Instituto Tecnológico Autónomo de México

Machine Learning

Proyecto Final

Profesor:
Fernando Esponda

Equipo:

Rodrigo Andrés Morales Mendoza
124341

Mariana Godina Abasolo
113682

Sonia Mendizábal Claustro
105720

7 de diciembre de 2016

Índice

1. Introducción

Entre 2005 y 2006, Taiwán enfrentó una crisis económica causada por la falta de pago en tarjetas de crédito. Esta deuda acumuló un total de casi $270 billones de dólares, y más de un millón de personas se declararon incapaces de liquidar la deuda. Esto generó una crisis social en la población de Taiwán: de acuerdo a datos del Departamento de Salud[1] la tasa de sucidio en 2006 aumentó 22.0 % en comparación con el año anterior. Según la misma fuente, la causa principal fue el desempleo y la deuda en tarjeta de crédito, lo que colocó al territorio como el segundo país con mayor tasa de suicidio en el mundo.

Es importante mencionar que esta crisis fue provocada por la disminución de rigor por parte de las instituciones bancarias taiwanesas para otorgar tarjetas de crédito. En consecuencia, se generó una gran desconfianza en el sistema crediticio de Taiwán, lo que posteriormente representó un reto para los tarjetahabientes e instituciones bancarias. Es por esto que se hizo relevante en la agenda de las instituciones bancarias y económicas estudiar las fallas de pago desde distintas perspectivas, con el objetivo de reducir la incertidumbre.

Aunque es difícil comparar a México con Taiwán, la motivación de este trabajo surge de la situación actual en México. Recientemente Banco de México anunció que el país está en desaceleración, y aunado a esto, hace pocas semanas redujeron la esperanza de crecimiento del PIB para el 2017: el pronóstico de crecimiento pasó de un intervalo entre 2.3 % y 3.3 % a un intervalo entre 2.0 % y 3.0 %.

Esto es relevante, porque en épocas de crisis es común que la deuda de las persona aumente. Según el Banco de México en 2014, las tarjetas de crédito emitidas a marzo de este año sumaron 27.2 millones y el 56.9 % de los tarjetahabientes no pagaban el total del saldo. Lo que representaría un reto para el país, las instituciones bancarias y los consumidores mexicanos.

Motivados por esta problemática que enfrenta el país, decidimos estudiar si es posible predecir la probabilidad de impago de un cliente. Los resultados ayudarían en primer lugar a que un banco pudiera tomar mejores decisiones de entregar o no crédito a sus clientes. Más aún, el alcance de dichas predicciones podrían funcionar como un indicador de riesgo financiero para una economía, ya que indicarían, del total de una cartera el porcentaje potencial que haría impago. Cabe recalcar que el impago no es sinónimo de pobreza. Hay clientes que conscientemente deciden no pagar en una fecha, y hay algunos que incluso lo olvidan, aun cuando tuvieran la solvencia para hacerlo. Es importante esta consideración al momento de revisar las métricas del desempeño de nuestros modelos. Sin embargo, el contar con una cuantificación del nivel de impago de un cliente podría ser un camino para conocer la sanidad de una entidad o sistema financiero, a la par de permitir a la misma tomar mejores decisiones, sin considerar siquiera factores como rentabilidad de largo plazo o factores externos.

[1] Health Statistics in Taiwan (2008)

2. Metodología

2.1. Datos

Se realizó una búsqueda en diversas páginas de bases de datos abiertas que nos pudieran servir para dicho análisis.

La mejor base de datos que se encontró fue una con el fallo de pago (o impago, en inglés, "default") de tarjetas de crédito en Taiwan de 2005. La estructura de los datos nos pareció completa, afín a lo visto en el curso, satisface las características del proyecto, y nos pareció un reto interesante.

