Predicción de Potenciales Clientes de Inmuebles para Aitué, Basado en Datos Históricos de sus Clientes.

Joaquín Antonio Cárdenas Liebethal
Departamento de Ingeniería Informática y Ciencias de la Computación
Universidad de Concepción

Profesor Patrocinante: Guillermo Cabrera
Comisión: Javier Vidal y Diego Seco

11 de Abril de 2019

Resumen

Para esta memoria de título se trabajó con la inmobiliaria Aitué. Se propuso como objetivo principal predecir la probabilidad de que un cotizante se convierta en cliente. Para lograr esto, se utilizaron datos de cotizaciones de 5 proyectos que actualmente están en venta. Se usó como marco de trabajo la metodología CRISP-DM. Aitué maneja una base de datos de personas que cotizan un inmueble y otra de quienes realizan negocios. A partir de estos datos se generó un nuevo dataset, que describe el comportamiento de las personas durante sus cotizaciones y su disposición a entregar datos personales. Además asigna dos etiquetas "negocio" y "compra", como los atributos a predecir, los cuales representan el inicio de un proceso de venta y una venta finalizada respectivamente. Para el estudio se seleccionó el proyecto "San Andrés Del Valle", el cual cuenta con la mayor cantidad de datos. Se abordaron 2 tareas, predecir un cliente para un cotizante nuevo y predecir un cliente en base su historial de cotizaciones. Para ambas tarea se utilizaron y compararon 5 técnicas de clasificación usadas en la literatura: *Logistic Regression(LR), Decision Tree (DT), Random Forest (RF), XGBoost (XGB) y Support Vector Machine (SVM)*. Para elegir el mejor modelo se utilizó como métrica comparativa "el área bajo la curva" (AUC) del gráfico *Receiver Operating Characteristics* (*ROC)* y el puntaje *f1-score,* privilegiando el primero. En la tarea de predecir un cliente cuando es un nuevo cotizante (Tarea 1) y variable objetivo "negocio", el mejor modelo fue XGB con un *AUC* de 0.827 y un *f1_score* de 0.652 y la variable objetivo "compra" el modelo RF fue el mejor, con un *AUC* de 0.926 y *f1-score* de 0.492. En la tarea de predecir un cliente cuando es un cliente histórico (Tarea 2) con variable objetivo "negocio", fue *LR* con un *AUC* de 0.833 y un *f1_score* de 0.598 y para la variable objetivo "compra" el modelo RF obtuvo el mejor puntaje con un *AUC* de 0.807 y *f1-score* de 0.304. Ningún modelo superó un *Lift* de 1.5. Además se destacan como los atributos más significativos para la tarea 1 son: si entrega la dirección, la fecha de nacimiento y el número de integrantes de la familia, mientras que en la tarea 2 fueron la cantidad de cotizaciones y si es "recontacto". Con el fin de utilizar los resultados de forma práctica, se realizó un prototipo de una aplicación web, el cual busca simular el actual sistema de ingreso de cotizantes y cotizaciones utilizado en la empresa, integrando los modelos aprendidos.

Índice general

1. Introducción

La magnitud de datos que se generan en las distintas industrias y administración pública ha incrementado durante los últimos años a nivel mundial. Se reporta que cada día el mundo produce cerca de 2.5 quintillones de bytes de datos, con un pronóstico de generación y consumo de 40 zettabytes para el 2020 [1].

Debido a este crecimiento de los datos, se necesita incorporar nuevos métodos para tratarlos y generar conocimiento. Esto, con tal de que los administradores de negocios e industrias puedan beneficiarse de los datos que ellos mismo producen y aquellos que vienen desde otras fuentes. Estos datos van desde los datos de transacciones diarias dentro de la empresa, hasta información recolectada desde plataformas sociales [2]. Desde la perspectiva del administrador empresarial, la importancia de la analítica avanzada de datos radica en su capacidad para proporcionar información de valor, sobre la cual basar las decisiones. El análisis avanzado de datos tiene mucho potencial para la inteligencia y toma de decisiones en el marketing, como también en las áreas de venta al por menor, banca y telecomunicaciones [3]. El uso de analitica avanzada en el área de marketing y ventas se usa para perfilar y segmentar clientes en función de las diferentes características socioeconómicas, esto aumenta los niveles de satisfacción y retención del cliente y permite comercializar productos en diferentes segmentos en función de las preferencias de los usuarios [4]. También se puede realizar análisis de opinión sobre los productos y servicios, pudiendo generar reportes de los clientes cuando quedan insatisfechos o cambian a diferentes productos [5]. El economista Erik Brynjolfsson demostró que la toma de decisiones basada en datos es sumamente beneficiosa, aumentando el retorno de la inversión, el valor de las acciones y uso eficiente de los recursos [27].

La región del BioBío es conocida por su gran nivel de producción especialmente por sus industrias de manufactura, agricultura, silvicultura y pesca, comercio al por mayor, entre otros. Esto significa que existe un gran potencial de distintas aplicaciones de inteligencia artificial y ciencias de datos en la zona. Particularmente el nicho de la construcción y venta de inmuebles será siempre un negocio lucrativo debido al continuo crecimiento poblacional. Según el portal Data Chile, alimentado con los datos del Instituto Nacional de Estadística de Chile (INE), el crecimiento estimado para la región del BioBío es de un 0.6% para el 2020, se estima que habrá alrededor de 320 mil personas entre las edades de 25 - 34, muchos de ellos con poder adquisitivo, pudiendo aumentar la demanda por viviendas [8].

Para esta memoria se trabajó con Aitué. Aitué es una empresa dedicada a la construcción y venta de inmuebles principalmente en la región del BioBío, se fundó en la década de los noventa y desde entonces ha crecido y se ha posicionado como una de las principales inmobiliarias y constructoras en la región.

Este trabajo busca potenciar el área de marketing y ventas de la empresa aplicando técnicas de ciencias de datos. Aitué maneja una base de datos histórica de personas que cotizan y otra de aquellas que entran en un proceso de negocio para comprar un inmueble, este trabajo busca construir modelos predictivos desde los datos de cotizaciones y negocio para predecir si un cotizante comprará un inmueble.

1.1. Aitué

Actualmente Aitué cuenta con 14 proyectos que se distribuyen entre Concepción, La Florida, Los Ángeles y Rancagua, dentro de los cuales 5 corresponden a casas, 8 a departamentos y 1 a oficinas. La venta de los inmuebles se realiza “en blanco”, “en verde” o de manera “inmediata”. La venta “en blanco” se da cuando se compra en un proyecto que cuenta con el permiso de edificación, pero aún no se han iniciado las obras, la venta “en verde” se da cuando se compra en un proyecto, el cual está siendo construido, sin disponibilidad de piloto. Por último, la venta “inmediata” se da cuando el proyecto ya está terminado y puede ser habitado inmediatamente luego de firmada la escritura pública.

Existen dos fases principales a la hora de vender un inmueble, el proceso de captación y el de la venta misma. La figura 1 muestra un diagrama del proceso de cotización y venta. La actividad de captación de clientes se realiza en base a la experiencia e intuición del nicho objetivo del proyecto por parte de los jefes de marketing, quienes reciben un presupuesto anual. Este presupuesto está fijado por el coste total del proyecto y la estimación de ingresos anual.

Actualmente no existe un mecanismo tecnológico sistemático que respalde estas ideas, como lo sería un manejo avanzado de consultas a bases de datos. La desventaja de este “método” de intuición es que si bien ha funcionado, se pueden estar gastando recursos en medios que no estén captando cotizantes que signifiquen una potencial ganancia en relación a lo invertido, y otros medios con menor llegada puede estar rentando positivamente.

La actividad de venta se puede dividir en 2 subprocesos, el de cotización y el de negocio. El proceso de cotización se puede realizar vía web, por medio de portales como *enlacebiobio*, *portalinmobiliario* o directamente en la web de Aitué llenando un formulario preliminar. Luego de realizar las cotizaciones vía internet, un ejecutivo comercial se contacta vía telefónica con el cotizante. El ejecutivo llena un nuevo formulario usando un sistema interno, agregando al cliente en la base de datos y un formulario de cotización. Una persona puede realizar múltiples cotizaciones durante el tiempo. Una vez que se decide por un inmueble, la cotización correspondiente procede a la etapa de negocio.

La etapa de negocio conlleva 4 subetapas: reserva, promesa, escritura y entrega, los cuales llevan un estado comercial asociado. Se entra a la subetapa reserva, cuando el cotizante ya decidió y visitó el piloto, el cliente avisa a su ejecutivo de ventas que desea que le reserven la propiedad, esto quiere decir que dentro de 5 días Aitué no mostrará u ofrecerá esta propiedad a otro cliente. Esto, sin ningún costo adicional. Dentro de estos 5 días el cliente deberá conseguir un crédito hipotecario, con el cual podrá firmar ante notario la promesa, entrando a la subetapa promesa. La promesa constituye un compromiso económico, donde el cliente debe abonar el 10% del total del inmueble. 45 días previos a la entrega de la vivienda, el cliente debe activar el crédito hipotecario, en caso contrario se anula la transacción. Con lo anterior en orden, se prosigue a firmar la escritura pública. Este, es un documento legal entre el cliente, el banco y Aitué, que es firmada y ratificada por un notario que se inscribe en el Registro de la Propiedad[1],

[1] Registro de Propiedades en Chile. Guía de Instituciones, Trámites y Formularios. Directorio de instituciones y servicios públicos para registrar propiedades y bienes inmuebles en Chile. Fuente(https://registronacional.com/chile/propiedad.html)

y es un reflejo de lo ya pactado en el contrato de compraventa. Luego de unos días el ejecutivo se comunicará con el cliente para coordinar la entrega de la vivienda. La tabla 2 muestra las combinaciones de las distintas etapas con los estados comerciales correspondientes. Los pares escritura-normal y entregado-normal, corresponden a ventas, o sea, compra por el cotizante.

Tabla 2. Descripción Estado y Estado Comercial Proceso Negocio.

Estado	Estado Comercial	Observación	Descripción
Reservado	Normal	Negocio Reservado	La persona está con una reserva hecha, tiene 5 dias para pasar a promesa
Reservado	Desistimiento	Se desiste la reserva de forma forzada	La persona desiste de promesar
Anulado	Normal	Han pasado los días de vigencia de la reserva	Pasaron los días de la reserva.
Promesado	Normal	Negocio Promesado	La persona se encuentra con la vivienda promesada. Pago el pie y firmó el contrato.
Promesado	Resciliacion	Se rescilia la promesa de forma forzada	La persona desiste de seguir con la compra, por distintos motivos.
Escriturado	Normal	Negocio Escriturado	La persona consigue el crédito y firman documentos pertinentes ante notario.
Entregado	Normal	Negocio Escriturado + Entregado Postventa	Se le hace entrega a la persona de las llaves de su vivienda.

Tabla 2. Combinaciones de subetapas con estados comerciales del proceso de negocio. Estado y Estado Comercial corresponden a atributos de los datos de negocios. Fuente(Christian Zapata, empleado Aitué, área de tecnologías de la información)

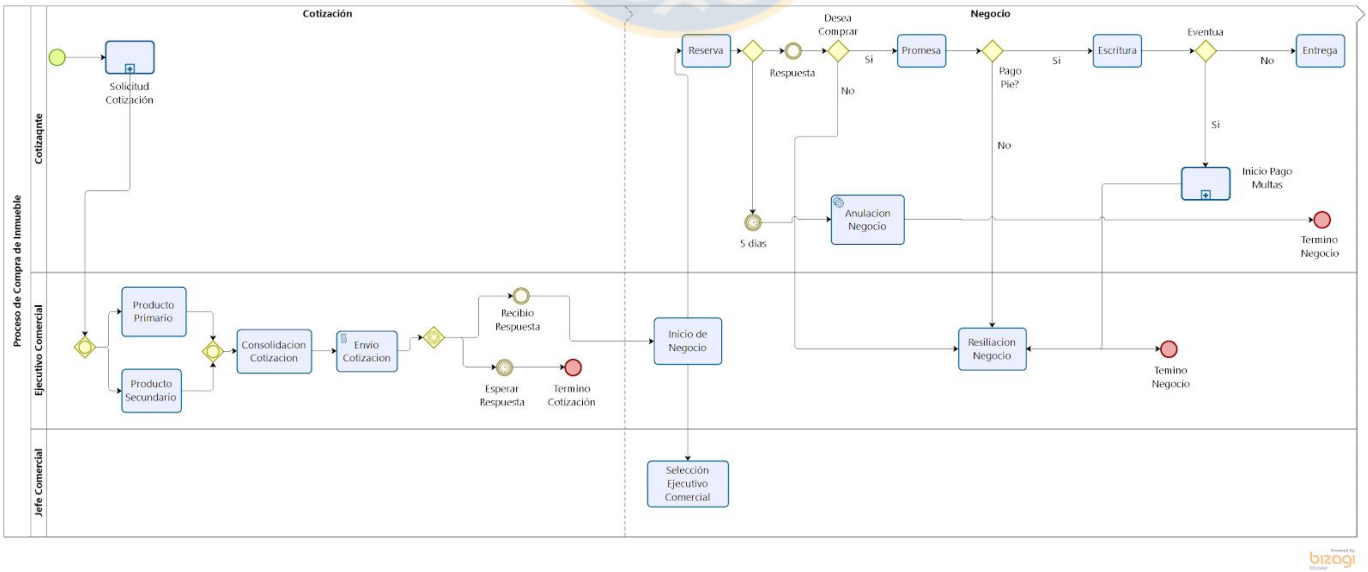

Figura 1. Muestra el flujo del proceso de venta de una vivienda en la inmobiliaria Aitué. Fuente(Elaboración propia)

1.2. Hipótesis

La creación de modelos de predicción binaria entrenados sobre un dataset de personas que han cotizado inmuebles en Aitué derivado de los datos de cotización y negocio, son capaces de predecir a un cliente cuando es un nuevo cotizante o predecir a un cliente cuando es un cotizante histórico obteniendo un valor de la métrica F1-score por sobre un 60% y un valor del área debajo de la curva ROC (“Receiver Operator Curve”) por sobre un 80%.

1.3. Objetivo General

Utilizar los datos históricos de negocios y cotizaciones de Aitué para generar un modelo de clasificación binaria de cotizantes a clientes que además estime la probabilidad de que un cotizante se convierta en cliente.

1.4. Objetivos específicos

- Explorar y describir los datos.
- Identificar personas que hayan entrado al proceso de negocio y personas que han comprado inmuebles.
- Entrenar y optimizar modelos de clasificación supervisada.
- Comparar los distintos modelos aplicados y seleccionar el mejor en base a las métricas de desempeño.
- Descubrir atributos claves que favorecen la predicción de clientes.
- Implementar un prototipo de software que permita hacer uso de los modelos, orientado al ejecutivo de ventas.

2. Discusión Bibliográfica

Sin duda el negocio de viviendas es uno de los más importantes que hay alrededor del mundo (en EEUU corresponde al 20%), donde intervienen bancos, entidades gubernamentales, el sector privado y personas naturales [15]. Las características que poseen las viviendas y el entorno que los rodea influyen directamente en su valor y hacen que su precio fluctúe, hacen que comprar un inmueble ya sea para vivir o como inversión para generar rentabilidad sea un decisión largoplacista y bastante meticulosa.

Una de las áreas de investigación más recurrentes en el área, es la estimación y predicción de precios de las viviendas. Algunos estudios se basan en las características inherentes de la vivienda, otros se basan en sus atributos temporales, otros en los geoespaciales y otros los mezclan [9]. Los métodos más utilizados en estas investigaciones han usado *Support Vector Machine* (*SVM*) , *Regression Trees* (*RT*) y *Linear Regression* (*LinReg*)[10, 13, 14, 16], pero son superados por estudios donde utilizan *XGB* [12], *Ensembles* [12, 14, 16] y *Procesos Gaussianos* (*GP*) [9].

Hay otros estudios [11] que desarrollan técnicas de visión computacional para proveer más features con los que mejorar la prediciones. Uno de los últimos estudios [15, 18] desarrolló una técnica de estimación de precio basada en *Convolutional Deep Neural Networks* (*CDNN*) utilizando imágenes de portales online de ventas como datos de entrada.

En cuanto a la utilización específica en el área de marketing en la industria inmobiliaria, de los documentos revisados solo [11, 15, 18] orientaron sus esfuerzos en esta área, asumiendo como supuesto que para las personas la fachada de un inmueble es un identificador importante al momento de valorar una vivienda. En la búsqueda no se encontró ningún trabajo orientado directamente a la segmentación de clientes y la posibilidad de identificar rasgos de los compradores aplicado a las inmobiliarias.

El uso de técnicas de aprendizaje de máquina en el mundo de los negocios es bastante extensa, y comprende distintas negocios que van desde la detección de fraude, predicción de stock hasta relación con el cliente. [20] identifica como la principal tarea entre todas ellas, la de aprendizaje supervisado. En el área de relación con los clientes (*CRM*), su principal tarea es entender los comportamientos de éstos con el fin de mejorar la relación de la organización con su medio y potenciales nuevos compradores. En el área de *CRM* existen distintos problemas a resolver y comprende 4 estrategias principales: identificación de cliente, atracción de clientes, retención de clientes y desarrollo de clientes. Las 2 primeras estrategias buscan captar nuevos clientes que generen un mayor beneficio y clientes que tengan una alta probabilidad de entrar al negocio e, identificar atributos que caracterizan a una subpoblación dentro de un segmento [22]. Las 2 últimas buscan sacarles el mayor provecho a los clientes existentes y evitar que se cambien de organización. Los principales algoritmos usados son: Logistic Regression, Decision Trees, Support Vector Machine, Multi Layer Perceptron, K-Nearest Neighbors, Random Forest, ADABoost, Naive Bayes, entre otros [20, 21, 22]. Además [22] muestra que los algoritmos de *ensemble*, logran mejores resultados que los otros. Un área en la cual han aplicado bastante aprendizaje automático orientado al CRM en los últimos años, es el sector de telecomunicaciones y retail con sistemas de recomendaciones. Los principales algoritmos

usados en las estrategias de identificación y atracción del cliente se pueden ver en la tabla 1. [20, 21]. Dentro de identificación de clientes, en la tarea de clasificación, se pueden sumar las técnicas mencionadas anteriormente.

Tabla 1. Tecnicas Aprendizaje de Máquinas Área Marketing

Estrategia	**Tarea**	**Tarea de Data Mining**	**Tecnica Aplicada**
Identificación de clientes	Segmentacion	Classification	*DT, Self-organizing-map(SOM), Markov chain model.*
		Clustering	K-means, Data-envelopment analysis, *SOM*, DT, Pattern based cluster.
	Target Customer Analysis	Regression Classification	*LR.*
Customer Attraction	Direct Marketing	Clustering Visualization	*SOM, Customer Map, DT, Genetic Algorithm, Neural Network.*
		Clustering	*Outlier Detection.*

Tabla 1. Estrategias de marketing orientadas al cliente, segmentadas por objetivo y objetivo de data mining. Fuente([21])

3. Conceptos

Este trabajo busca ejecutar estrategias de aprendizaje automático con el fin de potenciar el área de marketing de Aitué, potenciando la adquisición de clientes y su identificación.

3.1. Mineria de Datos

La minería de datos se describe como el proceso de descubrimiento de nueva información, patrones y tendencias en grandes volúmenes de datos de manera rápida y automática, mediante el uso de herramientas estadísticas, inteligencia artificial y bases de datos dando una ventaja competitiva a las empresas.
Si bien las técnicas aplicadas han existido desde hace años, el reciente crecimiento del poder de cómputo y disponibilidad de datos ha permitido aprovechar estas técnicas. El proceso de minería de datos es ampliamente usado en industrias como retail, bancos, manufactura, telecomunicaciones y aseguradoras. Este proceso comprende una serie de fases que se enmarcan principalmente en tres áreas: datos, descubrimiento y producción, en las cuales se integra las fuentes de datos, se descubren patrones y se decide hacer uso de la información descubierta. [19, 20]

3.2. Metodologia

Se utilizó la metodología “Cross-industry standard process for data mining”, también conocido como CRISP-DM para llevar a cabo el proyecto. La metodología CRISP-DM es un proceso iterativo y un modelo de trabajo dividido en fases, describe técnicas generales usadas por la industria para enfrentar proyectos de data mining dentro de empresas. La figura 2 muestra un diagrama conceptual de la metodología y cómo interactúan sus fases. CRISP-DM se divide en 6 fases que se describen brevemente a continuación. [28]

Las principales fases de la metodología usadas en este trabajo son:

- Entendimiento de los Datos: se procede a realizar actividades que permitan identificar las propiedades de los datos
- Preparación de los Datos: fase de construcción del dataset que servirá para alimentar los modelos.
- Modelamiento: Aplicación y calibración de los distintos modelos, con el fin de llevarlos al óptimo rendimiento.
- Evaluación. Evaluación de los modelos con tal de asegurar que el modelo cumple con los objetivos del negocio.
- Producción: Organización y presentación de los resultados de tal manera que el cliente pueda usarlos como reporte o como herramienta tecnológica.

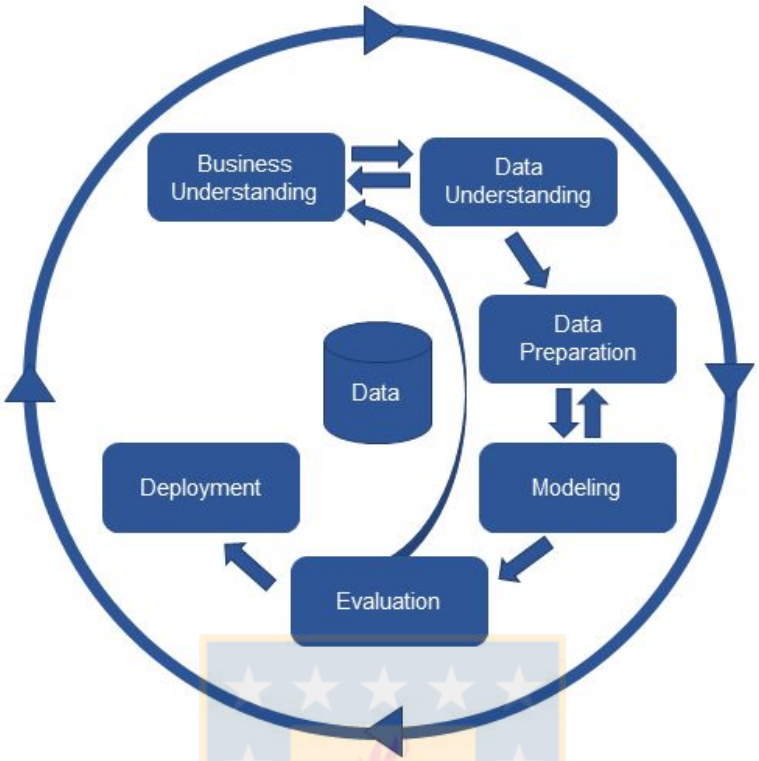

Figura 2. Imagen que describe la naturaleza iterativa dentro de las distintas fases de la metodología CRISP-DM. Fuente (https://www.kdnuggets.com/2017/01/four-problems-crisp-dm-fix.html)

3.3. Aprendizaje de Máquinas

El aprendizaje de máquinas es un subcampo de la inteligencia artificial que mezcla distintas disciplinas como la estadística, optimización matemática, computación y conocimiento específico del dominio del problema a resolver. El aprendizaje de máquinas mediante el uso de algoritmos, dota a las máquinas de la capacidad de aprender y predecir utilizando datos históricos.

