UNIVERSIDAD TORCUATO DI TELLA

Departamento de Economía

Maestría en Economía Aplicada

Modelo de scoring crediticio bancario para calificar Pequeñas y Medianas empresas: Aplicación empírica en Argentina

Alumno: Konstantin Kuznetsov

Legajo: 17M1963

Tutor: Hernán Ruffo

Fecha: 31 de mayo del 2019

Resumen

Este trabajo de aplicación propone un modelo de calificación crediticia para clientes Pymes (Pequeñas y Medianas Empresas) de una entidad financiera como en este caso un banco. Para llevar a cabo la estimación se propone utilizar un modelo de *regresión logística* que tiene reconocido uso en la industria financiera para la calificación de la Banca Individuos (crédito de consumo). Este modelo tiene como objetivo estimar una probabilidad de Default (cesación de pagos) tomando como *inputs* características observables del cliente y parámetros del crédito para separar a los créditos *Buenos* de los *Malos* y entonces decidir una política de crédito de la institución financiera.

Dicha estimación, se hace con datos históricos en dónde se encuentran las variables que usaremos como *inputs* junto al resultado binario de si el crédito ha sido *Bueno* o *Malo* para cada caso a través de una *regresión logarítmica*. Luego de la modelización, el ajuste de esta se evaluará a través de la curva ROC, en donde se analizará su *Sensibilidad, Precisión y Especificidad* en base a su capacidad predictiva.

El desafío planteado en el presente trabajo es testear con datos reales si el mismo modelo es aplicable para los clientes de créditos comerciales, ya que en esencia se diferencian de los créditos de consumo al menos en el sistema bancario argentino. En primer lugar, los créditos comerciales son dirigidos en su mayoría a Personas Jurídicas mientras que los créditos de consumo hacia Personas Físicas. Además, hay diferencia de

magnitud en los montos solicitados, lo que hace exigir una garantía a contraparte para los préstamos comerciales, mientras que los de consumo en su mayoría son *a sóla firma.*

Habiendo hecho las estimaciones pertinentes, se arriban a resultados predictivos alentadores que indican la posibilidad de empezar este tipo de metodología en la industria bancaria complementando el análisis crediticio tradicional.

Palabras clave: *credit scoring, calificación crediticia, empresas, préstamo bancario.*

Contenido

1. Introducción

El presente trabajo propone crear un modelo de *credit scoring* alternativo más no substituto para el crédito bancario comercial. La metodología propuesta replica a la que se usa actualmente para el crédito de consumo (Banca de Individuos) con la salvedad de elegir *variables* observables que comprendan las diferencias entre Personas Físicas y Jurídicas.

Los modelos de *credit scoring* son de gran utilidad para el sector financiero ya que permiten una apropiada estimación y valoración del riesgo asociado a una operación de crédito. Es una metodología puesta en práctica masivamente por el sistema bancario no hace más de 15 años, aunque que existe hace 40 en la industria de las finanzas. La misma ha visto un desarrollo exponencial en la última década con el aluvión de la gran cantidad de datos mesurables a través de técnicas de *data mining* y mayor poder de cálculo vía poder computacional. Primeramente, ha sido fomentada fuertemente en más de 100 países gracias a la promulgación del 2do Acuerdo de Basilea acerca de la regulación y legislación bancaria en el año 2004[1]. En dicho acuerdo, se ponía en foco la necesidad de analizar la calidad crediticia del deudor para reducir el riesgo de incumplimiento del crédito evitando así el quebranto en cadena dado el nivel alto de deuda del sistema financiero (curiosamente pocos años después, en el 2008, hemos sufrido uno de los grandes colapsos financieros de la historia). Es por eso por lo que hoy en día, combinando avanzadas técnicas estadísticas

[1] https://www.bis.org/publ/bcbs128_es.pdf Convergencia internacional de medidas y normas de capital, junio de 2006

con bases de datos internas y externas, es una herramienta fundamental a nivel mundial para toda institución financiera a la hora de adoptar una política crediticia.

Entendemos el **riesgo crediticio** como una eventualidad que ante la incertidumbre del futuro, puede comprometer el patrimonio de una institución debido a la incapacidad del deudor de cumplir en tiempo y forma con el acuerdo contractual del crédito. Cabe aclarar, que el cálculo del riesgo de crédito no se compone solamente de la *probabilidad de default(PD)*, sino que es compuesto además por: la *pérdida por default* (LGD: Loss Given Default) y la *exposición al default* (EAD: Exposure at Default)[2]. En este trabajo, sólo nos ocuparemos de la construcción de un modelo de scoring crediticio para la estimación de la *PD.* Tarea que igualmente no es sencilla, sabiendo que existen muchas técnicas para su modelización: *regresión lineal, regresión logística, árboles de decisión, análisis discriminante, bosques aleatorios, redes neuronales, máquinas de vectores de soporte.*

Para esta aplicación, se decidió implementar un modelo de *regresión logística* que entre otras muchas funciones, permite determinar el monto, plazo y tasa de interés óptima para que un crédito sea repagado por un determinado cliente. Este tipo de regresión arroja un resultado comprendido en el conjunto **[0; 1],** lo cual simplifica la interpretación probabilística del *output*. Un valor estimado cercano al 0 nos indicaría una **baja** probabilidad

❖ 2 Anderson, R. (2007). *The credit scoring toolkit: theory and practice for retail credit risk management and decision automation*. Oxford University Press. Pág. 507

de Default (PD), mientras que una estimación cercana a 1 indica una probabilidad **alta**. (Nota: Por definición arbitraria, esto ocurre si en nuestra base de datos previamente trabajada hemos asignado con valor "1" a la ocurrencia de *default*).

Actualmente en Argentina, este tipo de metodologías son muy usadas en la industria bancaria/fintech para la estimación de la **cartera de Individuos**. Aun así, es anormal que entidades financieras apliquen ello en el **segmento de Empresas**. Según un estudio realizado por FELABAN en el año 2008, tan sólo el 30% de los bancos latinoamericanos realizaba análisis de *scoring crediticio* a Pymes, mientras que el 95% lo hacía a través del análisis de estados financieros y flujo de caja[3]. Este análisis para clientes con actividad comercial es singular, basándose en fundamentos propios de cada empresa de acuerdo con literatura de finanzas corporativas. Es así como la calificación se procede de acuerdo con parámetros contables preestablecidos que cada entidad financiera elija usar de acuerdo con su política crediticia de riesgo. Por ejemplo, entre los parámetros contables a analizar con mayor consenso en la industria bancaria podemos encontrar: ventas del último período, resultado bruto y neto del balance, ratio de liquidez, ratio de solvencia, proyección de inversión, patrimonio neto. Además de esto, es muy común el pedido de garantía como contraparte del empréstito para reducir el nivel de riesgo. Para la elaboración de estos análisis, el pedido de documentación respaldatoria y el pedido de garantía crean barreras significativas para el acceso al crédito que conllevan: por un lado, mayores demoras en el

[3] Sanguinetti, P., Arreaza, A., Rodríguez, P., Álvarez, F., Ortega, D., & Penfold, M. (2011). RED 2011: Servicios financieros para el desarrollo. Promoviendo el acceso en América Latina (Reporte de Economía y Desarrollo (RED)). Caracas: CAF. Retrieved from http://scioteca.caf.com/handle/123456789/170 Pág. 148

otorgamiento, y por otro, restricciones de crédito para aquellas empresas con falta de colateral o tienen corta vida comercial no contando con un año de facturación (requisito esencial en todos los bancos argentinos). En la misma línea, un estudio realizado por Pasquini y De Giovanni (2010) muestra que el 7% de las Pymes que solicitan crédito son rechazadas mientras que el 37% se autoexcluyen del mercado[4]. Ante esta restricción, las empresas que buscan financiación se ven obligadas a recurrir al mercado de capitales, obtener una garantía a través de Sociedades de Garantía Recíproca u otra vía de financiación informal, o aún peor, ninguna.

Entonces, la propuesta de este trabajo es intentar diseñar un modelo de *credit scoring* que complemente los métodos tradicionales de evaluación crediticia dadas las restricciones financieras que impiden el acceso al crédito de las Pequeñas y Medianas Empresas.

[4] Sanguinetti, P., Arreaza, A., Rodríguez, P., Álvarez, F., Ortega, D., & Penfold, M. (2011). RED 2011: Servicios financieros para el desarrollo. Promoviendo el acceso en América Latina (Reporte de Economía y Desarrollo (RED)). Caracas: CAF. Retrieved from http://scioteca.caf.com/handle/123456789/170 Pág. 142

2. Revisón de la literatura

Uno de los primeros trabajo en *credit scoring* ha sido el de James H. Myers & Edward W. Forgy (1963)[5] en dónde plantean la necesidad de realizar una calificación crediticia para los prestatarios por parte de los prestadores. Analizan la capacidad de repago de créditos para consumo a través del método de *Análisis Discriminante* con una muestra de 600 casos con 41 variables observables. Del total de esas variables, encuentran 21 estadísticamente significativas a un nivel de confianza del 5% para predecir la capacidad de repago de clientes Individuos. Así, sientan un precedente para trabajos venideros.

En orden cronológico, el trabajo de Yair E. Orgler (1970)[6] plantea un modelo de *credit scoring* para préstamos comerciales de clientes empresas. El mismo propone un modelo de regresión logarítmica tomando tan sólo 5 variables que corresponden a un anáisis de estado financiero: *Liquidez, Rentabilidad, Nivel de endeudamiento, Actividad del cliente, Colateral*. Toma una muestra de 300 casos, con lo que obtiene un bajo nivel predictivo llegando a la conclusión de que su modelo será aplicable para un cliente específico de una industria específica. Aún así, una de sus contribuciones importantes, es haber determinado que la responsabilidad del repago de un crédito recae en el prestatario

[5] James H. Myers & Edward W. Forgy (1963) The Development of Numerical Credit Evaluation Systems, Journal of the American Statistical Association, 58:303, 799-806

[6] Orgler, Y. E. (1970). A credit scoring model for commercial loans. *Journal of money, Credit and Banking*, *2*(4), 435-445.

y no en el acreedor por las condiciones del préstamo, cuestión que hasta ese tiempo estaba en discusión.

Con el correr del tiempo y el desarrollo de nuevos métodos estadísticos fueron apareciendo nuevas metodologías para el análisis de *credit scoring*.

El paper de Hand, D. J., & Henley, W. E. (1997)[7] sabe resumir óptimamente los métodos que se estaban empleando hasta ese entonces. Con una muestra de 5000 casos de crédito de consumo, compara los diferentes métodos de estimación existentes. Arriban a la conclusión de que no existe un método óptimo, sino que depende en las características del problema a analizar: cómo es la estructura de datos y qué variables son elegidas. A su vez, hacen una contribución esencial de criterio para la evaluación de los modelos que es la curva de ROC inspirándose en la curva de Lorenz. La misma nos mostrará qué tan buen poder predictivo tiene un determinado modelo mostrando a 2 ejes la propoción de *Buenos* créditos aceptados contra los *Malos* aceptados.

Habiendo recopilado la teoría, es hora de abordar la práctica. Para armar el modelo con datos empíricos reales, usaremos el libro de R Anderson (2007)[8] acerca de modelos de *credit scoring* con ayuda del libro de teoría estadística de G James, D Witten, T Hastie, R

[7] Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *160*(3), 523-541.
[8] Anderson, R. (2007). *The credit scoring toolkit: theory and practice for retail credit risk management and decision automation*. Oxford University Press.

