Tecnológico de Monterrey

ITESM- Campus Puebla

NAATIK

Momento de Retroalimentación: Reto Datos

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: 11/11/2022

I. Expliquen qué herramientas y tecnologías van a usar para trabajar con los datos, y por qué creen que estas sean la mejor combinación.

Como tecnología principal se hizo uso de Python, utilizando las librerías de Pandas que nos ayudará a trabajar con los datos en formato de Dataset, lo que permite que sea más fácil de manipular, Numpy que nos ayudará a realizar operaciones matriciales con los data frames y Scikit-Learn cuya API contiene diversas herramientas que ayudan a realizar el aprendizaje automático, incluyendo clasificación, regresión, y reducción de dimensionalidad.

Se optó por la utilización de estas tecnologías debido a las especificaciones del socio formador, quién aseguró que no se contaría con sets de datos de tipo Big Data y que optó por una ejecución rápida. La razón por la cual no se decidió usar Pyspark es porque ésta herramienta es mayormente utilizada para procesar grandes volúmenes de datos y diferentes tipos de datos y no es lo que se trabajará en este reto.

Por otro lado, para la implementación de modelos más robustos (random forest y convolutional neural networks) se utilizó la herramienta TensorFlow, pues esta herramienta nos permite hacer uso de la GPU para un procesamiento más rápido.

II. Genere el modelo de almacenamiento de los datos relacionados al reto.

Dados los requerimientos del socio formador, en los cuales se especifica que el sistema debe poder ser utilizado de forma local en una computadora de gama media, los datos fueron almacenados de forma local como archivos .csv.
Se puede utilizar el sistema con una base de datos local; sin embargo, para el almacenamiento de la información recomendamos el uso de una base de datos más robusta en línea, de acuerdo a la criticalidad de los datos. Una alternativa a una base de datos tradicional es un Data Warehouse pues los conjuntos de datos del proyecto tienen una estructura definida y esta es ideal para el análisis de métricas de rendimiento clave (justo el propósito del proyecto).

Estructura del almacenamiento de los datos
Dado un nuevo conjunto de datos, se creará su carpeta correspondiente (con el mismo nombre del archivo .csv) en la carpeta 'data'. La estructura de esta nueva carpeta se define de la siguiente forma:

Archivos en la carpeta raíz:
- Archivo original (cargado por el usuario).
- Archivo de predicción (nuevo conjunto de datos, se agrega una vez se haya completado el ETL y el entrenamiento de un modelo).

- Archivo resultado de predicción.

Carpeta 'train:'
- Archivo *'original_train.csv'* (conjunto de datos para el entrenamiento).
- Archivo *'x_train.csv'* (conjunto de datos de las variables independientes con SMOTE*).
- Archivo *'y_train.csv'* (conjunto de datos de la variable dependiente).

III. De acuerdo con la cantidad de datos que se tiene determine si es necesario utilizar un enfoque orientado a Big Data o no.

Realizando un análisis de los datos nos dimos cuenta que la cantidad de datos es mínima para ser considerada big data, ya que para que se considere big data, los datos tienen que ser tan grandes, rápidos o complejos que es difícil o imposible procesarlos con los métodos tradicionales. Los datos actuales no cuentan con las 5Vs para ser considerado un proyecto de esta magnitud.

En primer lugar, nos encontramos con un set de archivos que no tiene un gran Volumen, ya que cuenta con poco más de un millón de datos. Así mismo no se cuenta con Velocidad, es decir, no se adquieren nuevos datos en poco tiempo. De igual manera no se tiene Variedad, ya que sólamente contamos con datos de tipo float, integer y object. El set de datos fue adquirido de Kaggle, pero no hay una forma de saber qué tan Veraz es. Finalmente se considera que el Valor de las variables será definido una vez realizado el modelo.

Así mismo, los socios formadores comentan que en ningún momento consideran trabajar con archivos que cumplan las condiciones de Big Data y como se trata de un set de datos cuyo contexto es el análisis de churn en una empresa, no se trabajará la solución como Big Data. Así mismo, los socios formadores no cuentan con servidores o herramientas que permitan el procesamiento de Big Data, ya que todo se trabajará en un equipo de la empresa de manera local. Debido a esta limitante, tampoco se puede emplear una solución que requiera de un alto poder computacional de procesamiento.