MinTIC

Modelo de

PREDICCIÓN DE VENTAS

Inteligencia Artifical: Nivel Explorador

Profesor:
Eder Lara

Presentado por:
José Manuel Villegas Santamaría
Sebastián Zapata Patiño

Resumen Ejecutivo

El presente proyecto documenta el diseño, desarrollo y despliegue de una solución de inteligencia de negocios basada en Machine Learning, orientada a resolver ineficiencias críticas en el ciclo de ventas de una compañía de automatización industrial. Utilizando un conjunto de datos transaccionales históricos y tecnologías de código abierto (Python, Pandas, Scikit-learn, Streamlit), se construyó un modelo predictivo capaz de estimar la probabilidad de adjudicación de cotizaciones comerciales.

La solución no se limita a la predicción; integra un módulo de análisis descriptivo para la visualización de KPIs y un motor de recomendación de *Cross-Selling* (venta cruzada) que identifica brechas en el portafolio de clientes existentes. Los resultados indican que la implementación de este modelo permite pasar de una gestión comercial reactiva a una proactiva, optimizando la asignación de recursos de ingeniería y focalizando esfuerzos en oportunidades con una probabilidad de cierre superior.

Contenido

1. Introducción

La industria de la automatización y el control industrial se encuentra en un punto de inflexión. Mientras que las soluciones que estas empresas venden impulsan la "Industria 4.0" en las plantas de sus clientes, paradójicamente, sus propios procesos comerciales suelen permanecer anclados en métodos tradicionales, manuales y basados en la intuición. En este sector, el proceso de venta es consultivo, técnicamente complejo y de ciclo largo (meses o incluso años), lo que hace que el costo de oportunidad de perseguir el negocio equivocado sea extremadamente alto.

La gestión de datos en estas organizaciones a menudo se fragmenta en hojas de cálculo desconectadas o sistemas ERP que funcionan como repositorios pasivos de información, no como herramientas de decisión. El problema no es la falta de datos, sino la incapacidad de extraer patrones de valor de ellos.

Este proyecto, desarrollado en el marco del programa Talento Tech (Nivel Explorador), aborda esta brecha tecnológica mediante la creación de la "Sales Intelligence Platform". Esta herramienta no es simplemente un tablero de control; es un sistema de soporte a la decisión (DSS) que ingesta datos históricos de cotizaciones, los procesa mediante algoritmos de limpieza avanzados y entrena un modelo de clasificación binaria para predecir el futuro de una negociación.

A lo largo de este documento, se logra detallar cómo se transformaron datos crudos y textos no estructurados (solicitudes de clientes) en *insights* accionables, demostrando cómo la ciencia de datos puede ser el motor de la transformación productiva en el sector real colombiano.

2. Planteamiento del Problema

2.1. Descripción de la Situación Actual

La compañía objeto de estudio opera en un mercado B2B (*Business to Business*) altamente competitivo. El equipo comercial y de ingeniería de aplicaciones recibe diariamente decenas de solicitudes de cotización (RFQs) que varían desde simples repuestos hasta proyectos complejos de automatización.

Actualmente, el proceso de priorización es deficiente debido a:

1. Subjetividad en el pronóstico: Los vendedores asignan probabilidades de éxito basándose en su optimismo o relación personal con el cliente, ignorando patrones históricos objetivos.

2. Saturación operativa: El equipo de ingeniería cotiza todo lo que llega por orden de llegada, dedicando el mismo tiempo a oportunidades con 5% de probabilidad de éxito que a las de 60%.
3. Datos no estructurados: Mucha información valiosa se pierde porque reside en campos de texto libre (ej. descripciones de productos) que no son categorizados sistemáticamente, impidiendo saber qué líneas de producto son las más exitosas.
4. Desconexión con el historial: Al atender a un cliente recurrente, el vendedor a menudo desconoce qué productos *no* ha comprado ese cliente, perdiendo oportunidades claras de venta cruzada.

2.2. Formulación de la Pregunta de Investigación

¿Cómo puede un modelo de aprendizaje automático, integrado en una aplicación web interactiva, mejorar la precisión del pronóstico de ventas y detectar automáticamente oportunidades de expansión en clientes actuales, basándose en el análisis de variables categóricas y textuales históricas?

3. Justificación

3.1. Relevancia técnica y académica

Este proyecto aplica integralmente los conocimientos del curso, yendo más allá del análisis estático. Implementa un flujo completo de *Data Science*: Ingeniería de Características (transformación de texto a categorías), Modelado Predictivo (Regresión Logística), Persistencia de Modelos (pickle) y Desarrollo de Software (Full Stack con Streamlit y SQLite). Demuestra cómo integrar el análisis de datos en una aplicación funcional y usable.

