PROYECTO DE PASANTÍA

Implementación y desarrollo de un sistema de priorización y predicción de leads para vehículos livianos en Andes Motor

Maruzzella Paz Sothers Falco

Proyecto para optar al título de Ingeniería Civil Industrial de la Facultad de Ingeniería y Ciencias de la Universidad Adolfo Ibáñez.

Profesor guía: *Raimundo Sánchez*

Santiago, Chile

2023

Resumen Ejecutivo

Los avances en inteligencia artificial están transformando la forma en que se toman las decisiones, especialmente en el comercio electrónico. Este proyecto presenta la implementación de un modelo avanzado de aprendizaje automático (Machine Learning) en Andes Motor, dirigido a un sistema que mejore la priorización y predicción de leads en el mercado automotriz.

A través de la metodología CRISP-DM y un extenso análisis de datos, se diseñó un sistema predictivo que mejora la precisión en la identificación de leads con alta probabilidad de conversión en ventas. El sistema se apoya en algoritmos de aprendizaje automático, con CatBoost destacándose por su adaptabilidad y rendimiento superior. De esta manera, se logró una precisión del pronóstico mayor al 80% en el problema de predicción de leads que se convierten en ventas.

La implementación del modelo demostró un aumento significativo en la eficiencia de las operaciones de ventas, reflejado en la mejora de la tasa de conversión y en la efectividad del proceso de selección de leads. Este enfoque estratégico no solo agiliza la toma de decisiones comerciales, sino que también sienta las bases para futuras mejoras y adaptaciones al cambiante entorno empresarial de la industria automotriz.

De esta manera, las técnicas de aprendizaje automático son claves para una eficiencia operativa, pues se tiene una rápida priorización de leads, mejorando sustancialmente las tasas de conversión de ventas. Además, este proyecto enriquece la comprensión sobre cómo la inteligencia artificial y el análisis de datos pueden ser fundamentales para afinar estrategias empresariales y mejorar los procesos decisionales dentro del mercado automotriz.

Palabras clave: Aprendizaje Automatizado, Priorización de Leads, Mejora de Ventas, Industria Automotriz, CatBoost, CRISP-DM.

Abstract

The advances in Artificial Intelligence are transforming the way decisions get made, especially in e-commerce. This project presents the execution of an advanced Machine Learning Model in Andes Motor, aiming to prioritize and predict leads in the automotive industry.

Designing a predictive system through data mining, CRISP-DM and an extensive data analysis, the program is in charge of bettering the chances of identifying leads with a higher conversion rate in sales. It leans on ML algorithms, CatBoost being the one to stand out for its adaptability and superior performance. This way, the model has over 80% accuracy towards predicting the conversion rate in sales.

Carrying out the model, showed a significant rise in sales operations efficiency, this being reflected on the conversion rate and the leads process selection efficacy. This approach not only agilizes commercial decision making, but also sets a precedent for future improvements and adaptations on the ever changing environment of the automotive industry.

This way, Machine Learning techniques are key for an operative efficiency, it being fast at prioritizing leads, substantially improving the sales conversion rates. Besides, this project favors how AI and data analysis are viewed and how they can help tune in different business strategies and improve decision making within the automotive market.

Keywords: Machine Learning, Lead Prioritization, Sales Improvement, Automotive Industry, CatBoost, CRISP-DM.

Índice

1. Introducción

En la introducción de este proyecto de pasantía en Andes Motor, se abordarán tres aspectos clave. En primer lugar, se presentará el contexto de la empresa, es decir, la visión general de Andes Motor, destacando su rol en la industria automotriz y sus modelos de venta. En segundo lugar, se analizará el contexto del problema, en donde se describe el desafío al que se enfrenta la empresa. En tercer lugar, se identifica el contexto de la oportunidad, en donde se observará la oportunidad de mejorar el problema por el que está pasando la empresa.

a. Contexto de la empresa

Comercial Motores de los Andes SPA, es una empresa de la industria automotriz chilena, la cual constituye una parte fundamental del Grupo Empresas Kaufmann. Este conglomerado empresarial ha sido parte fundamental del mercado chileno de automóviles durante más de siete décadas. Además, su presencia se extiende a varios países latinoamericanos, con operaciones en Perú, Nicaragua, Costa Rica, Panamá y Colombia.

La empresa Andes Motor fue fundada en 2008, acumulando así más de 15 años en la industria automotriz chilena. Durante este periodo, ha construido una amplia red de cobertura que se extiende de Arica hasta Punta Arenas, ofreciendo un amplio abanico de soluciones de movilidad. Dentro de estos destacan camiones, camionetas, buses, maquinarias pesadas, vans, vehículos comerciales y de pasajeros, además de vehículos eléctricos.

Hoy en día, las instalaciones se encuentran ubicadas en la Región Metropolitana de Santiago, en donde las oficinas comerciales y de venta, así como su taller, están establecidos en la comuna de Pudahuel. Además, el centro de almacenamiento de los vehículos se encuentra en el Centro Logístico de Kaufmann, situado en la comuna de Lampa.

La empresa es conocida por representar y comercializar una amplia gama de marcas automotrices internacionales. Las ocho marcas que representa son:

1- Maxus: Una marca con fuerte presencia en el segmento comercial y especializada en la venta de vehículos eléctricos.

2- Iveco: Ofrece vehículos comerciales livianos, medianos y pesados, consolidándose como un referente en el mercado.
3- Foton: Especialista en camiones y uno de los principales actores en la venta de buses eléctricos en Chile, contribuyendo así a la movilidad sostenible.
4- Sany: Reconocida a nivel mundial como el fabricante número uno de maquinaria pesada en China, brindando soluciones de alta calidad.
5- Agrale: Enfocada en el transporte público en regiones, comprometida con la movilidad urbana.
6- Karry: Una firma especializada en camiones de ciudad, que atiende las necesidades de trabajadores urbanos y rurales.
7- Kaiyi: Ofrece vehículos de pasajeros equipados con tecnología inteligente, brindando una experiencia avanzada.
8- Jetour: Se especializa en la comercialización de vehículos de pasajeros, los cuales cuentan con tecnología inteligente y se distinguen por su carácter de lujo.

Esta amplia gama de marcas ilustran la diversidad del portafolio de productos que ofrece, en donde se satisfacen las diferentes necesidades de los clientes del mercado automotriz chileno.

Dentro de la organización de Andes Motor, se distinguen diferentes áreas que desempeñan roles específicos. Este proyecto, se enfocará en el área comercial, específicamente, en el segmento de vehículos livianos que alberga las marcas Karry (vehículo comercial), Kaiyi (vehículo de pasajeros) y Jetour (vehículos de pasajeros).

Es relevante destacar que Karry comenzó a existir en el área de vehículos livianos en Septiembre de 2022, mientras que las marcas Kaiyi y Jetour fueron introducidas y comenzaron a ser distribuidas y comercializadas por Andes Motor en Febrero y Junio de 2023, respectivamente.

El liderazgo de esta área recae en el Gerente Comercial de Vehículos Livianos, el cual está a cargo de supervisar y dirigir la gestión del área. Dentro de sus responsabilidades, se

encuentra el aseguramiento de objetivos y metas de venta, además de la supervisión del desempeño y las responsabilidades de los colaboradores de esta área.

El equipo está conformado por tres Jefes de Ventas. Cada uno de ellos está asignado a una de las tres marcas mencionadas anteriormente. Además, el colaborador responsable de Jetour desempeña la labor de Product Manager de las tres marcas. Como parte del equipo, paralelamente, se encuentran tres ingenieros de Servicios que brindan soporte a Karry, Kaiyi y Jetour. Su labor es abordar y gestionar las fallas y necesidades de servicio técnico que surgen en los vehículos.

La comercialización de los vehículos livianos se lleva a cabo a través de 18 diferentes concesionarios, los cuales se encuentran ubicados a lo largo de todo Chile. Estas ventas operan bajo un modelo de ventas al por mayor (WholeSale), en donde las transacciones son gestionadas por el área comercial de Andes motor a través de un portal de concesionarios. Inicialmente, se generan órdenes de compra que se cargan en SAP, donde se crean fichas de venta. Luego, la facturación es autorizada y se coordina la entrega de los vehículos a los concesionarios. Este método permite que los concesionarios mantengan un inventario adecuado (Stock) para realizar ventas al por menor (RetailSale), es decir, ventas directas al público.

En cuanto a las ventas al por menor, estas son gestionadas por los propios concesionarios quienes se encargan de contactar a posibles clientes. Los clientes pueden provenir tanto de aquellos que se comunican directamente con el concesionario o de aquellos que solicitan cotizaciones a través de páginas web o redes sociales de la marca. (Figura 1.a.1)

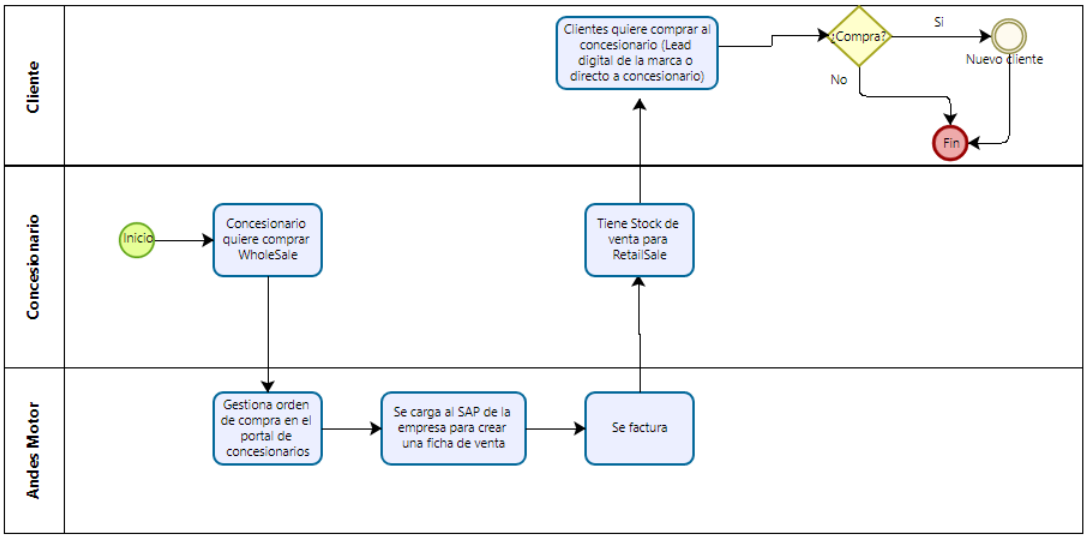


Figura 1.a.1: Diagrama del proceso de venta de vehículos livianos. (Elaboración propia)

Dado que la adquisición de vehículos no es una compra periódica normalmente, es esencial mantener una constante captación y conversión de nuevos clientes. Para lograr esto, Andes Motor ha adoptado una estrategia de marketing digital para comunicarse y atraer clientes de manera online. Esta estrategia se lleva a cabo a través de canales digitales, tanto gratuitos como de pago, con el objetivo de incrementar el número de posibles clientes. En este contexto, es fundamental el concepto de *leads*, término inglés que se refiere a los individuos que han manifestado interés en los vehículos de Andes Motor. Estos *leads* representan una oportunidad inicial de contacto o venta, ya que son potenciales compradores que han mostrado una predisposición a conocer más sobre los vehículos que tiene la empresa.

Actualmente, Andes Motor emplea dos métodos principales para la generación de leads. Por un lado, se aprovechan las herramientas ofrecidas por Google, destacando la "Optimización de Motores de Búsqueda" (SEO) y el "Marketing en Motores de Búsqueda" (SEM). Estas técnicas están diseñadas para mejorar la visibilidad en línea de las marcas. El SEO, una estrategia sin costo, se centra en aumentar la visibilidad orgánica en los resultados de búsqueda. Por otro lado, el SEM implica una inversión financiera, ya que promueve la posición del sitio web de las marcas en los primeros lugares de los resultados de búsqueda.

De forma complementaria, Andes Motor también realiza generación de leads a través de las redes sociales, colaborando estrechamente con Meta. En esta plataforma, se utilizan dos tipos de anuncios: los anuncios principales (Lead Ad) y anuncios de enlace (Link Ads). Los Lead Ads permiten a los usuarios completar un formulario de cotización directamente dentro de la aplicación, mientras que los Link Ads redirigen a los usuarios a la página web de la marca para completar el formulario.

En conclusión, el proceso de adquisición de clientes comienza con una estrategia de marketing digital que incluye tanto herramientas de Google como de Meta. Como se mencionó antes, estas estrategias generan leads, los cuales son gestionados a través de Bloomreach, una plataforma utilizada para administrar el contenido de páginas web y redes sociales.

Los datos de los leads captados se registran en una base de datos dentro de Bloomreach. Esta base está perfectamente sincronizada con Amazon Web Services (AWS), lo que facilita el alojamiento de los datos en el portal de concesionarios y permite su integración automática en los sistemas CRM de cada uno de ellos. Una vez almacenados los datos, los concesionarios proceden a establecer comunicación con todos los leads. Sin embargo, actualmente no existe un sistema de priorización para estos leads, lo que implica que los concesionarios deben interactuar con los leads en diferentes momentos, dependiendo de su disponibilidad. Finalmente, es decisión de los leads si desean proseguir con su proceso de compra o no, tal como se muestra en la Figura 1.a.2.

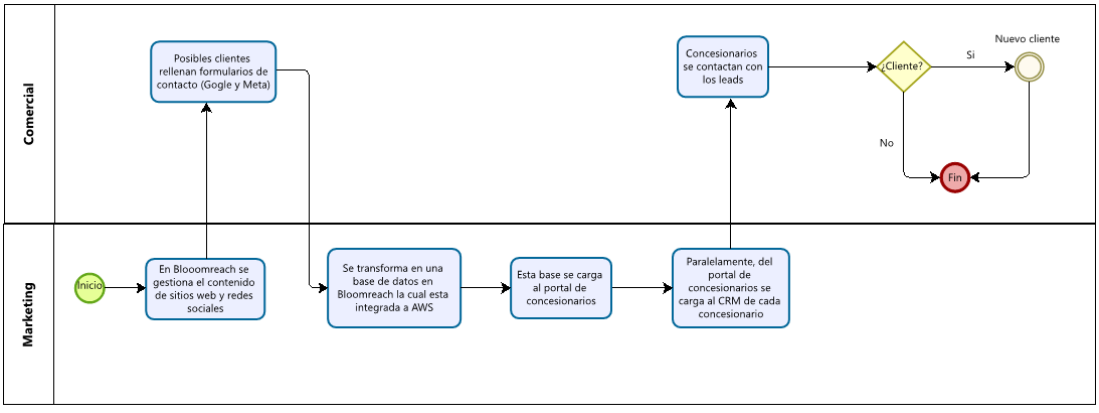


Figura 1.a.2: Diagrama del proceso de leads de vehículos livianos. (Elaboración propia)

En el marco de sus estrategias de marketing digital, Andes Motor asigna inversiones publicitarias específicas a Google y Meta en función de la marca (Tabla 1.a.1). La distribución de recursos se ajusta mensualmente según las ventas y leads obtenidos en el mes anterior. Esta táctica estratégica ha llevado a que la mayoría de las inversiones para las tres marcas se enfoquen en las Redes Sociales a través de Meta, reflejando una clara preferencia por este canal en la atracción de potenciales clientes.

