Predicción del gasto total en telefonía móvil por manzana censal para Santiago de Chile

Gonzalo Silva
Estudiante de Ciencia de datos
Barcelona, 27/06/2023

Resumen

El presente trabajo intenta a partir de un conjunto de variables sociodemográficas, predecir el gasto total en telefonía móvil de los hogares de algunos municipios de Santiago de Chile a nivel de manzana. En una primera parte y mediante código python se explora el dataset para graficar y respresentar la realidad del territorio en cuestión. Una segunda parte dedicada al machine learning y a los modelos de regresión, preprocesa, analiza y produce los principales resultados de la investigación. Evaluando y concluyendo con la mayor objetividad posible las mejoras a realizar.

1. INTRODUCCIÓN

La industria de las telecomunicaciones es uno de los sectores más competitivos de las últimas décadas en Chile. Cinco actores compiten ferozmente por arrebatarle una parte de la cuota de mercado a sus competidores. Con la evolución de las tecnologías, las telecomunicaciones hoy son parte del consumo básico de la población, sin lugar a dudas. El nivel de competitividad descrito anteriormente, sobre una ciudad con altos niveles de segregación socioespacial, desigualdad en la distribución de la riqueza, y un alto consumismo, consolidan un mercado muy atractivo para empresas que expanden su operación sobre mercados poco regulados y bajas tazas impositivas. En este contexto, la lucha por conseguir nuevos clientes se vive a pie de calle, y con ello el valor de analizar datos a escala de manzana censal.

El objetivo de la actual investigación, es explorar técnicas de machine learning para lograr predecir el gasto total en telefonía móvil por manzana censal, de manera de obtener un modelo capaz de ser aplicado no solo en la región Metropolitana de Chile sino que en el resto de ciudades donde normalmente la data disponible es insuficiente. De esta manera, se podría obtener una buena fuente de conocimiento para poder apoyar el direccionamiento estratégico y comercial en el negocio de las telecomunicaciones.

2. ESTADO DEL ARTE

Predecir cuánto gastará la gente en diversos tipos de bienes o servicios no es algo nuevo. Las predicciones más conocidas para servicios como por ejemplo, en telecomunicaciones, normalmente se realizan a escalas de análisis muy grande[1]. Predicciones a nivel mundial o por pais son un nivel de análisis al que normalmente estamos acostumbrados. Por otra parte y

[1] https://es.statista.com/estadisticas/636470/prevision-del-gasto-mundial-en-telefonia-inalambrica--2019/

desde la rama del geomarketing o micromarketing[2], las empresas logran realizar estimaciones de cuota de mercado según target y traducirlos a nivel territorial, para finalmente predecir gasto por servicios que una familia u hogar normalmente demanda.

3. METODOLOGÍA

a. Objetivo de la Investigación

El objetivo del presente trabajo es la predicción del gasto total por manzana censal en telefonía Móvil para el Grupo Socioeconómico C2. La varable target o variable a predecir es "G_TM_C2". Es una variable numérica y de tipo flotante, que refleja en pesos chilenos la estimación de la sumatoria de lo que todos los hogares o familias están dispuestos a pagar por este servicio de telefonía móvil de manera mensual.

b. Base de datos

La metodología de la presente investigación incia con una base de datos producida de manera personal y en colaboración con otros profesionales del sector de la analítica de datos geográficos, para un proyecto desarrollado para Santiago de Chile para el ámbito de las telecomunicaciones. Dicha base de datos consta de 70.000 registros (manzanas censales), agrupando mas de 30 municipios de la ciudad de Santiago de Chile.

c. Área de estudio

Se escogieron 8 municipios que representan de manera heterogénea la diversidad socioeconómica de la ciudad de Santiago y de Chile en general. Los 8 municipios son: Cerro Navia, La Pintana, Pudahuel, El Bosque, Maipú, La Florida, Las Condes y Providencia. La estratificación GSE desarrollada por la asociación de investigadores de mercado más importante de Chile (GSE[3]), es sin duda la característica más importante del dataset en cuanto a diferenciación del perfil social y económico de las manzanas censales. Es importante considerar que la manzanas censales de estos municipios disponibles para el dataset representan más del 98% de la población municipal.

d. Estructura del trabajo y código.

