Mapa de oportunidades comerciales de Buenos Aires utilizando modelos de aprendizaje automático

PCA - Modelo de Recomendación - LightFM

Abstract: Realizamos un proyecto de machine learning, aplicando un modelo de recomendación, y utilizando como data set el mapa de oportunidades comerciales de la Capital Federal. Con esto llegamos a concretar nuestro objetivo, que es poder solucionar una necesidad de las personas. Si uno quiere abrir un local en una zona de la Capital Federal, nuestro modelo le recomienda cual es la zona más indicada.

1 Introducción

Este proyecto surgió al indagar los datos abiertos del moc de mapa de oportunidades comerciales encontrados en la página del gobierno de la Ciudad Autónoma de Buenos Aires, y el cual nos hizo plantearnos la siguiente pregunta: "Si busco abrir un negocio, ¿En qué zona de la misma podría tener mayor éxito?".

A lo largo de este trabajo explicaremos el data set utilizado, las distintas herramientas y métodos utilizados para interpretar nuestros datos, los modelos de machine learning tanto supervisados como no supervisados utilizados, y para finalizar el resultado, el modelo elegido y la conclusión.

Utilizaremos un modelo de recomendación para un inversionista basado en 18 rubros y 161 zonas en la cual dividimos la Capital Federal. Los rubros analizados son "Insumos para el hogar", "Bares y cafés", "Carnes y verduras", "Comida al paso", "ferretería y construcción", "Fiambrerías y dietéticas", "Instituciones deportivas", "Heladerías", "Kioscos y loterías", "Música y librerías", "Ópticas y joyerías", "Panaderias", "Indumentaria", "Restaurantes", "Salud y cosmética", "Tratamiento estético", "Supermercados y almacenes" y "Veterinario".

Este proyecto se realizó con 4 modelos diferentes, de los cuales solo 1 tuvo el éxito buscado, por la complejidad del data set y de la respuesta buscada. El modelo utilizado fue un modelo de recomendación basado en las variables de cada rubro por cada zona.

2 Metodología

El data set seleccionado fue "Rubros" contiene información del año 2017 de los distintos rubros comerciales distribuidos en la geolocalización de la Ciudad Autónoma de Buenos Aires. Una vez seleccionamos, seguimos con la idea de comprender nuestros datos averiguando la composición del mismo, llegando al resultado de 2898 muestras junto con 22 variables.

Sin embargo, al conocer nuestro data set, llegamos a la conclusión que para llegar a nuestro objetivo de interacción entre las zonas y los distintos rubros, era necesario realizar un preprocesamiento de la información. Nuestro primer paso fue detectar la proporción de Nans o "datos faltantes" y tomar una estrategia que nos permita definir la forma de tratarlos y no afecte significativamente nuestro data set y las variables que tendrían mayor relación con nuestro campo de aplicación.

Nuestra decisión con respecto al preprocesamiento de nuestro data set, fue procesarlo de manera diferenciada.

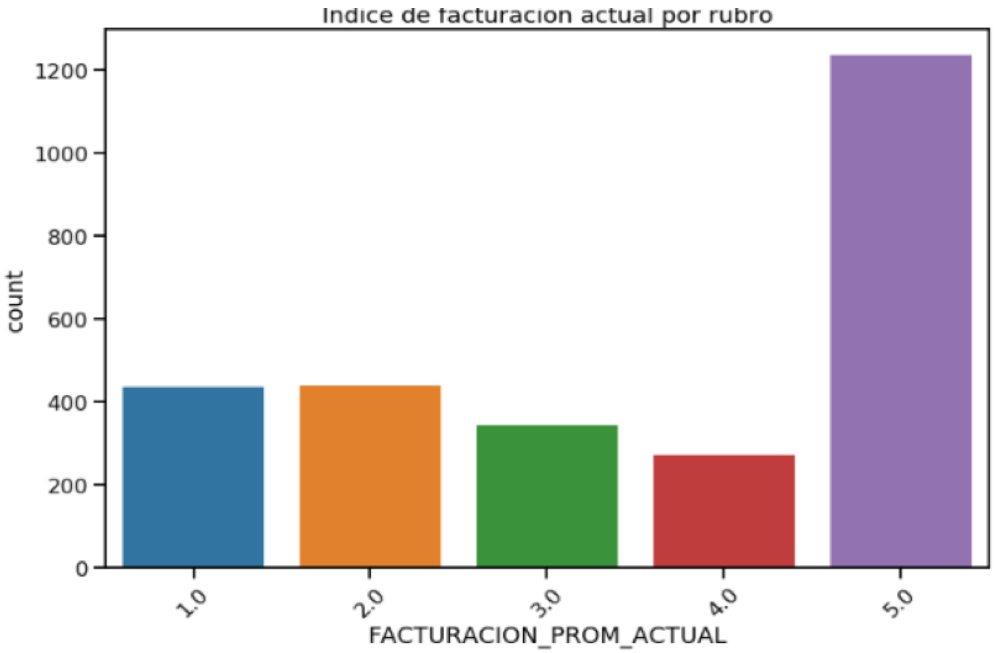
Luego de esto, realizamos un PCA ("Principal component analysis") sobre nuestro data set, ya preprocesado. La decisión de hacer un PCA [1] fue para mejorar los resultados obtenidos en nuestro modelo de recomendación. El PCA es un método que permite reducir la dimensionalidad de datos en alta dimensión, de manera tal de proyectar los datos originales en nuevas dimensiones que maximicen la variabilidad y perdiendo la menor cantidad de información original.