Año/Mes	Liferay	RRSS
2022	**$ 19.850.000**	**$ 72.600.000**
10	**$ 11.850.000**	**$ 24.300.000**
Karry	$ 11.850.000	$ 24.300.000
11	**$ 4.000.000**	**$ 24.300.000**
Karry	$ 4.000.000	$ 24.300.000
12	**$ 4.000.000**	**$ 24.000.000**
Karry	$ 4.000.000	$ 24.000.000
2023	**$ 62.318.350**	**$ 203.739.355**
1	**$ 3.000.000**	**$ 24.000.000**
Karry	$ 3.000.000	$ 24.000.000
2	**$ 8.369.698**	**$ 23.478.227**
Kaiyi	$ 3.721.509	$ 4.886.038
Karry	$ 4.648.189	$ 18.592.189
3	**$ 8.369.698**	**$ 23.478.227**
Kaiyi	$ 3.721.509	$ 4.886.038
Karry	$ 4.648.189	$ 18.592.189
4	**$ 6.510.422**	**$ 16.096.880**
Kaiyi	$ 3.721.509	$ 4.886.038
Karry	$ 2.788.913	$ 11.210.842
5	**$ 8.500.000**	**$ 15.500.000**
Kaiyi	$ 4.000.000	$ 3.000.000
Karry	$ 4.500.000	$ 12.500.000
6	**$ 7.974.152**	**$ 22.037.826**
Jetour	$ 243.530	$ 498.799
Kaiyi	$ 4.903.439	$ 6.942.529
Karry	$ 2.827.183	$ 14.596.498
7	**$ 11.896.529**	**$ 38.473.793**
Jetour	$ 414.370	$ 4.413.861
Kaiyi	$ 5.512.911	$ 15.114.426
Karry	$ 5.969.248	$ 18.945.506
8	**$ 7.697.852**	**$ 40.674.403**
Jetour	$ -	$ 8.548.828
Kaiyi	$ 3.503.168	$ 14.337.636
Karry	$ 4.194.684	$ 17.787.939
Total general	**$ 82.168.350**	**$ 276.339.355**

Tabla 1.a.1: Inversión por mes para cada una de las marcas, respecto a la publicidad de Google y Meta. (Elaboración propia)

Asimismo, se dispone de datos sobre el origen de los leads, los cuales se encuentran desglosados por marca (Tabla 1.a.2), donde se abarcan los leads digitales de cada marca desde que se insertaron al mercado chileno. En particular, se destaca que la mayoría de los leads que se generan provienen mayormente de Redes Sociales (Meta). Es importante mencionar que para Karry, este tipo de leads digital comenzó en Enero 2023, seguido por WhatsApp en Junio 2023. Para Kaiyi y Jetour, esta fuente se habilitó en Junio 2023, y se observó un aumento notable en los meses Julio y Agosto.

	2022			2023								Total general
Lead/Marca	10	11	12	1	2	3	4	5	6	7	8	
JETOUR									760	6274	2682	9716
Liferay									436	1687	569	2692
RRSS									324	4587	2113	7024
KAIYI				132	1373	286	826	484	1224	2958	2942	10225
Liferay				132	1373	286	826	484	629	565	1024	5319
RRSS									595	2393	1918	4906
KARRY	779	798	510	2345	1519	2284	1703	1569	1428	1619	1209	15763
Liferay	779	798	510	616	398	648	782	803	281	286	359	6260
RRSS				1729	1121	1636	921	766	1038	1122	705	9038
WHATSAPP									109	211	145	465
Total general	779	798	510	2477	2892	2570	2529	2053	3412	10851	6833	35704

Tabla 1.a.2: Leads obtenidos para las tres marcas en Google y Meta durante 2022 y 2023. (Elaboración propia)

b. Contexto del problema

El mercado automotriz es conocido por la alta competencia que tiene, el cual ha enfrentado desafíos en cuanto a sus cifras de ventas en los últimos años. Además, es fundamental tener en consideración que este sector se ha visto afectado debido a las restricciones de financiamiento, una disminución de la liquidez y una debilidad en la actividad económica, tanto en Chile como a nivel global.

Dichos factores han contribuido a la contracción de las ventas de vehículos livianos y comerciales según la Asociación Nacional Automotriz de Chile (ANAC). Pues al comparar los datos correspondientes a los años 2022 y 2023, se observa una caída de las ventas. (Figura 2.b.1).

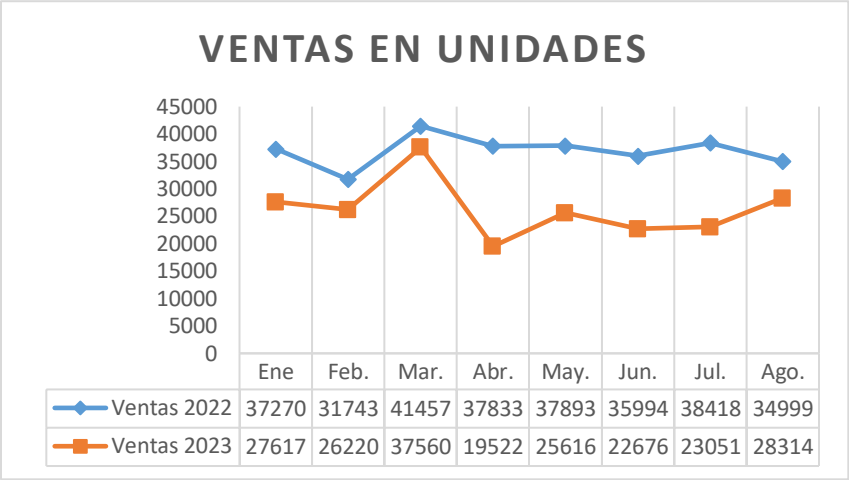


Figura 2.b.1: Unidades vendidas de vehículos en 2022 y 2023. Fuente ANAC. (Elaboración Propia)

Se puede observar la línea azul que ilustra las ventas de automóviles correspondientes al año 2022, mientras que la línea naranja indica las ventas del año 2023. Esta recesión en las ventas puede atribuirse a factores como la inflación e incertidumbre económica, en donde han impactado a la decisión de los consumidores a la hora de adquirir vehículos.

Las ventas durante el año 2023 en los vehículos livianos de Andes Motor (Figura 2.b.2) se han mantenido sin cambios significativos, a pesar de algunas fluctuaciones. En particular, las ventas para Karry se han mantenido estables. Por otro lado, Kaiyi y Jetour, que son marcas que se han incorporado al mercado este año, muestran una tendencia al alza en sus ventas.

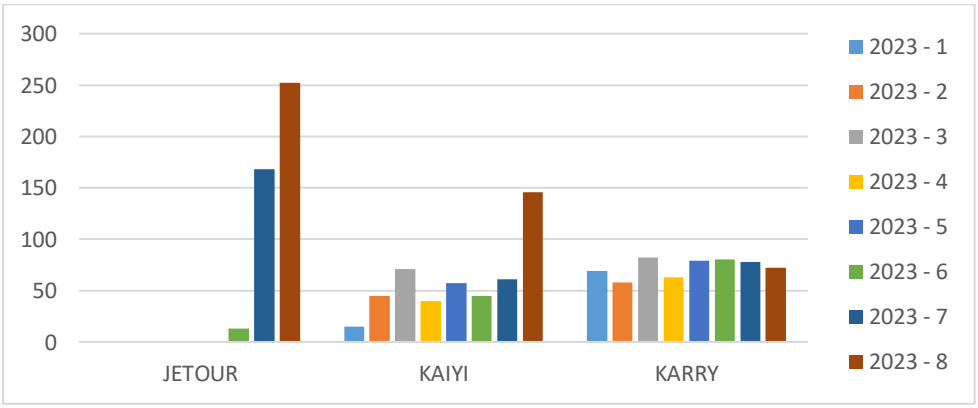


Figura 2.b.2: Unidades vendidas de vehículos en 2023 para Jetour, Kaiyi y Karry. (Elaboración Propia)

En consecuencia, las estrategias de marketing, adquisición de clientes y generación de leads, se han vuelto cruciales para Andes Motors. Pues, estas estrategias buscan asegurar un flujo constante de posibles clientes en medio de estas fluctuaciones económicas y las cambiantes de las preferencias de los consumidores.

Actualmente, el proceso de gestión de leads se lleva a cabo de manera intuitiva, utilizando información obtenida de una base de datos, en donde se encuentran los datos de posibles clientes, la concesionaria a la cual le gustaría comprar y el origen de estos leads, es decir, si provienen de la web o redes sociales. Sin embargo, a pesar de contar con esta información básica, carece de un método de análisis efectivo para predecir y priorizar las probabilidades reales de conversión en clientes.

La falta de un enfoque estructurado en la predicción de la priorización de los leads, genera que, si bien un potencial cliente pueda mostrar un interés inicial en un vehículo, muchas veces este abandona el proceso y no se realiza la compra. En consecuencia, la desconexión entre la generación de leads y su conversión efectiva en clientes resulta en un bajo rendimiento del proceso actual para las ventas en la empresa.

En la práctica, se refleja que del total de los leads digitales generados mensualmente que finalmente son clientes (tasa de conversión de leads) es extremadamente baja, con un valor promedio mensual de 0,97% para las tres marcas en lo que lleva del año. Esta cifra, claramente es insatisfactoria, ya que se sitúa por debajo del potencial de la empresa y no aprovecha plenamente las oportunidades de negocio. Es pertinente destacar, que en promedio el 6% de los leads se convierten en clientes, sugiriendo un objetivo claro para mejorar la tasa de conversión de leads (Dunkan y Elkan, 2015).

En lo que respecta a Jetour, su tasa de conversión de leads a ventas es en promedio del 1% (Tabla 2.b.1), lo cual según la cantidad de leads que genera es muy bajo. Por ejemplo, para el mes de junio de los 5561 leads generados, sólo 48 de esos se convirtieron en ventas. Además, la correlación entre leads y ventas es indeterminada, con un coeficiente de correlación fluctuando entre 0 y 1, lo que impide establecer una relación directa entre ambas variables.

Mes	leads_jetour	Inversion	ventas_jetour	ventas_leads_jet	tasa_conve
Junio	660	$ 742.329	13	7	1,06%
Julio	5561	$ 4.828.231	168	48	0,86%
Agosto	2482	$ 8.548.828	251	27	1,09%

Tabla 2.b.1: Leads y Ventas para Jetour 2023. (Elaboración propia)

Respecto a Kaiyi, la tasa de conversión de leads a ventas posee un promedio de 0,92% (Tabla 2.b.2), lo que en comparación a los leads que genera es extremadamente bajo. Por ejemplo, para el mes de Julio se generaron 2682 leads, de los cuales solo 11 de ellos fueron compras efectuadas. Asimismo, el análisis de correlación muestra una relación directa entre la inversión y los leads (Figura 2.b.3), indicando que un aumento en la inversión se corresponde con un incremento en los leads obtenidos.

Mes	leads_kaiyi	Inversion	ventas_kaiyi	ventas_leads_ka	tasa_conve
Enero	94	-	15	1	1,06%
Febrero	1260	$ 8.607.547	45	10	0,79%
Marzo	273	$ 8.607.547	71	4	1,47%
Abril	744	$ 8.607.547	40	7	0,94%
Mayo	430	$ 7.000.000	57	5	1,16%
Junio	1083	$ 11.845.968	45	8	0,74%
Julio	2682	$ 20.627.337	61	11	0,41%
Agosto	2689	$ 17.840.804	146	22	0,82%

Tabla 2.b.2: Leads y Ventas para Kaiyi 2023. (Elaboración propia)

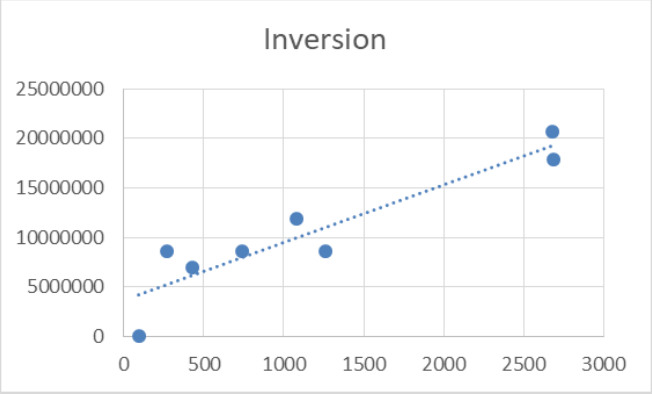


Figura 2.b.3: Correlación entre la inversión y leads en Kaiyi. (Elaboración propia)

Finalmente, para Karry, la tasa de conversión de leads a ventas es en promedio 0,99% (Tabla 2.b.3), una cifra que sigue la tendencia baja observada en las otras marcas.

Por ejemplo, se puede observar que para el mes de Mayo se generaron 1412 leads, de los cuales 13 de ellos resultaron en ventas. Adicionalmente, el análisis de correlación entre los leads digitales generados y la inversión (Figura 2.b.4), ha evidenciado una correlación negativa, indicando que un aumento en la inversión no necesariamente se traduce en un incremento proporcional de leads, sino todo lo contrario.

Mes	leads_karry	Inversion	ventas_karry	ventas_leads_ka	tasa_conve
Enero	575	$ 27.000.000	69	2	0,35%
Febrero	373	$ 23.240.377	58	6	1,61%
Marzo	488	$ 23.240.377	82	6	1,23%
Abril	1502	$ 13.999.755	63	9	0,60%
Mayo	1412	$ 17.000.000	79	13	0,92%
Junio	1283	$ 17.423.681	80	12	0,94%
Julio	1434	$ 24.914.754	78	10	0,70%
Agosto	1095	$ 21.982.623	72	17	1,55%

Tabla 2.b.3: Leads y Ventas para Karry 2023. (Elaboración propia)

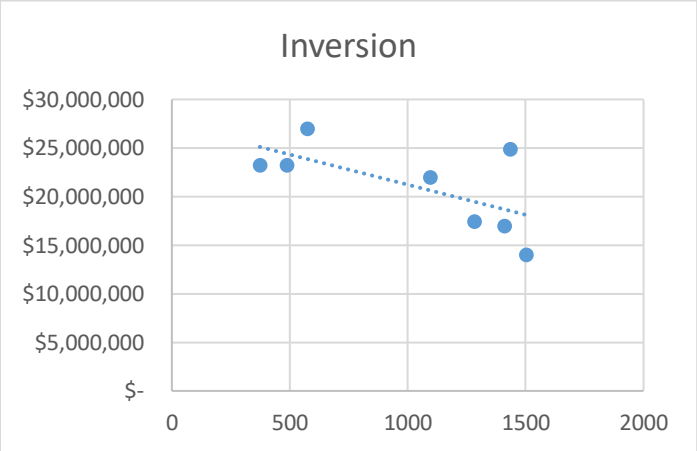


Figura *2.b.4: Correlación entre la inversión y leads en Karry. (Elaboración propia)*

En este contexto, resulta evidente la existencia de una problemática sustancial en lo que respecta a la ausencia de una predicción y priorización de leads, pues se busca establecer comunicación con todos los leads, no obstante, las ventas efectivas se mantienen en niveles bajos. Este problema se fundamenta en tres causas primordiales. (Figura 2.b.5)

En primer lugar, se halla vinculado a los procesos, es decir, a la carencia de un análisis exhaustivo de los leads, lo que dificulta la determinación de cuales tienen una mayor

probabilidad de conversión en ventas. Esta insuficiencia, provoca una asignación poco eficaz de recursos y esfuerzo de ventas.

