Tecnológico de Monterrey

Instituto Tecnológico de Estudios Superiores de Monterrey

Analítica de datos y herramientas de inteligencia artificial II

Actividad 6:

Regresión Lineal Múltiple y No Lineal

Profesor:

Alfredo García Suárez

Equipo 4:

María Fernanda Ledesma Martínez	A01734203
Karla Yamila Pagés Mejía	A01733409
César Ricardo Gastelú Parra	A01735328
Agustín Ibarra Sota	A01552618

Fecha de entrega: 12/10/2023

Regresión lineal múltiple y no lineal

A lo largo de esta actividad se analizan datos de un dataframe en distintas etapas, desde su limpieza, extracción de datos, hasta el análisis de regresiones lineales y no lineales; con el objetivo de sacar información valiosa que ayude a describir los datos, sus relaciones entre variables e interpretar los resultados para la toma de decisiones posterior.

Se trabajó con la base de datos *"BD_Socio formador (**TrainingDataComplete**)",* la cual consiste en 252,000 registros y 13 columnas que tratan de describir ciertas características de personas que piden un préstamo, y su riesgo a ser morosos basados en sus datos históricos. Se observan variables como el tipo en la casa (ya sea rentada o propia), su ingreso, años trabajando en su trabajo actual, edad, años de experiencia, profesión, estado civil, si tiene carro o no, estado y ciudad, así como su riesgo a ser deudor basado en sí la persona tuvo alguna deuda en el pasado.

Limpieza de datos

Se destaca que el conjunto de datos no presenta datos nulos ni la aparición de registros atípicos; por lo que, no se realizó ninguna manipulación de los datos y se dejó tal cual para su análisis.

Extracción de características

Para este punto se realizó la extracción de características de las columnas: Age, Experience, Married/Single, House_Ownership, Profession, CITY, CURRENT_JOB_YRS, CURRENT_HOUSE_YRS, y Risk_Flag con el objetivo de conocer un poco más de información sobre cada una. Para ello, se creó una copia del data frame original solo con las columnas requeridas y todas las variables se pasaron a tipo objeto para poder tratar a cada una como categórica y hacer el análisis correspondiente.

A continuación, se presentan las descripciones de los hallazgos obtenidos de cada una de las variables luego de realizar distintas gráficas con el fin de tener una mayor comprensión y visualización de estas

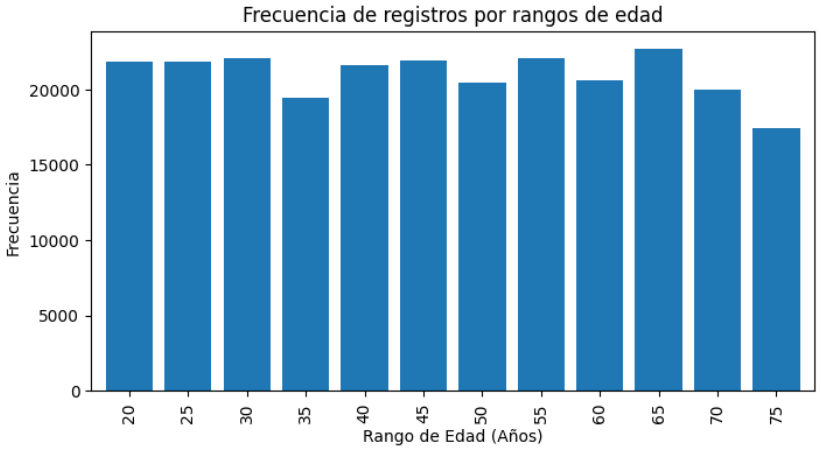
1. **Age**

   Para la variable de edad se obtuvo una gráfica de barras que mostrará la frecuencia de las edades y con ello, poder concluir cuáles son las edades que más buscan este tipo de servicios financieros. Para esto, se realizó previamente un reagrupamiento de las edades para facilitar su visualización; dividiéndolas en rangos de 5 años, en dónde la edad mínima es 20 y la máxima es 75.

Como se observa en la *Figura 1*, las personas que más tienden a pedir un préstamo son los que se encuentran entre los 20 y 30 años; sin embargo, también destacan las personas entre 60 y 65 años.

Figura 1

Frecuencia por edades



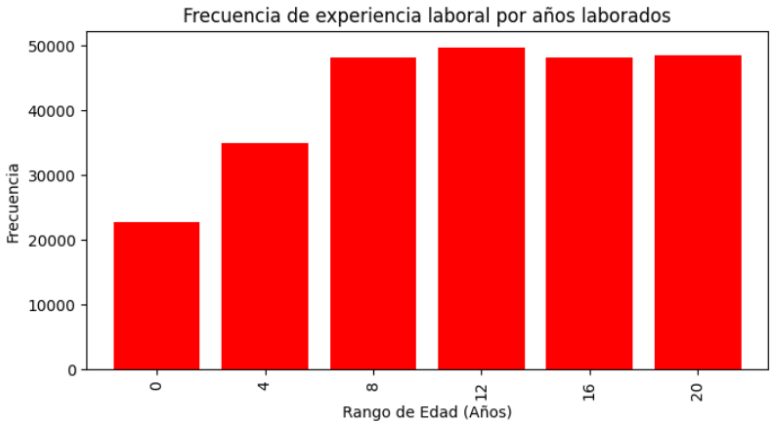
2. Experience

Para la variable de *Experience*, al igual que con *Age*, se realizó un agrupamiento de los años para facilitar su visualización; en este caso, los rangos se dividieron por: 0-1, 1-4, 4-8, 8-12, 12-16 y 16-20 años. Considerando que existen personas que tienen menos de un año de experiencia hasta las personas con el máximo de 20 años.

