Tecnológico de Monterrey

FASE 1 | Avance de proyecto ML

MLCanvas para Insurance Company Benchmark (CoIL 2000)

MNA - AAI
Operaciones de aprendizaje automático.
Dr. Gerardo Rodríguez Hernández

Desarrollo PREDICCIÓN DE CLIENTES POTENCIALES PARA SEGUROS DE CARAVANA (Casa rodante).

Equipo 58

Data engineer: Genaro Mateu A00822623
Data scientist: Diego Ramirez Moreno a01795699
ML engineer Ali Alfonso Rico Vázquez A01350630

Devops: Jorge Santana Mendoza A01376306
Software Engineer: Angel Garcia Ortega A01796653

MACHINE LEARNING CANVAS

Designed for:

Designed by:

Date / Iteration:

Prediction Task

What is the type of task?
Which entity are predictions made on?
What are the possible outcomes to predict?
When are outcomes observed?

Tipo de tarea:
Clasificación supervisada binaria.

Entidad de predicción:
Cliente individual (registro en dataset).

Posibles resultados:
1 → Cliente con interés / probabilidad alta de contratar seguro de caravana.
0 → Cliente sin interés o baja probabilidad.

Momento de observación:
Datos históricos (train) ya incluyen si contrató o no el producto.
Predicciones se realizan antes de enviar nuevas campañas (batch mensual o trimestral).

Decisions

How are predictions turned into actionable recommendations or decisions for the end-user? (Mention parameters of the process / application for this.)

Cómo se traducen las predicciones en acciones:
Se genera un ranking ordenado por probabilidad de compra.
Marketing selecciona el top-800 clientes con mayor score para incluirlos en campañas de mailing o llamadas.
Los resultados alimentan un panel de marketing para medir impacto por campaña.

Parámetros clave:
Tamaño del mailing (800).
Umbral de probabilidad o top-k ajustado al presupuesto.
Costo promedio por contacto vs. ingreso esperado por póliza.

Value Proposition

Who is the end beneficiary, and what specific pain points are addressed?
How will the ML solution integrate with their workflow, and through which user interfaces?

Beneficiario:
Equipo de Marketing Directo y Dirección Comercial.

Pain Points:
Campañas masivas ineficientes, alto costo por adquisición y bajo ROI.

Propuesta de valor:
Priorizar clientes con mayor probabilidad de compra para optimizar campañas, reducir costos y aumentar conversiones.

Integración en flujo de trabajo:
Exportación semanal de lista de clientes priorizados (.csv o dashboard).

Interfaz: CRM interno o Power BI con tabla de "clientes recomendados".

Data Collection

How is the initial set of entities and outcomes sourced (e.g., database extracts, API pulls, manual labeling)?
What strategies are in place to update data continuously while controlling cost and maintaining freshness?

Origen del dataset:
Fuente pública UCI: Insurance Company Benchmark (CoIL 2000).

Características:
86 variables (uso de productos financieros, datos socioeconómicos por código postal).
Target: interés en seguro de caravana (CARAVAN).

Estrategias de actualización:
Incorporar nuevos registros históricos cada trimestre (simulación de actualización).
Automatizar pipelines de ingestión y limpieza mediante DVC + Python.

Data sources

Where can we get data on entities and observed outcomes? (Mention internal and external database tables or API methods.)

Internos: CRM y bases de clientes históricos (simulados en dataset).
Externos (hipotéticos): fuentes de enriquecimiento socioeconómico (por código postal o INEGI).
Versionado: data/raw → data/clean → data/preprocessed (controlado por DVC).

Impact simulation

What are the cost/gain values for (in)correct decisions?
Which data is used to simulate pre-deployment impact?
What are the criteria for deployment?
Are there fairness constraints?

Valores costo/beneficio:
Costo medio de mailing por cliente: $5 MXN.
Ingreso promedio por póliza contratada: $1,200 MXN.
Break-even estimado: 1 venta por cada 240 envíos.

Simulación pre-despliegue:
Evaluar con datos test cuántos clientes del top-800 efectivamente comprarían.
Calcular Precision@800 y Lift.

Criterios de despliegue:
Precision@800 ≥ 0.20
Lift ≥ 3
Calibración estable y sin sesgos regionales graves.

