UNIVERSIDAD TORCUATO DI TELLA

Departamento de Economía

Maestría En Econometría

Una Comparación entre Regresión Logística y Random Forrest para Estimar la Probabilidad de Default

María Paula Busto

Tutor: Martín González Rozada

2 de Diciembre de 2013

Introducción:

Hoy en día, la técnica más utilizada por las entidades bancarias y financieras para intentar detectar el riesgo de incumplimiento de una obligación asumida es la regresión logística.

Esta técnica presenta muchas ventajas pero también tiene algunas debilidades; es por ello que en el presente trabajo se decide comparar dicha técnica con la de Random Forrest para tratar de determinar cuál de ellas arroja mejores resultados.

Objetivos del trabajo

Los objetivos del estudio son dos:

1. Comparar la Regresión Logística y Random Forrest con respecto a las variables seleccionadas por cada técnica.
2. Comparar la Regresión Logística y Random Forrest con respecto a la precisión de clasificación de cada técnica

Materiales y Métodos:

Para poder realizar dicha comparación, se utilizará una base de datos de un banco paraguayo y se construirán un modelo para aquellos individuos que ya tienen una vinculación previa con la entidad (clientes del banco) y que solicitan un crédito. La base contiene datos de 46026 solicitudes de crédito para un segmento muy particular del mercado de crédito. Se posee poca información de los individuos analizados por lo que en general suelen quedar marginados de otras instituciones de crédito. El porcentaje de morosidad en este data set es de 7,6% aproximadamente y las solicitudes analizadas abarcan el período que va desde enero 2009 hasta julio de 2010.

Se utilizará para la construcción del modelo ambas técnicas: Random Forest y Regresión Logística.

La variable dependiente utilizada en los modelos es: Malo 60 indeleble. Es decir, si el individuo alcanza o no un máximo de 60 días de atraso alguna vez, más allá de que luego se ponga al día (de ahí la marca de indeleble), y se utilizan diversas variables independientes.

En el caso de la regresión logística, se trabajarán las variables originales de forma de eliminar los missing values y outliers. Para cada variable dependiente seleccionada se toma alguna decisión pertinente teniendo en cuenta la cantidad de valores missings y su significancia debido a que la regresión logística no tolera missing values y es sensible a los outliers. También se realizaron distintos tramados de las variables y muchas de las variables continuas se incorporan en la regresión como piecewise. Se realizaron varias regresiones eligiendo variables

hasta obtener resultados deseados. Se presenta el mejor de los resultados obtenidos en cuanto a la calidad de sus indicadores y al sentidos de los betas. La regresión se correrá utilizando la opción fowardstep. Que permite hacer una selección de variables.

En el caso de la construcción del modelo de Random Forest, se realizan dos corridas. Una primera corrida con una gran cantidad de variables, incluso con variables redundantes, con el propósito de descartar y seleccionar las variables más relevantes y una segunda corrida seleccionado variables en base a los resultados de la primera corrida. En ambas corridas se incorporaron tanto variables originales de la base como variables transformadas que también se utilizan en la regresión logística. La técnica random forest permite trabajar tanto con missing values como con outlies de manera sencilla, por lo que no es necesario hacer ningún tratamiento previo a las variables.

Resultados y Conclusiones

Los resultados de las estimaciones mejoran en las corridas de random forest.

Mientras para la logística, se obtiene un valor de ROC y de K-S de 72,17% y 32,61% respectivamente, las dos corridas de random forest mejoran este resultado, aunque podría decirse que no sustancialmente. La corrida que incluye todas las variables obtiene un valor de ROC y K-S de 74,54% y 36,09% y la corrida con las variables seleccionadas de random forext obtiene valores para la ROC y K-S de 73,63% y 34,57% respectivamente. El gini también mejora en los modelos de random forest respecto al de regresión logística, siendo en este último un 44,34% y en los modelos de random forest 49,09% para el que incluye todas las variables y 47,26% para el que incluye variables seleccionadas.

Si bien los resultados de las corridas del random forest con todas las variables parecieran arrojar mejores resultados, se debe tener en cuenta que existen muchas variables redundantes que podrían generar mucha correlación, por lo que es mejor tener en cuenta los resultados de la corrida con las variables seleccionadas.

Con respecto a las variables seleccionadas por cada modelo, algunas son similares entre random forest y regresión logística. Pero sin embagro, random forest selecciona algunas otras variables que no se vieron involucradas en la regresión logística.

Índice

Introducción

Contexto de Análisis

Como consecuencia de la informalidad en el empleo que tiene lugar en Paraguay, existe un grupo de solicitantes desatendidos por no tener la información que se requiere para acceder a los créditos que se encuentran actualmente en el mercado. Sin embargo, si bien se posee en estos casos menos información de la usual a la hora del otorgamiento de un crédito, sí se tiene de forma directa o indirecta datos de la familia, de donde vive, de donde trabaja y del bureau externo (Informconf).

A partir de esa información se puede colocar a los miembros de este grupo de solicitantes dentro de un perfil sociodemográfico, se tiene antecedentes a partir del bureau, se tiene información geográfica de vivienda o de la sucursal donde hace la solicitud y se sabe si tuvo antes algún producto del banco y su historia respecto a estos productos.

A partir de esta información, se construye un modelo con el fin de valorizar y calificar la calidad de riesgo de los solicitante[1].

Es conveniente destacar que los datos corresponden a un segmento de crédito y de individuos muy particular, como se señaló anteriormente. Con poco facilidad de acceso a los datos, que en otras entidades no conseguirían acceso al crédito.

Como resultado de la inestabilidad en el mercado financiero, en Junio del 2004 el Comité de Supervisión Bancaria de Basilea publicó el documento "Convergencia Internacional de medidas y normas de capital: marco revisado", más conocido como Basilea II. Basilea II tiene por objetivo construir una base sólida para la regulación prudente del capital, la supervisión y la disciplina de mercado, así como perfeccionar la gestión del riesgo y la estabilidad financiera.

Teniendo en cuenta las directivas enunciadas por Basilea II, los bancos y supervisores se vieron en la necesidad de evaluar la solidez y precisión de las medidas internas del riesgo de crédito. Por ello, surge la necesidad de contar con metodologías para validar los sistemas internos y externos de puntuación que se le otorga al riesgo.

Basilea II permite que las instituciones financieras desarrollen modelos propios (IRB[2]) con el fin de determinar cuáles son los niveles de riesgo de su cartera. Por lo tanto, los supervisores se ven obligados a validar el proceso con el que los bancos miden el riesgo.

[1] En la práctica este modelo podría ser útil para ajustar las políticas de oferta de producto para este grupo de solicitantes, de todas formas, éste modelo construido en particular para este trabajo, es como una muestra tomada que no fue la real utilizada para construir el modelo definitivo con el que se trabajó realmente con el cliente (por motivos de confidencialidad) y los resultados obtenidos para este trabajo son a modo de ejemplo y son ficticios, es decir, no se utilizan en la práctica.

Siguiendo los lineamientos del Banco Central de Paraguay, se hacen un modelo de scoring que clasifica a los individuos de acuerdo a la probabilidad de mora estimada. El principio básico de los modelos de scoring es asignar a cada cliente una puntuación con el fin de separar morosos de no morosos. Entonces, los modelos de scoring pueden verse como una herramienta de clasificación ya que da información respecto al comportamiento que se espera que el cliente tenga en un futuro. De esta manera, el poder discriminante del modelo de scoring indica la habilidad del modelo para separar morosos de no morosos. Por ello, medir el poder discriminante del modelo es una parte importante del proceso de validación.

Con este fin, a ambos modelos se les hace el test de Kolmogorov-Smirnov y se calcula la curva COR como indica Basilea II.

Existen diferentes metodologías que permiten medir el riesgo crediticio. Entre estas se pueden distinguir las técnicas estadísticas, multivariadas, análisis de modelos de clasificación, árboles de decisión, modelos de elección cualitativa (PROBIT Y LOGIT) y el análisis de matrices de transición.

Hoy en día, la técnica más utilizada por las entidades bancarias y financieras para intentar detectar el riesgo de incumplimiento de una obligación asumida (riesgo de default) es la regresión logística.

Esta técnica presenta muchas ventajas pero también tiene algunas debilidades; es por ello que en el presente trabajo se decide comparar dicha técnica con la de Random Forrest para tratar de determinar cuál de ellas arroja mejores resultados.

El objetivo del estudio son dos:

1) Comparar la Regresión Logística y Random Forrest con respecto a las variables seleccionadas por cada técnica.
2) Comparar la Regresión Logística y Random Forrest con respecto a la precisión de clasificación de cada técnica

Para poder realizar dicha comparación, se utilizará una base de datos de un banco paraguayo y se construirán un modelo para aquellos individuos que ya tienen una vinculación previa con la entidad (clientes del banco) y que solicitan un crédito.

La variable dependiente utilizada en los modelos es: Malo 60 indeleble. Es decir, si el individuo alcanza o no un máximo de 60 días de atraso alguna vez, más allá de que luego se ponga al día (de ahí la marca de indeleble), y se utilizan diversas variables independientes.

[2] "Internal-ratings based approaches".

Revisión Bibliográfica

Ohlson (1980) fue el primero en utilizar la regresión logística para predecir la bancarrota de empresas. Utilizó un data set de los años 1970 a 1976. La base contiene 105 empresas que entraron en bancarrota y 2058 que no lo hicieron. Compara sus resultados con estudios previos que utilizan otras técnicas y encuentra lagunas ventajas de este método.

Wiginton (1980) también está dentro de los primeros en utilizar la regresión logística para determinar el riesgo crediticio. Compara la regresión logística con el análisis discriminante y encuentra que la regresión logística funciona mejor.

Si bien la técnica de Random Forest está comenzando a ser usada para la modelización del scoring de crédito, la literatura más extensa en materia de comparación de ambas técnicas se da en el campo de la medicina , epidemiología, etc. Es por ello que muchos de los papers analizados en esta sección se relacionan más con ese campo.

Lee et al. (2005) se basan en el paper de Dudoit et al.(2002) y comparan 21 métodos de clasificación dentro de las cuáles encontramos a las técnicas que nos interesan analizar en este trabajo. Utilizan siete datasets de microarray de distintos tamaños. Los autores encuentra que random forest tiene una mejor performance que la regresión logística.

Ruldolfer et al. (1998) comparan la regresión logística y los árboles de decisión para el diagnóstico de el síndrome del túnel carpiano. Encuentran que no existen mayores diferencias entre ambas técnicas para el data set analizado.

Delen et al. (2004) usan también regresión logística y árboles de decisión para predecir el la probabilidar de sobrevida del cáncer de mama. Utilizan un data set de 1973 a 2000 que contenía 202932 casos y 16 variable dependientes. Encuentran que los árboles tienen un mejor poder de clasificación.

Lin and Wu et al. (2004) compararon la regresión logística y random forest para predecir interacciones en proteínas en un data set grande. Llegaron a la conclusión de que random forest tiene una mejor performance que la logística.

Ming Geng (2006) compara la regresión logística y random forest. El objetivo es estimar los factores asociados con la etapa del diagnóstico y si la raza es una variable influyente comparando ambas técnicas en cuanto al poder de predicción y selección de variables.

Utilizan un data set con 960 casos y encuentran que ambas técnicas tienen un poder de predicción similar y eligen de manera similar a las variables.

Anne Ruiz-Gazen et al.(2008) utilizan una comparación entre regresión logística y random forest para predecir tormentas con un data set desbalanceado. Dividen el data set en dos, uno entre junio y agosto de 2004 que posee 11803 observaciones y otra entre los mimos meses de 2005 que posee 20998 observaciones. El primer set se utiliza para construir los modelos y el segundo como prueba. Encuentran que el hecho de tener un data set desbalanceados les trae problemas de estimación. Y no obtienen resultados muy distintos entre la regresión logística y

el random forest, es más, debido a que el algoritmo del random forest toma un tiempo considerablemente largo en ejecutarse en comparación a la regresión logística, los autores sugieren que es preferible, dado que no encuentran mejoras con el otro método, la utilización de la regresión logística en el campo de la meteorología.

Materiales y Análisis

Definición de la Muestra

Sólo se analizarán los individuos que estén adquiriendo un préstamo.

Este modelo se utiliza para evaluar individuos que ya eran clientes del banco cuando solicitaron un nuevo préstamo.

El modelo de clientes incluye aquellas solicitudes con préstamo otorgado y que el cliente tenía al menos un préstamo activo o inactivo.

Para la construcción de este modelo se cuenta no sólo con variables del individuo (como edad, sexo, etc.) sino también otras variables relevantes como la vinculación del individuo con la institución que también serán útiles a la hora de construir el modelo.

Se excluirán del universo de análisis inicial los individuos con productos de refinanciación, los individuos sin producto otorgado.

El periodo de análisis es del 2 de enero de 2009 al 30 de julio de 2010. El modelo de Clientes se construyó a partir de los datos de 49.026 solicitudes.

Madurez

Se utilizaron en el análisis las solicitudes con más de 5 meses de antigüedad al momento de la fecha de la extracción, las solicitudes que aunque hayan tenido menos tiempo ya habían incurrido en mora y aquellas que han pagado al menos el 75% de su deuda.

Morosidad

Se consideró que un individuo es moroso si tuvo un atraso mayor a 60 días en el cumplimiento de sus obligaciones. Como se utiliza el criterio de mora indeleble, la marca de "malo" se mantiene aunque el sujeto haya regularizado luego su situación.