El código esta desarrolado bajo lenguaje Python y utilizando como interpretador Jupyter notebook. La estructura del código podemos resumirla de la siguiente manera:

Parte 1	Contexto
Parte 2	Exploración preliminar de los datos
Parte 3	Eliminación de columnas previo al preprocesado
Parte 4	Preprocesado
Parte 5	Aplicación de Modelos de Regresión.
Parte 6	Graficando la predicción
Parte 7	Visualizando la predicción en un dataframe
Parte 8	Comentarios y conclusiones

e. Exploración de datos y eliminación de columnas.

Una primera parte del trabajo desarrolla en detalle la caracterización sociodemográfica del área de estudio. Son 24 variables inciales, donde el 70% de ellas solo serviará para poner en contexto la problemática y sel

[2] https://www.unica360.com/micromarketing-y-prediccion-de-demanda-%C2%BFcuanto-gastan-en-libros

[3] https://aimchile.cl/gse-chile/

30% para predecir el target. Mediante gráficos y tablas, podemos ver tanto la distribución porcentual del los diferentes GSE del dataset (Fig. 1 y Fig. 2) y también cómo el grupo socioeconómico C1 concentra la mayor proporción del gasto en telefonía móvil.

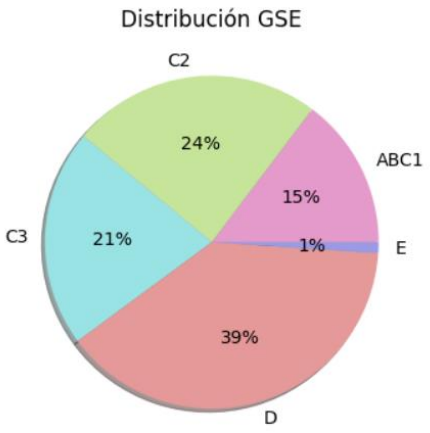

Fig 1. Gráfico que representa la disrtibución de las manzanas del dataset según GSE predominante.

Gasto total acumulado en TM según GSE y municipio

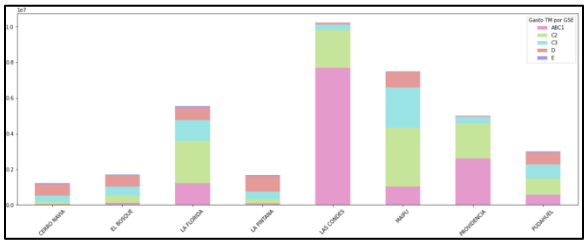

Fig 2. Representación de la distribución del gasto total acumulado por telefonía móvil según GSE por municipio

La exploración final del dataset nos plantea que el grupo C2 se alza como una conjunto interesante puesto que la media de gasto es la mas alta del bloque (todos los GSE). Es decir, este grupo tendría los clientes de mayor valor y tenderían a pagar un plan más que costoso que el resto de grupos.

Finalmente y luego de la exploración se elimina una importante catidad variables por dos motivos principales. El primer grupo de variables eliminadas no representan gran valor si quiera para explorar el dataset. Y por el otro, muchas de las variables no aportan nuevo valor a la explicación del target y no es necesario contar con ellas o que por otro lado han sido producidas utilizando inputs que también se encuentran en la producción de la variable target, por tanto estarían explicando el problema tanto en el train como en el test (al momento de aplicar machine learning).

f. Tipo de distribución y Outliers

Como ya sabemos, una etapa muy importante para el preprocesado posterior, es determinar qué variables presentan distribución gausiana y cuáles no. Por otra parte, es importantísimo saber cualés variables presentan outliers. Las 8 variables númericas han presentado outliers y arrojaron negativo el test de hipótesis de Shapiro al chequear la hipótesis sobre si su distribución es gausiana o no.

g. División Train test.

Primero de todo se ha separado la variable *target* (G_TM_C2) de las *features* en X e Y y luego, se ha utilizado una división de 70% del dataset para entrenamiento del modelo y un 30% para el testeo. Esto porque hemos querido contar con una buena cantidad de datos para entrenar dicho modelo.

h. Preprocesado.