Siendo así, se observa en la Figura 2 que la mayoría de la población cuenta con más años de experiencia, y estos son los que más piden préstamos. Además, el rango en dónde se ve más concentrado es el de 8 a 12 años.

Figura 2

Frecuencia de años de experiencia profesional

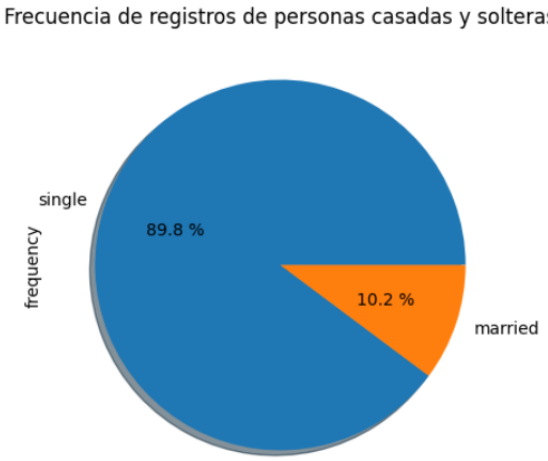


3. **Married/Single**

Para esta columna sobre el estado civil de las personas, se decidió hacer un gráfico de pastel por ser pocas instancias dentro de la variable. De esta manera rápidamente se puede visualizar la proporción de personas casadas sobre solteras; que en este caso es casi un 90% de la población que son solteras.

Figura 3

Frecuencia de registros que se encuentran solteros o casados

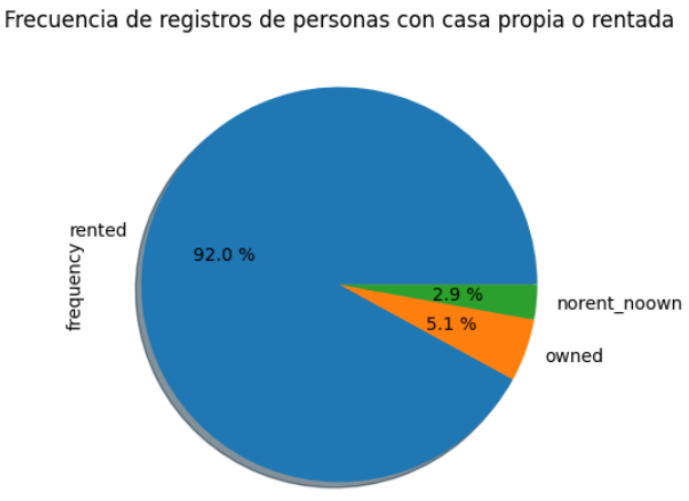


4. **House_Ownership**

Esta variable describe si la persona vive en casa propia, rentada o ninguna de las dos. También se eligió una gráfica de pastel al ser pocas instancias y poder visualizar rápidamente los datos. Se observa en la Figura 4 que la mayor parte de la población que pide un préstamo vive en casas rentadas y solo un 5% en casa propia.

Figura 4

Frecuencia de registros que cuentan con una casa propia, rentada o ninguna de las dos.

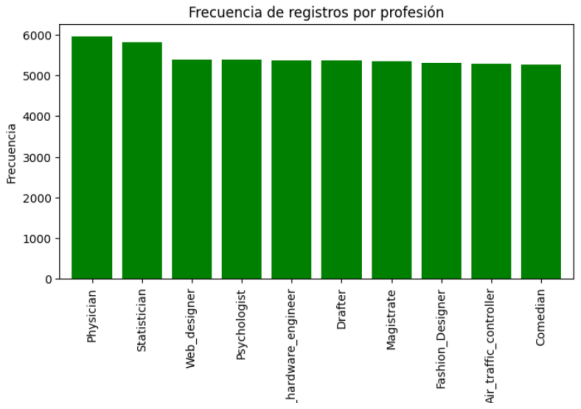


5. Profession

Esta variable tiene muchas instancias, por lo que decidimos solo visualizar las 10 más populares y con ello, sacar información de qué profesiones son las que más adquieren servicios financieros. Sin embargo, como se observa en la *Figura 5*, la proporción de personas que estudian una carrera es casi igual en todas; aun así, las carreras más destacadas en este rubro son las de: físicos y estadísticos (que de hecho podrían pertenecer a la misma área).

Figura 5

Frecuencia de registros que practican una profesión específica.

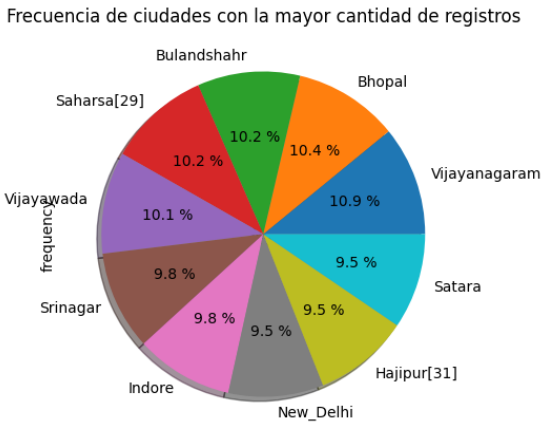


6. CITY

Para esta variable se realizó un filtro que mostrara las 10 ciudades más populares entre las personas que piden algún préstamo; esto, debido a que existen más de 300 ciudades en total en el dataframe. Para poder visualizar los datos y sacar la información más valiosa se realizó un gráfico de pastel como se muestra en la Figura 6. Adicional a esto, se observó que al ser tantas ciudades la proporción de cada una era muy baja con respecto al total de los datos, al mismo tiempo que muy similar entre todas.