Metodología

Modelo Logístico

El objetivo del modelo es adjudicar una probabilidad de mora (P_i) a cada Persona(i) evaluada, a través de un modelo logístico que otorga una calificación de cero a uno a cada persona. Cuánto más cercano a uno es el valor, mayor la probabilidad de que el cliente sea moroso.

El modelo logístico se define de la siguiente manera:

$$Z_i = \alpha + \beta_j X_{ji}$$
$$P_i = \frac{1}{1+e^{-Z_i}} = \frac{1}{1+e^{-(\alpha+\beta_j X_{ji})}}$$

Donde

- X_{ij} es el valor que toman los regresores (variables explicativas) del modelo para el individuo i,
- β_j es el coeficiente asociado a la variable explicativa X_j ,
- α es la constante.

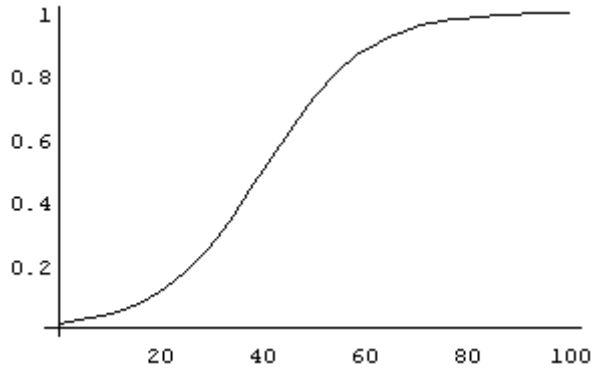
Como la variable dependiente está restringida entre cero y uno, la relación entre los regresores (X) y la variable dependiente no es lineal.

Si se toma el logaritmo natural de la razón de probabilidades se obtiene:

$$L_i = ln\left(\frac{P_i}{1-P_i}\right) = Z_i = \propto_i + \beta X_i$$

Entonces, L_i resulta lineal en X_i y también en los Parámetros. Es decir,β es la pendiente y mide el cambio en L_i ocasionado por un cambio unitario en X_i, es decir, dice cómo el logaritmo de la probabilidad de ser moroso cambia a medida que $X_i$cambia en una unidad.

La forma típica de esta función es:



El eje de las ordenadas representa la probabilidad de pertenecer a uno de los dos grupos de interés: moroso o no moroso en este caso. Si el coeficiente asociado a una variable explicativa es positivo, el valor en el polinomio es mayor y la probabilidad de mora más alta, es decir, hay

un castigo asociado a dicha variable. En cambio, si el coeficiente asociado a una variable explicativa es negativo, la variable implica una disminución en la probabilidad de mora de la solicitud, es decir, hay un beneficio asociado a dicha variable[3].

Random Forrest

El random forest es una técnica que se base en la idea de árboles de clasificación y regresión (CART) desarrollada por Leo Breiman et al. (1985). En CART el nodo raíz contiene todas las observaciones y es dividido en dos nodos hijo eligiendo para su división alguna variable del data set, por ejemplo sexo=femenino, edad >= 65 años. Este método, sin bien sencillo, tiene problemas en la precisión de clasificación del modelo y el modelo puede verse afectado fuertemente por cambios en el data set (Leo Breiman et al. (1985)).

Es por ello que Leo Breiman y Adele Cutler desarrollan el Random Forest, una técnica basada en *ensemble learning* para clasificación y regresión. Extienden el método CART a un conjunto de árboles solucionando de esta forma el problema de precisión en la clasificación. El método combina la idea de bagging de Breiman (Breiman (1994)) y la selección aleatoria de atributos, introducida de forma independiente por Ho (1995) Amit et al. (1997) para construir una colección de árboles de decisión con variación controlada.

La idea de este método es hacer crecer muchos arboles de clasificación. Cada árbol da como resultado una clasificación y decimos que los arboles "votan" por una clase. El bosque elige la clasificación que recibe la mayoría de los votos.

Cada árbol se construye de la siguiente forma:

a) Si el data set contiene N datos, seleccionar N de forma aleatoria con reposición. Éste será el training set. De esta forma se eligen aproximadamente 2/3 de los datos del total del data set para la construcción de cada árbol y el 1/3 restante (oob: out of the bag) se utilizan para obtener errores de clasificación insesgados y para calcular la importancia de las variables.
b) Si el data set contiene M variables, se seleccionarán m<<M. De forma tal que en cada nodo, m variables sean seleccionadas de forma aleatoria y el mejor Split es utilizado para dividir el nodo (Breiman L. (1999)).
c) Cada árbol se construye hasta su mayor extensión posible.

En cada rama, se divide los árboles utilizando una partición binaria, es decir, cada nodo se divide en no mas que dos nodos hijo.

Se hace crecer cada árbol al menos parcialmente at random.

[3] En el anexo i encontramos los distintos tipo de variables que se utilizarán con este modelo y un ejemplo de aplicación del peso en el caso de las variable piecewise y su construcción.

Se introduce la aleatoriedad haciendo crecer cada árbol sobre una sub-muestra aleatoria distinta.

La aleatoriedad se inyecta en el proceso de selección de división de modo que el divisor en cualquier nodo está determinado en parte al azar

El error de clasificación del bosque depende básicamente de:

a) La correlación entre dos árboles cualesquiera del bosque. A mayor correlación, mayor error.
b) El peso de cada árbol en el bosque. Un árbol con una baja tasa de error es un clasificador fuerte. El aumento de la fuerza de los árboles individuales disminuye la tasa de error de bosque.
c) El parámetro a controlar es m. éste controla tanto la correlación entre los árboles como la fuerza de clasificación. Al aumentar m, aumentan ambos valores, y al reducirlo, se reducen ambos. El objetivo es encontrar el valor óptimo de m. Se puede utilizar el oob error rate para determinar el valor de m.

Características de Random Forest

- No es superable en precisión, de entre los algoritmos actuales.
- Funciona de manera eficiente en grandes bases de datos.
- Puede manejar grandes cantidades de variables
- Provee estimaciones de qué variables son importantes en la clasificación
- Tiene un método eficaz para la estimación de los missing values[4] y logra mantener la precisión en la estimación.
- Los bosques generados se pueden almacenar para ser utilizados con otros datos
- No tiene problemas de overfit

Cómo funciona el algoritmo

Cuando se selecciona aleatoriamente el training set, se dejan fuera 1/3 de los casos (OOB) que son utilizados para obtener estimadores insesgados del errores de clasificación, además, se utiliza también para obtener estimaciones de la importancia de las variables.

Una vez que se construye un árbol, se utilizan los datos para computar las proximities[5]. Estas son utilizadas para reemplazad missing values y en la detección de outliers.

[4] Ver sitio de Breiman para una mejor referencia:
http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#missing1

[5] Ver sitio de Breiman para mayor información:
http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#prox

OOB: Estimación de error

En Random Forest, no hay necesidad de cross-validation debido a que cada árbol se construye utilizando una muestra boostrap de los datos originales.

Aproximadamente 1/3 de los casos se dejan fuera de este muestreo y no son utilizados en la construcción de cada árbol. Los datos que son dejados fuera de la construcción de un árbol específico, son utilizados para obtener una clasificación y por lo tanto una validación de la fuerza de clasificación consistencia en la construcción.

Como ya se dijo, las observaciones que se dejan fuera son utilizadas para obtener errores insesgados. El proceso se realiza de la siguiente forma:

Una vez construido el árbol k, se utilizan las observaciones que se dejaron fuera de las construcción del árbol k para ver si la clase que se predijo cuando se construyó el árbol k es la misma que se predice con estas observaciones. Se hace esto con todos los árboles construidos. Los resultados se promedian y se obtiene el error.

Importancia de la Variables

Para calcular la importancia de las variables, se utiliza la muestra oob.

La idea básica es calcular la precisión del modelo antes y después de "romper" el vínculo que tiene la variable con la clase y se calcula el promedio de todos los árboles del bosque. Este valor nos indica la importancia de esa variable y se calcula, básicamente de la siguiente forma.

Primero se toman estos datos y para cada árbol construido y se cuentan los votos para la clases correctamente predicha.

Luego se permutan aleatoriamente los valores de m en la muestra oob y se vuelven a contar los votos.

Luego se restan la cantidad de votos de ambas corridas. Este proceso se realiza para cada árbol y se obtiene un promedio. Este valor es la importancia de la variable m (4).

Además, se puede calcular otra medida de importancia (*gini importance*, que es similar a esta medida y en general arroja resultado también similares[6]).

Comparación entre Regresión Logística y Ramdom Forrest

[6] En la sección de: Resultados Obtenidos con Random Forest se pueden ver estas similitudes.

Cuadro 1: Comparación entre Random Forest y Regresión Logística

Característica	Regresíon Logística	Random Forrest
Técnica Paramétrica	Si	No
Permite calcular probabilidades (como odss ratio)	Si	No
Testea Interacción	Si	Si
No requiere un supuesto especial del conjunto de datos	No	Si
Robusto a los valores atípicos	No	Si
No requiere transformación de variables continuas	No	Si
No se ve afectado por el over-fitting	No	Si
Selecciona las variables relevantes de forma automática	No	Si
Maneja automáticamente los missing values	No	Si
Puede trabajar eficientemente con datasets desbalanceados	No	Si

La función logística tiene tanto ventajas como desventajas.

Como ya se vio, una de las razones por las cuáles es popularmente utilizada en estos análisis es porque su rango es entre cero y uno y por lo tanto puede ser utilizada para modelar la probabilidad, en este caso de que un individuo entre en default.

Además, la regresión logística permite las interacciones y confundings (Hosmer, D.W., et al.(2000)).

Debido a que esta técnica permite calcular el odds ratio, nos permite calcular la probabilidad de que un individuo entre en default de manera sencilla.

A pesar de sus ventajas, la este método presenta también algunos inconvenientes.

Es necesario hacer ciertos supuestos que en la práctica pueden no cumplirse (Hosmer, D.W., et al.(2000))(Collett, D. (2003)): la relación entre la media de la variable dependiente y las variables independientes sigue una distribución logística y los errores siguen una distribución binomial.

Los outliers pueden tener una influencia importante en los resultados del modelo (Collett, D. (2003)).

Los problemas de colinealidad también son frecuentes en este tipo de modelos (Hosmer, D.W., et al.(2000)).

Existe también el problema de over fitting en data sets con gran cantidad de variables.

Debido a que la selección de variables y las interacciones que pueden tener entre ellas es un problema frecuente, muchos paquetes permiten la opción stepwise para la selección de variables. Sin embargo, es pertinente señalar que este método ha sido bastante criticado (Harrell, F.E., et al (1996)) (Harrell, F.E., et al. (1984)).

El random forest por otro lado, es una técnica que se base en la idea de árboles de clasificación y regresión (CART) desarrollada por Leo Breiman et al. (1985). En CART el nodo raíz contiene todas las observaciones y es dividido en dos nodos hijo eligiendo para su división alguna variable del data set, por ejemplo sexo=femenino, edad >= 65 años. Este método, sin bien sencillo, tiene problemas en la precisión de clasificación del modelo y el modelo puede verse afectado fuertemente por cambios en el data set (Leo Breiman et al. (1985)).

Es por ello que Leo Breiman desarrolla el Random Forest, una extensión del método CART a un conjunto de árboles. Y soluciona el problema de la precisión de clasificación.

La técnica posee varias ventajas en relación a la regresión logística. Es capaz de seleccionar de un gran conjunto de variables aquellas que sean más relevantes (14), y esta forma forma de seleccionar las variables es distinta a la forma en que lo hace el stepwise lo cual la hace mas deseable debido a que el stepwise presenta varios problemas (Ming Geng (2006)). No tiene problemas de over fitting y es capaz de trabajar muy bien con valores missing .

Uno de los problemas, o desventajas de esta técnica, es que la estructura del árbol es una suerte de caja negra. No podemos, como en la regresión logística, ver la relación que hay entre algún nivel de alguna variable y la variable dependiente.

Poder Discriminante

Curva COR

Para medir la capacidad discriminante del modelo, es decir, la capacidad de separar morosos y no morosos, se utiliza la Curva COR que mide la eficiencia del modelo para agrupar los casos deseados frente a los no deseados.

El eje denominado susceptibilidad acumula el porcentaje de los casos morosos, para este análisis, rechazados por el modelo (aciertos). El eje de las abscisas indica el porcentaje de buenos pagadores rechazados (erróneamente) por el modelo. La diagonal de 45 grados indica el resultado que surgiría de un modelo de aprobación aleatoria simple, mostrando que, por ejemplo, para rechazar al 40% de los casos morosos, también habría que rechazar al 40% de los buenos pagadores.

Cuanto mayor resulte la capacidad predictiva de un modelo, menor será la proporción de buenos pagadores rechazados, para cada nivel de morosos rechazados. De aquí que cuanto mayor distancia entre la curva COR y la recta de 45 grados para un modelo dado, mayor capacidad predictiva habrá implícita. Es decir, cuanto más se acerque la curva al extremo superior izquierdo de la cuadricula mayor será el poder predictivo del modelo.

El Área Bajo la Curva COR es la probabilidad de clasificar correctamente a los morosos.

Kolmogorov-Smirnov (K-S)

El test de Kolmogorov-Smirnov es un test no paramétrico que compara si la distribución de una variable es la misma en dos muestras independientes, en este caso, la muestra de solicitudes morosas y la muestra de solicitudes sin mora.

