UNIVERSIDAD DE COSTA RICA

ESCUELA DE MATEMÁTICA

DEPARTAMENTO DE MATEMÁTICA PURA Y CIENCIAS ACTUARIALES

DISTRIBUCIÓN DE PÉRDIDAS

Proyecto Distribución de Pérdidas

ANTEPROYECTO

Grupo 05

Realizado por

Daniel Núñez - B85667

Andrés González - B83413

UNIVERSIDAD DE COSTA RICA

EMat Escuela de Matemática

Índice general

Capítulo 1

Objetivos

1. Objetivo General

Determinar la pérdida esperada de un banco a la hora de decidir la elegibilidad de una persona para un crédito.

2. Objetivos Específicos

1. Implementar un modelo de clasificación para la eligibilidad de una persona solicitante de crédito.
2. Estimar la correlación y dependencia de la elegibilidad de crédito y las variables significativas obtenidas por el modelo de clasificación mediante un modelo de cópulas.
3. Deducir la pérdida esperada del banco a través de la construcción de una función de densidad conjunta utilizando medidas de riesgo como el VaR y el CVaR.

Capítulo 2

Marco Teórico

Para un primer acercamiento a contestar la pregunta de investigación, antes de poder determinar correlaciones entre las variables de estudio para determinar las pérdidas del banco por impago crediticio se necesita una manera de determinar alguna medida como el score crediticio por individuo. Es por esto que primero se va a realizar un modelo de clasificación de elegibilidad de crédito, para lo cual se implementó un modelo de regresión logística. Este tipo de regresión como menciona James, a diferencia de los métodos de regresión, se teoriza un nuevo tipo de regresión llamada regresión logística, la cual sirve para clasificar de manera binaria una variable. Entonces en vez de entrenar un modelo para determinar si hay una correlación entre la variable dependiente y las covariables, se entrena para clasificar en alguna de dos categorías a la variable dependiente de acuerdo a sus covariables. Esta definición del modelo es similar a la que hace Chitarroni en su artículo ya que lo define como un instrumento de análisis multivariado y dependiendo del enfoque se puede utilizar para realizar predicciones o inferencia. Se menciona que es muy útil cuando la variable dependiente es de carácter dicotómico (binario). Asimismo, se aclara que cuando las covariables son categóricas estas deberían de recibir una transformación y convertirlas en variables "dummy", es decir, variables simuladas.

Para el modelo de regresión logística se va a utilizar la siguiente forma:

$$p(\boldsymbol{X}) = \frac{e^{\beta_0 + X_1\beta_1 + \cdots + X_p\beta_p}}{1 + e^{\beta_0 + X_1\beta_1 + \cdots + X_p\beta_p}} \tag{2.1}$$

En donde se cumple $0 < p(X) < 1$ y X_k con $k = 1, ..., p$ corresponden a las covariables con las que se entrena el modelo. (James et al., 2021)

Sin embargo, en el artículo expuesto por Chitarroni se le da más peso a las pruebas de significancia de variables así como a la interpretabilidad de los resultados enfocado más en un modelo predictivo de manera que especifica que este tipo de modelo se puede utilizar de manera predictiva, el cual no solo determina la probabilidad de la variable dependiente sino el peso que tienen las covariables para realizar la predicción, lo cual también ayuda a nivel interpretativo pues se pueden determinar que variables son más significativas que otras para el modelo. Mientras que en su libro James también menciona diagnósticos de los modelos de clasificación que ayudan a determinar si un modelo es mejor que otro o está mejor ajustado como por ejemplo el concepto de especificidad y sensibilidad los cuales miden la tasa de falsos positivos y los falsos negativos respectivamente. Y como de acuerdo al tipo de problema para el cual se está realizando el modelo, se debe considerar ajustes del modelo para aumentar estas medidas. Otro diágnostico bastante utilizado es el de la curva ROC (Receiver Operation Curve), la cual sirve para graficar la tasa de falsos positivos con la sensibilidad del modelo.

