DataMinds Solutions

Desbloqueando el Potencial Empresarial a través de Ingeniosas Ideas e Innovación en Big Data

ÍNDICE

1. Quiénes somos:

DataMind Solutions nació de la visión de tres jóvenes emprendedores apasionados por el potencial transformador de los datos en el mundo empresarial (Fer, Juan y Nico). Después de graduarse de la Academia BootCamp de Henry en ciencias de la computación y estadísticas, se dieron cuenta de que querían construir algo propio, algo que desafiara el status quo y aprovechara las oportunidades emergentes en el campo de la analítica de datos.

Con una determinación inquebrantable y una idea innovadora en mente, fundaron DataMind Solutions en un pequeño apartamento en el corazón de la ciudad de North Korea. Equipados con Pc´s antiguas, café y mate, y un sinfín de ideas, se lanzaron a la aventura empresarial con un enfoque claro: ayudar a las empresas a desbloquear el verdadero valor de sus datos.

A pesar de ser una empresa nueva en el mercado, el equipo de DataMind Solutions estaba lleno de ambición y determinación. Trabajaron incansablemente para construir una reputación sólida, centrada en la calidad, la innovación y la satisfacción del cliente. A medida que el boca a boca sobre sus habilidades y enfoque único se extendía, comenzaron a atraer la atención de empresas locales y multinacionales que buscaban soluciones de análisis de datos de vanguardia, y asi, lograron expandirse a la tan valorada ciudad de Silicon Valley, llevándose consigo a dos integrantes más quienes fueron sus compañeros en la academia (Tiago y Nacho)

Con grandes objetivos en mente y un equipo dedicado, DataMind Solutions se embarcó en su viaje empresarial con la convicción de que estaban destinados a dejar una marca indeleble en el mundo de la consultoría de datos.

2. Estructura de la empresa:

DataMind Solutions está compuesta por un equipo multidisciplinario de expertos en análisis de datos, ingeniería de datos y machine learning, con amplia experiencia en diversos sectores industriales.

Su grupo de trabajo, liderado por Fer Abraham como el Chief Executive Officer, o CEO de la empresa, comprende la cobertura en diferentes departamentos, tales como Data Engineer Department el cual es comandado por Tiago Sepulveda (Data Analyst), Business Intelligence Department a cargo de Ignacio Waukuluk (Data Engineer), Enterprise Architecture Department liderado por Nicolas Mussante (Data

Architect) y Research and Developer Department al frente de Juan Mendoza (Data Science).

3. Pedido realizado por el cliente:

Nuestro cliente, parte de un conglomerado de empresas de restaurantes y negocios afines, nos ha contratado como consultora de datos para realizar un análisis exhaustivo del mercado estadounidense. En particular, desean obtener información detallada sobre la opinión de los usuarios en dos plataformas clave: Yelp y Google Maps. El objetivo es entender cómo perciben los usuarios los distintos negocios relacionados con el turismo y el ocio, como restaurantes, hoteles y otros establecimientos afines.

Para lograr esto, se espera que realicemos un análisis de sentimientos de las reseñas en ambas plataformas, identificando tendencias y patrones en las opiniones de los usuarios. Además, el cliente busca predecir cuáles serán los rubros de los negocios que experimentaron un crecimiento o un decrecimiento en el futuro cercano, utilizando modelos predictivos basados en los datos recopilados de Yelp y Google Maps.

Además, nuestro cliente está interesado en determinar las ubicaciones óptimas para la apertura de nuevos locales de restaurantes y negocios afines, basándose en la información obtenida de las reseñas y otros datos relevantes del mercado. Esto implica realizar un análisis geoespacial y de densidad de negocios en diferentes áreas.

Finalmente, el cliente desea desarrollar un sistema de recomendación de restaurantes y otros establecimientos para los usuarios de ambas plataformas. Este sistema deberá utilizar la información recopilada de las reseñas para ofrecer recomendaciones personalizadas a los usuarios, permitiéndoles descubrir nuevos lugares basados en sus preferencias y experiencias previas.

4. Proceso de trabajo localmente:

4.1 Ingesta de archivos:

Se proporcionaron datos provenientes de reseñas en plataformas como Google Maps y Yelp, almacenados en Google Drive para su consumo. Estos datos son relevantes para abordar la problemática planteada por el cliente y son de gran

importancia para el análisis posterior a realizarse. En consecuencia, se procedió a la descarga manual para su primer acercamiento.

