22-3-2024

Yelp & Google maps- Reviews and Recommendations

Proyecto Final- Sprint 1

Kensit Cortes
Juan Carlos Sánchez
Facundo Blanco
José Efraín Pazos
Rafael Miranda

DATA FIVEBUSINESS CONSULTANS

Contenido

DATA FIVE
BUSINESS CONSULTANTS

Introducción

En el mundo actual, la información generada por las opiniones de los usuarios en plataformas con Yelp y Google Maps se ha convertido en una fuente invaluable de datos pata comprender las preferencias y opiniones de los consumidores, siendo el análisis de estos un factor crucial para el éxito de los negocios. En este contexto, nosotros como consultora nos enfrentamos al desafío de realizar un análisis del mercado estadounidense, específicamente, los restaurantes de las 5 ciudades con mayor afluencia turística.

Este proyecto tiene como objetivo proporcionar a nuestro cliente, parte de un conglomerado de empresas de restaurantes, una comprensión del panorama actual y futuro del mercado, con el fin de optimizar sus estrategias comerciales y mejorar la experiencia del cliente

Objetivo General

Desarrollar e implementar un sistema de gestión de datos integral y eficiente que permita la recopilación, depuración y disponibilidad de información proveniente de diversas fuentes, entre esas Yelp, con el fin de facilitar la creación de un DataWarehouse robusto y funcional. Este sistema también permitirá el análisis y la generación de reportes significativos sobre las opiniones de los usuarios en Estados Unidos en cuanto a restaurantes, así como el entrenamiento y despliegue de un modelo de machine learning enfocado en un sistema de recomendación.

Objetivos específicos

- Recopilar, limpiar y estructurar los datos de diversas fuentes como Yelp y Google maps
- Implementar técnicas de depuración de los datos, incluyendo la identificación y tratamiento de valores atípicos, nulos y la estandarización de formatos con el fin de garantizar calidad y consistencia de los datos almacenados
- Diseñar e implementar un DataWarehouse que permita almacenar y organizar eficientemente los datos recopilados de diferentes fuentes, utilizando tecnologías de almacenamiento de datos adecuadas
- Integrar el modelo de machine learning en el sistema de gestión de datos para permitir la generación de recomendaciones en tiempo real basadas en las preferencias individuales de los usuarios.
- Realizar un dashboard intuitivo e interactivo, con el fin de proporcionar métricas, gráficos y brindar los resultados del análisis de datos.

Alcance del proyecto

Alcance temporal

- Consta de 6 semanas, con duración de los sprint de 2 semanas.
  Sprint 1
  Desde: 4 de maro de 2024
  Hasta: 22 de marzo de 2024
  Sprint 2
  Desde: 25 de marzo de 2024
  Hasta: 5 de abril de 2024

  Sprint 3
  Desde: 8 de abril de 2024
  Hasta: 20 de abril de 2024

Recursos

- Data scientist (una persona)15 horas de trabajo a la semana durante 6 semanas
- Data engineer (dos personas)10 horas de trabajo a la semana durante 6 semanas
- Data analyst (dos personas)15 horas de trabajo a la semana durante 6 semanas

Entregables

Sprint 1

- 3 KPI's
- Documentación general del proyecto
- EDA de los datos
- Repositorio en GitHub
- Stack tecnológico seleccionado para trabajar
- Metodología de trabajo
- Diseño detallado
- Roles y responsabilidades del equipo de trabajo
- Cronograma general de trabajo, representado con un diagrama de Gantt
- Análisis preliminar de calidad de datos

Sprint 2

- ETL completo
- Estructura del DataWarehouse
- Pipeline ETL automatizado

DATA FIVE
BUSINESS CONSULTANTS

- Diseño del Modelo Entidad Relación detallado, incluye tablas, tipos de datos, llaves primarias y llave foránea
- Pipelines para alimentar el DataWarehouse
- Data Warehouse
- Automatización y validación de los datos
- Documentación
- Diccionario de datos
- Producto Mínimo Viable

Sprint 3

- Diseño de Reportes
- KPIs implementados en el dashboard
- Modelos de Machine Learning
- Modelo de ML en producción
- Documentación
- Selección del modelo, feature engineering
- Informe de análisis
- Video del proyecto realizado

Alcance financiero

Sin presupuesto asignado, utilizando herramientas gratuitas y recursos disponibles

