Reporte técnico: Modelo de riesgo de crédito

Amilder Stewin Ospina Tobón

Nicolás Pérez Vásquez

John Stiven Mejía Lopera

Juan Paulo Cepeda Zúñiga

Docente

Juan David Ospina Arango

Fundamentos de analítica

Universidad Nacional de Colombia

Sede Medellín

Facultad de Minas

2023

Contenido

Definición del problema

La base de datos Credit Risk Dataset (Tse, n.d.) muestra un caso donde se simula la participación de personas dentro del buró de crédito.

Dicha base de datos contiene la siguiente información recolectada, separada en columnas:

- Edad (años): person_age
- Ingreso anual ($): person_income
- Tipo de vivienda (renta, hipoteca, propia u otra): person_home_ownership
- Duración de empleo (años): person_emp_length
- Intención de préstamo (educativa, médica, empresarial, personal o consolidada): loan_intent
- Grado de préstamo (A, B, C, D o E): loan_grade
- Cantidad de préstamo ($): loan_amnt
- Interés (%): loan_int_rate
- Estado del préstamo (no incumplido o incumplido): loan_status
- Porcentaje de ingreso (%): loan_percent_income
- Histórico de incumplimiento (sí o no): cb_person_default_on_file
- Duración de historial crediticio (años): cb_person_cred_hist_length

Sin embargo, la base de datos muestra valores no adecuados o incoherentes entre sí, y bajo el contexto en que están siendo analizados. Es por ello que se busca realizar un modelo de riesgo crediticio basado en los datos presentados, y con él, poder tomar decisiones dentro del buró de crédito que se está manejando.

Para ello, es necesario filtrar adecuadamente las variables, los datos y poder realizar un análisis adecuado de estos. Se presentará entonces una página web que muestre el comportamiento del modelo propuesto.

Metodología

Para el desarrollo del análisis, se utilizan los datos suministrados por Tse (n.d.), de la plataforma de ciencia de datos, Kaggle. La base de datos utiliza las variables descritas anteriormente.

Para el análisis y manejo de los datos, se utiliza Python como lenguaje de programación que permite desarrollar lo propuesto.

De igual manera, se trabaja en un scorecard, el cual será utilizado para mostrar el modelo propuesto, mediante una página web.

Análisis descriptivo e hipótesis

En primer lugar, se debe realizar una limpieza de datos atípicos, o lo que se le llamará un pre-procesamiento de los datos, el cual seguirá los siguientes pasos:

1. Se eliminan los valores faltantes de la variable person_emp_length, correspondiente a la duración de empleo de la persona, es decir, aquellos registros cuyo valor sea nulo. Esto, a primera vista, no es viable, puesto que, al realizarlo, siguen existiendo 3048 datos nulos para esta columna. Por lo tanto, se debe revisar la significancia de la variable en el modelo propuesto.

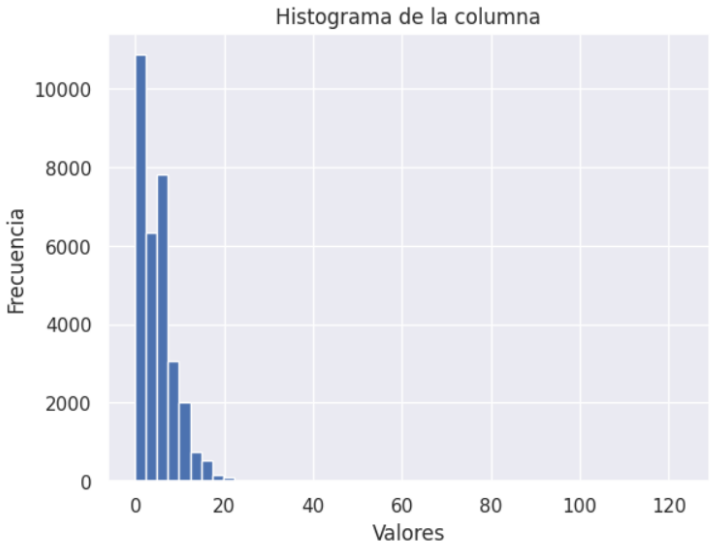

Figura 1. Histograma de la columna person_emp_length

La existencia de estos datos nulos indica que no es viable eliminar los registros, ya que representan casi el 10% de los datos. Por ende, se debe remuestrear o imputar estos valores.

En principio, se observa la viabilidad de realizar esta imputación de los datos por medio de un modelo KNN, pero esto no llegaría a ser muy viable en el modelo debido a la precisión que podría tener en el contexto.

2. Se revisa ahora la variable loan_int_rate, correspondiente al interés asociado, buscando eliminar también valores nulos.

```
df2 = df.copy()
filas_nulas = df2['loan_int_rate'].isnull()
valores_aleatorios = np.random.normal(11, 3.24, filas_nulas.sum())
df2['loan_int_rate'].fillna(pd.Series(valores_aleatorios, index=df.index[filas_nulas]), inplace=True)

print(len(df2))
```

Figura 2. Eliminación primaria de valores nulos en columna loan_int_rate

De acuerdo con ello, se construye un histograma de los datos:

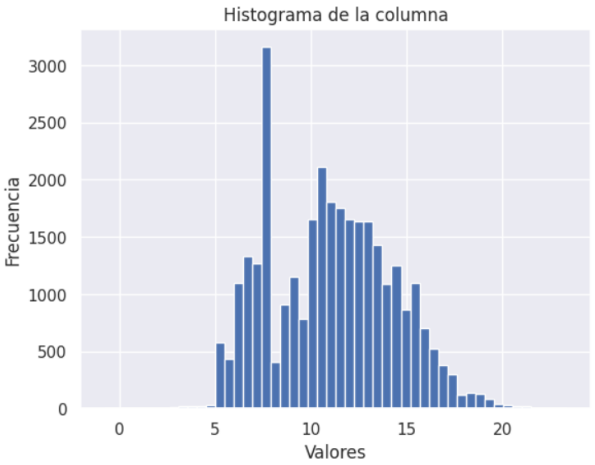

Figura 3. Histograma para la columna loan_int_rate

Y, antes de proseguir, se realiza una eliminación de datos atípicos según el contexto, como es la edad de las personas cuando es mayor a 100 años, y los datos donde las personas tengan una edad mayor al tiempo que han sido empleadas.

```
[12] #se toma las edades menores a 100
     mascara_age = df2['person_age'] <= 100
     df3 = df2[mascara_age]
```

```
[13] #crear mascara que tome del dataframe solo las personas que su edad sea mayor que el tiempo que haya sido empleado
     mascara = df3['person_age'] > df3['person_emp_length']
     df4 = df3[mascara]
     df4.head()
```

Figura 4. Eliminación de datos atípicos en las variables person_age y person_emp_length

Se presenta un gráfico de correlaciones, el cual se presenta con valores entre -1 y 1, donde 0 indicará correlación nula, 1 indica una correlación directa positiva y -1 indica una correlación directa negativa.