4.2 EDA inicial:

Utilizando la plataforma de Visual Studio Code y las herramientas como Pandas y Jupyter Notebooks en el lenguaje de programación de Python, se realizó un Análisis Exploratorio de Datos (EDA) inicial para comprender la estructura de los datos y su contenido para así evaluar las futuras transformaciones a realizar en el proceso de limpieza.

Se creó un nuevo proyecto o directorio de trabajo en Visual Studio Code para organizar los archivos y scripts relacionados con el análisis de datos. Se instaló Python en el sistema y se configuró el entorno de Python en Visual Studio Code. Se instaló la biblioteca Pandas utilizando el administrador de paquetes pip de Python para la manipulación y análisis de datos.

Se creó un nuevo Jupyter Notebook en Visual Studio Code para comenzar el análisis exploratorio de datos (EDA). El notebook se guardó en el directorio de trabajo del proyecto para facilitar el acceso y la gestión de los archivos relacionados con el análisis de datos.

En el Jupyter Notebook, se importaron las bibliotecas necesarias, incluyendo Pandas, para el análisis de datos. Se cargaron los conjuntos de datos provenientes de las reseñas en plataformas como Google Maps y Yelp utilizando Pandas, asegurando una adecuada manipulación y exploración de los datos.

4.3 ETL:

Se llevó a cabo el proceso de Extracción, Transformación y Carga (ETL) para limpiar, transformar y preparar los datos para su análisis. Durante este proceso, se observó la calidad de los datos a utilizar en el Análisis Exploratorio de Datos (EDA) previo, lo que permitió tomar decisiones como realizar diferentes transformaciones, eliminar columnas que son irrelevantes para nuestro análisis, entre otros.

La etapa de Extracción implicó la obtención de los datos de las fuentes correspondientes, como los conjuntos de datos provenientes de reseñas en plataformas como Google Maps y Yelp. Estos datos se cargaron en el entorno de trabajo utilizando las herramientas y bibliotecas adecuadas, como Pandas en Python.

En la etapa de Transformación, se aplicaron diversas técnicas para limpiar y preparar los datos para su posterior análisis. Esto incluyó la identificación y manejo de valores faltantes, la estandarización de formatos de datos, la codificación de variables categóricas, y la eliminación de duplicados. Además, se tomaron decisiones sobre qué columnas eran relevantes para el análisis y se eliminaron aquellas que no contribuían a los objetivos establecidos.

Finalmente, en la etapa de Carga, los datos transformados y preparados fueron almacenados en el formato adecuado para su posterior análisis. Esto podría implicar guardar los datos en un nuevo archivo o base de datos, o simplemente mantenerlos en la memoria para su uso inmediato en el proceso de análisis.

Este proceso de ETL aseguró que los datos estuvieran limpios, coherentes y listos para ser analizados en el siguiente paso del proyecto. Permitió una comprensión más profunda de la calidad y estructura de los datos, lo que a su vez facilitó la toma de decisiones informadas durante el análisis exploratorio y el desarrollo de modelos posteriores.

4.4 EDA final:

El Análisis Exploratorio de Datos (EDA) más detallado se llevó a cabo para profundizar en los insights obtenidos durante la fase inicial y preparar los datos para su análisis en los dashboards y el modelo de machine learning.

1. Exploración de variables clave:
    - Se examinaron en detalle las variables clave del conjunto de datos, incluyendo las reseñas de los usuarios, las calificaciones, la ubicación de los negocios, las categorías de los establecimientos, entre otros.
    - Se realizó un análisis estadístico más profundo de estas variables, incluyendo medidas de tendencia central, dispersión, y distribuciones.
2. Visualización de datos:
    - Se crearon una variedad de visualizaciones gráficas para explorar las relaciones entre diferentes variables y entender mejor los patrones y tendencias presentes en los datos.

