

INSTITUTO TECNOLÓGICO DE BUENOS AIRES – ITBA
ESCUELA DE POSTGRADO

PROPENSIÓN A ADQUIRIR PLAZOS FIJOS EN UN BANCO TRADICIONAL

AUTOR: Sprio, Agustina Beatriz/s (Leg. N°104.348)

TUTOR: Denicolay, Gustavo

TRABAJO FINAL PRESENTADO PARA LA OBTENCIÓN DEL TÍTULO ESPECIALISTA EN
CIENCIA DE DATOS

BUENOS AIRES
SEGUNDO CUATRIMESTRE, 2019

Índice

1. Introducción.....	3
2. Estado del arte	3
3. Definición del problema	5
4. Justificación del Estudio	6
5. Alcances del trabajo y limitaciones	6
6. Hipótesis.....	7
7. Objetivos de la investigación.....	8
8. Metodología a utilizar.....	9
9. Desarrollo de la investigación.....	11
10. Resultados obtenidos	20
11. Conclusiones	24

1. Introducción

El producto Plazo Fijo es un depósito que los bancos o instituciones financieras ofrecen con una tasa de interés previamente acordada, la cual va a ser cobrada por cliente junto con el dinero involucrado en el depósito en un determinado período de tiempo.

Cada entidad bancaria conoce diversa información acerca de sus clientes, alguna relacionada a características inherentes a él como en dónde vive, qué medios de contacto posee y a qué se dedica. Además, conoce la relacionada a sus transacciones, a sus operaciones en el sistema financiero, a productos asociados, a sus reclamos realizados, entre otras.

Toda esta información compone la historia de los clientes, la cual será analizada tanto en clientes que aún no poseen el producto plazo fijo, como en los que ya lo adquirieron para aprender de su comportamiento y predecir la propensión que poseen a adquirir el producto.

2. Estado del arte

Las personas destinan sus ingresos al consumo y al ahorro. Atucha (2018) [1] define al consumo como a la intención del gasto en bienes y servicios para satisfacer diversas necesidades, y al ahorro como al deseo de guardar parte de tal ingreso para ser consumido en un futuro.

Tal como explica Fisher et. Al. (1989) [2], las decisiones de las economías domésticas sobre el porcentaje que desean consumir son al mismo tiempo decisiones sobre lo que desean ahorrar. Relacionado a los objetivos que motivan a la intención de ahorrar, Bernake, B. y Frank, R. (2007) [3], distinguen el ahorro del ciclo vital del ahorro precaución y adicionalmente, al ahorro legado. El primer motivo se refiere a que la personas ahorran para cumplir objetivos a largo plazo como tener una buena jubilación, pagar los estudios universitarios de sus hijos o el deseo de proveerse de medios para hacer frente a determinadas necesidades como la compra de un auto o una vivienda. El segundo motivo mencionado, es el que se relaciona con el deseo de acumular reservas para imprevistos como la pérdida de empleo o una enfermedad. Por último, el ahorro legado es el que se realiza con el objetivo de dejar una herencia a parientes o instituciones.

El ahorro está directamente relacionado a la inversión, dado que la va a llevar a cabo un perfil que tenga la capacidad de ahorrar y no la de destinar todos sus ingresos a consumos inmediatos. Podemos asegurar que las decisiones de inversión se relacionan con el futuro y las expectativas que generan, por lo que son más inestables que los motivos que guían a los consumidores, generando que sea más difícil establecer una función que demuestre el comportamiento de la inversión, que la del consumo. No obstante, como lo expresa Heilbroner, R. y Thurow, L. (1987) [4]: “El gasto de inversión no es una variable al azar. Existen patrones en la inversión, aunque puedan ser desplazados por variaciones repentinas e imprevisibles en el gasto total de inversión”.

De manera general, los factores que afectan a la inversión son la confianza, la tasa de interés, el costo de los bienes de capital y el nivel de ingreso de la economía. Sin embargo, cuando nos referimos particularmente a la inversión de plazo fijos en el ámbito bancario, existen muchos otros factores que también se deberían tener en cuenta.

La industria bancaria compone un importante sector de la economía social, ofreciendo diversos productos y servicios a personas, incluyendo como productos la venta de Plazos Fijos. Este producto presenta un particular interés en el sector por ser la principal fuente de

ingresos que poseen las entidades bancarias tradicionales [5], generando que cada vez más se intente conocer cómo lograr que los clientes decidan invertir adquiriendo este tipo de productos.

En pleno crecimiento exponencial del volumen de la información, los datos son la base para entender el comportamiento de los clientes. Los datos correctamente modelados, permiten generar el conocimiento, basado en la historia de cada cliente en el banco, relacionada por ejemplo a qué productos posee, los reclamos realizados, ingresos que posee en sus cuentas, consumos realizados con sus productos, deudas adquiridas, la actividad económica a la cual se dedican, entre otros. Involucrando todos estos datos, es posible identificar grupos de clientes con características homogéneas entre los integrantes y heterogéneas entre los grupos, con el objetivo de segmentarlos y realizarles un seguimiento acorde a las características que los definen, método denominado Clustering. Esto compone el análisis de la situación actual, una explicación de la realidad en un momento dado. Por otro lado, utilizando herramientas como la minería de datos es posible conocer aún más de lo que en este momento está ocurriendo, sino que apunta a predecir con un porcentaje de efectividad aceptable, cómo responderán los clientes antes de haberlos contactado, en función a cómo actuaron clientes similares analizando su historia. La minería de datos o Datamining consiste en un conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto.

Q.R. Zhuang, Y.W. Yao, y O. Liu (2018) [6] analizan cómo se deberán aplicar acciones de Marketing directas a clientes objetivo de la primera compra de Plazo Fijos. En “nuevos compradores de Plazo Fijo”, explican, la mayor virtud es la paciencia, la duración del contacto tiene un impacto significativo en la ratio de éxito del telemarketing. Lo segundo más importante a tener en cuenta, es la cantidad de individuos a incluir en cada acción de Marketing, debe estar lo más correctamente direccionada posible. Adicionalmente, también se deberá controlar la intensidad con la que se enviarán las comunicaciones a los clientes, es decir, si en poco tiempo se lo intenta contactar muchas veces, se obtendrá un resultado negativo, no solamente con el producto en sí a ofrecer, sino también con la entidad en su totalidad. En cuanto a la manera de contactarlo, si el cliente brindó un número de línea y uno de un teléfono celular, se deberá intentar primero con el de línea, ya que está considerado como menos privado que el celular. Por último, aprovechar la segmentación realizada de clientes para aplicarles campañas de Marketing específicas que generen una mayor empatía con los mismos.

