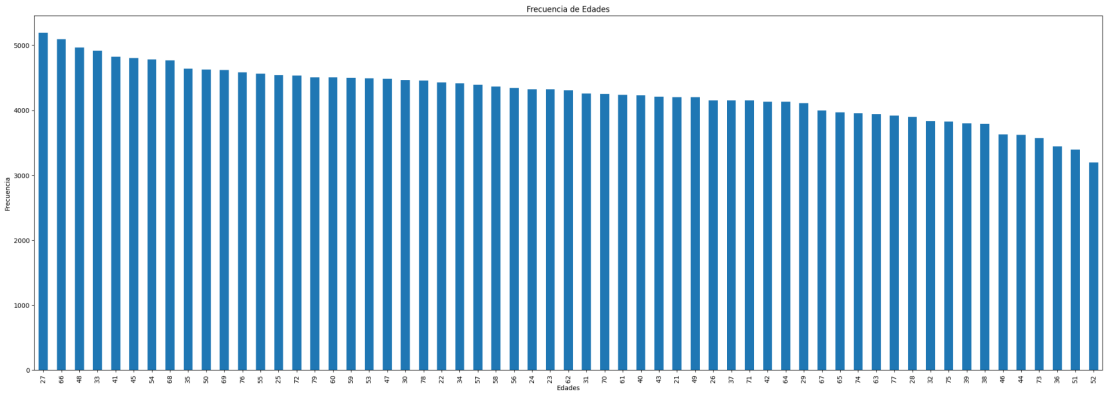
Leticia Castineyra Coutiño – A01173881
Aremy Díaz Cárcamo – A01552123
Edgar Díaz Ibáñez – A01552600
Carlos Morales Díaz – A01731899

Actividad 6 - Regresión Lineal Múltiple y No Lineal.

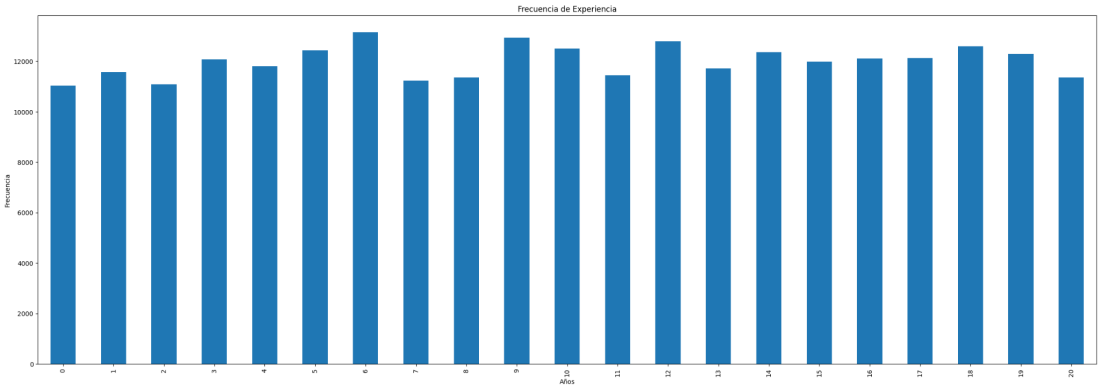
Extracción de características.

Con la base de datos original completa se analizó un total de nueve variables categóricas: Age, Experience, Married/Single, House_Ownership, Profession, CITY, CURRENT_JOB_YRS, CURRENT_HOUSE_YRS y Risk_Flag, donde se encontraron los siguientes hallazgos.

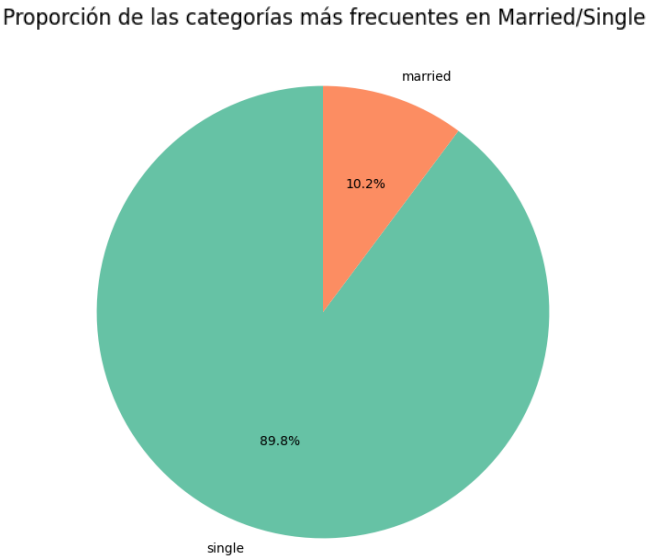
1. Age: si bien esta variable está clasificada como columna numérica, su naturaleza permite tratarla como variable categórica y analizar la frecuencia de sus posibles valores. En un inicio se realizó un histograma que graficaba en orden ascendente los valores de la edad; sin embargo, dado que no se encontró ningún patrón interesante se graficaron los valores en orden descendente de su respectiva frecuencia. Esta es a continuación insertada y permite identificar a la edad de 27 años como la más frecuente dentro de los usuarios disponibles en la base de datos y a la edad de 52 años como la menos frecuente. Desafortunadamente, en esta nueva representación tampoco se detectó patrón alguno.



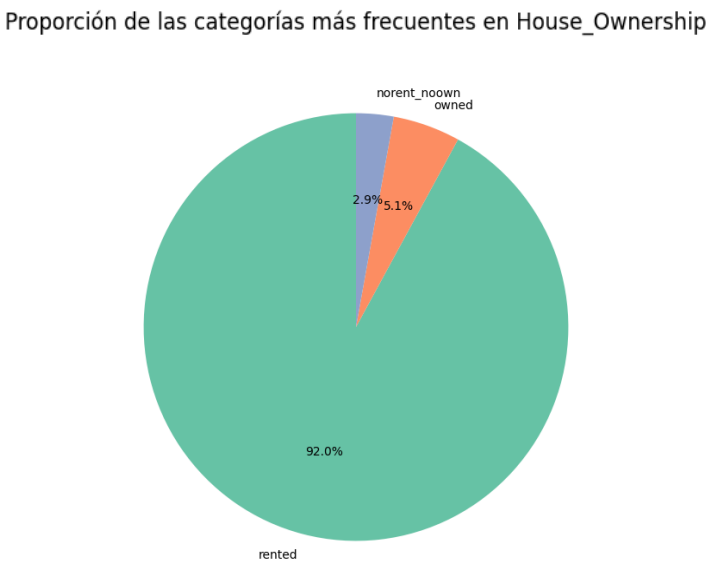
2. Experience: esta columna originalmente es de tipo numérica; sin embargo, su rango permite tratarla como variable de tipo cadena y analizar la frecuencia de sus registros. Si los valores se ordenan de manera ascendente, no se detecta ningún patrón relevante en la distribución de los datos; esto se aprecia en la imagen a continuación presentada.



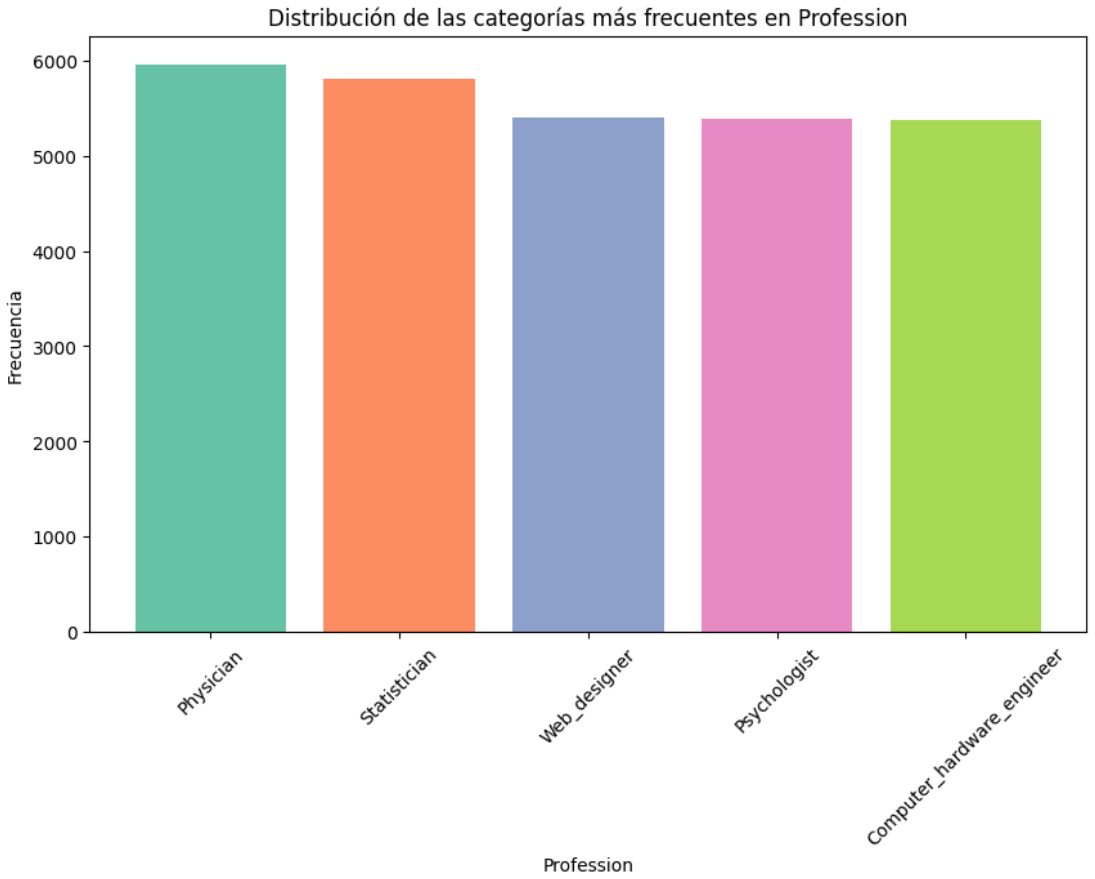
3. Married/Single: esta columna registra como valores únicos las dos opciones de estado civil mencionadas en el nombre de la variable. La frecuencia de sus valores no es uniforme pues del total de 252 mil registros solo el 10.2% de los registros pertenecen a la categoría Married y el 89.8% restante a la categoría Single como se observa en la siguiente imagen.