Desde el punto de vista del negocio, el aprendizaje de máquinas es una extensión del análisis de datos descriptivo, y busca predecir eventos futuros utilizando modelos y algoritmos, también conocido como análisis predictivo [6].

Tom M. Mitchell [7] definió de manera formal este conjunto de algoritmos como: “Un programa de computador se dice que aprende de una experiencia E con respecto a una clase de tarea T y una métrica de rendimiento P si este realiza una tarea T medida por P que mejora con experiencia E”. Existen distintas tareas que realiza el aprendizaje de máquinas como lo son: aprendizaje supervisado, aprendizaje semi-supervisado, aprendizaje activo, aprendizaje no supervisado y aprendizaje por reforzamiento. Este trabajo se centra en la utilización de técnicas de aprendizaje supervisado para realizar clasificación binaria. A continuación se explicará el aprendizaje supervisado.

3.4. Aprendizaje Supervisado & Clasificación binaria

El aprendizaje supervisado es la tarea de aprender una función que mapea un set de datos D como entrada a una variable objetivo t como salida, también conocida como etiqueta. La variable objetivo t se encuentra contenida en el set de datos, es una variable que se conoce y se quiere predecir. El objetivo del aprendizaje supervisado es producir un modelo que pueda generalizar una predicción t para un input x que no está en D [39].

D={(xi, ti)}, donde i va de 1 a Xn muestras, xi representa un vector de valores de atributos xi = {(a1,..., am)} donde m es el número de atributos, mientras que ti representa la variable objetivo o variable a predecir.

La clasificación binaria específicamente determina una función de mapeo f: X -> T, dado un set de datos D etiquetados, que predice un output ti para un dato xi nunca antes visto. Donde el output ti puede pertenecer a 2 clases, T= {0,1}. En la clasificación multi-clase el output ti pertenece a K clases, T={0,1,2,...,K}.

Un ejemplo clásico de un aprendizaje supervisado para clasificación binaria, es la clasificación de imágenes de perros y gatos. Se tienen un modelo M de clasificación binaria que es entrenado con un conjunto de datos D que contienen filas con atributos que describen características de perros y gatos como: tamaño, peso, largo orejas, etc.. Cada fila tiene una etiqueta indicando si dichos atributos representan a un perro o a un gato. El modelo una vez entrenado puede asignar una etiqueta de perro o gato a una lista de atributos de un animal no observado.

3.5. Logistic Regression

Logistic Regression es un algoritmo estadístico de aprendizaje supervisado que busca modelar la probabilidad de que una observación pertenezca a una categoría (1 ó 0). La intuición de *LR* es que modela la respuesta f(x) de un modelo lineal *(1)* probabilísticamente, es decir lleva desde un espacio $[-\infty,+\infty]$ a uno de $[0,1]$ usando como mapeador la función sigmoide (5), usando las ecuaciones (2) y (3) e igualando con (1) se tiene la relación (3), luego aplicando la ecuación (5), se puede conseguir la probabilidad de una serie de variables independientes/atributos desde una combinación lineal. En la fase de entrenamiento, Logistic Regression por medio del algoritmo de *Maximum Likelihood Estimation* y los datos de entrenamiento, estima los coeficientes w, estos luego son usados cuando se quiere predecir para una nueva observación.

El plano (discriminador lineal) separa el espacio en 2 regiones, una que agrupa los datos de la clase 1 y otro que agrupa los datos de la clase 0 idealmente. Para una nueva instancia, mientras esté más lejos de la línea de separación, más alta será su probabilidad de pertenecer a la clase de la región en la que se encuentra, en el caso contrario, si se aleja de la línea de separación de manera inversa, tiene una probabilidad más alta de pertenecer a la clase contraria. Mientras más cerca de la linea mas baja la probabilidad de pertenencia a una clase, esto porque es más difícil diferenciar dicho punto.

Para predecir, LR toma una muestra descrita por un vector de atributos (x1,x2,...,xn), los coeficientes w aprendidos y calcula su probabilidad con la ecuación (5), conociendo $e^{f(x)}$ podemos usar la ecuación (3) para calcular su probabilidad.

$$f(x) = w_1 + w_1x_1 + w_2x_2 + ... + w_nx_n \; ; \; f(x) \; \varepsilon \; [-\infty, +\infty] \quad (1)$$

$$odds(p) = \left(\frac{p}{1-p}\right) \quad (2)$$

$$logit(p) = ln(odds(p)) = ln\left(\frac{p}{1-p}\right) \quad (3)$$

$$ln\left(\frac{p}{1-p}\right) = f(x) \quad (4)$$

$$p = \frac{e^{f(x)}}{e^{f(x)}+1} = \frac{1}{1-e^{f(x)}} \; ; \; p \; \varepsilon \; [0, 1] \quad (5)$$

3.6. Decision Tree

El algoritmo *Decision Tree* [30, 31] es un algoritmo supervisado de clasificación que separa sistemáticamente los datos en subconjuntos, basándose en los atributos del dataset de entrenamiento. El algoritmo, durante la fase de entrenamiento, divide un set (dataset) recursivamente de manera binaria, realizando una prueba sobre un atributo. Esta prueba busca separar el espacio en dos subespacios, donde, idealmente los subespacios resultantes sean lo más homogéneos posibles en relación a la variable objetivo, volviendo más fácil la determinación de la clase contenido en el subespacio. Este paso se va repitiendo para subsets cada vez más chicos, en caso de que existan subespacios no homogéneos, hasta llegar en lo posible a subsets que no se puedan separar más (máxima pureza) [38].

El algoritmo va creando ramificaciones desde el nodo raíz, el nodo raíz representa el conjunto completo de los datos, luego este set es dividido recursivamente por los atributos restantes, generando a su paso nodos de decisión que llevan hasta las hojas, las cuales representan las clases a predecir. La clase asignada al nodo, es la clase de la mayoría de los datos en el subespacio de la hoja.

Existen métricas que miden la heterogeneidad de un subespacio, conocidos como índices de impureza. Los más usados son *Gini Impurity* y *Entropy*. Un valor cercano a 0 de impureza para un subespacio indica pureza, mientras que un valor cercano a 1 indica impureza. *Gini Impurity* [21] dice que para que un subset sea puro, la probabilidad o frecuencia con la que podemos seleccionar del conjunto un valor es 1, es decir todo el set pertenece a la misma clase, resultando en un valor de 0 de *gini*.

$$Gini\ Impurity = 1 - \sum_{i=1}^{C} p_i^2$$

$$C : clases\ totales\ ,\ p_i : probabilidad\ de\ la\ clase\ i$$

Según [24] la Entropía (Entropy) se define como la cantidad de información contenida en una variable.

$$Entropy(c) = \sum_{c=1}^{C} p_i log_2(p_i)$$

$$c \ : \ Nodo \ ; \ p_i \ : \ probabilidad \ valor \ i \ para \ nodo \ c$$

Para crear un nodo y elegir el atributo que lo represente, el algoritmo calcula la cantidad de información que se gana (*Information Gain*) separando el set por un atributo, esto lo calcula para todos los atributos restantes. El atributo ganador es aquel que entrega mayor información en relación al nodo padre o, en otras palabras reduce la impureza de su padre. Además asigna un puntaje de importancia a cada atributo realizando la prueba anterior [23].

$$Information\ Gain\ = I(p)\ -\ I(c)$$

$$I: \ metrica \ de \ impureza; \ p \ : \ Nodo \ Padre \ ; \ c: \ Atributo \ por \ el \ cual \ se \ está \ dividiendo$$

Con el árbol y sus reglas creadas, al momento de predecir, el nuevo dato entra por el nodo raíz y decanta hacia una hoja tomando las decisiones de los nodos por los que pasa, basado en los valores de sus atributos.

3.7. Random Forest

Random Forest [31, 32] es un algoritmo de aprendizaje automático perteneciente al grupo de *ensembles* por su nombre en inglés, o "ensamblados". Los *ensembles* se construyen usando múltiples clasificadores, con el fin de obtener un mejor desempeño que un clasificador normal. El Random Forest es una combinación de Decisions Trees.

Específicamente *Random Forest* se construye usando *bagging/bootstrap aggregating* y clasificadores de Decisión Tree. El procedimiento de *bagging/bootstrap aggregating* toma muestras (con reemplazo[2]) desde un dataset de entrenamiento D, generando n nuevos subconjuntos de entrenamiento, luego se entrenan n clasificadores (DT), cada uno con un muestreo distinto. Adicionalmente, durante el entrenamiento de cada DT, cuando se crean los nodos decisión, se toma un subconjunto de los atributos disponibles. Con estas dos variaciones en el muestreo y en la prueba de los nodos, cada DT que se genera es distintos a los demás. Como output, en una tarea de clasificación supervisada, *Random Forest* escoge la etiqueta que fue más "votada" por los distintos *Decision Trees*. La figura 3 muestra un esquema de este procedimiento.

Una de las ventajas de usar métodos de *Bagging*, en el caso de *Random Forest* es que reduce el sobre-ajuste a los datos de entrenamiento que sufren los *Decision Tree* utilizados de manera independiente.

[2] Un valor muestreado puede repetirse en el conjunto de muestras total.

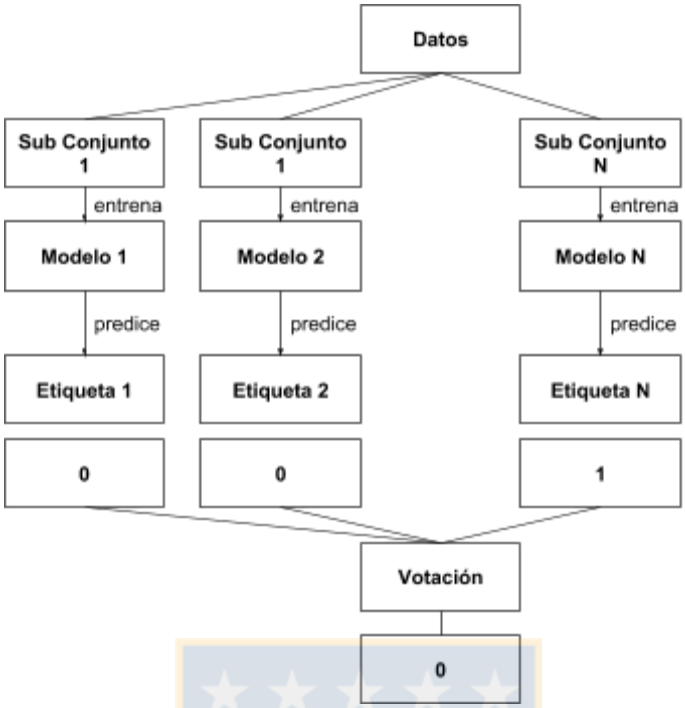

Figura 3. Procedimiento de *Bagging*. Se generan n subconjuntos con reemplazo, los cuales son utilizados para entrenar n modelos, cada modelo predice una etiqueta para una nueva observación, la etiqueta final es la etiqueta d e la mayoría de clasificadores Fuente(elaboración propia)

3.8. Métricas de Evaluación

El objetivo de un modelo de aprendizaje automático, es aprender, desde un conjunto de datos, patrones que permitan generalizar la predicción a datos nunca antes vistos. Para evaluar un modelo se divide el conjunto original en 3 partes, el set de entrenamiento, set de validación y set de pruebas. El set de entrenamiento es usado para "construir" el modelo (encontrar sus parámetros), el set de validación se usa para evaluar el modelo entrenado con el set de entrenamiento mientras se ajustan los parámetros del modelo en entrenamiento, y por último el set de prueba se usa para ver qué tan bien lo hizo el modelo, con los parámetros ajustados, sobre datos no visto antes. Existen distintas métricas para evaluar el rendimiento de un clasificador, una de ellas, la más utilizada es el "accuracy" o exactitud. La exactitud es la razón entre las clasificaciones correctas hechas por el clasificador y el total de clasificaciones hechas.

Para entender mejor el Accuracy primero presentamos la matriz de confusión (tabla 3) que permite visualizar el desempeño de un algoritmo de clasificación. La matriz de confusión expone las clasificaciones correctas y aquellas en las que se equivocó, se usa regularmente que las columnas representan las etiquetas reales y las filas las etiquetas predichas del dato. Para un problema de clasificación binaria se tiene una matriz de 4x4, donde se generan 4 pares, aquellas observaciones clasificadas correctamente como positivas se llaman *True Positives (TP)* o verdaderos positivos y las clases negativas correctamente predichas se llaman *True Negatives (TN)*, las clases positivas clasificadas incorrectamente se llaman *False Negatives (FN)* o falsos negativos y las negativas clasificadas incorrectamente se llaman *False Positives (FP)* o falsos positivos.

Tabla 3. Matriz de Confusión

	Etiqueta Real		
Etiqueta Predicha		**Positivo**	**Negativo**
	Positivo	True Positive (**TP**)	False Negative **(FN**)
	Negativo	False Positive **(FP**)	True Negative (**TN**)

Tabla 3. Matriz de confusión. Fuente(Elaboración propia)

La exactitud se usa para medir el rendimiento de manera muy general de un clasificador, pero no siempre refleja su desempeño real, especialmente en datos con etiquetas desbalanceadas.

Otras métricas que se usa para medir mejor el desempeño es el *recall* (*porcentaje de positivos predichos correctamente de todos los positivos reales*), *precision (*porcentaje de positivos predichos correctamente de todos los positivos predichos*), f1-score* (media armónica entre precision y recall), otras métricas de error que derivan de esta matriz (ver tabla 3) son: *false positive rate* (*FPR*, porcentaje de falsos positivos predichos) y *true positive rate* (*TPR* ó *Recall*, porcentaje de verdaderos positivos predichos).

La métrica *Receiver Operator Characteristics (ROC)* y su área bajo la curva (*AUC*, Total de etiquetas predichas positivas con mayor probabilidad estimada por sobre etiquetas predichas negativas [36]) son frecuentemente usadas para elegir un modelo sobre el otro. La figura 4, muestra un ROC como ejemplo, para 5 clasificadores distintos, cada punto es el balance entre el TPR (eje y) y el FPR (eje x) para distintos clasificador, esta curva suele ser usada para elegir un clasificador por sobre otro [25, 26].

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$TPR = Recall = \frac{TP}{TP + FN}$$

$$F1 - Score = \frac{2 * Precision * Recall}{(Precision + Recall)}$$

$$FPR = \frac{FP}{FP + TN}$$

$$AUC = \int_0^1 \frac{TP}{P} d\frac{FP}{N} = \frac{1}{PN} \int_0^N TP\, dFP$$

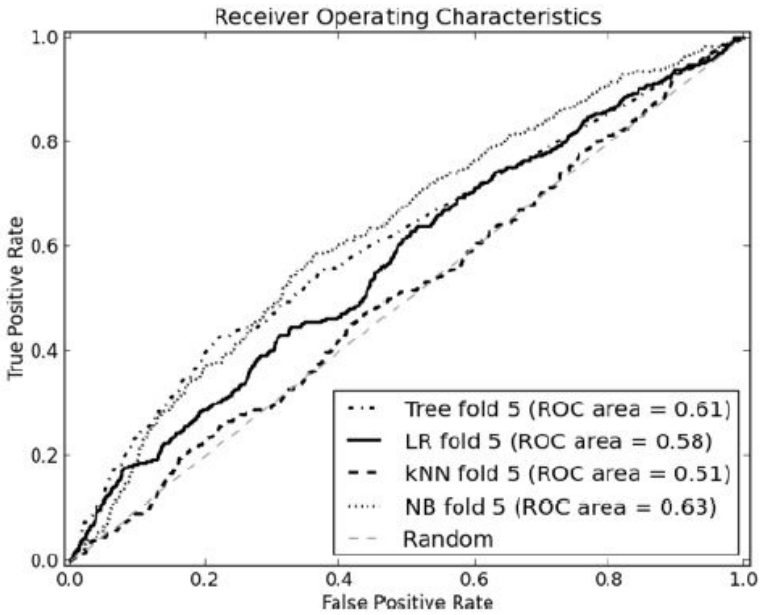

Figura 4. Ejemplo de una curva ROC, gráfica los TPR (eje x) y FPR (eje y) de 5 clasificadores diferentes con un Kfold cross-validation de 5. Además incluye una línea diagonal (x = y) que representa un muestreo de un clasificador aleatorio. Fuente(Data Science for Business, Chapter 8, Foster Provost & Tom Fawcett)

Otras métricas de evaluación utilizadas bastantes en las áreas de marketing [34, 35] son los gráficos *Cumulative Response Curve* (*CRC*) o Curva de ganancia acumulada y el *Lift Curve* (*LC*) o curva de elevación. Ambas sirven para elegir un modelo por sobre el otro. Se tiene una población *P*, se predice sus clases y su probabilidad estimada de pertenecer a la clase, luego se ordena *P* según su probabilidad estimada. La figura 5 muestra la *CRC*, que grafica la suma acumulada de los *TPR* (eje y) en función del un porcentaje *x* de *P*, en otras palabras nos entrega el total de *TP* que existen dentro de porcentaje x de P, entonces si se quiere obtener z *TP*s, se debe apuntar a un *x* % de *P*. El *Lift* indica cuánto mejor predice un clasificador en comparación con una selección aleatoria y se define según la ecuación (5) [36]. La figura 6 muestra una LC, está, grafica el lift de un clasificador (eje y) contra el porcentaje x de P (eje x). La *LC* es otra forma de visualización de la *CRC*, la *LC* compara la mejora de los clasificadores versus un clasificador aleatorio dada una población, mientras que la *CRC* muestra el % de *TP* entregados por un clasificador dada una población.

$$P = TP + FN;\ S = TP + FP + FN + TN\ ;$$

$$Sensitivity = \frac{TP}{P}\ ;$$

$$Yrate = \frac{TP + FP}{S}$$

$$Lift = \frac{Sensitivity}{Yrate}\ (5)$$

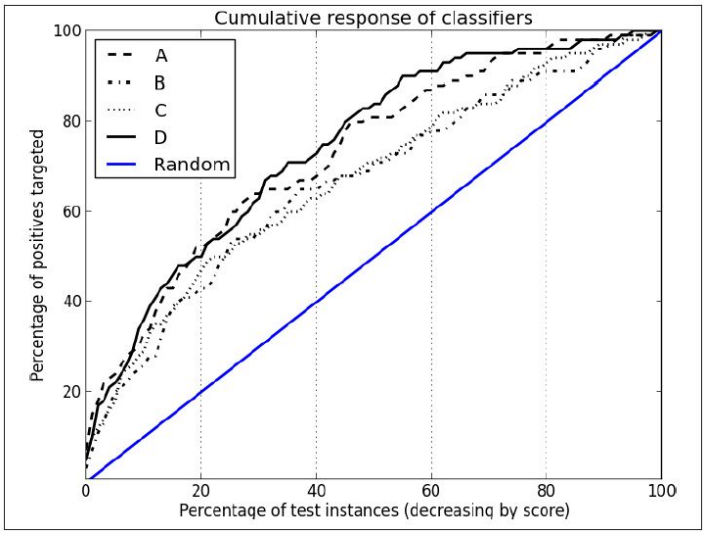

Figura 5. Ejemplo de una curva CRC, gráfica los TPR (eje y) en función del porcentaje de la población (eje x) de 5 clasificadores (hipotéticos) diferentes. Además incluye una línea diagonal que representa un elección aleatoria. Fuente(Data Science for Business, Chapter 8, Foster Provost & Tom Fawcett)

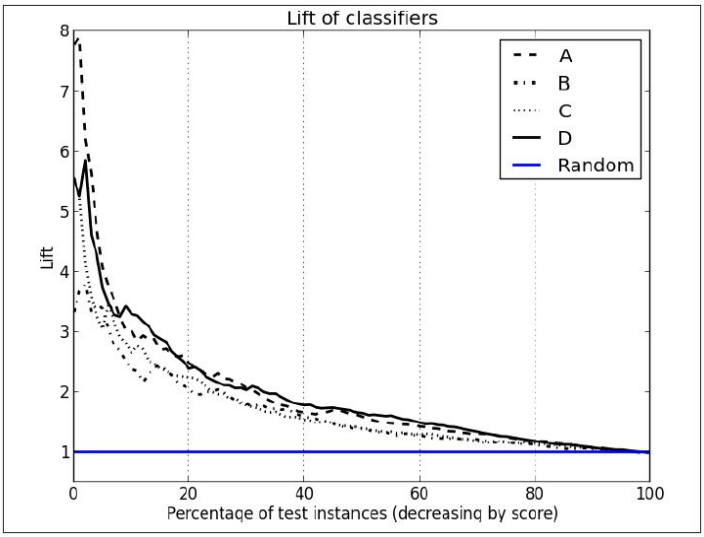

Figura 6. Ejemplo de una curva LC, gráfica Lift (eje x) y en función del porcentaje de la población (eje x) de 5 clasificadores (hipotéticos) diferentes. Además incluye una línea horizontal (y=1) que representa un lift como base rate. Fuente(Data Science for Business, Chapter 8, Foster Provost & Tom Fawcett)

3.9. Cross Validation

La validación cruzada, es el proceso de entrenar, validar y computar las métricas de evaluación sobre distintos segmentos de los conjuntos de entrenamiento y validación, mencionados anteriormente, evaluando el rendimiento general de un clasificador con los parámetros fijos. Buscando que el resultado no sea una casualidad del conjunto de entrenamiento y validación usados.

Una metodo de *Cross Validation* es la validación cruzada *KFold*. La figura 7 muestra como KFold divide el set de datos en k subconjuntos disjuntos de igual tamaño. Con los datos originales separados, se procede a entrenar el modelo con k-1 subconjuntos, utilizando el subconjunto restante para la etapa de validación. Este proceso se repite k veces, donde en cada paso se intercambia el conjunto de prueba con uno de entrenamiento, ver figura 8.

