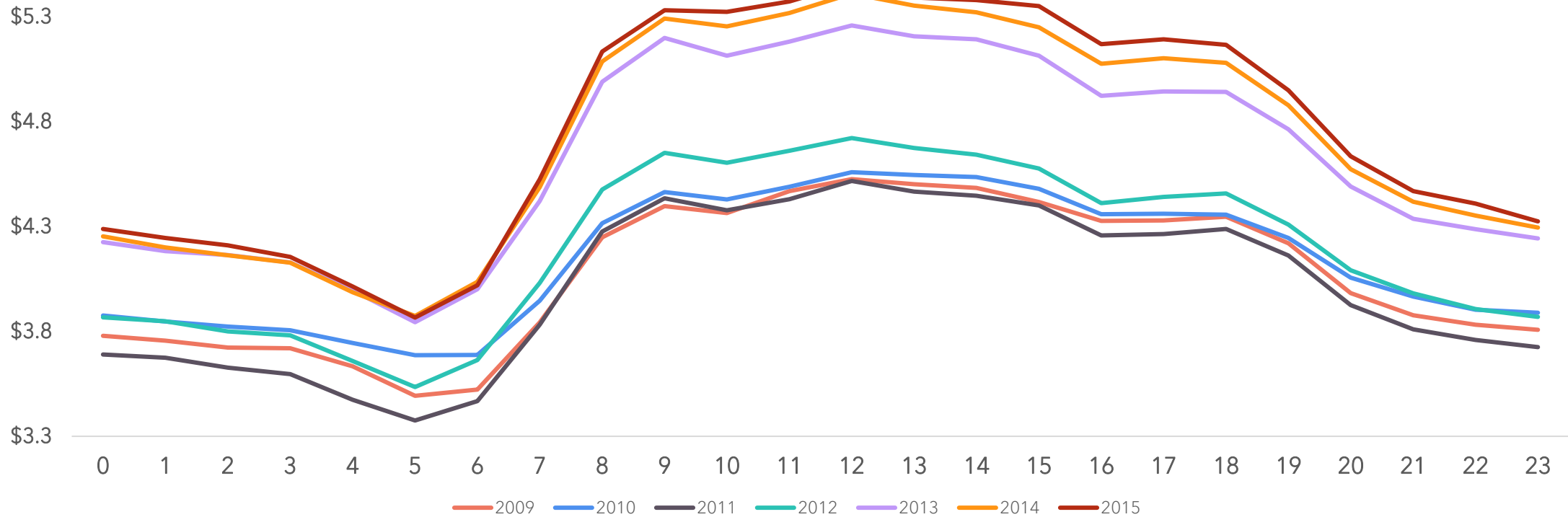


Entre los \$40 y \$60 dlls
existe una gran cantidad de
viajes que no importa la
distancia recorrida, el costo
es el mismo.

Pueden ser tarifas con
precios fijos.

