

Trabajo Práctico N°2

Aprendizaje Automático

Detección del perfil de riesgo crediticio de clientes tomadores de préstamos personales en una entidad financiera con métodos de ensamble

Comisión II

Grupo N°: 18

MORENO, Federico Nicolás

PERINI, Sofía Clara

PICCHETTI, Bianca

Buenos Aires, 26 de Julio de 2019

Facultad de Ciencias Exactas y Naturales
Maestría en Explotación de Datos y Descubrimiento del Conocimiento
Aprendizaje Automático

Resumen

Se presentan en el siguiente informe los resultados obtenidos durante el proceso de desarrollo de un modelo predictivo para una entidad financiera, que busca clasificar a sus potenciales clientes tomadores de préstamos en “cumplidores” o “deudores” de acuerdo a determinadas características personales, condiciones del crédito e información de contacto brindada. Para el aprendizaje automático de los modelos empleados se utiliza un conjunto de datos que consta de ejemplos de clientes de ambas clases.

Luego de efectuar pruebas con diversos algoritmos de ensamble, se obtuvo el más adecuado para clasificar el perfil del cliente: *RandomForest*, con árboles de decisión de altura 8 calculado con *InformationGain*, realizando *bootstrap* para el muestreo y un valor máximo para la separación de variables igual a la raíz cuadrada del total. Fue el que mejor desempeño demostró en el conjunto de test.

Basándose en este mejor método se pudo concluir que la variable con más peso a la hora de clasificar el comportamiento de pago de un cliente es la cantidad de cuotas del préstamo. Le siguen en importancia la edad y el ingreso formal, y posteriormente el importe de la cuota y el interés mensual. En concordancia con lo observado en la investigación previa, no se observa un gran peso para la clasificación de las características personales del cliente, como género, puesto laboral o información de contacto.

Las medidas de performance obtenidas del mejor método de ensamble (*RandomForest*) fueron menores a las obtenidas con el mejor modelo del estudio anterior: un único árbol de decisión. Sumado a la fácil interpretabilidad de este último, resulta más interesante continuar empleando árboles de decisión para encontrar variables que expliquen el comportamiento deudor o riesgoso de un cliente, y emplear métodos de ensambles para contrastar los resultados de los árboles y tener más solidez en los hallazgos. Por último, se proponen algunas líneas de investigación a futuro adicionales a las propuestas previamente, que incluyen una búsqueda exhaustiva de parámetros para los métodos de ensamble, así como la implementación de métodos divisivos (no abordados en el presente trabajo) para brindar solidez a las predicciones.

1. Introducción

Uno de los principales inconvenientes que suelen enfrentar las empresas financieras a la hora de otorgar préstamos a sus clientes es el riesgo crediticio. Este último consiste en “el riesgo de pérdida de capital causado por la falta de pago en tiempo y forma por parte de un acreedor” (*Econlink*¹, 2014). Este incumplimiento de las obligaciones crediticias, ya sea por falta total de pago, pago parcial o pago fuera de término, es una preocupación fundamental de las entidades financieras dado que ocasiona pérdidas indeseadas. Es por ello, que habitualmente los bancos y demás instituciones financieras suelen hacer exhaustivos análisis de sus clientes y su capacidad de pago al momento de otorgar créditos y definir los montos a otorgar a cada uno.

1.1. Objetivos

En este sentido, se propone como objetivo del presente trabajo identificar el perfil de cliente deudor o cumplidor de una pequeña empresa financiera, contrastando los resultados obtenidos en la

¹Econlink (06 de Ene de 2014). “El Riesgo Crediticio”. [en línea] Dirección URL: <https://www.econlink.com.ar/riesgo-crediticio> (Consultado el 30 de Abr de 2019)

investigación anterior (*TP1*). Se intentará predecir si un cliente supone un riesgo crediticio (deudor) o no para dicha empresa, mediante el empleo de técnicas de ensamble de modelos.

La predicción del perfil de cliente (deudor o cumplidor) será la función objetivo que guiará esta investigación. Un cliente² se considera deudor cuando se encuentra en situación de mora. Esta última se define como el retraso o la falta de pago con respecto a las obligaciones fijadas contractualmente. De acuerdo con la política de la empresa, se considera retraso a partir de los 160 días del vencimiento de una cuota.

De tal forma, el trabajo se encuentra dividido en tres secciones; la primera corresponde a la presente introducción e incluye el objetivo del trabajo, la preparación de los datos y la metodología de investigación. Se determina lo que se va a investigar, incluyendo una serie de hipótesis acerca de los atributos que orientarán el análisis a los efectos de encontrar los determinantes del riesgo crediticio (cliente deudor).