Fairness constraints:
Evitar discriminación geográfica (por código postal o nivel socioeconómico).

Making predictions

Are predictions made in batch or in real time?
How frequently?
How much time is available for this (including featurization and decisions)?
Which computational resources are used?

Modo de predicción:
Batch semanal o mensual, dependiendo del ciclo de campañas.
Tiempo de inferencia bajo (segundos).

Recursos computacionales:
Servidor o contenedor Docker (CPU).
Posible integración con API FastAPI para scoring en tiempo real futuro.

Building Models

How many models are needed in production?
When should they be updated?
How much time is available for this (including featurization and analysis)?
Which computation resources are used?

Modelos en producción:
1 modelo principal (XGBoost / Random Forest).
1 modelo baseline (Regresión Logística).

Actualización:
Reentrenamiento mensual o trimestral.

Recursos:
Entrenamiento en entorno local o nube (Colab / AWS).
Versionado de experimentos con MLflow (métricas + parámetros).

Features

What representations are used for entities at prediction time?
What aggregations or transformations are applied to raw data sources?

Representación:
Variables numéricas escaladas, categóricas codificadas (One-Hot / Ordinal).

Transformaciones:
Imputación de nulos.
Normalización de distribuciones.
Creación de ratios o interacciones relevantes.
Selección de features mediante importancia de modelos (SHAP / Gain).

Preprocesamiento controlado:
Pipeline reproducible con scikit-learn Pipeline.

Monitoring

Which metrics and KPIs are used to track the ML solution's impact once deployed, both for end-users and for the business?
How often should they be reviewed?

Métricas técnicas:
Precision@800, ROC-AUC, PR-AUC, Lift@k.

Métricas de negocio:
Costo de mailing / ROI campaña.
Incremento en conversión.

Frecuencia de revisión:
Mensual (después de cada campaña).

Monitoreo MLOps:
Drift de distribución.
Consistencia en tasas de predicción.
Alertas en dashboards Prometheus + Grafana.

Prediction Tasks

01

Tarea de ML: Clasificación Binaria.

02

Pregunta a Responder: Para un cliente dado, ¿cuál es la probabilidad de que contrate una póliza de seguro para caravana?

03

Salida del Modelo: Una puntuación de probabilidad entre 0.0 y 1.0 para cada cliente. Un umbral se usará para clasificar a un cliente como "alto potencial".

Decisions

Principal

¿A qué clientes debe contactar el equipo de marketing esta semana para la campaña de seguros de caravana?

Secundarias

Segmentación de Campaña: Crear diferentes estrategias para clientes con probabilidad alta (oferta directa) vs. media.

Prevención de Abandono: Identificar si los clientes con alta probabilidad de comprar un producto nuevo tienen riesgo de abandonar otros.

Value Proposition

Desarrollar un modelo de Machine Learning que asigne un valor con base en la probabilidad de cada cliente para adquirir un seguro de Caravanas y que después que permita seleccionar esos clientes para su aprovechamiento con campañas personalizadas.

Beneficiarios principales

El equipo de Marketing directo, ya que es el responsable de ejecutar campañas de mailling para comercializar el seguro para Caravanas.

Otros beneficiarios son el area comercial (Mejores tasas de conversión) y el area de análisis de datos (Optimización de tiempos de análisis)

¿Qué problema resolvemos?

El marketing masivo es ineficiente y costoso. La aseguradora gasta recursos en clientes que no están interesados en una póliza de caravana.

Baja efectividad (ROI) de campañas directas	Falta de segméntación predictiva	Baja trazabilidad para toma de deciones

Nuestro modelo identifica con precisión los clientes más propensos a contratar el seguro caravana, resolviendo estos pain points.

¿Cómo se integra la solución en el flujo de trabajo?

La solución se integrará en el proceso de marketing y CRM, exportando semanalmente una lista de clientes con su probabilidad de compra (Score). El equipo de marketing solo necesita definir el tamaño de la campaña para generar un listado exportable en CSV.

La lista puede cargarse en una visualizador de datos (Tableau) para analizar y monitorear desempeño. Ademas de un archivo CSV automatizado para consumirse en plataformas de mailing o CRM interno.

El modelo corre en un pipeline reproducible (DVC + MLflow + Docker), garantizando actualizaciones automáticas y consistentes.

