INFORME PRÉSTAMOS LENDING CLUB (MEJORADO)

Abstract

En este informe se estudia una gran base de datos (más de diez millones de datos) que habla sobre las características de los prestatarios de la plataforma Lending Club (una plataforma de préstamos online en Estados Unidos). La información caracteriza en cierta parte a los prestatarios (se conoces sus ingresos, su nivel de deuda, sus propiedades, sus impagos) y además habla sobre su comportamiento con el préstamo concedido (cuánto lleva pagado, si se encuentran o no al corriente de pagos). El informe presenta una brevísima introducción a los datos y la construcción de varios modelos de predicción para prever si es o no conveniente conceder un crédito a un potencial cliente. Esta problemática presenta una gran importancia desde la creación de las plataformas P2P[1] pues busca garantizar a los prestamistas la seguridad de sus inversiones así como permitir un reparto eficaz del crédito (en el sentido de que no haya personas rechazadas que realmente si serían solventes).

Antes de comenzar con la construcción del modelo se procede a una depuración de la base de datos. En esta depuración se rechazarán una serie de variables de nula utilidad para la construcción del modelo como pueden ser la identificación de cada préstamo, el motivo, la URL asociada…

Tras ello se realiza un primer acercamiento al estudio de los datos. Observando por ejemplo cómo se distribuyen los préstamos según su importe (**Figura 1**). Se puede observar que la mayoría de préstamos tienen un importe menor de 20.000€ con una gran concentración en torno a los 10.000.

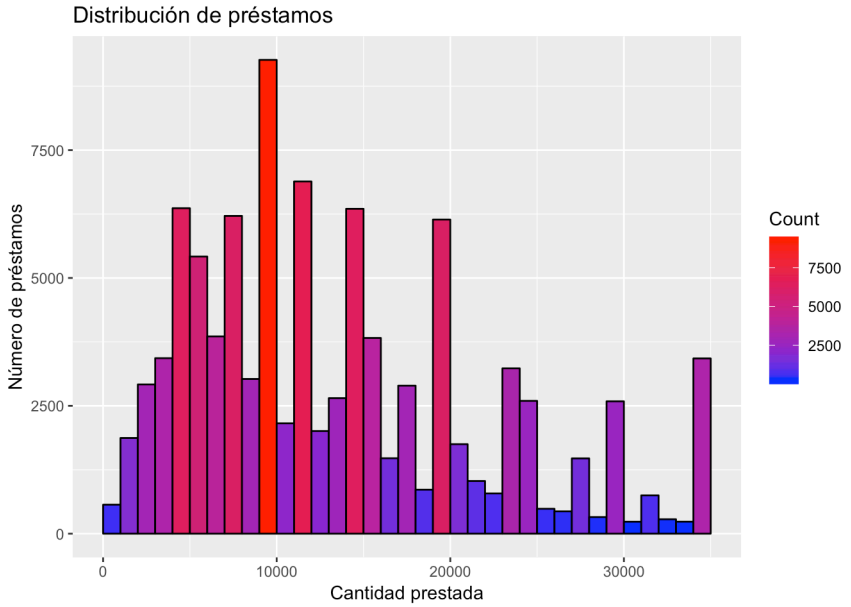


Figura 1. Distribución de los préstamos según su importe.

[1] Riza Emekter, Yanbin Tu, Benjamas Jirasakuldech & Min Lu (2015) Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending, Applied Economics, 47:1, 54-70, DOI: 10.1080/00036846.2014.962222

Es también interesante la relación entre el grado de riesgo y el número de préstamos concedidos. El gráfico de frecuencia (**Figura 2**) muestra que el mayor número de préstamos se concede a clientes de clase C. Esto no quiere decir que los prestatarios de grado C tengan ventajas sobre los de grado A si no que simplemente existe un mayor número de demandantes de grado C.