    - Se utilizaron gráficos de dispersión, histogramas, diagramas de caja, y mapas geoespaciales, entre otros, para visualizar diferentes aspectos de los datos y descubrir insights adicionales.
3. Análisis de correlación:
    - Se evaluaron las correlaciones entre las variables para identificar posibles relaciones y dependencias entre ellas.
    - Se calculó la matriz de correlación y se visualizó utilizando mapas de calor para identificar patrones de asociación entre las variables.
4. Identificación de outliers y valores atípicos:
    - Se identificaron posibles outliers y valores atípicos en los datos utilizando técnicas estadísticas como la desviación estándar y los rangos intercuartílicos.
    - Se evaluó la naturaleza de estos outliers y se tomó una decisión sobre cómo manejarlos, ya sea eliminándolos o tratándose de manera especial en el análisis posterior.

En resumen, este Análisis Exploratorio de Datos (EDA) más detallado proporcionó una comprensión más completa de la estructura, contenido y características de los datos, permitiendo identificar insights adicionales y preparar los datos de manera más efectiva para su análisis en los dashboards y el modelo de machine learning.

5. Proceso de trabajo en la nube:

5.1 Ingesta de archivos:

Después de una exhaustiva evaluación del uso de diferentes plataformas digitales de servicios en la nube, como Amazon Web Services, Google Cloud y Azure, el equipo tomó la decisión de utilizar los servicios de Google Cloud Services. Esta decisión se basó en varios factores, incluyendo el costo, la operatividad en el manejo de diferentes servicios alojados en la plataforma, y la familiaridad y simplicidad de uso que ofrece Google Cloud.

La simplicidad de uso y la familiarización con los servicios de Google Cloud resultaron especialmente útiles para proyectos de este estilo, lo que facilitó la toma de decisiones y la implementación efectiva de soluciones.

Para la ingesta de archivos y su posterior automatización, se decidió utilizar Google Drive como referencia, ya que el proveedor que nos contrata nos indicó que los datasets estarán alojados en una carpeta en Google Drive. Luego, mediante Cloud

Functions, los archivos se trasladan al bucket principal de nuestro proyecto en Google Cloud, creando así un Data Lake con la documentación a utilizar.

Este proceso se automatiza mediante una consulta por HTTP y el manejo de Schedule para su programación de carga, lo que garantiza la eficiencia y la puntualidad en la actualización de los datos en el entorno de Google Cloud.

En resumen, la elección de Google Cloud Services para la gestión de datos y la automatización de procesos se basó en su costo, operatividad, simplicidad de uso y la familiarización del equipo con sus servicios, lo que garantiza una implementación exitosa y eficiente de soluciones para el proyecto.

5.2 ETL con Cloud Functions:

Se implementaron funciones en la nube para automatizar y escalar el proceso de ETL, logrando así la automatización completa del mismo. Estas funciones se diseñaron con el objetivo de optimizar la carga, transformación y limpieza de los datos, asegurando su calidad y coherencia antes de ser almacenados en la nube.

Además, se desarrollaron funciones con restricciones específicas que se encargan de cargar archivos según su nombre y estructura, lo que garantiza la consistencia de los datos y facilita su procesamiento posterior. Estas restricciones se basaron en el análisis de EDA realizado localmente y en las transformaciones previas aplicadas durante el proceso de ETL en el entorno local.

El resultado final de este proceso es la obtención de datos de óptima calidad, los cuales fueron subidos a diferentes tablas de SQL utilizando el servicio de BigQuery. BigQuery fue seleccionado por su capacidad para garantizar la eficiencia y la disponibilidad de los datos, así como por su escalabilidad y su capacidad para manejar grandes volúmenes de datos de manera rápida y efectiva.

En resumen, la implementación de funciones en la nube para el proceso de ETL, junto con las restricciones específicas y la utilización de BigQuery para el almacenamiento de datos, aseguró la calidad, eficiencia y disponibilidad de los datos, proporcionando una base sólida para análisis y toma de decisiones posteriores.

5.3 BigQuery:

Se utilizó BigQuery como plataforma de almacenamiento y consulta, aprovechando su capacidad como Data Warehouse para realizar análisis a gran escala de los datos. En BigQuery se crearon diferentes tablas, poblándose con la información relevante obtenida de los procesos de ETL y limpieza previos.

Estas tablas fueron diseñadas de manera óptima para permitir consultas eficientes y análisis detallados de los datos. Se estructuraron teniendo en cuenta las necesidades específicas del proyecto y la naturaleza de los datos almacenados, garantizando la coherencia y la integridad de la información.