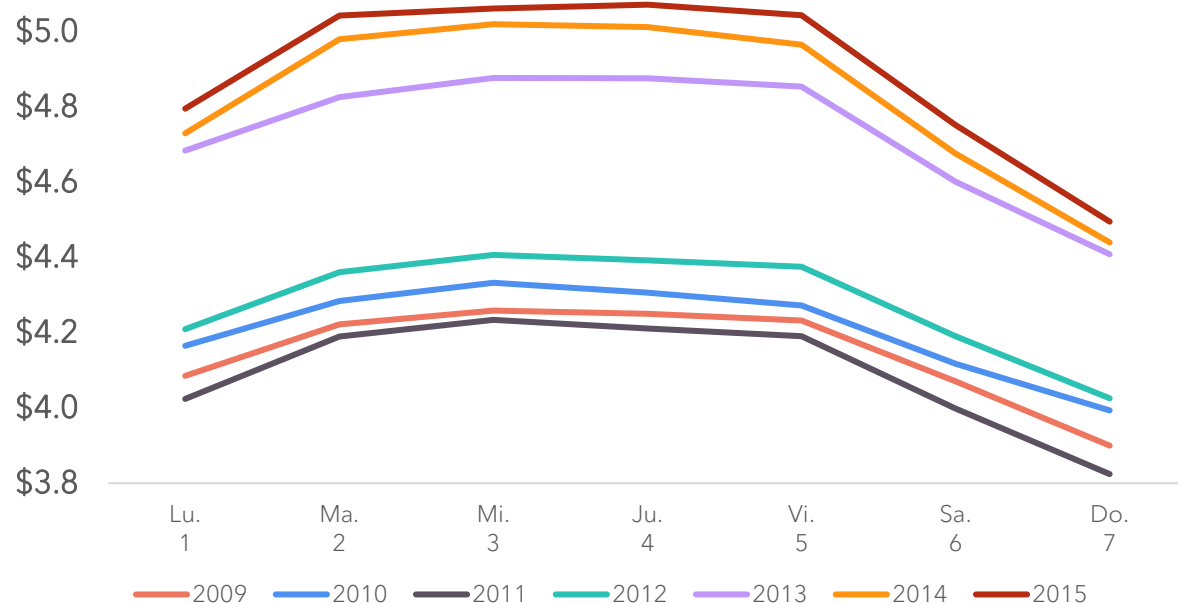
Parece ser que existe un **relación** entre la distancia y la tarifa del viaje.

Tarifa por KM y hora del viaje



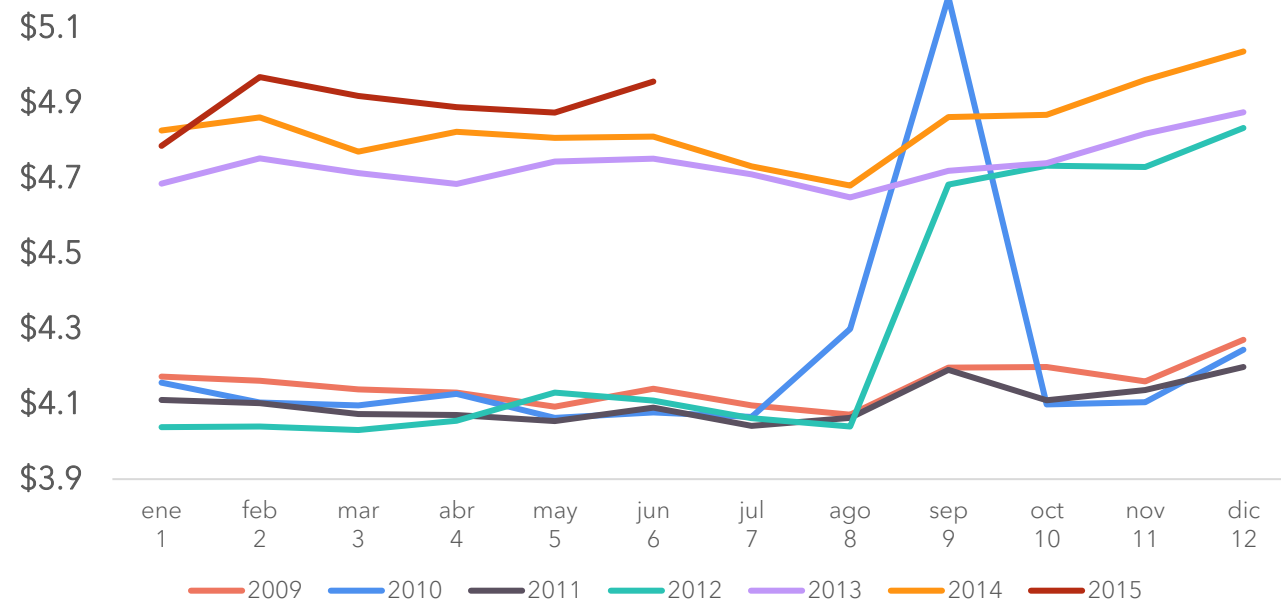
Existe un **patrón no lineal** entre la hora del viaje y el costo de la tarifa por KM, siendo las madrugadas las tarifas más bajas y partir de las 5AM el costo comienza a incrementar.

