UNIVERSIDAD TECNOLÓGICA NACIONAL

INGENIERÍA INDUSTRIAL

Ciencia de Datos

TRABAJO PRÁCTICO FINAL

Mapa de oportunidades comerciales

Equipo de Trabajo:

Latorre, Tomas Ignacio
Legajo: 155.557-1
Latorretomas49@gmail.com

Libertun, Federico Lucas
Legajo: 149.929-4
fedeliber_23@hotmail.com

Caputo, Lucas
Legajo: 144.917-5
lucas.93.lc@gmail.com

Objetivo

El objetivo de nuestra investigación es crear un modelo que permita recomendarle a un inversor en qué rubro y en que zona de la Ciudad Autonóma de Buenos Aires le conviene invertir, para ello, aplicaremos distintos modelos que afirmen con mayor seguridad la decisión del mismo.

1. Introducción

Se nos pide armar un modelo de clasificación y recomendación para un inversionista basado en 18 rubros de Capital Federal. Los rubros analizados son insumos para el hogar, bares y cafes, carnes y verduras, comida al paso, salud y comestica, ferretería y construcción, fiambrerías y diéteticas, instituciones deportivas, heladerías, kioscos y loterías, música y librerías, ópticas y joyerías, panaderías, tratamiento estéticos, restaurantes, supermercados y almacenes, indumentaria, y veterinaria.

2. Data Set

Para el análisis de los mismos, elegimos el Data Set del gobierno de la ciudad de mapa de oportunidades comerciales[1], el cual tiene 5 Data Sets adentro:

- **Apertura:** Este Data Set contiene información de los datos cuatrimestrales de las aperturas por rubro en cada zona. Tiene un total de 5210 samples por 5 features, las cuales son:
  - Rubro: Tipo de rubro del local
  - Moc_zonas_id: ID de la zona
  - Año: Año del dato (2016 o 2017)
  - Cuatrimestre: Cuatrimestre del dato (1, 2 o 3)
  - Nivel: Cantidad de aperturas representados en niveles del 1 al 5.

- **Cierre:** Este Data Set contiene información de los datos cuatrimestrales de los cierres por rubro en cada zona. Tiene un total de 3114 samples por 5 features, las cuales son:
  - Rubro: Tipo de rubro del local
  - Moc_zonas_id: ID de la zona
  - Año: Año del dato (2016 o 2017)
  - Cuatrimestre: Cuatrimestre del dato (1, 2 o 3)
  - Nivel: Cantidad de cierres representados en niveles del 1 al 5

- **Demografía:** Contiene los datos socioeconómicos por zona. Tiene un total de 2254 samples por 8 features, las cuales son:

[1] https://data.buenosaires.gob.ar/dataset/mapa-oportunidades-comerciales-moc

- Moc_demografia_id: ID univoco de los registros de MOC demografía
- Moc_zonas_id: ID univoco de los registros de MOC zonas.
- Pk_rango_etario_id: ID del rango etario
- Pk_genero_id: ID del genero
- Poblacion_viviente: Cantidad de personas que viven en dicha zona
- Poblacion_trabajadora: Cantidad de personas que trabajan en dicha zona
- Rango_etario: Descripción del rango etario
- Género: Genero de la persona

- **Zonas:** Este Data Set contiene la información de las características de la zona. Tiene un total de 161 samples por 25 features, las cuales son:
  - Moc_zonas_id: ID univoco de los registros de MOC zonas.
  - Pk_tiempo_id: ID periodo
  - Poblacion_flotante: Población que pasa por la zona
  - Poblacion_viviente: Cantidad de personas que viven por zona
  - Poblacion_trabajadora: Cantidad de personas que trabajan en esa zona
  - Cantidad_hogares: Cantidad de hogares que hay por zona
  - Precio_promedio_alquiler_local: Promedio de alquiler en dicha zona
  - Precio_max_alquiler_local: Precio máximo de alquiler por zona por local
  - Precio_min_alquiler_local: Precio mínimo de alquiler por zona por local
  - Supercie_m2_promedio_alquiler: Precio promedio de m2 en alquiler
  - Supercie_m2_max_alquiler: Precio máximo de m2 en alquiler
  - Supercie_m2_min_alquiler: Precio mínimo de m2 en alquiler
  - Rubro_predominante: Rubro predominante en dicha zona
  - Facturacion_prom_rubro_predominante: Facturación promedio del rubro predominante en la zona
  - Facturacion_prom_rubro_menos_predominante: Facturación promedio del rubro menos predominante en la zona
  - Precio_promedio_venta_local: Precio promedio de la venta de local
  - Precio_max_venta_local: Precio máximo de la venta de local
  - Precio_min_venta_local: Precio mínimo de la venta de local
  - Superficie_m2_promedio_venta: Precio promedio del Metro cuadrado en venta
  - Superficie_m2_maximo_venta: Precio máximo del Metro cuadrado en venta
  - Superficie_m2_min_venta: Precio mínimo del Metro cuadrado en venta
  - Nivel_locales_rubro_predominante: Cantidad de locales del rubro predominante en la zona en niveles

- Nivel_locales_rubro_menos_predominante: Cantidad de locales del rubro menos predominante en la zona en niveles
- Fecha: Fecha de actualización.