Tibshirani (2013)[9]. Complementando, utilizaremos los siguientes papers para el armado del modelo: M Bensic y N Sarlija(2005)[10] dónde que afirman que los ratios financieros no son determinantes para la estimación del repago de una pequeña empresa y sí la actividad lo es. Y además, el trabajo de M. Ansen(2017)[11] que expone diferentes técnicas de análisis multivariado.

[9] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

[10] Bensic, M., Sarlija, N., & Zekic-Susac, M. (2005). Modelling small-business credit scoring by using logistic regression, neural networks and decision trees. *Intelligent Systems in Accounting, Finance & Management: International Journal*, *13*(3), 133-150.

[11] Mathew, Ansen, "CREDIT SCORING USING LOGISTIC REGRESSION" (2017). Master's Projects. 532

3. Métodología

Habiendo presentado el estado del arte, la metodología que emplearemos para intentar discernir entre un cliente *Bueno* o *Malo* será a través de un **Modelo de Regresión Logística.** Como hemos anticipado previamente, se trata de una técnica que a través de variables observables del cliente permitirá estimar su comportamiento a través de una variable categórica: caer en *default o no*[12].

3.1. Regresión Logística Múltiple

Este tipo de regresión nos permite establecer una relación entre una variable dependiente $\boldsymbol{Y}$ no cuantitativa, y en este caso binaria, con un conjunto de variables independientes $(\boldsymbol{X_1} + \cdots + \boldsymbol{X_p})$ que pueden ser tanto cuantitativas como cualitativas. El ajuste de la ecuación se hace a través de la estimación de los parámetros $(\boldsymbol{\beta_1} + \cdots + \boldsymbol{\beta_p})$ con la idea de predecir el comportamiento de la variable **Y**.

Ante la imposibilidad de obtener resultados binarios con un modelo de regresión lineal simple, buscamos obtener un *output* entre [0,1] a través de la función logarítmica.

[12] Anderson, R. (2007). *The credit scoring toolkit: theory and practice for retail credit risk management and decision automation*. Oxford University Press. 163-164.

Consideramos la siguiente ecuación para la predicción de una respuesta binaria usando múltiples predictores:

(1) $$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

Dónde *p* indica la cantidad de variables predictoras.

Transformando *(1)* podemos obtener:

(2) $$g(x) = log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Dónde $g(X)$ es una transformación del modelo $p(X)$ habiendo aplicado logaritmo para ello (De ahí el nombre dado).

Para estimar los coeficientes de $\beta_0, \beta_1, \ldots, \beta_p$ se utilizará el método de *Máxima Verosimilitud* que tiene buen ajuste con modelos no lineales:

(3) $$l(\beta_0, \beta_1) = \prod_{i:y_{i=1}} p(x_i) \prod_{i\prime:y_{i\prime=0}} (1 - p(x_{i\prime}))$$

En este caso, $\widehat{\boldsymbol{\beta}}_{\boldsymbol{0}}$ y $\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}$ son elegidos para maximizar la *función de Máxima Verosimilitud* que expresa la probabilidad de las observaciones como una función de parámetros desconocidos.

No es motivación de este trabajo hacer el desarrollo matemático de la *función de Máxima Verosimilitud* pero podemos afirmar que los estimadores obtenidos maximizan esta función, por lo que garantizarán el mejor ajuste del modelo cuando su distribución es desconocida[13].

Estimación

Como insumo de nuestro modelo, usaremos una base de datos real de una entidad financiera pública con **4.362** casos de préstamos a empresas que sucedieron entre los años 2017 y 2018. Hemos condicionado para que las empresas tengan una Facturación Anual Estimada entre $500.000-$500.000.000 y tengan menos de 500 empleados (ver ANEXO I).

Para elaborar esta muestra de créditos, hemos tomado un solo registro por cliente ya que es muy común que un cliente comercial tenga varios que usa en su operatoria comercial. Para hacer esta selección se utilizaron los siguientes criterios:

1) Si el cliente no tuvo ningún préstamo en mora pero tiene varios préstamos durante el período de la muestra, se selecciona el préstamo de mayor monto.

[13] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. 133-134.

2) Si el cliente tiene un préstamo en mora y otros en situación *Normal,* se toma sólo el primer caso.

3) Si el cliente posee varios préstamos sin pagar, se toma aquél de mayor antigüedad ya que se entiende que fue el primero en dar muestra de la cesación de pagos.

Cabe mencionar, que todos las empresas de la muestra poseen una antigüedad en actividad superior a 1 año; ya que es requisito mínimo para una calificación crediticia tener un balance contable. Es por esto, que este tipo de modelo, no nos servirá *a priori* para calificar crediticiamente a emprendedores que requieran un crédito para empezar su actividad.

Variables propuestas

Definiremos 2 grupos de variables explicativas ***X***: en el primero tendremos los parámetros del crédito otorgado y el segundo estará compuesto por aquéllas inherentes al cliente por sus características observables. Por supuesto, no nos olvidemos de nuestra variable ***Y***, la más importante, que nos indica si el crédito incurrió en falta de pagos o no.

Moneda: "En Pesos" o "En Usd"	Moneda en la que fue dado el préstamo: **1** para pesos, **0** para dólares.
Destino: "Capital Trabajo" o "Inversion"	Destino del préstamo: **1** para Capital de Trabajo, **0** para Inversión.
Importe Otorgado	Monto total del crédito.
TNA	Tasa Nominal Anual del crédito. No tiene en cuenta comisiones.
Plazo meses	Plazo del crédito expresado en cantidad de meses.
Default	**1** si incurrió en Default del préstamo en el Banco, **0** si no ha incurrido.

Anos Cliente	Antigüedad del cliente del Banco expresada en años.
Zona Geográfica	**“CABA”, ”CONURBANO” o “INTERIOR”:** respectivo a la localidad de actividad comercial del cliente.
Cantidad de Empleados	Cantidad de empleados registrados.
Cuentassueldo	Cantidad de cuentas de haberes que tiene la empresa abiertas para sus empleados en el banco.
Facturacion Anual Estimada	Al no tener acceso al balance de cada empresa, se utilizará una variable proxy construida de acuerdo con los movimientos al crédito en la cuenta corriente en 1 año.
Tamanio Empresa	**“Microempresa”, “Pequena Empresa” o “Mediana Empresa”:** de acuerdo con el criterio de facturación establecido por el Ministerio de Producción y Trabajo.
Actividad BCRA 1	**“Act Bcra1 Agroindustria”, “Act Bcra1 Comercio”, “Act Bcra1 Industria” o “Act Bcra1 Servicios”:** de acuerdo con la clasificación de sectores económicos de actividad nivel 1 del BCRA.
Actividad BCRA 3	**“Act Bcra3 Agropecuario”, “Act Bcra3 Alimentos”, “Act Bcra3 Automotor”, “Act Bcra3 Combustibles”, “Act Bcra3 Construccion”, “Act Bcra3 Espectaculos”, “Act Bcra3 Manufacturera”, “Act Bcra3 Metales”, “Act Bcra3 Mineria”, “Act Bcra3 Petroquimica”, “Act Bcra3 Plastico”, “Act Bcra3 Servicios”, “Act Bcra3 Textilycuero” o “Act Bcra3 Turismo”:** de acuerdo con la clasificación de sectores económicos de actividad nivel 3 del BCRA.
Posee Bancainternet	**1** si posee un Homebanking empresarial del Banco, **0** si no lo tiene.
Posee Pfdolar	**1** si posee un Plazo Fijo en dólares en el Banco, **0** de no ser así.
Posee Pfpesos	**1** si posee un Plazo Fijo en pesos en el Banco, **0** de no ser así.
Posee Prestamoanterior	**1** si tiene un préstamo vigente con el Banco, **0** de no ser así.
posee seguroART	**1** si posee el servicio de ART del Banco, **0** de no ser así.
Posee Tarjetacorporativa	**1** si posee Tarjeta Corporativa del Banco, **0** de no ser así.

Para tener en cuenta, todas las variables aquí descriptas están presentadas al momento del otorgamiento del préstamo. Por eso, cuenta con consistencia intertemporal a pesar de tener un contexto macroeconómico fluctuante condicionado por la alta inflación. Demos un ejemplo: De acuerdo con el Ministerio de Producción y Trabajo, en el año 2017 una empresa Agropecuaria que facturaba anualmente hasta $3 Millones[14] era considerada como Microempresa. Para agosto del 2018, la misma etiqueta le correspondería si su facturación anual no superaba los $4,8 Millones[15]. Esto habilita la comparación (aunque no perfecta) de resultados entre una Microempresa que solicita un préstamo en 2017 contra una en 2018.

[14] http://servicios.infoleg.gob.ar/infolegInternet/anexos/65000-69999/66187/texact.htm
[15] https://www.argentina.gob.ar/noticias/actualizacion-de-categorias-para-ser-pyme

3.2. Validación y *cross validation*

En la búsqueda del modelo adecuado, primero debemos asegurarnos de que este sea estadísticamente significativo. Para ello, utilizaremos el test conocido como *Test de Razón de Verosimilitud:*

Se define $\lambda = \frac{L_R}{L_M}$ dónde L_M es la verosimilitud del modelo completo estimado y L_R es un modelo reducido. Nuestra tarea es averiguar si hay diferencia entre probabilidades de obtener los valores observados con el modelo logístico completo y las probabilidades de hacerlo con un modelo sin relación entre las variables. Calculamos la significatividad de diferencia de residuos entre el modelo con predictores y el modelo reducido.

$$H_0)\ \beta_1 = \cdots = \beta_p = 0$$

$$H_1)\ algún\ \beta_k \neq 0, k = 1, \dots, p$$

La hipótesis nula plantea que el conjunto de los parámetros estimados no tiene significatividad estadística. La hipótesis alternativa indica que hay al menos uno que tiene significatividad estadística.

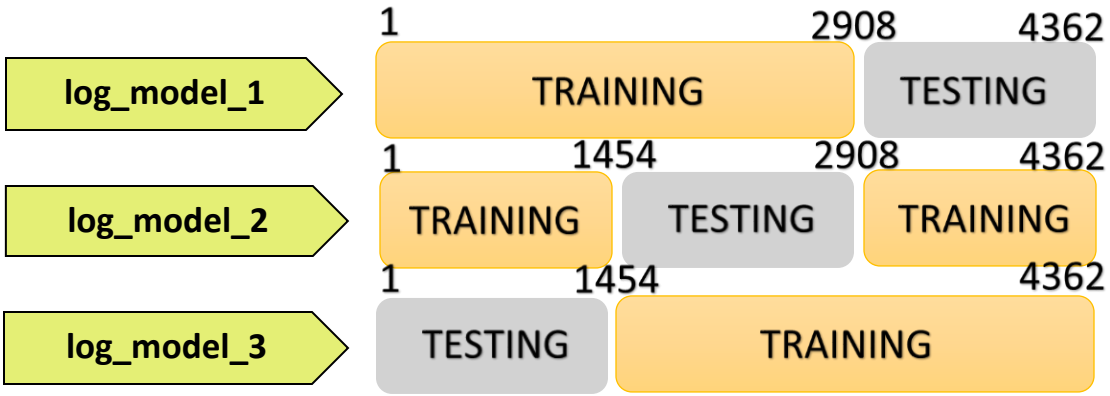
Planteadas las hipótesis, utilizaremos el estadístico $-2\ln(\lambda)$ que tiene una distribución *chi-cuadrado* $X^2_{(p+1-q),\alpha}$, siendo ($p+1$) y q la cantidad de parámetros incluidos en el modelo completo y el modelo reducido respectivamente.