En segundo lugar, se encuentra la limitación tecnológica, en donde la empresa carece de herramientas tecnológicas eficaces para predecir y priorizar la conversión de leads. Esto restringe la capacidad de la organización de tomar decisiones basadas en datos y obtener información precisa, lo que supone una limitación significativa en las estrategias comerciales.

Finalmente, la tercera causa va vinculada a la disponibilidad y comprensión de la información, pues Andes Motors se enfrenta a un desafío en términos de comprensión de las características que influyen en la conversión de algunos leads en ventas, mientras que otros no se convierten. Esta falta de entendimiento, dificulta las iniciativas de mejora en las estrategias de conversión, lo que puede resultar en reacciones tardías ante las oportunidades de ventas.

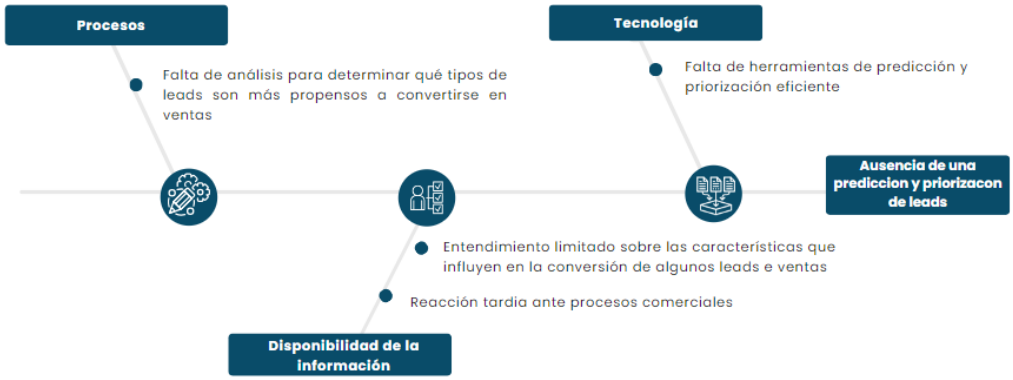


Figura *2.b.5: Espina de Ishikawa. (Elaboración propia)*

c. Contexto de la oportunidad

En el entorno altamente competitivo de la industria automotriz, la capacidad de convertir leads en clientes es esencial para el éxito empresarial. Por consiguiente, en Andes Motor se ha identificado que existe una baja tasa de conversión de leads a ventas, debido a la ineficiente predicción y priorización de leads.

De esta manera, se plantea la necesidad dentro de la empresa de implementar y desarrollar un sistema que permita tener un análisis efectivo de leads, el cual de acceso a identificar correctamente el comportamiento del lead.

La carencia de un sistema predictivo adecuado ha resultado en una asignación ineficiente de recursos y en tasas de conversión que no alcanzan el potencial de la empresa. Mensualmente, el área de vehículos livianos genera un promedio de 3265 leads para sus tres marcas de los cuales solo alrededor de 29 se transforman en ventas. La oportunidad existente a causa de este problema, es ver qué leads son los que tienen mayores probabilidades de ventas, lo que generaría ahorros para saber dónde distribuir según el origen de manera correcta.

2. Objetivos

En el contexto de los desafíos enfrentados por el área de vehículos livianos en Andes Motor, el objetivo principal de este proyecto es: “Automatizar la predicción y priorización de leads para vehículos livianos en Andes Motor, generando así un beneficio en el aumento de las ventas”.

a. Objetivo general

El objetivo general es crear un sistema de análisis predictivo que mejore la predicción y priorización de leads, con la meta de aumentar la tasa de conversión de leads a ventas en un 3%. Utilizando datos históricos y tecnología avanzada, se busca perfeccionar la identificación de clientes potenciales con altas posibilidades de compra. El proyecto aspira a mejorar significativamente la eficacia de la gestión comercial y se proyecta alcanzar este objetivo en un plazo de 6 meses.

b. Objetivos específicos

El objetivo general vendría siendo la suma de los objetivos específicos en este proyecto. En primer lugar, identificar el comportamiento actual que tienen los leads, según su origen y conversión en ventas. En segundo lugar, evaluar la capacidad del modelo predictivo para identificar patrones relevantes en la conversión de leads. En tercer lugar, generar una predicción y priorización de los leads, para así conocer la precisión del modelo respecto a la conversión de los leads.

c. Medidas de desempeño

Para evaluar el desempeño del proyecto de análisis predictivo de leads en Andes Motor, se utilizarán cuatro indicadores de rendimiento (KPI).

En primer lugar, se medirá el aumento en la cantidad de ventas que pueden atribuirse específicamente a los leads digitales. Esta se basa en la tasa de conversión de leads que indica los leads que se han convertido en ventas. Este se puede observar a través del siguiente KPI:

$$Tasa\ conversión\ de\ leads = \frac{números\ de\ leads\ convertidos}{número\ de\ leads\ totales} \times 100 \quad (1)$$

En segundo lugar, se evaluará la eficacia del modelo en la identificación correcta de leads con alta probabilidad de conversión. Esta se basa en el en la Tasa de aciertos (Hit Rate), en la que mide la proporción de predicciones correctas frente a las predicciones totales. Este se puede observar a través del siguiente KPI:

$$Tasa\ de\ aciertos = \frac{números\ de\ predicciones\ correctas}{número\ de\ Total\ de\ predicciones} \times 100 \quad (2)$$

En tercer lugar, se evaluará la precisión del modelo de análisis predictivo para comparar las predicciones generadas por el sistema con los resultados reales, utilizando Precisión del Pronóstico (Forecast Accuracy). Este se puede observar a través del siguiente KPI:

$$Precisión\ del\ pronóstico\ = \left(1 - \frac{\sum|Prediccion - Real|}{\sum Real}\right) \times 100 \quad (3)$$

3. Estado del arte

En este capítulo, se presentará el estado del arte, explorando el marco teórico y las investigaciones relevantes sobre analítica de datos y predicción de leads. Este análisis detallado establecerá el contexto necesario para el desarrollo de nuestra solución específica en Andes Motor.

a. Marco teórico

El núcleo del problema radica en la inexistencia de un sistema eficaz para predecir y priorizar leads, una carencia reflejada en la baja tasa de conversión de leads a ventas en proporción a su volumen. El proyecto se enfocará en analizar los datos de leads y las ventas efectuadas durante el año 2023, abarcando tanto el interés de potenciales clientes en adquirir un vehículo como la proporción que finalmente realiza una compra.

Para la recolección de datos de leads, se utiliza Amazon QuickSight, una herramienta de inteligencia empresarial basada en la nube que suministra información vital al equipo de marketing y ventas de Andes Motor. Por otro lado, los datos de ventas se registran manualmente en Excel para cada marca, detallando las especificaciones tanto del comprador como del concesionario.

b. Investigaciones

En la construcción de un proyecto enfocado en la predicción y priorización de leads, es imperativo realizar una investigación y análisis de la literatura existente y de los desarrollos previos. Sin embargo, tales investigaciones no están exentas de limitaciones. Tradicionalmente, el proceso de calificación de leads se ha manejado de manera manual (Figura 3.b.1), confiando en la experiencia de los especialistas en marketing para evaluar las características demográficas y de comportamiento de los clientes potenciales. Esta práctica, se enfrenta a críticas debido a su potencial ineficiencia y la posibilidad de sesgo (Jadli, 2022).

Figura 3.b.1: Proceso de gestión de leads de manera manual. (Jadli, 2022)

Por otro lado, nos encontramos con la calificación de leads de manera predictiva (Figura 3.b.2), la cual utiliza algoritmos de aprendizaje automático (Machine Learning) para descubrir patrones en datos históricos de ventas. Estos modelos tienen el potencial de revolucionar la calificación de leads al proporcionar un método más eficiente para predecir probabilidad de conversión. (Jadli, 2022).

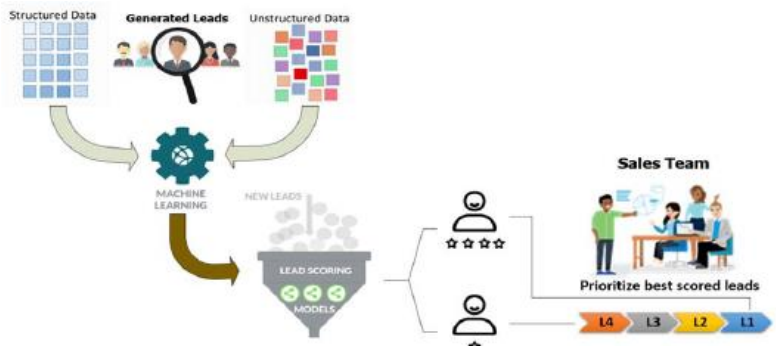


Figura 3.b.2: Proceso de gestión de leads de manera predictiva. (Jadli, 2022).

Respecto a la literatura relevante, hay una escasez de investigaciones que exploren la implementación de modelos automatizados de calificación de leads en el ámbito empresarial (Nygard y Mezet, 2020). Estas tecnologías podrían mejorar significativamente la eficiencia en la identificación y captación de leads, algo esencial para Andes Motor.

Dentro de este contexto, se examinó en primer lugar un estudio donde utilizan datos públicos “X Education”. El propósito de este estudio es evidenciar los beneficios de la automatización en el proceso de priorización de leads. Este proceso empieza con un análisis del conjunto de datos para determinar las variables relevantes, sus dimensiones y

categorías, así como la ejecución de un procedimiento de limpieza de datos, asegurando la precisión en el experimento subsiguiente.

Posteriormente, el estudio aborda una comparativa entre cinco algoritmos distintos de aprendizaje automático (ML), cada uno con características y aplicaciones específicas en la calificación y priorización de leads.

1. K-Vecinos Más Cercanos (K-Nearest Neighbors): Este algoritmo se basa en la proximidad para la clasificación, asignando etiquetas de clase a una instancia según la votación mayoritaria de sus vecinos más cercanos. Es especialmente útil para clasificar leads según su similitud con casos previamente exitosos. (Jadli, 2022). La selección de métricas de distancia en KNN, detallada en la Ecuación 3.b.1, es vital y se adapta según la naturaleza y distribución de los datos.

   - Euclidean Distance: $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$
   - Manhattan Distance: $\sum_{i=1}^{k}|x_i - y_i|$
   - Minkowski Distance: $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

   Ecuación 3.b.1: Métricas de distancia para KNN. (Jadli, 2022)

2. Naive Bayes (NB): Es un modelo de probabilidad basado en el teorema de bayes, ya que puede estimar la probabilidad de conversión de un lead basándose en características observadas. (Jadli, 2022). Este modelo permite una clasificación basada en la probabilidad de que una instancia pertenezca a una clase específica, tal como se observa en la Ecuación 3.b.2, donde $\rho(\ C_k\ |x)$ es la probabilidad de la clase C_k dado el vector de características *x.*

   $$\rho(C_k|x) = \frac{\rho(x|C_k)\rho(C_k)}{p(x)}$$

   Ecuación 3.b.2: Métricas de probabilidad para NB. (Jadli, 2022)

3. Máquina de Vectores de Soporte (Support Vector Machine): Es un clasificador binario que busca el hiperplano óptimo de separación entre dos clases (Jadli, 2022). Se utiliza para distinguir entre leads potencialmente valiosos y aquellos que no lo son, con una optimización detallada en la Ecuación 3.b.3 en donde w es el vector de pesos del hiperplano, ξi son las variables de holgura, y C es el parámetro de penalización.

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{I} \xi_i$$
$$w^T \phi(x_i) + b \geq 1 - \xi_i, \quad if\ y_i = m;$$
$$w^T \phi(x_i) + b \leq -1 + \xi_i, \quad if\ y_i \neq m;$$
$$\xi_i \geq 0$$

Ecuación 3.b.3: Métricas de probabilidad para SVM, (Jadli, 2022)

4. Árboles de Decisión (Decision Tree) y Bosques Aleatorios (Random Forest): Los Árboles de Decisión son modelos que funcionan como grafos de decisión, comenzando el proceso con la división del conjunto de datos en subconjuntos más pequeños. Este no solo divide los datos, sino que también memoriza la regla de clasificación en cada nivel, lo cual es crucial para construir progresivamente un árbol de decisión. (Jadli, 2022). La eficacia de un Árbol de Decisión se evalúa a través de la entropía de múltiples atributos (Ecuación 3.b.4), en donde para un conjunto de entrenamiento T y un atributo X, la probabilidad de clase es $P(c)$ y la entropía de clase es $E(c)$.

$$E(T, X) = \sum_{c \in X} P(c) E(c)$$

Ecuación 3.b.4: Métricas de probabilidad para DT (Jadli, 2022)

Avanzando hacia un enfoque más sofisticado, encontramos el algoritmo Bosques Aleatorios, que es una técnica de ensamble construida sobre la base de múltiples Árboles de Decisión. Cada árbol dentro de este "bosque" se desarrolla y realiza predicciones de forma independiente, la cual permite mejorar significativamente la precisión (Jadli, 2022). Al combinar diversas perspectivas y enfoques de clasificación, Bosques Aleatorios reduce de manera efectiva problemas comunes como la varianza y el sobreajuste. Esta característica hace que Bosques Aleatorios sea una mejor herramienta para la identificación efectiva de leads cualificados.

5. Regresión Logística (Logistic Regression): Es un método de clasificación binaria y lineal, en donde predice la probabilidad de conversión de un lead de una manera eficiente. El funcionamiento de esta se basa en la asignación de probabilidades a las instancias para determinar su pertenencia a una categoría específica (Jadli, 2022). En un modelo de clasificación multiclase de Regresión Logística, la probabilidad de que una muestra xi pertenezca a una categoría Ci se puede observar en la Ecuación 3.b.5.

$$\rho(C_i|x) = \frac{e^{w_i^T x + w_{0i}}}{\sum_{j=1}^{K} e^{w_i^T x + w_{0j}}}, i = 1, \dots, K$$

Ecuación 3.b.5: Métricas de probabilidad para LR (Jadli, 2022)

En este estudio, se emplean modelos de evaluación para comparar diferentes algoritmos. Primero, se utiliza la matriz de confusión (Tabla 3.b.1) para evaluar el rendimiento de los modelos en el conjunto de prueba.

