Proyecto Movilidad Eléctrica para Taxis en Nueva York, EEUU.

Introducción:

Los taxis y servicios como Uber, que han revolucionado la forma en que nos movemos, generan a la vez una avalancha de datos que, bien aprovechados, pueden ser la clave para un futuro más verde. Este proyecto busca expandir la flota hacia la movilidad eléctrica, tomando decisiones informadas basadas en un análisis profundo del movimiento de los taxis, su relación con la contaminación y la viabilidad de la electrificación. Este proyecto combina innovación, análisis y responsabilidad social para transformar el panorama del transporte en la ciudad, reduciendo la huella de carbono, mejorando la imagen pública y optimizando costos.

Planteo de Objetivos.

Objetivos Principales:

Descifrando la viabilidad de los autos eléctricos en Nueva York:

Abordar la pregunta central: ¿Es viable implementar una flota de autos eléctricos para el transporte de pasajeros en la ciudad de Nueva York?

Brindar información sólida para la toma de decisiones estratégicas para la implementación de la flota de autos eléctricos.

Empoderando la toma de decisiones informadas:
Realizar un análisis exhaustivo del problema, considerando aspectos ambientales, económicos y logísticos.
Generar insights accionables que guíen a la empresa hacia un camino sostenible y rentable.

Objetivos Técnicos.
Construyendo la autopista de datos en la nube:
Desarrollar un pipeline y arquitectura de datos robustos en la nube para procesar y almacenar de manera eficiente la gran cantidad de datos disponibles.

Dashboard interactivo:
Crear un dashboard interactivo y fácil de usar que presente información relevante para la toma de decisiones sobre la implementación de la flota eléctrica.
Facilitar la comprensión y el análisis de datos complejos para stakeholders de todos los niveles.

Modelo de machine learning No supervisado/ no predictivo:
Entrenar y desplegar un modelo de machine learning para identificar las ubicaciones óptimas para las estaciones de carga de vehículos eléctricos.
Optimizar la ubicación de las estaciones de carga para maximizar la eficiencia, la rentabilidad y la satisfacción del cliente.

Productos.
Un panel de control para la era eléctrica:
Desarrollar un dashboard interactivo que presente información crucial para la fase de implementación de la flota eléctrica.
Incluir análisis espaciales, temporales y técnicos/logísticos para una visión completa del panorama.

El mapa del futuro: Un modelo de machine learning para la ubicación de estaciones de carga:

Implementar un modelo de machine learning preciso para identificar las ubicaciones óptimas para las estaciones de carga de vehículos eléctricos.

Equipo:
- Rodrigo Nahuel Castro: Data Engineer
- Maximiliano Javier Lizarraga: Cloud Engineer
- Lucía Teresa Escobedo Villafane: Data Analyst
- Amelia Cristina Herrera Briceño: Data Analyst
- Nicolás Hernández Díaz: Machine Learning Engineer

Alcance del Proyecto:
Área de movilidad: Logística para soluciones de movilidad de pasajeros.
Área geográfica: Ciudad de Nueva York, EEUU.
Fuentes de datos: NYC Open Data, TLC Trip Record Data, EPA NY, entre otras.
Período de tiempo: 2019 - 2023.

KPIs:

1.- Tasa de Cambio en la Demanda de Taxis:
Objetivo: Medir el cambio porcentual en la demanda de taxis mes a mes.
Fórmula: ((DemandaActual − DemandaAnterior) / DemandaAnterior × 100)
Meta: Lograr un crecimiento mensual constante en la demanda de taxis de al menos un 5%.

2.- Reducción Porcentual de Emisiones de CO_2:
Objetivo: Calcular la reducción potencial de CO_2 al implementar vehículos eléctricos.
Fórmula: ((EmisionesCO_2vehiculoConvencional − EmisionesCO_2vehiculoElectrico) / EmisionesCO_2vehiculoConvencional × 100)
Meta: Alcanzar una reducción del 10% anual en las emisiones de CO_2 por milla.

3.- Porcentaje de Crecimiento en la Base de Usuarios de Servicios de Taxi:
Objetivo: Medir el crecimiento en el número de usuarios de servicios de taxi.
Fórmula: ((NumerodeUsuariosalfinaldelPeriodo − NumerodeUsuariosaliniciodelPeriodo) / NumerodeUsuariosaliniciodelPeriodo)
Meta: Lograr un incremento del 5% en la base de usuarios al final del período de análisis.

