UNIVERSIDAD DE COSTA RICA

ESCUELA DE MATEMÁTICA

DEPARTAMENTO DE MATEMÁTICA PURA Y CIENCIAS ACTUARIALES

DISTRIBUCIÓN DE PÉRDIDAS

Proyecto Distribución de Pérdidas

BITÁCORAS

Grupo 05

Realizado por

Daniel Núñez - B85667

Andrés González - B83413

UNIVERSIDAD DE COSTA RICA

EMat Escuela de Matemática

Índice general

Capítulo 1

Bitácora 1

1. Planificación

Idea principal

La idea inicial que se planteó va de la mano con calcular la pérdida esperada de un Banco a la hora de otorgar créditos basado en el perfil de la persona. Bajo este contexto se pueden encontrar dos tipos de riesgos:

1) El primer riesgo que se distingue es aquel en el que la persona presente una probabilidad alta de que no pague el crédito y que aún así el banco lo apruebe.
2) El segundo haciendo alusión al riesgo de que una persona tenga una probabilidad de pagar de vuelta el crédito pero el banco lo niegue, lo que resulta en una pérdida para la entidad bancaria.

Cuatro formas diferentes de reenplantearse la idea anterior son:

1) ¿Cuál sería la pérdida esperada de un banco a la hora de determinar la elegibilidad de una persona para un crédito?
2) ¿Por qué es importante conocer las probabilidades de pago o de impago de una persona solicitante?
3) ¿Cuáles son los factores que más importan para determinar la probabilidad de impago de un individuo?
4) ¿Cómo se relacionan las pérdidas con un buen sistema de elegibilidad para créditos?

Y las posibles respuestas de estas preguntas serían:

1) La pérdida esperada del banco del banco se puede estimar utilizando los datos de una persona y con esto se elige para crédito o no (Lógica). Para esto se recopilan datos de individuos optando por crédito de manera confidencial (Ética). Lo que facilitará al banco la toma de decisión de otorgar un crédito minimizando sus pérdidas (Emocional).
2) Es importante conocer las probabilidades de pago o impago de una persona pues esto determina si el banco otorga un crédito o no (Lógica). Sin embargo se promueve realizar un estudio aparte para cada individuo pues la decisión final de otorgar el crédito debería ser dado por un experto en el tema y no solo en los resultados del modelo (Ética). Esto le sirve al banco o institución financiera para conocer clientes potenciales para obtener ganancias (Emocional).

3) Los factores que más afectan la probabilidad de impago de un individuo son los determinados por el modelo realizado (Lógica). Esto es de gran utilidad para el banco pues se pueden realizar campañas especializadas en la obtención de clientes que cumplan dichos factores (Emocional). Asimismo se recomienda tomar estos resultados como ayuda a la toma de decisiones no como algo definitivo (Ética).

4) Las pérdidas esperadas por impago de créditos están directamente relacionadas con un sistema robusto de elegibilidad (Lógica), pues si se otorgan créditos a personas con alta probabilidad de pago y se niegan a los que no, minimizaría las pérdidas por impago de la institución (Emocional). Siempre siguiendo los estatutos y leyes para el otorgamiento y denegación de crédtios (Ética).

Base de 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 los datos no están delimitados temporalmente.

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.

Ahora se procede a dar 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 nuevo, 2 si es para un auto de segunda mano, 3 si es para una casa/apartamento 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. Sin embargo, se tratará de reducir la cantidad de variables 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 y Random Forest para determinar la elegibilidad de las personas.

Principios y/o teorías

Para el desarrollo de esta investigación se planea utilizar el modelo de cópulas para relacionar los riesgos mencionados ya que se está buscando correlacionar las pérdidas asociadas a ambos. Para esto resulta útil implementar un modelo de cópulas puesto que como mencionan Escarela y Hernández, estos modelos ayudan a obtener una distribución conjunta a partir de varias funciones de distribución asociadas a variables aleatorias con cierta relación. 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 (Escalera & Hernández, 2009).

Como este trabajo se basa en estimar las pérdidas que puede incurrir la institución financiera considerando los riesgos que conlleva el impago de créditos y así determinar la elegibilidad de posibles clientes, se va a recurrir a la teoría de valores extremos. Esta teoría tiene muchas aplicaciones tanto en ámbitos financieros como en otros, puesto que generalmente se utiliza para modelar eventos catastróficos, con pocas ocurrencias como pueden ser desastres naturales y con más relación a seguros,

estos se utilizan para modelar riesgos que no son tan frecuentes y que si suceden, representan una suma considerable de dinero. Este tipo de teoría se basa en estudiar las distribuciones después de un cierto valor o umbral establecido, por lo que generalmente se basa más en el estudio de las colas de la distribución y las pérdidas que se pueden incurrir por montos que exceden dicho umbral (Smith, 2009).

Fichas bibliográficas

- **Título:** The Grouped t-copula with an Application to Credit Risk
- **Autor(es):** Stéphane Daul, Enrico De Giorgi, Filip Lindskog & Alexander McNeil
- **Año:** 2003
- **Nombre del tema:** Cópulas
- **Forma de organizarlo:**
  -**Cronológico:** 2003
  -**Metodológico:** Grouped t-Copulas
  -**Temático:** Métodos para explicar la depedencia de covariables
  -**Teoría:** Uso de cópulas en el riesgo crediticio
- **Resumen en una oración:**
- **Argumento central:** Lo esencial de este artículo es la justificación del uso en particular de la t-Cópulas agrupadas en problemas de riesgo crediticio.
- **Problemas con el tema:** Los parámetros de los grados de libertad calculados por lo autores fueron considerablemente altos, sin embargo, aún así siguen dando incrementos importantes en el aumento del riesgo comparado con el modelo con cópula Gaussiana.
- **Resumen en un parráfo:** Se 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.

- **Título:** Application of Vine Copulas to Credit Portfolio Risk Modeling
- **Autor(es):** Marco Geidosch & Matthias Fischer
- **Año:** 2016

- **Nombre del tema:** Cópulas
- **Forma de organizarlo:**

  -**Cronológico:** 2015

  -**Metodológico:** Vine Copulas

  -**Temático:** Métodos para explicar la depedencia de covariables

  -**Teoría:** Uso de cópulas para modelar problemas relacionados con riesgo
- **Resumen en una oración:** Dado lo poco flexibles que pueden llegar a ser los modelos convencionales de cópulas se propone el uso de Vine Cópulas para modelar problemas multivariados relacionados con riesgo
- **Argumento central:** Mostrar la superioridad de las Vine Cópulas sobre las cópulas convencionales para la modelación del riesgo
- **Problemas con el tema:** La flexibilidad que presentan las Vine Cópulas viene con la desventaja de que hay una abundante número de estructuras de las que se puede escoger y a priori no se sabe cuál es la que mejor describe los datos.
- **Resumen en un párrafo:** 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 os 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.

- **Título:** Extreme value theory as a risk management tool
- **Autor(es):** Paul Embrechts, Sidney I. Resnick y Gennady Samorodnitsky
- **Año:** 1999
- **Nombre del tema:** Teoría de valores extremos
- **Forma de organizarlo:**

  -**Cronológico:** 1999

  -**Metodológico:** Distribuciones con valores extremos

  -**Temático:** Teoría de valores extremos y aplicaciones

  -**Teoría:** Teoría de valores extremos
- **Resumen en una oración:** Teoría de valores extremos con su fundamento teórico y las distribuciones usuales para esta metodología junto con sus aplicaciones.
- **Argumento central:** Base teórica y posibles aplicaciones a la teoría de valores extremos en manejo de riesgo.

- **Problemas con el tema:** El autor no menciona problemas.
- **Resumen en un párrafo:** Se presentan ejemplos de como eventos catastróficos costosos de predecir como terremotos, incendios, entre otros, que supusieron un gasto enorme no presupuestado por parte de las compañías aseguradoras. Es por esto que se empezaron a realizar estudios y métodos de preveer escenarios como estos para estar preparado a asumir estas pérdidas, como el VaR. Estos eventos extremos en el mundo de las finanzas se pueden ver como pérdidas masivas en industria o caídas considerables en mercados bursátiles, para los cuales la teoría de valores extremos provee metodologías robustas y con fundamento para cuantificar estos eventos y sus consecuencias. Es por esto que se explican diferentes herramientas que provee la teoría de valores extremos en el ámbito del manejo de riesgo.

2. Escritura

La pregunta a desarrollar durante la investigación es, ¿Cuál sería la pérdida esperada de un banco a la hora de determinar la elegibilidad de una persona para un crédito? por lo que se planea primeramente realizar un modelo de clasificación para determinar si una persona es elegible a crédito o no. Tomando en cuenta los dos riesgos asociados de una mala clasificación. Una vez con esto, y con los datos esocgidos realizar un modelo de estimación de pérdidas combinando ambos riesgos utilizando un modelo basado en cópulas y a su vez realizar un análisis tomando en cuenta la teoría de valores extremos para aquellos riesgos que supongan una pérdida mayor a un cierto umbral.

En cuanto a lo estudiado hasta el momento, se tiene una base de datos con información sobre la elegibilidad de personas a un crédito bancario. Lo primero que se piensa implementar es un modelo de clasificación como regresión logística o bosques aleatorios (Random forests) y de esta manera asociar el riesgo de clasificar un individuo como no elegible cuando este si debería ser elegible y viceversa. Esto debido a que los riesgos mencionados representarían pérdidas al banco, por lo que para encontrar una distribución de pérdidas que contemple ambos se debe de alguna manera encontrar una distribución conjunta.

Para resolver este problema de encontrar la distribución conjunta se ha investigado sobre los modelos de cópulas que ayudan a correlacionar estas distribuciones individuales univariadas en una distribución multivariada, junto con las relaciones y dependencias de estas variables entre sí. Sin embargo, la mayoría de distribuciones de cópulas están hechas para tomar en cuenta solo dos variables y se menciona que para más variables estas distribuciones presentan ciertas limitaciones. Es por esto que se ha indagado en otro tipos de modelos como el de vine-copulas que funciona para más de dos variables. Sin embargo, no es la única metodología de cópulas estudiadas, puesto que también se investigó los modelos de t-copulas y t-copulas agrupadas que permiten más flexibilidad a la hora de estructurar la dependencia entre las variables. Aun así, todavía no se tiene cual de los modelos implementar aunque se está considerando realizar varias para poder realizar una comparación con los resultados obtenidos por los distintos modelos y tener conclusiones más completas.

Como ya se mencionó, se planean realizar modelos para encontrar la distribución de pérdidas del banco conjunta, pero el otro tema a considerar es como ajustar esta distribución a posibles pérdidas altas que se pueden dar por otorgar un crédito a individuos que no van a incurrir en el pago. Es por esto, que se planea utilizar la teoría de valores extremos y distintas metodologías de esta teoría para poder realizar el análisis de las pérdidas esperadas en estos casos extremos. Por lo tanto este trabajo se piensa estructurar en tres grandes pasos, el primero el de realizar un modelo de elegibilidad de crédito, después encontrar una distribución de pérdidas conjunta para el banco y finalmente complementar esta estimación de pérdidas en caso de valores extremos.

Capítulo 2

Bitacora 2

1. Fichas bibliográficas nuevas

- **Título:** Modelado de parejas aleatorias usando cópulas
- **Autor(es):** Gabriel Escarela & Angélica Hernández
- **Año:** 2009
- **Nombre del tema:** Modelos de cópulas
- **Forma de organizarlo:**
  -**Cronológico:** 2009
  -**Metodológico:** Modelos de cópulas
  -**Temático:** Cópulas bivariadas
  -**Teoría:** Teoría de cópulas
- **Resumen en una oración:** Definición formal de las cópulas bivariadas junto con ejemplos de su utilidad
- **Argumento central:** Teoría de cópulas y su construcción teórica
- **Problemas con el tema:** Solo funcionan cuando se tienen 2 variables.
- **Resumen en un párrafo:** 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.

- **Título:** An Introduction to Statistical Learning with Applications in R - Chapter 4: Classification

- **Autor(es):** Gareth James, Daniela Witten, Trevor Hastie & Robert Tibshirani
- **Año:** 2021
- **Nombre del tema:** Métodos de clasificación
- **Forma de organizarlo:**
  -**Cronológico:** 2021
  -**Metodológico:** Métodos de clasificación
  -**Temático:** Regresión logística para clasificación
  -**Teoría:** Realizar modelos de clasificación
- **Resumen en una oración:** Implementar métodos de clasificación y sus pruebas para determinar la calidad del modelo.
- **Argumento central:** Métodos de clasificación y sus pruebas con ejemplos programados en R.
- **Problemas con el tema:** No es un problema como tal, pero se menciona que hay que tener cuidado pues hay varios métodos de clasificación y tal vez el escogido no sea el más conveniente.
- **Resumen en un párrafo:** 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. También se mencionan 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.