Tal como se detalló anteriormente, es importante en la decisión de invertir el contexto del país que rodea al individuo. Particularmente en Argentina, existen indicadores que orientan al inversor tales como el Riesgo País, la tasa de las Letras de Liquidez más conocidas como “LELIQ” y la cotización de la moneda Dólar Estadounidense. La empresa de servicios financieros JP Morgan calcula desde principios de la década del '90 el “Riesgo País” de la gran mayoría de los países del mundo. Este indicador es utilizado como referencia para determinar el riesgo de que no se paguen las deudas que hay en las distintas economías del mundo, en otras palabras, es un indicador simplificado e imperfecto de la situación de una economía. No se mide en porcentajes, sino que se mide en puntos básicos y corresponde a la sobretasa que paga un país por sus bonos en relación a la tasa que paga el Tesoro de Estados Unidos. Es decir, es la diferencia que existe entre el rendimiento de un título público emitido por el gobierno nacional y un título de características similares emitido por el Tesoro

de los Estados Unidos. Las letras de liquidez (LELIQ) son instrumentos creados en el 2018 a descuento en pesos emitidos por el BCRA diariamente a través de subastas. Los bancos colocan dinero a una tasa LELIQ por ejemplo del 58% y ofrecen a los ahorristas plazos fijos a una tasa promedio de entre 40 al 45 por ciento, de lo cual se deduce su ganancia. Así, el Banco Central busca que los pesos que circulan en el mercado se utilicen de esta manera y que no vayan hacia el dólar, con el objetivo de reducir la presión sobre la divisa estadounidense y que no se dispare su valor. Si aumenta la tasa LELIQ, aumenta la tasa de los plazos fijos que los bancos otorgan a sus clientes, y viceversa, impactando directamente en la decisión de los ahorristas relacionada a si conviene o no comprar este producto y cuándo es el mejor momento. A partir de fines del 2015 en Argentina está permitido comprar dólares a modo de ahorro o inversión sin límite en cuanto al monto, generando un medio más de inversión para los clientes, el cual compite directamente con la compra de plazo fijo. Aunque el valor de la compra de dólares continúa en alza desde el levantamiento del cepo, en determinados períodos generó menos ganancia que el plazo fijo, sobre todo si consideran inversiones a corto plazo. El aumento del valor del dólar estadounidense es complejo de analizar tal como indica Daniel Artana (2018) [8], "el dólar como cualquier precio de cualquier producto del mercado depende de la oferta y la demanda. Es complejo establecer una causalidad simple sobre la oferta y la demanda de dólares porque "intervienen un montón de factores". El economista Fernando Marengo (2018) [9] destaca como motivo principal de la suba que el peso se deprecia cuando baja la tasa del Banco Central, situación que genera que suban todas las demás monedas. Esto quiere decir que los ahorristas o pequeños inversores harán colocaciones en activos en pesos, siempre y cuando, no superen las expectativas de depreciación de tipo de cambio. En otras palabras, que la tasa de interés mensual sea mayor a la expectativa de la tasa de cambio.

Dado el enfoque relacionado al contexto que se contemplará, el análisis no solamente tendrá el objetivo de conocer la historia de los clientes y de su comportamiento de manera histórica sino también relacionado a los indicadores de la situación económica y financiera de Argentina en un momento dado. Interpretar el conjunto de datos obtenido para armar un modelo predictivo de propensión a nuevas compras del producto Plazo Fijo además de conocer qué características son las principales en cuanto a las reglas de decisión obtenidas.

3. Definición del problema

Dado que en un banco tradicional existe una gran cantidad de acciones de venta, las áreas de Marketing pretenden focalizar las ofertas que se le comunican al cliente por los diferentes canales de contacto (llamada telefónica, SMS, email, Push Notification, entre otros) con el objetivo de mejorar la eficiencia, es decir, a menor cantidad de contactos realizados, mayores ventas. Ningún cliente quiere recibir muchas comunicaciones al día, por lo que analizan cuál es la mejor oferta para cada cliente y el mejor medio de comunicación a utilizar, actualmente a través de técnicas sencillas de segmentación.

Estas técnicas consisten principalmente en identificar quienes creen que podrían estar interesados, y seleccionarlos agrupándolos por características básicas, por ejemplo, utilizando la edad para determinar si corresponde a un segmento Joven, Adulto o Mayor.

4. Justificación del Estudio

Particularmente, tal como explican James Chen y Chris B Murphy (2019) [7] en su artículo, el producto Plazo Fijo es el más importante para un banco porque si un cliente deposita dinero en el mismo, el banco puede invertirlo en otros productos financieros que pagan un rendimiento más alto de lo que el banco le está pagando al cliente. Entonces, eligen priorizar la oferta de este producto por la de otros teniendo como premisa que por lo general los clientes no leen todas las comunicaciones enviadas por el banco.

El desarrollo de un modelo predictivo para la propensión a adquirir Plazos Fijos, permite darle un giro a la definición de segmentación tradicional, y cambiar el rumbo de las comunicaciones de Marketing a uno más inteligente. Aprender del comportamiento histórico que tuvieron los clientes permite conocer cómo es probable que se comporten otros en un futuro.

Los modelos tienen asociado un porcentaje de efectividad, el cual se encuentra relacionado a qué clientes tienen una mayor probabilidad que otros y por qué, característica que responde a la necesidad de obtener los n clientes más propensos, siendo n la cantidad de objetivos que se quieren incluir en las campañas comerciales.

5. Alcances del trabajo y limitaciones

La compra del producto plazo fijo por parte de clientes a una entidad financiera tradicional puede no solamente depender de características inherente al cliente, también del contexto país que al mismo lo rodea y de los ingresos y egresos que posea en sus cuentas. Es por ello que el análisis se abordará de esa manera, dado un plazo fijo comprado en un día en particular, se analizará el contexto que tuvo y su situación económica para ese momento y los días anteriores. Adicionalmente, se agregarán características del mismo que permitirán conocer aún más si deberán ser tomadas en cuenta o no.

El producto que se estará evaluando es el “Plazo Fijo Tradicional en Pesos”, a partir de este momento “PF”, dado que la entidad otorga también otros en pesos, como los UVA y adicionalmente en moneda Dólar. Serán excluidos del análisis, dado que se entiende que los motivos por cuales llevan al cliente a comprarlos no son los mismos que el producto en cuestión. También serán excluidos del análisis productos que correspondan a una renovación de uno anterior el día del vencimiento, ya que la probabilidad de que saquen uno nuevo a partir del mismo es mucho mayor que la de los productos nuevos.

Con respecto a los clientes seleccionados para el análisis, se tomaron en cuenta a las personas físicas, descartando a los empleados del banco y a las jurídicas por tener diferentes condiciones al momento de obtener en producto. A los integrantes marcados como compradores, los llamaré de acá en más “positivos”, mientras que a los que no compraron “negativos”.

En relación al período a seleccionar, se utilizará únicamente lo ocurrido entre el 01/02/2019 hasta el 30/06/2019. Durante ese momento hubo días en los que tanto el banco, como el mercado financiero no operaron, entonces, el análisis se redujo a los días hábiles. Esto evitará inconsistencias futuras en cuanto a la evaluación de la historia de días anteriores, fijando un día como principal.

Se limitará la cantidad de clientes a analizar en función a la cantidad de positivos y negativos que existan en el conjunto de datos. La cantidad de integrantes en dichos grupos no puede ser muy dispar a la otra, error de concepto que podría generar resultados finales no válidos,

con un porcentaje de eficacia muy bueno. Adicionalmente, se recuperará la historia de un único banco, y el modelo predictivo servirá para que sea utilizado únicamente allí. No se podrá asegurar que el modelo sea válido para otras entidades.

Con la finalidad de determinar el contexto país que rodea a cada integrante del modelo se tendrá en cuenta el riesgo país otorgado por el BCRA, la cotización del dólar otorgada por la misma entidad bancaria y la tasa las Leliq.

Se entrenará un modelo predictivo llamado “Random Forest” de adquisición de plazo fijo a partir de la historia de los clientes generada en una entidad bancaria tradicional y adicionalmente al contexto país que lo rodea teniendo en cuenta los períodos febrero, marzo y abril para el entrenamiento, y mayo y junio para evaluar los resultados (año 2019).

Esta división de los datos permitirá probar el modelo de manera real, es decir que se predecirá qué hicieron los clientes en mayo y junio, habiendo ya ocurrido dichos meses. Es decir que, se conoce realmente como se comportaron los clientes a utilizar en la etapa de “testing”, valores que se compararán con los resultados obtenidos por el modelo. Por lo general se utiliza un 70% del conjunto de datos para “training” y un 30% para la etapa de “testing”, por lo que se consideró oportuno utilizar tres meses para el primer conjunto y dos meses para el segundo.