Confusion Matrix		Actual values	
		Positive	Negative
Predicted Value	Positive	TP	FP
	Negative	FN	TN

Tabla 3.b.1: Matriz de confusión. (Jadli, 2022)

Además, se analizan métricas de evaluación de modelos, incluyendo precisión, sensibilidad, especificidad y F1-score, como se detalla en la Ecuación 3.b.6.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ no.of\ \ predictions}$$

$$Precision = \frac{True\ positives}{Total\ predicted\ positives} = \frac{True\ positives}{True\ positives\ +\ Fasle\ positives}$$

$$Recall = \frac{True\ positives}{Total\ actual\ positives} = \frac{True\ positives}{True\ positives\ +\ False\ negatives}$$

$$F_1\ Score = 2 \times \frac{Precision\ \times\ Recall}{Precision\ +\ Recall}$$

Ecuación 3.b.6: Métricas de desempeño de modelos. (Bateman, 2020)

Se incorpora también la técnica de validación cruzada K-Fold, una metodología estadística para probar la capacidad de generalización de los modelos. Esta técnica divide los datos en varios subconjuntos y utiliza cada uno de ellos, a su vez, como conjunto de prueba. Finalmente, se utiliza la curva ROC (Receiver Operating Characteristic) para medir la habilidad de los modelos en diferenciar entre clases, enfocándose en obtener el mejor puntaje AUC (Area Under Curve).

En conclusión, el estudio revela que la eficacia de los diversos modelos, más en específico, para la data estudiada los modelos Bosques Aleatorios y Árboles de Decisión ofrecen un

rendimiento superior y consistente, siendo el Bosques Aleatorios el más preciso y Árboles de Decisión el óptimo en términos de tiempo de ejecución.

En segundo lugar, se examina el caso de la empresa portuguesa HUUB, especializada en comercio digital, que se embarcó en el desarrollo de un modelo predictivo de calificación de leads utilizando técnicas de aprendizaje automático (ML), ya que busca mejorar la priorización de sus contactos para aumentar la eficiencia y las tasas de conversión.
La metodología utilizada es CRISP-DM (Cross Industry Standard Process for Data Mining), la cual consta de seis etapas. (Figura 3.b.3)

Figura 3.b.3: Metodología CRISP-DM (Hotz, 2023)

La flexibilidad de CRISP-DM permite una adaptación fluida a las exigencias del proyecto, permitiendo la integración de las fases de comprensión de los datos y preparación de datos. (Pereira, 2021)

1. Entendimiento del negocio: Colaboración cercana con el equipo de ventas de HUUB para recopilar información crucial sobre el proceso de ventas y establecer requisitos clave para el modelo de puntuación de leads.
2. Entendimiento de los datos y preparación de datos: Análisis detallado del CRM, resolviendo problemas de calidad de datos y seleccionando variables esenciales. Posteriormente, se realiza un análisis utilizando herramientas como Excel y Power BI para hacer un muestreo y ver que variables son más importantes. Limpieza de datos y enriquecimiento con el atributo "Resultado de Conversión".
3. Modelado: En esta etapa, se seleccionó e implementó un algoritmo de aprendizaje automático, eligiendo los árboles de decisión en RStudio. Este algoritmo se valoró

altamente debido a su capacidad para ofrecer una representación visual intuitiva, lo que resultó especialmente beneficioso para usuarios como el equipo de ventas, que pueden no estar familiarizados con la complejidad de los modelos de aprendizaje automático (Pereira, 2021). Los árboles de decisión también tienen la ventaja de manejar tanto datos numéricos como categóricos, además de realizar una selección de características durante el proceso de construcción del modelo. Se utilizó el algoritmo CART (Classification and Regression Tree), un enfoque común para crear árboles de decisión, que utiliza el Índice de Gini para medir la pureza de un conjunto de datos. Este enfoque ayudó a minimizar el riesgo de sobreajuste, una preocupación común con los árboles de decisión. Para mitigar este riesgo, se emplearon estrategias como la poda del árbol y se consideraron parámetros como la profundidad máxima y el número mínimo de muestras en un nodo hoja. Además, se prepararon conjuntos de entrenamiento y prueba, utilizando métodos de muestreo aleatorio para mantener una distribución equitativa de las clases de la variable de salida, lo cual es esencial para mantener la integridad y la representatividad del modelo. Estos pasos sólidos establecieron la base para una predicción precisa de los resultados de conversión de leads.

4. Evaluación: La evaluación del modelo se centró en métricas estándar de rendimiento, incluyendo precisión, sensibilidad, especificidad y precisión, basados en la matriz de confusión. Estos indicadores son esenciales para evaluar la eficacia del modelo en la predicción de los "Resultados de Conversión". Posteriormente, se realizó un análisis detallado de los resultados del proyecto para comprender la eficacia del modelo y su contribución al proceso de toma de decisiones. Aunque el modelo no alcanzó la perfección ni representó la solución óptima para los datos específicos del caso, su implementación marcó un avance significativo. Esto se reflejó no solo en una mejora cuantitativa de las tasas de conversión, sino también en la calidad del proceso de selección de leads.
5. Despliegue: Definición de estrategias de mantenimiento y monitoreo en colaboración con el equipo de ventas para garantizar la efectividad continua del modelo.

El estudio demuestra como un modelo de puntuación de leads puede implementarse y evaluarse en un entorno empresarial real, proporcionando una herramienta útil que ha realizado mejoras significativas en la tasa de conversión de leads y la eficiencia de ventas.

En tercer lugar, se examina el estudio de Fei Qu, en donde se destaca el análisis del algoritmo metaheurístico Grey Wolf Optimizer (GWO), aplicado a la predicción de ventas de automóviles. Este enfoque innovador utiliza la Regresión de Vectores de Soporte (SVR), optimizada por GWO, para modelar y predecir las ventas. La metodología se desarrolla en varias etapas clave:

1. Determinación de Factores Influyentes: Se identifican los factores económicos, de precios y de materia prima que inciden en las ventas de automóviles, complementándose con información sobre la valoración de los consumidores.
2. Fuente de Datos y Pretratamiento: Se recogen datos de fuentes estadísticas oficiales y registros de ventas de automóviles. Este proceso incluye la corrección de anomalías y la normalización de los datos para su posterior análisis.
3. Construcción del Modelo de Predicción de Ventas Multifactorial: Se emplea la SVR, una técnica de aprendizaje automático para realizar regresiones. En este contexto, se utiliza para analizar la relación entre los factores influyentes y las ventas de automóviles. La SVR se expresa mediante la Ecuación 3.b.7, donde $f(x)$ es la función de predicción, $\langle w, x \rangle$ representa el producto escalar entre los pesos y las variables, y b es un término de sesgo.

$$f(x) = \langle w, x \rangle + b$$

Ecuación 3.b.7: Regresión de Vectores de Soporte. (Qu, 2022)

4. Optimización de Parámetros del Modelo SVR: Se aplica el GWO, inspirado en el comportamiento de caza de los lobos grises, para afinar los parámetros del modelo SVR. Este proceso incluye tanto el parámetro de regularización como los del Kernel, en caso de utilizar un Kernel no lineal. La optimización imita el comportamiento de caza de los lobos, ajustando iterativamente la posición de las soluciones hacia la óptima, como se muestra en la Ecuación 3.b.8.

$$X(t+1) = X_{prey} - A \cdot D$$

Ecuación 3.b.8: Algoritmo Grey Wolf Optimizer. (Qu, 2022)

Esta metodología, integra el análisis estadístico con técnicas avanzadas de aprendizaje automático, ofrece un enfoque preciso y robusto para la predicción de ventas en la industria

automotriz. La combinación de SVR y GWO proporciona un modelo potente y flexible, capturando relaciones complejas entre múltiples variables y las ventas, mejorando la precisión como la robustez de las predicciones y abordando la complejidad y variabilidad de los datos en el sector automotriz.

4. Solución

Entre las múltiples investigaciones revisadas, se identifican diversas soluciones posibles que podrían abordar eficazmente el problema que enfrenta Andes Motor. En el contexto de este proyecto, se expondrán tres opciones distintas de soluciones derivadas de estos estudios, para luego escoger la solución final.

a. Alternativas de solución

La primera alternativa de solución deriva de la investigación realizada por Jadli en 2022, la cual sugiere experimentar con múltiples algoritmos de aprendizaje automático para la calificación de leads. Esta metodología incluye el uso de algoritmos como K-Vecinos más Cercanos (KNN), Naive Bayes (NB), Máquina de Vectores de Soporte (SVM), Árbol de Decisión (DT), Bosque Aleatorio (RF) y Regresión Logística (LR), en donde utiliza diversas métricas para determinar al más eficaz. El probar diferentes tipos de algoritmos permite identificar de mejor manera cuál es el que se adapta mejor a los datos y a la predicción de leads. Además, al evaluar con métricas asegura una evaluación de cada modelo, lo que facilita cual algoritmo es el más preciso para la predicción de leads. Sin embargo, no sigue un proceso metodológico estandarizado, lo que puede generar un desafío en la ejecución del proyecto. Además, la diversidad de algoritmos puede conllevar una mayor suma de tiempo, lo que puede generar retrasos en la implementación.

La segunda alternativa de solución proviene de la investigación de Pereira en 2021, la cual sugiere aplicar la metodología CRISP-DM para desarrollar un modelo de puntuación de leads utilizando árboles de decisión. Esta metodología proporciona un marco claro y sistemático, desde la comprensión de la empresa hasta el desarrollo del proyecto. Además, aplicar el algoritmo de árboles de decisión es una buena estrategia en empresas, ya que proporcionan una representación visual intuitiva, lo que para el equipo de ventas le será un beneficio al comprender fácilmente. Asimismo, son flexibles con los datos numéricos y categóricos y no asumen ninguna distribución de datos. Sin embargo, son muy inestables, ya que pequeñas variaciones, pueden generar árboles muy diferentes. Por otro lado, la elección de un solo algoritmo podría no ser lo más efectivo, pues se pueden pasar por alto modelos potencialmente más eficaces para la predicción de leads.

La tercera opción propuesta se deriva del estudio de Fei Qu en 2022, que explora la combinación del algoritmo metaheurístico Grey Wolf Optimizer (GWO) y la Regresión de Vectores de Soporte (SVR) para predecir las ventas de automóviles. Esta metodología se caracteriza por un doble enfoque, pues utiliza SVR para analizar cómo diversas variables influyen en las ventas y aplica GWO para afinar los parámetros de SVR. Si bien este método ofrece un análisis exhaustivo y una optimización efectiva, presenta retos significativos. Estos incluyen la complejidad de su implementación y el riesgo de sobreajuste en situaciones con datos escasos, como en la priorización de clientes potenciales. Además, requiere de una inversión considerable en recursos computacionales y acceso a infraestructura tecnológica avanzada, lo que puede resultar prohibitivo en términos de coste y tiempo.

b. Solución escogida

Considerando los aspectos positivos y negativos de las alternativas antes expuestas, se opta por una solución híbrida que combina la primera y segunda alternativa de solución, es decir, la metodología CRISP-DM de Pereira con la experimentación de múltiples algoritmos propuesta por Jadli. Este enfoque integrado aprovecha el enfoque sistemático de CRISP-DM y también experimenta con una variedad de algoritmos de aprendizaje automático (ML) que identifican la mejor predicción y priorización de leads.

Esta solución combinada ofrece un equilibrio entre una metodología robusta y la flexibilidad para explorar diversas opciones tecnológicas, maximizando así las posibilidades de éxito en la predicción y priorización efectiva de leads. La implementación de esta solución es altamente viable, ya que no se requieren recursos externos y se utilizara una herramienta tecnológica ya disponible, como RStudio. Se espera que esta solución aborde el desafío de predicción de leads en Andes Motor, mejorando significativamente su capacidad para identificar y priorizar oportunidades de negocio de manera efectiva.

5. Metodologías

En este capítulo, se explorará la metodología adoptada para desarrollar y ejecutar la solución propuesta en este proyecto. Se detalla el enfoque sistemático y las etapas específicas que guiarán el proceso de desarrollo, desde la conceptualización inicial hasta la implementación final.

a. Metodología para desarrollar la solución

Para asegurar una implementación y desarrollo efectivos de este proyecto, se seleccionará la metodología CRISP-DM (Figura 3.b.3). Esta metodología consta de seis fases clave y se elige por su estructura organizada y su capacidad de adaptarse a las necesidades del proyecto.

1. Comprensión del negocio: Identificar objetivos y necesidades específicas del área de vehículos livianos en Andes Motor, explorando las herramientas tecnológicas disponibles.
2. Comprensión de datos: Analizar y recolectar datos existentes, como leads y ventas, evaluando su formato, calidad, relevancia y relaciones.
3. Preparación de datos: Seleccionar y procesar variables clave, limpiar datos y consolidar bases de datos para formar un conjunto unificado que incluya información relevante. También se examinan las probabilidades de ocurrencia y las relaciones de dependencia entre las variables.
4. Modelado: Construir y evaluar distintos modelos predictivos utilizando algoritmos adecuados para los datos de leads y ventas.
5. Evaluación: Comparar el rendimiento de los modelos utilizando métricas específicas para determinar su efectividad (Ecuación 3.b.6).
6. Implementación: Determinar y preparar el entorno más adecuado para la integración del modelo predictivo en la empresa.

b. Desarrollo del proyecto

Para un correcto desarrollo del proyecto, este se divide en tres etapas principales:

1. **Etapa Inicial:** Comprende la comprensión integral del negocio y un análisis detallado de los datos existentes.
2. **Etapa de Desarrollo:** Se centra en la preparación de los datos, seguida de la construcción y evaluación de los modelos predictivos.
3. **Etapa Final:** Implica la implementación del sistema predictivo en el entorno operativo.

b.1. Etapa inicial

Se dispone de información del año 2023 para las tres marcas en estudio, que incluyen los leads generados y las ventas de Karry, Kaiyi y Jetour, los cuales serán examinados y modelados en RStudio.

Respecto a los datos de leads, estos son accesibles desde abril en Amazon Quicksight debido a un cambio de plataforma. De esta manera, se obtuvieron los datos de los meses previos de otra plataforma y se fusionó con la data existente proporcionada por Amazon Web Services. Implicando así que la data de leads (Tabla 5.b.1.1) para todo el año 2023 conste de 44.570 leads y 11 columnas con información de cada uno de ellos. Sin embargo, esta información está desorganizada y cuenta con pocas variables relevantes como el origen del lead, el RUT del cliente y el concesionario con sucursal, lo que indica una falta de datos detallados para un análisis predictivo eficaz.

createdate	marca	modelo	origen	sucursal	rut	nombre	apellido	email	telefono	lead_id
14/02/2023 11:31:00	KAIYI	kyx3	Liferay	fronza - quillota	10094048-5	luis eduard	fabres vera	lfabres@agropuelma	988397733	6
04/02/2023 20:18:00	KAIYI	kyx3	Liferay	circulo autos - linares	14466447-7	Cristian	avila	CRISTIAN.ELECTRICI	946541108	22
15/02/2023 18:30:00	KAIYI	kyx3	Liferay	maritano y ebensperger - chillán	15160382-3	eulalia	manriquez	Laly-1982@hotmail.c	981867906	26
jul. 24, 2023 6:32pm	KARRY	Q22	WHATSAPP	MARITANO Y EBENSPERGER - CONCEPCIÓN	9505012-3	Jorge	Maldonado	jorgemaldonadocid0	+569978974	003d92b7-
jul. 5, 2023 5:54am	KARRY	Q22 Cargo	WHATSAPP	MARITANO Y EBENSPERGER - CHILLÁN	19946439-6	Alexander	NULL	Juancarloechaiz@g	+569332927	008b5a63-
jul. 14, 2023 7:45pm	KARRY	Q22	WHATSAPP	MARITANO Y EBENSPERGER - CONCEPCIÓN	15224293-k	Guillermo	NULL	gctransportes.obras	+569335998	01062990-
jun. 28, 2023	KARRY	Q22	WHATSAPP	ANTIVERO - SAN FERNANDO	27620831-4	Yasmerys	NULL	yasmerysmemdez0	+569781492	010e5a94-
sep. 23, 2023	KARRY	Q22	WHATSAPP	AUTOMOTRIZ CARMONA - LA SERENA	19948304-8	Ronaldo	NULL	rgeraldo980@gmail.	+569341386	0120aa77-
jul. 11, 2023	KARRY	Q22	WHATSAPP	AUTOMOTRIZ TECNOSUR - VALDIVIA	14510466-1	Evaristo	NULL	evaristoreyesdiaz40	+569726852	0280fc1b-
jul. 25, 2023 6:08pm	KARRY	Q22	WHATSAPP	H. MOTORES - ANTOFAGASTA	81734800-9	Juan varas	NULL	Secretaria@autoclu	+569593259	03b809f5-
sep. 27, 2023	KARRY	Q22	WHATSAPP	AUTOMOTRIZ CARMONA - COPIAPÓ	9458157-5	Wilson	NULL	Olivareswilson71@h	+569907714	03ff68c0-

Tabla 5.b.1.1: Base de datos leads durante el periodo 2023.

