Tareas calificadas por los compañeros: Proyecto Capstone - La Batalla de los Vecindarios (Semana 1)

**Introducción/Problema Comercial**

El problema que debemos abordar es determinar la ubicación no recomendada para abrir un restaurante en la ciudad de Nueva York. Esta es una pregunta crucial para los emprendedores y propietarios de restaurantes, ya que la ubicación puede tener un impacto significativo en el éxito del negocio, sobretodo en ciudades con un flujo tan activo de personas residentes y turistas.

Adicionalmente, es necesario comprender las distintas oportunidades y potenciales amenazas que puede presentar abrir un local comercial en una ubicación "A" o en una ubicación "B", dependiendo de las disponibilidades que presenta nuestro usuario.

El desarrollo de este proyecto permitirá identificar potenciales locaciones ya ocupadas para que el usuario que busque abrir un restaurante nuevo en una zona distinta en la ciudad de New York, de tal forma que pueda comparar las alternativas ofrecidas y conseguir una ubicación que cumpla con sus expectativas.

**Datos**

Para poder desarrollar este proyecto, se han considerado los siguientes datos:

Lugar (Place): Representado por el campo "name" (nombre), indica el nombre del lugar. Ejemplos: "Thai Diner", "Manhatta", "Fish Cheeks", etc.

Categorías (Categories): Representado por el campo "categories", contiene información sobre las categorías del lugar. Cada categoría tiene un ID y un nombre. Ejemplos: "Thai Restaurant", "American Restaurant", "Italian Restaurant", etc.

Ubicación (Location): Representado por el campo "location", contiene información sobre la dirección y la ubicación geográfica del lugar. Incluye campos como "address" (dirección), "locality" (localidad), "postcode" (código postal), "region" (región), etc.

Distancia (Distance): Representado por el campo "distance", indica la distancia en metros desde la ubicación de referencia hasta el lugar.

Coordenadas geográficas (Geocodes): Representado por el campo "geocodes", contiene las coordenadas de latitud y longitud del lugar. Se proporcionan tanto las coordenadas principales (main) como las del techo (roof).

Enlaces (Link): Representado por el campo "link", proporciona un enlace al lugar en el servicio de Foursquare.

Zona horaria (Timezone): Representado por el campo "timezone", indica la zona horaria del lugar.

**A modo de ejemplo, se ha considerado el siguiente:**

Variable 1: Lugar (Place)

Nombre: "Thai Diner"

Variable 2: Categorías (Categories)

ID: "4bf58dd8d48988d149941735"

Nombre: "Thai Restaurant"

Variable 3: Ubicación (Location)

Dirección: "123 Main St"

Localidad: "Ciudad"

Código postal: "12345"

Región: "Estado"

País: "País"

Variable 4: Distancia (Distance)

Valor: 200 metros

Variable 5: Coordenadas geográficas (Geocodes)

Coordenadas principales (main):

Latitud: 40.123456

Longitud: -74.654321

Coordenadas del techo (roof):

Latitud: 40.123459

Longitud: -74.654325

Variable 6: Enlaces (Link)

Enlace: "https://foursquare.com/venue/abcd1234"

Variable 7: Zona horaria (Timezone)

Zona horaria: "America/New\_York"

Análisis Exploratorio:

La base de datos inicial dispone un total de 9 variables, siendo estas las siguientes:

* Fsq\_id: variable correspondiente al id asignado a cada local en foursquare
* Categories: variable que contiene más información respecto a la categoría del local
* Chains: variable en la que se indica si el local corresponde a una cadena de locales
* Distance: variable que corresponde a la distancia existente entre locales
* Geocodes: variable que corresponde al código geográfico del local
* Link: enlace correspondiente al perfil de foursquare
* Location: locación del local
* Name: Nombre del local
* Related\_places: lugares relacionados cercanos
* Timezone: horario en el que se realizó la consulta

A screen shot of a computer

Description automatically generated with medium confidence

Seguidamente, se empleó un análisis descriptivo para evaluar el comportamiento de las variables. Se han considerado los 50 registros disponibles de Foursquare y se obtuvo la siguiente información:

A picture containing text, font, screenshot

Description automatically generated

Como se puede apreciar, la media de datos se encuentra en 4882.78, posiblemente debido a los valores de la distancia entre locales. Esto también se ve reflejado en el valor máximo, ascendiendo a 37598; y al valor mínimo, ascendiendo a 498.