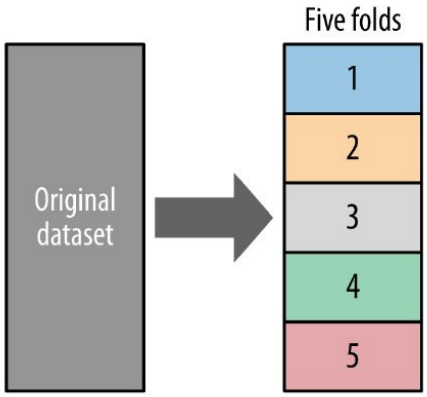

Figura 7. Muestra la separación en el dataset original en 5 partes. Fuente(Data Science for Business, Chapter 5, Foster Provost & Tom Fawcett)

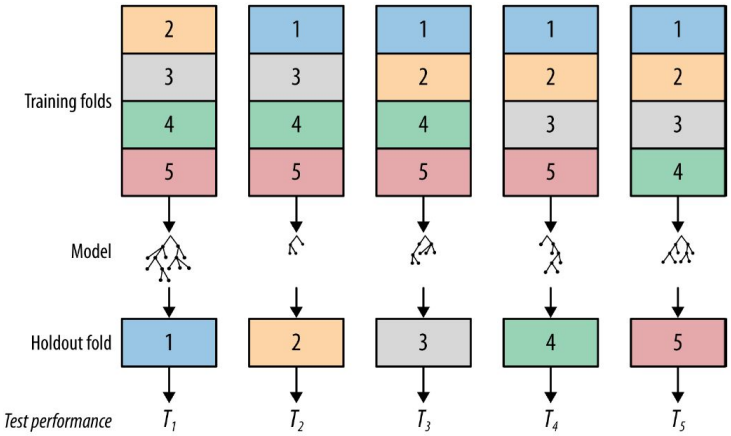

Figura 8. Proceso de validación cruzada KFold, se itera k veces, cambiando el conjunto de prueba (Holdout Fold) por un subconjunto de entrenamiento (Training Fold) en cada iteración. Fuente(Data Science for Business, Chapter 5, Foster Provost & Tom Fawcett)

4. Experimentos y Resultados

4.1. Entendimiento de los Datos

La fase de entendimiento de datos comienza con la recopilación de datos inicial y continúa con actividades que permiten familiarizarse con los datos, identificar la calidad de los datos, descubrir conocimientos en los datos y detectar subconjuntos de datos o atributos interesantes para el problema.

Recolección de Datos
Aitué cuenta con diversas fuentes de datos que son almacenados por una empresa que presta servicios informáticos, de nombre “E-Corebusiness”.

Datos cotizaciones. Colección de datos de cotizaciones recolectados por distintos ejecutivos de venta. Existen atributos como: fecha de nacimiento, actividad, profesión. Lo malo es que muchos de estos valores son nulos, lo que hace pensar que la facilitación de estos datos explica una mayor motivación por comprar un inmueble por parte de un cotizante.

Datos negocios. Colección de datos de cotizantes que comienzan con el proceso de comprar un inmueble. Estos datos describen un proceso y la completitud de los datos va variando según el estado del negocio. Los atributos de un cliente en etapa de “entrega” está más lleno que un cliente en la etapa de “reserva”. La colección añade atributos no presentes en los datos de cotización.

Datos proyectos. Colección de datos sobre los proyectos de Aitué. Estos datos son introducidos por los jefes de proyecto. Estos datos no fueron proporcionados por lo que no se tiene conocimiento de sus atributos y alcance.

Datos productos. Colección de datos sobre las características de los departamentos y casas que están en venta. Estos datos no fueron proporcionados por lo que no se tiene conocimiento de sus atributos y alcance. Con estos datos se podría realizar una recomendación cruzada entre cotizantes con características similares.

Hasta la fecha Aitué desconoce la capacidad de adquirir datos desde otras fuentes, pero tampoco está dentro de sus planes a mediano plazo adquirir bases de datos.

4.2. Descripción de los Datos

Aitué dispuso de 2 datasets, "cotizaciones" y "negocios" de 5 proyectos inmobiliarios. Las tablas 4 y 5 muestran un resumen que describe los datos proporcionados por Aitué divididos por proyectos. La tabla 4 muestra el resumen de los datos de "cotizaciones". Como ejemplo: El proyecto de "Altos del Valle", que ofrece departamentos, cuenta con 496 filas y 41 columnas, los registro van desde la fecha "2017-07-27" hasta "2018-05-20", casi un año de registros. Un total de 257 personas han cotizado en este proyecto y el promedio de cotizaciones por personas es de 1.35, y la persona que más cotizó, lo hizo 10 veces.

Tabla 4. Resumen Datos Cotizaciones

proyecto	AltosDelValle	Junge	Mil610	SanAndresDelValle	Urban
filas	496	830	1791	10558	2917
cols	41	41	41	41	41
fecha_min	2017-07-27 15:46:54	2016-01-19 08:53:16	2016-09-22 16:36:03	2013-08-19 16:58:26	2017-03-24 18:36:24
fecha_max	2018-05-20 17:58:24.067000	2018-05-21 17:33:47	2018-05-22 13:33:02	2018-05-22 13:42:53	2018-05-22 13:27:15
ruts_unicos	257	335	775	3816	1365
cot_ruts_mean	1.35603	2.40607	1.89111	2.78069	1.64508
cot_ruts_max	10	20	16	45	13
productos	deptos	deptos	deptos	casa	deptos
direccion	Calle Nueva 820	LOS CASTAÑOS 1533	Avenida Nahuelbuta 1610	Parques de Carriel 5250	Orompello 1470
sector	Las Monjas, Lomas de San Andrés	QUINTA JUNGE	Península de Andalué	esquina Tierras Coloradas	Centro
comuna	Concepción	Concepción	San Pedro de la Paz	Concepción	Concepción

Tabla 4. Muestra el resumen de los proyectos en cotizaciones. Fuente(Elaboración propia)

Tabla 5. Resumen Datos Negocio

proyecto	AltosDelValle	Junge	Mil610	SanAndresDelValle	Urban
filas	108	101	87	1848	203
cols	49	49	49	49	49
fecha_min	2017-07-27 15:46:58.750000	2016-01-25 13:21:02.013000	2016-09-26 16:12:56.703000	2013-08-19 18:35:56.620000	2017-03-27 17:03:29.563000
fecha_max	2018-05-07 16:29:32.157000	2018-04-18 18:25:01.797000	2018-05-17 11:55:49.837000	2018-05-21 17:12:31.210000	2018-05-20 18:10:29.167000
ruts_unicos	63	52	54	820	124
ruts_mean	1.71429	1.94231	1.61111	2.25366	1.6371
ruts_std	1.31282	1.48738	0.877747	1.80734	1.17795
ruts_max	7	7	5	17	8
#es_escriturado	0	27	29	201	0
#es_entregado	0	0	0	34	0
#vendidos	0	27	29	235	0
%es_anulados	0.694444	0.524752	0.517241	0.659091	0.492611
%es_promesas	0.240741	0.0792079	0.0804598	0.094697	0.349754
%es_reserva	0.0648148	0.128713	0.0689655	0.119048	0.157635
%es_entregado	0	0	0	0.0183983	0
%esc_desistido	0.0462963	0.128713	0.0689655	0.126623	0.147783
%esc_normal	0.944444	0.80198	0.873563	0.830087	0.837438

Tabla 5. Muestra el resumen de los datos de negocio, separándolos por proyecto. Fuente(Elaboración propia)

La tabla 5 muestra el resumen de los datos de "negocio" de los 5 proyectos. Se añaden campos como *#es_escriturado*, *#es_entregado*, *#vendidos*, *#anulados*, *#es_promesa*, *#es_reserva* que describen el número de personas que se encuentran en las distintas subetapas que tiene el proceso de compra de un inmueble. Las filas *ruts_mean*, *ruts_std* y *rut_max* indican el número de la media de cotizaciones por persona, la desviación estándar del número de cotizaciones y el número de cotizaciones máximo hecho por una persona.

Los datos de "cotizaciones" cuentan con 7 grandes grupos de atributos. Datos de la cotización (*Id*, *Fecha Cotización*), Medio(*Medio*, *Tipo Medio*, *Jefe Comercial*, *Ejecutivo Comercial*, *Presencial*, *Remoto*), Proyecto (*Etapa*), Productos(*Productos*, *Total Productos*, *Descuentos*, *Valor Final Venta*), datos personales básicos del cotizante(*Rut*, *Nombre*, *Apellido 1*, *Apellido 2*, *Nombre Completo*, *Teléfono*, *Celular*, *Correo Electrónico, Tipo Cliente, Razón Social*), datos demográficos del cotizante (*Dirección*, *Región*, *Comuna*, *Provincia*, *Sexo*, *Estado Civil*, *Rango Edad*, *Fecha de Nacimiento*, *N° Grupo Familiar*, *Actividad*, *Giro*) y datos acerca de su profesión (*Cargo*, *Situación Laboral*, *Antigüedad Laboral*, *Profesión*), mientras que los datos de "negocio" cuentan con estos grupos y suma los campos de *Fecha Promesa*, *Fecha Desistimiento*, *Fecha Escritura*, *Fecha Entrega*, *Estado*, *Estado Comercial*, *Motivo Desistimiento*, *Descripción Desistimiento* y *Resciliación por Modificación*. La tabla 6 muestra en resumen estos atributos.

Tabla 6. Atributos Datasets

Dataset	Grupo	Atributos
"cotizaciones"	Cotización	*Id*, *Fecha Cotización*
	Medio	Medio, Tipo Medio, Jefe Comercial, Ejecutivo Comercial, Presencial, Remoto
	Proyecto	*Etapa, Proyecto*
	Productos	*Productos*, *Total Productos*, *Descuentos*, *Valor Final Venta*
	Cliente Basico	*Rut*, *Nombre*, *Apellido 1*, *Apellido 2*, *Nombre Completo*, *Teléfono*, *Celular*, *Correo Electrónico, Tipo Cliente, Razón Social*
	Cliente Demografico	*Dirección*, *Región*, *Comuna*, *Provincia*, *Sexo*, *Estado Civil*, *Rango Edad*, *Fecha de Nacimiento*, *N° Grupo Familiar*, *Actividad*, *Giro*
	Cliente Profesión	*Cargo*, *Situación Laboral*, *Antigüedad Laboral*, *Profesión*
"negocios"	Se repiten los grupos de "cotizaciones"	Se repiten los atributos de "cotizaciones"
	Fechas Etapas	*Fecha Promesa*, *Fecha Desistimiento*, *Fecha Escritura*, *Fecha Entrega*
	Etapas	*Estado*, *Estado Comercial*
	Descripción Etapas	*Motivo Desistimiento*, *Descripción Desistimiento* y *Resciliación por Modificación*

Tabla 6. Atributos que componen los set de datos de "cotizaciones" y "negocios". Fuente (Elaboración propia)

Durante el proceso de cotización los ejecutivos de ventas por sistema están obligados a introducir los siguientes campos acerca del cotizante: *Rut*, *Nombre*, *Apellido Paterno*, *Apellido Materno*, *Email*, *Región*, *Provincia* y *Comuna*. Los campos *Tipo*, *Giro*, *Dirección*, *Número*, *Depto*, *Rango de edad* , *Sexo*, *Fecha de nacimiento*, *Estado Civil*, *N° Grupo Familiar*, *Actividad*, *Cargo*, *Situación Laboral*, *Empleador* y *Antigüedad Laboral* son campos opcionales, si bien algunos atributos como *Tipo* , *Actividad*, *Sexo*, *N°Grupo Familiar*, *Rango de Edad* muchas veces son llenados bajo la estimación visual del ejecutivo de ventas.

Los datos de "cotizaciones" tiene en total 16592 filas y 41 atributos, mientras que los datos de "negocios" tienen 2347 filas y 49 columnas. Existe un sobrelapamiento en los atributos de las personas en ambos dataset ya que ambos fueron generados desde la base de datos que manejan. Los datos de "cotizaciones" fueron generados combinando las tablas Cliente, Cotización y Ejecutivo, mientras que los datos de "negocios" fueron generados combinando Cliente, Cotización, Ejecutivo y Negocio.

Los atributos en su mayoría son del tipo categórico nominales, mientras que los demás son del tipo fecha y numérico. Una descripción más detallada de los atributos se puede ver en las tablas 7 y 8.

Tabla 7. Tipo Datos Cotizaciones

Tipo de Dato	Atributo
String	Jefe Comercial, Ejecutivo Comercial, Rut, Nombre, Apellido 1, Apellido 2, Nombre Completo, Telefono, Celular, Dirección, Correo Electrónico, Cargo, Situación Laboral, Empleador, Profesión
Categorico	Medio,Tipo de Medio, Proyecto, Etapa, Presencial, Remoto, Región, Comuna, Provincia, Sexo, Estado Civil, Tipo Cliente, Razón Social, Giro, Nacionalidad, Actividad, Situación Laboral.
Int	ID, Total Productos, Descuentos, Valor Final Venta, Antigüedad Laboral
Fecha	Fecha Cotización.

Tabla 7. Información acerca del tipo de dato de los atributos en “cotizaciones”. Fuente(Elaboración propia)

Tabla 8. Tipo Datos Negocios

Tipo de Dato	Atributo
String	*Jefe Comercial, Ejecutivo Comercial, Rut, Nombre, Apellido 1, Apellido 2, Nombre Completo, Telefono, Celular, Dirección, Correo Electrónico, Cargo, Situación Laboral, Empleador, Profesión*
Categorico	*Medio,Tipo de Medio, Proyecto, Etapa, Región, Comuna, Provincia, Sexo, Estado Civil, Tipo Cliente, Razón Social, Giro, Nacionalidad, Actividad, Situación Laboral.*
Int	*ID, ID Cotizacion, Total Productos, Descuentos, Valor Final Venta, Antigüedad Laboral*
Fecha	*Fecha Cotización, Fecha Promesa, Fecha Desistimiento, Fecha Escritura, Fecha Entrega*

Tabla 8. Información acerca del tipo de dato de los atributos en “negocios”. Fuente(Elaboración propia)

4.3. Exploración y Verificación de los Datos

Muchos de los valores en ambos dataset son ingresados por el ejecutivo de ventas via formulario a la base de datos, por lo que en general en los campos personales obligatorios existen ruts que no son válidos, nombres y apellidos que contienen puntos, comas, espacios o algún otro caracter inválido cuando el cotizante no desea entregar esta información. El rango de edad muchas veces es estimado por parte del ejecutivo de ventas o derechamente colocado con el valor “0”, mientras que los datos de la profesión y demográficos son en general más escasos, ya que no son obligatorios en el proceso de cotización y la gente es más reacia a entregarlos. Las figuras 11 y 12, en las cuales se puede ver la cantidad de valores nulos para los atributos en ambos set de datos. Se grafica la suma total de valores nulos que contiene cada atributo comparado con el total de filas en el conjunto de datos para “cotizaciones” y “negocios”.

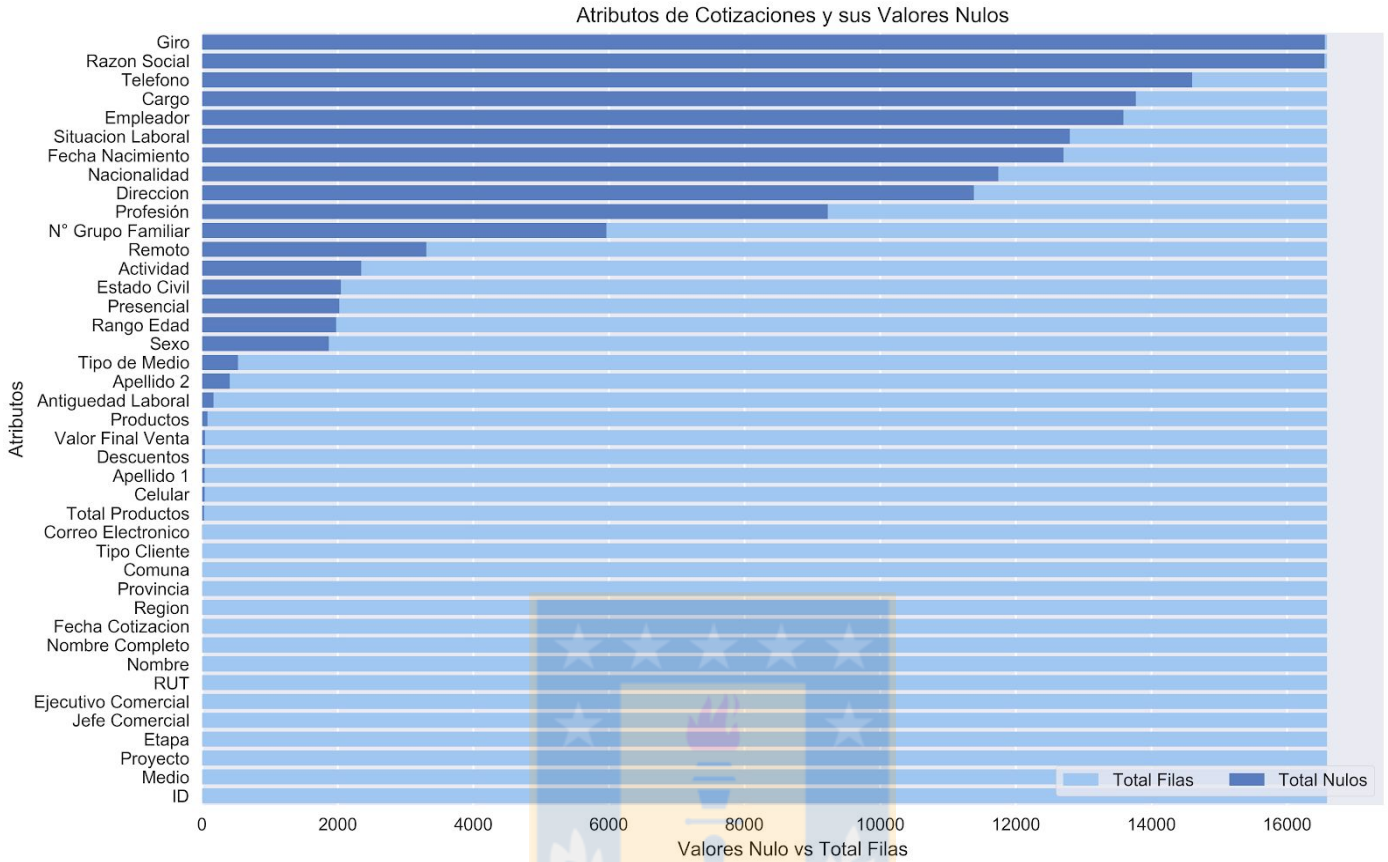

Figura 11. Cantidad de valores nulos por atributo (Azul oscuro) versus el total de filas disponibles en el dataset cotizaciones (Azul claro). Fuente(Elaboración propia)

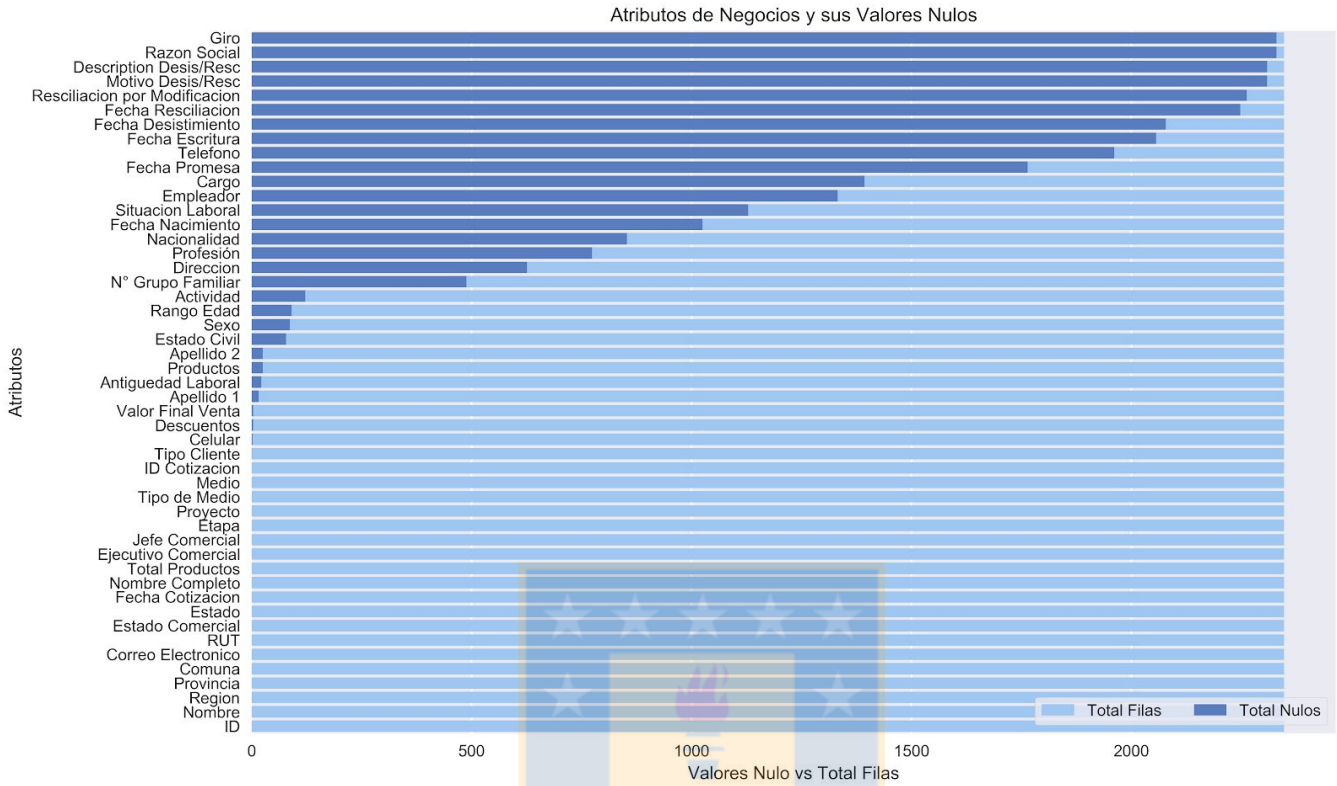

Figura 12. Cantidad de valores nulos por atributo versus el total de filas disponibles en el dataset negocios. Fuente(Elaboración propia)

Para saber qué cotizantes entraron al proceso de negocio o hicieron una compra, se cruzaron los datasets de “cotizaciónes” con los de “negocios” utilizando el atributo *ID Cotizacion* presente en “negocios”. Para identificar si un cotizante estuvo en negocio este debia de estar presente en “negocios”, mientras que para identificar si realizó una compra, debe de tener una fila en “negocios” con el atributo *Estado* con el valor “escriturado” o “entregado”. Existe un total de 6184 personas en el dataset de “cotizaciones, de los cuales 1102 entraron al proceso de negocio y de las personas que cotizaron que entraron a negocio, un total de 288 se encuentran con la escritura realizada o con la vivienda entregada, es decir con una compra. Existen 3 filas en el dataset de “negocio” cuyos “*ID Cotizacion”* no se encuentra en el dataset de “cotizaciones”. En la figura 13 se puede ver un gráfico que muestra la cantidad de valores entregados por atributo en el proceso de cotización por persona.