- **Rubros:** Este Data Set contiene características de la zona. Tiene un total de 2898 samples y 22 features, las cuales son:
  - Moc_rubros_id: ID univoco de los registros MOC Rubros
  - Moc_zonas_id: ID univoco de relación con la tabla MOC zonas
  - Rubro: tipo de rubro del local
  - Nivel_riesgo: nivel de riesgo del 1 al 5
  - Facturacion_prom_actual: Facturación promedio año en curso del 1 al 5
  - Indice_crecimiento: nivel del crecimiento de la zona y el rubro del 1 al 5
  - Indice_estabilidad: nivel de estabilidad de la zona y el rubro del 1 al 5
  - Indice_apertura: nivel de apertura de la zona y el rubro del 1 al 5
  - Indice_cierre: nivel de cierre de la zona y el rubro del 1 al 5
  - Ind_ap_act_vs_ind_ap_anio_ant: Variación del índice de apertura entre el año pasado y el actual
  - Ind_cl_ act_vs_ind_ap_anio_ant: Variación del índice de cierre entre el año pasado y el actual
  - Sup_menos_1: % de locales que tienen menos de 1 año de duración
  - Sup_entre_2_y_3: % de locales que tienen entre 2 y 3 años de duración
  - Sup_entre_3_y_4: % de locales que tienen entre 3 y 4 años de duración
  - Sup_entre_4_y_5: % de locales que tienen entre 4 y 5 años de duración
  - Sup_mas_5: % de locales que tienen más de 5 año de duración
  - Facturacion_prom_anio_ant: Facturación promedio del año anterior en niveles del 1 al 5
  - Nivel_locales: Performance del local en dicha zona.
  - Indice_cierre_anio_ant: Indice de cierre del rubro y de la zona con respecto a la totalidad del rubro y de la zona
  - Indice_apertura_anio_ant: Indice de apertura del rubro y de la zona con respecto a la totalidad del rubro y de la zona.

Para el EDA (Análisis exploratorio de datos) que significa como está compuesto nuestro dataset, y como vamos a modificarlo para poder identificar y analizar los datos, se utilizó un Data Set llamado **Rubros**, la cual nos pareció la que tiene la información más interesante y útil para el propósito buscado por nuestro TP

3. Exploratory Data Analysis → EDA

Nosotros separamos el EDA en 2 distintos, dependiendo el método que aplicamos.

Modelo de clasificación y clustering

El primer EDA fue para evaluar cómo se desempeñaban estos métodos en nuestro Data set. Nos dimos cuenta que, en el Data Set de Rubros, había muchos datos NaN y adicional a eso, había columnas que no nos interesaban, por lo cual procedimos a eliminar las columnas que no nos aportaban información relevante, a eliminar los samples que no contenían las features relevantes para nosotros y a completar el restante con la media de la columna.

	Total	Percent
INDICE_CIERRE_ANIO_ANT	1510	0.52
IND_CL_ACT_VS_IND_CL_ANIO_ANT	1502	0.52
INDICE_APERTURA_ANIO_ANT	832	0.29
IND_AP_ACT_VS_IND_AP_ANIO_ANT	831	0.29
INDICE_SUPERVIVENCIA	360	0.12
INDICE_CRECIMIENTO	186	0.06
FACTURACION_PROM_ACTUAL	170	0.06
NIVEL_LOCALES	155	0.05
FACTURACION_PROM_ANIO_ANT	154	0.05
INDICE_ESTABILIDAD	149	0.05
NIVEL_RIESGO	149	0.05
SUP_MAS_5	113	0.04
SUP_ENTRE_3_Y_4	113	0.04
SUP_ENTRE_2_Y_3	113	0.04
SUP_ENTRE_1_Y_2	113	0.04
SUP_MENOS_1	113	0.04

Descartamos todas las features relacionadas al año anterior.
Descartamos los registros que tenían NaN en facturación promedio actual

	Total	Percent
INDICE_SUPERVIVENCIA	216	0.08
INDICE_CRECIMIENTO	16	0.01
NIVEL_LOCALES	5	0.00
NIVEL_RIESGO	1	0.00
INDICE_ESTABILIDAD	1	0.00

Completamos los restantes con la media de cada feature

Modelo de recomendación:

Realizamos otro EDA para el método de recomendación:

En este modelo, en lugar de eliminar las NaNs las reemplazamos por cero ya que nos interesa conservar todas las filas.
Si en las features Nivel_riesgo, Facturacion_prom_actual, Indice_crecimiento o Indice_estabilidad aparece 0 para una sample, eso representa que falta información.

