Entregable Final: Inteligencia artificial avanzada para la ciencia de datos II

Cutberto Arizabalo Nava
A01411431
ITC
ITESM
Santiago de Querétaro, Querétaro
A01411431@tec.mx

Jose Pablo Cobos Austria
A01274631
IRS
ITESM
Santiago de Querétaro, Querétaro
A01274631@tec.mx

Carolina Herrera Martínez
A01411547
ITC
ITESM
Santiago de Querétaro, Querétaro
A01411547@tec.mx

Diego Arturo Padilla Domínguez
A01552594
ITC
ITESM
Santiago de Querétaro, Querétaro
A01552594@tec.mx

Keyuan Zhao
A01366831
ITC
ITESM
Santiago de Querétaro, Querétaro
A01366831@tec.mx

1. Introducción

En la ciudad de Santiago de Chile, la empresa telefónica Movistar ofrece sus servicios al 25 % del mercado de telefonía móvil [5]. Para brindar su servicio, Movistar cuenta con antenas telefónicas repartidas estratégicamente a lo largo de las distintas comunas de Santiago.

Utilizando los registros de un día aleatorio del año 2021 de las conexiones de los dispositivos móviles con dichas antenas telefónicas, nos fue posible aplicar técnicas de Data Science para conocer los viajes que se realizan entre las distintas comunas.

El presente proyecto busca identificar cuáles son los principales atractores de viajes hacia una comuna mediante la creación de un modelo predictor utilizando técnicas de Machine Learning. Para la creación de dicho modelo se utilizaron dos fuentes principales de información: la cantidad de viajes realizados a cada hora entre las comunas de Santiago, así como información recabada sobre las características de cada comuna.

Asimismo, se busca crear un método accesible de acceder a la información obtenida con Data Science sobre los viajes entre las comunas de la ciudad de Santiago.

2. Estado del Arte

Al buscar información relacionada con las técnicas y modelos del estado del arte actual, encontramos un método de modelación que nos fue de gran utilidad: XGBoost. Asimismo, encontramos también una buena estrategia de refinamiento automatizado de modelos, el algoritmo Grid Search CV.

Basados también en el estado del arte, elegimos como métrica de evaluación el MAPE, así como también seleccionamos K-Fold Cross Validation como nuestro método de validación. A continuación, se describe detalladamente cada uno de los elementos mencionados.

Selección de un modelo óptimo

Para la modelación se utilizó el método XGBoost (Extreme Gradient Boosting). Este es un método de ensamble para aprendizaje automático supervisado, el cual se basa en árboles de decisión y representa una mejora con respecto a métodos como el Random Forest gracias a la implementación de diversos métodos de optimización.

El método de XGBoost consiste en iniciar la modelación con una predicción inicial, para posteriormente calcular residuales en función de la predicción y el valor actual. Con base en estos residuales se crea un árbol de decisión, y a través de este árbol se busca la reducción de estos residuales. Al final, la salida de este árbol servirá para construir un nuevo árbol, y este proceso se repetirá hasta que ya no se reduzcan los residuales, o hasta que se alcance el número máximo de iteraciones [14].

Refinamiento automatizado del modelo

Para el refinamiento del modelo se utilizó un algoritmo llamado Grid Search CV. Este algoritmo nos permite hallar los valores óptimos de los hiperparámetros para obtener

la mayor precisión posible. Para encontrar dichos valores óptimos, el algoritmo realiza una búsqueda de fuerza bruta probando todas las posibles combinaciones y comparando la precisión obtenida con cada una de las distintas combinaciones probadas.

Métrica de evaluación del modelo

Para verificar la calidad y validez de los datos se utilizó como métrica el porcentaje medio de error absoluto de cada dataset (MAPE).

$$M = \frac{1}{n}\sum_{t=1}^{n} = |\frac{A_t - F_t}{A_t}| \quad (1)$$

Se decidió optar por el uso MAPE porque se utiliza a menudo en la práctica debido a su interpretación muy intuitiva en términos de error relativo, permitiendo mostrar resultados que todos los stakeholders puedan comprender sin necesidad de ver todo el análisis previo[8].