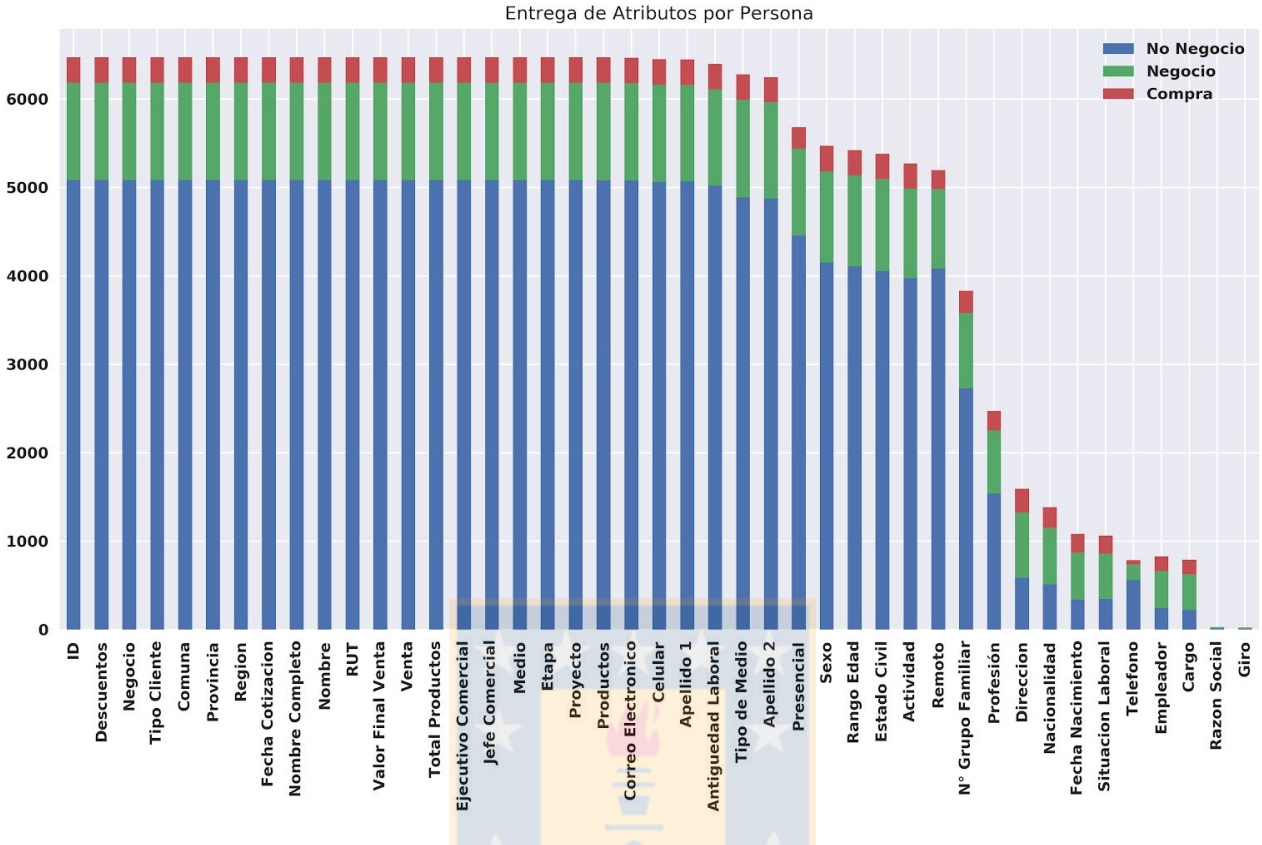

Figura 13. Cantidad de personas (Eje y) que entregaron el dato para el atributo (Eje x). Fuente(Elaboración propia)

Observando la figura 13 fácilmente se puede apreciar que los datos personales *Presencial, Sexo*, *Rango de Edad*, *Actividad* y *Remoto* comienzan a escasear ligeramente, mientras que desde los atributos *Nro Grupo Familiar* en adelante los datos son escasos y en gran proporción entregados por personas que entraron en el proceso de negocio para comprar un inmueble.

4.4. Preparación de los Datos

Para esta fase se decidió realizar un dataset nuevo y no modelar sobre el existente, esto por la condición de que una persona puede realizar múltiples cotizaciones y como el fin es predecir sobre la conducta de una persona, se usará, atributos agregados derivados del dataset de "cotizaciones" para generar un nuevo dataset de personas.

4.4.1. Selección de Datos

Selección de Filas. Se usaron todas las filas presentes en el dataset de "cotizaciones" para poder usarlos en la transformación a un dataset de personas. Este nuevo dataset refleja disposición de una persona a entregar datos personales y su historial de cotizaciones.

4.4.2. Limpieza de Datos y Construcción Dataset Personas.

En general los mayores problemas presentes en los valores de los datos son:

Acentos. Eliminación de acentos en las palabras.

Minúsculas. Se dejó todo en minúsculas.

Rut. Valores que no cumplen con la norma de formación de un rut, es decir no son válidos. Fueron buscados en el portal *rutificador.cl* y reemplazados a mano.

Sexo. Existen filas cuyos valores son "sin información", para aquellos ruts se utilizó el framework *scrapy*[3] para buscar automáticamente en el portal *rutificador.cl*[4] el sexo de aquellos ruts. Los que no fueron encontrados se llenaron a mano, tomando como base el nombre de la persona para identificar su sexo.

4.4.3. Construcción e Integración de nuevos Datos.

Es posible derivar atributos conductuales en base a las distintas cotizaciones de cada persona.

Atributos Preservados. Los siguientes atributos fueron preservados tal cual se encontraban en el set de datos de "cotizaciones":

- *Tipo Cliente* se renombró como: *tipo_cliente*
- *Region* se renombró como: *loc_region*
- *Provincia* se renombró como: *loc_provincia*
- *Comuna* se renombró como: *loc_comuna*
- *Sexo* se renombró como: *sexo*

Atributos Derivados. La tabla 9 muestra los atributos derivados desde los datos de "cotizaciones". Estos atributos se derivaron por persona ocupando las distintas cotizaciones que realizó a lo largo del tiempo.

[3] Scrapy es un software abierto, que se usa para la recolección de datos desde la web. (https://scrapy.org/)
[4] Pagina web en el cu

Tabla 9. Descripción Atributos Derivados.

Tipo de Variable	Nombre Columna	Descripción	Derivado del Atribnuto
Boolean	is_descuento	Si la persona ha recibido descuento en alguna cotización	Total Descuento
	is_recontacto	Si la persona ha realizado cotizaciones por el *Medio* "RECONTACTO". Derivado desde *Medio*	Medio
	valid_rut	Si la persona entregó un rut válido	RUT
	negocio	Si la persona ha participado en un proceso de negocio.	Cruzando datos Negocios y Cotizaciones
	compra	Si la persona realizó una compra de un inmueble.	Cruzando datos Negocios y Cotizaciones
Int	nro_proyectos	En cuantos proyectos a cotizado.	Proyecto
	Altos del Valle	Cuantas cotizaciones ha realizado en el proyecto Altos del Valle.	Proyecto
	Edificio Urban 1470	Cuantas cotizaciones ha realizado en el proyecto Edificio Urban 1470.	Proyecto
	Edificio Mil610	Cuantas cotizaciones ha realizado en el proyecto Edificio Mil610.	Proyecto
	Edificio Junge	Cuantas cotizaciones ha realizado en el proyecto Edificio Junge.	Proyecto
	San Andres del Valle	Cuantas cotizaciones ha realizado en el proyecto San Andres del Valle.	Proyecto
Float	mean_cot_bod	Promedio de cotizaciones por bodegas.	Productos
	mean_cot_vivienda	Promedio de cotizaciones por vivienda.	Productos
	mean_cot_estu	Promedio de cotizaciones por estudios.	Productos
	mean_cot_esta	Promedio de cotizaciones por estacionamientos.	Productos
	nro_cot_bod	Número de cotizaciones por bodegas.	Productos
	nro_cot_vivienda	Número de cotizaciones por vivienda. Derivado de *Productos*.	Productos
	nro_cot_estu	Número de cotizaciones por estudio.	Productos
	nro_cot_esta	Número de cotizaciones por estacionamientos.	Productos
	precio_cot_media	Promedio de los valores totales por los que cotizó la persona.	Valor Total
	precio_cot_median	Mediana de los valores totales por los que cotizó la persona.	Valor Total
	precio_cot_std	Desviación estándar de los valores totales por los que cotizó la persona.	Valor Total
	tiempo_cot_media	Promedio del tiempo que se demoró en realizar distintas cotizaciones. Medido en días.	Fecha Cotizacion
	tiempo_cot_median	Mediana del tiempo que se demoró en realizar distintas cotizaciones. Medido en días.	Fecha Cotizacion
	tiempo_cot_std	Desviación estándar del tiempo que se demoró en realizar distintas cotizaciones. Medido en diás.	Fecha Cotizacion

Tabla 9. Descripción de los atributos derivados desde los datos de "cotización". Fuente(Elaboración propia)

Nuevo dataset personas. El nuevo dataset de personas cuenta con 6155 filas y 31 atributos, los cuales fueron derivados en el paso anterior y algunos atributos que se mantuvieron desde "cotizaciones". La tabla 10 muestra el detalle.

Tabla 10. Resumen Dataset Personas

Tipo variable	Conjunto Atributos	Total Atributos
Boolean	is_apellido1, is_apellido2, is_celular, is_direccion, is_fnac, is_nombre, is_nombrecompleto, is_nrofam, is_presencial, is_profesion, is_recontacto, is_remoto, is_telefono, is_descuento, is_remoto, is_recontacto, valid_rut, negocio, compra.	20
Int	nro_cot_bod, nro_cot_vivienda, nro_cot_estu, nro_cot_esta, nro_proyectos, Altos del Valle, Edificio Urban 1470, Edificio Mil610, Edificio Junge, San Andres del Valle.	10
Float	Tiempo_cot_media, tiempo_cot_median, tiempo_cot_std, precio_cot_media, precio_cot_median, precio_cot_std, mean_cot_bod, mean_cot_vivienda, mean_cot_estu, mean_cot_esta.	10
Categorical	Loc_comuna, loc_region, loc_provincia, Sexo, tipo_cliente.	5

Tabla 10. Muestra los atributos que componen el dataset de personas y el tipo de variable. Fuente(Elaboración propia)

4.5. Modelamiento y Evaluación

Durante la fase de modelamiento y evaluación se busca resolver 2 tareas: la tarea de predecir un cliente cuando es un nuevo cotizante, desde ahora referenciada como la **tarea 1** y la tarea de predecir un cliente cuando es un cliente histórico, desde ahora referenciada como la **tarea 2**. Cada Tarea busca clasificar las variables objetivo (v.o.) “negocios” y “compras”, por lo que existen 2 modelos para la tarea 1 y tarea 2.

4.5.1. Datos Modelamiento

Datos para experimentación.

De las 3799 personas que cotizaron en San Andrés del Valle , un total de 235 efectivamente concretaron una compra. Dentro del mismo grupo se observa que otras 831 personas participaron en un negocio. En comparación a los otros proyectos fuera del estudio, existen dos de ellos que no cuentan con ventas concretadas, mientras que otros dos observan ventas de aproximadamente 30 unidades. La tabla 11 muestra los atributos que se usaron como descriptores y la variables objetivo a clasificar para la tarea 1, la tabla 12 es su símil para la tarea 2. El dataset personas cuenta con una proporción de 1:15 para la variable objetivo “compra” y 1:3 para la variable objetivo “negocio”, ver figura 14 y 15.

Tabla 11. Atributos Predictores Tarea 1

Uso	Atributos
Descriptores: Atributos para de entrenamiento	'actividad', 'is_apellido1', 'is_apellido2', 'is_celular', 'is_direccion', 'is_fnac', 'is_nombre', 'is_nombrecompleto', 'is_nrofam', 'is_presencial', 'is_profesion','is_recontacto', 'is_remoto', 'is_telefono', 'loc_comuna', 'loc_provincia', 'loc_region', 'medio_inicial', 'sexo'
Objetivo	Compra, negocio

Tabla 11. Descriptores y variables objetivo para los experimentos de cliente nuevo. Se realizaron 2 modelos, compra y negocio. Fuente(Elaboración propia)

Tabla 12. Atributos Predictores Tarea 2

Uso	Atributos
Descriptores: Atributos para de entrenamiento	is_recontacto, is_remoto,is_descuento,valid_rut, loc_comuna, loc_provincia, loc_region, sexo, tipo_cliente, mean_cot_bod, mean_cot_depto, mean_cot_esta, mean_cot_estu, medio_inicial,

	nro_cot_bod, nro_cot_depto, nro_cot_esta, nro_cot_estu, nro_proyectos, precio_cotizacion_media, precio_cotizacion_median, precio_cotizacion_std, tiempo_cotizacion_media, tiempo_cotizacion_median, tiempo_cotizacion_std, Altos del Valle, Edificio Urban 1470, Edificio Mil610, Edificio Junge
Objetivo	Compra, negocio

Tabla 12. Descriptores y variables objetivo para los experimentos de cliente histórico. Se realizaron 2 modelos, compra y negocio. Fuente(Elaboración propia)

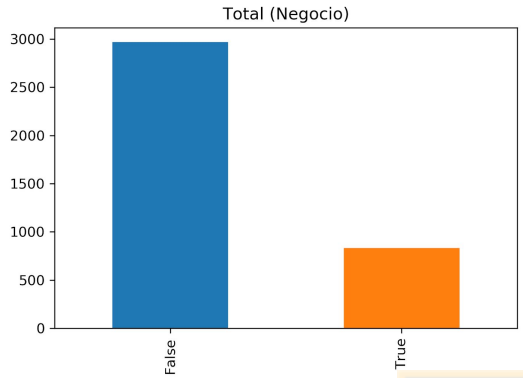

Figura 14. Proporción de etiquetas "negocio" para el dataset de personas que cotizaron en "San Andrés Del Valle", la etiqueta *True* indica si una persona hizo *Negocio*. Las relacion es de cerca del 3:1. Fuente(Elaboración propia)

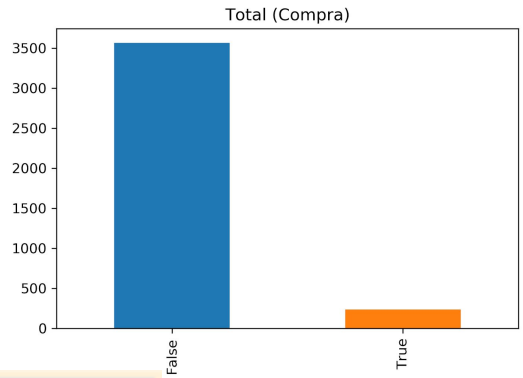

Figura 15. Proporción de etiquetas "compra" para el dataset de personas que cotizaron en "San Andrés Del Valle", la etiqueta *True* indica si una persona hizo *Compra*. Las relacion es de cerca del 15:1. Fuente(Elaboración propia)

Entrenamiento, validación y prueba

El dataset de personas que cotizaron en el proyecto de "San Andrés Del Valle" cuenta con 3799 filas y se dividió en 2 partes, un 80% (3039 filas) para las pruebas de validación cruzada (entrenamiento y validación) y un 20% (760) para el set de prueba. El set de prueba de "compra" contiene 45 datos de compras hechas y el set de prueba de "negocio" contiene 165 datos de negocios hechos. Para la evaluación del desempeño del modelo se utilizó la prueba de validación cruzada *KFold*. Se busca que los modelos produzcan altos *f1-scores* y *AUC scores*, privilegiando el primero por sobre el segundo. Esto último porque segun [25, 26] para que una curva ROC c1 sea indiscutiblemente superior a una curva c2, c1 debe producir un AUC score superior y no debe cruzarse en algún punto de la curva con c2.

4.5.2. Modelos

Selección de modelos. Cuando se resuelve un problema mediante el uso de modelos de machine learning, se recomienda entrenar varios modelos y luego comparar sus rendimientos, ya que no existe un modelo que sea el mejor para todos los problemas [42]. Según la literatura revisada, los modelos más usados para el problema de clasificación son: Logistic Regression, Decision Trees, Random Forest, Support Vector Machine y XGBoost por ser un clasificador altamente usado y con buenos resultados [41]. A continuación se configuran y evalúan.

Configuración de Parámetros. Cada modelo cuenta con hiper-parámetros únicos, los cuales deben ser especificados previo a entrenar un modelo. Se usó el método de *Gridsearch*[5], disponible en la librería de *scikit-learn* [33]. Este método busca dentro de un espacio de

[5] GridSearchCV genera candidatos de manera exhaustiva a partir de un conjunto de valores de parámetros especificados en un arreglo o diccionario. .(https://scikit-learn.org/stable/modules/grid_search.html)

parámetros, una combinación de estos, que produce el mejor rendimiento para cada clasificador. Los parámetros que recibe el método son: el clasificador, el método de la validación cruzada, la proporción de separación de los sets de entrenamiento y prueba, y un conjunto de parámetros que calibran el clasificador a optimizar. Durante la búsqueda para un modelo, se prueba cada combinación que exista en el set de parámetros del clasificador. Para cada combinación se realiza una validación cruzada. *GridSearch* escoge como mejor modelo junto con sus parámetros, aquel modelo que logró la mejor exactitud promedio en las distintas pruebas de validación cruzada.

4.5.3. Resultados Validación Cruzada

A continuación se muestran los resultados de la validación cruzada de los modelos utilizando la combinación de parámetros encontrados por *GridSearch* en las tareas 1 (predecir un cliente cuando es un nuevo cotizante) y 2 (predecir un cliente cuando es un cliente histórico).

Resultados Tarea 1 Negocio. Para la tarea 1 y la clasificación de la variable objetivo "negocio", los resultados de la validación cruzada muestran que el rendimiento más alto fue obtenido por *XGB* con una exactitud de 0.864*,* seguido de *SVM* y *RF*. Desde el punto de vista de las métricas de *f1-score* y *AUC,* el clasificador que obtuvo los puntajes más altos en ambas métricas, fue también *XGB* con un *f1-score* de 0.652 y un *AUC* de 0.827 seguido por *SVM*. Los detalles se pueden ver en la tabla 13, la figura 16 y figura 17.

Tabla. Resultados Validación Cruzada Tarea 1 y variable objetivo "Negocio"

Lugar	Modelo		Accuracy	Precision	Recall	F1-Score	AUC
1	XGB	Test	0.864 +- 0.016	0.750 +- 0.052	0.579 +- 0.047	0.652 +- 0.042	0.827 +- 0.026
2	SVM	Test	0.844 +- 0.019	0.645 +- 0.046	0.653 +- 0.044	0.648 +- 0.042	0.819 +- 0.029
3	RF	Test	0.842 +- 0.022	0.643 +- 0.058	0.643 +- 0.043	0.642 +- 0.046	0.825 +- 0.028
4	LR	Test	0.859 +- 0.013	0.749 +- 0.047	0.545 +- 0.040	0.630 +- 0.034	0.820 +- 0.027
5	DT	Test	0.831 +- 0.021	0.613 +- 0.043	0.628 +- 0.073	0.620 +- 0.055	0.803 +- 0.031

Tabla 13. Resultados de la validación cruzada con un Kfold de 6 para predecir la variable objetivo de "negocio". Fuente(Elaboración propia)

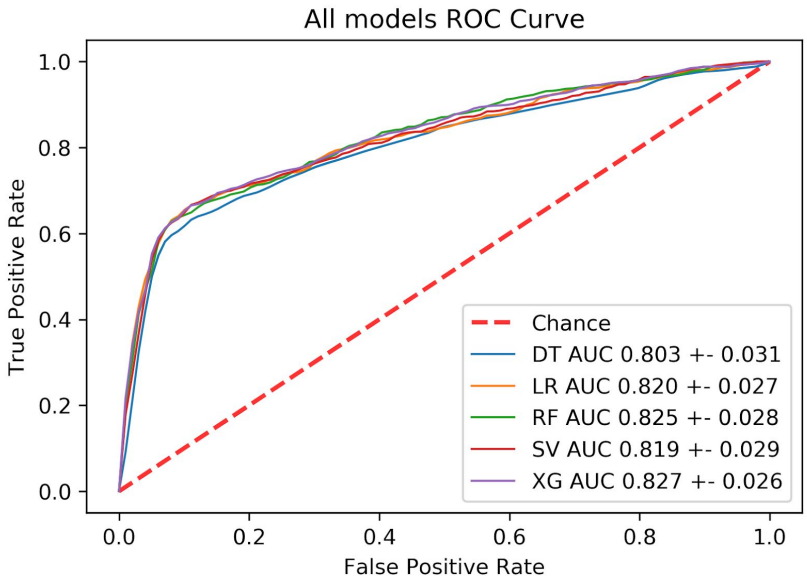

Figura 16. Curvas ROC ponderadas de cada clasificador durante la validación cruzada para la tarea 1 con la variable objetivo "negocio". Cada color representa un clasificador y su leyenda incluye su puntaje AUC de la validación cruzada. Ninguna curva es indiscutiblemente superior a otra. Fuente (Elaboración propia)

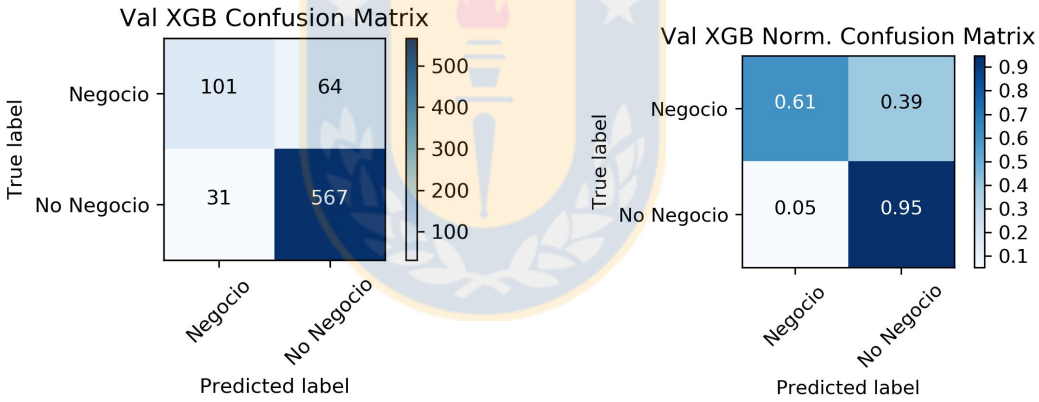

Figura 17. Matrices de confusión para la tarea 1, con la variable objetivo "negocio", del clasificador *XGB* sobre el set de prueba. La figura de la izquierda muestra las cantidades predichas, mientras que la de la derecha se encuentra normalizado. Fuente(elaboración propia)

Resultados Tarea 1 Compra. Para la tarea 1 y la clasificación de la variable objetivo "compra", los resultados muestran que el rendimiento más alto fue obtenido por *RF* con una exactitud de 0.891, seguido del *SVM* con un 0.821. Desde el punto de vista de las métricas de *f1-score* y *AUC*, el clasificador que obtuvo los puntajes más altos en ambos fue *RF* con un *f1-score* de 0.492 y un *AUC* de 0.926 seguido por *SVM* con un *f1-score* de 0.478 y un *auc* de 0.905 en el test de validación cruzada. Coronando a RF como el mejor clasificador para esta tarea. Los detalles se pueden ver en la tabla 14, la figura 18 y figura 19.