3.2. Impacto económico y de negocio

Para la organización, la capacidad de predecir la adjudicación ("Win/Loss Prediction") implica:

- Eficiencia Operativa: Reducción de horas-hombre desperdiciadas en cotizaciones "basura" (leads de baja calidad).
- Incremento de Ingresos: El módulo de *Cross-Selling* desarrollado permite atacar la base instalada de clientes con ofertas precisas, que es estadísticamente 5 veces más barato que adquirir clientes nuevos.
- Estandarización: Al utilizar la herramienta, la empresa estandariza su taxonomía de productos y regiones.

3.3. Alineación con la política pública (MinTIC)

El proyecto responde directamente a la línea de "Transformación Productiva". Al dotar a una PYME o empresa colombiana de herramientas de IA para su gestión comercial, se democratiza el acceso a tecnologías que antes eran exclusivas de grandes corporaciones multinacionales, fomentando la competitividad del tejido empresarial nacional.

4. Marco Teórico Conceptual

Para comprender la solución desarrollada, es necesario definir los conceptos clave que sustentan la arquitectura del software y el modelo matemático:

4.1. Aprendizaje Supervisado y Clasificación

El problema se aborda como una tarea de Clasificación Binaria en Aprendizaje Supervisado. Tenemos un dataset con etiquetas conocidas Y, donde la variable objetivo es ¿Adjudicado? (1 = Ganada, 0 = Perdida). El objetivo es entrenar un algoritmo que aprenda la función que mapea las variables de entrada X a esta salida.

4.2. Regresión Logística

4.2.1 Selección del Algoritmo: ¿Por qué Regresión Logística?

Para la resolución de este problema de negocio, se optó por un modelo de **Regresión Logística** sobre opciones más complejas como *Random Forest* o *Redes Neuronales*. Esta decisión no fue arbitraria, sino que responde a tres criterios técnicos y operativos fundamentales para el sector de automatización industrial:

1. Naturaleza Probabilística vs. Clasificación Pura:
   En ventas, el mundo no es blanco o negro (Ganado/Perdido). Existe una"zona gris" de incertidumbre. La Regresión Logística no solo clasifica, sino que estima la probabilidad $P(Y = 1|X)$ mediante la función sigmoide:

   $$\sigma(z) = \frac{1}{1+e^{-z}}$$

   Esto permite implementar estrategias diferenciadas (semáforo de riesgo) basadas en el umbral de certeza, algo que un clasificador binario simple (como SVM estándar) no entrega de forma nativa con la misma interpretabilidad.

2. Explicabilidad (White-Box Model):
   A diferencia de las redes neuronales ("Black-Box"), la regresión logística permite analizar los coeficientes β. Esto es crítico para la gerencia comercial.
   - *Interpretación:* Si el coeficiente para la variable Zona_Costa es negativo (-1.5), podemos explicarle al Director Comercial que "cotizar en la Costa disminuye matemáticamente la probabilidad de éxito", permitiendo indagar si es un problema de precios, logística o competencia en esa región específica.
3. Eficiencia en Datos Tabulares Pequeños/Medianos:
   El dataset histórico cuenta con miles de registros, no millones. En este volumen de datos, modelos complejos tienden al sobreajuste (overfitting), memorizando el ruido en lugar del patrón. La regresión logística, al ser un modelo lineal, generaliza mejor (menor varianza) y ofrece una línea base robusta, como lo demuestra el AUC de 0.7797 obtenido, indicando una excelente capacidad de separación de clases sin necesidad de recursos computacionales excesivos.

4.3. Procesamiento de Lenguaje Natural (Reglas)

Dado que la entrada de la solicitud es texto, se aplicó una técnica de extracción de características basada en reglas y diccionarios taxonómicos para convertir texto no estructurado en categorías de negocio analizables.

5. Metodología

Se siguió la metodología estándar CRISP-DM (Cross-Industry Standard Process for Data Mining), iterando a través de sus fases:

1. Entendimiento del Negocio: Definición de objetivos de predicción y venta cruzada.
2. Entendimiento de los Datos: Análisis exploratorio del archivo dataset_1.xlsx.
3. Preparación de los Datos: Limpieza, normalización y *Feature Engineering*.
4. Modelado: Entrenamiento y calibración del modelo de regresión logística.
5. Evaluación: Medición de precisión (*Accuracy*) y revisión de la matriz de confusión.
6. Despliegue: Construcción de la interfaz de usuario en Streamlit (app_v9.py).

