Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Puebla

Inteligencia Artificial para la ciencia de datos TC3007C

Reto
CRISP-DM Business Understanding

Fernando Jiménez Pereyra	A01734609
Daniel Flores Rodríguez	A01734184
Alejandro López Hernández	A01733984
Daniel Munive Meneses	A01734205

5 de octubre de 2022

Project Objectives

Background

Naatik AI Solutions, con la colaboración directa de Pablo Ibargüengoytia, quienes son una empresa enfocada en el desarrollo y aplicación de Inteligencia Artificial y Ciencia de Datos. Enfocados a brindar soluciones a diferentes sectores de industria y servicios. Con colaboradores con más de 30 años de experiencia y con presencia en más de 150 empresas. El socio formador se enfrenta frente al problema de poder tener un producto tangible que permita demostrar las ideas que la empresa tiene y quiere ofrecer

Business Objectives

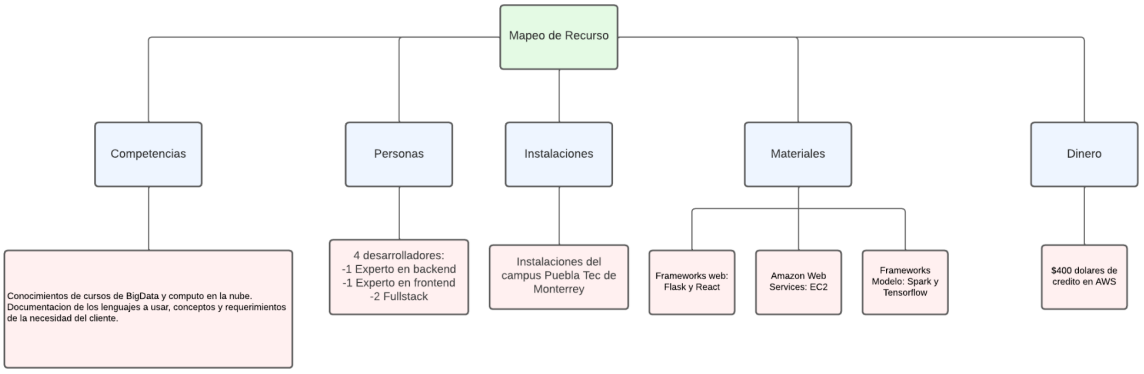
El objetivo principal del proyecto desde el punto de vista de Naatik es poder vender sistemas basados en modelos de predicción a empresas de diferente giro pero con un factor en común, que sigan con un modelo de negocio de suscripciones de usuario. Empresas de telefonía, internet, servicios de streaming, entre otras pueden ser potenciales clientes de Naatik.

Business Success Criteria

El principal criterio de éxito es que el sistema sea funcional a lo largo del tiempo y vea un resultado positivo en las potenciales ventas que se puedan generar. Lo importante de este aspecto es saber cómo vender ante las empresas; lo que se traduce a resaltar los aspectos más destacados de los resultados empleando un determinado set de datos.

Assess Situation

Inventory of Resources



Requirements, Assumptions & Constraints

Requerimientos funcionales

Modelos

1. El sistema deberá recibir un set de datos determinado y aplicar el procesamiento ETL correspondiente.
2. Se debe considerar un modelo de clusterización que segmente la información dada en N cantidad de clústers o grupos. Dicha cantidad de clústers será dinámica de acuerdo al dataset dado.
3. Se deberá contar con un modelo entrenado de clasificación que trabaje con cada uno de los clústers generados por el modelo anterior; dicho modelo debe determinar si un determinado cliente abandonaría la empresa o no. Tentativamente se trataría de la implementación de árboles de decisión.

Interfaz web

1. El sistema contará con una interfaz web que despliegue los resultados de la implementación de los modelos entrenados previamente mencionados.
2. La interfaz deberá mostrar un resumen sustancial de los datos más importantes de los resultados del modelo de clasificación; haciendo uso

de aspectos visuales como gráficas, imágenes, comparaciones estadísticas, etcétera.

3. Se debe considerar la opción de poder modificar los parámetros de entrada de los modelos a emplear; es decir, se debe poder modificar aspectos como el rango de clasificación.