El modelo a implementar fue un modelo de recomendación a través de la librería conocida como "LightFM". Básicamente, un modelo de recomendación [2] tiene dos items fundamentales. Por un lado, se encuentran los elementos y por el otro lado, se encuentran los usuarios. Llevándolo a nuestro caso de estudio, los usuarios estarían representados por los distintos rubros comerciales presentes en nuestro data set [3] y por el otro, los elementos serían las diferentes formas geométricas que se encuentra distribuido la Ciudad Autonóma de Buenos Aires. De esta manera, se busca un modelo que nos permita conocer las principales preferencias de los rubros comerciales a la hora de elegir las distintas zonas. Para poder llegar a este objetivo, como punto inicial se realizó "Variables selection" eligiendo las labels que mejor describan, desde nuestro punto de vista, en que zona se puede tener más éxito, estas fueron:

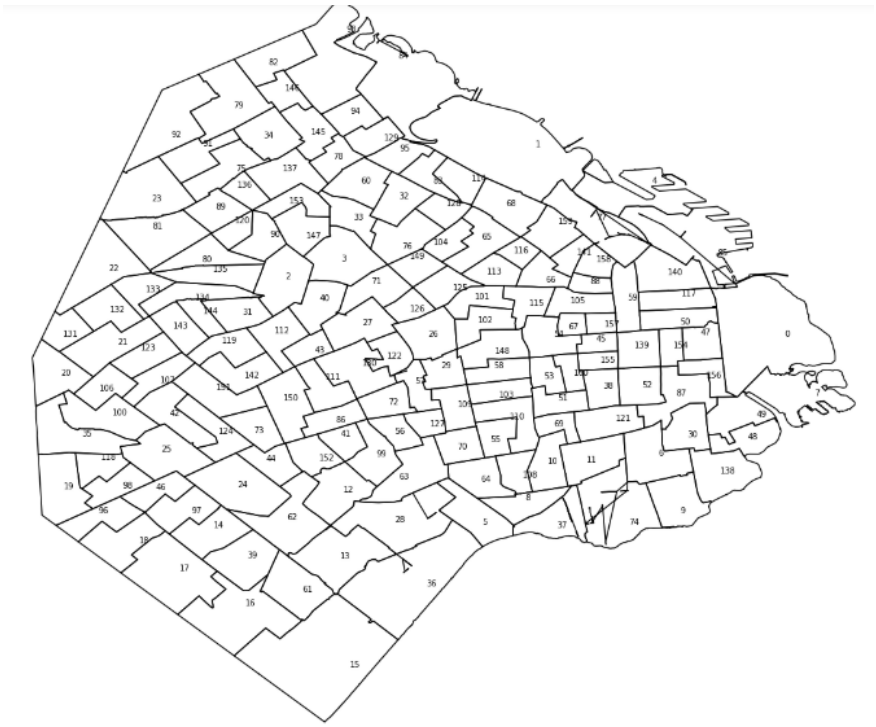
- Indice crecimiento
- Indice estabilidad
- Facturación_prom_actual
- Nivel de riesgo

A su vez, con los valores faltantes o Nans, se buscó llenar estos faltantes con 0, de manera de no reducir la cantidad de muestras de nuestro data set.

Continuando con las variables, para determinar cómo se miden estos índices, elegimos uno de ellos, en este caso "Índice de facturación actual por rubro" obteniendo la siguiente visualización:



A su vez, para conocer mejor los elementos (zonas) se decidió plotear el mapa:



Entendiendo que los modelos de recomendación utilizan la matriz interactiva para poder vincular los usuarios y sus preferencias de elementos, se busca crear un índice de recomendación. Este índice de recomendación funcionaría como un

rating de ese rubro para esa determinada zona. Para ello, creímos conveniente antes de la confección de nuestra matriz, modificar la variable "Nivel de riesgo", invirtiendo su valoración del 1 al 5, para poder conformar un data frame que denominamos "Nivel de seguridad". Con ello, multiplicamos los valores de las 4 variables, generando como resultado el "índice de recomendación"

Una vez, obtenido el mismo creamos la factorización matricial con una tabla pivote donde las filas son los distintas rubros comerciales y las columnas son las distintas zonas. Con este input del modelo de factorización matricial, el diccionario de usuario, el diccionario de elementos, el modelo va a buscar la lista ID de zonas que los rubros comerciales pueden estar interesados en interactuar.

