Can Large Language Models Predict Credit Risk? An Empirical Study on Consumer Loans in Chile

Diego Beas L.
División de Regulación de Solvencia y Liquidez - Bancos
CMF

David Díaz S.
Associate Professor, AI and Data Analytics, FEN
University of Chile

IX Conferencia Anual de la Comisión para el Mercado Financiero (CMF)
"100 años de supervisión bancaria en Chile“
Junio 2025

Introducción - motivación

Los Modelos de Lenguaje de Gran Escala (LLMs, por sus siglas en inglés) han emergido rápidamente como un foco central de la investigación en inteligencia artificial (IA), con aplicaciones en numerosos sectores de la economía.

En el sector financiero, las técnicas de aprendizaje automático (ML) ya han mostrado mejoras sustanciales en la modelación predictiva para evaluar la probabilidad de que un deudor no cumpla con sus obligaciones financieras.

La modelación del riesgo crediticio, en particular la estimación de la Probabilidad de Incumplimiento (PD), es crucial para las instituciones financieras, ya que ayuda a determinar requerimientos normativos, así como realizar definiciones estratégicas.

Los LLMs representan una técnica prometedora en este sentido, ya que están diseñados para procesar tanto datos estructurados como no estructurados, como información de clientes y documentos financieros.

A pesar de este potencial, el uso de LLMs en la predicción del riesgo crediticio ha sido limitado. Estudios recientes han comenzado a explorar la aplicabilidad de los LLMs en dominios relacionados. Por ejemplo, Bakumenko et al. (2024) demostraron que los LLMs pueden mejorar el rendimiento en la detección de anomalías por fraude en transacciones financieras.

Objetivos del trabajo

- **Experimentar con LLMs como predictores del incumplimiento en un portafolio de créditos, para:**

1. Evaluar su desempeño utilizando métricas tradicionales de clasificación binaria (AUROC y AUC de Precisión-Recall).

2. Comparar su rendimiento con otros algoritmos clásicos de Machine Learning:
a) Regresión Logística,
b) Light Gradient Boosting Machines (LightGBM),

3. Evaluar los efectos del diseño de prompts (*Prompt Engineering*):
a) Ejemplos en el prompt: comparación entre Zero Shots, Few Shots y RAG con *prompts* dinámicos.
b) Otras estrategias de prompts: juegos de rol, cadena de pensamiento (Chain-of-Thought), tomar una pausa, estímulos emocionales.