En la segunda sección, se detallarán los principales resultados de la investigación, producto de la aplicación de las técnicas indicadas en la metodología, comparando los modelos empleados con distintas métricas para lograr el objetivo buscado.

En la última parte, se hará hincapié en las principales conclusiones, corroborando el cumplimiento (o no) de las hipótesis planteadas sobre los atributos a fin de contribuir a discernir cuáles son los determinantes del riesgo crediticio que ayudaron a entrenar el modelo para predecir los perfiles de clientes deudores.

En este sentido, se plantean las siguientes hipótesis de investigación de acuerdo con los resultados obtenidos en la investigación previa, que se intentarán corroborar a lo largo del presente trabajo:

- El atributo que más separa los casos de estudio para su clasificación en deudores o pagadores es la cantidad de cuotas.
- También resultan importantes para explicar el target, el interés mensual y la edad de los clientes.
- Las características personales del cliente, como género, puesto laboral o información de contacto no tendrían gran peso en la determinación del cliente riesgoso o no riesgoso.
- El ingreso formal y la zona de residencia podrían brindar información sobre el comportamiento de pago.

1.2. Selección y preparación de Datos

Los análisis llevados a cabo en la presente investigación se realizan a partir de una base de datos de alrededor de 800 observaciones (casos, ejemplos) correspondientes a préstamos otorgados a clientes durante el período 2016-2018, propiedad de la empresa financiera objeto de estudio. Cabe aclarar que los clientes de la base de datos se clasifican como deudores si cumplen la condición de morosidad (citada previamente) a abril del 2019.

El *dataset* con el que se trabajó se encuentra ya transformado y preprocesado. El original contaba con 40 atributos numéricos y categóricos, correspondientes a información sobre las condiciones del préstamo (monto, cuotas, interés, fecha) y características del cliente (ingresos, edad, sexo, zona de residencia, puesto laboral, información de contacto), entre otros. No obstante, previo al procesamiento de la base, debieron efectuarse una serie de transformaciones, a fin de preparar/accondicionar los da-

² Según la política de la empresa financiera objeto de estudio, los clientes corresponden a personas que trabajan en relación de dependencia con más de un año de antigüedad y que al momento de requerir un préstamo personal no poseen embargos judiciales ni informes en el veraz. Es decir, la entidad toma clientes que al momento de otorgarles crédito no presentan mora con otras instituciones financieras. Todos ellos deben pasar por un filtro de solvencia y la cuota a otorgar no puede superar el 30% del sueldo.

tos para el análisis, entre los que cabe mencionar: anonimización de datos, ingeniería de atributos, imputación de faltantes y balanceo de datos(para más información sobre el preprocesamiento realizado, ver TP1).

El *dataset* final que se utilizó para la experimentación con distintos tipos de ensambles, cuenta con 806 casos correspondientes a 18 atributos, de los cuales exactamente la mitad (ejemplos) corresponden a cada categoría del *target*: deudor y pagador.

1.3. Metodología de investigación

De acuerdo con lo detallado previamente, se dispone de información de la calificación de clientes cumplidores y deudores (no riesgosos y riesgosos), en función de si han cumplido sus obligaciones crediticias o no, según la definición de mora previamente mencionada. Con estos datos, se puede emplear un método de aprendizaje automático supervisado para la inferencia inductiva de aquellos que potencialmente podrían ser riesgosos. La construcción de los modelos se realiza en un contexto de aprendizaje supervisado ya que se dispone de un conjunto de datos anotados donde se conocen tanto los datos como su clasificación.

En la primera investigación se habían empleado *Árboles de Decisión* y *Naive Bayes* como clasificadores, experimentando en el primer caso con distintos hiperparámetros, y utilizando diversas técnicas (*grid search* y *random search*) para obtener las mejores condiciones del árbol de decisión. En contraste, en esta oportunidad se utilizaron tres métodos distintos de ensambles planos que permiten analizar el caso: *RandomForest*, *AdaBoost* y *XGBoost*.

Los métodos de ensamble consisten en el uso de un conjunto de modelos que sirven para construir un meta-modelo, combinando sus estimaciones (por votación o ponderaciones). La idea de base de estos modelos es usar conocimiento de distintas fuentes al tomar decisiones. En particular, se seleccionaron ensambles planos, lo que tiene por detrás la intuición del “Comité de expertos” donde se reúnen expertos en la misma temática y toman decisiones. Por un lado, *RandomForest* propone una mejora al *bagging*; esta técnica consiste en aplicar repetidas veces el mismo modelo con diferentes sets de datos -con el fin de disminuir la varianza- y, en el caso de clasificación, seleccionar por votación. Es particularmente útil para clasificadores inestables como los árboles de decisión con los que trabajaremos en la presente investigación.

