Universitat de les Illes Balears

Facultad de Economía y Empresa

Aprendizaje Estadístico y Toma de Decisiones I

Problema de Clasificación

"Defaults of Credit Cards Clients"

Master en Análisis de Datos Masivos

Carlos Khali Delgado Moujahid

2017

DNI del alumno: 53261880K

Departamento de Economía Aplicada, Historia e Instituciones Económicas

Palabras clave del trabajo: Análisis de datos, predicción, modelos econométricos, clasificación, aprendizaje supervisado, estadística, toma de decisiones.

Índice de Contenidos

Resumen

Gobiernos, empresas, agencias y universidades de todo el mundo conocen la importancia de los datos, más ahora en el siglo XXI y después de vivir la revolución de la información, los datos son una realidad para poder describir, analizar y sobretodo predecir el futuro.

En el presente trabajo se va a tratar la importancia de los datos desde el punto de vista académico, donde el objetivo va a ser valorar las predicciones en dos tipos de bases de datos y con diferentes objetivos. En la base de datos *"default of credit card clients.xls"* el objetivo será predecir la probabilidad que tiene un cliente de pagar o no su tarjeta de crédito (esta probabilidad la podremos usar para clasificar a clientes), los datos están recogidos en Taiwán.

Finalmente, se mostrará los resultados y el modelo que mejor predice y optimiza nuestras variables objetivo.

Abstract

Governments, companies, agencies and universities around the world know the importance of data, more now in the XXI century and after the information revolution, data is a reality to be able to describe, analyze and over all predict the future.

In this document the data will be treated from an academic point of view, where the objective will be to value the predictions in two types of databases with different objectives. In the database *"default of credit card clients.xls"* the objective will be to predict the probability that a customer has to default his credit card (this probability can be used to classify clients), the data are collected in Taiwan.

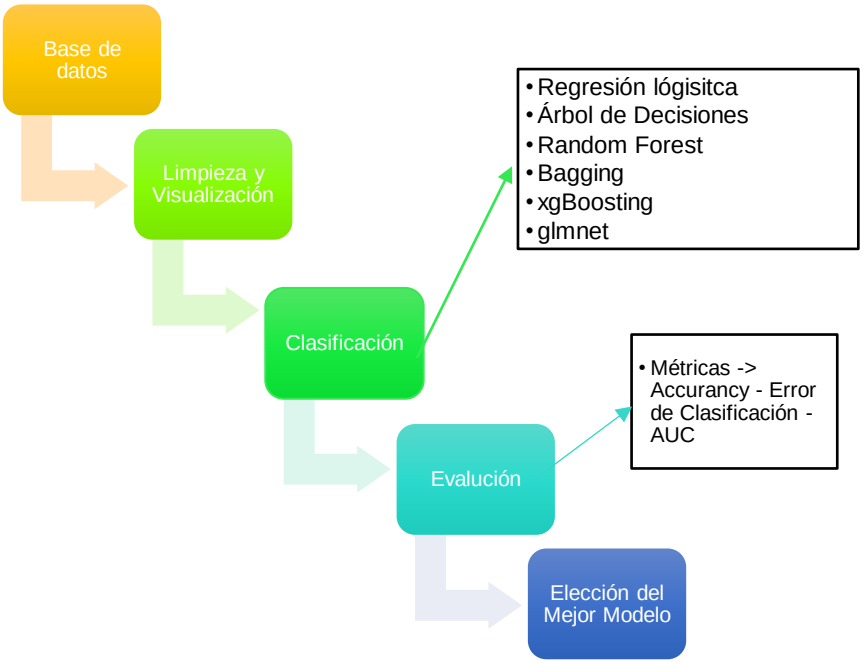
Finally, it will be show the results and the model that predicts and optimizes our target variables on the best way.

Descripción de la metodología

La metodología consistirá en seguir el siguiente procedimiento para el análisis de los datos:

1. Descripción de la base de datos.
2. Análisis la base de datos.
3. Limpieza de la base de datos.
4. Proposición de una teoría económica.
5. Aplicación de técnicas de aprendizaje estadístico.
6. Comparación de resultados.
7. Toma de decisión.
8. Conclusión.

En rasgos generales, se clasifican los datos usando diferentes clasificadores y se compara su exactitud, precisión, el error de entrenamiento y prueba para encontrar el mejor clasificador para los datos.



También vamos a utilizar una validación cruzada de $k = 10$ para evitar el sobreajuste de los datos en nuestros.

Herramientas usadas

En todo momento se utilizarán RStudio (versión 1.1.383 para MacOS, equivalente a la versión 3.4.2) para analizar los datos. Junto con el trabajo se adjuntará dos RScripts uno para cada caso, donde se podrá visualizar el código, los *"packages"* y los métodos utilizados.

Análisis "*Default of Credit Card Clients*"

Introducción

El objetivo es predecir el cumplimiento o no cumplimiento del pago de las tarjetas de crédito en Taiwán, además compararemos la precisión predictiva de la probabilidad de incumplimiento entre diferentes métodos utilizados en aprendizaje estadístico y econometría.

Cuando una persona solicita y recibe una tarjeta de crédito, está asumiendo una gran responsabilidad. La compañía bancarias que emiten las tarjetas de crédito evalúa la solvencia crediticia y les otorga una línea de crédito acorde a las características y capacidades que tiene cada cliente, en este momento la compañía emisora está asumiendo un riesgo inherente a la contratación donde puede ocasionarse impagos. Es cierto, que la mayoría de personas usarán su tarjeta de crédito para hacer compras que luego podrán pagar a la entidad que les ha ofrecido la tarjeta de crédito, pero por alguna razón u otra algunos clientes no cumplen los pagos estipulados por las entidades de crédito y finalmente realizan *"Default"* e incumplen con el pago de la tarjeta de crédito.

"Credit Card Default" es el término que se usa para describir lo que ocurre cuando un usuario de tarjeta de crédito realiza compras al cargarlas en su tarjeta de crédito y que luego no paga su factura.

La probabilidad de que un cliente incumpla con su deuda de tarjeta de crédito es muy importante para los bancos, ya que puede resultar un gran coste para la entidad si el nivel de *Default* es muy elevado, con un modelo predictivo preciso, los bancos pueden evitar emitir tarjetas de crédito a clientes que posteriormente incumplirán, y también los bancos pueden evitar rechazar clientes que realmente cumplirán con los pagos a tiempo.

