ITESM- Campus Puebla

NAATIK | CRISP DM

Inteligencia artificial avanzada para la ciencia Datos II

Integrantes Equipo 1:

Myroslava Sánchez Andrade A01730712
José Antonio Bobadilla García A01734433
Karen Rugerio Armenta A01733228
Alejandro Castro Reus A01731065

Fecha: 07/10/2022

Índice

1 Business Understanding

1.1 Determine Business Objectives

Background

La telefonía fija es el servicio de telecomunicaciones que proporciona la capacidad de comunicación de voz entre usuarios. Este servicio se da a través de líneas telefónicas conectadas a una central de conmutación automática que permite establecer la comunicación.

Actualmente, las compañías de telefonía fija ofrecen más servicios a sus usuarios como: internet, televisión por cable, servicios de streaming, etc.. Esto ha atraído a un mayor número de clientes, pero a su vez ha incrementado la competencia entre compañías por una mejor calidad de servicios.

Como parte de nuestro proyecto final de la materia *Inteligencia Artificial Avanzada para la Ciencia de Datos,* se nos asignó una problemática dada por la empresa NAATIK. Esta es una empresa enfocada en el desarrollo y aplicación de Inteligencia Artificial y Ciencia de Datos para brindar soluciones.
La problemática planteada consiste en el análisis de un conjunto de datos de una compañía telefónica que contiene los siguientes datos: servicios contratados de clientes, información de la cuenta de los clientes, información demográfica de los clientes, y la permanencia del cliente en el último mes (booleano).

La entrega para esta problemática consiste en la predicción de la permanencia de un grupo cliente dados sus datos de una empresa de telefonía, y análisis de clientes para el desarrollo de un modelo de retención.

Objetivos de negocio

Dada la problemática de nuestro cliente, podemos identificar 4 objetivos:

- Predicción de la permanencia de un cliente (modelo de machine learning).
- Identificación de los clientes con alto riesgo de abandono (segmentación).
- Análisis de clientes con alto riesgo de abandono y sus características.
- Interpretación de los modelos y el análisis.

Criterios de éxito

Para determinar el éxito y satisfacción del proyecto, se deben de cumplir con los siguientes puntos:

- Obtener un porcentaje mínimo del 80% de precisión en el modelo de predicción.
- Hacer una segmentación de los clientes por grupos con características afines.
- Correcto análisis e interpretación de la segmentación de clientes.

1.2 Assess Situation

Inventory of Resources

En este apartado se enlistan los recursos que están disponibles (o que fueron brindados por la institución educativa) para el desarrollo del proyecto.

- **Expertos:**
  - Dr. Benjamín Valdés Aguirre (área: Inteligencia Artificial).
  - Dr. Ismael Solís Moreno (áreas: BigData y Cloud Computing).
  - Dr. Carlos Alberto Dorantes Dosamantes (área: Estadística).
  - Dr. Juan Manuel Ahuactzin Larios (área: Inteligencia Artificial).

- **Datos:**
  - Archivo "telecom_churn_me.csv" extraído del sitio web: https://www.kaggle.com/datasets/mark18vi/telecom-churn-data

- **Recursos de cómputo:**
  - **Procesador**: Intel(R) Core(TM) i7-1065G7 CPU @ 1.30GHz, 1498 Mhz, 4 Core(s), 8 Logical Processor(s).
    Memoria RAM: 12 GB
    Almacenamiento: 1 TB
    Tarjeta Gráfica: Intel(R) Iris(R) Plus Graphics
    Sistema operativo: Windows 10

  - **Procesador**: Arm M1
    Memoria RAM: 8GB
    Almacenamiento: 500 GB
    Tarjeta Gráfica:
    Sistema operativo: iOS

  - **Procesador**: Intel core i5 10th generation.
    Memoria RAM: 16 GB
    Almacenamiento: 2TB
    Tarjeta Gráfica: NVIDIA GTX 1650

Sistema operativo: Windows 10

- **Software:**
  - Data Engineering (PySpark, Pandas, Numpy)
  - Machine learning libraries (TensorFlow, Keras, MLlib)

Requirements, Assumptions and Constraints

Este documento se puede encontrar en el repositorio de Github del proyecto como "**Requerimientos One Page.pdf".**

Risks and Contingencies

Es necesario realizar la identificación de los riesgos, para poder proponer una solución en caso de que ocurran.