La información almacenada en BigQuery se utilizó de dos maneras principales:

1. Consumo local para la elaboración de dashboards con Power BI:
    - Los datos almacenados en BigQuery fueron consumidos localmente para la elaboración de dashboards utilizando herramientas como Power BI. Estos dashboards proporcionan visualizaciones interactivas y análisis detallados de la información, permitiendo a los usuarios explorar y comprender los datos de manera intuitiva.
2. Consumo en la misma plataforma para la elaboración y puesta en marcha del modelo de Machine Learning:
    - Además, los datos almacenados en BigQuery fueron consumidos directamente en la misma plataforma para la elaboración y puesta en marcha del modelo de Machine Learning. BigQuery proporciona integraciones y herramientas que facilitan el desarrollo y la implementación de modelos de Machine Learning a gran escala, permitiendo realizar predicciones y análisis avanzados basados en los datos almacenados.

En resumen, el uso de BigQuery como plataforma de almacenamiento y consulta permitió aprovechar la capacidad de un Data Warehouse para realizar análisis a gran escala de los datos, mientras que su integración con herramientas como Power BI y las capacidades de Machine Learning facilitaron la elaboración de dashboards interactivos y la implementación de modelos predictivos avanzados.

6. Análisis de datos en dashboards y KPI's:

Se han desarrollado dos dashboards interactivos utilizando herramientas como Tableau o Power BI, con el propósito de visualizar y analizar los datos relacionados con la empresa, las opiniones de los clientes y las oportunidades de inversión.

El primer dashboard está enfocado en la clientela, ofreciendo diversas funcionalidades para que puedan seleccionar entre una amplia gama de opciones, desde encontrar el mejor lugar para comer hasta lugares para ejercitarse u otros intereses específicos. Los usuarios pueden filtrar por zonas y categorías que se ajusten a su presupuesto, lo que les permite tomar decisiones informadas y personalizadas.

Por otro lado, el segundo dashboard está dirigido a futuros inversores, quienes pueden explorar diferentes categorías alineadas con sus intereses y estrategias de inversión. También cuentan con la posibilidad de filtrar por estados de EE.UU. para identificar las ubicaciones más convenientes para su inversión.

En cuanto a los Key Performance Indicators (KPIs), se han implementado tres:

1. **Tasa Promedio de Reviews Anuales:** Este indicador es esencial para evaluar la satisfacción del cliente a lo largo del tiempo, proporcionando información valiosa sobre la calidad y el desempeño de los productos o servicios ofrecidos. Un aumento en este indicador puede indicar una mejora continua en la satisfacción del cliente, mientras que una disminución podría señalar áreas de oportunidad para la empresa.

   Fórmula

   (Total de stars / total de reseñas) * 100

2. **Tasa de Crecimiento de Inversiones Anuales:** Esta métrica clave mide la variación porcentual en el monto total de las inversiones realizadas durante un año específico en comparación con el año anterior. Es fundamental para evaluar la dinámica y evolución de las inversiones a lo largo del tiempo, proporcionando información valiosa sobre la dirección y el ritmo del crecimiento financiero.

Fórmula

(Inversiones del año actual - Inversión del año anterior / Inversión del año anterior) * 100

3. **Tasa de Crecimiento de Reseñas Anuales:** Este indicador mide el cambio porcentual en el número total de reseñas o comentarios recibidos durante un año específico en comparación con el año anterior. Es valioso para evaluar la evolución de la interacción de los clientes y la percepción pública de la empresa, sus productos o servicios a lo largo del tiempo.

   Formula

   (Reseñas del año actual - Reseñas del año anterior / Reseñas del año anterior) * 100

7. Análisis e implementación de ML:

Se construyó un modelo de machine learning de recomendación utilizando técnicas como collaborative filtering o content-based filtering, para proporcionar recomendaciones personalizadas de lugares gastronómicos a los clientes.

7.2 Análisis Predictivo: Uso de series temporales para determinar crecimiento o decrecimiento de un rubro particular.

Mediante el uso de la herramienta Vertex AI de Google, se puso en marcha el entrenamiento de un modelo de series temporales para hacer Forecasting Univariable (predecir el valor de una única variable en el futuro).

