



FICHA TÉCNICA

Hackathon FIS 2025

1. INFORMACIÓN GENERAL

Nombre del reto: DEFAULT PREDICTION

Equipo: FYRONYX

Integrantes (4): Jean Carlos Reyes, Juan David Jojoa, Camilo Enrique Correa, Julián Andrés Arcila

Filiación institucional: Estudiantes pregrado

Programa académico - pregrado: Ingeniería de Sistemas

Fecha: 11 de septiembre de 2025

2. CONTEXTO Y PROBLEMA

Descripción del reto:

Este proyecto surge de una competencia lanzada por un banco en mayo de 2022. La propuesta era sencilla pero ambiciosa: utilizar todos los datos disponibles sobre sus clientes (comportamiento de pagos, perfiles, historial) para desarrollar un modelo que prediga de manera más efectiva que los existentes quiénes podrían caer en mora. Lo interesante es que no se impusieron restricciones metodológicas, lo que permitió explorar diversos enfoques. TIMI AMERICAS SAS planteó este desafío al identificar una oportunidad real para mejorar los modelos actuales. A pesar de que las entidades financieras ya cuentan con modelos operativos, la realidad es que la disponibilidad de datos está en constante aumento y las técnicas de aprendizaje automático han avanzado considerablemente.

Para este proyecto, decidimos utilizar TIMi Suite, que incluye Anatella para la limpieza y procesamiento de datos, y TIMi Modeler para la construcción del modelo predictivo. Estas herramientas están específicamente diseñadas para gestionar grandes conjuntos de datos, que es precisamente lo que requeríamos. La ventaja es que todo se integra, facilitando su posterior implementación en producción.

Nuestro objetivo final no es solo desarrollar un modelo que prediga con mayor precisión, sino también comprender realmente qué factores son más determinantes en el incumplimiento de pagos, así como explorar la posibilidad de segmentar la población de una manera que sea útil para el negocio.



Problema que se busca resolver:

El problema de fondo

Los bancos siempre han tenido que determinar a quién conceden préstamos y a quién no. Aunque parece sencillo, la realidad es que es extremadamente complejo. Los modelos que utilizan en la actualidad son funcionales, sí, pero presentan limitaciones significativas que, al final, generan costos para las entidades y reducen las oportunidades para los clientes. El asunto es que el comportamiento de las personas en relación con el crédito es mucho más intrincado de lo que los modelos tradicionales pueden captar. Existen nuevos patrones que emergen, especialmente tras eventos como la pandemia, cambios económicos o incluso alteraciones en las costumbres de pago. Los modelos antiguos no se adaptan con rapidez a estas transformaciones.

Los problemas específicos que encontramos

Errores en las predicciones: Cuando el modelo se equivoca, pasan dos cosas malas. Primera: le dicen que no a gente que en realidad sí iba a pagar (falsos positivos). Esto significa perder clientes buenos y negocio. Segunda: le aprueban crédito a gente que no va a pagar (falsos negativos). Esto significa pérdidas directas de dinero.

Procesos demasiado rígidos: La mayoría de los bancos tienen procesos muy estándar para aprobar créditos. Si tu puntaje está por encima de X, aprobado. Si está por debajo, negado. Pero la realidad es más matizada que eso. Hay gente con puntajes bajos que son excelentes pagadores por otras razones, y gente con puntajes altos que pueden tener problemas específicos.

Gestión reactiva en lugar de preventiva: Los modelos actuales son bastante rudimentarios para anticipar problemas. Generalmente, se percatan de que alguien enfrentará dificultades cuando ya es demasiado tarde. Si contáramos con mejores predicciones, se podrían implementar estrategias preventivas: contactar al cliente antes de que se retrase, ofrecerle opciones de pago, reestructurar la deuda, etc.

Limitaciones técnicas: Muchas entidades enfrentan problemas técnicos para procesar toda la información que podrían utilizar. Poseen datos en sistemas distintos que no se comunican entre sí, procesos de limpieza de datos que no son muy avanzados, y herramientas que no están diseñadas para el volumen de información que manejan.

Por qué esto importa para el negocio

Pérdidas económicas directas: Cada cliente que no paga representa una pérdida. Los costos de cobranza son altos, los procesos legales son largos y costosos, y al final mucha cartera vencida termina siendo irrecuperable.



Oportunidades perdidas: Por cada cliente bueno que rechazan por error, pierden no solo el negocio inmediato sino también la relación a largo plazo. Ese cliente probablemente se va con la competencia y es difícil recuperarlo después.

Ineficiencia operacional: Los recursos humanos y tecnológicos que se dedican a manejar cartera problemática podrían estar mejor utilizados en crecer el negocio con clientes buenos.

Experiencia del cliente: Los procesos lentos y poco personalizados afectan la satisfacción del cliente. En un mercado competitivo, esto puede ser la diferencia entre retener un cliente o perderlo.

ODS relacionados:

ODS 1: Fin de la Pobreza

Un modelo más preciso puede ayudar a identificar personas que tradicionalmente han estado excluidas del sistema financiero formal pero que en realidad sí pueden manejar crédito responsablemente. Mucha gente queda fuera del sistema no porque sea mal pagador, sino porque los modelos actuales no capturan bien su situación particular.

ODS 8: Trabajo Decente y Crecimiento Económico

Mejores decisiones crediticias significan que el capital financiero se asigna de manera más eficiente. El dinero llega a quien realmente lo va a usar productivamente y lo va a devolver. Esto estimula la actividad económica general.

ODS 9: Industria, Innovación e Infraestructura

Este proyecto es básicamente innovación aplicada al sector financiero. Usar herramientas como TIMi Suite para procesar grandes volúmenes de datos y construir modelos predictivos más sofisticados es exactamente el tipo de desarrollo tecnológico que fortalece la infraestructura industrial del país.

ODS 10: Reducción de las Desigualdades

Los modelos de crédito tradicionales muchas veces tienen sesgos implícitos que discriminan contra ciertos grupos de población. Un modelo más objetivo, basado en datos y patrones de comportamiento real, puede reducir estos sesgos.



ODS 16: Paz, Justicia e Instituciones Sólidas

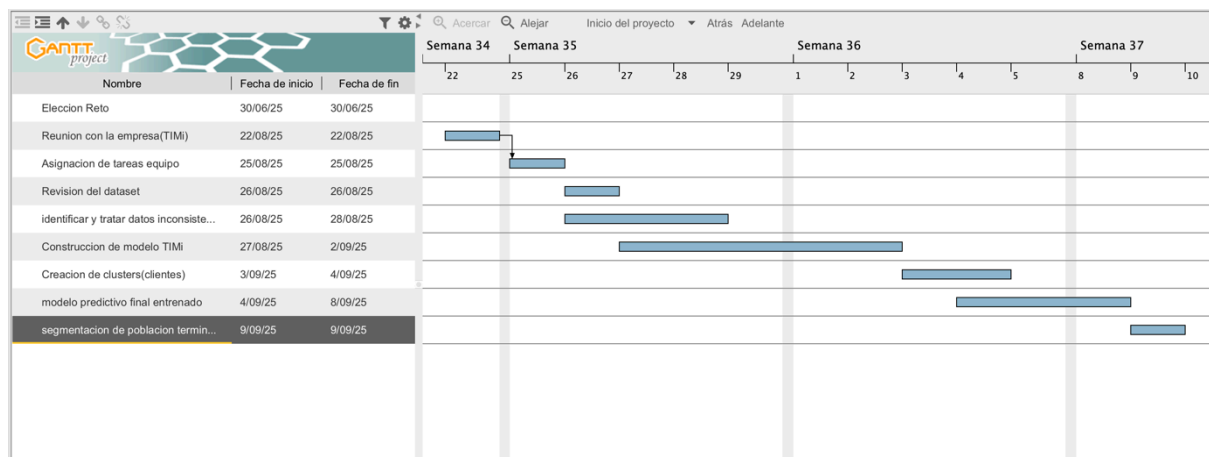
Tener procesos de evaluación crediticia más transparentes y objetivos contribuye a fortalecer la confianza en las instituciones financieras. Cuando las decisiones se basan en datos y criterios claros en lugar de sesgos o criterios arbitrarios, el sistema es más justo.

ODS 17: Alianzas para Lograr los Objetivos

Este proyecto es un ejemplo perfecto de colaboración entre diferentes sectores: academia, sector financiero, y empresa de tecnología. Este tipo de alianzas son clave para resolver problemas complejos que requieren diferentes tipos de experiencia.

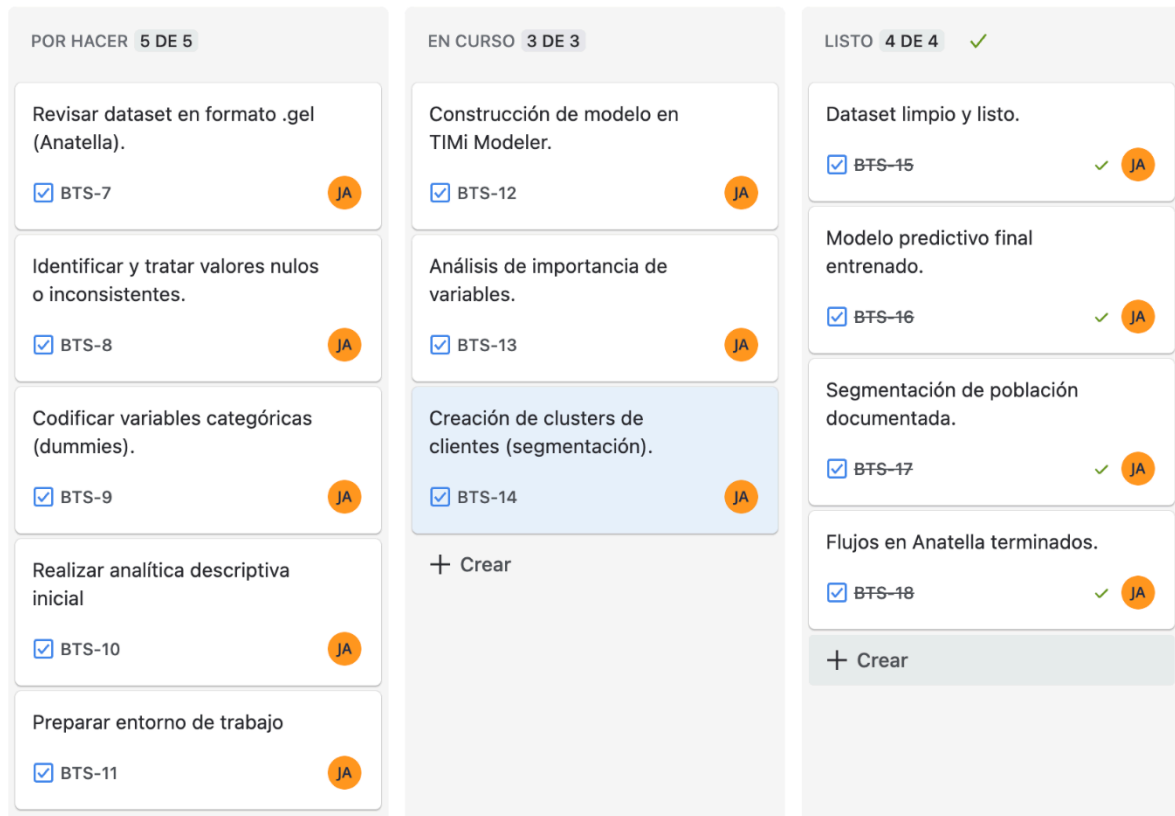
3. PLANEACIÓN DEL PROYECTO

Cronograma:





Tablero Kanban:



Metodología aplicada:

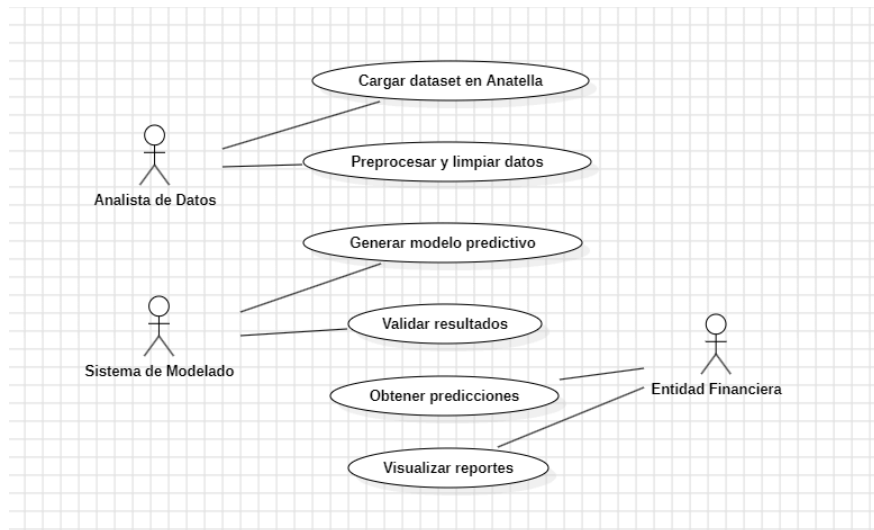
- **SCRUM:** para organizar el trabajo y mejorar la comunicación del equipo.



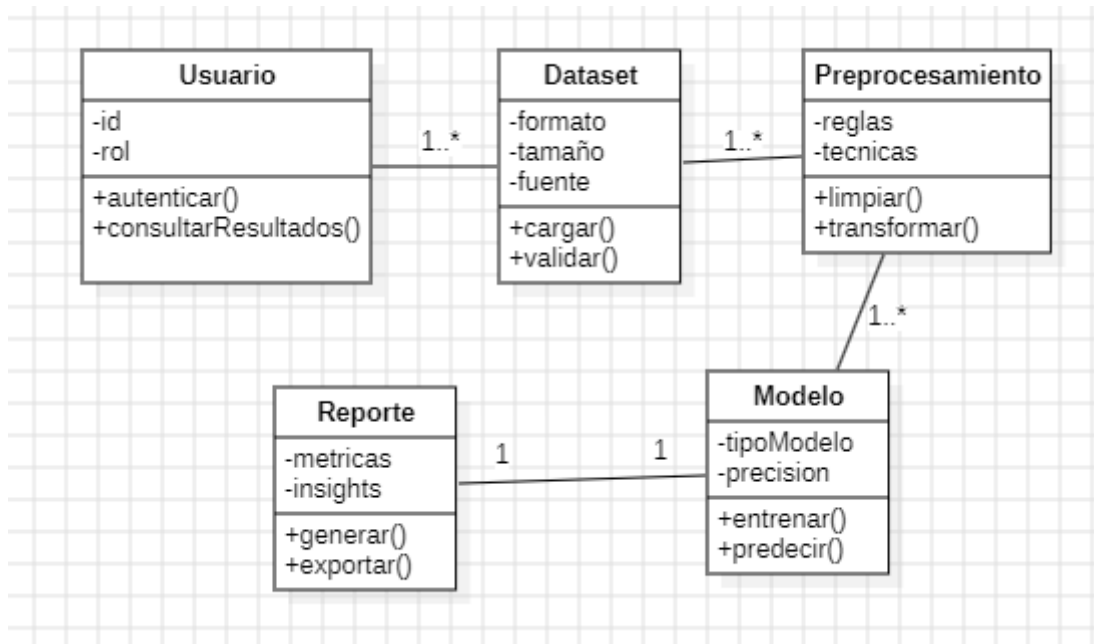
4. DISEÑO Y PROTOTIPO

Diagramas :

Casos de uso:

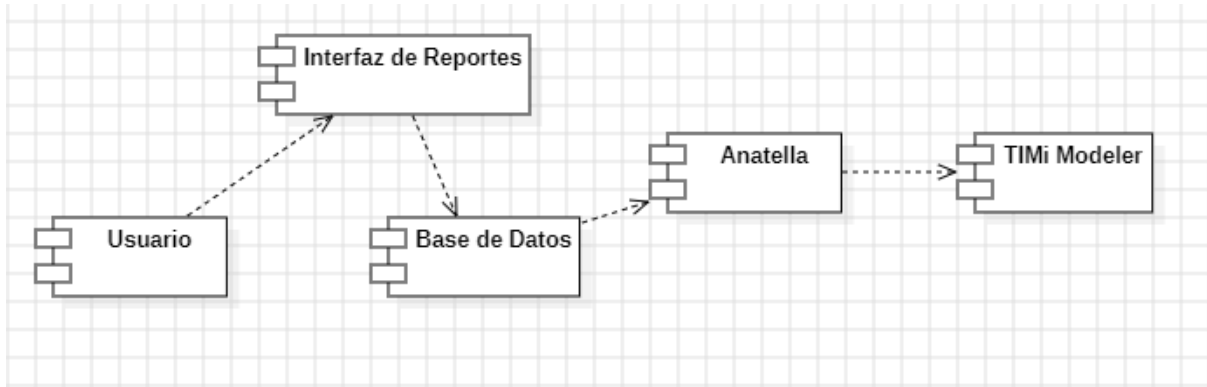


Clases:





Componentes:





Mockups de la solución / Modelo de datos / Representación del flujo de información:

Home Page:

Fyronyx

AcercaMetodologíaEquipoDashboard

Modelo Predictivo de Riesgo Crediticio

Un análisis avanzado de más de 5.5 millones de registros, diseñado para mejorar las decisiones de crédito con un modelo robusto e interpretable.

Explorar Dashboard

600 × 400

5.5M+ Registros procesados

AUC 0.84 Precisión en test

8 Variables finales

Acerca del proyecto

Este proyecto fue desarrollado en el marco de la Hackatón 2025-2 con el objetivo de construir un modelo de riesgo crediticio utilizando más de 5.5 millones de registros.

El modelo final logra un AUC de 0.84, lo que significa una alta capacidad para distinguir clientes buenos de clientes de alto riesgo.

Metodología

Depuración: de 825 variables iniciales → 185 relevantes → 8 finales.
Evitar fugas de información: eliminación de variables sospechosas (.MEAN con AUC > 0.9).
Validación: bootstrap y regularización para asegurar robustez.

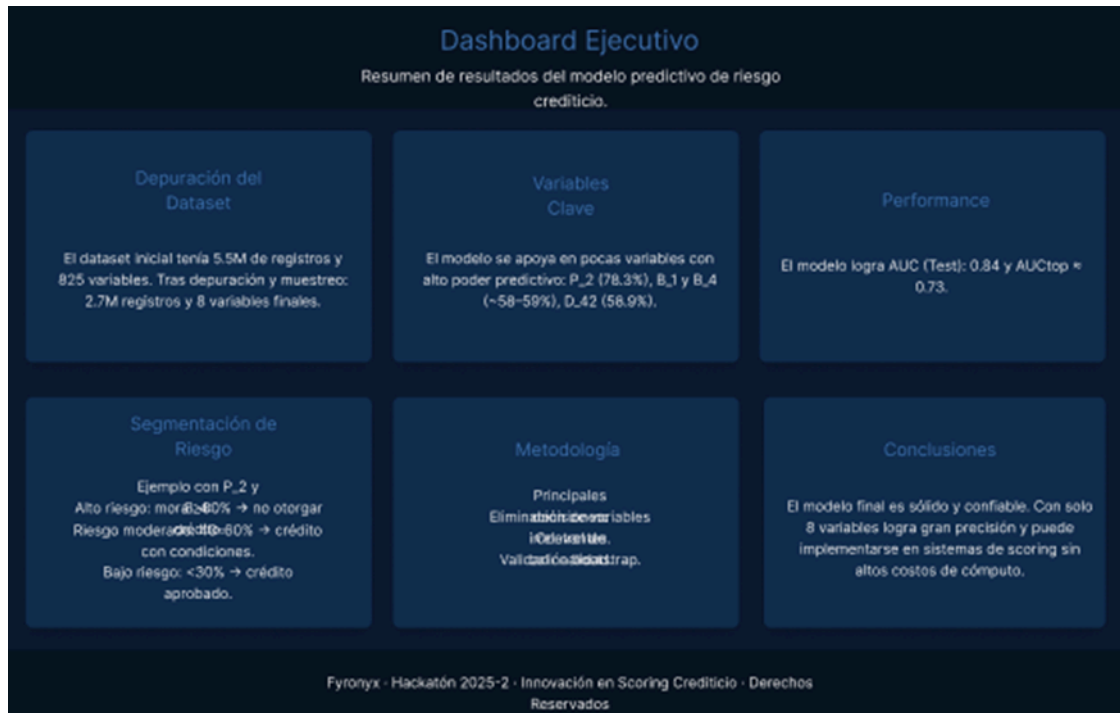
Equipo Fyronyx

Somos un grupo de estudiantes de Ingeniería de Sistemas con enfoque en analítica de datos y soluciones de negocio. Este proyecto busca demostrar capacidades técnicas con aplicabilidad real en el sector financiero.