Después de eso queremos comparar esas 4 columnas. Para eso armamos una nueva llamada Nivel_de_seguridad a partir de la feature NIVEL_RIESGO, transformando sus valores al revés, es decir si tenía 5 ahora tengo 1; si mi dato era 4, ahora vale 2, y así. Si tenía 0 sigue valiendo 0.

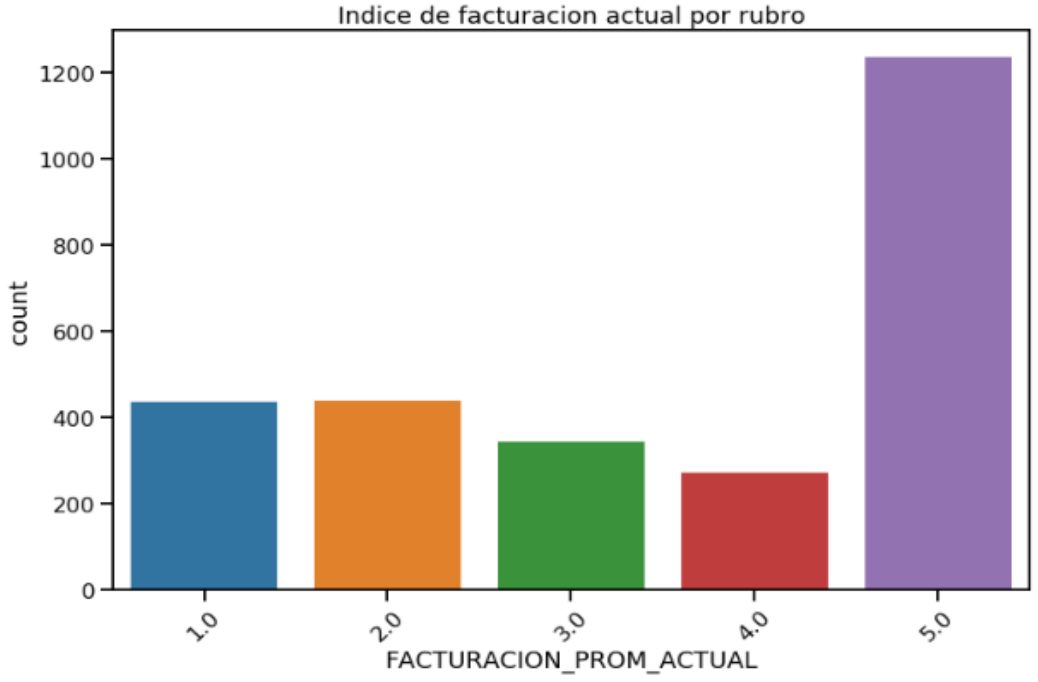
_ID	RUBRO	NIVEL_RIESGO
1	INSUMOS PARA EL HOGAR	5.0
2	INSUMOS PARA EL HOGAR	0.0
3	INSUMOS PARA EL HOGAR	1.0
4	INSUMOS PARA EL HOGAR	1.0
5	INSUMOS PARA EL HOGAR	0.0

