UNIVERSITAS
Miguel Hernández

Trabajo Fin de Grado
Grado en Estadística Empresarial

Análisis de las tendencias de compra a lo largo del tiempo: Estudio de los datos de un mercado estadounidense.

Autora: Aitana Bataller García

Tutora: María Asunción Martínez Mayoral

Facultad de Ciencias Sociales y Jurídicas de Elche

Universidad Miguel Hernández

Curso 2024/2025

UNIVERSITAS Miguel Hernández

Índice de contenidos

1. Resumen

El objetivo de este estudio es estudiar las tendencias de compra en un supermercado de Estados Unidos utilizando datos de ventas entre los años 2011 y 2015. Se analizarán patrones en los beneficios y tiempos de entrega para optimizar la toma de decisiones empresariales.

Se aplicaron técnicas de análisis exploratorio y modelos predictivos como regresión lineal, Lasso, Ridge, árboles de regresión y bosques aleatorios. Los resultados muestran que el modo de envío es el principal factor que afecta el tiempo de entrega, mientras que la ubicación de la compra (estado/ciudad) influye significativamente en la rentabilidad. Además, se observó que los descuentos elevados pueden reducir los márgenes de ganancia y que ciertas categorías de productos presentan variaciones en la demanda según la temporada. Estos hallazgos pueden servir como base para estrategias comerciales más eficientes, optimizando la gestión de inventario y los recursos logísticos.

Palabras clave

Análisis de mercados, decisiones basadas en datos, tiempos de entrega, beneficios por ventas, machine learning.

2. Antecedentes

Las empresas de hoy en día se basan en la recopilación y análisis de datos para tomar decisiones estratégicas. Los avances tecnológicos han facilitado el análisis de grandes volúmenes de información, permitiendo a las organizaciones descubrir información valiosa y detectar tendencias emergentes en el mercado. Tal conocimiento permite a las empresas optimizar sus operaciones, personalizar la experiencia del cliente y predecir el comportamiento del consumidor, generando así una ventaja competitiva sostenible. Al identificar áreas de mejora y oportunidades de crecimiento, las organizaciones pueden optimizar sus recursos y aumentar su rendimiento.

Este trabajo fin de grado, enfocado al análisis de una base de datos relativa a ventas de diferentes tipos de productos en Estados Unidos, pretende aportar luz sobre cómo el análisis de datos puede contribuir a una gestión más inteligente y estratégica de la información disponible. Este informe presenta un análisis exhaustivo de las ventas de la base de datos de una empresa de venta online. El análisis tiene como objetivo descubrir patrones valiosos que puedan ayudar a la compañía a comprender mejor la dinámica de sus ventas, el comportamiento de los clientes y la rentabilidad, poniendo especial interés en la predicción de los beneficios y de los tiempos de entrega, cuestiones cruciales en lo que se refiere a ventas online, dadas las condiciones habituales de descuentos y los costes de diferentes tipos de transporte.

Para llevar a cabo este análisis, se empleó un conjunto de datos obtenido de la reconocida plataforma de datos públicos Kaggle. Específicamente, se seleccionó el dataset titulado "Superstore", importado por Ishan Shrivastava. Este conjunto de datos, conformado por 9994 observaciones y 21 variables, proporciona una rica fuente de información sobre las

ventas realizadas entre los años 2011 y 2015 pertenecientes a un supermercado de Estados Unidos. La elección de este dataset se basó en su relevancia para el abordaje de un análisis exhaustivo de mercados. Además, los 26 usuarios que han trabajado con este dataset es un punto más a favor de la relevancia del mismo para el análisis de mercados.

Con esta base de datos trataremos de identificar patrones y tendencias en el comportamiento de los beneficios y tiempos de entrega, con el fin de ilustrar al máximo las posibles estrategias de venta en el futuro.

3. Objetivos

3.1 Objetivos generales

El objetivo general de este trabajo es analizar una base de datos pública, con datos de compras realizadas en Estados Unidos durante un determinado periodo de tiempo, para investigar las tendencias de compra y profundizar en los factores que afectan a los tiempos de entrega y a los beneficios obtenidos en las ventas. Este análisis nos permitirá una comprensión profunda del comportamiento del mercado estadounidense, así como sacar a la luz las estrategias más efectivas para la optimización de los beneficios y tiempos de entrega.

3.2 Objetivos específicos

Como objetivos específicos trataremos de investigar:

1. Cómo varía el número total de ventas por año, estado y categoría.
2. Qué productos son los más vendidos en cada año.
3. Las diferencias en el tiempo de entrega con respecto a los productos, por año y estado.
4. Si el modo de envío afecta al tiempo de entrega del pedido o al tipo de cliente.
5. Estudiar si difieren los tiempos de entrega a lo largo de los meses o siguen un patrón.
6. Qué tipo de envío se utiliza más en cada región o estado.
7. Cuál es la relación entre la tasa de descuento aplicada y el beneficio obtenido, diferenciando por categoría y subcategoría de producto.
8. Estudiar si hay productos que proporcionan más rentabilidad que otros y si hay algunos que generan pérdidas.

4. Información disponible

La base de datos elegida, “Superstore”, ha sido extraída de la plataforma “Kaggle”. Ésta, fue subida el 10 de febrero del 2024, por Ishan Shrivastava. Recopila en una tabla de Excel toda la información relacionada con las ventas online de un supermercado estadounidense entre los años 2011 y 2015. Observamos que el total de filas o registros es 9994 y consta de 21 columnas o variables, por lo que afirmamos que no se trata de una base masiva, aunque

sí voluminosa. No tenemos información sobre cómo se han recopilado los datos, o si se trata de la totalidad o no de las ventas en ese periodo de tiempo.

Esta base de datos, ha sido utilizada 26 veces para diferentes estudios de mercado utilizando métodos de aprendizaje automático (machine learning) como árboles de regresión o validación cruzada, entre otros.
Como hemos afirmado anteriormente, contamos con 22 variables. Sin embargo, sólamente hemos utilizado 6 variables numéricas y 7 categóricas.

Entre las variables numéricas encontramos:

- OrderDate: Se trata de una variable tipo fecha, que muestra el día, mes y año en que el cliente realizó su pedido.

- ShipDate: Igual a la anterior, se trata de otra variable de tipo fecha en la que se recopila el día, mes y año en que el pedido llega al cliente.

- Sales: Precio unitario del producto vendido (en dólares).

- Quantity: Cantidad de unidades de producto vendido.

- Discount: Porcentaje de descuento aplicado en el producto para la venta.