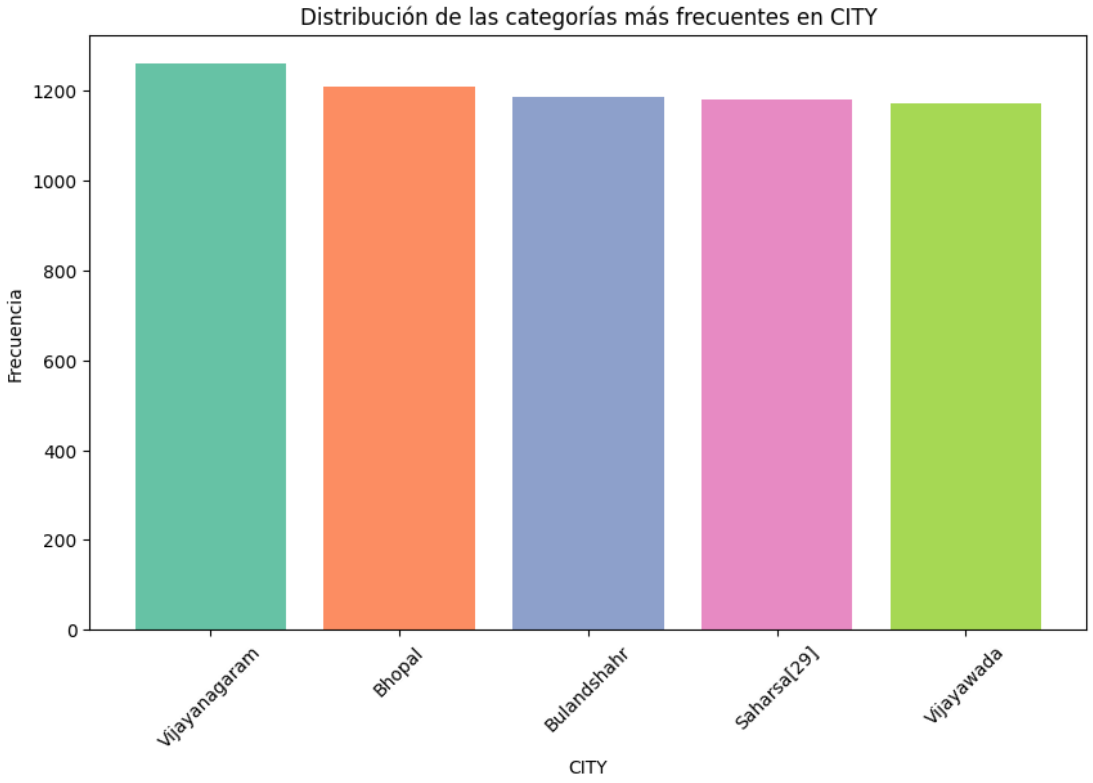
4. House_Ownership: esta variable registra tres valores únicos dependiendo de si el usuario tiene casa propia, rentada o ninguna de esas dos previas opciones. Se reconoce que el 92% de los registros vive en una casa rentada, el 5.1% en casa propia y el 2.9% restante en unidades habitacionales no propias ni rentadas como se puede observar en la siguiente imagen.



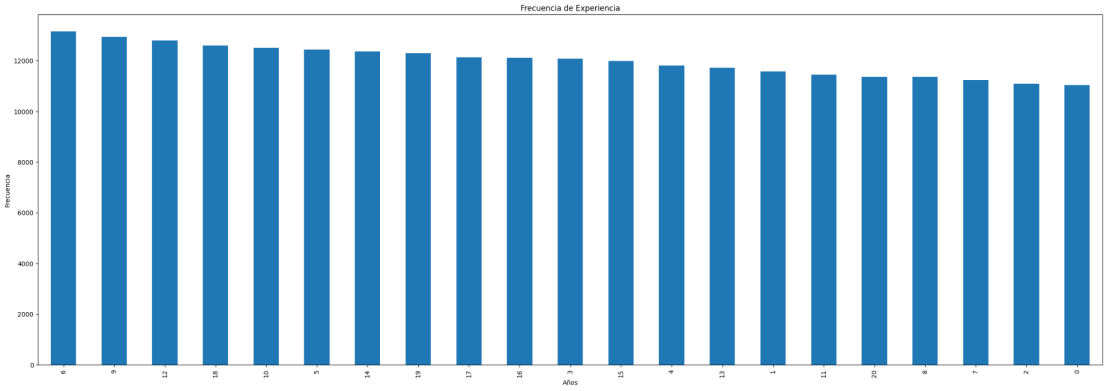
5. Profession: esta variable registra la profesión que ejerce cada uno de los usuarios; no obstante, por motivos del análisis y de la gran cantidad de valores únicos se decidió extraer las 5 categorías más frecuentes. En la siguiente imagen se pueden apreciar las cinco profesiones más comunes dentro de la base de datos.



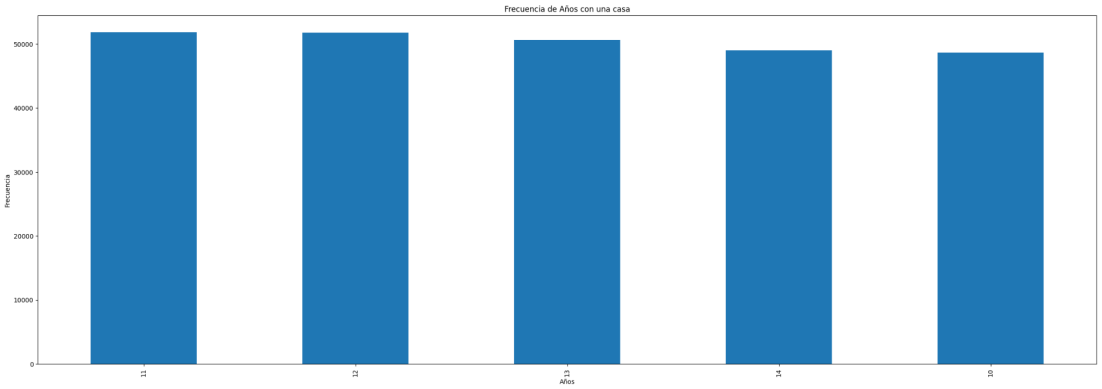
6. CITY: esta columna es la que registra más valores únicos de todas las analizadas; no obstante, se tomó la decisión de filtrar las cinco ciudades más frecuentes en la base de datos para delimitar el análisis. Estas se encuentran en la imagen a continuación insertada, cabe destacar que el archivo cuenta originalmente con 252 mil registros.



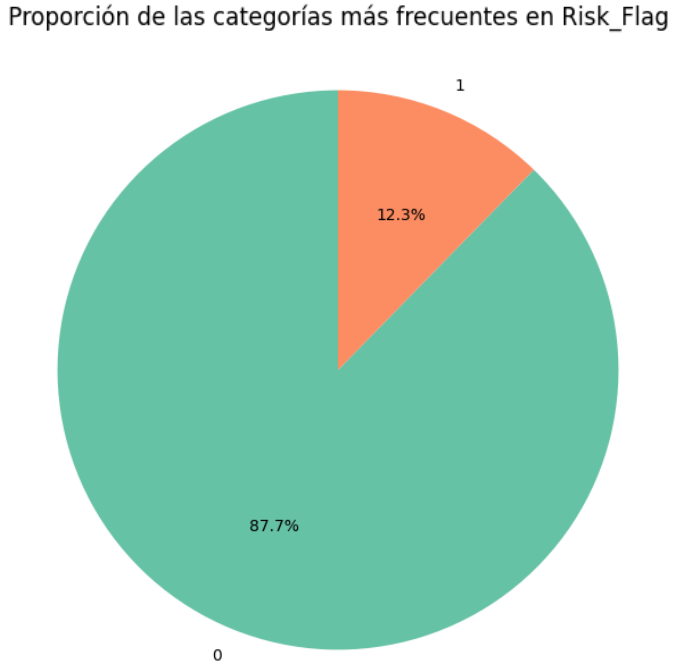
7. CURRENT_JOB_YRS: inicialmente esta variable era de tipo numérica; sin embargo, el marcado rango de sus valores permite tratarla como variable de tipo cadena. En un inicio se analizó la distribución de la variable ordenando los valores de menor a mayor, pero no se logró detectar nada relevante. Conservado el orden por frecuencia, se logró detectar que como moda el valor 6 y 0 como valor menos frecuente. Este hallazgo surgió al observar la siguiente imagen.