- **Título:** La regresión logística
- **Autor(es):** Horacio Chitarroni
- **Año:** 2002
- **Nombre del tema:** Regresión logística
- **Forma de organizarlo:**
  -**Cronológico:** 2002
  -**Metodológico:** Métodos de clasificación
  -**Temático:** Regresión logística para clasificación
  -**Teoría:** Realizar modelos de clasificación
- **Resumen en una oración:** Implementar métodos de clasificación con ejemplos e interpretación de resultados
- **Argumento central:** Método de regresión logística y su implementación

- **Problemas con el tema:** El autor no menciona problemas.
- **Resumen en un párrafo:** El modelo de regresión logística es 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. 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.

---

2. Ordenamiento de literatura

La literatura se va a ordenar de manera metodológica por lo que se tienen dos grupos, la literatura ligada a los métodos de clasificación, y la literatura enfocada en cópulas.

Cuadro 2.1: Clasificación Binaria

Tipo de grupo	Nombre del grupo	Nombre del tema	Título	Año	Autor(es)
Metodológico	Regresión logística	Métodos de clasificación	An Introduction to Statistical Learning	2021	James et al.
Metodológico	Regresión logística	Regresión logística	La regresión logística	2002	Horacio Chitarroni

Cuadro 2.2: Cópulas

Tipo de grupo	Nombre del grupo	Nombre del tema	Título	Año	Autor(es)
Metodológico	Cópulas	Grouped t-Copulas	The Grouped t-copula with an Application to Credit Risk	2003	Daul et al.
Metodológico	Cópulas	Vine Copulas	Application of Vine Copulas to Credit Portfolio Risk Modeling	2016	Geidosch & Fischer
Metodológico	Cópulas	Cópulas bivariadas	Modelado de parejas aleatorias usando cópulas	2009	Escarela & Hernández

3. Análisis estadístico

Dado que la mayor'ia de variables son categ'oricas, primero se realizar'a un proceso de depuración de la base con el fin de reducir el n'umero 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 tabla de frecuencias según el Próposito del préstamo.

Cuadro 2.3: Distribución de la variable Próposito del Crédito

Propósito	Cantidad de observaciones
1	284
2	364
3	12
4	340

Note:
Elaboración propia con datos extraídos de Kaggle

La tabla 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. Visto de manera gráfica:

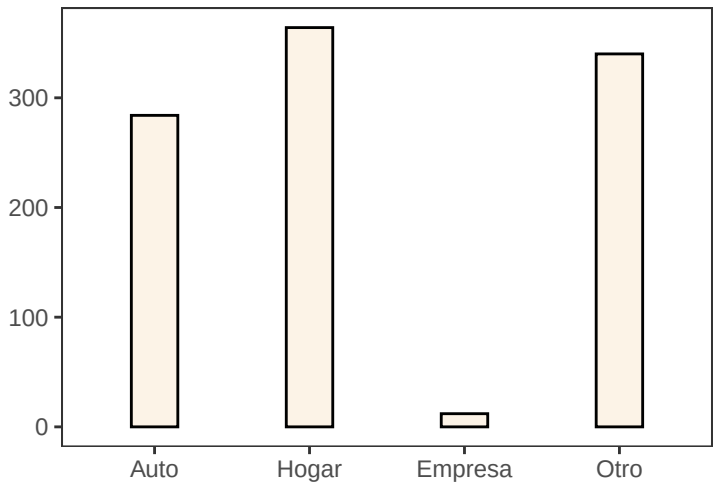
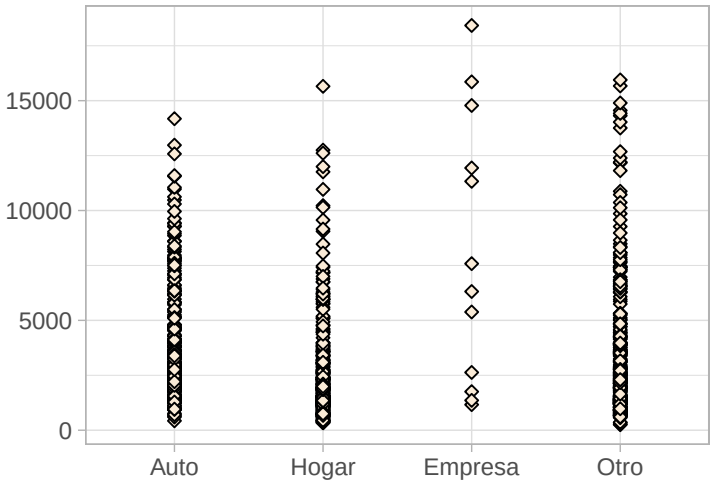


Figura 2.1: Distribución de la variable Próposito del crédito

Asimismo, en el siguiente gráfico se muestra la relación que existe entre el monto del crédito según su razón, 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 la prósito "Empresa" con alguna otra, pues es de esperarse que haya información importante que pueda revelar.



Elaboración propia con datos extraídos de Kaggle

Figura 2.2: Distribución del Monto del Crédito según su próposito

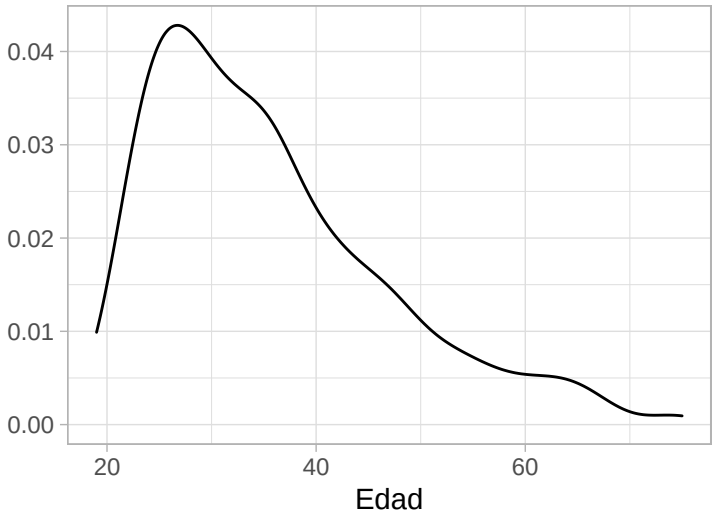
Haciendo una análisis similar para las edades, se llega a la conclusión de que existe una tendencia mayor en las personas cuyas edades estén en el rango de 20 a 40 años a solicitar préstamos como lo muestra la siguiente tabla:

Cuadro 2.4: 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 derecho 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 2.3: 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 2.5: 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, sin embargo, el siguiente cuadro muestra información sobre las variables de carácter continuo:

Cuadro 2.6: Resumen de 5 números

	Min	Q1	Mediana	Q3	Max
Monto del crédito	250	1365.5	2319.5	3972.25	18424
Duración en meses del crédito	4	12.0	18.0	24.00	72
Edad	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 de los solicitantes con la variable Monto del Crédito, sería importante describir de manera exhaustiva esta variable. A continuación un histograma que muestrala distribución de los montos:

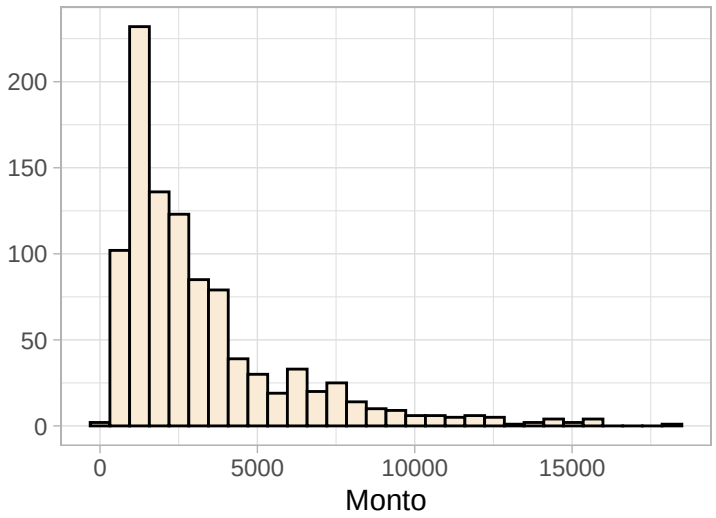


Figura 2.4: Histograma de la variable Monto del Crédito

El histograma muestra una fuerte asimetría hacia su derecho 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:

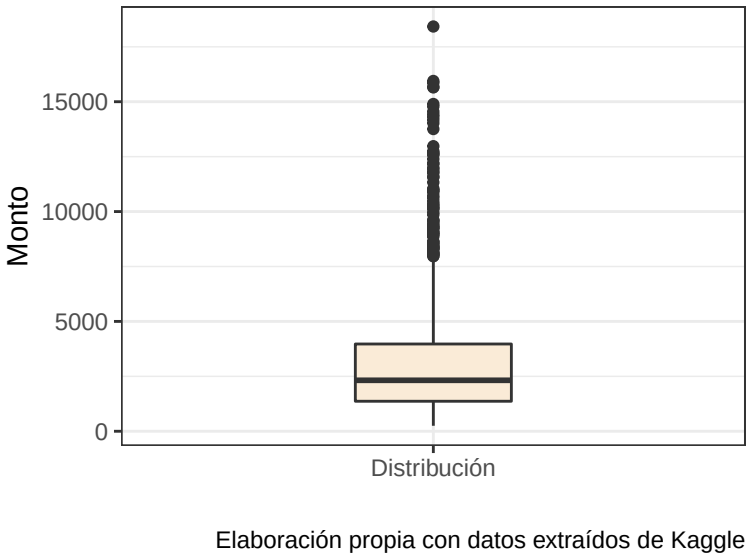


Figura 2.5: Diagrama de Caja y Bigotes de la variable Monto del Crédito

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, Savings/Stock Value, Length of Current Employment, Type Apartment, Most Valuable Asset, Concurrent Credits. 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:

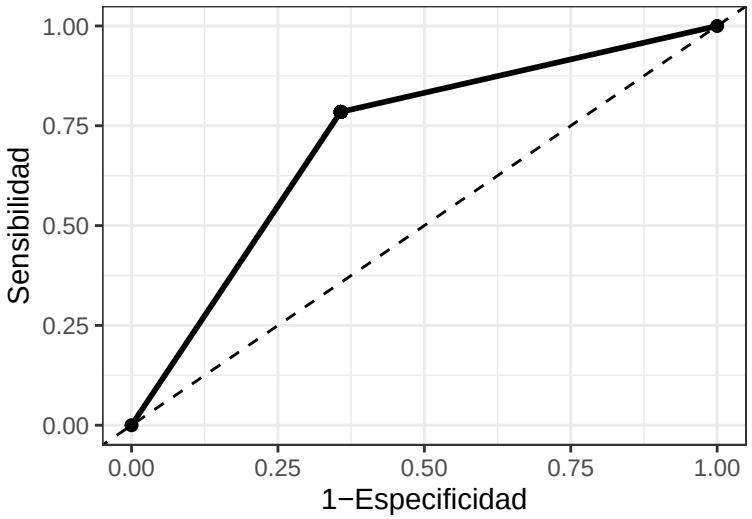


Figura 2.6: Curva ROC

4. Fichas de Resultados

- **Nombre Hallazgo:** Importancia de la re-categorización de la variable Purpose
- **Resumen en una oración:** En un estudio similar, la categorización de la variable es agrupada de manera diferente y se llega a una precisión del modelo menor a la que se llega de la forma propuesta
- **Principal característica:** La forma en la que se agrupen las diferentes categorías influye significativamente en el modelo
- **Problemas o posibles desafíos:** Para la segmentación se uso un criterio subjetivo por lo que si se probarán diferentes combinaciones la significancia de la variable puede aumentar o disminuir.
- **Resumen en un parráfo:** En un estudio en el que se usa la base de datos estudiada, se propone una categorización que toma el valor de 1 si es para un auto nuevo, 2 para auto de segunda mano, 3 si es relacionado al hogar y 4 para cualquier otra razón. La categorización propuesta consistió en combinar las categorías 1 y 2 (referente a la compra de un auto) y dejar una categoría para todos aquellos créditos destinados al sector empresarial. Se tomó esa decisión debido a que tanto la categoría 1 y 2 comparten una misma naturaleza la cual es la compra de un auto. Como consecuencia, mediante la aplicación de las pruebas Chi-Cuadrado se llegó a que la variable Purpose no era significante por lo que se decidió descartar del modelo de regresión logística. Cabe resaltar que en el estudio consultado, sí toman en cuenta esta variable para la regresión y el resultado es una precisión menor a la encontrada sin usar la variable.

- **Nombre Hallazgo:** Pruebas de significancia sobre las variables usadas
- **Resumen en una oración:** La base presenta múltiples variables que se buscan reducir mediante la aplicación de pruebas con el fin de llegar a un modelo más optimizado.
- **Principal característica:** La realización de la Prueba Chi-Cuadrado y la Prueba t arrojan resultados importantes a la hora de la selección de variables para la regresión.
- **Problemas o posibles desafíos:** Las variables continuas no cumplen con los supuestos de normalidad para aplicar la Prueba t por lo que no se pudo verificar su significancia a la hora de determinar si un solicitante es buen deudor o no.
- **Resumen en un parráfo:** La base cuenta con 20 diferentes variables predictoras. Se busca reducir esta cantidad de variables mediante la aplicación de pruebas como la Chi-Cuadrado y la Prueba t. Para la Prueba t se necesita un supuesto de normalidad en la distribución de los datos lo cual no cumplieron las variables de carácter continuo por lo que se descartó esta prueba para determinar la significancia de estas variables cuantitativas. Por otro lado, mediante la aplicación de la Prueba Chi-Cuadrado se descartaron un total de 10 variables que no parecían tener significancia con la elegibilidad de una persona solicitante. Entre las variables descartadas destacan si la persona tenía un teléfono o no (Telephone), el estado civil de la persona, si la persona era extranjera o no, la duración que lleva el solicitante viviendo en su ubicación actual, entre otras.