- Profit: Margen de beneficio obtenido con la venta del producto.

Por otra parte, entre las variables categóricas podemos encontrar:

- ShipMode: Tipo de entrega, pudiendo diferenciar entre First Class y Standard Class, es decir, envío rápido o normal.

- Segment: Diferenciación de los clientes por segmento de mercado, pudiendo ser Consumer, Corporate o Home Office.

- City: Ciudad desde la que el cliente realiza la compra, también relacionada

- State: Estado al que pertenece la ciudad en la variable anterior.

- Región: Zona geográfica donde se encuentra la ciudad desde donde se realiza la compra (procedente de la agrupación de estados). De esta forma, diferenciamos tres regiones, Este, Central y Oeste.

- Category: Tipo de producto comprado.

- SubCategory: Categoría inferior o subtipo de producto. Viene anidada en Category.

Por otra parte, en la base de datos contamos con diversas variables que no utilizamos en el modelo final porque no añaden valor significativo al análisis o por su contenido confidencial, como RowID, OrderID, CustomerID, CustomerName, Country, PostalCode, ProductID y ProductName.

5. Metodología

En esta sección se presentan los procedimientos utilizados para el análisis, desde el exploratorio, al inferencial basado en modelización estadística, y se describe previamente cómo se ha realizado el procesado de los datos.

5.1 Procesado de los datos

Al procesar los datos hemos filtrado por los 5 estados con más ventas, para adaptar el análisis a las capacidades computacionales disponibles. Con esto, la base de datos se ha reducido de 9994 registros a sólo 5000. Además, hemos “creado” nuevas variables:

- Tiempo de entrega: extrayendo el mes y el año de la variable OrderDate (fecha de pedido);
- calculado el tiempo de envío, con la diferencia entre la fecha de entrega y la fecha de pedido;
- calculada la tasa de beneficio, dividiendo la variable Profit (beneficio) entre Sales (Ventas) y multiplicando por 100.

Es posible que existan patrones y tendencias específicas en los estados no incluidos en el análisis, que no se reflejarán en las conclusiones finales. Esto puede llevar a una comprensión incompleta del fenómeno estudiado.

Añadir también que nuestra base de datos no disponía de valores perdidos o faltantes, por lo que no ha sido necesario realizar ningún tratamiento previo al análisis para eliminarlos o imputarlos. Por otra parte, los posibles valores anómalos se identifican por medio de los gráficos descriptivos correspondientes.

5.2 Análisis estadístico

En el presente trabajo se lleva a cabo un análisis exhaustivo de las ventas de un supermercado a través de la aplicación de técnicas estadísticas de regresión. Con el objetivo de obtener una comprensión profunda de los datos y extraer información relevante, se desarrolla un proceso que combina el análisis exploratorio de datos y la construcción de modelos.

Por una parte, el análisis exploratorio de datos nos permitirá familiarizarnos con la estructura y características de nuestro conjunto de datos. Mediante la aplicación de diversas técnicas descriptivas y gráficas, buscaremos identificar patrones, tendencias y posibles relaciones entre las variables. Este primer acercamiento nos proporcionará una base sólida para formular hipótesis y seleccionar las técnicas de modelización más adecuadas.

En la segunda parte de este estudio, nos centraremos en la modelización estadística. A través de la construcción y evaluación de modelos de regresión, podremos cuantificar las relaciones entre las variables de interés y realizar inferencias sobre la población de estudio. La selección del modelo más apropiado se hará en base a los criterios propuestos a continuación.

5.2.1 Análisis exploratorio

Para una visualización más efectiva de los datos, se ha optado por una representación gráfica diversa. Los gráficos de barras se han utilizado para comparar las magnitudes de variables categóricas, como el total de ventas por producto, entre otras. Esta elección gráfica permite evaluar rápidamente las categorías con mayor o menor frecuencia.

Por otro lado, las variables continuas, como los tiempos de entrega, se han representado mediante boxplots. Estos diagramas proporcionan una visión general de la distribución de los datos, incluyendo la mediana, los cuartiles y los valores atípicos. De esta manera, es posible identificar si la distribución de los tiempos es simétrica.

Finalmente, para analizar la relación entre variables numéricas, como los descuentos y los beneficios, entre otras, se han utilizado tanto diagramas de cajas y bigotes como diagramas de dispersión. Los primeros permiten comparar los beneficios promedio asociados a diferentes niveles de descuento, mientras que los segundos revelan si existe una tendencia lineal o no lineal entre las variables.

5.2.2 Modelización estadística

La modelización estadística es una herramienta poderosa que nos permite comprender, analizar y predecir fenómenos complejos a partir de datos. Consiste en crear modelos matemáticos con incertidumbre, que representan la relación entre diferentes variables. A partir de estos modelos, es posible obtener conclusiones significativas y realizar inferencias sobre la población de estudio.

Tras el análisis exploratorio para identificar patrones, relaciones y posibles problemas, se elige un modelo estadístico, en nuestro caso de regresión, que se ajuste a la naturaleza de los datos y a las preguntas de la investigación. A continuación se estiman los parámetros del modelo a partir de los datos, se valida y se evalúa su capacidad explicativa para predecir los datos disponibles.Y por último, se utilizan los resultados de predicción para compararlo con otros modelos y elegir el mejor.

5.2.2 Modelización

Los modelos de predicción que utilizaremos, todos basados en la regresión, tienen como objetivo predecir el valor de una variable dependiente, denominada Y, basándose en el valor

de otras variables conocidas o independientes, recopiladas en una matriz predictora X. Los modelos que proponemos son los siguientes:

- **Regresión Lineal**. Está basada en asumir una relación lineal entre la respuesta Y y los predictores X. Cuando tenemos un único predictor hablamos de regresión lineal simple, y de múltiple cuando son dos o más los predictores, como es el caso en nuestro estudio, en el que además, los predictores son de tipo numérico y categórico. Trabajamos pues, con modelos de tipo ANCOVA. En estos modelos, la selección de variables es crucial para identificar los predictores más relevantes que afectan a la variable dependiente. Para ello, se utilizan técnicas de regularización como Lasso y Ridge. Además, con tal de cuantificar la variabilidad explicada por las variables predictoras, se emplea la tabla de ANOVA. Este tipo de modelo plantea la relación entre 2 o más variables independientes X_1, X_2, … , X_p , algunas de las cuales pueden ser variables dummy (0/1) derivadas de la descomposición de variables categóricas, y la correspondiente variable dependiente, Y, mediante la fórmula matemática siguiente:

  $$Y=\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n+\varepsilon$$

  Podemos interpretar los componentes de la fórmula como:

  Y: Variable dependiente
  β_0: Representa el valor de la Y cuando todas las X son cero.
  β_n: Coeficientes de regresión de las variables independientes que queremos estimar, y que explican la importancia de cada predictor en la estimación de la respuesta, o lo que es lo mismo, el grado de asociación lineal.
  X_n: Variables independientes,
  n= Número de variables independientes del modelo.
  ε:Variable aleatoria, residuo, épsilon o error. En este tipo de regresión, los residuos se comportan de forma aleatoria y normal con media cero y varianza 1.