Para la segunda parte del trabajo, donde ya se plantea el hecho de encontrar las pérdidas del banco, se planean utilizar modelos que involucren cópulas. Estos modelos sirven para encontrar distribuciones conjuntas que generalmente tienen un alto grado de correlación entre sí, por lo que analizarlas por separado no es lo más recomendable. Pues como menciona Escarela, las cópulas bivariadas son funciones que intentan correlacionar dos distribuciones univariadas por lo que estos modelos ayudan a obtener una distribución conjunta a partir de varias funciones de distribución asociadas a variables aleatorias con una cierta relación entre sí. Es decir, con este método se construye una distribución multivariada a partir de las distribuciones univariadas de las variables respectivas. Este tipo de modelo también resulta en una forma de estructurar la dependencia de estas parejas de variables aleatorias en distribuciones conjuntas. Es por esto que son tan populares puesto que tienen una gran flexibilidad para encontrar distribuciones conjuntas a partir de cualquier pareja aleatoria, lo que es usual tener en muchas disciplinas.

Por lo que es importante definir el concepto de cópula bidimensional, la cual es una función bivariada de un vector aleatorio $\boldsymbol{V} = (V_1, V_2)$ cuyas marginales V_1 y V_2 son uniformes en el intervalo $\boldsymbol{I} = (0, 1)$. Por lo que la cópula es una función $C : \boldsymbol{I}^2 \to \boldsymbol{I}$ que satisface las siguientes 2 condiciones:

- Acotamiento

$$\lim_{v_j \to 1^-} C(v1, v2) = v_{3-j} \tag{2.2}$$

$$\lim_{v_j \to 0} C(v1, v2) = 0 \tag{2.3}$$

con $j = 1, 2$ y $(v_1, v_2)^T \in \boldsymbol{I}^2$

- Incremento

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0 \tag{2.4}$$

para toda $u_1, u_2, v_1, v_2 \in \boldsymbol{I}$ tal que $u_1 \leq u_2$ y $v_1 \leq v_2$

Sin embargo para el enfoque del trabajo estas no son de mucha utlidad pues como menciona Geidosch el uso de cópulas convencionales presenta dos limitantes, la primera es que cuando el número de dimensiones es mayor que

dos, el número de familias de cópulas aplicables se limita al caso elíptico y arquimediano. La segunda limitante es que solo dependencias simétricas y simples pueden ser modeladas y eso está lejos de ser considerado un escenario real. De ahí nace la necesidad de usar Vine Cópulas. Las Vines Cópulas se usan para describir cópulas de varias variables usando como base cópulas bivariadas o "pair-copulas" . La idea que existe de fondo con los Vines Cópulas es que distribuciones multivariadas se pueden descomponer secuencialmente en descomposiciones bivariadas y para llevar acabo dicho proceso se hace uso de los D-vine y los R-vine. Lo que se hace es tomar la densidad multivariada de la distribución y se transforma en funciones de densidad de las cópulas bivariadas.

Estas limitantes son las que motivan también a Daul a encontrar modelos más generales ya que en su artículo extiende el concepto de t-Cópula para generar una nueva t-Cópula agrupada que describe de manera más efectiva la dependencia que existe entre factores de riesgo en el ámbito financiero. En el artículo también se explica el proceso de cálculo de los nuevos parámetros de las t-Cópulas agrupadas. Básicamente la idea que hay de fondo es juntar los diferentes factores de riesgo en esta t-Cópula agrupada donde las variables aleatorias dentro de esta cópula tienen una t-cópula con diferentes grados de libertad. Como resultado, esto da una estructura de dependencia más flexible que se ajusta mejor para modelos de riesgos que tengan múltiples factores. A modo de conclusión, utilizan el value-at-risk para comparar el modelo usando t-Cópulas con el modelo de t-Cópulas agrupadas y este segundo siempre arroja resultados con medidas de riesgo mayores, por lo que se tiene un modelo más conservador y preciso.