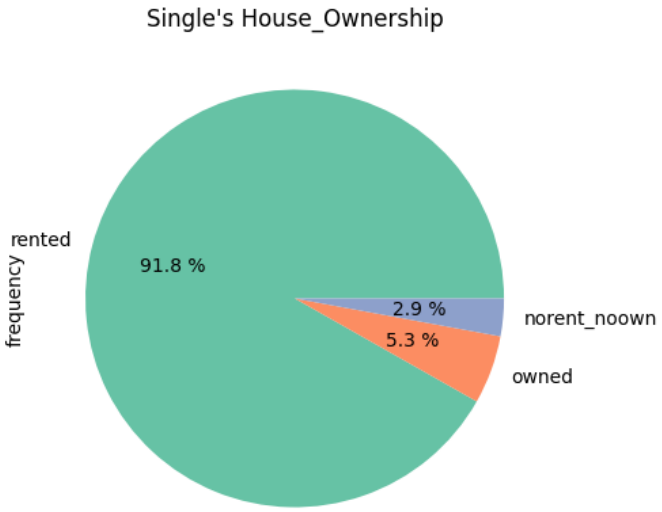
8. CURRENT_HOUSE_YRS: los valores de esta variable también a pesar de ser numéricos pueden ser analizados como categóricos. Al realizar la extracción de sus características se detectó que contaba con un total de cinco valores únicos que tienen un rango de 10 a 14 años. Analizando la frecuencia de los posibles valores se halló que el valor más frecuente dentro de los registros son 11 años; lo aquí descrito se aprecia en la siguiente gráfica.



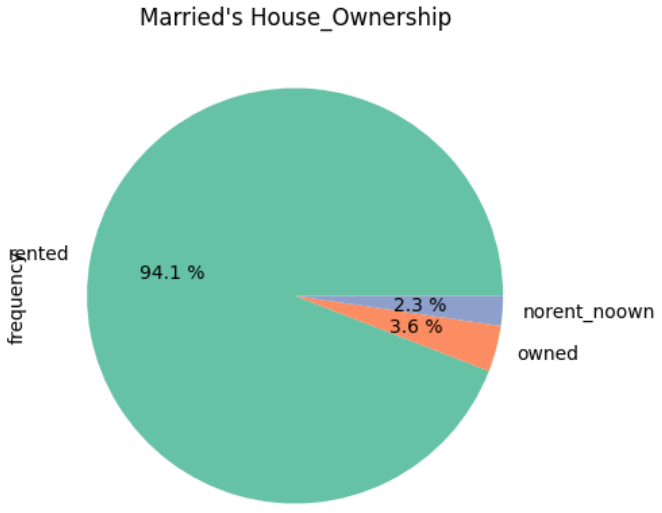
9. Risk_Flag: en un inicio esta columna también era detectada como de tipo numérico, pero debido al número de valores únicos fue analizada como variable categórica. Los valores registrados en esta columna son booleanos: 0 y 1. Mediante la extracción de categorías se descubrió que la diferencia entre el total de registros de cada categoría es bastante amplia: el 87.7% de los registros tienen un risk flag igual a 0 y el 12.3% restante fueron etiquetados como riesgosos. Esta proporción se aprecia en la siguiente imagen.



Una vez analizadas las variables de manera individual se realizaron dos mezclas para intentar tener hallazgos adicionales. La primera mezcla surgió al filtrar Single como estado civil y conocer la proporción de propiedad de casa en estos registros. En la siguiente gráfica se representa visualmente cómo el 91.8% de los solteros rentan casa, solo el 5.3% tiene casa propia y el 2.9% no renta ni tiene casa propia. Estas proporciones están representadas visualmente en la siguiente gráfica pastel.



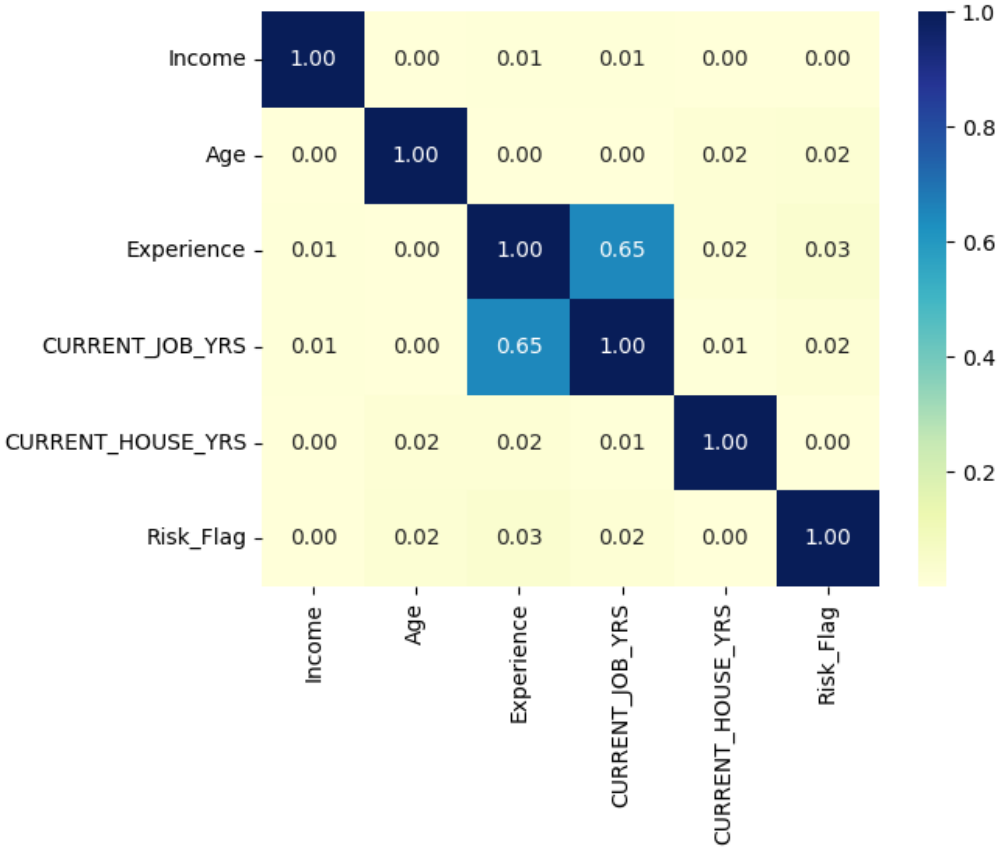
La siguiente mezcla de variables surgió al analizar la propiedad de casa ahora de las personas casadas. Se detectó que se conserva el orden de las proporciones con un 94.1% de personas que rentan, 3.6% de registros con casa propia y solo un 2.3% de registros que no rentan ni tienen casa propia. Estas proporciones se aprecian en la imagen insertada a continuación.



Modelos de regresión lineal simple.

Con la finalidad de analizar la correlación que existe entre todas las variables numéricas del dataframe (Income, Age, Experience, CURRENT_JOB_YRS, CURRENT_HOUSE_YRS y Risk_Flag) se visualizó un mapa de calor. Como la variable de

interés en el dataframe es Risk_Flag se seleccionaron sus tres correlaciones más altas (ignorando la diagonal) para la creación de esos modelos lineales simples. El mapa de calor realizado y evaluado es presentado a continuación.

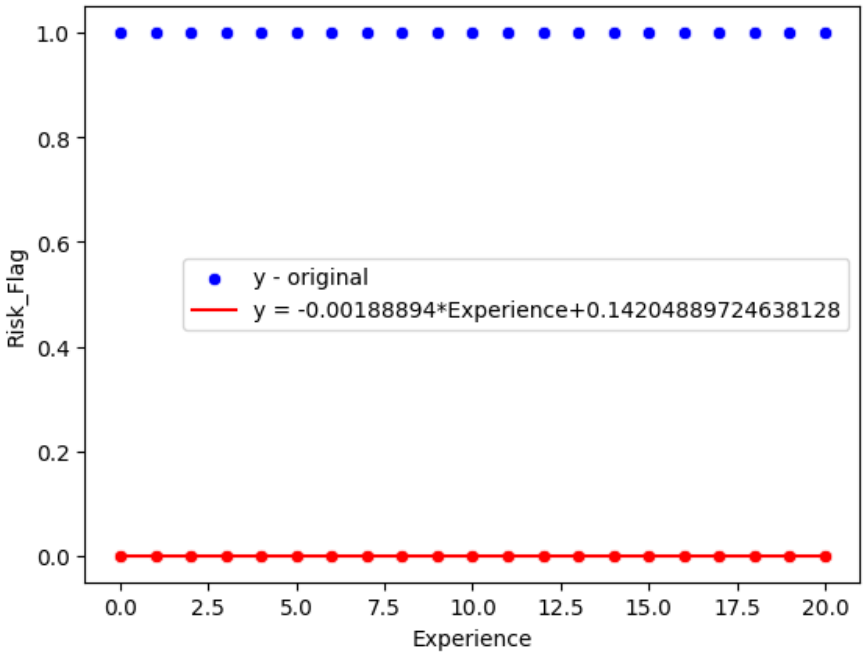


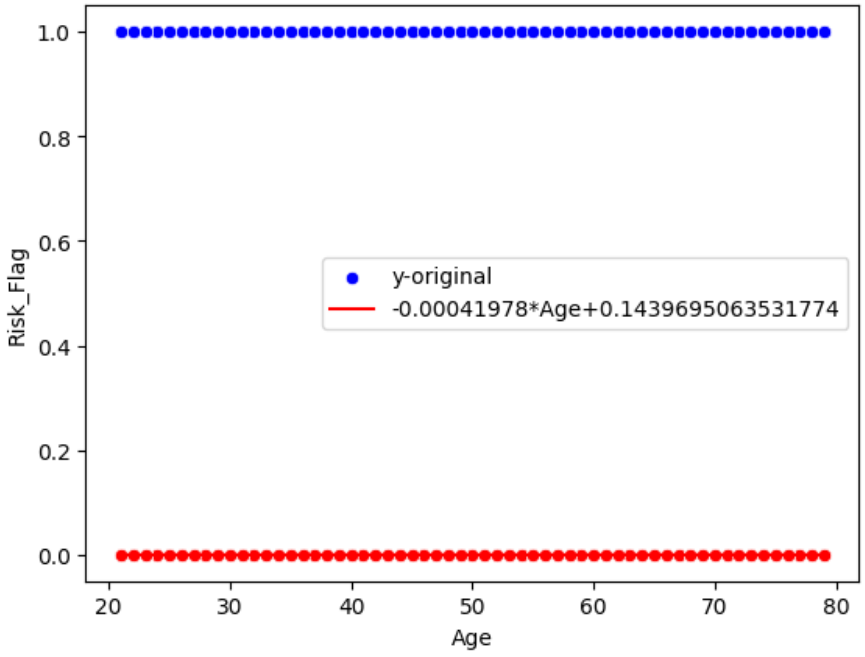
A pesar de los bajos coeficientes de correlación de Risk_Flag, se seleccionaron los tres más altos para el diseño de tres modelos de regresión lineal simple que intentan explicar su comportamiento. Los resultados obtenidos se presentan en la siguiente tabla.