Una de sus características es que sus distribuciones de error subyacentes de estas medidas únicamente tienen valores positivos y no tienen límite superior, los errores porcentuales son muy propensos a la asimetría a la derecha en la práctica real, pero esto no debe de ser algo de qué preocuparnos, ya que por el contexto no es una desventaja[11].

Una de sus principales ventajas es que aunque haya un cambio en la escala de los datos, la escala de la métrica sigue siendo la misma, siendo esto muy útil al momento de hacer distintas iteraciones en los modelos.

El MAPE ofrece las mismas propiedades que el MSE y el RMSE, pero se expresa en porcentajes[13],

Otro factor para no elegir otra métrica es que se tiene planteado el uso de variables indicatrices (dummie), este tipo de variables provocan que el uso de otras métricas como R2 se vean afectadas, provocando que su valor no sea correcto.

Validación del modelo

La validación cruzada ayuda a evaluar los modelos de aprendizaje automático. Este método estadístico ayuda a comparar y seleccionar el modelo en el aprendizaje automático aplicado. En la validación cruzada k-fold, el conjunto de datos se divide en un número K de partes. Una vez realizadas estas divisiones, se realiza un ciclo de evaluaciones que se describe a continuación [3].

Para la primera iteración, se toma el primer slice del dataset para hacer validación, y el resto del dataset se utiliza para entrenamiento. Después, se califica el rendimiento del modelo y se finaliza la iteración. Este proceso se repite en cada iteración, utilizando el slice correspondiente al número de iteración para validar y dejando el resto de slices para entrenamiento [3].

Una vez que se han completado las K iteraciones, se promedian los resultados de cada fase de validación para así conocer el desempeño del modelo de forma más completa.

3. Método

La metodología de trabajo utilizada en el presente proyecto fue CRISP-DM, pasando por las fases de:

- Entendimiento del Negocio
- Entendimiento de los Datos
- Preparación de los Datos
- Modelación
- Evaluación
- Despliegue

A continuación, se describe la forma en que se generaron los dos pilares de este proyecto: el tablero de visualización de viajes y la modelación de los atractores de viajes.

Obtención de los viajes realizados entre las comunas de Santiago

Como punto de partida se utilizaron los registros de conexión de dispositivos móviles con las antenas de Movistar durante las 24 horas de un día aleatorio del año 2021. Para convertir estos datos en información, fue necesario delimitar parámetros que nos permitieran definir lo que es un viaje dentro de nuestro contexto.

Entre los distintos supuestos que se hicieron para poder definir los viajes, se encuentran los siguientes:

- Un viaje inicia cuando un usuario cambia de la antena A a la antena B
- Un viaje finaliza cuando el usuario tiene dos o más conexiones consecutivas a la misma antena que sumen 20 minutos o más.
- Un viaje no puede durar más de 2 horas. Si después de 2 horas se siguen registrando cambios, se debe hacer un corte y registrar los eventos ocurridos después de dos horas como un nuevo viaje.
- Un viaje no puede tener una distancia recorrida superior a los 25 kilómetros. Si después de 25 kilómetros se sigue en viaje, se debe hacer un corte y registrar los eventos ocurridos después de los 25 kilómetros como un nuevo viaje.
- En un viaje no se pueden tener velocidades superiores a los 125 km/h. Como parte de la limpieza de datos, se deben remover los datos de aquellas conexiones que indiquen un movimiento con velocidades iguales o superiores a los 125 km/h.

Siguiendo estos supuestos, se generó una matriz de viajes con el siguiente formato

Origen	Destino	Hora de Inicio	Hora de Fin	Distancia
(Ejemplo) Buin	Lo Barnechea	17:58:35	18:33:50	12.5 km

Tabla 1. FORMATO DE MATRIZ DE VIAJES

Visualización de los viajes realizados entre las comunas de Santiago

Para crear una forma accesible de acceder a la información de los viajes entre comunas, se construyó un tablero interactivo en formato de Heatmap utilizando la herramienta Tableau. En dicho tablero informativo es posible ver la cantidad de viajes realizados entre comunas, así como aplicar filtros de acuerdo a atributos como la cantidad mínima y máxima de viajes, las comunas a analizar, entre otras cosas.