4. Evaluar los efectos del *fine tuning*.

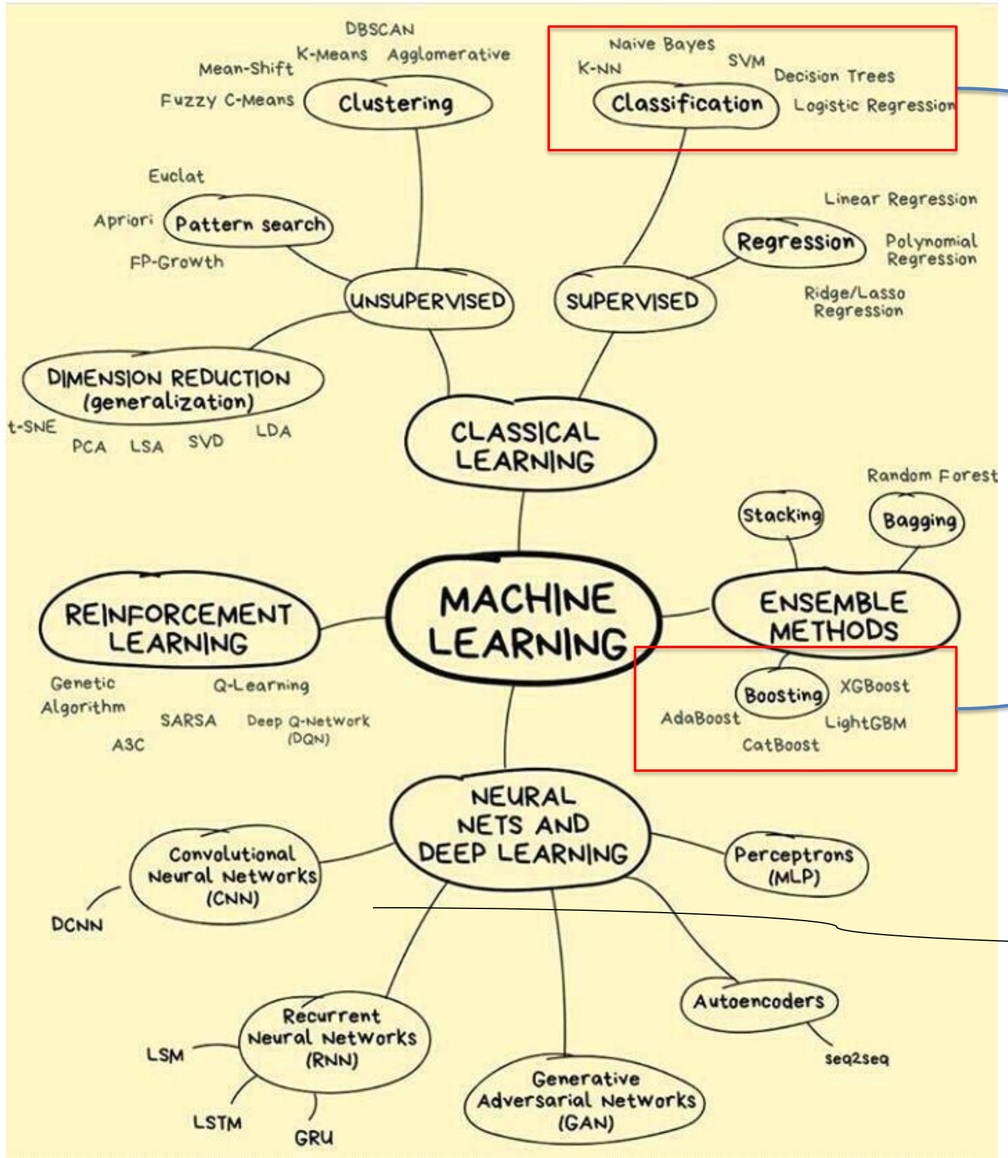
5. Evaluar la estabilidad de las predicciones debido al hiperparámetro "*Temperature*".

6. Evaluación cualitativa (humana) de las justificaciones entregadas por los LLMs y análisis de su equidad y capacidad de explicación (*explainability*).

Datos y variables utilizadas

- La investigación se centra en operaciones de crédito de consumo otorgadas por bancos en Chile entre 2010 y 2020. Se utilizó una muestra de 10.000 observaciones **anonimizadas** de deudores.
- La variable objetivo de este estudio es la Probabilidad de Incumplimiento (PD), la cual toma valor 1 si un deudor entra en incumplimiento dentro de los siguientes 12 meses a la observación de las variables, y 0 en caso contrario.
- El acceso estuvo limitado a variables predictoras y la etiqueta objetivo, donde las variables predictoras (V1 a V10, anexo) representan indicadores ampliamente reconocidos en modelos de riesgo de crédito, enfocándose en aspectos como comportamiento de pago y razones de endeudamiento.
- **El conjunto de datos se dividió en 10 subconjuntos estratificados y asignados aleatoriamente, manteniendo en cada uno la proporción constante de deudores en incumplimiento (10%). Esto nos permite realizar evaluación con validación cruzada para determinar un intervalo de confianza del desempeño de cada modelo-configuración.**
- Para los modelos que no requerían entrenamiento (como los LLMs sin *fine tuning*), se evaluó cada modelo-configuración 10 veces sobre las 1.000 muestras de prueba de cada subconjunto.