$-2\ln(\lambda) = -2(lnL_R - lnL_M)$, haciendo que la hipótesis nula sea rechazada para el nivel de significación α cuando $-2\ln(\lambda) > X^2_{(p+1-q)}$.

Si rechazamos la hipótesis nula, entonces obtendremos significatividad estadística del modelo.

Además de utilizar este test, se propone que el modelo sea válido en distintos tramos de la base de datos para evitar sesgos que surjan a la hora de tomar un muestreo aleatorio. Este tipo de técnica se denomina como *cross validation* o "validación cruzada".

La idea es separar la base de datos en 2 grupos diferentes: un grupo de *training* y otro grupo de *testing*. En esta aplicación, se tomará el 66,6% de los registros para la base que arme el modelo, y el 34,4% restante para testear las predicciones que este produzca. A su vez, estos 2 grupos serán compuestos de manera diferente a lo largo de toda la muestra. A continuación se presentará un cuadro que resuma la distribución de los grupos a través de la etiqueta ***Id Cliente***:

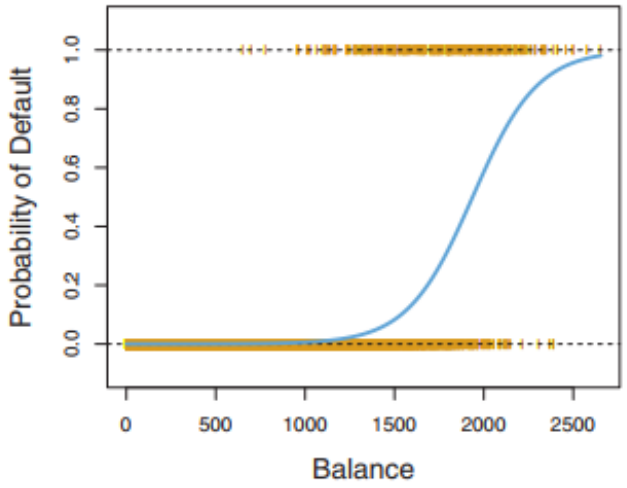


Los números indican el tramo de la muestra tomada de la población total según el código de identificación de cada cliente. Este muestreo se realizó con total aleatoriedad y no se

ordenó bajo ningún criterio. Obtendremos entonces, que en cada grupo de *Training* tenemos 2908 observaciones que serán de utilidad para calibrar el modelo y luego poder probar su poder de predicción en las siguientes 1454 observaciones de los grupos de *Testing*.

Este tipo de validación cruzada nos permitirá aumentar o disminuir el nivel de aprobación del modelo para cuando este se enfrente a nuevos datos por fuera de la muestra.

Recordemos una vez más, que el tipo de regresión que vamos a implementar es no lineal, y que por lo tanto los criterios para evaluar qué tan bien ajusta el modelo son diferentes:



FUENTE: James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Página 131. Figure 4.2.

Este gráfico sirve como ejemplo para mostrar cómo se distribuyen los resultados de las estimaciones a través de una regresión logarítmica simple entre el estado del **Balance de una cuenta corriente** y la **Probabilidad de Default.**

3.3. Punto de corte y Bondad de Ajuste

Habiendo estimado los parámetros del modelo con el tramo *Training*, podremos hacer predicciones con las variables explicativas del grupo *Testing* obteniendo valores que se encuentren en el conjunto [0,1], consiguiendo ser interpretados como una probabilidad. El problema siguiente, es elegir un punto de corte (conocido como ***umbral t***) entre esos valores para discriminar al crédito *Bueno* del *Malo.* A primera intuición, diríamos que el valor debería ser la media entre 0 y 1, o sea 0.5. Los valores debajo de ese punto deberían ser considerados como *Buenos* y análogamente por arriba del punto serían *Malos.* Pues bien, al elegir un valor arbitrario como ese, podríamos caer en la trampa de predecir los créditos *Buenos* como *Malos* y viceversa. Es decir, obtendríamos muchos *Falsos Positivos* y *Falsos Negativos*.

En el caso de ser un banco, tener muchos *Falsos Positivos* no es de alta gravedad ya que a lo sumo se perderá cierta rentabilidad al perder oportunidades de negocio. Pero tener una alta cantidad de estimaciones con *Falsos Negativos* es crítico, ya que estaríamos prediciendo que un cliente es *Bueno* cuando es más probable que caiga en *default.* La comparación de predicciones con los datos reales se muestra a través de una *matriz de confusión*:

		Predicted class		
		− or Null	+ or Non-null	Total
True class	− or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N*	P*	

FUENTE: James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Página 148. Table 4.6. Possible results when applying a classifier or diagnostic test to a population.

Curva de ROC

Un método muy utilizado para evaluar la calidad de ajuste de un modelo determinando así el corte óptimo es a través de una *curva ROC (Receiver Operating Characteristic)* que muestra de forma gráfica el desempeño del modelo a través de la variación de su comportamiento.

Tomando como base la matriz de confusión podemos construir 3 indicadores de ajuste:

$$\boldsymbol{Precisión} = \frac{TruePositive + TrueNegative}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

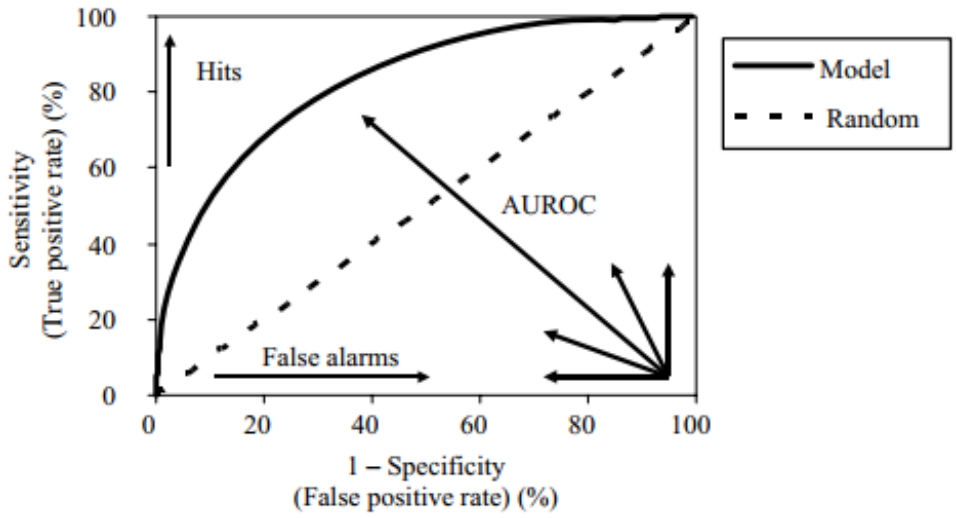
$$\boldsymbol{Sensibilidad} = \frac{TruePositive}{TruePositive + FalseNegative}$$

$$\boldsymbol{Especificidad} = \frac{TrueNegative}{TrueNegative + FalsePositive}$$

El gráfico de la *curva ROC* muestra la tasa de verdaderos positivos en el eje vertical **(Sensibilidad)** y la tasa de falsos positivos en el eje horizontal **(1 – Especificidad)** cuando el umbral de clasificación (valores predichos) va variando en el rango [0,1]. Es una curva que resume la información en una función de distribución acumulada de los puntajes de ambos grupos. Se puede pensar como una completa representación del funcionamiento de la función de clasificación[16].

[16] Krzanowski, W. J., & Hand, D. J. (2009). *ROC curves for continuous data*. Chapman and Hall/CRC.

A continuación podremos observar un ejemplo de curva ROC:



FUENTE: Anderson, R. (2007). *The credit scoring toolkit: theory and practice for retail credit risk management and decision automation*. Oxford University Press. Página 207, Figure 8.6. **ROC curve.**

Como se observa en el gráfico, sobre la recta punteada (0,0) a (1,1) el modelo no puede discriminar entre créditos *Buenos* y *Malos*. Allí, las tasas *Verdadera Positiva* y *Falsa Positiva* son iguales. Sin embargo, a medida que nos vamos acercando a la esquina izquierda de arriba, vamos aumentando la tasa *Verdadera Positiva* y reduciendo la *Falsa Positiva* por lo que el ajuste va mejorando considerablemente. Es decir, que el desempeño del modelo será mejor en cuanto a clasificación.

El problema que se presenta a continuación es cómo elegir entre 2 o más curvas de modelos diferentes cuando visualmente el contraste no es tan obvio.

Para hacer este tipo de comparaciones se utiliza un indicador llamado *AUC (Area Under the Curve)* que no es ni más ni menos el cálculo del área debajo de la curva:

$$\boldsymbol{AUC} = \int_0^1 y(x)dx$$

La *AUC* es la Tasa Positiva Promedio, tomada de manera uniforme sobre todas las posibles Tasas de Falsos Positivos en el rango [0,1]. Entonces, proporciona una medida del desempeño del modelo para discriminar entre observaciones que se hace comparable al ser resumida en un valor numérico continuo.

La regla general básica acerca de la interpretación del valor de la *AUC* establece[17]*:*

- Si **AUC = 0.5** entonces no existe discriminación por lo que el modelo es malo.
- Si **0.7 < AUC < 0.8** el ajuste y poder predictivo del modelo son aceptables.
- Si **0.8 < AUC < 0.9** el ajuste y poder predictivo del modelo son muy buenos.
- Que **AUC > 0.9** es poco probable que suceda.

Este criterio nos permitirá comparar entre diferentes curvas de modelos para poder elegir al mejor. Pero aún, no hemos determinado cuál es el umbral óptimo que logre maximizar la *Tasa Verdadera Positiva* y minimice la *Falsa Positiva.*

[17] Blanco, J. (2006). Introducción al análisis multivariado. *IESTA. Montevideo*.23.

3.4. Elección del umbral óptimo

En este punto, necesitamos establecer el valor del ***umbral óptimo t*** para usarlo como criterio para evaluar futuras observaciones. Es decir, calificar nuevos clientes que quieran solicitar un crédito a partir de la probabilidad estimada con nuestro modelo.

Uno de los procedimientos utilizados es minimizar la distancia entre el punto (0,1) que es la esquina superior izquierda y un punto de la curva de ROC obteniendo así el valor óptimo de ***t***. No obstante, esto no tiene un consenso general, por lo que algunos autores aseguran que este procedimiento puede llegar a introducir una mayor tasa de error de clasificación si el punto se ubica muy a la derecha[18].

Otro de los procedimientos más utilizados dentro de la industria bancaria es el estadístico *Kolmogorov-Smirnov (K-S)*[19]. Para calcularlo, se debe obtener para cada ***t*** la diferencia entre la *Tasa de Verdaderos Positivos* y *Falsos Positivos*. Así, el umbral óptimo de corte será aquél cuya diferencia sea la máxima:

$$K - S = max(\frac{TruePositive}{TruePositive+FalseNegative} - (1 - \frac{TrueNegative}{TrueNegative+FalsePositive})$$

$$K - S = max(\text{TasaVerdaderosPositivos} - \text{TasaFalsosPositivos})$$

[18] Krzanowski, W. J., & Hand, D. J. (2009). *ROC curves for continuous data*. Chapman and Hall/CRC.
[19] Hand, D. J. (2005). Good practice in retail credit scorecard assessment. *Journal of the Operational Research Society*, *56*(9), 1109-1117.