El test presenta la diferencia Extrema Absoluta, Positiva y Negativa observada entre la frecuencia acumulada de una muestra y la frecuencia acumulada de la otra muestra. Cuanto más grande sea esta diferencia mejor discriminará el modelo.

Resultados

Resultados Obtenidos con Regresión Logística[7]

Modelo Logístico

Las siguientes tablas muestran las variables que se incluye en el modelo. La última tercer columna muestra el coeficiente estimado (las betas del modelo).

[7] Para una mejor referencia a la descripción de las variables del modelo logístico, ver anexo ii.

Cuadro 2: Resultados del Modelo Lógistico

Variables Categóricas				
Variable	**Categoría**	**Coeficiente**	**Significatividad**	**Importance**
Categoría Segmento y Antecedentes en informconf	Ant+-Segm Muy Alto	0,0000	0,0000	4,91%
	Ant+-A*-CompDir+-	0,758	0,052	
	Ant+-B*-Coop/CCom/Tel/Bancbajo	0,811	0,037	
	Ant+-C*-AspiracMezcla	1,006	0,010	
	Ant+-D*-Coop peq/sin operfin-ComDir--	1,204	0,002	
	E*-Inf pero Sin Antec	1,068	0,007	
	F1aF4-Ant Malos sin atrasos	1,515	0,000	
	F5-Ant malos con atrasos	1,902	0,000	
	Sin Informe	1,352	0,001	
Vinculación y Experiencia	Muy Baja	0,000	0,000	34,67%
	Baja	-0,328	0,000	
	Media	-0,424	0,000	
	Alta	-0,911	0,000	
	Atraso Leve	0,456	0,000	
	Malo	0,972	0,000	
Cantidad de Dependientes	1	0,000	0,000	1,73%
	2	0,058	0,168	
	>2	0,215	0,000	
Región	Asuncion	0,000	0,000	1,87%
	Ciudad del Este	-0,111	0,188	
	Norte, Este y Centro	-0,268	0,000	
	Chaco y Sur	-0,107	0,198	
Trabajo	Público	0,000	0,000	4,09%
	Jubilado	-0,001	0,994	
	Privado con IPS	-0,186	0,000	
	Privado sin IPS	-0,120	0,015	
	Independiente	0,546	0,000	
	Rent/ Estud/ AmaCasa	0,289	0,198	
	Sin Informar	1,519	0,000	
Tiene Teléfono	Sin Teléfono Fijo	0,0000	0,0000	5,62%
	Con Teléfono Fijo	-0,224	0,000	

Variables Piecewise				
Variable	**Categoría**	**Coeficiente**	**Significatividad**	**Importance**
Edad en años	Edad > 17 and Edad <= 29	-0,084	0,000	17,48%
	Edad > 29 and Edad <= 37	-0,031	0,001	2,37%
	Edad > 37 and Edad <= 45	-0,022	0,045	0,88%
	Edad > 45 and Edad <= 60	-0,038	0,000	4,03%
Antiguedad en el Trabajo (en meses)	>2 y <=3	-0,103	0,067	0,73%
	>3 y <+5	-0,134	0,000	3,03%
	>5 y <=10	-0,054	0,001	2,40%
	>10 y <=20	-0,019	0,042	0,90%
Plazo en meses	Plazo > 6 and Plazo <= 9	0,113	0,052	0,82%
	Plazo > 9 and Plazo <= 12	0,190	0,000	4,15%
	Plazo > 12 and Plazo <= 13	0,163	0,013	1,35%
	Plazo > 13 and Plazo <= 15	0,133	0,000	8,95%

VALOR DE LA CONSTANTE	-2,861	0,000	

Se puede observar que todas las variables presentan el signo esperado.

Todas las variable son significativas, tanto las continúas como las discretas (en conjunto). Algunas categorías de las variables discretas aparecen como no significativas (por ejemplo la categoría jubilado en la variables trabajo). Esto puede deberse a la baja cantidad de casos en algunas categorías. De todas formas, los signos son los esperados, y las variable es significativa en conjunto.

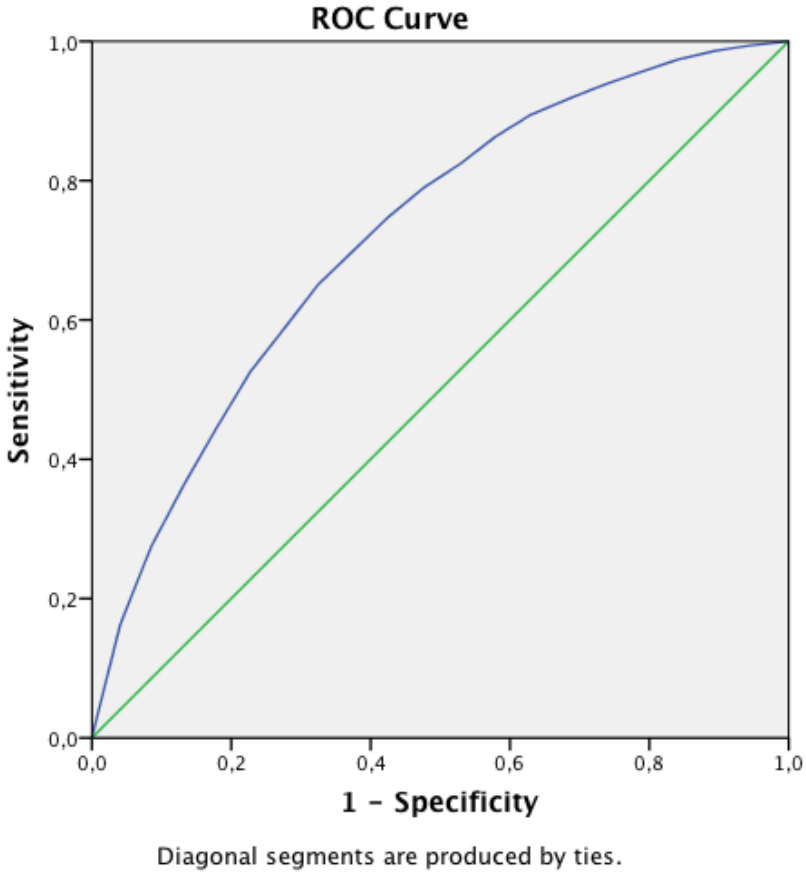
En la última columna se calcula la importancia de las variables a través del test de wald. Si bien a través del fstep es posible seleccionar de las variables que se incluyen en la regresión, aquellas relevantes, con esta medida de importancia podemos ordenarlas.

A la hora de correr la regresión, se incluyeron 27 variables, y a través de ete método se seleccionaron 18 variables como relevantes.

Poder Discriminante: Curva COR y K-S

El área debajo de la curva COR correspondiente a este modelo es igual a 0.722. Es decir, se espera clasificar correctamente a los morosos con una probabilidad de 72,2%.

Gráfico 1: Curva COR



Por otro lado, el K-S en este modelo indica que la mayor diferencia entre la frecuencia acumulada de la muestra de morosos y la frecuencia acumulada de la muestra de no morosos es 0,326.

Indicadores de desempeño

A continuación se presenta un cuadro con los indicadores de desempeño.

Se toma el resultado de la estimación del modelo. En este caso la probabilidad de mora estimada por el modelo, se la ordena y se la divide, en este caso en 20 grupos, para ver cómo se distribuyen morosos y no morosos.

Cuadro 3: Indicadores de Desempeño

Nivel de Riesgo	Malos			Buenos		Diferencia
	%	%F	% malos hasta este punto	%	%F	%F.buenos-%F.Malos
1	,0086	*0,6%*	*0,9%*	,9914	*5,4%*	4,81%
2	,0126	*1,4%*	*1,1%*	,9874	*10,7%*	9,32%
3	,0192	*2,6%*	*1,3%*	,9808	*16,0%*	13,38%
4	,0273	*4,4%*	*1,7%*	,9727	*21,3%*	16,86%
5	,0277	*6,2%*	*1,9%*	,9723	*26,6%*	20,31%
6	,0318	*8,3%*	*2,1%*	,9682	*31,8%*	23,47%
7	,0339	*10,5%*	*2,3%*	,9661	*37,0%*	26,48%
8	,0490	*13,7%*	*2,6%*	,9510	*42,2%*	28,43%
9	,0600	*17,7%*	*3,0%*	,9400	*47,3%*	29,60%
10	,0514	*21,0%*	*3,2%*	,9486	*52,4%*	31,37%
11	,0641	*25,2%*	*3,5%*	,9359	*57,5%*	32,24%
12	,0738	*30,1%*	*3,8%*	,9262	*62,5%*	32,43%
13	,0738	*34,9%*	*4,1%*	,9262	*67,49%*	32,61%
14	,0967	*41,2%*	*4,5%*	,9033	*72,4%*	31,18%
15	,0955	*47,5%*	*4,8%*	,9045	*77,3%*	29,83%
16	,1212	*55,4%*	*5,3%*	,8788	*82,0%*	26,66%
17	,1257	*63,6%*	*5,7%*	,8743	*86,8%*	23,17%
18	,1354	*72,5%*	*6,2%*	,8646	*91,5%*	19,00%
19	,1722	*83,7%*	*6,7%*	,8278	*95,9%*	12,21%
20	,2489	*100,0%*	*7,6%*	,7511	*100,0%*	0,00%
% malos general	7,64%				*K-S*	*32,61%*
Diferencia 5% mejor-peor	24,03%				*COR*	*72,17%*
					Gini	*44,3%*

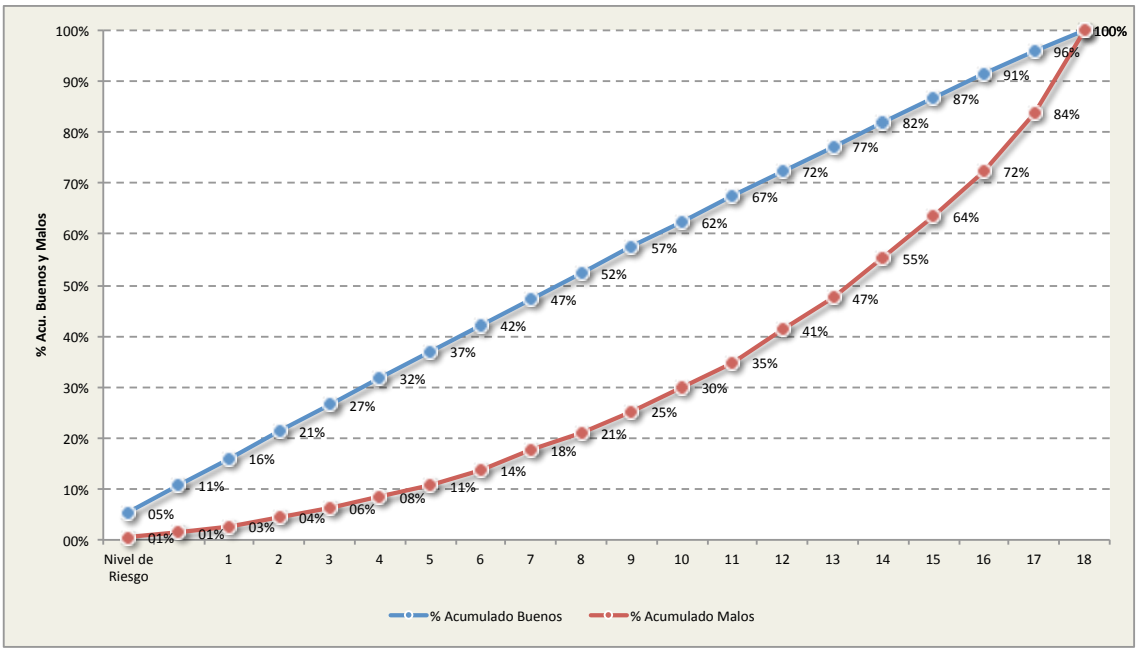
Se puede observar que el modelo ordena bien tanto a los morosos como a los no morosos. A pesar de ser un modelo con pocos malos. Es decir, los niveles de riesgo más bajos tiene menos porcentaje de morosos y más porcentaje de no morosos y viceversa.

En el cuadro anterior, también se observar el valor del K-S (32,61%). Como se explicó anteriormente, este valor indica la máxima separación en este caso, entre morosos y no morosos. En el caso de este modelo, esa separación máxima se logra en el nivel de riesgo definido como número 13. Estos niveles de riesgo suelen utilizarse a la hora de tomar decisiones sobre el punto de corte para el otorgamiento ya que dependiendo de cual se elija,

se dejarán fuera cierto porcentaje de potenciales morosos a costa de rechazar también un porcentaje de potenciales no morosos. El punto de corto no necesariamente coincide con el de mayor separación entre morosos y no morosos. Muchas veces esto tiene que ver con decisiones comerciales de cuánto mora se puede tolerar y cuántos buenos clientes potencialmente buenos se está dispuesto a perder.

En el gráfico Siguiente se puede ver el acumulado de morosos y no morosos graficado.

Gráfico 2: % Acumulado Morosos y No morosos por Nivel de Riesgo



Resultados Obtenidos con Random Forest[8]

A continuación se presentan los resultados obtenidos a través del método Random Forest.

Se realizan dos corridas. Una primera corrida con una gran cantidad de variables, incluso con variables redundantes, con le propósito de descartar y seleccionar las variables más relevantes y una segunda corrida seleccionado variables en base a los resultados de la primera corrida.