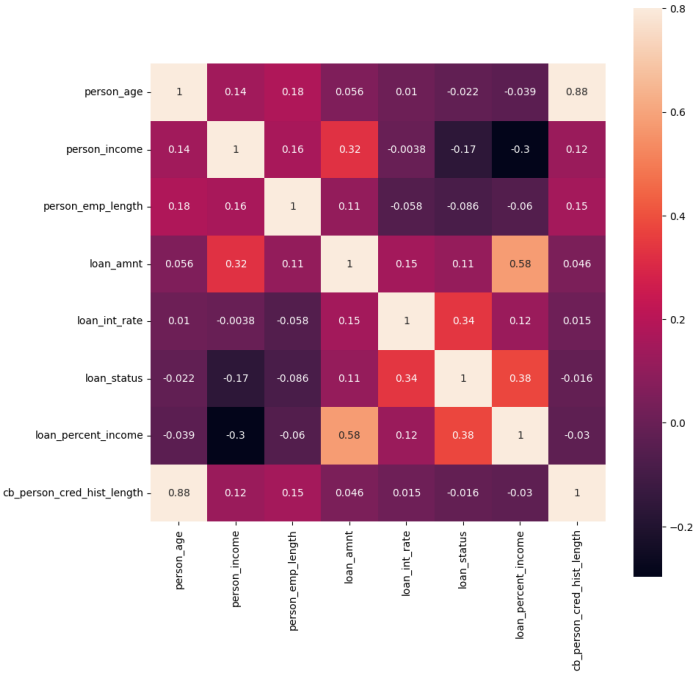

Figura 5. Gráfico de correlación entre variables de la base de datos

Se observa que las variables con mayor correlación parecen ser la edad de la persona y la duración de historial crediticio. Las otras variables parecieran no estar altamente relacionadas, así exista un grado de correlación.

A partir de esta filtración, se obtendrían los siguientes datos:

	person_age	person_income	person_emp_length	loan_amnt	loan_int_rate	loan_status	loan_percent_income	cb_person_cred_hist_length
count	31679.000000	3.167900e+04	31679.000000	31679.000000	31679.000000	31679.000000	31679.000000	31679.000000
mean	27.730673	6.649010e+04	4.782064	9659.962436	11.031626	0.215442	0.169610	5.809211
std	6.213427	5.276879e+04	4.034948	6334.360554	3.232160	0.411135	0.106269	4.059710
min	20.000000	4.000000e+03	0.000000	500.000000	-0.785578	0.000000	0.000000	2.000000
25%	23.000000	3.936600e+04	2.000000	5000.000000	7.900000	0.000000	0.090000	3.000000
50%	26.000000	5.600000e+04	4.000000	8000.000000	10.990000	0.000000	0.150000	4.000000
75%	30.000000	8.000000e+04	7.000000	12500.000000	13.470000	0.000000	0.230000	8.000000
max	94.000000	2.039784e+06	41.000000	35000.000000	23.397272	1.000000	0.830000	30.000000

Figura 6. Resumen de los datos filtrados

Sin embargo, siguen existiendo datos atípicos, lo cual se muestra en el gráfico de caja de la serie, mostrado a continuación:

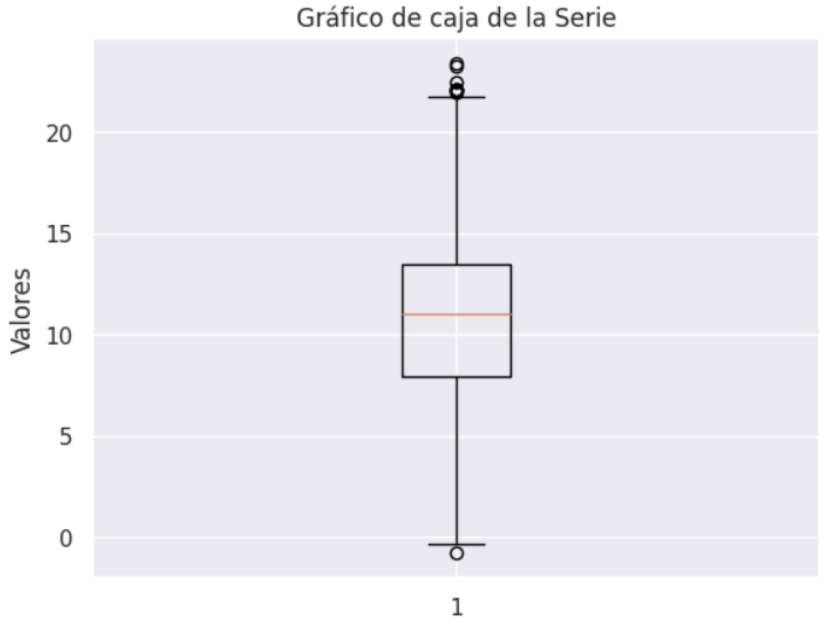

Figura 7. Boxplot para los datos de la nueva serie

En este punto, se toman los valores mínimos y máximos hallados en el gráfico de caja (figura 7), y a partir de ellos, se filtran los valores típicos, de manera que los valores puedan existir dentro de la máscara de aceptación definida en la serie.

```
[18] Q1 = tasa_interes.quantile(0.25)
     Q3 = tasa_interes.quantile(0.75)
     IQR = Q3-Q1
     lim_sup = (Q3+1.5*IQR)
     lim_inf = (Q1-1.5*IQR)
     # Imprime los valores mínimo y máximo de la caja (sin contar los valores atípicos)
     print("Valor mínimo de la caja en el gráfico de caja:", lim_inf)
     print("Valor máximo de la caja en el gráfico de caja:", lim_sup)

     Valor mínimo de la caja en el gráfico de caja: -0.45500000000000007
     Valor máximo de la caja en el gráfico de caja: 21.825000000000003
```

```
[19] tasa_tipicos = tasa_interes[(tasa_interes >= 0.455) & (tasa_interes <= 21.825)]
     mascara = (tasa_interes > 22) | (tasa_interes < 0.4)
     mascara1 = (tasa_interes >= 0.4) & (tasa_interes <= 22)

     # Sumar los valores que cumplen con la máscara
     suma_valores_deseados = tasa_interes[mascara1]
     print(len(suma_valores_deseados))
     print(len(tasa_interes))

     31669
     31679
```

Figura 8. Filtración de datos típicos, a partir del gráfico de caja

Aquí, se podrán descartar 7 valores como atípicos y, finalmente, el dataframe con el que se trabajará será definido como df4.