Siguiendo las indicaciones de (Krzanowski y Hand,2009), usaremos como base el índice *K-S* pero no será un criterio absoluto ya que es mejor complementarlo con juicio propio viendo caso por caso.

Una vez elegido el umbral de corte para cada mejor modelo estimado de la validación cruzada, se procederá a construir una nueva *matriz de confusión* para mostrar los resultados obtenidos. A partir de allí, podremos recalcular los indicadores de *Precisón, Sensibilidad* y *Especificidad* para evaluar el desempeño del modelo general.

La *Precisión* nos dirá el ajuste general del modelo entre observaciones reales versus los resultados estimados.

La *Sensibilidad* nos indicará la capacidad de nuestro modelo de identificar como morosos a los realmente tales.

La *Especificidad* nos indicará la capacidad de nuestro modelo de identificar a los buenos pagadores como realmente tales.

4. Resultados

Resumen general del procedimiento realizado

En primer lugar, hemos depurado la base de datos empírica de ciertos outliers de las variables de **Facturación Anual Estimada**, **Cantidad de empleados** y principalmente de *missing values.* Dejando así, sólo aquellas observaciones que consideramos pertinentes a las características de la empresa y las condiciones del crédito para el armado del modelo.

De esta manera, nos hemos quedado con un total de **4.362 observaciones** entre 2017 y 2018 en dónde la variable a estimar era si la empresa entró en *default* de su crédito con el Banco o no a través de una regresión logística múltiple. El resto de las variables se ha elegido de acuerdo con la literatura vigente.

El principal software que se ha utilizado para la manipulación de los datos y realización de las estimaciones ha sido el **RStudio**.

Se ha empleado una técnica de validación cruzada en dónde la muestra se ha partido en un 66.6% de empresas para *Training* y un 33.3% para *Testing.* A su vez, la misma muestra se ha tomado de 3 formas diferentes cambiando de lugar la composición de los grupos. Por *Id de Cliente* vamos a tener en el primer grupo *Training 1* aquellos clientes desde el nro 1 hasta 2.908. En el segundo grupo, *Training 2*, tendremos empresas con *Id* desde el nro 1 hasta 1.454 y luego desde 2.909 hasta 4.362. Y por último el *Training 3* estará conformado por *Id*s

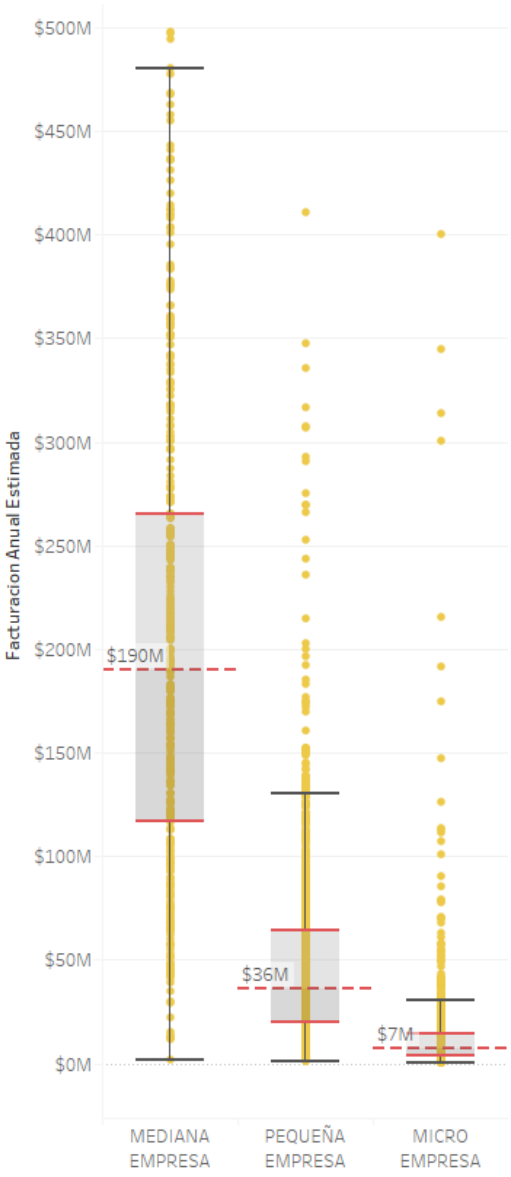
ente 1455 hasta 4.362. Los grupos *Testing* van a ser el resto de las observaciones de cada grupo.

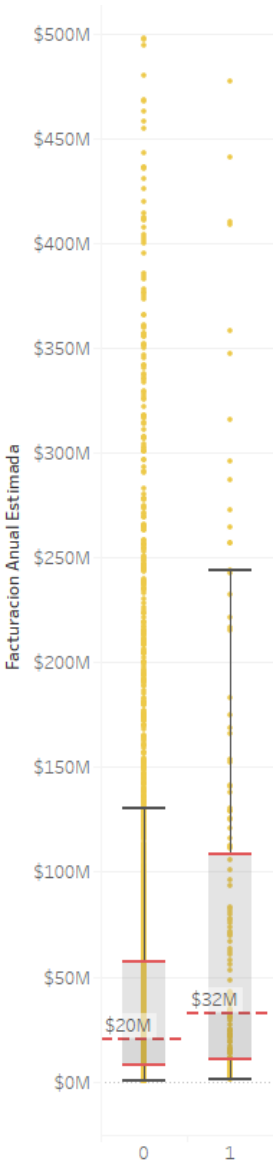
Se han propuesto 3 modelos diferentes para cada tramo de la validación cruzada: El primero contempla todas las variables incluidas la clasificación sectorial del BCRA en nivel 1 y 3. El segundo contempla las mismas pero sólo del nivel 1. Y el tercer modelo está compuesto por las mismas variables pero descarta totalmente el sector económico de la empresa. Resumiendo, se han hecho en total 9 estimaciones, 3 para cada tramo de la muestra. Estas mismas se han realizado a través de la *función de Máxima Verosimilitud*.

Haciendo predicciones con datos de cada grupo de *Testing,* se eligió el mejor modelo para cada tramo de acuerdo con los criterios de ajuste que vimos en la sección de *Metodología (*ver *curva ROC y AUC)*. Para discriminar los préstamos entre *Buenos* y *Malos*, se buscó el **umbral óptimo** para el corte de los *outputs* obtenidos a través del estadístico *K-S* y la posterior calibración manual. Habiendo elegido el **umbral óptimo**, se evaluó el modelo de acuerdo con su poder predictivo a través de los indicadores, calculados de la matriz de confusión, de *Precisión, Sensibilidad* y *Especificidad*.

Calibración del modelo para cada Tramo de la muestra

Modelo para el Tramo 1 (log_model_1)





Podemos observar que la mediana de **Facturación Anual Estimada** es acorde al tamaño de la empresa. Por otro lado, vemos que en los clientes morosos, la mediana de facturación anual es más elevada de los que no han incurrido en falta de pago.

DEFAULT	0	1	Grand Total
MEDIANA EMPRESA	343 91.47%	32 8.53%	375 100.00%
PEQUEÑA EMPRESA	1,175 94.53%	68 5.47%	1,243 100.00%
MICRO EMPRESA	1,239 96.05%	51 3.95%	1,290 100.00%
Grand Total	2,757 94.81%	151 5.19%	2,908 100.00%

En porcentaje, las empresas *Medianas* son las que más han defaulteado.

En toda esta muestra tenemos un 5.19% de créditos en *default.*

A continuación presentamos los resultados de las regresiones para el primer tramo:

Las variables *Antigüedad del cliente, Tasa de interés, Condición de si posee préstamo anterior con la entidad bancaria y Moneda del crédito* son significativas individualmente a un nivel de confianza del 99.9% en los 3 modelos. La variable *Cantidad de Empleados,* es significativa estadísticamente a un nivel de confianza del 99% también en los 3 modelos.

Según los criterios **AIC** y **BIC** el modelo con mejor ajuste es **log_model_1_v3.** Las variables de Actividad Sectorial no son significativas invididualmente en los primeros dos modelos.

	log_model_1	log_model_1_v2	log_model_1_v3
(Intercept)	-19.89	-20.30	-20.43
	(603.77)	(599.79)	(606.91)
anos_cliente	-0.08 ***	-0.08 ***	-0.08 ***
	(0.02)	(0.02)	(0.02)
FACTURACION_ANUAL_ESTIMADA	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)
TNA	0.15 ***	0.15 ***	0.15 ***
	(0.02)	(0.02)	(0.02)
PLAZOMESES	-0.01	-0.01	-0.01
	(0.01)	(0.01)	(0.01)
ImporteOtorgado	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)
posee_pfdolar	-0.01	-0.11	-0.03
	(1.10)	(1.09)	(1.08)
posee_pfpesos	-1.32	-1.38	-1.34
	(1.03)	(1.03)	(1.03)
posee_prestamoanterior	4.21 ***	4.20 ***	4.19 ***
	(1.01)	(1.01)	(1.01)
posee_tarjetacorporativa	0.14	0.12	0.12
	(0.22)	(0.22)	(0.22)
posee_bancainternet	14.75	14.85	14.75
	(603.77)	(599.79)	(606.91)
posee_seguroART	-0.01	-0.01	-0.02
	(0.22)	(0.21)	(0.21)
Cuentassueldo	-0.01	-0.01	-0.01
	(0.01)	(0.01)	(0.01)
CABA	0.18	0.20	0.24
	(0.31)	(0.31)	(0.30)
CONURBANO	0.26	0.27	0.39
	(0.25)	(0.24)	(0.21)
PEQUENA_EMPRESA	0.43	0.42	0.43
	(0.23)	(0.22)	(0.22)
MEDIANA_EMPRESA	0.30	0.28	0.31
	(0.44)	(0.43)	(0.43)
capital_trabajo	0.32	0.30	0.26
	(0.34)	(0.34)	(0.34)
en_pesos	-4.42 ***	-4.44 ***	-4.43 ***
	(0.51)	(0.51)	(0.50)
cantidad_empleados	0.01 **	0.01 **	0.01 **
	(0.00)	(0.00)	(0.00)
ACT_BCRA1_AGROINDUSTRIA	-0.37	-0.25	
	(0.47)	(0.27)	
ACT_BCRA1_COMERCIO	-0.37	-0.46	
	(0.44)	(0.29)	
ACT_BCRA1_SERVICIOS	-0.69	-0.75	
	(0.64)	(0.55)	
ACT_BCRA3_AGROPECUARIO	-0.19		
	(0.59)		
ACT_BCRA3_ALIMENTOS	-0.25		
	(0.55)		
ACT_BCRA3_AUTOMOTOR	-0.70		
	(0.64)		
ACT_BCRA3_COMBUSTIBLES	-14.54		
	(2013.38)		
ACT_BCRA3_CONSTRUCCION	-0.57		
	(0.53)		
ACT_BCRA3_MANUFACTURA	-0.40		
	(0.42)		
ACT_BCRA3_METALES	-0.44		
	(0.51)		
ACT_BCRA3_MINERIA	0.38		
	(1.38)		
ACT_BCRA3_PETROQUIMICA	-0.30		
	(0.68)		
ACT_BCRA3_PLASTICO	-0.14		
	(0.58)		
ACT_BCRA3_SERVICIO	-0.41		
	(0.51)		
N	2908	2908	2908
AIC	940.48	921.32	919.51
BIC	1143.64	1058.75	1039.01
Pseudo R2	0.31	0.30	0.30

