ANÁLISIS DEL MODELO GLM EN R STUDIO

Máster en Data Science

Ricardo Queralt

Marta Divassón Carribero

Fecha: 2 noviembre 2018

Índice:

Introducción

Este proyecto tiene como objetivo analizar modelos líneas de generalización mediante la regresión de una serie de variables basados en el dataset "*Lending Club*". La información consta de una serie de datos sobre préstamos otorgados con variables notorias como el estado de los préstamos, variable que se tomará como dependiente. Debido a la larga cantidad de datos informativos, con este informe se pretende simplificar para obtener aquellas variables predictivas mediante un análisis exploratorio.

Resumen ejecutivo

A partir de los datos obtenidos se persigue construir un modelo lineal de generalización con una variable discreta, aquella que es binaria y toma un rango de valores, y una serie de variables continuas. Las variables predictivas que reflejen si los individuos pagarán o no los préstamos y para ello se deben buscar las categorías que más afectan al modelo.

La mejor forma de ver esto es con tablas de frecuencia y representaciones gráficas antes de establecer el modelo para observar la relación que hay antes de desarrollar un modelo final. Es decir, estas observaciones pueden ser útiles para ver si las variables independientes que se escogen se relacionan con la variable "respuesta" y así ser futuros predictores.

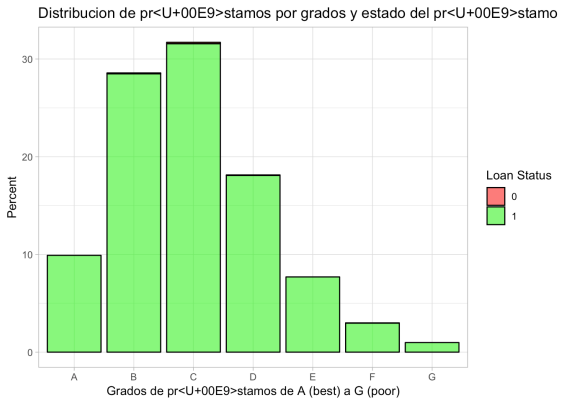
Tras la exploración de datos y la limpieza de ellos se procede a hacer un modelo de prueba y de entrenamiento (*test* y *train*), aquel que coge los datos de manera independiente sin conocer el modelo inicial y que predice una serie de datos que luego se ajustarán con una precisión determinada.

Por tanto, se lleva a cabo una investigación de:

- El crecimiento de los prestamos en cuestión de dólares y volumen
- La intención del préstamo
- El cambio de tendencia del interés
- Los grados para el pago de prestamos
- Análisis de correlación
-

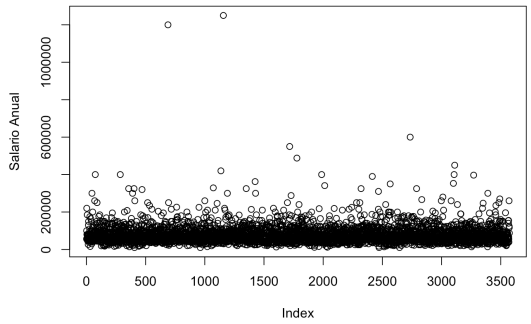
Análisis

En el proceso de Data Mining (extracción de información útil para la elaboración del modelo) se observa un *dataframe* con 99122 observaciones y 111 variables. De esas 111 variables, parecen ser de gran utilidad 19 variables, que se seleccionan para tener mayor concisión. Estas son de diferentes clases, desde numéricas hasta factores, por tanto, se debe hacer una conversión para poder trabajar con la misma base y tratas los "*NAs*" mediante una sustitución de la mediana de cada parámetro. Cuando tomamos la variable "*loan status*" o estado del préstamo podemos reflejarla con el grado de riesgo de los préstamos, que es de gran utilidad para los inversores a la hora de analizar la historia crediticia de estos prestamos; siendo los de mejor grado los que se llegan a pagar (aquellos que están entre A y C y que corresponde a más del 50% de los préstamos).



Posteriormente, en un análisis más intensivo se aprecia que la mayoría de los préstamos se utilizan para pagar deuda pendiente por aquellas personas que lo solicitan. Otro ejemplo, es que a medida que aumenta el interés el crédito recibe una nota más baja, con mas improbabilidad de ser pagado.