Capítulo 3

Descripción de los Datos

Para efectos de la investigación se va a trabajar con una base de datos de un banco alemán que contiene información de personas solicitantes de un crédito y con base en esta información se determina su elegibilidad. Los datos fueron recuperados de kaggle y originalmente fueron obtenidos de Penn State Eberly College of Science. Los datos son públicos y son de libre acceso, sin embargo, por motivos de confidencialidad el nombre del banco nunca se menciona. Vale la pena mencionar que en el sitio de donde se extrajeron los datos, no se indica un contexto temporal de los mismos.

La población de estudio se define como las personas que solicitaron un crédito en esta entidad bancaria ubicada en Alemania mientras que la unidad estadística se define como la persona solicitante de un crédito en el banco alemán estudiado. La muestra para el desarrollo de dicha investigación consta de 1000 individuos. Asimismo, la base cuenta con 21 variables de interés donde en su columna matriz se encuentra la variable binaria de elegibilidad, que toma el valor de 1 si fue elegible y el valor de 0 si no lo fue.

Además, es importante resaltar que cuando se intenta desarrollar un modelo de score crediticio es común estudiar muchas características que pueden ser relevantes para determinar la probabilidad de impago de un individuo, sin embargo, a nivel estadístico usualmente no es lo más apropiado. Debido a esto, se realiza un proceso de depuración de la base con el fin de reducir la cantidad de categorías presentes en cada variable categórica. Más adelante se detallan los pormenores de este proceso. A continuación, una leve explicación de cada una de las variables que conforman esta base:

- Elegibilidad: toma el valor de 1 si fue elegible y el valor de 0 si no lo fue.
- Accout Balance: variable categórica que toma el valor de 1 si la persona no cuenta con ninguna cuenta en el banco, el valor de 2 si no tiene un balance pendiente con el banco y el valor de 3 si sí tiene un balance pendiente.

- Duration: la duración en meses del crédito solicitado
- Payment Status of Previous Credit: variable categórica que toma el valor de 1 si el individuo presenta problema con el pago del crédito anterior, el valor de 2 si ya lo pagó y el valor de 3 si no tiene problemas con el crédito anterior
- Purpose: variable categórica que toma el valor de 1 si es para un auto, 2 si es préstamos relacionados a vivienda, 3 si es para un crédito empresarial y 4 si es para cualquier otra cosa.
- Credit Amount: el monto del crédito solicitado en "Deutsche Mark" (DM), que es la unidad monetaria usada en la base.
- Saving/Stock value: toma el valor de 1 si no tiene nada de ahorros o de stock, de 2 si el valor es menor a los 100 DM, 3 si se está en el intervalo [100, 500[DM , 4 si está en [500, 1000[DM y 5 si está arriba de los 1000 DM.
- Length of currrent employment: variable categórica que toma el valor de 1 si es desempleado, de 2 si tiene menos de año, de 3 si es de 1 a 4 años, de 4 si es de 4 a 7 años y de 5 si es mayor a 7 años.
- Sex/marital status: toma el valor de 1 si es un hombre divorciado, el de 2 si es hombre soltero, el de 3 si es hombre casado/viudo y el de 4 si es mujer
- Guarantors: variable binaria que toma el valor de 1 si la persona tiene un fiador y de 0 si no lo tiene.
- Duration in current address: variable categórica que determina cuánto tiempo lleva la persona viviendo en la última dirección registrada. Toma el valor de 1 si es menos de un año, el de 2 si lleva entre uno y 4 años, el de 3 si lleva entre 4 y 7 años y el de 4 si lleva más de 7 años.
- Most valuable asset: toma el valor de 1 si no tiene ninguno, el de 2 si es un carro, el de 3 si es un seguro de vida y el de 4 si son bienes raíces.
- Edad: edad en años
- Tarjetas de créditos: toma el valor de 1 si el aplicante tiene tarjetas con otros bancos, el valor ed 2 si tiene tarjetas de créditos con empresas y el de 3 si no tiene nada.
- Type of department: toma el valor de 1 si no paga renta/hipoteca, el de 2 en caso de pague renta o hipoteca y el de 3 si es dueño de la vivienda/apartamento.
- No. of credits at this bank: toma el valor de 1 si solo tiene 1, el de 2 si tiene entre 2 y 3 créditos, el de 3 si tiene entre 4 y 5 créditos y el de 4 tiene más de 6 créditos.
- Occupation: 1 en caso de que sea desempleado o no calificado, 2 en caso de que sea un residente permanente no calificado, 3 en caso de que sea una persona calificada y 4 en caso de que sea un ejecutivo/a.
- No. of dependents: número de personas que mantiene. Toma el valor de 1 si son más de 3 y de 2 si son entre 0 y 2 personas.
- Foreign worker: variable binaria que toma el valor de 0 en caso de que sea un trabajador extranjero y 1 en caso de que no lo sea.