Nivel_de_seguridad
1.0
0.0
5.0
5.0
0.0

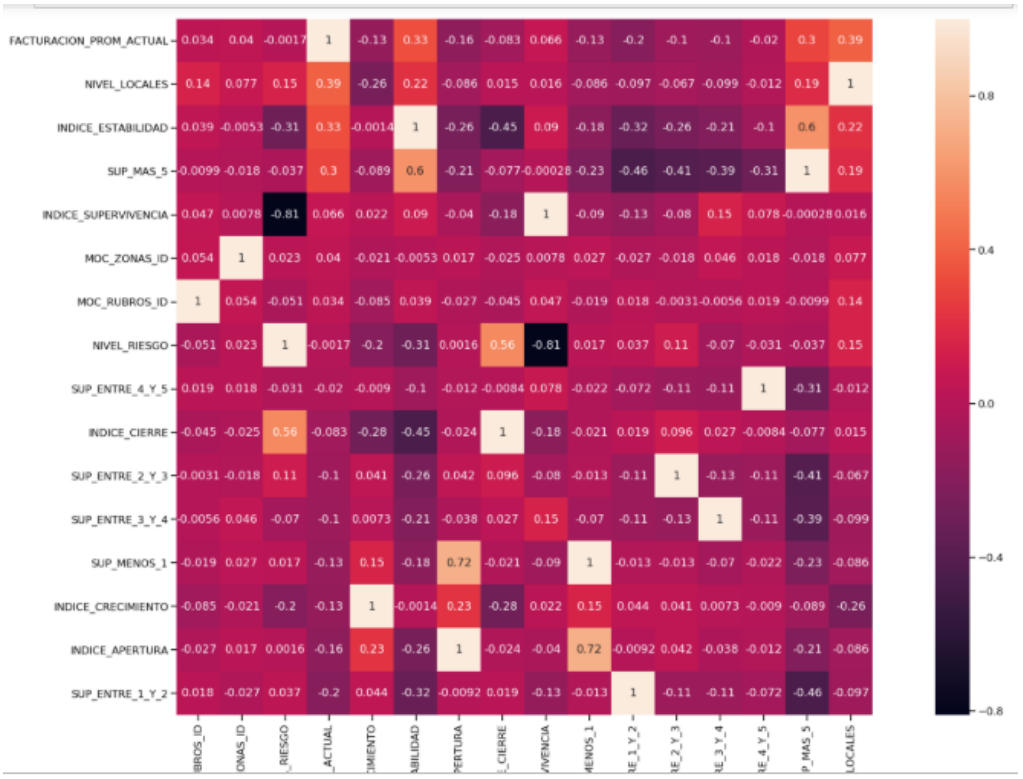
4. Análisis Gráficos

Para visualizar mejor los datos y su dispersión se realizaron una serie de gráficos de los cuales se pudieron extraer varias conclusiones e ideas para tener una mejor comprensión del Data Set.

Se realizó un gráfico para ver la distribución que había entre los índices de facturación actual por rubro. Y nos dimos cuenta que en la mayoría de rubros hay un nivel de facturación muy alto. Este dato es relevante, el único problema es que no tenemos un dato muy preciso de que significa a nivel monetario ese índice 5 y ver la dispersión interna que hay entre ese mismo índice.



Se realizó otro grafico para ver la correlación entre las features, intentando ver alguna correlación inesperada. De este grafo pudimos determinar que hay correlación entre las variables esperadas. Indice de apertura y el de supervivencia de menos de un año (ya que fueron las que abrieron ese último año). Correlación negativa entre Indice de supervivencia y de riesgo (ya que mientras más riesgo menos pueden sobrevivir). Indice de cierre con el índice de riesgo (ya que mientras más riesgo más cierran). Indice de estabilidad con el indicador de supervivencia de más de 5 años. Entre otros. Esto lo realizamos con un heatmap, para ver de una forma más representativa esta correlación.



5. Feature Engineering

Modelo de clasificación y clustering

5.1 Dummies Creation

Primero se dividió la creación de Dummies, dependiendo el modelo a aplicar. Para los modelos de clasificación se generó Dummies para la feature categórica Rubros, ya que se unió posteriormente al Data Set de Rubros, para poder aumentar el accuracy de nuestro clasificador.
Para los modelos de clustering se realizó la creación de Dummies para la feature categórica Zonas, ya que adicionamos esta última con la creada de Rubros al Data Set y aumentamos el silhouette de nuestros modelos.

5.2 Partición X & Y

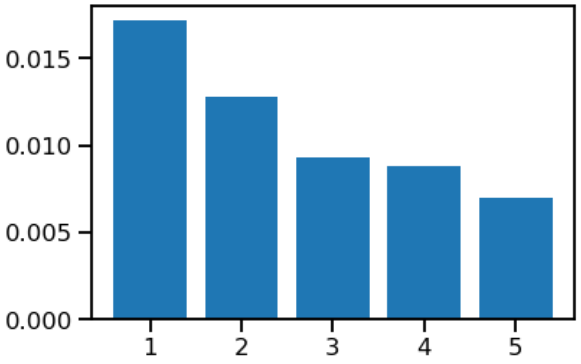
Se separó el Data Set primero (utilizado para clasificar) en un 40% para entrenamiento de modelos y el restante 60% para la evaluación de los mismos.

En el caso del segundo Data Set (utilizado para clustering) no se realizó Split entre test y train ya que no es necesario.

Las features fueron estandarizadas (media = 0 y desviación estándar = 1) para ambos Data Sets.

6. Dimensionality Reduction Analysis

Con la expectativa de poder analizar esta investigación mediante modelos no supervisados, se optó por realizar una reducción dimensional.
Infortunadamente el porcentaje de variación del Data Set que capta dicha reducción es muy poco y no sirve para un análisis confiable.



Modelo de Recomendación:

En el modelo de recomendación, en cambio, creamos una nueva feature llamada Índice_de_Recomendación que será el resultado de multiplicar las features: Facturacion_prom_actual, Indice_crecimiento, Indice_estabilidad y Nivel_de_seguridad. Con esta nueva feature lo que buscamos es tener en cuenta todos los índices en mi modelo de recomendación y no solo uno, para tener una recomendación más certera.

Índice_de_Recomendación
3.0
0.0
160.0
160.0
0.0

7. Modelos de Clasificación

A continuación, se presentarán los modelos que fueron utilizados para performar la clasificación de rubros por zona.

Para medir la performance de cada modelo se optó por utilizar las métricas de **Accuracy.** Íbamos a evaluar también el AUC, o área debajo de la curva, pero en este caso al ser multiclase el modelo, no era posible calcularlo.

- ## Logistic Regression

Este modelo busca la probabilidad de que un sample X corresponda a un label Y. Como resultado se posee la probabilidad de cada label para cada sample (la suma de las probabilidades es igual a 1) y se selecciona la respuesta con mayor probabilidad.

