Predicción de solicitud de subsidio “Tarjeta AlimentAR”

Ramiro Paz

Universidad Tecnológica Nacional –
Facultad Regional Buenos Aires

***Resumen* — El objetivo de este proyecto fue aplicar modelos de aprendizaje supervisado para la predicción de altas en el plan nacional AlimentAR. En particular, se utilizaron los modelos de regresión lineal y Support Vector Regression (SVR), para entender cual ajustaba mejor a los datos muestreados.**

Palabras Clave — AlimentAR, Machine Learning, Regresión

I. INTRODUCCIÓN

La Tarjeta Alimentar es un instrumento del Plan Argentina contra el Hambre, una política integral que impulsa la Nación en articulación con las provincias, y los municipios. El programa está orientado a garantizar a las familias el acceso a la canasta básica alimentaria. Permite comprar todo tipo de alimentos, a excepción de bebidas alcohólicas.

Está dirigido a madres y padres con hijos e hijas de hasta 6 años que reciben la Asignación Universal por Hijo (AUH). El método de uso del dinero disponible es a través de una tarjeta entregada por el Banco Nación, de la cual no se permite extraer dinero en efectivo ni realizar transferencias.

Actualmente, los montos mensuales disponibles en la tarjeta son de $4000 pesos argentinos para una madre o padre con un chico de hasta 6 años y de $6000 para quienes tengan más de un hijo. A partir de diciembre del corriente año, los valores se actualizan a $8000 y $12000 respectivamente.

II. OBJETIVO Y MÉTODOS UTILIZADOS

El objetivo que tuvo este trabajo fue el de entender la base de datos subida al portal de datos abiertos de la nación (1) y, a través de modelos de aprendizaje supervisado, entender si, con los datos provistos, se podía armar un modelo predictivo de suscripciones por día al programa. Los modelos utilizados fueron Regresión Lineal y Support Vector Regression (SVR)

A. Jupiter Notebook

Para poder analizar los datos y llegar al objetivo propuesto, se utilizó el entorno *Jupiter Notebook*. En el mismo se desarrollo el Análisis Exploratorio de Datos (EDA) y se aplicaron los modelos de machine learning.

Las librerías utilizadas fueron Numpy (para el armado de matrices), Pandas (para cálculos aritméticos), Matplotlib (para visualización de datos), Scikit-Learn (para los algoritmos de regresión) y Geopandas (para la referenciación geográfica de los datos).

B. Regresión Lineal

La Regresión Lineal es un algoritmo de aprendizaje supervisado. Lo que persigue es encontrar una aproximación que modele la relación entre la variable escalar dependiente Y (label/etiqueta) y una o más variables explicativas X (simples/muestras) [1].

La función lineal en este caso se construye de la siguiente manera:

$$Y = mX + b$$

Donde Y es el resultado, X la variable, m es la pendiente de la recta y b la constante (donde X = 0)

Las medidas de error para evaluar la *performance* de una regresión son las siguientes:

- $MAE = \frac{|\sum(yt' - yt)|}{n}$
- $MSE = \frac{\sum(yt' - yt)^2}{n}$
- $RMSE = \sqrt{MSE}$
- $R^2 = \frac{TSS - RSS}{TSS}$

Para el caso del R^2, el TSS (Total Sum Squares) mide la varianza total de las etiquetas “y”. El R^2 explica la proporción de la varianza de “y” que explica el modelo de regresión. El R^2 toma valores de entre 0 y 1. El RSS es la suma de los residuos al cuadrado.

C. Support Vector Regression (SVR)

El otro modelo utilizado en el trabajo fue SVR, el cual busca construir una función lineal o hiperplano separador que generalice de la mejor manera posible el fenómeno observado [2]. Específicamente, determina un margen como función de costo y trata de que todas las muestras estén dentro de ese margen. Caso de que el problema no presente una solución lineal, con kernels se puede mapear no linealmente las muestras a otro espacio donde el hiperplano funcione. Los hiperparámetros en este caso serán que tanto penalizo al modelo en dejar muestras por fuera del margen establecido y, caso que aplique, el kernel a utilizar (gaussiano, lineal, polinomial).

Las medidas de error para evaluar la performance del modelo son las mismas que se expresaron anteriormente en la regresión lineal.

III. DATASET ELEGIDO

El *DataSet* elegido se encuentra en la página oficial del gobierno nacional, en el apartado de datos abiertos [3]. El mismo cuenta con 1.2M filas de registros únicos de personas que se acercaron a retirar la tarjeta AlimentAR, la cual da acceso al monto de dinero antes mencionado para cambiar por bienes alimenticios, en los lugares convenidos por el plan.