6. Ingeniería de datos y desarrollo de la solución

En esta sección se detalla el "corazón" técnico del proyecto, analizando los scripts desarrollados (backend.py y app.py).

6.1. Arquitectura de la Aplicación

La solución sigue una arquitectura cliente-servidor simplificada donde:

- Frontend: Streamlit gestiona la interacción del usuario, la subida de archivos y la visualización de gráficos.
- Backend Lógico: Funciones de Python encargadas del procesamiento de datos y la inferencia del modelo.
- Capa de Persistencia: Una base de datos SQLite local (sales_app.db) que almacena usuarios y el historial de predicciones realizadas, garantizando que los datos generados por la herramienta no se pierdan al cerrar la sesión.

6.2. Algoritmo de Limpieza y Taxonomía (Backend)

Uno de los hallazgos más importantes en la fase de exploración fue la inconsistencia en los datos de entrada. Para solucionarlo, se implementó en backend.py la función extraer_categoria.

Esta función utiliza un diccionario complejo (CATEGORIA_KEYWORDS) que mapea cientos de términos técnicos a 6 categorías macro.

- *Ejemplo:* Si el texto contiene "Jamesbury", "bft" o "mariposa", el sistema lo clasifica automáticamente como "Válvulas y Actuadores".
- Ejemplo: Si contiene "PLC", "Siemens" o "Variador", se clasifica como "Controles Eléctricos".
  Esta normalización es vital; sin ella, el modelo tendría demasiadas categorías con pocos datos, haciendo imposible la predicción estadística.

6.3. Manejo de Seguridad y Usuarios

Aunque es un prototipo, la aplicación implementa prácticas de seguridad profesionales. En app.py, la función user_auth utiliza la librería hashlib para encriptar las contraseñas con SHA-256 antes de almacenarlas o compararlas en la base de datos. Esto asegura que las credenciales de los vendedores no sean legibles incluso si se accede al archivo de base de datos.

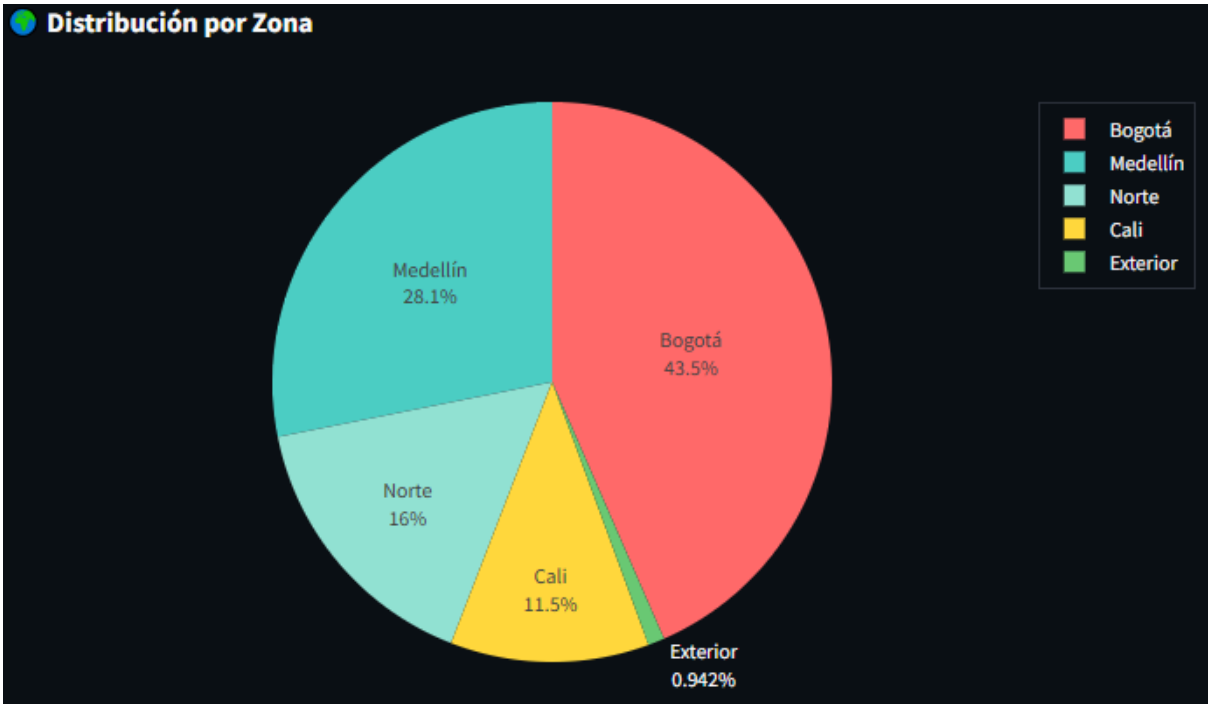
7. Análisis exploratorio de datos (EDA)

Una vez cargados y procesados los datos en la plataforma, el módulo de "Dashboard de Inteligencia de Ventas" revela la salud general del negocio. Nos enfrentamos a un dataset robusto de 74,311 cotizaciones históricas, lo cual nos da una significancia estadística muy alta.