3. Se realiza una prueba de normalidad para los datos típicos, obteniendo que no cumplen con una distribución normal, como se muestra en la figura 9.

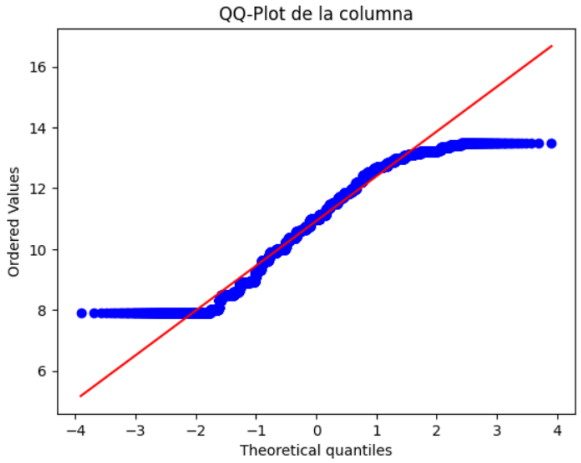

Figura 9. QQ-Plot para datos típicos

Esto inclusive se comprueba con pruebas teóricas estadísticas, donde se afirma la inexistencia de evidencia suficiente para afirmar que los datos se distribuyen normalmente.

```
[ ] import scipy.stats as stats

    # Supongamos que tienes una Serie llamada 'serie' con tus datos
    resultado, p_valor = stats.shapiro([tasa_interes])

    # Imprimir el resultado de la prueba y el p-valor
    print("Estadística de prueba:", resultado)
    print("P-valor:", p_valor)

    # Interpretar los resultados
    if p_valor > 0.05:
        print("Los datos parecen seguir una distribución normal.")
    else:
        print("Los datos no siguen una distribución normal.")
```

```
Estadística de prueba: 0.9768064022064209
P-valor: 0.0
Los datos no siguen una distribución normal.
/usr/local/lib/python3.10/dist-packages/scipy/stats/_morestats.py:1882: UserWarning: p-value may not be accurate for N > 5000.
  warnings.warn("p-value may not be accurate for N > 5000.")
```

Figura 10. Prueba Valor P para columna loan_int_rate

Significancia de variables

Habiendo realizado el pre-procesamiento de datos, se comienza a trabajar en un modelo basado en los datos filtrados y sus respectivas variables..

Inicialmente, se toma una decisión, y es la de la función objetivo. Se escoge la variable de loan_status, correspondiente al estado del préstamo. Esto debido a que se busca conocer la proporción de pagos y no pagos ante los préstamos dentro del buró crediticio, y cómo esto se ve afectado por las variables mencionadas.

Para ello, se observa una proporción de 78,34% versus 21,66%, que representan las personas con estatus de pago y de no pago. Esto se observa en la figura 11.

```
[21] # explore the unique values in loan_status column
     df4['loan_status'].value_counts(normalize = True)
```

```
0    0.783389
1    0.216611
Name: loan_status, dtype: float64
```

Figura 11. Proporción del estado de pago frente a los préstamos, siendo 0 pagado y 1 no pagado

Con estos datos, es factible comenzar a evaluar la significancia de todas las variables en el modelo, frente a la variable objetivo. Para esto, lo primero que se realiza, es definir los datos entre variables categóricas y numéricas.

De los datos analizados, se encuentran cuatro variables categóricas, las cuales son person_home_ownership, loan_intent, loan_grade y cb_person_default_on_file. También se definen siete variables numéricas, las cuales serían representadas por aquellas variables restantes.

```
[22] # split data into 80/20 while keeping the distribution of bad loans in test set same as that in the pre-split dataset
     X = df4.drop('loan_status', axis = 1)
     y = df4['loan_status']
     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2,
                                                         random_state = 42, stratify = y)

     # hard copy the X datasets to avoid Pandas' SetttingWithCopyWarning when we play around with this data later on.
     # this is currently an open issue between Pandas and Scikit-Learn teams
     X_train, X_test = X_train.copy(), X_test.copy()

     #Se eliminan las columnas debido al IV
     X_train2 = X_train.drop(['person_age','cb_person_cred_hist_length'], axis = 1)
     col=[col for col in X_train2]

     X_test2 = X_test.drop(['person_age','cb_person_cred_hist_length'], axis = 1)
```

```
[23] # first divide training data into categorical and numerical subsets
     X_train_cat = X_train.select_dtypes(include = 'object').copy()
     X_train_num = X_train.select_dtypes(include = 'number').copy()
     X_test_num = X_test.select_dtypes(include = 'number').copy()
```

Figura 12. Definición de datos categóricos y numéricos

Una vez definidas las variables, se procede a realizar cada una de las pruebas de significancia, obteniendo los siguientes resultados a partir de la tabla ANOVA:

	Numerical_Feature	F-Score	p values
0	loan_percent_income	3767.409615	0.000000
1	loan_int_rate	2975.113851	0.000000
2	person_income	636.893996	0.000000
3	loan_amnt	287.591259	0.000000
4	person_emp_length	154.476200	0.000000
5	person_age	13.526869	0.000236
6	cb_person_cred_hist_length	4.872463	0.027298

Figura 13. Tabla ANOVA para significancia de las variables

Con un nivel de significancia de 0.05, se toman los valores P mostrados en la tabla, y se observa que todos están por debajo de este. Se podría afirmar entonces que no existe evidencia suficiente para afirmar que las variables sean insignificantes dentro del modelo.