Descripción de la Base de Datos

Los datos se centran en la relación entre el "Default" y los resultados de las decisiones financieras que los consumidores toman dentro de las limitaciones de los términos contractuales establecidos por los emisores de tarjetas de crédito.

Nuestro conjunto de datos contiene la información muy detallada incluso sobre el comportamiento y las características de uso de la tarjeta de crédito.

Data set credits: I-Cheng Yeh

Department of Information Management, Chung Hua University, Taiwan; Department of Civil Engineering, Tamkang University, Taiwan.

UCI Machine Learning Repository http://archive.ics.uci.edu/ml. Irvine, CA: University of California, School of Information and Computer Science.

Tamaño Dataset: 30,000 filas; 25 columnas (11 columnas de valores enteros (Integer), 14 columnas numéricas)
Es decir, la base de datos está formada por 30,000 observaciones y 25 variables que se describen a continuación:

default.payment.next.month: La **variable respuesta** (pago es *Default* en el próximo mes) es binaria, es decir: (*0→ No Default, 1→ Default*).

ID: Identificación de cada cliente

LIMIT_BAL*:* monto del crédito concedido (NT dólar): incluye tanto el crédito individual al consumidor como el crédito familiar (suplementario).

SEX: Sexo (1 → hombre, 2 → mujer).

EDUCATION: Educación (1 → Postgrado, 2 → Universidad, 3 → Secundaria, 4 → otros).

MARRIAGE: Estado Civil (0 → Divorciado, 1 → Casado, 2 → Soltero, 3 → Otros).

AGE: Edad (en años).

PAY_0 - PAY_6: Historial del pagos, es decir, estado del reintegro desde Abril a Septiembre de 2005, es decir, *PAY_0 → Situación de devolución en Septiembre, PAY_2 → Situación de devolución en Agosto,…, PAY_6 → Situación de devolución en Abril de 2005,* respectivamente. La escala de medición para el estado de devolución del crédito es la siguiente:

-2 → no ha habido consumo;
-1 → pagado correctamente;
0 → uso del crédito renovable;
1 → retraso de pago de un mes;
2 → retraso de pago de dos meses; . . .; 8 → retraso de pago de ocho meses;
9 → retraso de pago de nueve meses o más.

BILL_AMT1 - BILL_AMT6: Balance de la cuenta corriente (NT dólar), desde Abril a Septiembre 2005, es decir, *BILL_AMT1 → Balance en Septiembre, BILL_AMT2 → Balance de Agosto, ..., BILL_AMT6→ Balance de Abril de 2005*, respectivamente.

PAY_AMT1 - PAY_AMT6: Importe de pagos previos (NT dólar), desde Abril a Septiembre de 2005, es decir, *PAY_AMT1 → Importe pagado en Septiembre, PAY_AMT2 → Importe pagado en Agosto,…, PAY_AMT6 → Importe pagado en Abril de 2005,* respectivamente.

Modelo Teórico-Económico

El enfoque básico utilizado por las entidades de crédito es que si hay un pago adeudado en la cuenta corriente correspondiente y/o si el pago no se realiza en el periodo estipulado, se establece el indicador de retraso correspondiente.
Por lo tanto, si hay indicadores de demora establecidos y el pago no se realiza finalmente, el cliente se clasificará como *"Default"*, en caso contrario, el cliente se clasifica como *"No Default"*.

Desde un punto de vista económico-teórico los indicadores que nos pueden ayudar a predecir el impago podrían ser el sexo *(SEX)*, nivel educativo *(EDUCATION)*, estado civil *(MARRIAGE)* serán variables cualitativas significativas en el modelo y combinado con el nivel de crédito concedido, forma de consumir y los retrasos en los pagos *(PAY_X)* serán variables significativas.

En el caso del balance en la cuenta corriente *(BILL_AMTX)*, es una variable que la lógica nos puede llevar a pensar que es importante en el modelo ya que nos mide la riqueza, no obstante, el balance mes a mes no es significativo y si recibes ingresos normalmente son gasto o ahorro y mes a mes la variable suele estar muy correlacionada ya que depende del saldo inicial, gasto y ahorro, por este motivo creo que no es una variable importante en nuestro modelo.

Por último los pagos previos pueden ser interesantes para ver si el cliente cumple con regularidad con los pagos, si los pagos previos *(PAY_AMTX)* coinciden con el consumo de la tarjeta de significará *No Default*, si las cantidades de los pagos son inferiores al consumo de la tarjeta de crédito mes a mes aumentará la probabilidad de *Default.*

Limpieza de Datos y Visualización

En este paso he realizado la limpieza y visualización de algunas cosas que pueden ser interesante para nuestro estudio.

Proceso de limpieza y visualización:

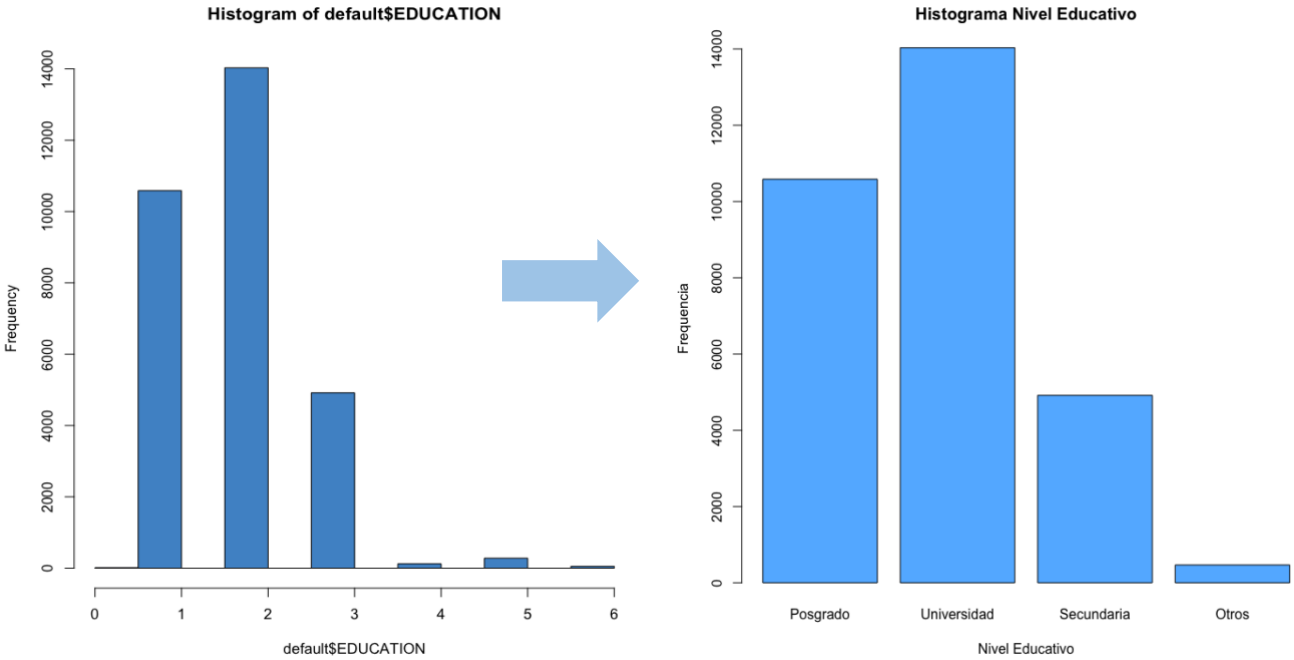
1. Eliminar la columna ID (no es necesaria)