*** p < 0.001; ** p < 0.01; * p < 0.05.

Corremos el test de *Razón de Verosimilitud* arribando a los siguientes p-values:

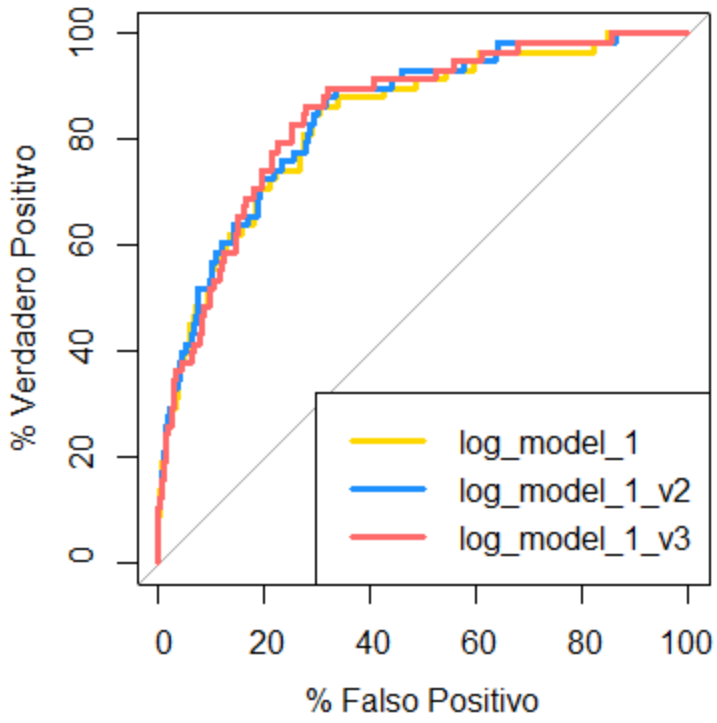
p_value_1=0.00000000000000000000000000000000000000000000001041658

p_value_1_v2=0.00000000000000000000000000000000000000000000000004472613

p_value_1_v3=0.00000000000000000000000000000000000000000000000500390

El test de *Razón de Verosimilitud* rechaza la hipótesis nula en todos los modelos a un nivel de confianza del 99% lo que indica significatividad conjunta de las variables.

Ahora graficamos las curvas de ROC para cada modelo:



A simple vista la curva ROC Roja del tercer modelo parece estar mejor ubicada que los primeros 2 modelos. Verificando las *AUC:*

AUC log_model_1: 0.836

AUC log_model_1_v2 0.8454

AUC log_model_1_v3: 0.8469

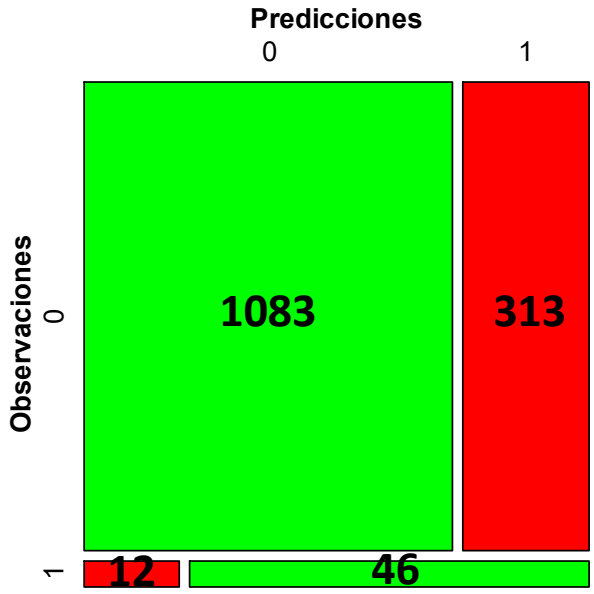
Estamos en condición de afirmar que el mejor modelo es **log_model_1_v3** ya que tiene una mayor *AUC* por lo que tiene mayor *Tasa VerdaderaPositiva* frente a la *FalsaPositiva* que el resto de los modelos.

Eligiendo el modelo **log_model_1_v3** hemos de seleccionar el umbral óptimo de corte. Generamos la tabla para el análisis de corte:

ID	tpp	fpp	thresholds	KS
1096	79.31034	22.42120	0.04971094	56.88914
1046	82.75862	25.85960	0.04174776	56.89902
.	.	.	.	.
1012	86.20690	28.15186	0.03714730	58.05503
1013	86.20690	28.08023	0.03718812	58.12667
1014	86.20690	28.00860	0.03721420	58.19830

Elegimos el ***umbral t*** en **0.049** ya que tiene un alto *K-S*, y además la tasa de *FalsosPositivos* es menor a un 25% mientras que la tasa de *VerdaderosPositivos* está casi en un 80%.

Tomando en cuenta el umbral obtenido, reformulamos una matriz de confusión para el set de *Testing:*



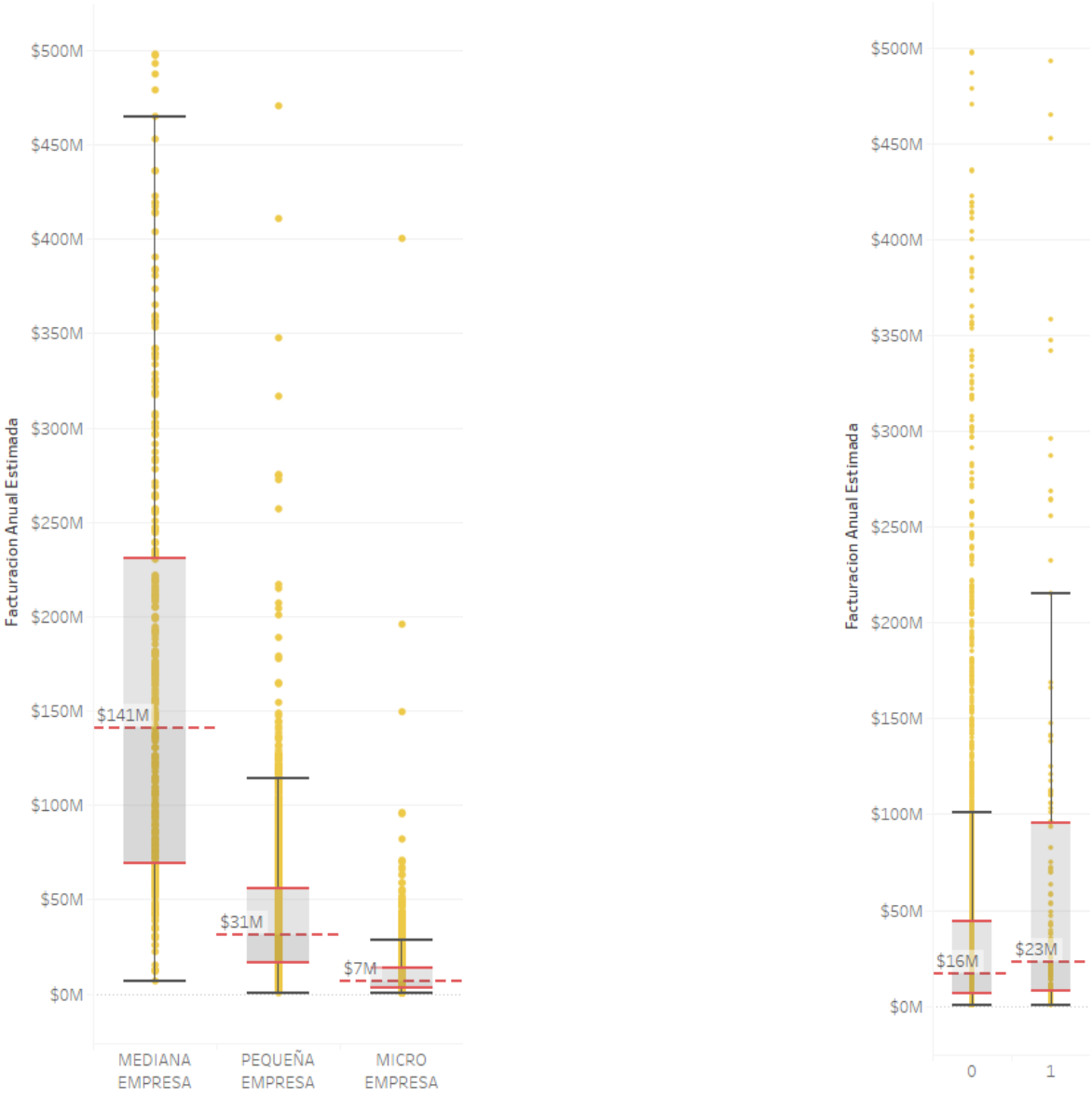
Con la nueva matriz pasamos a calcular los índices de ajuste del modelo.

$$Precisión = \frac{46 + 1083}{46 + 313 + 1083 + 12} = 77.64\%$$

$$Sensibilidad = \frac{46}{46 + 12} = 79.31\%$$

$$Especificidad = \frac{1083}{1083 + 313} = 77.57\%$$

Modelo para el Tramo 2 (log_model_2)



Podemos observar que la mediana de **Facturación Anual Estimada** es acorde al tamaño de la empresa. Por otro lado, vemos que en los clientes morosos, la mediana de facturación anual es más elevada de los que no han incurrido en falta de pago.

DEFAULT	0	1	Grand Total
MEDIANA EMPRESA	293 90.71%	30 9.29%	323 100.00%
PEQUEÑA EMPRESA	1,131 95.77%	50 4.23%	1,181 100.00%
MICRO EMPRESA	1,351 96.23%	53 3.77%	1,404 100.00%
Grand Total	2,775 95.43%	133 4.57%	2,908 100.00%

En porcentaje, las empresas *Medianas* son las que más han defaulteado.