- **Nombre Hallazgo:** Resultados de la Regresión Logística
- **Resumen en una oración:** Mediante la aplicación de la regresión logística no solo se logró el objetivo de clasificar al solicitante como buen deudor o no, sino que también se obtuvieron las variables más significantes que determinan a un solicitante como buen deudor.
- **Principal característica:** Para la regresión se usaron las variables que fueron significativas en las pruebas hechas anteriormente. Inicialmente se obtuvieron resultados bastantes decentes.
- **Problemas o posibles desafíos:** Un desafío sería aumentar aún más la precisión del modelo.
- **Resumen en un parráfo:** Se desarrollaron una regresión logística que arroja un área debajo de la curva de alrededor del 70 %. La precisión es de un 75 % y la sensibilidad del 82 % aproximadamente. Asimismo, entre las variables más significantes de este modelo destaca el balance actual, el estado de pago actual del pasado crédito solicitado, la duración del crédito solicitado y el tiempo que lleva el solicitante en su actual empleo. Por otro lado, entre las variables que arrojan menos significancia destacan la edad del solicitante, el monto del crédito, el tipo de apartamento u hogar que tenga y si tiene créditos en otros bancos o instituciones. A partir de estos resultados se comenzará a usar modelos de cópulas para verificar la independecia existente entre la clasificación recién hecha y las variables significativas del modelo.

5. Enlaces de Literatura

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.

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.

Tomando en consideración ambos enfoques se puede ver que con los resultados obtenidos en el modelo realizado se logra desarrollar una regresión logística que arroja un área debajo de la curva de alrededor del 70 %. La precisión es de un 75 % y la sensibilidad del 82 % aproximadamente. Asimismo, entre las variables más significantes de este modelo destaca el balance actual, el estado de pago actual del pasado crédito solicitado, la duración del crédito solicitado y el tiempo que lleva el solicitante en su actual empleo. Por otro lado, entre las variables que arrojan menos significancia destacan la edad del solicitante, el monto del crédito, el tipo de apartamento u hogar que tenga y si tiene créditos en otros bancos o instituciones. A partir de estos resultados se comenzará a usar modelos de cópulas para verificar la independecia existente entre la clasificación recién hecha y las variables significativas 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.

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

Bitacora 3

1. Ajuste del modelo

Para poder ajustar un modelo de cópulas bivariadas, primero se tienen que saber las funciones de distribución marginales univariadas para cada variable de estudio. En este se necesitan encontrar dos de estas marginales pues el modelo de cópulas a implementar va a considerar como dos variables, el monto del crédito solicitado y el valor de la regresión logística asociado a la probabilidad de elegibilidad asignado con el modelo.

Por lo que primeramente se van a mostrar los gráficos de la distribución empírica de las variables de estudio, mediante histogramas.

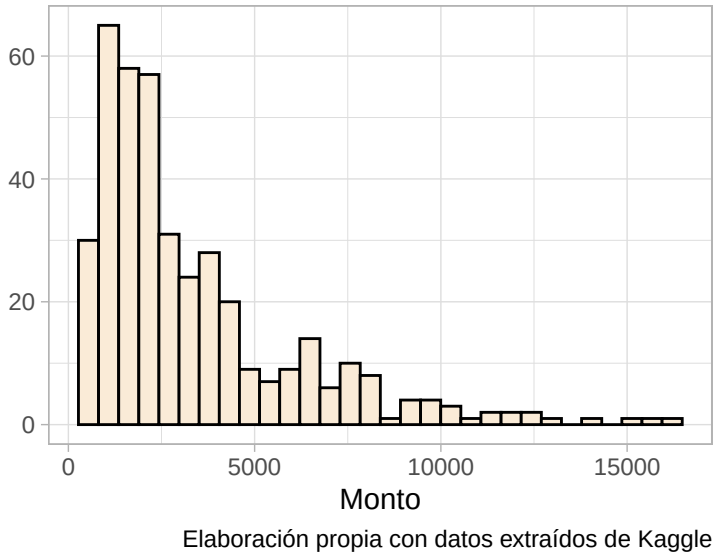


Figura 3.1: Histograma de los montos de crédito de la base de prueba

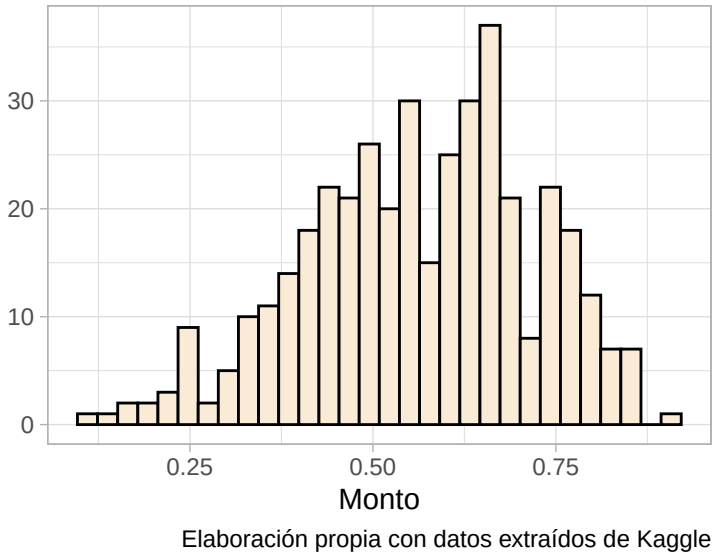


Figura 3.2: Histograma de los probabilidades de elegibilidad

Con estas figuras se pueden observar ciertas tendencias de como se distribuyen las variables, lo que permite buscar ajustar una distribución paramétrica de acuerdo a su forma. Bajo este hilo, se sigue que el monto del crédito es bastante asimétrica hacia la derecha, mientras que la probabilidad de elegibilidad sigue una tendencia más simétrica.

Para el siguiente desarrollo se va a utlizar el lenguaje de programación R para realizar los modelos, cálculos y estimaciones. Primeramente, se va van a realizar distintos modelos para las distribuciones marginales de las variables de estudio donde por métodos como AIC y estimación de parámetros por máxima verosimilitud. Bajo esta metodología, se consigue que la mejor distribución que se ajusta para el monto del crédito es una distribución $Gamma(1{,}8115402, 1849)$ bajo una parametrización de forma y escala. Mientras que, al realizar este ajuste con la probabilidad de elegibilidad el modelo que mejor ajusta es una distribución $N(0{,}5625845, 0{,}1549248)$.

Una vez, con estas distribuciones marginales se procede a modelar la correlación entre estas variables mediante cópulas para lograr estimar una función de distribución bivariada considerando las variables de estudio.

Como se mencionó en el marco teórico, el coeficiente de correlación que se va a utilizar es el de Kendall, conocido como tau de Kendall. Para los datos empíricos se obtiene que $\hat{\tau}_K = -0{,}3$ lo que indica una correlación negativa entre las variables, es decir que conforme el monto del crédito aumenta, las probabilidades de elegibilidad disminuyen. Con esta información, se procede a escoger una función de cópulas que más se ajuste a los datos observados para poder construir su función de distribución. Bajo métodos de escogencia de una función de cópulas como el AIC y la estimación de parámetros por medio de máxima verosimilitud se llega a un modelo con el mejor ajuste que sería una cópula de Frank con parámetro $\theta = -3{,}6$ y un $\tau_K = -0{,}36$, lo cual es bastante cercano a la correlación empírica calculada anteriormente. Es decir, se obtiene una función de cópulas (Arquimediana) que mantiene la correlación de los datos, lo cual también es lo buscado.

Una vez con este modelo, y las marginales univariadas calculadas al principio, se procede a construir la función de distribución y densidad bivariada. Que es lo que se planea contestar con la pregunta de investigación planteada. Para su visualización se utilizan técnicas de graficación en tres dimensiones en dos dimensiones por lo que el siguiente gráfico muestra el contorno de la densidad bivariada.

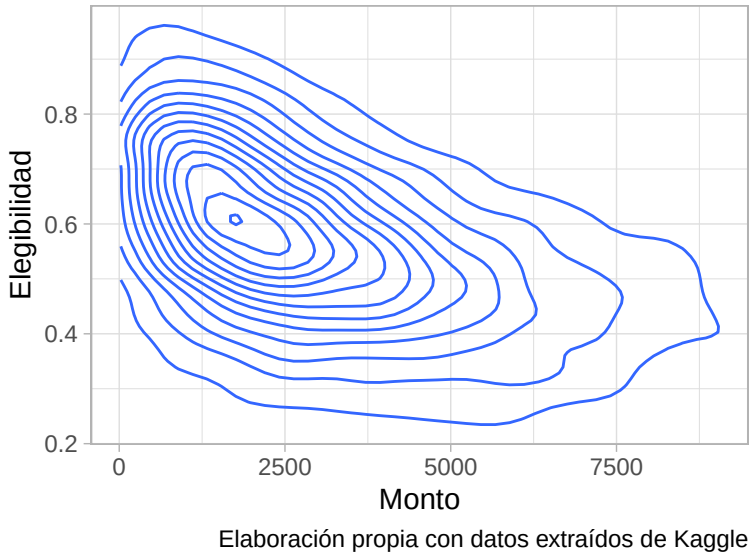


Figura 3.3: Contorno de la función de densidad bivariada

2. Diagnósticos del modelo

Como se mencionó anteriormente, la escogencia de la cópula fue mediante AIC sin embargo para poder determinar que tan bien se ajusta una cópula, se utilza el método de *Goodness-of-fit test* o *Gof Test*. El cual se basa en calcular el estadístico de Cramér-von-Mises o S_n. Por lo que, ne la siguiente tabla se muestran los resultados de las pruebas realizadas.

Cuadro 3.1: AIC de las diferentes familias de cópulas

Cópula	AIC
Frank	-63.46017
Normal	-55.04813
t	-53.04770
Clayton	-10.97244

Cuadro 3.2: AIC de las diferentes familias de cópulas

Cópula	Estadístico Cramér-von-Mises
Frank	0.0266808
Normal	0.0269870
t	0.0426066
Clayton	0.6172515

Como se puede observar, tanto el AIC como el S_n dan más bajos en el caso de una cópula de la familia Frank. Por lo que debido a estas pruebas se llega a la conclusión que el mejor modelo de cópulas para las variables de estudio es una cópula de Frank. También, vale la pena recalcar que las peores métricas se obtienen considerando un modelo de cópula de Clayton, lo cual corrobora el hecho

que si se presenta una correlación negativa este modelo es peor y en algunos algoritmos ni se considera intentar ajustar con este modelo.

Finalmente, también se realizó una serie de datos simulados con la función de distribución bivariada contra los datos reales para de manera visual poder analizar tendencias y comportamientos entre el modelo y la realidad.

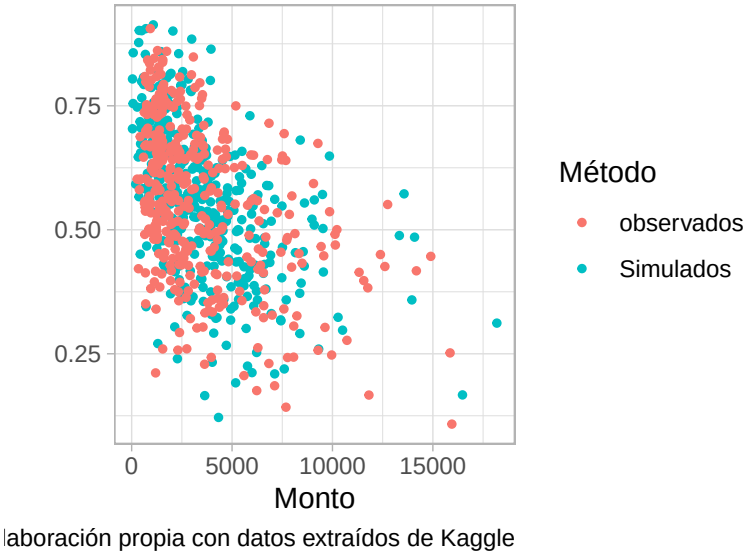


Figura 3.4: Datos simulados con la función de distribución vs. datos observados

3. Fichas de resultados

Hallazgo # 1

- **Nombre del hallazgo:** Correlación empírica negativa entre el monto del crédito y la probabilidad de elegibilidad
- **Resumen en una oración:** Se muestra una correlación negativa lo que indica que ambas variables se mueven en direcciones opuestas.
- **Principal característica:** La correlación se calculó utilizando la asociación de Kendall, también conocido como tau de Kendall
- **Problemas con el tema:** Dependiendo del método con el que se calcule la correlación, la misma puede variar. Por ejemplo, para el caso de la correlación de Pearson el resulado es de -0.46, mientras que con Kendall es de -0.30 aproximadamente.
- **Resumen en un párrafo:** La correlación empírica de ambas variables es negativa y da un valor de -0.30 aproximadamente. Esta correlación se calculó con el método del tau de Kendall ya que es el método más popular para comparar dependencia entre modelos de cópulas. Este resultado indica que si el monto a solicitar por parte del cliente es muy alto, disminuyen las probabilidades de ser elegido. Lo mismo pasa de manera análoga cuando el monto del crédito es bajo, en donde las probabilidades de ser elegido aumentan. Vale la pena mencionar que no se tiene una correlación perfecta, esto se debe a que hay muchos más factores que influyen en la elegibilidad de los solicitantes, no solo el monto.