2. Renombrar dos variables: *default.payment.next.month* $\rightarrow$ *DEFAULT* y *PAY_0* $\rightarrow$ *PAY_1*

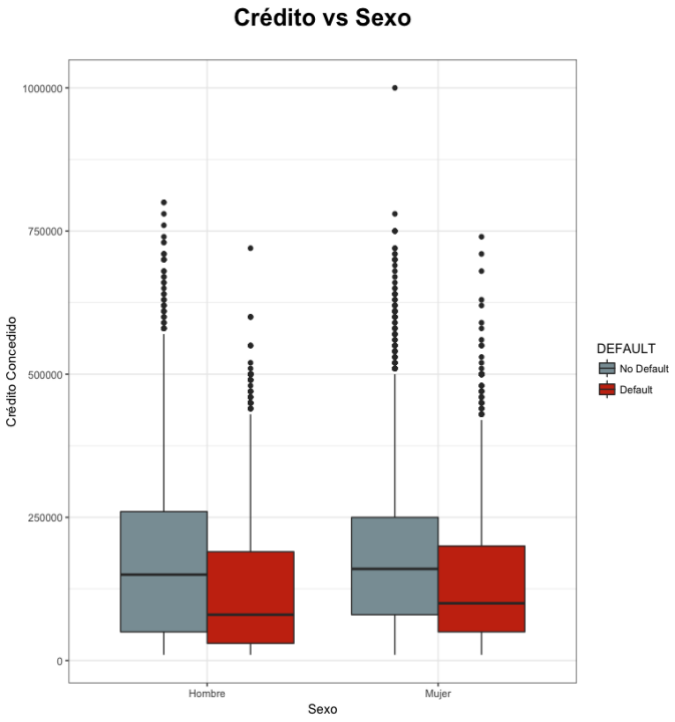
3. Reasignar factores no definidos en la variable *EDUCATION* en concreto los valores 0, 5 y 6 no están definidos. Como son pocos los voy a reasignar con el valor 4 que pertenece a la categoría *"Otros"*.

4. Factorizar y asiganar la clase que les corresponde a cada variable cuantitativa: *SEX, MARRIAGE, EDUCATION y DEFAULT.*

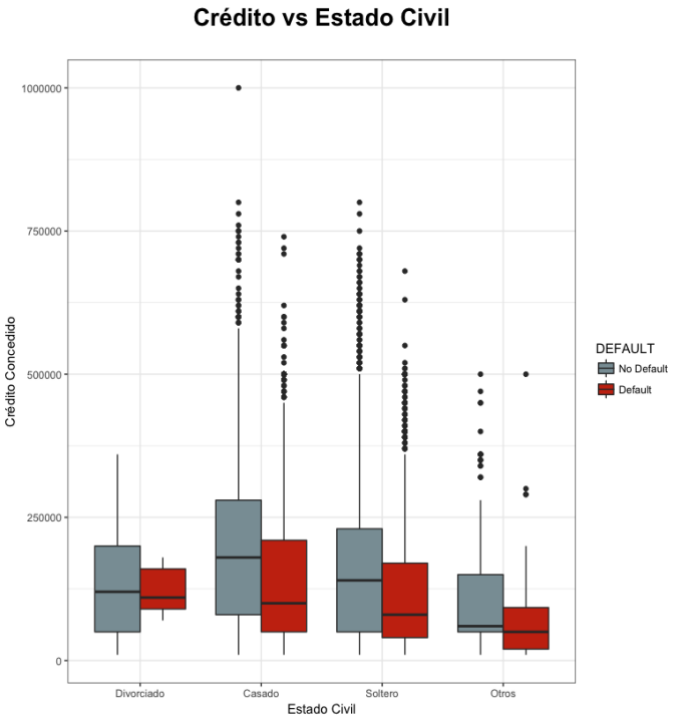
Ejemplo del paso 3 y 4:



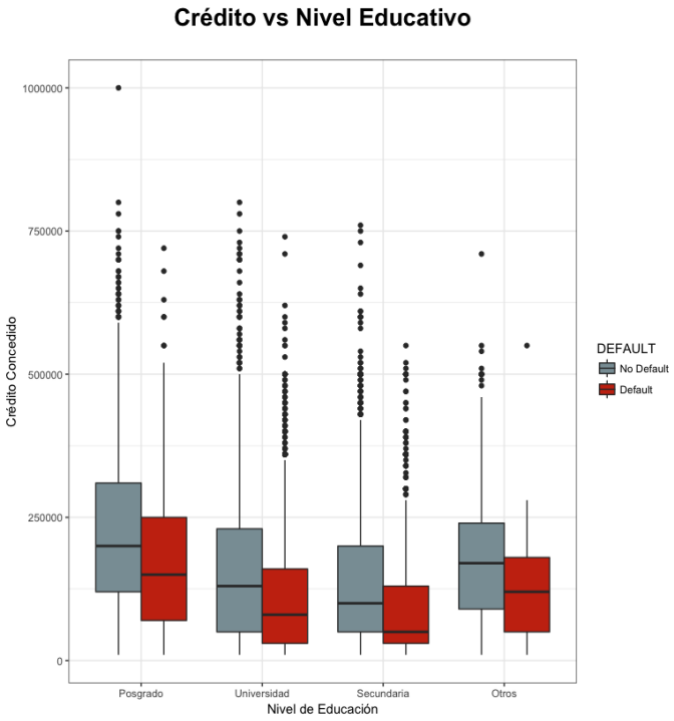
5. Realizar varias gráficas para visualizar por categorías como se distribuyen el número de *Default* y *No Default* sobre las variables categóricas vs su nivel de Crédito Concedido (LIMIT_BAL).