Por otro lado, los modelos *AdaBoost* y *XGBoost* son métodos de *boosting*, en los cuales cada modelo se arma utilizando información del modelo anterior. El primero, *AdaBoost*, lo hace asignando pesos en cada iteración, dando mayor importancia a los casos mal clasificados. En cambio, el segundo, *XGBoost* utiliza el error de clasificación como valor esperado en el modelo siguiente. Estos últimos es recomendable usarlos para clasificadores simples y estables como podría ser *Naive Bayes*.

De tal modo, luego de particionar los datos en conjuntos de entrenamiento, validación y test, se emplearon técnicas de *grid* y *random search* en la etapa de experimentación con distintos hiperparámetros a fin de encontrar los mejores modelos clasificadores del *target*. Una vez hallados los mejores de cada método de ensamble, se procedió a su contrastación para definir el de mejor desempeño. Las proporciones que se utilizaron para particionar el conjunto de desarrollo y test, y luego entrenamiento y validación fueron 80% y 20%, respectivamente.

Los métodos fueron implementados principalmente con árboles de decisión, por lo que, en la mayoría de los casos, se utilizaron los parámetros del mejor árbol obtenidos en el estudio anterior (TP1) como punto de partida. Dada la cantidad de parámetros con los que cuenta cada uno de estos métodos, se ajustó una cantidad limitada de ellos. De cada método se seleccionó el conjunto de parámetros de mejor *performance* con los datos de validación. El detalle de la búsqueda de parámetros para cada uno se encuentra en la Sección 2.1.

Una vez finalizada la etapa de experimentación y encontrados los mejores parámetros de cada método para aproximar la función objetivo, se procedió al testeo de los mismos con el conjunto de test.

Cabe destacar que, para todos los modelos entrenados, se calcularon las siguientes métricas: verdaderos positivos (*true positive*, TP), falsos positivos (*false positive*, FP), verdaderos negativos (*true negative*, TN) y falsos negativos (*false negative*, FN), como también *Accuracy*, *Precision*, *Recall* y F1. Esta última medida fue la utilizada para evaluar la performance con *5-fold cross-validation*, ya que es la media armónica entre *precision* y *recall* permitiendo una contribución relativa equitativa de ambas.

2. Resultados

2.1. Etapa de Experimentación

Con el objeto de avanzar en la selección del mejor algoritmo para clasificar el *target* (perfil del cliente), se detallan a continuación algunos de los principales resultados del proceso de experimentación con distintos modelos y parámetros, los cuales fueron evaluados con el conjunto de validación.

RandomForest

Para encontrar los hiperparámetros óptimos de *RandomForest* se partió del mejor árbol obtenido en el estudio anterior (*TP1*): un árbol de profundidad 8 con *Information Gain* como criterio. A partir de ello, se generaron 3 alternativas distintas; primero, se realizó un clasificador con estos parámetros fijos, utilizando los demás por defecto³ (*Tabla 1, RF1*).

Luego, si bien el modelo cuenta con varios parámetros para ajustar, se decidió encontrar los valores óptimos de *bootstrap* -dado que contamos con un *dataset* pequeño la aplicación de esta técnica puede resultar favorable-, cantidad de árboles y el número máximo de variables a considerar para la separación (*max_features*). En este sentido, se efectuó una búsqueda *random* fijando la profundidad del árbol (*max_depth=8*) y el criterio (*Information Gain*), variando *bootstrap* (*true*, *false*), *max_features* (raíz cuadrada, logaritmo base dos o máximo) y cantidad de árboles (entre 10 y 1.100). Esta búsqueda se realizó repetidas veces utilizando *f1-score 5-folds cross validation*. En todos los casos se obtuvieron como mejores parámetros: *bootstrap* y raíz cuadrada en cantidad de variables, entrenando así un segundo modelo (*Tabla 1, RF2*).

A partir de estos resultados, se generó un tercer clasificador fijando dichos parámetros con los mejores resultados obtenidos y realizando una búsqueda *Grid* para la cantidad de árboles, obteniendo 13 como el mejor valor (*Tabla 1, RF3*).

De los tres modelos entrenados, aquel que mejor performance presentó, analizando todas las métricas calculadas, fue el segundo: de altura máxima 8 (con *Information Gain*), raíz cuadrada en cantidad de variables y *Bootstrap true*, con 800 estimadores. Los resultados de las métricas obtenidas con estos tres modelos se pueden ver en la *Tabla 1*.

Adicionalmente, se calculó para este modelo la importancia de cada variable en el proceso de clasificación de este modelo (*Figura A1* en Anexo). Se observa que la de mayor importancia es “Ingreso Formal”, siguiendo “Importe Cuota”, “Edad” y “Cantidad de Cuotas”, las tres con valores cercanos.