Como se pudo observar, la base cuenta con una cantidad considerable de variables de interés, donde leyendo la descripción de cada una se puede entender e intuir el posible impacto que tengan en la elegibilidad de las personas, pero vale la pena distinguir variables como "Payment Status" que es de esperarse que tenga un nivel significancia importante dentro del modelo.

Debido a ese exceso de variables, se tratará de reducirlas haciendo un análisis exploratorio de los datos de tal manera que se pueda eliminar variables que estén correlacionados entre sí. Para llevar a cabo dicho proceso, se

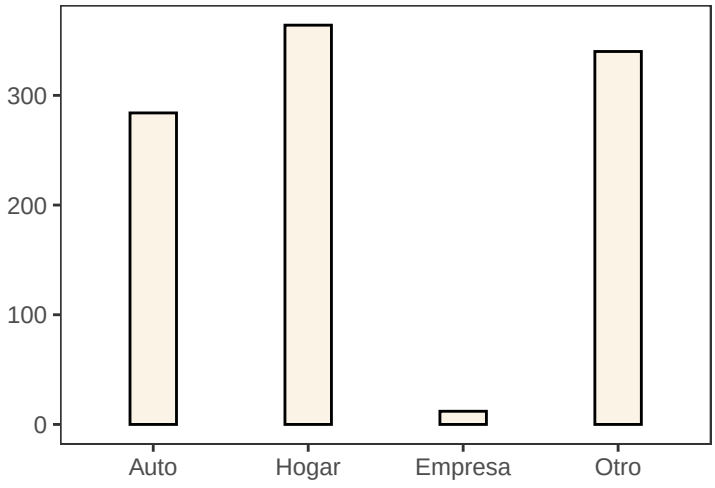
utilizara el modelo de cópulas para eliminar o unificar este tipo variables y se utilizarán modelos predictivos como la regresión logística para determinar la elegibilidad de las personas.

Capítulo 4

Análisis de Datos

Dado que la mayoría de variables son categóricas, primero se realizó un proceso de depuración de la base con el fin de reducir el número de variables estudiados. Para ello, primeramente se reducirán la cantidad de categorías que existen en cada uno de las variables combinando categorías que compartan carácterísticas o presenten muy pocas observaciones. Por ejemplo, en el caso de la variable del Próposito del Crédito, hay 11 categorías diferentes pero se simplificó de tal manera que solo hubiera 4 categorías. La primera son los préstamos relacionados a la compra de un Automóvil, ya sea nuevo o de segunda mano. La segunda a los préstamos realizados al Hogar, ya sea para compra de muebles o remodelaciones. Una tercera catogoría relacionada a créditos Empresariales y una última categoría que incluyera todos los préstamos cuya razón de solicitud no entre en las categorías anteriores.