Tabla de Resultados Validación Cruzada Tarea 1 Compra.

Lugar	Modelo		Accuracy	Precision	Recall	F1-Score	AUC
1	RF	Test	0.891 +- 0.006	0.346 +- 0.014	0.853 +- 0.071	0.492 +- 0.022	0.926 +- 0.014
2	SVM	Test	0.897 +- 0.012	0.351 +- 0.035	0.753 +- 0.081	0.478 +- 0.043	0.905 +- 0.030
3	LR	Test	0.863 +- 0.009	0.304 +- 0.014	0.927 +- 0.046	0.457 +- 0.015	0.931 +- 0.019
4	DT	Test	0.864 +- 0.023	0.306 +- 0.030	0.891 +- 0.092	0.452 +- 0.022	0.881 +- 0.028
5	XGB	Test	0.929 +- 0.006	0.399 +- 0.079	0.258 +- 0.064	0.310 +- 0.059	0.925 +- 0.019

Tabla 14. Resultados de la validación cruzada con un Kfold de 6 para predecir la variable objetivo de "compra". Fuente(Elaboración propia)

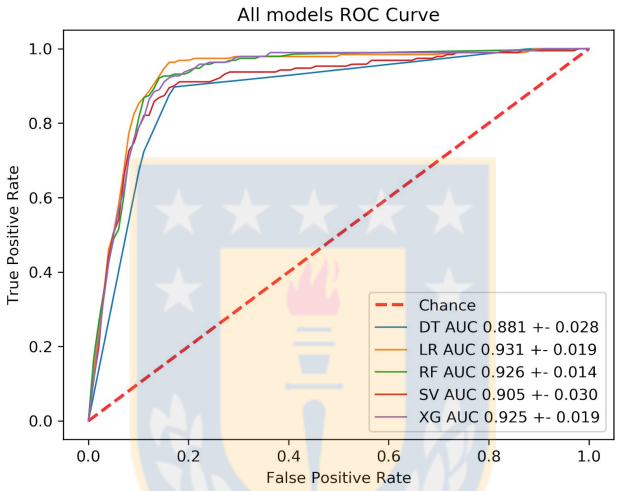

Figura 18. Curvas ROC ponderadas de cada clasificador durante la validación cruzada para la tarea 1 con la variable objetivo "compra". Cada color representa un clasificador y su leyenda incluye su puntaje AUC de la validación cruzada. Ninguna curva es indiscutiblemente superior a otra. Fuente (Elaboración propia)

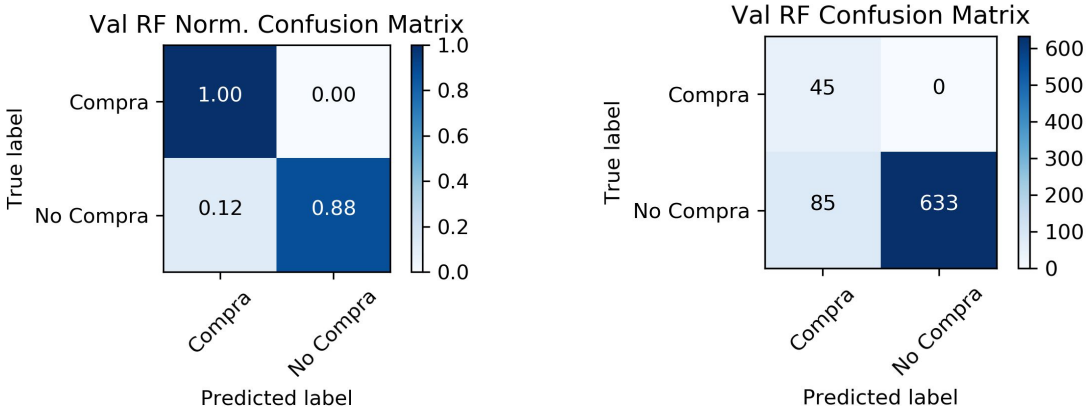

Figura 19. Matrices de confusión para la tarea 1 el set de prueba, con la variable objetivo "compra", del clasificador *RF*. La figura de la izquierda muestra las cantidades predichas, mientras que la de la derecha se encuentra normalizado. Fuente(elaboración propia)

Resultados Tarea 2 Negocio. Para la tarea 2 y la clasificación de la variable objetivo "negocio", los resultados de la validación cruzada muestran que el rendimiento más alto fue obtenido por *LR* con un accuracy de 0.796, seguido del *RF* con un 0.795 y *DT* con un 0.790. Desde el punto de vista de las métricas de *f1-score* y *auc,* el clasificador que obtuvo los puntajes más altos en ambos fue *LR* con un *f1-score* de 0.598 y un *auc* de 0.810 seguido por *RF* con un *f1-score* de 0.587 y un *auc* de 0.810. Ver tabla 15. El modelo marcó los atributos "sexo_femenino" y "is_nrofam" como aquellos atributos que mejor ayudan a clasificar entre un cliente que compra y otro que no. Ver figura 18, la figura 20 y figura 21. Coronando a LR como el mejor clasificador para esta tarea.

Tabla de Resultados Tarea 2 Negocio.

Lugar	Modelo		Accuracy	Precision	Recall	F1-Score	AUC
1	**LR**	**Test**	0.796 +- 0.016	0.528 +- 0.028	0.692 +- 0.052	0.598 +- 0.032	0.833 +- 0.018
2	**RF**	**Test**	0.795 +- 0.023	0.530 +- 0.043	0.661 +- 0.048	0.587 +- 0.032	0.810 +- 0.017
3	**DT**	**Test**	0.790 +- 0.010	0.519 +- 0.018	0.606 +- 0.035	0.559 +- 0.025	0.730 +- 0.016
4	**XGB**	**Test**	0.829 +- 0.010	0.689 +- 0.033	0.408 +- 0.045	0.511 +- 0.040	0.826 +- 0.015
5	**SVM**	**Test**	0.332 +- 0.048	0.240 +- 0.012	0.935 +- 0.020	0.382 +- 0.013	0.512 +- 0.083

Tabla 15. Resultados de la validación cruzada con un Kfold de 6 para predecir la variable objetivo de "negocio". Fuente(Elaboración propia)

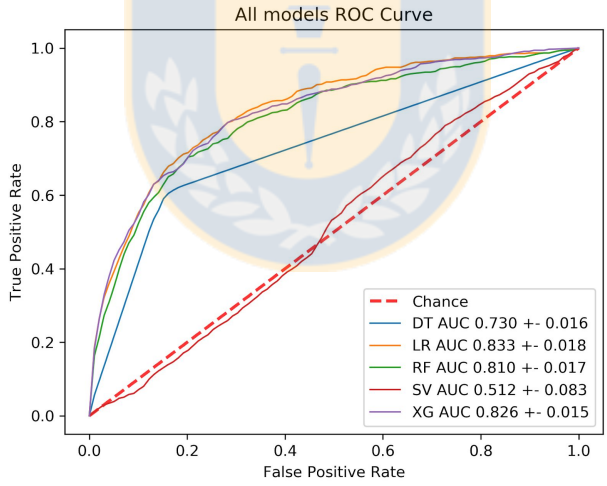

Figura 20. Curvas ROC ponderadas de cada clasificador durante la validación cruzada para la tarea 2 con la variable objetivo "negocio". Cada color representa un clasificador y su leyenda incluye su puntaje AUC de la validación cruzada. Ninguna curva es indiscutiblemente superior a otra. Fuente (Elaboración propia)

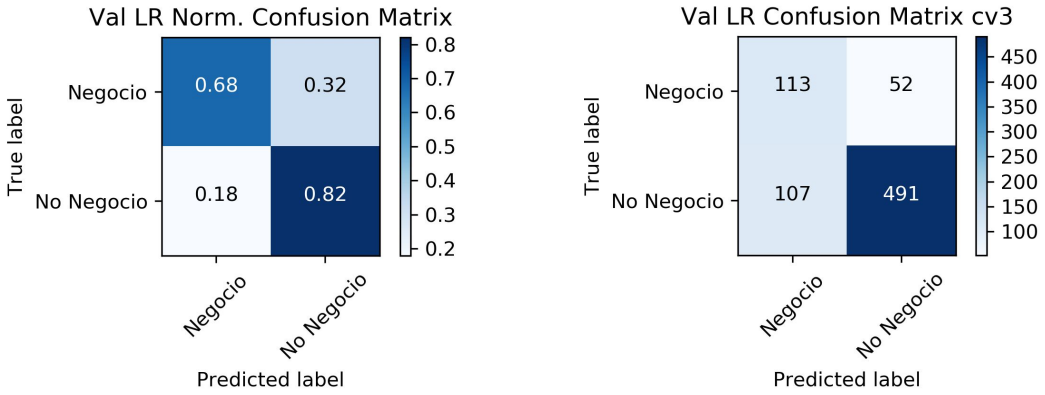

Figura 21. Matrices de confusión para la tarea 2 con la variable objetivo "negocio" del clasificador *RF* tarea 2 sobre el set de prueba. La figura de la izquierda muestra las cantidades predichas, mientras que la de la derecha se encuentra normalizado. Fuente(elaboración propia)

Resultados Tarea 2 Compra. Para la tarea 2 y la clasificación de la variable objetivo "compra", los resultados de la validación cruzada muestran que el rendimiento más alto fue obtenido por *XGB* con un 0.929 +- 0.006, seguido del *RF* con un 0.821. Desde el punto de vista de las métricas de *f1-score* y *auc,* el clasificador que obtuvo los puntajes más altos en ambos fue *RF* con un *f1-score* de 0.304 y un *auc* de 0.807 seguido por *LR* con un *f1-score* de 0.299 y un *auc* de 0.825 en el test de validación cruzada. Coronando a RF como el mejor clasificador para esta tarea. Los detalles se pueden ver en la tabla 16, la figura 2 y figura 23.

Tabla de Resultados Validación Cruzada Compra Cliente Histórico Tarea 2

Lugar	Modelo		Accuracy	Precision	Recall	F1-Score	AUC
1	**RF**	**Test**	0.821 +- 0.010	0.201 +- 0.018	0.631 +- 0.075	0.304 +- 0.027	0.807 +- 0.033
2	**LR**	**Test**	0.809 +- 0.008	0.194 +- 0.017	0.657 +- 0.095	0.299 +- 0.030	0.825 +- 0.052
3	**DT**	**Test**	0.773 +- 0.004	0.178 +- 0.013	0.737 +- 0.070	0.287 +- 0.022	0.765 +- 0.037
4	**XGB**	**Test**	0.929 +- 0.006	0.269 +- 0.177	0.089 +- 0.061	nan +- nan	0.820 +- 0.019
5	**SVM**	**Test**	0.326 +- 0.023	0.071 +- 0.004	0.816 +- 0.043	0.131 +- 0.007	0.501 +- 0.056

Tabla 16. Resultados de la validación cruzada con un Kfold de 6 para predecir la variable objetivo de "compra". Fuente(Elaboración propia)

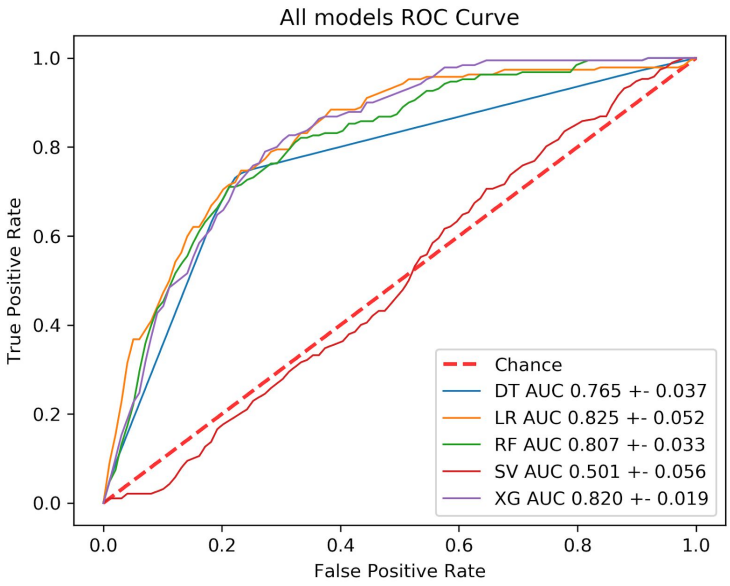

Figura 22. Curvas ROC ponderadas de cada clasificador durante la validación cruzada para la tarea 2 con la variable objetivo "compra". Cada color representa un clasificador y su leyenda incluye su puntaje AUC de la validación cruzada. Ninguna curva es indiscutiblemente superior a otra. Fuente (Elaboración propia)

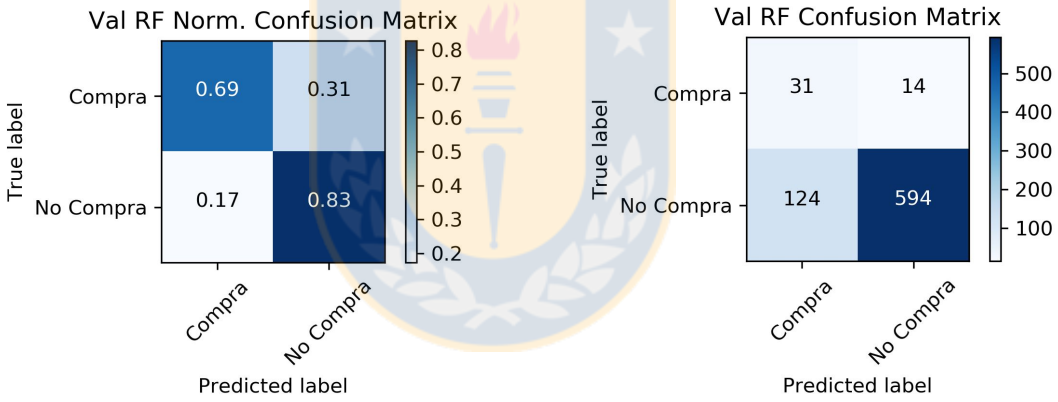

Figura 23. Matrices de confusión para la clasificación entre *Compra y No Compra* del clasificador *LR* Tarea 2. La figura de la izquierda muestra las cantidades predichas, mientras que la de la derecha se encuentra normalizado. Fuente(elaboración propia)

4.5.4. Evaluación según Lift

Para dar mayor soporte a la elección del mejor modelo, se realizó un análisis de Lift. Lo que se muestra en las figuras 24, 23, 24 y 25, son los Lifts, para las 2 tareas y sus variables objetivos ("negocio" y "compra"). Cada curva, es la curva Lift ponderada de la validación cruzada de un modelo. En los 4 casos, se puede apreciar que ningún modelo obtuvo un lift mayor que otro (Lift en % Población cercana a 0), obteniendo como máximo un 1.3 para las tareas con v.o. "negocio" y un 1.07 para las tareas con v.o. "compra".

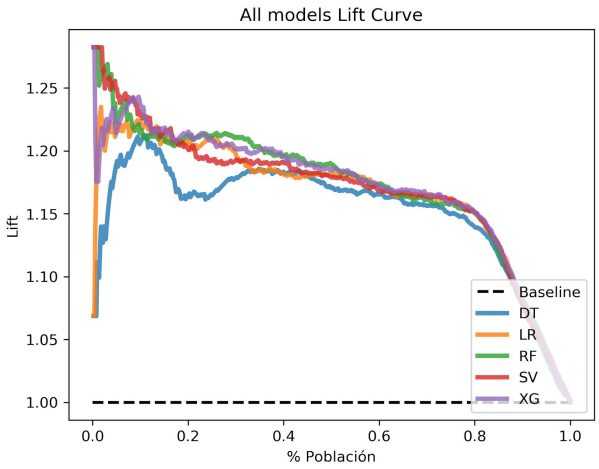

Figura 24. Lift Curve ponderado de la validación cruzada para los distintos modelos aplicados en la tarea 1 con la variable objetivo de “negocio”. Fuente(elaboración propia)

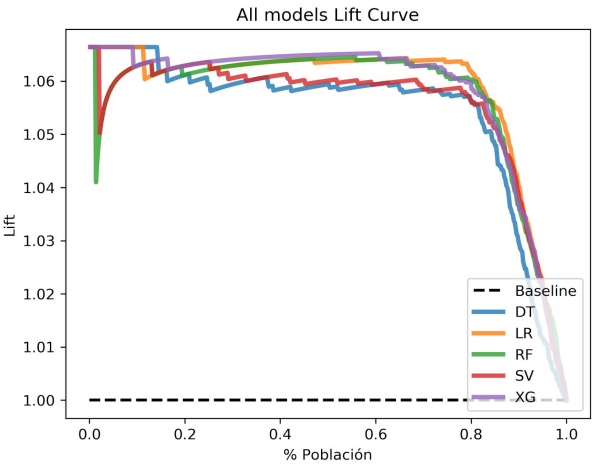

Figura 25.Lift Curve ponderado de la validación cruzada para los distintos modelos aplicados en la tarea 1 con la variable objetivo de “compra”. Fuente(elaboración propia)

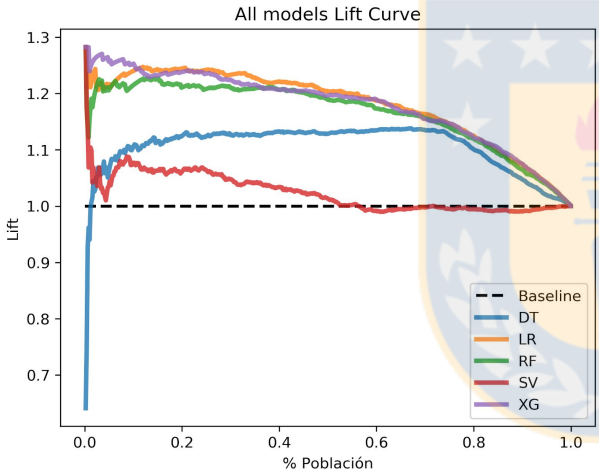

Figura 26. Lift Curve ponderado de la validación cruzada para los distintos modelos aplicados en la tarea 2 con la variable objetivo de “negocio”. Fuente(elaboración propia)

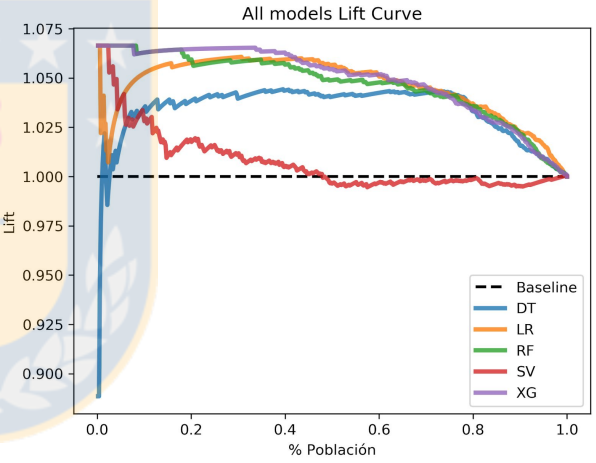

Figura 27. Lift Curve ponderado de la validación cruzada para los distintos modelos aplicados en la tarea 2 con la variable objetivo de “compra”. Fuente(elaboración propia)

En Aitué durante una jornada normal para los ejecutivos de ventas, existen tramos de tiempo de poco flujo de cotizantes, esto debido a que en general, los interesados en buscar propiedades trabajan en jornadas diurnas, por lo que se les recomienda a los ejecutivos captar clientes vía telefónica. Los ejecutivos llaman a los cotizantes quienes mostraron interés o indecisión, y les ofrecen una nueva cotización.

En general, el modelo que peor rendimiento tuvo, para las tareas y v.o distintas, fue DT y, SVM para la tarea 2 en general. Los 4 gráficos nos dicen que es una mala elección utilizar los modelos mencionados anteriormente, el TPR que se obtiene de ese tamaño de población no es mucho mejor que una elección aleatoria. Para las tareas 1 y 2 para la v.o. “negocio” presentan, en un inicio, un *lift* cercano a 1.3 y, dependiendo del % de la población que se quiera alcanzar, se puede ocupar entre RF, LR, SVM y XGB, para la tarea 1 y v.o “negocio”, mientras que en la tarea

2 y v.o. “negocio” las elecciones quedan solo en RD, LR y XGB, siendo LR, el que mantiene la superioridad, abarcando desde una población mayor al 18%. En general para la tarea 1 y 2, con v.o. “compra”, la mejora obtenida por los modelos no super a un 6%. Se mantiene el análisis, con respecto de la utilización de los modelos de LR, RF y XGB, para las tarea 1 y 2, con v.o “compra”.

La métrica Lift es altamente usada en marketing para resumir el desempeño de un clasificador, si bien el análisis anterior no entregó un lift por sobre un 1.5 para las tareas de clasificación de “compra”, esta métrica se ve afectada por el balanceo en las clases, como depende de la población, mientras más grande sea la diferencia en el tamaño de ambas clases, el lift irá disminuyendo, esto porque, aun clasificando correctamente aquellos cotizantes que son clientes, estos, están “sumergidos” en una gran población de la clase contraria, habiendo menos clases positivas por segmentos de la muestra total (porcentajes).

4.5.5. Atributos más Importante

Los modelos como RandomForest, Decision Trees y XGBoost tienen la facultad de entregar de manera bastante sencilla aquellos atributos que mejor “ayudan” a clasificar la variable objetivo, entregando una métrica de importancia. RF y DT asignan dicha importancia según la información que se gana separando por dicho atributo el set inicial, mientras que XGB asigna la importancia según qué tan utilizado fue ese atributo para construir reglas durante el aprendizaje del algoritmo. A continuación se señalan dichos atributos por tarea y se hace un breve análisis univariado de los primeros 3 atributos más importantes por modelo.

Resultados Tarea 1 Negocio. El clasificador XGB indicó los atributos “is_nrofam”, “is_telefono” y “is_fnac” como aquellos atributos que mejor ayudan a clasificar a un nuevo cotizante que entra al proceso de negocio y uno que no. La tabla 17 muestra que el atributo que entrega mayor discriminación según XGB es si una persona entrega su numero de grupo familiar, donde un 36% de aquellas personas que entregaron el atributo “nro_fam” entraron a un proceso de negocio, mientras que solo el 0.13% de las personas que no entregaron el dato lo hicieron. El segundo atributo más importante si una persona entrega su número de teléfono, donde un 30% de aquellas personas que entregaron lo entregaron entraron a un proceso de negocio, mientras que el 0.21% de las personas que no entregaron el dato lo hicieron. Por último, de las persona que entregaron su fecha de nacimiento (“is_fnac”) un 69% entraron a un proceso de negocio. La figura 28 grafica univariadamente estos atributos.

Tabla 17. Feature Importance Tarea 1 Negocio.