2.2. Descripción de los Datos

Los datos obtenidos incluyen características demográficas, historial crediticio, información de crédito y saldo de tarjetas de crédito desde abril de 2005 a Septiembre de 2005 correspondientes a 30,000 tarjetahabientes. La dimensión de la base es de 30,000 observaciones, con 23 variables explicativas.

- Edad: años cumplidos
- Género: 1 = hombre, 0 = mujer.
- Educación: 1 = posgrado, 2 = universidad, 3 = preparatoria, 4=otros, 5=desconocido).
- Estado Civil: 1=casado, 2=otros.
- `LIMIT_BAL`: Monto de crédito otorgado en dólares de Nuevo Taiwán (NTD ó NT$).
- $\texttt{PAY}_i$: Historial de pagos atrasados durante los 6 meses anteriores.
- $\texttt{BILL\_AMT}_i$: Historial de monto de deuda en la cuenta durante 6 meses anteriores.
- $\texttt{PAY\_AMT}_i$: Historial de monto de pagos a la cuenta durante 6 meses anteriores.

Con $i = 1, \ldots, 6$ equivalente a los meses septiembre, agosto, julio, junio, mayo y abril respectivamente, es decir, un mes antes, dos meses antes y así sucesivamente.

2.3. Limpieza y transformación

La base de datos era completa, por lo que realmente no hubo que hacer limpieza de los datos, sino más bien transformación de algunas variables para hacerlas compatibles con los modelos en los sistemas usados.

La transformación de la base consistió en recodificar variables para poder usarlas como input en los modelos. En particular, la variable Estado Civil se modificó para únicamente considerar si es casado o no lo es. Las variables monto de crédito otorgado, monto de deuda y monto de pagos se estandarizaron para quitar problemas de escala.

Además se crearon nuevos rasgos considerando la naturaleza de cada una de las variables:

- $\texttt{AMT}_i\texttt{\_PBILLIM}$: proporción de monto de deuda respecto al crédito otorgado en el i-ésimo mes.
- $\texttt{AMT}_i\texttt{\_PPAYBIL}$: proporción de pago respecto monto de deuda en el i-ésimo mes.
- `maximum.delay`: número máximo de pagos no realizados considerando el historial de pagos atrasados.

Posteriormente, estas variables se incluyeron en un Random Forest de clasificación con el objetivo de obtener la importancia de Gini, para una selección de variables.

Además, en algún momento fue necesario transformar algunas variables a no-negativas para que pudieran funcionar como input en el modelo Bayes Ingenuo Gaussiano (Gaussian NB()).

2.4. Modelos

Como primer acercamiento, se realizó un modelo Multinomial de Bayes Ingenuo con todas las variables. Como es necesario para este modelo, se tuvieron que usar variables categóricas, por lo que fundamentalmente se usó:

- Género
- Educación
- Edad (para ésta se hicieron grupo de edad para no tener excesivas categorías), nos basamos en los histogramas y proporciones de sí-no en estas edades para definir los grupos. Fueron 11 categorías de edad en total, con interavlos iguales, salvo para los más grandes.
- Estado civil
- Pay_0 a Pay_6: historial de pagos atrasados. Al final quedaron 10 grupos en cada variable.
- BillAmt_0 a `BillAmt_6` y `PayAmt_0` a PayAmt_6: historial de pagos atrasados y de pagos durante meses anteriores. Al igual que para edad, se hicieron grupos basados en distribuciones. En total se hicieron 4 categorías para cada variable.

Se obtuvo el siguiente resultado preliminar:

Multinomial NB() Performance over the training set de Categorical Variables:	
precision =	0.57
recall =	0.37
Accuracy Score =	0.79

Multinomial NB() Performance over the TEST set de Categorical Variables:	
precision =	0.57
recall =	0.37
Accuracy Score =	0.80

Dado los resultados prometedores, se pasó a usar otros modelos más robustos, como Random Forest o SVM. Se obtuvieron resultados cercanos a los del Multinomial NB() - se omiten los cuadros, excepto el del mejor modelo obtenido, por no ser tan relevantes, con variables categóricas o numéricas sin mayor "feature engineering".