Fyronyx · Hackatón 2025-2 · Innovación en Scoring Crediticio · Derechos Reservados



Dashboard:



Modelo de datos:

DataTable (11 363 762 rows - 190 columns) (complete)																			
	customer_ID	S_2	P_2	D_39	B_1	B_2	R_1	S_3	D_41	B_3	D_42	D_43	D_44	B_4	D_45	B_5	R_2		
6716252	97849dbbe70b5e65...	2019-03-15	0.6235	0.002373	0.06335	0.811	0.0044	0.1779	0.009346	0.00063	0.0839		0.00928	0.00896	0.01506	0.003084	0.006275	0.46	
6716253	97849dbbe70b5e65...	2019-04-27	0.66	0.00874	0.0006275	1.007	0.00667	0.1754	0.002495	0.0085	0.05466		0.00524	0.005306	0.02151	0.02356	0.006527	0.4;	
6716254	9784aa1ecb0deb9...	2018-10-22	0.4973	0.2065	0.893	0.0282	0.2507	0.2312	0.003504	0.815		0.2805	0.6313	0.802	0.345	0.02089	0.00767	0.96	
6716255	9784aa1ecb0deb9...	2018-11-22	0.5034	0.00399	0.8633	0.02078	0.256	0.2405	0.00814	0.898		0.2426	0.63	0.9146	0.355	0.02249	0.002262	0.75	
6716256	9784aa1ecb0deb9...	2018-12-22	0.47	0.00931	0.864	0.02248	0.2532	0.2374	0.001629	0.974		0.2155	0.757	0.9233	0.355	0.01596	0.00892	0.9;	
6716257	9784aa1ecb0deb9...	2019-01-22	0.4866	0.2112	0.8794	0.0255	0.2554	0.2292	0.004505	0.9946		0.1996	1.01	0.9927	0.351	0.02266	0.0003023	1.1;	
6716258	9784aa1ecb0deb9...	2019-02-27	0.4868	0.005783	0.8623	0.02768	0.258	0.2335	0.00883	1.013		0.1936	1.001	1.069	0.3545	0.01685	0.00957	1.0;	
6716259	9784aa1ecb0deb9...	2019-03-22	0.5015	0.2119	0.8755	0.0281	0.2593	0.2339	0.005863	1.054		0.1813	0.6255	1.095	0.3599	0.01648	0.000533	0.96	
6716260	9784aa1ecb0deb9...	2019-04-13	0.5234	0.001046	0.8545	0.02316	0.2593	0.1658	0.003473	1.048		0.175	0.6343	1.095	0.3572	0.01794	0.008705	0.86	
6716261	9784aa1ecb0deb9...	2019-05-22	0.4917	0.001942	0.848	0.02798	0.7573	0.1583	0.003544	1.063		0.1704	0.878	0.9487	0.362	0.02441	0.003784	0.8;	
6716262	9784aa1ecb0deb9...	2019-06-30	0.549	0.0004187	0.908	0.0241	0.758	0.2115	0.004547	1.053		0.1603	0.7583	0.951	0.3643	0.01813	0.004387	0.96	
6716263	9784aa1ecb0deb9...	2019-07-22	0.545	0.00538	0.9014	0.02193	0.754	0.2064	0.003164	1.037		0.1578	0.758	0.989	0.3674	0.01698	0.00877	0.7;	
6716264	9784aa1ecb0deb9...	2019-08-14	0.549	0.005585	0.913	0.02383	0.7573	0.1871	0.007133	1.044		0.149	0.7593	0.9775	0.368	0.01671	0.007744	0.7;	
6716265	9784aa1ecb0deb9...	2019-09-29	0.555	0.006252	0.9077	0.02176	0.7563	0.1763	0.002298	1.039		0.0887	0.751	0.988	0.3726	0.01483	0.005516	0.7	
6716266	9784aa1ecb0deb9...	2019-10-04	0.5537	0.004128	0.913	0.02037	0.7505	0.1807	0.001025	1.044		0.0889	0.752	0.9893	0.3757	0.01595	0.004166	0.6;	
6716267	9784bdf9a8f638489...	2018-10-03	0.3599	0.4805	0.0495	1.001	0.003891	0.1627	0.004993	0.0168	0.1925	0.713	0.3792	0.159	0.01971	0.005787	0.001753	0.4;	
6716268	9784bdf9a8f638489...	2018-11-17	0.3423	0.0372	0.02545	0.4255	0.003065	0.1353	0.00747	0.0237	0.1973	0.532	0.3772	0.1821	0.01825	0.01204	0.002615	0.5;	
6716269	9784bdf9a8f638489...	2018-12-13	0.375	0.00292	0.05362	0.4292	0.00815	0.169	0.008736	0.02153	0.1918	0.4844	0.3813	0.1888	0.0249	0.0119	0.008	0.5;	
6716270	9784bdf9a8f638489...	2019-01-19	0.3853	0.03635	0.03952	0.3064	0.004303	0.1648	0.005005	0.02733	0.1873	0.4312	0.377	0.1799	0.03354	0.01819	0.00795	0.4;	
Copy Copy with Column Names #Row: 50 Go																			

Dataset contenedor de registros bancarios.

D_* = Delinquency variables (variables de morosidad / incumplimiento)

S_* = Spend variables (variables de gasto)

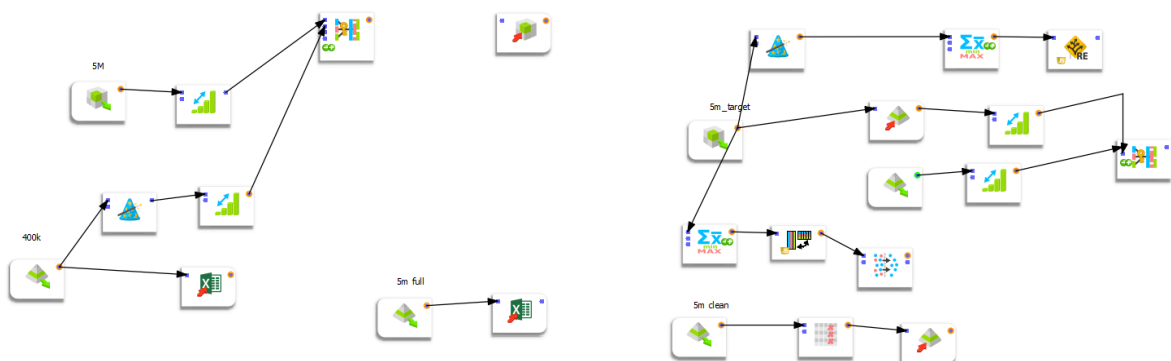
P_* = Payment variables (variables de pago)



DataTable (458 913 rows - 2 columns) (complete)		
	customer_ID	target
1	0000099d6bd597052cdcd90ffabf56573fe9d7c79be5fbac11a8ed792feb62a	0
2	00000fd6641609c6ece5454664794f0340ad84ddce9a267a310b5ae68e9d8e5	0
3	00001b22f846c82c51f6e3958ccd81970162bae8b007e80662ef27519fcc18c1	0
4	000041bd8a6ecadd89a52d11886e8eaaec9325906c9723355abb5ca523658edc	0
5	00007889e4fcd2614b6cbe7f8f3d2e5c728eca32d9eb8ad51ca8b8c4a24cefed	0
6	000084e5023181993c2e1b665ac88dbb1ce9ef621ec5370150fc2f8bdca6202c	0
7	000098081fde4fd64bc4d503a5d6f86a0aedc425c96f5235f98b0f47c9d7d8d4	0
8	0000d17a1447b25a01e42e1ac56b091bb7cbb06317be4cb59b50fec59e0b6381	0
9	0000f99513770170a1aba690daeeb8a96da4a39f11fc27da5c30a79db61c1e85	1
10	00013181a0c5fc8f1ea38cd2b90fe8ad2fa8cad9d9f13e4063bdf6b0f7d51eb6	1
11	0001337ded4e1c2539d1a78ff44a457bd4a95caa55ba1730b2849b92ea687f9e	1
12	00013c6e1cec7c21bede7cb319f1e28eb994f5625257f479c53ad6e90c177f7c	1

DataTable (5 531 451 rows - 24 columns) (complete)																			
	[1A] customer_ID	B_38	D_114	B_2	B_3	B_7	B_9	B_22	D_131	D_44	D_46	D_48	D_51	D_52	D_62	D_75	D_77		
988	000cc98607442c5074d...	6	1	0.0808	0.3962	0.3357	0.5513	0.00353	0.9824	0.2507	0.4465	0.629	0.003006	0.00252	0.002047	0.1356		0	
989	000cfb5aac8db5018589...	3	1	0.7656	0.011566	0.1305	0.03934	0.001982	0.005295	0.3845	0.716	0.666	0.00568	0.3198	0.1744	0.4724	0.1621	0	
990	000cfb5aac8db5018589...	2	1	0.9365	0.01164	0.1621	0.04138	0.007122	0.00725	0.2502	0.4858	0.5693	0.335	0.3215	0.1775	0.408	0.1692	0	
991	000cfb5aac8db5018589...	3	1	0.8867	0.01045	0.1192	0.0281	0.00449	0.007538	0.259	0.6465	0.6265	0.3389	0.3206	0.1805	0.4033	0.1631	0	
992	000cfb5aac8db5018589...	3	1	0.8867	0.010544	0.1053	0.03894	0.007015	0.007046	0.2502	0.6543	0.632	0.3376	0.321	0.1754	0.406	0.161	0	
993	000cfb5aac8db5018589...	2	1	1.009	0.01277	0.1278	0.05624	0.002346	0.00795	0.2554	0.4888	0.622	0.3335	0.32	0.1781	0.401	0.165	0	
994	000cfb5aac8db5018589...	3	1	0.751	0.014824	0.1807	0.0982	0.00754	0.001063	0.3838	0.3005	0.618	0.343	0.3186	0.171	0.401	0.166	0	
995	000cfb5aac8db5018589...	2	1	1.001	0.00605	0.1382	0.0651	0.001341	0.00542	0.2534	0.3223	0.6396	0.3403	0.3184	0.1725	0.3335	0.1643	0	
996	000cfb5aac8db5018589...	2	1	1.005	0.002903	0.1466	0.04822	0.007496	0.00405	0.2563	0.332	0.642	0.3425	0.3228	0.01165	0.4023	0.009315	0	
997	000cfb5aac8db5018589...	2	1	1.003	0.00657	0.1432	0.03314	0.001717	0.00797	0.255	0.3286	0.6406	0.3413	0.3162	0.0134	0.4053	0.01475	0	
998	000cfb5aac8db5018589...	2	1	1.0	0.007122	0.1428	0.1265	0.007145	0.006855	0.1329	0.0749	0.549	0.3406	0.3152	0.1794	0.3347	0.1698	0	

Integración en Anatella y preparación del dataset final.