Hallazgo # 2

- **Nombre del hallazgo:** Una correlación negativa disminuye el número de cópulas que podrían hacer un buen ajuste.
- **Resumen en una oración:** Hay ciertos tipos de cópulas que están hechas para modelar únicamente correlaciones positivas, por lo que una relación negativa hace que se reduzca los tipos de cópulas a escoger
- **Principal característica:** La correlación negativa no es compatible con las cópulas de Clayton, Gumbel, Joe, BB1, BB6, BB7 y BB8.
- **Problemas con el tema:** El hecho de que se esté reduciendo la variedad de cópulas de dónde escoger para ajustar a los datos no es lo óptimo ya que el modelo ajustado puede que no vaya a ser el mejor.
- **Resumen en un párrafo:** La asociación negativa indicada anteriormente funciona como una especie de filtro a la hora de escoger la cópula que mejor ajusta los datos. Esto sucede porque no todas las cópulas pueden modelar relaciones negativas. Esto trae ciertas desventajas ya que se reduce el número de cópulas de dónde escoger para modelar la distribución conjunta.

Hallazgo # 3

- **Nombre del hallazgo:** La Frank Copula mostró ser la mejor cópula para los datos
- **Resumen en una oración:** Se probaron múltiples cópulas entre las cuales distinguen la gaussiana, la t-student,la Frank y la Gumbel.
- **Principal característica:** El AIC fue el método usado para determinar la mejor cópula.
- **Problemas con el tema:** La cópula que muestra ser la mejor depende directamente de las probabilidades calculadas con la regresión logística, es decir, si se decide volver a predecir las probabiliades arrojadas por la regresión logística se tendría un nuevo vector de probabilidades por lo que probablemente el modelo de cópula óptimo pueda variar.
- **Resumen en un párrafo:** Se probaron múltiples combinanciones de cópulas para determinar cuál era la que mejor se adaptaba a los datos. Entre las principales usadas se distinguen la gaussiana, la t-student y la Frank. La Frank se escogió como la mejor con un AIC de -63.46, mientras que las demás tuvieron AIC de -55.04 y -53.04 respectivamente.

Hallazgo # 4

- **Nombre del hallazgo:** Relación entre la correlación empírica y la teórica calculada por el modelo
- **Resumen en una oración:** La correlación empírica era de -0.30 aproximadamente mientras que la correlación teórica calculada por el modelo de cópula de Frank es de -0.36.
- **Principal característica:** El modelo mantiene la correlación de los datos observados.

- **Problemas con el tema:** Como se había mencionado anteriormente, hay múltiples formas de calcular la correlación por lo que la relación entre el estadístico teórico y empírico depende y varía con la forma en la que se calcula la correlación.
- **Resumen en un párrafo:** Otra buena forma de medir qué tan bien ajustado está el modelo a los datos es comparar sus correlaciones. Lo ideal sería que la dependencia teórica calculada por el modelo sea bastante similar a la dependencia empírica. En este caso, se da que la empírica corresponde a un valor aproximado de -0.30 mientras que la correlación teórica calculada por el modelo de cópula de Frank es de -0.36. Hay una diferencia de 0.06 puntos porcentuales entre ambas correlaciones, pero a groso modo, ambos explican el fenómeno de la asociación entre ambas variables bastante parecido indicando el buen ajuste del modelo.

Hallazgo # 5

- **Nombre del hallazgo:** Distribución marginal de la variable de Monto de Crédito
- **Resumen en una oración:** La distribución que mejor se ajusta a la variable Monto de Crédito es una distribución Gamma.
- **Principal característica:** La distribución se ajusta como una $Gamma(1{,}8115,\ 1849)$
- **Problemas con el tema:** Este ajuste se realizó utilizando métodos estadísticos y podría tener errores por el método de cálculo.
- **Resumen en un párrafo:** Para poder construir las funciones de probabilidad bivariadas aparte de la función de cópulas, se necesitan las funciones de distribución para cada variable, a estas funciones se les llama marginales. La distribución que mejor se ajusta al monto de crédtio bajo el método de AIC y de estimación de parámetros mediante máxima verosimilitud es una distribución $Gamma(1{,}8115, 1849)$ bajo una parametrización de forma y escala.

Hallazgo # 6

- **Nombre del hallazgo:** Distribución marginal de la variable de Elegibilidad
- **Resumen en una oración:** La distribución que mejor se ajusta a la variable Elegibilidad es una distribución Normal.
- **Principal característica:** La distribución se ajusta como una $N(0{,}5626,\ 1549)$
- **Problemas con el tema:** Este ajuste se realizó utilizando métodos estadísticos y podría tener errores por el método de cálculo.
- **Resumen en un párrafo:** Una vez con la distribución estimada de la variable para el monto de crédito, se procede a realizar un ajuste similar esta vez para la variable restante que es la elegibilidad de la persona. Así, la distribución que mejor se ajusta para la elegibilidad es una distribución $N(0{,}5626, 0{,}1549$. Por lo que ahora si se puede proceder a construir las funciones de probabilidad considerando las marginales y la función de cópulas estimadas.

Hallazgo # 7

- **Nombre del hallazgo:** La cópula de Frank consigue las mejores métricas
- **Resumen en una oración:** Bajo métodos de diagnóstico como el AIC y S_n se escoge la cópula de Frank como más adecuada.
- **Principal característica:** La cópula del modelo pasa los diágnosticos y es la que mejor se ajusta.
- **Problemas con el tema:** No se encontraron muchos diágnosticos para modelos de cópulas, por lo que los resultados se basan solo en AIC y el test S_n.
- **Resumen en un párrafo:** Otro método en el cual se basó para determinar cual cópula es la que mejor se ajusta, se utilizó un *Goodness-of-fitness test*, el cual se basa en el cálculo del estadístico de Cramér-von-Mises, también conocido como S_n. Este estadístico se puede interpretar similar al AIC en el sentido que entre menor es el valor se obtiene un mejor modelo, pues el *valor-p* asociado va a ser mayor y esto resulta en que no haya evidencia suficiente para rechazar la hipótesis nula, que es lo que se está buscando. Por lo que bajo esta prueba también se obtiene como resultado que la cópula que mejor se ajusta es una de Frank. Similarmente se puede ver como la evidencia estadística sugiere lo hallado con respecto a que si se tiene una correlación negativa la función de Clayton no es recomendable y esto se puede observar en como este modelo es el que peor métricas obtiene de los cuatro modelos.

Hallazgo # 8

- **Nombre del hallazgo:** Las funciones de distribución y densidad bivariadas
- **Resumen en una oración:** Se encuentra la función de distribución y densidad bivariadas considerando las marginales y la cópula estimada.
- **Principal característica:** La forma de las funciones de probabilidad bivariada que mejor ajusta a los datos observados.
- **Problemas con el tema:**Estas funciones por construcción dependen de las marginales y la función de cópulas estimadas por lo que cualquier error que se pueda haber presentado en esas estimaciones pueden afectar considerablemente estas funciones.
- **Resumen en un párrafo:** Una vez con las marginales y la función de cópulas estimadas se procede a la construcción de funciones de densidad y distribución de las funciones. Estas funciones son las que se buscan para contestar la pregunta de investigación por lo que es uno de los resultados más importantes del trabajo. Con estas funciones se pueden crear simulaciones de datos nuevos y a su vez permite una manera de representar las variables de estudio. Estas funciones al ser bivariadas representan funciones de dos dimensiones por lo que su resultado o probabilidad se da en una tercera dimensión. Esto complica la visualización usual de gráficos de densidad y distribución. Para esto se procede a utilizar técnicas como gráficos de contornos para facilitar la interpretación de estos resultados. A su vez, estas funciones también permiten calcular medidas de riesgo como el VaR o CVaR, que es lo que se piensa estimar en futuras bitácoras.

---

4. Estructra del proyecto

A continuación, se clasifican los principales elementos de la investigación en primarios y secundarios de acuerdo con su peso para contestar la pregunta de investigación:

Elementos del reporte	
Primarios	**Secundarios**
Justificación de la investigación	Idea intuitiva de la regresión logística
Análisis de datos	Idea intuitiva de las cópulas
Cópulas	Regresión Logística
Frank Copulas	Resultados de Regresión Logística
Hallazgos 3 y 4	Hallazgos 1 y 2
Hallazgo 7 y 8	Tau de Kendall
	Hallazgo 5 y 6

Cuadro 3.3: Clasificación de los elementos principales de la investigación

Asimismo, el ordenamiento de la literatura de acuerdo con la clasificación ya hecha:

Ordenamiento de la literatura	
Sección	**Temas a tratar**
Introducción	- Justificación de la investigación (P) - Idea intuitiva de la regresión logística (S) - Idea intuitiva de las cópulas (S)
Metodología	- Descripción y análisis de los datos (P) - Regresión logística (S) - Cópulas (P) - Frank Cópulas (P) - Tau de Kendall (S)
Resultados	- Resultados de la regresión (S) -Hallazgo 1 y 2 (S) -Hallazgo 3 y 4 (P) -Hallazgo 5 y 6 (S) -Hallazgo 7 y 8 (P)

Cuadro 3.4: Ordenamiento de la literatura de acuerdo a la clasificación hecha

5. Introducción

Día con día personas se presentan a entidades bancarias con el fin de solicitar préstamos. A pesar de lo que se pueda pensar, esto es un negocio muy rentable para los bancos, sin embargo, el otorgamiento de esos préstamos debe realizarse de manera responsable y considerando múltiples criterios. Y dado el contexto actual en donde se está atravesando por una crisis económica importante en todo el mundo, la solicitud de préstamos se ha vuelto una forma común de financiamiento para afrontar la crisis. Este puede ser uno de los principales problemas a lo que se enfrentan las entidades financieras, determinar a cuáles personas es rentable prestarles dinero.

Este proyecto se encarga de estudiar ese fenómeno con herramientas estadísticas como lo es la regresión logística, sin embargo, para dar una respuesta aún más completa hará faltar hacer de otros métodos estadísticos. Esto debido a que este tipo de modelos por lo general no contemplan si existe relación entre las variables estudiadas. Esa posible dependencia hará que el modelo se vea

afectado. Mediante este proyecto se propone realizar un análisis de dependencia adicional para intentar relacionar esa probabilidad de elegibilidad del individuo con otras variables de interés como el monto del crédito. De esta forma, se distinguen dos principales etapas en el trabajo, la primera es la construcción de esas probabilidades de elegibilidad mediante una regresión logística. Para esta etapa, se usaron datos de un banco alemán que mantiene su nombre en anonimato y hay muchas variables que este banco considera de interés para otorgar un préstamo dentro de las cuales se destaca la edad, el monto del préstamo, la cantidad de préstamos que tiene, entre otras. El análisis descriptivo de las variables se realiza más adelante en la investigación.

Una vez calculadas esas probabilidades de elegibilidad, se procede con la segunda etapa del trabajo, la cual es la construcción de una función marginal que explique esos datos. La idea principal sería comparar esta distribución marginal con la distribución marginal de otras variables de interés como el monto del crédito. Para ello, se pretende la creación de una sola función de distribución que explique la relación y dependencia entre ambas variables. Es precisamente en esta etapa en la que usan cópulas para la creación del producto esperado que vendría siendo la distribución conjunta.

La construcción de esta distribución conjunta es necesaria para extraer conclusiones en donde se esté contemplando la relación que existe entre ambas variables. La idea del trabajo es extraer estas conclusiones a partir de la probabilidad de elegibilidad y su relación con el monto del crédito.

6. Metodología

Para un primer acercamiento a contestar la pregunta de investigación, se tienen que definir conceptos importantes como el de crédito que se define como "una operación de financiación donde un 'acreedor' presta una cierta cifra monetaria a un 'deudor', quien garantiza al acreedor que retornará esta cantidad solicitada más una cantidad adicional, llamada 'intereses' " (Montes de Oca, J, 2015). Además, se tiene que definir la pérdida esperada como "el valor esperado de pérdida por riesgo crediticio en un horizonte de tiempo determinado" (Funding Circle, s.f.). Y finalmente el concepto de elegibilidad que según la RAE es la cualidad de la persona que pueda ser elegida para algo (Real Academia Española, 2014).