Figura 6

Top 10 ciudades.

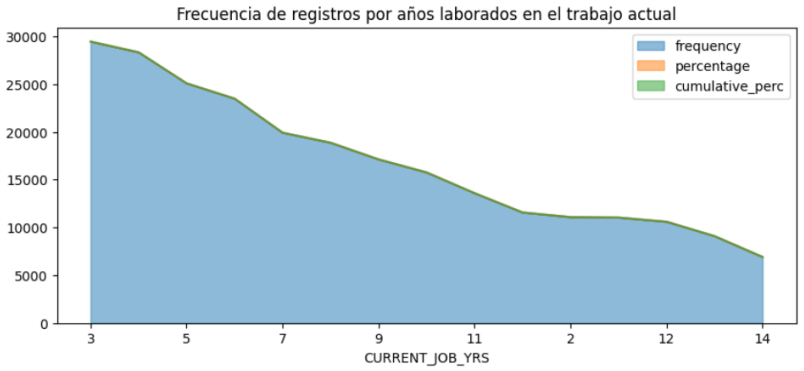


7. CURRENT_JOB_YRS

Como se muestra en la Figura 7, esta variable muestra la frecuencia de registros por años laborados en el trabajo actual, siendo 3 años el valor mínimo y 14 el valor máximo. Así mismo, se observa que entre menores los años, son más las personas.

Figura 7

Frecuencia de registros por años laborados en un trabajo actual.



8. CURRENT_HOUSE_YRS

Esta variable mide los años que las personas llevan viviendo en su domicilio actual; al ser pocas categorías, que van de los 10 años hasta los 14, se decidió hacer una gráfica de pastel que pudiera mostrar la proporción de los años y sus diferencias en frecuencia. Como se puede observar en la *Figura 8*, la proporción de cada una de las categorías es muy similar entre ellas, no habiendo mucha diferencia, puesto que los años son continuos.

Figura 8

Frecuencia de registros por año viviendo en un domicilio actual.

9. **Risk_Flag**

Esta variable indica el riesgo de la persona es ser o no morosa, basada en si en el pasado tuvo alguna experiencia de este tipo, siendo 1 si contiene riesgo o 0 en caso contrario.

Así como variables anteriores con pocas categorías, se prefirió utilizar una gráfica de pastel para visualizar los datos, como se observa en la *Figura 9*.

En este sentido, se percibe que la mayoría de los casos que van a pedir alguno de estos servicios, son personas sin riesgo a ser morosos.

Figura 9

Frecuencia de registros de personas que son candidatas o no al crédito.

Análisis de correlación

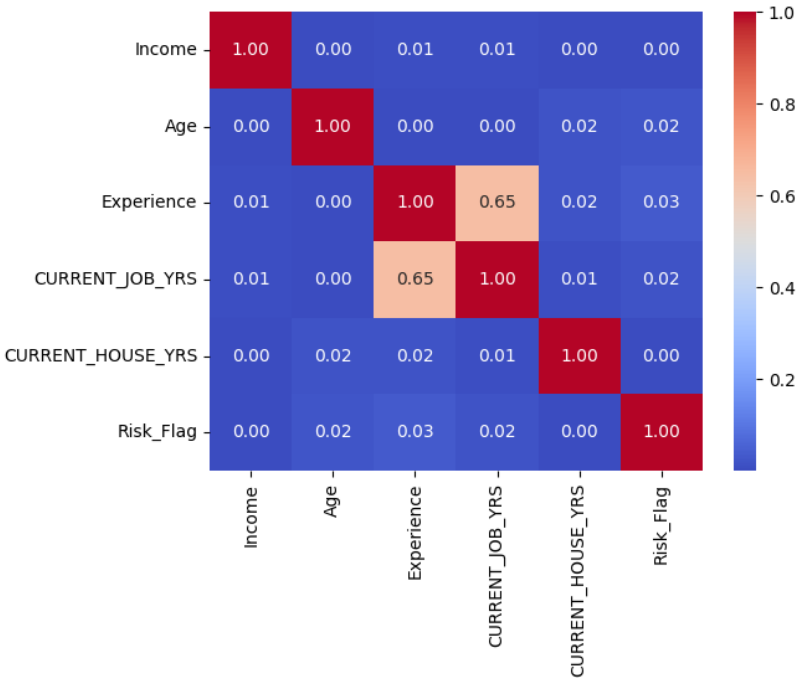
Para este punto también se creó una copia del data frame original, pero solo con las variables numéricas, con el objetivo de sacar la correlación de cada de ellas y los mejores modelos de regresión lineal simple, múltiple y no lineal y con ello, poder sacar predicciones.

- **Regresión lineal simple**

Como se observa en la *Figura 10*, la correlación entre las variables es casi nula o nula, a excepción de la correlación que existe entre el par de variables de “Experience” y “CURRENT_JOB_YRS”, las cuales ambas describen los años de las personas trabajando. Además, siendo la única correlación aceptable para la creación de un modelo de regresión.

Figura 10

Mapa de calor con correlaciones de datos reales.



