Reporte técnico

Asesor:

Gerardo Jesús Camacho

Equipo

Cycle - Sense

Andreina Cardenas Arredondo (A010219917)

Sue Mi Zamarrón Orrantia (A01781507)

Valeria Alejandra Parraga Muñoz (A01029428)

Andrea Ochoa Zuluaga (A01660137)

Fecha de entrega:

09/06/2023

Introducción

Citibike es un sistema y programa de bicicletas compartidas en la ciudad de Nueva York que busca proveer un sistema de bicicletas compartidas seguro, eficiente y sostenible para los residentes y visitantes de la ciudad. Asimismo, busca contribuir con la reducción de tráfico, promover un estilo de vida saludable y sostenible siendo un componente integral de la infraestructura de transporte de la ciudad (Motivate International Inc., 2023).

Para cada vez acercarse más a sus objetivos y visión, Citibike deberá crear estrategias inteligentes que lo permitan crecer. Esto se hará por medio de un análisis de datos. Se utilizarán los datos de "rides" de años pasados. Por medio del análisis descriptivo se explorarán los datos de Citibike y se encontrarán insights que apoyen aún mejor entendimiento del negocio, con el objetivo de apoyar a definir estrategias.

En este reporte, se presentan las soluciones a preguntas de negocio utilizando modelos de machine learning, tanto supervisados como no supervisados. Se realiza un análisis exhaustivo del conjunto de datos proporcionado por Citibike y se generan predicciones para brindar apoyo a los tomadores de decisiones. El reporte destaca el análisis de datos realizado y cómo se aplicaron los conocimientos de aprendizaje automático para abordar los desafíos planteados.

Metodología

Para nuestro proyecto de análisis de datos enfocado en el crecimiento y mejor toma de decisiones de Citibike, empleamos la metodología CRISP. Tiene 4 fases: entendimiento de negocio, comprensión de datos, preparación de los datos, fase de modelado, evaluación, implementación.

Entendimiento de Negocio

Para el primer paso de la metodología CRISP se hizo un estudio de la empresa para poder obtener objetivos concretos para lograr que el análisis de los datos promueva el crecimiento de la empresa Esto se hizo definiendo la misión, visión y objetivos de la empresa. Asimismo, se estudió la historia de la empresa. También se hizo un análisis del AS IS para hacer un estudio de cómo está la empresa en este momento y cómo son sus procesos. Después de obtener un panorama completo, se definieron las siguientes preguntas de negocio a resolver:

- ¿Cómo se puede minimizar los costos?
    - ¿Cómo se puede optimizar y hacer más eficiente el proceso de mantenimiento de las bicicletas?
    - ¿Hay alguna estación que valga la pena consolidar?
- ¿Qué tipos de clientes se tienen?

Comprensión de los datos

Dentro de esta etapa examinamos y buscamos comprender en detalle cada variable de nuestros datos. Nos enfocamos en entender qué representa cada variable, cuáles son sus características principales y cómo se comporta en el contexto de nuestro estudio. Esta fase es de suma importancia antes de aplicar cualquier modelo de aprendizaje automático, ya que nos brinda información clave para tomar decisiones fundamentadas.
Lo que encontramos fue lo siguiente, 30 variables las cuales se divididos en los siguientes datatype: 3 datetime, 12 flotantes, 2 entero16, 3 entero32, 2 entero8 y 9 objetos

