Desafio Final

Grupo 6

Barbara Malamud
Juan Molinari
Marcos Garcia Verdier
Leonardo Campos

Agenda

1. Presentar objetivo del desafío final
2. Bases de datos utilizadas
3. Análisis exploratorio de la información
4. Algoritmos de ML aplicados
5. Análisis de features
6. Test
7. Conclusiones

1. Objetivo del desafio final

OBJETIVO:

Desarrollar un clasificador para entender qué estaciones de servicio ubicadas sobre rutas y autopistas nacionales podrían ser adquiridas por su alta rentabilidad.

CONOCIMIENTO DE NEGOCIO

- Para que una estación de servicio ubicada sobre una ruta o autopista sea altamente rentable debe vender al menos 500 m3 / mes de combustible(1 m3 = 1000 lts) .

2. Bases de datos utilizadas

Las bases de datos que utilizamos fueron:

- Resolución 1104 de la secretaria de energia.
  - Reportes de ventas mensual histórico de todas las estaciones del país.
  - Fue utilizado para dar las etiquetas al modelo
    - 1 si la estacion vende mas de 500 m3 / mes
    - 0 si vende menos de 500 m3 / mes
  - Localidad, razon social, bandera, direccion y ventas mensuales
- CENSO 2010 INDEC
  - Población por provincia
- Tránsito medio diario anual (TMDA) de Vialidad Nacional
  - TMDA = tránsito pasante medio que pasa diariamente por año por cierto punto
  - Tránsitos anuales de todas las rutas nacionales del país del 2017
  - Longitud de tramos, transito de tramos
- Parque automotor histórico por provincia
  - Delta del parque circulante por provincia histórico
- Resolución 314 de la secretaria de energia
  - Reportes de precios mensuales histórico de todas las estaciones del país.
- Superficie total