Impacto esperado

- Increménto de efectivdad de campaña (+15% tasa de conversión).
- Reducción de costos de mailing (35%).
- Increménto de ROI al focalizar esfuerzos en clientes con mayor propensión.
- Identificar insights de audiencias asociados al interes de clientes.
- Reproducibilidad mediante practicas de MLOps.

Data Collection

Fuentes de Datos:

- La fuente principal es el dataset insurance_company_modified.csv, que representa un extracto del sistema de clientes de la aseguradora.

Data Sources

Proceso de Recolección (en un entorno real):

- Los datos se extraerían periódicamente de la base de datos de clientes de la compañía (Data Warehouse), que consolida información demográfica y de las pólizas activas de cada cliente.

Impact Simulation

Simulación pre-despliegue:

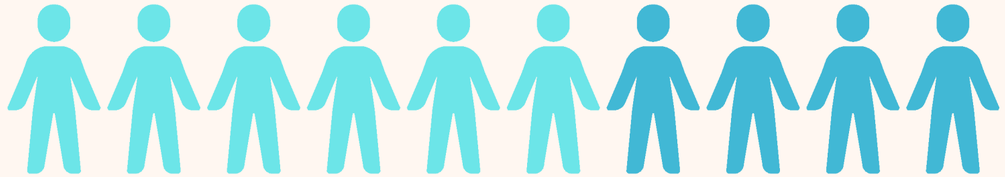
Evaluar con datos test cuántos clientes del top-800 efectivamente comprarían.
Calcular Precision@800 y Lift.

Criterios de despliegue

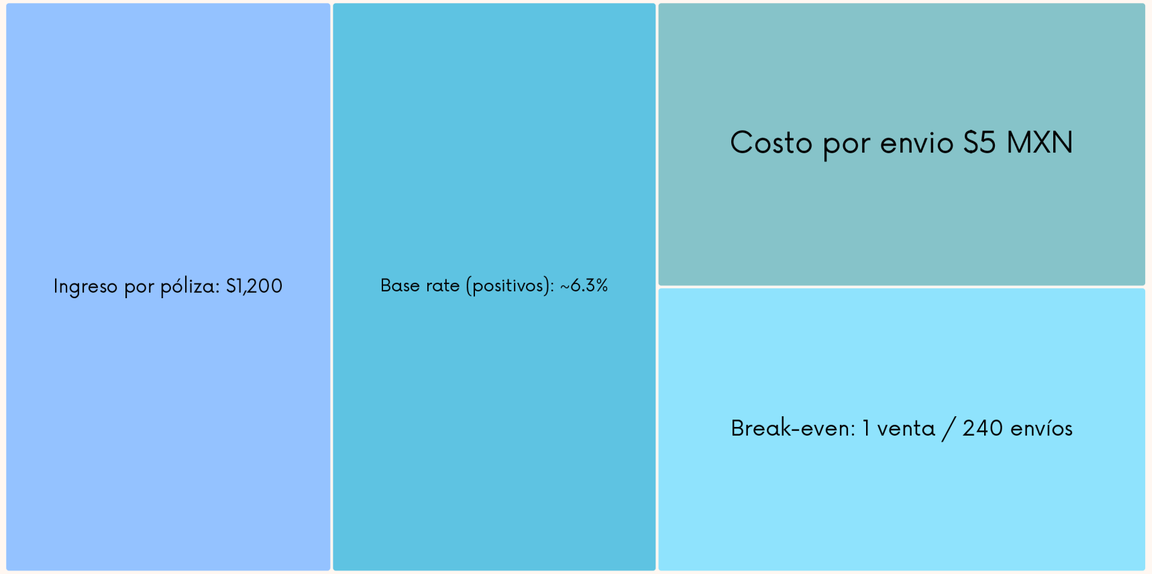
Precision@800 ≥ 0.20
Lift ≥ 3
Calibración estable y sin sesgos regionales graves

Fairness constraints:

Evitar discriminación geográfica (por código postal o nivel socioeconómico).



KPIs



Making Predictions

Vamos a generar predicciones mediante:

Modo de Operación: Inferencia Batch (por lotes)

No se necesita una respuesta en tiempo real.

Frecuencia

El modelo se ejecutará de forma programada (ej. semanalmente) sobre la base de datos actualizada de clientes.