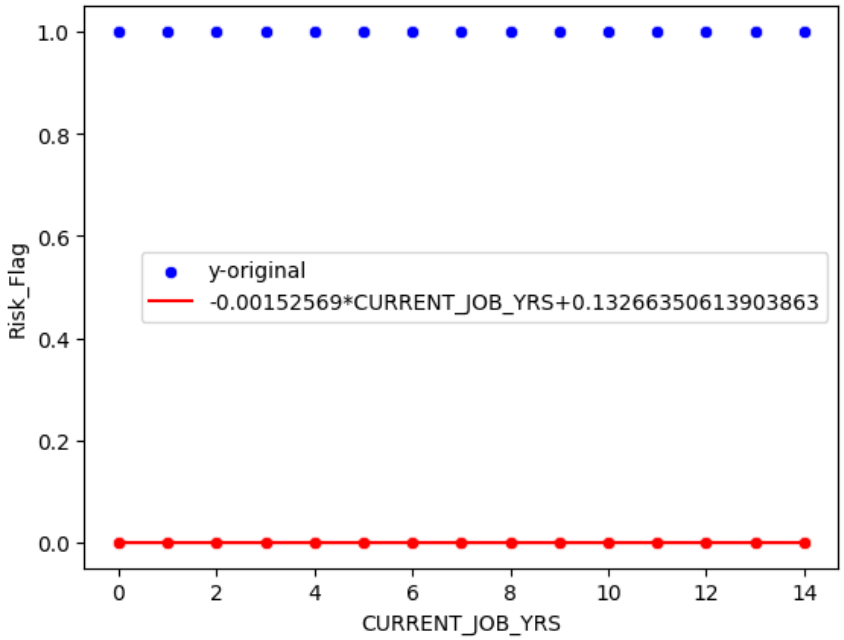
Tabla 1. Modelos de regresión lineal simple.

Dependientes	Independientes	Determinación	Correlacion
Risk_Flag	Expirience	0.001191810801	0.03452261289
Risk_Flag	Age	0.0004756445219	0.02180927605
Risk_Flag	CURRENT_JOB_YRS	0.0002870172024	0.01694158205

Como se puede observar los coeficientes son muy bajos, lo que es reflejo de la ineficiencia que ofrecen para describir el comportamiento de Risk_Flag. De hecho al graficarlos se puede apreciar de manera visual cómo los modelos no logran representar correctamente el comportamiento de los datos reales.


1.0
0.8
0.6
0.4
0.2
0.0
Risk_Flag
y - original
y = -0.00188894*Experience+0.14204889724638128
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Experience


1.0
0.8
0.6
0.4
0.2
0.0
Risk_Flag
y-original
-0.00041978*Age+0.1439695063531774
20
30
40
50
60
70
80
Age



Los modelos de regresión lineal simple en estos casos no son efectivos para predecir las variables dependientes a partir de las variables independientes. Estos resultados recomiendan la necesidad de explorar enfoques de modelado más avanzados, como la regresión lineal múltiple o no lineal, para comprender mejor la relación entre estas variables.

Modelos de regresión lineal múltiple.

Tabla 2. Valores de determinación y correlación de la variable dependiente "Risk_Flag".

n Variables	Dependiente	Variables independientes	Determinación	Correlación
5	Risk_Flag	('Income', 'Age', 'Experience', 'Current_Job_Yrs', 'Current_House_Yrs')	0.001745	0.041772
4	Risk_Flag	('Age', 'Experience', 'Current_Job_Yrs', 'Current_House_Yrs')	0.001736	0.041670
4	Risk_Flag	('Income', 'Age', 'Experience', 'Current_Job_Yrs')	0.001728	0.041571

Tabla 3. Valores de determinación y correlación de la variable dependiente "Current_House_Yrs".

n Variables	Dependientes	Variables independientes	Determinación	Correlacion
5	Current_House_Yrs	(Income', 'Age', 'Experience', 'Current_Job_Yrs', 'Risk_Flag')	0.0008855691515	0.02975851393
4	Current_House_Yrs	('Age', 'Experience', 'Current_Job_Yrs', 'Risk_Flag')	0.0008792634884	0.02965237745
4	Current_House_Yrs	('Income', 'Age', 'Experience', 'Current_Job_Yrs')	0.0008688150175	0.02947566823

Tabla 4. Valores de determinación y correlación de la variable dependiente "Current_Job_Yrs".

n Variables	Dependientes	Variables independientes	Determinación	Correlacion
5	Current_Job_Yrs	Income', 'Age', 'Experience', 'Risk_Flag', 'Current_House_Yrs'	0.4175375184	0.6461714311
4	Current_Job_Yrs	('Income', 'Experience', 'Risk_Flag', 'Current_House_Yrs')	0.4175293722	0.6461651276
4	Current_Job_Yrs	('Age', 'Experience', 'Risk_Flag', 'Current_House_Yrs')	0.4175291357	0.6461649447

Tabla 5. Valores de determinación y correlación de la variable dependiente "Experience".

n Variables	Dependientes	Variables independientes	Determinación	Correlacion
5	Experience	Income', 'Age', 'Risk_Flag', 'Current_Job_Yrs', 'Current_House_Yrs'	0.4182563842	0.646727442
4	Experience	('Age', 'Risk_Flag', 'Current_Job_Yrs', 'Current_House_Yrs')	0.4182530114	0.6467248344
4	Experience	('Income', 'Risk_Flag', 'Current_Job_Yrs', 'Current_House_Yrs')	0.4182490707	0.6467217877

Tabla 6. Valores de determinación y correlación de la variable dependiente "Age".

n Variables	Dependientes	Variables independientes	Determinación	Correlacion
5	Age	ome', 'Risk_Flag', 'Experience', 'Current_Job_Yrs', 'Current_House_Y	0.000901647666	0.03002744854
4	Age	('Risk_Flag', 'Experience', 'Current_Job_Yrs', 'Current_House_Yrs')	0.0009010506946	0.03001750647
4	Age	('Income', 'Risk_Flag', 'Current_Job_Yrs', 'Current_House_Yrs')	0.0008890871648	0.0298175647

Tabla 7. Valores de determinación y correlación de la variable dependiente "Income"

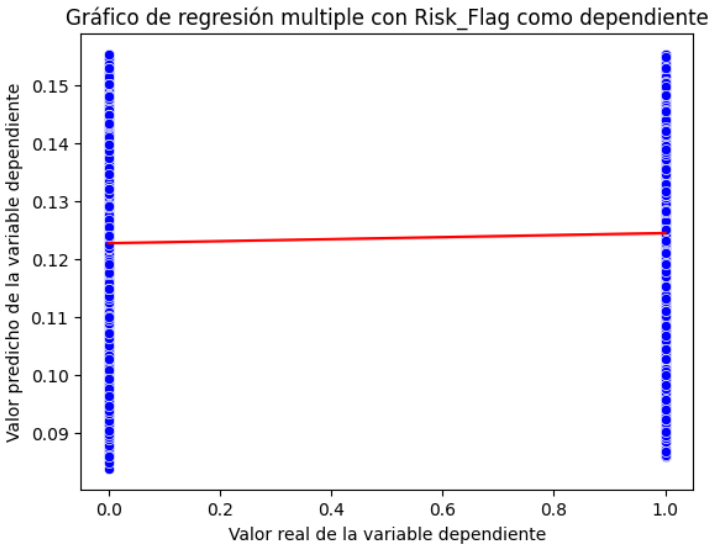
n Variables	Dependientes	Variables independientes	Determinación	Correlacion
5	Income	Risk_Flag', 'Current_House_Yrs', 'Current_Job_Yrs', 'Experience', 'Age'	0.00007086061456	0.008417874706
4	Income	('Risk_Flag', 'Current_House_Yrs', 'Current_Job_Yrs', 'Experience')	0.00007026314673	0.008382311538
4	Income	('Risk_Flag', 'Current_House_Yrs', 'Current_Job_Yrs', 'Age')	0.00006506316486	0.008066174115

En los datos presentados, se ha realizado un análisis exhaustivo de modelos de regresión lineal múltiples, para explorar la relación entre las variables cuantitativas: "Income", "Age", "Experience", "CURRENT_JOB_YRS", "CURRENT_HOUSE_YRS" y "Risk_Flag." Sin embargo, se ha encontrado que ninguno de los modelos ha resultado satisfactorio en la explicación o predicción de las variables dependientes.