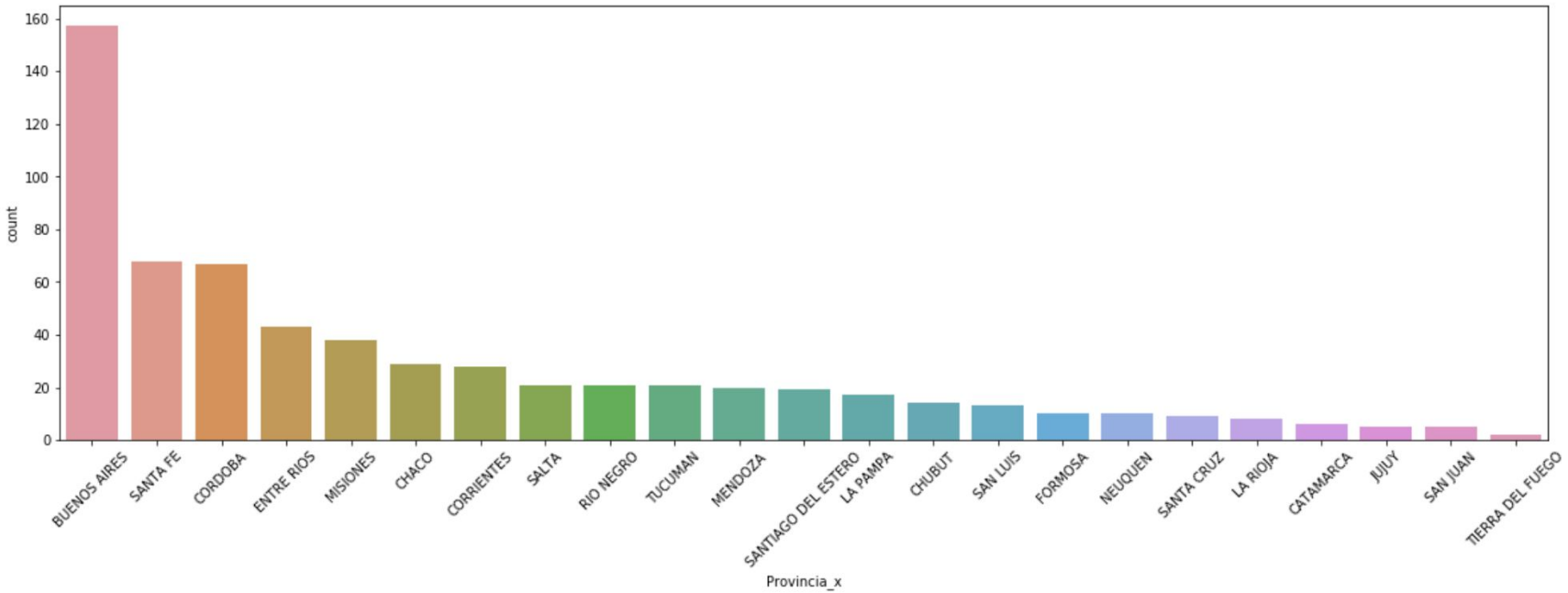
2. Bases de datos utilizadas

Dataset Final

- Shape:
  - 620 estaciones
  - 400 features
    - Dummies de bandera
    - Dummies de provincia
    - Dummies de rutas
    - Dummies de ubicación geoespacial del tramo asociado al transito
    - KPIS
      - Cantidad de estaciones por tramo
      - Distancia de los tramos
    - Superficie de provincias
    - Poblacion por provincia
    - Parque automotor
    - Ultimo precio vigente del 2019 de la nafta super de cada estacion
    - Hectareas sembradas de oleaginosas
    - Etc

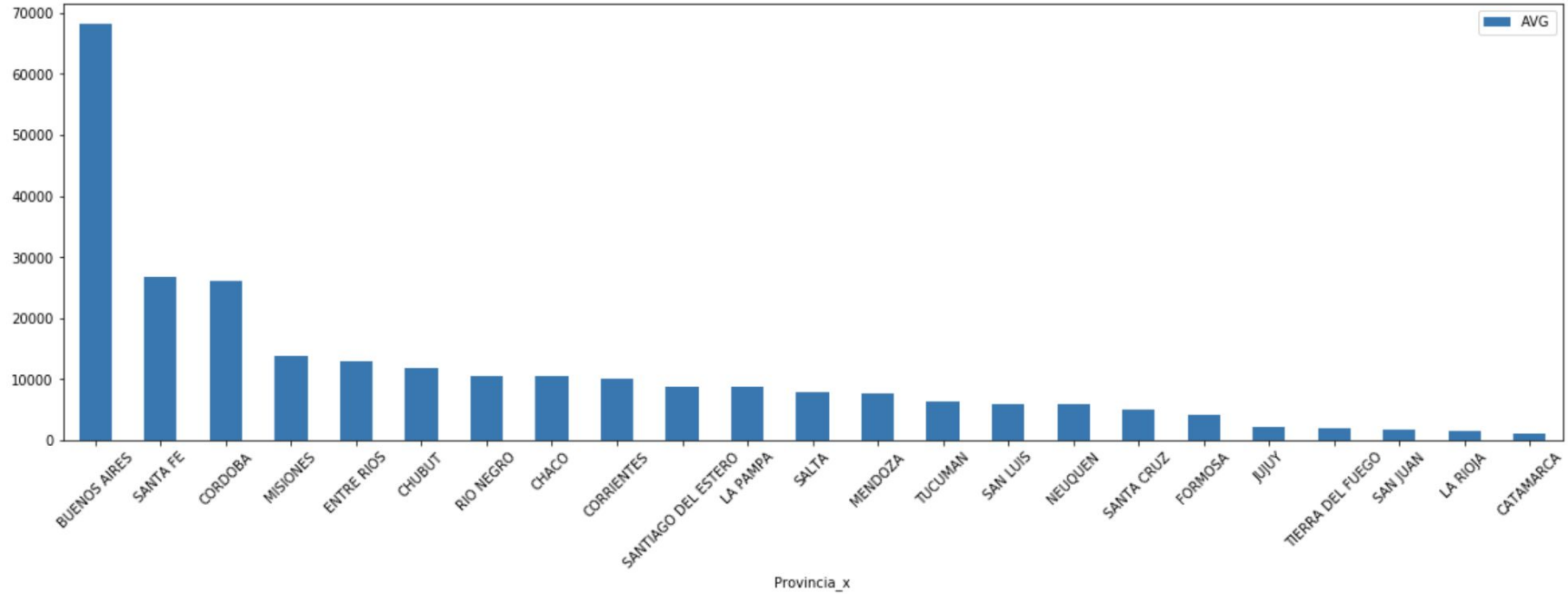
3. Analisis exploratorio de la informacion