Otra variable que puede ser significativa es el salario anual de cada individuo, situándose entre 10,000 y 1,250,000 dólares, aunque la cifra máxima sesga el estudio ya que hay muy pocas personas con un suelo superior a 1 millón de dólares, consideradas outliers. Esto puede observarse en el siguiente gráfico.



El último paso en el análisis exploratorio es interpretación con respecto a la multicolinealidad. Esto se hace para ver de forma cercana cuales son las variables que se incluirán en el modelo mediante correlaciones. La existencia de correlaciones negativas entre variables serán determinantes para excluir esas características del modelo. La cantidad del préstamo está muy relacionada con "*installment*" (0.9603) seguida por su relación con salario anual (0.3623). Esto se aplica a que vamos a excluir variables como "installment" y sub_grade por estar muy correlacionados con el grado de los riesgos.

Modelo

Al efectuar el modelo, se precisa observar el *pvalor*, si este se sitúa por debajo de 0.05 indica que es significativo y explica un elevado porcentaje de la varianza (que se observa en el Rsquare).

Por tanto, tras un análisis exploratorio, el modelo incluirá las siguientes variables independientes: *grade, int_rate, open_acc, pub_rec, dti, revol_bal, revol_util, delinq_2yrs, inq_last_6mths, emp_cat, annual_inc, home_ownership, purpose and loan_amnt, teninedo como variabla dependiente loan_status.*

Se escoge como variable dependiente porque el estado del préstamo (loan_status) es una variable binaria que muestra si se ha pagado o no el préstamo o esta en curso y va a ser pagada. Para saber todo esto queremos estimar como afectan las variables independientes sobre esto.

Pero, antes de concluir con un modelo final de regresión logística, hay que desarrollar un modelo entrenamiento y de prueba, mediante la división de datos. Primero mediante el entrenamiento en un set de entrenamiento, luego se prueba el nivel de ajuste de ese set en nuestro set de prueba donde se desconoce de donde provienen los datos (ya que se ha hecho un *set.seed*). Luego se prueba el performance de lo desarrollado mediante un análisis y la curva ROC, para ver si el modelo se ajusta.

Ahora bien, la mejor forma de llevar a cabo los modelos de entrenamiento y prueba es mediante "*cross validation*", con diferentes datos que se van probando. Al ser un método bastante largo, se divide el modelo en 2, siendo el 80% el de entrenamiento y 20% el de prueba.

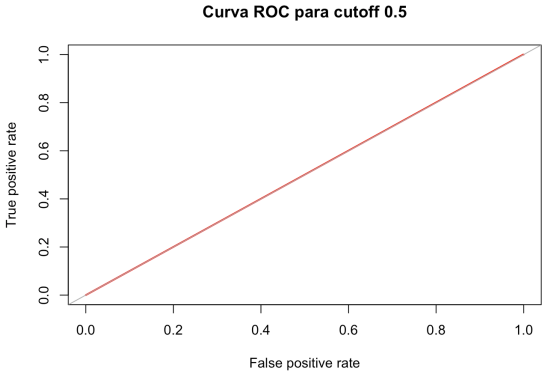
El modelo es significativo cuando se utilizan todos las variables para predecir, entre ellas las más significativas son:

- **Dti**: esto explica la cantidad de deuda que tienen por su salario (es un ratio), excluyendo variables como la hipoteca.
- **Pub_rec:** son los "public records", es decir, aquellos informes públicos desfavorables sobre los individuos a los que se les condece un préstamo.
- **Revol_bal:** es decir, también afecta el revolving balance, que es aquella porción de un préstamo que queda sin pagarse cuando vence el préstamo. En una línea de crédito, mes a mes se acumula la deuda de una línea de crédito.

Estas variables son significativas ya que afectan de forma negativa a la hora de pagar un crédito y son buenas a la hora de predecir.