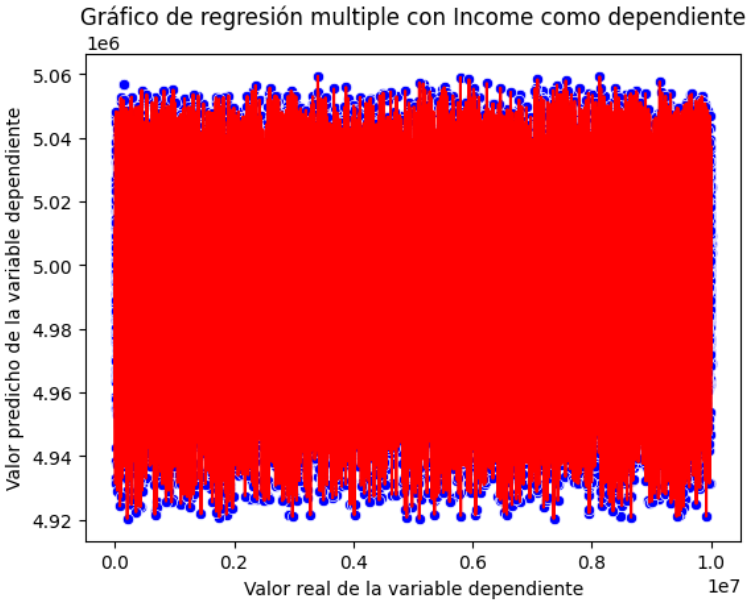
El mejor resultado obtenido se refiere a un modelo de regresión múltiple que incluye la variable "Experience" junto con cinco variables independientes. A pesar de este mejor escenario, el coeficiente de determinación (R^2) alcanza 0.418, lo que indica que solo se explica el 41.8% de la variabilidad en "Experience." La correlación (R) también se mantiene en 0.647, lo que sugiere una correlación moderada pero no excepcional. Esto significa que aún existe una cantidad sustancial de variabilidad en "Experience" que no se ha capturado.

En otros casos, los modelos de regresión lineal múltiple no han tenido éxito en explicar las relaciones entre las variables. Los valores de R^2 y R son bajos o nulos, lo que indica que no se logra una buena correspondencia entre las variables independientes y dependientes.

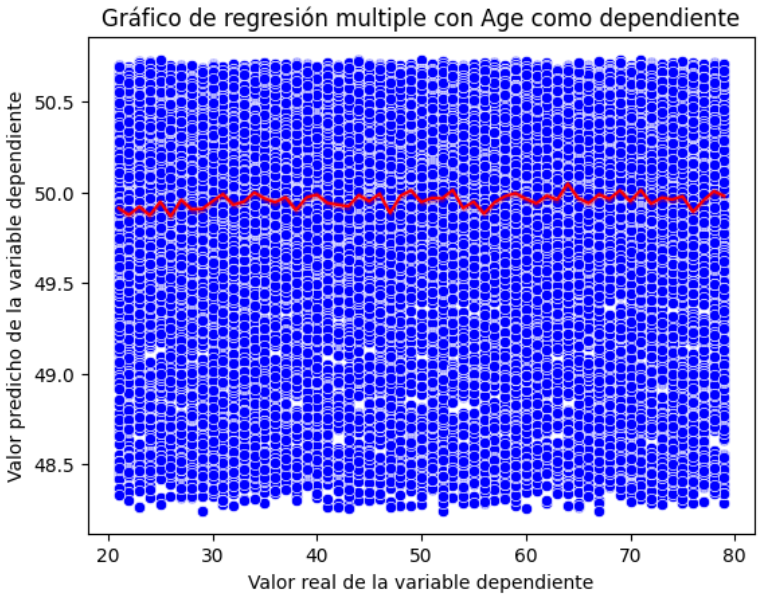
Dado que ninguno de los modelos lineales, ni siquiera el mejor caso, ha proporcionado resultados satisfactorios, se plantea la necesidad de explorar enfoques de modelado no lineal para comprender mejor la relación entre estas variables.



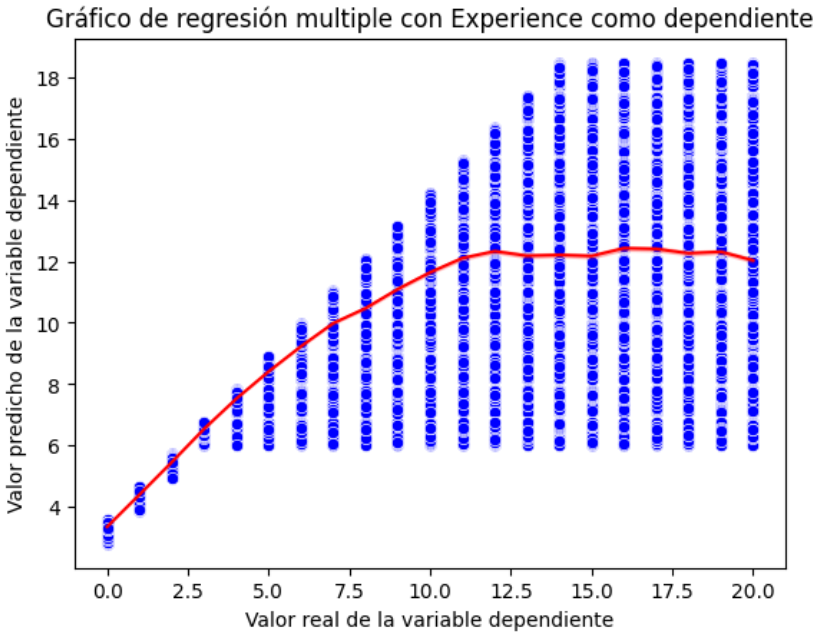
Dentro de la gráfica presente, se puede observar que los datos representados por el color azul no coinciden con la predicción establecida que es representada por la línea roja, esto es acorde a los resultados de correlación y determinación obtenidos.



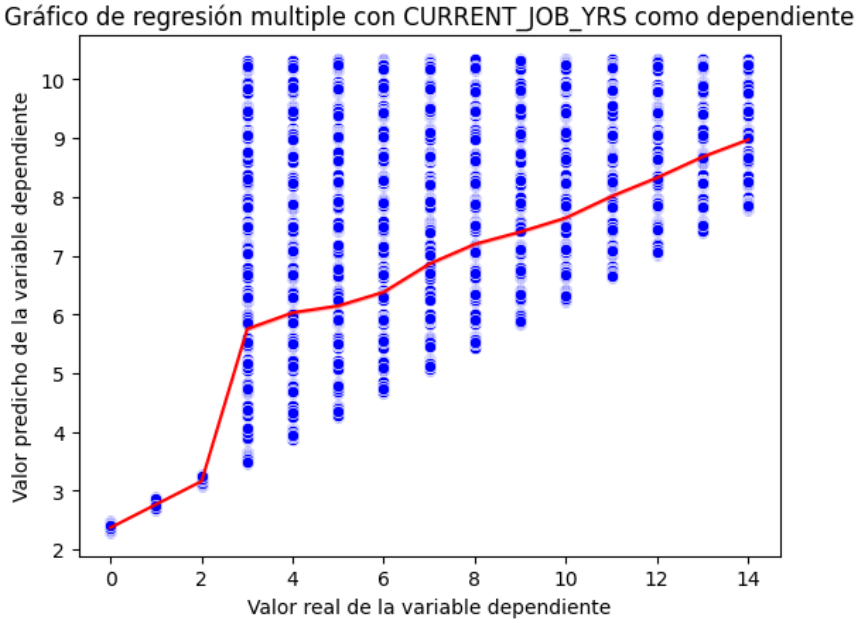
Dentro de la gráfica de Income se puede apreciar que igualmente está acorde a los resultados de correlación y determinación, pero a diferencia de la gráfica anterior, la variabilidad es tan grande que las predicciones realizadas no logran formar una línea recta como es común en este tipo de regresión.



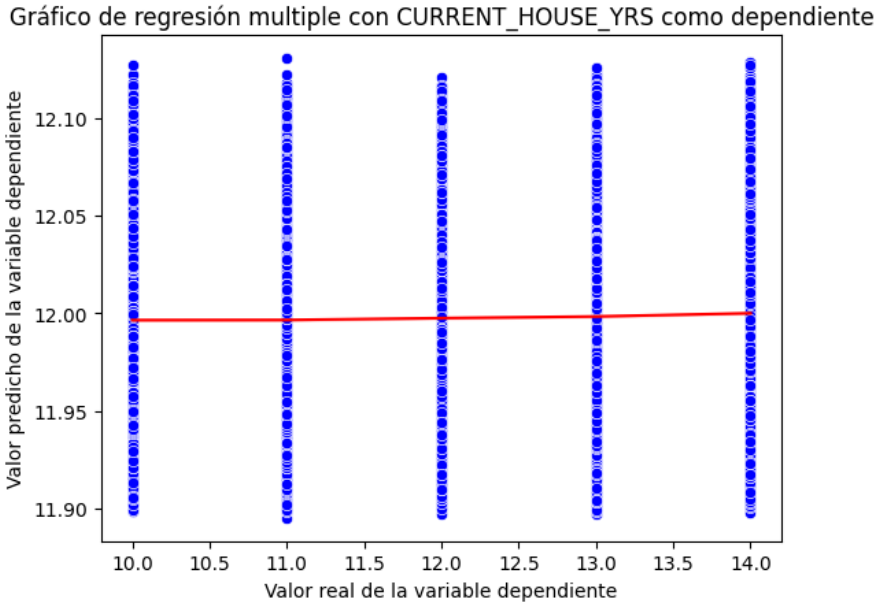
Dentro de la gráfica de cambio se puede observar que hay demasiada variabilidad en los datos, esto provoca que la predicción en este caso no sea del todo lineal como lo es representado en la imagen, esto es reforzado al ver los resultados de la correlación como de la determinación.



