TECNOLÓGICO DE MONTERREY®

Inteligencia artificial avanzada para la ciencia de datos

Reto

Etapa Deployments

Diego Arturo Padilla Domínguez - A01552594

Keyuan Zhao - A01366831

Carolina Herrera Martínez - A01411547

Cutberto Arizabalo Nava - A01411431

Jose Pablo Cobos Austria - A01274631

Campus Querétaro

30 de noviembre de 2022

Deployments

Plan de entrega

Obtuvimos dos tipos de resultados, modelos y hallazgos, y para cada uno de ellos se llevara a cabo un plan distinto.

Modelos
Lo que se va a entregar es lo siguiente:

- Matriz de viajes
- Modelo de predicción de viajes

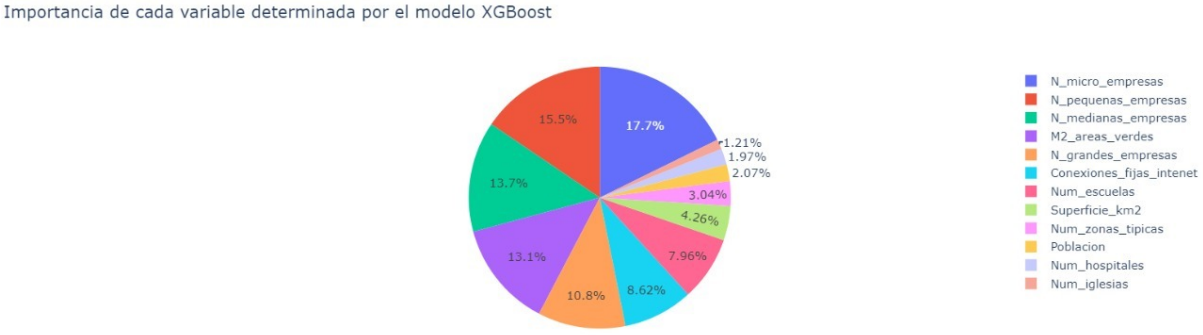
Plan de entrega:

- Se entregará la documentación y codificación para la generación de la matriz
- Se entregará la documentación y codificación para la generación y entrenamiento del modelo

Estas entregas se llevarán a cabo por medio del repositorio de GitHub en donde está almacenado todo lo mencionado previamente.

Hallazgos

- La mayor cantidad de viajes dentro de Santiago de Chile se llevan a cabo dentro de la misma comuna.
- Representación de los viajes entre distintas comunas.
- Los principales atractores de viajes son:



La entrega de los hallazgos se harán por medio de una presentación y la entrega de un reporte en donde dichos hallazgos estarán mayormente detallados.

Beneficios
Los beneficios de los entregables provienen de aquel uso que les haga el socio formador a los modelos y hallazgos encontrados, debido a que no se tiene conocimiento sobre cuáles son los siguientes pasos del socio con respecto al trabajo realizado. Por este motivo solo se harán propuestas sobre como estos pueden llegar a ser utilizados:

- Toma de decisiones al momento de realizar nuevas construcciones, ya que con esto sabrán que tanto afectaría la construcción de estas

- En caso de querer implementar nuevas rutas de transporte se puede ver realmente en donde son más necesarias.
- Incremento del conjunto de datos a más días del año para tener mayor certeza de las predicciones.

Plan de monitoreo y mantenimiento

Dado que no es un producto o servicio que se va a instaurar dentro de la organización, su monitoreo y mantenimiento no será necesario.

Reporte final

Reporte final

Presentación final

https://www.canva.com/design/DAFTZioGvnA/tB82oxEnhLYiSfWG2JXjIA/view?utm_content=DAFTZioGvnA&utm_campaign=designshare&utm_medium=link&utm_source=publishsharelink

Revisión del proyecto

¿Qué salió bien?

- Se pudieron lograr los objetivos de negocio y de minería de datos definidos al inicio del proyecto, brindando así la satisfacción del socio formador.
- El tiempo de finalizar las fases fue mejor a lo estimado en el plan, esto nos permitió obtener una retroalimentación temprana sobre los entregables, permitiéndonos implementar aquellas recomendaciones que nos hizo el socio formador en tiempo y forma.
- La implementación de CRISP-DM en el desarrollo del proyecto.
- Entendimiento de los datos, esto fue necesario antes de iniciar el procesamiento de ellos.