En toda esta muestra tenemos un 4.57% de créditos en *default.*

A continuación presentamos los resultados de las regresiones para el segundo tramo:

Las variables *Antigüedad del cliente, Tasa de interés, Condición de si posee préstamo anterior con la entidad bancaria y Moneda del crédito* son significativas individualmente a un nivel de confianza del 99.9%. La variable *Cantidad de Cuentas Sueldo en el Banco,* es significativa estadísticamente a un nivel de confianza del 95% en los modelos **log_model_2_v2** y **log_model_3_v2**. El *Importe Otorgado* es significativo individualmente a un nivel del 99% en **log_model_2** y al 95% en los otros**.**

Según los criterios **AIC** y **BIC** el modelo con mejor ajuste es **log_model_2_v3.** Las variables de Actividad Sectorial no son significativas invididualmente en los primeros dos modelos.

	log_model_2	log_model_2_v2	log_model_2_v3
(Intercept)	-19.88	-20.41	-20.73
	(593.30)	(593.76)	(591.29)
anos_cliente	-0.10 ***	-0.11 ***	-0.10 ***
	(0.02)	(0.02)	(0.02)
FACTURACION_ANUAL_ESTIMADA	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)
TNA	0.15 ***	0.15 ***	0.15 ***
	(0.02)	(0.02)	(0.02)
PLAZOMESES	-0.00	-0.00	-0.00
	(0.01)	(0.01)	(0.01)
ImporteOtorgado	0.00 **	0.00 *	0.00 *
	(0.00)	(0.00)	(0.00)
posee_pfdolar	-14.15	-14.40	-14.48
	(953.23)	(977.56)	(975.34)
posee_pfpesos	-0.76	-0.76	-0.78
	(1.03)	(1.02)	(1.02)
posee_prestamoanterior	4.17 ***	4.15 ***	4.16 ***
	(1.02)	(1.02)	(1.02)
posee_tarjetacorporativa	-0.01	-0.02	-0.02
	(0.22)	(0.22)	(0.22)
posee_bancainternet	14.89	14.92	14.94
	(593.30)	(593.76)	(591.28)
posee_seguroART	-0.10	-0.08	-0.08
	(0.24)	(0.23)	(0.23)
Cuentassueldo	-0.02	-0.02 *	-0.02 *
	(0.01)	(0.01)	(0.01)
CABA	0.08	0.15	0.17
	(0.32)	(0.31)	(0.30)
CONURBANO	-0.12	-0.05	0.02
	(0.27)	(0.26)	(0.24)
PEQUENA_EMPRESA	0.28	0.27	0.23
	(0.24)	(0.24)	(0.24)
MEDIANA_EMPRESA	0.63	0.62	0.54
	(0.43)	(0.43)	(0.41)
capital_trabajo	0.60	0.62	0.61
	(0.36)	(0.36)	(0.35)
en_pesos	-4.34 ***	-4.28 ***	-4.32 ***
	(0.57)	(0.57)	(0.55)
cantidad_empleados	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)
ACT_BCRA1_AGROINDUSTRIA	-0.10	-0.27	
	(0.56)	(0.33)	
ACT_BCRA1_COMERCIO	0.00	-0.34	
	(0.56)	(0.31)	
ACT_BCRA1_SERVICIOS	-0.20	-0.48	
	(0.57)	(0.33)	
ACT_BCRA3_AGROPECUARIO	-0.68		
	(0.66)		
ACT_BCRA3_ALIMENTOS	-0.46		
	(0.57)		
ACT_BCRA3_AUTOMOTOR	-8.94		
	(3956.18)		
ACT_BCRA3_COMBUSTIBLES	-4.03		
	(2.70)		
ACT_BCRA3_CONSTRUCCION	-15.02		
	(1232.91)		
ACT_BCRA3_MANUFACTURA	-0.78		
	(0.57)		
ACT_BCRA3_METALES	-14.21		
	(2052.15)		
ACT_BCRA3_MINERIA	-3.41		
	(3.64)		
ACT_BCRA3_PETROQUIMICA	0.19		
	(1.22)		
ACT_BCRA3_PLASTICO	-15.08		
	(2224.93)		
ACT_BCRA3_SERVICIO	-0.73		
	(0.61)		
N	2908	2908	2908
AIC	840.97	828.02	824.33
BIC	1044.12	965.45	943.84
Pseudo R2	0.32	0.31	0.31

*** p < 0.001; ** p < 0.01; * p < 0.05.

Corremos el test de *Razón de Verosimilitud* arribando a los siguientes p-values:

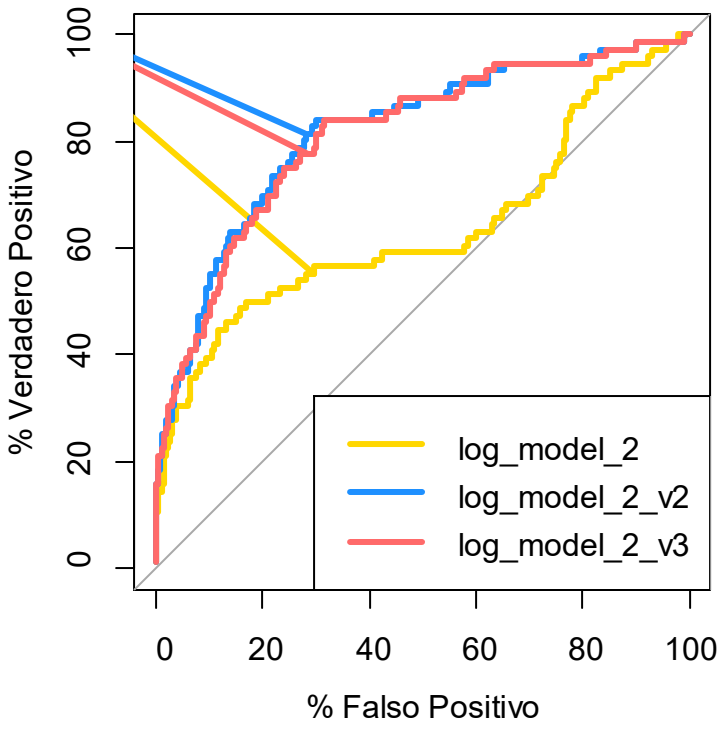
p_value_2= 0.00000000000000000000000000000000000000000000002928291

p_value_2_v2= 0.00000000000000000000000000000000000000000000000259868

p_value_2_v3= 0.00000000000000000000000000000000000000000000000127958

El test de *Razón de Verosimilitud* rechaza la hipótesis nula en todos los modelos a un nivel de confianza del 99% lo que indica significatividad conjunta de las variables.

Ahora graficamos las curvas de ROC para cada modelo:



Las curvas del 2do y 3er modelo son muy parecidas. Mención aparte al pésimo ajuste del primer modelo que tiene una muy mala capacidad de predicción. Para elegir entre el 2do y 3er modelo vemos las medidas de las *AUC:*

AUC log_model_2: 0.6366

AUC log_model_2_v2: 0.818

AUC log_model_2_v3: 0.8126

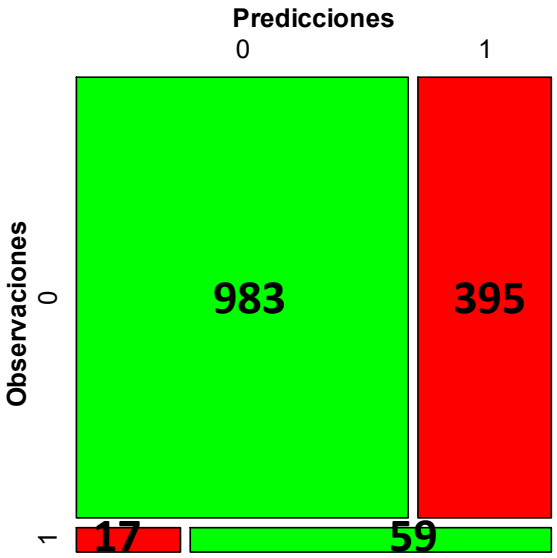
Estamos en condición de afirmar que el mejor modelo es **log_model_2_v3** ya que tiene una mayor *AUC* por lo que tiene mayor *Tasa VerdaderaPositiva* frente a la *FalsaPositiva* que el resto de los modelos.

Eligiendo el modelo **log_model_2_v3** hemos de seleccionar el umbral óptimo de corte. Generamos la tabla para el análisis de corte:

ID	tpp	fpp	thresholds	KS
968	81.57895	30.84180	0.03294574	50.73715
.	.	.	.	.
1069	75.00000	23.87518	0.04618870	51.12482
.	.	.	.	.
953	84.21053	31.78520	0.03152589	52.42533
954	84.21053	31.71263	0.03166624	52.49790

Elegimos el ***umbral t*** en **0.0461** ya que tiene un alto *K-S*, y además la tasa de *FalsosPositivos* es menor a un 23.87% mientras que la tasa de *VerdaderosPositivos* está casi en un 75%.

Tomando en cuenta el umbral obtenido, reformulamos una matriz de confusión para el set de *Testing:*



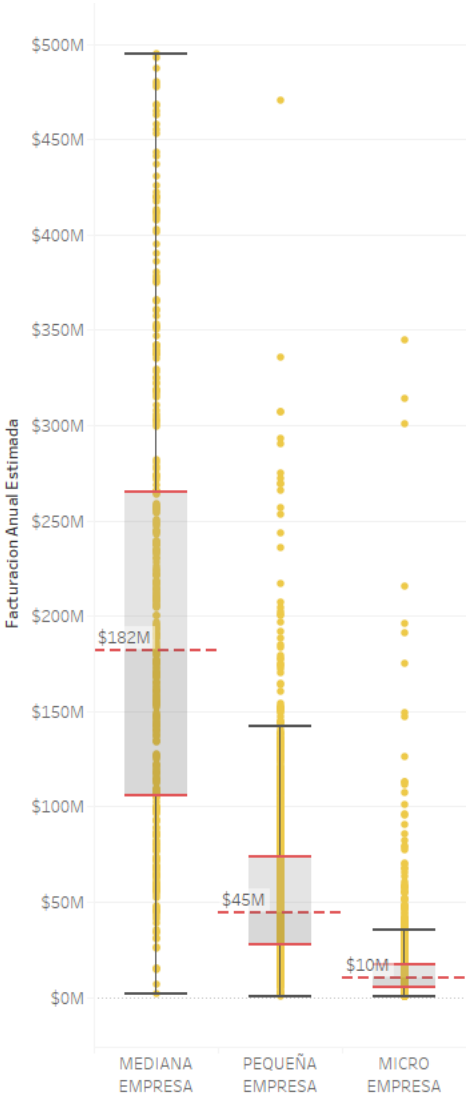
Con la nueva matriz pasamos a calcular los índices de ajuste del modelo.

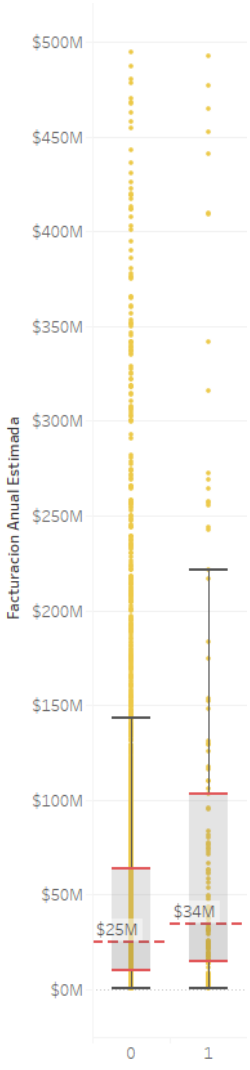
$$Precisión = \frac{59 + 983}{59 + 395 + 983 + 17} = 71.66\%$$

$$Sensibilidad = \frac{59}{59 + 17} = 77.63\%$$

$$Especificidad = \frac{983}{983 + 395} = 71.33\%$$

Modelo para el Tramo 3 (log_model_3)





Podemos observar que la mediana de **Facturación Anual Estimada** es acorde al tamaño de la empresa. Por otro lado, vemos que en los clientes morosos, la mediana de facturación anual es más elevada de los que no han incurrido en falta de pago.

DEFAULT	0	1	Grand Total
MEDIANA EMPRESA	346 92.51%	28 7.49%	374 100.00%
PEQUEÑA EMPRESA	1,122 95.08%	58 4.92%	1,180 100.00%
MICRO EMPRESA	1,306 96.45%	48 3.55%	1,354 100.00%
Grand Total	2,774 95.39%	134 4.61%	2,908 100.00%

En porcentaje, las empresas *Medianas* son las que más han defaulteado.