Random Forest, Performance sobre Training set de Categorical values:	
precision =	0.97
recall =	0.84
Accuracy Score =	0.96

Random Forest, Performance over the TEST set de Categorical values:	
precision =	0.58
recall =	0.33
Accuracy Score =	0.80

Los resultados obtenidos sugirieron que la información que tenemos no permite una mejor predicción o bien que una componente importante del default en las tarjetas de crédito está explicada por alguna variable omitida en el modelo (por carecer de ella).

Llama poderosamente la atención que los modelos que incorporaban a las variables numéricas no mostraban un mejor comportamiento que las categóricas, por lo que hasta aquí, el mejor modelo aparentemente era Random Forest con categóricos. Sin embargo, esto no siguió una metodología estricta, sino que formó parte de una primer prueba exploratoria de los datos.

Es importante mencionar que para hacer una selección de las variables más explicativas existen distintos métodos. Nuestra base de datos cuenta con una cantidad considerable de variables, y aún más si se considera el número de variables (categóricas y numéricas) que habíamos agregado en la etapa de feature extraction (recordar que hay 2^n combinaciones posibles sólo considerando a las variables a tomar en el modelo). Es por esto que se decidió realizar un Random Forest para seleccionar variables. Primero se realizó un modelo que incluye las variables centradas numéricas y demográficas. Se decide excluir variables siguiendo la medida de importancia de Gini, que en este caso principalmente excluye las variables demográficas, que a excepción de la edad, parecen no aportar información para discriminar entre el pago y no pago de la deuda.

En el segundo Random Forest se incluyen únicamente las variables centradas y los nuevos rasgos. De esta selección se dejaron únicamente las 11 variables con importancia de Gini más alta.

Con estas variables se realizaron modelos de clasificación. Los dos modelos con mejores resultados fueron con Redes Neuronales (NN) y Máquinas de Soporte Vectorial (SVM).

Curvas de Aprendizaje

Para saber con qué tamaño de conjunto de datos debíamos trabajar para optimizar nuestros resultados, hicimos curvas de aprendizaje, básicamente sobre todos los modelos y conjuntos de datos. Prácticamente todos nos dieron el mismo resultado: el conjunto de datos no aportaba más información arriba de 10mil o 15mil datos, por lo que nuestros conjuntos de aprendizaje estaban bien hechos, pero con un poco extra de información innecesaria. Esto es importante porque en una implementación, sabríamos aproximadamente qué muestreo usar para optimizar los tiempos de entrenamiento.

Mostramos en la Figura 1 solamente la curva Aprendizaje de Árboles Aleatorios para variables categóricas, por poner un ejemplo. En el anexo se pueden ver otras gráficas.

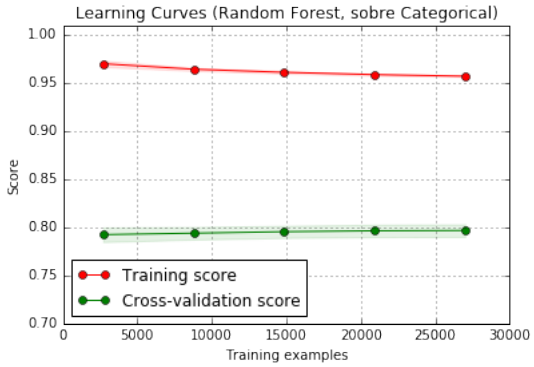

Figura 1: Random Forest, Categorical variables

Redes Neuronales

El modelo de redes neuronales se ajustó con las variables selecciondas previamente mencionadas. Se probaron varios modelos con distintos parámetros. En la primera prueba se usaron los siguientes parámetros:

- 11 nodos de entrada
- 1 nodo de salida
- 0 capas ocultas
- Función de perdida cuadrática
- Función sigmoide
- 10,000 corridas
- gradiente descendiente .01

con estos parámetros se obtiene un ajuste de 0.78 de *accuracy*.
En la segunda prueba se usaron los mismos parámetros que el anterior, excepto por la función sigmoide, que se cambia por la función tangente hiperbólica (tanh) y se obtiene un ajuste de 0.22, lo cual es una pérdida significativa.
Se decidió entonces usar la función sigmoide, probar el modelo con la función de pérdida de entropía y aumentar el gradiente descendiente a 0.05. Con estos parámetros, el modelo alcanza un ajuste de 0.81, lo cual se acerca a RF y SVM.
Se ajustó el modelo con 1 capa oculta extra y distintos nodos intermedios, es decir, 6 y 8 pero el ajuste sólo mejora por .02, lo cual bien se podría deber al muestreo aleatorio del training y test set.

Por último, para construir el perfil demográfico de los deudores (6,636 observaciones) se usó la técnica no supervisada clusters jerárquicos considerando únicamente variables demográficas. El objetivo es contrastar la probabilidad obtenida del mejor modelo e identificar grupos de mayor riesgo. La clasificación se realizó con clusters jerárquicos y se obtuvieron distintos dendogramas variando el método de asociación: simple, completo y ward. En cada dendograma se realizaron distintos cortes para obtener de 3 a 6 grupos distintos. Finalmente, el método de asociación ward con 3 grupos es el que se seleccionó por tener una interpretación sencilla. En la Figura 5 al final del documento, se incluyen los dendogramas para los diferentes métodos con diferentes cortes.

3. Resultados

El primer Random Forest realizado tiene ajuste o *accuracy* de prueba de 79 % y precisión 60 %. En la Gráfica 4 en el apéndice al final del documento, se presenta la importancia de Gini para cada variable de este modelo (Random Forest 1). Profundizando en el hallazgo de omitir variables demográficas, se realiza un pequeño descriptivo de las variables demográficas contra el hallazgo y se presenta en la Figura 2. No se observa una diferencia importante entre grupos.

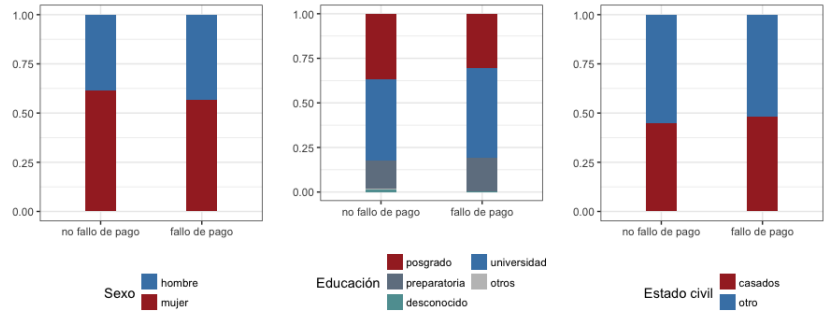

Figura 2: Descriptivo de demográficos.

De la misma forma que el primer Random Forest, se usa la importancia de Gini para evaluar la similitud de las variables, que se presenta en la Figura 4 incluida en el apéndice (Random Forest 2). Este modelo tiene un ajuste de prueba de 81 % y precisión 65 %.

Máquinas de Soporte Vectorial

Los resultados obtenidos con SVM después de volver a transformar las variables y seleccionar las 11 mejores fueron:

SVM con kernel Gaussiano y sobre Training set de variables selected:	
precision =	0.70
recall =	0.32
Accuracy Score =	0.82

SVM con kernel Gaussiano y sobre Test set de variables selected:	
precision =	0.70
recall =	0.34
Accuracy Score =	0.81

Así que podemos ver una mejoría, pero no es un cambio significativo con respecto al Random Forest sobre variables categóricas. La mayor mejoría se ve con *precision*, que aumenta casi 15 por ciento, por lo que recomendamos usar este modelo o bien Redes Neuronales, con la desventaja de que fue el modelo que más se tardó en entrenar. Dado que los tiempos de predicción son muy rápidos tanto para SVM como para NL, y porque en este contexto podemos esperar que el hecho de que se tome un largo tiempo entrenando no es problema, puesto que este tipo de datos se podría entrenar mensualmente sin mayores pérdidas para la compañía, recomendamos usar estos modelos para la predicción de préstamos.