Entrenamiento del modelo en TIMi Modeler:



TIMi Modeler Config - D:/_User/Desktop/SEXTO SEMESTRE/HACKATON/MODELER/FULLDATASET_TARGET_J.CfgXML

File Tools Additional panels Windows Help

TIMI Select Variables Select Lines / Segment Target Definition CREATE MODEL Apply model Execution Log

Filter 1 ☐ Activate ** Invert ☐ Filter 2 ☐ Activate ** Invert ☐ Filter 3 ☐ activate List Var Invert ☐ 825

Activate/Ignore Recoding - Monotonicity Discretization - Penalization

Recoding	Iniv.Imp	Variable Name	Modality	Group	Monotonicity	Ignore	Illustrativ
Dummy	12.09	D_124	missing	VALUE	N/A	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Dummy	12.09	D_124	0.04544 <x<= 0.05273	VALUE	N/A	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Dummy	12.09	D_124	0.0909 <x<= 0.0982	VALUE	N/A	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Dummy	12.09	D_124	0.1009 <x<= 0.1436	VALUE	N/A	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Dummy	12.09	D_124	0.1818 <x<= 0.1891	VALUE	N/A	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Dummy	12.09	D_124	0.1918 <x<= 0.2345	VALUE	N/A	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Dummy	12.09	D_124	0.2373 <x<= 0.2798	VALUE	N/A	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Dummy	12.09	D_124	0.318 <x<= 0.3254	VALUE	N/A	<input checked="" type="checkbox"/>	<input type="checkbox"/>
RTT	7.46	D_125		VALUE	none	<input type="checkbox"/>	<input type="checkbox"/>
PASSTH...	7.46	D_125		VALUE	N/A	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Dummy	7.46	D_125	missing	VALUE	N/A	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Dummy	7.46	D_125	0 <x<= 0.01	VALUE	N/A	<input checked="" type="checkbox"/>	<input type="checkbox"/>

Select All Recoding: Monotonicity:

TIMi Modeler Config - D:/_User/Desktop/SEXTO SEMESTRE/HACKATON/MODELER/FULLDATASET_TARGET_J.CfgXML

File Tools Additional panels Windows Help

TIMI Select Variables Select Lines / Segment Target Definition CREATE MODEL Apply model Execution Log

Line/Segment Selection

☒ Boolean Expression ☐ External file ☐ StarDust Segmentation Model

Standard Parameters Help

All Columns:

Column Name	Type
B_30	V
B_31	V
B_38	V
D_114	V
D_116	V
D_117	V
D_120	V
D_126	V

Filter: *

??? parser

☐ add '_n_' as input var

Functions on Numbr

Expression: ☐ Truncate table on first failed row

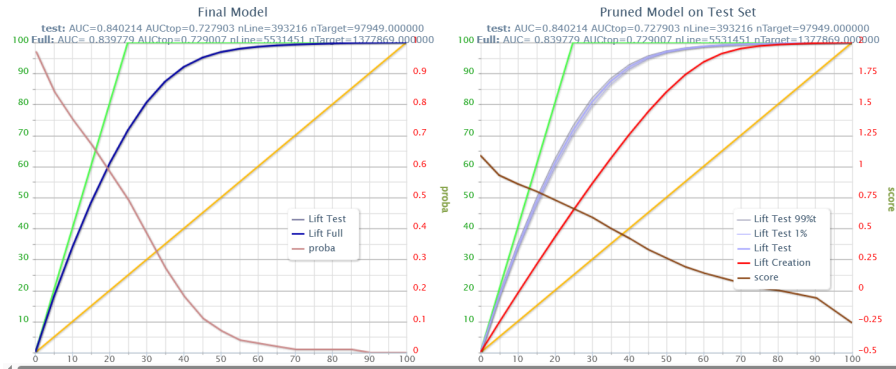
Debug ☒ Value (debug):

Resultados (AUC 84%):

Hackathon FIS 2025

	[%]	[%]	Target [%]	Model [%]	RMSE	AUROC [79]	RMSE	Discrim.	INDIVIDUALS
1P_2	6.56	78.30	22.64	100.00	1.59		208.48	-0.459<v<=-0.1727	
2B_1	3.39	59.36	36.38	43.68	35.65		193.42	1.08<v<=1.1	
3D_45	3.16	44.98	21.55	33.51	27.38		157.09	0.02313<v<=0.03078	
4B_4	2.91	57.12	51.88	32.15	28.56		193.99	1.628<v<=19.8	
5D_41	2.91	19.74	5.58	15.77	87.80		205.48	2.148<v<=8.99	
6D_43	2.89	39.29	51.88	18.26	51.81		198.33	1.913<v<=10.11	
7D_42	2.67	58.88	8.78	22.61	7.56		198.26	0.9033<v<=1.132	
8D_64	2.24	23.92	36.38	10.86	75.04		137.94	v=-1	

Lifts



Activar Windows
Ve a Configuración para activar Windows.



5. IMPLEMENTACIÓN

Arquitectura propuesta:

Enfoque arquitectónico general

Para este proyecto, optamos por implementar una arquitectura fundamentada en el ecosistema TIMi Suite, lo que nos permite gestionar todo el pipeline de datos de forma integrada. La idea principal era establecer un flujo de trabajo sólido que abarque desde los datos en bruto hasta un modelo de predicción listo para su producción, todo dentro de un entorno controlado y escalable. La arquitectura que proponemos sigue una estructura de capas bien definidas, donde cada capa tiene una responsabilidad específica y puede evolucionar de manera independiente. Esto es crucial, ya que en proyectos de machine learning, las circunstancias cambian considerablemente durante el desarrollo, y se requiere flexibilidad para realizar ajustes sin comprometer la integridad del sistema.

Tecnologías principales

TIMi Suite como plataforma central Toda la implementación se centra en TIMi Suite, que comprende dos componentes principales:

Anatella: Esta herramienta se encarga de todo el procesamiento de datos. Es extremadamente eficiente para trabajar con grandes conjuntos de datos, ya que fue diseñada específicamente para tal fin. La ventaja radica en su capacidad para procesar millones de registros sin enfrentar problemas de memoria o rendimiento.

TIMi Modeler: En esta parte es donde construimos y entrenamos los modelos predictivos. Posee algoritmos especializados para problemas de clasificación como el nuestro, y lo destacable es que está optimizado para trabajar con los datos generados por Anatella.

La decisión de utilizar TIMi Suite no se basó únicamente en sus capacidades técnicas, sino también en su diseño orientado a implementaciones industriales. Muchas herramientas de machine learning funcionan adecuadamente para prototipos, pero se complican cuando se requiere escalar o implementar algo en producción.

Formato de datos: Los datos se presentan en formato .gel, que es el formato nativo de Anatella. Esto simplifica considerablemente las cosas, ya que no tenemos que preocuparnos por conversiones de formato o problemas de compatibilidad. Anatella puede leer estos archivos directamente y procesarlos de manera muy eficiente.

Capa de datos crudos: En la base se encuentran los datos tal como los proporciona la entidad bancaria. Esta información incluye el perfil de los clientes, su comportamiento en los pagos y datos temporales de las transacciones. Todo está anonimizado, por supuesto, pero mantiene la estructura necesaria para el análisis.



Esta capa es de solo lectura. Los datos originales nunca se alteran, lo cual es una buena práctica, ya que siempre puedes regresar al punto de partida si algo falla en el procesamiento.

Capa de procesamiento de datos: Aquí es donde Anatella realiza su labor. Esta capa se ocupa de:

Limpieza de datos: identificar y gestionar valores nulos, inconsistencias y outliers extremos.

Transformaciones: crear variables dummy para categóricas, normalizar variables numéricas y establecer ventanas temporales.

Feature engineering: generar nuevas variables que puedan ser útiles para el modelo, como ratios, promedios móviles e indicadores de tendencia.

Validación de calidad: garantizar que los datos transformados tengan sentido y se encuentren dentro de rangos esperados.

Lo interesante de Anatella es que todo esto se lleva a cabo mediante flujos visuales. En lugar de escribir código, arrastras y conectas bloques que representan diferentes operaciones. Esto hace que el proceso sea más transparente y fácil de documentar.

Capa de modelado: TIMi Modeler toma los datos procesados y construye los modelos predictivos. Esta capa incluye:

Partición de datos: separar automáticamente en conjuntos de entrenamiento, validación y prueba.

Entrenamiento de modelos: probar diferentes algoritmos y configuraciones.

Evaluación: calcular métricas de desempeño, matrices de confusión y curvas ROC.

Selección de variables: identificar qué variables son más relevantes para la predicción.

Optimización: ajustar hiperparámetros para mejorar el rendimiento.

Una ventaja de TIMi Modeler es que maneja automáticamente muchos aspectos técnicos del modelado, como el balanceo de clases (importante en problemas de default donde la mayoría de los clientes sí pagan) y la validación cruzada.

Capa de análisis y segmentación: Además del modelo principal, hemos implementado un análisis de segmentación para comprender mejor a la población. Esto abarca:

Clustering de clientes basado en patrones de comportamiento.

Análisis de la importancia de variables por segmento.

Identificación de perfiles de riesgo específicos.



Recomendaciones diferenciadas por cluster.

Capa de resultados: Los resultados del sistema incluyen:

Modelo predictivo que ha sido entrenado y validado.

Scores de probabilidad de default para cada cliente.

Análisis de la importancia de variables.

Segmentación de la población.

Reportes sobre el desempeño del modelo.

Consideraciones de seguridad

A pesar de que trabajamos con datos anonimizados, hemos implementado diversas medidas de seguridad:

Aislamiento del ambiente: Todo el procesamiento se realiza dentro del entorno TIMi Suite, sin exportar datos intermedios a sistemas externos.

Control de acceso: Solo el equipo del proyecto tiene acceso a los datos y modelos. TIMi Suite permite configurar permisos detallados.

Trazabilidad: Todos los flujos de Anatella y los modelos de TIMi Modeler se documentan automáticamente, por lo que siempre sabemos qué transformaciones se aplicaron y cuándo.

Versionado: Mantenemos versiones de los flujos y modelos para poder reproducir resultados y realizar un rollback si es necesario.