A partir de esto, se aceptan todas y se utilizan dentro del modelo.

Seguidamente, se tendrán que volver las variables categóricas en numéricas, asociando valores de 0 y 1 dependiendo de la categoría que se esté evaluando.

```
[27] # function to create dummy variables
     def dummy_creation(df, columns_list):
         df_dummies = []
         for col in columns_list:
             df_dummies.append(pd.get_dummies(df[col], prefix = col, prefix_sep = ':'))
         df_dummies = pd.concat(df_dummies, axis = 1)
         df = pd.concat([df, df_dummies], axis = 1)
         return df

     # apply to our final four categorical variables
     X_train = dummy_creation(X_train, ['person_home_ownership', 'loan_intent', 'loan_grade', 'cb_person_default_on_file'])
     X_test = dummy_creation(X_test, ['person_home_ownership', 'loan_intent', 'loan_grade', 'cb_person_default_on_file'])
     # reindex the dummied test set variables to make sure all the feature columns in the training set are also available in the test set
     X_test = X_test.reindex(labels=X_train.columns, axis=1, fill_value=0)

     X_test
```

Figura 14. Conversión de variables categóricas en numéricas

Una vez realizado esto, se hace un análisis de Weight of Evidence (WoE) e Information Value (IV), con el siguiente código:

```
# function to calculate WoE and IV of categorical features
# The function takes 3 arguments: a dataframe (X_train_prepr), a string (column name), and a dataframe (y_train_prepr).
def woe_discrete(df, cat_variabe_name, y_df):
    df = pd.concat([df[cat_variabe_name], y_df], axis = 1)
    df = pd.concat([df.groupby(df.columns.values[0], as_index = False)[df.columns.values[1]].count(),
                    df.groupby(df.columns.values[0], as_index = False)[df.columns.values[1]].mean()], axis = 1)
    df = df.iloc[:, [0, 1, 3]]
    df.columns = [df.columns.values[0], 'n_obs', 'prop_good']
    df['prop_n_obs'] = df['n_obs'] / df['n_obs'].sum()
    df['n_good'] = df['prop_good'] * df['n_obs']
    df['n_bad'] = (1 - df['prop_good']) * df['n_obs']
    df['prop_n_good'] = df['n_good'] / df['n_good'].sum()
    df['prop_n_bad'] = df['n_bad'] / df['n_bad'].sum()
    df['WoE'] = np.log(df['prop_n_good'] / df['prop_n_bad'])
    df = df.sort_values(['WoE'])
    df = df.reset_index(drop = True)
    df['diff_prop_good'] = df['prop_good'].diff().abs()
    df['diff_WoE'] = df['WoE'].diff().abs()
    df['IV'] = (df['prop_n_good'] - df['prop_n_bad']) * df['WoE']
    df['IV'] = df['IV'].sum()
    return df
```

Figura 15. Análisis WoE e IV

El análisis de WoE e IV muestra qué tan precisa puede llegar a ser una predicción hecha con los datos presentes, y de acuerdo a sus valores, se pueden tomar decisiones de su uso para un modelo de predicción (Kinden Property, 2019).

Teniendo en cuenta esto, se obtienen los siguientes valores para las variables discretas:

- Tipo de vivienda: Los datos de vivienda propia muestran un valor de IV entre 0.02 y 0.1, indicando que su poder de predicción es débil. Los datos de hipoteca y otro tipo de vivienda tienen un valor de IV 0.1 y 0.3, dándole un poder de predicción medio.

	Bin	Count	Count (%)	Non-event	Event	Event rate	WoE	IV	JS
0	[OWN]	1760	0.076839	1650	110	0.062500	1.422402	0.099255	0.011456
1	[MORTGAGE]	9409	0.410784	8215	1194	0.126900	0.643004	0.139619	0.017158
2	[OTHER, RENT]	11736	0.512377	8079	3657	0.311605	-0.493024	0.141456	0.017505
3	Special	0	0.000000	0	0	0.000000	0.0	0.000000	0.000000
4	Missing	0	0.000000	0	0	0.000000	0.0	0.000000	0.000000
Totals		22905	1.000000	17944	4961	0.216590		0.380330	0.046119

Figura 16. Análisis WoE e IV para variable person_home_ownership

- Intención de préstamo: Los datos en intención educativa, personal, de vivienda y médica muestran un IV muy bajo (menor a 0.02), indicando que no serán influyentes en la predicción de un modelo. Adicionalmente, la intención empresarial y consolidada muestran un IV de entre 0.02 y 0.1, indicando que su poder de predicción será débil.

	Bin	Count	Count (%)	Non-event	Event	Event rate	WoE	IV	JS
0	[VENTURE]	3979	0.173718	3418	561	0.140990	0.521441	0.040359	0.004989
1	[EDUCATION]	4570	0.199520	3786	784	0.171554	0.289008	0.015305	0.001906
2	[PERSONAL]	3912	0.170792	3126	786	0.200920	0.094904	0.001497	0.000187
3	[HOMEIMPROVEMENT]	2536	0.110718	1887	649	0.255915	-0.218338	0.005603	0.000699
4	[MEDICAL]	4249	0.185505	3108	1141	0.268534	-0.283574	0.016104	0.002006
5	[DEBTCONSOLIDATION]	3659	0.159747	2619	1040	0.284231	-0.362077	0.023057	0.002867
6	Special	0	0.000000	0	0	0.000000	0.0	0.000000	0.000000
7	Missing	0	0.000000	0	0	0.000000	0.0	0.000000	0.000000
Totals		22905	1.000000	17944	4961	0.216590		0.101925	0.012654

Figura 17. Análisis WoE e IV para variable loan_intent

- Grado de préstamo: El grado C muestra un IV menor a 0.02, indicando que no tendrá fuerza en la predicción del modelo. El grado B muestra que habrá un poder de predicción débil. El grado A, con un IV entre 0.1 y 0.3, muestra que tendrá un poder de predicción medio. Y, por último, la combinación de los datos de los grados D, E, F y G muestran un IV mayor a 0.5, dando indicios de un poder de predicción fuerte, pero posiblemente atípico.