4.- Accesibilidad espacial de las estaciones de carga:
Objetivo: Medir la accesibilidad espacial de las estaciones de carga desde los puntos finales de los viajes de vehículos eléctricos.
Fórmula: [D = \frac{1}{n} \sum_{i=1}^{n} \min_{j \in [1, m]} \text{Distancia}(PuntoFinal_i, EstacionDeCarga_j)]
Meta: Reducir la distancia promedio a la estación de carga más cercana en un 10%.

Implementación Detallada.
Configuración de Databricks:
Las actividades de ingesta y transformación de datos, se realizaron utilizando Databricks.

Transformación de datos en Databricks:
Se utilizó Databricks para realizar transformaciones avanzadas (ETL) en los datos utilizando PySpark.
Se transformaron los datos en un formato adecuado y almacenados en Azure SQL Database. De acuerdo a la **arquitectura Medallion**, se estructuran los datos en tres capas: **bronce** (datos originales, en bruto, sin transformaciones significativas), **plata** (aquí se validan los datos; se aplican reglas de calidad y limpieza para asegurar que los datos sean confiables y precisos) y **oro** (en esta capa, los datos están listos para análisis avanzados). Esto busca

garantizar la calidad de los datos a medida que avanzan desde su estado original hasta su uso en análisis y toma de decisiones.

Almacenamiento de datos curados en Azure SQL Database:
Configurar Azure SQL Database para recibir y almacenar datos curados.
Asegurar que los datos están limpios, transformados y listos para análisis y modelado.

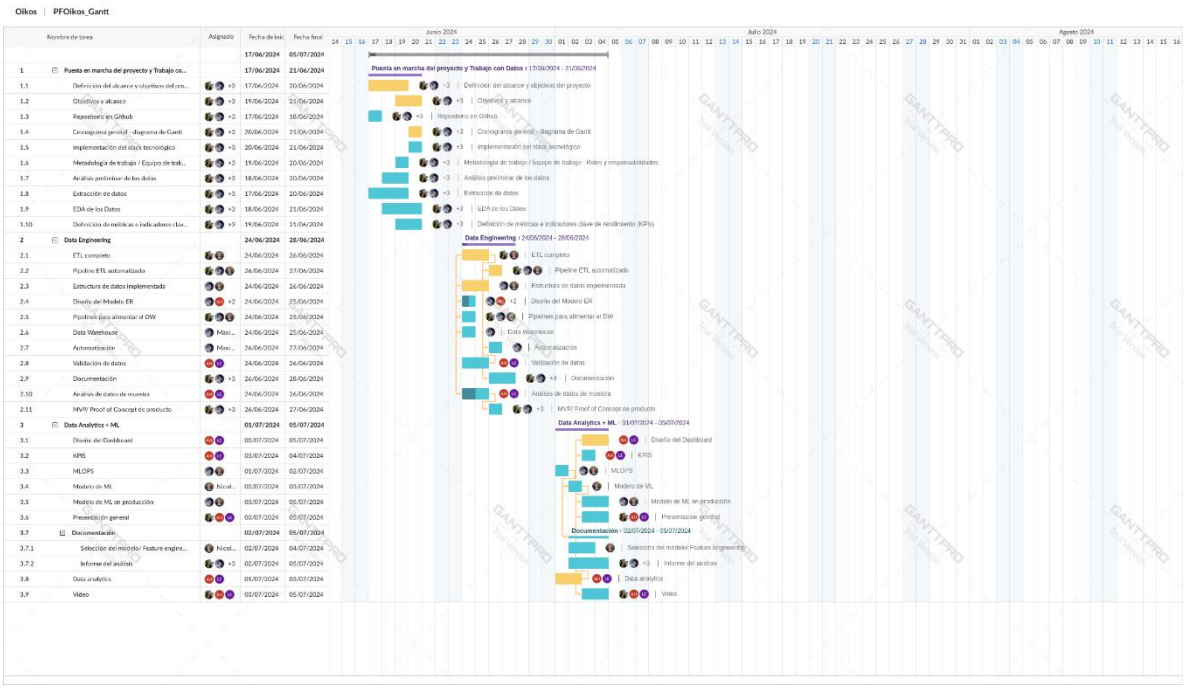
Construcción y despliegue de modelos en Azure Machine Learning:
Usar los datos almacenados en Azure SQL Database para entrenar modelos de machine learning en Azure DataBricks. Desplegar los modelos entrenados y procesar la información, encontrar patrones en los datos y mejorar las operaciones y decisiones empresariales.

Visualización de datos y resultados en Power BI:
Crear dashboards e informes en Power BI para visualizar datos y resultados de modelos de machine learning.