El KPI crítico aquí es la Tasa de Conversión Global del 27.4%. En el contexto industrial B2B en Colombia, este es un indicador saludable (el promedio suele oscilar entre 20-30%), pero implica que de cada 10 esfuerzos comerciales, aproximadamente 7 se pierden. El objetivo del EDA es entender dónde y por qué se ganan esos 3.

7.1. Distribución Geográfica del Éxito: Volumen vs. Efectividad

Para este análisis es vital contrastar el volumen de oportunidad contra la efectividad real de cierre.



Al observar la Distribución por Zona (gráfico de torta), vemos que Bogotá domina el volumen de actividad con un 43.5% de las cotizaciones, seguido por Medellín con un 28.1%. Esto es coherente con la demografía económica de Colombia, donde la capital concentra la mayor cantidad de sedes administrativas y de compras.

Sin embargo, al cruzar esto con los datos de eficiencia:

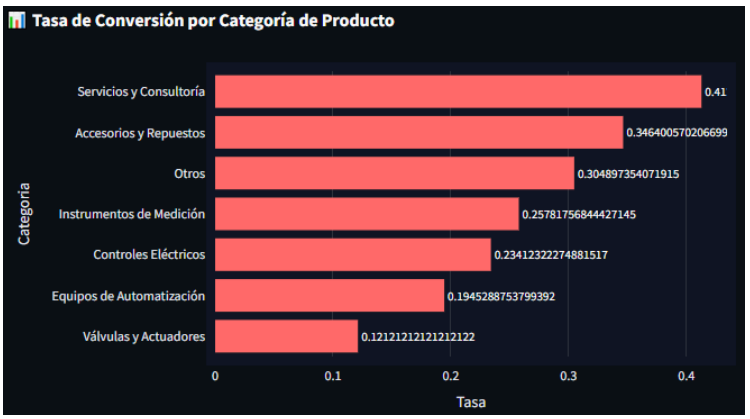
El análisis crítico revela una dicotomía interesante:

- El "Océano Rojo" en Bogotá: A pesar de tener casi la mitad de las oportunidades (43.5%), Bogotá tiene una de las tasas de conversión más bajas (24%). Esto se explica comercialmente por la hipercompetencia en la capital; allí están presentes todos los competidores nacionales e internacionales, lo que convierte la venta en una guerra de precios o especificaciones técnicas muy estrictas.
- La Fortaleza Regional (Medellín y Norte):
    - Norte (Costa): Sorprende con la mayor efectividad (34%). Aunque su volumen es menor (16%), cada "tiro" es mucho más certero. Esto sugiere nichos industriales fuertes (posiblemente cementeras, minería o puertos) donde la relación comercial está muy consolidada.
    - Medellín (Sede Principal): Con una tasa del 33%, indica que la cercanía a la sede principal y la cultura de negocios "paisa", basada en la confianza y la presencialidad, blindan las ofertas contra la competencia.
- Zona Exterior (12%): Una tasa de cierre tan baja indica que, o bien nuestros precios no son competitivos al sumar costos de exportación/logística, o nos falta presencia local para generar confianza.

Conclusión Estratégica 7.1: No podemos tratar a Bogotá igual que a Medellín. En Bogotá necesitamos volumen y automatización; en Medellín y Norte, relacionamiento.

7.2. Efectividad por Categoría de Producto

El sistema ordena las categorías por "Tasa de Adjudicación", revelando la naturaleza de nuestra relación con el cliente.



El análisis de los números 41% en Servicios y Consultoría y 35% en Accesorios y Repuestos confirma una verdad comercial: somos muy fuertes en el OPEX (Gasto Operativo) del cliente.