Por ejemplo, se utilizaron en esta corrida varias variables que indican la zona. Como variable pura o como una combinación con alguna otra variable. Se elegirá para la segunda corrida del random forest, a la variable de zona que posea una mayor importancia.

[8] Se utilizó el programa gratuito R (disponible en: http://www.r-project.org/) para correr el random forest, debido a que sus autores programaron originalmente la sintaxis en este lenguaje. Hay muchas versiones distintas de random forest y distintos paquetes. En este trabajo se utilizará el comando original programado por los autores: RandomForest. Se puede obtener el manual en: http://cran.r-project.org/web/packages/randomForest/randomForest.pdf

Resultados Obtenidos Con Todas las Variables[9]

Se realizó una primera corrida que incluía un total de 90 variables. El data set incluía tanto variables continuas como variables discretas. Muchas de ellas eran variables originales del data set y que no habían sufrido modificaciones o que habían sufrido modificaciones leves. Muchas de las variables contenían missing values.

Errores OOB Modelo con Todas as Variables

A medida que se van calculando los árboles, se puede ir observando el OOB error y los errores en cada clase. En este caso, se pueden observar los errores cada 100 árboles:

Cuadro 4: errores OOB

ntree	OOB	1	2
100:	38,18%	38,91%	29,33%
200:	37,92%	38,71%	28,42%
300:	37,67%	38,49%	27,70%
400:	37,37%	38,27%	26,42%
500:	37,32%	38,25%	26,07%
600:	37,25%	38,19%	25,86%
700:	37,14%	38,08%	25,73%
800:	37,25%	38,21%	25,73%
900:	37,20%	38,14%	25,81%
1000:	37,19%	38,10%	26,18%

Como se puede observar en la tabla anterior, se construyen en total, 1000 árboles. Los errores pueden fluctuar a medida que se realizan las corridas, como es el caso observado, no siempre disminuyen.

Se presenta además la matriz de confusión que resume los resultados de la tabla anterior:

Cuadro 5: Matriz de Confusión

	OOB estimate of error rate: 37.19%			
	Buenos	Malos		
		0	1	class.error
Buenos	0	28027	17252	0,3810155
Malos	1	981	2766	0,2618094

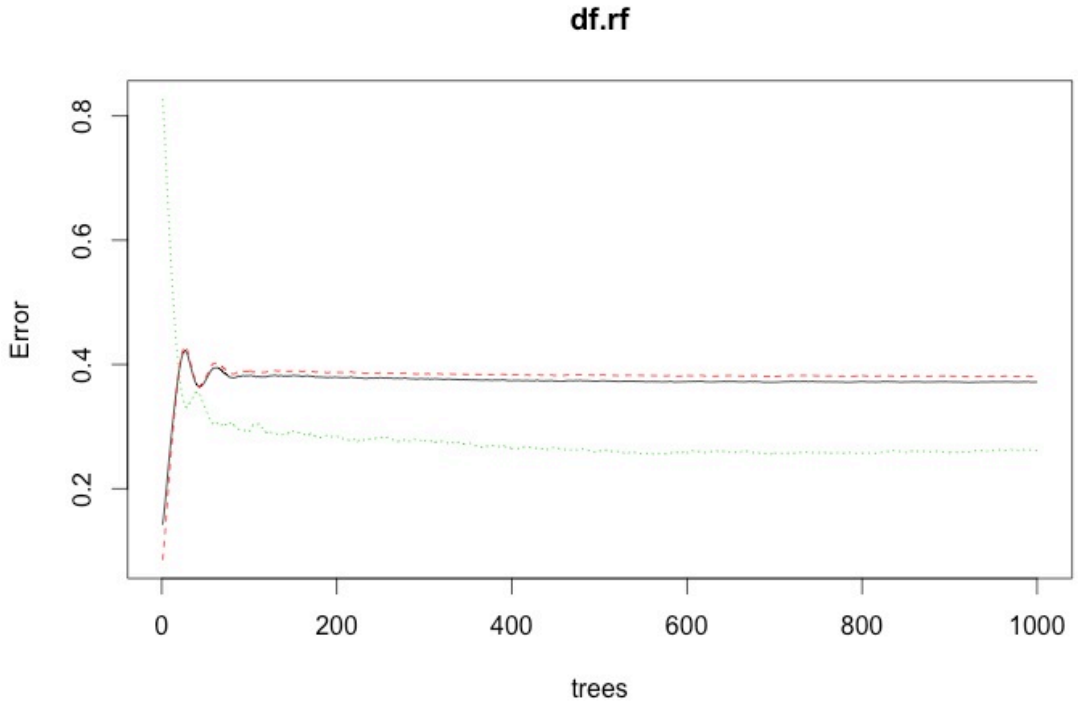
En ambas tablas se pueden observas los errores de clasificación tanto globales del modelo como de cada clase. Se observa que el error de clasificación global que se comete con el

[9] La descripción de las variables utilizadas en los modelos de random forest no se incluyen en el trabajo como se incluyó la descripción de las variables del modelo logístico debido a que la mayoría de las variables que aparecen en estos modelos puede inferirse su contenido por el nombre o bien debido a que muchas de ellas son similares a las que aparecen en el modelo de regresión logística.

modelo es de un 37,19%. Se clasifican erróneamente 38,10% de los buenos y un 26,18% de los malos. Que la clase más desbalanceada no esté peor clasificada tiene que ver con que se utilizan classweights en la corrida. Siguiendo las sugerencias de Breiman, se setean estos pesos de manera inversamente proporcional al peso de la clase.

El gráfico siguiente muestra la misma información grafica para árbol.

Gráfico 3: Error OOB



Se puede observar como el error global y el error de clasificación de la clase de los buenos tienden a acercarse. Esto no es algo que ocurra necesariamente.

Importancia de las variables

A continuación se muestran las dos medidas de importancia calculadas que nos permitirán seleccionar variables relevantes:

Cuadro 6: Importancia de las Variables

	MeanDecreaseAccuracy	MeanDecreaseGini
o_cod_sucursal	90,018516	139,3511957
pfl_montoing	69,946136	221,07971
NivelVinculacion	69,540369	98,56622534
pfl_CiudadesImportantes	65,835182	100,1439389
CodCiudadSUCURSAL	65,826119	99,68595933
Informconf_cat_consultas	60,169307	100,862882
IngE_CuotaIngIntegrado	58,663415	219,9791456
s_dpto_sucursal	58,318556	65,47677274
sb_categoria_consultas	56,2809	93,73938091
i_crc_deuda_sistema_financiero	55,604603	172,6064064
Plazo_Mensual_r	55,297002	123,1275894
TipoVincExp	52,160057	67,15205016
pfl_TipoCiuLoc	50,765207	65,72591821
Cod_Departamento	50,004344	61,96064172
pfl_monto	47,910946	75,4273844
Ingreso_Relativo_t	47,721932	78,02139168
TipoVincExp_Agrup	44,433237	53,82357935
s_cod_ocupacion	44,04725	33,24038245
s_cant_dependientes	43,974547	65,98327581
s_estado_civil	43,749918	53,26563372
s_cant_dependientes_t	43,217404	53,37925531
s_Edad	42,373398	180,8681899
pfl_ingresomensual	42,020042	72,86844826
TipoVinculacion	36,904353	40,82167306
seg_2_CategoriaSegmento	36,73616	76,3557019
pfl_ingreso	36,452257	57,27198608
pfl_trabajoIPS	35,820164	50,71881505
TuvoAlgunaconsulta	35,65206	85,66739792
EstadoCivil_r	34,512315	46,52331718
seg_Antecedentes	29,88535	26,96189035
s_tel_correo	29,256034	24,33968848
s_cod_propiedad	28,742071	30,30690222
pfl_zona	28,630306	32,94514114
s_zona	28,520694	30,61917465
pfl_antiglab	28,27705	39,53349184
pfl_TipoTelefono	27,55975	36,81235924
s_tel_celular	27,420466	12,9786137
pfl_ocupacion_t	27,406973	27,65885865
pfl_vivienda	27,178164	25,41388933
pfl_ocupacion	27,051114	28,12432709
pfl_EsPublicoConIPS	27,023027	16,1067961
i_estado_civil	26,873383	23,88973142
pfl_tipociudad	26,648457	40,91016419
i_tiene_tele_part	26,466015	21,29270334
pfl_edad	25,998424	40,801191
i_tiene_lugar_trabajo	25,800613	23,48905661
SegAntecedentesYAtrasos	25,052597	22,37909157
i_tiene_tele_lab	24,394436	22,99632541
s_zona_sucursal	23,903162	40,92754784
pfl_TieneIPS	22,932261	20,54315861
s_tel_particular	22,554037	16,81777181

	MeanDecreaseAccuracy	MeanDecreaseGini
Tiene_ips	21,590243	20,33311449
o_cod_tipo_canal	21,387305	54,60727423
pfl_region_t	20,697587	21,71274191
Cod_RegionDepartamento	20,521911	29,8196048
pfl_EsPublicoSinIPS	20,128009	19,84585462
o_cod_tipo_canal_t	19,354075	54,13505859
s_cod_actividad_eco	19,34555	45,57599996
pfl_tiene_telefono	18,586553	16,69990515
i_TieneAntecedentesNegativos	18,417391	11,73012541
TieneAtrasosInformConf	18,007635	18,79995733
s_cant_trab_activos	17,953135	35,65210383
pfl_region2	17,297756	27,9210666
pfl_trabajo	16,74641	10,62992806
pfl_EsPrivadoSinIPS	16,57498	16,38903325
pfl_grupocanal	16,541799	48,63055369
pfl_region1	16,200794	23,9942943
pfl_tiposolicitante	16,044584	36,89681815
pfl_EsPrivadoConIPS	15,62885	15,55183854
TipoVinculacionAgrup	15,295941	11,94290332
Sexo_r	15,271718	32,17137686
s_tel_auxiliar	15,234118	31,73032162
s_cod_ciudad_t	14,75386	7,46144044
pfl_region	14,223036	9,15918506
ConPrestamoAnteriorVigente	13,898935	12,03156199
ve_cant_prest_vigentes	12,382554	21,59495047
o_cod_comercio	11,302327	10,58577381
pfl_EsIndependiente	10,271809	3,20504084
s_separacion_bienes	9,56044	6,60507247
sb_TieneConsultaBureau	9,11251	3,02476458
pfl_EsJubilado	8,92941	2,99836709
i_cant_atrasos	8,91258	5,44216118
seg_Cat2	7,765875	4,39808493
TieneSegmentoInformConf	6,472258	4,89588018
ConPrestamoAnteriorCancelado	5,946468	5,99435593
s_cod_nacionalidad	4,256752	7,26557255
s_ind_no_conf	3,879969	3,93579975
i_TieneInformconf	1,990433	1,52516901
s_ind_no_ubicable	1,068129	0,07778283
ConPrestamoAnterior	-5,224553	1,42715121

Ambas medidas tienen una interpretación similar. A medida que aumenta el valor calculado de dicha medida, aumenta la importancia relativa de la variable en el conjunto.

Se puede observar que en general, las medidas son bastante consistentes. Suelen clasificar como menos y más importante de manera similar a las variables.

Para una mejor lectura, a continuación se presentará la información anterior de forma gráfica.