Las bases de datos de ventas para las marcas Karry, Kaiyi y Jetour en Andes Motor registran 643, 557 y 611 ventas respectivamente, con 14 variables en cada una, incluyendo día, mes y año. Sin embargo, estas bases se limitan a datos básicos como concesionario, sucursal de venta y RUT del cliente. La falta de variables más detalladas limita la capacidad de análisis predictivo. La Tabla 5.b.1.2 proporciona una visión estructurada de estas bases de datos de

ventas.

MODELO	FECHA_WHOLE	AÑO_VEN	MES_VEN	DIA_VEN	MODELO	A_HOMOLO	FECHA_RV	Fecha_de_Ve	CONCESIONARIO	VIN	SUCURSAL	NOMBRE_CLIENTE	RUT CLIENTE
KYX3	17-01-2023	2023	1	17	KYX3 1.5L M	Solicitado	19-01-2023	17-01-2023	VESPUCIO NORTE - MO	LUYJB1G20RA000203	VESPUCIO	0	(
KYX3	17-01-2023	2023	1	20	KYX3 1.5L M	14-02-2023	13-02-2023	20-01-2023	DIFOR CHILE SOCIEDA	LUYJB1G23RA000227	PLAZA TOBALAB	MARCELO Cristian CANALES	17516501-0
KYX3	17-01-2023	2023	1	20	KYX3 1.5L M	03-02-2023	02-02-2023	20-01-2023	DIFOR CHILE SOCIEDA	LUYJB1G29RA000264	RANCAGUA	Mauricio Eduardo Lara Mec	12253511-8
KYX3	20-01-2023	2023	1	24	KYX3 1.5L M	10-02-2023	08-02-2023	24-01-2023	AUTOMOTRIZ CORDIL	LUYJB1G22RA000221	MOVICENTER	ZENAIDA FLORES SANDOVAL	23505814-6
KYX3	19-01-2023	2023	1	24	KYX3 1.5L M	14-02-2023	07-02-2023	24-01-2023	AUTOMOTRIZ CORDIL	LUYJB1G20RA000220	MOVICENTER	KATHERINE SOLEDAD ZENTE	15485373-1
KYX3	19-01-2023	2023	1	24	KYX3 1.5L M	08-02-2023	06-02-2023	24-01-2023	AUTOMOTRIZ CORDIL	LUYJB1G29RA000233	MOVICENTER	FRANCISCO LEOPOLDO MON	25707892-2
KYX3	30-01-2023	2023	1	25	KYX3 1.5L C	03-02-2023	03-02-2023	25-01-2023	AUTOMOTRIZ DANIEL	LUYJB2G25RA000302	LAS CONDES	RODRIGO ANDRES IBARRA H	12659620-0
KYX3	17-01-2023	2023	1	26	KYX3 1.5L C	09-03-2023	28-02-2023	26-01-2023	MARITANO Y EBENSPE	LUYJB2G29RA000299	LOS ANGELES	ALFONSO ENRIQUE ARANEDA	17450937-9
KYX3	19-01-2023	2023	1	27	KYX3 1.5L C	27-02-2023	17-02-2023	27-01-2023	AUTOMOTRIZ CORDIL	LUYJB2G21RA000300	PLAZA NORTE	DARIO ALBERTO SANDOVAL	25210861-0
KYX3	02-02-2023	2023	1	28	KYX3 1.5L C	08-02-2023	08-02-2023	28-01-2023	AUTOMOTRIZ DANIEL	LUYJB2G25RA000297	LAS CONDES	SCARLETT VIVIANA HUECHU	19362835-4
KYX3	19-01-2023	2023	1	31	KYX3 1.5L M	14-02-2023	13-02-2023	31-01-2023	AUTOMOTRIZ CORDIL	LUYJB1G28RA000255	MOVICENTER	FRANCISCO LEOPOLDO MON	25707892-2
KYX3	31-01-2023	2023	1	31	KYX3 1.5L M	14-02-2023	08-02-2023	31-01-2023	AUTOMOTRIZ CORDIL	LUYJB1G24RA000219	MOVICENTER	JESSICA ANDREA BORQUEZ C	15459291-1
KYX3	31-01-2023	2023	1	31	KYX3 1.5L M	14-02-2023	08-02-2023	31-01-2023	AUTOMOTRIZ CORDIL	LUYJB1G26RA000240	MOVICENTER	CLAUDIA GARCIA SUAREZ DE	26296388-8
KYX3	31-01-2023	2023	1	31	KYX3 1.5L M	14-02-2023	08-02-2023	31-01-2023	AUTOMOTRIZ CORDIL	LUYJB1G24RA000236	MOVICENTER	KHEVYN ABRAHAM NAVARR	26312389-1
KYX3	19-01-2023	2023	1	31	KYX3 1.5L M	14-02-2023	08-02-2023	31-01-2023	AUTOMOTRIZ CORDIL	LUYJB1G21RA000226	PLAZA NORTE	JOSE ELIAS MARTINEZ CANE	20138422-2
KYX3	17-01-2023	2023	2	2	KYX3 1.5L M	08-02-2023	08-02-2023	02-02-2023	AUTOMOTRIZ DANIEL	LUYJB1G22RA000266	LAS CONDES	CECILIA JOVITA SEGUEL PER	12281237-5
KYX3	17-01-2023	2023	2	2	KYX3 1.5L M	07-02-2023	06-02-2023	02-02-2023	AUTOMOTRIZ TECNOC	LUYJB1G25RA000214	OSORNO	CARLOS ANDRES SANHUEZA	17068642-K
KYX3	17-01-2023	2023	2	3	KYX3 1.5L M	14-02-2023	08-02-2023	03-02-2023	DIFOR CHILE SOCIEDA	LUYJB1G2XRA000273	PLAZA TOBALAB	FRANCISCO ALEJANDRO RAI	15163317-K
KYX3	17-01-2023	2023	2	9	KYX3 1.5L C	15-02-2023	15-02-2023	09-02-2023	AUTOMOTRIZ DANIEL	LUYJB2G23RA000296	LAS CONDES	CLAUDIA CAROLINA NU 16550825-4	
KYX3	17-01-2023	2023	2	9	KYX3 1.5L C	21-02-2023	21-02-2023	09-02-2023	AUTOMOTRIZ DANIEL	LUYJB2G28RA000293	LAS CONDES	ANGELICA DEL TRANSITO NU	69810240-3
KYX3	17-01-2023	2023	2	11	KYX3 1.5L M	03-03-2023	28-02-2023	11-02-2023	MARITANO Y EBENSPE	LUYJB1G23RA000275	CONCEPCION	JARED SAMA QUEZADA FAU	19577620-2
KYX3 PRO	14-02-2023	2023	2	13	KYX3 PRO 1.	08-03-2023	07-03-2023	13-02-2023	MARITANO Y EBENSPE	LUYJB2G20RA000501	CONCEPCION	PATRICIA DEL PILAR ARROYO	7503470-9

Tabla 5.b.1.2: Base de datos de ventas de Kaiyi para el año 2022.

De esta manera, se comienza con la etapa de preparación de las cuatro datas que utilizaremos para realizar el modelo predictivo. Para estructurar adecuadamente los datos históricos de leads, se adoptó un proceso detallado y riguroso. Tal como se muestra a continuación:

1. **Exclusión de Nombres de Prueba:** Elimina filas con nombres que sugieren ser datos de prueba. (como "test", "prueba", etc.).
2. **Dividir Nombres:** Se divide la columna nombre en dos nuevas columnas: ‘primer_nombre’ y ‘segundo_nombre’, en donde de acuerdo al primer_nombre se le asigna un género (Male/Female) utilizando la función genero(). Creando así una nueva columna ’género’.
3. **Limpieza de RUT:** Se limpia la columna ‘rut’ eliminando puntos, guiones y espacios. Luego, se renombra esta columna a RUT_CLIENTE.
4. **Identificación de Empresas:** Identifica si el RUT pertenece a una empresa basándose en su valor numérico, es decir, a aquellos RUT que tienen un número mayor a 40 millones, asumiendo que estos corresponden a empresas. De esta manera, se agrega a la columna género donde se asigna “Empresa”.
5. **Agrupamiento por RUT**: Se agrupan los RUT en grupos numéricos de cada 5 millones. Estos grupos numéricos, forman una nueva columna ‘grupo_rut’.
6. **Estandarización de Sucursales:** Se convierte a mayúsculas y se limpian los nombres de la ‘sucursal’, dividiéndolos en ‘concesionario’ y ’sucursal_limpia’.
7. **Asignación de Regiones:** Se asignan regiones basadas en el nombre de la sucursal.

8. **Renombrar concesionarios y regiones:** Se renombran variables para eliminar espacios y puntos en los nombres de concesionarios y sucursales, para que estén todas en el mismo formato.
9. **Renombrar Modelos:** Se renombran modelos de vehículos a un formato estándar y se filtran aquellos que son válidos.
10. **Traducción de Fechas:** Se transforman las fechas para que queden separadas en nuevas columnas día, mes y año. Se aplica una estandarización al compilar distintas datas donde se encontraban en distintos formatos.
11. **Eliminación:** Se identifican y eliminan registros duplicados basados en el RUT del cliente. Además, se eliminan los datos donde género sea NA.

Por consiguiente, la data de los leads fue preparada y con todos estos cambios han quedado 25.767 leads, por lo que ahora la se fusionará con los datos de las ventas totales de las tres marcas con los de leads. (Tabla 5.b.1.3)

1. **Fusión de leads con Datos de Ventas:** Se fusionan los leads con los datos de ventas correspondientes, mediante 'RUT_CLIENTE'
2. **Creación de la Columna 'VENTA':** Se crea una nueva columna que indica si el lead resultó en una venta (Si/No), basándose en la presencia de un número VIN.
3. **Conversión de Variables a Factores:** Se convierten las columnas año y mes a factores, que son tipos de datos categóricos en R.

marca	modelo	origen	concesionario	sucursal	region	rut	RUT_CLIENTE	nombre	apellido	email	telefono	mes	año	Columna1	primer_nombre	genero	MODELO2	FECHA_WHOLESALE	AÑO_VENTA	MES_VENTA
JETOUR	Dashing	RRSS	DANIEL_ACHONDO	LAS_CONDES	RM_SANTIAGO	8323267-7	83223677	Angelo Rolando	Mo	angelo5085@gmail.com	999999999	6	2023	NA	Angelo	male	DASHING 1.5T 6DCT	2023-06-23	2023	8
JETOUR	DASHING	Liferay	SALAZAR_ISRAEL	MOVICENTER	RM_SANTIAGO	7557565-3	75575653	Cristian	Scheib	scheib@motorrad.cl	998223637	6	2023	NA	Cristian	male	DASHING 1.6T 7DCT	2023-06-22	2023	6
JETOUR	Dashing	RRSS	FRONZA	LA_CALERA	V_VALPARAISO	6553101-1	65531011	Daniel	Fernandez	divalenzuela.propiedades@gmail.com	56998956687	6	2023	NA	Daniel	male	DASHING 1.6T 7DCT	2023-06-30	2023	7
JETOUR	DASHING	Liferay	MARITANO_Y_EBENSPERGER	TEMUCO	IX_LA_ARAUCANIA	19929344-3	199293443	Antonia	Castro	ignaciocv25@gmail.com	956757060	6	2023	NA	Antonia	female	DASHING 1.5T 6DCT	2023-07-24	2023	7
JETOUR	DASHING	Liferay	H_MOTORES	MAIPÚ	RM_SANTIAGO	18694142-K	18694142K	Daniela	Sepúlveda	cataldo.s.daniela@gmail.com	950475752	6	2023	NA	Daniela	female	DASHING 1.5T 6MT	2023-06-29	2023	6
JETOUR	DASHING	Liferay	H_MOTORES	ANTOFAGASTA	II_ANTOFAGASTA	18371501-1	183715011	Jorge	Quijada	jorge.quijada.pasten@gmail.com	992527777	6	2023	NA	Jorge	male	DASHING 1.6T 7DCT	2023-06-23	2023	6
JETOUR	DASHING	Liferay	DIFOR	MALL_PLAZA_VESPUCIO	RM_SANTIAGO	17599754-7	175997547	Camila	Guzman	camila_0990@hotmail.com	958238382	6	2023	NA	Camila	female	X70 PLUS 1.5T 6DCT	2023-07-31	2023	8

DIA_VENTA	MODELO	FECHA_HOMOLOGACION	FECHA_RVM	Fecha_de_Venta	CONCESIONARIO	VIN	REGION	SUCURSAL	NOMBRE_CLIENTE	RUT CLIENTE	VENTA	grupo_rut	Cohorte
9	DASHING	45152	2023-08-11	1691539200	AUTOMOTRIZ CORDILLERA S.A.	HJRPBGFBXRB150771	RM - SANTIAGO	vitacura	ANGELO MARTINO ROLANDO MOCELLI	8.322.367-7	SI	2	jun. 2023
30	DASHING	45120	2023-07-07	1688083200	AUTOMOTRIZ CORDILLERA S.A.	HJRPBGGB5RB150868	V - VALPARAISO	vitacura	CRISTIAN JAVIER SCHEIB CAMPOS	7557565-3	SI	2	jun. 2023
31	DASHING	45148	2023-08-08	1690761600	AUTOMOTRIZ CORDILLERA S.A.	HJRPBGGBXRB150834	RM - SANTIAGO	Plaza Norte	DANIEL FRANCISCO VALENZUELA FE	6.553.101-1	SI	2	jun. 2023
28	DASHING	45154	2023-08-16	1690502400	MARITANO Y EBENSPERGER LTDA	HJRPBGFB8RB152146	IX - ARAUCANIA	Temuco	ANTONIA BELEN CASTRO BUSTAMANTE	19.929.344-3	SI	4	jun. 2023
30	DASHING	45113	2023-07-04	1688083200	H.MOTORES S A	HJRPBGFA5RB150703	RM - SANTIAGO	.ONQUEN	DANIELA ALEJANDRA CATALDO SEPULVEDA	18694142-K	SI	4	jun. 2023
30	DASHING	45113	2023-07-05	1688083200	H.MOTORES S A	HJRPBGGB3RB150867	II - ANTOFAGASTA	ANTOFAGASTA	JORGE IGNACIO QUIJADA PASTEN	18371501-1	SI	4	jun. 2023
26	X70 PLUS	NA	2023-08-28	1693008000	AUTOMOTRIZ CORDILLERA S.A.	HJRPBGFB5PF023111	RM - SANTIAGO	Las Condes	CAMILA IGNACIA GUZMAN FUENTES	17.599.754-7	SI	4	jun. 2023

Tabla 5.b.1.3: Fusión de bases de datos de leads y ventas. (Elaboración propia)

Finalmente, se hizo un análisis exploratorio de los datos, para entender mejor la distribución y las características de los leads. Esto incluyó análisis univariado, bivariado, multivariado y de cohorte, además de pruebas chi cuadrado.