- **Regresión lineal Múltiple**

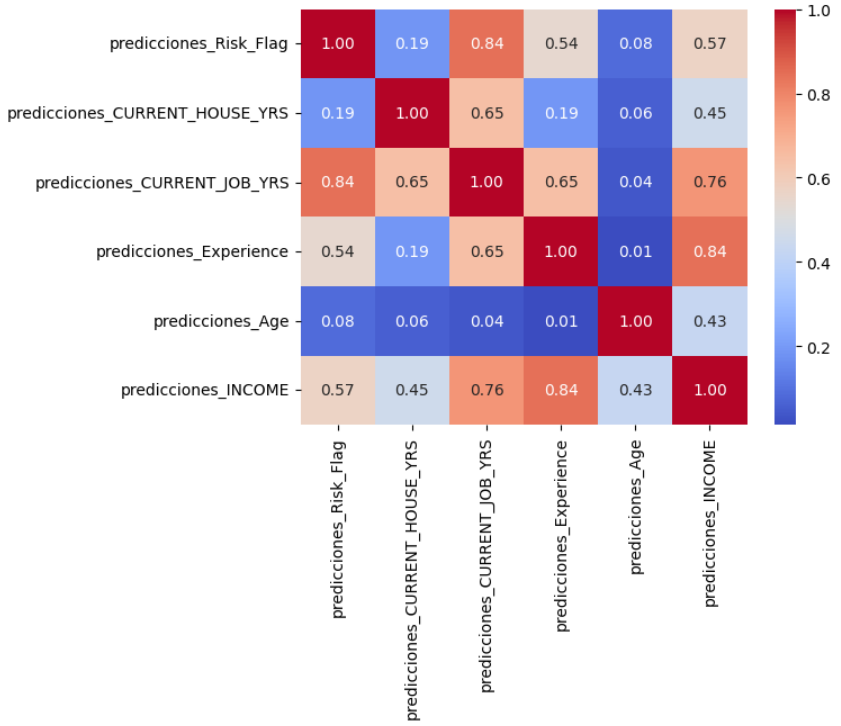
Para tener una mayor correlación entre las variables de nuestra base de datos numéricas, se realizó un ajuste de modelo mediante regresión lineal múltiple, en donde se evaluaron distintos casos para asegurar la relación entre la variable dependiente y las variables independientes.

Una vez obtenido el mejor modelo de regresión, es decir, con un coeficiente de correlación mayor al obtenido con la base de datos original, se procedió a realizar las predicciones de cada una de las instancias, almacenando estas en un nuevo data frame. Posteriormente, se realizó un mapa de calor con los coeficientes de correlaciones entre cada variable y otro mapa de calor con el coeficiente de determinación que hay entre cada variable.

A continuación, se presenta el mapa de calor de correlaciones (Figura 11) en donde podemos observar que a diferencia del mapa de calor con los datos reales, se ve una mejora significativa en estos valores del coeficiente, denotando que el arreglo con el ajuste de regresión lineal múltiple tuvo una mejora para cada variable cuantitativa.

Figura 11

Mapa de calor correlación regresión lineal múltiple.



Ahora bien, a continuación, se presenta la tabla 1 que resume los resultados de los modelos de regresión lineal múltiple. Cada fila de la tabla corresponde a un modelo diferente para cada una de las variables, y se incluyen la variable dependiente, las variables independientes utilizadas en el modelo, los coeficientes de correlación (R) y los coeficientes de determinación (R^2) asociados a cada uno de estos modelos:

Tabla 1. *Resultados de regresión lineal múltiple.*

Variable Dep.	Variables Indep.	Modelo de regresión Lineal Múltiple	Coeficientes
Income	Age Experience CURRENT_JOB_YRS CURRENT_HOUSE_YRS Risk_Flag	Ecuación de regresión Income = 5028710 - 130 Age + 1514 Experience + 3923 CURRENT_JOB_YRS - 5171 CURRENT_HOUSE_YRS - 25641 Risk_Flag	R-cuadrado 0.01 Correlación 0.1
Age	Experience CURRENT_HOUSE_YRS Risk_Flag	Age = 53.093 - 0.00422 Experience - 0.2464 CURRENT_HOUSE_YRS - 1.140 Risk_Flag	R-cuadrado 0.09 Correlación 0.3

Experience	Income Age CURRENT_JOB_YRS CURRENT_HOUSE_YRS Risk_Flag	Experience = 2.6280 + 0.000000 Income - 0.000952 Age + 1.06259 CURRENT_JOB_YRS + 0.06731 CURRENT_HOUSE_YRS - 0.4308 Risk_Flag	R-cuadrado 0.4182 Correlación 0.6466
CURRENT_JOB_YRS	Income Age Experience	CURRENT_JOB_YRS = 2.3262 + 0.000000 Income + 0.000615 Age + 0.392547 Experience	R-cuadrado 0.4175 Correlación 0.6461
CURRENT_HOUSE_YRS	Income Age Experience CURRENT_JOB_YRS Risk_Flag	CURRENT_HOUSE_YRS = 12.0546 - 0.000000 Income - 0.001654 Age + 0.006280 Experience - 0.00462 CURRENT_JOB_YRS - 0.01745 Risk_Flag	R-cuadrado 0.09 Correlación 0.3
Risk_Flag	Age Experience CURRENT_HOUSE_YRS	Risk_Flag = 0.17480 - 0.000422 Age - 0.001886 Experience - 0.000975 CURRENT_HOUSE_YRS	R-cuadrado 0.17 Correlación 0.4123

Tabla 1. Modelos de regresión lineal múltiple para variables cuantitativas

- **Regresión No Lineal**

Para describir los modelos de regresión no lineal, se utilizaron ecuaciones que describen comportamientos no lineales, en donde existe una sola variable independiente para cada variable dependiente.