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48636559 entries, 0 to 48636558
Data columns (total 31 columns):
 #   Column                   Dtype
---  ------                   -----
 0   TRIPDURATION             int32
 1   STARTTIME                datetime64[ns]
 2   STOPTIME                 datetime64[ns]
 3   START_STATION_ID         int16
 4   START_STATION_NAME       object
 5   START_STATION_LATITUDE   float64
 6   START_STATION_LONGITUDE  float64
 7   END_STATION_ID           int16
 8   END_STATION_NAME         object
 9   END_STATION_LATITUDE     float64
 10  END_STATION_LONGITUDE    float64
 11  BIKEID                   int32
 12  MEMBERSHIP_TYPE          object
 13  USERTYPE                 object
 14  BIRTH_YEAR               float64
 15  GENDER                   int8
 16  OBSERVATION_TIME         datetime64[ns]
 17  CITY_ID                  int32
 18  CITY_NAME                object
 19  COUNTRY                  object
 20  CITY_LAT                 float64
 21  CITY_LON                 float64
 22  CLOUDS                   int8
 23  TEMP_AVG                 float64
 24  TEMP_MIN                 float64
 25  TEMP_MAX                 float64
 26  WEATHER                  object
 27  WEATHER_DESC             object
 28  WEATHER_ICON             object
 29  WIND_DIR                 float64
 30  WIND_SPEED               float64
dtypes: datetime64[ns](3), float64(12), int16(2), int32(3), int8(2), object(9)
memory usage: 9.5+ GB
```

Preparación de los datos

En el proceso de creación de los modelos, se realizó una limpieza de los datos para garantizar su calidad. Se abordó el tratamiento de los valores faltantes (NA) en función del objetivo específico de cada modelo. En algunos casos, se optó por eliminar las filas o columnas que contenían valores faltantes, mientras que en otros casos se decidió llenar esos valores con técnicas como la media, la

Figura 1. Atributos de la base de datos

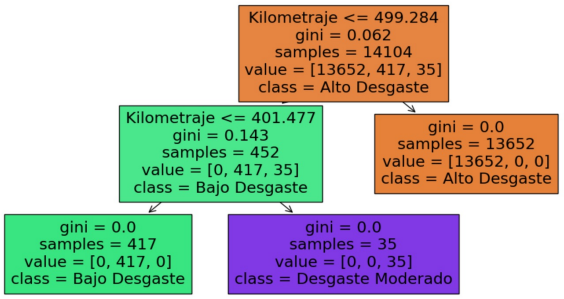
mediana o el valor más frecuente. Además, se llevaron a cabo acciones para mejorar la estructura de los datos. Se crearon nuevas variables que se consideraron relevantes para los modelos en desarrollo, y en algunos casos se eliminaron variables que se consideraron redundantes o que no aportaban información significativa.

También se realizaron tareas de normalización y estandarización de los nombres de variables, asegurando consistencia y claridad en la interpretación de los datos. Esto incluyó la eliminación de caracteres especiales, la estandarización de mayúsculas y minúsculas, y la corrección de posibles errores tipográficos.

Todas estas acciones se llevaron a cabo con el objetivo de adaptar los datos a las necesidades y requisitos específicos de cada modelo en desarrollo, buscando maximizar su rendimiento y su capacidad para brindar resultados precisos y útiles para la toma de decisiones.

Modelado

<u>Supervisado</u>



Se utilizó un modelo supervisado de árboles de decisión para poder estimar cuándo hacerle mantenimiento a las bicicletas teniendo su kilometraje. Se hizo esto calculando el kilometraje total de cada viaje y sumando esto por cada bicicleta para obtener el kilometraje total. Se obtuvieron 3 tipos de clasificaciones: bajo desgaste, desgaste moderado y alto desgaste. Esto para que Citibike calcule y

Figura 2. Árbol de decisión para el mantenimiento de bicicletas

planee los horarios de mantenimiento de mejor manera. Se recomienda checar las bicicletas una vez cumplan 500 km para prevenir a no llegar a un alto desgaste.

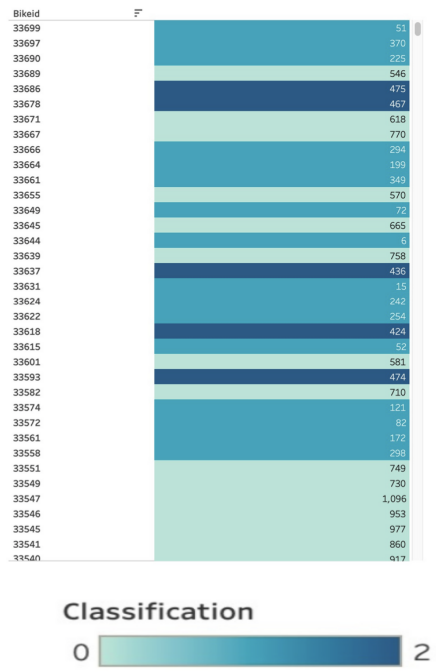


Figura 3. Visualización de clasificación de estado de Bicicletas Citibike

En la figura 3 se puede observar cómo en base al kilometraje, las bicicletas están o en bajo desgaste (0), desgaste moderado (1) o alto desgaste (2).

No Supervisado

Se utilizaron modelos no supervisados para analizar si hay estaciones que se puedan consolidar y si hay alguna estación que valga la pena que exista. Para definir si hay estaciones que se puedan consolidar, se utilizó un clustering jerárquico con K-means para agrupar a las estaciones con menor distancia. A partir de esto, se analizó la distancia entre el grupo más pequeño de clusters y se analizó el uso de cada una de ellas. Con las condiciones de si tenían poca distancia y poco uso, se debería considerar consolidar las estaciones.