A partir de ello, algunos datos importantes es que, se obtuvo que del total de leads generados en el año 2023, solo el 1,05% ha logrado a transformarse a venta. También, podemos observar que la mayoría de los leads provienen de la región Metropolitana, seguida por Biobío y Valparaíso, lo cual está directamente relacionado con las ventas resultantes de estos leads. (Figura 5.b.1.1)

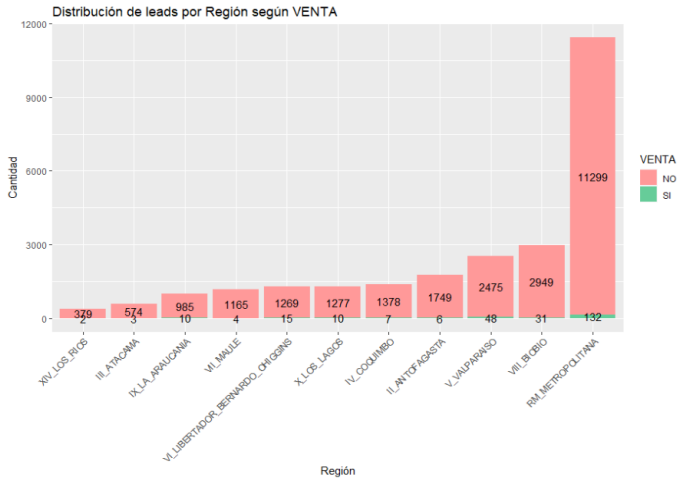


Figura 5.b.1.1: Distribución de leads por región según ventas 2023. (Elaboración propia)

Asimismo, se llevó a cabo un análisis de cohorte segmentando a los clientes por RUT en rangos de cinco millones, de 0 a 100 millones, para examinar su comportamiento en el tiempo. Se observa que los grupos de RUT 2, 3 y 4 son los que consistentemente han realizado más cotizaciones. (Figura 5.b.1.2)

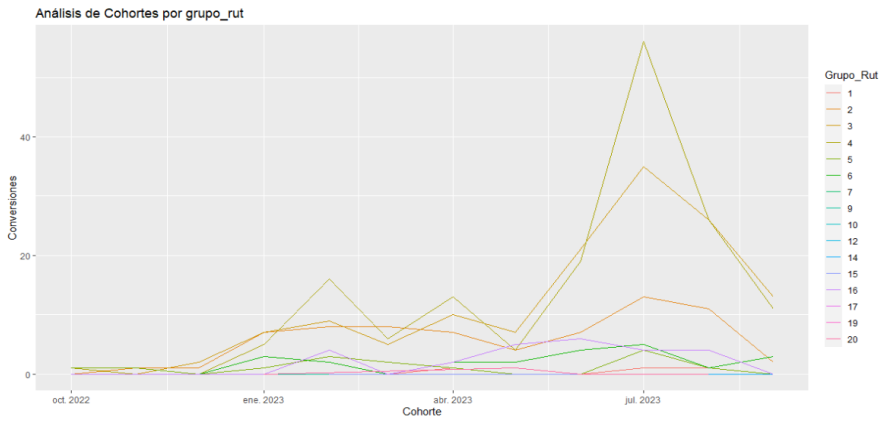


Figura 5.b.1.2: Distribución de leads por grupo_rut según ventas 2023. (Elaboración propia)

Además, se evidencia una asociación entre el mes de adquisición del lead y la conversión en

venta, sugiriendo una influencia estacional o mensual en las conversiones de venta. Adicionalmente, se identifica una relación significativa y significativa entre el modelo del vehículo y la conversión en venta. Por último, se destaca una asociación significativa entre el origen del lead y la conversión en venta. (Figura 5.b.1.3)

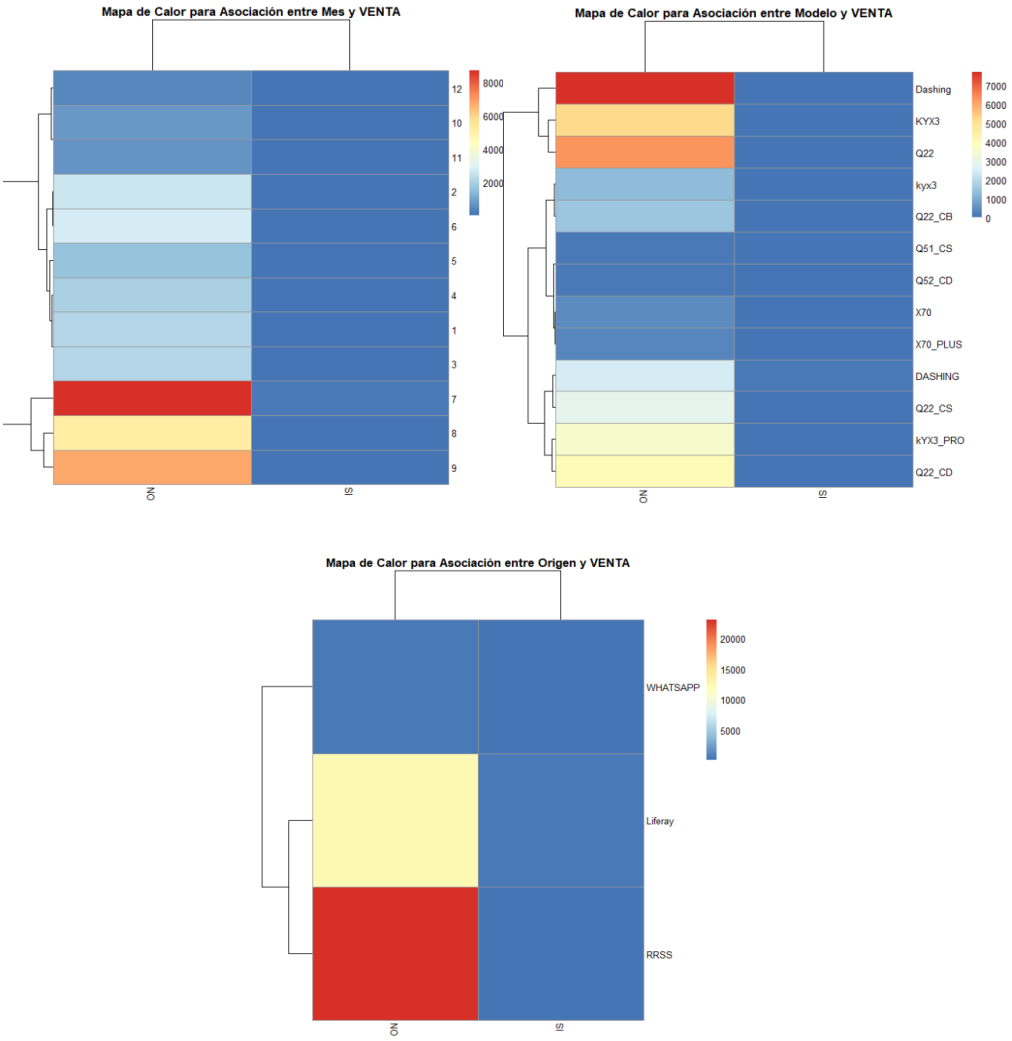


Figura 5.b.1.3: Mapas de asociación de mes, modelo y origen de leads con la conversión en venta. (Elaboración propia)

b.2 Etapa de desarrollo

En base al análisis previo, se procedió a la preparación de datos para construir un modelo clasificatorio destinado a predecir qué lead resulta en venta. En primera instancia, se seleccionan y retienen sólo aquellas variables con alta relevancia para el resultado de la venta. Además, se transforma la variable objetivo ´VENTA´, en donde se le colocan valores

binarios, asignando el valor 1 para “Si” y 0 para “No”. De esta manera, se crea la data *dummies* (Tabla 5.b.2.1) la cual es una tabla con variables binarias, que asignan el valor 1 o 0 dependiendo de la presencia o ausencia de un atributo para cada variable. Este nuevo conjunto de datos es esencial para la modelación, ya que permite a los algoritmos de aprendizaje automático interpretar y aprender de datos categóricos.

marcaKARRY	modeloDashing	modeloDASHING	modelokyx3	modeloKYX3	modeloKYX3_PRO	modeloQ22	modeloQ22_CB	modeloQ22_CD	modeloQ22_CS
1	0	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	1	0
1	0	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	1	0
1	0	0	0	0	0	0	0	0	1

Tabla 5.b.2.1: Parte de la base de datos dummies. (Elaboración propia)

Para la selección de un modelo de aprendizaje automático, se debe dividir la data final en dos subconjuntos: un 80% de los datos para entrenamiento y un 20% de los datos para pruebas. Esta subdivisión es esencial para desarrollar un algoritmo que no solo aprenda eficazmente, sino que también realice predicciones efectivas

Cabe mencionar que se identifica un desbalance de clases en el conjunto de entrenamiento, donde las ventas, siendo la clase minoritaria, presenta muy pocas muestras. Para abordar este desafío, se aplican técnicas de balanceo de clases. Entre estas, se incluyen la penalización para compensar las clases, subsampling en la clase mayoritaria y oversampling para generar muestras sintéticas de la clase minoritaria. Estas técnicas son fundamentales para garantizar que el modelo no esté sesgado hacia la clase mayoritaria.

Por consiguiente, de los diferentes tipos de algoritmos de predicción existentes, se utilizaron cinco modelos, incluyendo Regresión Logística, Árboles de Decisión, Bosques Aleatorios, XGBoost y CatBoost. Estos modelos fueron elegidos por su capacidad para manejar datos heterogéneos y proporcionar un balance óptimo entre rendimiento, interpretabilidad y eficiencia. Su implementación permite obtener predicciones precisas y facilita la toma de decisiones informadas basadas en datos.

Inicialmente, se evaluaron los modelos aplicando el balance de clases en los conjuntos de prueba. Para analizar cómo cada modelo se adapta a las diferentes técnicas de balanceo, se

emplearon herramientas de evaluación clave. Se utilizó la matriz de confusión (Tabla 3.b.1) para observar la distribución de las predicciones frente a los valores reales. Además, se calcularon métricas fundamentales como precisión, recall, accuracy y F1 score (Ecuación 3.b.5), que proporcionan una visión integral de la efectividad del modelo. Finalmente, se examinó el área bajo la curva (AUC) para medir la capacidad de los modelos para distinguir entre las clases.

En el caso específico de la Regresión Logística, se optó por la técnica de penalización para compensar las clases, dado que esta metodología mostró los mejores resultados respecto a las métricas. Para evaluar la capacidad del modelo de diferenciar entre clases positivas y negativas, se generó una curva ROC (Figura 5.b.2.1). Esta curva indica un desempeño razonablemente bueno, ya que se aproxima más al borde superior izquierdo que a la línea diagonal, lo cual es indicativo de una mayor capacidad de distinción. Sin embargo, el modelo no alcanza la perfección y existe un margen significativo para su mejora.

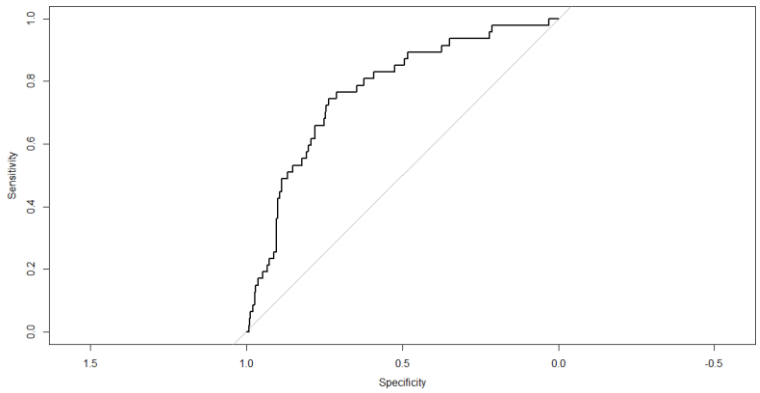


Figura 5.b.2.1: Gráfico curva ROC para regresión logística con penalización para compensar clases. (Elaboración propia)

Asimismo, para el algoritmo Bosque Aleatorio se optó por el balanceo de clases mediante oversampling de la clase minoritaria, estrategia que demostró ser la más eficaz para este modelo. Se observó que ciertas variables, especialmente 'origen', tienen una influencia significativa en la precisión de las predicciones del modelo. Además, el modelo asigna una importancia a determinadas regiones y modelos de vehículos, lo que resulta crucial para la exactitud de sus resultados (Figura 5.b.2.2.)

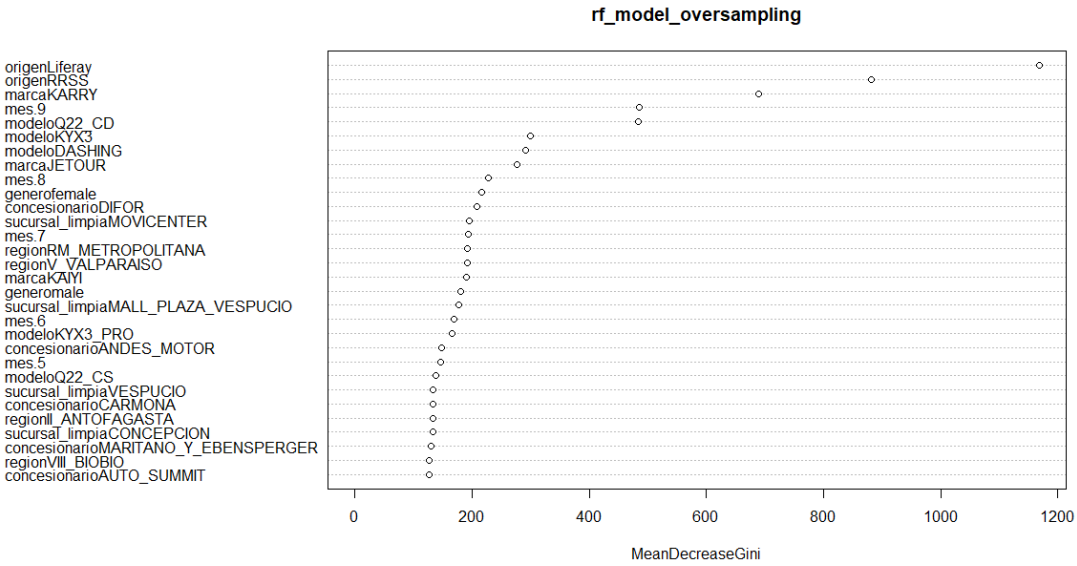


Figura 5.b.2.2: Importancia de las variables predichas por bosques aleatorios con oversampling en la clase minoritaria. (Elaboración propia)

De manera paralela, se observa que el modelo de Árbol de Decisión utiliza primordialmente la variable 'origen' para la primera división de los datos (Figura 5.b.2.3) lo que sugiere que es un predictor importante para la clasificación. Las subdivisiones subsiguientes parecen basarse en otras variables como 'marca' y 'región', indicando que también contribuyen a la clasificación final.

