

Riesgo de crédito

Predicción de impagos,
importancia relativa de las características

Javier Eduardo Hernández Sánchez

2022

Introducción

El crédito es una figura económica que permite que un actor, llamado acreedor, proporcione dinero o recursos a otro actor, que se conoce como deudor. El deudor no devuelve los recursos inmediatamente sino que promete reembolsarlos en una fecha posterior pagando una cantidad adicional.

La palabra crédito proviene del término latino *credĭtum*, una sustantivación del verbo *creer*, esto indica, desde el principio, la relación de confianza que debe existir entre acreedor y deudor debido a la posibilidad de impago.

El impago ha sido tratado de formas muy diferentes, durante la mayor parte ha sido tratado como un crimen: el código de Hammurabi, por ejemplo, indica que debe ser tratado de forma similar al robo o fraude y castigado con la esclavitud.

Aun así la posibilidad de impago era alta, y algunos estamentos, como la aristocracia, podían simplemente repudiar un crédito.

Durante siglos el riesgo que comportaba un crédito se midió con base en la observación de los deudores y sus negocios. Durante el siglo XIX la construcción de los ferrocarriles en los estados unidos hizo que el escenario se tornase mucho más complejo, esto hizo que se creasen las primeras agencias de rating así como la prensa financiera especializada.

La información de que se disponía era principalmente cualitativa, al menos hasta que W. Braddock Hickman publicó un estudio sobre los bonos corporativos en los estados unidos entre 1953 y 1960. Y no fue hasta la década de 1980 cuando se empezó a estudiar el crédito al consumo.

Avanzando hasta años recientes vemos que Basilea II requiere que se calcule la tasa de probabilidad de incumplimiento y la pérdida dado el incumplimiento. Aun así el cálculo de las tasas de recuperación sigue siendo más arte que ciencia.

El presente trabajo se inscribe en este ámbito y se pretende usar técnicas de data science para predecir impagos y para descubrir que datos son más relevantes en este cometido.

Descripción de los datos

Los datos escogidos para este trabajo se denominan “default of credit card clients Data Set” que presenta los datos de 30.000 tarjetas de crédito en Taiwan referentes a 2005.

Variable	Tipo	Descripción
LIMIT_BAL	Numérico	Crédito concedido en dólares
SEX	Categorico	1: masculino, 2: femenino
EDUCATION	Categorico	1: graduado escolar, 2: universidad, 3: instituto; 4: otros
MARRIAGE	Categorico	1: casado, 2: soltero, 3: otro
AGE	Numérico	Edad

PAY_0 – PAY_6	Categorico	Historia de pagos anteriores, contiene pagos de abril a septiembre de 2005. -1, -2, ..., -n: debidamente pagado durante los últimos n meses, 1-8: retraso de 1 a 8 meses, 9: retraso de 9 o más meses
BILL_AMT1 – BILL_AMT6	Numérico	Estado de la cuenta entre abril y septiembre de 2005 en dólares
PAY_AMT1 – PAY_AMT6	Numérico	Monto del pago del mes anterior en dólares
default payment next month	Categorico	Estado de impago 1: impagado, 0: pagado

Metodología

Exploración de los datos

Inicialmente se comprueba si hay valores nulos y si los datos se ajustan a la estructura del diccionario de datos.

Columna	Datos inválidos
EDUCATION	277
MARRIAGE	45
PAY1 – PAY6	5225

El resto de los valores se corresponde con el diccionario de datos, además se hacen algunas pruebas de calidad, como por ejemplo que la edad sea positiva y menor de 100 años, sin encontrar anomalías.

Outliers

Los outliers pueden afectar considerablemente los resultados que pueda obtener un modelo de Machine Learning, por tanto el paso siguiente es detectar los outliers que puedan estar presentes en el conjunto de datos.

Para ello usamos el test de Tukey, que define como outliers los puntos que estén en el siguiente rango:

$$[Q_1 - k(Q_3 - Q_1), Q_3 + (Q_3 - Q_1)]$$

Dónde Q_n es el cuartil n y k es una constante que puede valer 1,5 o 3. Se ha usado este método por su sencillez y por la posibilidad de visualizar los valores en un boxplot.

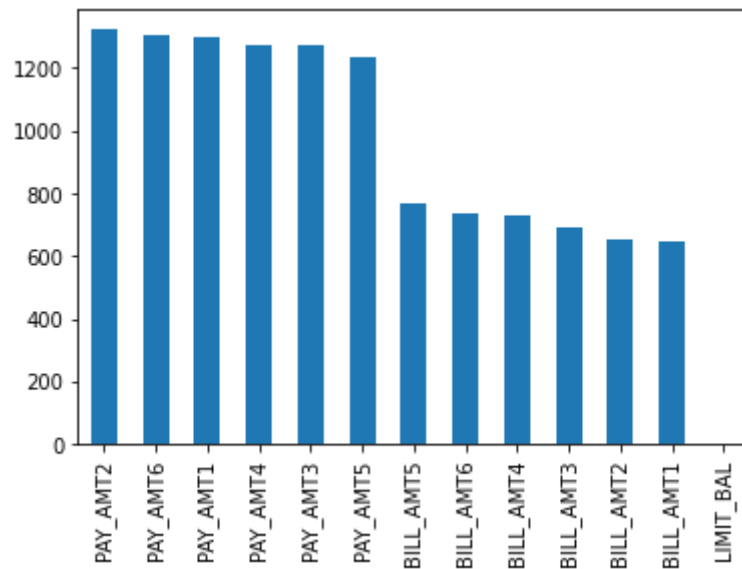


Ilustración 1: Cantidad de observaciones que se consideran outliers

Visualización de los datos

Para obtener una visión general del conjunto de datos se ha procedido hacer gráficos por cada una de las columnas.