Descripción y Análisis de 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:

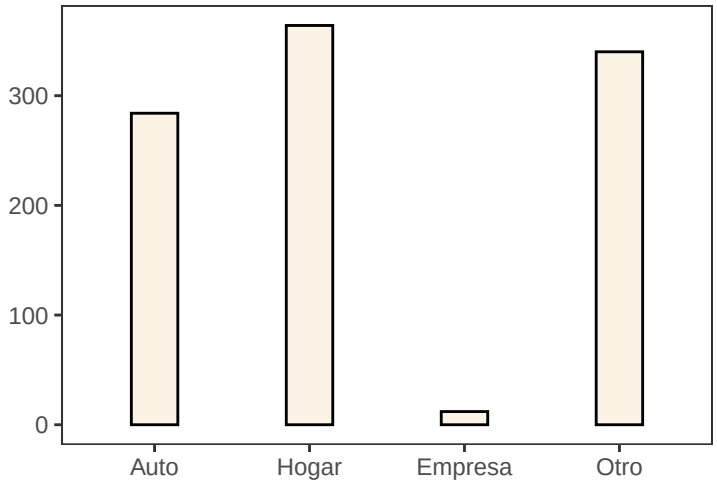
- Elegibilidad: toma el valor de 1 si fue elegible y el valor de 0 si no lo fue.
- Accout Balance: variable categórica que toma el valor de 1 si la persona no cuenta con ninguna cuenta en el banco, el valor de 2 si no tiene un balance pendiente con el banco y el valor de 3 si sí tiene un balance pendiente.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

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:

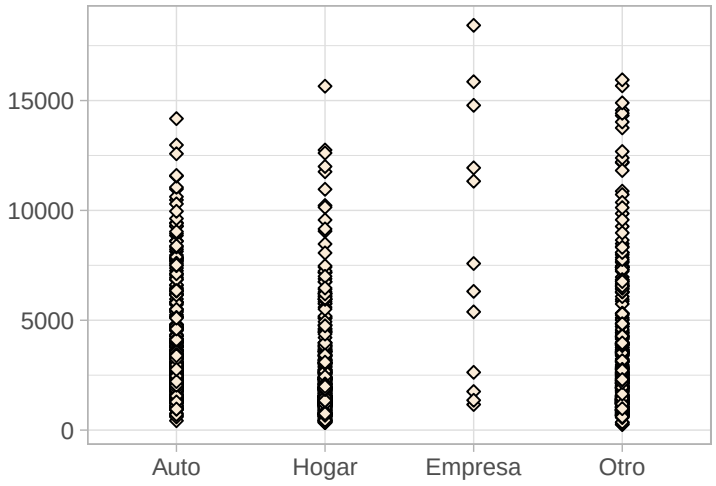


Elaboración propia con datos extraídos de Kaggle

Figura 3.5: 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 3.6: Distribución del Monto del Crédito según su próposito

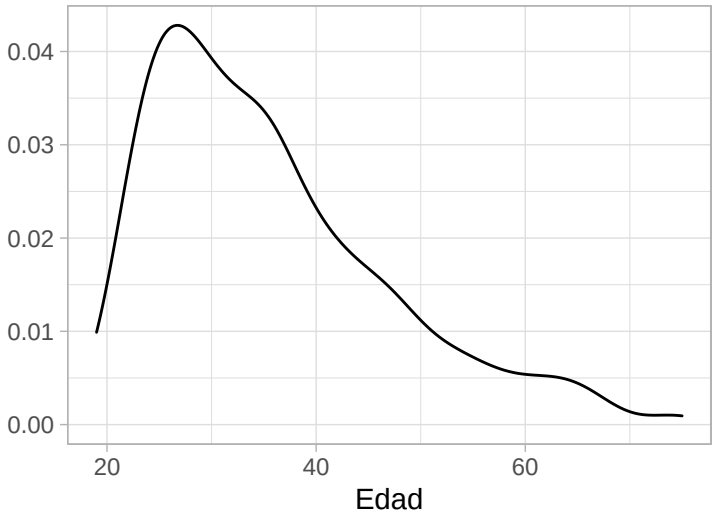
Haciendo una análisis similar para las edades, se llega a la conclusión de que existe una tendencia mayor en las personas cuyas edades estén en el rango de 20 a 40 años a solicitar préstamos como lo muestra la siguiente tabla:

Cuadro 3.5: 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 3.7: 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 3.6: 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

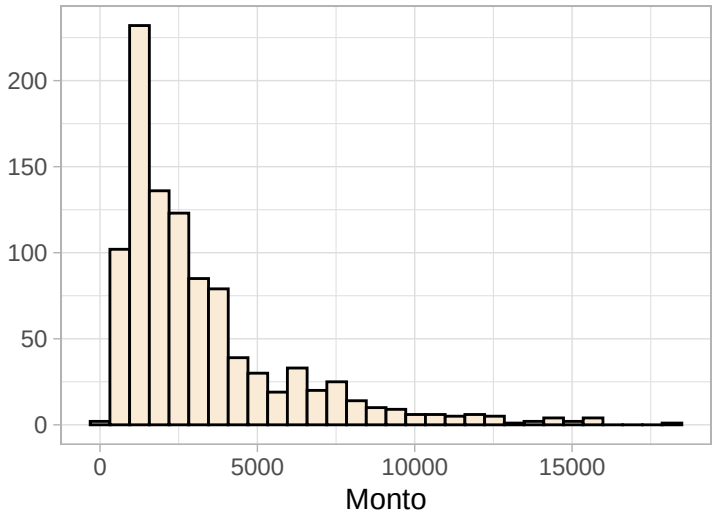
La mayoría de variables que se encuentran en la base son de carácter categórico, por lo que se presenta el siguiente cuadro que muestra información sobre las variables de carácter númerico continuo:

Cuadro 3.7: Resumen de 5 números

	Min	Q1	Mediana	Q3	Max
Monto del crédito	250	1365.5	2319.5	3972.25	18424
Duración en meses del crédito	4	12.0	18.0	24.00	72
Edad	19	27.0	33.0	42.00	75

Note:
Elaboración propia con datos extraídos de Kaggle

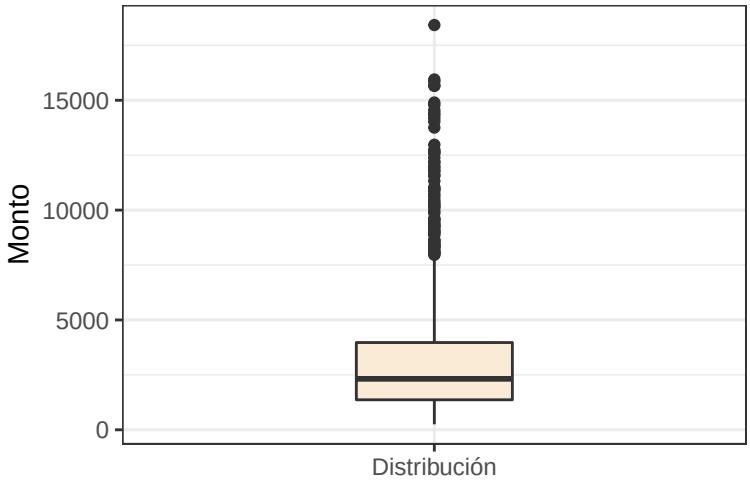
Dado a que eventualmente la idea es realizar un modelo de Cópulas entre los resultados del proceso de clasificación con el Monto del Crédito, sería importante describir de manera exhaustiva esta variable. A continuación un histograma que muestra la distribución de los montos:



Elaboración propia con datos extraídos de Kaggle

Figura 3.8: 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 3.9: 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.

Regresión Logística

Con estos términos claros se puede continuar con el intento responder la pregunta del trabajo, por lo que antes de poder calcular las pérdidas del banco por impago se necesita una manera de determinar 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, esta 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:

$$p(\boldsymbol{X}) = \frac{e^{\beta_0 + X_1\beta_1 + \cdots + X_p\beta_p}}{1 + e^{\beta_0 + X_1\beta_1 + \cdots + X_p\beta_p}} \tag{3.1}$$

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 y col., 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 ya que este tipo de modelo no solo determina la probabilidad de la variable dependiente sino el peso que tienen las covariables para realizar la predicción. Esto ayuda a nivel interpretativo pues se pueden determinar qué variables son más significativas. 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. De acuerdo al tipo de problema para el cual se está realizando el modelo, se debe considerar ajustes para aumentar estas medidas. Otro diagnóstico 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.

Cópulas

Para la segunda parte del trabajo, donde ya 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. 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:

$$\lim_{v_j \to 1^-} C(v1, v2) = v_{3-j} \tag{3.2}$$

$$\lim_{v_j \to 0} C(v1, v2) = 0 \tag{3.3}$$

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

- Incremento

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0 \tag{3.4}$$

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

Frank Copulas

Las Frank Copulas son cópulas arquimedianas, donde cabe distinguir que son de las más usadas para resolver problemas empíricos. Las Frank Copulas son útiles para este problema debido a que pueden expresar la relación entre dependencias tanto positivas como negativas, asimismo, tienen una estructura de dependencia simétrica. (Handini, Maruddani y Safitri, 2019)

Para entrar más en contexto con este tipo de cópulas, primero se procede a explicar lo que es un cópula arquimediana. Suponiendo una dimensión de d, una cópula se le llama arquimediana cuando es de la forma:

$$C(u_1, u_2, \ldots, u_d) = \varphi^{-1}(\varphi(u_1) + \cdots + \varphi(u_d)) \tag{3.5}$$

donde a la función φ se le conoce como función generadora, con el supuesto de que tiene solo un parámetro. Para el caso de las Frank Copulas, esta función generadora viene dada por:

$$\varphi(u) = \ln\left(\frac{e^{-\vartheta u} - 1}{e^{-\vartheta} - 1}\right) \tag{3.6}$$

con $\vartheta > 0$. Como se está trabajando en el caso bivariado, la función C se escribiría de la siguiente forma:

$$C(u_1, u_2) = -\frac{1}{\vartheta} \ln \left(1 + \frac{(e^{-\vartheta u_1} - 1)(e^{-\vartheta u_2} - 1)}{e^{-\vartheta} - 1} \right) \tag{3.7}$$

Estimación del parámetro

La estimación del parámetro ϑ se realiza mediante el método de máxima verosimilitud. Esta función para el caso de cópulas bivariadas se puede escribir como:

$$L = \prod_{i=1}^{2} c_{(u_1,u_2)} \left\{ F_1(x_1), F_2(x_2) \right\} f_1(x_1) f_2(x_2) \tag{3.8}$$

y al aplicar una transformación logarítmica se puede reescribir como:

$$\ln f(x_1, x_2; \vartheta, \rho) = \ln c(F_1(x_1), F_2(x_2); \rho) + \ln f_1(x_1; \vartheta) + \ln f_2(x_2; \vartheta) \tag{3.9}$$

Rho de Spearman

Para encontrar la dependencia existente en la distribución conjunta producto de la cópula, se utiliza el test de estructura de dependencia. Si la data consiste de $x_i = (x_{i1} + x_{i2} + \cdots + x_{id})$, donde d indica la cantidad de observaciones e i indica la i-ésima variable. Este test puede ser calculado con el *Coeficiente de Correlación Rho de Spearman* cuya fórmula viene dada por:

$$\rho = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)} \tag{3.10}$$

donde n es la cantidad de observaciones y d_i es la diferencia de rangos del i-ésimo elemento.

Un detalle importante a la hora de modelar cópulas es la transformación de los datos para que vivan dentro del intervalo $[0, 1]$. Para ello, y en caso de ser necesario, se mapean los datos de la siguiente manera:

El ρ de Spearman muestra la correlación lineal entre ambas variables, sin embargo, el tau de Kendall es otra medida que se ha vuelto más popular en el modelaje y comparación de cópulas.

Tau de Kendall

Este estadístico es un medida de correlación de rango y es muy usado para determinar el grado de dependencia entre dos o más variables. Una prueba de hipótesis τ es una prueba cuya hipótesis nula (H_0) es suponer una correlación nula entra ambas variables ($\tau = 0$) mientras que la hipótesis alternativa sería suponer cierto grado de correlación ($\tau \neq 0$) , en ese sentido es similar al test de dependencia de Spearman.

Esta prueba es no paramétrica ya que no toma en cuenta la distribución de las variables x_1 y x_2 para su desarrollo. (Genest, Quessy y Rémillard, 2006) Esta medida será la usada para el estudio del modelo. Su fórmula viene dada por:

$$\tau = 1 - \frac{4}{\vartheta} + \frac{4D_1(\vartheta)}{\vartheta} \tag{3.11}$$

con $D_1(\vartheta) = \int_0^{\vartheta} \frac{xdx}{e^x - 1}$. Este coeficiente igual arroja valores entre -1 y 1, y la interpretación es análoga al caso del ρ de Spearman.