Requerimientos no funcionales

1. Los colores, uso de imagen y multimedia será proporcionado por Naatik con base al aspecto visual de la empresa
2. Los modelos deberán ser implementados en el lenguaje de programación Python, con la librería Tensorflow.
3. La interfaz web será realizada en el framework React en términos de frontend; mientras que para el backend se hará uso de Flask
4. El sistema deberá estar montado remotamente con el uso de módulos de Amazon Web Services

Asunciones

1. Se asume que los dataset empleados para el sistema podrían variar, por lo que se debe considerar la flexibilidad de la clusterización dada
2. Se asume que los servicios en dónde montar el sistema como los módulos de AWS podrían no ser definitivos
3. Se asume que el acompañamiento por parte de Naatik será constante y cercana con tal de desarrollar un sistema óptimo a base de retroalimentación continua en las diferentes etapas del proyecto

Limitantes

1. El desarrollo del proyecto sólo será de 8 semanas aproximadamente.
2. El dataset dado para entrenar a los modelos del sistema no es lo suficientemente grande como para poder considerarlo “Big Data”.
3. Se debe seleccionar sólo la información importante para desplegarse en la plataforma, lo que podría omitir buena parte de la información dada en los resultados

Risk and Contingencies

Dentro del desarrollo del sistema, se pueden presentar una cantidad de situaciones o escenarios que pueden influenciar en el funcionamiento del equipo del desarrollo y afectar la calidad de los entregables del proyecto. Algunos escenarios son los siguientes, mencionando entre paréntesis la probabilidad de que éstos sucedan:

- Uno o más compañeros de equipo presenten un cuadro de COVID-19 y no puedan presentarse presencialmente a trabajar. (Medio)
- Mala comunicación entre miembros del equipo. (Bajo)
- Pobre implementación de los modelos o interfaz debido a las entregas individuales correspondientes a los módulos de clase. (Alto)
- Mala implementación de ETL con el set de datos correspondiente para el desarrollo. (Bajo)
- Mala configuración del modelo de clusterización, generando retrasos significativos en los tiempos de entrega. (Medio)
- Elección inadecuada del modelo de clasificación, generando retrasos significativos en los tiempos de entrega y calidad del modelo (Medio).
- Caída de servidores en donde puedan estar montados el backend y el frontend de la interfaz. (Alto)
- Poco o nulo acompañamiento por parte de la organización socio formadora, en este caso Naatik. (Bajo)

Terminology

- Churn rate: la tasa de abandono por parte de los clientes.
- Cluster: un conjunto de cosas que poseen cualidades muy similares entre sí.

- Amazon Web Service: es una nube que ofrece servicios infraestructura como servicio, en donde podremos realizar un entorno de desarrollo y simulación de despliegue para el producto final.
- Infrastructure as a service: en español infraestructura como servicio, consiste en un modelo de negocio en el cual se ofrece una infraestructura de hardware administrada por el proveedor.

Costs and Benefits

Actualmente no se plantea realizar nada que genere un gasto que no pueda ser cubierto más allá de los $400 dólares de crédito que tenemos disponible en Amazon Web Service.

Se plantea que el beneficio final sea el obtener una demo funcional de los sistemas que puede ofrecer Naatik con el fin de conseguir clientes interesados en desarrollar modelos similares.

Data Mining Goals

Goals

El objetivo de este proyecto es crear un sistema computacional con el fin de clasificar, como ejemplo para entrenar el sistema, a los clientes de una empresa de telefonía para así predecir la probabilidad de abandono de la empresa y poder invertir para poder tomar acciones que reduzcan la cantidad o porcentaje de este tipo de usuario.

Success Criteria

1. Implementar un modelo de clusterización con una cantidad de clústers flexible según el dataset a utilizar.
2. Implementar un modelo de predicción de abandono de clientes para cada clúster generado. Considerando que se trata de un problema de clasificación, una opción viable de modelo podría ser los árboles de decisión.
3. Detectar los perfiles de los clientes a partir de un porcentaje determinado de abandono o churn a partir de un set de datos dado.
4. Obtener las matrices de confusión para poder comprender la tendencia de falsos negativos o falsos negativos del modelo de predicción para poder así diseñar un uso interesante de dicha información desde el punto de vista del cliente.

Project Plan