Gráfico 4: Mean Decrease Accuracy

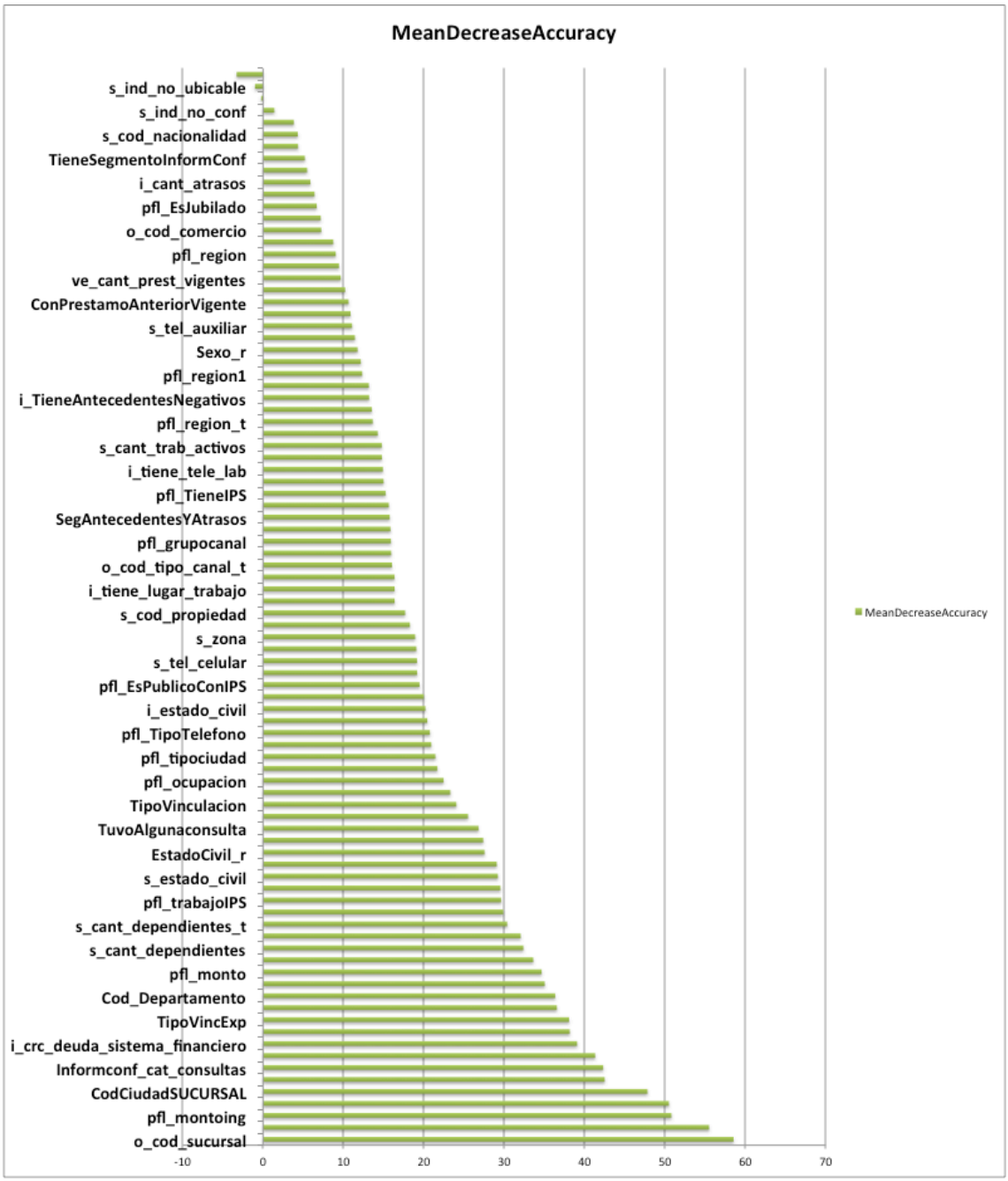
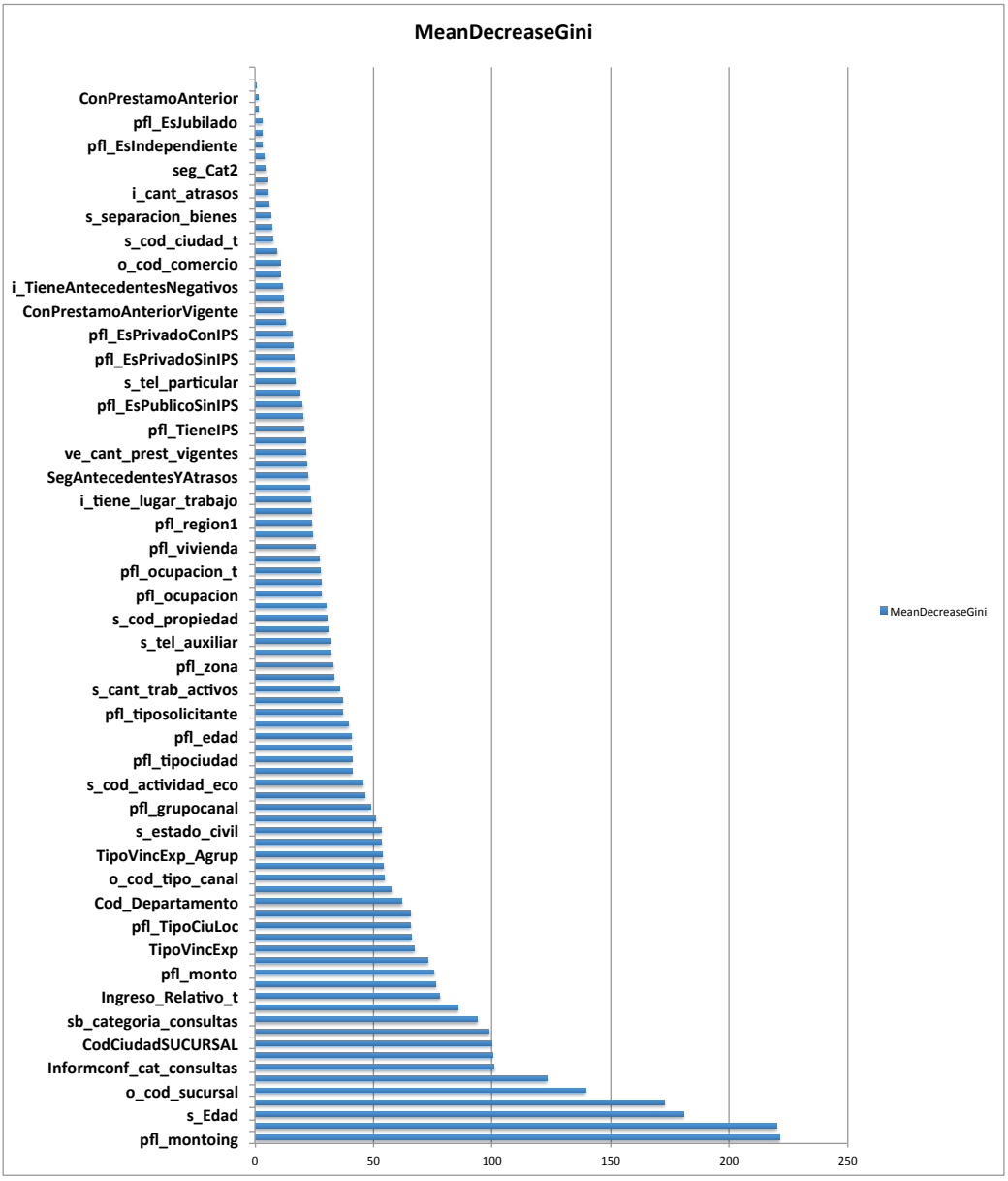


Gráfico 5: Mean Decrease Gini



Si bien las medidas no clasifican a las variables en el mismo exacto orden, lo hacen en un sentido parecido, es decir, ambas medidas encuentran como muy relevantes variables como montoingreso así como también la sucursal, las consultas en el sistema financiero y la deuda en el sistema financiero, entre otras.

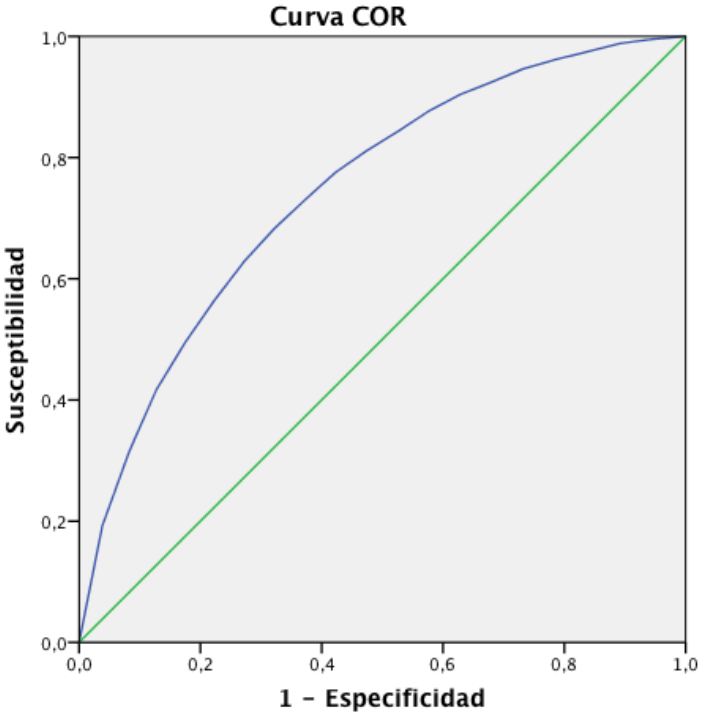
A pesar de que esta corrida pareciera arrojar resultados suficientemente interesantes, nos sirve para seleccionar en una primera instancia las variables, debido a que como puede verse, en el conjunto anterior de variables, hay muchas que son redundantes, y por lo tanto están muy correlacionadas. Por ejemplo, existen diversas variables que informan la ciudad, la ciudad, la región. Estos datos muchas veces provienen de distintas fuentes y muchas veces se realizan

distintos tratamientos con ellos (se los trama de distinta forma según la conveniencia). Es por ello que en una primera instancia, se decidió correr el random forest con todas las variables a considerar con el objetivo de permitirle al algoritmo seleccionar de aquellas variables redundantes, las más relevantes. Luego de este proceso, se realiza una selección de variables, por ejemplo. En el caso de la zona, se decide poner la variable de zona que haya salido como más importante, medido a través de estos indicadores, para la corrida. Se realiza el mismo tratamiento con todas las variables que se encuentren redundantes en la base, y se vuelve a correr.

Poder Discriminante: Curva COR y K-S

A continuación se presentan los resultados de la curva COR para este modelo.

Gráfico 6: Curva COR



Los segmentos diagonales son producidos por los empates.

Área bajo la curva

Variables resultado de contraste: RF -Total-tramos 5% de probmora RF

Área
,745

El área debajo de la curva COR correspondiente a este modelo es igual a 0.745. Es decir, se espera clasificar correctamente a los morosos con una probabilidad de 74,5%

A continuación se presenta el resultado del K-S para este modelo

Cuadro 7: K-S

Estadísticos de contraste[a]

		RF -Total- tramos 5% de probmora RF
Diferencias más extremas	Absoluta	,361
	Positiva	,361
	Negativa	,000
Z de Kolmogorov-Smirnov		21,233
Sig. asintót. (bilateral)		,000

a. Variable de agrupación: Malo60iConProd

Por otro lado, el K-S en este modelo indica que la mayor diferencia entre la frecuencia acumulada de la muestra de morosos y la frecuencia acumulada de la muestra de no morosos es 0,361.

Indicadores de Desempeño

A continuación se presenta un cuadro con los indicadores de desempeño para este modelo.

Cuadro 8: Indicadores de Desempeño

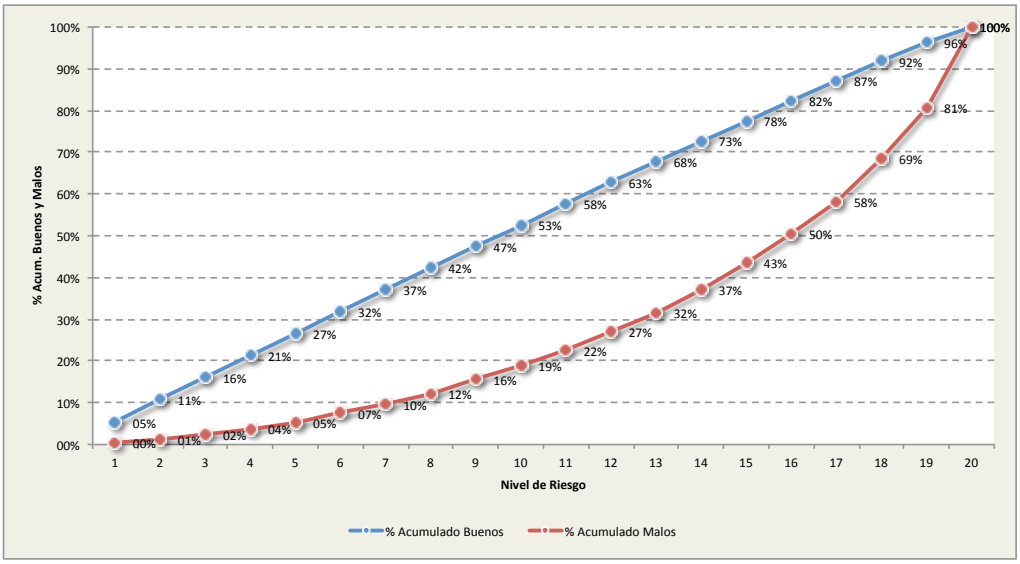
Nivel de Riesgo	Malos			Buenos		Diferencia
	%	%F	% malos hasta este punto	%	%F	%F.buenos-%F.Malos
1	,0070	***0,5%***	***0,7%***	,9930	***5,4%***	4,91%
2	,0105	***1,1%***	***0,9%***	,9895	***10,8%***	9,61%
3	,0206	***2,5%***	***1,3%***	,9794	***16,0%***	13,54%
4	,0201	***3,8%***	***1,5%***	,9799	***21,3%***	17,51%
5	,0230	***5,3%***	***1,6%***	,9770	***26,7%***	21,35%
6	,0336	***7,5%***	***1,9%***	,9664	***31,9%***	24,36%
7	,0314	***9,6%***	***2,1%***	,9686	***37,1%***	27,55%
8	,0419	***12,3%***	***2,4%***	,9581	***42,3%***	30,00%
9	,0514	***15,7%***	***2,7%***	,9486	***47,4%***	31,76%
10	,0489	***18,9%***	***2,9%***	,9511	***52,6%***	33,71%
11	,0543	***22,4%***	***3,1%***	,9457	***57,7%***	35,28%
12	,0694	***27,0%***	***3,4%***	,9306	***62,7%***	35,78%
13	,0720	***31,7%***	***3,7%***	,9280	***67,7%***	36,09%
14	,0826	***37,1%***	***4,0%***	,9174	***72,7%***	35,66%
15	,0978	***43,5%***	***4,4%***	,9022	***77,6%***	34,15%
16	,1064	***50,4%***	***4,8%***	,8936	***82,5%***	32,02%
17	,1194	***58,2%***	***5,2%***	,8806	***87,2%***	28,98%
18	,1583	***68,6%***	***5,8%***	,8417	***91,8%***	23,18%
19	,1864	***80,8%***	***6,5%***	,8136	***96,2%***	15,39%
20	,2938	***100,0%***	***7,6%***	,7062	***100,0%***	0,00%
% malos	7,64%				***K-S***	***36,09%***
Diferencia 5%	28,68%				***Power***	***74,54%***
					Gini	***49,1%***

Se puede observar que este modelo también ordena bien. Loa niveles de riesgo más bajos acumulan la menor cantidad de malos y la mayor cantidad de buenos.

