

Entendimiento del negocio

David Guzmán Leyva - A01706417
Enrique Santos Fraire - A01705746
Leonardo Alvarado Menéndez - A01705998
Oscar Enrique Delgadillo Ochoa - A01705935

Índice

Índice	2
Bitácora de cambios	3
Contexto	3
Determinar los objetivos de negocio	3
Background	3
Objetivos del negocio	3
Criterios de éxito	3
Evaluar la situación	4
Inventario de recursos	4
Requerimientos, supuestos y restricciones	5
Riesgos y contingencias	5
Terminología	6
Costos y beneficios	6
Determinar los objetivos de la minería de datos	6
Objetivos de minería de datos	6
Criterios de éxito de minería de datos	6
Producir el plan de proyecto	6
Plan de proyecto	6
Evaluación inicial de herramientas y técnicas	7
Referencias	8

Bitácora de cambios

Versión	Fecha	Autor(es)	Modificaciones
1.0	04/11/2022	David Guzmán Leyva Enrique Santos Fraire Leonardo Alvarado Menéndez Oscar Enrique Delgadillo Ochoa	Línea base
2.0	15/11/2022	David Guzmán Leyva Enrique Santos Fraire Leonardo Alvarado Menéndez Oscar Enrique Delgadillo Ochoa	Segunda iteración

Contexto

Se realizó una segunda iteración de este documento debido a que aunque se cumplían los objetivos anteriores, no se estaba utilizando realmente una solución de utilidad aplicando machine learning. Además que el valor otorgado no era significativo, con esta nueva iteración se busca llegar a una mejor solución explorando diversas alternativas.

Determinar los objetivos de negocio

Background

Dada la dificultad y las numerosas complejidades que traen consigo las encuestas de movilidad, se optó por utilizar datos pasivos captados mediante dispositivos móviles con el fin de obtener información de una manera más eficiente. De esta manera se espera poder estimar y visualizar los Viajes Origen-Destino de los usuarios de la ciudad de Santiago de Chile en su región metropolitana.

Dichos datos se obtienen de la empresa de telefonía móvil Movistar, y se trabajarán en conjunto con el centro de investigación de Data Science de Chile.

Objetivos del negocio

- Determinar las comunas con mayor presencia de home-office.
- Identificar los factores determinantes que sean indicadores porcentuales de presencia de home office.

Criterios de éxito

- Se encuentra una causalidad concreta y verificable de los indicadores, siendo validado por la Dra. Loreto Bravo del Instituto de Data Science de Chile y el Dr. Benjamin Valdés Aguirre.

Evaluar la situación

Aquí se identifican los factores que deben ser considerados para el logro de los objetivos y el plan de proyecto.

Inventario de recursos

Lista de los recursos disponibles para el proyecto:

- Expertos
 - MTI Eduardo Daniel Juárez Pineda.
 - Dr. Benjamin Valdés Aguirre.
 - Dr. Ismael Solís Moreno.
 - Dr. José Antonio Cantoral Ceballos.
 - Dr. Carlos Alberto Dorantes Dosamantes.
 - Dra. Loreto Bravo.
- Datos
 - Phone_ID
 - timestamp
 - bts_id
 - lat
 - lon
- Recursos de cómputo (hardware)
 - Laptops individuales
 - AWS
- Software
 - AWS
 - Tableau
 - Colaboratory notebooks
 - Python
 - Spark
 - Hadoop
 - Tensorflow
 - Sklearn
 - Pandas

Requerimientos, supuestos y restricciones

- Lista de todos los requerimientos del proyecto:
 - El proyecto se debe finalizar antes del 1° de diciembre del 2022.
 - Generar una matriz de viajes con al menos los siguientes atributos:
 - ID anonimizado: Se conserva el hash de los móviles anonimizados con un string de 64 caracteres.
 - Origin Zone: Zona de origen del viaje.
 - Origin time: Tiempo de inicio de viaje con desagregación de segundos (HH:MM:SS).
 - Destiny Zone: Zona de destino del viaje.
 - Destiny time: Tiempo de término de viaje con desagregación de segundos (HH:MM:SS).
 - Comuna de Origen: Comuna de origen del viaje.
 - Comuna de Destino: Comuna de destino del viaje.
 - Generar un modelo de predicción de home office por comuna según los indicadores de la zona a analizar.
- Lista de supuestos del proyecto sobre:
 - Los datos:
 - Contamos con un registro de teléfonos conectados a cada antena según la fecha y hora del día.
 - No contamos con la información de las comunas de Santiago de Chile.
 - Los datos se obtienen de la compañía Movistar en Chile y se encuentran anonimizados.
 - Se cuenta con el permiso de uso de los datos.
- Lista de restricciones del proyecto sobre:
 - Seguridad:
 - Evitar compartir el dataset con personas ajenas a la unidad de formación.
 - Disponibilidad de los recursos
 - Contamos con la información pública de Chile, además de los datos aportados por el Centro de Data Science de Chile.
 - Aspectos técnicos
 - Limitaciones en cuanto a la capacidad que tenemos de hardware para poder analizar, limpiar y modelar los datos obtenidos.
 - Limitaciones en conocimiento de frameworks útiles para obtener la matriz de viajes solicitada y posteriormente realizar alguna predicción que sea de valor para el centro de investigación.