Cantidad de estaciones por provincia:



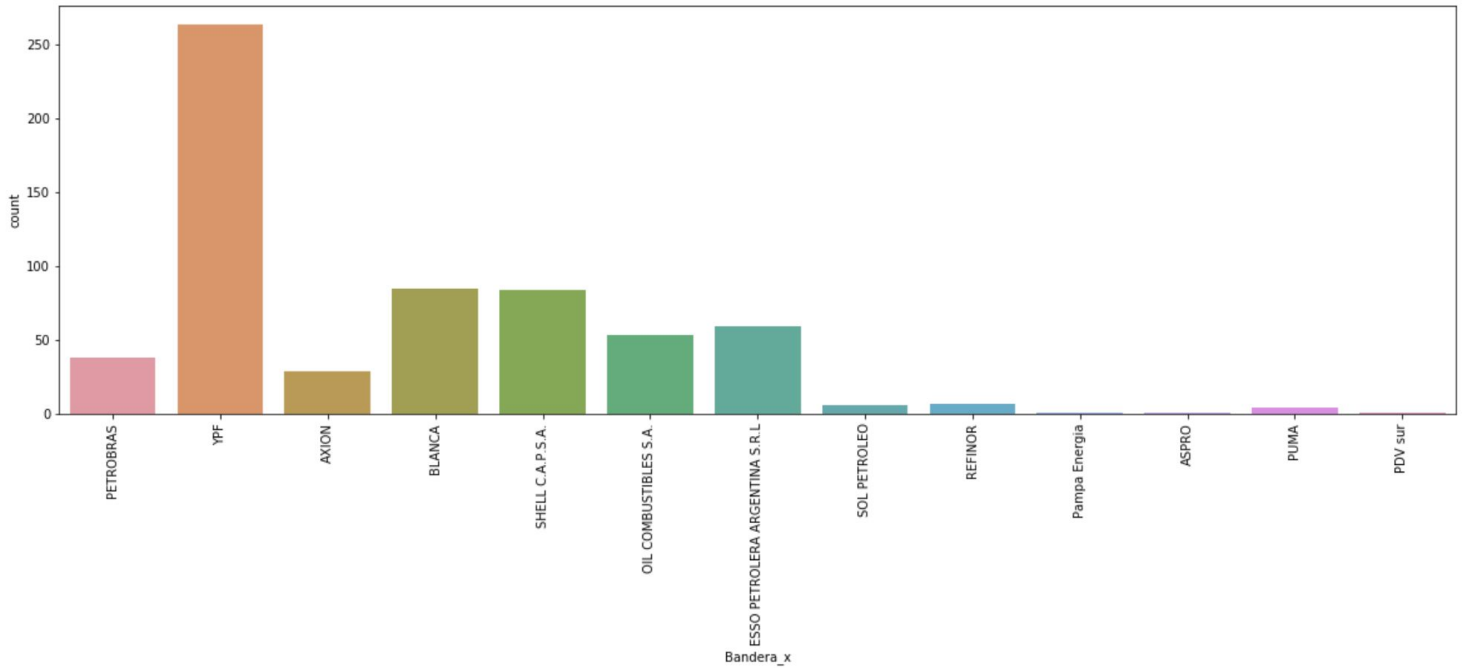
3. Analisis exploratorio de la informacion

Volumen de venta mensual por provincia:



3. Analisis exploratorio de la informacion

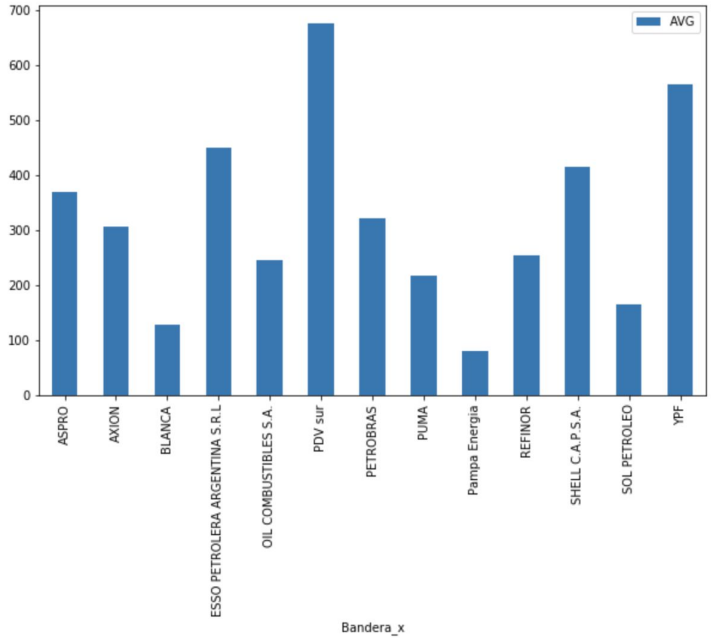
Cantidad de estaciones por bandera:



*sobre un total de 620 estaciones sobre rutas y autopistas nacionales

3. Analisis exploratorio de la informacion

Venta promedio por bandera:



3. Analisis exploratorio de la informacion

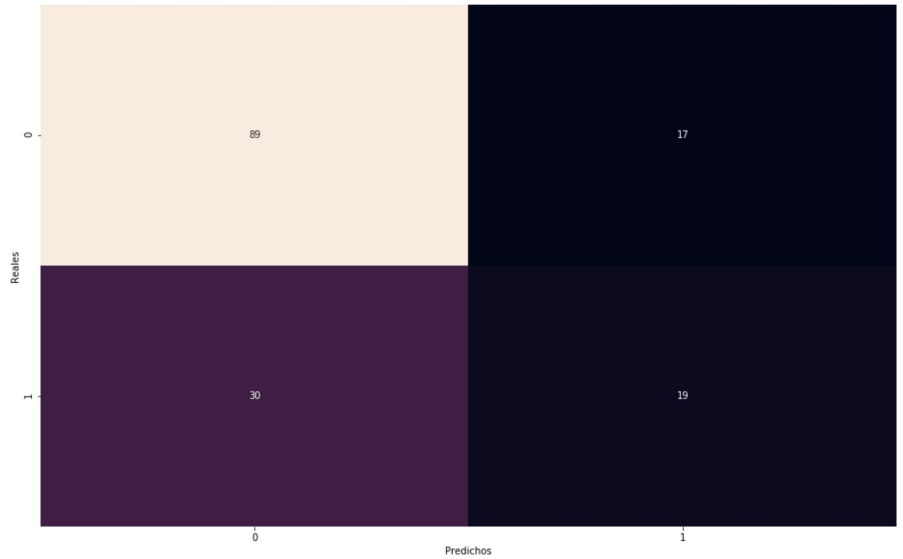
Geolocalización estaciones vs Tránsito

4. Algoritmos de ML aplicados:

KNN con Gridsearch

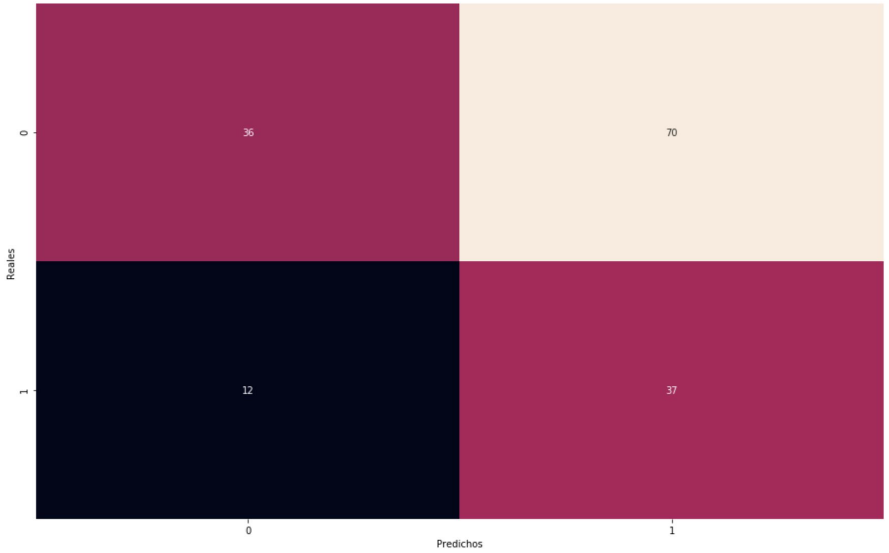
- Accuracy ⟶ 0.67

	precision	recall	f1-score	support
0	0.75	0.84	0.79	106
1	0.53	0.39	0.45	49
accuracy			0.70	155
macro avg	0.64	0.61	0.62	155
weighted avg	0.68	0.70	0.68	155



4. Algoritmos de ML aplicados:

	precision	recall	f1-score	support
0	0.75	0.34	0.47	106
1	0.35	0.76	0.47	49
accuracy			0.47	155
macro avg	0.55	0.55	0.47	155
weighted avg	0.62	0.47	0.47	155



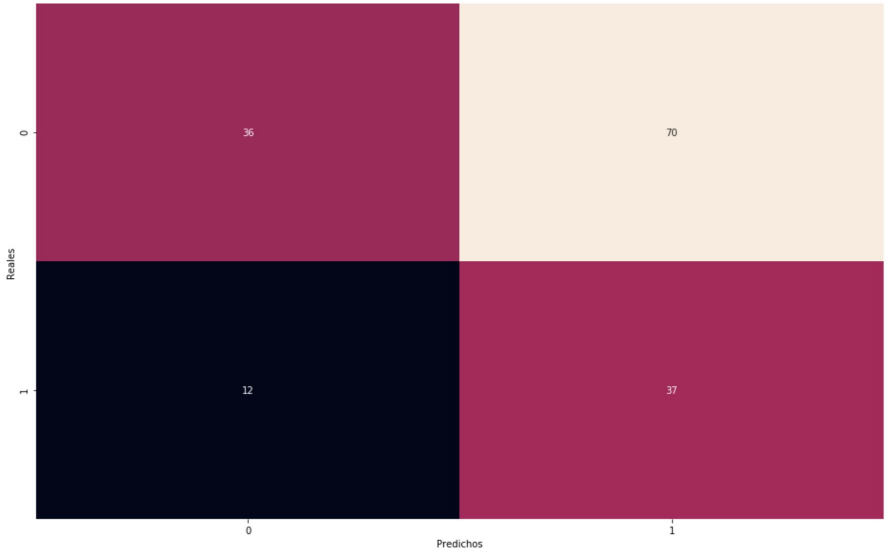
*sobre un total de 620 estaciones sobre rutas y autopistas nacionales

4. Algoritmos de ML aplicados:

Naive Bayes

- Accuracy ⟶ 0.47

	precision	recall	f1-score	support
0	0.75	0.34	0.47	106
1	0.35	0.76	0.47	49
accuracy			0.47	155
macro avg	0.55	0.55	0.47	155
weighted avg	0.62	0.47	0.47	155



4. Algoritmos de ML aplicados:

```
Rendimiento de Árbol de decisión:       0.65 ± 0.052
Rendimiento de AdaBoostClassifier:      0.647 ± 0.031
Rendimiento de GradientBoostingClassifier:      0.7 ± 0.015
```

Boosting c / GridSearch

```
Rendimiento de AdaBoostClassifier + GS: 0.726 ± 0.086
Rendimiento de GradientBoostingClassifier + GS: 0.712 ± 0.008
```