	Bin	Count	Count (%)	Non-event	Event	Event rate	WoE	IV	JS
0	[A]	7575	0.330714	6834	741	0.097822	0.936016	0.216675	0.026137
1	[B]	7293	0.318402	6146	1147	0.157274	0.393003	0.043744	0.005433
2	[C]	4560	0.199083	3632	928	0.203509	0.078859	0.001210	0.000151
3	[D, E, F, G]	3477	0.151801	1332	2145	0.616911	-1.762106	0.631084	0.070041
4	Special	0	0.000000	0	0	0.000000	0.0	0.000000	0.000000
5	Missing	0	0.000000	0	0	0.000000	0.0	0.000000	0.000000
Totals		22905	1.000000	17944	4961	0.216590		0.892713	0.101762

Figura 18. Análisis WoE e IV para variable loan_grade

- Histórico de incumplimiento: Aquellas personas que muestran un histórico negativo (N), presentan datos con un IV asociado de entre 0.02 y 0.1, indicando un poder de predicción débil. Mientras que aquellas personas con histórico positivo (Y), presentan datos con IV entre 0.1 y 0.3, indicando poder de predicción débil.

	Bin	Count	Count (%)	Non-event	Event	Event rate	WoE	IV	JS
0	[N]	18835	0.82231	15407	3428	0.182002	0.217196	0.036408	0.004542
1	[Y]	4070	0.17769	2537	1533	0.376658	-0.781893	0.131066	0.015978
2	Special	0	0.00000	0	0	0.000000	0.0	0.000000	0.000000
3	Missing	0	0.00000	0	0	0.000000	0.0	0.000000	0.000000
Totals		22905	1.00000	17944	4961	0.216590		0.167473	0.020520

Figura 19. Análisis WoE e IV para variable cb_person_default_on_file

Ahora se toman las variables continuas. Para esto, se tomarán intervalos para determinar la distribución de los datos y se llegan a los siguientes resultados:

- Porcentaje de ingreso: Los datos más bajos (entre -∞ y 0.16) presentan un IV entre 0.02 y 0.1, representando datos con poder de predicción débil. Los datos entre 0.16 y 0.31 presentan IV menores a 0.02, representando datos que no tendrán fuerza en la predicción. Y los datos más altos (entre 0.31 y +∞) muestran un IV entre 0.3 y 0.5, representando datos con poder de predicción fuerte.

	Bin	Count	Count (%)	Non-event	Event	Event rate	WoE	IV	JS
0	(-inf, 0.05)	2462	0.107487	2199	263	0.106824	0.837955	0.058267	0.007077
1	[0.05, 0.07)	1857	0.081074	1633	224	0.120625	0.70088	0.032138	0.003937
2	[0.07, 0.13)	6051	0.264178	5286	765	0.126425	0.647293	0.090867	0.011164
3	[0.13, 0.16)	2565	0.111984	2197	368	0.143470	0.501117	0.024183	0.002992
4	[0.16, 0.20)	2948	0.128706	2419	529	0.179444	0.234473	0.006607	0.000824
5	[0.20, 0.25)	2705	0.118096	2184	521	0.192606	0.147515	0.002462	0.000308
6	[0.25, 0.31)	1683	0.073477	1257	426	0.253119	-0.203605	0.003221	0.000402
7	[0.31, 0.38)	1459	0.063698	456	1003	0.687457	-2.073906	0.366593	0.039052
8	[0.38, inf)	1175	0.051299	313	862	0.733617	-2.298701	0.359315	0.037071
9	Special	0	0.000000	0	0	0.000000	0.0	0.000000	0.000000
10	Missing	0	0.000000	0	0	0.000000	0.0	0.000000	0.000000
Totals		22905	1.000000	17944	4961	0.216590		0.943652	0.102826

Figura 20. Análisis WoE e IV para variable loan_percent_income

- Interés al préstamo: Los datos comprendidos entre -∞ y 6.46, y entre 15.28 y +∞ muestran un IV entre 0.1 y 0.3, representando datos con poder de predicción medio. Los datos comprendidos entre 6.46 y 9.64 presentan un IV entre 0.02 y 0.1, indicando datos con poder de predicción débil. Y los datos entre 9.64 y 14.37 muestran un IV menor a 0.02, representando datos que no tendrán fuerza en la predicción.

	Bin	Count	Count (%)	Non-event	Event	Event rate	WoE	IV	JS
0	(-inf, 6.46)	1584	0.069155	1501	83	0.052399	1.609398	0.107699	0.012175
1	[6.46, 7.21)	1550	0.067671	1414	136	0.087742	1.055874	0.054258	0.006484
2	[7.21, 7.89)	2418	0.105566	2170	248	0.102564	0.883405	0.062670	0.007589
3	[7.89, 9.64)	2476	0.108099	2150	326	0.131664	0.600677	0.032499	0.004002
4	[9.64, 10.38)	1657	0.072342	1399	258	0.155703	0.404905	0.010511	0.001305
5	[10.38, 11.27)	2449	0.106920	2059	390	0.159249	0.378181	0.013665	0.001698
6	[11.27, 12.76)	3653	0.159485	3004	649	0.177662	0.246619	0.009024	0.001125
7	[12.76, 13.61)	2141	0.093473	1672	469	0.219057	-0.014475	0.000020	0.000002
8	[13.61, 14.37)	1306	0.057018	943	363	0.277948	-0.330985	0.006824	0.000849
9	[14.37, 15.28)	1341	0.058546	718	623	0.464579	-1.143725	0.097864	0.011607
10	[15.28, 16.31)	1156	0.050469	484	672	0.581315	-1.613822	0.175073	0.019781
11	[16.31, inf)	1174	0.051255	430	744	0.633731	-1.833904	0.231084	0.025416
12	Special	0	0.000000	0	0	0.000000	0.0	0.000000	0.000000
13	Missing	0	0.000000	0	0	0.000000	0.0	0.000000	0.000000
Totals		22905	1.000000	17944	4961	0.216590		0.801191	0.092033

Figura 21. Análisis WoE e IV para variable loan_int_rate

- Ingreso anual: Los datos comprendidos entre -∞ y 34990, y entre 88450 y +∞, presentan un IV entre 0.1 y 0.3, representando un poder de predicción medio. Los datos entre 34990 y 59982 presentan un IV mucho menor a 0.02, indicando que estos valores no serán de utilidad para la predicción. Y los restantes, comprendidos entre 59982 y 88450 presentan un IV entre 0.02 y 0.1, representando un poder de predicción débil.