³<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>

Parámetro/métrica	RF1	RF2	RF3
<i>Criterion</i>	<i>Entropy</i>	<i>Entropy</i>	<i>Entropy</i>
<i>Max_depth</i>	8	8	8
<i>Max_features</i>	<i>None</i>	<i>Sqrt</i>	<i>Sqrt</i>
<i>Bootstrap</i>	<i>True</i>	<i>True</i>	<i>True</i>
<i>N_estimators</i>	10	800	13
TP	49	47	49
FP	16	11	15
TN	55	60	56
FN	9	11	9
<i>F1 5-folds CV</i>	0.611	0.642	0.629
<i>Accuracy</i>	0.806	0.829	0.814
<i>Recall</i>	0.775	0.845	0.789
<i>Precision</i>	0.859	0.845	0.862
<i>F1-Score</i>	0.815	0.845	0.824

Tabla 1: parámetros y medidas de performance obtenidas para *RandomForest* en Validación.

AdaBoost

En el caso del modelo *AdaBoost*, se efectuaron también tres modelos para encontrar la mejor combinación de hiperparámetros. El primero, consistió simplemente en entrenar el modelo con sus parámetros por defecto (Tabla 2, AB1). Luego, se efectuó *Grid Search* para el estimador del ensamble, probando con: *Naive Bayes*, un árbol de decisión con sus parámetros por defecto y el árbol de decisión obtenido en el estudio anterior. El mejor estimador encontrado con la técnica mencionada fue un árbol de profundidad 8 con *Information Gain* como criterio y raíz cuadrada como máxima cantidad de variables de separación (Tabla 2, AB2). Por último, se repitió la búsqueda *Grid*, esta vez para la cantidad de estimadores (se buscó entre 5 y 500) y se obtuvo 198 como el mejor valor (Tabla 2, AB3).

De los tres modelos entrenados con el método de ensamble *AdaBoost* se observa que el de mejor desempeño fue AB3: el mismo árbol seleccionado para *RandomForest*, pero en este caso con 198 estimadores. Los valores obtenidos son muy cercanos a los de AB2, que emplea 50 estimadores. Estos resultados se pueden ver en la Tabla 2.

Parámetro/métrica	AB1	AB2	AB3
<i>Base_estimator</i>	<i>DecisionTreeClasifier</i>	<i>DecisionTreeClasifier</i>	<i>DecisionTreeClasifier</i>
<i>Criterion</i>	<i>Gini</i>	<i>Entropy</i>	<i>Entropy</i>
<i>Max_depth</i>	1	8	8
<i>Max_features</i>	<i>None</i>	<i>Sqrt</i>	<i>Sqrt</i>
<i>Learnnig_rate</i>	1.0	1.0	1.0
<i>N_estimators</i>	50	50	198
TP	40	51	50
FP	28	6	5
TN	43	65	66
FN	18	7	8
<i>F1 5-folds CV</i>	0.597	0.630	0.647
<i>Accuracy</i>	0.643	0.899	0.899
<i>Recall</i>	0.606	0.915	0.930
<i>Precision</i>	0.705	0.903	0.892
<i>F1-Score</i>	0.652	0.909	0.910

Tabla 2: parámetros y medidas de performance obtenidas para *AdaBoost* en Validación.

En la Figura A2 (en Anexo) se muestra la importancia de cada variable para la clasificación; se observa en este caso que la variable "Ingreso Formal" también es la de mayor importancia, pero a diferencia de RF2, la segunda en importancia es "Edad", presentando amplia diferencia con respecto a las siguientes.

XGBoost

Para el ensamble *XGBoost* se experimentó primero con todos sus parámetros por to⁴(*Tabla 3, XGB1*). Asimismo, se probó con un árbol de profundidad 8, *Information Gain* como criterio y raíz cuadrada como máxima cantidad de variables de separación (en este método este parámetro se llama *colsample_bytree*) (*Tabla 3, XGB2*). En este caso se realizaron diversas búsquedas *Grid*, ajustando los parámetros de a uno: *min_child_weight*, *gamma*, *learning_rate*, *n_estimators*, obteniendo los parámetros de la *Tabla, XGB3*.

El modelo *XGB3* fue el de mejor desempeño de los tres entrenados. Éste corresponde al mismo árbol de los modelos anteriores, con los parámetros específicos del algoritmo de ensamble definidos como 0.24 para la tasa de aprendizaje, 2 para *min_child_weight* (mínimo peso hessiano necesario para crear un nodo, controla la complejidad del modelo), 0.4 para *gamma* (valor mínimo necesario de la función de costo para generar una separación, también controla la complejidad del modelo) y 548 el número de estimadores. Esto se puede verificar en la *Tabla 3*.

