Analisis de calidad de servicio eléctrico

Mentoria: Diplomatura Ciencias de Datos, FaMAF, 2020

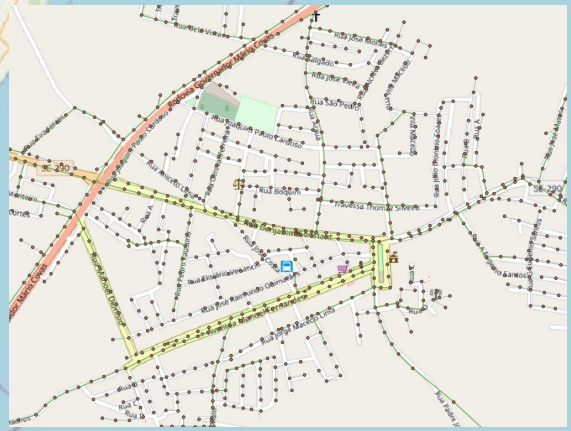
Mentor - Ramiro Caro

Participantes:

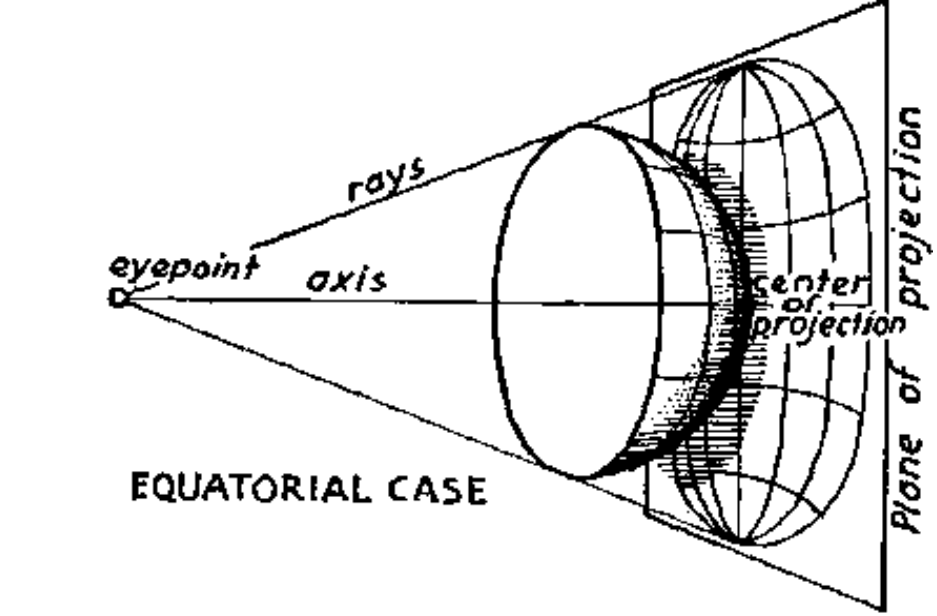
Javier Gallo

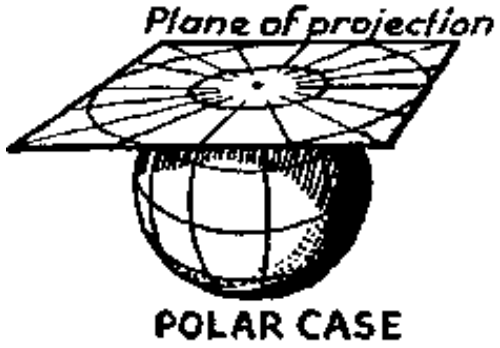
Brandon Janes

Guillermo Adolfo Coseani



https://github.com/georeference-ML-mentoria





ML sobre Información Georeferenciada

Predicción de frecuencia de cortes eléctricos

Desafíos

- Intentar predecir la calidad de servicio prestada a los consumidores basándonos en datos de la infraestructura eléctrica con información georeferenciada y las características de los consumidores.