Todavía más, dado que ambos modelos producen probabilidades como output, se puede fijar un valor de corte que aumente el *recall* y sacrifique un poco la precisión, si la cartera fuera tal que el costo de predecir erróneamente fuera muy alto. Sin embargo, sabemos que los bancos prefieren otorgar créditos con un poco

de riesgo, puesto que esto genera altos rendimientos en el mediano y largo plazo, ya que hay un mercado de carteras vencidas. Esta es la razón por la cual no nos preocupamos por aumentar el *recall*, mientras que sí procuramos mantener una precisión alta. Ésta todavía se podría aumentar más disminuyendo el *recall* con el mismo razonamiento, pero sin contar con una fórmula que represente los costos, consideramos que un rango de 60/30 es razonable para estas métricas. Queda la posibilidad de que algún método de validación cruzada con regularización aumente el desempeño del modelo. Esto no se hizo debido a que consideramos que los resultados son suficientemente buenos, toda vez que no contamos con una métrica que nos ayude a especificar un umbral de decisión óptimo.

Redes Neuronales

El modelo selecionado nos arroja las siguientes medidas:

Neural Network over Test Set of Selected Variables:	
precision =	0.65
recall =	0.38
Accuracy Score =	0.81

Lo cual es similar al SVM sobre Selected variables, con un poco mejor *recall*, pero perdiendo un poco de *precision*. Nuevamente, como no conocemos los costos de una cartera de clientes, no podemos determinar que uno u otro modelo sea mejor en función de estos números.

Demográficos

Los grupos demográficos obtenidos se describen a continuación, entre paréntesis se incluye el tamaño relativo:

- Grupo 1: (47 %) grupo de jóvenes de ambos sexos, menores de 34 años con nivel de educación universitario principalmente y la mayoría.
- Grupo 2: (34 %) grupo de adultos menores de 50 años, la mitad de los miembros casados y con nivel de educación preparatoria y universidad.
- Grupo 3: (19 %) grupo más viejo principalmente mayores de 48 años, con menor preparación escolar y en su mayoría casados.

Es interesante notar que el grupo deudor más grande es el de gente joven y principalmente mujeres.

Del ejercicio de grupos demográficos con la técnica no supervisada, finalmente no se observa una distribución distinta en el comportamiento de las probabilidades. En la gráfica de la Figura 3 se puede observar la distribución de la probabilidad obtenida por grupo. Aunque el tercer grupo tiene menor densidad en las probabilidades bajas, en general, se confirma lo visto en el primer modelo de selección de variables en el que no existe una relación fuerte con variables demográficas, incluso únicamente considerando la muestra de fallo.

4. Conclusiones

La base de datos original era buena, pues no se requirió de hacer una limpieza, sino más bien transformaciones a las variables para hacerlas legibles para los modelos usados.

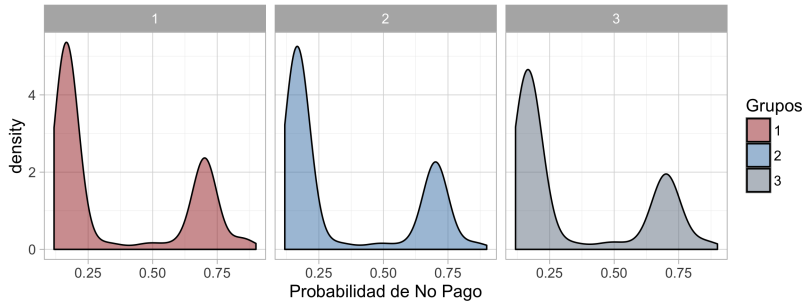

Figura 3: Densidad de probabilidades por grupo.

