CIENCIA DE DATOS E INTELIGENCIA DE NEGOCIOS

PROYECTO DE APLICACIÓN: CLASIFICACIÓN Y OPTMIZACIÓN DE MODELO.

Víctor Ramos Calderón y Eduardo Castillo Martínez

4 Abril 2017

Introducción

La finalidad de este proyecto es la aplicación y evaluación de los métodos de análisis de variables, creación y optimización de modelos de clasificación basados en regresión logística.

La base de datos que se analizará tiene información de transacciones de tarjetas de crédito, existen algunas transacciones que son clasificadas como fraudulentas. Entonces el objetivo de este proyecto es desarrollar un modelo logístico que logré identificar los movimientos que correspondan a fraudes (representados con el número 1) y de los movimientos que no son fraudes (representados con el número 0).

Base de datos

La base de datos que se utilizó cuenta con un total de 31 variables, la variable con nombre "Time", 28 variables sin nombre "V1...V28", y las variables "Amount" y "Class" siendo esta última, la variable binaria que nos indica si la transacción fue fraudulenta o no. La base de datos cuenta con un total de 284,807 filas, siendo cada una de ellas una transacción.

Selección de variables pre-modelado

El criterio de selección de variables que se utilizó fue el de componentes principales, este tipo de análisis consta en realizar la proyección de los datos sobre los ejes donde se tiene una mayor dispersión de los datos. Esto hace posible reducir la cantidad de atributos haciendo un mapeo a un espacio de menor dimensión, es decir, se reduce la dimensionalidad de los datos.

Tabla 1

0	2255124013.946	11	0.001	21	0.992
1	62554.860	12	0.113	22	0.955
2	3.704	13	0.106	23	0.544
3	2.405	14	0.166	24	0.903
4	2.005	15	0.230	25	0.776
5	1.879	16	0.245	26	0.716
6	1.800	17	1.013	27	0.767
7	1.601	18	0.367	28	0.836
8	1.430	19	0.402	29	0.659
9	1.212	20	0.502	30	0.674
10	1.195				

Valores propios para cada una de las variables

Como se aprecia en la tabla 1, la variable "tiempo" y la variable "V1" son las dos variables con valores propios notoriamente mayores a los demás, estas dos variables son las que más información aportan en esta base de datos.

Solución

Se realizaron diversas reducciones de variables y para cada una de se entrenó un modelo de regresión logística con el propósito de encontrar el número de variables resultantes después del ACP que produjeran los mejores resultados en la regresión.

Los criterios que se usaron para determinar que regresiones son mejores son:

- Accuracy: total de aciertos del modelo dividido entre el número total de datos.
- Precision: Precisión de los clasificados como"1" por el modelo.
- Recall: Proporción de cantidad de "1" originales encontrados por el modelo.
- F1: Similar a un promedio ponderado entre "precision" y "recall".
-

Regresiones con polinomio de 1er grado:

Gráfico 1

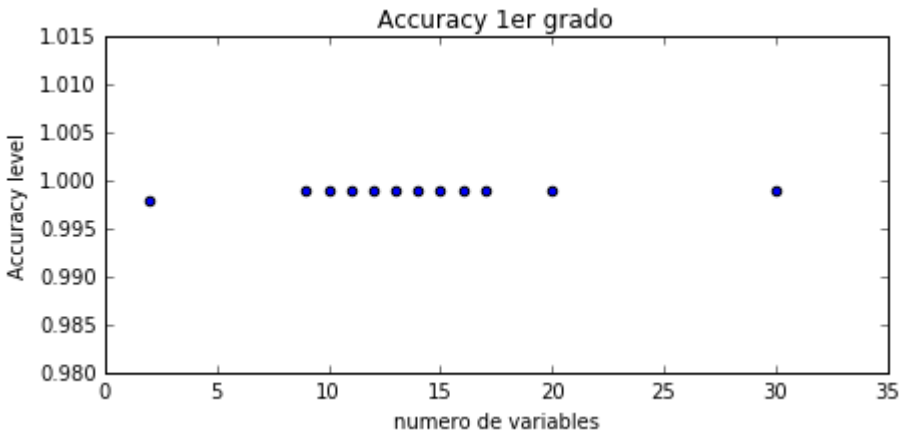


Gráfico de resultados en "accuracy" para diferente número de variables en el ACP, Pol.1er°

Dado que la base de datos solo cuenta con 0.17% de transacciones fraudulentas y 99.82% de transacciones no fraudulentas, el modelo de regresión logística tendrá un nivel de "accuracy" muy alto incluso si clasifica todas las transacciones como no fraudulentas (representadas con el número 0), es por esto que no nos enfocaremos en este criterio en este documento.

