Problema predictivo a resolver:

La entidad bancaria desea predecir el nivel de riesgo al que se encuentra expuesta con cada crédito que otorga a sus usuarios. Esto es importante, debido a que un grueso de la población no tiene acceso a los servicios bancarios. Por ende, no posee historial crediticio, o datos que corroboren su intención y capacidad de pago.

A raíz de esta situación, muchos prestamistas malintencionados se han aprovechado de ello, y han otorgado créditos falsos, o con altos intereses, a la población, haciendo que la brecha por acceder a servicios financieros formales sea más alta.

Es allí donde un modelo confiable y seguro se hace indispensable. Esta entidad bancaria no puede comprometer sus operaciones, por lo que tampoco puede asumir un gran riesgo al otorgar créditos. Si llegase a desembolsar sumas de dinero a personas con altas probabilidades de no pagar a tiempo, podría caer rápidamente en bancarrota.

Con el ánimo de obtener más clientes certeros (que paguen o con poca tendencia a no pagar), y evitar evaluar mal a un prospecto a crédito, se crea este modelo.

Dado que es un caso

El dataset a utilizar posee información en diferentes aspectos de cada uno de los prospectos del crédito. Allí se puede ver información personal tal como:

- Edad
- Género
- Nivel de estudios
- Si tiene vehículo propio
- Número de hijos
- Total de ingresos anuales
- Estado civil
- Días laborados o con trabajo fijo
- Sector en el que trabaja

Entre otros datos de interés para el banco o entidad crediticia. Dado que este es un caso donde los datos de entrada son varias de estas categorías, y el dato de salida pretende ser una evaluación específica en “apto para crédito” o “rechazado para crédito”, el caso se vuelve clasificatorio.

En resumen, se puede decir que el problema a resolver y evaluar es: dadas las condiciones de un usuario, como su nivel de ingresos, educación, bienes, etc, vamos a predecir si es apto para un crédito, o no, según el riesgo que puede asumir x entidad financiera que presta dinero.

Dataset a utilizar

El dataset consta de alrededor de 108 columnas, y 48.000 instancias (filas). De allí, tenemos columnas categóricas como CODE_GENDER (masculino, femenino, 0 - 1), OWN_CAR (Y – N, 0 - 1), y también

otros de carácter no discreto como INCOME_TOTAL (ingresos en integer - float), DAYS_EMPLOYED, CAR_AGE, etc. A continuación, se deja una muestra de los datos para hacerse una idea:

NAME_CO	CODE_GEI	FLAG_OW	FLAG_OW	CNT_CHIL	AMT_INC	AMT_CRE	AMT_ANN	AMT_GOO	NAME_TYI	NAME_IN	NAME_ED	NAME_FA	NAME_HO	REGION_P	DAYS_BIR	DAYS_EMI	DAYS_REG	DAYS_ID_I	OWN_CAF	FLAG_MOI	FLAG_EMF
Cash loans	F	N	Y	0	135000	568800	20560.5	450000	Unaccomp	Working	Higher edu	Married	House / ap	0.01885	-19241	-2329	-5170	-812		1	1
Cash loans	M	N	Y	0	99000	222768	17370	180000	Unaccomp	Working	Secondary	Married	House / ap	0.035792	-18064	-4469	-9118	-1623		1	1
Cash loans	M	Y	Y	0	202500	663264	69777	630000		Working	Higher edu	Married	House / ap	0.019101	-20038	-4458	-2175	-3503	5	1	1
Cash loans	F	N	Y	2	315000	1575000	49018.5	1575000	Unaccomp	Working	Secondary	Married	House / ap	0.026392	-13976	-1866	-2000	-4208		1	1
Cash loans	M	Y	N	1	180000	625500	32067	625500	Unaccomp	Working	Secondary	Married	House / ap	0.010032	-13040	-2191	-4000	-4262	16	1	1
Cash loans	F	Y	Y	0	270000	959688	34600.5	810000	Unaccomp	State serv	Secondary	Married	House / ap	0.025164	-18604	-12009	-6116	-2027	10	1	1
Cash loans	M	Y	Y	2	180000	499221	22117.5	373500	Unaccomp	Working	Higher edu	Married	House / ap	0.0228	-16685	-2580	-10125	-241	3	1	1
Cash loans	M	N	Y	0	166500	180000	14220	180000	Unaccomp	Working	Higher edu	Single / no	With parer	0.005144	-9516	-1387	-5063	-2055		1	1
Cash loans	F	N	Y	0	315000	364896	28957.5	315000	Unaccomp	State serv	Higher edu	Married	House / ap	0.04622	-12744	-1013	-1686	-3171		1	1
Cash loans	F	Y	Y	1	162000	45000	5337	45000	Family	Working	Higher edu	Civil marri	House / ap	0.018634	-10395	-2625	-8124	-3041	5	1	1
Cash loans	F	N	Y	0	67500	675000	25447.5	675000	Unaccomp	Pensioner	Secondary	Married	House / ap	0.003122	-23670	365243	-7490	-4136		1	0
Cash loans	F	N	Y	0	135000	261621	16848	216000	Unaccomp	Working	Secondary	Widow	House / ap	0.008019	-15524	-3555	-7833	-3985		1	1
Cash loans	F	N	Y	0	247500	296280	23539.5	225000	Unaccomp	Working	Secondary	Civil marri	House / ap	0.018634	-12278	-929	-6031	-4586		1	1
Cash loans	F	Y	Y	0	90000	360000	18535.5	360000	Unaccomp	Working	Secondary	Married	House / ap	0.01452	-19687	-3578	-10673	-3183	3	1	1
Revolving	M	N	Y	0	180000	157500	7875	157500	Spouse, pa	Working	Secondary	Civil marri	House / ap	0.031329	-12091	-1830	-1042	-4221		1	1
Cash loans	M	Y	Y	0	180000	296280	21690	225000	Unaccomp	Working	Secondary	Civil marri	House / ap	0.032561	-13563	-1007	-5719	-4044	14	1	1
Cash loans	F	Y	Y	0	202500	407520	26041.5	360000	Unaccomp	Working	Secondary	Single / no	House / ap	0.00702	-19375	-4739	-3035	-2895	11	1	1
Cash loans	M	Y	Y	0	90000	499221	22117.5	373500	Unaccomp	Pensioner	Secondary	Married	House / ap	0.003122	-22705	365243	-5107	-5990	15	1	0
Cash loans	F	Y	Y	1	225000	431280	23526	360000	Unaccomp	Commerci	Higher edu	Civil marri	With parer	0.025164	-10962	-1883	-99	-1721	8	1	1
Cash loans	F	Y	Y	0	175500	478498.5	46741.5	454500	Family	Working	Secondary	Married	House / ap	0.010147	-17955	-305	-10087	-1516	10	1	1
Cash loans	F	N	Y	0	99000	225000	19242	225000	Unaccomp	Commerci	Secondary	Civil marri	House / ap	0.007274	-10507	-2780	-5072	-2729		1	1
Cash loans	F	N	Y	0	157500	266652	16443	202500	Unaccomp	Working	Higher edu	Widow	House / ap	0.011657	-22557	-13294	-13154	-4187		1	1
Cash loans	F	N	Y	0	135000	540000	27702	540000	Family	Pensioner	Higher edu	Separated	House / ap	0.072508	-22784	365243	-5103	-5229		1	0
Cash loans	M	N	Y	1	337500	1313213	42493.5	1012500	Unaccomp	Commerci	Incomplet	Married	House / ap	0.02461	-11948	-1415	-5611	-4052		1	1
Cash loans	M	Y	Y	0	157500	539100	29245.5	450000	Unaccomp	Working	Secondary	Married	House / ap	0.031329	-13172	-3518	-6207	-4243	20	1	1
Cash loans	F	N	Y	0	76500	225000	12334.5	225000	Unaccomp	Working	Secondary	Civil marri	House / ap	0.010966	-15394	-967	-5389	-5433		1	1
Cash loans	F	N	Y	0	112500	256032	10759.5	180000	Unaccomp	Pensioner	Secondary	Widow	House / ap	0.019689	-21150	365243	-4157	-4658		1	0
Cash loans	F	N	Y	0	225000	501363	25726.5	418500	Family	Commerci	Higher edu	Married	House / ap	0.031329	-21040	-2467	-918	-3663		1	1

El dataset completo puede ser visto en https://www.kaggle.com/competitions/home-credit-default-risk/overview

Métricas de desempeño

Para medir la efectividad, se utilizará, fundamentalmente:

Accuracy: el algoritmo será entrenado con un caso ideal, donde una persona con x cantidad de ingresos, sin hijos, con propiedades y bienes, con estudios superiores, entre otras características. De allí, deberá buscar similitudes, y evaluar o compara cada usuario con este cliente ideal, para determinar cuánto porcentaje de pago pierde por cada requisito que no cumpla.

De aquí, se tendrá en cuenta los falsos positivos (FP), falsos negativos (FN), verdaderos positivos (TP) y verdaderos negativos (TN) que arroje el modelo.

Para accuracy, tenemos:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

Precision: Esta se calculará de la siguiente manera:

$$\text{Precision} = TP / (TP + FP)$$

Recall: será la medición de los verdaderos positivos, o aquellos aciertos que cumplen con las características esperadas:

$$\text{Recall} = TP / (TP + FN)$$

F1 Score: Adicionalmente, se usará este factor para medir si la muestra se encuentra muy desbalanceada:

$$F1 = 2 * ((\text{recall} * \text{precision})/(\text{recall} + \text{precision}))$$

Con lo anterior, el modelo debería estar en capacidad de predecir correctamente los casos en un 57%, teniendo en cuenta que si la precisión predice bien 2 de 3 casos, sería un 66.6% de efectividad; luego, el recall toma bien 2 de cada 4, con un 50%, y el F1 score hace lo siguiente: F1 de 2 * ((0,50 * 0,666)/(0,50 + 0,666)), que da como resultado 57,1%.

Se puede concluir que, si el modelo no logra predecir bien un 57% de los casos, no representa un esfuerzo útil para la entidad bancaria, ya que el riesgo que asume sería igual de incierto o con alta probabilidad de fallar.