6. Hipótesis

“La selección de clientes basada en modelos predictivos permite reducir la cantidad de comunicación de ofertas e incrementar la venta de Plazos Fijos en un banco tradicional”

Clasificación de variables:

La hipótesis planteada es del **tipo “Hipótesis de 2 o más variables”, multivariable**, dado que están involucradas 4 variables.

La relación entre las variables es de Causa-Efecto, porque se identifica que:

A (Selección de clientes basada en Modelos Predictivos) ->

C1 (Reducción de cantidad de Comunicación de ofertas) y C2 (Incremento de Venta de Plazo Fijos)

Variables:

- Selección de clientes basada en Modelos Predictivos
 - Tipo: Variable Independiente
 - Definición nominal: La selección de clientes basada en Modelos Predictivos significa utilizar la información histórica relacionada a los clientes y a su contexto para conocer cómo responderán nuevos clientes con características parecidas a los ya analizados.
- Cantidad de Comunicación de ofertas
 - Tipo: Variable Dependiente

- Definición nominal: Nos referimos a los estímulos que se le enviaron al cliente para contactarlo por diversos canales, como email, sms, push notification, entre otros.
- Definición Operacional: La Cantidad de comunicaciones es la sumatoria de estímulos enviados a los clientes, teniendo en cuenta todos los canales de contacto
- Venta de Plazos Fijos
 - Tipo: Variable Dependiente
 - Definición nominal: La venta de Plazo Fijo es un producto bancario el cual consiste en que uno o más clientes depositen dinero de manera prolongada y reciban por ello un interés acordado.
 - Definición Operacional: La venta de Plazo Fijos se mide en función a la cantidad de productos vendidos, independientemente del plazo y/o monto elegido para cada uno.
- Banco Tradicional
 - Tipo: Variable Contextual
 - Definición nominal: Un Banco es una empresa comercial que realiza operaciones financieras con el dinero procedente de accionistas y clientes. Con tradicional nos referimos a que posee sucursales físicas a las cuales los clientes pueden concurrir para realizar sus operaciones en persona, y no únicamente de manera virtual.

7. Objetivos de la investigación

Objetivo general

Crear un modelo predictivo que permita conocer la propensión de los clientes a adquirir un producto Plazo Fijo con un porcentaje de eficiencia superior al 75%.

Objetivos específicos

- Recuperar la historia de los clientes, tanto los que compraron un Plazo Fijo como los que no, desde Febrero del 2019 hasta Junio del 2019.
- Analizar la magnitud y la distribución de los datos obtenidos.
- Realizar un perfilamiento de los clientes y armar un conjunto de datos válido para modelos predictivos. Con válido, se refiere a que no puede tener información en formato texto, todos los campos deben ser numéricos, y no nulos.
- Entrenar el modelo predictivo “Cross Validation” utilizando los períodos Febrero-Marzo-Abril/2019. Probar el modelo utilizando Mayo-Junio/2019.
- Se considerará válido si tiene asociado un porcentaje de efectividad superior al 75%
- Interpretar los resultados obtenidos

8. Metodología a utilizar

a. Técnicas

Las técnicas estarán relacionadas a las tareas que se llevarán a cabo durante el proyecto:

1. Armado del conjunto de datos: A través de consultas SQL se interrelacionarán los datos, actualmente ubicados en diferentes bases de datos.
2. Perfilamiento de los datos: Con el conjunto de datos armado, se deberá realizar un entendimiento de los datos, columna por columna. Qué tipo de dato contiene cada una, y cuáles son los posibles valores que puede tomar. Tomar decisiones en cuanto a cómo se trabajar los campos vacíos, valores incorrectos, valores similares.
3. Feature engineering: Crear variables nuevas que sirvan para agregar relaciones entre uno o más campos o categorizaciones de manera de ayudar al modelo en cuestiones conceptuales, entendibles para los humanos. Por ejemplo, dividir las edades en rangos etarios, interpretar el nivel de estudios literal como una etapa numérica agrupando varios niveles, entre otros.
4. Limpieza el dataset: Todas las variables a utilizar en el modelo deberán ser numéricas, por lo que se deberán reemplazar las literales por numéricas en caso de que sea posible, y finalmente eliminarlas del conjunto de datos definitivo.
5. Aplicar el modelo predictivo “Cross-Validation” sobre el dataset: Dividir el conjunto de datos en “training” y “testing” para entrenar con un porcentaje de los datos, y probar con el resto de manera de obtener el porcentaje de efectividad de los resultados. Para esta división, utilizar el mes asociados al registro (fila de la tabla). Es decir, “training” si el mes es febrero, marzo o abril, y “testing” si es mayo y junio.

b. Herramientas

Las herramientas que se utilizarán para aplicar las técnicas anteriormente mencionadas son:

1. SQL Management Tools para ejecutar consultas básicas explotarias en la base de datos principal de la cual se obtendrán los datos. Además, los resultados parciales se guardarán en un SQL Server, al cual se accederá también con esta misma herramienta.
2. IBM SPSS Modeler para recuperar datos de diferentes orígenes, y armar el dataset preliminar
3. Dashboard Databricks para graficar el análisis de los campos que contienen el dataset
4. Notebook Jupiter en IBM Cloud, lenguaje Scala para el resto de las tareas:

- a. Analizar el conjunto de datos
- b. Aplicar los agregados y las conversiones de datos que sean necesarias.
- c. Procesamiento de los modelos predictivos
- d. Evaluación de los resultados

9. Desarrollo de la investigación

Armado del conjunto de datos

Para armar el dataset final se comenzó armando el conjunto de positivos, es decir clientes que compraron nuevos plazos fijos en pesos:

1. Personas físicas que correspondan a clientes de la entidad bancaria, los cuales hayan comprado un plazo fijo en pesos durante el período de Febrero a Junio del 2019.

Código de Cliente	Compró PF
12345	1
45677	1

2. A cada cliente se le asoció la sumatoria del saldo en pesos de la Caja de Ahorro y de la Cuenta Corriente que tuvo de manera diaria entre Febrero y Junio del 2019. Día por día, mientras el mismo haya sido uno hábil.

Código de Cliente	Fecha	Saldo del día	Compró PF (ese día)
12345	01/02/2019	\$ 100.000	0
12345	04/02/2019	\$ 120.000	1
...	...	...	...
45677	27/06/2019	\$ 30.000	0
45677	28/06/2019	\$ 15.000	1

3. Se propone utilizar Febrero, Marzo y Abril para realizar el training del modelo, mientras que Mayo y Junio para el testing. Dado esto, se agregó la variable adicional “tipo_registro” para dividirlos en los dos conjuntos.

Código de Cliente	Fecha	Saldo del día	Compró PF (ese día)	Tipo de Registro
12345	01/02/2019	\$ 100.000	0	Training
12345	04/02/2019	\$ 120.000	1	Training
...	...	...	...	...
45677	27/06/2019	\$ 30.000	0	Test
45677	28/06/2019	\$ 15.000	1	Test

4. Luego se excluyó a los clientes que **el mismo día que compraron un plazo fijo**, se produjo el vencimiento de uno anterior (en pesos o en dólares). Sólo se incluirán como positivos a aquellos clientes que no hayan comprado un plazo fijo a partir de uno anterior (compra conocida como “Renovación de Plazo Fijo”).

Armado del conjunto negativo:

Una vez armado el set de datos del conjunto positivos, para armar el de los negativos, se recuperaron clientes que no pertenecen al conjunto anterior. Sin embargo, los clientes en este grupo son más de 3.000.000, comparado a los 150.000 del conjunto positivo, la distribución está claramente desbalanceada. Para que esta situación no perjudique los resultados del modelo, se seleccionó de manera aleatorio 150.000 clientes negativos de los 3.000.000.