Las ecuaciones que se utilizaron para predecir la variable dependiente fueron:

- y = ax^2 + bx + c ("Función cuadrática")
- y = a*exp(bx) + c ("Función exponencial")
- y = 1/a*x ("Función inversa")
- y = a*np.sin(x) + b ("Función senoidal")
- y = a*np.tan(x) + b ("Función tangencial")
- y = a*np.abs(x) + b*x + c (Función Valor absoluto)
- y = (a*x**2 + b)/ c*x (Función cociente entre polinomios)

- y = a*np.log(x) + b (Función logaritmica)
- y = a*x + b*x + c*x (Función lineal con producto de coeficientes)
- y = 1/a*x**2 (Función cuadrática inversa)
- y = a/b*x**2 + c*x (Función polinomial inversa)

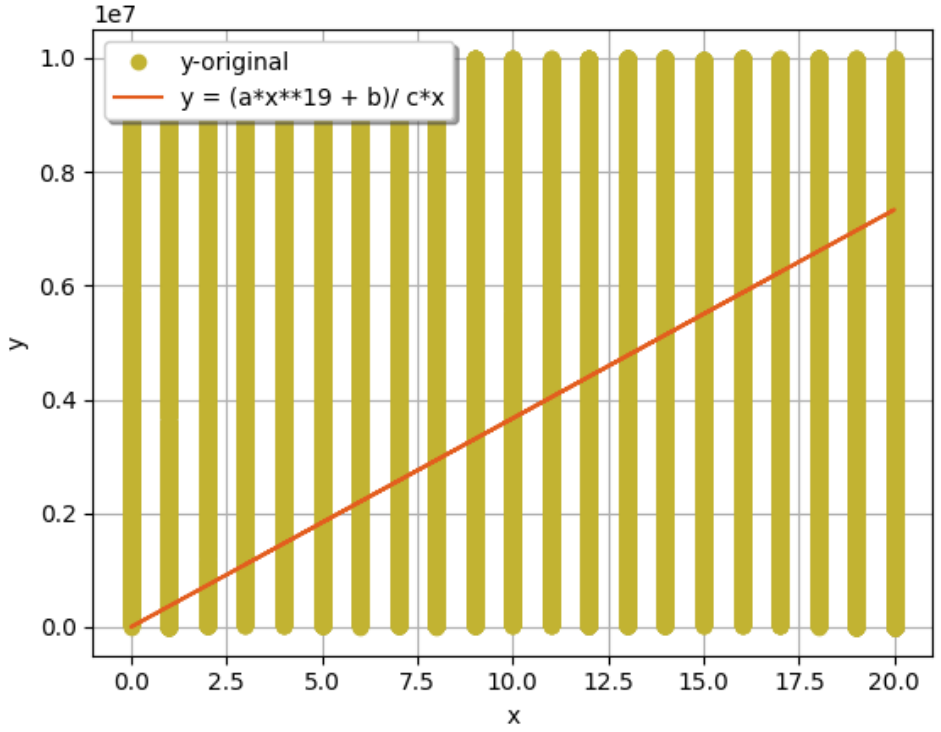
Para cada una de las variables numéricas se describieron modelos de regresión no lineal. Se hicieron 3 para cada una de ellas, en donde se cambiaron los exponentes y coeficientes de las ecuaciones de ajuste presentadas anteriormente. Se escogió uno de los tres modelos para cada variable, este, debía ser el que presentara el coeficiente de correlación más alto y dentro del rango aceptable.

A continuación se presenta el mejor modelo descrito por medio de una gráfica para cada una de las 6 variables dependientes.

- **Income**

Figura 12

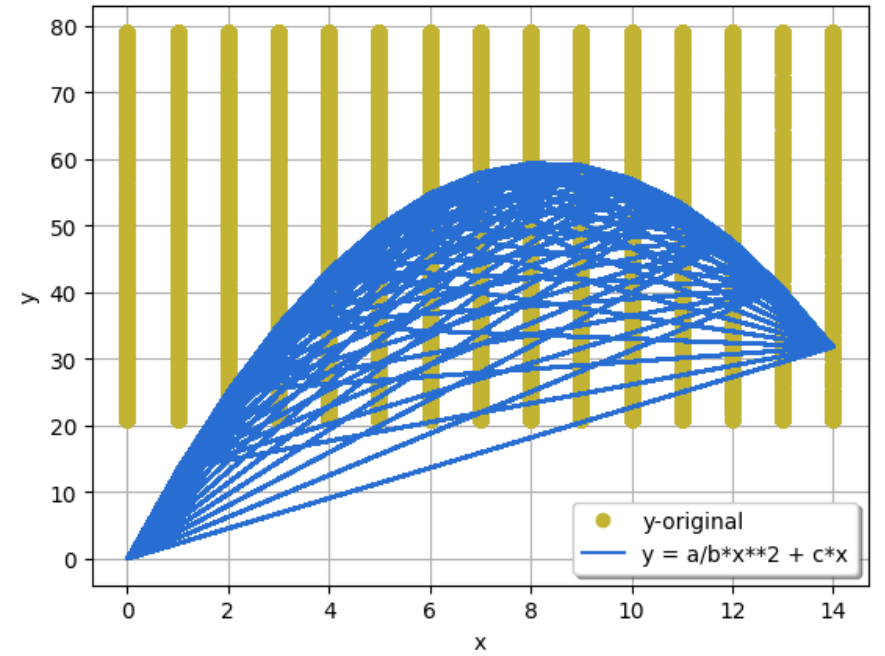
Ecuación de regresión no lineal predictora para la variable "Income".