Se ha utilizado la herramienta *Pipeline* para aplicar un preprocesado de forma mucho más automática y orgánica.
A través de *Column Transform* hemos categorizado la variable “GSE” como dummy por tanto hemos aplicado ‘One Hot

Encoder', mientras que para el resto de variables numéricas y sin distribución gausiana hemos aplicado Robust Scaler.

```
Pipeline(steps=[('prep',
                 ColumnTransformer(transformers=[('Dummie',
                                                  OneHotEncoder(drop='first',
                                                                handle_unknown='ignore'),
                                                  ['GSE']),
                                                 ('Robust', RobustScaler(),
                                                  ['POB_0_14', 'POB_30_59',
                                                   'POB_60_MAS', 'PXQ',
                                                   'G_TM_ABC1', 'G_TM_C3',
                                                   'G_TM_D', 'G_TM_E'])])),
                ('LR', LinearRegression())])
```

Fig 3. Pipeline aplicado al preprocesado. One Hot Encoder y Robust Scaler.

i. Modelos seleccionados.

Hemos utilizado dos modelos de regresión lineal múltiple para predecir el gasto total en telefonia móvil para el GSE C2 por manzana censal en Santiago de Chile. Los modelos han sido por lado *'LinearRegression' (LR)* y *'KNeighborsRegressor' (KNR)* y han sido escogidos Los dos modelos utilizados han sido escogidos por su alta flexibilidad y su fácil interpretabilidad.

j. Resultados Obtenidos.

Hemos obtenidos tres resultados de acuerdo a los 3 procesos y ajustes que se han aplicado. Los primeros resultados para el primer ajuste arrojaron tanto para la 'LR' como para el 'KNR' un R2 en promedio de casi 0.9, mientras que como error medio absoluto por un lado el 'LR' alcanza los 297 € y el 'KNR' 163 €. Porpuestos son valores bastante altos en ambas métricas (ver Fig.4).

```
Mean Absolute Error modelo LR: 297.7
R2 modelo LR: 0.887

Mean Absolute Error modelo KNR: 163.4
R2 modelo KNR: 0.917
```

Fig 4. Resultados del primer ajuste o primer *'Fit'*.

Luego, el segundo ajuste producto de la validación cruzada dividida en 5 partes arrojó los resultados de la figura 5.

```
LR
[0.87504242 0.89307247 0.76594154 0.8772583  0.84858363]
Mean Absolute Error: 326.2 +/- 84.4
R2 accuracy: 0.852 +/- 0.045

KNR
[0.88962874 0.84539833 0.90545645 0.8411315  0.74287352]
Mean Absolute Error: 202.7 +/- 63.9
R2 accuracy: 0.845 +/- 0.057
```

Fig 5. Segundo resultado luego de validación cruzada. Segundo *'Fit'*.

Finalmente y luego de la modificación de hiperparámetros, los resultados se mantienen dentro el mismo rango y alcanzaron los valores que se exhiben en la figura 6.

```
Mean Absolute Error modelo LR: 297.69
R2 modelo LR: 0.88675

Mean Absolute Error modelo KNR: 171.00
R2 modelo KNR: 0.89269
```

Fig 6. Tercera iteración post cálculo con los mejores hiperparámetros de los dos modelos. Tercer *'Fit'*.

k. Gráficando resultados.

Los resultados aparentemente son muy buenos tanto en R2 como en MAE. También hemos podido chequear las 5 particiones del dataset con validación cruzada y los resultados son aparentemente muy buenos. Sin embargo y como veremos en el siguiente gráfico ocurren algunos cosas que por ahora no es posible explicar (ver Fig 7). Si bien la curva del modelo parece ceírse bastante al target original, podemos ver que tanto el target como la predicción presentan valores negativos que no existen en el conjunto de datos. Puesto que el gasto no presenta un gasto negativo. Todos son números positivos y flotantes.

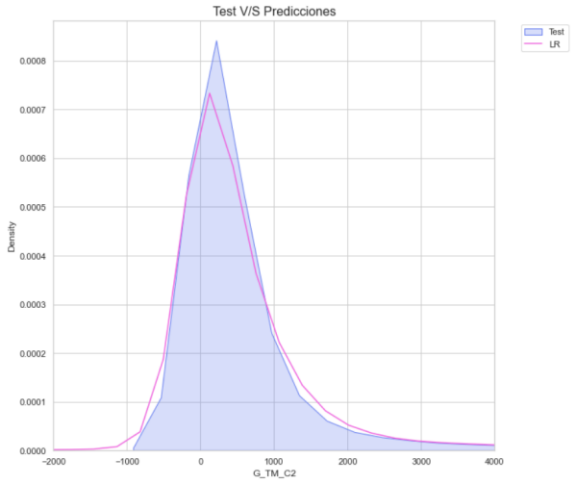

Fig 7. Gráfica de la predicción de la curva de predicción vs Target.