Posteriormente, hay que buscar el valor cutoff óptimo en regresión logística cuando los datos no están "en balance". Esto se hace porque los valores positivos de pago del préstamo son bajos (*paid off loans*). El valor de cutoff se suele establecer en 0.04 y 0.08. Cuando se hace el cutoff de 0.05, no aporta mucho al modelo ya que indica que todas las variables están por debajo de ese nivel. En el caso del AUC o "*area under the curve*", obtenemos un valor de 0.4992. Este numero debe estar entre 0 y 1, y cuanto mas alto mejor performance. Por tanto, esto no nos da mucha información. Al hacer un gráfico con la curva ROC. La curva ROC es otra forma de medida para decidir el cutoff, ya que se pretende hace run balance de datos entre los FNR y FPR (false negative y false positive rates). Lo que indica la grafica es que cuanto mas cerca este la línea a la esquina izquierda de la grafica mejor performance tiene el modelo. Se pueden comparar diferentes curvas ROC aunque con esto no sabemos con precisión qué modelo será el mejor. De igual forma ocurre con un cutoff de 0.8.

Se utiliza la matriz de confusión para ver los datos más detalladamente, ya que hacen un display de una tabla de contingencia, siendo 710 los datos en los "true negative" (TN).

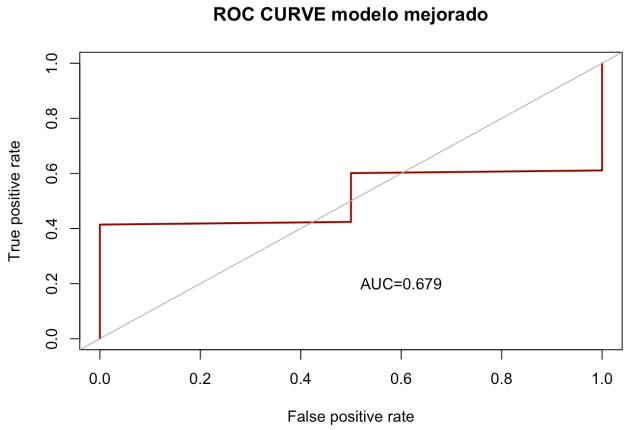


Por tanto, se elabora un nuevo modelo logístico, que sea mejor predictor de datos. Esto se hace con una estrategia en la que el cutoff sea en 0.8 y sobre el modelo de prueba. Se divide en los préstamos que se aceptan, que representan una probabilidad muy baja y se elabora una estrategia de aceptación según los préstamos que estén evaluados negativamente. Se deben aceptar, por tanto, aquellos que tengan un ratio de aceptación entre 0.2 y 0.4. Esto se hace para tener los datos balanceados, que haya un cutoff que este distribuido dejando el 50% de las variables mas o menos a cada lado. Al efectuar esto, quedan separadas 1491 y 1368 variables a cada lado. Siendo más preciso que el modelo anterior.

Las variables significativas, en este caso son las siguientes:

- **Open_acc:** es el numero de cuentas que tiene la persona que pide el préstamo.
- **Public records**: como en el caso anterior.
- **Dti** o deuda por salario: como en el caso anterior.
- **Delinq_2_years**: es el numero de incidentes del cliente en los últimos 30 dias.
- **Inq_last_6_months**: numero de "inquiries" en los últimos 6 meses.
- **Annual_inc** o salario anual
- **Loan_amnt:** es la cantidad solicitada para préstamo.
- **Revol_bal** como con anterioridad al igual que **revol_util**, esto es: es el ratio de utilización de la línea de crédito.

Y finalmente, al hacer un balance de los datos, la curva ROC tendría la siguiente forma.



Conclusion

La razón por la que se solicitan préstamos ese n mayor parte por acabar con la deuda que tienen los clientes (debt consolidation). El valor de corte es fundamental, porque los inversores miraran estos datos para decisiones de inversión en prestamos y también qué clientes obtendrán un préstamo y quienes no.

Después de evaluar el modelo, esta claro que no son los mejores estimadores y tienen un bajo rendimiento. Aquí, parece ser que la precisión de la clasificación es una medida que da problemas cuando hay desequilibrio, lo cual es normal en el entorno bancario, en el que reina la incertidumbre. Por tanto, la mejor forma de medir esto es mediante la curva ROC, mostrando las tasas verdaderas negativas contra las positivas. Por esta incertidumbre, también podemos concretar que hay situaciones extrínsecas al modelo que pueden llevar al cliente a hacer un "default" en el préstamo.

Parece ser que los modelos de riesgo de préstamo llevan asociados más variables "True Negatives" que "True Positives" porque hay una clasificación de por si que no está balanceada.

Github URL:
https://github.com/martadiva/PREDICCION2/blob/master/HW2PREDICTMARTA.R