  De forma matricial lo podemos expresar como:

  $$Y = X\beta + \varepsilon$$

- **Regresión Lasso.** Se trata de un método de estimación penalizada del modelo de regresión, que redunda en una reducción del número de variables en el modelo, y una selección de variables que no solapan información. Este enfoque es especialmente adecuado en problemas de multicolinealidad, donde los predictores están altamente correlacionados entre sí. Al seleccionar solo las variables más relevantes, el método ayuda a suavizar el impacto negativo de la multicolinealidad en el modelo, mejorando así la estabilidad y la interpretación de los coeficientes.

La estimación de la regresión Lasso se expresa a través de la minimización de la suma de cuadrados residual penalizada:

$$\sum_{i=1}^{n}(yi - \beta 0 - \sum_{j=1}^{p}\beta_j X_{ij}^2)+\lambda\sum_{j=1}^{p}|\beta j| = \text{suma residuos cuadrados} + \lambda\sum_{j=1}^{p}|\beta_j|$$

donde

λ: parámetro de regularización

En el modelo de regresión Lasso se realiza una validación cruzada con 10 particiones para encontrar el valor óptimo de λ, aquel que minimiza el error cuadrático medio (MSE). Este proceso asegura que el modelo no esté sobreajustado y selecciona automáticamente el nivel de penalización más adecuado.

Dado que un predictor con coeficiente de regresión cero no influye en el modelo, Lasso consigue excluir los predictores menos relevantes reduciendo hacia cero la estimación de los coeficientes. El grado de penalización está controlado por el hiperparámetro λ. Cuando λ=0, el resultado es equivalente al de un modelo lineal por mínimos cuadrados ordinarios. A medida que λ aumenta, mayor es la penalización y más predictores quedan excluidos.

- **Regresión Ridge**. Este tipo de regresión es conocida también por su capacidad para manejar la multicolinealidad, controlar el sobreajuste y reducir la complejidad del modelo, al disminuir los coeficientes hacia cero, sin eliminarlos por completo, a diferencia de la regresión Lasso. El grado de penalización está controlado por el hiperparámetro λ. Cuando λ=0, la penalización es nula y el resultado es equivalente al de un modelo lineal por mínimos cuadrados ordinarios. A medida que λ aumenta, mayor es la penalización y menor el valor de los predictores, aunque nunca llegan a ser 0.

La estimación de la regresión Ridge se expresa a través de la minimización de la suma de cuadrados residual, penalizada de la siguiente manera:

$$\sum_{i=1}^{n}(yi - \beta 0 - \sum_{j=1}^{p}\beta j\chi ij)^2 +\lambda\sum_{j=1}^{p}\beta j^2 = \text{suma residuos cuadrados} + \lambda\sum_{j=1}^{p}\beta j^2$$

donde

λ: parámetro de regularización

La principal ventaja de aplicar Ridge frente al ajuste por mínimos cuadrados ordinarios es la reducción de la varianza. Por lo general, en situaciones en la que la relación entre la variable respuesta y los predictores es aproximadamente lineal, las estimaciones por mínimos cuadrados tienen poco sesgo, pero aún pueden sufrir alta varianza. Este problema se acentúa conforme el número de predictores introducido en el modelo se aproxima al número de observaciones de entrenamiento, llegando al punto en que, si p>n, no es posible ajustar el modelo por mínimos cuadrados ordinarios. Empleando un valor adecuado de λ , el método Ridge es capaz de reducir varianza sin apenas aumentar el sesgo, consiguiendo así un menor error total.

La desventaja de aplicar Ridge es que, el modelo final incluye todos los predictores. Esto es así porque, si bien la penalización fuerza a que los coeficientes tiendan a 0, nunca llegan a ser cero, sólo si $\lambda=\infty$. Este método consigue minimizar la influencia sobre el modelo de los predictores menos relacionados con la variable respuesta pero en el modelo final, van a seguir apareciendo. Aunque esto no supone un problema para la precisión del modelo, sí lo es para su interpretación.

- **Árboles de regresión.** Son árboles de decisión que tienen como objetivo predecir una variable respuesta cuantitativa Y a partir de un conjunto de variables explicativas, que pueden ser cuantitativas o categóricas. Como su nombre indica, un árbol de regresión ayuda a tomar una decisión gracias a una serie de preguntas cuya respuesta, de sí o no, llevará a la decisión final. Un árbol de regresión se construye mediante la partición recursiva de la muestra en grupos cada vez más homogéneos, denominados nodos, hasta llegar a los "nodos terminales". La muestra se divide en función de los valores de una variable predictora, la cual se selecciona de acuerdo con un criterio de división binaria conun valor frontera que proporciona la máxima entropía entre nodos disjuntos, y la mayor homogeneidad entre las observaciones de cada nodo. Una vez que se ha construido un árbol, la respuesta para cualquier observación se puede predecir siguiendo la ruta desde el nodo raíz hasta el nodo terminal apropiado del árbol. Además, las variables que intervienen en la partición del árbol resultan ser las que se identifican como las más relevantes en la predicción de la respuesta.
  Por una parte, son robustos frente a datos atípicos y se pueden utilizar para la detección de los mismos. Por otro lado, son sensibles a los datos anómalos. Para combatirlo podemos utilizar un modelo combinando distintos árboles llamado random forest.

- **Random Forest.** Es una técnica de Machine Learning basada en el ajuste y combinación de una gran cantidad de árboles de clasificación. Cada árbol del conjunto puede tomar decisiones ligeramente diferentes debido a la aleatoriedad con que selecciona los datos y las variables que va a utilizar en las sucesivas particiones durante el entrenamiento. Cuando todos los árboles han sido entrenados, se combinan sus predicciones para generar una respuesta final, a partir del promedio de las estimaciones conseguidas en todos los árboles. Al utilizar un árbol pequeño se obtendrá un modelo con baja varianza y alto sesgo. Normalmente, al incrementar la complejidad del modelo se verá una reducción en el error de predicción debido a

un sesgo más bajo en el modelo, hasta llegar a un punto en el que éste será muy complejo y se producirá un sobreajuste del modelo, el cual empezará a sufrir de varianza alta. El modelo óptimo debería mantener un balance o equilibrio entre estos dos tipos de errores.
Con esto, podemos asumir que debido a su aleatorización y al hecho de que cada árbol se entrena en un subconjunto diferente de datos (y de variables), el modelo es más robusto frente al ruido y las anomalías en los datos.