- Crear nuevas features a partir de los datos georeferenciados que nos permitan mejorar las predicciones

- Seleccionar de muchos modelos de Aprendizaje Automático aquel que presente mejor performance

Descripción del dataset

- Contiene información sobre infraestructura eléctrica; características y consumo eléctrico de los consumidores de baja tensión

- Cada nivel de la infraestructura eléctrica poseía un dataset diferente por lo que fue necesario unirlos

- Dos variables objetivo posibles:

  DIC: Valores anuales de duración (en horas) de las interrupciones por unidad

  FIC: Frecuencia de las interrupciones individuales de la unidad

Descripción del dataset

- El dataset proporciona información sobre:

- Información georeferenciada de cada tipo de infraestructura eléctrica (consumidores, transformadores, líneas de transmisión). Esto nos permitió trabajar con la creación de nuevas features de distancia.

- Información del consumidor: consumo, actividad económica, si pertenece a un área urbana o rural, tipo de construcción, etcétera

GeoPandas

- En un paquete que presenta características y funciones especiales, útiles en GIS (geodataframes).

- Usa Pandas como plataforma y se provee de otros paquetes como Matplotlib, Shapely y Fiona

Variables georeferenciadas

Se trabajó con tres tipos de datos georeferenciados estos son puntos, líneas y polígonos.

Entre los puntos encontramos:

- Las unidades transformadoras de subestación (UNTRS).
- Las unidades transformadoras de distribución (UNTRD).
- Unidades consumidoras de baja tensión (UCBT).

Variables georeferenciadas

Entre las líneas encontramos:

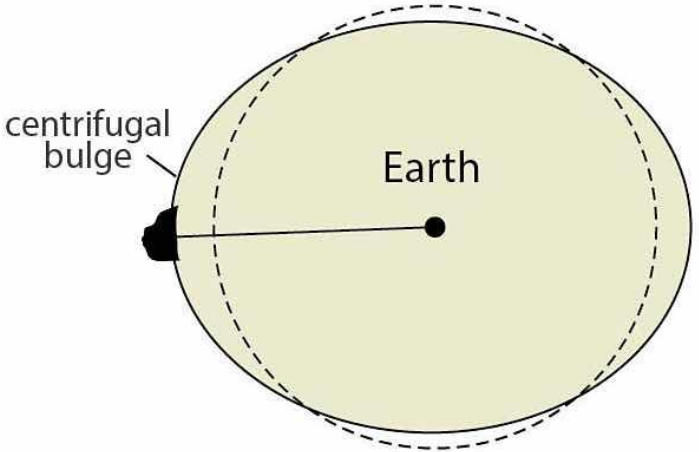
- Las líneas de transmisión de media tensión (SSMDT)

Finalmente se trabajó con un polígono (CONJ) que agrupa en un área un conjunto de unidades de consumo.

Con estos distintos tipos de datos georeferenciados se trabajó en la creación de nuevas features que pudieran servir en la aplicación de modelos de Aprendizaje Automático