Tanto a los clientes positivos como negativos, se los enriqueció con los siguientes datos:

1. Como se detalló en el punto 2, cada registro corresponde a una fecha en particular del calendario. A partir de la fecha, se creó la variable Día de la semana, para identificar si es un Lunes, Martes, Miércoles, Jueves o Viernes.
2. También se crearon las variables Cotización del dólar, Tasa Leliq, y Riesgo País.
 - a. Tanto para estas variables como para el saldo, se crearon variables adicionales para identificar que ocurrió con respecto a uno y a dos días anteriores.
3. Otros atributos asociados a la entidad cliente:
 - a. Edad
 - b. Provincia
 - c. Flag Jubilado
 - d. Ingresos (Monto en pesos)
 - e. Cliente en situación de Mora
 - f. Flag cliente digital: Indica si el cliente utiliza los medios digitales
 - g. Score: Utilizado en la entidad bancaria que indica cuán valorable es el cliente para venderle productos
 - h. Flag posee atrasos en préstamos
 - i. Flag posee bonificaciones en la Tarjeta de Crédito y en Débito
 - j. Flag Tarjeta de crédito con resumen, que indica si el cliente utiliza la tarjeta además de contar con el plástico
 - k. Flag membresía Aerolíneas Plus. Algunos clientes asocian a sus consumos el programa de millas, mientras que otros poseen el por defecto en la entidad, compuesto por puntos a canjear por productos existentes en un catálogo interno
 - l. Flag consume en 1ras marcas
 - m. Flags de tenencia de productos: Seguros, Títulos, Cuota partes en Fondos de Inversión
 - n. Flag índice de conflictividad

Feature Engineering

Se crearon nuevas variables que suman valor al modelo, para ayudarlo a interpretar relaciones tanto utilizando la existente en los mismos registros, como también con respecto a fechas anteriores.

Las variables que se agregaron son:

- Variación del saldo en cuentas correspondiente a uno y dos días hábiles anteriores
- Flag si tenía o no suficiente plata el día anterior para la compra del plazo fijo

- Variación de la cotización del dólar, riesgo país y tasa LELIQ correspondiente a uno y dos días hábiles anteriores
- División del saldo en rangos:
 - 0: Menos de \$ 300
 - 1: Entre \$ 300 y \$ 49.000
 - 2: Entre \$ 50.000 y \$ 99.999
 - 3: Entre \$ 100.000 y \$ 199.999
 - 4: Entre \$ 200.000 y \$ 499.999
 - 5: Más de \$ 500.000

Otras conversiones

- Dado que el modelo necesita que las variables de entrada sean de tipo numérico, se aplicaron adicionalmente las siguientes conversiones:
 - Valor numérico *outlier* o literal en campos con “nulos”
 - Sexo, “M” pasó a ser 0, mientras que “F” pasó a ser 1
 - Provincia, pasó a ser numérico utilizando la función `StringIndexer()`, la cual permite asignar un valor numérico de manera aleatoria a cada string diferente que exista en la columna del dataset:

provincia	provincia_index
SALTA	0.0
MISIONES	1.0
JUJUY	2.0
SANTA FE	3.0
CORDOBA	4.0
BUENOS AIRES	5.0
CAPITAL FEDERAL	6.0
MENDOZA	7.0
RIO NEGRO	8.0
CHUBUT	9.0
NEUQUEN	10.0
TIERRA DEL FUEGO	11.0
ENTRE RIOS	12.0
CORRIENTES	13.0
TUCUMAN	14.0
SANTA CRUZ	15.0

LA RIOJA	16.0
SANTIAGO DEL ESTERO	17.0
CATAMARCA	18.0
CHACO	19.0
LA PAMPA	20.0
SAN LUIS	21.0
SAN JUAN	22.0
FORMOSA	23.0
sin_datos	24.0

- Día de la semana, pasó a ser numérico utilizando la función StringIndexer(), la cual permite asignar un valor numérico de manera aleatoria a cada string diferente que exista en la columna del dataset:

dia_semana	dia_semana_index
Lunes	0.0
Viernes	1.0
Miercoles	2.0
Martes	3.0
Jueves	4.0

Análisis de la distribución de los datos

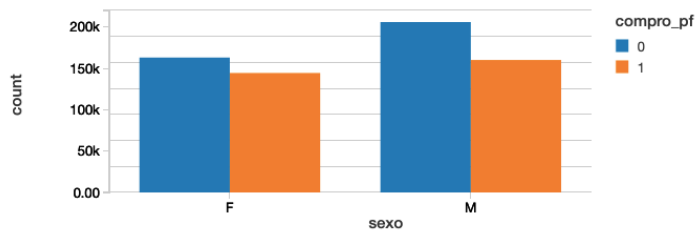
- Tabla resumen de cantidades del conjunto de datos a analizar:
 - Training: En la iteración 1 se incluyó a 369.698 registros, y para el segundo modelo se seleccionó de manera random a 30.000.
 - Testing: En la iteración 1 se incluyó a 303.785 registros, y para el segundo modelo se seleccionó de manera random a 10.000 registros

Datos generales			Training		Testing	
Iteración	Atributos	Registros	Positivos	Negativos	Positivos	Negativos
1	46	673.483	174.091	195.607	130.355	173.430
2	37	673.483	174.091	195.607	130.355	173.430

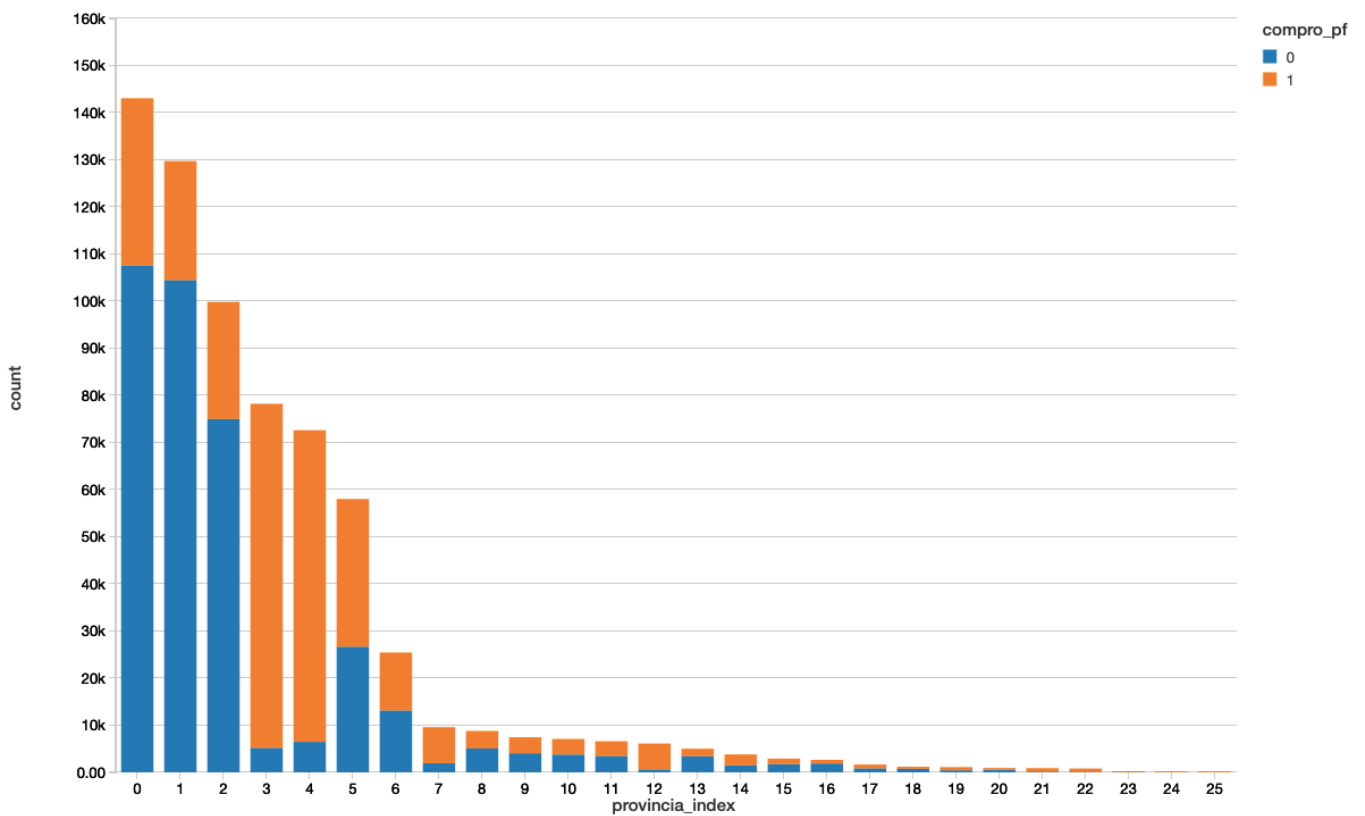
- Referencias de features
 - Se realizaron dos modelos, y en el definitivo se incluyeron las variables:

Número de Feature	Nombre de Variable
0	index
1	edad
2	flag_jubilado
3	ingresos
4	flag_cli_digital
5	score_cliente
6	tasa_leliq_bcra_hoy
7	var_leliq_bcra_0_1
8	var_leliq_bcra_1_2
9	valor_dolar_hoy
10	var_dolar_0_1
11	var_dolar_1_2
12	riesgo_pais_0
13	riesgo_pais_1
14	riesgo_pais_2
15	var_riesgopais_0_1
16	var_riesgopais_1_2
17	flag_atraso_prestamo
18	flag_bonificaciones_tc
19	flag_tc_con_resumen
20	flag_membresia_aa_plus
21	flag_cons_1ras_marcas
22	indice_confl_reclamos
23	flag_posee_bonif_td
24	flag_tc_consume_cuotas
25	flag_tc_consume_exterior
26	flag_tc_posee_adelantos
27	flag_tc_posee_consumos
28	flag_tenecia_seguro
29	flag_tenencia_titulos
30	flag_fc_inversion
31	flag_venta_me
32	flag_compra_me
33	sexo_index
34	provincia_index
35	dia_semana_index
36	cliente_en_mora_flag

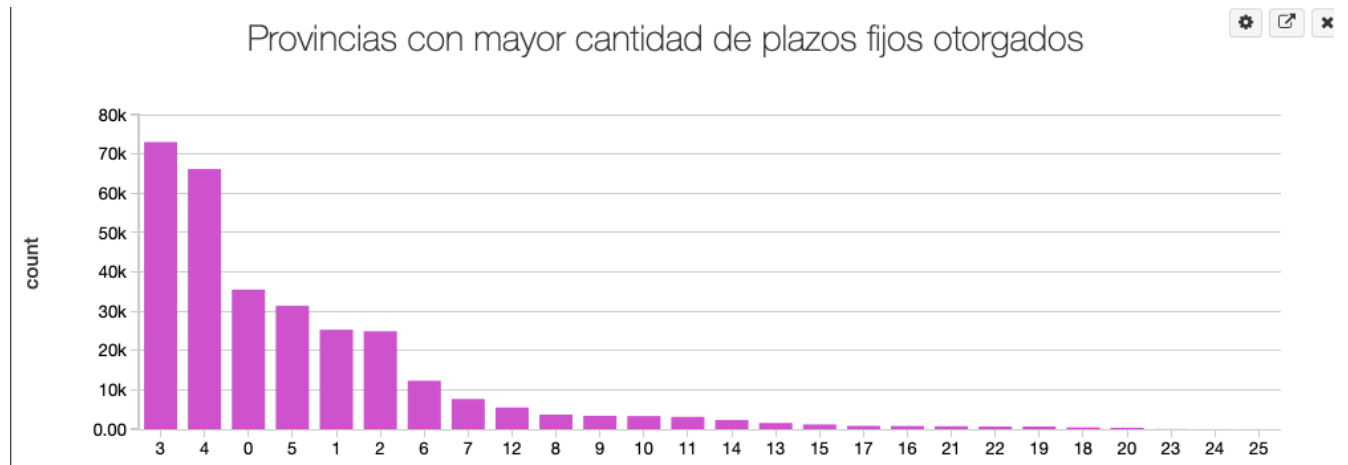
- **Sexo:** Cantidad de clientes distribuidos por sexo y por compra de Plazo Fijo.



- **Provincia:** Cantidad de clientes por provincia que posee el dataset, divididos por quienes sacaron y quienes no el Plazo Fijo.



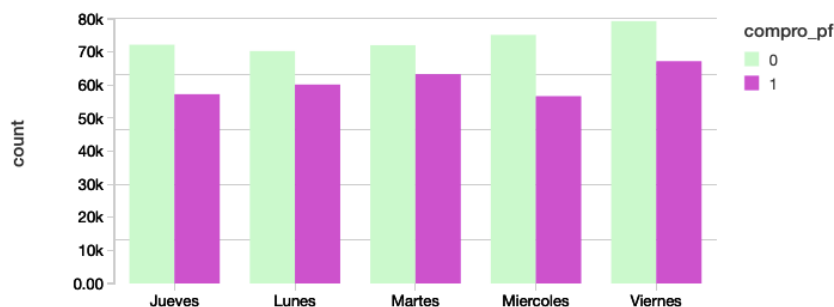
Las provincias que lideran el Ranking de Plazos Fijos son Santa Fe y Córdoba. En tercer lugar está Salta y recién en el cuarto, la provincia de Buenos Aires:



Referencia de provincias:

provincia_index	provincia	count
0	SALTA	143044
1	MISIONES	129721
2	JUJUY	99835
3	SANTA FE	78200
4	CORDOBA	72602
5	BUENOS AIRES	57994
6	CAPITAL FEDERAL	25428
7	MENDOZA	9577
8	RIO NEGRO	8785
9	CHUBUT	7453
10	NEUQUEN	7082
11	TIERRA DEL FUEGO	6597
12	ENTRE RIOS	6111
13	CORRIENTES	5002
14	TUCUMAN	3795
15	SANTA CRUZ	2919
16	LA RIOJA	2652
17	SANTIAGO DEL ESTERO	1659
18	CATAMARCA	1203
19	CHACO	1107
20	LA PAMPA	925
21	SAN LUIS	883
22	SAN JUAN	806
23	FORMOSA	86
24	sin_datos	15
25	NO DEFINIDA	2

- **Distribución con respecto a los días de la semana.** Cantidad de clientes de la muestra por día de la semana que sacaron y no sacaron un plazo fijo.

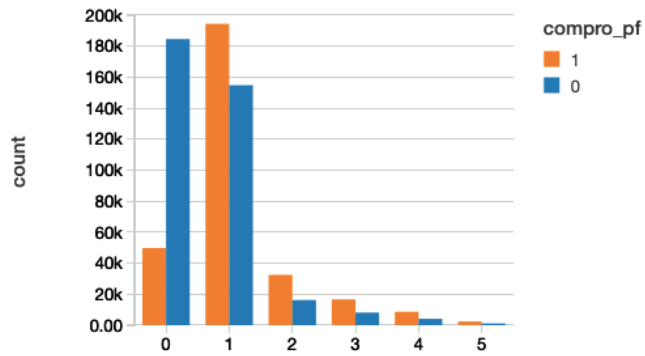


- **Rangos de saldo.** Se le asignó un rango del 0 al 5 a los clientes, en función al saldo que tenían el día de la compra y los días anteriores. El siguiente gráfico indica la cantidad de personas por rango de saldo, divididos por los que compraron y los que no.

