

PROBLEMÁTICA

Una empresa de transporte de pasajeros, actualmente enfocada en micros de media y larga distancia, busca expandirse al sector de transporte de pasajeros con automóviles. Con una visión de futuro sostenible y alineada con las tendencias de mercado, la empresa desea evaluar la relación entre los vehículos particulares y la calidad del aire, así como la contaminación sonora. El objetivo es estudiar la viabilidad de incorporar vehículos eléctricos a su flota, ya sea en su totalidad o en parte.

INTRODUCCIÓN

La empresa "NewYorkDrive" ha solicitado a nuestro equipo; realizar un estudio para respaldar su proceso de toma de decisiones, respecto a la implementación de una nueva línea de transporte. El enfoque principal es garantizar que esta nueva línea sea rentable, eficiente y respetuosa con el medio ambiente. Se llevará a cabo una investigación exhaustiva utilizando datos de alta calidad de fuentes clave en la ciudad de Nueva York para analizar las características fundamentales que influyen en la toma de decisiones.

EQUIPO DE TRABAJO

Carlos Hadla: Data Engineer
Natalia Queija: Data Analytic
Jair Alarcón: Machine Learning
Ismael Flores: Especialista en GCP

METODOLOGÍA DE TRABAJO

En este proyecto, adoptaremos la metodología ágil Scrum como marco de trabajo principal, respaldado por la plataforma colaborativa Trello para la gestión y seguimiento de las tareas. Scrum nos permitirá abordar las complejidades del proyecto mediante iteraciones cortas y entregas incrementales, enfocándonos en la flexibilidad, adaptación constante y colaboración entre los equipos. Utilizaremos Trello como una herramienta visual para organizar, priorizar y monitorear el progreso de las tareas en tiempo real, asegurando una comunicación fluida y una visión clara del avance del proyecto para todo el equipo

OBJETIVO GENERAL

Dado que esta unidad de negocio sería nueva, se busca realizar un análisis preliminar del movimiento de taxis en la ciudad de Nueva York. El objetivo es establecer un marco de referencia que permita tomar decisiones informadas durante el proceso de implementación de la nueva línea de transporte.

ALCANCE

En respuesta a la problemática global del cambio climático, que afecta a todas las poblaciones, la empresa "NewYorkDrive" tiene la intención de introducir una flota de vehículos eléctricos en la ciudad de Nueva York. Este proyecto considerará aspectos clave como energías alternativas, impacto ambiental y rentabilidad. Se enfocará especialmente en la viabilidad, cumpliendo con los acuerdos establecidos por la ciudad para el cambio tecnológico en el transporte, con el objetivo de reducir la huella de carbono y generar un impacto positivo en la comunidad a corto y mediano plazo.

STACK TECNOLÓGICO

- **Google Colab:**

Justificación: Google Colab proporciona un entorno de desarrollo de Python en la nube basado en Jupyter Notebooks, que permite colaborar en tiempo real en la escritura y ejecución de código Python. Al ser parte de Google Drive, facilita el acceso a conjuntos de datos almacenados en la nube y ofrece recursos de hardware como GPU y TPU de forma gratuita, lo que es especialmente útil para ejecutar modelos de Machine Learning que requieren potencia computacional adicional. Además, su integración con otros servicios de Google Cloud Platform simplifica la carga y el procesamiento de datos directamente desde Colab a otras herramientas como BigQuery o Cloud Storage. Esta herramienta es valiosa para la investigación, desarrollo y prototipado rápido de algoritmos antes de implementarlos en un entorno de producción en GCP.

- **Python:**

Justificación: Python es un lenguaje de programación versátil y poderoso para el análisis de datos. Su gran cantidad de bibliotecas (como Pandas, NumPy, Matplotlib) y su facilidad para manejar datos lo convierten en una opción excelente para manipular y analizar conjuntos de datos complejos.

- **Google Cloud Platform (GCP):**

Justificación: GCP proporciona una infraestructura escalable y confiable para alojar aplicaciones y servicios en la nube. Ofrece una variedad de herramientas para el almacenamiento, procesamiento y análisis de datos, lo que facilita el manejo de grandes volúmenes de información de forma eficiente y segura.

- **Cloud Function:**

Justificación: Las Cloud Functions permiten ejecutar código de manera automatizada en respuesta a eventos en la nube sin preocuparse por la administración de servidores. Esto es útil para realizar tareas específicas en el flujo de trabajo del análisis de datos, como procesamiento por lotes o acciones automatizadas.