Los riesgos se evaluarán siguiendo una métrica del 1 al 4 en cuanto a probabilidad y gravedad. Una probabilidad 1 significa que es poco probable que ocurra, mientras que una probabilidad de 4 significa que las posibilidades de que ocurran son altas, de igual manera, tener un 1 como gravedad significa que el riesgo es bajo y no tendrá un gran impacto en la entrega del proyecto, sin embargo una gravedad de 4 significa que si el riesgo ocurre el proyecto ya no será viable como se había planteado. A continuación se presentará una tabla del análisis de riesgos que pueden presentarse durante este proyecto.

El nivel de riesgo se puede calcular haciendo uso de una matriz de evaluación de riesgo, en el cual se multiplica la probabilidad de que el evento suceda por la gravedad del mismo y se le asigna un nivel de prioridad. Los riesgos con una prioridad de 1-3 se aceptan, los riesgos con prioridad entre 4-8 se mitigan con un plan de acción y los riesgos con un nivel mayor de 9 se evitan totalmente.

Riesgo	Probabilidad	Gravedad	Nivel de prioridad
Los requerimientos no cubren las necesidades del socio formador	1	4	4
Un desarrollador	3	2	6

no puede trabajar debido a una fuerza mayor			
El modelo elegido no cumple con el porcentaje de accuracy estipulado	2	4	8
Los perfiles de los clientes no son fácilmente distinguibles.	2	4	8
La interfaz web no es lo que los clientes esperan	2	4	8
Una librería o tecnología no cumple con las características necesarias y debe ser cambiada	2	2	4
Se asigna un nuevo dataset cuando ya se llevaba un avance en el ETL	4	3	12
Las variables del modelo están poco correlacionadas con el target	1	4	4
Más de un miembro del equipo se encuentra ausente al mismo tiempo	2	4	8

Terminology

Target - Resultado obtenido que se encuentra en el set de datos o resultado esperado al momento de realizar un modelo de predicción. En este caso el target se considera como la permanencia del cliente.

Inteligencia Artificial - Combinación de algoritmos planteados con el propósito de crear programas que puedan realizar aprendizaje similar al ser humano.

Ciencia de Datos - Metodología utilizada para unificar estadísticas, análisis de datos, aprendizaje automático, y sus métodos relacionados, a efectos de comprender y analizar fenómenos que ocurren en alguna situación determinada.

Machine learning - Es la subdivisión de la inteligencia artificial que se centra en desarrollar sistemas que aprenden, o mejoran el rendimiento, con base en los datos que consumen y su procesamiento.

Procesador - Es un componente del hardware que se puede encontrar dentro de un ordenador, teléfonos inteligentes, y otros dispositivos programables. Su función es interpretar las instrucciones de un programa informático mediante la realización de las operaciones básicas aritméticas, lógicas, y externas.

RAM - La memoria de acceso aleatorio es utilizada en un sistema de cómputo para cargar los programas o archivos que se están siendo utilizados.

Almacenamiento - Un dispositivo de almacenamiento de datos es un conjunto de componentes electrónicos habilitados para leer o grabar datos en el soporte de almacenamiento de datos de forma temporal o permanente.