Lugar	Atributo	Importancia
1	is_nrofam	0.074
2	is_telefono	0.069
3	is_fnac	0.067

Tabla 17. Fuente(Elaboración propia)

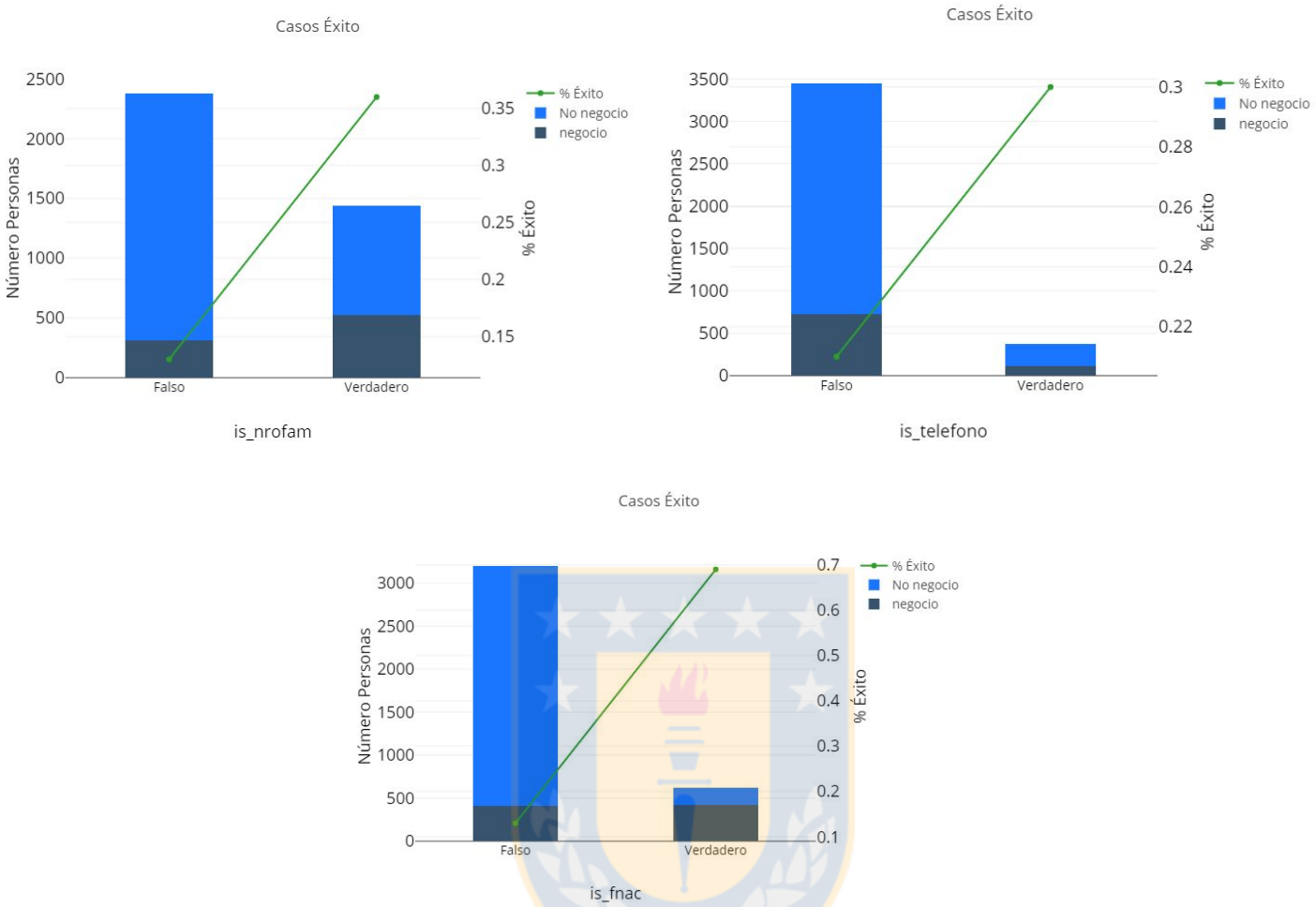

Figura 28. Gráficos que describen de manera univariada el top 3 de los feature más importantes identificados por el modelo *XGB* para tarea 2 y variable objetivo "negocio". Los elementos de barra (azul claro y azul oscuro) sumados son el total Fuente(Elaboración propia)

Resultados Tarea 1 Compra. El modelo RF indico los atributos "is_fnac", "is_nrofam" y "medo_inicial_finco" como aquellos atributos que mejor ayudan a clasificar a un nuevo cotizante que compra y uno que no. Ver figura 18. La tabla X muestra que el atributo que entrega mayor discriminación según RF es si una persona entrega su dirección, donde un 25% de aquellas personas que lo entregaron realizaron una compra, mientras que ninguna persona las personas que no lo entregó terminó comprando. El segundo atributo más importante si una persona entrega su fecha de nacimiento, donde un 31% de aquellas personas que lo entregaron concretaron una compra, mientras que el 1% de las personas que no entregaron el dato no compraron. Por último, de las persona que su grupo familiar ("is_fnac") un 12% entraron a un proceso de negocio. Figura 29 grafica univariadamente estos atributos.

Tabla 18. Feature Importance Tarea 1 Negocio.

Lugar	Atributo	Importancia
1	is_direccion	0.41
2	is_fnac	0.28
3	is_nrofam	0.05

Tabla 18. Fuente(Elaboración propia)

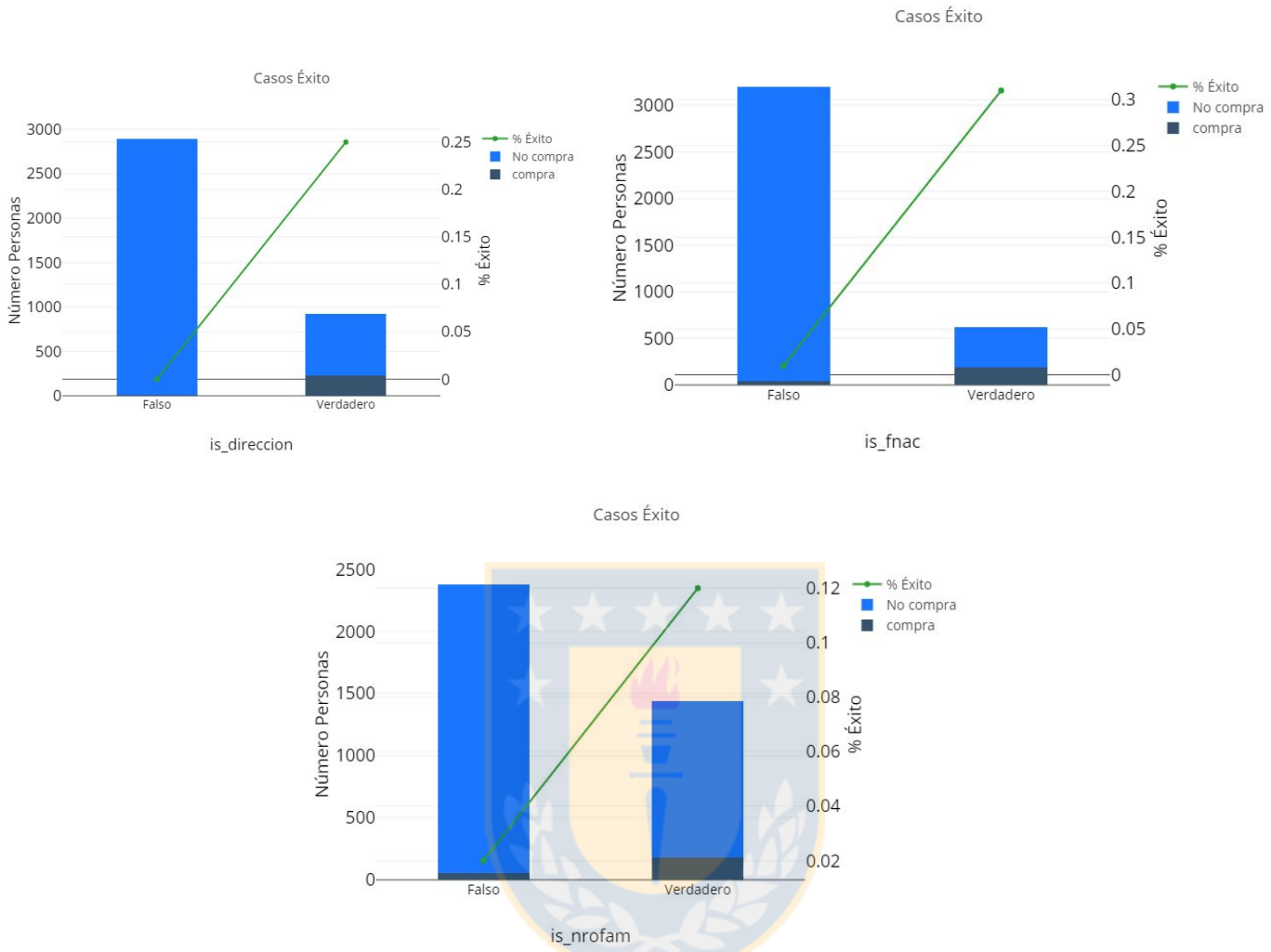

Figura 29. Gráficos que describen de manera univariada el top 3 de los feature más importantes identificados por el modelo *RF* para la tarea 1 y variable objetivo "compra". Los elementos de barra (azul claro y azul oscuro) sumados son el total Fuente(Elaboración propia)

Resultado Tarea 2 Negocio.

Para describir los atributos más importantes a la hora de clasificar se utilizó el RF por tener un desempeño parecido al del LR, los 3 atributos que mejor ayudan a clasificar son: *"nro_cot_depto", "is_recontacto"*, *"is_descuento"*. Ver figura 19. El atributo que entrega mayor discriminación según el RF es el número de cotizaciones por viviendas (nro_cot_depto), se agruparon las cotizaciones en 5 grupos, donde cada grupo agrupa cada 5 cotizaciones, es decir, el primero grupo considera desde las 0 cotizaciones hastas las 4 cotizaciones (existe gente que en sus cotizaciones incluye solo bodegas o estacionamientos), el segundo de 5 hasta 9 y así sucesivamente hasta las 45 cotizaciones. El primer grupo "0 - 5" que concentra el mayor número de personas (≈3000 personas) en general tiene un bajo porcentaje de conversión, de cerca del 13%, a medida que las cotizaciones aumentan también lo hace el porcentaje de conversión, pero disminuye drásticamente el número de cotizantes, con un 57% para "5 - 10" (≈ 500 personas), un 86% para "10 - 15" (104 personas), un 79% para "15 - 20" (24 personas) y un 8% para "20 - 45" (10 personas). En el caso del atributo *"is_recontacto"* tuvo 13% de conversión

para quienes no fueron re-contactados (≈2700 personas) en contraste con un 47% para quienes sí fueron contactados (≈1000 personas). Un comportamiento similar tuvo el atributo *"is_descuento"* con un 18% de conversión para quienes no recibieron descuento (≈2600 personas) y un 33% para quienes sí recibieron. La figura 30 grafica univariadamente estos atributos.

Tabla 19. Feature Importance Tarea 1 Negocio.

Lugar	Atributo	Importancia
1	nro_cot_depto	0.39
2	is_recontacto	0.25
3	is_descuento	0.03

Tabla 19. Fuente(Elaboración propia)

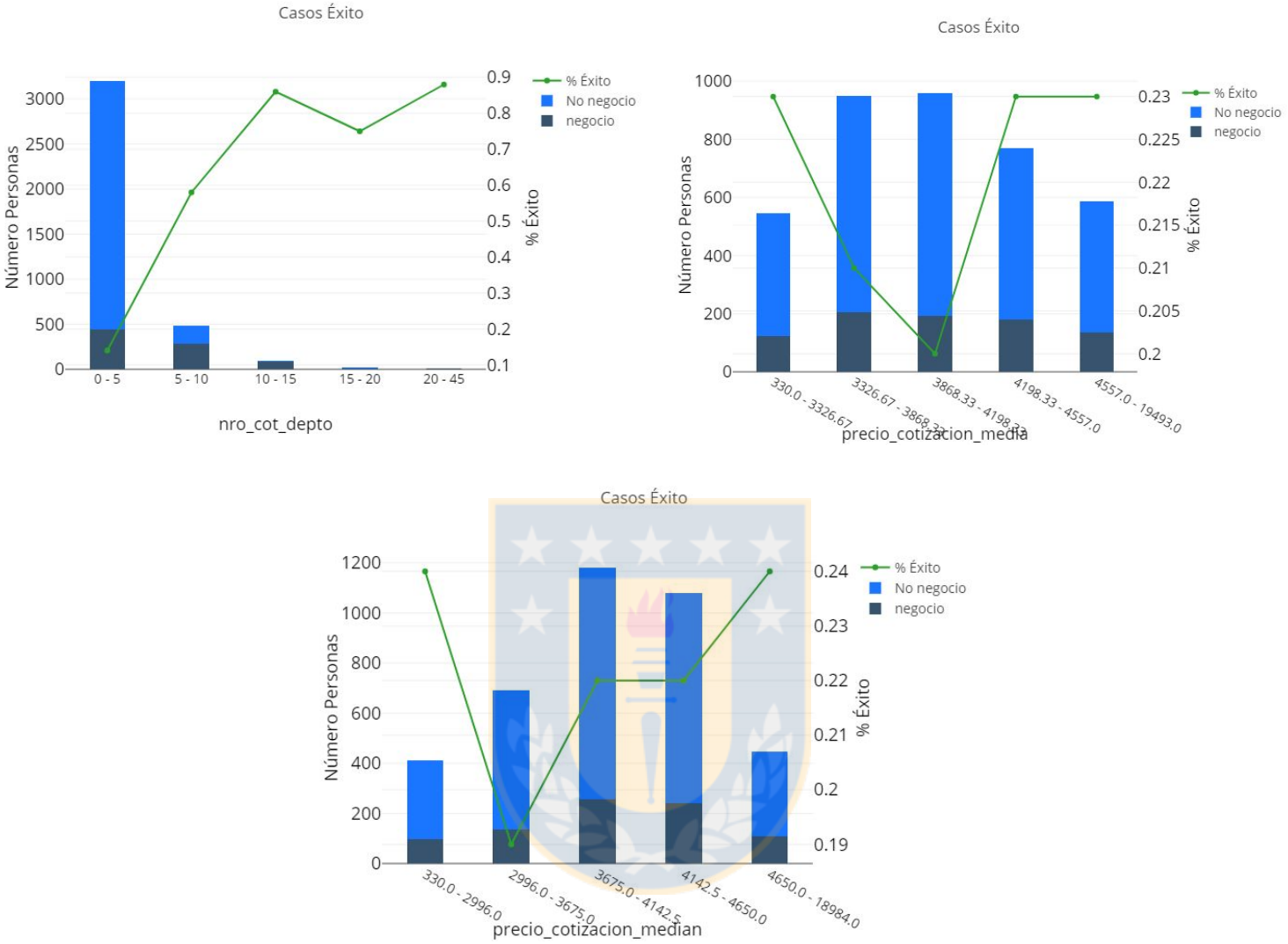

Figura 30. Gráficos que describen de manera univariada el top 3 de los feature más importantes identificados por el modelo *RF* para la tarea 2 y variable objetivo "negocio". Los elementos de barra (azul claro y azul oscuro) sumados son el total Fuente(Elaboración propia)

Resultado Tarea 2 Compra.

De la misma manera que con el modelo de *Negocio*, se describen a continuación , los 4 atributos que mejor ayudan a clasificar compradores, estos son: *"nro_cot_depto"*, *"is_recontacto"* y *"precio_cotización_media"*. El atributo que entrega mayor discriminación según el RF (tabla 20) a la hora de clasificar compradores es nuevamente el número de cotizaciones por viviendas. El primer grupo "0 - 5" que concentra el mayor número de personas ($\approx$3000 personas) tiene un bajo porcentaje de conversión, de cerca del 3%, a medida que las cotizaciones aumentan también lo hace el porcentaje de conversión, pero disminuye drásticamente el número de cotizantes, con un 18% para "5 - 10" ($\approx$500 personas), un 34% para "10 - 15" (104 personas), un 17% para "15 - 20" (24 personas) y un 40% para "20 - 45" (10 personas). En el caso del atributo *"is_recontacto"* tuvo 3% de conversión para quienes no fueron

re-contactados (≈2700 personas) en contraste con un 15% para quienes sí fueron contactados (≈1000 personas). Con respecto a *"precio_cotizacion_media"* ningún grupo supera el 7% con un mínimo de 6% en la conversión. Para ver mas detalle ver figura 31.

Tabla 20. Feature Importance Tarea 2 Compra.

Lugar	Atributo	Importancia
1	nro_cot_depto	0.25
2	is_recontacto	0.18
3	precio_cotizacion_media	0.09

Tabla 21. Fuente(Elaboración propia)

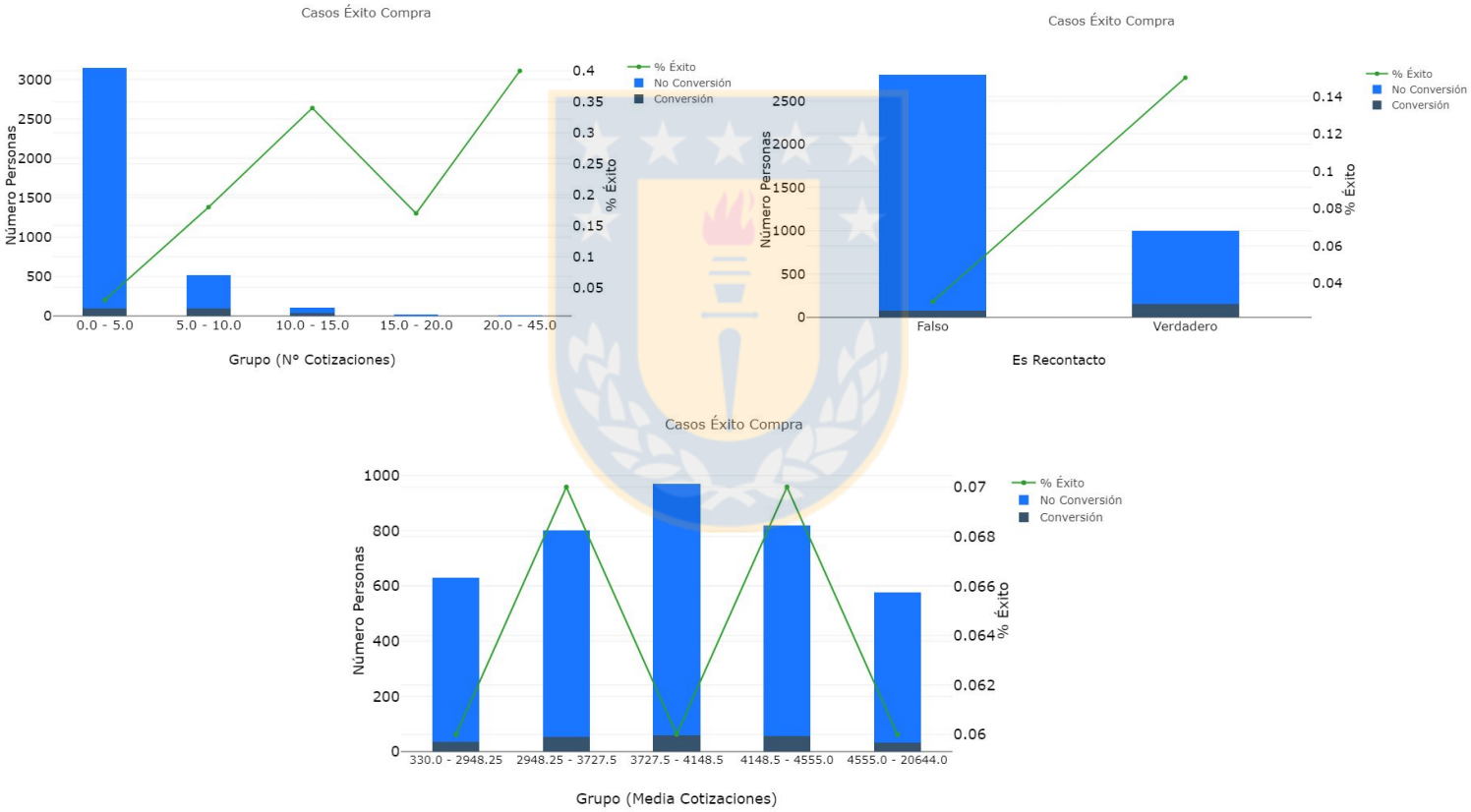

Figura 31. Gráficos que describen de manera univariada el top 3 de los feature más importantes rescatados por el modelo *RF* para la clasificación de *Compra* Tarea 2. Fuente(Elaboración propia)

4.6. Prototipo Software

Esta etapa busca implementar los modelos en un software orientado para los ejecutivos comerciales, para mejorar el procesos de captación de clientes de la empresa.

4.6.1. Software

El software de prototipo creado busca potenciar el área de marketing de la empresa, específicamente en identificar potenciales clientes que se acercan a las salas de ventas o son contactados por los ejecutivos de venta vía telefónica.

Aspectos técnicos. El prototipo se construyó usando tecnologías web. Se usó como base el lenguaje de programación python en su versión 3.6. Para crear el prototipo se utilizó un framework llamado Dash. Dash está construido sobre otros 2 frameworks, Flask y ReactJS. Flask es un *framework backend,* que funciona como servidor de la aplicación y sus datos, mientras que ReactJS es un *framework frontend* que funciona del lado del cliente en sus navegadores web. Dash hace uso de ambos frameworks y agregar componentes propios.

Software funcionalidad. El software cuenta con 3 funcionalidades principales, se puede apreciar más en detalle en la figura 32.

1. Agregar nuevo cliente, revisar sus datos personales, ver su historial de cotizaciones y visualizar su probabilidad de compra o negocio. Ver figura 33, figura 34 y figura 35.
2. Buscar un cliente existente, revisar sus datos personales, ver su historial de cotizaciones y visualizar su probabilidad de compra o negocio. Ver figura 36 y figura 37
3. Listar dos listas por proyecto, una para compra y otra para negocio. Las listas muestran de manera descendente los clientes existentes en la base de datos según su probabilidad de compra y negocio. Ver Figura 38.