Tarifa promedio por KM y día de la semana



Existen también un **patrón no lineal** entre el día y la **tarifa por KM**, siendo entre martes y viernes los precios más altos, y fines de semana en específico los domingos las tarifas más bajas

Tarifa promedio por KM y día de la semana



Analizando la tarifa por KM de forma mensual, hay algunos datos atípicos en los años 2010 y 2012, y no parece haber un relación directa con las tarifas.

Justificación modelado

Debido a la gran cantidad de información (millones), las relaciones lineales y no lineales entre las variables, junto con limitación de recursos técnicos y tiempo. Se optó por entrenar un modelo con **LightGBM (Gradient boosting machines)** ya que es eficiente en términos de:

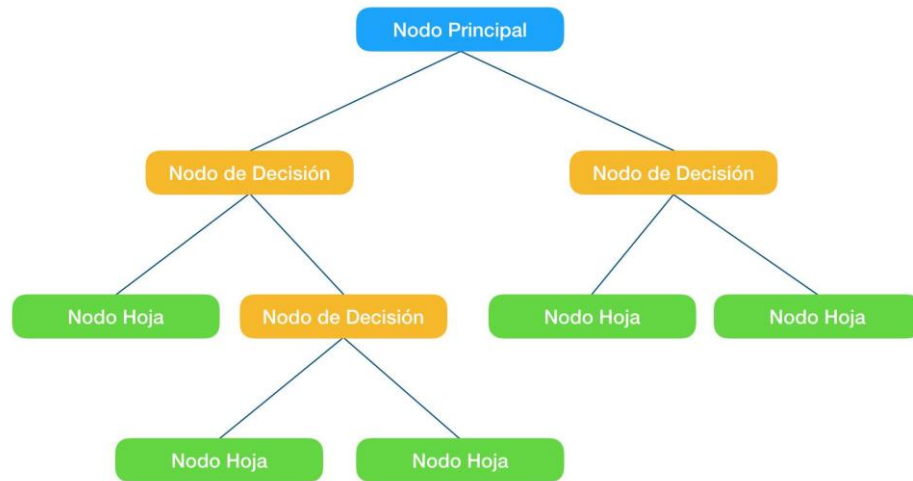
- Uso de recursos computacionales (Memoria - procesador)
- Precisión
- Compresión de datos complejos
- Soporte para bigdata

Existen otros modelos de machine learning, pero algunos de sus implementaciones se ven limitadas por los recursos disponibles, o no se pueden entrenar de forma paralela o pueden llevar más tiempo de su desarrollo.