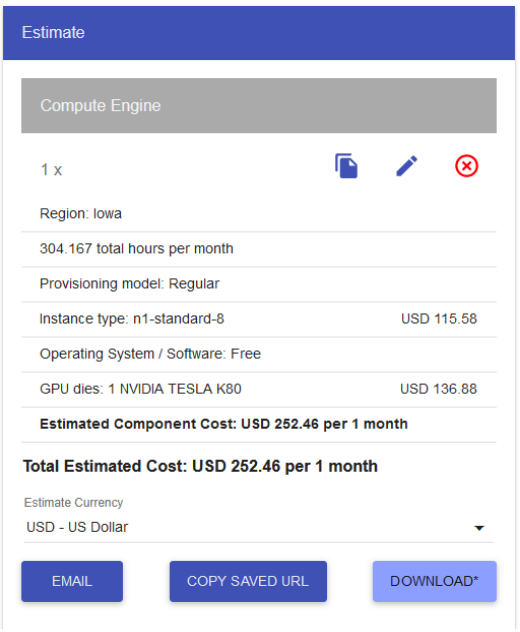
Tarjeta Gráfica - Es una tarjeta de expansión de la tarjeta madre o motherboard del computador que se encarga de procesar los datos provenientes del procesador y transformarlos en información comprensible y representable en el dispositivo de salida. Mayormente utilizada para representar gráficos en una pantalla, pero puede ser utilizada para otras tareas, como por ejemplo, resolver matrices de manera eficiente.

Sistema operativo - Es el programa que se encarga de administrar los recursos de la computadora.

Data mining - Es un campo de la estadística y las ciencias de la computación que se enfoca al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos.

Costs and Benefits

- 4 Laptops Asus: $25,199 * 4 = **$100,796 MXN**
    - 15.6 Pulgadas
    - Full HD
    - Intel Core i5
    - NVIDIA GeForce RTX 3050
    - 8 GB RAM

    - 512 GB SSD
- Sueldo promedio de ingeniero de ML al mes en México: $43, 218 MXN (Glassdor: https://www.glassdoor.com.mx/Sueldos/machine-learning-engineer-sueldo-SRCH_KO0,25.htm)
- $43,218 MXN * 4 ingenieros: **$172,872 MXN**
- Costo de procesamiento en la nube (Google Cloud): $252.46 dólares * mes = $5,034 MXN
- 

1.3 Determine Data Mining Goals

Data Mining Goals

- Implementar un modelo de árbol de decisión para realizar predicciones sobre la permanencia de un cliente con base en su historial desde 1998 hasta 2018.
- Realizar una segmentación de clientes para identificar a los clientes con alto riesgo de abandono.
- Realizar un análisis de los perfiles clasificados como clientes con alto riesgo de abandono, tomando en cuenta las variables que determinan esta condición.
- Plasmar en un documento la interpretación del modelo y los resultados obtenidos.

Data Mining Success Criteria

- El accuracy obtenido del decision tree debe ser mayor al 85% para ser considerado un modelo confiable.

- Realizar una matriz de confusión para evaluar el accuracy del modelo de clasificación.
- Los clientes considerados de alto riesgo comparten variables similares.

1.4 Produce Project Plan

Project plan

El plan del proyecto se encuentra estructurado en Project Libre y se puede encontrar en el repositorio de Github del proyecto como **ProjectPlan.pod**, en él, se definen las etapas en las que se realizará el proyecto, que en este caso son semanales, ya que se trabaja con una metodología Ágil, así mismo se define la duración de cada una de las actividades a realizar, la persona o personas responsables de las actividades y las dependencias que se presentan a lo largo del proyecto.

Initial Assessment of Tools, and Techniques

Este documento se puede encontrar en el repositorio de Github del proyecto como "**Mapeo de recursos y herramientas disponibles.pdf".**

2 Data Understanding

Referencias

- Naatik. (s/f). "Naatik AI Solutions". Naatik. Recuperado el 10 de octubre del 2022 del sitio web: https://www.naatik.ai/
- ift. (s/F). "Instituto Federal de Telecomunicaciones". ift. Recuperado el 10 de octubre del 2022 del sitio web: https://www.ift.org.mx/usuarios-y-audiencias/telefonia-fija