La distribución entre hombre y mujeres es muy parecida, lo cual nos sorprende ya que esperábamos que a mayor crédito concedido mayor probabilidad de *Default*, aun así se observa que el rango intercuartílico es superior en el caso de las mujeres que propician *Default* en comparación con los hombres como esperábamos. Por otro lado, los datos atípicos están situado en la parte superior, está claro que lo normal es tener un crédito que ronde la mediana de la población (ricos hay pocos).



Se observa que los casado se les concede créditos más elevado, la estabilidad familiar es muy importante a la hora de determinar la cantidad de crédito concedido. En cuanto a la probabilidad de *Default*, las categorías tienen una distribución parecida exceptuando los Divorciados, pero esta diferencia se debe a que hay pocos divorciados en la muestra.



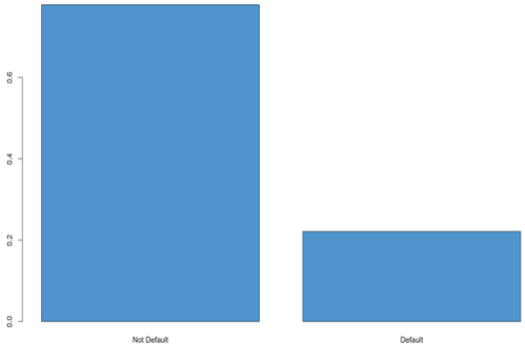
El nivel educativo nos indica que las personas que reciben mayor educación optan a créditos más altos, la distribución de *Default es muy* parecida entre clases y va acorde con el nivel de crédito concedido a los *No Default*.

6. Los datos son poco balanceados lo cual tendremos que tener en cuenta a la hora de trabajar con ellos. La solución pasa por utilizar alguna estrategia tipo ROSE, Over-Sampling o Under-sampling. Pero desde un punto de vista práctico y utilizando las métricas adecuadas podemos valorar los modelos de forma no balanceada.

En un caso de estudio sobre la misma base de datos, Vamsi Grandhi demuestra que los datos no balanceado funcionan mejor que las diferentes opciones para balancear los datos, como se demuestra en la siguiente tabla extraída de su caso:

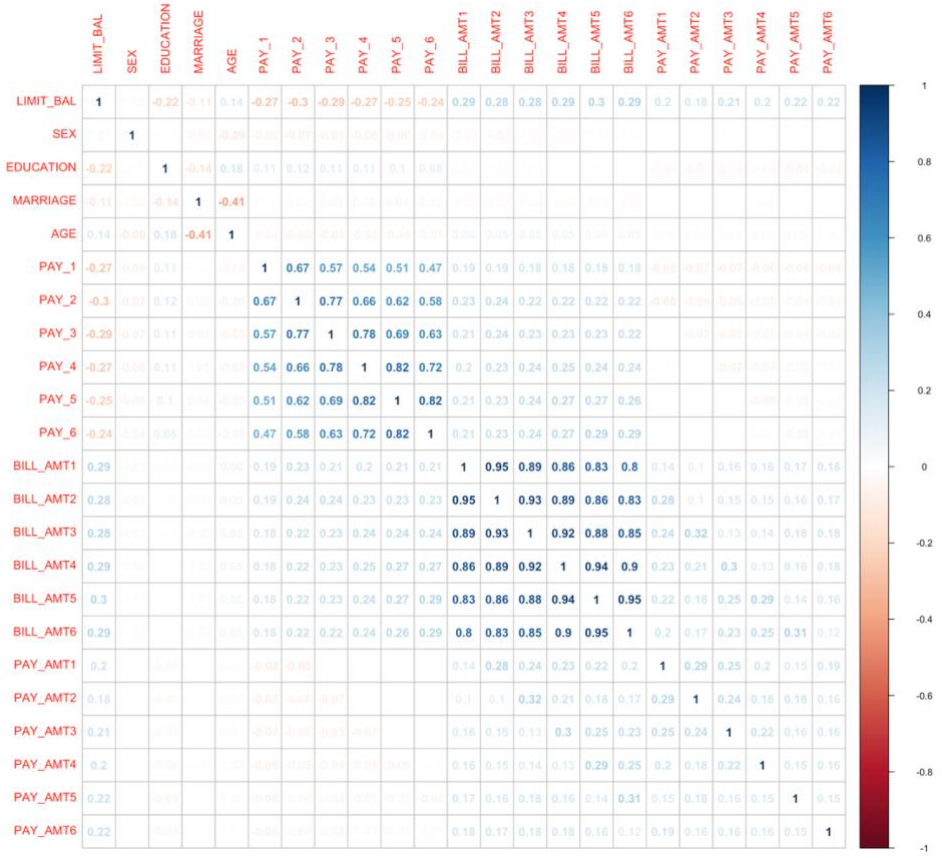
Model	naïve bayes		XG Boost	
Metrics	Accuracy %	Sensitivity %	Accuracy %	Sensitivity %
Unbalanced (train.task)	**77.55**	**48.65**	**82.43**	**68.19**
Oversample (train.over)	65.75	35.44	79.17	55.18
under sample (train.under)	60.41	32.18	76.99	47.4
Smote (train.smote)	42.36	25.84	80.74	58.31

Fuente: [Grandhi, V. 2017] - Aquí.



El 78% son *No Default* y el 22% son *Default*

7. Correlaciones de Pearson



Las correlaciones de Pearson no nos proporcionan una información muy útil, ya que las variable más correlacionadas son *"BILL_AMTX"* entre ellas y *"PAY_X"* entre ellas también, es de esperar que lo estén ya que suelen reflejar valores acumulados mes a mes.

Problemas en la Toma de Decisiones

1. Valores Extremos

Los valores extremos pueden distorsionar las predicciones y por lo tanto disminuir la precisión del modelo, en nuestro modelo hay valores a típico pero no se van a prescindir de ellos ya que pueden ayudar a mejorar la precisión del modelo, y por supuesto no interesará predecir también valores atípicos.