Hay una mejor separación entre el grupo con el 5% mejor de los casos y el grupo con el 5% peor de los casos. La mayor separación entre el porcentaje acumulado de buenos y malos se da en el nivel de riesgo 13 (K-S). Estos porcentajes se grafican a continuación.

Respecto a la acumulación de malos por deciles, la regresión logística parecería acumular en los primeros mayor cantidad de malos que el random forest. De todas formas las diferencias no son sustancialmente grandes.

Gráfico 7: % Acumulado Morosos y No morosos por Nivel de Riesgo



Resultados Obtenidos Con Variables Seleccionadas

Se realizó una segunda corrida con un total de 39 variables seleccionadas a partir de los resultados de la primer corrida del random forest.

Errores OOB Variables Seleccionadas

A continuación se presentan los resultados de correr un segundo modelo random forest con un set reducido de variables:

Cuadro 9: error OOB Variables Seleccionadas

ntree	OOB	1	2
100:	38,70%	39,40%	30,32%
200:	38,32%	39,10%	28,90%
300:	37,86%	38,65%	28,32%
400:	37,80%	38,62%	27,94%
500:	37,84%	38,66%	27,81%
600:	37,82%	38,64%	28,00%
700:	37,72%	38,56%	27,49%
800:	37,65%	38,49%	27,52%
900:	37,59%	38,43%	27,44%
1000:	37,52%	38,39%	27,01%

Se puede observar que los errores de clasificación en este modelo empeoran aunque no sustancialmente; de todas formas, la calidad del modelo mejora debido a se eliminan la mayoría de las redundancias y con ello las correlaciones[10].

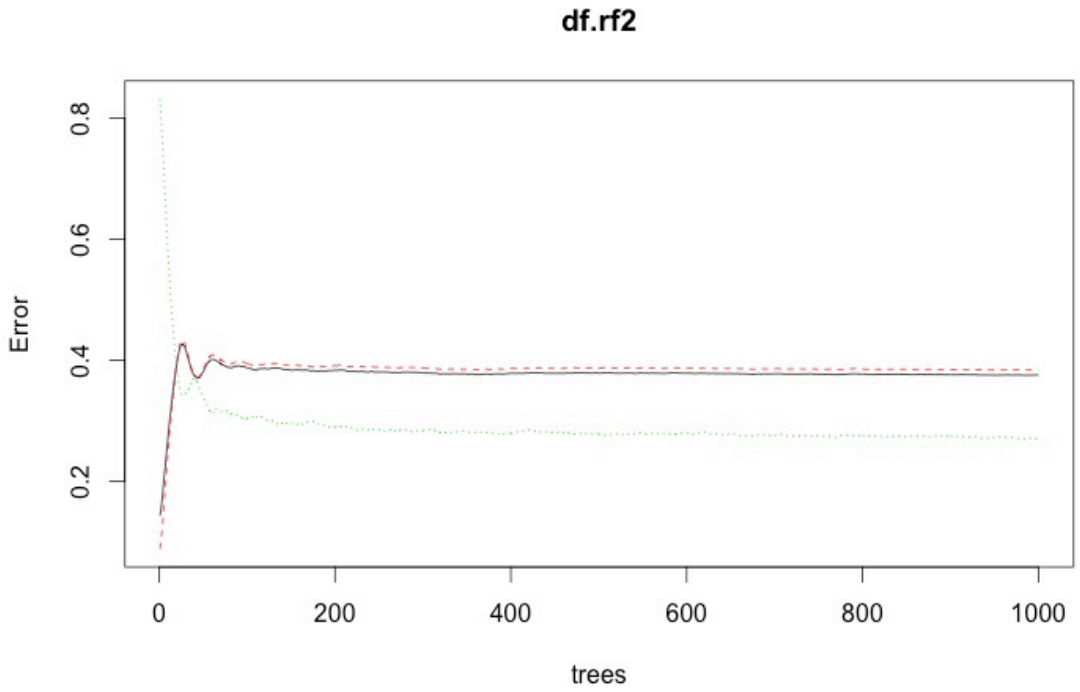
Cuadro 10: Matriz de confusión Variables Seleccionadas

		OOB estimate of error rate: 37.52%		
		Buenos	Malos	
		0	1	class.error
Buenos	0	27897	17382	0,3838866
Malos	1	1012	2735	0,2700827

[10] No se presentarán las matrices de correlación, ni si quiera en el caso de este random forest con 39 variables debido a que su análisis con tantas variables es complejo. Se ven algunas correlaciones entre algunas variables evidentes (edad y estado civil) y entre otra variables que no aparentaban estar correlacionadas. De todas formas, muchas de estas correlaciones resultan estadísticamente no significativas.

A continuación se presenta el gráfico de los errores presentados el el cuadro xx. Los errores se grafican para los 1000 árboles tanto para el modelo global como para cada una de las clases.

Gráfico 8: Error OOB



Como se observó anteriormente, el error de clasificación casi no varía respecto del modelo donde se utiliza la totalidad de las variables.

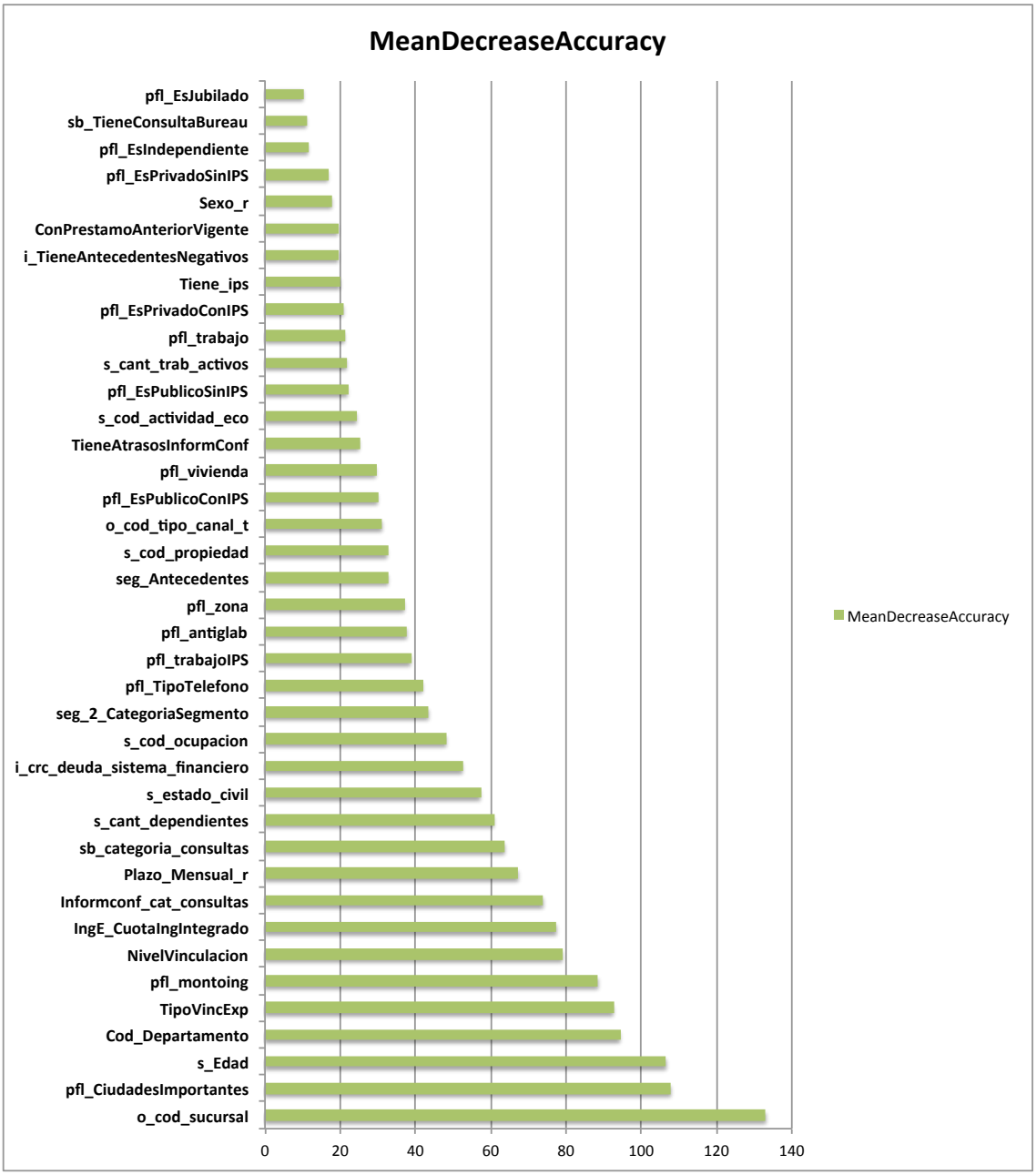
Importancia de las Variables

A continuación se presenta la importancia de las variables elegidas para la corrida:

Cuadro 11: Importancia de las Variables

	MeanDecreaseAccuracy	MeanDecreaseGini
o_cod_sucursal	133,1286	281,136883
pfl_CiudadesImportantes	107,906	200,644273
s_Edad	106,68643	293,291627
Cod_Departamento	94,35655	158,790657
TipoVincExp	92,59156	124,795433
pfl_montoing	88,59952	367,149717
NivelVinculacion	79,0888	155,886729
IngE_CuotaIngIntegrado	77,53854	367,077856
Informconf_cat_consultas	74,01245	117,656464
Plazo_Mensual_r	67,40652	188,1785
sb_categoria_consultas	63,80406	106,055575
s_cant_dependientes	60,85365	114,470425
s_estado_civil	57,31367	97,195982
i_crc_deuda_sistema_financiero	52,56312	266,387096
s_cod_ocupacion	48,2804	51,303534
seg_2_CategoriaSegmento	43,58829	87,837122
pfl_TipoTelefono	41,96332	69,07591
pfl_trabajoIPS	38,79831	73,881497
pfl_antiglab	37,81788	55,015343
pfl_zona	37,11612	76,595601
seg_Antecedentes	32,8431	33,961096
s_cod_propiedad	32,59635	42,583465
o_cod_tipo_canal_t	31,24314	138,10423
pfl_EsPublicoConIPS	30,05396	25,166184
pfl_vivienda	29,57446	34,180123
TieneAtrasosInformConf	25,13858	33,865114
s_cod_actividad_eco	24,49541	69,97035
pfl_EsPublicoSinIPS	22,02432	31,046836
s_cant_trab_activos	21,8175	55,524299
pfl_trabajo	21,2747	15,863259
pfl_EsPrivadoConIPS	20,98751	22,901896
Tiene_ips	20,08049	33,073953
i_TieneAntecedentesNegativos	19,5643	14,275417
ConPrestamoAnteriorVigente	19,48325	27,382763
Sexo_r	17,92487	52,591752
pfl_EsPrivadoSinIPS	16,8931	23,760916
pfl_EsIndependiente	11,59896	4,802897
sb_TieneConsultaBureau	11,16149	5,212397
pfl_EsJubilado	10,32643	4,120551