```
The following stations can be consolidated:
- 8D QC Station 01
- India St & East River
- Leonard St & Meeker Ave
- Broadway & W 25 St
- Penn Station Valet
- Elizabeth St & Hester St
- NYCBS Depot - PIT
- Lafayette St & Jersey St N
- Prototype Lab Motivate Headquarters
- Soissons Landing
- Sands St & avy St
- Gowanus Tech Station
- Montague St & Clinton St
- Lefferts Pl & Franklin Ave
- Expansion Warehouse 333 Johnson Ave
- Riverside Dr & W 72 St
- Riverside Dr & W 89 St
- E 71 St & 2 Ave
- MLSWKiosk
- Convent Ave & W 129 St
- E 98 St & Lexington Ave
- 21 St & 41 Ave
- 27 Ave & 4 St
- 8D OPS 01
- 333 Johnson TEST 1
```

Figura 4. Las estaciones que vale la pena que sean consolidadas

Para saber el tipo de clientes que se tienen se creó otra clusterización geográfica y jerárquica. Primero se creó la geográfica y después se analizó cuáles son los atributos de los usuarios que predominan en cada cluster. Esto con el objetivo de saber cuáles son los tipos de clientes que tiene Citibike.

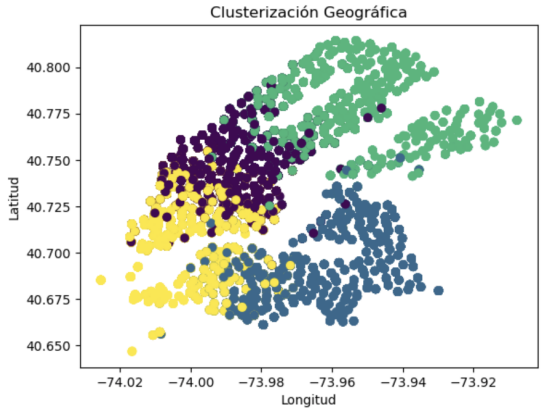
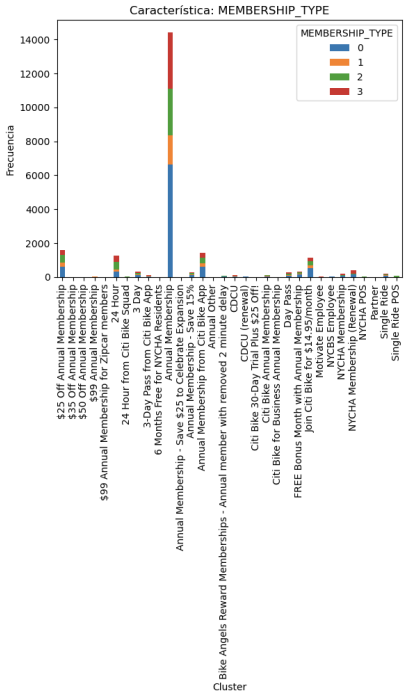


Figura 5. Clusterización geográfica por medio de la longitud y latitud



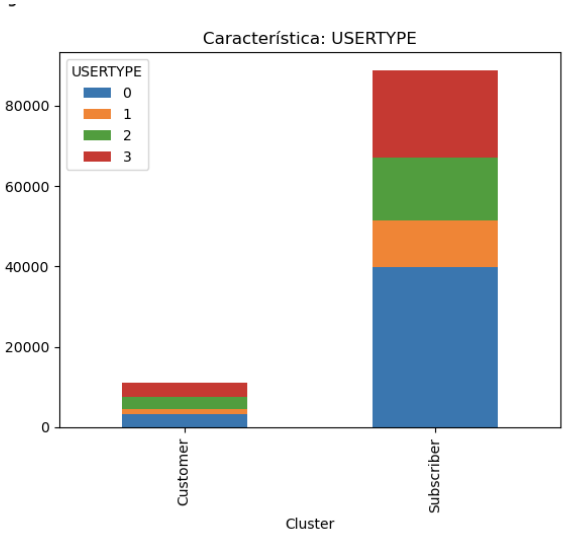


Figura 7. Tipos de usuarios que predominan en los clusters

Figura 6. Tipos de miembros que predominan en los clusters.

En las figuras anteriores se puede ver los atributos que predominan en cada cluster. Esto con el fin de tener un mejor entendimiento de con qué tipo de clientes se trabaja para poder adaptar las estrategias de marketing de manera más efectiva, ya que cada tipo de cliente tiene diferentes necesidades, preferencias y comportamientos. Asimismo, es posible identificar oportunidades de mejora y personalización en la oferta de productos o servicios para atraer a más clientes y aumentar el alcance de la empresa.

Evaluación

Métodos de verificación utilizados.
Para nuestros modelos no supervisados, utilizamos un método de verificación llamado el método del codo. El método del codo es una técnica utilizada para determinar el número óptimo de clusters en un algoritmo de clustering, como K-means. Este método tiene como objetivo encontrar el número de clusters que nos van a dar la mejor ganancia en la varianza explicada. Se basa en el concepto de que al aumentar el número de grupos, la varianza dentro de cada grupo tiende a disminuir.