Una vez que se cumplió con la consulta a través de la API de Foursquare, se optó por transformar los datos para obtener una mayor cantidad de información. De esta forma, se logró un nuevo dataframe que consta de un total de 29 variables, que son:

1. fsq\_id: El ID único del restaurante en Foursquare.
2. categories: Las categorías a las que pertenece el restaurante.
3. chains: Indica si el restaurante pertenece a una cadena o es independiente.
4. distance: La distancia del restaurante a una ubicación de referencia.
5. geocodes: Datos geográficos relacionados con la ubicación del restaurante.
6. link: Enlace relacionado con el restaurante.
7. location: Información detallada sobre la ubicación del restaurante.
8. name: El nombre del restaurante.
9. related\_places: La cantidad de lugares de interés cercanos al restaurante.
10. timezone: La zona horaria en la que se encuentra el restaurante.
11. address: La dirección del restaurante.
12. census\_block: El bloque censal asociado al restaurante.
13. country: El país en el que se encuentra el restaurante.
14. cross\_street: La calle transversal cercana al restaurante.
15. dma: Área de mercado designada del restaurante.
16. formatted\_address: La dirección del restaurante en un formato legible.
17. locality: La localidad en la que se encuentra el restaurante.
18. postcode: El código postal del restaurante.
19. region: La región o estado en el que se encuentra el restaurante.
20. address\_extended: Información adicional sobre la dirección del restaurante.
21. main.latitude: La latitud principal de la ubicación del restaurante.
22. main.longitude: La longitud principal de la ubicación del restaurante.
23. roof.latitude: La latitud del techo del restaurante.
24. roof.longitude: La longitud del techo del restaurante.
25. drop\_off.latitude: La latitud del punto de entrega del restaurante.
26. drop\_off.longitude: La longitud del punto de entrega del restaurante.
27. category\_1: La categoría principal del restaurante.
28. category\_2: La segunda categoría del restaurante.
29. category\_3: La tercera categoría del restaurante.

Habiendo identificado esto, se comenzó el análisis exploratorio, comenzando con las distintas categorías de restaurantes, de tal forma que se identifique cual es la más popular para considerarlo en nuestro estudio:

A picture containing text, screenshot, parallel, software

Description automatically generated

Como se puede apreciar, los restaurantes americanos son los más comunes dentro de la muestra seleccionada, ascendiendo a un total de 12 restaurantes. Asimismo, en segundo lugar, se encuentran los restaurantes italianos. Es importante precisar que existen muestras más pequeñas de otro tipo de locales como pizzerías, bares de cerveza, entre otros.

Para comprender mejor la información, se realizó un filtro para las 10 categorías más comunes:

A picture containing text, screenshot, plot, line

Description automatically generated

Una vez que se tiene mayor conocimiento entre categorías, se optó por evaluar el comportamiento de las distancias entre locales, buscando ver qué lugares están más cercanos a cuáles:

A picture containing text, screenshot, software, plot

Description automatically generated

Se puede apreciar que las comidas de origen asiático, así como las de comida rápida, tienen una menor distancia. En cambio, restaurantes mexicanos o americanos se encuentran considerablemente más alejados. Asimismo, se ha identificado que los bares de vino se encuentran radicalmente alejados de otro tipo de locales.

A continuación, se presenta un mapa para evaluar la posición de los distintos lugares, evaluando que estos efectivamente presentan una corta distancia entre ellos:

A map with blue pins

Description automatically generated with low confidence

Para concluir el análisis exploratorio, se determinó aquellas interacciones que presentan los locales con otros, de la siguiente forma:

A screen shot of a computer

Description automatically generated with medium confidence

Esta información ha sido recopilada para el primer restaurante del dataframe, contando con 10 lugares de interés cercano y 39 lugares que no son adecuados.

Modelado

Para el caso del modelado, se optó por identificar aquellos locales que son los adecuados para abrir un restaurante. De esta forma, se optó por seleccionar la distancia con otros locales, los puntos de interés y la cantidad de restaurantes en el área. A cada una de estas subvariables se les asignó un peso, siendo estos: 30% para la cantidad de lugares relacionados, 50% para la distancia entre otros locales y 20% para la cantidad de restaurantes en el área. Evaluando esta interacción, se desarrolló una nueva variable denominada “variable\_objetivo” que mide y evalúa si el lugar es adecuado o no:



Para el proceso de modelado, se empleó la técnica Random\_forest, buscando identificar la probabilidad de predecir un valor que no sea el adecuado para establecer un restaurante en New York. De esta forma, el modelo cumplió satisfactoriamente el proceso, prediciendo de forma eficiente aquellos valores que no son adecuados para la apertura de un restaurante.

A picture containing text, screenshot, font

Description automatically generated