Taxonomía tradicional de los modelos de machine learning



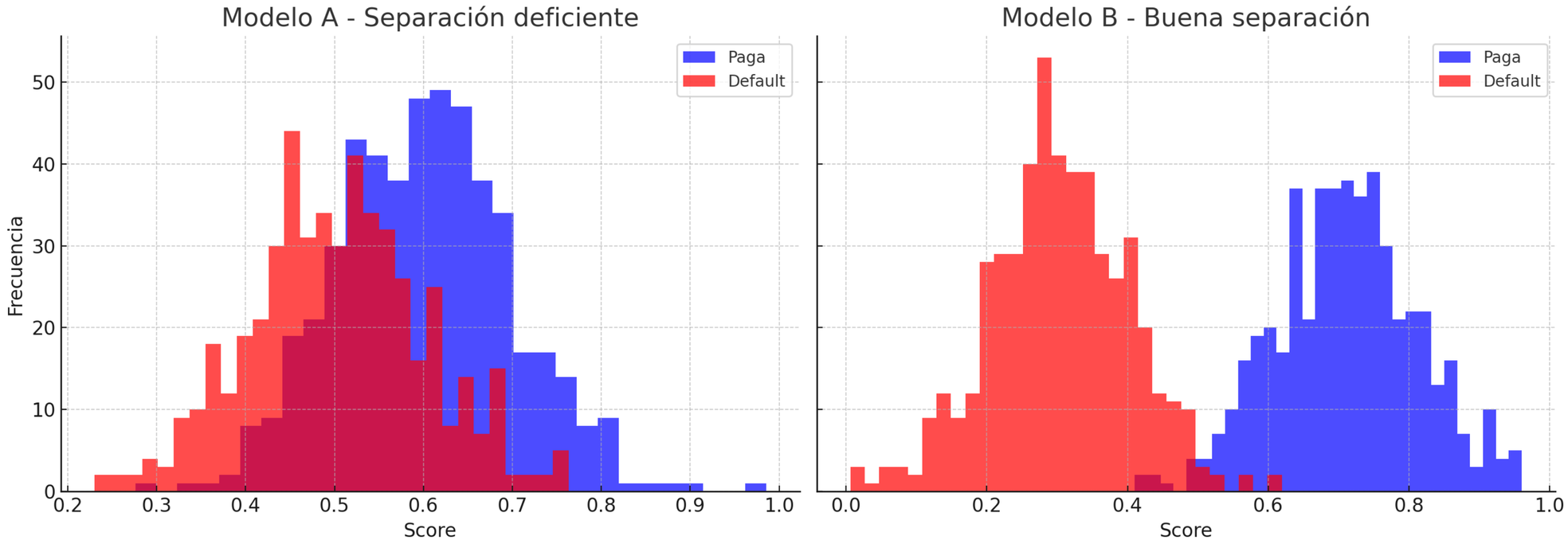
Diseñados para datos estructurados

Mejores con datos no estructurados

Transformers ChatGPT – BERT, Llama, …

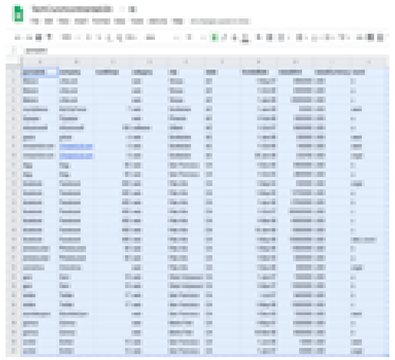
Resultado deseado en credit scoring

- Un modelo deficiente (MA), no logra diferenciar entre aquellos que pagan a tiempo versus aquellos que entran en incumplimiento.
- Habitualmente, se busca un modelo que logre separar y agrupar, en base al puntaje que nos arroja el modelo, a clientes que son buenos pagadores de aquellos que podrían incumplir (MB).



¿Cómo obtener este resultado si el modelo LLM sólo nos devuelve una etiqueta?