Construcción de los modelos preliminares para analizar atractores de viajes

En nuestra modelación para analizar los principales atractores de viajes hacia una comuna, se dividió el dataset total en dos partes: 85 % para entrenamiento y 15 % para pruebas. Se utilizaron las siguientes variables:

Variables independientes:

- Número de escuelas
- Número de hospitales
- Número de iglesias
- Número de zonas típicas
- 23 variables dummies para las horas del día

Variable dependiente:

- Número de viajes hacia la comuna

Para la generación del modelo final, se crearon múltiples modelos intermedios. El primer modelo intermedio fue nuestro modelo benchmark de regresión lineal, el cual nos sirvió para medir las mejoras obtenidas en cada nuevo modelo generado.

A raíz de este modelo se decidió explorar dos ramas; la de los modelos basados en redes neuronales y la de los modelos basados en árboles de decisiones.

Como primer paso, se crearon dos modelos nuevos: un modelo de Red Neuronal (Multi Linear Perceptor) y un modelo de Random Forest. Para ambos modelos se utilizó el algoritmo GridSearchCV para la optimización de sus hiperparámetros.

Al comparar el rendimiento de ambos modelos, hallamos que para nuestro dataset, el modelo de Random Forest nos brinda una precisión superior, ya que se tuvo un MAPE de 67.23 % en la Red Neuronal MLP, comparado con un MAPE de 15.05 % en el modelo de Random Forest

Debido a que la Red Neuronal MLP nos ofreció una precisión inferior con nuestro volumen de datos de entrenamiento, se tomó la decisión de no explorar más opciones relacionadas con redes neuronales, descartando así los modelos de Deep Learning también. Se decidió seguir explorando los modelos basados en árboles.

A continuación, se utilizó el método XGBoost para la creación de nuestro siguiente modelo. Para el refinamiento de este modelo se decidió complementar el resultado de Grid Search CV con un ajuste manual a los hiperparámetros, para de esta forma evitar el overfitting causado por la fuerza bruta del algoritmo de refinamiento. Este modelo nos brindó una mejora del 1 % en la precisión comparado con el modelo creado previamente con Random Forest.

Comparación de los modelos preliminares

En el siguiente cuadro comparativo podemos observar una comparación entre los resultados obtenidos por cada modelo, así como la configuración utilizada para cada uno de ellos.

Modelo	Hiperparámetros/ Configuración	Score Train (%MAPE)	Score Test (%MAPE)
Regresión Lineal	• 26 Variables (23 dummies)	63.55%	87.80%
Random Forest	• 27 Variables (23 dummies) • n_estimators: 100 • max_depth : None • min_samples: None	14.98%	15.05%
XGBoost	• 27 Variables (23 dummies) • colsample_bytree: 0.1 • learning_rate: 0.1 • max_depth: 3 • n_estimators: 20000	10.72%	14.89%
Red neuronal MLP	• 27 Variables (23 dummies) • random_state=1, • max_iter=1000000, • learning_rate="adaptive", • activation = "logistic"	58.31%	67.23%

Tabla 2. COMPARACIÓN DE LOS MODELOS PRELIMINARES

Derivado del rendimiento de cada modelo, se tomó la decisión de elegir XGBoost como nuestro modelo final, ya que este nos ofrece una mayor precisión que el resto de opciones.

Construcción del modelo Final

Como último paso para la obtención de la versión final del modelo, se decidió ampliar la cantidad de variables con la intención de identificar si existía alguna variable con una mayor capacidad para describir el fenómeno de los viajes recibidos en cada comuna.
Nuevo listado de variables independientes:

- Número de escuelas
- Número de hospitales
- Número de iglesias
- Número de zonas típicas
- Superficie en Km^2
- Población
- Conexiones fijas de internet
- M^2 de áreas verdes
- Número de microempresas
- Número de empresas pequeñas
- Número de empresas medianas
- Número de empresas grandes
- 23 variables dummies para las horas del día

Variable dependiente:

- Número de viajes hacia la comuna

De esta forma fue que se llegó al modelo final, el cual se construyó usando el método XGBoost. Para su creación, primero se midió el desempeño del modelo usando los valores por default de la implementación de XGBoost del framework SkLearn.