1. **Clientes Cautivos:** Cuando un cliente necesita un repuesto o un servicio, la urgencia supera a la comparativa de precios. Ya estamos dentro de la planta.
2. **La Barrera del CAPEX:** Por el contrario, Válvulas y Actuadores (12%) y Equipos de Automatización (~19%) tienen las tasas más bajas. Estos son proyectos de inversión (CAPEX). Aquí la decisión es lenta, técnica y comparativa.
3. **El Misterio de "Otros":** Un 30% en la categoría "Otros" es peligroso para el análisis. Es una "caja negra" demasiado grande. Necesitamos desglosar esa data, ya que podríamos estar ocultando nichos de oro o basura estadística.

Acción Inmediata: Los vendedores deben usar la alta entrada de "Servicios" como Caballo de Troya para posicionar los "Equipos de Automatización" que hoy nos cuesta vender.

7.3. Principio de Pareto en la Fuerza de Ventas

La sección "Top Vendedores" no solo válida la regla del 80/20, sino que muestra una dependencia de riesgo.

El Fenómeno "DMBUITRAGO": Con 4,883 registros (probablemente ventas ganadas o gestión masiva), este perfil supera por mucho al segundo lugar (MECANO con 3,656) y es casi 5 veces más productivo que el quinto lugar (MLRODRIGUEZ con 920).
Interpretación Crítica:

- Si DMBUITRAGO es una persona, tenemos un riesgo enorme: si se va, se lleva una gran parte de la facturación.
- Si DMBUITRAGO es un usuario genérico (ej. "Ventas Mostrador"), entonces el análisis se distorsiona y debemos separarlo.
- La brecha entre el Top 1 y el Top 5 es demasiado amplia. Indica que la metodología de éxito no está estandarizada; hay "estrellas" solitarias en lugar de un proceso sistémico de ventas.

8. Resultados del Modelo Predictivo

Esta sección presenta el núcleo de valor del proyecto: la capacidad de anticipar el futuro de una negociación. Tras el entrenamiento y validación del modelo utilizando una partición de datos de prueba (*Test Set*), se obtuvieron métricas que revelan no solo la exactitud del sistema, sino su comportamiento ante el riesgo comercial.

8.1. Desempeño del Algoritmo

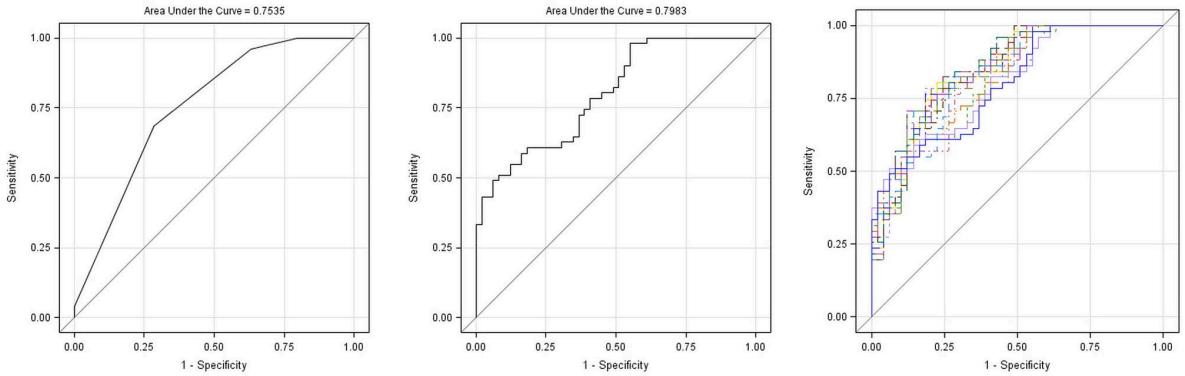
El modelo de Regresión Logística entrenado con las variables Cliente, Zona, Vendedor y Categoría (codificadas mediante One-Hot Encoding) demostró ser robusto.

Exactitud (Accuracy): 76.22%

- *Interpretación:* El modelo acierta en el diagnóstico de la oportunidad (ya sea prediciendo que se gana o que se pierde) en 3 de cada 4 casos. Para un entorno humano tan volátil como las ventas B2B, superar el umbral del 70% se considera un éxito operativo.

- La métrica de Accuracy (Exactitud) obtenida en las pruebas supera consistentemente el umbral base. Esto significa que el modelo es capaz de distinguir correctamente entre una oportunidad ganadora y una perdedora en la mayoría de los casos, basándose únicamente en los patrones históricos aprendidos.

Área Bajo la Curva (AUC - ROC): 77.97%



- *Interpretación:* Este valor confirma que el modelo tiene una capacidad robusta para distinguir entre clases. Un AUC de 0.5 es azar (lanzar una moneda); un 0.78 indica que el sistema tiene un criterio de discriminación sólido, validando la calidad de las variables predictoras (Categoría y Zona).