En toda esta muestra tenemos un 4.61% de créditos en *default.*

A continuación presentamos los resultados de las regresiones para el tercer tramo:

Las variables *Antigüedad del cliente, Tasa de interés, Condición de si posee préstamo anterior con la entidad bancaria y Moneda del crédito* son significativas individualmente a un nivel de confianza del 99.9%. La variable *Cantidad de empleados de la empresa,* es significativa estadísticamente a un nivel de confianza del 95% en los modelos **log_model_3** y **log_model_3_v2**; y a un nivel de 99% en el modelo **log_model_3_v3.**

Según los criterios **AIC** y **BIC** el modelo con mejor ajuste es **log_model_3_v3.** Las variables de Actividad Sectorial no son significativas invididualmente en los primeros dos modelos.

	log_model_3	log_model_3_v2	log_model_3_v3
(Intercept)	-16.86	-18.05	-18.40
	(535.41)	(532.13)	(539.73)
anos_cliente	-0.07 ***	-0.07 ***	-0.07 ***
	(0.02)	(0.02)	(0.02)
FACTURACION_ANUAL_ESTIMADA	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)
TNA	0.14 ***	0.13 ***	0.14 ***
	(0.02)	(0.02)	(0.02)
PLAZOMESES	0.01	0.00	0.00
	(0.01)	(0.01)	(0.01)
ImporteOtorgado	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)
posee_pfdolar	0.07	-0.07	-0.03
	(1.11)	(1.08)	(1.09)
posee_pfpesos	-0.60	-0.63	-0.54
	(0.77)	(0.76)	(0.76)
posee_prestamoanterior	3.33 ***	3.27 ***	3.27 ***
	(0.73)	(0.73)	(0.73)
posee_tarjetacorporativa	-0.22	-0.23	-0.23
	(0.22)	(0.22)	(0.22)
posee_bancainternet	14.34	14.38	14.32
	(535.41)	(532.13)	(539.73)
posee_seguroART	0.02	0.02	0.01
	(0.23)	(0.22)	(0.22)
Cuentassueldo	-0.01	-0.01	-0.01
	(0.01)	(0.01)	(0.01)
CABA	0.15	0.22	0.27
	(0.32)	(0.32)	(0.32)
CONURBANO	0.26	0.30	0.42
	(0.23)	(0.23)	(0.22)
PEQUENA_EMPRESA	0.36	0.32	0.32
	(0.25)	(0.24)	(0.24)
MEDIANA_EMPRESA	0.04	0.00	0.02
	(0.48)	(0.47)	(0.46)
capital_trabajo	0.23	0.23	0.17
	(0.37)	(0.37)	(0.37)
en_pesos	-5.17 ***	-5.13 ***	-5.03 ***
	(0.63)	(0.63)	(0.60)
cantidad_empleados	0.01 *	0.01 *	0.01 **
	(0.00)	(0.00)	(0.00)
ACT_BCRA1_AGROINDUSTRIA	-0.90	-0.77	
	(0.67)	(0.51)	
ACT_BCRA1_COMERCIO	-0.48	-0.41	
	(0.53)	(0.25)	
ACT_BCRA1_SERVICIOS	-0.51	-0.38	
	(0.55)	(0.27)	
ACT_BCRA3_AGROPECUARIO	-1.34		
	(1.70)		
ACT_BCRA3_ALIMENTOS	-15.10		
	(858.82)		
ACT_BCRA3_AUTOMOTOR	-1.63		
	(1.31)		
ACT_BCRA3_COMBUSTIBLES	-3.75		
	(2.43)		
ACT_BCRA3_CONSTRUCCION	-1.42		
	(1.24)		
ACT_BCRA3_MANUFACTURA	-1.18		
	(1.21)		
ACT_BCRA3_METALES	-1.29		
	(1.23)		
ACT_BCRA3_PETROQUIMICA	-1.45		
	(1.35)		
ACT_BCRA3_PLASTICO	-1.00		
	(1.26)		
ACT_BCRA3_SERVICIO	-1.12		
	(1.28)		
N	2908	2908	2908
AIC	869.01	855.17	853.66
BIC	1066.20	992.60	973.17
Pseudo R2	0.30	0.29	0.29

*** p < 0.001; ** p < 0.01; * p < 0.05.

Corremos el test de *Razón de Verosimilitud* arribando a los siguientes p-values:

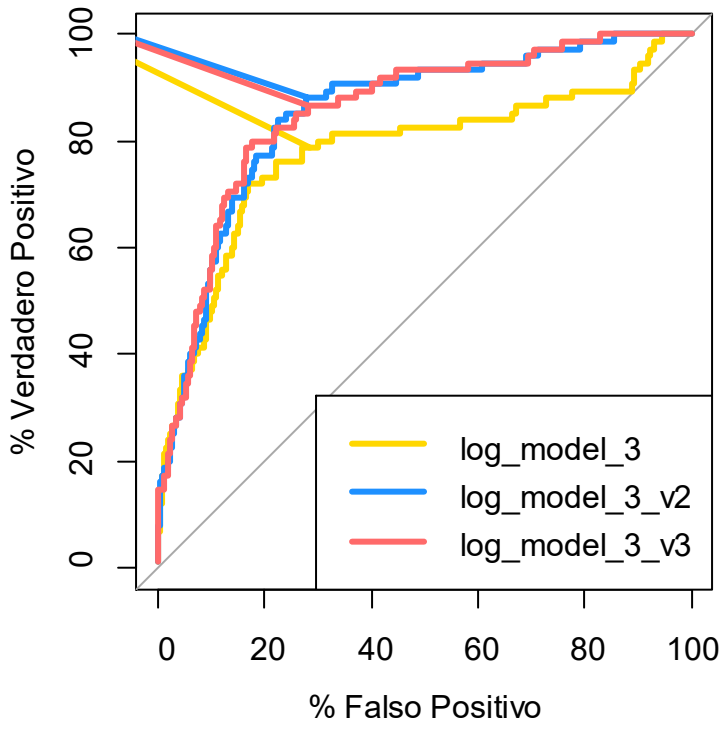
p_value_3= 0.0000000000000000000000000000000000000000004490747

p_value_3_v2= 0.0000000000000000000000000000000000000000000004756327

p_value_3_v3= 0.0000000000000000000000000000000000000000000007214178

El test de *Razón de Verosimilitud* rechaza la hipótesis nula en todos los modelos a un nivel de confianza del 99% lo que indica significatividad conjunta de las variables.

Ahora graficamos las curvas de ROC para cada modelo:



Definitivamente la curva del primer modelo tiene baja capacidad predictiva. Sin embargo, deberemos ahondar en las *AUC* del 2do y 3er modelo para establecer cuál es el mejor:

AUC log_model_3: 0.7787

AUC log_model_3_v2: 0.8522

AUC log_model_3_v3: 0.8558

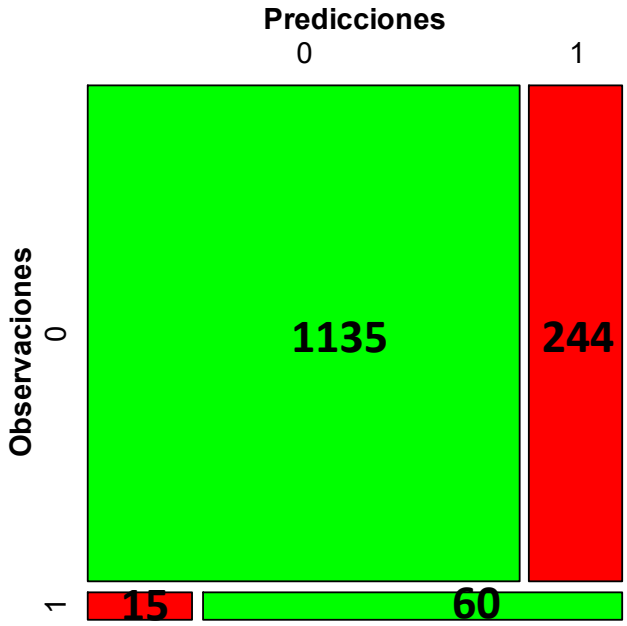
Estamos en condición de afirmar que el mejor modelo es **log_model_3_v3** ya que tiene una mayor *AUC* por lo que tiene mayor *Tasa VerdaderaPositiva* frente a la *FalsaPositiva* que el resto de los modelos.

Eligiendo el modelo **log_model_3_v3** hemos de seleccionar el umbral óptimo de corte. Generamos la tabla para el análisis de corte:

ID	tpp	fpp	thresholds	KS
1123	80.00000	19.72444	0.04602129	60.27556
.	.	.	.	.
1150	80.00000	17.76650	0.04900969	62.23350
1151	80.00000	17.69398	0.04910338	62.30602
1152	80.00000	17.62146	0.04932804	62.37854

Elegimos el ***umbral t*** en **0.0491** ya que tiene un alto *K-S*, y además la tasa de *FalsosPositivos* es menor a un 17.7% mientras que la tasa de *VerdaderosPositivos* está en un 80%.

Tomando en cuenta el umbral obtenido, reformulamos una matriz de confusión para el set de *Testing:*



Con la nueva matriz pasamos a calcular los índices de ajuste del modelo.

$$Precisión = \frac{60 + 1132}{60 + 244 + 1135 + 15} = 81.98\%$$

$$Sensibilidad = \frac{60}{60 + 15} = 80.0\%$$

$$Especificidad = \frac{1135}{1135 + 244} = 82.30\%$$

5. Conclusiones

Habiendo planteado 3 diferentes modelos para cada tramo de la muestra con datos reales podemos aseverar lo siguiente:

Las únicas variables que resultaron estadísticamente significativas a un nivel de confianza del 99.9% fueron la ***Antigüedad del cliente, Tasa de interés del Préstamo, Condición de si posee un préstamo anterior con la entidad bancaria y Moneda del Préstamo***. De forma no general, resultaron significativas a un nivel de confianza entre 95 y 99% las variables de ***Cantidad de Empleados, Cantidad de Cuentas Sueldo en el Banco e Importe Otorgado del Crédito***.

Podemos decir entonces, que las variables de características intrínsecas de las empresas como *Actividad, Ubicación y Tamaño* no fueron de significatividad estadística individual para el modelo. Particularmente la variable de *Actividad* en niveles 1 y 3 sólo ha desmejorado la capacidad predictiva de los modelos a lo largo de los 3 tramos testeados. Una vez habiendolas sacado del modelo, las predicciones mejoraron sustancialmente. Las causas de su no significatividad pueden ser múltiples: sesgo en la muestra, tamaño de muestra insuficiente para la cantidad de actividades, alta correlación entre variables, método de estimación inadecuado, etc. Los pasos a seguir para develar su significatividad como indica la literatura vigente son probar abarcar una muestra más grande en diferentes períodos de tiempo y/o explorar otros métodos de estimación.

A pesar de quedarnos con el modelo más reducido a lo largo de los 3 tramos de la muestra, éste nos dio mejor poder predictivo medido a través de la técnica de la curva ROC y el índice AUC. Esto responde muy bien al criterio de *parsimonia*; la solución más simple parece haber sido la mejor.

Los umbrales de probabilidad estimada que hemos elegido para el corte óptimo de separar un crédito *Bueno* de uno *Malo* en las 3 muestras se encuentran **entre 0.046 y 0.049**. Redondeando, si un cliente nuevo, por fuera de la muestra, querrá solicitar un crédito en el Banco obteniendo un *score* **superior al 5% sabremos que tiene una alta probabilidad de entrar en cesación de pagos en el futuro**. Por lo tanto, el Banco debería rechazar su petición. Caso contrario, si su *score* fuese menor al 5%, la entidad bancaria debería estar gustosa de otorgar dicho crédito. Claro está, que cada entidad bancaria define su riesgo a tolerar a través de su política, y que la *probabilidad de default* es sólo un componente junto a otras 2: *pérdida por default* (LGD: Loss Given Default) y la *exposición al default* (EAD: Exposure at Default).