Posteriormente, se aplicó el algoritmo GridSearchCV para la búsqueda de hiperparámetros óptimos. Esto resultó en un modelo con overfitting, por lo que procedimos a realizar un refinamiento manual a partir de los hiperparámetros encontrados por GridSearchCV para evitar el overfitting.

Una vez finalizado el refinamiento, se evaluó el modelo utilizando MAPE y K-Folds Cross Validation con un K value igual a 10. Se eligió esta cantidad de folds, ya que al usar 10 folds se suele tener un bias bajo y una varianza moderada en los resultados, representando muchas veces un punto óptimo [3].

4. Resultados

En esta sección se muestran los resultados obtenidos relacionados con los dos principales objetivos del proyecto: la generación del tablero de viajes y la evaluación y hallazgos del modelo para analizar los atractores de viajes.

Tablero de Viajes Origen-Destino

A continuación podemos observar el primer tablero interactivo generado para visualizar los viajes dentro de Santiago de Chile. El heatmap contiene los viajes realizados de manera interna en las comunas, así como los viajes entre distintas comunas.

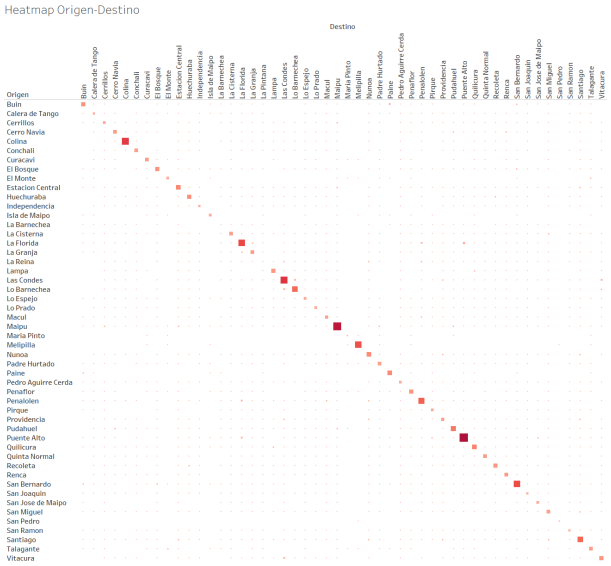

Figura 1. Heatmap de viajes dentro de la región metropolitana de Santiago de Chile.

En la siguiente imagen podemos observar el segundo tablero construido. Este heatmap contiene únicamente los viajes realizados entre distintas comunas, excluyendo los viajes internos.

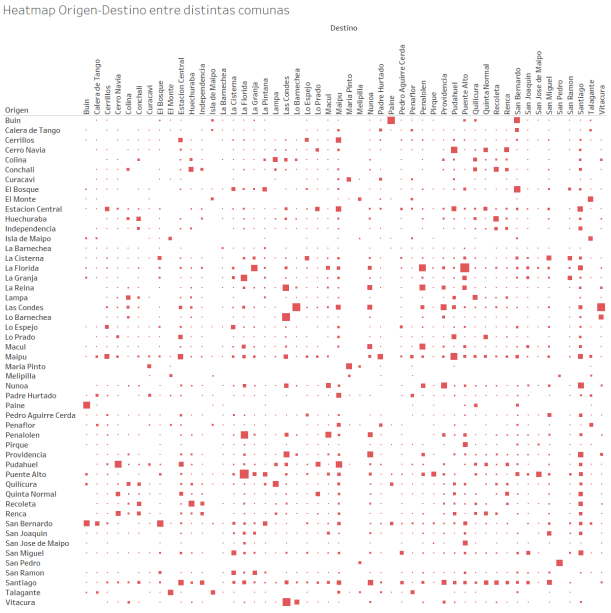

Figura 2. Heatmap de viajes entre distintas comunas dentro de la región metropolitana de Santiago de Chile.

Los hallazgos que obtuvimos son los siguientes:

- Hallazgo 1: En la Fig.1 podemos observar que los viajes son principalmente dentro de las mismas comunas, esto tiene alto sentido debido a que normalmente todas las personas buscan lugares de trabajo, estudio y/o recreación lo más cercano posible a su hogar.