Parámetro/métrica	XGB1	XGB2	XGB3
<i>Criterion</i>		<i>Entropy</i>	<i>Entropy</i>
<i>Max_depth</i>	3	8	8
<i>colsample_bytree</i>	1	0.2357	0.2357
<i>Min_child_weight</i>	1	1	2
<i>Gamma</i>	0	0	0.4
<i>Learnnig_rate</i>	0.1	0.1	0.24
<i>N_estimators</i>	100	100	548
TP	38	37	38
FP	34	29	24
TN	37	42	47
FN	20	21	20
<i>F1 5-folds CV</i>	0.614	0.628	0.640
<i>Accuracy</i>	0.581	0.612	0.659
<i>Recall</i>	0.521	0.592	0.662
<i>Precision</i>	0.649	0.667	0.701
<i>F1-Score</i>	0.578	0.627	0.681

Tabla 3: parámetros y medidas de performance obtenidas para XGBoost en Validación.

La variable más importante en este modelo es, al igual que los dos vistos anteriormente, “Ingreso Formal”, pero en contraste con ellos la importancia en las variables parece estar distribuida de forma más equitativa (*Figura A3*).

2.2. Etapa de Evaluación

Resultados para el conjunto de validación

Para cada método de ensamble se seleccionaron los mejores parámetros encontrados en la etapa de experimentación: RF2 para el método *RandomForest*, AB3 para el método *AdaBoost* y XGB3 para el método *XGBoost*. En la *Figura 1* se muestra la matriz de confusión para cada uno de ellos, al evaluarlos con el conjunto de validación.

En este caso, observando la performance en validación, el modelo *AdaBoost* resulta el que mejor desempeño muestra como clasificador del conjunto de entrenamiento tanto para la categoría de deudores como de pagadores. Las matrices de confusión mencionadas muestran estos resultados gráficamente. Se puede notar en todas las métricas analizadas un mejor desempeño de ese modelo,

⁴<https://xgboost.readthedocs.io/en/latest/parameter.html>

salvo cuando se lo entrena y evalúa efectuando *CrossValidation*, en este caso los tres métodos presentan valores cercanos (*Tabla 4*).

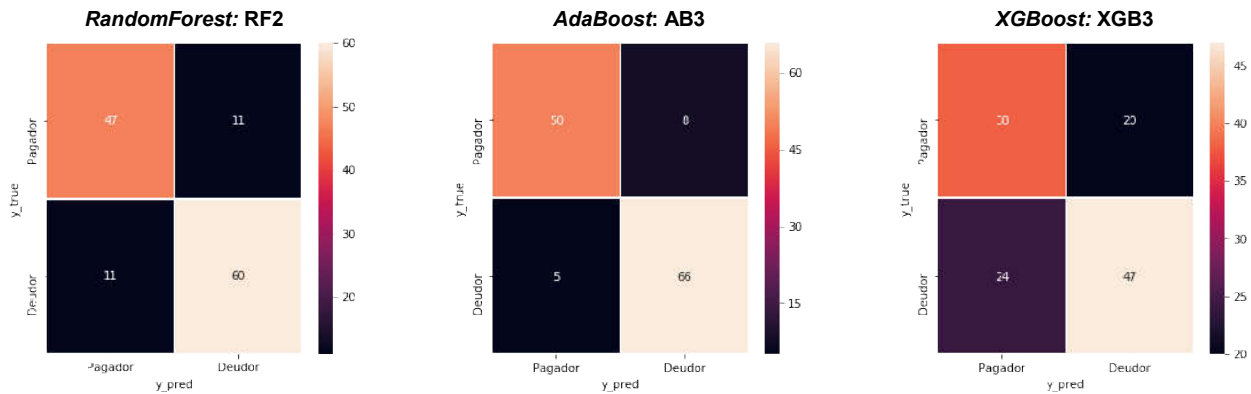


Figura 1: Matrices de confusión del conjunto validación de los modelos seleccionados.

Resultados con el conjunto de test

Luego de efectuadas las pruebas y seleccionados los modelos, se entrenaron los mejores con el conjunto de desarrollo y se evaluaron con el conjunto de test para su comparación (*Tabla 4*). A diferencia de los resultados obtenidos con el conjunto de validación en que el mejor ensamble con la mayoría de las métricas calculadas era *AdaBoost*, al entrenarlo con el conjunto de desarrollo y evaluarlo con el test, *RandomForest* aparece como el mejor.

En la *Figura 2* se pueden apreciar las matrices de confusión de los modelos evaluados con el conjunto de test. A diferencia de lo observado en validación, en este caso *RandomForest* clasifica mejor que el resto tanto los clientes pagadores como deudores.