- **Age**

Figura 13

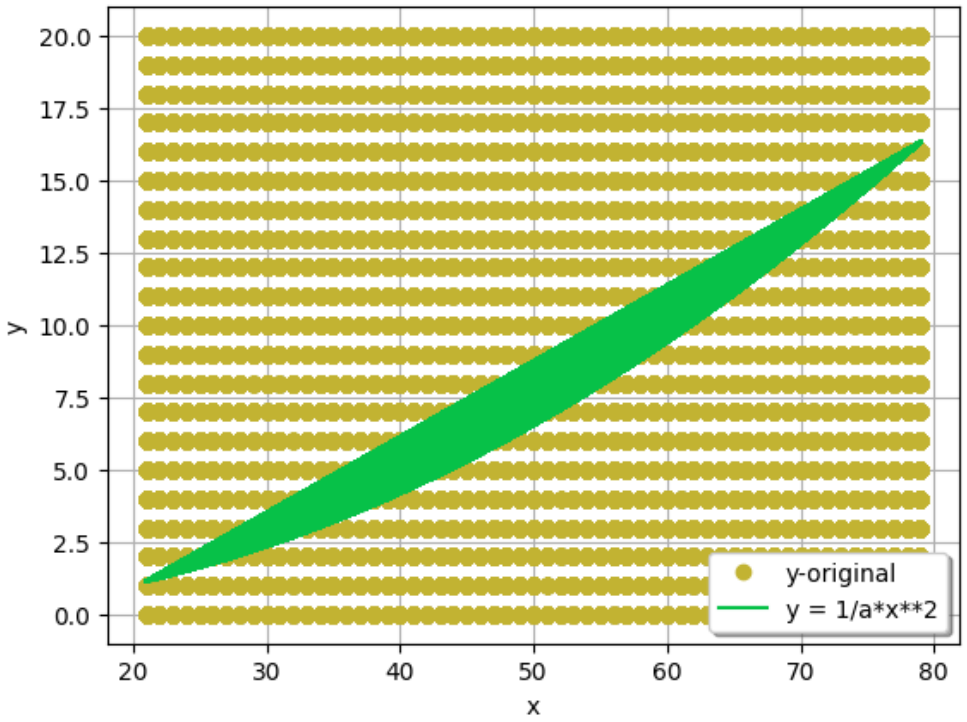
Ecuación de regresión no lineal predictora para la variable "Age".



- **Experience**

Figura 14

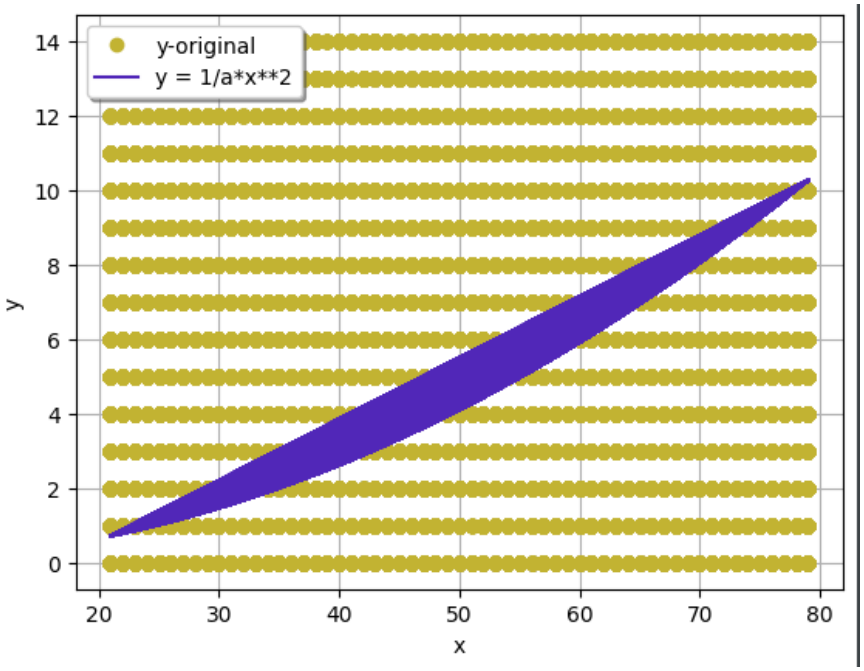
Ecuación de regresión no lineal predictora para la variable "Experience"