- Hallazgo 2: Debido a que solo se ve una clara representación de los viajes dentro de las mismas comunas y siguiendo el objetivo, se optó por crear un heatmap en donde estos viajes no estuviesen contabilizados (Fig. 2), con ello se observa de mejor manera aquellas comunas en las que hay más viajes entre sí.

Evaluación del modelo final e insights obtenidos

Con el último refinamiento del modelo final XGBoost, obtuvimos 10.4 % de error (MAPE) en train y 16.9 % de error (MAPE) en test. Asimismo, se obtuvo una precisión del 85.78 % mediante la validación cruzada de K-Folds.

A continuación se muestra un análisis de los hallazgos obtenidos en dicho modelo.

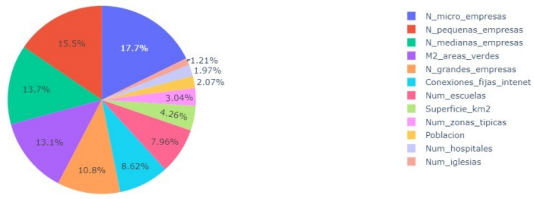

Figura 3. Importancia de cada variable en porcentaje

Observando la Fig.3 podemos observar que el principal atractor de viajes son las microempresas, considerando que los pequeños negocios suelen tardar más en adoptar el proceso de transformación digital, es más probable que los empleados de dichos negocios acudan a trabajos presenciales en época de pandemia. Asimismo, vemos que conforme el tamaño de la empresa es mayor, representa un menor impacto de movilidad, estando los 4 tamaños de empresas entre los primeros 5 lugares de factores con mayor impacto. Esto probablemente se debe a la adopción del home office por las empresas de mayor tamaño, reduciendo la movilidad de sus empleados en tiempos de dicha pandemia.

5. Discusión

Trabajo a futuro

Un acercamiento posible para analizar los atractores de viajes sería crear un modelo basado en series de tiempo. Para elaborar este modelo es necesario contar con información de un rango de tiempo mucho mayor a un día, ya que sería de mucho valor poder modelar utilizando meses, o años de información de movilidad.

Sin embargo, la creación de dicho modelo supone múltiples inconvenientes, siendo uno de los más importantes la privacidad de los datos. En el estado actual del dataset solo se cuenta con un día de registros de conexión, pero si se contara con la información de meses o años, se corre el riesgo de que se puedan identificar patrones de comportamiento de los usuarios anónimos, lo cual culminaría en el rompimiento de la anonimidad de las personas. Esto es un problema de privacidad grave, ya que nos permitiría saber cosas como la casa de una persona, su escuela/trabajo, el horario en el que se suele encontrar en casa, los lugares que frecuenta, etc.

Otro inconveniente es el volumen de los datos, por el hecho de que al analizar los registros de tan solo un día, el dataset pesa aproximadamente 5 GB. En caso de extrapolar esto, estamos hablando de aproximadamente 150GB de información para un mes, y aproximadamente 1.8 TB de información para un año. Si se desea trabajar con dicho volumen de datos, es necesario utilizar estrategias avanzadas de procesamiento de big data, además de necesitarse un poder de cómputo superior al de los ordenadores domésticos.

Sin embargo, contar con un modelo de ese nivel nos permitiría hacer predicciones mucho más precisas de la cantidad de viajes que recibirá cada comuna de acuerdo a las características de la misma, por lo que es importante evaluar si este beneficio supera a las desventajas y hace que valga la pena la realización de este nuevo modelo.

Cambios en el plan

Uno de los cambios más importantes en nuestro proyecto fue la modificación de un objetivo de negocio. En un inicio se tenía la meta de predecir cuál sería la mejor hora del día para hacer obras públicas, considerando que la hora del día con menor movilidad urbana sería la hora ideal.

Se tuvo en cuenta que el dataset actual solo abarca la información de un día, pero el modelo predictivo estaba pensado para ser escalable y poder recibir la información de múltiples días para hacer una modelación basada en series de tiempo.