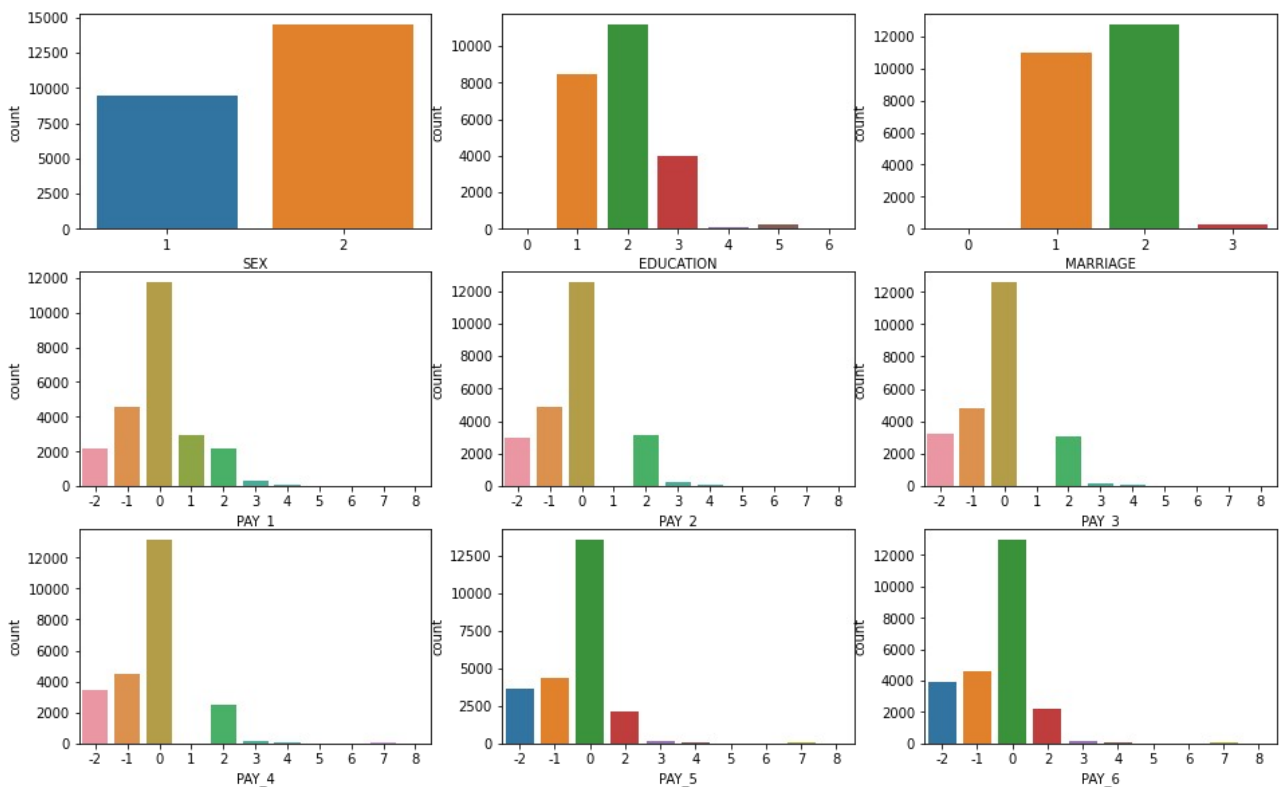


Ilustración 2: Datos categóricos

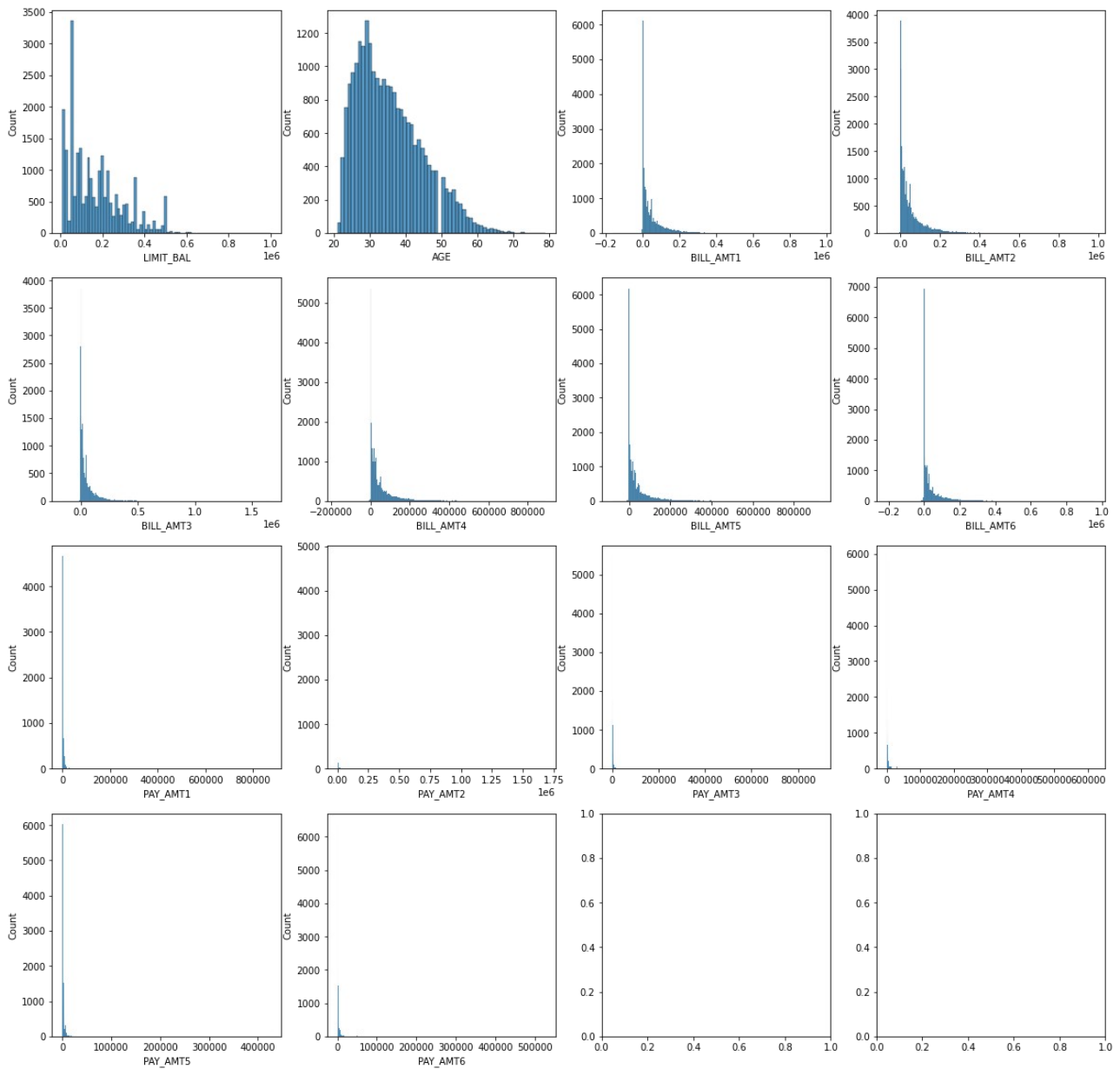


Ilustración 3: Datos categóricos

Como se puede observar, algunas de las columnas tienen valores extremos por lo que los gráficos no permiten observar la verdadera estructura de los mismos. Por ello, se muestran a continuación representados por un diagrama de cajas que permite identificar valores atípicos e intuir su morfología y simetría:

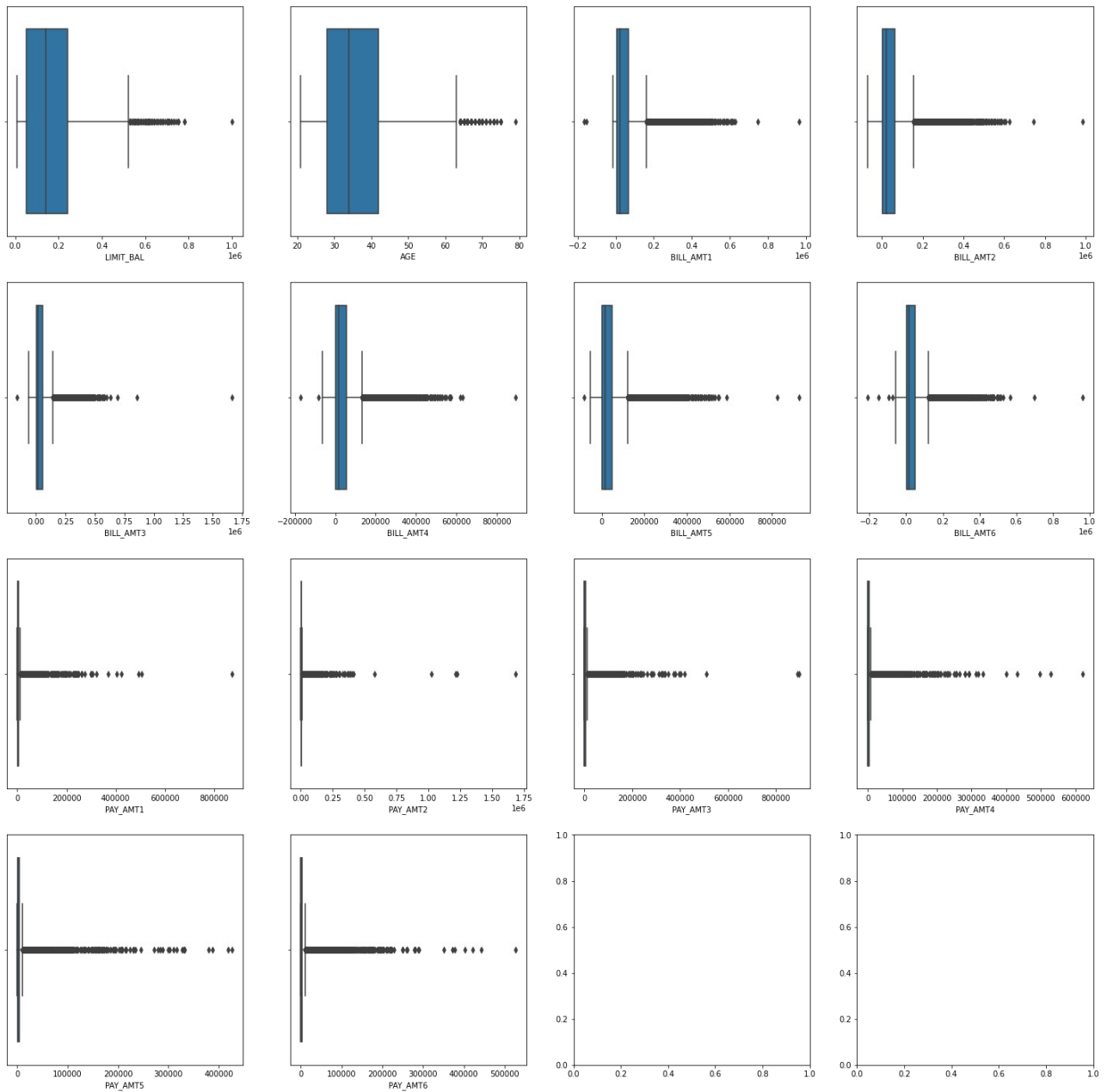