Arboles de decisión y gradient Boosting

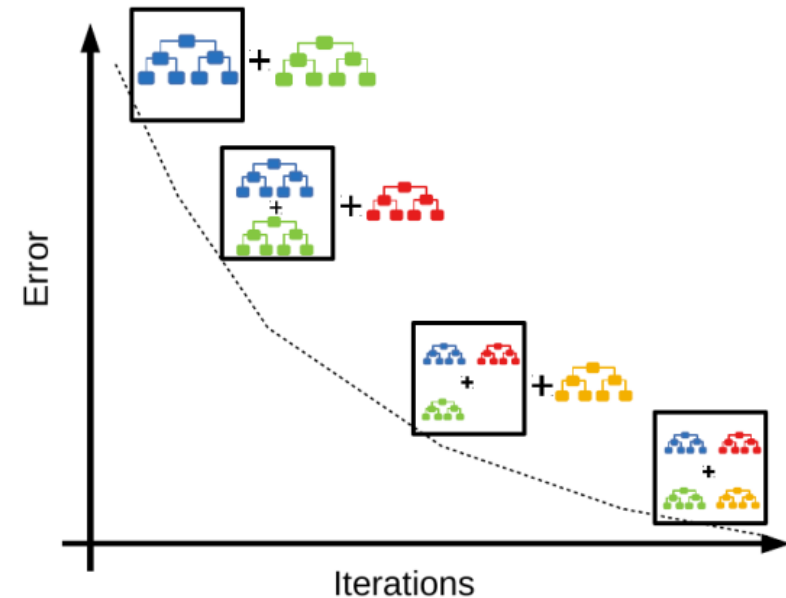
Arboles de decisión

Es una estructura jerárquica que representa decisiones. Tiene un nodo principal y sigue diferentes reglas/pruebas, que conducen a un resultado final.



Gradient Boosting

Consiste en combinar múltiples árboles de decisión de forma secuencial, donde cada nuevo modelo se enfoca en mejorar los errores del árbol anterior.



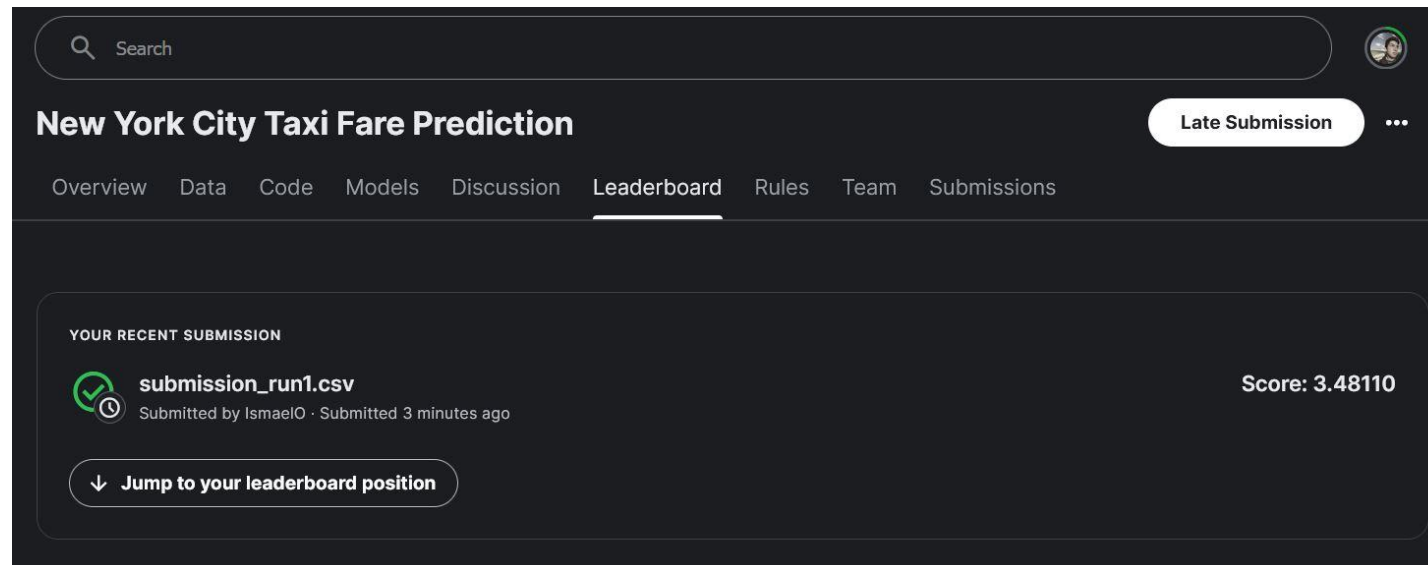
Resultados

Se usaron **8 millones** de registros para entrenar el modelo, adicional de otros 2 millones de hacer pruebas.

Para lograr una mejor precisión se hicieron varias iteraciones de entrenamiento para identificar los mejores parámetros que se ajustaran a los datos.