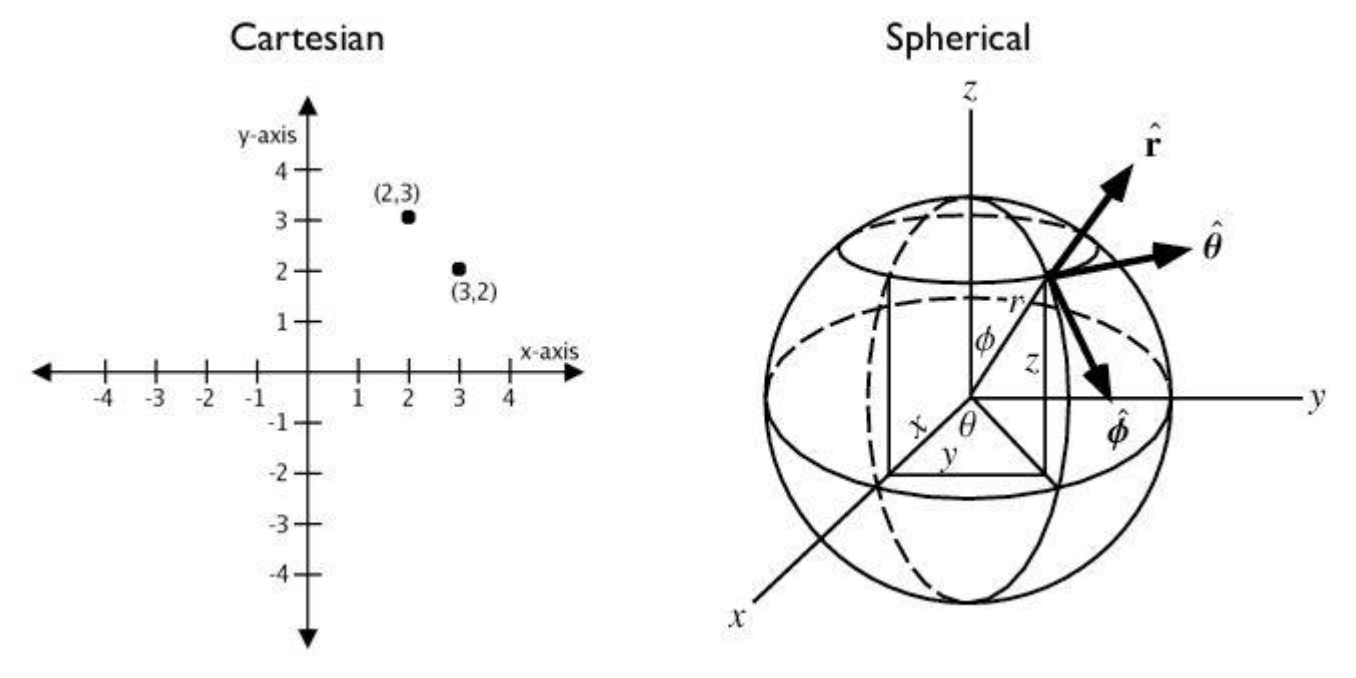
Mount Chimborazo

The summit is over 6,800 feet [2,072 meters] farther from Earth's center than Mount Everest's summit.



Proyección

Distancia Geométrica vs. Geográfica



GIS con Python

```
import geopandas as gpd
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
import networkx as nx
```

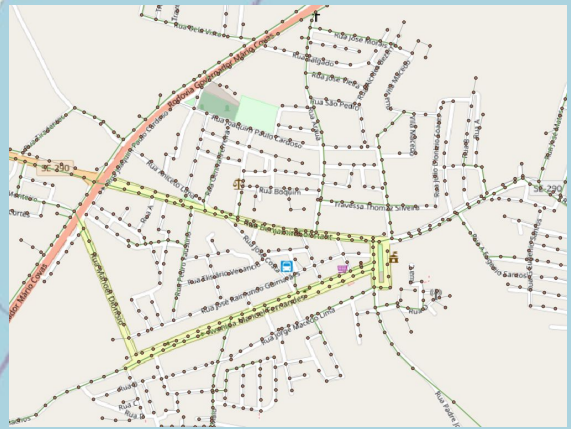
```
path = nx.shortest_path(SG,untrd.NODE[10], untrd.ENDNODE[10])
ponnot.set_index('COD_ID', inplace=True)
path_ponnot = ponnot.loc[path]
ax = path_ponnot.to_crs(epsg=3857).plot(figsize=(8,8))
ax.set_title('Grafo entre transformador 10 y la subestacion')
ctx.add_basemap(ax, source= ctx.sources.OSM_A)
```


```
from shapely.geometry import LineString, Polygon
```

```
l = LineString([(1,0),(-1,-4),(-4,-0.5)])
l
```

```
c = Polygon([(0,0),(0,1),(-1,-1),(2,-1),(2,1)])
c
```