	Bin	Count	Count (%)	Non-event	Event	Event rate	WoE	IV	JS
0	(-inf, 24228.00)	1343	0.058633	614	729	0.542815	-1.457327	1.642825e-01	1.889153e-02
1	[24228.00, 34990.00)	2818	0.123030	1676	1142	0.405252	-0.90202	1.233907e-01	1.492133e-02
2	[34990.00, 39937.50)	1609	0.070247	1180	429	0.266625	-0.273836	5.672330e-03	7.068342e-04
3	[39937.50, 49996.50)	3444	0.150360	2698	746	0.216609	-0.000108	1.754743e-09	2.193428e-10
4	[49996.50, 59982.00)	3114	0.135953	2471	643	0.206487	0.060585	4.904497e-04	6.129684e-05
5	[59982.00, 79946.00)	4821	0.210478	4084	737	0.152873	0.426596	3.371737e-02	4.183001e-03
6	[79946.00, 88450.00)	1337	0.058372	1192	145	0.108452	0.821006	3.054216e-02	3.714034e-03
7	[88450.00, inf)	4419	0.192927	4029	390	0.088255	1.049478	1.531385e-01	1.830953e-02
8	Special	0	0.000000	0	0	0.000000	0.0	0.000000e+00	0.000000e+00
9	Missing	0	0.000000	0	0	0.000000	0.0	0.000000e+00	0.000000e+00
Totals		22905	1.000000	17944	4961	0.216590		5.112340e-01	6.078756e-02

Figura 22. Análisis WoE e IV para variable person_income

- Cantidad de préstamo: Casi todos los datos presentan un IV menor a 0.02, por lo que no serán muy útiles para la predicción. Únicamente el rango entre 22150 y +∞ muestra un IV entre 0.02 y 0.1, presentando una fuerza de predicción débil.

	Bin	Count	Count (%)	Non-event	Event	Event rate	WoE	IV	JS
0	(-inf, 3262.50)	2939	0.128313	2325	614	0.208915	0.045832	0.000266	0.000033
1	[3262.50, 5462.50)	3847	0.167955	3119	728	0.189238	0.169318	0.004584	0.000572
2	[5462.50, 6437.50)	2001	0.087361	1706	295	0.147426	0.469283	0.016711	0.002070
3	[6437.50, 7387.50)	1357	0.059245	1133	224	0.165070	0.33533	0.006032	0.000751
4	[7387.50, 9237.50)	2598	0.113425	2106	492	0.189376	0.168419	0.003064	0.000383
5	[9237.50, 10612.50)	2501	0.109190	2009	492	0.196721	0.121265	0.001550	0.000194
6	[10612.50, 12837.50)	2128	0.092905	1675	453	0.212876	0.022028	0.000045	0.000006
7	[12837.50, 18087.50)	3123	0.136346	2298	825	0.264169	-0.261237	0.009988	0.001245
8	[18087.50, 22150.00)	1171	0.051124	790	381	0.325363	-0.556415	0.018235	0.002250
9	[22150.00, inf)	1240	0.054137	783	457	0.368548	-0.747199	0.036226	0.004426
10	Special	0	0.000000	0	0	0.000000	0.0	0.000000	0.000000
11	Missing	0	0.000000	0	0	0.000000	0.0	0.000000	0.000000
Totals		22905	1.000000	17944	4961	0.216590		0.096702	0.011929

Figura 23. Análisis WoE e IV para variable loan_amnt

- Duración del empleo: Nuevamente, se observa que los datos no son muy útiles para la predicción, debido a que presentan IV mucho menores a 0.02. Únicamente los datos comprendidos entre -∞ y 1.5 años cuentan con un IV un poco mayor a 0.02, indicando que su fuerza de predicción será débil.

	Bin	Count	Count (%)	Non-event	Event	Event rate	WoE	IV	JS
0	(-inf, 1.50)	5064	0.221087	3638	1426	0.281596	-0.349088	0.029568	0.003677
1	[1.50, 2.50)	2776	0.121196	2069	707	0.254683	-0.211858	0.005764	0.000719
2	[2.50, 4.50)	4572	0.199607	3640	932	0.203850	0.076758	0.001150	0.000144
3	[4.50, 5.50)	2209	0.096442	1797	412	0.186510	0.187202	0.003201	0.000399
4	[5.50, 7.50)	3487	0.152238	2840	647	0.185546	0.193565	0.005391	0.000673
5	[7.50, 11.50)	3241	0.141497	2654	587	0.181117	0.22315	0.006601	0.000823
6	[11.50, inf)	1556	0.067933	1306	250	0.160668	0.367615	0.008231	0.001023
7	Special	0	0.000000	0	0	0.000000	0.0	0.000000	0.000000
8	Missing	0	0.000000	0	0	0.000000	0.0	0.000000	0.000000
Totals		22905	1.000000	17944	4961	0.216590		0.059906	0.007459

Figura 24. Análisis WoE e IV para variable person_emp_length

- Edad: Todos los datos comprendidos en esta variable cuentan con un IV mucho menor a 0.02, indicando que no serán muy útiles para la predicción del modelo.

	Bin	Count	Count (%)	Non-event	Event	Event rate	WoE	IV	JS
0	(-inf, 22.50)	3388	0.147915	2526	862	0.254427	-0.210511	6.943510e-03	8.663396e-04
1	[22.50, 25.50)	7407	0.323379	5803	1604	0.216552	0.000226	1.652344e-08	2.065430e-09
2	[25.50, 28.50)	4598	0.200742	3627	971	0.211179	0.032186	2.060618e-04	2.575661e-05
3	[28.50, 29.50)	1173	0.051212	931	242	0.206309	0.061673	1.913809e-04	2.391882e-05
4	[29.50, 31.50)	1718	0.075005	1366	352	0.204889	0.070362	3.639346e-04	4.548244e-05
5	[31.50, 39.50)	3410	0.148876	2723	687	0.201466	0.091507	1.214271e-03	1.517309e-04
6	[39.50, inf)	1211	0.052871	968	243	0.200661	0.096522	4.790925e-04	5.986333e-05
7	Special	0	0.000000	0	0	0.000000	0.0	0.000000e+00	0.000000e+00
8	Missing	0	0.000000	0	0	0.000000	0.0	0.000000e+00	0.000000e+00
Totals		22905	1.000000	17944	4961	0.216590		9.398267e-03	1.173094e-03

Figura 25. Análisis WoE e IV para variable person_age

- Duración de historial crediticio: Al igual que en el caso anterior, los IV asociados a todos los datos de esta variable son menores a 0.02, por lo que no serán de utilidad para la predicción del modelo.

