Sprint #1: Puesta en marcha del proyecto y Trabajo con Datos

**Análisis del proyecto**

La ciudad de Nueva York, es una de las ciudades con mayor movilidad vial en el mundo, actualmente existen múltiples medios de transporte de pasajeros que permiten a los usuarios arribar a sus destinos, sin embargo, el colapso vial, aunado a la contaminación ambiental están obligando al sector a buscar soluciones de mejora que permitan elevar la eficiencia en el transporte, reduciendo costos y aumentando utilidades sin descuidar el cuidado al medio ambiente.

Nuestro proyecto se enfoca en este sector del transporte de pasajeros, con la intención de proponer estrategias que permitan a una empresa privada realizar una fuerte inversión para entrar a este mercado, analizando sus tendencias, escenario actual, fortalezas y debilidades, de tal forma que, permitan ejecutar un plan de acción que brinde los mejores resultados posibles.

Para ello, abordaremos la información desde un panorama general, donde se extraerá información de aquellos medios de transporte de pasajeros existentes en la ciudad, se analizarán las rutas actuales, el consumo que realizan los usuarios de estos servicios, utilidades generadas, horas picos de solicitudes, zonas y puntos estratégicos de inicio y fin de viajes, la huella de carbono generada por los automóviles de combustión, uso de autos eléctricos, entre otros.

Con dicha información analizada, se buscará brindar a la empresa que desea realizar la inversión, los insights necesarios que le permitan tomar la mejores decisiones para su entrada a mercado.

**Objetivos**

* Recopilar, depurar y disponibilizar la información.

En primer lugar, nos enfocaremos en la recopilación y depuración de datos de diversas fuentes. Esto incluirá información sobre la operación de taxis en la ciudad de Nueva York, las emisiones de CO2 asociadas a diferentes modelos de automóviles, la disponibilidad de estaciones de carga para vehículos eléctricos y otros conjuntos de datos relevantes.

* Desarrollar y presentar un Dashboard Interactivo para mejorar la visualización y comprensión de datos.

Presentaremos un Dashboard interactivo que permitirá a los interesados visualizar análisis de valor en función de los KPI (Indicadores Clave de Desempeño) que hemos propuesto.

* Comparar los costos y ganancias obtenidos por los autos de combustión y los autos eléctricos.

Presentaremos un análisis que muestre las utilidades que se pueden generar en los autos eléctricos y cómo es que sus costos de recorrido se reducen en comparación con los autos de combustión, apoyando la decisión de migrar de la combustión a la electricidad.

**Alcance**

En esta primera entrega, se pretende terminar un análisis preliminar del sector de transporte de pasajeros en la ciudad de Nueva York, que permita conocer el estado actual del sector, utilizando servicios de base de datos y Dashboards interactivos que permitan explotar la información de los análisis generados.

Se realizará la aplicación de un modelo de Machine Learning, con la intención de analizar las rutas más frecuentes utilizadas por los usuarios y nos ayude a generar nuevos recorridos que disminuyan los colapsos viales, contaminación ambiental y auditiva.

**KPIs Propuestos**

* Aumentar en un 5% las ventas promedio de los viajes en comparación con el mes anterior.
* Aumentar en un 10% el ingreso en comparación con el semestre anterior.
* Disminuir en un 40% las emisiones de CO2 por año en el servicio de taxis.
* Reducir en un 10% el gasto de combustible por año en comparación con el año anterior.

**Stack Tecnológico**

Se definen las herramientas tecnológicas que se utilizarán en nuestro proyecto, entre ellas destacan:

* Python: Es el lenguaje de programación para el análisis de información y aplicación de modelos de Machine Learning, por excelencia.
  + Pandas: Uitlizado en la creación de dataframes y manipulación de archivos csv.
  + Numpy: Se utiliza para el análisis estadístico de los datos
  + Seaborn:
  + Folium: Utilizada para la visualizción de mapas interactivos y geolocalizaciones.
  + Request: Se utiliza para realizar solitiudes HTTP
  + Os: Utilizada en la manipulación de archivos y rutas del sistema operativo
  + Matplotlib: Utilizada en la creación de los gráficos establecidos en el EDA
  + Nltk: Librería para la aplicación y análisis de modelos de Machine Learning.
  + Beautiful soup: Necesario para realizar el Web Scrapping y extracción de información
* MySQL Server workbench: Utilizada en la creación de bases de datos y administración de la información.
* Trello: Es nuestro organizador de tareas
* PowerBI: Necesario para la creación y presentación de Dashboards interactivos
* Google Cloud:
* Render:

**Metodología de trabajo**

Se utilizará una metodología ágil de tipo Scrum, en la que de lunes a viernes nos vamos a reunir con nuestro (MAXI), con la intención de identificar cuales son las tareas que está realizando cada uno de los integrantes del equipo, cuáles son las próximos pasos a realizar y si es que existe alguna actividad que esté generando complicaciones en el avance para que sea atacada en el momento.

**Diseño detallado**

**PENDIENTE DE TERMINAR**

**Equipo de trabajo - Roles y responsabilidades**

El equipo de trabajo se compone de cinco integrantes, los cuales asumirán los siguientes roles y responsabilidades:

Analisis de datos: Joyce Ruiz, Marco Delgado y Jovanni Escamilla.

Data Engineers: Joyce Ruiz, Jovanni Escamilla, Lucas Salzotto, Matias Barna

Machine Learning: Matias Barna, Marco Delgado, Lucas Salzotto

**Cronograma general de actividades – Gantt**

**PENDIENTE AGREGAR DIAGRAMA**

**Solución propuesta**

Deben detallar qué tareas harán para cumplir los objetivos de trabajo propuestos previamente y cómo lo harán (metodologías de trabajo, forma de organización, distribución de tareas, roles de cada uno dentro del equipo, etc). También, deben detallar qué productos surgirán de su trabajo y en qué etapa los presentarán, teniendo en cuenta los requerimientos generales (entregables esperados) para cada etapa del proyecto.

A su vez, deben realizar una estimación de tiempo para cada tarea, contemplando los tiempos de ejecución globales y los hitos previstos para cada semana; y plasmar esa estimación en un diagrama de Gantt.

Una parte muy importante de la solución propuesta, es con qué herramientas (**stack tecnológico**) van a realizar la arquitectura del proyecto. Para esto, lo que van a tener que hacer es seleccionar una pequeña porción de los datos que disponen y realizar un proceso de limpieza y transformación utilizando las herramientas que planean implementar. Esto les dará una idea de cómo funcionarán en el proyecto completo y les permitirá tener un mejor abordaje para futuras tareas. Hay que tener en cuenta que, como este ítem va a ser una presentación previa de lo que van a trabajar en el segundo sprint, el PO puede dar el OK o determinar cuál es el mejor camino para que tomen. Esto les va a permitir adelantar trabajo de la segunda semana, ya que no se va a tener que esperar hasta la segunda demo para verificar si la arquitectura cumple con los requisitos del PO.

Finalmente, como en Data es muy importante trabajar con datos de calidad, deberán incluir en su informe un análisis sobre los datos con los que van a trabajar (metadatos), detallándolos lo más posible: fuentes y confiabilidad de las mismas, qué representa cada columna de cada dataset, tipos de datos, método de adquisición, fecha de adquisición y última actualización, etc.

Entregables

* Documentación
* Stack elegido y fundamentación