**MEMORIA: Ganancias y gastos de familias filipinas**

Artem Vartanov y Mario Baldocchi

Titulos a tamaño 14, texto tamaño 12

**Introducción**

Los [datos](https://www.kaggle.com/datasets/grosvenpaul/family-income-and-expenditure) que estudiaremos reflejan las ganancias y gastos de diversas familias filipinas, además de ciertos aspectos importantes a la hora de reforzar nuestra idea sobre el poder adquisitivo de las mismas.

Esta base de datos se podría emplear en diversos ámbitos laborales, pero supondremos que nuestro cliente es un banco, al que le interesa poder saber cuándo conceder un préstamo o denegarlo. Para poder hacer esto último, una de las principales variables a tener en cuenta es el dinero total que ingresa cada unidad familiar. Sin embargo, en Filipinas mucha gente trabaja en negro, por lo que no puede uno realmente fiarse de lo que el cliente deposita en la cuenta. Es por ello que, basándonos en los gastos y otros datos personales, trataremos de predecir el sueldo

anual (en pesos filipinos, que a 30/11/2023 el cambio es de 1₱ = 0,017€) de una unidad familiar.

**Descripción de los datos**

La base de datos se conforma de 18022 clientes y 60 variables, de las cuales 43 son nominales y 17 categóricas. Entre las distintas variables nominales se puede observar que aproximadamente un 25% están relacionadas con desembolsos monetarios, como por ejemplo los gastos médicos, en ropa o en comida. El resto describe la familia y la vivienda en la que residen.

**Preprocesado**

Una vez importados los datos, lo primero que debe hacerse es echar un vistazo por encima. De lo primero que se percata uno, es que hay variables categóricas "escondidas", entre las que se encuentra, por ejemplo, Electricity. Y es que aunque en esta columna solo haya números (0 y 1), estos son utilizados para indicarnos si el cliente en cuestión tiene electricidad en casa o no, es decir, se podría sustituir el 1 por un "Sí" y el 0 por un "No". Consecuentemente, hay que ir columna por columna comprobando que ninguna variable categórica se haga pasar por otro tipo de variable, y en el caso de localizar alguna, realizaremos su conversión a factor con as.factor <líneas 20-36>.

Seguimos analizando. Vemos que hay solo dos columnas que tienen valores N/A´s, *Household.Head.Occupation* y *Household.Head.Class.of.Worker,* que representan el trabajo del miembro de la familia con las mayores ganancias y para quién trabaja respectivamente. Ante esta situación se nos presentan varias opciones. Por un lado, se puede optar por la eliminación de todas aquellas filas que contengan un N/A (3355 en total). Por el otro, podría ser interesante prescindir de ambas columnas. En este caso, elegiremos la segunda opción por dos motivos principales. En primer lugar, tanto la una como la otra no aportan información relevante frente al resto de variables, por lo que es preferible perder dicha información a "sacrificar" un cliente. En segundo lugar, al ser dos variables categóricas no formarán parte del modelo de regresión, el cual es una de nuestras prioridades. Es verdad que si estuviésemos tratando con un banco que facilitase muy pocos datos de los clientes (no es nuestro caso), cabría la posibilidad de optar por quitar todas aquellas filas con algún N/A para posteriormente convertir las variables categóricas en "dummy", pudiendo así ajustar mejor el modelo de regresión.

Para finalizar con el preprocesado, es recomendable prestar atención en cómo se distribuyen los datos. Para ello, podemos comenzar generando un histograma de cada una de las variables. De esta manera, podemos percatarnos de que todos aquellos atributos que representan algún gasto tienen algo en común: son asimétricos hacia la derecha. Esto es algo completamente lógico, pues el gasto que realiza una familia solo puede encontrarse en el intervalo (0, ∞). Además de la simetría, es de especial interés conocer si existen outliers o no. Para esto último, generaremos un "plot" para cada una de las variables. Como era de esperar, diversas columnas como .... presentan outliers.

Una vez realizados todos los pasos anteriormente mencionados, nos queda una base de datos limpia de N/A´s y outliers, donde la categoría de todas las variables es la correcta.

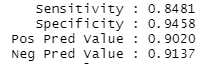
**Regresión Logística**

La semana pasada se ha producido uno de los peores terremotos de la historia de Filipinas, dejando gran parte del país destrozado. Consecuentemente, muchas familias han perdido sus casas. Este acontecimiento ha hecho que el banco haya visto una oportunidad de oro para aumentar su capital. Es por ello que ha puesto en marcha el PRVF (plan para la reconstrucción de viviendas en Filipinas), un préstamo de 90000₱ (1485€) al que cualquier familia filipina puede acceder siempre y cuando cumpla una condición: la renta per cápita mensual (ingresos mensuales de la familia / número de miembros en la familia) debe ser mayor o igual al 5% (4500₱) del préstamo.

Para poder realizar el modelo de regresión logística utilizaremos la misma base de datos que hasta ahora, con una pequeña diferencia, y es que *Total Household Income* se transforma en una variable categórica, donde los únicos posibles valores que puede tomar son “Accept”, si la familia cumple la condición, o “Deny” en caso contrario.

En el primer modelo vamos a tener en cuenta todos los atributos nominales disponibles en la base de datos. Sorprendentemente obtenemos un modelo con una gran precisión.

****

****

Antes de quedarnos con este modelo, comprobemos si hay uno mejor

**Correlaciones**

Al construir el modelo de regresión con las variables numéricas, es necesario eliminar las columnas categóricas. Antes de comenzar a estudiar las correlaciones, ese seleccionan todas aquellas variables que cuyaslas variables correlacionadas con la variable respuesta. Pero antes de ello, primero quitamos las columnas no numericas.

\*\*\*FALTA TEXTO\*\*\*

**Regresión lineal múltiple completa (RLM\_FULL)**

Generamos el modelo de regresión múltiple que usará todas las variable numéricas y todo el conjunto de datos. En la *Figura 1* aparece el summary del modelo completo. Vemos que no todas las variables tienen el p-valor < 0.05 (valor prefijado para rechazar hipótesis nulas). Esto quiere decir que no podemos estar seguros de que el coeficiente de esa variable no sea 0, en otras palabras, que debería estar en el modelo esa variable.

Cabe destacar que el R2 ajustado es 0.7981 que mide la cantidad de variabilidad explicada por el modelo, teniendo en cuenta la penalización por introducir más variables en el modelo(usamos este criterio para poder contrastarlo con modelos posteriores, pues tendrán diferente número de variables). En este caso, como tenemos 42 predictores, la penalización es bastante alta.

Después de calcular la raíz cuadrada del error del modelo nos sale 132910.2, lo cual parece ser un error crítico.