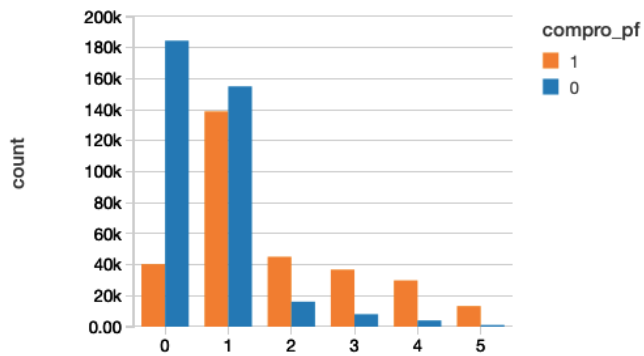
Rangos de saldo

1. Menos de \$300
2. Menos de \$50.000
3. Menos de \$100.000
4. Menos de \$200.000
5. Menos de \$500.000
6. Superioro igual a \$500.000

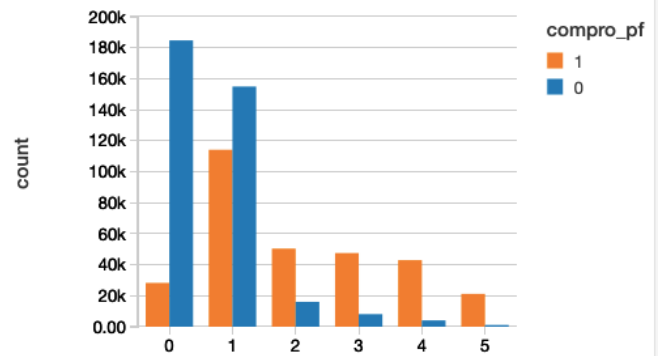
Día de la compra



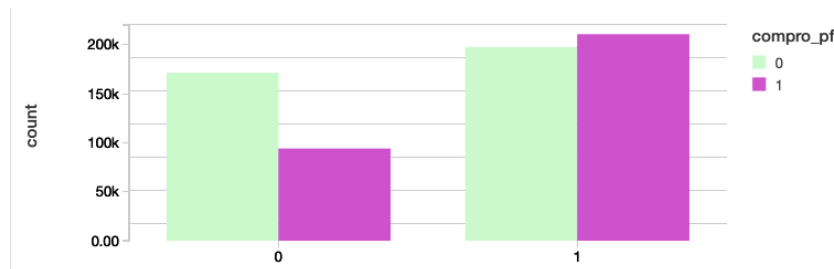
2 días anteriores a la compra



Día anterior a la compra



- **Distribución de la variable “flag_saldo_pf_1”.** Indica cantidad de clientes que el día anterior tenían plata para sacar el plazo fijo (1), o si no la tenían (0). Además se dividió por color, si el cliente además de tener o no la plata, compró o no un Plazo Fijo.



Modelo de predicción: Random Forest

Tal como detalla Nuria Victoria Castillo González (2015) [9], Random forest es una combinación de árboles de decisión en la que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos. Una ingresa al árbol y es sometida a una serie de test binarios en cada nodo (split) hasta llegar a una hoja, en la que se encuentra la respuesta. Por lo cual, esta técnica se puede pensar como una técnica para dividir un problema complejo en un conjunto de problemas simples. Particularmente en la fase de entrenamiento el algoritmo intenta optimizar los parámetros de las funciones de split a partir de las muestras de entrenamiento. Para esta fase, se utilizarán los meses de Febrero a Abril del 2019, mientras que luego, se probará la eficacia del modelo utilizando los meses Mayo y Junio del 2019.

Utilizar esta técnica, particularmente nos brinda las siguientes ventajas:

1. Se puede realizar tanto clasificación como regresión.
2. Se obtiene gran precisión en los resultados gracias a la generalización, en el cual muchos árboles ligeramente diferentes aportan información para obtener una medida más fiables.

El problema de determinar la propensión a adquirir un Plazo Fijo corresponde a uno de Clasificación, dado que esperamos como respuesta del modelo un 1 o un 0, lo cual indicaría:

- Predicción = 1, si el cliente es propenso a comprar
- Predicción = 0, si el cliente no es propenso a comprar

10. Resultados obtenidos

Para utilizar el algoritmo, todos los valores del set de datos deben ser numéricos, y agrupados en forma de vector, llamado “features”. Dentro de ese vector no se encuentra la variable clave para la predicción, en este caso “compro_pf”, sino que se encuentra en otra variable llamada “label”. Entonces, “features” y “label” de los meses de Febrero a Abril del 2019, ingresaron al modelo Random Forest a modo de “entrenamiento”.

Iteración 1 – Random Forest

Inicialmente, se utilizó un dataset de 46 variables para ser utilizado en el modelo Random Forest, el cual contenía las variables de saldos de los clientes, y su relación de como fue variando el mismo los días previos a la compra o no de un plazo fijo.

Esto hizo que el modelo aprenda de manera muy satisfactoria la situación de si un cliente tiene saldo alto en su cuenta hay un alto grado de probabilidad de que saque un plazo fijo, dejando atrás otros aprendizajes tal vez más profundos acerca de los clientes que se podrían llegar a encontrar.

Utilizando este primer modelo se obtuvo un porcentaje de aciertos igual al 92,5%, destacando las siguientes variables como principales:

Resultado predicción 1: 92.5% Accuracy

```
In [59]: val testData = dfInputTest.cache()

In [60]: val predictions = model.transform(testData)
val evaluator = new BinaryClassificationEvaluator().setLabelCol("label")
val accuracy = evaluator.evaluate(predictions)
println(f"Accuracy: $accuracy%2.3f")

Accuracy: 0.925

In [61]: printPredictionMetrics(predictions)

MSE:      0.14031963395164343
MAE:      0.14031963395164343
RMSE Squared: 0.37459262399524557
R Squared: 0.4272050547421661
Exp. Variance: 0.2409477940272644
```

Names	Values
saldo_ayer	0.20432079035055425
provincia_index	0.19856300122185672
saldo_ayer_rango	0.151796079989188
flag_saldo_pf_1	0.09500916356654743
edad	0.07585728847237455
saldo_antesdeayer	0.07093078056868503
saldo_dia	0.06367111586746974
saldo_dia_rango	0.04612836494883382
saldo_antesdeayer_rango	0.026541379813923023
var_saldo_1_2	0.018743029635027456

Iteración 2 – Random Forest

Luego, se redujo la cantidad de variables a 36, eliminando puntualmente las relacionadas al saldo del cliente.

En esta nueva ejecución se utilizó la siguiente configuración para el modelo:

- Máxima profundidad del árbol (MaxDepth) = 10. Por ejemplo, profundidad 0 significa 1 nodo de hoja; profundidad 1 significa 1 nodo interno + 2 nodos hoja.
- Máxima cantidad de árboles para entrenar (NumTrees) = 60.
- Máxima cantidad de “bins” (MaxBins) = 100. Se refiere al al máximo número de bins usados para discretizar características continuas y para elegir como “split” por característica (feature) en cada nodo. Mayor número de bins otorgan mayor

granularidad. Debe ser mayor o igual a dos y mayor o igual al número de categorías en cualquier variable categórica.