Riesgos y contingencias

Terminología

- Comuna: División territorial en Santiago de Chile.
- home-office: Trabajo realizado desde el hogar.
- seudominisación: tratamiento de datos personales de manera tal que ya no puedan atribuirse a un interesado.

Costos y beneficios

Tal y como observamos durante la presentación del proyecto con nuestra socia formadora. El análisis de datos mediante las encuestas de movilidad es sumamente costoso tanto en tiempo como en dinero, además de no poder realizarse de manera frecuente, por lo que los datos es probable que queden obsoletos rápidamente dependiendo del uso que se les den.

En cambio con la recolección de datos por medios pasivos como lo es a través de empresas de telefonía nos permite analizar un mayor volumen de datos y más recientes, con la capacidad de obtenerlos incluso con un día de antigüedad, pudiéndose considerar prácticamente como de tiempo real.

Determinar los objetivos de la minería de datos

Objetivos de minería de datos


- Identificar los indicadores y su correlación con la presencia de home office a través del entrenamiento del modelo.
- Conocer el porcentaje de home-office por cada comuna según los indicadores obtenidos previamente a través del modelo. Resultando así en las comunas con mayor presencia de home-office y qué depende de ello.

Criterios de éxito de minería de datos

- Se llega a un Mean Squared Error (MSE) menor a 5 tanto en el set de entrenamiento como en el de prueba o un Mean Absolute Error (MAE) menor a 2.
- Los indicadores establecidos mantienen una correlación verificable con el porcentaje de home office del área.

Producir el plan de proyecto

Plan de proyecto

 Plan de Trabajo.xlsx

Evaluación inicial de herramientas y técnicas

La herramienta primaria que se va a utilizar para llevar a cabo este proyecto de minería de datos es la Big Data, específicamente utilizaremos el framework Apache Spark. Dicho framework tiene las siguientes ventajas:

- Rapidez: Ejecuta las cargas de trabajo 100 veces más rápido que con Hadoop MapReduce. Con Spark, se tiene un alto rendimiento con los datos por lotes y de streaming gracias al programador de grafos acíclicos dirigidos de última generación, al optimizador de consultas y al motor físico de ejecución.
- Manejo de datos: Spark puede hacer uso de Resilient Distributed Datasets (RDD).
- Facilidad de uso: Spark cuenta con más de 80 operadores generales que facilitan el desarrollo de aplicaciones en paralelo. Puede ser utilizado de forma interactiva desde el shell de Scala, Python, R y SQL para escribir aplicaciones rápidamente.
- Uso general: Spark permite usar una pila de bibliotecas que incluye SQL, DataFrame, MLlib para aprendizaje automático, GraphX y Spark Streaming.
- Framework de código abierto: Además de ser gratis, cuenta con potencial colectivo para aportar más ideas, desarrollarlas más rápido y solucionar los problemas en cuanto aparecen.

Asimismo emplearemos algunas librerías de python necesarias para realizar ETL:

- Pandas: Sirve para la manipulación y análisis de datos, además ofrece estructuras de datos y operaciones para manipular tablas numéricas y series temporales.
- Numpy: Para realizar cálculos u operaciones con los datos.
- Geopy: Será utilizado para la obtención de comunas.

Entornos de ejecución del proyecto:

- Google Colab: Permite a cualquier usuario escribir y ejecutar código arbitrario de Python en el navegador. Es especialmente adecuado para tareas de aprendizaje automático, análisis de datos y educación.
- Jupyter notebook: Es un entorno de desarrollo interactivo basado en la web para cuadernos, códigos y datos.
- Anaconda: Es una distribución libre y abierta de los lenguajes Python y R, utilizada en ciencia de datos, y aprendizaje automático. En ella podemos crear Jupyter notebooks.

Manejo de versiones del proyecto:

- Github: es una forja para alojar proyectos utilizando el sistema de control de versiones Git.
- Link de repositorio: <https://github.com/Davidguzley/Movilidad-Chile>

Administración del proyecto:

- Google Drive: Es un servicio de alojamiento de archivos en el cual realizamos planes de trabajo y documentación del proyecto de manera colaborativa

Referencias

- *NumPy documentation — NumPy v1.23 Manual.* (s. f.). Recuperado 18 de octubre de 2022, de <https://numpy.org/doc/stable/>
- *pandas - Python Data Analysis Library.* (s. f.). Recuperado 18 de octubre de 2022, de https://pandas.pydata.org/getting_started.html
- *PySpark Documentation — PySpark 3.3.0 documentation.* (s. f.). Recuperado 18 de octubre de 2022, de <https://spark.apache.org/docs/latest/api/python/>
- *Welcome to GeoPy's documentation! — GeoPy 2.2.0 documentation.* (s. f.). Recuperado 18 de octubre de 2022, de <https://geopy.readthedocs.io/en/stable/>