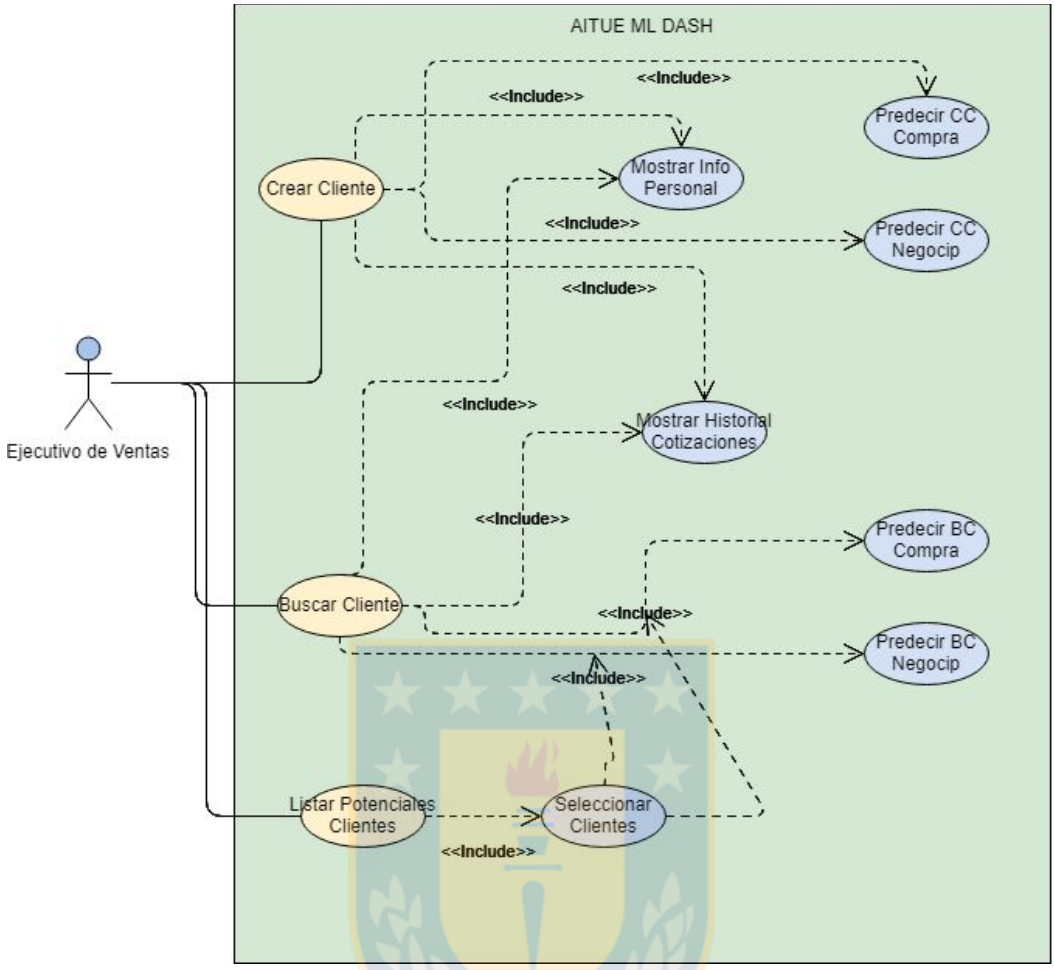

Figura 32. Diagrama de casos de uso del software. Tiene de principal actor el ejecutivo de ventas, quien es el único que interactúa con el sistema durante una cotización. Fuente(Elaboración propia)

4.6.2. Casos de Uso

La figura 32 muestra el diagrama de casos de uso, donde el título “Aitué ML Dash” representa el tema del uso de caso, donde el rectángulo representa los límites del sistema. A continuación se presentan las descripciones de los 3 casos de uso presentes en el diagrama de casos de uso.

4.6.2.1. Caso de uso Crear Cliente

Nombre	Crear Cliente
Actores	Ejecutivo de Ventas
Descripción	El ejecutivo llena un formulario con el fin de ingresar un cotizante a la base de datos del sistema. Con el formulario ingresado, el sistema automáticamente mostrará los datos personales, los datos del comportamiento histórico del cotizantes, un puntaje de acuerdo a probabilidad estimada por el modelo de la tarea 1 para "negocio" y un puntaje de acuerdo a probabilidad estimada por el modelo de la tarea 1 para "compra"
Precondición	Deben existir modelos entrenados
Postcondición	El cliente nuevo debe quedar en cache para acelerar su ingreso de cotizaciones
Flujo	1. Apretar el botón "Agregar Cliente". 2. Llenar el formulario según el cotizante disponga sus datos. 3. Apretar el botón "Agregar Cliente" dentro del formulario. 4. Esperar que el programa muestre la información personal, la información histórica, el ranking de negocio y el ranking de compra.
Flujo alternativo	Causa: error interno del programa. 3. El sistema no genera respuesta y se cancela el caso de uso.

4.6.2.1. Caso de uso Buscar Cliente

Nombre	Buscar Cliente
Actores	Ejecutivo de Ventas
Descripción	El ejecutivo ingresa un rut y luego consulta la base de datos por el cliente histórico. El sistema automáticamente mostrará los datos personales, los datos del comportamiento histórico del cotizantes, un puntaje de acuerdo a probabilidad estimada por el modelo de la tarea 2 para "negocio" y un puntaje de acuerdo a probabilidad estimada por el modelo de la tarea 2 para "compra"
Precondición	El cliente debe existir en la base de datos y el rut debe ser válido. Deben existir modelos entrenados
Postcondición	El cliente debe quedar en cache para acelerar su ingreso de cotizaciones
Flujo	1. Ingresar un rut en el buscador 2. Apretar el botón "Consultar". 3. Esperar que el programa muestre la información personal, la información histórica, el ranking de negocio y el ranking de compra.
Flujo alternativo	Causa: error interno del programa. 3. El sistema no genera respuesta y se cancela el caso de uso.

4.6.2.1. Caso de uso Buscar Cliente

Nombre	Listar Potenciales Clientes
Actores	Ejecutivo de Ventas
Descripción	El ejecutivo selecciona un proyecto y un conjunto de clientes. El sistema automáticamente mostrará 2 listas ordenadas de manera descendente según la probabilidad estimada de los modelos, una para negocio y otra para compra. Las listas son tuplas con información personal básica, de contacto y el puntaje asignado por el modelo.
Precondición	Deben existir modelos entrenados y conjuntos de clientes.
Postcondición	El cliente debe quedar en cache para acelerar su ingreso de cotizaciones
Flujo	4. Ingresar un rut en el buscador 5. Apretar el botón “Consultar”. 6. Esperar que el programa muestre la información personal, la información histórica, el ranking de negocio y el ranking de compra.
Flujo alternativo	Causa: error interno del programa. 3. El sistema no genera respuesta y se cancela el caso de uso.

4.6.3. Descripción Pantallas

Pantalla cliente nuevo. La figura 33 muestra una pantalla que integra 5 componentes:

- *Botón Agregar Cliente*: Este botón despliega un formulario en forma de modal. Los atributos son aquellos que corresponden al ingreso en el formulario oficial de la plataforma de Aitué.
- *Componente Información Personal*: Despliega la información personal del recién ingresado cotizante a la base de datos de clientes.
- *Componente Información Cotizaciones*: Despliega la información del comportamiento histórico del recién ingresado cotizante a la base de datos de clientes.
- *Componente Ranking Compra*: Muestra la probabilidad estimada por parte del modelo final seleccionado para la tarea 1 con variable objetivo “compra”
- *Componente Ranking Negocio*: Muestra la probabilidad estimada por parte del modelo final seleccionado para la tarea 1 con variable objetivo “negocio”

La figura 34 muestra el formulario en forma de modal que se despliega luego de apretar el botón agregar cliente. El formulario cuenta con un botón de enviar, el cual agrega el nuevo cliente a la base de datos y, desencadena una serie de operaciones que llevan a la figura 35. Esta, muestra los resultados de agregar un nuevo cliente, como: información personal, información de cotizaciones, puntaje de la v.o. “compra” y puntaje de la v.o. “Negocio”.

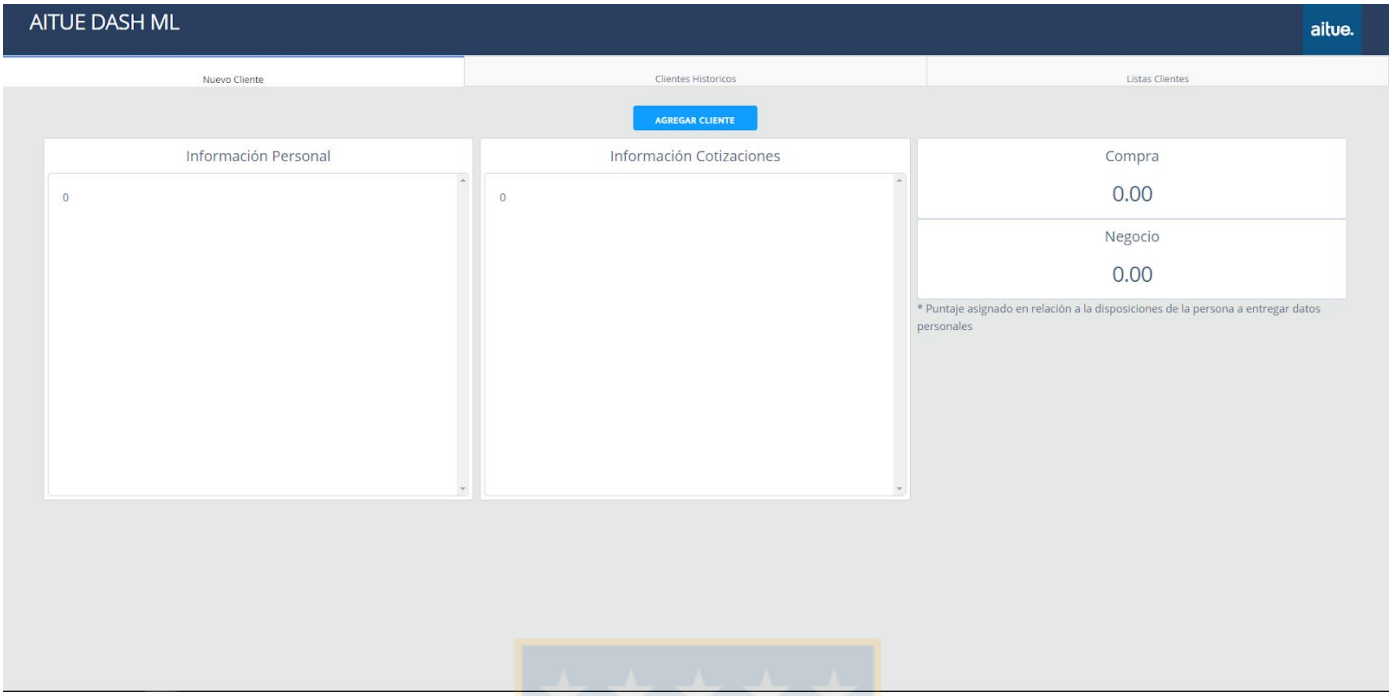

Figura 33.Vista del módulo de nuevo cliente. La interacción principal es con el botón “agregar cliente”.Fuente(Elaboración propia)

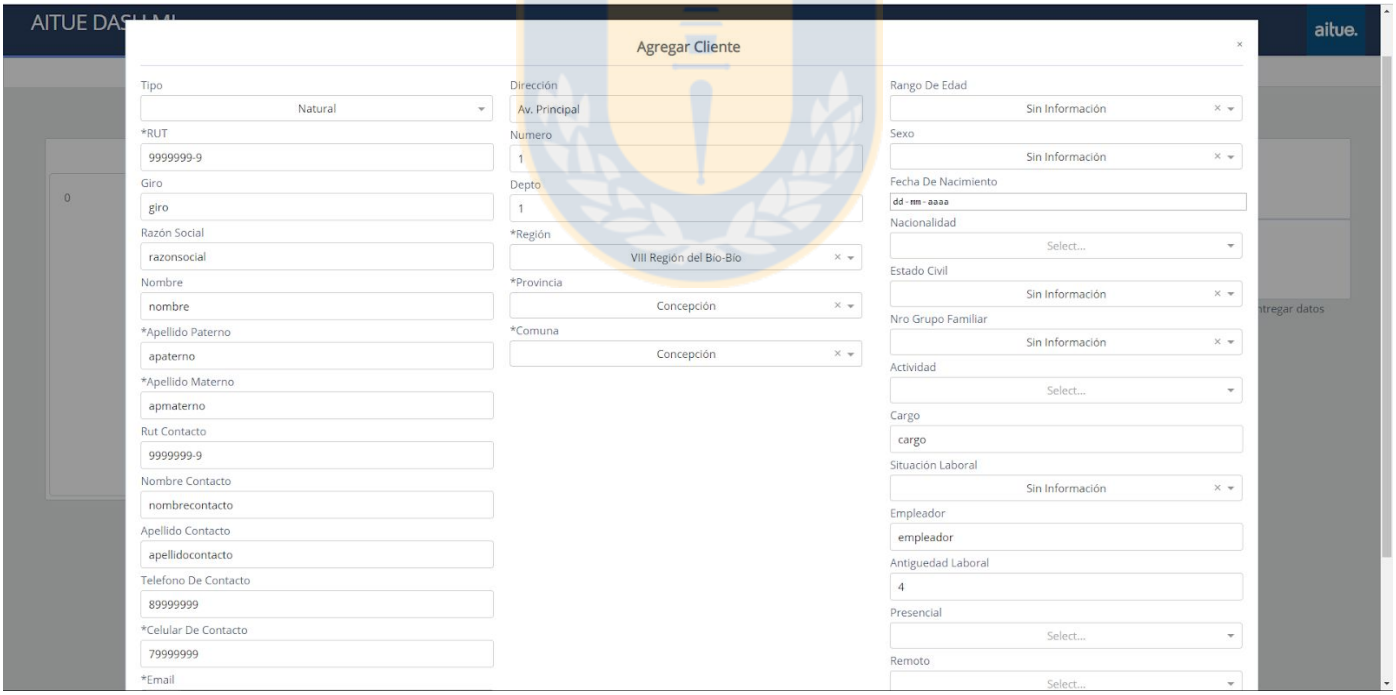

Figura 34. Modal con formulario para ingresar un nuevo cliente al sistema.Fuente(Elaboración propia)

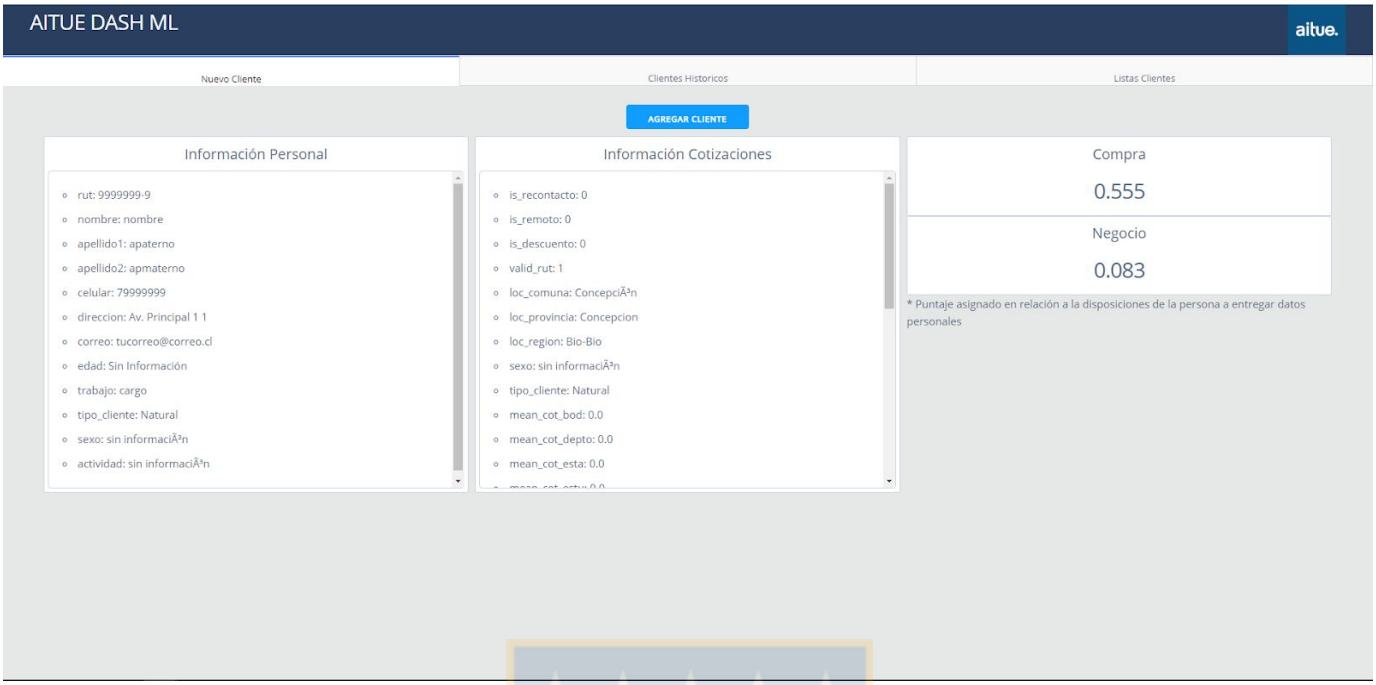

Figura 35. Vista “Nuevo Cliente” luego de haber agregado un cliente nuevo. Se puede ver su información personal, su historial de cotizaciones vacío y los puntajes asignados por los modelos.

Pantalla clientes históricos. La figura 36 muestra una pantalla que integra 7 componentes:

- *Input Rut Cliente*: Componente de input para ingresar el rut de un cliente histórico.
- *Dropdown Menu Estimador Proyecto*: Elige el modelo a aplicar para el cliente (Actualmente solo disponible modelos para el proyecto de San Andrés Del Valle)
- *Botón Buscar Cliente*: Realiza la consulta a la BD sobre el cliente, trayendo los datos del cliente.
- *Componente Información Personal*: Despliega la información personal del recién buscado cliente de la base de datos de clientes.
- *Componente Información Cotizaciones*: Despliega la información del comportamiento histórico del recién buscado cliente de la base de datos de clientes.
- *Componente Ranking Compra*: Muestra la probabilidad estimada por parte del modelo final seleccionado para la tarea 2 con variable objetivo “compra”
- *Componente Ranking Negocio*: Muestra la probabilidad estimada por parte del modelo final seleccionado para la tarea 2 con variable objetivo “negocio”

La figura 36 muestra un resultado desfavorable en “negocio” y “compra” para el cliente con rut “9571973-2”, mientras que la figura 37 muestra un resultado favorable para el cliente con rut “9996079-5”.

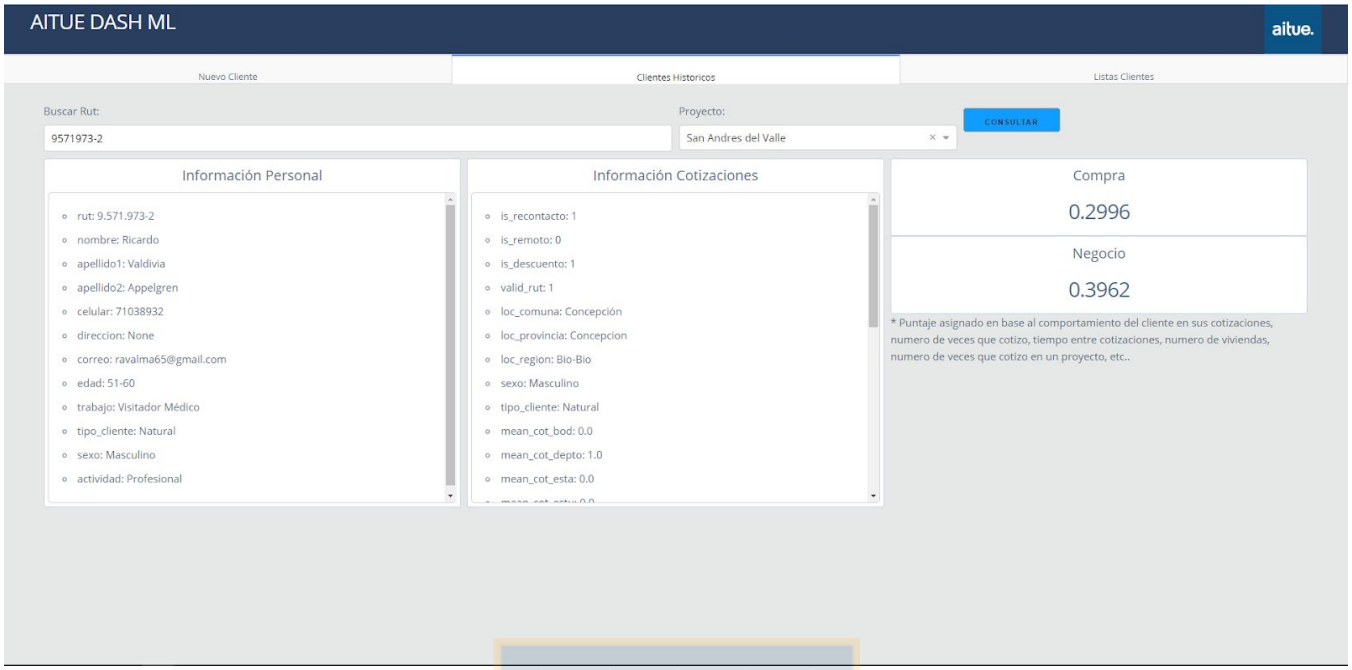

Figura 36. Muestra el resultado de buscar un cliente, el cual tiene baja probabilidad para *Compra* y *Negocio*. Fuente(Elaboración propia)

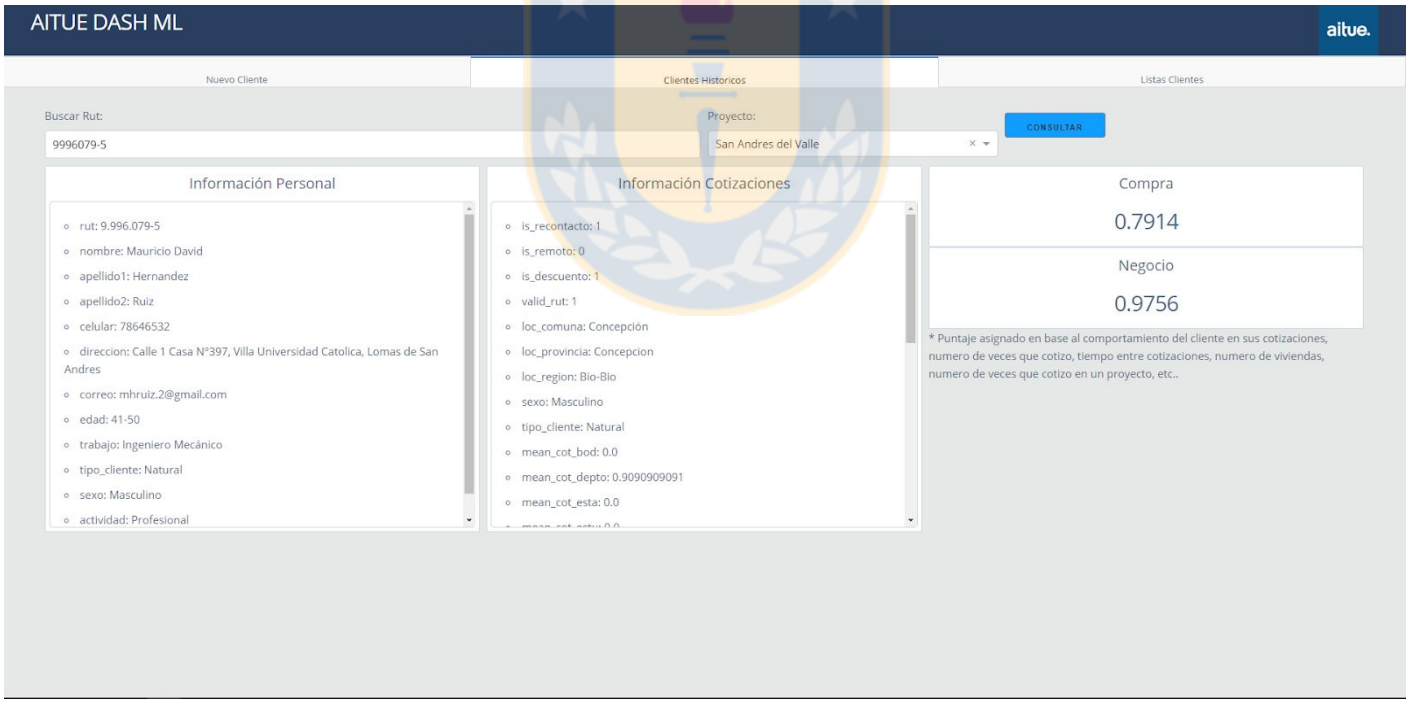

Figura 37. Muestra el resultado de buscar un cliente, el cual tiene una alta probabilidad. Fuente(Elaboración propia)

Pantalla lista clientes. La figura 38 muestra una pantalla que integra 4 componentes:

- Selección Modelo: Selección de los modelos de un proyecto que se aplicaran en la predicción de una lista de clientes. Los modelos que se ocupan son aquellos de la tarea 2. Actualmente esta disponible solo los modelos de San Andrés Del Valle.