Estrategia para obtener predicciones del modelo LLM



Structured input
Known Y and Xs

Transform Data

To unstructured

```
#### A bank customer has the following features:
MAX_MORA_T3_T0= 0.0
TEND_T3_T12= 1.654124208314593
BIN_DET30_SIST_T0= 0.0
MAX_MORA_T0_T0= 0.0
PROM_MORA_T12_T0= 0.0
BIN_DET30_SIST_T12_T0= 0.0
BIN_GAR_CHIP_T0= 0.0
DTOI_T0_T0_2= 0.6967880509029298
PROM_CF2_T12_T0_1= 0.0536255735227787
PROM_DTOI_T12_T0_2= 0.8177925605742992


####

Its status is:
```

Convert Xs into text
Unstructured input

"Example Based Prompting"
"Zero Shot learning"
"Few Shots learning"
"RAG and Dynamic Shot Learning"

Prompt Engineering

Prompt Strategies
"Role play"
"Chain-of-thought"
"Take a breather"
"Emotional Stimuli"

Make a Prediction: Is this customer going to default?

Large Language Model

Answer

Provide justification for its prediction

Human evaluation

Para aproximar el "score", en vez de tomar la etiqueta predicha, se toma la probabilidad (confianza) con la cuál el modelo LLM señala el valor. Luego, se transforma mediante la siguiente expresión:

$$s(x) = \begin{cases} l(label) & if\ label = \text{default}\ and\ l(label) > 0.5 \\ 1 - l(label) & if\ label = \text{normal}\ and\ l(label) > 0.5 \\ 0.5 & if\ \ l(label) \leq 0.5 \end{cases}$$

You are an intelligent credit evaluation system, and you will help me in evaluating the credit risk of debtors. For each client we have information on certain variables, whose definition is:
V1: …
…
V10: ..
Now, predict the status of the following debtors:
#ID = 10001:
V1 = 0.0
V2 = 0.0
V3 = 0.4223
V4 = 0.0
V5 = 0.1453
V6 = 0.0239
V7 = 0.7657
V8 = 0.0
V9 = 0.0
V10 = 0.0
…
The response you must generate is a table separated by ";" with the format "ID";"ESTATUS"; where ID is the ID of each debtor and STATUS is whether the debtor is predicted as normal or in default.
To identify the location of the table, there must be a { sign at the beginning and } at the end of the response. For example, you could answer {100, normal} where 100 is the ID and "normal" the status.
The answer must contain only the required table.

Context (System)

Shots (Zero, Few, Dynamic-RAG)

Task

Output format

Other Prompt Engineering (TaB, CoT, ES)

……
The response you must generate is a table separated by ";" with the format "ID";"ESTATUS"; where ID is the ID of each debtor and STATUS is whether the debtor is predicted as normal or in default.
To identify the location of the table, there must be a { sign at the beginning and } at the end of the response. For example, you could answer {100, normal} where 100 is the ID and "normal" the status.
The answer must contain only the required table.

Output format

Also, you must provide the three most important reasons that you use to get your prediction.
For each given reason, you must provide a short explanation and the variables with its values that are considered in the reason. Also, for each reason you must provide how the considered variables impact on your reasoning and what values you would have considered in the opposite direction.
For the reasoning part, you must generate and output with the following dictionary format: <begins>{r1:"reason1", r2:"reason2", ...}</begins> where r1 is the reason 1 and reason1 that includes the explanation, variables, impact and all the required analysis before.
At the beginning of the reasons part, please write the following title "#Reasons for prediction#".