QGIS

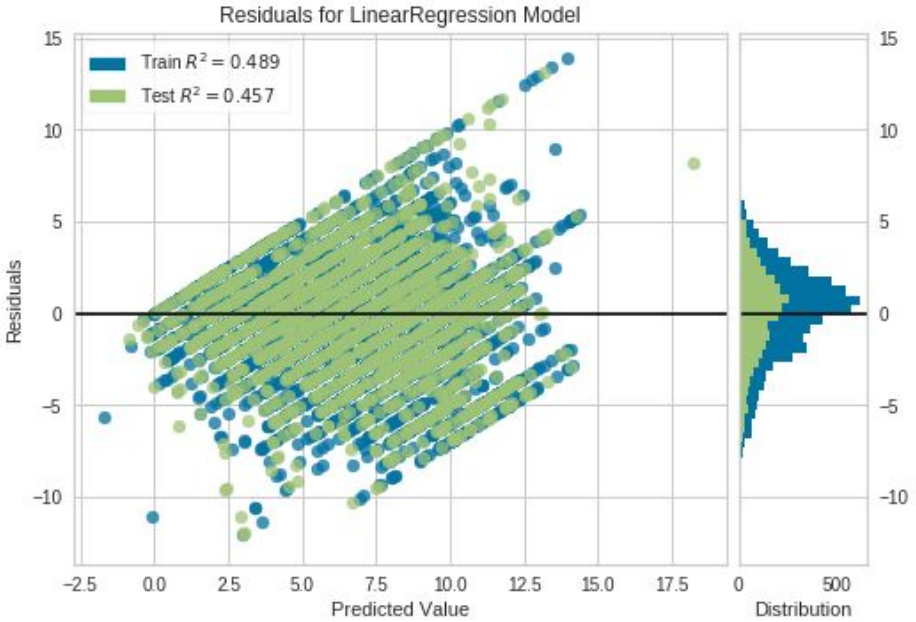


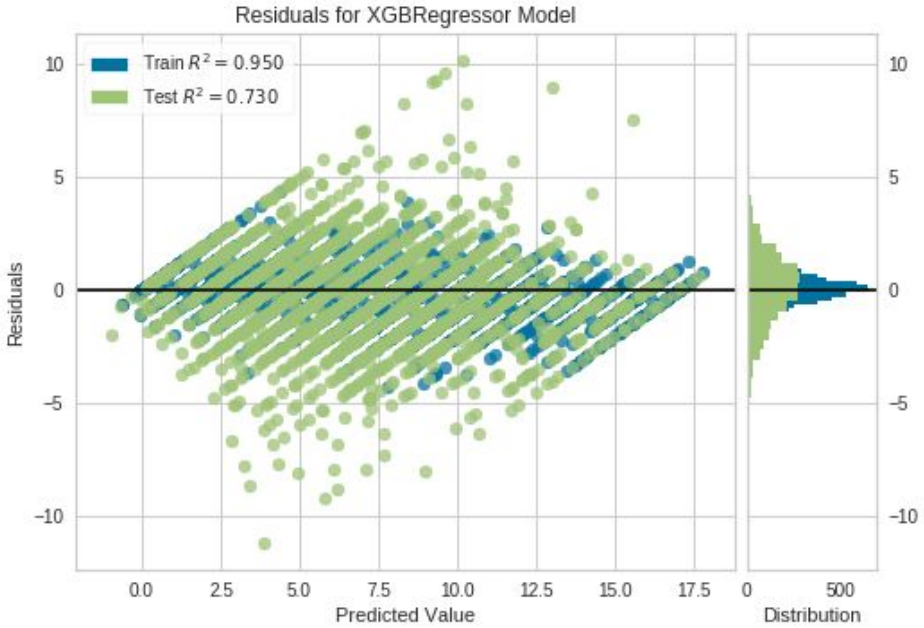
Aprendizaje Automático Supervisado

- Problema: predecir frecuencia de cortes
- Transformación del conjunto de datos
  - Queremos variables numéricas exclusivamente.
  - Quitamos, recodificamos y agregamos variables (codificación one-hot).

- División entre conjuntos de entrenamiento y de validación
- Pruebas con varios tipos de modelos **sin selección de variables** (96 características)
  - De Scikit Learn: **LinearRegression** (con datos escalados, sin regularización), **Lasso** y **Ridge** (lineales con regularización). Resultados más o menos decentes, similares, Lasso un poco peor.
  - De XGBoost: **XGBRegressor** (con parámetros por defecto, con y sin datos escalados -> resultados similares). Resultados mucho mejores!