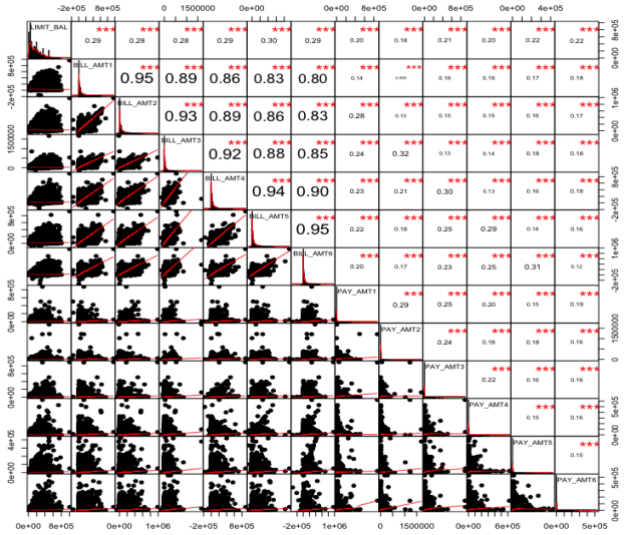
2. Over-fitting (Sobre-ajuste) o Under-fitting (Bajo-ajuste)

Un modelo presenta Over-fitting cuando la eficiencia en las predicciones en los datos de entrenamiento es buena, pero su precisión disminuye cuando aplicamos los datos de pruebas sobre nuestro modelo. Ocurre porque el modelo memoriza muy bien o se ajustó demasiado bien a los datos de entrenamiento, que cuando intenta generalizar el modelo entrenado con observaciones nuevas, este no predice eficientemente con los datos que no ha visto. Under-fitting es todo lo contrario, donde los datos no se ajustan lo suficientemente bien a los datos de entrenamiento ni a los datos de test.

En nuestro caso y los modelos que usaremos tenemos más riesgo de ocasionar sobre ajuste por ello, las medidas que tomaremos contra esto será dividir los datos en Test y Entrenamiento, y sobre los de entrenamiento se aplicaremos validación cruzada k-veces para evitar que haga sobreajuste con los datos de entrenamiento ya que será evaluado sobre el fold k-1.

3. Multicolinealidad

La multicolinealidad es cuando las variables explicativas Xi del modelo presentan correlación lineal entre ellas. Es decir, cuando las variables explicativas presentan multicolinealidad significa que las Xi no son linealmente independientes. Cuando hay dependencia lineal total o parcial entre algunas de las variables explicativas estamos incumpliendo uno de los supuestos básicos en la teoría del modelo y por lo tanto una relación casi lineal entre algunas de la variables explicativas, puede ocasionar problemas con la inferencia del modelo. No queda tan claro si afecta o interfiere con el objetivo de predicción por eso se realizará el estudio con ambos casos, si resulta que hay multicolinealidad entre variables, de ese modo observaremos como afecta a la precisión o al error de predicción.



La variables *BILL_AMTX* guardan un gran relación lineal entre ellas, por lo tanto es necesario que las variables predictores sean independientes entre sí. Para valorar este aspecto se utiliza el índice de VIF (Variance Inflation Factor).

$VIF = \frac{1}{1-R^2}$, donde R^2 es el coeficiente de determinación de la regresión de nuestro modelo frente a todos sus predictores (sin descartar ninguno).

Si el índice del VIF($\hat{\beta}_j$) > 10, entonces las variables presentan multicolinealidad elevada.

Variable Explicativas	**VIF($\widehat{\boldsymbol{\beta}}_j$)**	**VIF($\widehat{\boldsymbol{\beta}}_j$)***
LIMIT_BAL	1.52	1.51
SEX	1.02	1.02
EDUCATION	1.14	1.14
MARRIAGE	1.23	1.23
AGE	1.27	1.27
PAY_1	1.50	1.50
PAY_2	2.67	2.67
PAY_3	3.30	3.28
PAY_4	3.89	3.87
PAY_5	4.27	4.25
PAY_6	3.00	3.00
BILL_AMT1	24.21	3.72
BILL_AMT2*	40.73	Presenta Multicolinealidad
BILL_AMT3*	29.32	Presenta Multicolinealidad
BILL_AMT4*	27.04	Presenta Multicolinealidad
BILL_AMT5*	31.44	Presenta Multicolinealidad
BILL_AMT6	18.70	4.28
PAY_AMT1	1.47	1.16
PAY_AMT2	1.44	1.15
PAY_AMT3	1.46	1.14
PAY_AMT4	1.44	1.14
PAY_AMT5	1.52	1.16
PAY_AMT6	1.11	1.10

Efectivamente, las variables *BILL_AMTX* presentan multicolinealidad. Se procede a eliminar las variables que presenta de multicolinealidad, el procedimiento consiste en quitar la variable con mayor multicolinealidad y volver a calcular VIF para evaluar cómo va disminuyendo la multicolinealidad del resto, así sucesivamente hasta que ninguna variable presente multicolinealidad.

Métricas para Evaluar la Clasificación

A partir de la matriz de confusión describiremos las métricas de clasificación que más nos interesa:

		Real	
Predicción	Verdadero Positivo - VP	Falso Positivo - FP (Tipo I)	PTP
	Falso Negativo – FN (Tipo II)	Verdadero Negativo - VN	PTN
	RTP	RTN	TOTAL

Accurancy es el porcentaje de instancias correctamente clasificadas de todas las instancias. Es más útil en una clasificación binaria que en los problemas de clasificación de clases múltiples. Saber más sobre Accurancy.

Kappa o el *Kappa de Cohen* es como el *"Accurancy"* para problemas de clasificación, excepto que está normalizado por la línea de base de probabilidad para una clasificación aleatoria de los datos. Es muy útil en problemas no balanceados en las clases (por ejemplo, 70-30 para las clases 0 y 1 y puede alcanzar el 70% de precisión al predecir que todas las instancias son para la clase 0). Saber más sobre Kappa.

Precision es el número de verdaderos positivos (VP) sobre el número de verdaderos positivos más el número de falsos positivos (PTP). Una baja precisión indica un gran número de predicciones incorrectas como Falsos Positivos.

Recall es el número de verdaderos positivos (VP) sobre el número de verdaderos positivos más el número de falsos negativos (RTP). Un bajo Recall indica un gran número de predicciones incorrectas como Falsos Negativos. Para saber más sobre Precision y Recall.

Sensitivity son las observaciones de la clase positiva ("No Default") que se clasifican correctamente.

Specificity son las observaciones de la clase negativa ("Default") que se clasifican correctamente. Para saber más sobre Sensitivity y Specificity.