El tau de Kendall también arroja valores entre -1 y 1 su interpretación es la siguiente:

- Un valor de $\rho = 1$ indica perfecta correlación positiva entre las variables
- Si $0 < \rho < 1$ muestra una correlación positiva
- Un valor de $\rho = 0$ indica que no existe correlación entre las variables
- Si $-1 < \rho < 0$ muestra una correlación negativa
- Un valor de $\rho = -1$ indica perfecta correlación negativa entre las variables

7. Resultados

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:

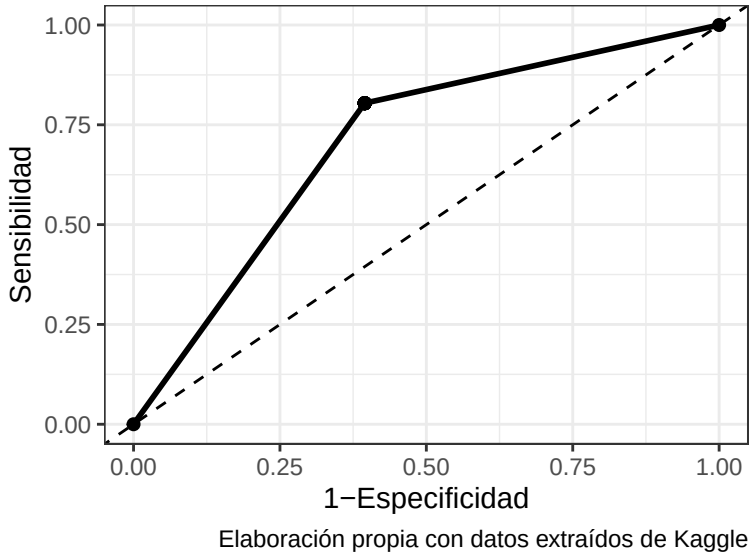


Figura 3.10: Curva ROC

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.

El fijar el umbral para determinar la curva ROC se hizo meramente para mostrar que la regresión logística resulta ser buena, por lo que realmente este es un resultado secundario de la investigación. Cabe recordar que el propósito del trabajo es comparar la distribución marginal de las probabilidades de elegibilidad con variables como el monto del crédito, con el fin de calcular una distribución conjunta usando cópulas.

Dicho esto, si se calcula la correlación empírica entre las probabilidades de elegibilidad con el monto del crédito se llega a una correlación de -0.30. Esta correlación se calculó con el método del tau de Kendall ya que es el método más popular para comparar dependencia entre modelos de cópulas. Este resultado indica que, si el monto a solicitar por parte del cliente es muy alto, disminuyen las probabilidades de ser elegido. Lo mismo pasa de manera análoga cuando el monto del crédito es bajo, en donde las probabilidades de ser elegido aumentan. Vale la pena mencionar que no se alcanza una correlación perfecta, en parte, esto se debe a que hay muchos más factores que influyen en la elegibilidad de los solicitantes, no solo el monto. Además, dependiendo del método con el que se calcule la correlación, la misma puede variar. Por ejemplo, para el caso de la correlación de Spearman el resultado es de -0.44.

Lo anterior expuesto se puede identificar como el primer hallazgo y resulta ser muy útil ya que esta asociación negativa funciona como una especie de filtro a la hora de escoger la cópula que mejor ajusta los datos. Esto sucede porque no todas las cópulas pueden modelar relaciones negativas. Esto trae ciertas desventajas ya que se reduce el número de cópulas de dónde escoger para modelar la distribución conjunta. La dependencia negativa no es compatible con las cópulas de Clayton, Gumbel, Joe, BB1, BB6, BB7 y BB8.

De las cópulas posibles, se probaron múltiples combinaciones entre las cuales distinguen la gaussiana, la t-student y la Frank. Todo esto para determinar cuál era la que mejor se adaptaba a los datos. El AIC fue el método usado para determinar la mejor cópula. El hallazgo primario aquí sería que la Frank se escogió como la mejor con un AIC de -63.46, mientras que las demás tuvieron AIC de -55.04 y -53.04 respectivamente. Además, otro hallazgo primario que rectifica la escogencia de esta cópula para este problema es la relación que existe entre la correlación empírica y la teórica calculada por la cópula. Lo ideal sería que la dependencia teórica calculada por el modelo sea bastante similar a la dependencia empírica. En este caso, se da que la empírica corresponde a un valor aproximado de -0.30 mientras que la correlación teórica calculada por el modelo de cópula de Frank es de -0.36. Hay una diferencia de 0.06 puntos porcentuales entre ambas correlaciones, pero a groso modo, ambos explican el fenómeno bastante parecido indicando el buen ajuste de la cópula con los datos.

Otro método en el cual se basó para determinar cual cópula es la que mejor se ajusta, se utilizó un *Goodness-of-fitness test*, el cual se basa en el cálculo del estadístico de Cramér-von-Mises, también conocido como S_n. Este estadístico se puede interpretar similar al AIC en el sentido que entre menor es el valor se obtiene un mejor modelo, pues el *valor-p* asociado va a ser mayor y esto resulta en que no haya evidencia suficiente para rechazar la hipótesis nula, que es lo que se está buscando. Por lo que bajo esta prueba también se obtiene como resultado que la cópula que mejor se ajusta es una de Frank. Similarmente se puede ver como la evidencia estadística sugiere lo hallado con respecto a que si se tiene una correlación negativa la función de Clayton no es recomendable y esto se puede observar en como este modelo es el que peor métricas obtiene de los cuatro modelos.

Para poder construir las funciones de probabilidad bivariadas aparte de la función de cópulas, se necesitan las funciones de distribución para cada variable, a estas funciones se les llama marginales. La distribución que mejor se ajusta al monto de crédtio bajo el método de AIC y de estimación de parámetros mediante máxima verosimilitud es una distribución $Gamma(1{,}8115, 1849)$ bajo una parametrización de forma y escala.Una vez con la distribución estimada de la variable para el monto de crédito, se procede a realizar un ajuste similar esta vez para la variable restante que es la elegibilidad de la persona. Así, la distribución que mejor se ajusta para la elegibilidad es una distribución $N(0{,}5626, 0{,}1549)$. Por lo que ahora si se puede proceder a construir las funciones de probabilidad considerando las marginales y la función de cópulas estimadas.

Una vez con las marginales y la función de cópulas estimadas se procede a la construcción de funciones de densidad y distribución de las funciones. Estas funciones son las que se buscan para contestar la pregunta de investigación por lo que es uno de los resultados más importantes del trabajo. Con estas funciones se pueden crear simulaciones de datos nuevos y a su vez permite una manera de representar las variables de estudio. Estas funciones al ser bivariadas representan funciones de dos

dimensiones por lo que su resultado o probabilidad se da en una tercera dimensión. Esto complica la visualización usual de gráficos de densidad y distribución. Para esto se procede a utilizar técnicas como gráficos de contornos para facilitar la interpretación de estos resultados. A su vez, estas funciones también permiten calcular medidas de riesgo como el VaR o CVaR, que es lo que se piensa estimar en futuras bitácoras.

8. Parte de reflexión

A través del proceso de análisis de datos y del primer intento de modelaje, se tomó la decisión de modificar el rumbo de la investigación donde la misma ya no se enfocará en determinar la pérdida esperada del banco. Ahora su eje principal girará alrededor de encontrar la función de densidad conjunta entre la probabilidad de ser elegido para un crédito y el monto de dicho crédito. Se construirá esta función con cópulas, esto con el fin de contemplar la dependencia entre ambas variables. Una vez con esta función se pueden calcular probabilidades relacionadas con riesgo como el Value-at-Risk.

Capítulo 4

Bitácora 4

1. Resumen

La investigación propuesta consiste en la elaboración de un modelo de clasificación para determinar probabilidad de elegibilidad de un individuo para un crédito considerando diversas variables de interés como la edad y el monto solicitado. Hay diversos estudios que siguen una línea similiar, sin embargo, esta investigación se diferencia del resto por el hecho de que se va a realizar un análisis de dependencia de las variables y se marca como objetivo la creación de una función de distribución bivariada que contenta la probabilidad de elegibilidad.

De esta forma, se distinguen dos principales etapas en el trabajo, la primera es la construcción de esas probabilidades de elegibilidad mediante una regresión logística. Sin embargo, en esta etapa se presenta un inconveniente con la naturaleza de los datos, ya que hay muchas variables y cada variable tiene múltiples categorías. Para solventar esto, se realiza una depuración para combinar categorías que tengan relación y se usa la prueba Chi-Cuadrado para estudiar la significancia que estas variables tienen con la elegibilidad.

La segunda etapa del trabajo consiste en la elaboración de una distribución bivariada para las probabilidades calculadas en el punto anterior; la otra variable de interés será el monto del crédito. El primer paso es encontrar la distribución marginal para cada variable y para el desarrollo de la distribución conjunta se utilizarán cópulas. Como se expondrá más adelante, hay muchos tipos diferentes de cópulas, por lo que el primer será encontrar la que más se ajuste a los datos por lo que habrá primero que hacer un análisis entre diferentes estadísticos para determinar la mejor cópula. Los estadísticos y diagnósticos que se utilizarán son el AIC y el Goodness-of-fit test.

Una vez concluidas ambas etapas, se destina una sección para exponer los resultados alcanzados. Por un lado se tiene una regresión logística con buena sensibilidad y una precisión de aproximadamente el 76 %. Asimismo, la distribución marginal para el monto del crédito resultó ser una $Gamma(1{,}8115, 1849)$ y la para la probabilidad de elegibilidad es una $N(0{,}5626, 0{,}1549)$. Con estas distribuciones se prueban varias cópulas y se determina que el mejor ajuste se logran mediante una Frank Cópula, donde se logra una AIC de -63.46 y un Goodness-of-fit de 0.0266 que comparados con las otras cópulas, resultan ser los valores más pequeños.

2. Introducción

Día con día personas se presentan a entidades bancarias con el fin de solicitar créditos. A pesar de lo que se pueda pensar, esto es un negocio muy rentable para los bancos, sin embargo, el otorgamiento de los mismos debe realizarse de manera responsable y considerando múltiples criterios. Y dado el contexto actual en donde se está atravesando por una crisis económica importante en todo el mundo,

la solicitud de créditos se ha vuelto una forma común de financiamiento para afrontar esta recesión económica. Este puede ser uno de los principales retos a lo que se enfrentan las entidades financieras, determinar a cuáles personas es rentable prestarles dinero. Dicho esto, se distingue una relevancia social bastante más importante más de allá de aplicar algoritmos matemáticos para contestar una pregunta, ya que con estudios de esta índole se puede contribuir a mejorar la capacidad económica de la población.

Este proyecto se encarga de estudiar ese fenómeno con un modelo de clasificación como lo es la regresión logística, sin embargo, para dar una respuesta aún más completa hará faltar hacer de otros métodos estadísticos. Esto debido a que este tipo de modelos por lo general no contemplan si existe relación entre las variables estudiadas. Esa posible dependencia hará que el modelo se vea afectado. Mediante este proyecto se propone realizar un análisis de dependencia adicional para intentar relacionar esa probabilidad de elegibilidad del individuo con otras variables de interés como el monto del crédito. Se usaran datos de un banco alemán que mantiene su nombre en anonimato y hay muchas variables que este banco considera de interés para otorgar un préstamo dentro de las cuales se destaca la edad, el monto del crédito, la cantidad de créditos que tiene, entre otras. Dado la alta sensibilidad de la información, la página de dónde se extrajeron los datos no identifica un contexto temporal definido.

La presente investigación consta de dos principales etapas,la primera es la construcción de modelo para la elegibilidad de los usuarios y la segunda es una análisis de dependencia de la probabilidad de elegibilidad. La primera etapa de investigación se basará en estudios ya existentes en las cuales se utiliza un modelo de clasificación logística para determinar la elegibilidad para un crédito. Por su contraparte, no se encontró literatura relacionada al análisis de dependencia propuesto entre la probabilidad de elegibilidad y variables de interés como el monto del crédito.

Con este proyecto, el principal objetivo es identificar la relación que existe entre la probabilidad de elegibilidad para un crédito con el monto de dicho crédito, más allá de simplemente calcular un coeficiente de correlación lineal que en muchas ocaciones no muestra la relación en su totalidad. Asimismo, se busca construir una función de distribución que explique dicha relación entre las variables y permita realizar cálculos como un Value at Risk. Para la construcción de la distribución bivariada se usarán cópulas, en donde primero se hará una análisis exploratorio para determinar la cópula que mejor explica la relación.

3. Metodología

Antes de entrar de lleno con los métodos estadísticos propuesto para contestar la pregunta, se realiza un análisis de los datos por utilizar. Los base fue recuperada de kaggle y originalmente fue obtenida 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. Asimismo, no se indica un contexto temporal.