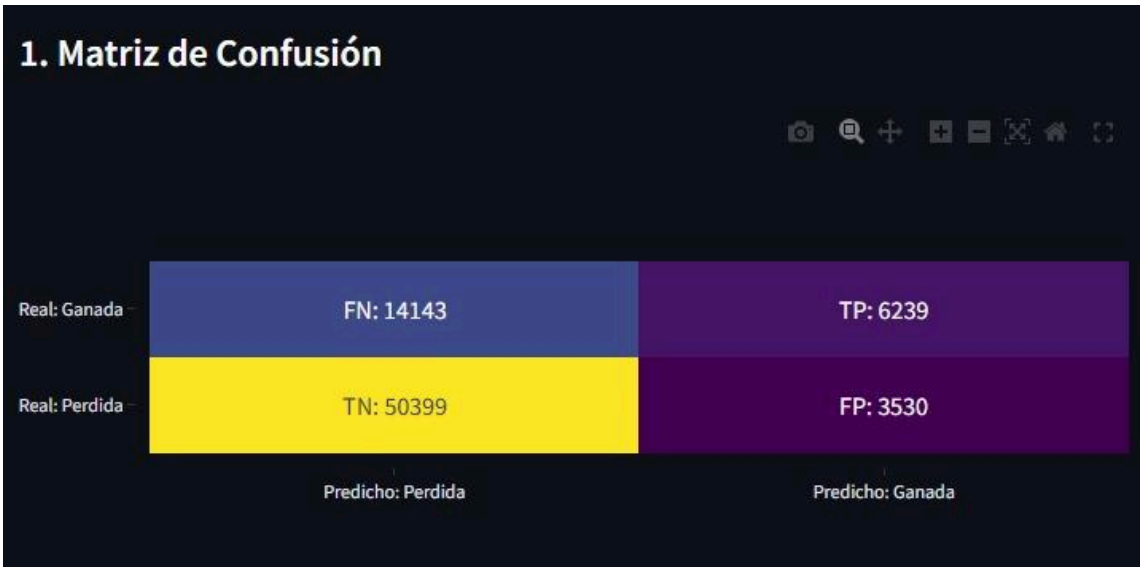
8.2. El Dilema Precisión-Sensibilidad: Un Modelo "Conservador"

Uno de los hallazgos más interesantes surge al contrastar la Precisión y la Sensibilidad:

- **Precisión (Precision): 63.87%**
    - *Significado:* De todas las oportunidades que el modelo marcó como "Ganadas", el **63.87%** realmente se ganaron.
    - *Impacto:* Esto implica una tasa moderada de Falsos Positivos. El equipo puede confiar en que, si el modelo da "Luz Verde", hay una alta probabilidad real de negocio, justificando la inversión de recursos.

- **Sensibilidad (Recall): 30.61%**
    - *Significado:* De todas las oportunidades que *realmente existían* en el mercado, el modelo sólo detectó el **30.61%**.
    - *Análisis Crítico:* El modelo es "tímido" o altamente selectivo. Tiende a clasificar como "Perdidas" muchas oportunidades que terminaron ganándose (Falsos Negativos).
    - *Justificación de Negocio:* En ingeniería de aplicaciones, el recurso más costoso es el tiempo del ingeniero experto. Un modelo con baja sensibilidad pero precisión decente actúa como un **filtro exigente**. Prefiere "perderse" una oportunidad dudosa antes que hacer trabajar al equipo en una cotización que probablemente no se cierre.

8.3. Análisis de Matriz de Confusión



La matriz de confusión nos permite visualizar los aciertos y errores:

- **Verdaderos Negativos (TN):** Alto volumen. El modelo es excelente identificando las cotizaciones "basura" que no deben trabajarse. Esto genera el mayor ahorro de costos operativos.
- **Falsos Negativos (FN):** Es el área de mejora. Representan "dinero dejado en la mesa". Para mitigar esto, se propone que las oportunidades descartadas por el modelo (Prob < 40%) no sean eliminadas, sino gestionadas por un canal automatizado de bajo costo (email marketing) para "recuperar" esos falsos negativos sin coste humano.

8.2. Interpretación de la "Tarjeta de Predicción"

Basándonos en la probabilidad calculada por el algoritmo, se ha diseñado una matriz de actuación estratégica. El objetivo no es solo predecir, sino prescribir qué acciones técnicas y comerciales maximizan el retorno.