Los resultados de la gráfica presente muestra una distribución variable de datos, esto afecta en el cálculo de la variación como puede ser presenciado, normalmente este debería ser lineal, pero la variabilidad es tal que la predicción parece ser logarítmica debido a su curvatura que es cóncava hacia abajo, estos serian resultados aceptables en caso de realmente ser logarítmica, pero no en este caso.



La gráfica que toma en cuenta CURRENT_JOB_YRS igualmente es irregular debido a la distribución de los datos. Debido a los resultados que se obtuvieron respecto a la correlación y la determinación, se puede ver que llega a ser bastante comprensible que las predicciones no sigan la linealidad deseada.



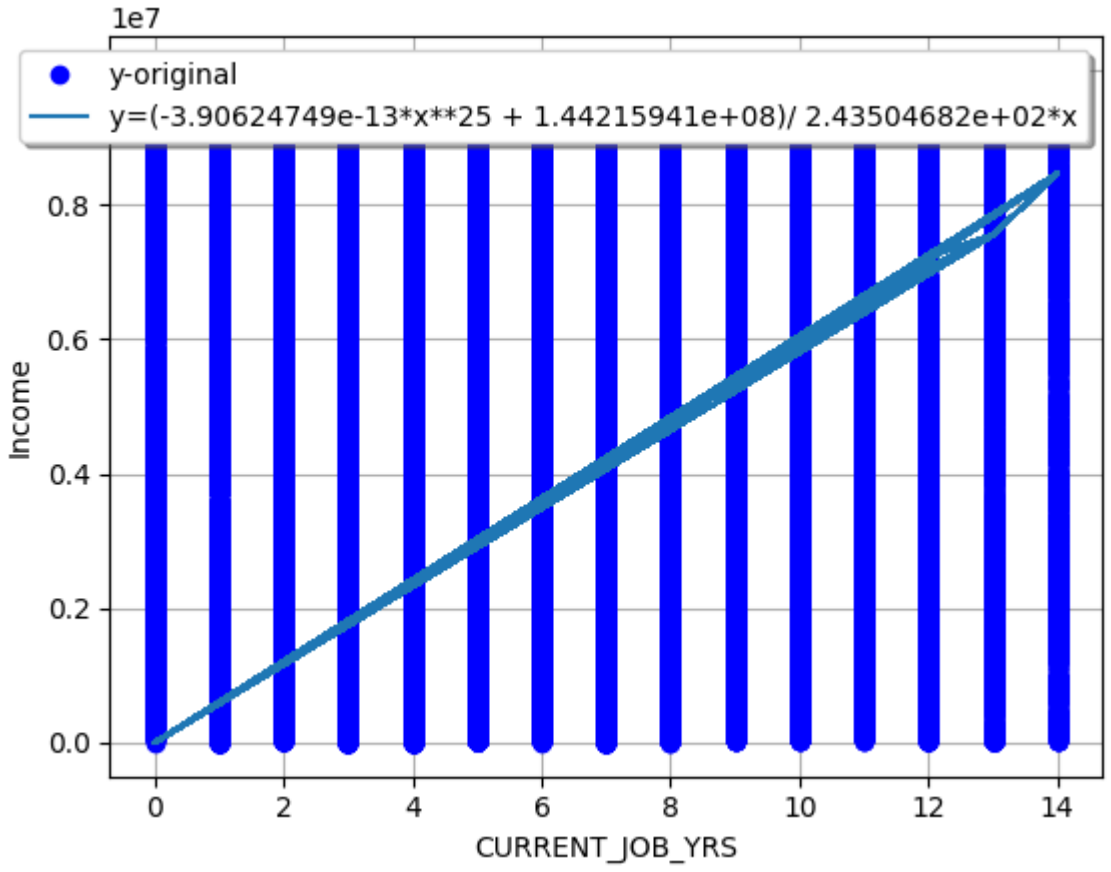
En el gráfico de CURRENT_HOUSE_YRS se puede presenciar una predicción con una mayor linealidad, esto puede ser algo positivo de no ser porque los datos están

distribuidos de una manera demasiado variable, esto llega a ser confirmado debido a los resultados obtenidos en cuanto a la correlación y a la determinación.

Modelos de regresión no lineal.

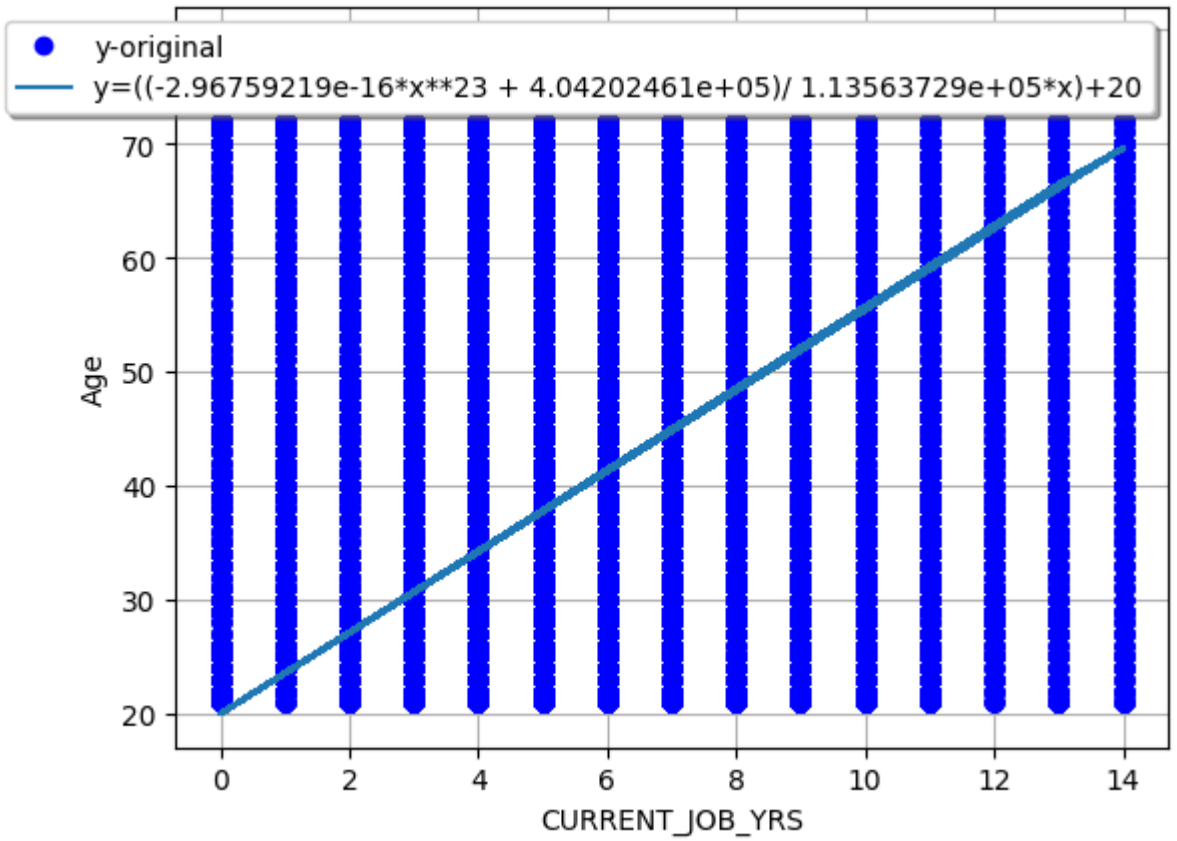
Se realizó un total de seis modelos no lineales donde cada variable numérica tuvo la oportunidad de ser la variable dependiente buscando mejorar los coeficientes obtenidos por los modelos lineales (simples y múltiples). Para ello se visualizó una matriz de diagramas de dispersión considerando todas las variables numéricas con la finalidad de proyectar la gráfica de aquella función no lineal que abarque el mayor área posible para disminuir la suma del error sobre la dispersión de las variables.

Para el primer modelo se utilizó Income como variable dependiente de CURRENT_JOB_YRS. Se utilizó una función cociente entre polinomios que se estructura como y = (-3.90624749e-13*x**2 + 1.44215941e+08)/ 2.43504682e+02*x. La siguiente gráfica presenta como puntos la dispersión real entre Income y CURRENT_JOB_YRS, y como recta el comportamiento del modelo no lineal propuesto.