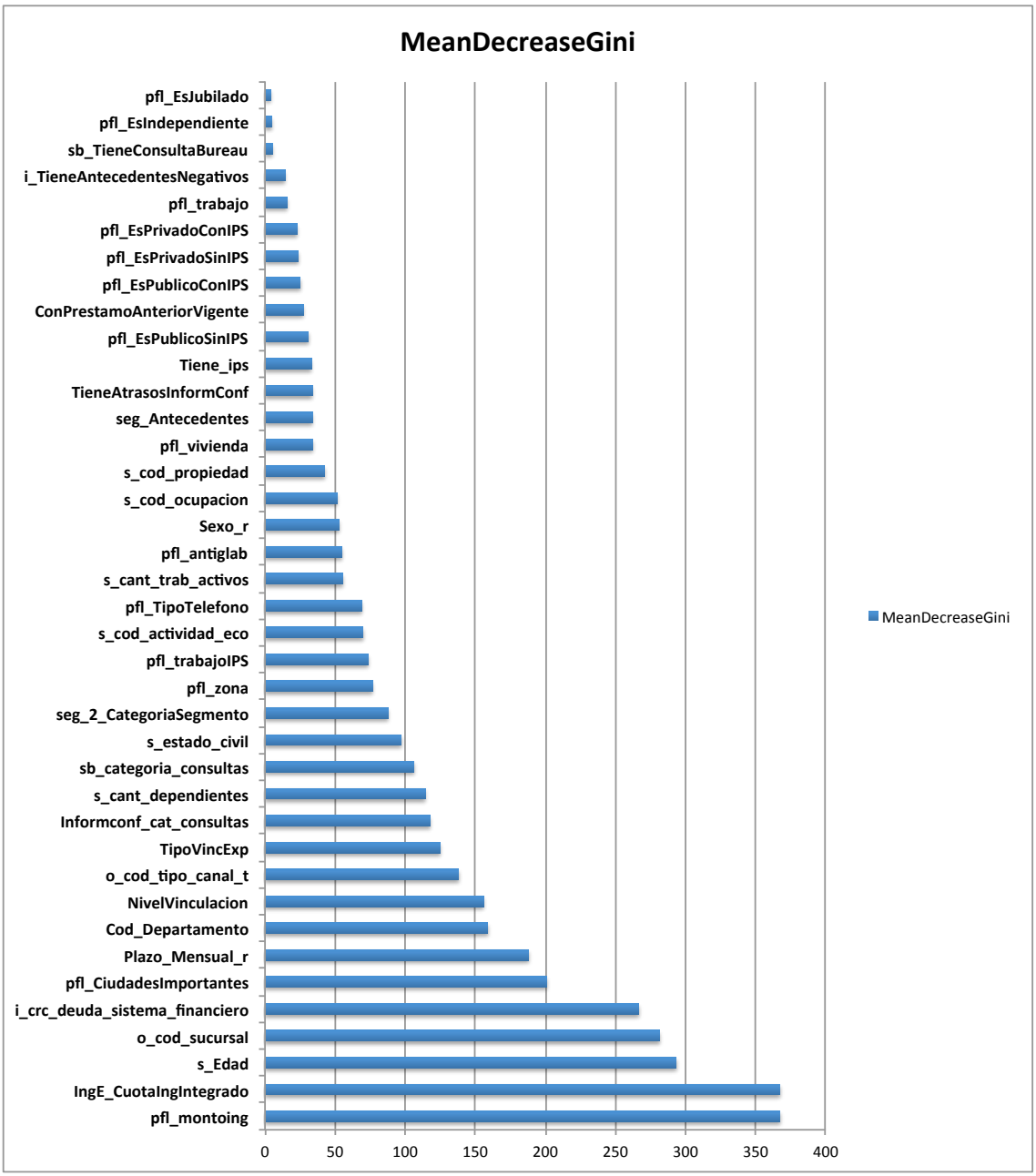
Gráfico 9: Mean Decrease Accuracy



Vemos que las variables que eran importantes en la primer corrida (la corrida que incluía todas las variables) siguen siendo importantes ahora. Para esta corrida, además de descartarse las variables redundantes se descartaron también las variables de menor importancia (todas aquellas cuyo mean decrease accuracy fuese menor a 5. Se sugiere en general descartarlas cuando esta medida sea menor a 3, pero dado que algunas de esas variables todavía tenían superposiciones con otras más relevantes y otras parecían tener poca importancia, ya que

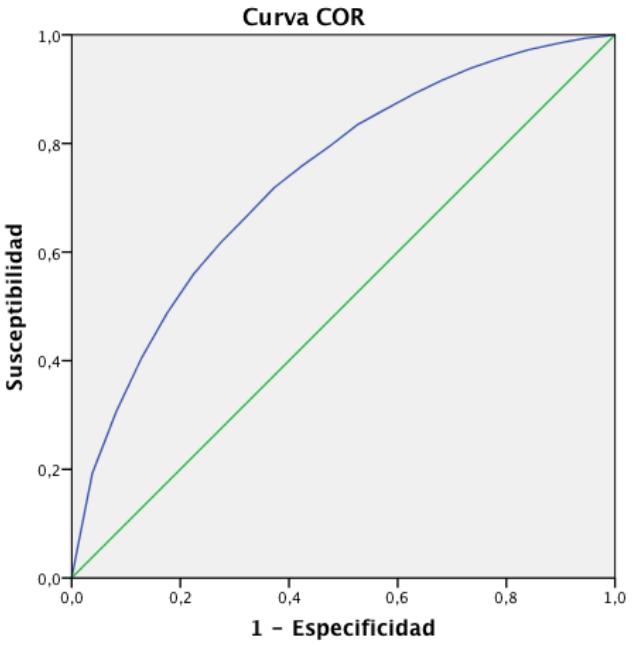
tenían mucho peso en alguna de sus categorías y muy poco en otras, se decidió sacarlas de todas formas).

Gráfico 10: Mean Decrease Gini



Poder Discriminante: Curva COR y K-S

Gráfico 11: Curva COR



Los segmentos diagonales son producidos por los empates.

Área bajo la curva

Variables resultado de contraste: RF -Total-tramos 5% de probmora RF (Vars elegidas)

Área
,736

El área debajo de la curva COR correspondiente a este modelo es igual a 0.736. Es decir, se espera clasificar correctamente a los morosos con una probabilidad de 73,6%.

A continuación se presenta el K-S para este modelo:

Cuadro 12: K-S

Estadísticos de contraste[a]

		RF -Total- tramos 5% de probmora RF (Vars elegidas)
Diferencias más extremas	Absoluta	,346
	Positiva	,346
	Negativa	,000
Z de Kolmogorov-Smirnov		20,337
Sig. asintót. (bilateral)		,000

a. Variable de agrupación: Malo60iConProd

Por otro lado, el K-S en este modelo indica que la mayor diferencia entre la frecuencia acumulada de la muestra de morosos y la frecuencia acumulada de la muestra de no morosos es 0,346.

Indicadores de Desempeño

A continuación se presenta una cuadro con los indicadores de desempeño para este modelo.

Cuadro 13: Indicadores de Desempeño

Nivel de Riesgo	Malos			Buenos		Diferencia
	%	%F	% malos hasta este punto	%	%F	%F.buenos-%F.Malos
1	,0090	*0,6%*	*0,9%*	,9910	*5,4%*	4,77%
2	,0156	*1,6%*	*1,2%*	,9844	*10,7%*	9,05%
3	,0177	*2,8%*	*1,4%*	,9823	*16,0%*	13,26%
4	,0245	*4,4%*	*1,7%*	,9755	*21,3%*	16,94%
5	,0271	*6,1%*	*1,9%*	,9729	*26,6%*	20,41%
6	,0337	*8,4%*	*2,1%*	,9663	*31,8%*	23,45%
7	,0384	*10,9%*	*2,4%*	,9616	*37,0%*	26,15%
8	,0428	*13,7%*	*2,6%*	,9572	*42,2%*	28,53%
9	,0437	*16,5%*	*2,8%*	,9563	*47,4%*	30,85%
10	,0603	*20,5%*	*3,1%*	,9397	*52,5%*	31,99%
11	,0556	*24,1%*	*3,3%*	,9444	*57,6%*	33,47%
12	,0608	*28,1%*	*3,6%*	,9392	*62,6%*	34,57%
13	,0788	*33,2%*	*3,9%*	,9212	*67,6%*	34,41%
14	,0776	*38,3%*	*4,2%*	,9224	*72,6%*	34,32%
15	,0877	*44,0%*	*4,5%*	,9123	*77,6%*	33,52%
16	,1088	*51,2%*	*4,9%*	,8912	*82,4%*	31,23%
17	,1269	*59,5%*	*5,3%*	,8731	*87,1%*	27,65%
18	,1489	*69,2%*	*5,9%*	,8511	*91,7%*	22,52%
19	,1753	*80,7%*	*6,5%*	,8247	*96,2%*	15,51%
20	,2956	*100,0%*	*7,6%*	,7044	*100,0%*	0,00%
% malos general	7,64%				*K-S*	*34,57%*
Diferencia 5%	28,66%				*COR*	*73,63%*
					Gini	*47,3%*

Se puede observar que este modelo también orden de forma adecuada. Y como el anterior modelo de random forrest, sigue acumulando peor que el modelo de regresión logística, si bien la diferencia no es sustancial.

El K-S de este modelo es 34,57%.

Resumen y Conclusiones

En este trabajo, se compararon dos técnicas: random forest y regresión logítica respecto a las variables seleccionadas por cada una y a la precisión en la predicción de cada modelo.

Se utilizó como base de prueba un date set de solicitudes de un banco paraguayo que con tenía 49026 solicitudes de crédito para clientes de un segmente particular.

A continuación se presenta un cuadro que resume los resultados obtenidos por todos los modelos:

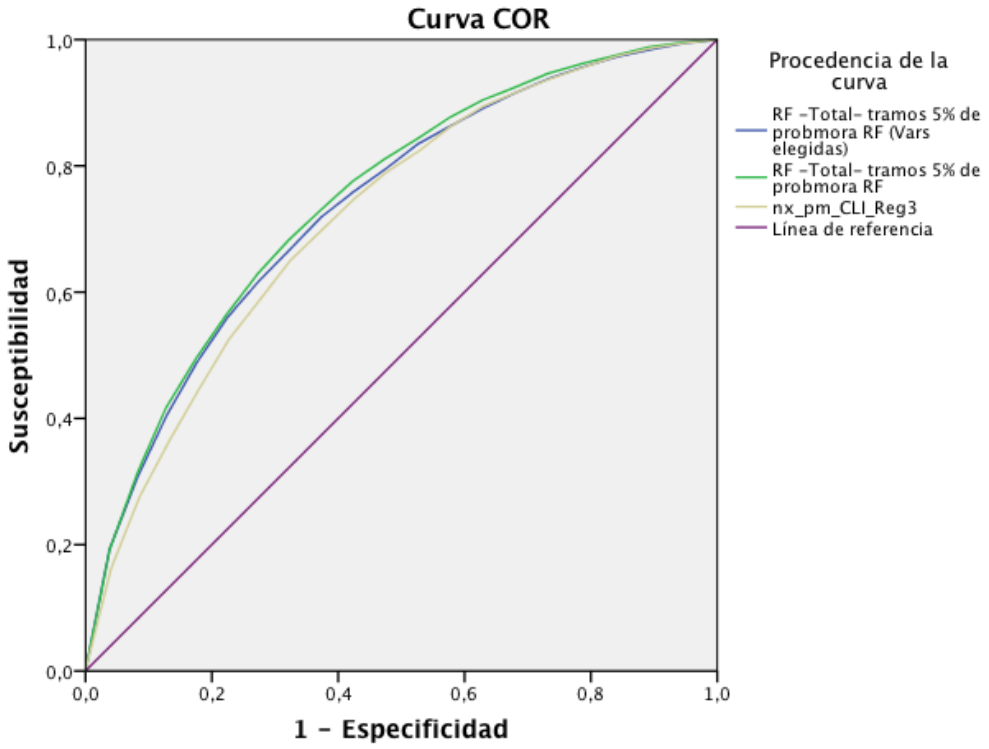
Cuadro 14: Resultados Obtenidos con los modelos

	Modelos		
	Regresión Logística	Random Forest (Todas las Vars.)	Random Forest (Vars. Elegidas)
K-S	32,61%	36,09%	34,57%
COR	72,17%	74,54%	73,63%
GINI	44,34%	49,09%	47,26%
Diferencia 5% mejor-peor	24,03%	28,68%	28,66%
Cant. Vars.	27	90	39

Como puede verse en el cuadro anterior, con la utilización del random forest los resultados mejoran en todos los indicadores, ya sea con el modelo que utiliza todas las variables, aún las redundantes, o el modelo que utiliza las variables elegidas.
Si bien los resultados del modelo de random forest utilizando todas las variables sin selección previa pareciera arrojar mejores resultados, es conveniente aclarar que muchas de estas medidas son sensibles a las correlaciones (ver Metz et al. (1984) para mayor información).

En el gráfico siguiente puede observarse el resultado de la curva COR para los tres modelos.

Gráfico 12: Curva COR Comparación Modelos



Los segmentos diagonales son producidos por los empates.

El gráfico muestra las diferencias entre las tres curvas COR de los modelos.

Se puede observar que si bien en la tabla anterior se notaba una diferencia entre los dos modelos random forest (curvas verde y azul), la diferencia no está tan clara en este gráfico.

Pero si existe una diferencia mayor con el modelo de regresión logística (curva marrón).

Como puede observarse en el cuadro siguiente, se presentan las principales variables detectadas como importantes para cada modelo.

Cuadro 15: Algunas Variables Importantes para los modelos

Modelos		
Regresión Logística	Random Forest (Todas las Vars.)	Random Forest (Vars. Elegidas)
TipoVincExp	o_cod_sucursal	o_cod_sucursal
Edad	pfl_montoing	pfl_CiudadesImportantes
pfl_tiene_telefono	NivelVinculacion	s_Edad
seg_2_CategoriaSegmento	pfl_CiudadesImportantes	Cod_Departamento
plazo	Informconf_cat_consultas	TipoVincExp
pfl_trabajoIPS	IngE_CuotaIngIntegrado	pfl_montoing
	Plazo_Mensual_r	Plazo_Mensual_r

El cuadro anterior muestra algunas de las variables más importantes para cada modelo. Si bien el orden difiere, variables como la vinculación y la experiencia que tiene el individuo en la utilización de productos del banco, la edad, el plazo y las consultas en inforconf entre otras.
Si bien no selecciona la variable tramada de la misma manera, o en mucho casos selecciona la variable sin transformar (por ejemplo la edad, en el caso del modelo logístico se la incluye como piecewise, pero en el modelo random forest seleccionó como mas relevante la variable original continua), muchas de las variables presentes en el modelo de regresión logística. Algunas otras variables que no fueron tomadas en cuenta por este último aparecen como relevantes en el random forest, como la sucursal o la ciudad.

Como conclusión podemos decir que los resultados con la técnica de random forest mejoran, aunque para este data set parecieran no mejorar sustancialmente. De todas formas, hay que tener en cuenta que los indicadores pueden verse afectados por la correlación en la variables, por lo que es aconsejable tomar estos resultados y hacer un análisis más profundo de correlación, seleccionar un poco mejor las variables, y hacer algunas otras corridas.

En cuanto a las variables seleccionadas, algunas de ellas fueron similares en ambos métodos, pero dado que otras variables fueron seleccionadas como relevantes para el random forest y estas no estaban incluidas en una primera instancia en la regresión logística, por no haber arrojado resultados favorables en los análisis previos, es aconsejable realizar alguna corridas de la regresión logística incluyendo las variables que resultaron relevantes para el random forest para chequear si los resultados mejoran.