Escalabilidad y rendimiento

La arquitectura está hecha para manejar el volumen de datos que maneja una entidad bancaria real:

Procesamiento eficiente: Anatella está optimizada para conjuntos de datos grandes. Puede procesar millones de registros utilizando la memoria de manera inteligente.

Paralelización: Tanto Anatella como TIMi Modeler pueden aprovechar múltiples núcleos para acelerar el procesamiento.

Optimización de memoria: El formato .gel es muy eficiente en cuanto a espacio, y las herramientas están diseñadas para no cargar todo en memoria al mismo tiempo.

Modularidad: Los flujos están divididos en módulos que se pueden ejecutar de forma independiente. Esto permite optimizar partes específicas sin afectar el resto.

Integración y deployment



Una ventaja clave de usar TIMi Suite es que está diseñado para entornos de producción:

APIs nativas: TIMi Suite puede exponer los modelos como servicios web, facilitando la integración con sistemas bancarios existentes.

Procesamiento por lotes: Los modelos pueden ejecutarse en lotes para manejar grandes volúmenes de solicitudes de crédito.

Monitoreo: La plataforma incluye herramientas para supervisar el rendimiento de los modelos en producción y detectar cuándo necesitan reentrenamiento.

Gobernanza: Hay controles para versionar modelos, documentar cambios y mantener auditorías de las decisiones del modelo.

Ventajas de esta arquitectura

Simplicidad: Al usar una plataforma integrada, evitamos problemas de compatibilidad entre herramientas diferentes.

Rendimiento: Las herramientas están optimizadas para el tipo de datos y volúmenes que manejamos.

Mantenibilidad: Los flujos visuales de Anatella son más fáciles de entender y modificar que código tradicional.

Escalabilidad: La arquitectura puede crecer con las necesidades del negocio sin requerir cambios fundamentales.

Tiempo de implementación: Al usar herramientas especializadas, el desarrollo es más rápido que construir todo desde cero.

Esta arquitectura nos permitió enfocarnos en resolver el problema de negocio (predecir defaults) en lugar de gastar tiempo en problemas técnicos de infraestructura e integración.

6. ESTÁNDARES Y BUENAS PRÁCTICAS

Normas ISO/IEEE aplicadas

Para el desarrollo del proyecto *DEFAULT PREDICTION* se han considerado diferentes normas internacionales aplicables al ámbito del software y los sistemas. Tras el análisis, se determinó que las más adecuadas para este caso son **ISO/IEC 25010** e **ISO/IEC/IEEE 29148**, debido a su relevancia directa con los objetivos y necesidades del reto planteado.



En primer lugar, la norma **ISO/IEC 25010** resulta pertinente ya que establece un modelo de calidad del software basado en características fundamentales como **funcionalidad, fiabilidad, usabilidad, seguridad, mantenibilidad y eficiencia**. Dado que el proyecto involucra el desarrollo de un modelo de machine learning para la predicción de incumplimiento crediticio, es indispensable garantizar no solo la **precisión de las predicciones**, sino también la **seguridad en el manejo de datos sensibles**, la **robustez del sistema en diferentes escenarios** y la **capacidad de actualización del modelo** frente a la evolución de la información disponible.

Por otro lado, la norma **ISO/IEC/IEEE 29148** es clave para la **definición y documentación de requisitos** tanto funcionales como no funcionales del sistema. Esta norma permite estructurar de manera clara y verificable aspectos como:

- **Requisitos funcionales**, por ejemplo, que el sistema sea capaz de calcular la probabilidad de incumplimiento de un cliente.
- **Requisitos no funcionales**, como el tiempo máximo de procesamiento de solicitudes o el nivel mínimo de precisión esperado en las predicciones.

Prácticas ágiles adoptadas:

En primer lugar, se implementó la práctica de **gestión visual mediante un tablero Kanban**, donde las tareas se organizaron en columnas de *Por hacer*, *En progreso*, *En revisión* y *Hecho*. Este enfoque permitió al equipo mantener un control constante sobre el estado de las actividades, identificar bloqueos tempranos y priorizar las tareas críticas del proyecto, como la limpieza del dataset, el entrenamiento del modelo y la validación de métricas.

Adicionalmente, se realizaron **iteraciones cortas de trabajo**, similares a *sprints*, en las que se definieron objetivos específicos para cada bloque de días hábiles. Esto facilitó la entrega incremental de resultados, por ejemplo, tener el dataset limpio en una fase intermedia, contar con un modelo baseline en una etapa posterior y culminar con el modelo ajustado y validado para la entrega final.

También se aplicó la práctica de **retroalimentación continua**, revisando avances en cada fase antes de pasar a la siguiente. Este enfoque evitó retrabajos y garantizó que los entregables cumplieran con los requisitos funcionales y no funcionales previamente definidos.

Finalmente, se promovió la **colaboración activa entre los integrantes del equipo**, asignando responsables por tareas y fomentando la comunicación constante. Esto permitió optimizar el uso del tiempo disponible y asegurar que cada entregable estuviera alineado con los objetivos del reto.



7. RESULTADOS ESPERADOS

Producto Mínimo Viable (MVP):

Componentes técnicos entregables:

Modelo predictivo funcional: Un modelo de machine learning implementado en TIMi Modeler que tome como input los datos de un cliente (perfil, historial de pagos, comportamiento transaccional) y devuelva como output un score de probabilidad de default junto con la clasificación de riesgo (alto, medio, bajo). Este modelo debe demostrar una mejora estadísticamente significativa comparado con los modelos actuales en métricas como AUC, precisión y recall, y debe ser capaz de procesar tanto solicitudes individuales como lotes masivos de evaluación.

Pipeline de datos automatizado: Flujos completos implementados en Anatella que transformen automáticamente los datos crudos (.gel) en el formato requerido por el modelo. Esto incluye procesos de limpieza de datos, manejo de valores nulos, creación de variables dummy para categóricas, normalización, feature engineering (creación de ratios, promedios móviles, indicadores de tendencia), y validación de calidad. Los flujos deben ser reutilizables y ejecutables de manera automática cuando lleguen datos nuevos.

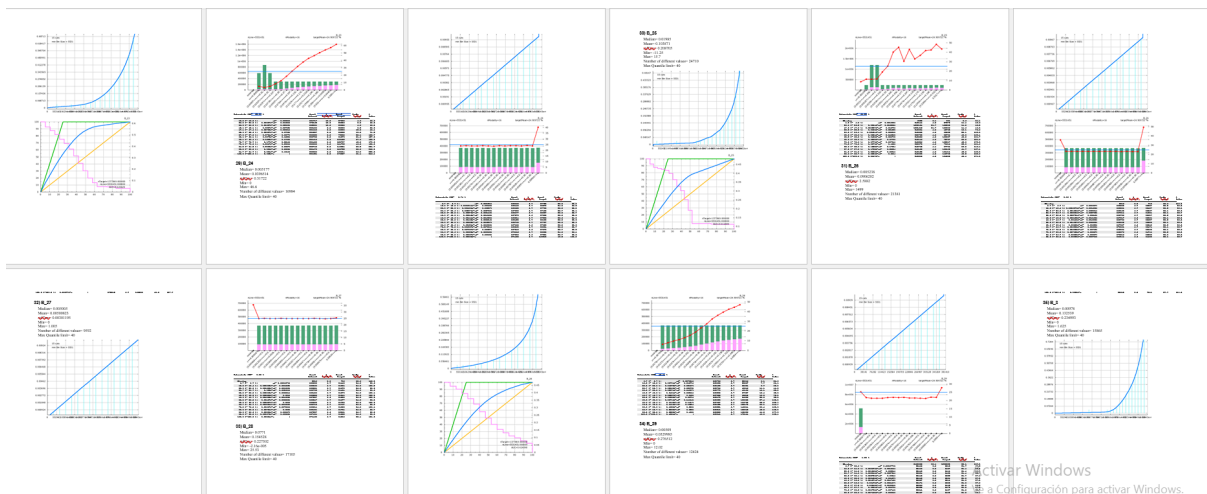
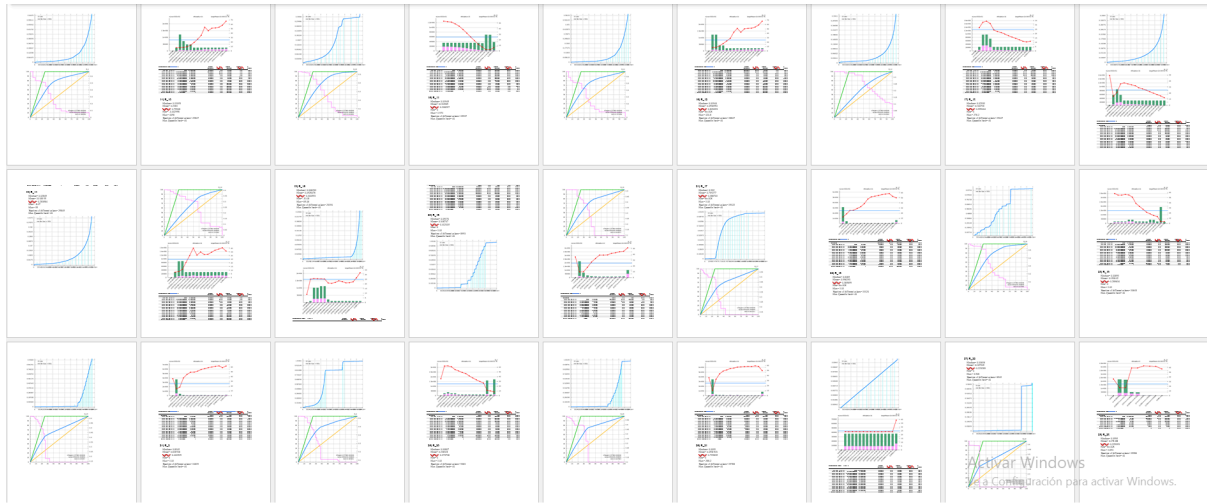
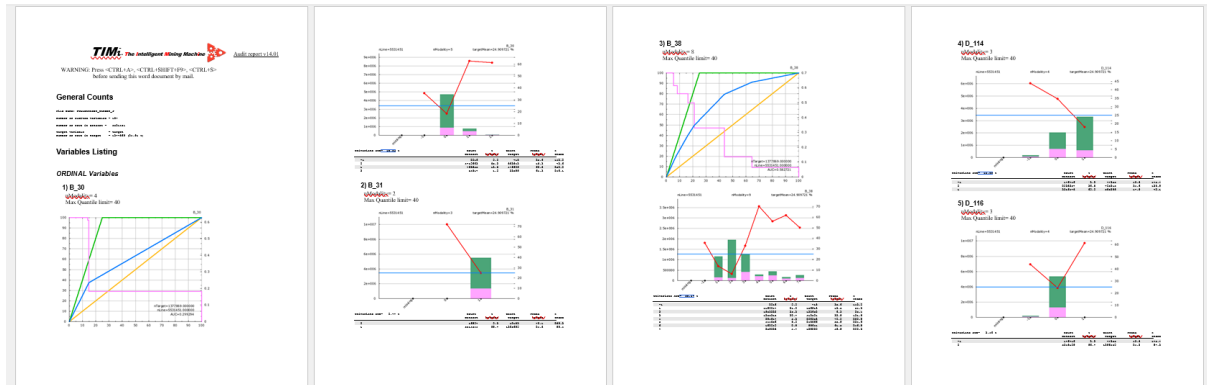
Sistema de segmentación de clientes: Análisis de clustering implementado que agrupe clientes en segmentos homogéneos basados en patrones de comportamiento crediticio. Cada segmento debe tener características definidas, perfiles de riesgo específicos, y recomendaciones diferenciadas de tratamiento comercial. Este sistema debe integrar con el modelo predictivo para ofrecer predicciones personalizadas por segmento.