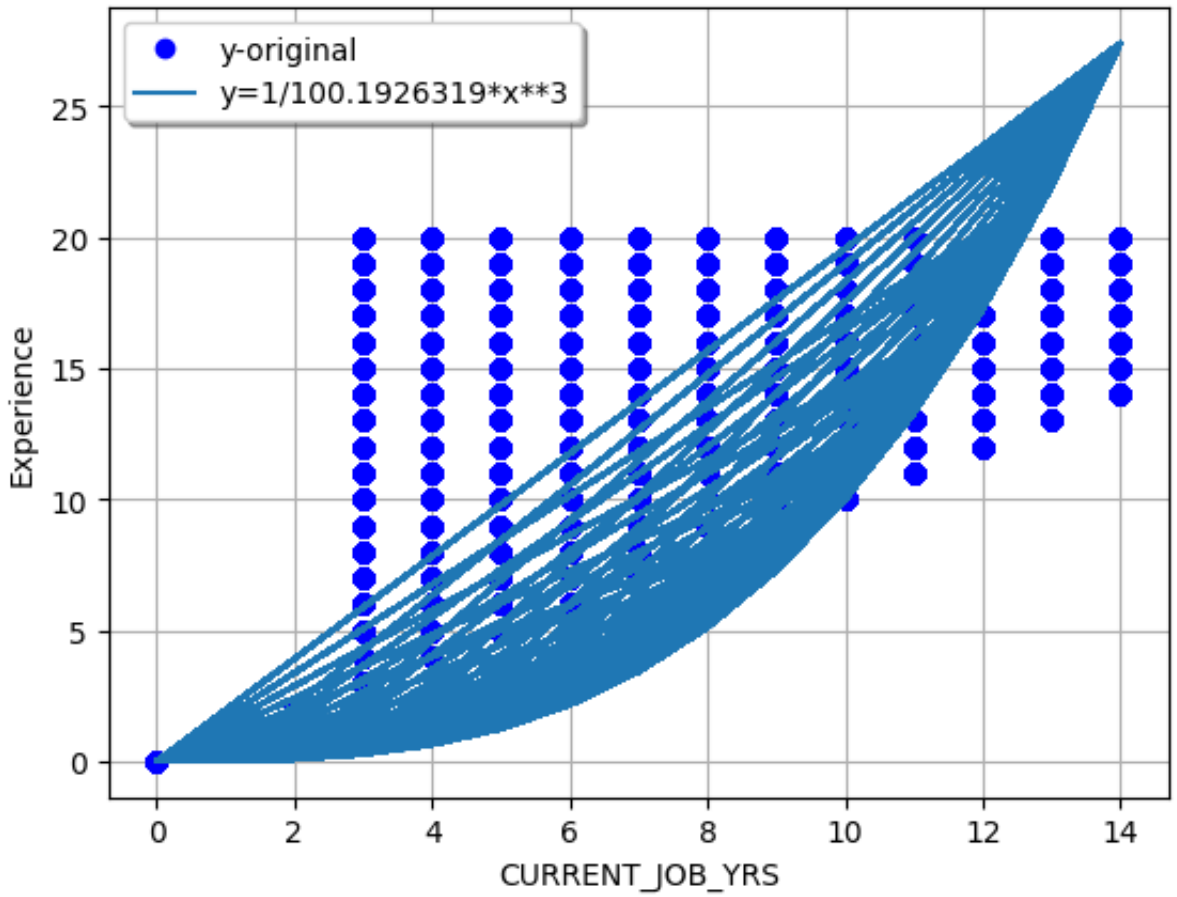
A pesar de no apreciarse visualmente el área de cobertura que ofrece el modelo sobre los datos reales en la gráfica anterior, se obtuvo un coeficiente de determinación de -0.7395 y un coeficiente de correlación de 0.8599, siendo el mejor modelo encontrado para explicar Income comparado con los modelos lineales antes mencionados.

En el segundo modelo de regresión no lineal se usó la misma variable independiente para explicar el comportamiento de Age con la siguiente función cociente entre polinomios: y=((-2.96759219e-16*x**23 + 4.04202461e+05) / 1.13563729e+05*x)+20. A continuación se inserta la gráfica que representa como puntos de dispersión los datos reales y como gráfica de línea el modelo resultante de la ecuación propuesta.



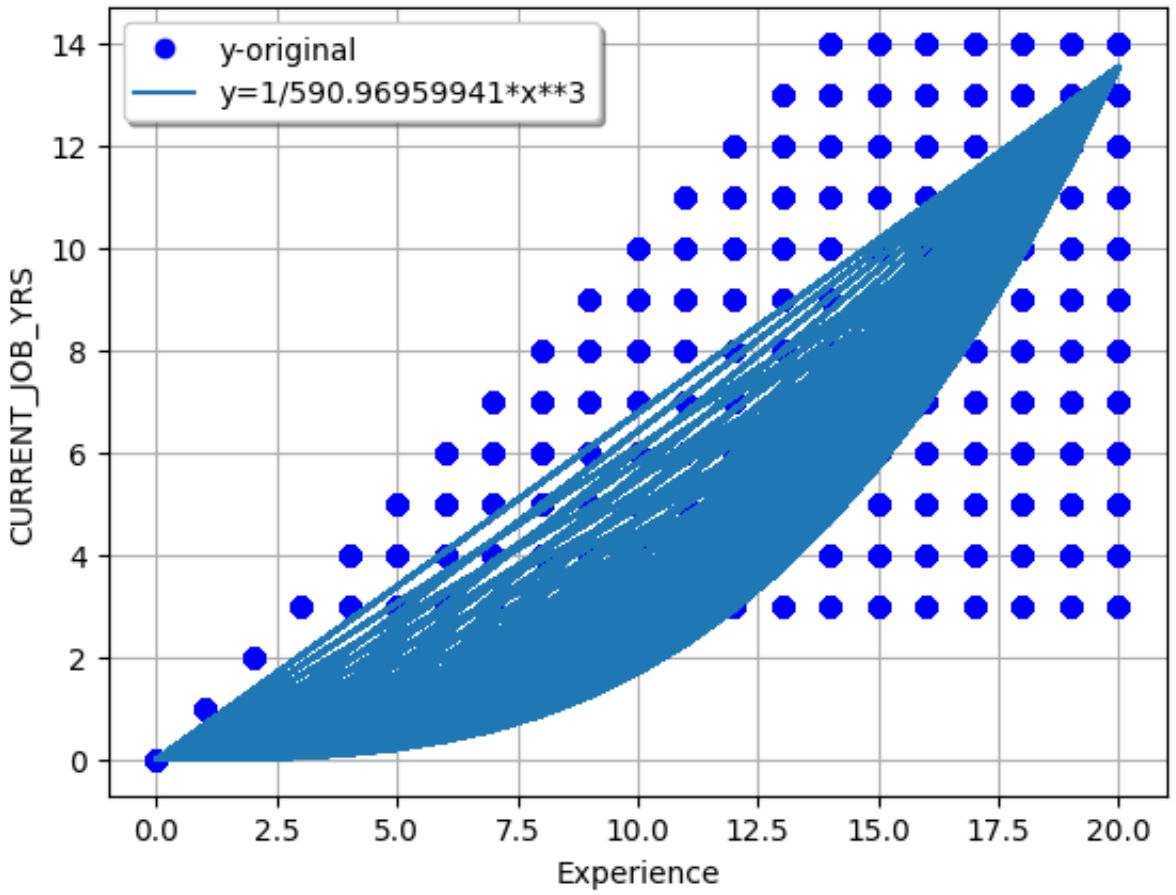
A pesar de no apreciarse visualmente el área de cobertura que ofrece el modelo sobre los datos reales en la gráfica anterior, se obtuvo un coeficiente de determinación de -0.76396 y un coeficiente de correlación de 0.874, siendo el mejor modelo encontrado para explicar Age comparado con los modelos lineales mencionados anteriormente.

En el tercer modelo de regresión no lineal se usó la misma variable independiente que en los dos modelos anteriores para explicar el comportamiento de Experience con el inverso de una función cúbica representado por la siguiente ecuación: y=1/100.1926319*x**3. A continuación se inserta la gráfica que representa como puntos de dispersión los datos reales y como gráfica de línea el modelo resultante por la ecuación propuesta.



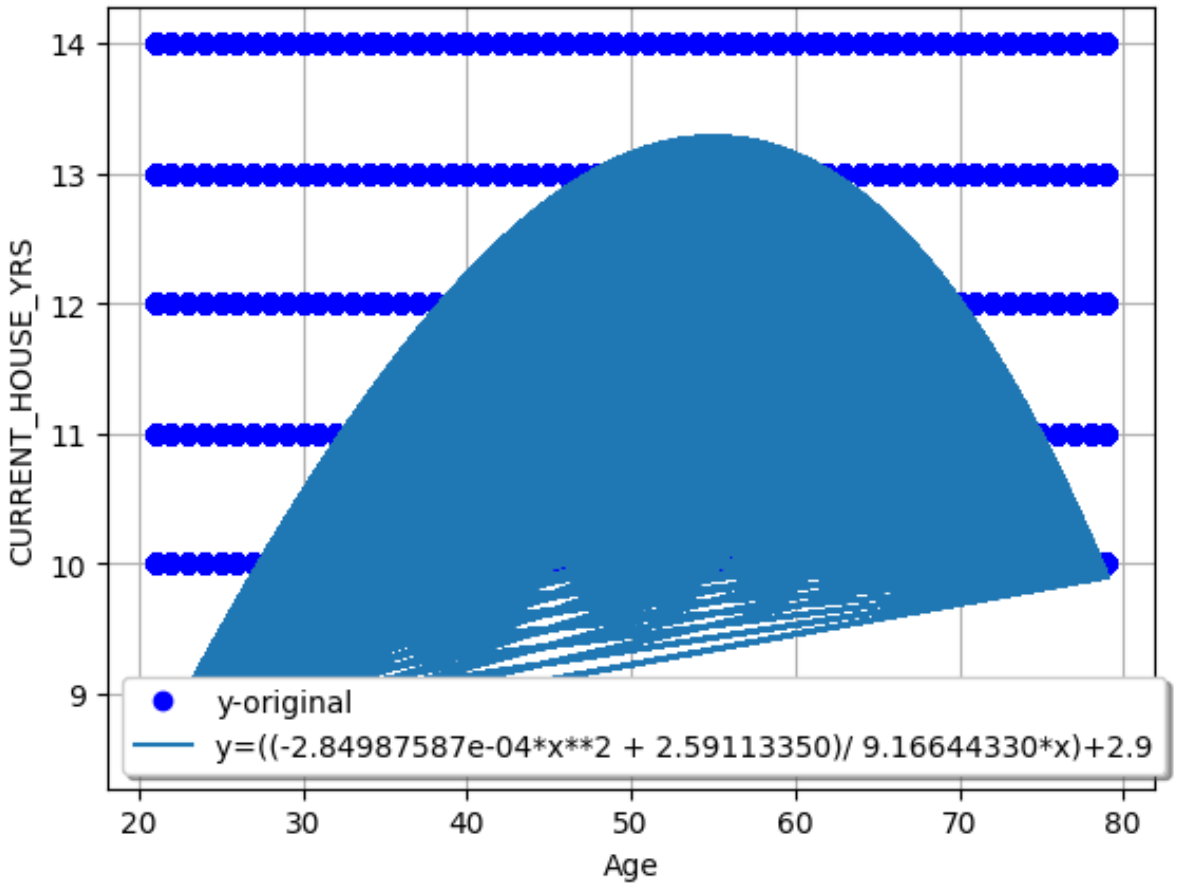
En este caso se logra apreciar una mayor cobertura del modelo sobre los datos reales comparado con las gráficas anteriores y, a pesar de poder detectar datos que no son considerados dentro del modelo (lo que aumenta el margen de error), se obtuvo un coeficiente de determinación de -0.7869 y un coeficiente de correlación de 0.8871 ofreciendo mejores resultados que los obtenidos por los modelos lineales que de por sí ya eran aceptables.