En línea generales, el modelo tiene un desempeño aceptable de acuerdo con los criterios generales de *Precisión, Sensibilidad y Especificidad.* Calculando sobre las matrices de confusión generadas, se obtuvo una *Precisión* promedio del 77.09%, *Sensibilidad* del 78.98% y *Especificidad* del 77.06%. Es decir, el ajuste general del modelo es más que aceptable, y además tiene mejor resultados a la hora de predecir *VerdaderosPositivos* (o sea morosos) que *VerdaderosNegativos* (o sea buenos pagadores) aunque ambos indicadores son bastante aceptables.

Podemos afirmar entonces, que este método sirve como complemento del análisis crediticio tradicional pero aún no puede ser un sustituto perfecto. De todas maneras, podría relajar la necesidad de un análisis crediticio exhaustivo o desechar la necesidad de un colateral para préstamos de hasta cierto monto.

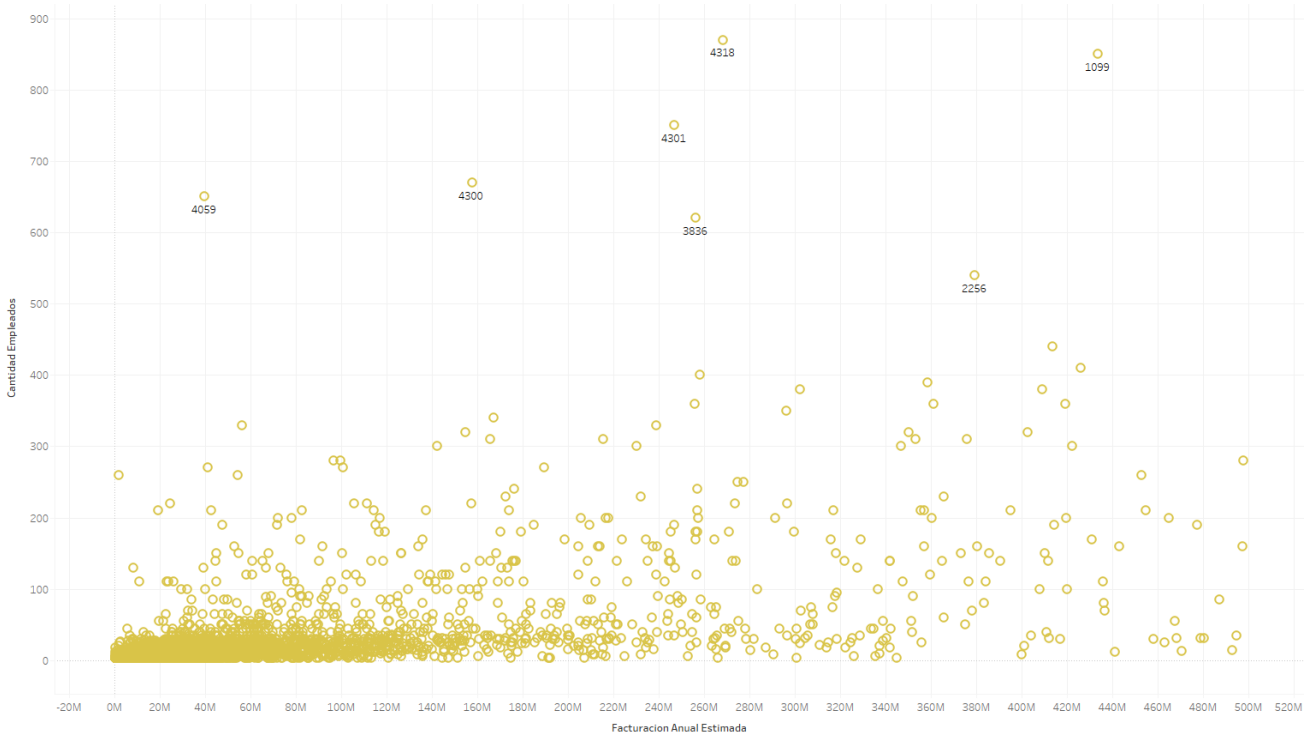
Para ahondar en futuros trabajos, este monto podría estimarse a partir del mismo modelo dando por resultado el *monto cierto* que podría otorgársele a cada cliente sin necesidad de un colateral financiero y que este podría devolver con una alta probabilidad sin caer en default. Además, se deberían explorar otras técnicas estimativas como *Árboles de decisión, Análisis discriminativo, Redes Neuronales y Análisis de Clusters* a ser aplicadas con una base de datos más grande para incluir mayor cantidad de variables como las de Actividad y que estas resulten significativas.

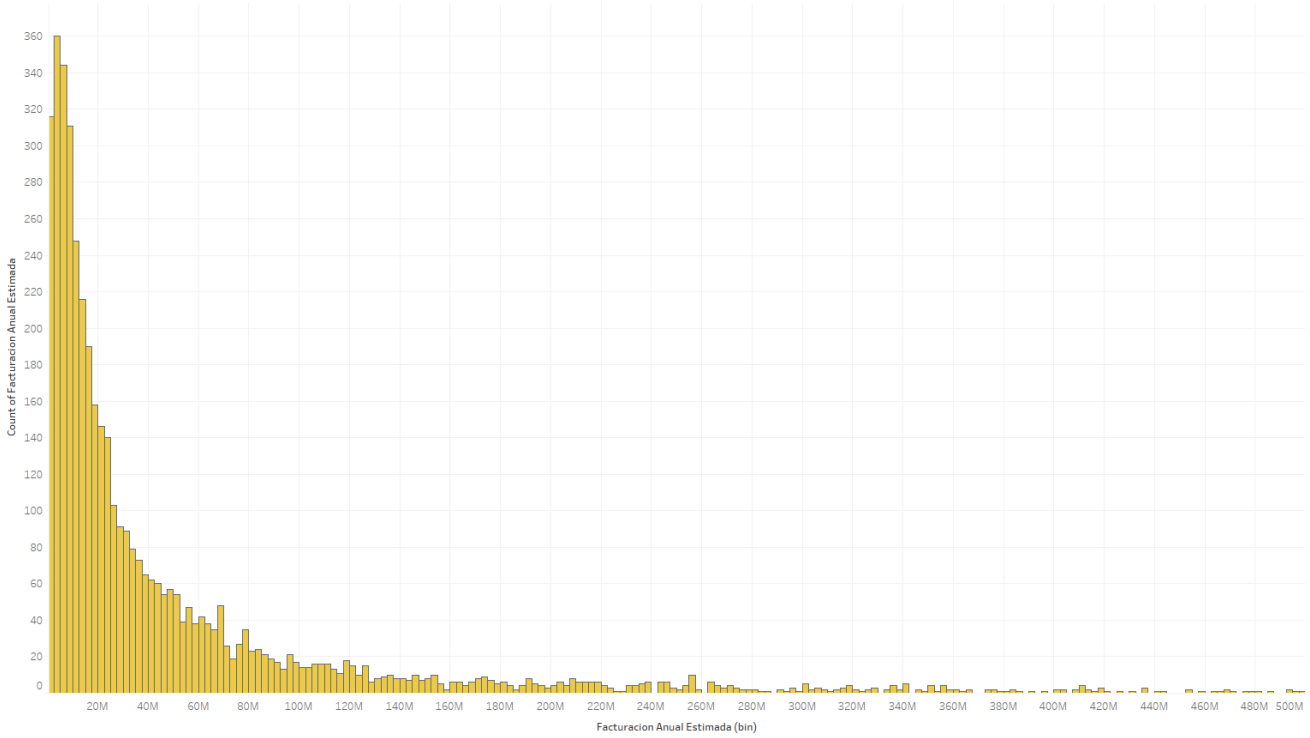
6. Bibliografía

- Anderson, R. (2007). *The credit scoring toolkit: theory and practice for retail credit risk management and decision automation*. Oxford University Press.
- Bensic, M., Sarlija, N., & Zekic-Susac, M. (2005). Modelling small-business credit scoring by using logistic regression, neural networks and decision trees. *Intelligent Systems in Accounting, Finance & Management: International Journal*, *13*(3).
- Blanco, J. (2006). Introducción al análisis multivariado. *IESTA. Montevideo*.23.
- Bolton, C. (2009). *Logistic regression and its application in credit scoring* (Doctoral dissertation, University of Pretoria).
- Caracota, R. C., Dimitriu, M., & Dinu, M. R. (2010). Building a scoring model for small and medium enterprises. *Theoretical and applied economics*, *9*(9), 117.
- Hand, D. J. (2005). Good practice in retail credit scorecard assessment. *Journal of the Operational Research Society*, *56*(9).
- Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *160*(3)
- James H. Myers & Edward W. Forgy (1963) The Development of Numerical Credit Evaluation Systems, Journal of the American Statistical Association, 58:303, 799-806
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

- Krzanowski, W. J., & Hand, D. J. (2009). *ROC curves for continuous data*. Chapman and Hall/CRC.
- Mathew, Ansen, "CREDIT SCORING USING LOGISTIC REGRESSION" (2017). Master's Projects.
- Orgler, Y. E. (1970). A credit scoring model for commercial loans. *Journal of money, Credit and Banking*, *2*(4)
- Sanguinetti, P., Arreaza, A., Rodríguez, P., Álvarez, F., Ortega, D., & Penfold, M. (2011). RED 2011: Servicios financieros para el desarrollo. Promoviendo el acceso en América Latina (Reporte de Economía y Desarrollo (RED)). Caracas: CAF.

7. ANEXO I


Cantidad Empleados
Facturacion Anual Estimada
4318
1099
4301
4059
4300
3836
2256


Count of Facturacion Anual Estimada
Facturacion Anual Estimada (bin)

CATEGORÍA	ACTIVIDAD				
	Construcción	Servicios	Comercio	Industria y minería	Agropecuario
Micro	$ 5.900.000	$ 4.600.000	$ 15.800.000	$ 13.400.000	$ 3.800.000
Pequeña	$ 37.700.000	$ 27.600.000	$ 95.000.000	$ 81.400.000	$ 23.900.000
Mediana tramo 1	$ 301.900.000	$ 230.300.000	$ 798.200.000	$ 661.200.000	$ 182.400.000
Mediana tramo 2	$ 452.800.000	$ 328.900.000	$ 1.140.300.000	$ 966.300.000	$ 289.300.000

FUENTE: https://www.argentina.gob.ar/noticias/nuevas-categorias-para-ser-pyme Mayo-2018

SECTOR / CATEGORÍA	AGROPECUARIO	INDUSTRIA Y MINERÍA	COMERCIO	SERVICIOS	CONSTRUCCIÓN
MICRO	$ 3.000.000	$ 10.500.000	$ 12.500.000	$ 3.500.000	$ 4.700.000
PEQUEÑA	$ 19.000.000	$ 64.000.000	$ 75.000.000	$ 21.000.000	$ 30.000.000
MEDIANA TRAMO 1	$ 145.000.000	$ 520.000.000	$ 630.000.000	$ 175.000.000	$ 240.000.000
MEDIANA TRAMO 2	$ 230.000.000	$ 760.000.000	$ 900.000.000	$ 250.000.000	$ 360.000.000

FUENTE: https://www.infobae.com/economia/2017/04/03/cuales-son-los-nuevos-montos-de-facturacion-para-saber-que-empresas-son-pyme/ Abril-2017