Documentación y entregables: Manual técnico de operación del sistema, reporte ejecutivo con hallazgos principales y recomendaciones de implementación, análisis de importancia de variables con interpretación de negocio, validación estadística del modelo con métricas comparativas, y presentación final con insights y aplicaciones prácticas de los resultados.

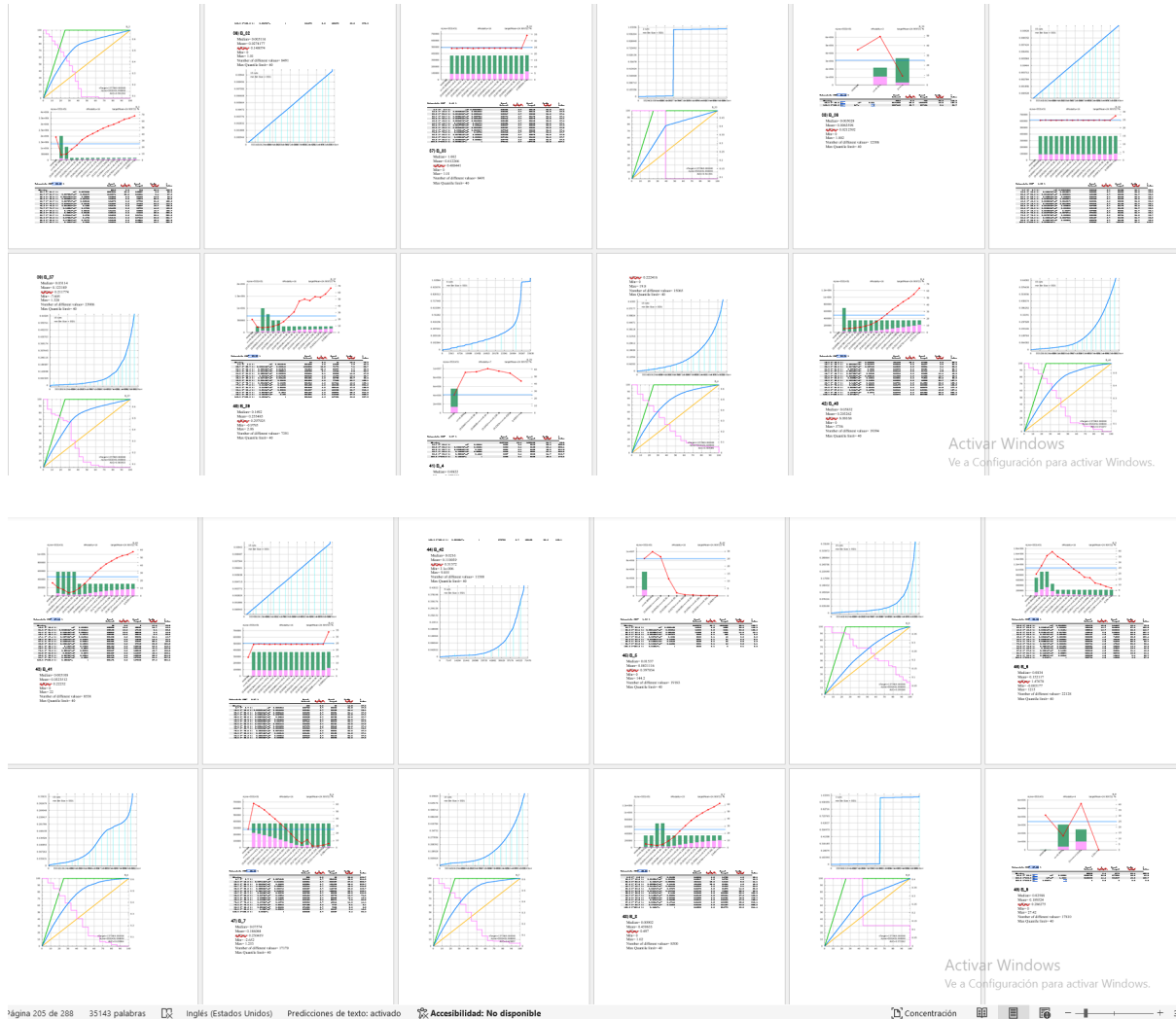
Evidencias:

Análisis individual de cada variable.

Hackathon FIS 2025



Hackathon FIS 2025

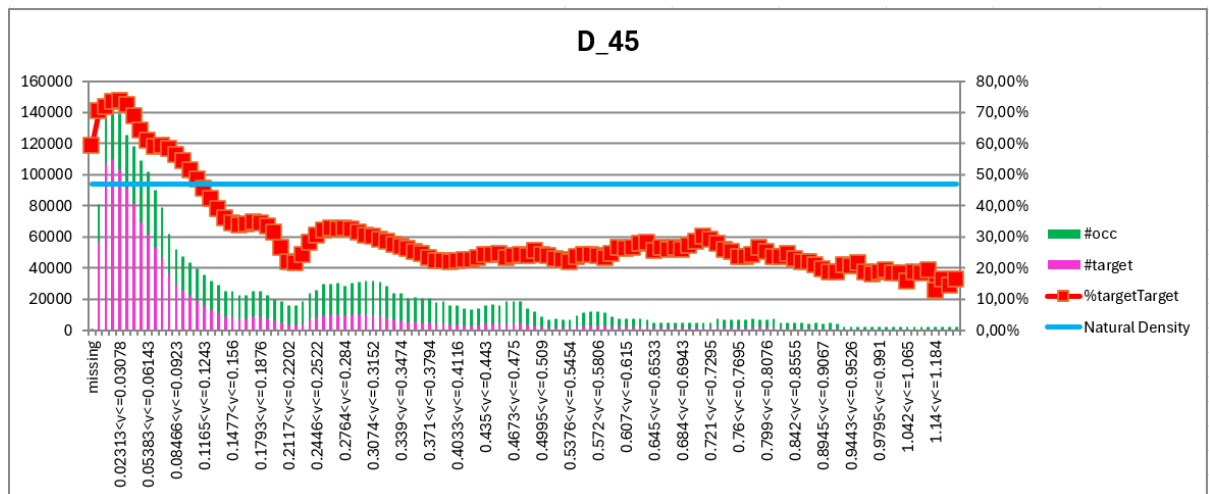
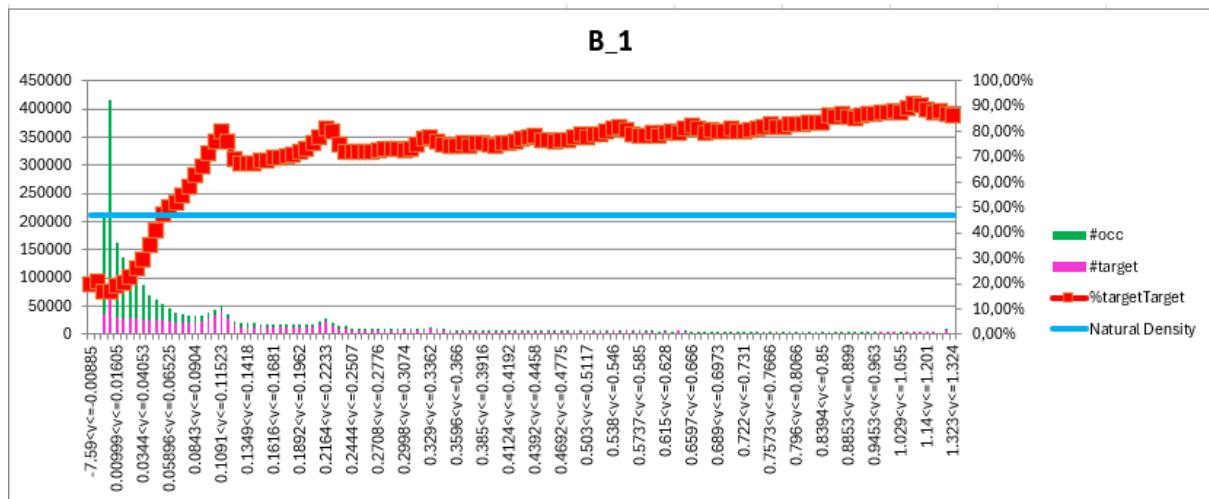
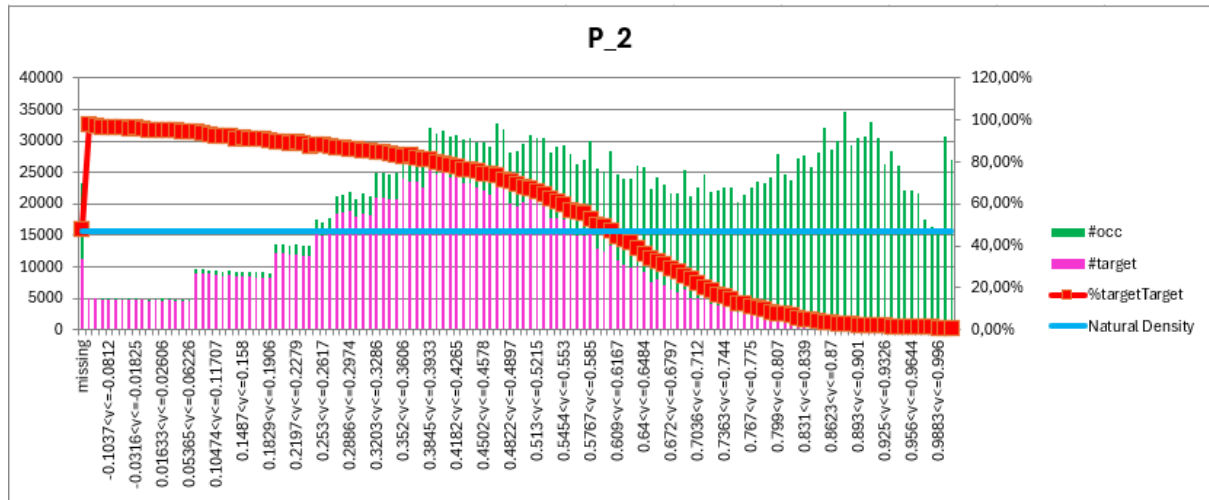