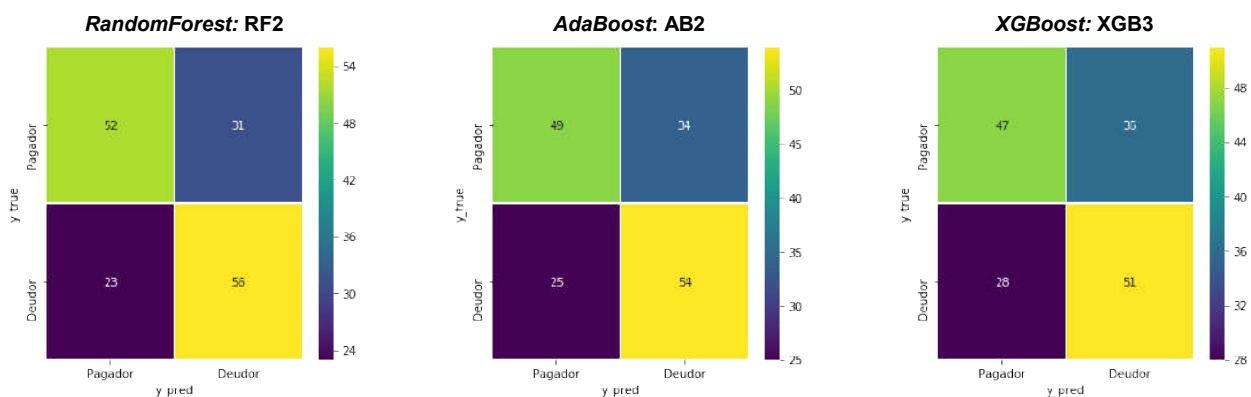


Figura 2: Matrices de confusión del conjunto Test de los modelos seleccionados.

El ensamble realizado con el método *XGBoost* presenta la peor *performance* de los tres modelos comparados, bajo cualquier métrica que se analice. *AdaBoost* si bien resultaba muy bueno para clasificar el conjunto de entrenamiento, no lo demostró con el conjunto de test. De esta forma, el modelo *RandomForest* presenta valores más altos en todas las métricas analizadas para el conjunto de test, es decir, que resulta el mejor modelo de los tres.

En este sentido, al analizar la importancia de cada variable para la clasificación del *target* con este modelo, se observa que el atributo más relevante es la "Cantidad de cuotas" del préstamo. Le

siguen en importancia la “edad” y el “ingreso formal” del cliente, y con menor importancia, el “importe de la cuota” y el “interés mensual” (Ver *Figura 3*).

Métrica	Random Forest	AdaBoost	XGBoost
Conjunto Validación			
F1 5-folds CV	0.642	0.647	0.640
Accuracy	0.829	0.899	0.659
Recall	0.845	0.930	0.662
Precision	0.845	0.892	0.701
F1-Score	0.845	0.910	0.681
Conjunto Test			
F1 5-folds CV	0.616	0.598	0.560
Accuracy	0.667	0.636	0.605
Recall	0.709	0.684	0.646
Precision	0.644	0.614	0.586
F1-Score	0.675	0.647	0.614

Tabla 4: Comparación de medidas de *performance* de cada ensamble con los conjuntos de validación y test.

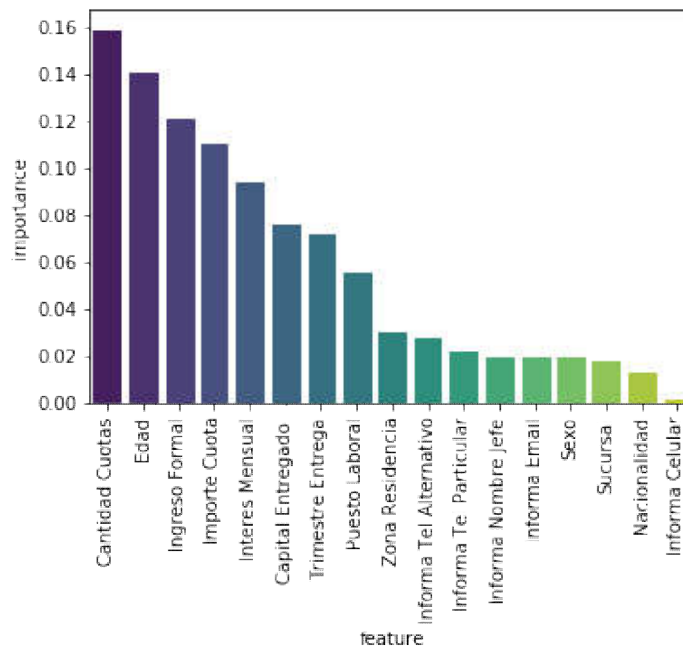


Figura 3: Importancia de variables para el mejor método de ensamble: *RandomForest*.