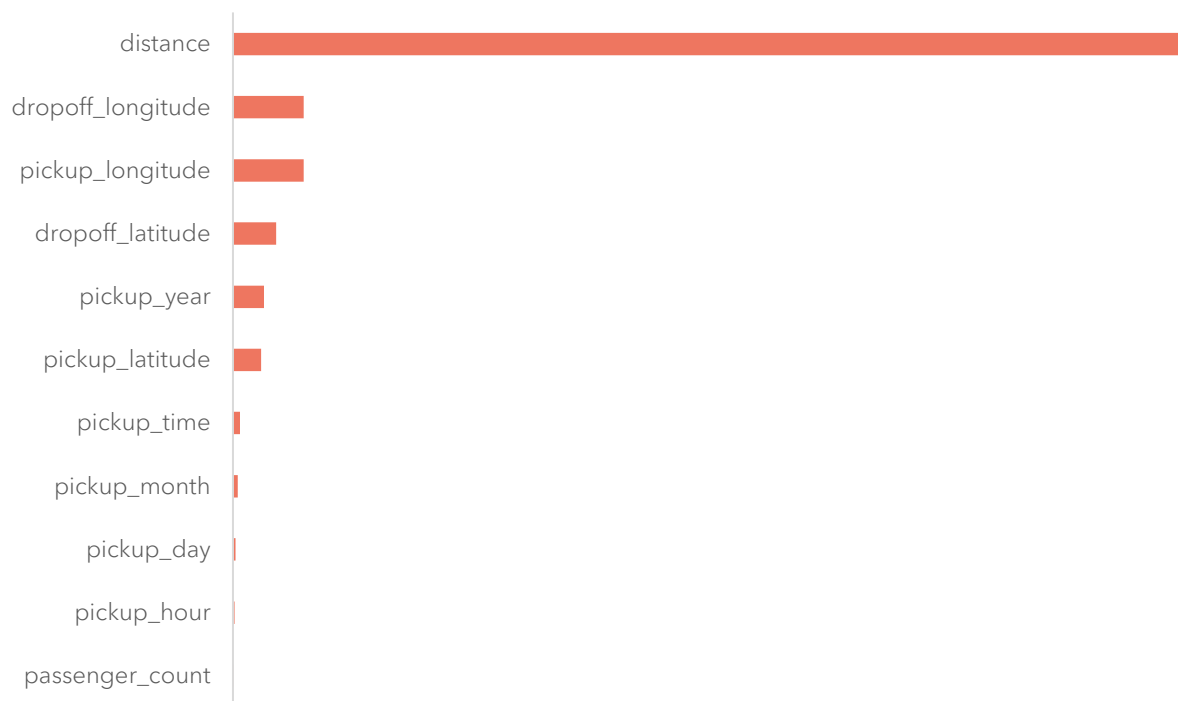
Dentro de los datos de prueba se obtuvo una **variación de \$2.5 dlls** entre las predicciones del modelo y los datos reales.

En la competencia de kaggle se obtuvo un **score de 3.48**



The screenshot shows the Kaggle competition interface for 'New York City Taxi Fare Prediction'. At the top, there is a search bar and a user profile icon. Below the competition title, there are tabs for 'Overview', 'Data', 'Code', 'Models', 'Discussion', 'Leaderboard' (which is selected), 'Rules', 'Team', and 'Submissions'. A 'Late Submission' button is visible on the right. The main content area displays 'YOUR RECENT SUBMISSION' with a green checkmark icon, the filename 'submission_run1.csv', and the text 'Submitted by IsmaelIO · Submitted 3 minutes ago'. The score 'Score: 3.48110' is shown on the right. At the bottom, there is a button that says '↓ Jump to your leaderboard position'.

Relevancia de cada variable



La gráfica nos indica la influencia de cada variable en las predicciones del modelo, siendo la distancia la de mayor relevancia, cosa que podíamos intuir desde la grafica entre el la distancia y tarifas.

Seguido de variables de coordenadas que a su vez se relacionan con la distancia. Y finalmente las variable relacionadas con el fecha, y casi sin relevancia la cantidad de pasajeros