Más allá de que los resultados de los indicadores no sean radicalmente distintos entre random forest y resgresión logística, es importante resaltar que la mayoría de las variables utilizadas en la regresión logística tienen una gran cantidad de trabajo invertido. Es decir, para poder llegar a esos resultados se invirtió una gran cantidad de tiempo, mientras que muchas de las variables que se utilizaron en el random forest, y la mayoría que el random forest reconoció como relevantes tienen poco tratamiento. Lo cual implica que a priori valdría la pena seguir investigando esta técnica debido a que con poco trabajo en las variables de todas formas se

consiguieron mejores resultados.
Además, la técnica ofrece una manera sencilla de tratar outliers y missing values que pareciera a priori funcionar satisfactoriamente.

Blibliografía

Amit, Yali; Geman, Donald, 1997. *Shape quantization and recognition with randomized trees*. Neural Computation 9 (7): 1545–1588.

Breiman, L., Friedman, J., Olshen, R., and Stone, C., 1985.Classication and Regression Trees , Wadsworth/BrooksCole,Monterey, CA.

Breiman, L., 1994. *Bagging Predictors*. Technical Report No. 421. Department of Statistics.

University of California Berkeley, California

Breiman, L., 1999. Random Forest – Random features Statistics Department. Technical Report 567. University of California Berkeley, California.

Breiman , L., 2001. *Random Forests*. Machine Learning, 45, 5-32.

Collett, D., Modelling Binary Data. Second ed. 2003, Boca Raton: Chapman &Hall/CRC.

Delen, D., Walker, G., Kadam, A., 2005. *Predicting breast cancer survivability: a comparison of three data mining methods*. Artificial intelligence in medicine 34, 113-127.

Dudoit, S., Fridlyand, J., Speed, P., 2002. *Comparison of discrimination methods for classification of tumors using gene expression data*. J. Amer. Statist. Assoc. .97, 77–87.

Harrell, F.E., Lee, K.L., Califf, R.F., et al., 1984. Regression modeling strategies for improved prognostic prediction. Stat. Med, 1984. 3: p. 143-152.

Harrell, F.E., Lee, K.L., Mark, D.B., et al, 1996. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat. Med. 15: p. 361-387

Ho, Tin Kam, 1995. *Random Decision Forest* . Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.

Hosmer, D.W., Lemeshow, S, 2000. *Applied Logistic Regression*. Second ed. New York: Wiley.

Lee, J. W., Lee, J. B., Park, M., et al., 2005. *An extensive comparison of recent classification tools applied to microarray data*. Computational statistics & data analysis 48, 869-885.

Metz CE, Wang PL, Kronman HB, 1984. A new approach for testing the significance of differences between ROC curves measured from correlated data. In: Deconinck F, editor. Information Processing in Medical Imaging. The Hague: Nijhoff. 432–45.

Ming Geng, 2006. A Comparison of Logistic Regression to Random Forests for Exploring Differences in Risk Factors Associated with Stage at Diagnosis Between Black and White Colon Cancer Patients. University of Pittsburgh

Nan Lin, Baolin Wu, Ronald Jansen, Mark Gerstein and Hongyu Zhao, 2004.*Information assessment on predicting protein-protein interactions*. BMC Bioinformatics, 5:154

Anne Ruiz-Gazen, Nathalie Villa, 2008. *Storms Prediction: Logistic Regression vs Random Forest for Unbalanced Data*. Institut de Mathématiques de Toulouse.

Anexos:

i. Tipos de Variables Utilizadas en la Regresión Logística

Se utilizarán tres tipos de variables en el modelo: Discretas, continuas y piecewise. Cada tipo de variables es distinta en su naturaleza y en la forma en la que califican a la solicitud.

- **Discretas**

Las variables discretas son aquellas que toman un conjunto finito de valores. Las variables discretas utilizadas pueden ser "originalmente" discretas (por ejemplo el *estado civil*, que se puede seleccionar una alternativa entre un grupo acotado de opciones) o variables que originalmente eran continuas pero fueron tramadas en rangos y por lo tanto "discretizadas". Existe una categoría considerada base y el resto de las categorías son comparadas con la base.

Por lo tanto, las categorías con coeficiente negativo se encuentran "favorecidas" ya que implican una menor probabilidad de mora.

- **Continuas**

Se multiplica el coeficiente del tramo por el valor que toma la variable. Entonces, si una variable tiene coeficiente asociado negativo, a mayor valor en la variable, menor será la probabilidad de mora.

- **Piecewise**

Las Variables Piecewise son un conjunto de variables construidas a partir de una variable continua original. El objetivo de utilizar estas variables es poder captar cambios de pendientes al interior de la variable continua. A partir de una variable continua se genera un conjunto de variables piecewise, cada uno representa a un tramo de la variable continua original.

Dado que son variables muy similares a las continuas (continuas por tramos), la forma de puntuar es análoga a la de una continua pero para cada tramo.

Si a partir de una variable continua como la edad se hacen 3 variables piecewise (pw) en los tramos:

1. Pw1: <=20
2. Pw2: >20 y <=28,
3. Pw3: >28 y <=35 y
4. Pw4: >35 y <=45
5. Pw5:>45

Las variables piecewise construyen de la siguiente forma:

Primero se fijan todas las variables en cero. Luego se remplaza el valor en caso de que cumplan las condiciones que se especifican a continuación.

- si edad>0 y <=20, entonces: pw1 = edad
  si edad >20 , entonces: pw1 = 20
- si edad>20 y <=28, entonces: pw2 = edad-20
  si edad >28 , entonces: pw2 = 28-20
- si edad>28 y <=35 , entonces: pw3 = edad - 28
  siedad >35 , entonces: pw3 = 35-28
- si edad >35 y <=45 , entonces: pw4 = edad - 35
  siedad >45 , entonces: pw4 = 45-35
- si edad >45 , entonces: pw5 = edad – 45

Suponiendo que el beta para cada piecewise es beta1, beta2, beta3, beta4 y beta5 respectivamente,

- Si la edad=22. Las variables piecewise toman los siguientes valores:
  - pw1=20
  - pw2=22-20=2
  - pw3=0
  - pw4=0
  - pw5=0

Luego, para calcular la probabilidad de mora (si todos los tramos entran en el modelo), el valor que entra en Z_edad=beta1*20+beta2*2+beta3*0+beta4*0+beta5*0.

- Si la edad=36. Las variables piecewise toman los siguientes valores:
  - Pw1=20
  - Pw2=28-20=8
  - Pw3=35-28=7
  - Pw4=36-35=1
  - Pw5=0

Luego, Si todos los tramos entran en el modelo el resultado que entra como sumando en Z_edad=beta1*20+beta2*8+beta3*7+beta4*1+beta5*0.

En el caso de que algún tramo no entra en el modelo no se pone tampoco como sumando para obtener el resultado que va a Z.

ii. Variables en la Regresión

Estimación del Ingreso

Dado que no todos los individuos tienen empleos formales, este dato muchas veces es difícil de conseguir. Es por ello, que muchas veces se recurre a estimaciones del mismo. Los

resultados obtenidos con las estimaciones del ingreso a lo largo de los años para los modelos son satisfactorias por lo cual suele ser una práctica recurrente su estimación.

Debido a que a veces los clientes tienen ingreso declarado, según cual sea el tipo de ocupación del cliente se tomará como válido dicho ingreso o no en el momento de calcular el "Ingreso Integrado". Por cuestiones de confidencialidad, no se mostrará cómo se calcula esta variable. Pero si se aclara que en el caso en que se poseen datos de ingresos declarados se toma este valor, sino, el dato es estimado con una fórmula teniendo en cuenta diversas características socio-demográficas del individuo.

Categoría Segmento según Informconf

El segmento del cliente se basa en la información desagregada de InformConf sobre las consultas realizadas por los clientes a distintos grupos de empresas.

Se tipificaron las empresas en donde el cliente había consultado antes de hacer la solicitud en el cliente. La tipificación se realizó de acuerdo a cómo puede ser vista cada empresa por los clientes de este Banco en particular en relación al banco: competencia, bancos de primera línea, etc.

Se estudió la cantidad de consultas realizadas por los solicitantes, la frecuencia y el tiempo pasado entre ellas.

Esta variable se ingresa al modelo como variable categórica.

Categoría Segmento y Antecedentes en informconf

Como InformConf informa en qué empresas el cliente consultó y si tiene o no antecedentes negativos, se crea una variable que combina esta información. Entonces, en la regresión también entra una variable que da una puntuación a la combinación Tipo de Empresa Consultada y Comportamiento en el pasado.

Esta variable se ingresa al modelo como variable categórica.

Vinculación y Experiencia

Esta variable se construye de acuerdo a la cantidad de productos, la antigüedad que el cliente tiene con dichos productos y el comportamiento que ha mostrado con ellos. Esta variable ingresa al modelo en la forma de cuatro variables dicotómicas:

- **"Muy Baja"** implica alguna de las siguientes situaciones:
- El cliente no ha tenido atraso, tiene únicamente Tarjeta de Crédito y la antigüedad de la tarjeta es menor a 12 meses.
- El cliente no ha tenido atraso, tiene únicamente un crédito cancelado y la antigüedad del crédito es menor a 12 meses.
- El cliente no llegó a atrasarse 60 días pero sí se atrasó más de 2 veces al menos 30 días. No es Malo 60 pero sí cae en mora leve de forma recurrente.

- **"Baja"** implica alguna de las siguientes situaciones:

- El cliente no ha tenido atraso, tiene únicamente Tarjeta de Crédito y la antigüedad de la tarjeta es mayor a 12 meses.
- El cliente no ha tenido atraso, tiene únicamente un crédito cancelado, la antigüedad del crédito es mayor a 12 meses y ha tenido hasta 3 o menos préstamos anteriores.
- El cliente no ha tenido atraso, tiene Tarjeta de Crédito y crédito cancelado. Además, la antigüedad de los productos es inferior a 12 meses.
- El cliente no ha tenido atraso, tiene únicamente un crédito activo y la antigüedad del crédito es menor a 12 meses.
- El cliente no ha tenido atraso, tiene Tarjeta de Crédito y crédito activo. Además, la antigüedad de los productos es inferior a 12 meses.

- "**Media**" implica alguna de las siguientes situaciones:
- El cliente no ha tenido atraso, tiene únicamente un crédito cancelado, la antigüedad del crédito es mayor o igual a 12 meses y ha tenido hasta 3 o menos préstamos anteriores.
- El cliente no ha tenido atraso, tiene Tarjeta de Crédito y crédito cancelado. Además, la antigüedad de los productos es superior o igual a 12 meses y ha tenido hasta 3 o menos préstamos anteriores.
- El cliente no ha tenido atraso, tiene únicamente crédito activo. Además, la antigüedad del crédito es superior o igual a 12 meses y ha tenido hasta 3 o menos préstamos anteriores.
- El cliente no ha tenido atraso, tiene Tarjeta de Crédito y crédito activo. Además, la antigüedad de los productos es superior o igual a 12 meses y ha tenido hasta 3 o menos préstamos anteriores.

- **"Alta"** implica alguna de las siguientes situaciones:
- El cliente no ha tenido atraso, tiene Tarjeta de Crédito y crédito cancelado. Además, la antigüedad de los productos es superior o igual a 12 meses y ha tenido más de 3 préstamos anteriores.
- El cliente no ha tenido atraso, tiene únicamente crédito activo. Además, la antigüedad del crédito es superior o igual a 12 meses y ha tenido más de 3 préstamos anteriores.
- El cliente no ha tenido atraso, tiene Tarjeta de Crédito y crédito activo. Además, la antigüedad de los productos es superior o igual a 12 meses y ha tenido más de 3 préstamos anteriores.

Región

Debido a que la región en la que vive el cliente es fuente de información de perfil, se incluye una variable con la región correspondiente al cliente en la regresión logística.

En el caso en que no esté disponible este dato, se le asigna la región en donde recibe correspondencia. En caso de no tener información de residencia ni de correspondencia se incluye la de la sucursal donde hizo la solicitud. Esta variable se construye únicamente utilizando la información que se obtuvo de la base del Banco.

Esta variable se ingresa al modelo como variable categórica.

Cantidad De Dependientes

Cantidad de personas que dependen del solicitante

Trabajo

Cuál es el trabajo del solicitante. Se toma de la solicitud y se lo agrupa en categoría según si es trabajo público, privado, con o sin IPS (aportes).

Tiene Teléfono Fijo

Esta es una variable dummy que indica si la persona tiene o no teléfono fijo en su domicilio.

Es de esperar que aquellos que tengan teléfono fijo tengan mejor comportamiento dado que facilita la función de recupero y las personas que son más fácilmente accesables tienden a tener un mejor comportamiento.

Edad en Años

Debido a que los individuos se comportan de distinta forma de acuerdo al rango de edad en el que se encuentren, esta variable va a ingresar al modelo como variable piecewise.

Antigüedad en el Trabajo (en meses)

Es de esperar que aquellas personas que tengan una mayor permanencia en sus puestos de trabajo tengan un mejor comportamiento.

Plazo en meses

Se prevé que los individuos que soliciten créditos con plazos más largos van a ser más morosos, por ello, se castiga más cuánto más largo es el plazo en meses.

Esta variable también ingresa como piecewise.