El cuarto modelo de regresión no lineal invierte las columnas del modelo anterior como variables dependiente e independiente; es decir, ahora se modela CURRENT_JOB_YRS como variable dependiente de Experience con el siguiente inverso de una función cúbica: y=1/590.96959941*x**3. La siguiente gráfica es una representación visual donde los puntos de dispersión son los datos reales y la gráfica de línea el modelo resultante por la ecuación propuesta.



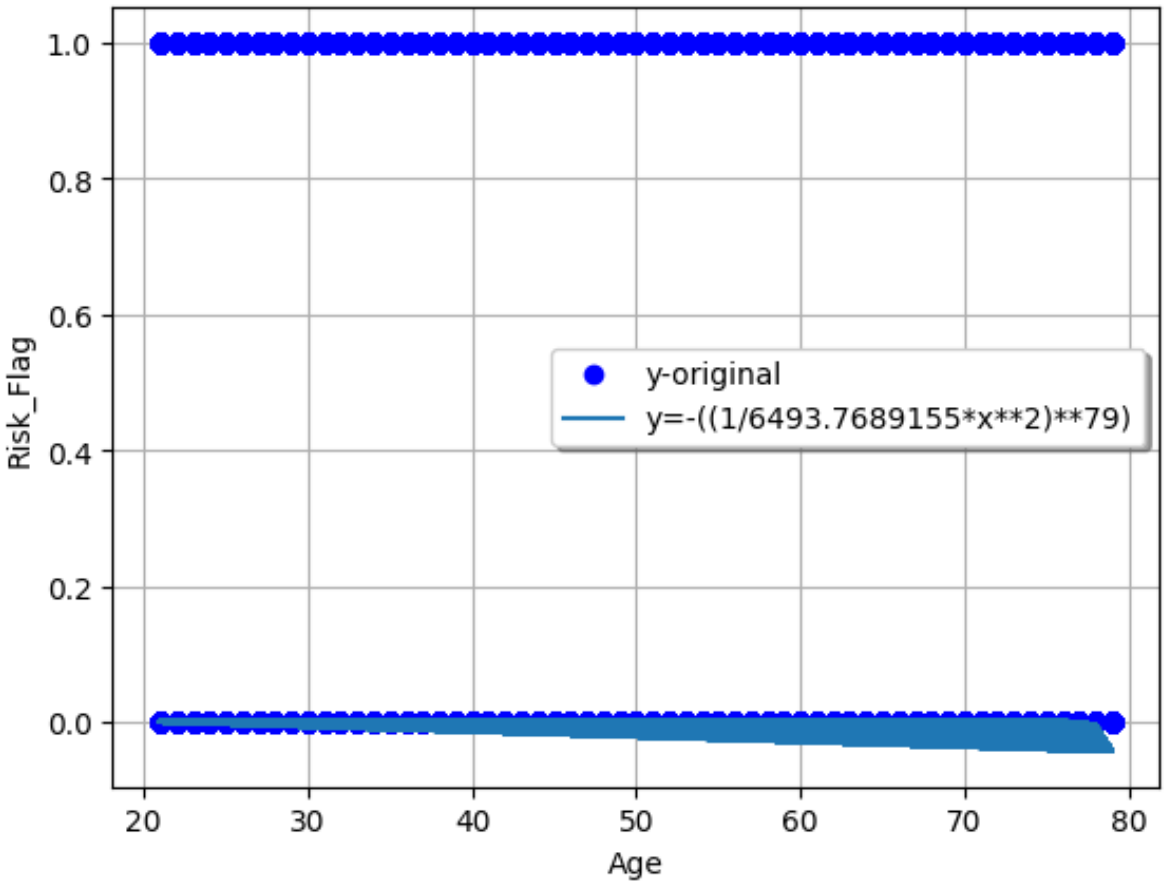
En este caso también se logra apreciar una buena cobertura del modelo sobre los datos reales y, a pesar de poder detectar datos que no son considerados dentro del modelo (lo que aumenta el margen de error), se obtuvo un coeficiente de determinación de -0.7966 y un coeficiente de correlación de 0.8925 ofreciendo mejores resultados que los obtenidos por los modelos lineales que de por sí ya también eran aceptables.

El quinto modelo de regresión no lineal considera como variable independiente a Age y como variable dependiente CURRENT_HOUSE_YRS y es explicado por la siguiente función cociente entre polinomios: y=((-2.84987587e-04*x**2 + 2.59113350)/ 9.16644330*x)+2.9. La gráfica a continuación insertada es una representación visual donde los puntos de dispersión son los datos reales y las líneas el modelo resultante por la ecuación propuesta.



Si bien el área de cobertura del modelo no cubre la totalidad de los datos reales, este modelo obtuvo un coeficiente de determinación de -0.98457 y un coeficiente de correlación de 0.9922, siendo el mejor modelo encontrado para explicar una variable numérica dentro de las columnas disponibles del conjunto de datos. Debido al inesperado hallazgo de un modelo tan eficiente se insertó la columna de predicciones generadas por este modelo para la variable dependiente y se calculó la diferencia entre los valores reales y dichas predicciones. El promedio del error de estimación es de 0.2084 años lo que confirma la precisión de predicción del modelo de regresión no lineal propuesto.

El sexto y último modelo considera a Risk_Flag como variable dependiente de Age bajo el siguiente reflejo de una función cuadrática inversa: y=-((1/6493.7689155*x**2)**79). La presencia de un signo negativo al inicio de la función sirve para reflejar el modelo con respecto al eje x, esto con la finalidad de tener una predicción exacta para cero y poder condicionar todos aquellos valores de la variable dependiente menores a 0 con el valor 1 al tratarse de una variable booleana. La siguiente gráfica es una representación visual donde los puntos de dispersión son los datos reales y las líneas el modelo resultante por la ecuación propuesta.



Se decidió conservar el modelo a una altura que favorece las predicciones de datos reales iguales a 0; ya que al ser la moda de la variable se reduce el margen de error general del modelo y se obtienen los coeficientes de determinación y correlación más altos posibles, siendo estos -0.14247 y 0.37745 respectivamente. Se reconoce que dichos coeficientes reflejan la deficiencia del modelo, por lo que se concluye que para predecir esta variable se necesita algo más complejo que un modelo de regresión no lineal.

Conclusiones.

Después del diseño de los modelos expuestos (lineales y no lineales) a lo largo de este reporte se realizó una comparación de los coeficientes obtenidos por cada uno de ellos para las seis variables numéricas analizadas. Dicha comparación es representada por el siguiente mapa de calor.

		Determinación	Correlación
Income	RLS_y1		
	RLM_y1	0.00007	0.00842
	RNL_y1	0.73947	0.85992
Age	RLS_y2		
	RLM_y2	0.00090	0.03003
	RNL_y2	0.76396	0.87405
Experience	RLS_y3		
	RLM_y3	0.41826	0.64673
	RNL_y3	0.78693	0.88709
CURRENT_JOB_YRS	RLS_y4		
	RLM_y4	0.41754	0.64617
	RNL_y4	0.79659	0.89252
CURRENT_HOUSE_YRS	RLS_y5		
	RLM_y5	0.00089	0.02976
	RNL_y5	0.98457	0.99226
Risk_Flag	RLS_y6	0.00119	0.03452
	RLM_y6	0.00175	0.04177
	RNL_y6	0.14247	0.37745

El principal hallazgo de este ejercicio fue la mejora de resultados al implementar modelos no lineales sobre los modelos lineales, tanto simples como múltiples, entendiendo como resultados a los coeficientes de determinación y correlación de los modelos. Esto se debe a que en todos los casos un modelo lineal no era suficiente para explicar el comportamiento de las variables; por lo que fue necesario el uso de modelos no lineales para aportar flexibilidad al modelo y mejorar su eficiencia al reducir el margen de error entre los datos reales y las predicciones.

Referencias.

Reflecting Functions: Examples. (s. f.). [Vídeo]. Khan Academy. https://www.khanacademy.org/math/algebra2/x2ec2f6f830c9fb89:transformations/x2ec2f6f830c9fb89:reflect/v/reflecting-functions-examples#:~:text=We%20can%20reflect%20the%20graph,applied%20to%20solve%20various%20problems.