Gráfico 2

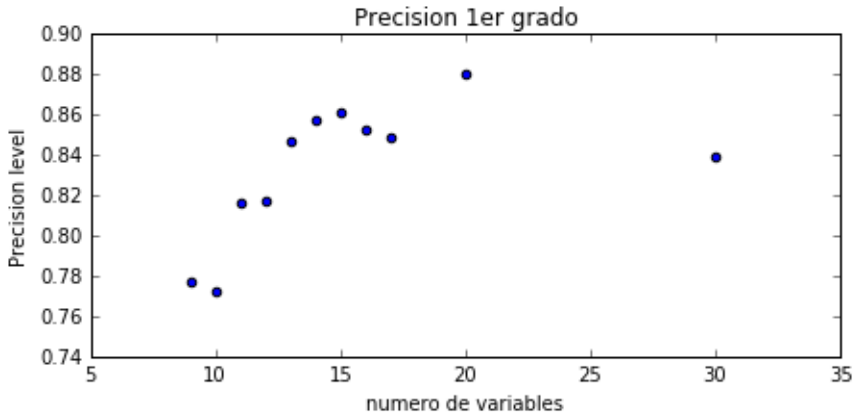


Gráfico de resultados en "Precision" para diferente número de variables en el ACP, Pol.1er°

Gráfico 3

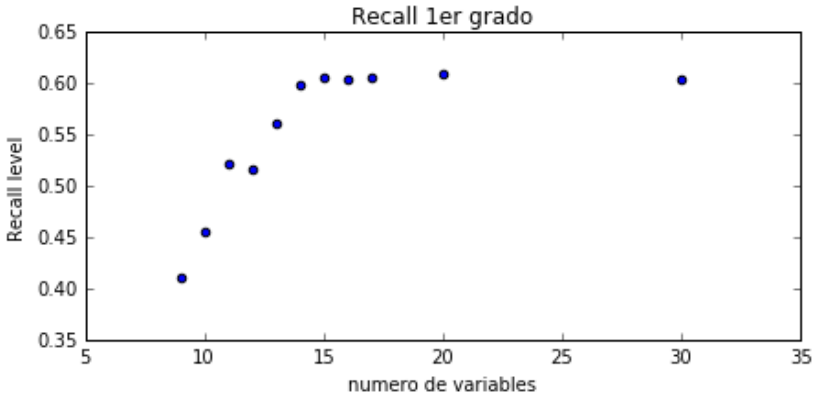


Gráfico de resultados en "Recall" para diferente número de variables en el ACP, Pol.1er°

Gráfico 4

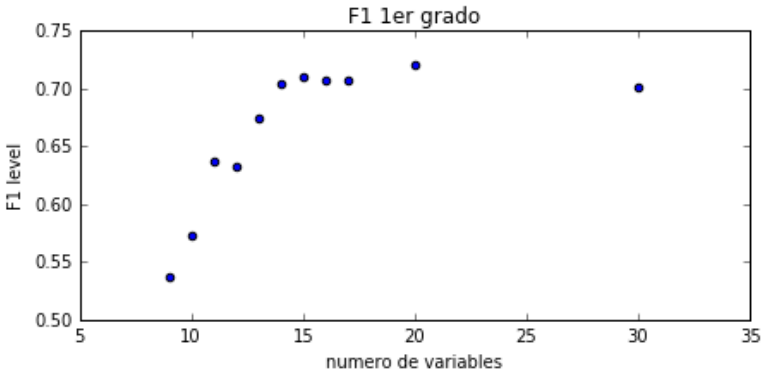


Gráfico de resultados en "F1" para diferente número de variables en el ACP, Pol.1er°

Regresiones con polinomio de 2do grado:

Gráfico 5

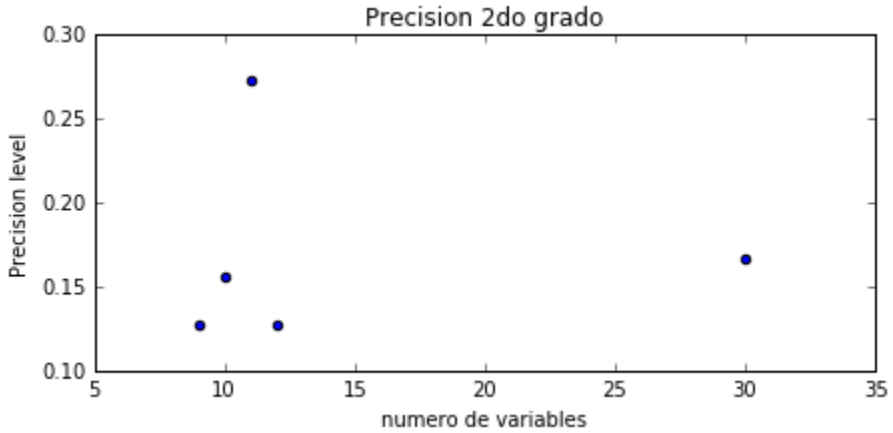


Gráfico de resultados en "Precision" para diferente número de variables en el ACP, Pol.2do°

Gráfico 6

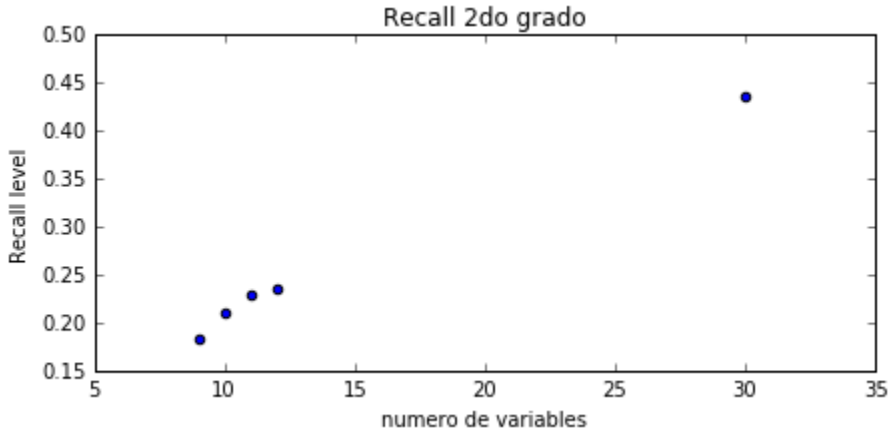


Gráfico de resultados en "Recall" para diferente número de variables en el ACP, Pol.2do°

Gráfico 7

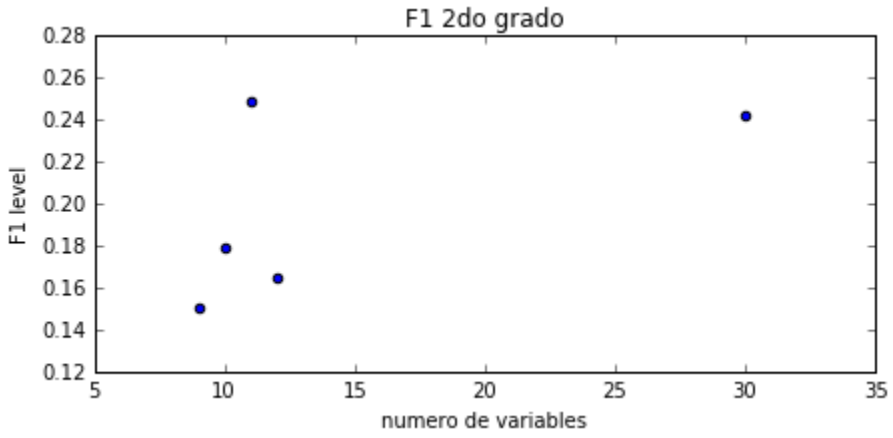


Gráfico de resultados en "F1" para diferente número de variables en el ACP, Pol.2do°

En estos gráficos se puede apreciar que las regresiones con polinomio de 2do grado tuvieron un desempeño muy por debajo de los modelos con polinomio de 1er grado en los 3 criterios que usaremos: Precision; Recall; F1. Esto puede ser debido a un "overfitting" en el modelo. La memoria de la computadora que se utilizó para correr los códigos no fue suficiente para correr algún modelo con polinomio de 3er grado. De aquí en adelante nos enfocaremos en los modelos logísticos con polinomio grado uno.