Asking Justification and Variable Importance

Resultados (1/4)

					Prompt Engineering (PE)				RAG		
Model	N° exp.	N° shots	Chunk size	T	RP	CoT	TaB	ES	DS	AUROC	AVGPREC
RL	1	-	-		-					79,1% ± 1,2%	39,3% ± 2,9%
L-GBM	2	-	-		-					**81,7% ± 1,3%**	**45,2% ± 2,1%**
Llama 3 8B (without fine tuning)	3	0	1	D	✓					73,6% ± 1,9%	27,3% ± 2,5%
	4	5	1	D	✓					73,4% ± 1,2%	28,5% ± 2,6%
	5	**5**	**1**	**D**	✓	✓	✓	✓		**73,5% ± 2,0%**	**28,6% ± 2,8%**
	6	5	1	D	✓	✓	✓	✓	✓	66,2% ± 1,5%	22,3% ± 2,1%
	7	5	20	D	✓					71,0% ± 1,7%	22,9% ± 1,5%
	8	10	1	D	✓					73,1% ± 2,0%	28,9% ± 3,0%
	9	20	1	D	✓					69,4% ± 1,5%	26,3% ± 2,3%
GPT 3.5 (without fine tuning)	10	0	1	D	✓					69,3% ± 1,9%	23,3% ± 2,1%
	11	0	20	D	✓					72,8% ± 1,9%	25,2% ± 1,9%
	12	0	100	D	✓					71,3% ± 1,7%	24,7% ± 1,9%
	13	5	1	D	✓					72,6% ± 2,1%	26,4% ± 2,1%
	14	**5**	**20**	**D**	✓					**73,4% ± 1,6%**	**24,8% ± 1,7%**
	15	5	100	D	✓					71,2% ± 2,1%	22,9% ± 2,3%
	16	5	1	D	✓	✓	✓	✓		69,4% ± 2,5%	21,2% ± 2,6%
	17	5	20	D	✓	✓	✓	✓		73,2% ± 1,9%	24,4% ± 2,0%
	18	5	20	D	✓	✓	✓	✓	✓	68,5% ± 2,0%	26,5% ± 2,5%
	19	10	1	D	✓					71,9% ± 2,2%	23,9% ± 1,7%
	20	10	100	D	✓					69,7% ± 1,9%	22,4% ± 3,1%

Symbology:

T: temperature value, where "D" refers to the default setting.

Prompt strategies:

- **RP:** role play.
- **CoT:** chain of thoughts.
- **TaB:** take a breather
- **ES:** emotional stimuli.
- **DS:** dynamic shot or RAG.

Resultados (2/4)

					Prompt Engineering (PE)				RAG		
Model	N° exp.	N° shots	Chunk size	T	RP	CoT	TaB	ES	DS	AUROC	AVGPREC
RL	1	-	-		-					79,1% ± 1,2%	39,3% ± 2,9%
L-GBM	2	-	-		-					81,7% ± 1,3%	45,2% ± 2,1%
GPT 4.0o (without fine tuning)	23	0	1	D	✓					69,6% ± 0,9%	24,6% ± 1,4%
	24	5	1	D	✓					70,2% ± 1,6%	18,4% ± 1,4%
	25	5	20	D	✓					75,3% ± 1,7%	28,9% ± 2,5%
GPT 4.0o mini (without fine tuning)	26	0	1	D	✓					69,6% ± 0,9%	24,6% ± 1,4%
	27	5	1	D	✓					65,7% ± 1,5%	19,6% ± 1,1%
	28	5	20	D	✓					70,0% ± 1,9%	19,6% ± 1,3%
GPT 3.5 (with fine tuning)	29	0	1	D	✓					80,2% ± 1,6%	40,7% ± 2,2%
	30	0	1	0	✓					80,3% ± 1,6%	40,7% ± 2,1%
GPT 4.0o mini (with fine tuning)	31	0	1	D	✓					80,1% ± 1,2%	40,8% ± 2,6%
	32	0	1	0	✓					80,0% ± 1,1%	40,7% ± 2,6%
Llama 3 8B (with fine tuning)	33	0	1	D	✓					78,9% ± 1,1%	37,7% ± 2,9%
	34	0	1	0	✓					73,1% ± 1,1%	21,4% ± 1,0%

Symbology:

T: temperature value, where "D" refers to the default setting.

Prompt strategies:

- **RP:** role play.
- **CoT:** chain of thoughts.
- **TaB:** take a breather
- **ES:** emotional stimuli.
- **DS:** dynamic shot or RAG.