Implementación

Debido a que no se tuvo contacto directamente con Citibike, sino que se utilizó su base de datos con lo que permiten que el público pueda realizar ejercicios como este, no se puede observar claramente la fase de implementación. Sin embargo, en esta fase se comunicaría los hallazgos con Citibike para que ellos puedan tomar mejores decisiones en base a el análisis de datos para promover el crecimiento de la empresa.

En este caso, igual se pudo sacar diferentes hallazgos y conclusiones relevantes como fue visto en los modelos anteriores. Asimismo, dentro de la fase de exploración de datos se

hallaron oportunidades de mejora para la recopilación de los datos (estaciones con nombres mal escritos, longitudes y latitudes erróneas, entre otras). Cómo también ideas para analizar los datos para maximizar las ventas, cómo hallar cuáles son las rutas más utilizadas y cuál es el inventario necesitado por cada estación.

Conclusión

En un mundo impulsado por la información, el análisis de datos estructurados se ha convertido en la clave para desbloquear el potencial de crecimiento óptimo de cualquier empresa. Al aprovechar el poder de los datos, se pueden revelar patrones ocultos, identificar oportunidades y tomar decisiones estratégicas fundamentadas. Los proyectos de análisis de datos estructurados son la piedra angular de la excelencia empresarial en la era digital, impulsando la innovación, mejorando la eficiencia y brindando una ventaja competitiva insuperable. Citibike, al tomar decisiones estratégicas en base a los hallazgos, podrá fomentar el crecimiento de su empresa alineado con su misión, visión y objetivos.

Fuentes de Referencia

- Citibike NYC. (n.d.). Citi Bike NYC. Retrieved Month Day, Year, from https://citibikenyc.com
- Géron, A. (2019). Hands-on machine learning with scikit-learn, keras and tensorflow: Concepts, tools, and techniques to build Intelligent Systems. O'Reilly.
- Sngular. (2022). Sngular Ventures. Retrieved June 9, 2023, from https://ventures.sngular.com