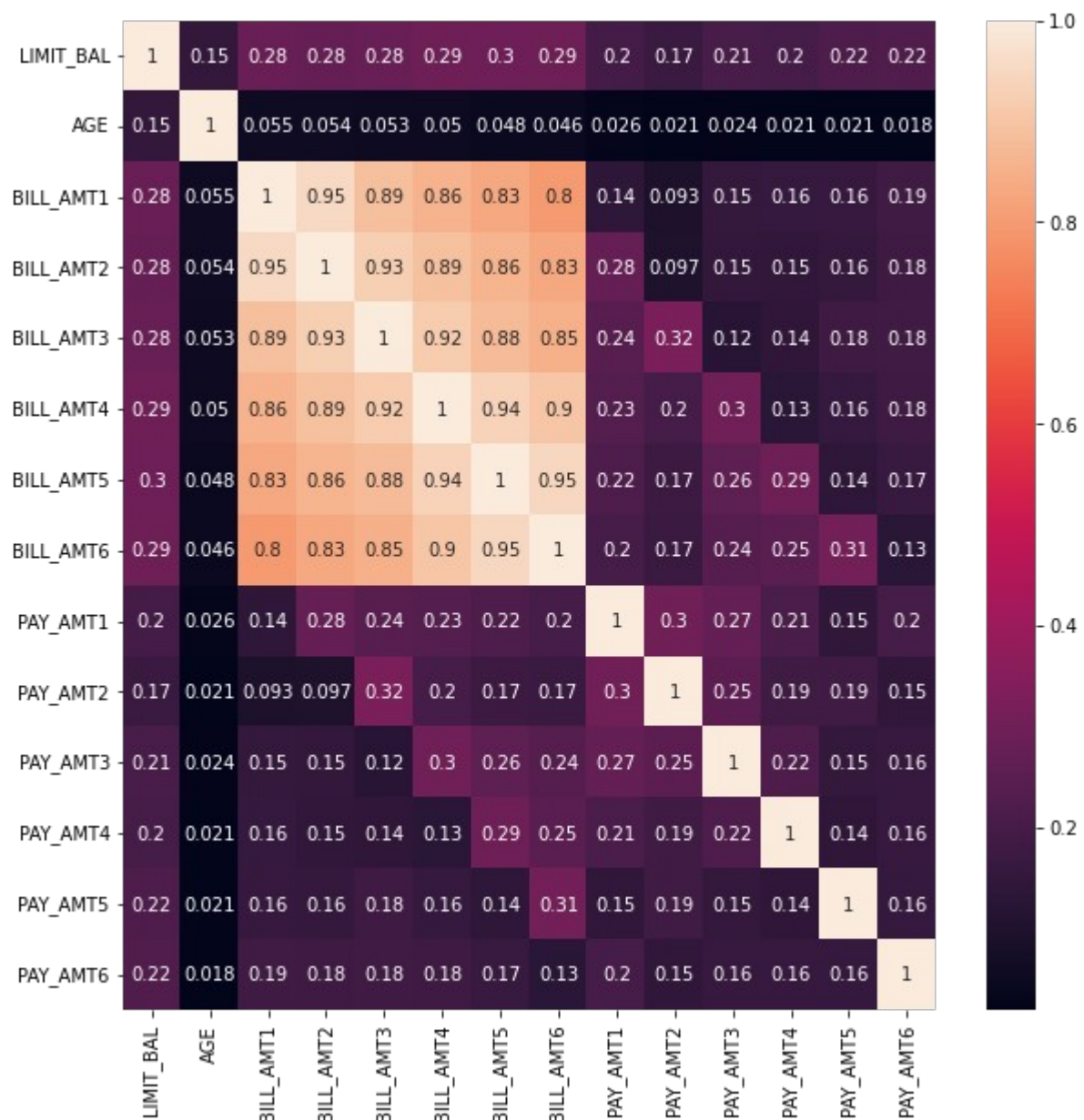