Con respecto a selección de variables, se encontró que las variables permiten hacer una predicción razonable de los deudores. Sin embargo, incluso después de un trabajo de selección de variables, éstas tienen un límite en los modelos. Además, este límite hace que sea muy difícil rebasar un umbral de alrededor del 80 % de *accuracy*, 30 % de *recall* y 65 % de *precision*.

El perfil demográfico no es fuerte para discriminar riesgos de no pago, lo cual es contra-intuitivo. Sin embargo sí se observó que en particular los jóvenes tienden a caer en la falla de pago. Esto abre un área de oportunidad para incentivar a los jóvenes y encontrar medidas que induzcan al pago de los créditos, lo que disminuiría el riesgo.

En segundo lugar, se puede ver que el modelo que se elija no es el tema crucial en la predicción, ya que todos funcionaron razonablemente bien desde el principio.

Como es habitual, el modelo de Bayes Ingenuo logra un ajuste muy bueno, dada la simpleza del modelo. Asimismo, el SVM logra resultados muy buenos, similares a los de una Red Neuronal, pero en un tiempo significativamente menor.

Es interesante observar que las variables categóricas funcionaron consistentemente mejor que las numéricas, sugiriendo que no es necesario tener un desglose de micro-datos para entrenar a los modelos, además de que éstas permiten una interpretación mucho más sencilla de los resultados, a costa de una posible pérdida de información, principalmente en la selección del tamaño de los intervalos y por ende del número de categorías. Sin embargo, como se ve en los resultados, una selección educada permite un balance justo entre información relevante e interpretabilidad.

Cabe la posibilidad de que la selección de variables no haya sido la mejor, pues basamos nuestra decisión en el coeficiente de Gini, que en esencia tomará en consideración correlaciones entre las variables y la respuesta. Podría ser que una combinación particular de variables permitiera un ajuste muy bueno para SVM, por ejemplo, pero esto no es detectado por el Random Forest, ya que toma las más importantes.

Se logra la predicción de riesgo de impago deseado, además de que en el camino se descubren características importantes que permiten interpretar mejor los resultados y que dan cabida a un conocimiento más profundo de los clientes en un banco.

4.1. Sugerencias de trabajo futuro

En primer lugar, se podría hacer selección de variables tipo backwards y forward para ver si se obtienen mejores resultados con algún modelo. Esto resultaría muy costoso (en tiempo) para modelos como el SVM

o Neural Network, pero sería factible con Random Forest o Bayes Ingenuo, y después se podría probar esas variables con el SVM y NN para ver si se mejora el desempeño.

Además de esto, no descartamos que puedan existir más transforamciones de datos que podrían resultar interesantes, como los pagos entre la edad. Dichas transformaciones apuntarían a condensar información en menos variables, por lo que una técnica natural si se quisiera esto, sería usar PCA sobre variables correlacionadas entre sí.

Finalmente, como meta-proyecto, se podría usar estas predicciones para predecir riesgo sistémico de bancos, lo cual serviría como un indicador del riesgo país si se considerara el número total de deudores con respecto a la población total, y sería un indicador interesante de la estabilidad financiera en una economía. Hasta donde tenemos conocimiento, no existe dicho indicador, al menos no en los términos en que lo proponemos.