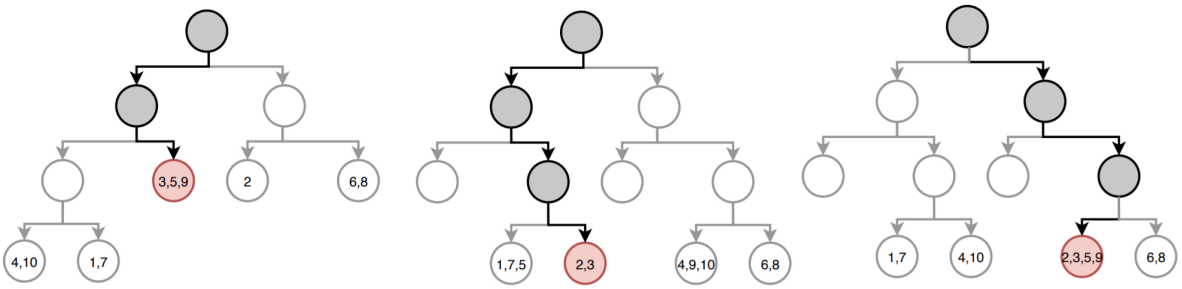

Figura 1: *Esquema explicativo de los árboles de regresión.*

5.3.3 Métricas de evaluación

Con el fin de evaluar la calidad y la precisión de un modelo de aprendizaje automático como son los modelos de regresión propuestos, empleamos métricas de evaluación. Estas nos proporcionan una puntuación numérica que indica cómo de bien predice nuestro modelo los datos observados. Al comparar estas métricas, podemos identificar cuál de los modelos ajustados tiene una mayor calidad predictiva. Las métricas más comunes para comparar modelos en regresión son el error cuadrático medio y el coeficiente de determinación. También utilizamos el criterio AIC para comparar modelos con una misma formulación paramétrica, de la regularización en los modelos de regresión como método para seleccionar variables, y de los índices de importancia en los bosques aleatorios, para identificar las variables más relevantes en la predicción.

MSE o Error Cuadrático Medio
El error cuadrático medio o MSE es la desviación estándar de los valores residuales, que se calculan como la diferencia entre los valores observados y los valores predichos del modelo. Estos, son una medida de la distancia que nos sirve para cuantificar cuánto se desvían nuestras predicciones de los datos disponibles. El objetivo con este indicador es reducirlo al máximo.

El MSE se calcula mediante la siguiente fórmula:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

donde y_i representa cada uno de los datos observados en la variable respuesta e $\widehat{y_i}$ su predicción.

Sabiendo esto, podemos concluir que un MSE más bajo indica que las predicciones del modelo están, en promedio, más cerca de los valores reales, lo que sugiere un modelo más preciso. Por el contrario, un MSE más alto sugiere que las predicciones del modelo están, en promedio, más lejos de los valores reales, por lo que podemos deducir que el modelo no será muy preciso. No existe sin embargo, una medida de referencia para juzgar la validez de un ajuste, si bien resulta útil para comparar ajustes con modelos distintos sobre un mismo banco de datos.

R2

El coeficiente de determinación R2 es la proporción de la varianza total explicada por la regresión. Este coeficiente, también llamado R cuadrado, refleja pues, la bondad del ajuste de un modelo. Cuanto mayor sea el R2, mayor será la variabilidad explicada por el modelo de regresión, considerando que el valor oscila entre 0 y 1.

- **R^2=1:** El modelo explica toda la variabilidad de la variable dependiente alrededor de su media. Un R^2 de 1 indica un ajuste perfecto del modelo a los datos.
- **R^2=0:** El modelo no explica nada de la variabilidad de la variable dependiente alrededor de su media. Un R^2 de 0 indica que el modelo no tiene ningún poder explicativo.
- **0 < R^2 < 1:** El modelo explica parcialmente la variabilidad de la variable dependiente.

El R^2 se calcula mediante la siguiente fórmula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \widehat{y_i})^2}{\sum_{i=1}^{n} (y_i - \bar{y}i)^2}$$

donde $\bar{y}i$ es el promedio de todas las observaciones de y_i.

El problema del coeficiente de determinación es que su valor aumenta cuando lo hace el número de predictoras, aunque estas sean poco relevantes.

5.3.4 Selección de variables

Puesto que es un objetivo de este estudio conseguir modelos sencillos con los que predecir los tiempos de entrega y las tasas de beneficio por la venta de productos, es relevante tener

en consideración cómo vamos a eliminar variables que sean redundantes o no expliquen suficientemente la respuesta. En función del modelo, discriminamos las variables relevantes y no relevantes mediante: el criterio AIC y la regularización en el modelo de regresión, la propia construcción del árbol de decisión, que no contiene variables no relevantes en las particiones, y los índices de importancia en los bosques aleatorios.

Criterio AIC

El criterio de información de Akaike (AIC) busca un equilibrio entre el número de datos disponibles, el ajuste obtenido, en términos de verosimilitud, y el número de parámetros utilizados. Con esto, buscamos el mejor modelo eliminando y añadiendo variables mediante un proceso de selección hacia delante y hacia atrás. Aplicar este criterio nos facilita encontrar un modelo que se ajuste bien a los datos sin ser demasiado complejo.

La fórmula del AIC es la siguiente:

AIC = 2k - 2Ln($\widehat{L}$),

donde

k: número de parámetros del modelo
$\widehat{L}$: es el máximo valor de la verosimilitud estimada para el modelo

El término 2k penaliza los modelos más complejos, es decir, con más parámetros. Además, por el término 2Ln($\widehat{L}$), a menor valor, mejor será el ajuste del modelo a los datos.