A fines de depurar el *DataSet* para analizar correctamente los datos se procedió a:

- Verificar registros duplicados (los cuales no se encontraron)
- Encontrar valores nulos (había una columna de “fecha_baja” completamente vacía)
- Pasar de formato *string* a *float* a todos los valores numéricos
- Agregar una columna de monto liquidado total por persona.

Se agregaron, a modo de complemento, un .csv con la población total de cada provincia y otro con la incidencia % según la población total, para el armado de un mapa de calor del país.

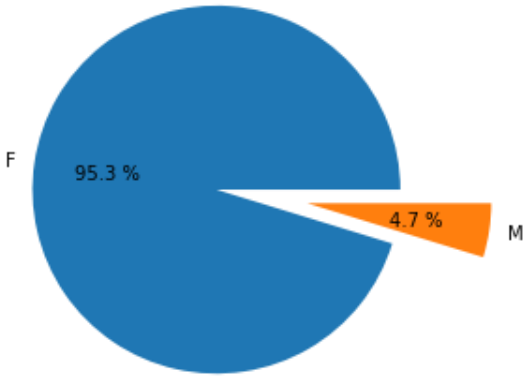
IV. ANÁLISIS EXPLORATORIO DE DATOS (EDA)

Se obtuvieron varios hallazgos del *DataSet* a la hora de realizar el análisis exploratorio de datos.

En primer lugar, se detecto que el 95% de registros únicos del *DataSet* pertenecen al sexo femenino, es decir, son las madres de familia las que se acercan a obtener el beneficio AlimentAR. Por el contrario, solo un 5% pertenecen al sexo masculino. Esto se puede observar en la Figura 1.

Figura 1

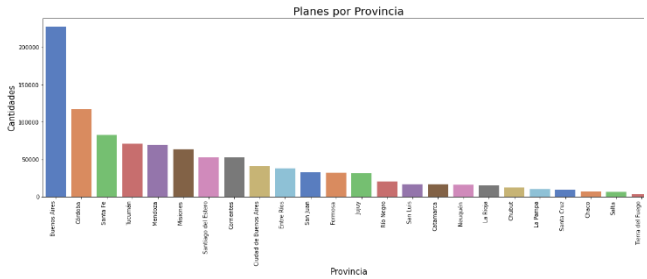
Porcentaje de Hombres y Mujeres del DataSet elegido



Por otro lado, se realizo un estudio de cuales son las provincias que más personas adheridas al plan tienen. En la Figura 2 podemos observar aquellas de mayor relevancia nominal.

Figura 2

Cantidad de personas por Provincia con Tarjeta AlimentAR

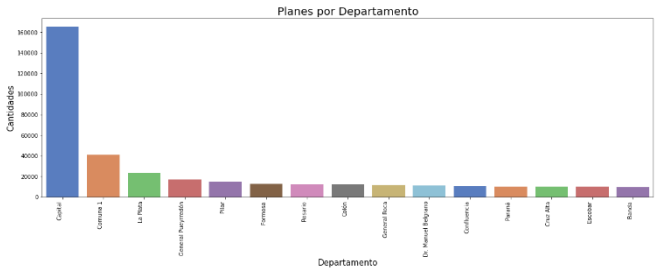


Es para destacar que Buenos Aires, Córdoba y Santa Fe cuentan con más del 40% de las personas adheridas al plan, con alrededor de 450.000 registros.

En la Figura 3 se puede ver el mismo análisis, pero a nivel departamento.

Figura 3

Cantidad de personas por Departamento con Tarjeta AlimentAR

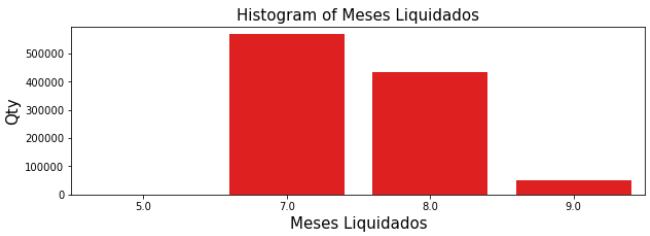


En particular se observa que el departamento que mayor relevancia tiene dentro del plan es “Capital”, el cual hace alusión a las capitales de Córdoba, Santa Fe y Ciudad Autónoma de Buenos, por mencionar las más relevantes.