5. Apéndice

5.1. Importancias

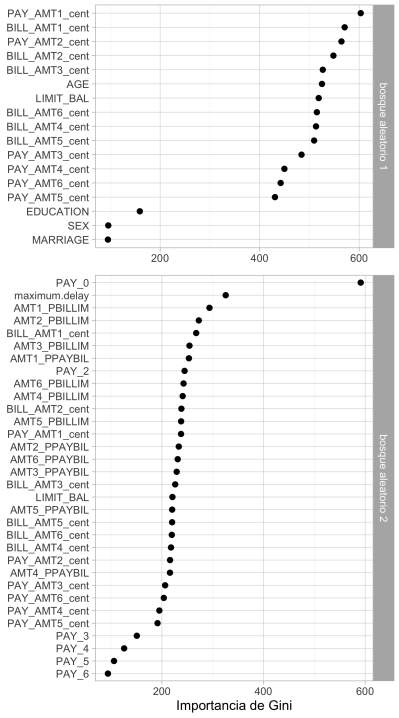

Figura 4: Importancias de Gini para aelección de variables.

5.2. Dendogramas

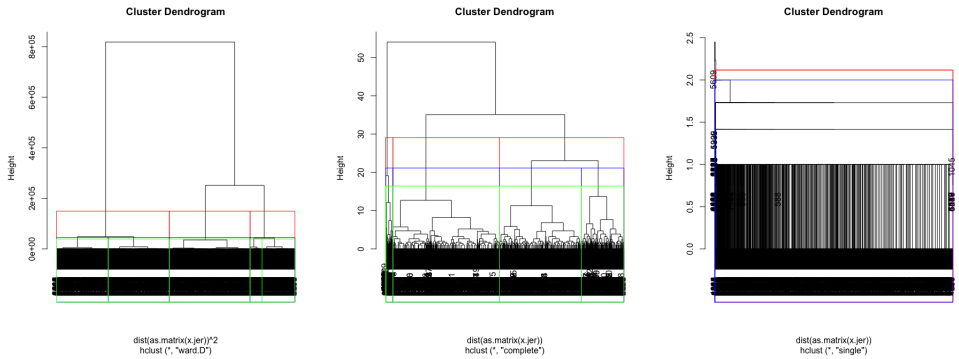

Figura 5: Dendogramas con diferentes métodos de asociación.

5.3. Desempeño de los modelos

Multinomial NB()Performance over the training set de Categorical Variables:	
confusionmatrix =	[[16100, 1415]
	[3129, 1856]]
precision =	0.567410577805
recall =	0.372316950853
Accuracy Score =	0.798044444444

Multinomial NB() Performance over the TEST set de Categorical Variables *- the important one -* :	
confusionmatrix =	[[5388, 461]
	[1032, 619]]
precision =	0.573148148148
recall =	0.37492428831
Accuracy Score =	0.800933333333

———————-

Random Forest, Performance sobre Training set de Categorical values:	
confusionmatrix =	[[17387, 134]
	[808, 4171]]
precision =	0.96887340302
recall =	0.837718417353
Accuracy Score =	0.958133333333

Random Forest, Performance over the TEST set de Categorical values *- the important one -* :	
confusionmatrix =	[[5451, 392]
	[1108, 549]]
precision =	0.583421891605
recall =	0.331321665661
Accuracy Score =	0.8

———————-

SVM con kernel Gaussiano y sobre Training set de variables selected:	
confusionmatrix =	[[16889, 680],
	[3324, 1607]]
precision =	0.70266724967205951
recall =	0.32589738389778949
Accuracy Score =	0.82204444444444447

SVM con kernel Gaussiano y sobre Test set de variables selected: :	
confusionmatrix =	[[5551, 244],
	[1123, 582]]
precision =	0.70460048426150124
recall =	0.34134897360703814
Accuracy Score =	0.81773333333333331

5.4. Learning curves

Se muestran las curvas de aprendizaje de los modelos con mejor desempeño.