Stack Tecnológico:
Databricks: Transformación avanzada de datos utilizando PySpark.
Azure SQL Database: Almacenamiento de datos estructurados.
Power BI: Visualización y creación de dashboards interactivos, para la presentación del análisis de datos.

Data: Fuentes de datos para análisis y modelado en NYC Open Data & TLC Trip Record (además de los datos entregados desde un comienzo).

Carta Gantt: Movilidad en la Gran Manzana: Taxis, autos compartidos y la carrera hacia la sostenibilidad



1.- Objetivo: Apoyar a la empresa de transporte de pasajeros en la toma de decisiones informadas sobre la viabilidad de implementar vehículos eléctricos en su flota.

Sprint 1 (1 semana)

Semana 1

- Definición del alcance y objetivos del proyecto.
- Repositorio GITHUB.
- Cronograma general, Carta Gantt.
- Implementación de stack tecnológico.
- Metodología de trabajo, equipo de trabajo, repartición de tareas.
- Análisis preliminar de los datos para comenzar a pensar en las KPIs y modelo ML (EDA y definición de métricas).

Herramientas:

- Lenguajes de programación para ciencia de datos (Python)
- Bibliotecas de análisis de datos (Pandas, NumPy, scikit-learn)
- Herramientas de visualización de datos (Matplotlib, Seaborn, Plotly)
- Entornos de desarrollo integrados (Jupyter Notebook, Studio visual code)
- Plataformas de cloud computing (Microsoft Azure)
- Software de gestión de proyectos (Ganttpro)
- Herramientas de comunicación (App.gather.town, Zoom)
- Presentaciones finales (PowerBI)

Recursos adicionales:

- DATASET DE HENRY
- NYC Taxi & Limousine Commission: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page
- Kaggle dataset on CO2 emissions by country and year: https://www.kaggle.com/datasets/debajyotipodder/co2-emission-by-vehicles
- NYC sound dataset: https://on.nyc.gov/OpenData
- NYC air quality dataset: https://www.nyc.gov/site/doh/health/health-topics/air-quality.page
- Microsoft Azure: https://azure.microsoft.com/en-us

2.- Objetivo: Completar la arquitectura, el ETL, la estructura de datos, el modelo de entidad-relación, la automatización, la documentación, el MVP del proyecto, esbozo del dashboard y del modelo de Machine Learning.

Sprint 2 (1 semana):

Semana 2

- Revisión de ETL de los datos completos.
- Automatización de pipeline de ETL.
- Implementación de la estructura de los datos.
- Diseño del Modelo ER.
- Pipelines para alimentar al DW.
- Data warehouse.
- Validación de los datos.
- Documentación de cada punto anterior.
- Análisis de datos de muestra.
- Prueba del concepto de producto MVP.

Herramientas:

- **Lenguajes de programación:** Python.
- **Bibliotecas de análisis de datos:** Pandas, NumPy, Scikit-learn.
- **Herramientas de visualización de datos:** Matplotlib, Seaborn, Plotly.
- **Entornos de desarrollo integrados:** Jupyter Notebook, SVC.
- **Plataformas de cloud computing:** Microsoft Azure.
- **Software de gestión de proyectos:** Ganttpro.
- **Herramientas de comunicación:** App.gather.town, Zoom.
- **Herramientas de ETL:** Pandas, Numpy, etc.
- **Herramientas de modelado de datos:** K-means, clustering.
- **Herramientas de documentación:** Readme.md, Markdown, pdf.
- **Herramientas de dashboarding:** Power BI.
- **Herramientas de Machine Learning:** Scikit-learn.

3.- Objetivo: Completar el análisis de datos, el diseño del dashboard, la implementación de los KPIs, el MLOps, la puesta en producción del modelo de Machine Learning, la presentación general, la documentación final y la selección del modelo y el Feature Engineering.

Sprint 3 (1 semana):

Semana 3

- Diseño del dashboard.
- KPIs.
- MLOPs.
- Modelo de ML.
- Modelo de ML en producción.
- Presentación general (informe del análisis).

Herramientas:

- **Herramientas de análisis de datos:** Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, Plotly
- **Herramientas de dashboarding:** Power BI
- **Herramientas de MLOps:** Scikit-learn
- **Herramientas de presentación:** PowerBI, Github.
- **Herramientas de documentación:** Readme.md, Markdown, pdf
- **Recursos implementados Gestión del proyecto:** Google meet.

- **EDA, ETL, SQL BD:** Python, Beautiful Soup, Pandas, Matplotlib, Seaborn, PySpark, Azure SQL.
- **Business Intelligence & Machine Learning:** Plotly, PowerBI, Scikit-learn.
- **Cloud:** Microsoft Azure.//.