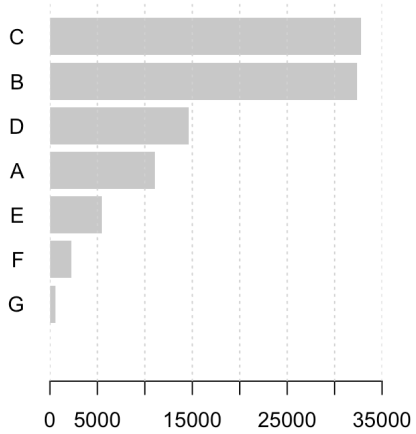


Figura 2. Gráfico de frecuencias de los préstamos según la clase del prestatario.

Una vez realizada una primera aproximación a los datos se plantea la construcción de un modelo que permita predecir la concesión de créditos. Para ello en primer lugar se deben amoldar las variables para trabajar con ellas, en este caso, este proceso de depuración constará de la eliminación de observaciones incompletas (que poseen algún NA) y la adaptación de las variables de manera conveniente (transformación de porcentajes en ratios, normalización del dti), dicotomización de algunas variables… Todo este proceso se encuentra perfectamente detallado en el código presente en GitHub. En una primera versión se produjeron ciertos problemas con la dicotomización del estado del crédito llevando a unos modelos que presentaban un enorme overfitting. En esta versión se plantean dos variantes: que los préstamos en proceso (current) se consideren como exitosos o que, por el contrario, se descarten y no se consideren como casos a estudiar.

En este informe se desarrolla esta segunda versión pues se sigue disponiendo de más de 5000 observaciones, no obstante si al lector le interesa en GitHub dispone de la versión en la que se consideran los current (*versión alternativa*) siendo la interpretación de los datos obtenidos en esta análoga a la que se realiza en el informe.

Tras completar este proceso se plantean las siguientes variables para la construcción del modelo: cantidad prestada (loan_amnt), cantidad financiada(funded_amnt), cantidad financiada invertida (funded_amnt_inv), ratio de deuda e ingresos (dti), uso de la revolving credit line (revol_util), posesión de un hogar (home_ownership), estado de verificación del crédito (verification_status) y riesgo para predecir el pago o impago de cada crédito.

Al llevar acabo la construcción de este primer modelo y estudiarla destaca la importancia de algunas variables como revol_util (con un p-valor asociado inferior a $2{\cdot}10^{-16}$), home_ownership (con un p-valor asociado de $7.89{\cdot}10^{-07}$) y verification_status(con un p-valor asociado inferior a $2{\cdot}10^{-16}$), con menor relevancia aparece dti (con un p-valor asociado de 0.00123).

Así se opta por un modelo construido a partir de las variables anteriores. Al comparar ambos modelos una vez construido se observa que el modelo de menor tamaño (en el sentido de que requiere menos variables) es mejor (su criterio informativo de Akaike es 24806 siendo la del mayor 39607). Además el alto valor de Intercept indica que una de las clases (crédito concedido o no) va a poseer muchos más elementos que otra. En este caso es correcto porque la mayor parte

de los elementos de la muestra se encuentran o aun en proceso (se dan por buenos) o completados correctamente.

Con este nuevo set de estudio se procede al entrenamiento del modelo eligiendo un subconjunto de entrenamiento del formado por el 80% de la muestra (se probaron otras configuraciones que dieron resultados peores o muy similares) dejando un 20% de la muestra para el testeo.

Tras entrenar este modelo se obtiene la siguiente matriz de confusión:

		Observado
Predicho	20	427
	13	1055

Matriz de confusión del modelo lineal.

Partiendo de este modelo inicial cabe preguntarse cómo puede ser mejorado. Algunas de las opciones posibles son modelos de cresta, lasso o redes elásticas:

Modelo cresta

El modelo se entrena y mediante un proceso de cross validation se obtiene un cuttoff óptimo de 0.1221244 algo que se puede visualizar de manera clara en la siguiente gráfica (nótese que ln(0.1221244)= -2.1027):