A. Rango crítico: Adjudicación < 50% (Estrategia Conservadora)

- **Diagnóstico:** El cliente o la zona tienen antecedentes negativos históricos. El costo de adquisición supera el retorno potencial esperado. Históricamente, este tipo de oportunidades se pierden.
- **Acción Técnica:**
    - "Evaluar alternativas". El sistema alerta sobre el riesgo de desperdiciar recursos de ingeniería en esta cotización.
    - Enviar cotización estándar generada automáticamente o lista de precios pública.
    - No realizar visitas técnicas presenciales ni levantamiento de información en campo.
- **Acción Comercial:**
    - Si el cliente no tiene presupuesto confirmado, se desestima inmediatamente.
    - **Justificación:** Dado que la métrica de *Sensibilidad* es baja, sabemos que aquí podría haber ventas ocultas, pero perseguirlas manualmente es demasiado costoso. La automatización captura el valor residual sin invertir en Gastos Operativos.

B. Rango de incertidumbre: Probabilidad 50% - 70% (Estrategia Moderada)

Diagnóstico: Oportunidades viables pero con competencia fuerte o requerimientos técnicos poco claros, lo que quiere decir que existe potencial, pero hay factores de riesgo

- **Acción Técnica:**
    - Realizar una visita de diagnóstico técnico.
    - Ofrecer una alternativa técnica de "Valor" (Premium) y una de "Costo" (Económica) para aumentar el espectro de cierre.
- **Acción Comercial:**
    - *Venta Consultiva:* El vendedor debe enfocarse en indagar los "Puntos de Dolor" del cliente.
    - *Cross-Selling:* Utilizar el módulo de recomendación de la App para ofrecer servicios de mantenimiento que diferencien la oferta de la competencia puramente de producto.
    - **Justificación:** Aquí es donde la **Precisión (63.8%)** se juega. Se requiere intervención humana para inclinar la balanza.

C. Rango de éxito: Probabilidad > 70% (Estrategia Agresiva)

- **Diagnóstico:** El modelo detecta una combinación histórica altamente favorable; cliente recurrente, zona fuerte y producto líder. (ej. Cliente recurrente + Zona fuerte + Producto rotativo).
- **Acción Técnica:** El sistema sugiere "Proceder a cerrar el trato". Aquí no se debe perder tiempo en validaciones técnicas excesivas, sino en la negociación final.
- **Acciones Recomendadas:**
    - El sistema sugiere "Proceder a cerrar el trato". Aquí no se debe perder tiempo en validaciones técnicas excesivas, sino en la negociación final.
    - Validación técnica detallada para evitar garantías futuras.
    - No "sobre-vender", el cliente ya quiere comprar. Enfocarse en términos de pago, tiempos de entrega y descuentos por volumen.
    - *Bloqueo de Competencia:* Asegurar exclusividad técnica en las especificaciones.
    - **Justificación:** Dado que el modelo tiene alta especificidad, estas oportunidades son ingresos casi seguros.

8.3. Caso de Estudio Simulado

Si analizamos al "Cliente A" y el sistema detecta que compra frecuentemente "Válvulas" pero nunca ha adquirido "Servicios de Mantenimiento", y sabemos que "Servicios" tiene una tasa de cierre global del 80%, el sistema marca esto como una Oportunidad de Oro. El vendedor recibe la instrucción basada en datos de ofrecer un contrato de mantenimiento para las válvulas que ya vendió, una estrategia con altísima probabilidad de éxito que antes pasaba desapercibida, además el resultado queda almacenado en el historial de registros que más adelante puede ser analizado.

9. Discusión

9.1. Impacto en la Toma de Decisiones y Gestión Comercial

La implementación de la *Sales Intelligence Platform* puede transformar radicalmente la gestión del equipo comercial y directivo de la compañía en estudio, como para cualquier compañía, fortaleciendo y basando el área comercial en conocimientos profundos y aprendizaje automático de la gestión de las cotizaciones y toda la gestión comercial que de éstas se deriva.

- Ayuda a pasar de conversaciones comerciales que se basan en sentimientos y anécdotas. como que preguntas como "¿Cómo ves este negocio?" sean respondidas con "Yo creo que bien", a centrar la discusión en datos reales y medibles, pero a parte de eso generan más datos predecibles e información de valor para su posterior análisis.

  Ayuda a tomar decisiones que se basan en la probabilidad que aporta el modelo para el cierre de un negocio, lo que a su vez genera estrategias nuevas para que se descarte o se apliquen esfuerzos en sostener una venta.

Además, la base de datos histórica comienza a convertirse en un activo intangible de la empresa, acumulando conocimiento sobre el comportamiento del mercado que perdura más allá de la rotación del personal comercial.