CURRENT_JOB_YRS

Figura 15

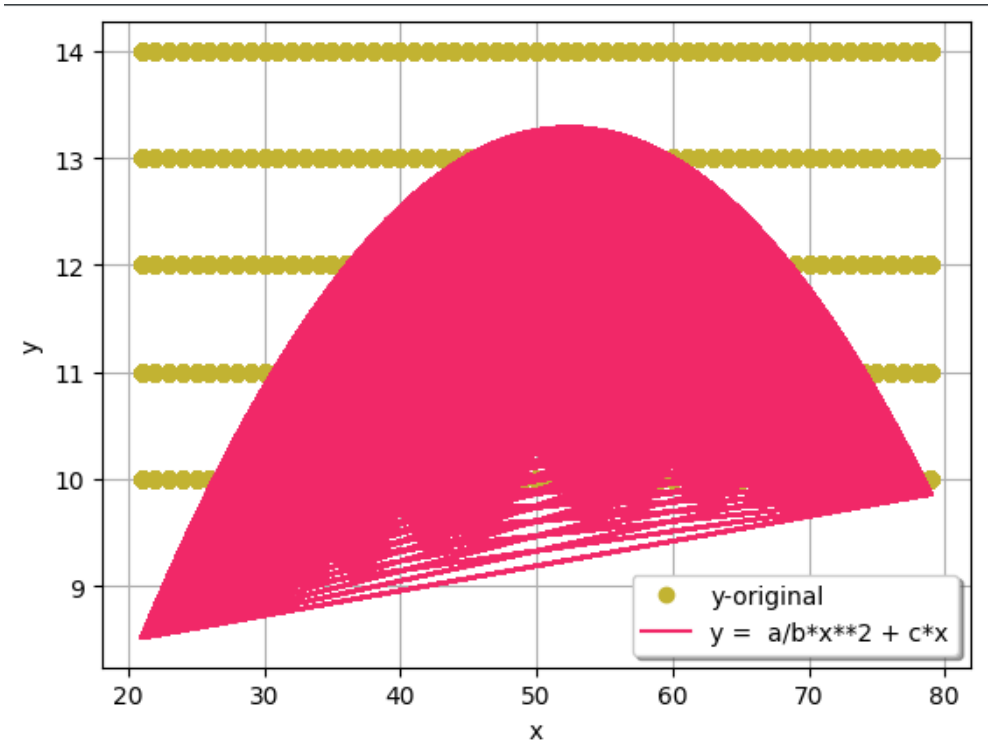
Ecuación de regresión no lineal predictora para la variable "CURRENT_JOB_YRS"



- **CURRENT_HOUSE_YRS**

Figura 15

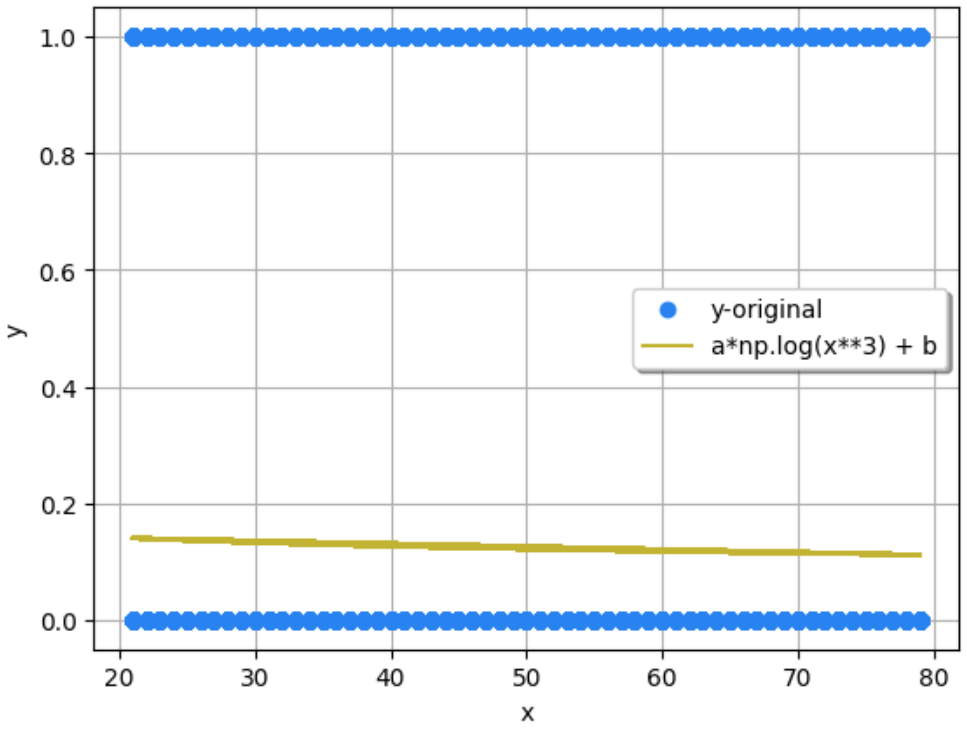
Ecuación de regresión no lineal predictora para la variable "CURRENT_HOUSE_YRS"



- **Risk_Flag**

Figura 16

Ecuación de regresión no lineal predictora para la variable "Risk_Flag"



En la tabla 2 a continuación, se presenta cada uno de los modelos presentados anteriormente, con sus respectivos coeficientes:

Tabla 2. *Resultados de regresión no lineal múltiple.*

Variable Dep.	Variables Indep.	Modelo de regresión Lineal Múltiple	Coeficientes
Income	Experience	Ecuación de regresión Income = 4.05669955e-15 (Experience)^16 + 2.17256772e+08 / 5.92475905e+02 (Experience)	Correlación 0.88
Age	CURRENT_JOB_YRS	Age = -0.51996886/0.60653883 * (CURRENT_JOB_YRS)^2 + 14.27229173 * CURRENT_JOB_YRS	Correlación 0.96
Experience	Age	Experience = 1 / 382.39569028 * (Experience)^2	Correlación 0.89
CURRENT_JOB_YRS	Age	CURRENT_JOB_YRS = 1/ 607.74108478 * (Age)^2	Correlación 0.91

CURRENT_HOUSE_YRS	Age	CURRENT_HOUSE_YRS = 0.11873093 / 24.5326461 * (CURRENT_HOUSE_YRS)^2 + 0.50714925 * (CURRENT_HOUSE_YRS)	Correlación 0.97
Risk_Flag	Age	Risk_Flag = 0.00716136 * log((Risk_Flag)^3) + 0.20561261	Correlación 0.02

Tabla 2. Modelos de regresión no lineal para variables cuantitativas

Posteriormente, se realizó una comparación entre las correlaciones de los modelos de regresión múltiple con los no lineales. Como se aprecia en las figuras 17 y 18, es visible que los modelos de regresión no lineal tienen una mejor predicción de los valores buscados que los de regresión múltiple.