Alcance geográfico

Del país de Estados unidos, se tomarán las 5 ciudades con mayor numero de turismo que son: New York, Miami, Orlando, Los Ángeles y San Francisco. Se tomarán empresas de restaurantes

Formulación de KPI´s

KPI	Descripción	Formula	Periodicidad	Objetivo
Índice de Satisfacción del cliente	Porcentaje de clientes satisfechos con su experiencia	$\left(\frac{numero\ de\ reseñas\ positivas}{Total\ de\ reseñas}\right) * 100$	Mensual	Aumentar el índice de satisfacción un 10% en 6 meses
Tasa de retención de clientes	Porcentaje de clientes que regresan a un restaurante	$\left(\frac{Clientes\ que\ repiten\ sitio}{Total\ de\ clientes}\right) * 100$	Trimestral	Aumentar la tasa de retención un 5% en 1 año
Valoración promedio de las reseñas	Calificación promedio que recibe un negocio en las reseñas	$\frac{Suma\ de\ las\ calificaciones}{total\ reseñas}$	Trimestral	Aumentar la valoración promedio 5 puntos en 1 año

Repositorio en GitHub

https://github.com/FacuSB/PF_Grupal

DATA FIVE
BUSINESS CONSULTANTS

Stack tecnológico

	Herramienta	Utilidad
Lenguaje de programación	Python	Se utilizará para la creación del ETL y el EDA, y demás manipulaciones requeridas en la base de datos
Librerías	Numpy	Se utilizará para cálculos numéricos
	Pandas	Se utilizará para la manipulación y análisis de los datos
	Scipy	Se utilizará para estadísticas y modelado en Python
	Matlotlib	Se utilizará para la generación de gráficos
	Seaborn	Se utilizará para visualización estadística en Python
	Scikit learn	Se utilizará para técnicas de machine learning
Entornos de desarrollo	Jupyter Notebook	Se utilizará en la exploración interactiva de los datos y creación de informes
Bases de Datos	SQL	Para crear y gestionar bases de datos relacionales
Herramienta de inteligencia de negocio	Power BI	Se utilizará para la creación de tableros
Máquina virtual	Google Cloud Platform	Se utilizará la plataforma en la nube de Google para procesamiento de datos
Automatización y control de versiones	Git	Control de versiones

Metodología de trabajo

Marco de trabajo ágil Scrum

Diseño detallado

Sprint Planning: Sesión de planificación del Sprint

Se realiza una reunión previa al comienzo del Sprint, donde se definen el enfoque y alcance del proyecto, objetivos, entregables, actividades a realizar y roles de los integrantes del equipo

Comienzo del Sprint de Scrum:

Cada Sprint cuenta con una duración de dos semanas, donde cada miembro del equipo trabajara en las tareas pendientes que se establecieron en la sesión de planificación del sprint

Daily Stand Up:

Se organizan reuniones de 15 minutos los días lunes, miércoles y viernes con el _____ para informar con respecto al trabajo que se este realizando e identificar cualquier obstáculo e inquietudes que hayan surgido

Sprint Review:

Se presenta el trabajo una vez terminado el Sprint, es decir el viernes de la segunda semana. Tendrá una duración de 45 minutos, donde el equipo se reunirá con el cliente para hacer una revisión del Sprint y verificar que el producto entregable coincida con los objetivos trazados al comienzo de cada sprint