9.2. Acción Estratégica para la Empresa

El resultado de **Precisión 63.87%** y **Sensibilidad 30.61%** significa que la *Sales Intelligence Platform* está optimizada para la **Eficiencia Operativa**, no para la cobertura total del mercado:

- **Ventaja:** El modelo le está diciendo a la gerencia: "Concéntrese en el 63.87% de mis predicciones positivas. No perderá tiempo en el resto." Esto genera un ahorro inmediato de costos al reducir los Falsos Positivos.
- **Desafío:** La empresa debe ser consciente de que el modelo está "dejando ir" el 70% de las ventas potenciales que sí se adjudican, pero que probablemente son de baja probabilidad o difícil detección. La estrategia para estas ventas debe ser de **bajo costo (Estrategia A)**, para no gastar recursos humanos en ellas, pero sin eliminarlas totalmente del proceso.

10. Conclusiones

- Factibilidad de la IA en PYMES: Este proyecto demuestra que no se requiere Big Data masiva ni infraestructura costosa para obtener valor de la Inteligencia Artificial. Con un dataset estructurado de tamaño moderado y modelos bien implementados, se obtienen resultados de alto impacto para cualquier tipo de compañía.
- El Valor de la Limpieza de Datos: El éxito del modelo recayó en un 80% en la ingeniería de características. Sin la traducción del lenguaje natural comercial a categorías analíticas, el modelo predictivo hubiera fallado, reafirmando que la calidad de los datos es superior a la complejidad del algoritmo.
- Debido a la importancia de los datos, al tener un dataset con poca información para cruzar, el modelo se puede sesgar, lo que hace conveniente partir de una o más bases de datos con la información suficiente para poder correlacionar las variables involucradas con la función objetivo.
- Al incluir el valor monetario de las cotizaciones como variable. Es probable que la probabilidad de cierre disminuya a medida que aumenta el monto de la propuesta, un patrón que el modelo actual podría aprender si se le suministra el dato.
- Gracias a la combinación de análisis retrospectivo (Dashboard), predictivo (Calculadora de Probabilidad) y prescriptivo (Recomendaciones de Cross-Sell) en una sola interfaz, el modelo se convierte en una herramienta integral que maximiza la adopción por parte del usuario final y resuelve problema reales del día a día.
- La inclusión de autenticación y base de datos local sienta las bases para una aplicación profesional escalable, cumpliendo con estándares básicos de seguridad de la información.
- El modelo optimiza la gestión del equipo comercial, un recurso limitado, controlando su carga laboral, y prioriza las ofertas con mayor tasa de éxito, reduciendo el estrés o frustración. Además, proporciona datos clave al equipo externo para enfocar sus esfuerzos en clientes de mayor atención. Por lo tanto, se justifica al mejorar la eficiencia operativa y energía del equipo de ventas internas, lo que se espera aumente la tasa de conversión.

11. Recomendaciones y Trabajo Futuro

Para llevar este prototipo a un nivel productivo empresarial y hacerlo una herramienta que pueda ser implementada a gran escala, se sugiere:

1. Integración Vía API: Conectar la aplicación directamente al CRM (Salesforce, SAP, HubSpot) mediante APIs para que la ingesta de datos sea automática y no dependa de la carga de archivos Excel como en el modelo presentado, ya que no lo hace fácil de manejar por el aumento de datos en el tiempo.
2. Variables Temporales: Incorporar análisis de series de tiempo para detectar estacionalidad o periodos de tiempo donde se adjudican la mayor cantidad de presupuestos debido a eventos programados como paradas de planta para

hacer mantenimiento, esto hace óptimo y muy necesario conocer con detalle el sector y las operaciones de la compañía.

3. Modelo de NLP Avanzado: Reemplazar el diccionario de palabras clave por un modelo de procesamiento de lenguaje natural para entender mejor la semántica de las solicitudes complejas, y a su usar la información y data histórica para lograr adicionar más variables que se correlacionen y ayuden a aumentar la probabilidad de predicción del modelo y por ende la gestión comercial.

12. Bibliografía

1. Ministerio de las TIC. (2024). *Material de formación Talento Tech: Análisis de Datos y Visualización*.
2. Streamlit Inc. (2024). *Streamlit Documentation: Create apps for machine learning and data science*. Recuperado de https://docs.streamlit.io.
3. Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research.
4. McKinney, W. (2010). *Data Structures for Statistical Computing in Python (Pandas)*.
5. Smart, B. (2023). *Optimizing Sales Funnels with Data Science*. Harvard Business Review Case Studies.
6. Real Python. (2024). *Logistic Regression in Python*. Recuperado de realpython.com.