Una vez hecha estas modificaciones, se puede extraer información interesante como la cantidad de préstamos solicitados por propósito cuyo gráfico se muestra a continuación:



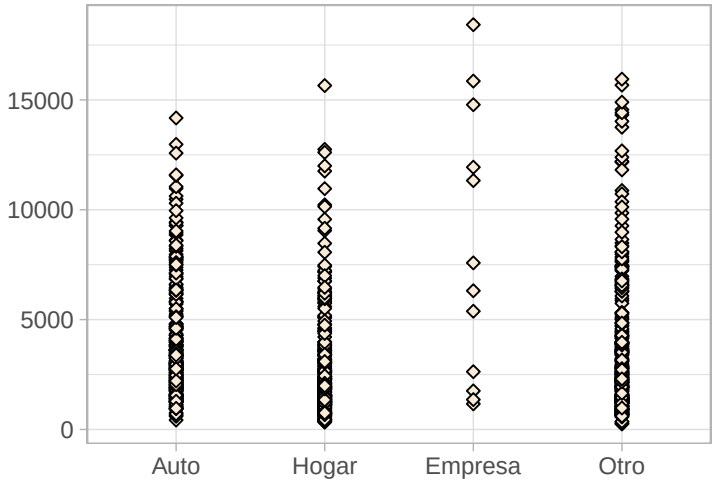
Elaboración propia con datos extraídos de Kaggle

Figura 4.1: Distribución de la variable Próposito del crédito

El gráfico anterior revela que la mayoría de préstamos solicitados son con asuntos relacionadas al hogar, mientras que hay una porción muy bajo de préstamos destinados al área empresarial.

Asimismo, en el siguiente gráfico se muestra la relación que existe entre el monto del crédito según su propósito, donde cabe destacar que aunque los motivos empresariales es la razón menos frecuente en la base de datos, el

crédito solicitado de mayor monto tiene como razón dicho motivo. Mediante este análisis fue que se descartó la posibilidad de combinar lel próposito "Empresa" con alguna otra categoría, ya que hay información importante que pueda revelar.



Elaboración propia con datos extraídos de Kaggle

Figura 4.2: Distribución del Monto del Crédito según su próposito

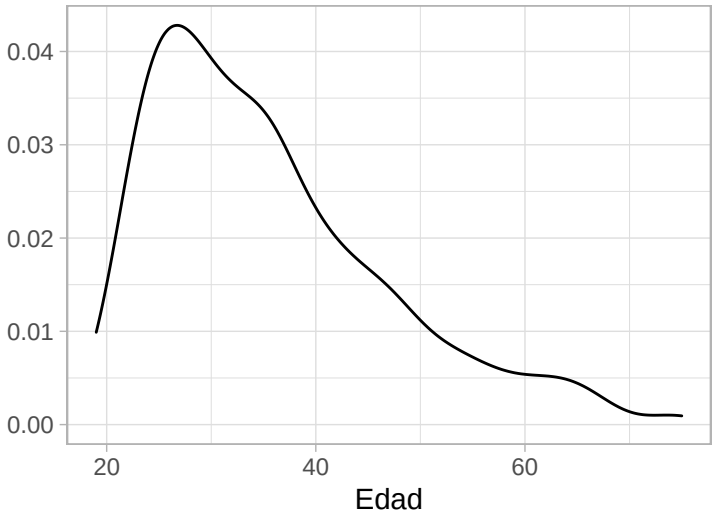
Haciendo una análisis similar para las edades, se llega a la conclusión de que existe una tendencia mayor en las personas cuyas edades estén en el rango de 20 a 40 años a solicitar préstamos como lo muestra la siguiente tabla:

Cuadro 4.1: Distribución de las edades

Rango de edad	Cantidad de observaciones
(10,20]	16
(20,30]	393
(30,40]	319
(40,50]	159
(50,60]	68
(60,70]	39
(70,80]	6

Note:
Elaboración propia con datos extraídos de Kaggle

De manera gráfica, la densidad de la variable edad viene dada por el siguiente gráfico donde se nota una asimetría hacia a la derecha que deja en evidencia lo que se habló anteriormente donde hay una tendencia mucho mayor en las personas entre los 20 y los 40 años de solictar préstamos.