Ya haciendo un análisis de los montos, podemos observar en la Figura 4 cuantos fueron los meses liquidados a cada registro único del *DataSet.*

Figura 4

Cantidad de Meses Liquidados por Persona

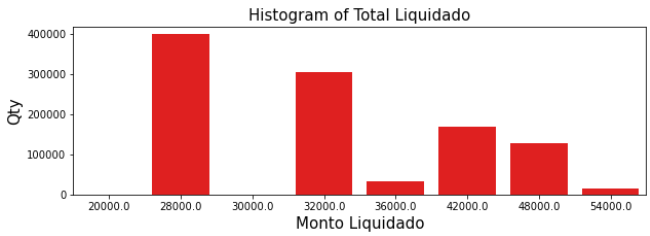


De las más de 1.2M de personas registradas, solo 450 hicieron uso del plan por 5 meses, siendo que la mayor concentración esta entre 7 y 8 meses.

En cuanto a los montos liquidados, se observa en la Figura 5 un histograma del total liquidado por registro único.

Figura 5

Monto total liquidado por Registro Único

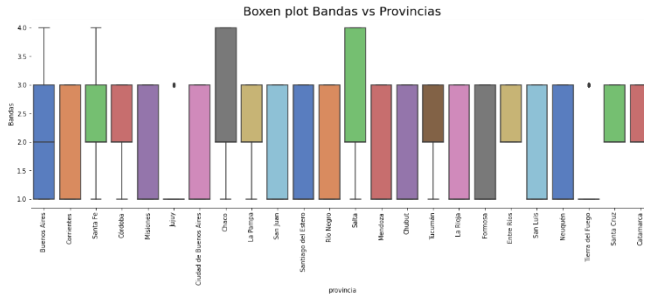


Los valores más representativos son el de $28.000 y $32.000 ARS respectivamente, mientras que pocas personas cobraron los montos del rango inferior y superior de nuestros datos. Haciendo una estadística descriptiva básica de estos valores, el monto medio otorgado por el Estado fue de $34.479 con un desvío de $7.401 ARS.

Relacionado a los montos cobrados, se hizo un gráfico de caja y bigote relacionando a las provincias con bandas de cobro en $ARS. Las bandas propuestas fueron 4 (>20.000 =<30.000, >30.000 =<40.000, >40.000 =<50.000, >50.000). En la Figura 6 podemos observar los resultados.

Figura 6

Boxen Plot de Bandas vs Provincias

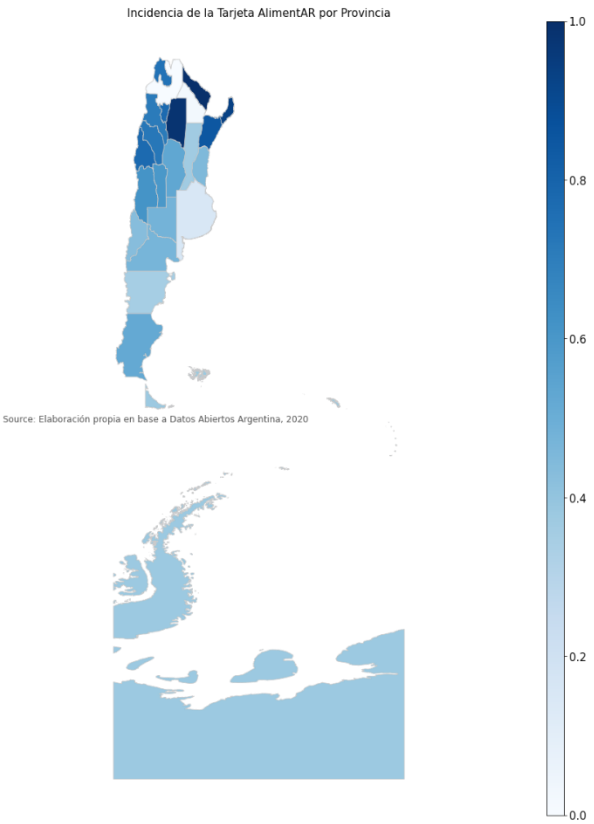


La única provincia que cuenta con valores en las 4 bandas es Buenos Aires, mientras que provincias como Santa Cruz y Catamarca tienen sus valores concentrados en la banda 2 y 3. Es también relevante observar que hay provincias como Santa Fe y Córdoba que casi no presentan valores en la banda 1, aunque sean las que más personas tienen adheridas al beneficio de AlimentAR. Hay comportamientos bien diferentes provincia por provincia, por lo que es interesante entender los resultados antes mostrados.

Finalmente, se realizo un mapa de calor del país en su totalidad, representando con distintos tonos de azul la incidencia del plan AlimentAR en relación a la población total de la provincia. En la Figura 7 podemos observar los resultados.