Regularización

Por otra parte, la elección del valor óptimo de λ en los modelos Lasso y Ridge es fundamental para obtener un buen rendimiento del modelo, por lo que elegiremos un modelo de selección que se adapte bien a nuestros datos, como la validación cruzada. Con este método, se divide el conjunto de datos en varios subconjuntos. Para cada valor de λ, se ajusta el modelo en uno de ellos y se evalúa sobre los restantes. Finalmente se promedian los resultados de validación, una vez se ha ajustado el modelo sobre cada uno de los subconjuntos de validación. El valor de λ que produce el mejor rendimiento en la validación cruzada se selecciona como óptimo.

En resumen, tanto Lasso como Ridge son técnicas de regularización que ayudan a prevenir el sobreajuste en modelos lineales. La principal diferencia entre ellas radica en la forma en que penalizan los coeficientes:

- Por una parte, Lasso realiza una selección de variables más agresiva, llegando a reducir los coeficientes a cero y mejorando la interpretabilidad del modelo.

- Por otra parte, Ridge reduce el tamaño de todos los coeficientes, pero no realiza una selección tan agresiva, pues nunca llegan a ser 0. Mejora la estabilidad del modelo y puede ser útil cuando hay alta correlación entre las variables predictoras.

Árboles e índices de importancia

En los árboles de regresión la selección de variables es automática, pues sólo intervienen en el árbol si aportan información al dividir la muestra, teniendo en cuenta la que aportan el resto de variables. Uno de los problemas que podemos encontrarnos es el sobreajuste. Esto ocurre cuando un modelo se ajusta demasiado bien a los datos de entrenamiento, capturando incluso el ruido aleatorio, lo que resulta en un modelo que no generaliza bien a nuevos datos. Cuando esto sucede, utilizamos la llamada poda de los árboles, que consiste en recortar o eliminar los nodos del árbol que no contribuyan significativamente a la predicción. Esto mejora la interpretabilidad del árbol.

El índice de importancia en un árbol de regresión es una métrica que nos indica cuán relevante es cada variable independiente en la predicción de la variable dependiente. En otras palabras, nos dice qué variables están contribuyendo más a la construcción del modelo y a la calidad de las predicciones.Existen diferentes formas de calcular el índice de importancia, pero una de las más comunes en árboles de regresión está basada en la reducción de la impureza. La impureza mide la heterogeneidad de los valores de la variable objetivo dentro de un nodo del árbol. Al dividir un nodo en dos subnodos, se reduce la impureza. La importancia de una variable se calcula sumando la reducción de impureza en todas las divisiones del árbol donde se utilizó esa variable. Que una variable tenga una alta importancia significa que al dividir los datos en función de los valores de esa variable, se logra una reducción significativa en la impureza de los nodos hijos. Esto indica que esa variable es muy útil para separar las observaciones en grupos con valores similares de la variable objetivo.

La importancia de las variables se suele representar gráficamente mediante un gráfico de barras. La altura de cada barra indica la importancia relativa de la variable correspondiente. Esta medida es muy relevante para describir las predictoras más relevantes en los random forests.

5.3 Software y hardware.

Para realizar este estudio ha sido necesario del uso de Rstudio, un editor de un lenguaje de programación especializado en análisis de datos y R. De estos programas, hemos utilizado las versiones 2024.9.0.375 y 4.4.1, respectivamente. Dentro de Rstudio, se han utilizado un conjunto de librerías específicas para cada objetivo y su respectivo gráfico.

Las librerías que hemos utilizado según la funcionalidad son:

- Con el fin de importar la base de datos desde una hoja de Excel: *readxl*.
- Para extraer los descriptivos: *summarytools, psych, skimr, janitor* y *rstatix*,

- Para la personalización de las tablas en las que incluir estos descriptivos: *KableExtra*, *flextable.*
- Con tal de facilitar la manipulación de los datos: *dplyr* y *tidyr*
- Para centrarnos específicamente en las variables con formato fecha/hora: *lubridate*
- En el momento de representar gráficamente los datos y personalizarlos: *ggplot2, plotrix* y *RColorBrewer*,
- Para organizar los gráficos en una sola página: *gridExtra*.
- Para los árboles de regresión: *tree* y *rpart*.
- Bosques aleatorios: *sample, randomForest, ranger, caret, h2o.*

6. Resultados

Se presentan en este apartado los resultados obtenidos al llevar a cabo el análisis exploratorio y la modelización estadística para responder a los objetivos planteados.

6.1 Estadísticas descriptivas

En este apartado se muestran un conjunto de gráficos utilizados para presentar de manera clara y concisa la información contenida en el conjunto de datos. Con esto, facilitamos la identificación de patrones mediante parámetros de tendencia central y métricas y detectamos anomalías.

Ventas

Representamos en la Figura 2 el número total de ventas realizadas entre 2011 y 2014 para los cinco estados considerados en nuestro estudio: Washington, Texas, Pennsylvania, New York y California, que se corresponden con los estados que más ventas acumulan de modo individualizado y que representan a un 20% de la base de datos total.

Observamos que el estado con más ventas es California, con una diferencia de casi 1000 ventas con el segundo estado, que es New York. El siguiente, a la par con el segundo, es Texas, y por último, Pennsylvania y Washington.

Esta distribución sugiere una posible relación entre la cantidad de ventas y la población de cada estado, ya que California, Texas y Nueva York son los estados más poblados del país, lo que implica una mayor cantidad de potenciales compradores. Por otro lado, Pennsylvania y Washington, que tienen menos población en comparación con los anteriores, presentan un menor número de ventas, lo que refuerza la idea de a mayor población, mayores son las ventas.