3. Conclusiones

Luego de la realización de diversos modelos de ensamble, se observa que el atributo más relevante para la clasificación en deudores o pagadores es la cantidad de cuotas del préstamo (según el ensamble de mejor desempeño: *RandomForest*). Le siguen en importancia la edad y el ingreso formal del cliente, y posteriormente, el importe de la cuota y en menor medida el interés mensual (Ver *Figura 3*).

Contrastando las hipótesis de investigación, se pudo corroborar que el atributo que más separa los casos de estudio para su clasificación en deudores o pagadores es la cantidad de cuotas, similar conclusión a la que se había arribado con árboles de decisión en la primera investigación (*TP1*). Por

otro lado, el interés mensual no tiene tanta relevancia como se habría observado en el estudio previo, no así la edad que se vuelve a erigir como *feature* importante para determinar la clasificación del perfil del cliente.

Asimismo, en coincidencia con lo observado previamente no se registra un gran peso para la clasificación de las características personales del cliente, como género, puesto laboral o información de contacto. En este caso parece tener más relevancia el puesto laboral que la zona de residencia, o incluso otras características personales. En cambio, la edad y el ingreso formal sí se corrobora que podrían brindar información sobre el comportamiento de pago del cliente.

Las medidas de *performance* obtenidas del mejor método en el conjunto de test (*RandomForest*) fueron menores a las obtenidas en el estudio anterior, utilizando un único árbol (*TP1*); el hecho de aumentar la cantidad de árboles disminuye la varianza, pero esto no necesariamente ocurre para el sesgo (dilema Bias-Varianza).

Si bien estos métodos de ensamble resultan interesantes y demuestran niveles de *performance* similares al árbol de decisión del *TP1*, no permiten gran interpretabilidad de los resultados para el problema de investigación. En este sentido, siguen siendo válidas las líneas de investigación a futuro planteadas en la investigación previa. Sería de particular interés agregar una búsqueda exhaustiva de parámetros para los métodos de ensamble, contemplando rangos más amplios y variando de a dos o más en simultáneo. Además, han quedado fuera de este estudio los métodos divisivos; podría ser una estrategia apropiada combinar, mediante el método de *stacking*, *RandomForest* con diferentes algoritmos de clasificación para intentar obtener medidas de *performance* que superen significativamente a las obtenidas con un solo árbol.

Parece resultar más interesante a efectos de encontrar variables que expliquen el comportamiento deudor o riesgoso de un cliente, continuar empleando árboles de decisión, en vez de estos modelos dado que tuvieron mejor *performance* y a su vez permitieron arribar a conclusiones más específicas. Quizás podrían emplearse ensambles para contrastar los resultados de los árboles para tener más solidez en los hallazgos. Sin embargo, parecería necesario tener un modelo que permita entender cómo afectan los atributos a la clasificación para poder sacar conclusiones que sean de mayor utilidad para la financiera en estudio.

El presente análisis permitió corroborar cuáles son las principales variables que explican el comportamiento crediticio de los clientes, como también sentar bases para futuros análisis y metodologías más apropiadas para el problema de investigación. Se proponen algunas líneas de investigación a futuro, que implican análisis más específicos o más enfocados, que deberán ser evaluadas por la entidad financiera de acuerdo a sus intereses.

Anexo I. Figuras y Tablas

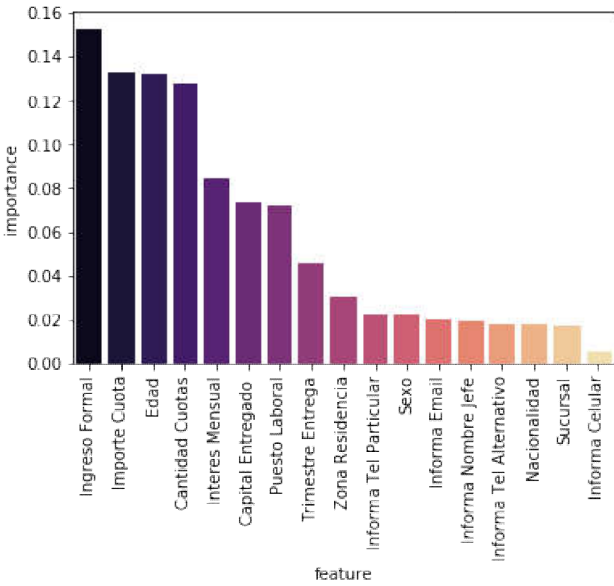


Figura A1: Importancia de variables para el método de ensamble RF2 en el conjunto de validación.