Figura 7

Heatmap según la incidencia del plan AlimentAR en Argentina



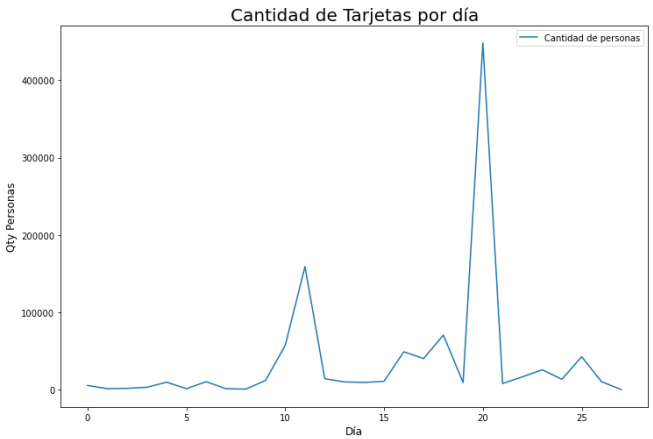
Vemos como provincias que anteriormente mencionamos como las más representativas a nivel nominal del plan dejan de ser tan relevantes vs la población total del territorio. Aquellas que están con un azul más fuerte (Formosa, Corrientes, Santiago del Estero y Misiones) son las que, en relación a sus habitantes, más planes obtienen del Estado (entre un 5% y 7% aproximadamente). Provincias como Buenos Aires y Santa Fe no llegan al 2,5%, por lo que tienen un color de azul más claro.

V. Resultados de Modelos de Regresión

El análisis que se propuso realizar fue el de partir al *DataSet* en cantidad de personas que fueron a buscar su tarjeta AlimentAR por día, para entender si se podía armar un modelo predictivo de personas a retirar el plan a futuro, y así preparar las localidades en relación a los resultados. Como vemos en la Figura 8, la cantidad de personas por día que se presentaron a retirar el plan pertenece a un rango de datos bien disperso.

Figura 8

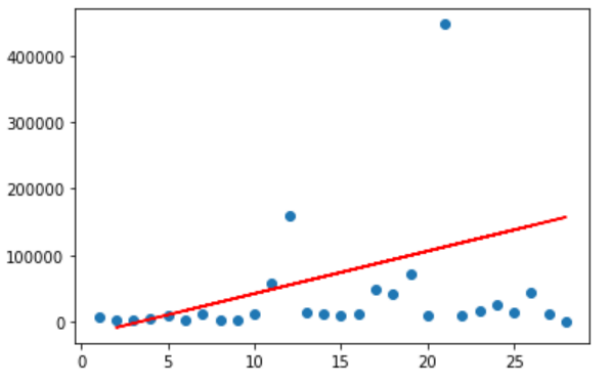
Cantidad de personas que retiraron la Tarjeta AlimentAR por día



En tan solo dos días se encuentra aproximadamente el 45% de las personas que retiraron la tarjeta, por lo que este fue un primer punto en contra para realizar la regresión por ambos modelos. Por mencionar uno de los resultados obtenidos en la regresión lineal, se observa en la Figura 9 que la regresión no generaliza bien el fenómeno observado.

Figura 9

Modelo de Regresión Lineal aplicado a la muestra



Vemos que hay valores diarios que se van demasiado del rango medio de los datos, siendo estos *Outliers* un problema a la hora de generalizar una función de decisión que nos ayude a predecir la cantidad de personas diarias que, en el futuro, buscaran su tarjeta AlimentAR.

Los resultados según las métricas de medición del éxito de los modelos se pueden ver en la Figura 10.

Figura 10

Métricas de Regresión aplicadas al DataSet

Modelo	RMSE	MAE	R^2
Regresión Lineal	85.056	69.342	-15,52
Regresión Lineal s/outliers	23.837	18.618	-0,19
SVR	37.327	34.669	-2,17
SVR s/outliers	22.783	16.369	-0,09

En todos los casos el R^2 no es representativo, ya que no esta acotado entre 0 y 1. Por otro lado, se intento hacer el análisis quitando los *outliers*, más allá que no sea correcto por la representatividad de los mismos.

VI. CONCLUSIONES

Los resultados obtenidos no fueron los esperados, pero se podrían lograr mejorar métricas si el *DataSet* subido por el gobierno nacional contará con mayor horizonte de tiempo, y así conseguir mayor cantidad de muestras para entrenar al modelo.

VII. REFERENCIAS

[1] Obi Tayo, Benjamin (2020). Recuperado de https://medium.com/towards-artificial-intelligence/how-do-i-calculate-accuracy-for-regression

[2] Bishop, C.M. (2006). Pattern Recognition and Machine Learning. Springer Science + Business Media

[3] Titulares de la Tarjeta AlimentAR (2020). Recuperado de https://datos.gob.ar/dataset/desarrollo-social-titulares-tarjeta-alimentar