Si observamos por otra parte los datos en el dataframe creado, podremos ver que si bien algunos valores tiene un buen ajuste y adecuado al error y el R2, otros son exactamente idénticos (ver Fig 8.).

Dataframe final con la predicción

G_TM_C2	G_TM_C3	G_TM_D	G_TM_E	GSE	G_TM_C2_PRED
2082.91	370.89	47.60	20.51	C2	1735.76
0.00	0.00	404.58	0.00	D	0.00
486.01	556.33	166.59	10.25	C3	486.01
416.58	927.22	166.59	20.51	C3	347.15
208.29	1019.94	118.99	0.00	C3	208.29
347.15	370.89	713.96	112.80	D	347.15
208.29	185.44	452.17	61.53	D	208.29
208.29	185.44	380.78	51.27	D	208.29
486.01	463.61	356.98	20.51	D	127.96
624.87	139.08	23.80	10.25	C2	624.87

Fig 8. Dataframe final con target (izquierda) y predicción (derecha).

4. COMENTARIOS FINALES Y CONCLUSIONES

Como ya hemos visto en el apartado anterior, hemos seguido el procedimiento para alcanzar la predicción de un target, sin embargo hemos encontrado al final errores evidentes en el cálculo. A pesar del alto R2 calculado y un bajo MAE, la información que nos muestra la representación gráfica a través del análisis de densidad, nos advierte que alguna parte del proceso ha fallado. En primer lugar, los valores arrojados incluyen cifras negativas, cuando los valores de entrada para 'G_TM_C2' (gasto en Telefonía Móvil) toman como mínimo el 0. Es decir, el modelo no refleja del todo cifras coherentes.

Qué podría estar ocurriendo?

Multicolinialidad no abordada de manera adecuada en el preprocesado:

En efecto y según la documentación relativa al machine learning y los algoritmos de aprendizaje automatizado, es necesario un profundo conocimiento de los datos sobre todo si estos presentan características complejas (cómo por ejemplo, la realidad socioeconómica de la población de una ciudad como Santiago de Chile). Valores máximos muy altos y valores mínimos muy bajos, una fracción importantísima de outliers y concentraciones de datos en ciertas manzanas que un modelo no muy robusto podría pasar por alto. Podríamos añadir un chequeo de las hipótesis que estamos entregando me diante las siguientes técnicas de análisis. Un análisis de componentes principales sería una herramienta a utilizar en un futuro para evaluar (sin dejar de perder mucha data) cómo una reducción de dimensionalidad podría aportar a eliminar el ruido que generan ciertos valores de nuestros atributos del dataset. Además, analizar la multicolianidad a través del método VIF y por otra parte aplicar un método Feature Importance, serían otras opciones interesantes a contemplar en una segunda fase de mejora de este proyecto.

Hemos escogido los modelos adecuados?:

Otra parte importante a analizar en este trabajo es la evaluación del o los modelos correctos para la futura predicción de los valores de gasto. En esta oportunidad hemos considerado dos modelos cuyas principales características son la flexibilidad y la facilidad de interpretación. Sin embargo es muy posible que por las características del actual dataset (mucho outliers), un Random Forest podría ser una elección adecuada ante gran catidad de valores atípicos y variables que pudieran tener alta correlación entre ellas. También, los algoritmos de Gradient Boosting podrían servir para adaptarse a datasets como el actual, dado que entre sus características se encuentra ,al igual que el Random Forest, robustass herramientas para evaluar valores atípicos.

Para cerrar, es importante aclarar que como primera experiencia en el cálculo predictivo el resultado ha dejado muchos aprendizajes, los cuales serán puestos en valor y aplicados para rectificar los resultados, bajando quizás la correlación pero entregando cifras mucho mas consistentes.