¿Qué salió mal?

- El tardar en encontrar una solución al lento procesamiento de los datos, esto provocó que el tiempo de desarrollo fuera más al esperado. Esto fue solucionado dividiendo el dataset en diferentes partes y con ello dividir el procesamiento entre más dispositivos.
- El acotarse a un alcance corto para la solución planteada, esto se vio reflejado en las recomendaciones que se nos hicieron al momento de validar el modelo.
- La mala organización de versiones de los datos generados.

¿Qué se hizo bien?

- La comunicación con los stakeholders fue muy buena, ya que no hubo confusiones con las necesidades de cada uno de ellos, permitiendo que no hubiese retrasos con el desarrollo de los productos.
- La toma de decisiones para solucionar los problemas encontrados a lo largo del proyecto.
- Uso de librerías que nos permitieron encontrar los mejores hiperparametros de manera eficiente.

¿Qué se necesita mejorar?

- No salirse de lo estipulado dentro del plan.
- Tomar acciones correctivas al ver que el plan no se está siguiendo.
- Involucramiento de todos los miembros del equipo en la realización de tareas.

Enfoques engañosos

- Eliminar las velocidades mayores a 150 km/h de forma incorrecta, este error se solucionó a través del análisis del resultado de dicha operación.
- Usar framework no apto para procesar Big Data.

Pistas para seleccionar técnicas adecuadas para minería de datos

- Cuando el procesamiento de datos sea muy difícil para los dispositivos disponibles, es mejor dividir los datos y trabajar cada una de las divisiones en distintos dispositivos.
- Buscar los mejores hiperparametros dentro de aquel modelo que tenga mejor precisión con sus valores base.

Reportes individuales

- Diego Arturo Padilla Domínguez: Como experiencia, aquello que fue más impactante fue la preparación de los datos, dado que en un inicio pensé que esta etapa sería la más sencilla del proyecto, pero, por el contrario, fue la etapa que más tiempo consumió dentro del desarrollo. Hubo partes en la construcción de datos en donde había confusiones en la lógica al momento de la codificación, además de que por el problema del tiempo de procesamiento de los datos no solo se tenía que buscar un algoritmo que solucionara el problema, sino que también se debía de buscar el que fuese el más eficiente para reducir el tiempo de ejecución lo más posible.

  Todo esto me permitió aprender a buscar soluciones y desarrollar el pensamiento lógico más allá de lo que había hecho en proyectos anteriores.

- Carolina Herrera Martínez: Personalmente considero que la parte de modelación es de las partes más retadoras del proyecto, ya que me permitió junto a mis compañeros a buscar diferentes alternativas para de esta manera observar cual es la que se adecuaba más con el proyecto, Algunos de los tipos de modelos que probamos fueron: Regresión lineal, Red neuronal MLP, XGBoost, etc. Siendo XGBoost el ganador. Tomando en cuenta que es una mejora de un árbol de decisión, vimos que nos dio una mayor precisión en las predicciones de los viajes.
  Otro punto de vital importancia y de gran impacto dentro de mi equipo es como integramos Crisp DM en el desarrollo del proyecto, como cada una de sus fases

desde el Business Understanding hasta el Deployment. Ya que sin esta metodología presente en nuestro proyecto, no hubiéramos obtenido el éxito del proyecto en su totalidad. Nos brinda organización, calidad y nos ayudó a evitar retrasos y contratiempos.
De igual manera creo que es de vital importancia mencionar la constante comunicación que se debe tener con el Stakeholder, aun cuando el stakeholder tiene una baja disponibilidad de tiempo. En base a esto, aprendimos a optimizar nuestras reuniones para de esta manera comunicarnos eficientemente.
En cuanto a la preparación de datos creo que nos retrasó como equipo, ya que no contabamos que se iba a tomar tanto tiempo en hacer entender la lógica a la computadora y la aplicara a un volumen muy grande de datos. Por lo cual nos obligó a buscar diferentes alternativas de solución de las cuales dividir el dataset en 5 partes fue la más óptima. Una vez más esto me enseñó a buscar diferentes soluciones.