F Score: A menudo es conveniente **combinar la Precisión y el Recall en una sola métrica** para comparar de forma sencilla dos clasificadores. En vez de calcular una media de la Precisión y el Recall, se calcula su **media harmónica**. De esta forma se da más peso a los valores bajos por lo que sólo se conseguirá un F-score alto si ambas, Precisión y Recall, son altas. Para saber más sobre F Score.

ROC – AUC es en realidad el área bajo la curva ROC o AUC. El AUC representa en cierto modo la habilidad de los modelos para discriminar entre clases positivas y negativas (pude ser muy útil en nuestro modelo. Un área de 1 representa un ajuste perfecto, es decir, todas las observaciones están correctamente clasificadas. Un área de 0.5 representa un modelo tan bueno como aleatorio. Saber más sobre ROC.

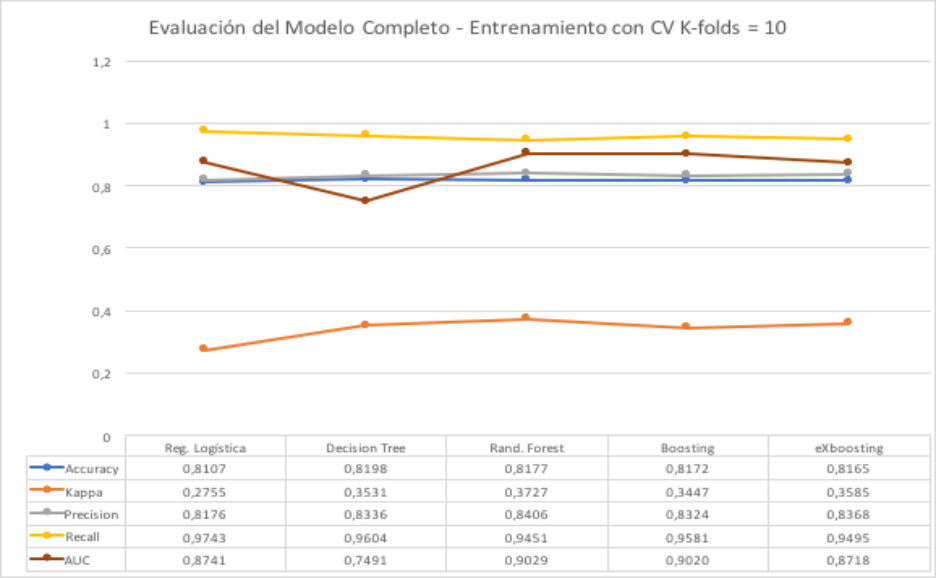
Resultados

Para el entrenar los modelos se va a utilizar la librería *Caret*, ya que nos permite aplicar validación cruzada de una forma fácil y sencilla sobre los datos, a la vez que nos permite usar "tuneGrid" para especificar parámetros variables en algunos métodos y evaluarlo cada vez y seleccionar los mejores parámetros para nuestro modelo.

Línea Base – Se sitúa en un ***accurancy* de 0.78** que es la probabilidad de clasificar siempre la clase mayoritaria, desde este punto podemos decir que nuestro modelo es capaz de mejorar la decisión de siempre predecir la clase mayoritaria.

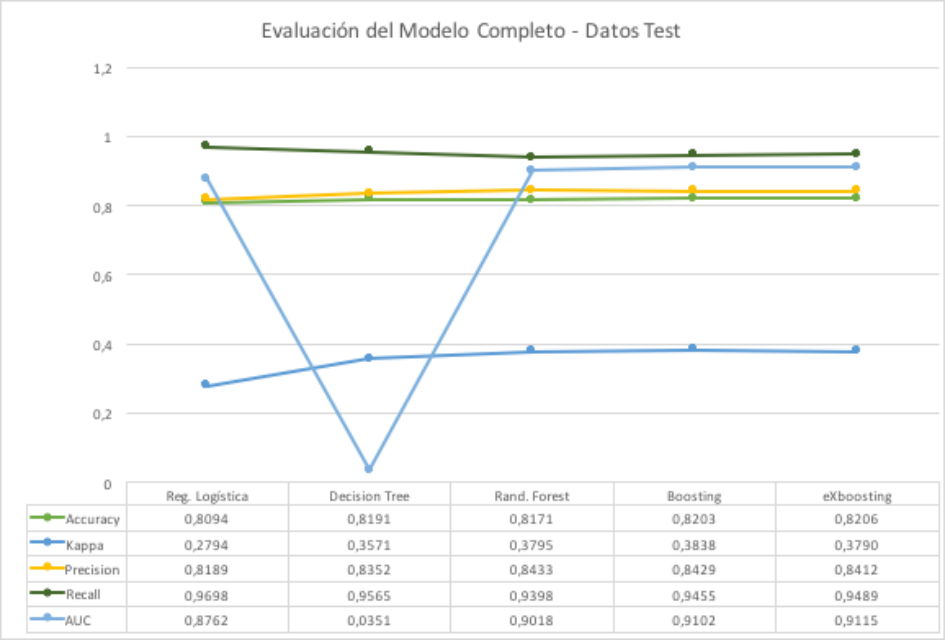
Modelo Completo – Datos de Entrenamiento con CV = 10

Clasificador	Accuracy	Kappa	Precision	Recall	F Score	Specificity	Sensitivity	AUC
Reg. Logística	0.8107	0.2755	0.8176	0.9743	0.8891	0.2346	0.9743	0.87407
Decision Tree	0.8198	0.3531	0.8336	0.9604	0.8925	0.3250	0.9604	0.74908
Rand. Forest	0.8177	0.3727	0.8406	0.9451	0.8898	0.3693	0.9451	0.90289
Boosting	0.8172	0.3447	0.8324	0.9581	0.8908	0.3210	0.9581	0.90201
eXboosting	0.8165	0.3585	0.8368	0.9495	0.8896	0.3481	0.9495	0.87184



Modelo Completo – Datos de Test

Clasificador	Accuracy	Kappa	Precision	Recall	F Score	Specificity	Sensitivity	AUC
Reg. Logística	0.8094	0.2794	0.8189	0.9698	0.8880	0.2447	0.9698	0.87624
Decision Tree	0.8191	0.3571	0.8352	0.9565	0.8917	0.3352	0.9565	0.03512
Rad. Forest	0.8171	0.3795	0.8433	0.9398	0.8889	0.3849	0.9398	0.90183
Boosting	0.8203	0.3838	0.8429	0.9455	0.8913	0.3794	0.9455	0.91023
eXboosting	0.8206	0.379	0.8412	0.9489	0.8918	0.3688	0.9489	0.91153