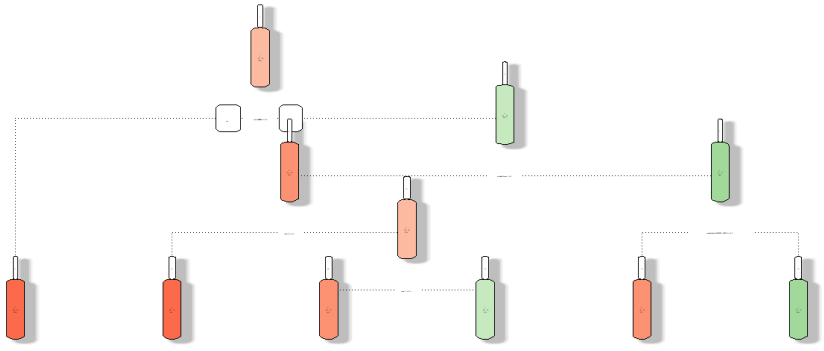


Figura 5.b.2.3: Decisiones tomadas en Árboles de Decisión para clasificar los datos. (Elaboración propia)

En cuanto al modelo XGBoost, se optó por el oversampling de la clase minoritaria, dada su efectividad demostrada en el rendimiento del modelo. Un análisis revela que ciertas

variables ejercen una influencia notable en las predicciones (Figura 5.b.2.4). Esto indica que el modelo se apoya significativamente en variables como 'origen', 'modelo', 'mes' y 'género' para efectuar sus clasificaciones.

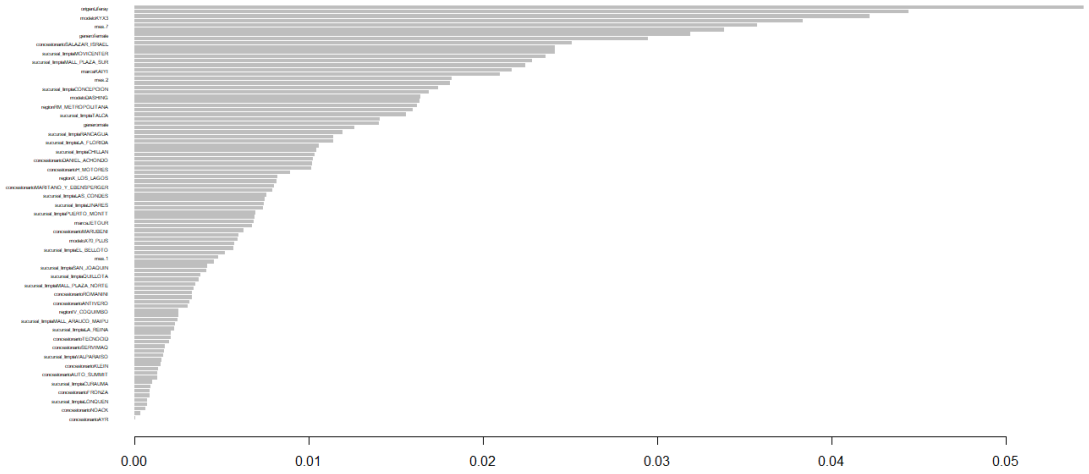


Figura 5.b.2.4: Importancia de las variables por XGBoost con oversampling en la clase minoritaria. (Elaboración propia)

Por otro lado, para el modelo CatBoost se decidió implementar el oversampling de la clase minoritaria, estrategia que ha mostrado un excelente rendimiento en las pruebas realizadas. Se evaluó el modelo mediante una curva ROC (Figura 5.b.2.5), en donde se presenta una sensibilidad cercana al 40% (eje y) y la especificidad cercana al 100% (eje x), lo que sugiere en una gran precisión al identificar no ventas, pero con margen de mejora en la detección de ventas efectivas.

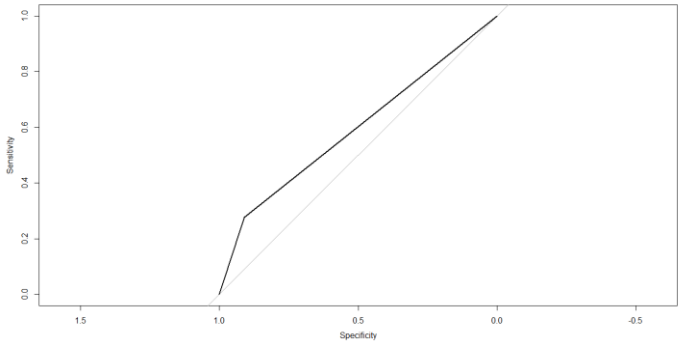


Figura 5.b.2.5: Gráfico curva ROC para CatBosost con oversampling de la clase minoritaria. (Elaboración propia)

Las métricas de la Tabla 5.b.2.2 revelan que, si bien todos los modelos tienen un rendimiento adecuado, XGBoost y CatBoost sobresalen. Por lo tanto, se ha elegido utilizar Catboost para la predicción de leads en Andes Motor debido a su excepcional manejo de variables categóricas y su robustez frente al sobreajuste. A pesar de que XGBoost muestra un rendimiento ligeramente superior, CatBoost se alinea mejor con la estructura de datos empresariales y las exigencias prácticas de la industria, ofreciendo un equilibrio óptimo entre precisión y practicidad operativa. Estas cualidades hacen que CatBoost sea la elección preferida para la implementación en nuestro entorno de negocio.

Modelo	Accuracy	Precision	Recall	F1_Score
XGBoost Oversampling	0,9909	0,9909	1,0000	0,9954
CatBoost Oversampling	0,9042	0,9927	0,9099	0,9495
Bosques Aleatorios Oversampling	0,8229	0,9943	0,8259	0,9023
Regresión Logística Penalizada	0,7452	0,9963	0,7456	0,8530
Árbol de Decisión Oversampling	0,7253	0,9968	0,7251	0,8395

Tabla 5.b.2.2: Evaluación métricas de algoritmos. (Elaboración propia)

b.3 Etapa final

La fase final del proyecto marca la implementación del modelo de predicción para la empresa, en donde se instala el programa R en los computadores de los tres encargados de cada marca de vehículos livianos. De esta manera, se permite el acceso directo al modelo mediante Shiny Server de R, que proporciona una interfaz web interactiva.

Pese a contemplar el uso de servicios en la nube como Amazon Web Services para mejorar la accesibilidad, la integración a esta es compleja y está sujeta a las políticas de Kaufmann. Por ahora, Andes Motor ha decidido mantener la solución interna mientras se realiza un análisis económico para una posible futura migración a la nube, dependiendo de las decisiones estratégicas y logísticas de la corporación.

La implementación de Shiny Server permite a los usuarios de la empresa cargar un archivo Excel con los datos de los leads en una interfaz web (Figura 5.b.3.1). Una vez cargada la

información, se presenta una tabla interactiva que permite filtrar variables según necesidades específicas y exportar la información procesada en varios formatos como CSV, Excel y PDF.

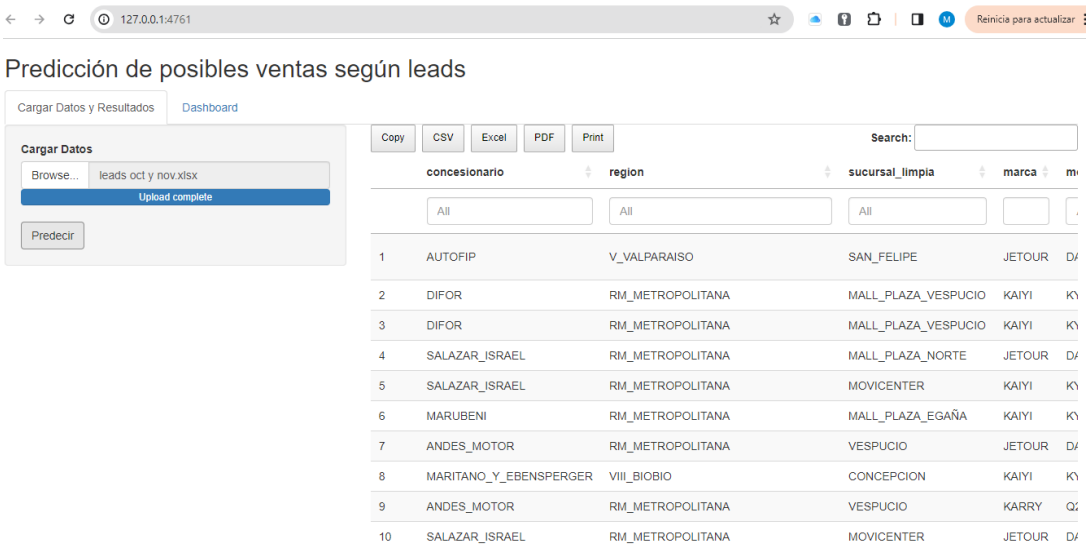


Figura 5.b.3.1: Interfaz web que entrega la predicción de ventas según leads. (Elaboración propia)

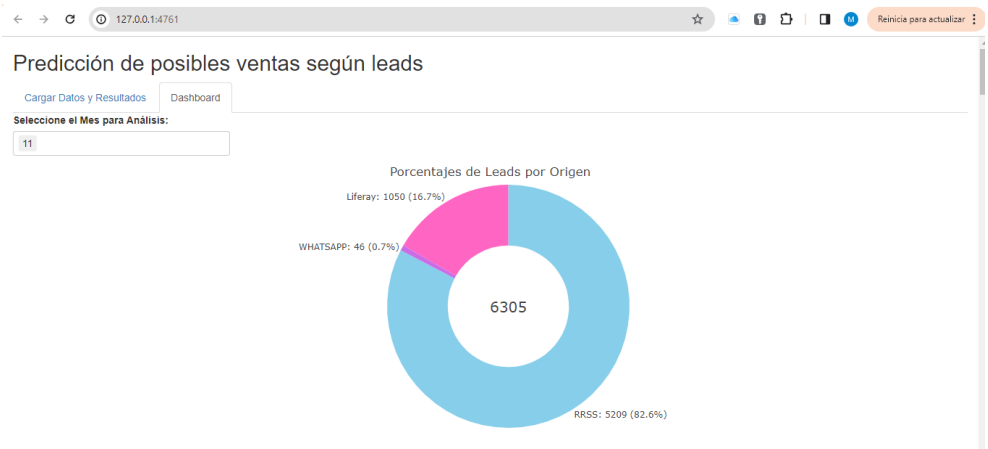
La tabla resultante incluye variables clave; concesionarios, región, sucursal_limpia, marca, modelo, origen, rut, nombre, apellido, género, email, teléfono, lead_id, día, mes y predicción venta. La columna objetivo "Predicción_Venta" muestra un valor binario que predice la probabilidad de conversión del lead, en donde un '1' indica una alta probabilidad de venta (lead calificado positivamente) y un '0' sugiere una baja probabilidad (lead calificado negativamente), permitiendo una segmentación eficiente para la estrategia de ventas. (Figura 5.b.3.2)

	telefono	lead_id	dia	mes	Prediccion_Venta
	All	All			All
pumil07@gmail.com	+56988869358	678381104479813	30	11	1
usbastardo@gmail.com	973501890	13987	30	11	1
iezpiro@gmail.com	+560950864527	6902979073150338	30	11	0
la750@gmail.com	+56965733968	276227531646052	30	11	0
gmail.com	+56958235599	883867746279212	30	11	1
@hotmail.com	+56994415865	344951224826459	30	11	0
.parra@gmail.com	+56948555732	380165087686398	30	11	1
nhueza@yahoo.com	+56974291918	1038978097363903	30	11	0
25.12.2017@gmail.com	+56983513439	7115317821822581	30	11	0
ıandez80@gmail.com	+56954193615	858876722905233	30	11	0
@gmail.com	+56933577147	331722789596154	30	11	1

Figura 5.b.3.2: Interfaz web que entrega la predicción de ventas según leads. (Elaboración propia)

La implementación de este modelo predictivo representa un avance significativo para la empresa, ya que permite a los concesionarios identificar con precisión qué leads priorizar, basado en su probabilidad de convertirse en ventas, mejorando así el proceso de toma de decisiones.

Además, se enriqueció la interfaz web con una pestaña de 'Dashboard', proporcionando un panel de control para un mejor análisis, el cual permite a los usuarios examinar métricas mes a mes. Proporciona visualizaciones claras a través de gráficos que desglosan los leads por origen, además de la distribución y predicciones de ventas por concesionario, región, modelo y marca, tal como se pueden observar en algunos gráficas de la Figura 5.b.3.3.



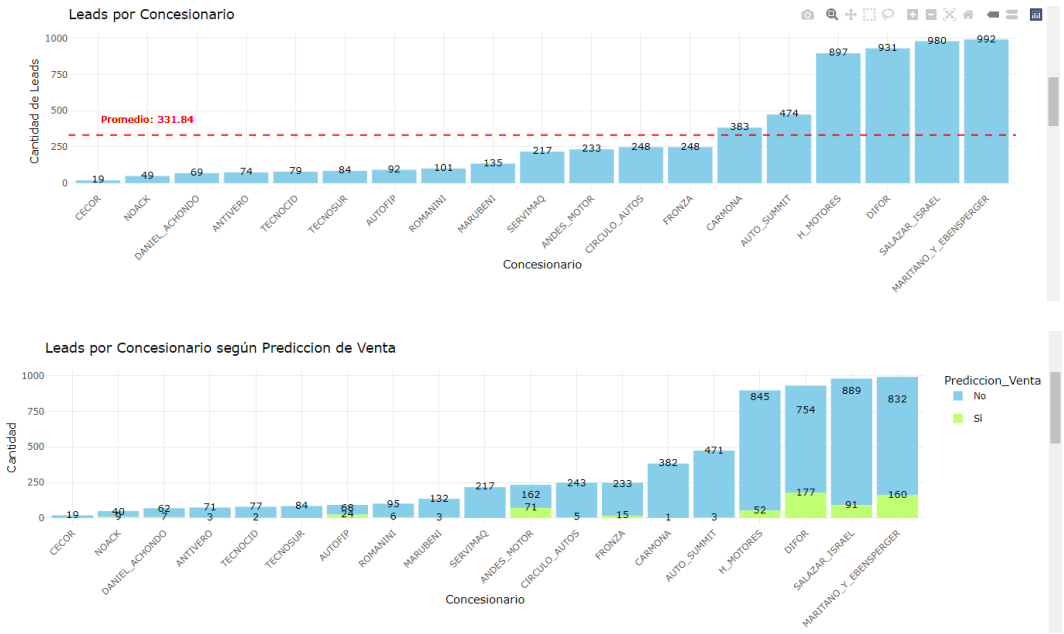


Figura 5.b.3.3: Gráficas de la interfaz web según leads y predicciones de venta.