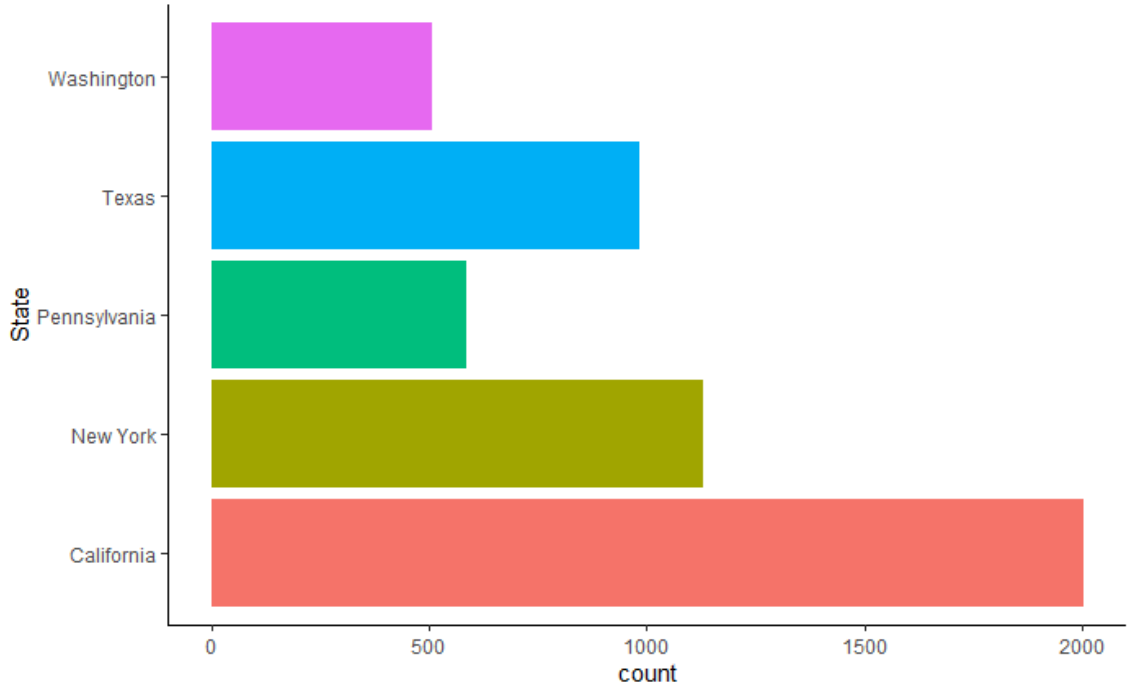

Figura 2: *Total de ventas realizadas entre 2011 y 2014 en cada uno de los estados en la base de datos.*

Observamos en la Figura 3, que la tendencia general de las ventas entre los estados es alcista. Aún así, en 2013 vemos que Washington decae hasta no tener casi ventas y New York empieza a aumentar sus ventas ya en el último año. Por otra parte, las ventas de Texas decaen en 2012 y se mantienen durante los dos años siguientes. En Washington se registra una caída considerable en 2013, seguida de un crecimiento relevante en 2014.

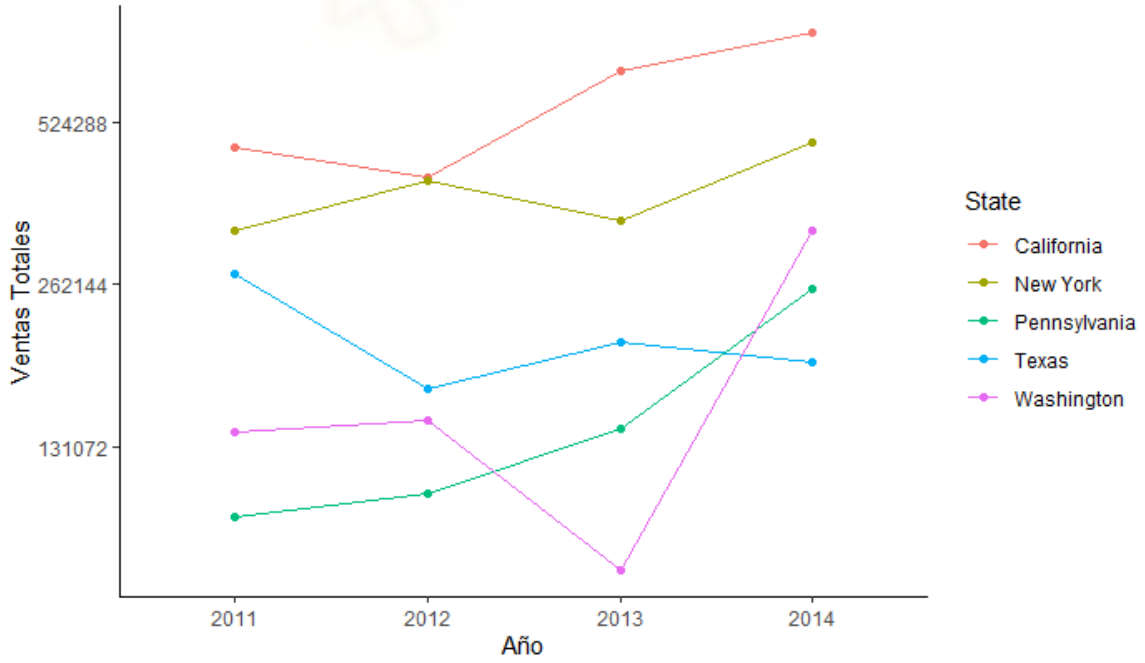

Figura 3 *Número de ventas realizadas por año entre 2011 y 2014 en cada uno de los estados en la base de datos.*

La Figura 4 nos muestra el total de ventas realizadas en las 8 ciudades con más registros de los estados mencionados anteriormente. Podemos ver que New York City, seguida de Los Ángeles, es la ciudad con más ventas realizadas. Lo cual es coherente con el hecho de que ambas son las ciudades más pobladas de Estados Unidos. Vemos que las siguientes tienen aproximadamente el mismo número de ventas, menos las dos últimas, que distan bastante del resto.