Métrica de LightFM

Luego de crear la co-matrix (matriz pivote en el entorno de Light FM), definir los hiper parámetros, separar el modelo en train y test y entrenarlo continuaremos con la evaluación del mismo. Para esta evaluación de los resultados que nos devuelve el modelo vamos a utilizar la métrica MRR (Mean Reciprocal Rank) o Rango Recíproco Medio [4]. El rango recíproco medio se basa en la siguiente fórmula:

$$\mathrm{MRR} = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{\mathrm{rank}_i}$$

Esta métrica lo que va a permitir es medir en las distintas devoluciones del modelo, donde se va a ubicar el mejor elemento para el usuario seleccionado, es decir, de los rubros seleccionados donde se va a ubicar la zona que mejor índice de recomendación obtenemos. De esta manera, al sacar con cada uno de los usuarios su métrica correspondiente, se hace un promedio general de todas las respuestas y se obtiene la métrica correspondiente.

3 Resultados

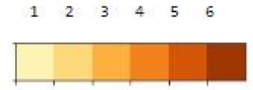
LightFM

Teniendo en cuenta que los resultados de este modelo de recomendación, va a depender en gran medida en la eficiencia de la métrica elegida, nosotros obtuvimos como resultado del proceso:

```
Train reciprocal rank: 0.97
Test reciprocal rank: 0.97
```

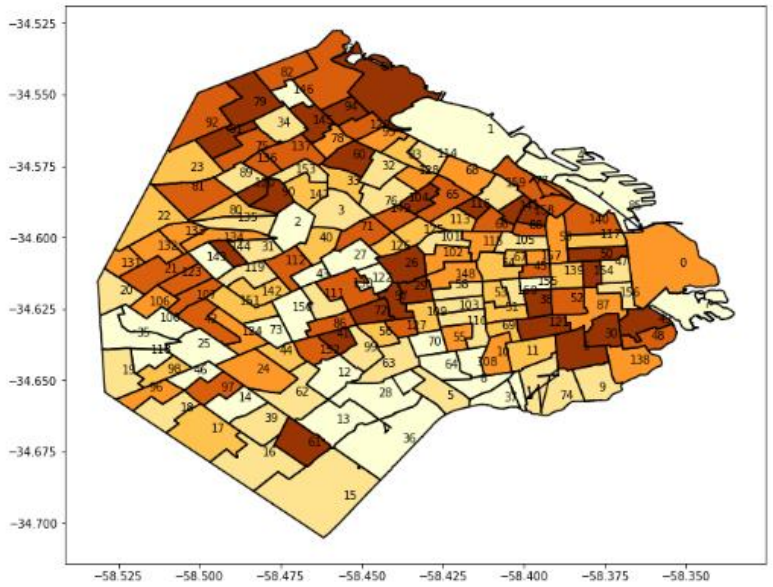
Esto quiere decir que gran parte de los resultados que devuelve el modelo de recomendación, nos sugiere una gran posibilidad de obtener el mejor resultado posible.

A su vez, para poder visualizar nuestros resultados, decidimos implementar en el ploteo inicial un heatmap para cada rubro con los resultados en las distintas zonas mostradas inicialmente. A continuación, presento un ejemplo del rubro "Bares y Café" teniendo en cuenta la delimitación presentada:



<Figure size 792x648 with 0 Axes>

HeatMap de zonas recomendadas para el rubro Bares y Cafes



1: 0 => 17
2: 18 => 47
3: 48 => 71
4: 72 => 119
5: 120 => 199
6: 200 => 499

4 Discusión y Conclusión

Como conclusión final del trabajo destacamos que con Light FM se logró obtener resultados elevados en varios indicadores por lo cual es recomendable utilizar este método para este modelo de recomendación. Adicionalmente destacamos que es muy útil el PCA, para mejorar los resultados de tu modelo final. Por otro lado, todos los rubros del dataset pueden visualizarse en el mapa de Capital Federal indicando las mejores localizaciones para cada uno de acuerdo a una escala de colores.

5 Referencias

[1] *Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. Chemometrics and intelligent laboratory systems, 2(1-3), 37-52.*

[2] *Rubtsov, V., Kamenshchikov, M., Valyaev, I., Leksin, V., & Ignatov, D. I. (2018). A hybrid two-stage recommender system for automatic playlist continuation. In Proceedings of the ACM Recommender Systems Challenge 2018 (pp. 1-4).*

[3] Fumega, S. (2014). Opening cities: open data in Buenos Aires, Montevideo and Sao Paulo; report-city of Buenos Aires, Open Government Data initiative.

[4] *Wu, Y., Mukunoki, M., Funatomi, T., Minoh, M., & Lao, S. (2011, August). Optimizing mean reciprocal rank for person re-identification. In 2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (pp. 408-413). IEEE.*