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

Los resultados de este modelo fueron los siguientes:

- **Accuracy Train**: 0.00%
- **Accuracy Test**: 0.00%

- ## K Nearest Neighbors

El modelo es entrenado con un set de features y labels. Cada nuevo sample que ingrese al modelo será clasificado al label cuyas features compartan la mayor relación (cercanía). Uno especifica la cantidad de puntos que se desean tomar en cuenta al clasificar.

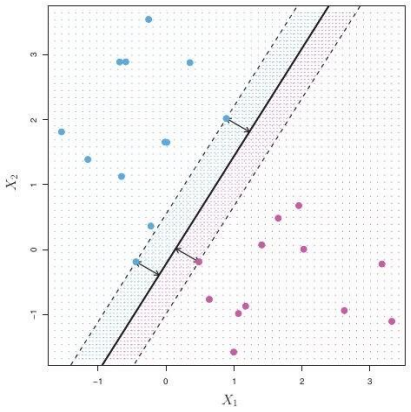
$$\Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j)$$

Los resultados de este modelo fueron los siguiente:

- **Accuracy Train**: 1.83%
- **Accuracy Test**: 0.18%

- Support Vector Classification

Dado un set de samples con sus features y labels correspondientes, se busca el hiperplano que separe los diferentes labels posibles en N dimensiones. Alrededor de este hiperplano se encuentran los vectores de soporte que se crean una región de separación entre diferentes labels.

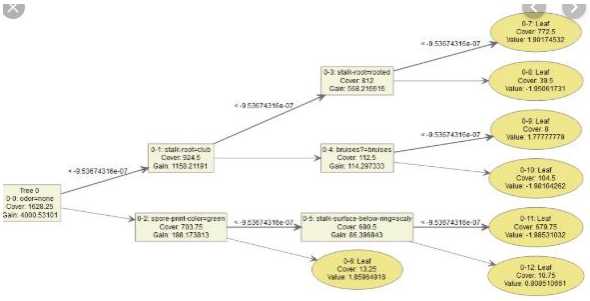


Los resultados de este modelo fueron los siguiente:

- **Accuracy Train**: 99.08%
- **Accuracy Test**: 0.18%

- Xgboost

Es un método es una implementación avanzada de un algoritmo de aumento de gradiente con un modelo de árbol como modelo base. Los algoritmos de aumento conocen de forma iterativa los clasificadores débiles y, a continuación, los añaden a un clasificador fuerte final.



Los resultados de este modelo fueron los siguiente:

- **Accuracy Test**: 1.10%

8. Clustering

A continuación, se presentarán los modelos que fueron utilizados para performar la clusterizacion del Data Set Rubros, con el objetivo de ver si hay algún clustering oculto dentro de este Data Set.
Para medir la performance de cada modelo se optó por utilizar las métricas de **Silhouette score.**

- K-means

Es un método que busca la distancia euclidiana cuadrática entre muestras como medida de similaridad

$$d(x_i, x_{i'}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = ||x_i - x_{i'}||^2$$

Luego utiliza como función objetivo la de minimizar esta distancia entre cada muestra y cada centroide

$$\min_{C, \{m_k\}_1^K} \sum_{k=1}^{K} N_k \sum_{C(i)=k} ||x_i - m_k||^2$$

Y por último asigna las muestras al centroide más cercano (Teniendo en cuenta que hay un centroide por clúster).

$$C(i) = \underset{1 \le k \le K}{\operatorname{argmin}} ||x_i - m_k||^2$$

Utilizamos un For para poder buscar cual es el número de clúster óptimo para nuestro modelo y el resultado fue el siguiente:

- **Clusters sin PCA:** 9
- **Silhouette:** 1.90%
- **Clusters con PCA**: 2
- **Silhouette:** 24.77%

- Hierarchical clustering

Es un método que construye representaciones jerárquicas en donde los clusters de cada nivel de jerarquía son creados agrupando los clusters del nivel inmediatamente inferior. En el nivel más bajo posible, cada clúster contiene una sola muestra.
Como en el K-means utilizamos un For para poder buscar el numero óptimo de clusters y el resultado fue el siguiente:

- **Clusters sin PCA:** 5
- **Silhouette:** 0.35%
- **Clusters con PCA:** 2
- **Silhouette:** 19.00%

9. Recommendation

A continuación, se presentarán los modelos que fueron utilizados para performar la clasificación de rubros por zona.