- Conjunto a Evaluar: Selección de clientes a los que se evaluaran, actualmente solo esta disponible el conjunto de prueba.
- Lista Ranking Negocio: Muestra una lista de clientes, ordenada descendentemente según su probabilidad estimada de pertenecer a la clase negocio.
- Lista Ranking Compra:Muestra una lista de clientes, ordenada descendentemente según su probabilidad estimada de pertenecer a la clase compra.

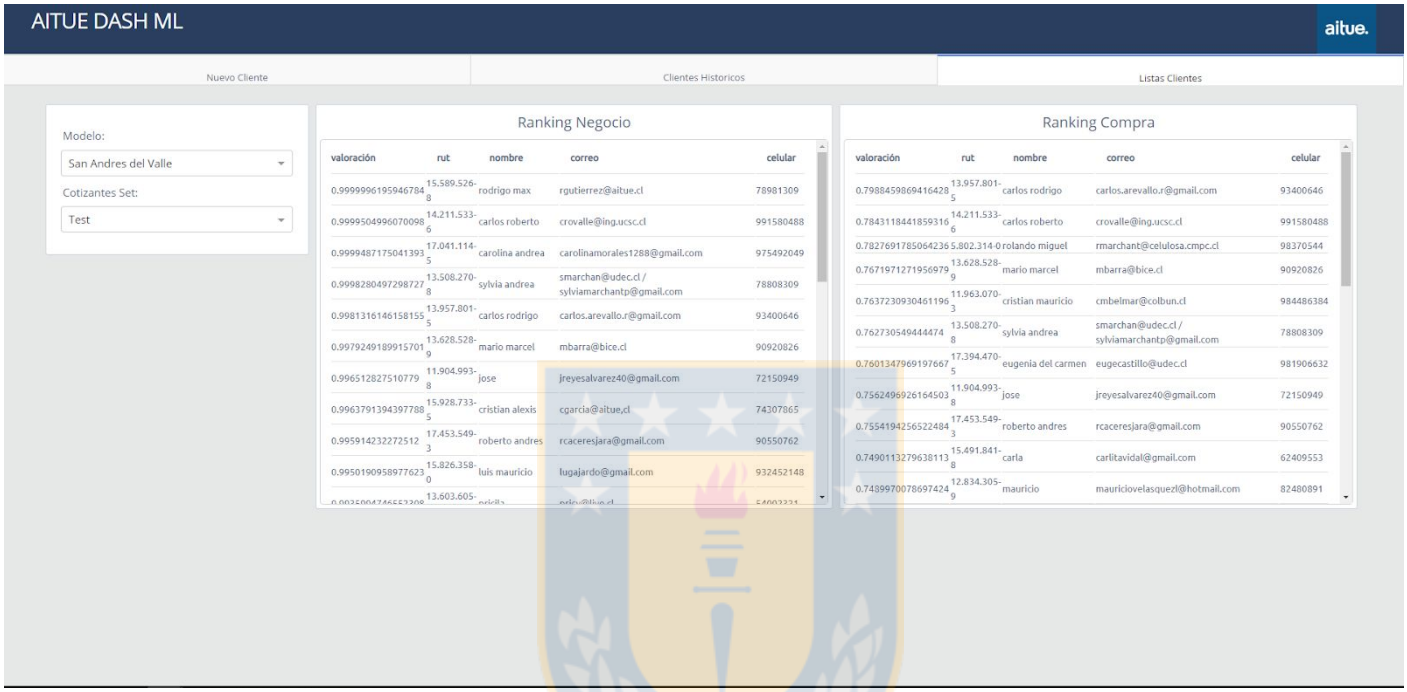

Figura 38. Muestra dos listas rankeadas de clientes para el proyecto de “San Andrés Del Valle”. Rankea de manera descendente a los cotizantes en los datos de Test.

4.6.4 Alcance Prototipo

El software consiste en una prueba de concepto de como funcionaria un software final. El software no se implementó en la empresa por el alcance del mismo, pero en el Anexo II se presentan una guia inicial para lograr una correcta implementación e integración con los sistemas de Aitué.

5. Conclusión

En este trabajo se exploró la aplicación de distintos algoritmos de aprendizaje de máquina altamente usados en ambientes de marketing, más específicamente, se usan para estrategias de identificación y captación de clientes. Estos algoritmos fueron *Support Vector Machine*, *XGBoost (XGB)*, *Random Forest*, *Decision Tree* y *Logistic Regression*. Se abordaron 2 tareas, predecir un cliente para un cotizante nuevo, la tarea 1, y predecir un cliente en base su historial de cotizaciones, la tarea 2. Para cada tarea se realizaron 2 modelos, uno para clasificar si un cotizante entra a el proceso de negocio (variable objetivo "negocio") y un segundo modelo para clasificar si un cotizante realizará una compra (variable objetivo "compra"). Los resultados de los mejores modelos para las tareas son: *XGB* con una exactitud de 0.86, un *F1_score* de 0.65 y un *Area Under the Curve (AUC) del grafico Reciever Operating Curve* de 0.83 para la tarea 1 y v.o. "negocio", *RF* con una exactitud de 0.89, un *F1_score* de 0.5 y un *AUC* de 0.93 para la tarea 1 y v.o. "compra", *LR* con una exactitud de 0.80, un *F1_score* de 0.6 y un AUC de 0.83 para la tarea 2 y v.o. "negocio" y, *RF* con una exactitud de 0.82, un *F1_score* de 0.3 y un *AUC* de 0.81 para la tarea 2 y v.o. "compra". Respecto al Lift, el análisis no entregó información concreta de un modelo indiscutiblemente ganador, más bien, nos indicó que algunos modelos pueden usarse intercambiadamente, dependiendo del tamaño de la muestra que queramos ocupar o si queremos alcanzar a un número determinado de potenciales clientes (*TP)*, exceptuando la tarea 2 con v.o. "negocio", donde, si de querer alcanzar una población superior al 15%, LR supera ligeramente a los demás modelos. La Tarea 1 con v.o. "compra" tuvo bajos errores de falsos negativos, atrapando a los clientes que efectivamente podrían llegar a realizar una compra, esto desde el punto de vista del negocio es más beneficioso que dejar pasar potenciales clientes. Además la métrica *AUC* está por sobre 0.90, lo que es bueno, significa que los *TP* , están sobresaliendo por encima de los FN. El *F1_score* se ve bajo por el hecho de que los etiquetas para esta tarea, la diferencia numérica es mucha. En la tarea 2 con v.o "compra" y "negocio", los modelos produjeron clasificaciones de falsos positivos y verdaderos positivos en una tasa similar, Dentro del conocimiento ganado se encuentran ciertos atributos que fueron reveladores al momento de clasificar, los atributos que entrega mayor información fueron: el número de integrantes de la familia, si entregan la fecha de nacimiento, si entregan la dirección y si entregan el teléfono, los primeros dos apareciendo en ambos modelos, mientras que la tarea 2, los modelos arrojaron que el número de cotizaciones, que a medida que las personas realizan más cotizaciones, su probabilidad estimada para "negocio" y "compra" aumentan. Otros atributos destacados en la tarea 2 fueron: Las personas que recibieron un descuento o fueron contactadas posteriormente a sus cotizaciones. Por último se realizó un prototipo de software que implementa estos modelos con el fin de ayudar en la tarea diaria de los ejecutivos de contactar cotizantes históricos y evaluar nuevos prospectos. De esta manera, hay un directo beneficio al utilizar los modelos de Negocio y Compra, ya que ayudan a los ejecutivos a llegar a la cuota trimestral o anual, proporcionándoles información invaluable para la captación de potenciales nuevos clientes.

Referencias

[1] Uthayasankar Sivarajah, Muhammad Mustafa Kamal. ***Critical analysis of Big Data challenges and analytical methods***. Journal of Business Research, 2016. (https://www.researchgate.net/publication/306051907_Critical_analysis_of_Big_Data_challenges_and_analytical_methods)

[2] Nada Elgendy, Ahmed Elragal. ***Big Data Analytics: A Literature Review Paper***. Department of Business Informatics & Operations, German University in Cairo, 2014.

[3]Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: ***Big Data: The Next Frontier for Innovation, Competition, and Productivity***. McKinsey Global Institute Reports, 2011.

[4] Russom, P. ***Big Data Analytics. In: TDWI Best Practices Report***. 2011.

[5]Economist Intelligence Unit,: ***The Deciding Factor: Big Data & Decision Making***.
Capgemini Reports, 2012.

[6] Nyce, Charles, **Predictive Analytics White Paper**, American Institute for Chartered Property Casualty Underwriters/Insurance Institute of America, p. 1, 2017

[7] Mitchell, T. (1997). **Machine Learning. McGraw Hill**. p. 2. ISBN 978-0-07-042807-2.

[8] https://es.datachile.io/geo/biobio-8#demography

[9] James, Henry, Stephen, A A, Crosby, Henry Jarvis, Stephen A, **A spatio-temporal, Gaussian process regression, real-estate price predictor**, 2016

[10] Li, Da Ying, Xu, Wei, Zhao, Hong, Chen, Rong Qiu, **A SVR based forecasting approach for real estate price prediction**, Proceedings of the 2009 International Conference on Machine Learning and Cybernetics, Volume 2, July 2009.

[11]Henn, A. Römer, Christoph, Gröger, Gerhard ,Plümer, Lutz, **Automatic classification of building types in 3D city models**, GeoInformatica, Volume 16, 2012.

[12]Trawinski. Bogdan, Telec. Zbigniew, Krasnoborski. Jacek, Piwowarczyk. Mateusz, Talaga. Michal, Lasota Tedeusz, Sawilow, Edward, **Comparison of expert algorithms with machine learning models for real estate appraisal,** 2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications, 2017.

[13]Graczyk, Magdalena Trawinski, Bogdan, **Computational Collective Intelligence. Semantic Web, Social Networks and Multiagent Systems**, Volume 5796, 2009.

[14]Horino, Hiroki Nonaka, Hirofumi Claire, Elisa Carreón, Alemán Hiraoka, Toru, **Development of an Entropy-Based Feature Selection Method and Analysis of Online Reviews on Real Estate**, 2017.

[15]You, Quanzeng Pang, Ran Cao, Liangliang Luo, Jiebo, **Image Based Appraisal of Real Estate Properties**, 2016. (https://arxiv.org/pdf/1611.09180.pdf)

[16]Sangani, Darshan Erickson, Kelby Hasan, Mohammad Al, **Predicting Zillow Estimation Error Using Linear Regression and Gradient Boosting**, 2017 IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS), 2017

[17]Wang, Lingjing Qian, Cheng Kats, Philipp Kontokosta, Constantine Sobolevsky, Stanislav, **Structure of 311 service requests as a signature of urban location**, PLoS ONE, Volume 12, 2017.

[18], Poursaeed, Omid Matera, Tomáš Belongie, Serge, **Vision-based real estate price estimation**, Machine Vision and Applications, Volume 29, 2018

[19] www.sas.com,Data Mining From A to Z, White Paper, SAS INstitute, 2016.

[20]Wei-Chao Lin, Shih-Wen Ke, Chih-Fong Tsai, **Top 10 Data Mining Techniques in Business Applications: A Brief Survey**, Kybernetes, https://doi.org/10.1108/K-10-2016-0302

[21]E.W.T. Ngai, Li Xiu, D.C.K. Chau, **Application of data mining techniques in customer relationship management: A literature review and classification**, Expert Systems with Applications, Elsevier, 2009.

[22]Sahar F, **Machine-Learning Techniques for Customer Retention: A Comparative Study**, International Journal of Advanced Computer Science and Applications, Vol 9, N°2, 2018.

[23] Nikita Patel, Saurabh Updahyay. Study of Varioys **Decision Tree Pruning Method with their Emprirical Comparison in Weka**. International Journal of Computer Applications, Volume 60 No.12, 2012.

[24] Sriram Vajpeyam, **Understanding Shannon's Entropy metric for Information**. Arxiv.org. 2014 (https://arxiv.org/abs/1405.2061)

[25] Provost FJ, Fawcett T, Kohavi R. **The case against accuracy estimation for comparing induction algorithms**. In: International Conference on Machine Learning, ICML. San Francisco: Morgan Kaufmann Publishers Inc, 1998. (https://pdfs.semanticscholar.org/7770/3a2783f64dfceb638aa9eebd9c9c501bb835.pdf)
[26] Mauno Vihinen. **How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis**. BMC Genomics. 18 June 2012 descargado (https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-13-S4-S2)
[27]Brynjolfsson, Hitt, & Kim, **Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?**, 2011.
[28] Shearer**, CRISP-DM**, 2000
[29] Misha Denil, David Matheson, Nando de Freitas, Narrowing the Gap: **Random Forests In Theory and In Practice**,, 2011
[30] L. Breiman and J. Friedman and R. Olshen and C. Stone, **Classification and Regression Trees**, Wadsworth and Brooks, Monterey, CA, 1984,
[31] Carl Kingsford, Steven L Salzberg. **What are decision trees?**. Department of Computer Science, Institute for Advandced.Nat Biotechnol Computer Studies and Center for Bioninformatics and Computational Biology, University of Maryland, College Park, USA. 2008 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2701298/pdf/nihms92537.pdf)
[32] L. Breiman, **Random Forests**, Kluwer Academic Publishers. Manufactured in The Netherlands. 2011(https://link.springer.com/content/pdf/10.1023%2FA%3A1010933404324.pdf)
[33] https://scikit-learn.org/stable/about.html
[34] Charles X, Chenghui L, **Data Mining for Direct Marketing: Problems and Solutions**, KDD, 1998
[35]Mohamed B. Kheliouane D, Akrouf A **Evaluation Measures for Models Assessment over Imbalanced Data Sets**, Journal of Information Engineering and Applications, 2013
[36] Jonathan Burez Dirk, Van den Poel, **Handling class imbalance in customer churn prediction**, expert systems with applications. 2009
[37] Foster Provost, Tom Fawcett, **Data Science for Business**,O'Reilly Media, 2013
[38] Ilyes Jenhani, Nahla Ben Amor, Zied Elouedi. **Decision trees as possibilistic classifiers**, Science Direct, Elsevier. LARODEC, Institut Supérier de Gestion, Tunis, Tunisia. 2007
[39] Jiawei Han, MIcheline Kambre, Jian Pei, **Data Mining Concepts and Techniques**, Morgan Kaufmann Publishers, 2011
[40]Osvaldo Simeone, **A Brief Introduction Machine Learning for Engineers**, Department of Informatics, King's College London, arxiv, 2018.
[41] Tianqi Chen, Carlos Guestrin, **XGBoost: A Scalable Tree Boosting System**, 2016
[42] Wolpert, D.H., Macready, W.G. (1995), **No Free Lunch Theorems for Search**, Technical Report SFI-TR-95-02-010

Anexo I

Tabla Descriptiva Datos Cotización

	Tipo Columna	Valores Nulos	Personas Total	Filas Valores Unicos	Filas Total	Filas Disponibles	Personas Entregaron	Personas Negocio	Personas No Negocio	Personas Compra	Personas No Compra
ID	int64	0	6184	16592	16592	16592	6184	1102	5082	288	5896
Comuna	object	0	6184	106	16592	16592	6184	1102	5082	288	5896
Provincia	object	0	6184	38	16592	16592	6184	1102	5082	288	5896
Region	object	0	6184	15	16592	16592	6184	1102	5082	288	5896
Negocio	bool	0	6184	2	16592	16592	6184	1102	5082	288	5896
Nombre Completo	object	0	6184	6179	16592	16592	6184	1102	5082	288	5896
Nombre	object	0	6184	2481	16592	16592	6184	1102	5082	288	5896
RUT	object	0	6184	6184	16592	16592	6184	1102	5082	288	5896
Ejecutivo Comercial	object	0	6184	62	16592	16592	6184	1102	5082	288	5896
Venta	bool	0	6184	2	16592	16592	6184	1102	5082	288	5896
Jefe Comercial	object	0	6184	3	16592	16592	6184	1102	5082	288	5896
Fecha Cotizacion	datetime64[ns]	0	6184	16589	16592	16592	6184	1102	5082	288	5896
Medio	object	0	6184	12	16592	16592	6184	1102	5082	288	5896
Proyecto	object	0	6184	5	16592	16592	6184	1102	5082	288	5896
Etapa	object	0	6184	12	16592	16592	6184	1102	5082	288	5896

Tipo Cliente	object	1	6184	2	16592	16591	6184	1102	5082	288	5896
Correo Electronico	object	5	6184	6082	16592	16587	6180	1102	5078	288	5892
Total Productos	float64	30	6184	2195	16592	16562	6184	1102	5082	288	5896
Celular	object	38	6184	6037	16592	16554	6164	1100	5064	288	5876
Apellido 1	object	39	6184	1865	16592	16553	6160	1091	5069	286	5874
Valor Final Venta	float64	44	6184	2539	16592	16548	6184	1102	5082	288	5896
Descuentos	float64	44	6184	339	16592	16548	6184	1102	5082	288	5896
Productos	object	83	6184	3534	16592	16509	6182	1102	5080	288	5894
Antiguedad Laboral	float64	172	6184	35	16592	16420	6110	1091	5019	287	5823
Apellido 2	object	407	6184	1665	16592	16185	5962	1087	4875	286	5676
Tipo de Medio	object	532	6184	31	16592	16060	5990	1102	4888	288	5702
Sexo	object	1869	6184	3	16592	14723	5184	1034	4150	286	4898
Rango Edad	object	1980	6184	7	16592	14612	5137	1031	4106	285	4852
Presencial	object	2027	6184	2	16592	14565	5438	982	4456	242	5196
Estado Civil	object	2049	6184	9	16592	14543	5094	1040	4054	286	4808
Actividad	object	2350	6184	6	16592	14242	4987	1015	3972	285	4702
Remoto	object	3310	6184	2	16592	13282	4984	903	4081	211	4773
N° Grupo Familiar	object	5964	6184	9	16592	10628	3579	851	2728	251	3328
Profesión	object	9226	6184	129	16592	7366	2251	714	1537	220	2031

Direccion	object	11384	6184	1291	16592	5208	1324	741	583	269	1055
Nacionalidad	object	11744	6184	8	16592	4848	1149	639	510	236	913
Fecha Nacimiento	datetime64[ns]	12701	6184	840	16592	3891	871	535	336	210	661
Situacion Laboral	object	12795	6184	3	16592	3797	857	512	345	206	651
Empleador	object	13588	6184	509	16592	3004	662	422	240	165	497
Cargo	object	13771	6184	444	16592	2821	623	401	222	165	458
Telefono	object	14598	6184	590	16592	1994	734	174	560	52	682
Razon Social	object	16552	6184	23	16592	40	25	12	13	3	22
Giro	object	16559	6184	20	16592	33	21	12	9	3	18

Tabla 19. Descripción Datos cotizaciones, disponibilidad de datos y cantidad de personas que entregaron datos sea en negocio o compra .

Anexo II

Alcance Prototipo y Guia de Integración

El prototipo hace uso de los datos provistos por Aitué y los carga de manera local. Aitué hoy hace uso de un sistema externo de servicios de tecnologías de información. La integración de este prototipo dentro de la organización necesita un esfuerzo extra, que requiere mayor personal y recursos, por lo que está fuera del alcance de este proyecto de título.

La base de datos que maneja Aitué cuenta con numerosas tablas. Los datos utilizados en este proyecto y para la creación del prototipo son una vista generada desde estas tablas de manera manual por un empleado de Aitué y no existe un canal que conecte a la base de datos que manejan sus proveedores. A continuación se propondrán soluciones para su integración dentro de la organización.

Las soluciones propuestas a continuación asumen como base el costo de integración del prototipo dentro de un servidor mantenido por Aitué o por un servicio externo como lo es un servidor cloud.

La solución número 1 consta de habilitar desde el prototipo una funcionalidad de subida de datos. Los datos deben ser generados manualmente cada vez que se desee utilizar el software de manera actualizada.

La solución número 2 consta de automatizar la generación de la vista de la base de datos desde el prototipo, reemplazando el archivo que se aloja localmente en la aplicación, para su posterior uso. Esta vista puede ser generada semanal o diariamente para no cargar los servidores cada vez que se desee hacer uso del prototipo.

Las soluciones 1 y 2 son rápidas de implementar y de bajo costo, pero presentan un problema fundamental en relación a una de las posibles funcionalidades del prototipo, esta es ingresar manualmente nuevos clientes. Las soluciones 1 y 2 proponen una “mensajería“ de un sentido, en dirección del prototipo. El prototipo consumiria información desde la plataforma principal (base de datos), desalineando los datos de nuevos clientes entre ambos.

La solución más completa sería hacer una integración de mensajerías bidireccional entre el prototipo y la bases de datos principal (bases de datos externa), por medio de APIs (Application Programming Interface). Si bien esta solución es la más completa, es a su vez las más costosa en términos de tiempo, involucra las áreas de TI de Aitué y de la empresa externa que maneja la base de datos. Se puede realizar una integración completa o parcial, esto se refiere que, no es necesario generar todas las llamadas APIs para todas las tablas, pero si se quiere escalar siempre es mejor tener más acceso y control. Otra alternativa es es que la empresa externa genere la vista y la envie via *api* al prototipo y habilitar una llamada para ingresar nuevos clientes.