Con esta nueva iteración se obtuvo un **porcentaje de aciertos igual a 87.18%**.

```
1 val predictions = model.transform(testData)
2 val evaluator = new BinaryClassificationEvaluator().setLabelCol("label")
3 val accuracy = evaluator.evaluate(predictions)
4 println(f"Accuracy: $accuracy%2.3f")
```

► (4) Spark Jobs

►  predictions: org.apache.spark.sql.DataFrame = [features: udt, label: double ... 3 more fields]

Accuracy: 0.872

El modelo pondera la importancia que tuvo cada variable para evaluar cada registro a qué clase correspondía.

10 variables más importantes utilizadas en las reglas de los árboles y su importancia asociada:

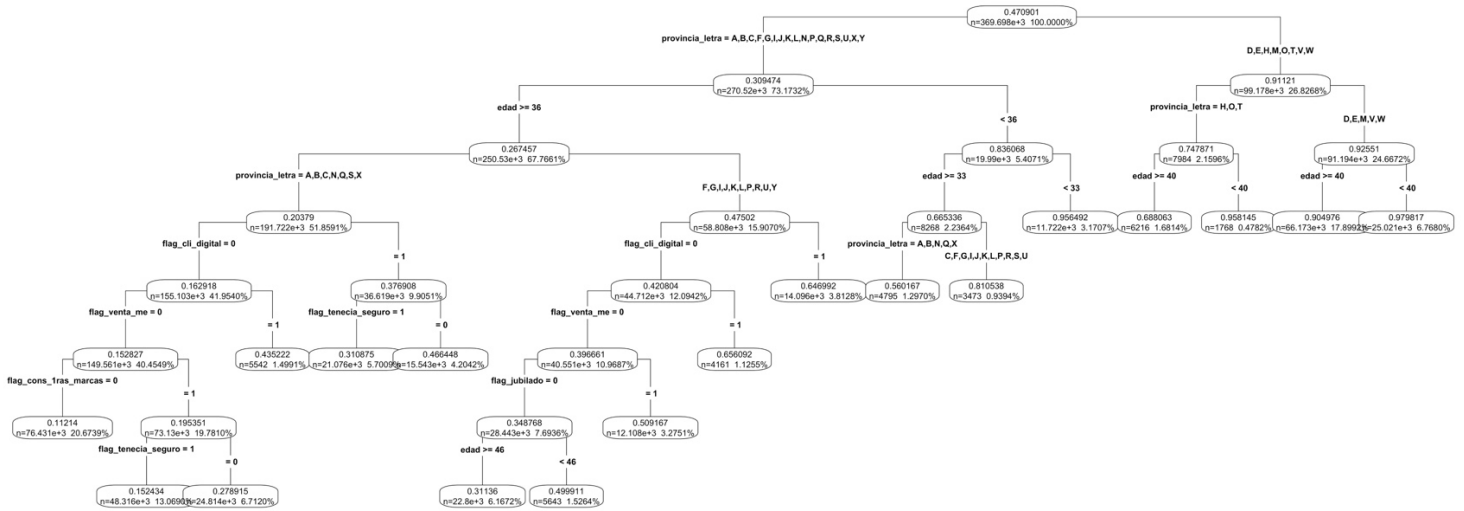
Names	Values
provincia_index	0.5530685357045759
edad	0.17506043873506275
flag_cli_digital	0.09117026526547664
flag_venta_me	0.03501117852925291
flag_cons_1ras_marcas	0.03194680616286941
ingresos	0.01998732081739423
flag_tenecia_seguro	0.014327604615048546
flag_fc_inversion	0.012229233017034552
flag_membresia_aa_plus	0.009464855031946792
flag_tc_consume_exterior	0.008245675393460246

A continuación, se visualiza un árbol de decisión del modelo, utilizando únicamente las variables que consideró más importantes:

Propensión a adquirir Plazos Fijos a través de un Banco tradicional

Tesis – Especialización en Ciencia de Datos - ITBA

Agustina Beatriz Sprio



Relación de provincias:

provincia	provincia_index	Provincia_
SALTA	0.0	A
MISIONES	1.0	B
JUJUY	2.0	C
SANTA FE	3.0	D
CORDOBA	4.0	E
BUENOS AIRES	5.0	F
CAPITAL FEDERAL	6.0	G
MENDOZA	7.0	H
RIO NEGRO	8.0	I
CHUBUT	9.0	J
NEUQUEN	10.0	K
TIERRA DEL FUEGO	11.0	L
ENTRE RIOS	12.0	M
CORRIENTES	13.0	N
TUCUMAN	14.0	O
SANTA CRUZ	15.0	P
LA RIOJA	16.0	Q

SANTIAGO DEL ESTERO	17.0	R
CATAMARCA	18.0	S
CHACO	19.0	T
LA PAMPA	20.0	U
SAN LUIS	21.0	V
SAN JUAN	22.0	W
FORMOSA	23.0	X
sin_datos	24.0	Y

En el árbol se observa cómo nivel a nivel selecciona los features principales que dividen a los clientes positivos de los negativos desde una raíz (variable más importante) hasta llegar a las hojas (menos importantes). La manera de leer el gráfico es la siguiente:

- La principal variable es la Provincia. En la raíz se indica que de el 100% de los clientes, el 47% es de la clase positiva.
- En el segundo nivel, el nodo de la derecha, muestra que, dentro de los clientes que radican en las provincias SANTA FE, CORDOBA, MENDOZA, TUCUMAN, CHACO, SAN LUIS Y SAN JUAN, el 91% sacaron un plazo fijo.
- De ese grupo, si se atomiza aún más, si la provincia es MENDOZA, TUCUMAN o CHACO, el porcentaje de positivos es 74%, mientras que si es del resto corresponden a un 92%
- Por último, en ambos casos, tienen mayor probabilidad de sacar un Plazo Fijo los que además cumplan con tener menos de 40 años de edad.

11. Conclusiones

A lo largo del desarrollo del proyecto, se ha logrado aceptar la hipótesis:

“La selección de clientes basada en modelos predictivos permite reducir la cantidad de comunicación de ofertas e incrementar la venta de Plazos Fijos en un banco tradicional”

El modelo seleccionado, denominado “Random Forest”, consiste en el armado de árboles de decisión que a través de las variables del conjunto de datos de entrada selecciona los clave al momento de evaluar la cantidad de clientes que sacaron o no un plazo fijo, y generando así una probabilidad de ocurrencia.

Principalmente, se realizaron dos modelos predictivos de compra de Plazo Fijos. En el primero se incluyeron 46 variables, de los cuales 9 estaban relacionadas al saldo que tuvieron una determinada fecha, hayan o no sacado un plazo fijo. Entre estas variables, se encontraban las que informaban el saldo de un día en particular, y el de uno y dos días anteriores. También se incluía la relación entre dichos saldos, y adicionalmente, cómo influía en si alcanzaba o no para la compra de cualquier plazo fijo. Este conjunto de datos aseguraba como muy poca probabilidad de error (un 92,5% de aciertos, es decir un 7,5% de error) que cualquier cliente