ACA PEGATE EL INDICE DE GINI
ENTROPIA Y ESAS COSAS

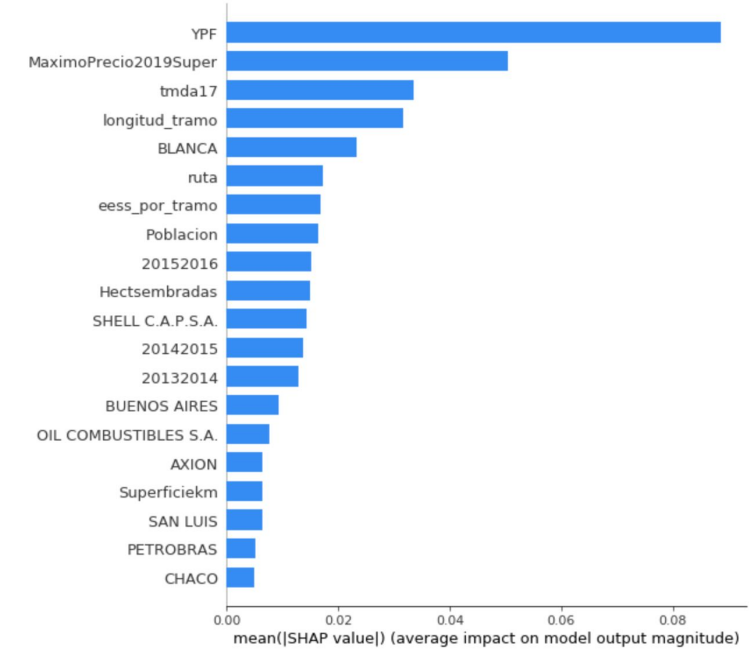
*sobre un total de 620 estaciones rutas y autopistas nacionales

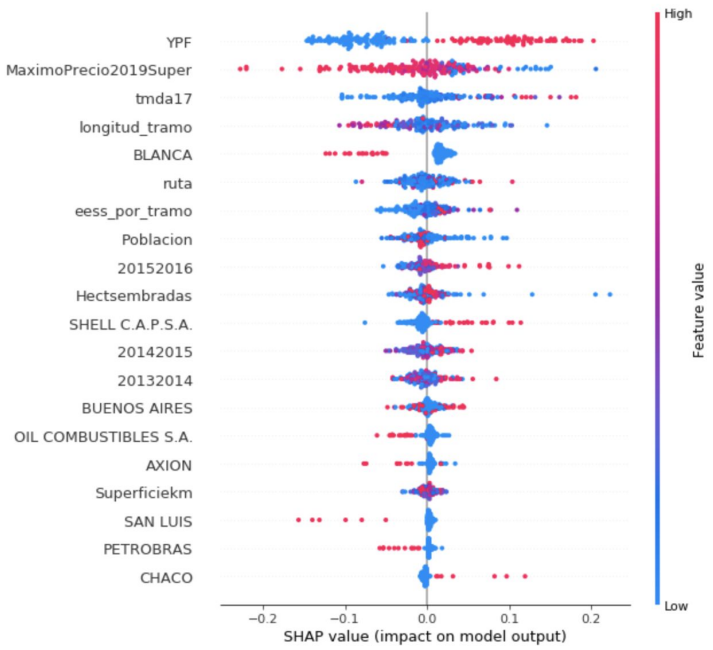
4. Algoritmos de ML aplicados:

ALGORITMO	ACCURACY
KNN	0,67
Random Forest	0,67
Naive Bayes	0,47
Arbol de decision	0,65
Gradient Boosting sin GS	0,7
Ada Boosting sin GS	0,647
Gradient Boosting con GS	0,712
Ada Boosting con GS	0,726
XG Boost	0,699
Cat Boost	
Light LGBM	0,704

5. Análisis de features

Utilizamos la libreria de Shap para analizar que features eran determinantes de nuestro modelo con Ada Boost ya que es el que mejor accuracy dio:





6. Test

- Testeamos 10 estaciones que nos dejamos del data set original para chequear el accuracy de forma real
- Los resultados fueron bastante coherentes con el accuracy.. Sabiamos que 2 de las 10 estaciones que le dimos en el modelo eran mayores a 500 y solo fallo en una de esas dos.

Conclusiones

- La mas obvia es que aprendimos mucho con modelos por categoría
- Nos tomo mucho tiempo componer el dataset en especial en el uso de geolocalización como elemento de merge
- Folium tiene dificultades técnicas con Google, se recomienda utilizar Firefox para notebooks que trabajen con imágenes grandes
- Entendemos que como datos exploratorios sirve pero como modelo de toma de decisiones necesitamos mas variables especificas que se le puede sumar