Diagrama de correlación

Finalmente nos queda comprobar las relaciones entre las variables, para ello se usa el diagrama de correlación.



Clasificación

En el problema que se intenta solucionar se pretende predecir si un crédito va o no a producir impago, esto es equivalente a clasificar los créditos en impagados y pagados, por lo que utilizaremos algoritmos de clasificación para ello.

Para medir la calidad de un clasificador lo compararemos con el clasificador naïve que se limita a escoger siempre la clase mayoritaria. Se elige una amplia variedad de algoritmos disponibles en la librería *sklearn* y se ejecuta validación cruzada sobre todos ellos y se compara su exactitud.

En concreto se han probado los siguientes algoritmos: (i) logistic regresion, (ii) linear discriminant analisys, (iii) quadratic discriminant analisys, (iv) k neighbors classifier, (v) classification and regression trees, (vi) Naive Bayes, (vii) AdaBoost, (viii) stochastic gradient descent, (ix) Bernoulli Naive Bayes, (x) Ridge Classifier, (xi) Calibrated Classification Model, (xii) Suport Vector Machine Classifier - linear, (xiii) Support Vector Machine Classifier – rbf, (xiv) Multi-layer Perceptron Classifier, (xv) Random Forests.

Búsqueda de hiperparámetros

Sobre los algoritmos que presentan un mejor rendimiento se realiza una búsqueda de hiperparámetros para intentar mejorar sus resultados y obtener una clasificación lo más ajustada posible. Los resultados de los modelos ajustados se presentan en la sección de resultados.

Importancia de las variables

Uno de los objetivos principales de este trabajo es determinar que variables ofrecen información para determinar la probabilidad de un futuro impago. Para ello se han usado dos métodos diferentes.

El primero de ellos se conoce como RFE (Recursive Feature Elimination) y es un algoritmo de selección de características. Es decir, su objetivo es seleccionar un subconjunto de las características disponibles que sea óptimo para la tarea de clasificación. Como un efecto lateral de la ejecución de RFE se produce una puntuación (scoring) que representa la importancia relativa de las características. Esto se hace típicamente usando el propio algoritmo de clasificación.

Es decir, el algoritmo toma el conjunto de características y va eliminándolas una a una y comprobando qué conjunto resultante es el que tiene una mejor puntuación, así la primera característica descartada será la que tiene peor puntuación mientras que la última en descartarse tendrá la mejor puntuación. El método se ha elegido porque es fácil de configurar, entender y efectivo al seleccionar las características que son más relevantes en la predicción.

El segundo método se llama SHAP (Shapley Additive exPlanations) que se basa en los valores de Shapley para indicar la contribución relativa de cada característica a una predicción dada. El método fue originalmente pensado para calcular la contribución de cada jugador en un juego de colación.

Supongamos que hay un conjunto de S jugadores cuyo resultado actual es $v(S)$. Al unirse un nuevo jugador i queremos calcular el valor marginal aportado por dicho jugador. Para ello haremos la media de la contribución de dicho jugador en todas las posibles permutaciones con que se puede formar la coalición y este será el valor de la aportación del nuevo jugador. El valor calculado es usado como puntuación del jugador o en nuestro caso, la característica.

El método fue escogido por tres razones principales, (i) Interpretabilidad global: es capaz de mostrar las relaciones positivas o negativas de cada característica con el resultado. (ii) Interpretabilidad local: se pueden obtener las relaciones para una observación data y (iii) es global: se puede calcular para cualquier método¹.

Resultados

Respecto a la clasificación

De entre los algoritmos probados se han elegido los que presentan mejores resultados y el clasificador logístico a modo de línea base.

¹ Para métodos no basados en árboles se usan algoritmos sustitutos o métodos basados en kernels.

Hay 4 métodos que sobresalen claramente: LDA (linear discriminant analysis), ADAB (AdaBoost), RC (Ridge Classifier) y RF (Random Forests) tal y como se puede comprobar en la ilustración 9.