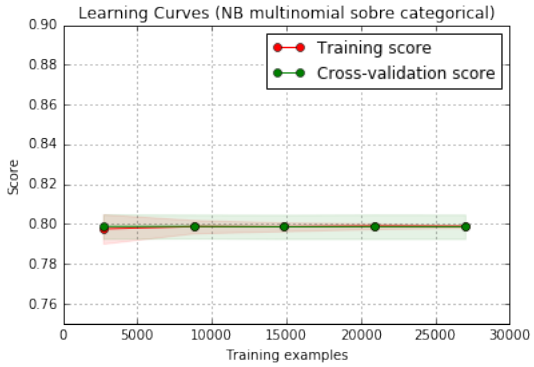

Figura 6: Bayes Ingenuo, Multinomial, Categorical variables

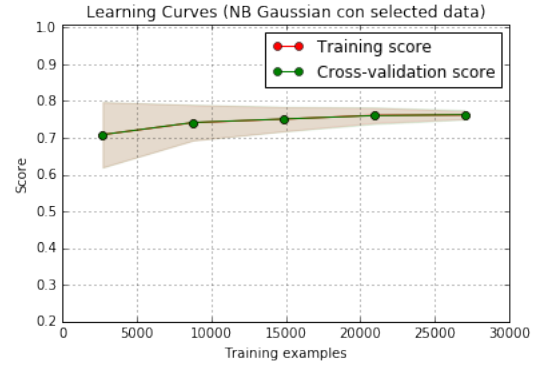

Figura 7: Bayes Ingenuo, Gaussian, Selected variables

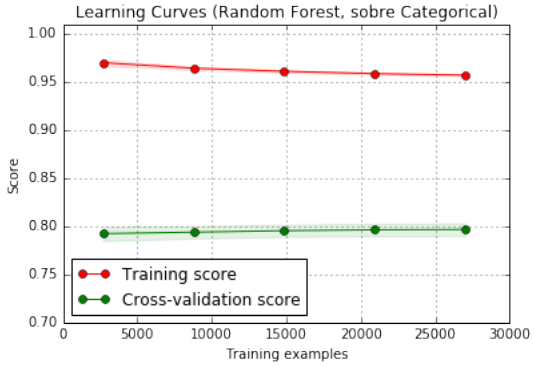

Figura 8: Random Forest, Categorical variables

Figura 9: Random Forest, Selected variables

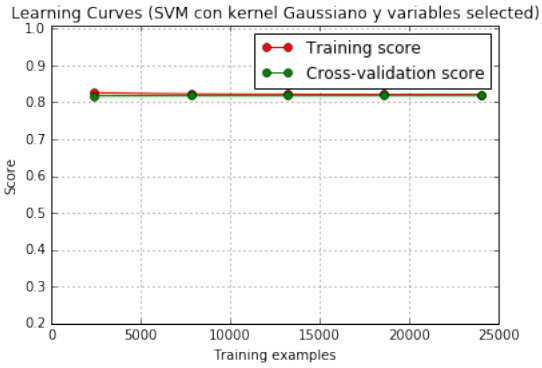

Figura 10: Support Vector Maschine, Selected variables

6. Referencias

1. Lichman, M. (2013). *UCI Machine Learning Repository*. http://archive.ics.uci.edu/ml. Irvine, CA: University of California, School of Information and Computer Science.

2. Yeh, I. C., and Lien, C. H. (2009). *The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients.* Expert Systems with Applications.

3. Kaggle. *Default of Credit Card Clients Dataset* Default Payments of Credit Card Clients in Taiwan from 2005.

   https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset/version/1.

4. Taiwanese Department of Health (2006) *Health Statistics in Taiwan 2006.*

5. Wang, Erik (2006) *Taiwan's Credit Card Crisis.*

   http://sevenpillarsinstitute.org/case-studies/taiwans-credit-card-crisis

6. Forbes Staff. *La lista negra de las tarjetas de crédito*. Forbes Online 2014.

   <http://www.forbes.com.mx/la-lista-negra-de-las-tarjetas-de-credito/#gs.kE1BQSs>