- **Google Cloud Scheduler:**

Justificación: Google Cloud Scheduler permite programar y automatizar tareas en GCP, ejecutando servicios de manera periódica o según horarios específicos. Es útil para la

programación de trabajos de procesamiento de datos, activación de flujos de trabajo y gestión de tareas recurrentes en la nube.

- **Google Cloud Storage:**

Justificación: Google Cloud Storage proporciona un almacenamiento escalable y duradero en la nube. Permite almacenar grandes cantidades de datos de forma segura y acceder a ellos de manera eficiente. Es ideal para el almacenamiento de datos de análisis, archivos de respaldo, contenido multimedia y conjuntos de datos estructurados o no estructurados.

- **BigQuery:**

Justificación: BigQuery es un servicio de almacenamiento y análisis de datos completamente administrado que permite consultar grandes conjuntos de datos de manera rápida y eficiente. Su escalabilidad y capacidad para manejar grandes volúmenes de datos lo hacen ideal para realizar consultas y análisis complejos.

- **Power BI:**

Justificación: Power BI es una herramienta de visualización de datos que permite crear informes interactivos y paneles dinámicos. Al integrarse con otros servicios de datos como BigQuery, facilita la creación de visualizaciones efectivas para presentar y comunicar los resultados del análisis.

- **Vertex AI:**

Justificación: Vertex AI es una plataforma de Machine Learning completamente administrada que permite el desarrollo, entrenamiento y despliegue de modelos de manera eficiente. Su integración con otras herramientas de GCP simplifica el proceso de implementación de modelos de aprendizaje automático en la infraestructura de la nube.

VIDA DEL DATO

Obtención del dato:

- El proceso de adquisición de datos se llevó a cabo a partir de diversas fuentes para recopilar información clave. Se emplearon técnicas de web scraping desde el portal oficial del gobierno de la ciudad de Nueva York para extraer datos relacionados con los taxis amarillos, verdes y servicios de transporte de aplicaciones como Uber, Lyft, entre otros. Además, se obtuvo un dataset consolidado que abarcaba los promedios mensuales de estos datos desde el año 2013 hasta la actualidad.
- Además de la obtención de datos mediante web scraping, se utilizó el acceso a APIs para obtener información valiosa. Esto incluyó datos sobre la contaminación a lo largo de varios años, así como detalles específicos sobre los tipos de vehículos y su nivel de emisión en distintas circunstancias, ya sea por milla recorrida, por hora, etc. Otro dataset obtenido a través de API fue el relacionado con los 'medallion', que contiene información detallada sobre las acreditaciones de taxis en Nueva York.

EDA (Análisis Exploratorio de Datos):

- Cada dataset se sometió a un análisis minucioso. Se llevó a cabo una revisión exhaustiva de cada conjunto de datos para entender su estructura, la calidad de la información, la presencia de valores atípicos o nulos, así como para determinar qué variables eran relevantes para los objetivos del proyecto. Para facilitar este proceso, se creó un diccionario detallado para cada dataset, documentando todas las variables y sus características, lo que permitió tomar decisiones informadas sobre qué datos retener y cómo manejarlos en las siguientes etapas del proceso.

Automatización:

- Una de las piezas clave en este proyecto fue la automatización del flujo de datos. Esto se logró utilizando las herramientas de Google Cloud Platform (GCP) para establecer un sistema automatizado robusto. Se programaron cronogramas en GCP que activaban disparadores (triggers) en Cloud Functions para descargar datos desde las diferentes fuentes. Estos datos se almacenan inicialmente en Google Cloud Storage (GCS). Posteriormente, se implementó un segundo disparador en Cloud Function para iniciar la limpieza de los datos recién descargados. Esta limpieza incluyó la eliminación de datos redundantes, la corrección de datos incorrectos o faltantes, y la agrupación según las decisiones tomadas durante la exploración de datos previa.
- El proceso de limpieza dejó los datos listos y preparados para su carga en BigQuery de GCP. Los archivos limpios se almacenaron en un nuevo bucket de GCS en espera de ser agregados a las tablas de BigQuery, preparadas y estructuradas para recibir esta información.

Data Warehouse:

- La transferencia de datos a BigQuery se llevó a cabo utilizando DataTransfer, también configurado en GCP. Este proceso se ejecutó minutos después de que se completaran todas las operaciones de limpieza y preparación de datos en GCS. Una vez que los datos se transfirieron a BigQuery, se organizaron en sus tablas correspondientes.
- Posteriormente, se realizaron actualizaciones en los diccionarios de datos para reflejar las nuevas tablas y la estructura en BigQuery. Esto facilitó enormemente la implementación de técnicas de machine learning, ya que se contaba con un marco de datos bien definido y organizado. Además, se sentó la base para la construcción de un dashboard interactivo que permitiría visualizar los datos actualizados mes a mes en tiempo real.
- Este enfoque meticuloso y automatizado no solo garantizó la adquisición de datos de múltiples fuentes, sino que también permitió su limpieza y organización eficientes. La estructura establecida en BigQuery y la actualización constante de los datos sentaron las bases para futuros análisis, implementaciones de machine learning y presentaciones visuales para la toma de decisiones fundamentadas.