En los gráficos 2,3 y 4 se puede observar que los números de variables utilizadas en el ACP que mejores resultados tuvieron en la regresión fueron 15 y 20 variables, incluso con mejores resultados que la regresión logística de 1er grado con todas las variables.

Modelo 1. Regresión logística 1er grado con todas las variables (30 variables)

Sin utilizar la reducción por componentes principales y usando un polinomio de 1er grado para la regresión logística (RL), se obtuvieron los siguientes resultados:

Tabla 2

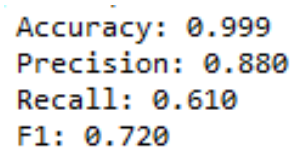
```
Accuracy: 0.999
Precision: 0.839
Recall: 0.604
F1: 0.702
```

Resultados RL, Pol.1er°, con todas las variables

Modelo 2. Regresión logística 1er grado con 20 variables en el ACP.

Para el modelo de RL con las 20 variables seleccionadas por el ACP y un polinomio de 1er grado,se obtuvieron los siguientes resultados:

Tabla 2



Resultados RL, Pol.1er°, con 20 variables.

De los resultados de estos dos modelos se puede decir que es posible hacer un análisis más "limpio" de los datos al reducir la dimensionalidad de estos. El nivel de "precision" de los categorizados fraudes por el modelo de RL aumentó en 4.1 pts. porcentuales, pasando del 83.9 al 88 por ciento al reducir la dimensionalidad. Mientras que el "Recall", que nos dice que porcentaje de los ya sabidos fraudes fueron encontrados por el modelo de RL, aumentó 0.6 pts porcentuales a 61%.

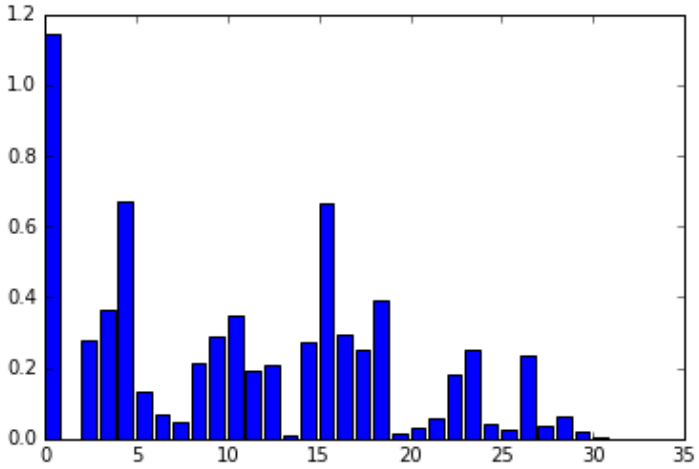
Selección de variables post-modelado

Este criterio de selección es sencillo, consta de fijar un nivel de umbral que cada quien considere adecuado, eliminando todas las variables cuyo coeficiente en el modelo de RL esté por debajo del umbral seleccionado.

Una vez que se reduce el número de variables se entrena el modelo una vez más con las variables restantes.

Modelo 3. Regresión logística 1er grado con todas las variables y reducción post-modelado

Gráfico 8



Coeficientes absolutos en la RL para las 30 variables originales más una variable de "unos"

Se seleccionó un umbral de 0.05 para el criterio de reducción de variables post-modelado, con este umbral se eliminaron 10 variables: [2,8,14,20,21,25,26,28,30,31].