- LIGHT FM

Un **modelo híbrido de recomendación** de representación latente. El modelo aprende incrustaciones (representaciones latentes en un espacio de alta dimensión) para usuarios y elementos de una manera que codifica las preferencias del usuario sobre los elementos. Cuando se multiplican juntas, estas representaciones producen puntajes para cada elemento para un usuario dado; Es más probable que los elementos con una puntuación alta sean interesantes para el usuario.[2]

Este Módulo que contiene funciones de evaluación adecuadas para juzgar el rendimiento de un modelo LightFM ajustado[3]. Para esta evaluación se eligieron las siguientes herramientas:

- **PRECISIÓN K – MÉTRICA:** la fracción de positivos conocidos en las primeras k posiciones de la lista clasificada de resultados. Se obtiene la matriz Numpy que contiene puntajes de precisión k para cada usuario. Si no hay interacciones para un usuario determinado, la precisión devuelta será 0. Con los siguientes resultados:

```
Train precision: 0.97
Test precision: 0.92
```

- **EVALUACIÓN RECALL:** Mida el recuerdo en la métrica k para un modelo: el número de elementos positivos en las primeras k posiciones de la lista clasificada de resultados dividido por el número de elementos positivos en el período de prueba. Una puntuación perfecta es 1.0.

```
Train recall: 0.03
Test recall: 0.12
```

- **AUC ROC:** Mide la métrica ROC AUC para un modelo: la probabilidad de que un ejemplo positivo elegido al azar tenga una puntuación más alta que un ejemplo negativo elegido al azar. Una puntuación perfecta es 1.0.

```
Train AUC score: 0.68
Test AUC score: 0.86
```

- **RANKING RECÍPROCO:** Mida la métrica de rango recíproco para un modelo: 1 / el rango del ejemplo positivo mejor clasificado. Una puntuación perfecta es 1.0.

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

En este modelo de recomendación, esta todavía en fase de desarrollo la predicción que permita indicar en cada zona que rubro es el más acorde para invertir (presentación a llevarse a cabo durante la exposición), en reemplazo a este faltante, consideramos oportuno realizar un “heatmap” que introduciremos a continuación y explicaremos las conclusiones en la fase final de este documento:

[2] https://lyst.github.io/lightfm/docs/lightfm.html

[3] https://lyst.github.io/lightfm/docs/lightfm.evaluation.html

Para ello, armamos una matriz pivote con los datos de Índice_de_Recomendación. En las filas se encuentran los 18 rubros y en las columnas las 161 zonas.

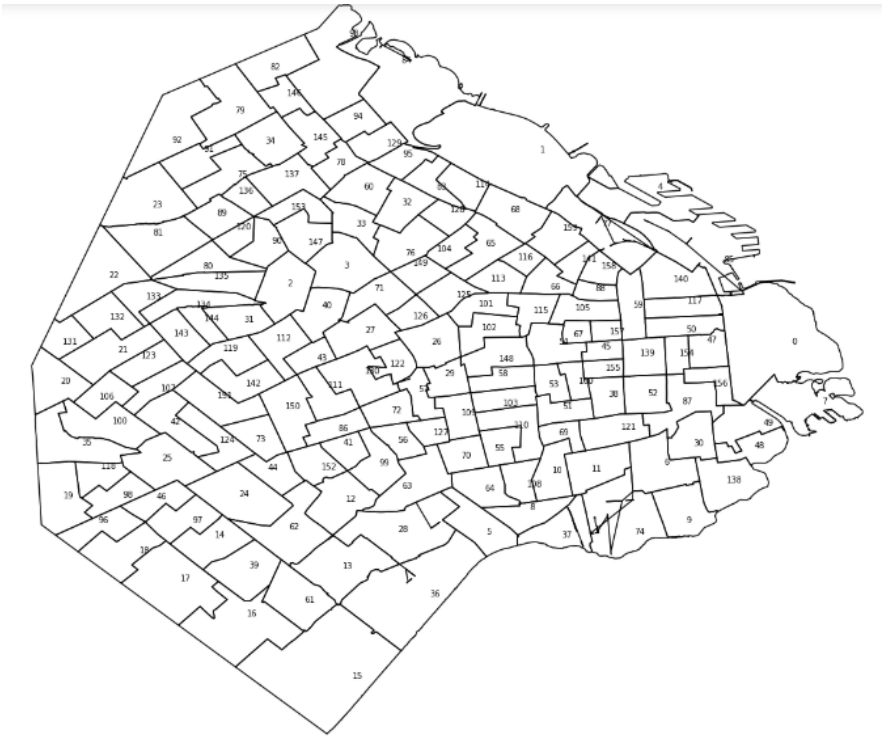
MOC_ZONAS_ID RUBRO	1	2	3	4	5	6	7	8	9	10	...	152	153	154	155	156	157
BARES Y CAFES	0.240	0.00	0.000	0.090	0.0	0.120	0.600	0.00	0.006	0.080	...	0.144	0.320	0.096	0.480	0.004	0.060
CARNES Y VERDURAS	0.400	0.00	0.192	0.400	0.0	0.048	0.500	0.00	0.240	0.600	...	0.320	0.020	0.192	0.120	0.180	0.400
COMIDA AL PASO	0.600	0.40	0.160	0.800	0.0	0.040	0.400	0.00	0.480	0.144	...	0.240	0.032	0.480	0.016	0.080	0.750
FERRETERIA Y CONSTRUCCION	0.800	0.16	0.800	0.384	0.0	0.800	0.800	0.00	0.800	0.320	...	0.480	0.192	0.640	0.800	0.216	0.640
FIAMBRERIAS Y DIETETICAS	0.400	0.00	0.018	0.600	0.0	0.200	0.160	0.00	0.036	0.500	...	0.100	0.360	0.800	0.320	0.160	0.128
HELADERIAS	0.080	0.00	0.100	0.000	0.0	0.000	0.000	0.00	0.160	0.000	...	0.800	0.000	0.160	0.080	0.480	0.080
INDUMENTARIA	0.320	0.00	0.160	0.120	1.0	0.240	0.800	0.08	0.384	0.400	...	0.120	0.288	0.240	0.400	0.160	0.144
INSTITUCIONES DEPORTIVAS	0.080	1.00	0.120	0.060	1.0	0.000	0.080	0.00	0.800	0.800	...	0.160	0.040	0.002	0.480	1.000	0.240