Machine Learning:

- Creación, entrenamiento y despliegue de modelos predictivos o analíticos para extraer información útil de los datos, y así resolver distintas problemáticas.

Dashboard:

- Diseño y desarrollo de un dashboard interactivo que presenta de manera clara y concisa la información derivada del análisis de datos.

KPIs

Estos son los KPIs que estaremos desarrollando a lo largo del proyecto, esto con fines de realizar un seguimiento mensual y anual con diversos datos para llegar a conclusiones más concretas y confiables.

1. Retorno de inversión (ROI) esperado por la implementación de la flota de vehículos eléctricos:

El ROI se calcula teniendo en cuenta el beneficio neto obtenido por la inversión dividido por el costo de la inversión, expresado como un porcentaje.

$$\text{ROI} = \frac{\text{Beneficio Neto}}{\text{Costo de la Inversión}} \times 100$$

Podemos decir, por ejemplo, que el retorno esperado por la implementación de vehículos eléctricos va a ser de un 20% luego del período de 6 años.

También hacer lo mismo para vehículos a combustión interna para hacer una comparativa.

2. Reducción de emisión de carbono:

La Reducción de Emisiones (%) es un indicador clave de rendimiento que evalúa la efectividad de las medidas adoptadas para reducir las emisiones generadas por los vehículos de combustión. La fórmula es la siguiente:

$$\text{Reducción de Emisiones} = \frac{\text{Reducción de Emisiones}}{\text{Emisiones de Vehículos de Combustión}} \times 100$$

Al 50% mensual

3. KPI de Eficiencia de Rutas:

Nos enfocamos en evaluar la eficiencia de las rutas de los taxis, con el objetivo de optimizar la distancia promedio recorrida por cada viaje.

$$\text{Distancia promedio por viaje} = \frac{\text{Sumatoria de todas las distancias recorridas en viajes}}{\text{Número total de viajes}}$$

Podemos decir que buscamos reducir la distancia promedio por viaje en un 10% en los próximos seis meses, implementando estrategias de optimización de rutas y proporcionando orientación a los conductores para mejorar la eficiencia del servicio.

4. Calidad del aire en zonas de servicio

Relación entre la calidad del aire y la cantidad de viajes en diferentes zonas.

Revelar si existe alguna relación entre la cantidad de viajes de taxis en una zona específica y los niveles de contaminación del aire en esa área. Puede medirse en periodos de tiempo mensuales/ anuales.

$$\text{Calidad del aire en zonas de servicio} = \frac{\text{Nivel de contaminación Anterior} - \text{Nivel contaminación Actual}}{\text{Nivel de contaminación Anterior}} \times 100$$

Reducción del nivel de contaminación del aire en una zona específica

Se mediría la disminución porcentual en los niveles de contaminantes en comparación con un punto de referencia previo.

5. Eficiencia en Costos de Combustible

La Eficiencia en Costos de Combustible es un indicador clave que evalúa el ahorro económico derivado de la utilización de vehículos eléctricos en comparación con los vehículos de combustión en términos de costos operativos. La fórmula para calcular este KPI es la siguiente:

$$\text{Eficiencia en costo de combustible} = \frac{\text{Costos operativos promedio por vehículo eléctrico} - \text{Costos operativos promedio por vehículo de combustión}}{\text{Costos operativos promedio por vehículo eléctrico}} \times 100$$

CONCLUSIÓN

En resumen, este proyecto representa un paso significativo hacia la consecución de nuestros objetivos de implementar una solución de transporte sostenible y eficiente para la

empresa 'NewYorkDrive'. Con un equipo dedicado, tecnologías innovadoras y una metodología ágil, nos embarcamos en este viaje para transformar la movilidad urbana. Estamos comprometidos con la búsqueda de soluciones que no solo impulsen la rentabilidad del negocio, sino que también fomenten un impacto positivo en el medio ambiente y en la calidad de vida de la comunidad. El uso estratégico de tecnologías, la gestión eficaz de los datos y la definición de KPIs claros nos permitirán medir nuestro progreso y tomar decisiones informadas a lo largo del proyecto. Estamos emocionados por el camino que tenemos por delante y ansiosos por los resultados que lograremos juntos. ¡Gracias por su atención y apoyo en este proyecto transformador!