Estos modelos, tras haber sido previamente ajustados usando una búsqueda en sus hiperparámetros ofrecen los resultados que se muestran en la tabla de resultados de clasificación.

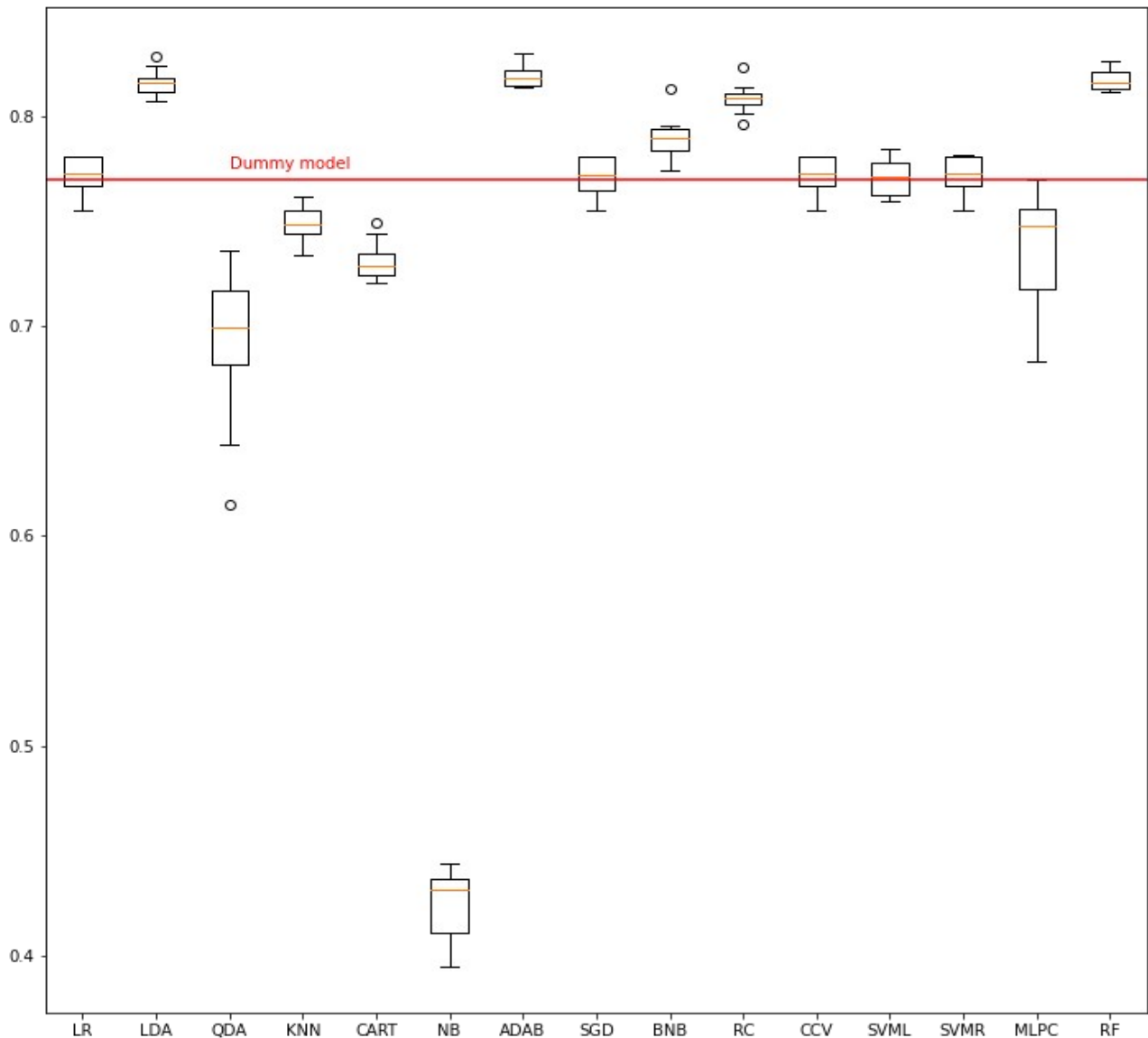


Ilustración 4: Exactitud de distintos métodos


Método	Exactitud	Precisión	Exhaustividad (recall)	f1-score
train				
LR	0,79	0,80	0,79	0,80
LDA	0,82	0,80	0,82	0,79
AdaBoost	0,82	0,81	0,82	0,80
RC	0,82	0,81	0,82	0,81
RF	0,89	0,89	0,89	0,88
test				
LR	0,79	0,80	0,79	0,80
LDA	0,81	0,80	0,81	0,79
AdaBoost	0,82	0,81	0,82	0,80
RC	0,82	0,81	0,82	0,80
RF	0,82	0,80	0,82	0,80

Respecto a la importancia de las variables

La tabla de importancias de variables muestra las diez variables más importantes para cada uno de los clasificadores elegidos usando RFE y SHAP. Además se han dividido las variables en categorías para poder interpretar mejor los resultados.