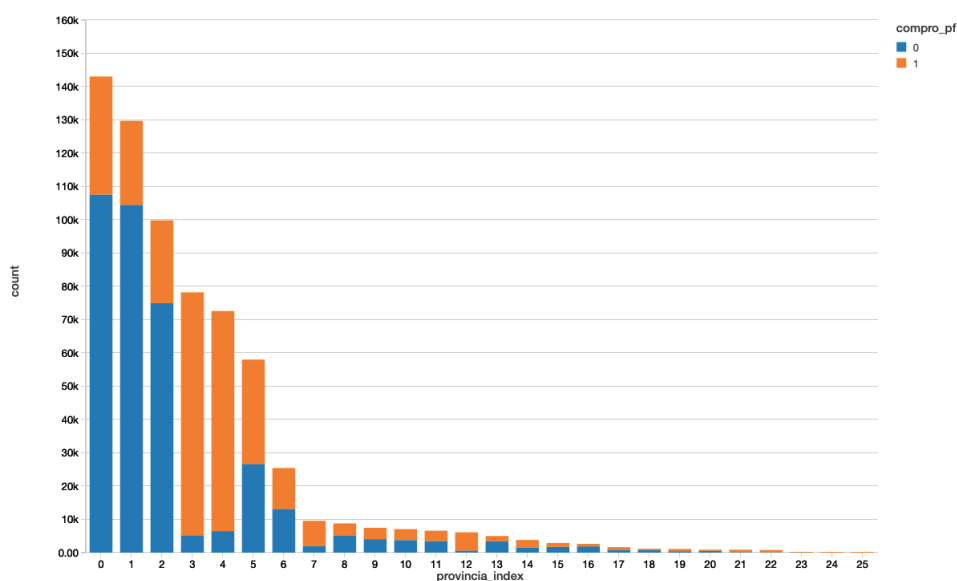
con saldo en su cuenta tiene alta probabilidad de compra de plazo fijo. Las variables relacionadas al saldo estaban muy correlacionadas con la clase positiva, por lo que impedía conocer otros insights acerca de los clientes de la entidad.

Con este escenario, se realizó un nuevo modelo, eliminando todas las variables de saldos anteriormente mencionadas, elaborando un conjunto de datos de 37 variables de entrada. Este segundo modelo asegura un **87,18% de aciertos** a la probabilidad de compra.

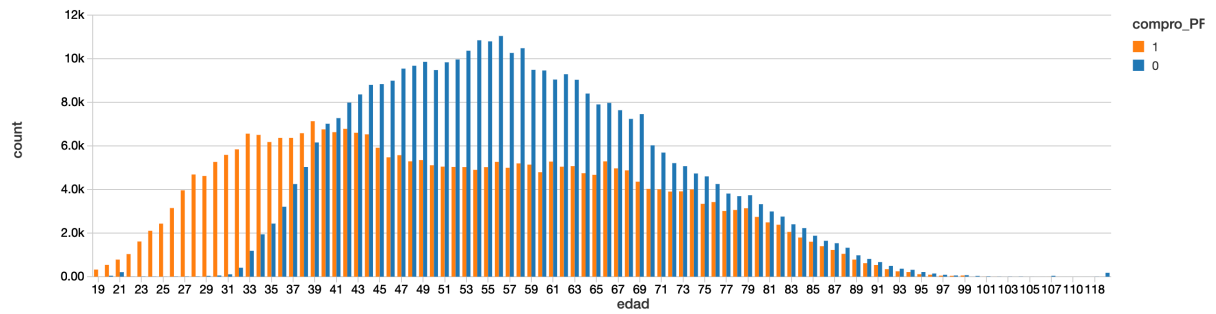
Las principales variables que utilizó el modelo para la toma de decisión resultaron ser las inherentes a los clientes, y no las asociadas a un contexto país particular, como lo son la provincia en donde el cliente radica, su edad, si utiliza los medios digitales para operar con sus cuentas bancarias, si posee un perfil inversor (compra moneda extranjera y cuota partes en fondos comunes de inversión), si consume con su tarjeta de crédito en las primeras marcas de Argentina, su nivel de ingresos, si posee seguros de cualquier tipo (auto, hogar, celular, entre otros) y si está asociado al programa Aerolíneas Plus para acumular millas en vez de puntos propios de la entidad.

Si analizamos la relación positivos/negativos de las principales variables del modelo, vemos que existe tal correlación:

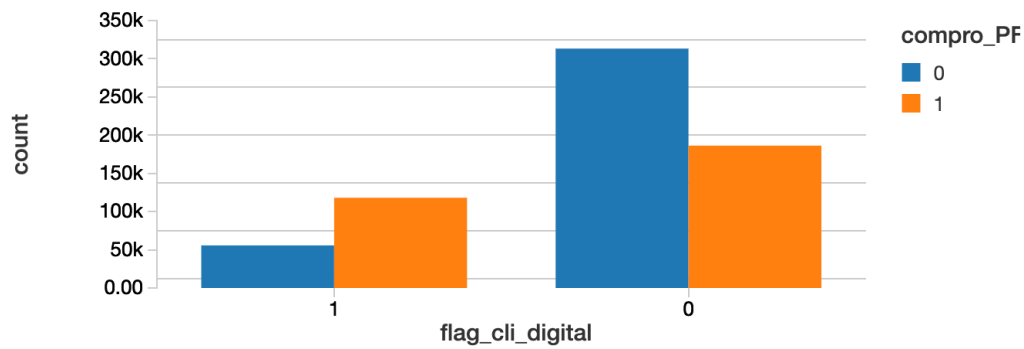
Provincia de radicación: Existen determinadas provincias en donde la mayoría de los clientes sacaron un plazo fijo, y provincias en donde ocurre totalmente lo contrario.



Edad: En el análisis por edad, se interpreta que la mayoría de los clientes entre 18 y 35 años sacaron un plazo fijo entre febrero y abril del 2019. A partir de los 36 años, empieza a emparejarse la distribución hasta los 40 que se invierte totalmente. A partir de los 40 años, la cantidad de clientes que no saca un plazo fijo es mayor a la que sí, y a pesar de esta relación, compran menos plazos fijos que los de edad más joven.



Flag cliente digital: Esta marca indica si el cliente opera por canales digitales. Hoy en día, es un punto clave el ingreso por parte de los clientes al Home Banking de la entidad bancaria desde dispositivos Desktop y Mobile para realizar la compra de Plazos Fijos directamente desde allí, en vez de hacerlo a través de las sucursales físicas. En el gráfico se muestra que la mayoría de los clientes que poseen clave digital sacaron un plazo fijo, mientras que los que no la tienen, en su mayoría no compraron uno.



Referencias Bibliográficas

[1] Atucha, A. J., & Gualdoni, P. (2018) [En línea]. Los gastos planeados en una economía: el consumo y la inversión.

<http://nulan.mdp.edu.ar/2899/1/atucha-gualdoni-2018.pdf>

[2] Fisher, S., Dornbusch, R., & Schmalensee, R. (1996). Economía. Ed. McGraw

[3] Bernake, B. & Frank, R. (2007). Principios de Economía. Ed. McGraw Hill

[4] Heilbronner, R. & Thurow, L. (1987). Economía. Séptima edición. Ed. Prentice Hal

[5] Khir, K., Gupta, L., & Shanmugam, B. (2007). Islamic banking: A practical perspective. Institut Bank-Bank Malaysia.

[6] Q.R. Zhuang, Y.W. Yao, y O. Liu (2018) [En línea]. Application of Data Mining in Term Deposit Marketing

http://www.iaeng.org/publication/IMECS2018/IMECS2018_pp707-710.pdf

[7] James Chen y Chris B Murphy (2019), [En línea]. Term Deposit

<https://www.investopedia.com/terms/t/termdeposit.asp>

[8] Daniel Artana (2018), [En línea]. Cinco motivos de por qué sube el dólar en Argentina

<https://www.perfil.com/noticias/economia/los-5-motivos-por-los-que-suba-el-dolar.phtml>

[9] (2015), [En línea]. Técnicas de Machine Learning para el Post-Proceso de la predicción de la Irradiancia.

http://masteres.ugr.es/moea/pages/curso201415/tfm1415/tfm_castillo_gonzalez_nuriav/!