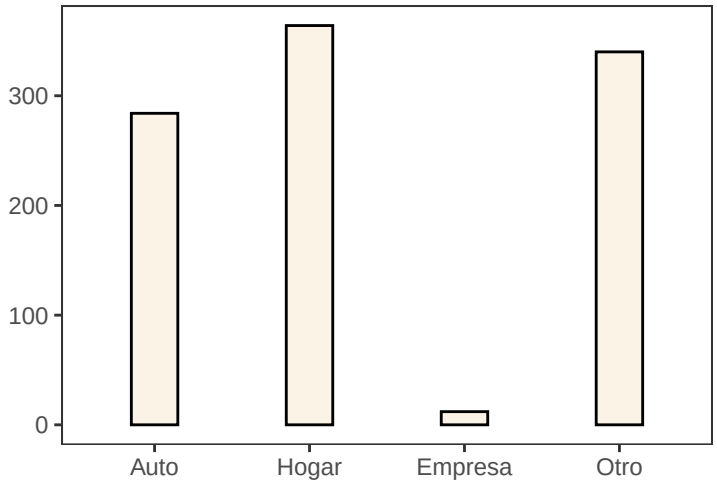
La población de estudio se define como las personas que solicitaron un crédito en esta entidad bancaria ubicada en Alemania mientras que la unidad estadística se define como la persona solicitante de un crédito en el banco alemán estudiado. La muestra para el desarrollo de dicha investigación consta 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 conveniente enfatizar que en 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, y debido a ese exceso de variables, se tratará de reducirlas haciendo un proceso de depuración de los datos. 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:

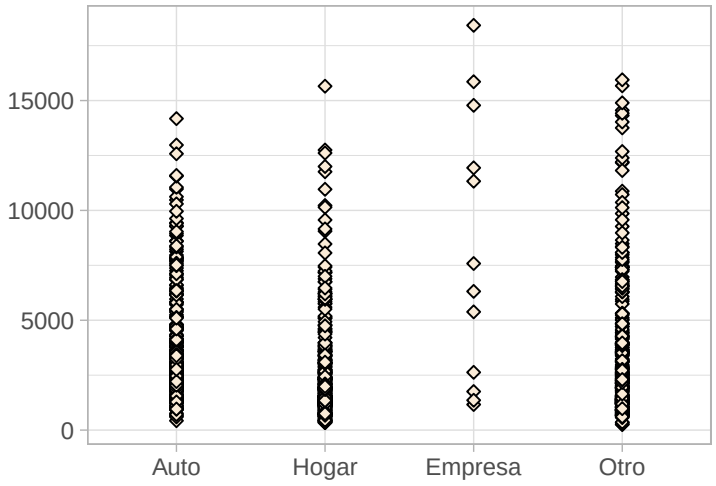


Elaboración propia con datos extraídos de Kaggle

Figura 4.1: Distribución de la variable Próposito del crédito

Lo 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

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 4.2: Resumen de 5 números

	Min	Q1	Mediana	Q3	Max
Monto del crédito	250	1365.5	2319.5	3972.25	18424
Duración en meses del crédito	4	12.0	18.0	24.00	72
Edad	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:

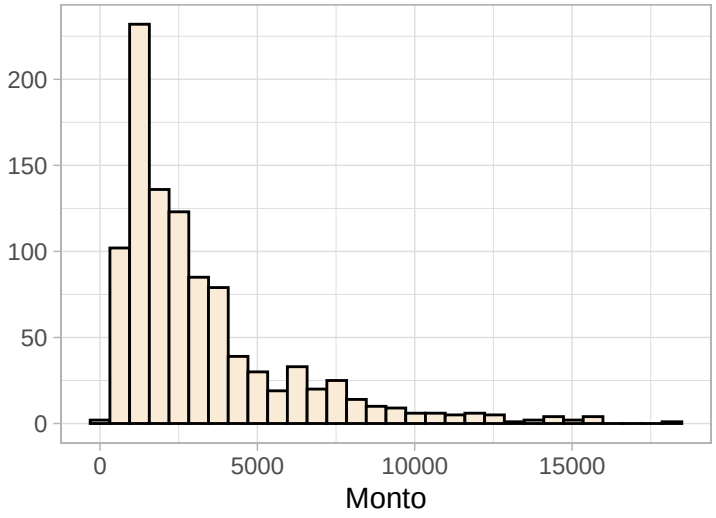
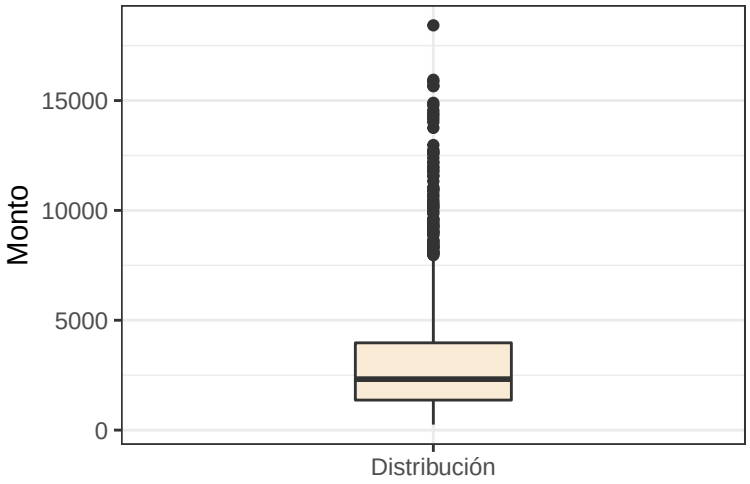


Figura 4.3: Histograma de la variable Monto del Crédito

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 4.4: 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.

Regresión Logística

Con estos términos claros se puede continuar con el intento responder la pregunta del trabajo y el primer paso es el desarrollo de 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, esta 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:

$$p(\boldsymbol{X}) = \frac{e^{\beta_0+X_1\beta_1+\cdots+X_p\beta_p}}{1+e^{\beta_0+X_1\beta_1+\cdots+X_p\beta_p}} \tag{4.1}$$

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 y col., 2021). James menciona diagnósticos para modelos de esta índole pero el que más destaca es el de la curva ROC (Receiver Operation Curve), la cual sirve para graficar la tasa de falsos positivos con la sensibilidad del modelo. En la sección de resultados se muestra dicha curva con las respectivas de sensibilidad de especificidad.

Cópulas

Para la segunda parte del trabajo ya se planean utilizar 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. 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 $\boldsymbol{I} = (0, 1)$. Entonces, se puede sintetizar el concepto de cópula como una función de la forma: $C : \boldsymbol{I}^2 \to \boldsymbol{I}$. Sin embargo, hay muchos estilos diferentes de cópulas de dónde escoger, hay algunas especiales para modelar correlaciones negativas, otras que se usan para valores extremos, etc. Para efectos de la investigación, la que resulta de interés es la Frank Cópula que se detalla a continuación:

Frank Copulas

Las Frank Copulas son cópulas arquimedianas, donde cabe distinguir que son de las más usadas para resolver problemas empíricos. Las Frank Copulas son útiles para este problema debido a que pueden expresar la relación entre dependencias tanto positivas como negativas, asimismo, tienen una estructura de dependencia simétrica. (Handini, Maruddani y Safitri, 2019)

Para entrar más en contexto con este tipo de cópulas, primero se procede a explicar lo que es un cópula arquimediana. Suponiendo una dimensión de d, una cópula se le llama arquimediana cuando es de la forma:

$$C(u_1, u_2, \ldots, u_d) = \varphi^{-1}(\varphi(u_1) + \cdots + \varphi(u_d)) \tag{4.2}$$

donde a la función φ se le conoce como función generadora, con el supuesto de que tiene solo un parámetro. Para el caso de las Frank Copulas, esta función generadora viene dada por:

$$\varphi(u) = \ln\left(\frac{e^{-\vartheta u}-1}{e^{-\vartheta}-1}\right) \tag{4.3}$$

con $\vartheta > 0$. Como se está trabajando en el caso bivariado, la función C se escribiría de la siguiente forma:

$$C(u_1, u_2) = -\frac{1}{\vartheta} \ln \left(1 + \frac{(e^{-\vartheta u_1} - 1)(e^{-\vartheta u_2} - 1)}{e^{-\vartheta} - 1} \right) \tag{4.4}$$

Estimación del parámetro

La estimación del parámetro ϑ se realiza mediante el método de máxima verosimilitud. Esta función para el caso de cópulas bivariadas se puede escribir como:

$$L = \prod_{i=1}^{2} c_{(u_1,u_2)} \left\{ F_1(x_1), F_2(x_2) \right\} f_1(x_1) f_2(x_2) \tag{4.5}$$

y al aplicar una transformación logarítmica se puede reescribir como:

$$\ln f(x_1, x_2; \vartheta, \rho) = \ln c(F_1(x_1), F_2(x_2); \rho) + \ln f_1(x_1; \vartheta) + \ln f_2(x_2; \vartheta) \tag{4.6}$$

Medidas de asociación

Existe estadísticos que detallan la relación que existe entre variables pero dependiendo del problema que se esté trabajando, la información de estos estadísticos puede resultar pobre. Sin embargo, pueden revelar información importante sobre qué tipo de cópula se puede usar. Entre las medidas de asociación más populares sobresalen:

ρ de Spearman

El ρ de Spearman muestra la correlación lineal entre ambas variables y es de las medidas más usadas. Para dos variables aleatorios con distribución uniforme en $[0, 1]$ se tiene que el ρ de Spearman viene dado por:

$$\rho_S(X_1, X_2) = 12 \cdot \mathbb{E} \left[F_1(X_1) F_2(X_2) \right] - 3 \tag{4.7}$$

Y cambiando $U = F_1(X_1)$ y $V = F_2(X_2)$ se reescribre como:

$$\begin{aligned} \rho_S(X_1, X_2) &= 12 \cdot \mathbb{E} \left[UV \right] - 3 \\ &= 12 \int_0^1 \int_0^1 uv \, dC(u, v) - 3 \\ &= 12 \int_0^1 \int_0^1 C(u, v) - 3 \end{aligned} \tag{4.8}$$

Si bien es cierto que esta es de la medidas más usadas, cuando de modelado de cópulas se habla, el τ de Kendall es más popular.

τ de Kendall

Este estadístico es un medida de correlación de rango y es muy usado para determinar el grado de dependencia entre dos o más variables. Se puede escribir como función de una cópula bivariada como:

$$\begin{aligned} \tau_K(X_1, X_2) &= 4 \cdot \int_0^1 \int_0^1 C(u, v) \, dC(u, v) - 1 \\ &= 4 \cdot \mathbb{E} \left[C(U, V) \right] - 1 \end{aligned} \tag{4.9}$$

El τ de Kendall arroja valores entre -1 y 1 su interpretación es la misma para el caso del ρ de Spearman la cual es:

- Un valor de $\rho = 1$ indica perfecta correlación positiva entre las variables
- Si $0 < \rho < 1$ muestra una correlación positiva
- Un valor de $\rho = 0$ indica que no existe correlación entre las variables
- Si $-1 < \rho < 0$ muestra una correlación negativa
- Un valor de $\rho = -1$ indica perfecta correlación negativa entre las variables

4. Resultados

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:

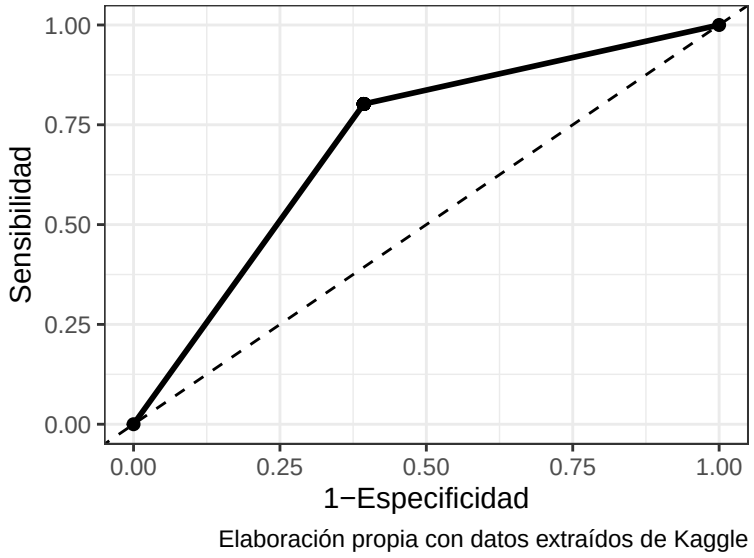


Figura 4.5: Curva ROC

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.

El fijar el umbral para determinar la curva ROC se hizo meramente para mostrar que la regresión logística resulta ser buena, por lo que realmente este es un resultado secundario de la investigación. Cabe recordar que el propósito del trabajo es comparar la distribución marginal de las probabilidades de elegibilidad con variables como el monto del crédito, con el fin de calcular una distribución conjunta usando cópulas.

Dicho esto, si se calcula la correlación empírica entre las probabilidades de elegibilidad con el monto del crédito se llega a una correlación de -0.30. Esta correlación se calculó con el método del tau de Kendall ya que es el método más popular para comparar dependencia entre modelos de cópulas. Este resultado indica que, si el monto a solicitar por parte del cliente es muy alto, disminuyen las probabilidades de ser elegido. Lo mismo pasa de manera análoga cuando el monto del crédito es

bajo, en donde las probabilidades de ser elegido aumentan. Vale la pena mencionar que no se alcanza una correlación perfecta, en parte, esto se debe a que hay muchos más factores que influyen en la elegibilidad de los solicitantes, no solo el monto. Además, dependiendo del método con el que se calcule la correlación, la misma puede variar. Por ejemplo, para el caso de la correlación de Spearman el resultado es de -0.44.