Las matrices de confusión se presentan a continuación:

Figura 17

Mapa de calor correlación regresión lineal múltiple

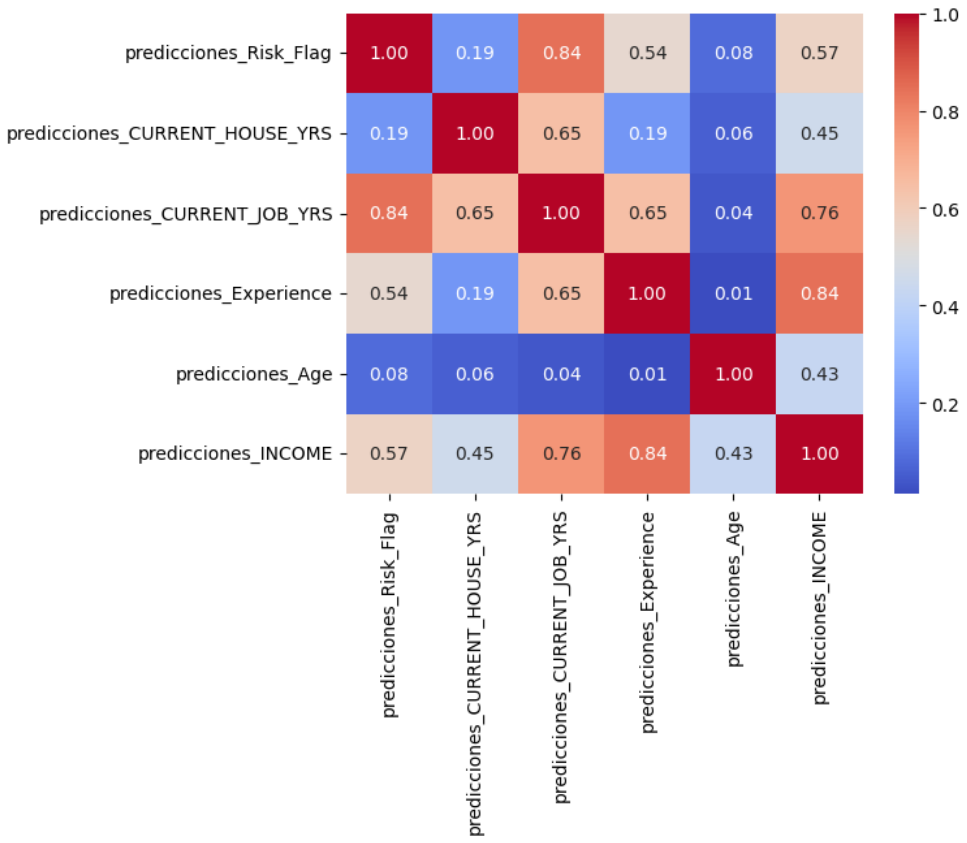
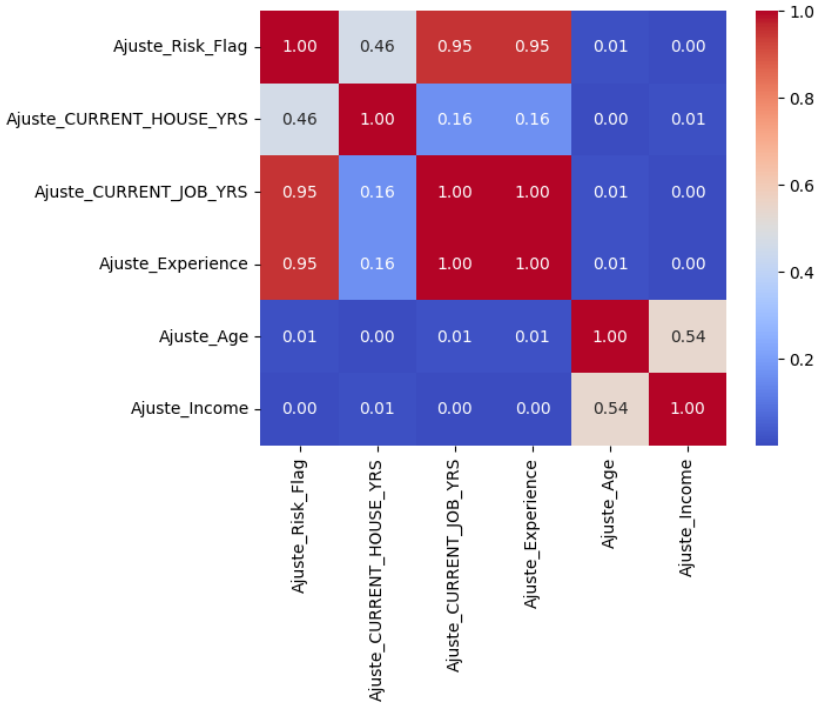


Figura 18

Mapa de calor correlación regresión no lineal



Como se puede observar, la matriz de correlación del modelo no lineal tiene coeficientes muy cercanos al 100%, por lo que es rápidamente concluido de manera visual que estos modelos, al tener grados exponenciales mayores a 1, son mejores que los de la regresión lineal múltiple.