Variables más importantes:

Variable Type	Variable Name	Multi.Importance	Univ.Importance	Normalized Weight in	linear Corr. with Target	Weight in Model
Discriminative	P_2	6,55793	78,3	100	22,6361	0,554484
Discriminative	B_1	3,39366	59,36	43,6839	36,3785	0,238038
Discriminative	D_45	3,16424	44,98	33,5075	21,5539	0,201039
Discriminative	B_4	2,9123	57,12	32,1511	51,8819	0,196149
Discriminative	D_41	2,90852	19,74	15,7682	5,58436	0,127537
Discriminative	D_43	2,89056	39,29	18,2576	51,8819	0,167965
Discriminative	D_42	2,67335	58,88	22,6122	8,77599	0,382885
Discriminative	D_64	2,24261	23,92	10,8645	36,3785	0,055096

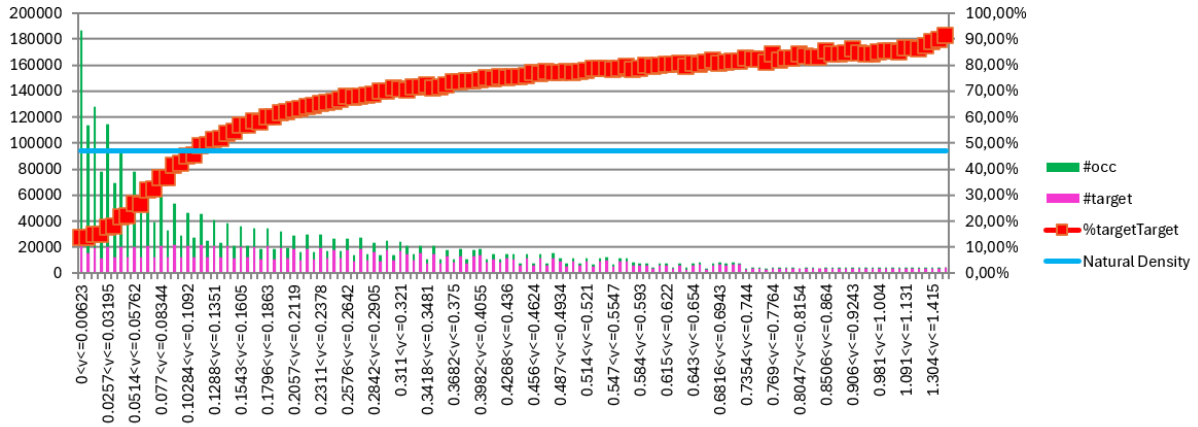
Hackathon

FIS 2025

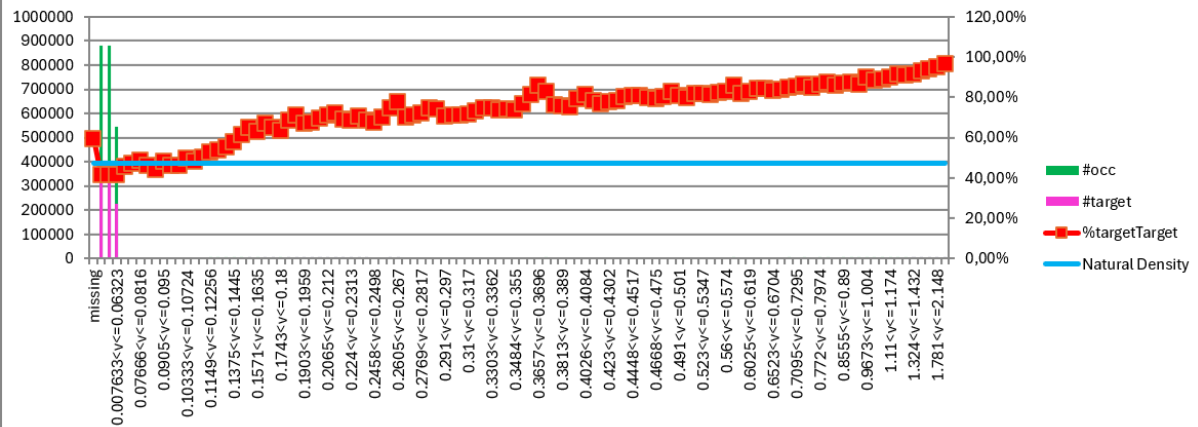




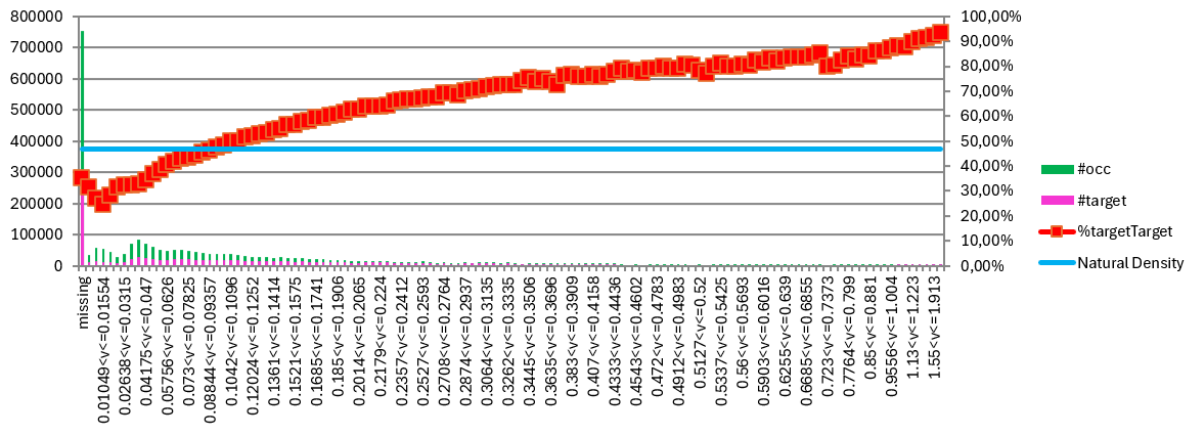
B_4

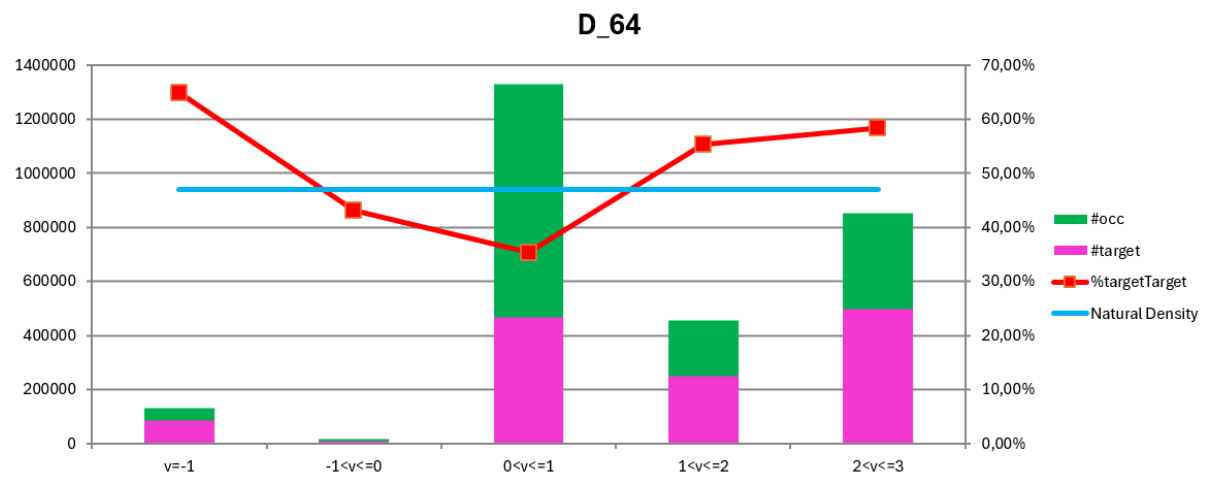
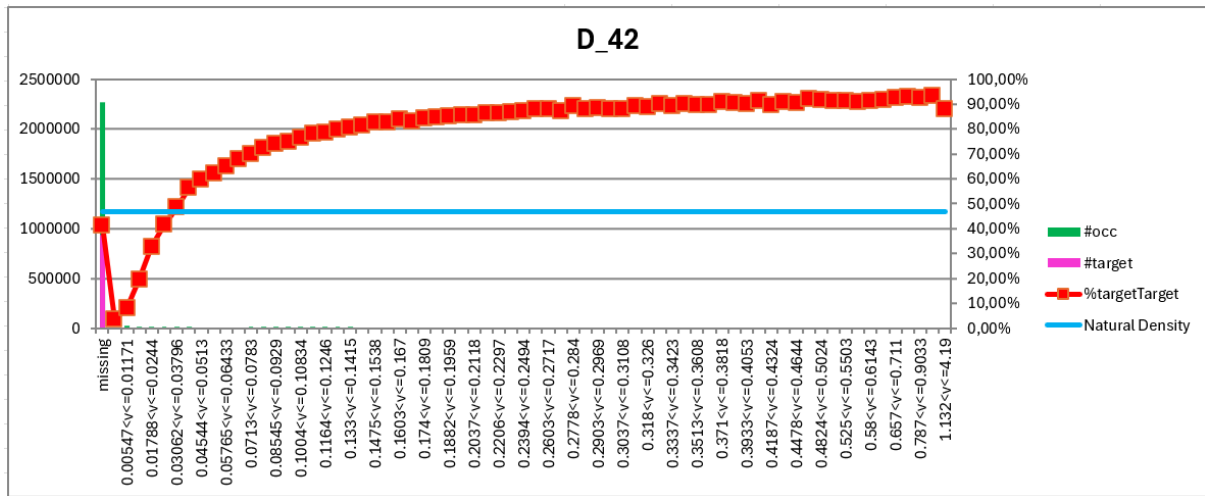