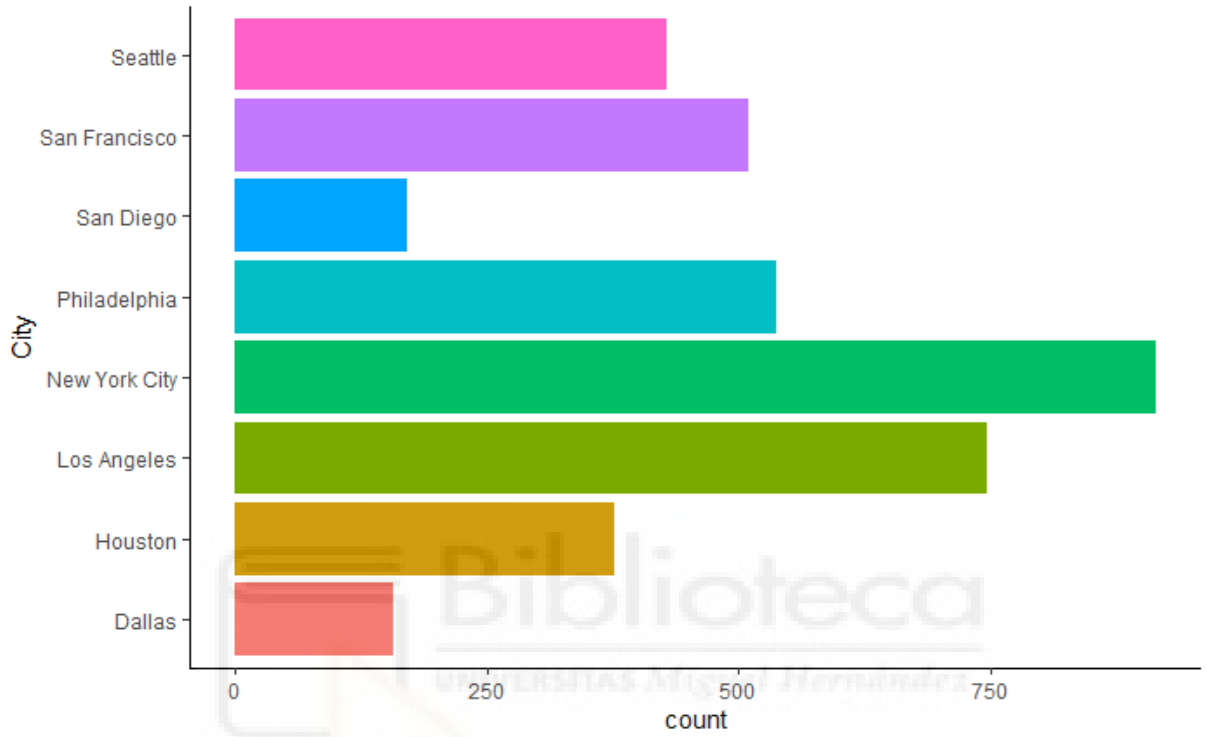

Figura 4: *Número de ventas realizadas por año entre 2011 y 2014 en las 8 ciudades con más ventas en la base de datos.*

En la Figura 5 se presentan con gráficos de barras, el total de ventas realizadas por año en cada una de las categorías de producto consideradas, y se han añadido los porcentajes relativos que representan dichos números respecto de las ventas realizadas cada año. Se observa una tendencia ascendente en el volumen absoluto de ventas en todas las categorías, si bien los porcentajes por categorías se mantienen estables. Vemos que el comportamiento de las ventas con respecto a la categoría de producto no varía demasiado a lo largo de los años: el mayor volumen de ventas se da en productos de Office Supplies, en torno al 60% cada año, seguido de Furniture en torno al 22%, y Technology, en torno al 18%.

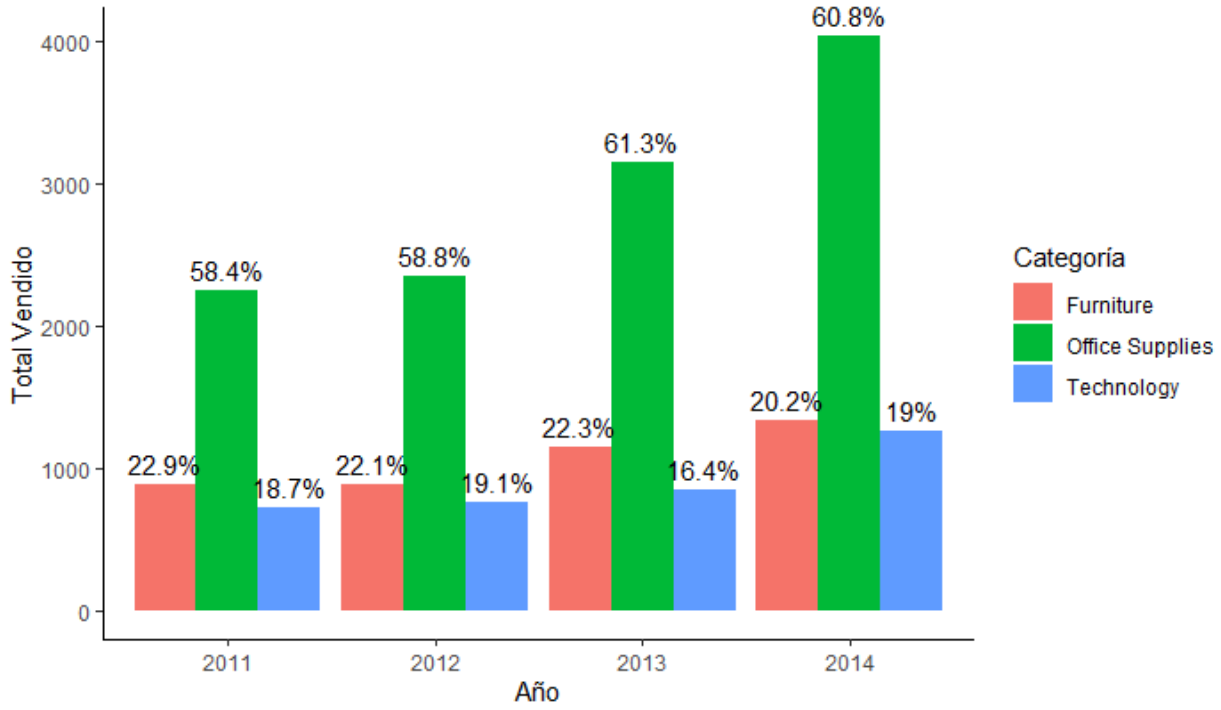

Figura 5: *Total de ventas realizadas entre 2011 y 2014 en las 6/8 ciudades con más ventas.*

Tiempos de entrega

En la Figura 6 se muestra el tiempo de entrega a lo largo de los años, diferenciado por categoría de producto. Podemos afirmar que no muestra asociación alguna con la categoría en que se clasifican los productos vendidos, ni varía a lo largo de los años, pues es igual todos los años en las tres. El mínimo se encuentra en 0 días y el máximo en 7. El percentil 50 se encuentra en las tres categorías en 4 días y los percentiles 25 y 75, en 3 y 5, respectivamente. El único cambio remarcable se encuentra en el percentil 25 del año 2014 de la categoría Furniture, pues disminuye a 2 días.