	LR		RC		RF		AdaBoost		LDA	
	rfe	shap	rfe	shap	rfe	shap	rfe	shap	rfe	shap
1	pay_1_3	pay_1_0	pay_1_3	pay_1_0	pay_1_4	pay_1_2	pay_amt1	pay_1_2	pay_6_8	pay_1_2
2	pay_1_2	limit_bal	pay_1_2	pay_6_0	pay_1_-1	pay_1_0	pay_amt2	limit_bal	pay_1_3	pay_1_0
3	pay_4_4	pay_6_0	pay_2_6	pay_5_0	pay_1_0	pay_2_2	bill_amt3	pay_amt2	pay_2_6	limit_bal
4	pay_5_2	pay_1_2	pay_1_4	pay_4_0	pay_4_8	limit_bal	bill_amt1	pay_amt1	pay_4_8	pay_3_2
5	pay_5_7	pay_1_-1	pay_6_8	pay_1_2	pay_6_0	pay_3_2	pay_amt3	pay_1_1	pay_5_6	pay_5_2
6	pay_2_1	education_2	pay_4_1	pay_1_-1	pay_2_6	pay_4_2	pay_amt4	bill_amt1	pay_5_7	pay_1_3
7	pay_1_0	bill_amt2	pay_2_1	pay_5_-1	pay_4_1	pay_amt1	limit_bal	pay_6_2	pay_4_7	pay_4_2
8	pay_1_-1	education_1	pay_1_0	pay_6_-1	pay_2_1	pay_amt2	bill_amt4	pay_4_2	pay_4_6	pay_6_2
9	pay_6_3	sex_1	pay_1_-1	pay_4_-1	pay_4_6	pay_5_2	pay_1_1	bill_amt3	pay_2_8	bill_amt2
10	education_4	pay_2_0	pay_4_8	limit_bal	pay_6_8	pay_2_0	pay_1_2	pay_amt4	pay_1_4	pay_2_2

 Indicador de retraso en el pago

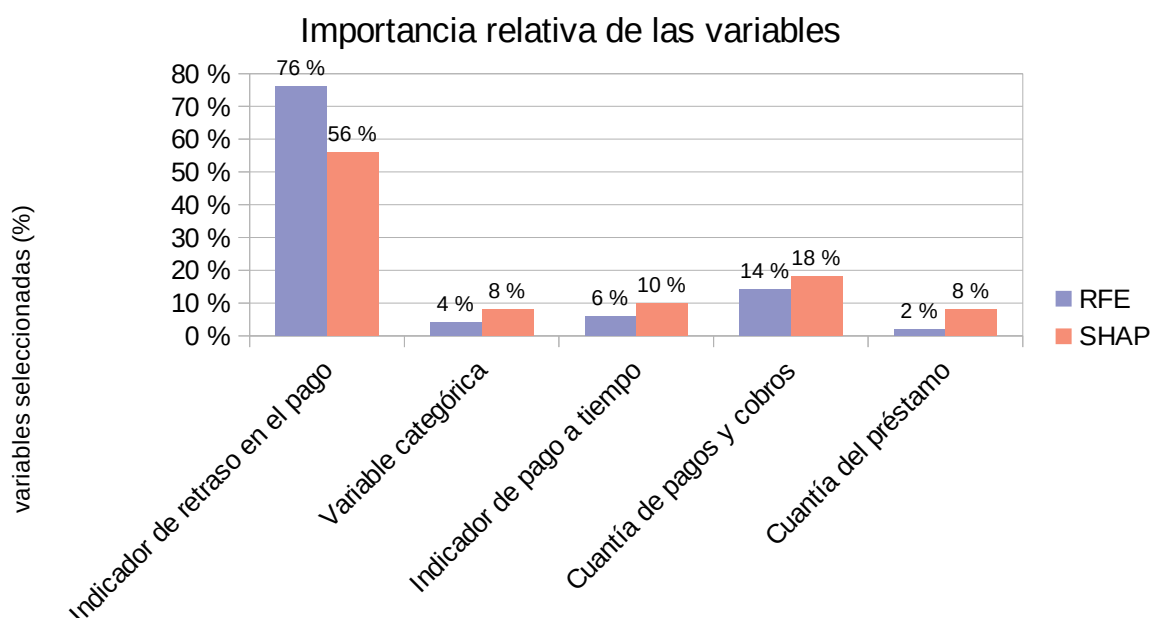
 Variable categórica

 Indicador de pago a tiempo

 Cuantía de pagos y cobros

 Cuantía del préstamo

En la ilustración correspondiente a la importancia relativa de las variables se muestra la puntuación que cada uno de los algoritmos da a los distintos grupos de variables.



Conclusiones

El primer objetivo de este trabajo es entrenar múltiples algoritmos de clasificación supervisada de forma que predigan el pago o impago de un crédito.

En primer lugar se realizó una exploración de los datos usando las técnicas de análisis habituales para posteriormente proceder a limpiar los datos y con ellos entrenar y probar un conjunto de algoritmos de clasificación.