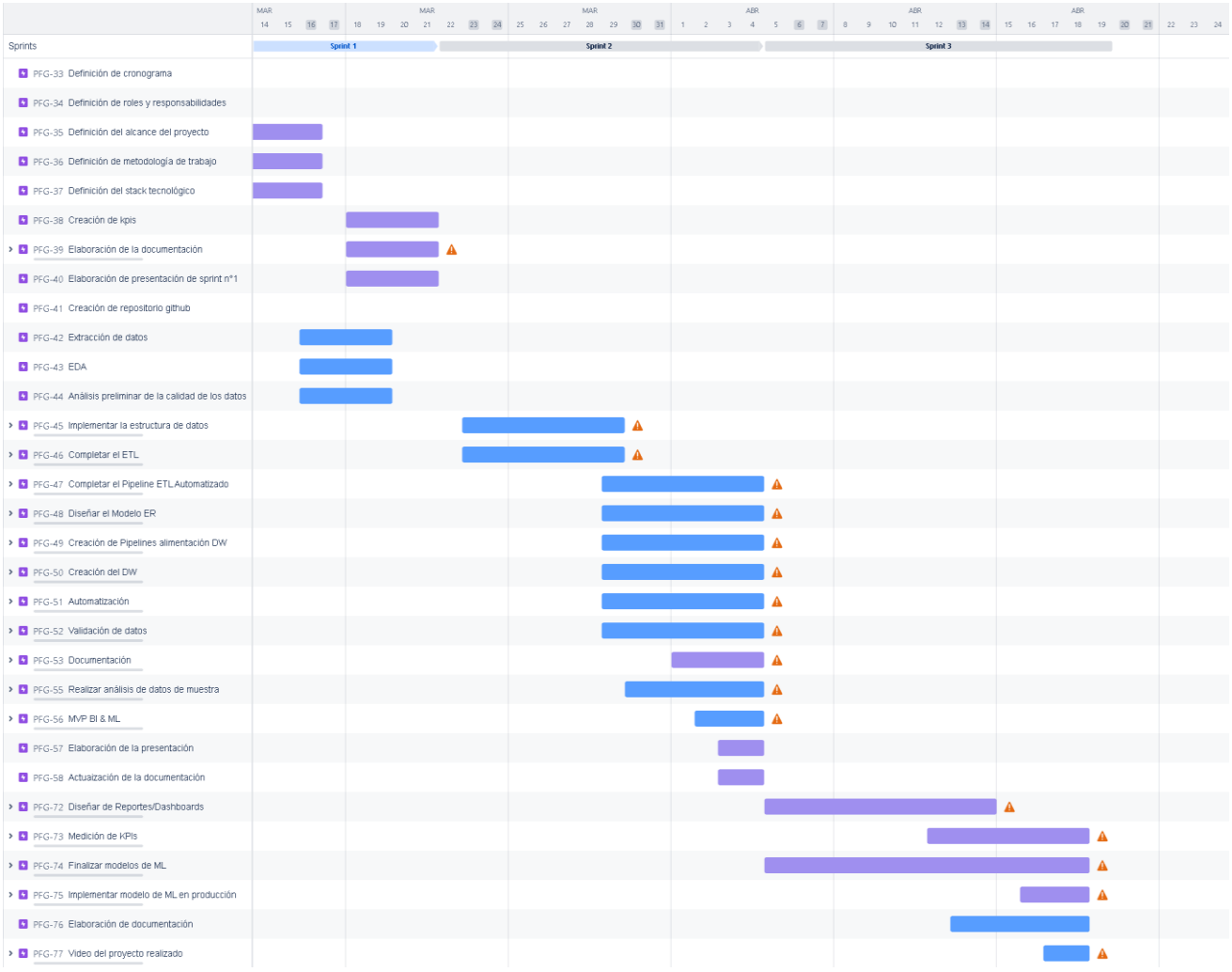
Equipo de trabajo

Nombre y rol del integrante	**Responsabilidades**
Facundo Blanco **Data Engineer**	- Diseñar y construir la arquitectura del DataWarehouse - Diseñar e implementar un modelo de machine learning que utilice técnicas de recomendación para predecir las preferencias de los usuarios y ofrecer recomendaciones personalizadas - Supervisar el procesamiento de los datos
Juan Carlos Sánchez **Data Engineer**	- Crear un modelo predictivo utilizando técnicas de machine Learning para realizar pronósticos sobre el crecimiento futuro de los negocios - Desarrollar el pipeline automatizado
Rafael Miranda **Data Analyst**	- Realizar análisis exploratorio de los datos - Identificar patrones, tendencias y relaciones utilizando técnicas estadísticas - Generar informes y visualizaciones para comunicar patrones y tendencias
José Efraín Pazos **Data Scientist**	- Explorar y limpiar los datos, identificar y tratar valores atípicos, manejar datos faltantes y transformar variables según sea necesario para mejorar la calidad de los datos

	- Realizar análisis descriptivos y la implementación del modelo predictivo - Seleccionar las tecnologías adecuadas para garantizar un procesamiento rápido y confiable de grandes volúmenes de datos
Kensit Cortes **Data Analyst**	- Obtener datos del DataWarehouse creado - Realizar análisis descriptivos con los datos - Crear Data Storytelling

Cronograma general

https://proyectofinalhenryg5pt6.atlassian.net/jira/software/projects/PFG/boards/1/timeline?timeline=WEEKS



DATA FIVE
BUSINESS CONSULTANTS