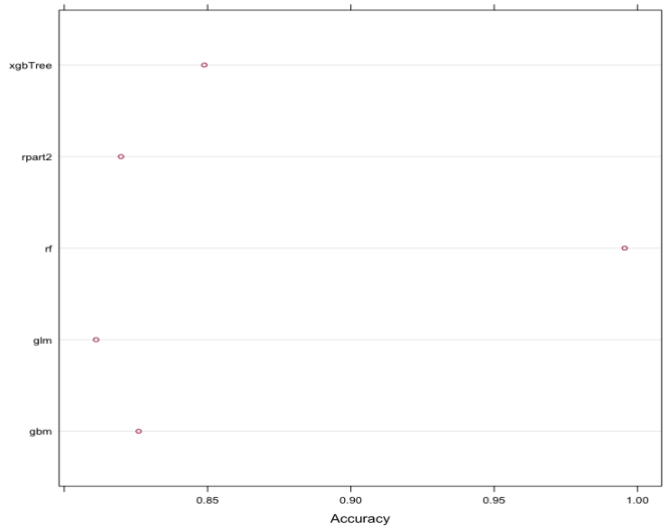


Conclusión

Todos los modelos superan el “Accurancy” de la línea base 0.78, pero sin destacar una gran mejoría. En los casos de entrenamiento “Decision Tree” aporta el mayor “Accurancy”, pero si tenemos que valorar el conjunto de modelos según el que mejor clasifica la clase minoritaria (“Default”), escogería “Random Forest” ya que es la que combina una mejor relación entre el total de métricas y el que mejor clasifica la clase minoritaria según los parámetros Kappa y Specificity.

Por otro lado, si nos fijamos en la predicciones sobre los datos de test observamos que los modelos más complejos predicen mejor, en concreto “Extreme Gradient Boosting”, es el que consigue un mayor “Accurancy”, hay que reconocer que la mejoría en la predicción respecto a otros modelos más simples no es muy destacables, teniendo en cuenta el coste computacional que conlleva este modelo podríamos plantearnos escoger alguno más simple y ahorrar el coste que conlleva el entrenamiento de modelos complejos. Optaría quizás por “Random Forest”, porque propociona un “Accurancy” bastante bueno comparado con “Extreme Gradient Boosting”, y el motivo con más peso es su nivel para clasificar la clase minoritaria “Default” sobre datos de test.

Evaluación de Modelos según *“Accurancy”* (Observaciones vs Predicciones)



Modelo Sin Multicolinealidad – Datos Entrenamiento con CV = 10

Clasificador	Accuracy	Kappa	Precision	Recall	F Score	Specificity	Sensitivity	AUC
Reg. Logística	0.8104	0.2739	0.8173	0.9744	0.8890	0.2331	0.9744	0.87347
Decision Tree	0.8198	0.3531	0.8336	0.9604	0.8925	0.3250	0.9604	0.75107
Rad. Forest	0.8172	0.3767	0.8422	0.9417	0.8892	0.3789	0.9417	0.90491
Boosting	0.8221	0.3860	0.8425	0.9489	0.8926	0.3756	0.9489	0.91171
eXboosting	0.8189	0.3717	0.8397	0.9485	0.8908	0.3625	0.9485	0.91002

Modelo Sin Multicolinealidad – Datos Test

Clasificador	Accuracy	Kappa	Precision	Recall	F Score	Specificity	Sensitivity	AUC
Reg. Logística	0.8099	0.2797	0.8188	0.9706	0.8883	0.2437	0.9706	0.87648
Decision Tree	0.8191	0.3571	0.8352	0.9565	0.8917	0.3352	0.9565	0.03512
Rad. Forest	0.8194	0.3885	0.8450	0.9406	0.8903	0.3925	0.9406	0.90339
Boosting	0.8199	0.3812	0.8423	0.9458	0.8911	0.3764	0.9458	0.91052
eXboosting	0.8194	0.3768	0.8411	0.9471	0.8909	0.3698	0.9471	0.91214

Conclusión

En el caso del modelo sin multicolinealidad funciona un poco mejor “Boosting” en el entrenamiento del modelo proporcionando unos valores levente superiores al anterior modelo, donde “Random Forest” proporcionaba una mejor estimación por el hecho de que clasificaba mejor la clase minoritaria. En este caso sin duda “Boosting” proporciona una buenos resultado en todas las métricas comparadas con los modelos alternativos, también es de los que mejor clasifican la clase minoritaria observando las métricas Kappa y Specificity.

Por otro lado, cuando probamos los modelos con nuestros datos de test obtenemos que “Boosting” es el que mayor “Accurancy” proporciona como en los datos de entrenamiento, pero en este caso “Random Forest” clasifica mejor con los datos de test tanto la clase minoritaria (“Default”) como la clase mayoritaria, consiguiendo un “Accurancy” muy cercano al que proporciona “Boosting”.

Para finalizar el presente estudio, me gustaría dar una opinión un poco más subjetiva y desde el punto de vista financiero y económico. Como el objetivo es prevenir y detectar la probabilidad de impago, pero a la vez no queremos denegar la posibilidad de conceder crédito a un cliente que si va a pagar debido a que la mayoría pagan, el estudio podría hacerse desde un punto de vista distinto y ofrecer una perspectiva no tanto de predicción sino de estudiar los casos desde la probabilidad que tiene un cliente de impago. Podríamos aplicar políticas, tarifas y estrategias comerciales para que los cliente muestren sus capacidades financieras y de ese modo se autodiscriminen mostrando sus verdaderas probabilidades de ocasionar un impago en el futuro o si de verdad son clientes solventes y con capacidad financiera. Con estas políticas podríamos hacer un análisis y ver cómo afecta a las probabilidades, y de este modo poder generar grupos con características que ocasionan impago y después sería ir mejorando las políticas para poder clasificar correctamente a dos clientes que con características idénticas uno comete impago y el otro no.

Bibliografía

Yeh, I. C., & 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, 36(2), 2473-2480.

Kuhn, M., 2017. The 'caret' Package. *GitHub Pages*. Available at: http://topepo.github.io/caret/model-training-and-tuning.html [Accessed December 5, 2017].