Para elegir los algoritmos de clasificación se ha entrenado un primer clasificador “dummy” contra el que se comparan todos los demás y se eligen sólo aquellos que presentar una mejora significativa. Entre ellos figuran LDA (linear discriminant analysis), ADAB (AdaBoost), RC (Ridge Classifier) y RF (Random Forests). Hay que añadir además el Logistic Regression Classifier que se usa a modo de línea base.

La exactitud obtenida varía desde 0,79 en el clasificador logístico hasta 0,89 en el random forest y la precisión entre el 0,80 y el 0,89. No se observan grandes diferencias entre los distintos algoritmos de clasificación.

El segundo de los objetivos es observar la importancia relativa de las distintas variables en la clasificación. Para ello se han usado los algoritmos RFE y SHAP para obtener una lista ordenada de la importancia de las variables en cada uno de los clasificadores usados. Estas variables se pueden agrupar en: (i) indicadores de retraso en el pago, (ii) variables categóricas, (iii) indicadores de pago a tiempo, (iv) cuantías de pagos y cobros y (v) cuantía total de préstamos.

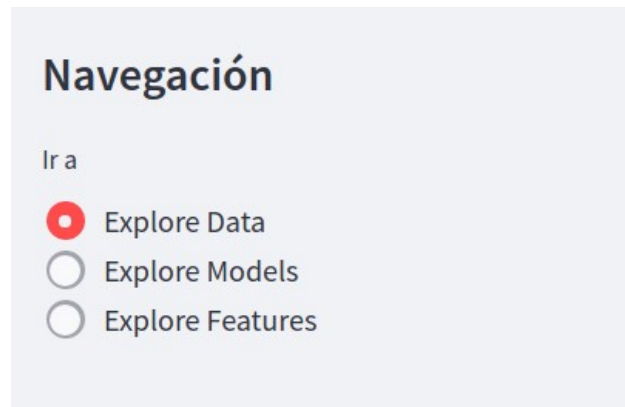
RFE indica que un 76% de las variables escogidas corresponde a indicadores de retraso en el pago mientras que SHAP indica un 56% siendo la categoría principal para ambos algoritmos. Algo que está en consonancia con información encontrada en la literatura [1].

Hay que considerar en todo caso que las variables que mejor predictibilidad presentan según la literatura, son las poblacionales, y no estaban incluidas en nuestro dataset [2].

Manual del Frontend

La aplicación de frontend es un informe dinámico del trabajo realizado en este TFM. Consta de tres partes diferentes:

1. Exploración de datos
2. Exploración de modelos
3. Exploración de características



El panel de navegación lateral permite cambiar de una a otra sección.

Exploración de datos

En esta sección se puede ver una muestra de los datos originales así como estos mismos datos tras ser procesados por la aplicación.

A continuación se presenta un desplegable con el nombre de todas las variables originales y al seleccionarla se generan dos gráficos: un histograma y un diagrama de caja para entender la distribución de las variables de forma visual.

Finalmente se muestra una tabla de correlaciones entre las variables que permite escoger entre los datos originales y los datos ya tratados. Esta tabla además está coloreada para mostrar más claramente las relaciones entre las variables.

Exploración de modelos

En esta segunda sección se puede elegir un modelo entre los siguientes:

- Dummy
- LR (Logistic Regression)
- AdaBoost
- RC (Ridge Classifier)
- RF (Random Forest)
- LDA (Linear Discriminant Analysis)

Tras seleccionarlo se mostraran en primer lugar los datos correspondientes a “accuracy”, “precision” y “recall” del modelo. A continuación las matrices de confusión calculados con los datos de entrenamiento y de prueba y finalmente la curva ROC.

Exploración de características

En esta sección se puede seleccionar uno de los modelos anteriormente mencionados y se presentará a continuación la clasificación de las variables de más a menos influyentes en la clasificación según SHAP seguido de un diagrama de la influencia positiva o negativa de dichas variables en el resultado.

Ejecución de la aplicación

Las instrucciones para la ejecución de la aplicación se encuentran en el fichero README.txt incluido dentro de la carpeta frontend.

Bibliografía

Rising Odegua, Predicting Bank Loan Default with Extreme Gradient Boosting

Luca Barbaglia, Sebastiano Manzan, Elisa Tosesti, Forecasting Loan Default in Europe with Machine Learning, *Journal of Financial Econometrics*, 2021

Jason Brownlee, Recursive Feature Elimination (RFE) for Feature Selection in Python (<https://machinelearningmastery.com/rfe-feature-selection-in-python/>)

Dr. Dataman, Explain Your Model with the SHAP Values (<https://towardsdatascience.com/explain-your-model-with-the-shap-values-bc36aac4de3d>)

Dr. Dataman, The SHAP with More Elegant Charts (<https://dataman-ai.medium.com/the-shap-with-more-elegant-charts-bc3e73fa1c0c>)

Dr. Dataman, Explain Any Models with the SHAP Values – Use the KernelExplainer (<https://towardsdatascience.com/explain-any-models-with-the-shap-values-use-the-kernelexplainer-79de9464897a>)