Se volvió a entrenar el modelo de RL con las 20 variables restantes más una variable de "unos" y se obtuvieron los siguientes resultados:

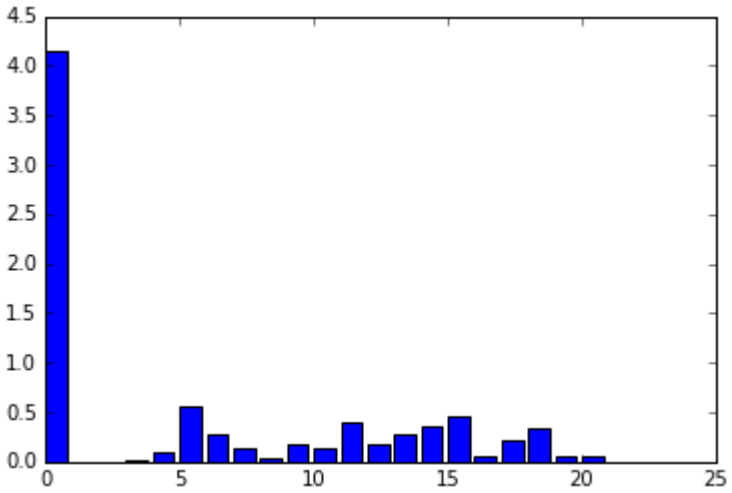
Tabla 3

Accuracy: 0.999
Precision: 0.887
Recall: 0.622
F1: 0.731

Resultados RL, Pol.1er°,con 30 variables iniciales y reducción a 20 post-modelado

Modelo 4. Regresión logística 1er grado con 20 variables seleccionadas en el ACP y reducción post-modelado

Gráfico 9



Coeficientes absolutos en la RL para las 20 variables del ACP más una variable de "unos"

Se seleccionó un umbral de 0.1 para el criterio de reducción de variables post-modelado, con este umbral se eliminaron 8 variables.

Se volvió a entrenar el modelo de RL con las 12 variables restantes más una variable de "unos" y se obtuvieron los siguientes resultados:

Tabla 3

Accuracy: 0.999
Precision: 0.888
Recall: 0.614
F1: 0.726

Resultados RL, Pol.1er°,con 20 variables en ACP y reducción a 12 post-modelado

Conclusiones

Al final el modelo que obtuvo el mejor resultado fue el modelo 3, el segundo lugar fue el modelo 4, a pesar de que el modelo 4 obtuvo una "precision" pobremente superior al modelo 3, el modelo 3 mostró mejor desempeño en el criterio de "Recall", dada la naturaleza de los datos analizados, se consideró que el principal criterio a juzgar sería justamente el "Recall" ya que es primordial que sean detectadas por el modelo las posibles transacciones fraudulentas. El modelo 3 tuvo un mejor desempeño que el 4 por 0.8 pts porcentuales, 0.008% que representa casi 4 fraudes del total de 492 fraudes de la base de datos completa. En tercer lugar, quedó el modelo 2 al que se le aplicó una reducción pre-modelado de 30 a 20 variables, finalmente en cuarto lugar, el modelo de RL con todas las variables sin ningún tipo de reducción.

Una conclusión que está a la vista es la practicidad y efectividad de la reducción de variables, ya sea post o pre modelado, gracias a esta reducción se reduce significativamente el poder computacional requerido, además de eliminar posible "ruido" por "overfitting" que resulta en un mejor entrenamiento del modelo de regresión logística. Por lo que un futuro siempre será considerado la reducción de variables al encontrarnos con bases de datos con tantas variables como esta.

Con la información obtenida en este documento, concluimos que aunque el modelo 3 con reducción post modelado fue mejor que el modelo 4 con reducción pre y post modelado, la mejora no fue tan significativa, teniendo una mejora en el "recall" de tan solo 0.8 pts porcentuales y una reducción de 0.1 pts porcentuales en la "precision" , pequeña mejora que debe ser pagada por un mayor nivel de procesamiento requerido.

El modelo 3 requirió que la primera regresión fuera con 30 variables, una regresión así requiere algo de poder computacional incluso para un polinomio de 1er grado, ni hablar si se aumentara el grado de polinomio, la segunda y última regresión de este modelo fue con 20 variables, número aún considerable.

Por el otro lado, en el modelo 4 la primera regresión fue ya con 20 variables, gracias a la reducción pre modelado, mientras que para la segunda solo se hizo con 12 variables, casi la mitad de las variables restantes en el modelo 3, a un costo bajo desde nuestro punto de vista. Al hacer reducción pre y post modelado con métodos claramente diferentes, se tiene una mayor oportunidad de reconocer la información verdaderamente representativa y valiosa, si la pérdida en el "recall", o en el criterio principal que se tenga en cada caso, no es tan grande, vale la pena considerar una reducción pre y post modelado para bases de datos grandes.