(Elaboración propia)

Este 'Dashboard' integra una tabla exportable que compara los datos actuales con los del mes anterior, evidenciando la variación en leads y predicciones de ventas, y exhibiendo el porcentaje de participación en el mercado (Figura 5.b.3.4).

Leads por punto de venta

Excel CSV Copy Print

Search:

	region	concesionario	sucursal_limpia	LeadsMesActual	LeadsMesAnterior	LeadsAmbosMeses	VariacionLeads	PrediccionVentaMesActual	PrediccionVentaMesAnterior	VariacionPrediccionVenta	PorcentajeParticipacionMercadoTotal
1	III_ATACAMA	CARMONA	COPIAPO	116	65	181	78.462	0	0		1.84
2	II_ANTOFAGASTA	H_MOTORES	ANTOFAGASTA	329	248	577	32.661	19	15	26.667	5.218
3	IV_COQUIMBO	CARMONA	LA_SERENA	221	208	429	6.25	0	0		3.505
4	IV_COQUIMBO	CARMONA	OVALLE	46	18	64	155.556	1	0		0.73
5	IX_LA_ARAUCANIA	MARITANO_Y_EBENSPERGER	TEMUCO	200	122	322	63.934	10	2	400	3.172
6	RM_METROPOLITANA	ANDES_MOTOR	VESPUCIO	233	202	435	15.347	71	39	82.051	3.695
7	RM_METROPOLITANA	AUTO_SUMMIT	MALL_PLAZA_SUR	357	314	671	13.694	0	0		5.662
8	RM_METROPOLITANA	AUTO_SUMMIT	SAN_JOAQUIN	117	98	215	19.388	3	3	0	1.856
9	RM_METROPOLITANA	CECOR	DEPARTAMENTAL	19	28	47	-32.143	0	0		0.301
10	RM_METROPOLITANA	DANIEL_ACHONDO	LAS_CONDES	69	56	125	23.214	7	4	75	1.094

Figura 5.b.3.4: Tabla de la interfaz web según leads y predicciones de venta.

(Elaboración propia)

La implementación de este modelo predictivo representa un avance significativo para la empresa, ofreciendo a los concesionarios una herramienta precisa para priorizar leads según su potencial de conversión en ventas y por ende, mejorando así el proceso de toma de

decisiones. Además, la interfaz interactiva 'Dashboard' no solo fortalece la gestión de leads, sino que también respalda las decisiones estratégicas de la empresa.

Este sistema aborda las deficiencias del QuickSight de Amazon (donde se cargan los leads), mejorando la asignación de leads cuando no se sincronizan correctamente con los CRM de cada concesionario. Pues, los filtros automatizados y las variables de región y sucursal integradas en la web, agilizarán la distribución y selección de información de concesionarios se vuelve más eficiente, haciendo este sistema un recurso valioso y ya operativo para la empresa.

c. Matriz de Riesgos

Para evaluar los riesgos asociados al proyecto, hemos desarrollado una matriz de riesgos para evaluar la probabilidad y severidad de seis riesgos identificados en el proyecto. La matriz también proporciona estrategias para mitigar estos riesgos, con el objetivo de prevenir o reducir su impacto (Tabla 5.c.1).

Tema	Riesgos	Probabilidad	Severidad	Impacto	Mitigación
Recuperación de datos	Datos faltantes y/o erróneos en la recopilación las datas	2	4	8	Corregir datos en la base de datos original y ejecutar una seleccion correcta de los datos
Uso de datos	Baja calidad de los datos utilizados	1	5	5	Verificar los datos y hacer limpieza de ellos
Organización	Resistencia al cambio por parte del equipo de ventas	2	5	10	Realizar sesiones de capcitacion sobre los beneficios del sistema y su uso.
Modelo Preedictivo	Problemas en la precisión del modelo	2	4	8	Probar distintos modelos de aprendizaje automatizado y una selección colecta de variables a utilizar
Desarrollo	Alto tiempo de creación del modelo predictivo y adaptarlo	3	4	12	Investigar los procesos de adaptación del modelo
Evaluación del modelo	Poco tiempo para medir el desempeño del modelo	5	4	20	Implementar el modelo de manera mas pronta y medir oportunidades a corto plazo

Tabla 5.c.1: Matriz de riesgos y mitigaciones. (Elaboración propia)

Se observa que tres riesgos tienen un impacto moderado y tres tienen un impacto alto, pero con las medidas de mitigación adecuadas, se pretende proteger la integridad y continuidad del proyecto.

6. Resultados

A continuación se presentarán los hallazgos clave obtenidos tras la puesta en marcha del proyecto, incluyendo el rendimiento del modelo predictivo y la evaluación económica del mismo.

a. Resultados

El primer resultado obtenido es la creación de una herramienta avanzada de priorización de leads, diseñada para proporcionar mayor claridad y precisión en la identificación de los leads más valiosos, permitiendo así que el equipo de ventas dirija sus esfuerzos de manera más efectiva.

Para comprender mejor el impacto del modelo predictivo, se realizó una comparativa entre dos periodos de tiempo, el mes de puesta en marcha del proyecto y un mes previo, para establecer una línea base de comparación. (Tabla 6.a.1)

Periodo	24 Nov - 06 Dic	24 Ago - 06 Sep
Leads generados	3997	2444
Leads evaluados por el modelo	2963	1941
Predicción como No venta	2678	1776
Predicción como Si venta	285	165
Ventas efectivas	10	5
Ventas efectivas predichas por el modelo	9	4
Tasa de Conversion	0,337%	0,257%
Tasa de Aciertos	59,20%	44%
Precisión del pronostico	88,8%	75%

Tabla 6.a.1: Resultados obtenidos con el modelo para dos periodos. (Elaboración propia)

El proyecto se inició el 24 de noviembre, y hasta el 6 de diciembre, se estudiaron 2963 leads de los cuales 2678 fueron marcados como no venta y 285 como posibles ventas. De estas oportunidades, 10 culminaron en ventas reales, lo que se traduce en una tasa de conversión del 0,337% durante ese periodo. Además, la tasa de aciertos fue del 59,20%, y la precisión del pronóstico del modelo alcanzó el 88,8%, prediciendo correctamente 9 de las 10 ventas.

Comparativamente, se puede observar que en el periodo comprendido entre el 24 de

agosto y el 6 de septiembre, en donde aún no existía el modelo, se examinaron 1941 leads, con 165 señalados como posibles ventas. Finalmente, 5 de estos leads se convirtieron en ventas efectivas, alcanzando una tasa de conversión de 0,257%. Adicionalmente, la tasa de aciertos para este periodo fue del 44%, y la precisión del pronóstico se situó en el 75%, con el modelo prediciendo acertadamente 4 de las 5 ventas.

Con la implementación del modelo, se denota una mejora significativa en la capacidad de identificar qué leads tienen una mayor probabilidad de convertirse en compras efectivas, enfocándose así en aquellos leads con mayores oportunidades de cierre exitoso, implicando un aumento en las ventas. Por otro lado, cuando no se tenía el modelo existía una pérdida en las oportunidades de venta, ya que el equipo de ventas no disponía de la información necesaria para priorizar adecuadamente su atención y esfuerzos.

Para evaluar la efectividad y precisión del modelo predictivo se llevó a cabo un análisis retrospectivo en los meses previos a su implementación. Este estudio se centró en la tasa de aciertos y la precisión del pronóstico durante los meses de julio y agosto, en donde los resultados obtenidos son destacables (Tabla 6.a.2).

Periodo	Agosto	Julio
Tasa de Aciertos	81,01%	84,4%
Precisión del pronostico	86,14%	85,4%

Tabla 6.a.2: Efectividad y precisión obtenidas con el modelo predictivo para dos periodos. (Elaboración propia)

En conclusión, los resultados obtenidos del modelo predictivo reflejan su alta efectividad y precisión en la predicción de leads potenciales. Se ha observado un incremento en la tasa de conversión de leads, lo que sugiere un mayor éxito en ventas futuras. Con tasas de aciertos y precisión del pronóstico superior al 80%, estos indicadores resaltan la habilidad del modelo para guiar de manera estratégica las actividades de venta de la empresa. Esta eficiencia en la selección de clientes potenciales basada en datos sólidos promete una gestión de ventas más exitosa y orientada a resultados.

b. Evaluación económica

La evaluación económica del proyecto de análisis predictivo de leads es un paso crítico para comprender su viabilidad financiera. Esta evaluación considera tanto los ingresos esperados como los costos asociados con la implementación y mantenimiento del proyecto.

La inversión inicial necesaria para el lanzamiento del sistema incluye el costo de desarrollo, específicamente el sueldo del desarrollador, el cual asciende a aproximadamente $280.000 mensuales durante un periodo de 5 meses, totalizando una inversión de $1.400.000.

Los costos están organizados en dos categorías principales: entrenamiento y mantenimiento.

1. Costos de Entrenamiento: Se destina a la capacitación del equipo, la cual consiste en una sesión de una hora para cuatro miembros del personal, en donde el costo por hora es de $13.333 por persona aproximadamente, sumando un total de $53.333.
2. Costos de Mantenimiento: Una vez operativo, el sistema requerirá un mantenimiento semestral para garantizar su eficiencia. Este mantenimiento será realizado por un analista de datos, con un costo de $4.444 por hora, resultando en un gasto anual de 8.888, el cual se encargará de las labores de mantenimiento y actualizaciones continuas del sistema.

Los ingresos del proyecto están directamente relacionados con el incremento en las ventas debido a la mejora en la tasa de conversión de leads. Se proyecta que la tasa de conversión de leads a clientes aumente en 0,67%, pasando de un 0,97% a un 1,65%, lo que representaría un incremento de $188.737.115 en las ventas.

Por consiguiente, para la evaluación financiera se utiliza una tasa de descuento mensual del 10%. Bajo estas condiciones, se calcula un Valor Actual Neto (VAN) de $820.540.989 y una Tasa Interna de Retorno (TIR) de 12986,50%. Estos indicadores financieros sugieren un rendimiento económico significativo del proyecto, destacando su potencial para generar valor y retorno de inversión en el corto y largo plazo.

La Tabla 6.b.1 resume los flujos de caja mensuales del proyecto, incluyendo la inversión inicial, los costos operativos y los ingresos proyectados:

Evaluación Economica							
Mes	0	1	2	3	4	5	6
Inversion							
Desarrollo de implementacion	$1.400.000						
Total	-$1.400.000	$0	$0	$0	$0	$0	$0
Costos							
Entrenamiento	$53.333						
Mantenimiento							$8.889
Total	-$53.333	$0	$0	$0	$0	$0	-$8.889
Ingresos							
Aumento de ventas		$188.737.115	$188.737.115	$188.737.115	$188.737.115	$188.737.115	$188.737.115
Total		$188.737.115	$188.737.115	$188.737.115	$188.737.115	$188.737.115	$188.737.115
Total	-$1.453.333	$188.737.115	$188.737.115	$188.737.115	$188.737.115	$188.737.115	$188.728.226

TSD	10%
VAN	$820.540.989
TIR	12986,50%

Tabla 6.b.1: Evaluación económica. (Elaboración propia)

7. Conclusión

Este proyecto de pasantía en Andes Motor culminó exitosamente con la implementación de un sistema avanzado de predicción y priorización de leads para vehículos livianos. A través de un exhaustivo análisis de datos y la implementación de modelos predictivos avanzados, hemos logrado identificar patrones clave que incrementan la probabilidad de conversión de leads en ventas efectivas.

Los resultados obtenidos reflejan una mejora significativa en la eficiencia del proceso de ventas, evidenciada por una mayor precisión en la identificación de leads potenciales y un incremento en las tasas de conversión.

La puesta en marcha del sistema demostró su eficacia en varios frentes. Primero, a través de la selección y priorización estratégica de leads, lo cual permitió dirigir los esfuerzos de ventas más efectivamente. Segundo, la herramienta proveyó claridad y precisión en la identificación de leads valiosos, lo que se tradujo en un aumento de las ventas efectivas en comparación con los períodos anteriores al proyecto.

Además, el análisis de los datos históricos y la comparación de los resultados de diferentes períodos reafirmaron la capacidad del modelo para mejorar la tasa de conversión de leads. El modelo no solo cumplió con las expectativas cuantitativas, sino que también proporcionó perspectivas cualitativas sobre las preferencias y comportamientos de los clientes, permitiendo a la empresa ajustar sus estrategias de marketing y ventas de manera más efectiva.

En conclusión, el proyecto ha generado un valor significativo, entregando herramientas y conocimientos que facilitarán una gestión comercial más eficiente y orientada a resultados. La integración de tecnologías de aprendizaje automatizado (ML) en los procesos de negocio no solo mejora la conversión de leads, sino que también ha sentado las bases para futuras innovaciones y mejoras continuas en la empresa.

8. Referencias

[1] Aprende Machine Learning (2023). *Clasificación con datos desbalanceados*. https://www.aprendemachinelearning.com/clasificacion-con-datos-desbalanceados/

[2] Bateman, B., Ranjan Jha, A., Johnston, B., & Mathur, I. (2020). *The Supervised Learning Workshop*. 2nd ed. Birmingham: Packt Publishing Ltd. 490 p. isbn: 978-1-80020-904-6.

[3] Duncan, B. A., & Elkan, C. P. (2015, August). *Probabilistic modeling of a sales funnel to prioritize leads.* In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1751-1758).

[4] Asociación Nacional Automotriz de Chile (2023). *Estudio de Mercado* https://www.anac.cl/category/estudio-de-mercado/

[5] Hotz, N. (2023). *What is CRISP DM? Data Science Process Alliance*. https://www.datascience-pm.com/crisp-dm-2/

[6] Jadli, A., Hamim, M., Hain, M., & Hasbaoui, A. (2022). *TOWARD A SMART LEAD SCORING SYSTEM USING MACHINE LEARNING*. Indian Journal of Computer Science and Engineering, 13(2), 433-443.

[7] Nygård, R., & Mezei, J. (2020). *Automating lead scoring with machine learning: An experimental study.*

[8] Pereira, R. M. M. (2021*). Building a predictive lead scoring model for contact prioritization: the case of HUUB* (Doctoral dissertation).

[9] Qu, F., Wang, Y. T., Hou, W. H., Zhou, X. Y., Wang, X. K., Li, J. B., & Wang, J. Q. (2022). *Forecasting of automobile sales based on support vector regression optimized by the grey wolf optimizer algorithm*. Mathematics, 10(13), 2234.

[10] Tangirala, S. (2020). *Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm*. International Journal of Advanced Computer Science and Applications, 11(2), 612-619.

[11] Wu, M., Andreev, P., & Benyoucef, M. (2023). *The state of lead scoring models and their impact on sales performance.* Information Technology and Management, 1-30.