Brownlee, J. (2017). *How to Build an Ensemble Of Machine Learning Algorithms in R (ready to use boosting, bagging and stacking) - Machine Learning Mastery*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/machine-learning-ensembles-with-r/ [Accessed 7 Dec. 2017].

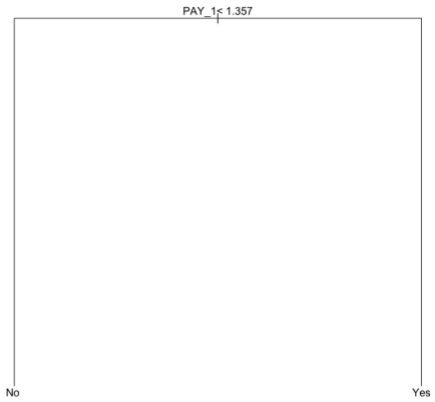
Grandhi, V. (2017). *Case of Study - Credit Default*. [online] GitHub. Available at: https://github.com/gvamsi01/Credit-card-default-data-UCI-ML-repoistory/blob/master/default%20credit%20card%20dataset.pdf [Accessed 21 Dec. 2017].

Brownlee, J. (2017). *Machine Learning Evaluation Metrics in R - Machine Learning Mastery*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/machine-learning-evaluation-metrics-in-r/ [Accessed 21 Dec. 2017].

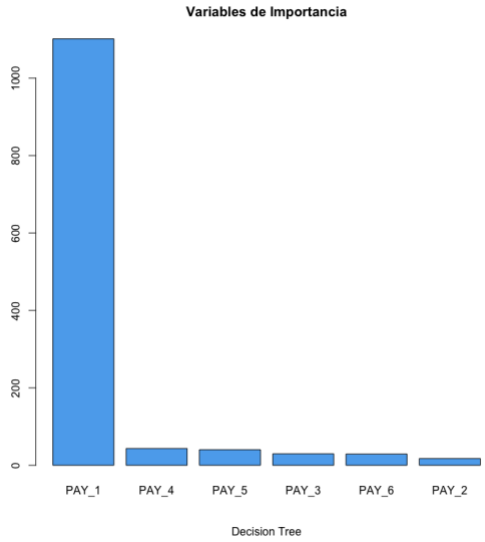
Data set:
UCI Machine Learning Repository: default of credit card clients Data Set. [Accessed December 05, 2017]. http://archive.ics.uci.edu/ml/datasets/default of credit card clients.

ANEXO DE PROCEDIMIENTOS DE MODELOS

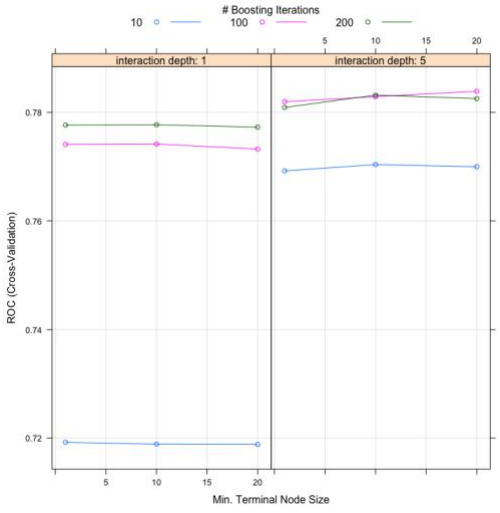
Árbol de decisiones – Modelo Entrenamiento – BestTune: Profundidad 1 – Parámetro decisivo para la decision PAY_1 < 1.357



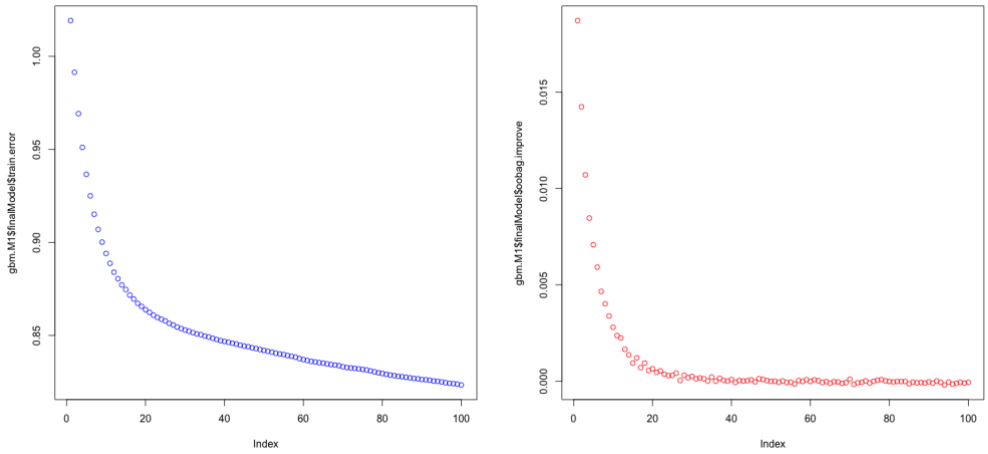
Variables por orden de importancia *"PAY_1"* en Decision Tree



Parámetros de utilizados en Boosting

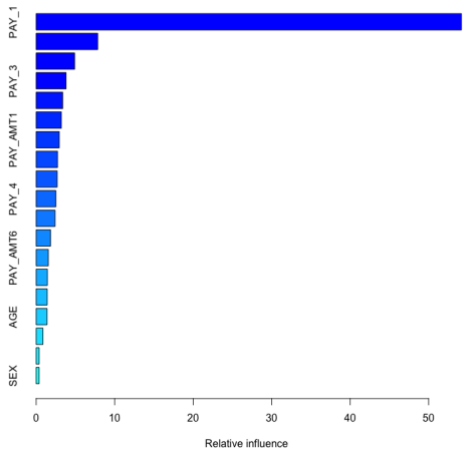


Train Error Evolution and OOBAG Improvement – Modelo Boosting



Observamos como "Bosting" con cada iteración va aproximando los errores a su mínimo, hasta que el índice oobag que indica la mejora que se obtiene por cada iteración se aproxima a 0, indicando que el modelos ya no puede mejorar más, aunque lo sigamos iterando.

Variables por orden de importancia: *"PAY_1"* – Modelo Boosting



Boosting también nos da proporciona como la variable más importante para clasificar es PAY_1 que coincide con el resto de métodos que también utiliza árboles de decisión en su proceso.