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

Índice general I

1 Bitácora 1 **1**
1. Planificación 1
Idea principal 1
Base de datos 2
Principios y/o teorías 3
Fichas bibliográficas 4
2. Escritura 6

2 Bitacora 2 **7**
1. Fichas bibliográficas nuevas 7
2. Ordenamiento de literatura 10
3. Análisis estadístico 11
4. Fichas de Resultados 16
5. Enlaces de Literatura 18

3 Anexos **20**
1. Uve de Gowin 20

Referencias **21**

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:

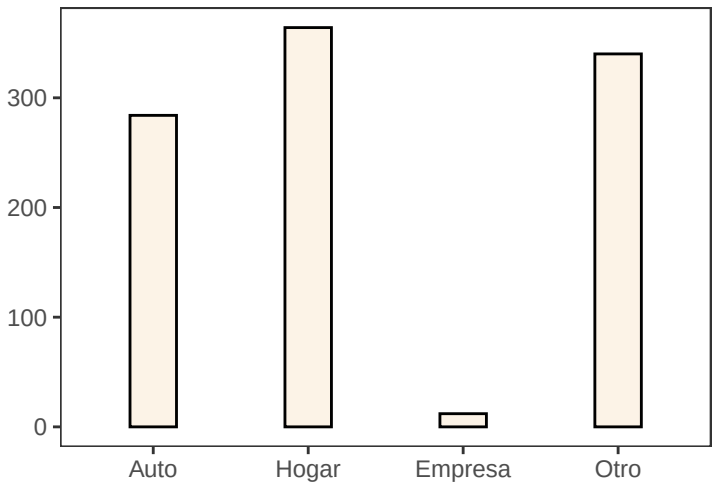
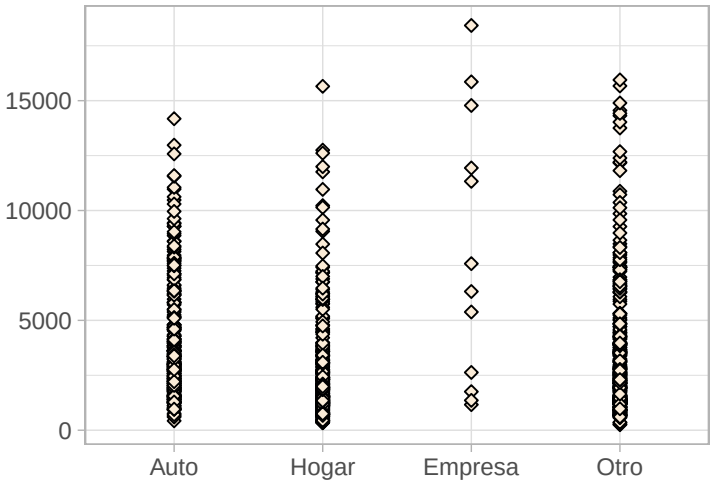


Figura 2.1: Distribución de la variable Próposito del crédito

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

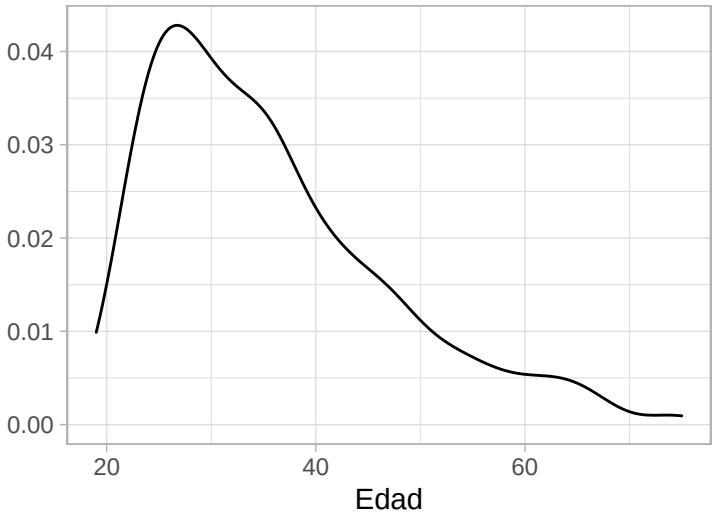
Haciendo una análisis similar para las edades, se llega a la conclusión de que existe una tendencia mayor en las personas cuyas edades estén en el rango de 20 a 40 años a solicitar préstamos como lo muestra la siguiente tabla:

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
250	1365.5	2319.5	3972.25	18424
4	12.0	18.0	24.00	72
19	27.0	33.0	42.00	75

Note:
Elaboración propia con datos extraídos de Kaggle

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

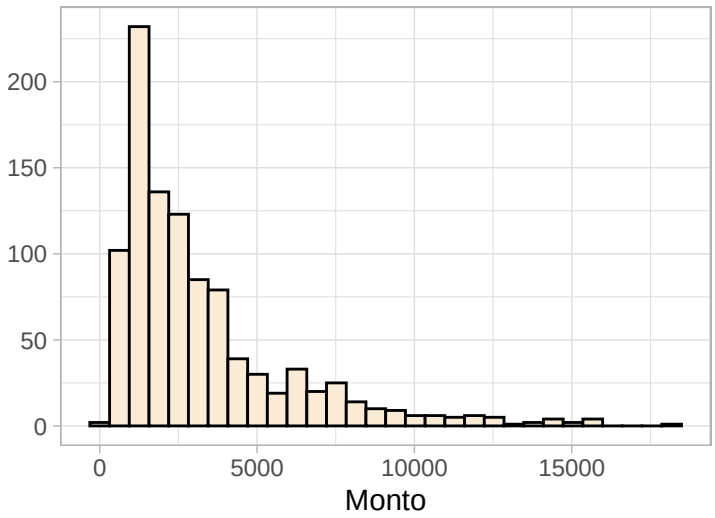


Figura 2.4: Histograma de la variable Monto del Crédito

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:

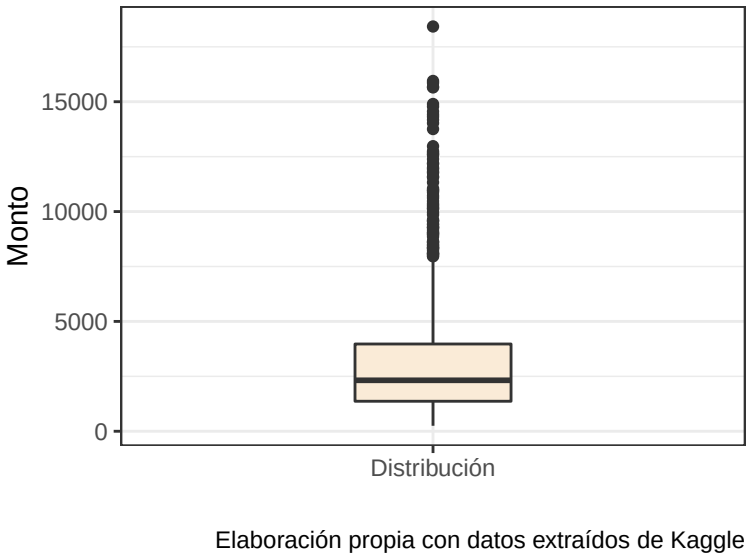


Figura 2.5: Diagrama de Caja y Bigotes de la variable Monto del Crédito

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:

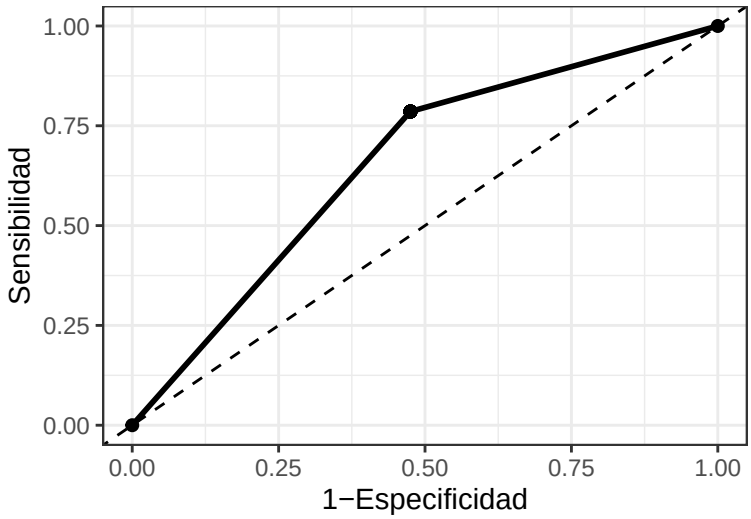


Figura 2.6: Curva ROC

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.

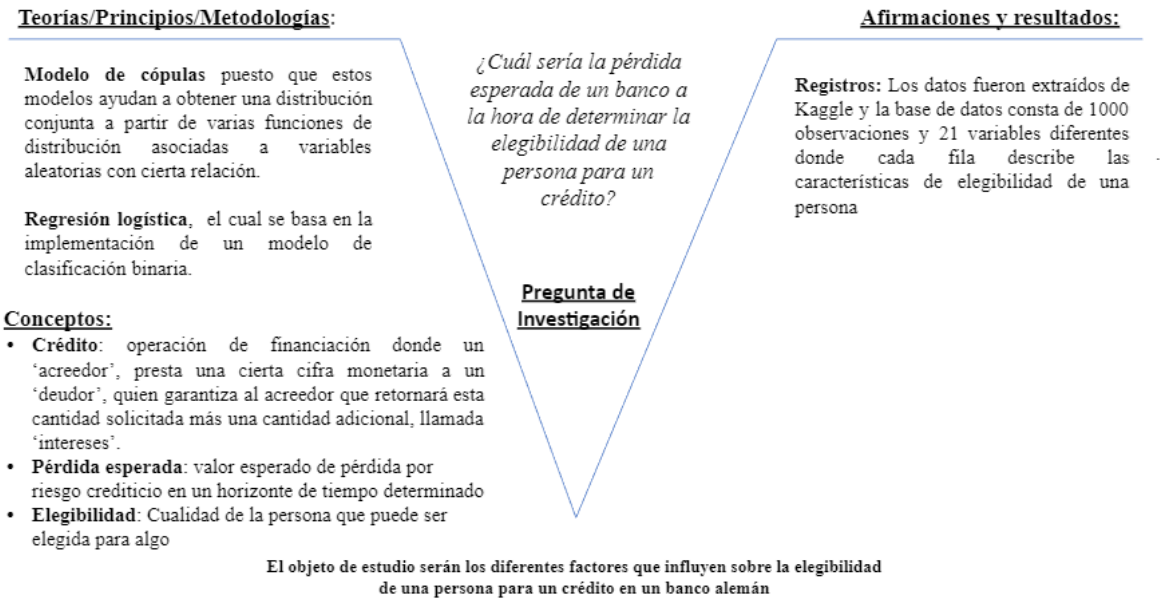
Sin embargo para el enfoque del trabajo estas no son de mucha utlidad pues como menciona Geidosch el uso de cópulas convencionales presenta dos limitantes, la primera es que cuando el número de dimensiones es mayor que dos, el número de familias de cópulas aplicables se limita al caso elíptico y arquimediano. 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

Anexos

1. Uve de Gowin



Referencias

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

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

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

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

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

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

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

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