	Bin	Count	Count (%)	Non-event	Event	Event rate	WoE	IV	JS
0	(-inf, 2.50)	4192	0.183017	3219	973	0.232109	-0.089207	0.001493	0.000187
1	[2.50, 4.50)	8366	0.365248	6514	1852	0.221372	-0.027961	0.000288	0.000036
2	[4.50, 5.50)	1349	0.058895	1065	284	0.210526	0.036107	0.000076	0.000009
3	[5.50, 6.50)	1313	0.057324	1045	268	0.204113	0.075137	0.000317	0.000040
4	[6.50, 7.50)	1322	0.057717	1063	259	0.195915	0.126374	0.000889	0.000111
5	[7.50, 8.50)	1338	0.058415	1064	274	0.204783	0.071014	0.000289	0.000036
6	[8.50, 11.50)	2976	0.129928	2355	621	0.208669	0.047316	0.000287	0.000036
7	[11.50, inf)	2049	0.089456	1619	430	0.209858	0.04013	0.000142	0.000018
8	Special	0	0.000000	0	0	0.000000	0.0	0.000000	0.000000
9	Missing	0	0.000000	0	0	0.000000	0.0	0.000000	0.000000
Totals		22905	1.000000	17944	4961	0.216590		0.003781	0.000472

Figura 26. Análisis WoE e IV para variable cb_person_cred_hist_length

De acuerdo al análisis realizado con las variables discretas y continuas, se observa que aquellas que, con seguridad, no serán de utilidad o aporte para predecir el modelo, son la variable person_age y cb_person_cred_hist_length (edad de la persona y duración del historial crediticio, respectivamente).

Es por ello que se decide eliminar estas dos variables o columnas de la base de datos, para proseguir con un modelo más acertado.

Modelo propuesto

A partir de las variables definidas anteriormente, se construye un modelo de regresión logística donde se inicia una instancia de la clase BinningProcess. De igual manera, se

presenta un estimador para el modelo, el cual permitirá hasta 1000 iteraciones para su funcionamiento, y se define el class_weight para atacar el problema del desbalanceo de clases

```
[57] binning_process = optbinning.BinningProcess(
         variable_names=cols,
         #selection_criteria=selection_criteria,
         categorical_variables=['person_home_ownership', 'loan_intent', 'loan_grade', 'cb_person_default_on_file']
     )
```

```
[59] estimator = linear_model.LogisticRegression(max_iter=1000, class_weight = 'balanced')
```

Figura 27. BinningProcess y estimador para el modelo de regresión logística

A partir de esto, se construye la scorecard, la cual se utilizará para calificar a los solicitantes del crédito. Esta recibe como entrada la probabilidad de incumplimiento crediticio y devuelve el puntaje ya escalado entre 300 y 800.

```
[60] # scorecard
     scorecard = optbinning.Scorecard(
         binning_process=binning_process,
         estimator=estimator,
         scaling_method="min_max",
         scaling_method_params={"min": 300, "max": 800},
         #scaling_method = "pdo_odds",
         #scaling_method_params = {"pdo": 20, "odds": 50, "scorecard_points": 100},
         #intercept_based=True,
         #reverse_scorecard=True
     )
```

Figura 28. Scorecard para modelo de regresión logística

Se ajusta entonces el modelo para aplicarlo en la Scorecard, y se obtiene la siguiente tabla con los valores que serán presentados de acuerdo a la variable de grado de préstamo:

	Variable	Bin id	Bin	Count	Count (%)	Non-event	Event	Event rate	WoE	IV	JS	Coefficient	Points
0	loan_grade	0	[A]	8231	0.324887	7457	774	0.094035	0.972155	0.226719	0.027274	-1.076543	111.531599
1	loan_grade	1	[B]	8186	0.323110	6853	1333	0.162839	0.344073	0.034530	0.004295	-1.076543	82.005870
2	loan_grade	2	[C]	5107	0.201579	4073	1034	0.202467	0.077763	0.001192	0.000149	-1.076543	69.486839
3	loan_grade	3	[D, E, F, G]	3811	0.150424	1497	2314	0.607190	-1.728696	0.603135	0.067216	-1.076543	-15.433640
4	loan_grade	4	Special	0	0.000000	0	0	0.000000	0.000000	0.000000	0.000000	-1.076543	65.831230
5	loan_grade	5	Missing	0	0.000000	0	0	0.000000	0.000000	0.000000	0.000000	-1.076543	65.831230

Figura 29. Comportamiento de las variables dentro del modelo de regresión para la variable loan_grade

Igualmente, los datos que se muestran en la figura 30, a continuación, son una ejemplificación de lo que se podrán observar en la Scorecard:

	person_income	person_home_ownership	person_emp_length	loan_intent	loan_grade	loan_amnt	loan_int_rate	loan_percent_income	cb_person_default_on_file	score
22193	20000	RENT	4.0	PERSONAL	C	6800	12.840000	0.34	N	454.812662
21896	76000	RENT	4.0	MEDICAL	B	6000	3.475426	0.08	N	638.260802
19618	35000	MORTGAGE	1.0	VENTURE	B	4000	10.990000	0.11	N	678.252362
6402	38004	RENT	0.0	DEBTCONSOLIDATION	B	6000	11.890000	0.16	N	581.948541
26892	117000	MORTGAGE	17.0	HOMEIMPROVEMENT	A	20000	6.170000	0.17	N	698.432928
...	...	...	...	...	...	...	...	...	...	...
3151	33000	RENT	0.0	MEDICAL	B	3500	11.710000	0.11	N	572.540924
7397	39000	RENT	3.0	MEDICAL	B	6600	11.988233	0.17	N	584.009878
13502	92700	MORTGAGE	5.0	VENTURE	C	1500	14.650000	0.02	Y	712.012838
11040	48000	RENT	1.0	DEBTCONSOLIDATION	B	4100	11.990000	0.09	N	601.292277
14623	110000	MORTGAGE	9.0	PERSONAL	A	20000	7.880000	0.18	N	715.580484

Figura 30. Comportamiento de algunos datos por cada variable en el modelo

Teniendo en cuenta esto, se presenta la figura 31, la cual presentará los scores obtenidos para el modelo de regresión propuesto.