Resultados (3/4)

Table 3: Standard Deviation of predicted Scores. “Normal” vs “Default” Predictions

Model	Label	Min	Q25	Q50	Avg.	Q75	Q90	Max
GPT 3.5 with fine tunning (temp. 0.5)	**Total**	**0,0%**	**0,0%**	**0,0%**	**0,6%**	**0,0%**	**0,5%**	**8,2%**
	Normal	0,0%	0,0%	0,0%	**0,3%**	0,0%	0,0%	7,0%
	Default	0,0%	0,0%	0,0%	**2,3%**	4,2%	6,4%	8,2%
GPT 3.5 with fine tunning (temp. 0)	**Total**	**0,0%**	**0,0%**	**0,0%**	**0,1%**	**0,0%**	**0,0%**	**3,5%**
	Normal	0,0%	0,0%	0,0%	**0,0%**	0,0%	0,0%	0,1%
	Default	0,0%	0,0%	0,0%	**0,4%**	0,2%	0,8%	3,5%
Llama 3 (8B) with fine tunning (temp. 0.5)	**Total**	**0,9%**	**9,5%**	**11,6%**	**10,9%**	**12,8%**	**13,4%**	**20,1%**
	Normal	3,1%	9,9%	11,7%	**11,2%**	12,8%	13,5%	20,1%
	Default	0,9%	7,7%	9,4%	**9,3%**	12,6%	12,9%	13,2%
Llama 3 (8B) with fine tunning (temp. 0)	**Total**	**0,0%**	**0,0%**	**0,0%**	**0,0%**	**0,0%**	**0,0%**	**0,0%**
	Normal	0,0%	0,0%	0,0%	**0,0%**	0,0%	0,0%	0,0%
	Default	0,0%	0,0%	0,0%	**0,0%**	0,0%	0,0%	0,0%

Resultados (4/4)

- Un experto humano tuvo la tarea de evaluar, en una escala de 1 (peor) a 7 (mejor), las justificaciones entregadas por los LLMs para una submuestra de 100 predicciones.
- La evaluación realizada por el experto humano arrojó un puntaje promedio de 5,5, con un intervalo de confianza del 95% entre 5,3 y 5,7.
- Estos resultados sugieren que GPT tiene el potencial de ofrecer análisis razonables en evaluaciones de riesgo crediticio.
- La supervisión humana sigue siendo crucial, ya que el modelo puede ocasionalmente cometer errores o presentar inconsistencias en ciertos casos.

Table 4: Example of reasons given by GPT for predicting an individual to be in default

VARIABLES	**EXPLANATIONS**
#ID=20522://V1=1.0//V2=0.0//V3=8.4844//V4=3.0//V5=0.5563//V6=0.5563//V7=1.3526//V8=3.0//V9=2.0//V10=0.0//	{r1:"High delinquency levels: The debtor has a delinquency of more than 90 days in the system and between 3 to 1 month before the observation (V10 = 0.0), which indicates a severe default risk. This is a strong indicator of financial distress and inability to meet debt obligations. Impact: High delinquency levels significantly increase the likelihood of default. Opposite direction: A value of 1 for V10 would have indicated a higher default risk.", r2:"Consistent delinquency history: The debtor has a high average delinquency in days in the bank in the last 3 months (V9 = 2.0) and a high delinquency in days in the bank in the last 6 months (V8 = 3.0), showing a pattern of ongoing payment issues. Impact: A consistent history of delinquency suggests a persistent inability to manage debt effectively. Opposite direction: Lower values for V9 and V8 would have implied a more stable payment behavior.", r3:"Lack of mortgage guarantee: The debtor does not have a mortgage guarantee (V2 = 0.0), which could indicate a lack of valuable collateral to secure the debt. Impact: Without a mortgage guarantee, the lender has less assurance of recovering the debt in case of default. Opposite direction: A value of 1 for V2 would have provided additional security for the lender and potentially lowered the default risk."}