Lo anterior expuesto se puede identificar como el primer hallazgo y resulta ser muy útil ya que esta asociación negativa funciona como una especie de filtro a la hora de escoger la cópula que mejor ajusta los datos. Esto sucede porque no todas las cópulas pueden modelar relaciones negativas.

Para poder ajustar un modelo de cópulas bivariadas, primero se tienen que saber las funciones de distribución marginales univariadas para cada variable de estudio. En este se necesitan encontrar dos de estas marginales pues el modelo de cópulas a implementar va a considerar como dos variables, el monto del crédito solicitado y el valor de la regresión logística asociado a la probabilidad de elegibilidad asignado con el modelo.

Por lo que primeramente se van a mostrar los gráficos de la distribución empírica de las variables de estudio, mediante histogramas.

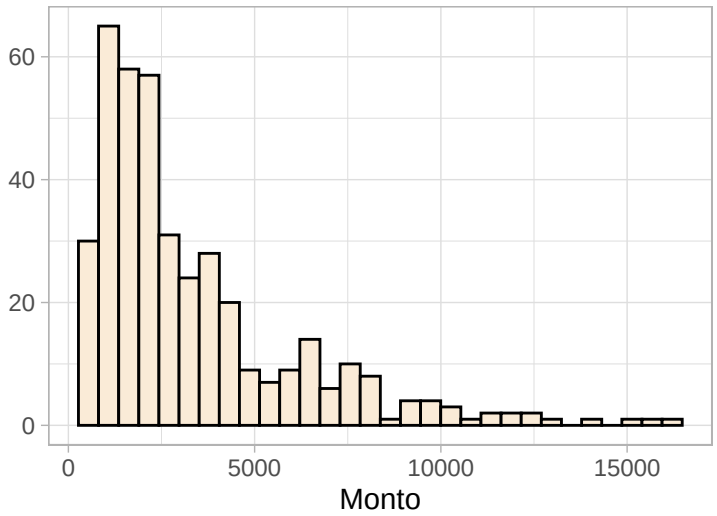


Figura 4.6: Histograma de los montos de crédito de la base de prueba

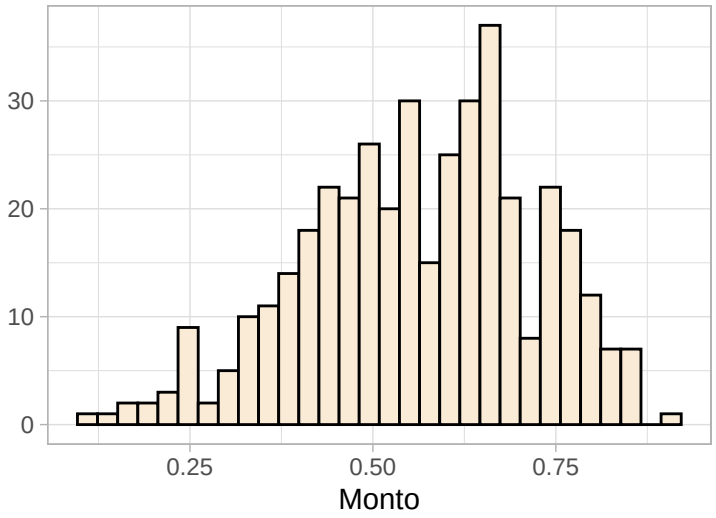


Figura 4.7: Histograma de los probabilidades de elegibilidad

Con estas figuras se pueden observar ciertas tendencias de como se distribuyen las variables, lo que permite buscar ajustar una distribución paramétrica de acuerdo a su forma. Bajo este hilo, se sigue que el monto del crédito es bastante asimétrica hacia la derecha, mientras que la probabilidad de elegibilidad sigue una tendencia más simétrica.

Para el siguiente desarrollo se va a utlizar el lenguaje de programación R para realizar los modelos, cálculos y estimaciones. Primeramente, se va van a realizar distintos modelos para las distribuciones marginales de las variables de estudio donde por métodos como AIC y estimación de parámetros por máxima verosimilitud. Bajo esta metodología, se consigue que la mejor distribución que se ajusta para el monto del crédito es una distribución $Gamma(1{,}8115402, 1849)$ bajo una parametrización de forma y escala. Mientras que, al realizar este ajuste con la probabilidad de elegibilidad el modelo que mejor ajusta es una distribución $N(0{,}5625845, 0{,}1549248)$.

Una vez, con estas distribuciones marginales se procede a modelar la correlación entre estas variables mediante cópulas para lograr estimar una función de distribución bivariada considerando las variables de estudio.

Como se mencionó en el marco teórico, el coeficiente de correlación que se va a utilizar es el de Kendall, conocido como tau de Kendall. Para los datos empíricos se obtiene que $\hat{\tau}_K = -0{,}3$ lo que indica una correlación negativa entre las variables, es decir que conforme el monto del crédito aumenta, las probabilidades de elegibilidad disminuyen. Con esta información, se procede a escoger una función de cópulas que más se ajuste a los datos observados para poder construir su función de distribución. Bajo métodos de escogencia de una función de cópulas como el AIC y la estimación de parámetros por medio de máxima verosimilitud se llega a un modelo con el mejor ajuste que sería una cópula de Frank con parámetro $\theta = -3{,}6$ y un $\tau_K = -0{,}36$, lo cual es bastante cercano a la correlación empírica calculada anteriormente. Es decir, se obtiene una función de cópulas (Arquimediana) que mantiene la correlación de los datos, lo cual también es lo buscado.

Una vez con este modelo, y las marginales univariadas calculadas al principio, se procede a construir la función de distribución y densidad bivariada. Que es lo que se planea contestar con la pregunta de investigación planteada. Para su visualización se utilizan técnicas de graficación en tres dimensiones en dos dimensiones por lo que el siguiente gráfico muestra el contorno de la densidad bivariada.

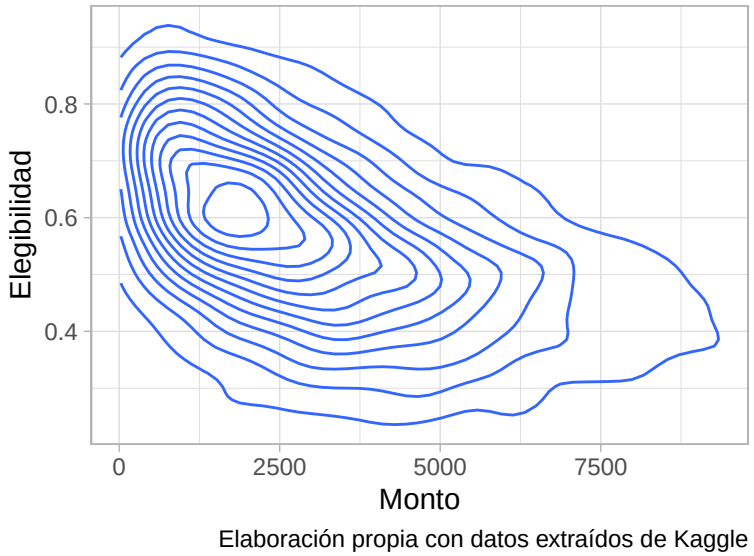


Figura 4.8: Contorno de la función de densidad bivariada

Como se mencionó anteriormente, la escogencia de la cópula fue mediante AIC sin embargo para poder determinar que tan bien se ajusta una cópula, se utilza el método de *Goodness-of-fit test* o *Gof Test*. El cual se basa en calcular el estadístico de Cramér-von-Mises o S_n. Por lo que, ne la siguiente tabla se muestran los resultados de las pruebas realizadas.

Cuadro 4.3: AIC de las diferentes familias de cópulas

Cópula	AIC
Frank	-63.46017
Normal	-55.04813
t	-53.04770
Clayton	-10.97244

Cuadro 4.4: AIC de las diferentes familias de cópulas

Cópula	Estadístico Cramér-von-Mises
Frank	0.0266808
Normal	0.0269870
t	0.0426066
Clayton	0.6172515

Como se puede observar, tanto el AIC como el S_n dan más bajos en el caso de una cópula de la familia Frank. Por lo que debido a estas pruebas se llega a la conclusión que el mejor modelo de cópulas para las variables de estudio es una cópula de Frank. También, vale la pena recalcar que las peores métricas se obtienen considerando un modelo de cópula de Clayton, lo cual corrobora el hecho que si se presenta una correlación negativa este modelo es peor y en algunos algoritmos ni se considera intentar ajustar con este modelo.

Finalmente, también se realizó una serie de datos simulados con la función de distribución bivariada contra los datos reales para de manera visual poder analizar tendencias y comportamientos entre el modelo y la realidad.

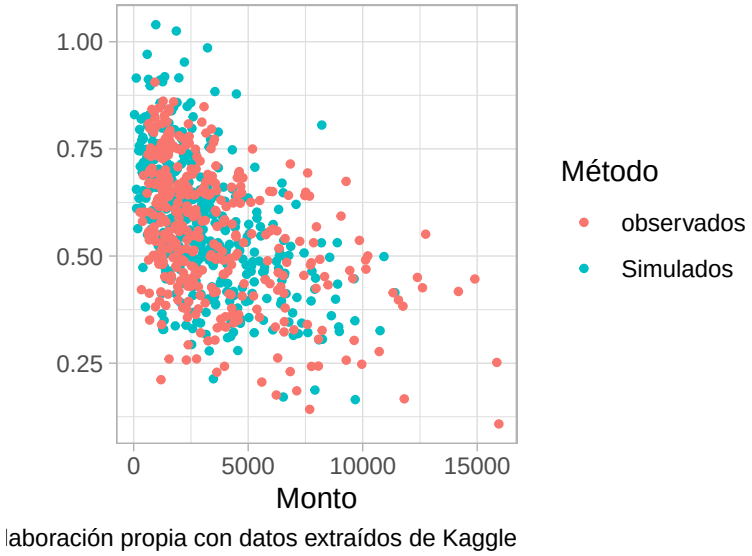


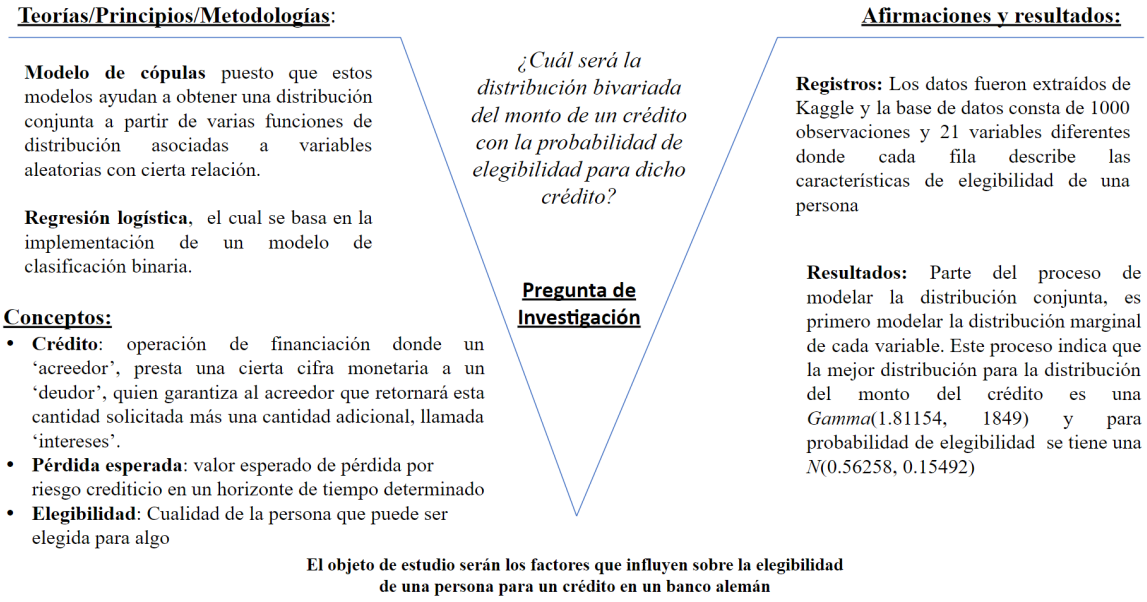
Figura 4.9: Datos simulados con la función de distribución vs. datos observados

5. Conclusión

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.

Funding Circle. (s.f.). *Pérdida esperada (PE)*.

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

Genest, C., Quessy, J.-F., & Rémillard, B. (2006). Goodness-of-fit procedures for copula models based on the probability integral transformation. *Scandinavian Journal of Statistics*, *33*(2), 337–366.

Handini, J., Maruddani, D., & Safitri, D. (2019). Frank copula on value at risk (VaR) of the construction of bivariate portfolio (case study: Stocks of companies awarded with the IDX top ten blue with stock period of 20 october 2014 to 28 february 2018). *Journal of Physics: Conference Series*, *1217*, 012078.

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

Klugman, S. A., Panjer, H. H., & Willmot, G. E. (2013). *Loss models: Further topics*. John Wiley & Sons.

Montes de Oca, J. (2015). Crédito. *Economipedia*.

Real Academia Española. (2014). *Elegibilidad* (23. ed.).

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