Al compartir la idea con el cliente, se resaltó el hecho de que los resultados obtenidos con el modelo entrenado con el dataset de un solo día no serían representativos.

A raíz de esto, comenzamos a darle más valor a los resultados que se pueden obtener utilizando la versión actual de los datos disponibles, dejando de lado las ideas que se basan en utilizar versiones del dataset que actualmente no están a nuestro alcance. Sin embargo, resulta interesante la idea de diseñar modelos escalables, los cuales puedan ser construídos usando una versión pequeña del dataset para posteriormente alimentarlo con una versión mucho más robusta del mismo.

Lecciones aprendidas

Durante la exploración de los diversos métodos para crear modelos de ML, aprendimos que no todos los modelos son aptos para resolver cualquier problema.

Existen modelos que destacan de acuerdo a las cualidades del dataset con el que se trabaja, por lo que es importante realizar una exploración amplia de las opciones existentes.

La viabilidad del procesamiento de los datos no debe de ser solamente teórica, dado que a pesar de que la teoría diga que el procesamiento de los datos será posible, no siempre es correcto, debido a que esto puede variar con respecto a cada dispositivo y/o herramienta utilizada, por ello siempre se deben de realizar pruebas en un inicio y buscar alternativas en caso de encontrarse con algún problema.

Aprendimos también a priorizar la creación de modelos que puedan darnos resultados con un significado que aporte valor en este momento, en lugar de priorizar aquellos modelos que dependan de información con la que no se cuenta actualmente.

Referencias

[1] Aasa A. Silm S. Tiru M Ahas, R. Daily rhythms of suburban commuters' movements in the tallinn metropolitan area: Case study with mobile positioning data, 2010. URL: https://www.sciencedirect.com/science/article/abs/pii/S0968090X09000400.

[2] Jiang S. Murga M. González M. C. Alexander, L. A new metric of absolute percentage error for intermittent demand forecasts, 2015. URL: https://www.sciencedirect.com/science/article/abs/pii/S0968090X1500073X?via\%3Dihub.

[3] Nisha Arya. Why use k-folds cross fold validation? kdnuggets, 2016. URL: https://bit.ly/3VngDpS.

[4] Freaza Nadia etc. Aón Laura, Giglio María. Los atractores de viajes como concepto operacional en el estudio de la movilidad urbana. revista de transporte y territorio 23, 2020. URL: https://bit.ly/3VInOsB.

[5] BnAmericas. El sector de telecomunicaciones de chile bajo la lupa, 2022. URL: https://bit.ly/3P3905P.

[6] Hombourger E. Olteanu-Raimond A. M. Smoreda Z Bonnel, P. Passive mobile phone dataset to construct origin-destination matrix: potentials and limitations, 2015. URL: https://bit.ly/3VBMpz9.

[7] Diao M. Di Lorenzo G. Ferreira Jr J. Ratti C Calabrese, F. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example, 2013. URL: https://bit.ly/3XHrn3T.

[8] B.; Le Grand B.; Rossi F De Myttenaere, A.; Golden. Mean absolute percentage error for regression models, 2017. URL: https://bit.ly/3Fe1C3W.

[9] Saez-Trumper D Graells-Garrido, E. In proceedings of the second international conference on iot in urban space, 2016.

[10] Peredo O.- García J Graells-Garrido, E. Sensing urban patterns with antenna mappings: the case of santiago, chile, 2016. URL: https://bit.ly/3Fe8ONs.

[11] Kim H Kim S. A new metric of absolute percentage error for intermittent demand forecasts, 2016. URL: https://bit.ly/3ijuDSR.

[12] s/a. Modalidad híbrida: La estrategia de algunos colegios de cara al regreso a clases en: https://www.facebook.com/teletrece, 2020. URL: https://bit.ly/3Vi2hXK.

[13] J.; Bryan-T.M Swanson, D.A.; Tayman. Mape-r: A rescaled measure of accuracy for cross-sectional, subnational forecasts, 2011. URL: https://bit.ly/3VBrnkf.

[14] Chen Tianqi y Guestrin Carlos. A new metric of absolute percentage error for intermittent demand forecasts, 2016.