Conclusiones

Rendimiento de los LLMs:

- Los LLMs ajustados (**fine-tuned**) igualan a los modelos tradicionales en evaluación de riesgo crediticio, alcanzando valores AUROC cercanos al 80% y AVGPREC alrededor del 40%.
- Rendimiento razonable de modelos fundacionales (sin ajuste). *Existe espacio para que pequeñas empresas implementen modelos de predicción de default con muy poco esfuerzo*.
- Impacto limitado de estrategias alternativas de prompt (“Take a Breather”, “Chain of Thought”, “Emotional Stimuli”).
- Impacto inesperado del tamaño de fragmento (“chunk size” o tamaño de lote), posiblemente relacionado con “few-shot indirecto”.

Estabilidad de resultados según temperatura:

- (Llama 3.1) muestra menor variabilidad de resultados cuando la temperatura = 0.
- (Llama 3.1) muestra mayor elasticidad a la temperatura en comparación con modelos como ChatGPT 3.5.

Explicabilidad:

- Los LLMs entregaron *insights* interpretables, ofreciendo alternativas rápidas a técnicas como SHAP, aunque requieren supervisión humana debido a inconsistencias ocasionales.

Implicancias regulatorias:

- La participación humana es crucial para la adopción de LLMs en finanzas, a fin de garantizar transparencia y cumplimiento regulatorio.

Limitaciones del estudio / Investigación futura:

- Basado en datos de Chile previos a la pandemia. Futuras investigaciones podrían explorar otras fuentes de datos no estructurados (por ejemplo, análisis de texto de evaluaciones de analistas), estudiar otros mercados, tipos de crédito y modelos más avanzados. También se debe examinar una nueva frontera: sistemas GenAI multiagente.

Can Large Language Models Predict Credit Risk? An Empirical Study on Consumer Loans in Chile

Diego Beas L.
División de Regulación de Solvencia y Liquidez - Bancos
CMF

David Díaz S.
Associate Professor, AI and Data Analytics, FEN
University of Chile

IX Conferencia Anual de la Comisión para el Mercado Financiero (CMF)
"100 años de supervisión bancaria en Chile“
Junio 2025

Anexo: variables

Name	Definition
V1	Corresponds to a binary variable that takes a value of 1 when the client has a delinquency of more than 30 days in the financial system and the month prior to the observation.
V2	Corresponds to a binary variable that takes a value of 1 when the client has a mortgage guarantee and 0 otherwise.
V3	Debt to income ratio for the month, without considering off balance sheet exposures and commercial debt. This definition is currently used in the Chilean framework for the determination of risk weighted asset for credit risk.
V4	Delinquency in days in the bank in the last 12 months according to the following coding: 0 equal to 0 days, 1 equal to between 1 to 30 days, 2 equal to between 31 to 60 days, 3 equal to between 61 to 89 days, and 4 or more is a delay greater than 90 days.
V5	Financial burden in the month, considering off balance sheet exposures but not commercial debt.
V6	Financial burden in the month, considering off balance sheet exposures and commercial debt.
V7	Corresponds to the ratio between the debt of the month of observation compared to the average of the last 12 months. It considers all the debt in the system.
V8	Delinquency in days in the bank in the last 6 months according to the following coding: 0 equal to 0 days, 1 equal to between 1 to 30 days, 2 equal to between 31 to 60 days, 3 equal to between 61 to 89 days, and 4 or more is a delay greater than 90 days.
V9	Average delinquency in days in the bank in the last 3 months according to the following coding: 0 equal to 0 days, less than or equal to 1 is between 1 to 30 days, less than or equal to 2 and greater than 1 is equal to between 31 to 60 days, less than or equal to 3 and greater than 2 is equal to between 61 to 89 days, and 3 or more is arrears greater than 90 days.
V10	Corresponds to a binary variable that takes a value of 1 when the client has a delinquency of more than 90 days in the system and between 3 to 1 month before the observation.