D_41



D_43





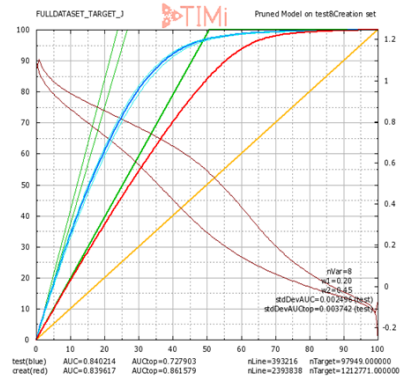


Discriminative Variables ranking

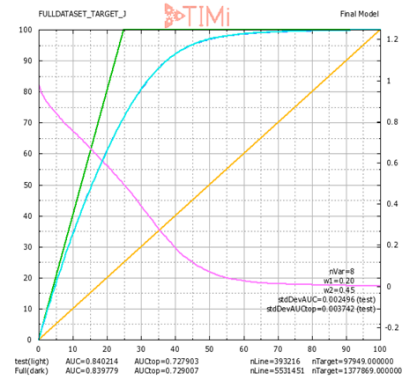
	Importance (%)	Univ. Import. (%)	correlation with Target (%)	Weight In Model (%)	INDEX (%)	Max	Highest Positive	Modalities
1 P 2	6.6	78.3	22.6	100.0	1.6	208.5	-0.459<v<=0.1727	
2 B 1	3.4	59.4	36.4	43.7	35.7	193.4	1.08<v<=1.1	
3 D 45	3.2	45.0	21.6	33.6	27.4	157.1	0.02319<v<=0.03078	
4 B 6	2.3	57.1	51.9	32.0	28.6	194.0	1.628<v<=19.8	
5 D 41	2.3	19.7	6.6	15.8	87.8	205.5	2.148<v<=8.99	
6 D 43	2.3	39.3	61.9	18.3	51.8	198.3	1.913<v<=10.11	
7 D 42	2.2	58.9	0.8	22.6	7.6	198.3	0.9003<v<=1.152	
8 D 64	2.2	23.9	36.4	10.9	75.0	137.9	v=-1	

Lifts

Pruned Model on Test Set























Final Model





Highest Univariate Quality of variables (100 first only)

	78.304 % : P_2
	69.161 % : D_48
	67.878 % : D_77
	64.135 % : D_61
	64.063 % : D_75
	63.875 % : B_42
	63.219 % : B_18
	62.877 % : B_7
	62.477 % : B_6
	62.212 % : B_23
	61.886 % : D_62
	60.670 % : D_44
	60.551 % : B_9
	60.461 % : B_10
	59.480 % : D_55
	59.357 % : B_1
	59.176 % : B_3
	59.135 % : B_37
	58.883 % : D_42
	58.795 % : B_2

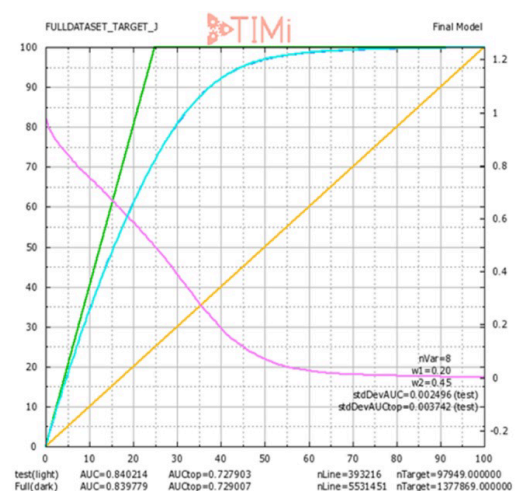
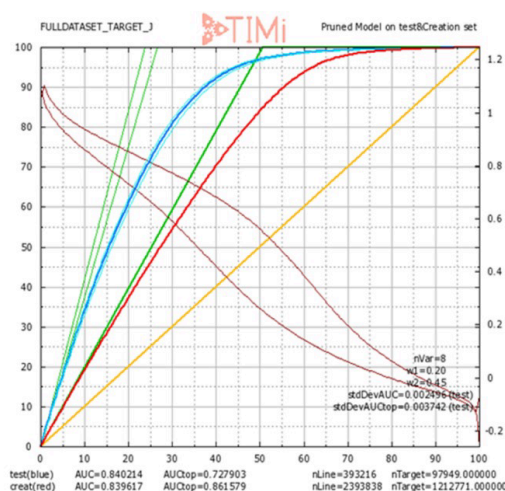


Impacto social / empresarial:

Impacto empresarial: Se anticipa una disminución considerable en las pérdidas por cartera vencida; una mejora del 5-10% en la precisión predictiva podría traducirse en ahorros de cientos de millones de pesos anuales para una entidad de tamaño medio. En consecuencia, una mejor identificación de clientes potencialmente buenos permitirá incrementar la generación de créditos sin aumentar proporcionalmente el riesgo, lo que generará mayores ingresos por intereses y comisiones, especialmente en segmentos que han sido tradicionalmente desatendidos. La eficiencia operativa se verá favorecida a través de procesos de aprobación más precisos que demanden menos revisión manual, lo que liberará recursos en el área de riesgo y permitirá que el equipo comercial se concentre en clientes con mayor probabilidad de aprobación.

Impacto social: Este proyecto facilita la democratización del acceso al crédito formal al identificar a individuos que los modelos convencionales excluyen de manera errónea, beneficiando en particular a jóvenes sin un historial crediticio extenso, a personas en regiones con escasa información disponible, y a emprendedores con proyectos viables, pero con perfiles no tradicionales. Así mismo contribuye a prevenir el sobreendeudamiento al identificar de manera más efectiva a aquellos que están en riesgo, protegiendo tanto a la entidad como al cliente. Esto fomenta la formalización de la economía al disminuir la necesidad de acudir a prestamistas informales que imponen tasas abusivas.

8. Resultados y análisis



1. Tipo de gráfica:

Curvas ROC y curvas acumuladas de ganancias.



- El eje X representa el porcentaje de clientes ordenados por el score del modelo (del más riesgoso al menos riesgoso).
- El eje Y representa el porcentaje acumulado de buenos/malos identificados.

Estas gráficas muestran *qué tan bien el modelo distingue entre clientes cumplidos y morosos.

2. Panel Izquierdo (Pruned Model)

El gráfico corresponde al modelo podado (pruned model), es decir, después de eliminar variables irrelevantes.

- Curva azul (test) y roja (creat) → son las curvas ROC para el conjunto de prueba y entrenamiento.
- Ambas alcanzan un $AUC \approx 0.84$, lo cual indica un buen nivel de discriminación.
- $AUC_{top} \approx 0.73$ significa que, en el segmento de clientes más riesgosos, el modelo prioriza correctamente alrededor del 73 % de los casos de mora.
- La podada redujo la cantidad de variables a solo *8*, pero mantuvo casi la misma precisión, lo que hace al modelo más interpretable y eficiente.

3. Panel Derecho (Final Model)

Este es el modelo final, listo para producción.

- Se confirman los resultados:
 - * AUC en test = 0.840
 - * AUC global = 0.839
 - * $AUC_{top} \approx 0.729$
- La curva azul clara (test) y la oscura (full dataset) prácticamente se superponen → esto indica que el modelo generaliza bien, sin sobreajustarse a los datos de entrenamiento.
- De nuevo, solo se utilizan 8 variables clave, lo que simplifica la interpretación y reduce el costo computacional.

el modelo final es estable, consistente y con un rendimiento muy bueno ($AUC \approx 84\%$).

4. Interpretación práctica de los resultados

- El AUC de 0.84 significa que, al comparar un cliente cumplido con uno moroso, el modelo los ordena correctamente en el 84 % de los casos.



- El AUCtop de 0.73 significa que, en el grupo más riesgoso identificado por el modelo, hay una alta concentración de clientes que efectivamente cayeron en mora.
- Esto es muy útil para el banco, porque puede segmentar clientes* en:
 - * Muy riesgoso → no otorgar crédito.
 - * Riesgo moderado → crédito con condiciones.
 - * Bajo riesgo → crédito aprobado sin restricciones.

9. Referencias

1. Scrum Guide. (2020). The Definitive Guide to Scrum: The Rules of the Game. Ken Schwaber & Jeff Sutherland. Disponible en: <https://scrumguides.org>
2. Kanban University. (2021). Kanban Methodology Overview. Disponible en: <https://kanban.university>
3. ISO/IEC 25010:2011. Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) System and software quality models. International Organization for Standardization.
4. ISO/IEC/IEEE 29148:2018. Systems and software engineering — Life cycle processes — Requirements engineering. International Organization for Standardization / IEEE.
5. Beck, K. et al. (2001). Manifesto for Agile Software Development. Disponible en: <https://agilemanifesto.org>
6. Cohn, M. (2010). Succeeding with Agile: Software Development Using Scrum. Addison-Wesley.
7. Anderson, D. J. (2010). Kanban: Successful Evolutionary Change for Your Technology Business. Blue Hole Press.
8. Sommerville, I. (2011). Ingeniería del Software (9ª edición). Pearson Educación. (Referencia general para buenas prácticas de ingeniería de software y gestión de proyectos).
9. Project Management Institute (PMI). (2021). Guía de los Fundamentos para la Dirección de Proyectos (Guía PMBOK®) – 7ª edición. Project Management Institute.

9. OBSERVACIONES

Observaciones:



El proyecto DEFAULT PREDICTION demuestra que la integración de datos masivos con técnicas avanzadas de machine learning puede transformar la forma en que una entidad financiera gestiona el riesgo crediticio. Nuestros resultados muestran un modelo sólido, interpretable y escalable, capaz de mejorar significativamente la precisión en la predicción de incumplimientos con un AUC de 0.84, lo que representa un avance frente a los modelos tradicionales.

La propuesta está diseñada para ser implementada de manera rápida y efectiva en entornos de producción, gracias a la arquitectura integrada de TIMi Suite, que garantiza seguridad, escalabilidad y gobernanza de datos.

Enlace de la solución (video YouTube):

<https://youtu.be/a9clAohAWvY>

Enlace de la API:

<https://linktr.ee/Fyronyx>