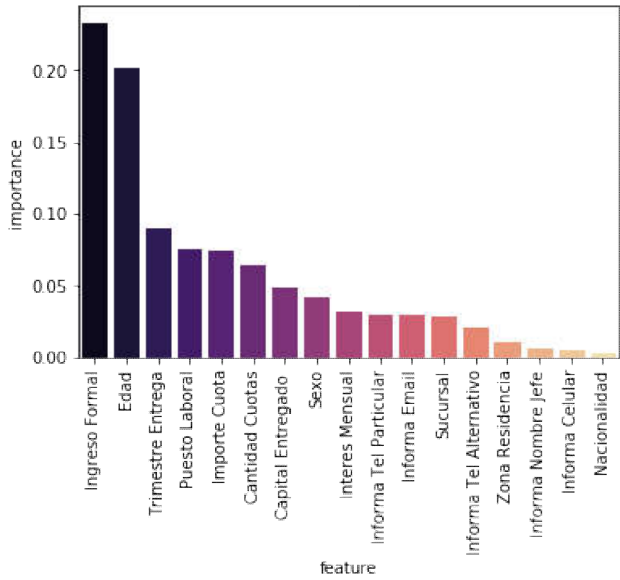


Figura A2: Importancia de variables para el método de ensamble AB2 en el conjunto de validación.

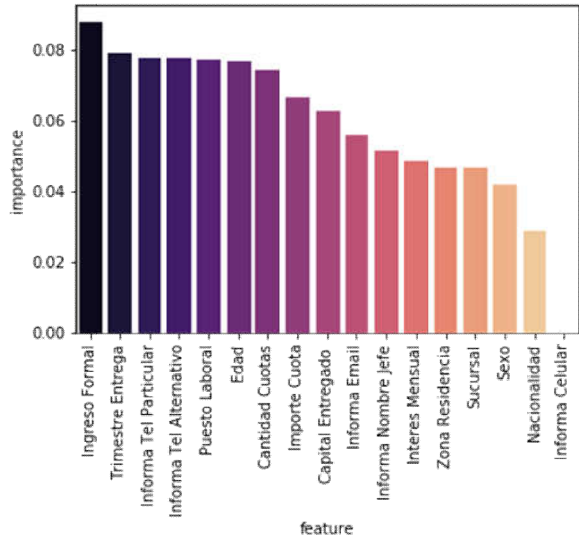


Figura A1: Importancia de variables para el método de ensamble XGB3 en el conjunto de validación

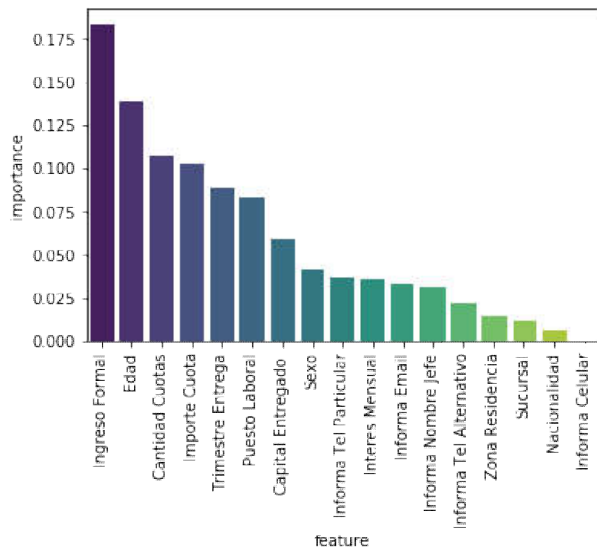


Figura A4: Importancia de variables para el método de ensamble AB2 en el conjunto de test.

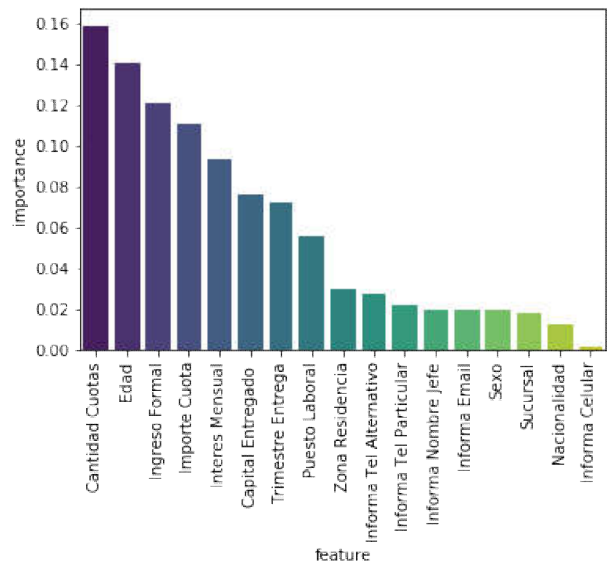


Figura A5: Importancia de variables para el método de ensamble XGB3 en el conjunto de test.