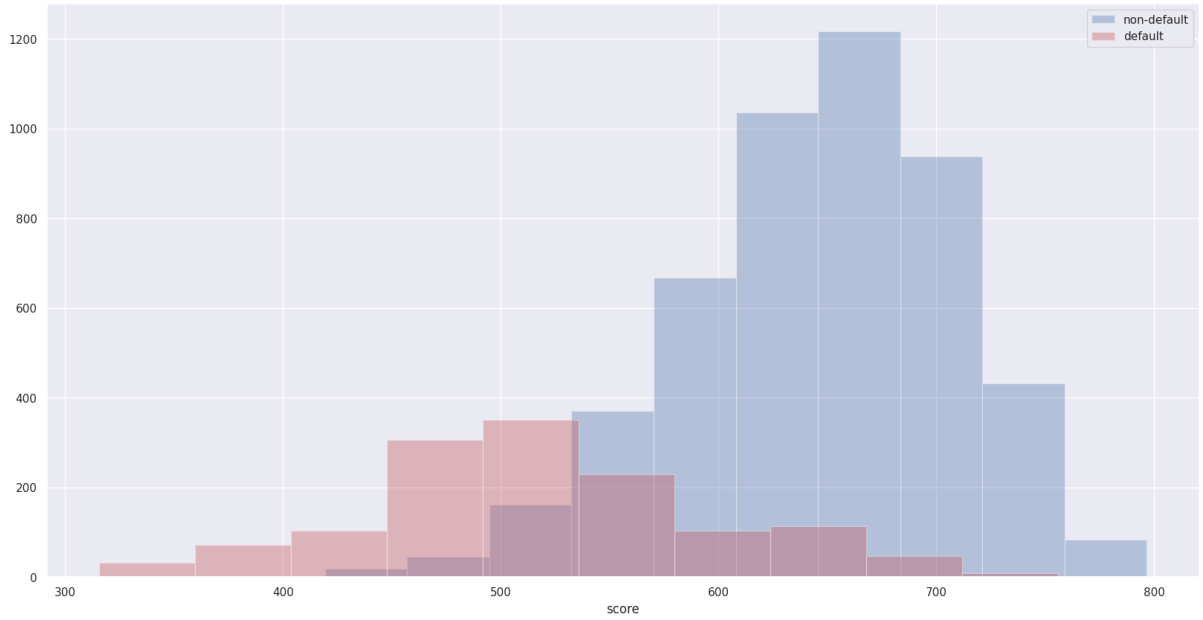

Figura 31. Distribución de scores en el modelo de regresión

Aprendizajes

Un modelo de regresión debe ser realizado, basado en datos que sean comunes y fácilmente explicados dentro del contexto en que se encuentren. Es por eso que, lo primordial en estos casos de estudio, será realizar un procesamiento previo de los datos, el cual permite que se prediga un modelo mayormente adecuado a lo que se busca.

Igualmente, el análisis de WoE e IV permite analizar la significancia de las variables, y denota un peso importante frente a las decisiones que se tomen con ellas. Si bien, existen las pruebas de significancia tradicionales, estas no estarían muy acertadas con el modelo y no permitirían asignarles valores de score a los datos y variables presentadas.

El IV juega entonces un papel importante al momento de definir modelos y series de regresión que permitan predecir con mayor peso lo que se desea observar. Gracias a ello, en el modelo propuesto anteriormente se pudieron eliminar con facilidad dos variables que parecían no ser de mucha importancia para realizar predicciones.

Casos de uso

El uso de metodologías WoE e IV está presente para análisis de datos extensos y típicos. Esto indica que puede utilizarse siempre y cuando haya un procesamiento adecuado de datos, sin importar estos de qué manera se distribuyen en el tiempo.

Entrando en el caso específico del buró crediticio analizado, asumiendo que los datos son confiables y genéricos para poder predecir comportamientos en una situación común, podría ser de utilidad este manejo de datos en diversas ocasiones.

Por ejemplo, para tomar decisiones bancarias de si realizar préstamos a clientes, qué tipos de préstamos podrían tenerse en cuenta, cuál es la esperanza de que se retorne todo el préstamo realizado, entre otras decisiones que se pueden tomar a partir de ellos.

Otro ejemplo estaría en el caso natural, donde la persona pueda observar con facilidad cómo está manejando sus préstamos a otras personas naturales y, así, generar confianza en el mundo crediticio descrito. Además de que puedan evaluar si les es factible realizar préstamos con entidades financieras, observando si les es viable o no.

Finalmente, es una fuente para que las empresas puedan también tomar decisiones de cómo están realizando sus movimientos de préstamo y deuda en el sector comercial y financiero interno, teniendo en cuenta también la relación con los bancos y el movimiento de sus pasivos.

Conclusiones

Realizar un modelo basado en datos existentes puede llevar a diversas conclusiones. En este caso, es importante recalcar que la interpretación de los datos y de las variables a utilizar son de dependencia de aquellos que estén evaluando el problema.

La forma en que se manejaron los datos fueron tenidas en cuenta frente a si eran acordes al contexto en el que se encontraban, si no eran valores muy alejados frente a los otros y si estos datos eran de peso para predecir.

Esto último se da con la ayuda de métodos WoE e IV, lo cual da una filtración de datos más detallada y permite evaluar si las variables son de peso para realizar una predicción. Esto indica que va más allá de las correlaciones encontradas entre variables, y ayuda a reducir la posible incertidumbre que presente la correlación o significancia de las variables dentro de un modelo.

Finalmente, es de importancia este análisis de peso para asignarle un score a cada variable, y que, de esta manera, sea de facilidad implementar los datos para un modelo de regresión con mayor certidumbre.

Referencias bibliográficas

Kinden Property. (2019, December 6). *Intro to Credit Scorecard: Step by Step Guide on How to build a simple Credit Scorecard*. Towards Data Science. https://towardsdatascience.com/intro-to-credit-scorecard-9afeaaa3725f

Tse, L. (n.d.). *Credit Risk Dataset*. Kaggle. https://www.kaggle.com/datasets/laotse/credit-risk-dataset