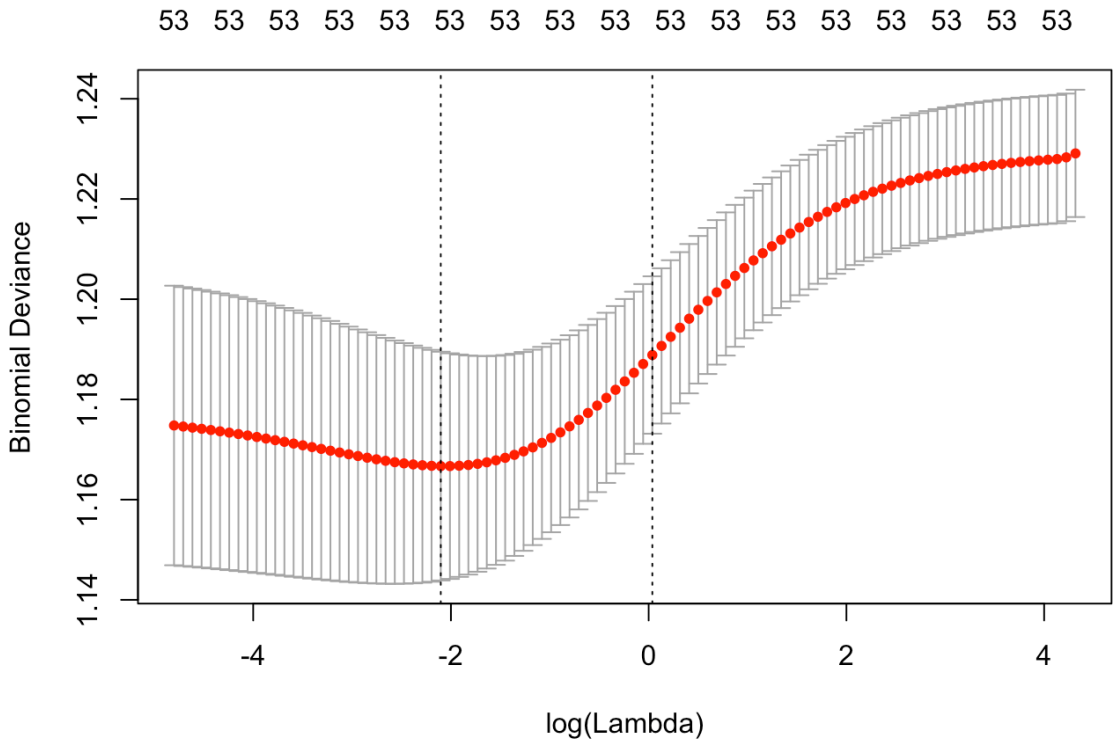


Figura 3. Gráfica de los distintos cutoff para el modelo cresta.

Si se construye el modelo con este cutoff y se calcula el número de errores cometido sobre el test set es 734 (es un valor orientativo) que en porcentaje sobre la muestra (2525) representa un 29,18%.

Este modelo se compara a continuación con el modelo lasso:

Modelo Lasso

Siguiendo un proceso similar al anterior (todo esto se encuentra perfectamente detallado en el código anexo) se construye un modelo lasso a partir de las variables y se calcula su cutoff óptimo; en este caso 0.010622. Tras ello se entrena el modelo y sus resultados en el testeo muestran una ligera mejora respecto al modelo cresta. En este caso, el número de predicciones erróneas asciende a 719, lo que representa un porcentaje sobre la muestra de validación del 28,48%; un 0,7% mejor que el modelo cresta.

Este modelo confirma además que las subgrades realmente no aportan una información mucho más relevante que las grades por lo que se pueden considerar redundantes (no obstante se conservan porque determinadas calificaciones como G2 y G3 parecen aportar cierta información).