- Cutberto Arizabalo Nava : Uno de los retos más destacables de este proyecto fue la fase de preparación de datos, ya que nunca había manejado un volumen de datos tan grande, y no dimensionaba la complejidad computacional de la tarea hasta que me enfrenté a scripts con tiempos de ejecución de más de 12 horas. Fue bastante retador en su momento, y esto me llevó a revisitar temas como la paralelización con multi threads, así como la aplicación de estrategias de divide y conquista. Otro aspecto retador fue aprender a interpretar los modelos generados, ya que hay una diferencia muy grande entre solo crear un modelo para predecir, y saber interpretar los elementos de un modelo para darle valor al mismo.
  Así como hubo cosas retadoras, también hubo herramientas que me ayudaron a afrontar el reto, como lo fue la metodología CRISP-DM, ya que gracias a ella tuvimos un proyecto que fluyó sin contratiempos. Considero que este tipo de aprendizajes me han preparado para desenvolverme mucho mejor en el contexto de la analítica de datos, minería de datos e inteligencia artificial.

- Keyuan Zhao : Considero que aprendí mucho en la forma de analizar los datos y buscar la manera de usar algún framework para procesarlos, ya que el volumen de los datos es muy grande y aumenta la complejidad del procesado. En esta parte no fue favorable para nuestro equipo porque tuvimos dificultades de hacer el procesamiento de datos, uno de los scripts se tardó más de 9 horas en procesar en vez de 1 hora según nuestros cálculos, al final tuvimos que dividir el dataset en partes pequeñas para “optimizar” el tiempo. Por otro lado, también me llevo de aprendizaje cómo adaptar el modelo de CRISP-DM a este reto, ya que sin tener conocimiento de las fases que contiene dicho modelo, podríamos entrar en caos sin saber cómo empezar.

- José Pablo Cobos Austria : Tras haber finalizado con el proyecto, ha sido una experiencia bastante interesante. Ya que hubo fases donde hubo problemas o situaciones adversas que no nos habíamos esperado encontrar, siendo esta específicamente en la fase de data preparation, en la cual para poder obtener los datos que necesitábamos para el modelo, inicialmente pensamos que iba a ser una tarea sencilla y que no iba a tomar mucho tiempo. El caso real fue totalmente

opuesto porque nos tomó mucho tiempo probar las opciones para tratar de disminuir este tiempo, consideramos varias opciones, desde dividir la información, hasta pypark. Además otro punto que en lo personal me costó trabajo, fue mucho la interpretación estadística de los modelos, ya que a veces los confundía o no los reconocía bien las métricas y eso me generaba problemas. Finalmente, lo que más captó mi atención y siento que genere más conocimiento fue en la parte de Deep Learning y en el mismo CRISP-DM. El primero porque se me hizo muy atractivo como era el funcionamiento de las redes funcionales y su aplicaciones, y el segundo, porque nunca en la carrera había tenido que basarme en una metodología de ese estilo y se me hizo muy útil para poder llevar un control eficaz y correcto de un proyecto

Retroalimentación

- El socio formador está satisfecho con el resultado del proyecto.
- Las mejoras solicitadas fueron implementadas, por ello no requieren mejoras dentro

¿Qué funcionaron bien?

- Dividir el dataset en partes correspondientes y procesarlo por separado, ya que esto fue la forma que encontramos para optimizar el proceso.

¿Qué error cometieron?

- Al no poder adaptar un framework correcto para procesar el dataset, tuvimos que dividir el dataset para reducir el tamaño en un solo ejecución.

¿Cuáles fueron las lecciones aprendidas?

- Buscar más alternativas para optimizar el procesamiento de los datos.
- Consultar y recibir retroalimentación de los expertos para avances del proyecto.
- Satisfacer la necesidad del socio formador con lo que requiere y no con lo que nosotros creemos que necesita.
- La implementación de una metodología es de gran importancia para el desarrollo de un proyecto, ya que nos permite tener una línea a seguir y no hacer las cosas sin un orden en específico y sin sentido.
- La integración entre todos los miembros del equipo es de gran importancia, debido a que si hay confianza se puede tener una mejor comunicación y como resultado la eficiencia del equipo se ve incrementada.

¿Cómo puede mejorarse el proceso y la experiencia?

- Implementando la mejora continua dentro del proceso, viendo continuamente que es lo que se puede cambiar durante la experiencia.
- Tener una retroalimentación más continua entre el socio formador y el equipo hubiese permitido tomar decisiones de manera más eficaz.