Elaboración propia con datos extraídos de Kaggle

Figura 4.3: Histograma de la variable Edad

Otra variable que es de esperarse que sea de importancia es el Balance Actual que el solicitante tiene. Esta variable es categórica y toma el valor de 0 en caso de que el solicitante no tenga ninguna cuenta abierta con el Banco, el valor 1 en caso de que sí tenga una cuenta con el banco pero no tenga saldo o balance, y por último, toma el valor de 3 en caso de que sí tenga balance. La distribución de frecuencias viene dada por:

Cuadro 4.2: Distribución de la variable Balance Actual

Balance Actual	Cantidad de observaciones
1	274
2	269
3	457

Note:
Elaboración propia con datos extraídos de Kaggle

La mayoría de variables que se encuentran en la base son de carácter categórico, por lo que se presenta el siguiente cuadro que muestra información sobre las variables de carácter númerico continuo:

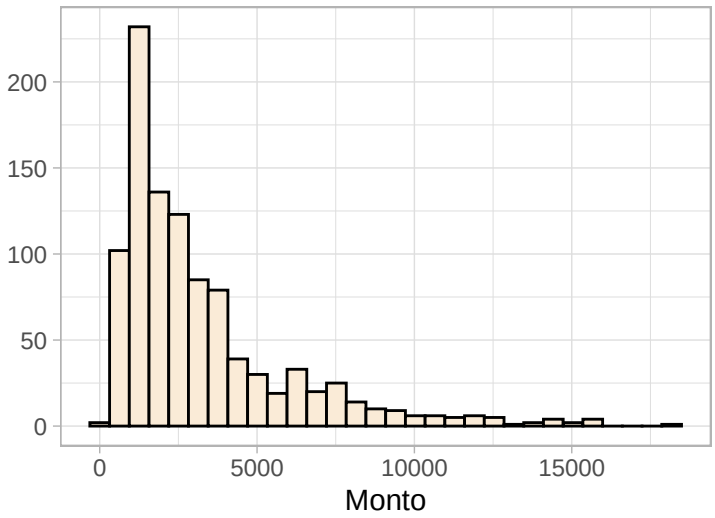
Cuadro 4.3: Resumen de 5 números

Min	Q1	Mediana	Q3	Max
250	1365.5	2319.5	3972.25	18424
4	12.0	18.0	24.00	72
19	27.0	33.0	42.00	75

Note:
Elaboración propia con datos extraídos de Kaggle

Dado a que eventualmente la idea es realizar un modelo de Cópulas entre los resultados del proceso de

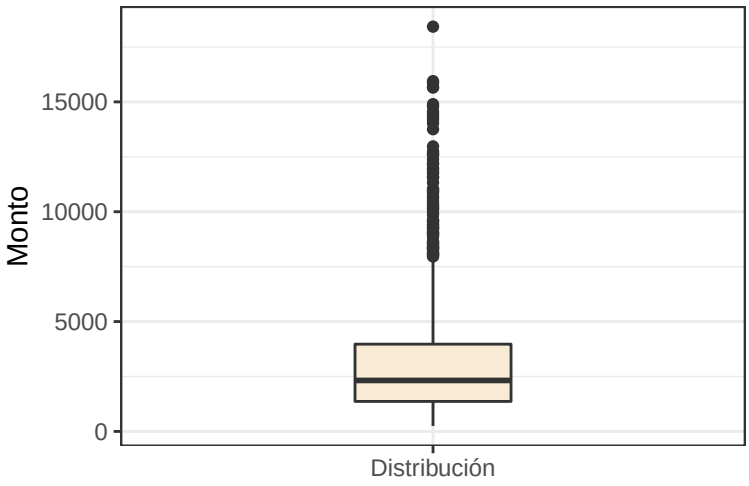
clasificación con el Monto del Crédito, sería importante describir de manera exhaustiva esta variable. A continuación un histograma que muestra la distribución de los montos:



Elaboración propia con datos extraídos de Kaggle

Figura 4.4: Histograma de la variable Monto del Crédito

El histograma muestra una fuerte asimetría hacia su derecha indicando la tendencia en la solicitud de crédito de montos bajos. Esto en efecto se puede ver el siguiente diagrama de caja y bigotes que también busca mostrar de manera más detallada la distribución intercuantílica de los montos:



Elaboración propia con datos extraídos de Kaggle

Figura 4.5: Diagrama de Caja y Bigotes de la variable Monto del Crédito

El gráfico es congruente con lo visto en el histograma y queda aún más en evidencia lo dicho anteriormente pues

alrededor del 50 % de los créditos solicitados fueron por montos menores a los 2500 DM que si comparamos este momento con el monto máximo solicitado de 18424 DM, se puede notar una gran diferencia.

Dada la cantidad de variables categóricas con las que cuenta la base, se realizarán pruebas de tal manera que se puedan distinguir las variables de mayor relevancia y, a partir de las mismas, descartar las menos relevantes. Para ello se utilizará el la prueba de independencia Chi-Cuadrado. Se busca que los $p - values$ sean cercanos a cero con tal de afirmar que las variables son estadísticamente significantes en el estudio.

Estas pruebas mostraron que las variables más significativas son: Acount Balance, Payment Status, Purpose of the credit, Savings/Stock Value, Length of Current Employment, Type Apartment y Most Valuable Asset A partir de estas variables se desarrolla un modelo de regresión logística en el toma un 60 % para entrenamiento y un 40 % para testing.

Para determinar si al solicitante se le dará el crédito, el umbral será del 0.5, por lo que si el resultado de la regresión es mayor a 0.5, se le da el crédito y en caso contrario no. El resultado de la regresión arroja una curva ROC como se muestra a continuación:

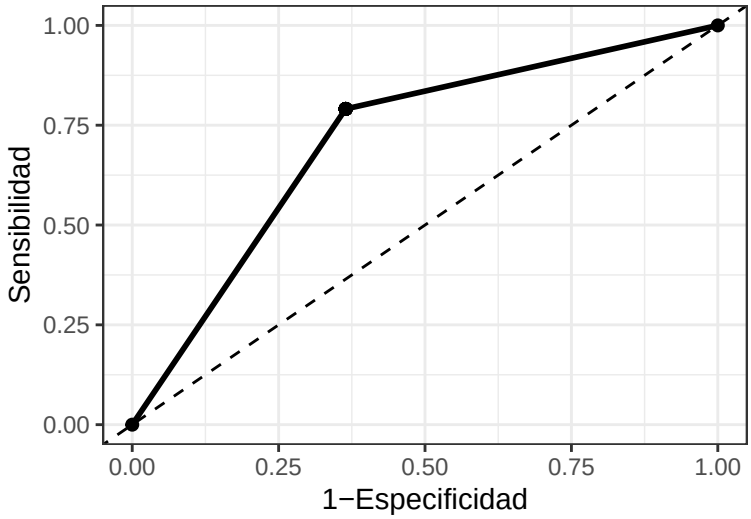


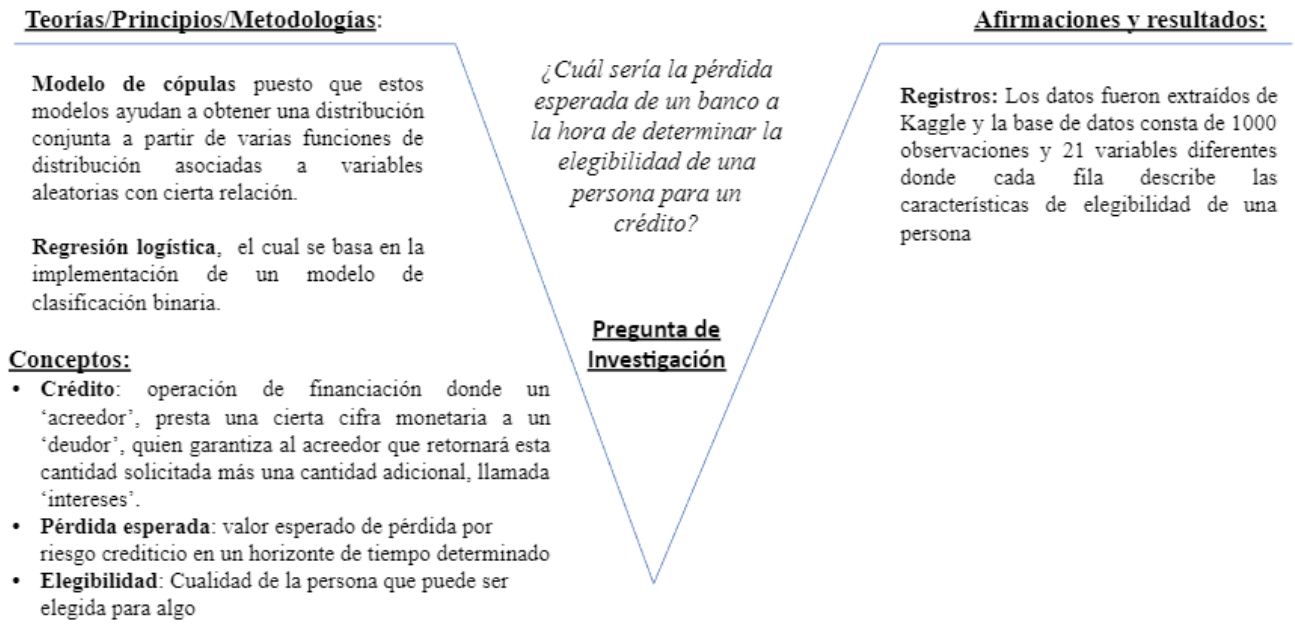
Figura 4.6: Curva ROC

Asimismo, este modelo arroja una precisión aproximada del 76 % con un intervalo de confianza de $]0{,}71, 0{,}80[$. Además se tiene una sensibilidad de 82 %, lo que indica que el modelo es bueno para clasificar a buenos deudores como buenos. Por otro lado, se tiene una especificidad de apenas el 61 % por lo que se concluye que el modelo no es óptimo para clasificar a malos deudores como malos.

Capítulo 5

Anexos

1. UVE de Gowin



Referencias

Cediel, Y. F. (2016). Regular vine cópulas: Una aplicación al cálculo de valor al riesgo. *Revista de Investigación En Modelos Financieros*, *2*, 30–64.

Chitarroni, H. (2002). La regresión logística. *IDICSO*.

Daul, S., De Giorgi, E. G., Lindskog, F., & McNeil, A. (2003). The grouped t-copula with an application to credit risk. *Available at SSRN 1358956*.

Embrechts, P., Resnick, S. I., & Samorodnitsky, G. (1999). Extreme value theory as a risk management tool. *North American Actuarial Journal*, *3*(2), 30–41.

Escarela, G., & Hernández, A. (2009). Modelado de parejas aleatorias usando cópulas. *Revista Colombiana de Estadística*, *32*(1), 33–58.

Geidosch, M., & Fischer, M. (2016). Application of vine copulas to credit portfolio risk modeling. *Journal of Risk and Financial Management*, *9*(2), 4.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in r*. Springer.

Smith, R. (2009). Extreme value theory. *Department of Statistics and Operations Research, University of North Carolina*.