El objetivo es lograr predecir el volumen de reviews en los seis meses siguientes a la última fecha registrada, para así poder determinar cuál será el área y lugar de mayor tráfico de usuarios. Se procede bajo la hipótesis de que un mayor volumen de reseñas es equivalente a tener más clientes potenciales y reales.

Si lo que se busca es hacer inversiones inteligentes, resulta de vital importancia conocer cuales son los rubros dentro del turismo y el ocio para los cuales se logra

evidenciar un alto índice de concurrencia, mismo que se ve reflejado en la cantidad de comentarios que se hace en sitios web como Yelp y Google Maps.

7.3 Modelo de recomendación: Análisis de similitudes entre usuarios para la recomendación de servicios.

De igual forma mediante la utilización de Vertex AI, se creó un modelo de recomendación basado en las opiniones de los usuarios. Con el objetivo de lograr un mayor flujo de clientela en el rubro se desarrolló un análisis de los usuarios basado en sus experiencias previas, luego se los comparó entre sí para inferir similitudes y poder generar recomendaciones personalizadas.

La hipótesis es que gracias a esto se dará un mayor engagement de los consumidores que, se verán alentados a probar las nuevas experiencias acordes a sus gustos personales, por lo que aumentará el consumo y a su vez el crecimiento de la industria.

El modelo de tipo filtro colaborativo, llevado a cabo con un vectorización de matrices para lograr un análisis completo y funcional a los propósitos de este proyecto y de cualquier persona que se encuentre, o pensase en estar, invirtiendo en la industria.

8. Conclusión:

En conclusión, el proyecto de análisis del mercado estadounidense para nuestro cliente, parte de un conglomerado de empresas de restaurantes y ocio, ha sido un desafío emocionante y gratificante para DataMind Solutions. Desde la fase inicial de comprensión de los requisitos del cliente hasta la implementación de soluciones avanzadas de análisis de datos y machine learning, nuestro equipo ha demostrado su compromiso con la excelencia y la innovación en cada etapa del proceso. Mediante la combinación de técnicas de análisis de datos local y en la nube, hemos logrado proporcionar a nuestro cliente información valiosa sobre la opinión de los usuarios, las tendencias del mercado y oportunidades de inversión. Además, la implementación de modelos predictivos y de recomendación de restaurantes ha agregado un valor significativo al proyecto, brindando a nuestro cliente herramientas poderosas para la toma de decisiones informadas. En resumen, este proyecto ejemplifica el enfoque integral y la capacidad de DataMind Solutions para ofrecer soluciones de análisis de datos a medida que impulsan el éxito empresarial de nuestros clientes.

9. Enlaces relevantes:

Se adjuntan enlaces a los dashboards desarrollados, así como documentación técnica y otros recursos relevantes para el cliente.

Información de contacto de los integrantes

Fer Abraham

-GitHub: https://github.com/wilson2905

-LinkedIn: https://www.linkedin.com/in/jorge-fernando-abraham-451a44290/

Juan Mendoza

-GitHub: https://github.com/Juan-Mendoza00

-LinkedIn: https://www.linkedin.com/in/juan-camilo-mendoza-lopez-97195a290/

Nico Musante

-GitHub: https://github.com/nicoMusante

-LinkedIn: https://www.linkedin.com/in/nicolas-musante123/

Tiago Sepulveda

-GitHub: https://github.com/Zzepu

-LinkedIn: https://www.linkedin.com/in/tiago-sepulveda-047102243/

Ignacio Sanchez Wakuluk

-GitHub: https://github.com/Ignacio-sw

-LinkedIn: https://www.linkedin.com/in/ignacio-s-wakuluk-4583a0291/

Enlace de GitHub propuesta de trabajo

https://github.com/soyHenry/PF_DS/blob/FULL-TIME/Proyectos/yelp-goople.md

Enlace de Github cuenta colaborativa Proyecto DMS

https://github.com/PFHenryGrupo6/PF_Grupo06_Henry

Enlace al Proyecto en Google Cloud Services

https://console.cloud.google.com/welcome?project=dms-pfh

Enlace Dashboards

https://drive.google.com/drive/folders/1Wy_kRJRUM71MEBnL9HkGRe68O6qMrkrG?usp=sharing

Enlace Modelo de Machine Learning

https://console.cloud.google.com/vertex-ai/models/locations/us-central1/models/4084551556769251328/versions/1/evaluations/4233594872372921470?project=dms-pfh