Aprendizaje Automático Supervisado

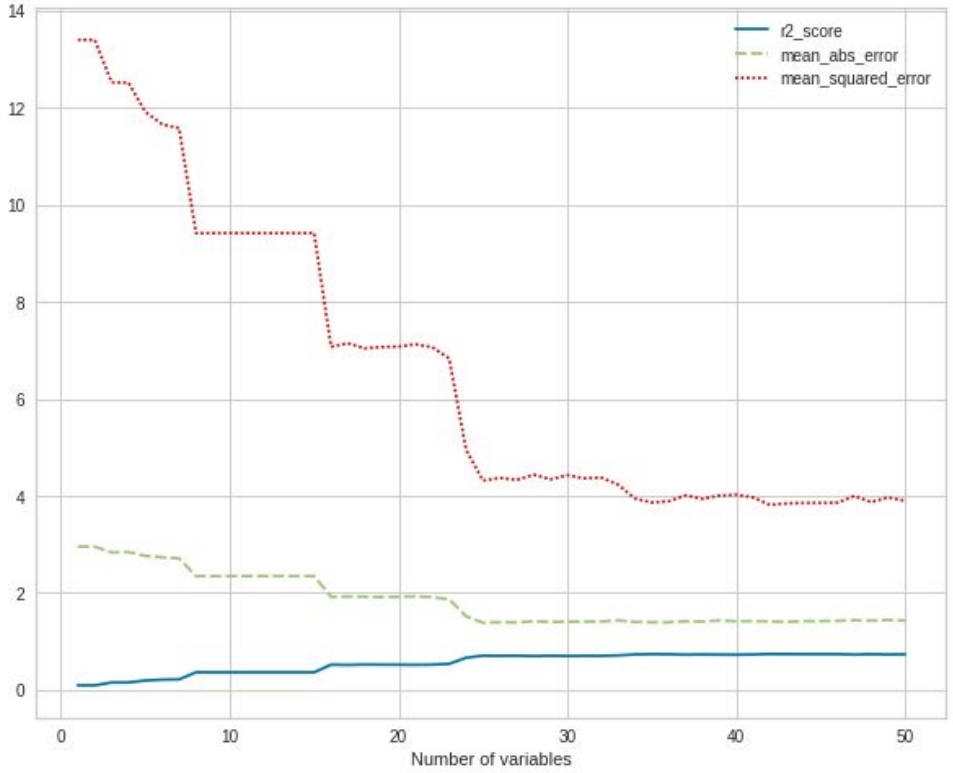




Aprendizaje Automático Supervisado

Más pruebas XGBRegressor, **con selección de variables** (35 características más importantes según XGBRegressor)

- **Sin ajuste de hiper-parámetros**: resultados aún mejores (error algo más bajo).
- **Con ajuste de hiper-parámetros**:
  - Scikit Learn: **RandomizedSearchCV** y **GridSearchCV** (mismos espacios -> resultados similares). Resultados significativamente mejores, pero Grid Search CV **súper** lento.
  - Hyperopt (optimización bayesiana): resultados similares.