Queremos llevar en un mapa los resultados, de tal manera que para cada rubro nos indique en qué zonas conviene ubicarlo.

Abrimos el archivo donde se encuentra almacenado el mapa de Buenos Aires con sus zonas.

	zone_id	link_zones	coords
154	155	001_15	[(-58.381510313660996, -34.61766127676578), (-...
66	67	002_25, 002_24, 002_21, 002_20, 002_19	[(-58.41602732363647, -34.59785442604267), (-5...
98	99	009_5	[(-58.519676471845564, -34.64779841811188), (-...
92	93	012_1	[(-58.491774485664024, -34.545615917341614), (...
29	30	006_18, 006_9, 006_19, 006_20	[(-58.43002862764531, -34.61544921414433), (-5...

Visualizamos las zonas:

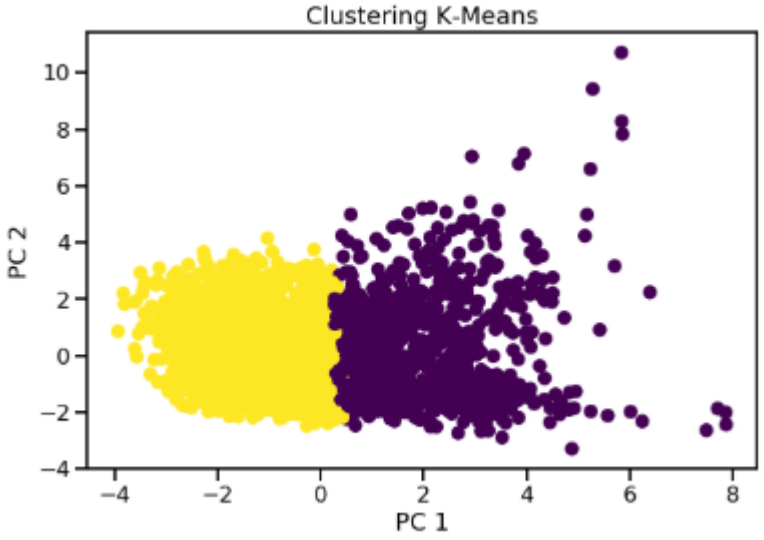


Y realizamos un “heatmap” para recomendar en qué zonas conviene ubicar cada rubro.

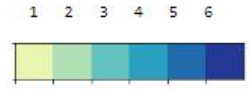
10. Conclusiones

Utilizando el método de clasificación llegamos a conclusión de que ningún método puede clasificar con un accuracy elevado, lo único que conseguimos fue un buen accuracy con un SVM, pero en el train y cuando lo probamos con el test, nos dimos cuenta que estaba muy overfiteado y no era generalizado.

Utilizando el método de Clustering, nos dimos cuenta que funcionaba mejor con un PCA (por más que no eliminamos ningún feature) que con el modelo original. Llegando a un 24.77% de Silhouette score, con un numero de clusters igual a 2. Esto no es un modelo muy preciso, pero se puede notar la división de los 2 clusters, aunque comparte muchos puntos de contacto

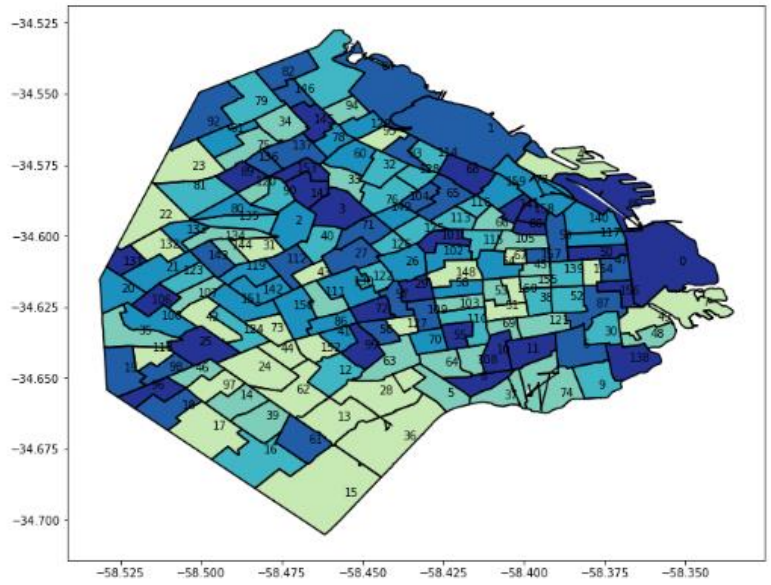


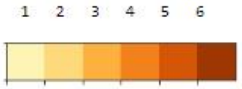
Utilizando método de recomendación podemos ver en un heatmap de Capital Federal en qué zona se recomienda ubicar cada rubro. Mostramos a continuación 3 ejemplos:



<Figure size 792x648 with 0 Axes>

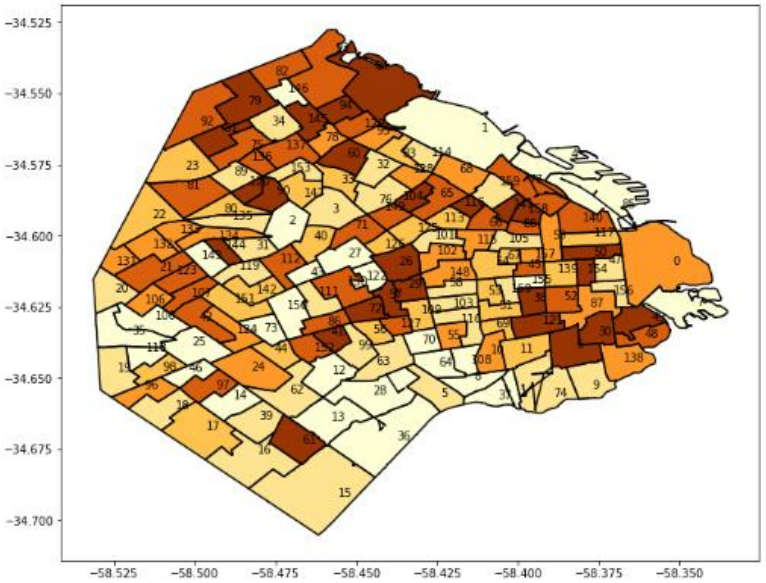
HeatMap de zonas recomendadas para el rubro Comida al paso

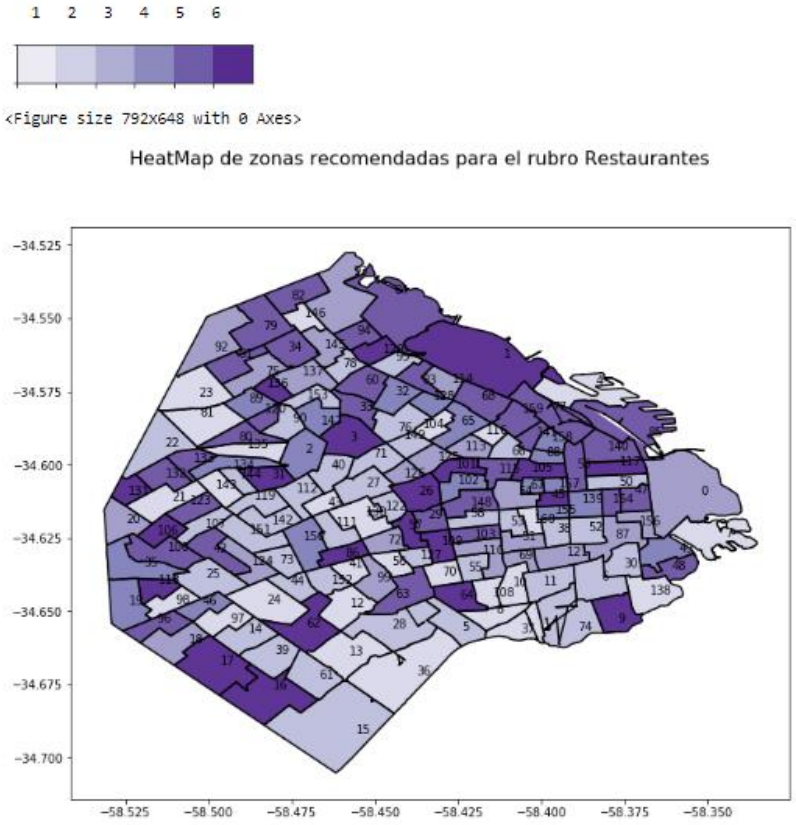




<Figure size 792x648 with 0 Axes>

HeatMap de zonas recomendadas para el rubro Bares y Cafes





Este método es el más adecuado para nuestro modelo y nos permite responder la pregunta más importante que nos hacemos con este Data Set que es, en cual zona invertir dependiendo el rubro. Además de esto, es el método que mejores resultados nos da. En un futuro nos gustaría adicionarle otro modelo de recomendación, para poder tener una comparación entre los modelos y ver cuales nos da un mejor rendimiento y un mapa que ilumine las 3 principales zonas recomendadas para el rubro elegido.