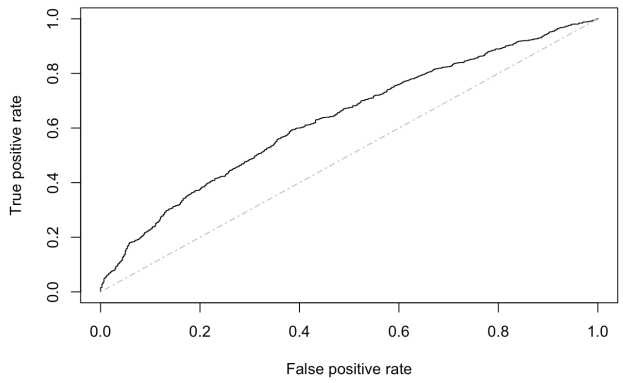
Modelo Elastic Net

Estos modelos buscan aproximar un alfa óptimo (entre 0, correspondiente al modelo cresta y 1 correspondiente al modelo lasso). En este estudio se aproxima solo a la décima obteniéndose el valor óptimo en 0.2 con 710 errores. Se prescinde de buscar una mejor precesión porque los valores en 0.1 y 0.3 son ya muy próximos a 720.

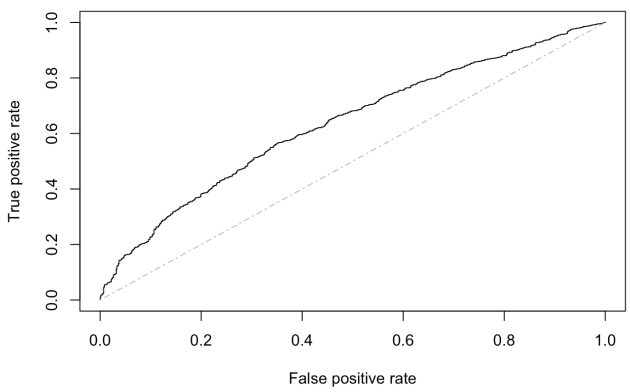
Se realiza así sobre este modelo el mismo proceso para encontrar el mejor lambda que aparece en 0.04017486. Y en este caso se obtienen 708 errores. Así se ha conseguido reducir finalmente el error a un 28,03%.

Por último se descarta la construcción de modelos polinomiales porque en este modelo influyen multitud de variables discretas (incluyéndose entre ellas la variables a predecir).

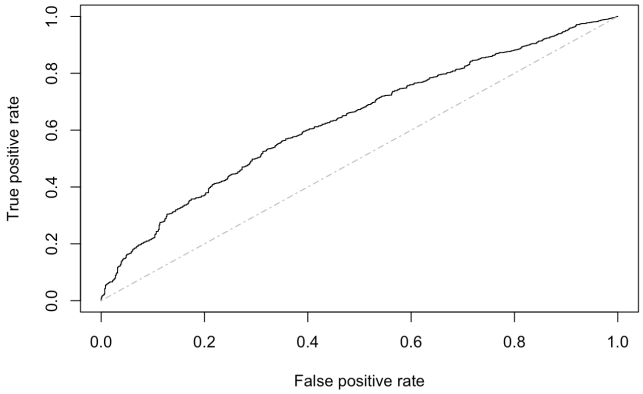
Para terminar el informe la **Figura 4** muestra las curvas ROC relativas a cada modelo confirmando que los tres son sustancialmente similares:



Modelo Ridge



Modelo Lasso



Modelo elastic net con cutoff=0.2

Conclusiones

Este informe mejora en varios aspecto el anterior presentado. En primer lugar, se logra introducir en el modelo variables no numéricas y más en concreto la grade que permite una predicción mucho más realista de los préstamos. En los modelos construidos parece no darse una gran importancia al ratio de deuda - ingresos. Aunque en un primer momento se planteo la eliminación de esta variable de los modelos se decidió conservarla porque los resultados eran ligeramente peores y diversos artículos como el antes citado prueban la gran importancia de esta variable. Su escasa relevancia en este caso puede deberse a los datos elegidos. (Recuerde el lector que en esta versión se trabaja solo con 5.000 préstamos descartando aquellos que están en proceso). Por lo demás estos modelos vienen a confirmar la hipótesis generalizada que subraya la importancia de la nota y el uso de la revolting credit line. Además en este caso aparece tan bien como una variable muy significativa el estado de verificación del crédito sobre todo cuando este aparece verificado desde fuente (source verified).

Anexo I

El siguiente enlace conduce al repositorio online donde se encuentran los códigos sobre los que se sostiene este informe:

https://github.com/ArturoSanchezPalacio/Prediction.git