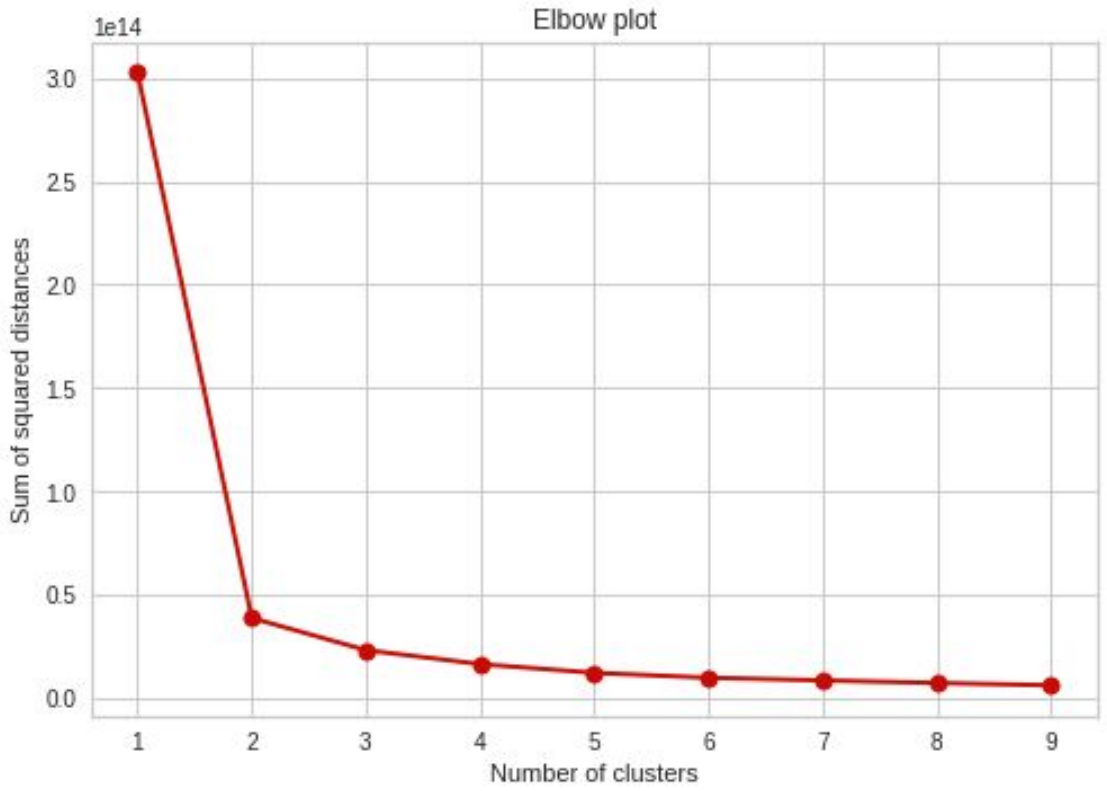


Aprendizaje Automático No Supervisado

Clusterización k-means para generación de features.

Se agrega columna nueva a conjuntos de entrenamiento y validación y se vuelve a entrenar XGBRegressor con optimización bayesiana.

El número óptimo de clusters parece ser 2, sin embargo obtuvimos resultados mejores con k=3.



Aprendizaje automático - comparación pruebas con XGBRegressor

¿Con selección de variables?	Fuente de hiper-parámetros	R^2	Error absoluto medio	Error cuadrado medio
No	Valores por defecto	0.729571	1.473470	4.014440
Sí	Valores por defecto	0.739747	1.397778	3.863390
Sí	Randomized Search CV	0.745935	1.354804	3.771527
Sí	Grid Search CV	0.753541	1.339098	3.658618
Sí	Optimización bayesiana	0.747511	1.336126	3.748129
Sí, con variable de cluster	Optimización bayesiana	0.752439	1.334594	3.674972

Conclusiones

- XGBoost es una herramienta interesante para este tipo de problemas
- La selección de variables fue de gran ayuda
- La búsqueda de hiperparámetros mejoró un poquito las métricas, GridSearchCV no recomendable.

Analisis de calidad de servicio eléctrico

Mentoria: Diplomatura Ciencias de Datos, FaMAF, 2020

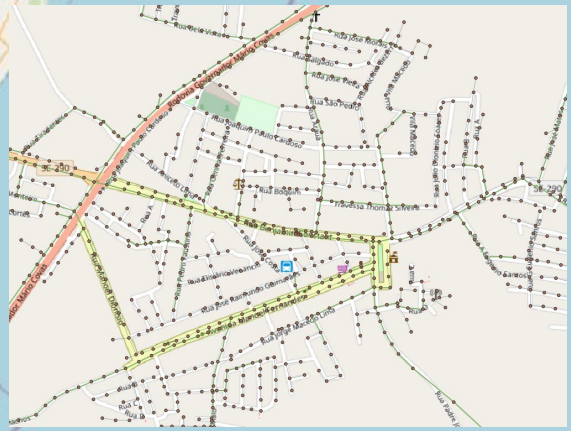
Mentor - Ramiro Caro

Participantes:

Javier Gallo

Brandon Janes

Guillermo Adolfo Coseani



https://github.com/georeference-ML-mentoria