Entrega

El resultado será una lista priorizada de los clientes con mayor probabilidad de compra, que se entregará al equipo de marketing (vía CSV, dashboard o integrado a su CRM).

Building Models

Diseñamos una arquitectura simple y gobernable: un modelo principal para capturar señal predictiva y un baseline para control y fallback. El ciclo de vida prioriza reproducibilidad, versionado y actualizaciones periódicas.

01

Modelos en producción:

E1 modelo principal (XGBoost / Random Forest).
E1 modelo baseline (Regresión Logística).

02

Actualización:

Reentrenamiento mensual o trimestral.

03

Recursos:

Entrenamiento en entorno local o nube (Colab / AWS).
Versionado de experimentos con MLflow (métricas + parámetros).

Features

01

¿Qué datos usamos para predecir?

Todas las variables socio-demográficas y de productos del cliente del dataset limpio.

02

Características Clave:

- Demográficas: Avg age, Avg size household, Income level, Education level.
- De Comportamiento: Customer main type, Number of car policies, Contribution life insurances.

03

Feature Engineering (Propuesta de mejora):

Se podrían crear nuevas características, como Total number of policies o Ratio of income to policies, para capturar patrones más complejos.

Monitoring

Para evaluar el impacto del modelo en producción, medimos desempeño técnico y de negocio y lo revisamos mensualmente tras cada campaña. El objetivo es maximizar conversión rentable y mantener estabilidad del sistema (calibración, ausencia de sesgos y salud operativa), con tableros y alertas automáticas.

01

Métricas técnicas

Precision@800, ROC-AUC, PR-AUC y Lift@k.

02

Métricas de negocio

Estimamos un ingreso medio de $1,200 tomando en cuenta un costo por envío $5. Monitoreamos el uplift vs. campañas pasadas y el break-even 1 venta / 240 envíos.

03

Frecuencia de revisión

Al menos de manera mensual despues de cada campaña.

sin embargo se trendrán nformes ejecutivos con: Precision@800, Lift, ROI y conversión.

Se realizarán revisiones ad-hoc si caen KPIs (>20% vs. histórico) o si se dispara alguna alerta de drift/calibración.

04

Drift de distribución

Seguimiento de PSI/KS para variables clave y para el score.
Semáforo: PSI ≤ 0.1 (verde), 0.1–0.2 (ámbar), > 0.2 (rojo). Ante rojo, activar re-entrenamiento o recalibración y análisis de causa.

05

Alertas y observabilidad (MLOps)

Dashboards en Prometheus + Grafana con alertas a Slack/Email.
Reglas sugeridas: Precision@800 < 0.20, Lift < 3, ROI(K=800) < 0, PSI > 0.2, caída de conversión o latencia/errores del pipeline.

Conclusiones finales

- Dataset + EDA/depuración:Se realizó una identificación de variables categóricas/ordinales/numéricas y eliminación de colinealidad ($\rho > 0.95$) para reducir redundancia y mejorar estabilidad.
- Preprocesamiento reproducible: ColumnTransformer + Pipeline; one-hot para nominales y escalado robusto para numéricas.
- Desbalance (≈6.3% positivos): foco en PR-AUC; balanceo de clases y selección de umbral por G-Mean para equilibrar sensibilidad y especificidad.
- Modelo base (Regresión Logística balanceada): PR-AUC ≈ 0.15 y ROC-AUC ≈ 0.66 en prueba.
- XGBoost (RandomizedSearchCV, 3-fold, métrica PR-AUC, scale_pos_weight): mejora clara — PR-AUC ≈ 0.19 (validación) / ≈ 0.15 (prueba), ROC-AUC ≈ 0.73; ↑ F1 y ↑ G-Mean.
- Impacto: mayor detección de verdaderos positivos con control de falsos positivos; mejor trade-off sensibilidad–especificidad.
- Siguientes pasos: calibración de probabilidades, ajuste fino del umbral según objetivos de negocio y prueba de CatBoost/LightGBM para posibles ganancias adicionales.

Gracias

Bibliografía

- MLC: Version 1.2. Created by Louis Dorard, Ph.D. Licensed under a Creative Commons Attribution-ShareAlike 4.0 International License
- Data Set: Putten, P. (2000). Insurance Company Benchmark (COIL 2000) [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5630S.