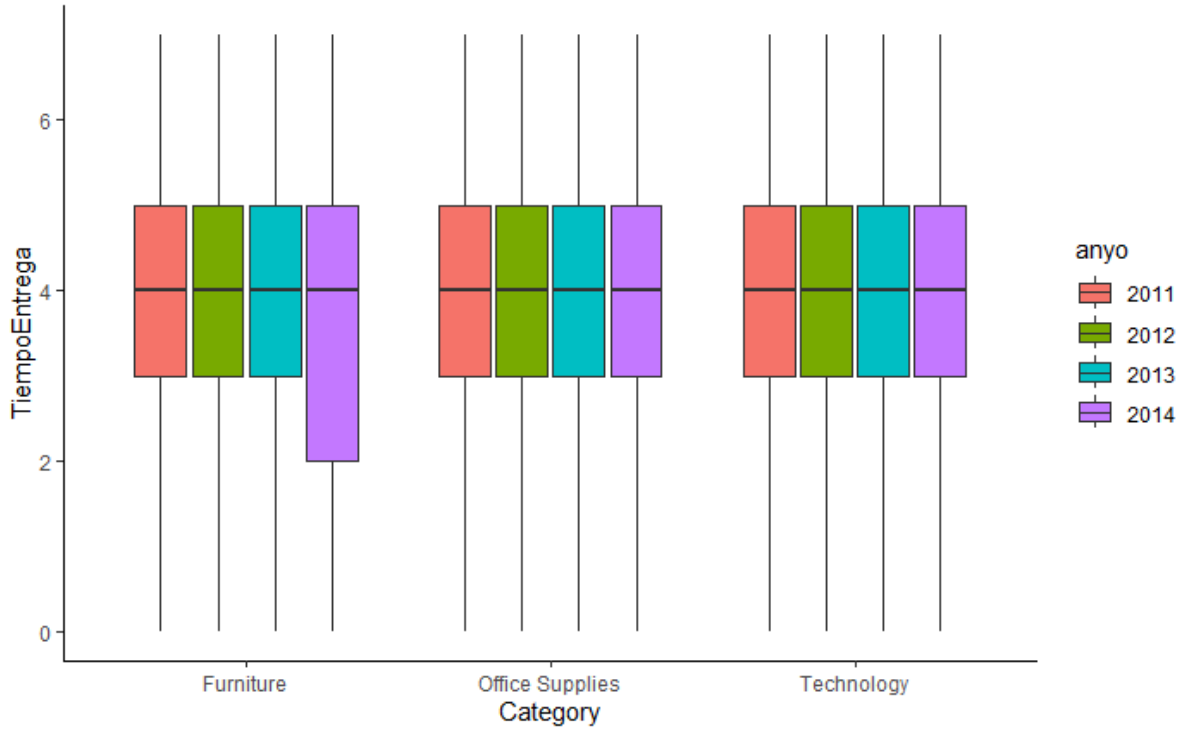

Figura 6:*Tiempo de entrega correspondiente a las ventas por año entre 2011 y 2014 en cada categoría de producto.*

Para estudiar los tiempos de entrega respecto de los estados, visualizamos estos en la Figura 7, durante los cuatro años de observación, El tiempo de entrega con respecto al año y al estado, tampoco sufre muchas variaciones. El máximo y el mínimo se encuentran en 0 y 7 días, respectivamente y en la mayoría de ellos, el percentil 75 se encuentra en 5 días, aunque el 25 varía un poco más, entre 2 y 3 días. La mediana se mantiene siempre en 4 días. En 2013 podemos observar un comportamiento extraño en el estado de Washington, pues aumenta el tiempo de entrega respecto de los demás estados, entre 4 y 6 días. Podemos encontrar este mismo comportamiento en New York en 2014.

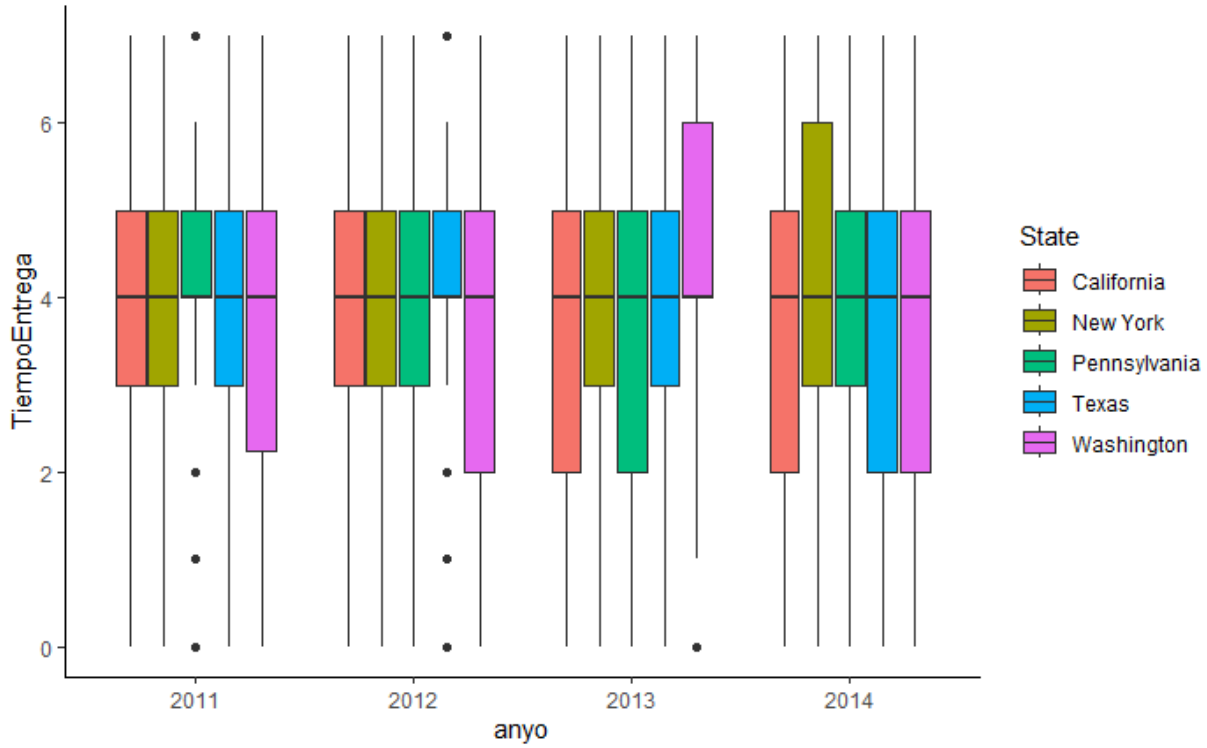

Figura 7:*Tiempo de entrega correspondiente a las ventas por año entre 2011 y 2014 en cada estado de la base de datos.*

En la Figura 8 mostramos el tiempo de entrega en función del tipo de envío. Observamos que el tipo de envio básico o estándar, es el que más tarda en llegar al consumidor (entre 4 y 7 días), luego se encuentra el Second Class, que rebaja un poco el tiempo (entre 2 y 4 días), el tiempo de entrega de primera clase responde siempre en 3 días como máximo, y la entrega en el mismo día es efectiva, salvo una excepción registrada, que se ha demorado en un día.

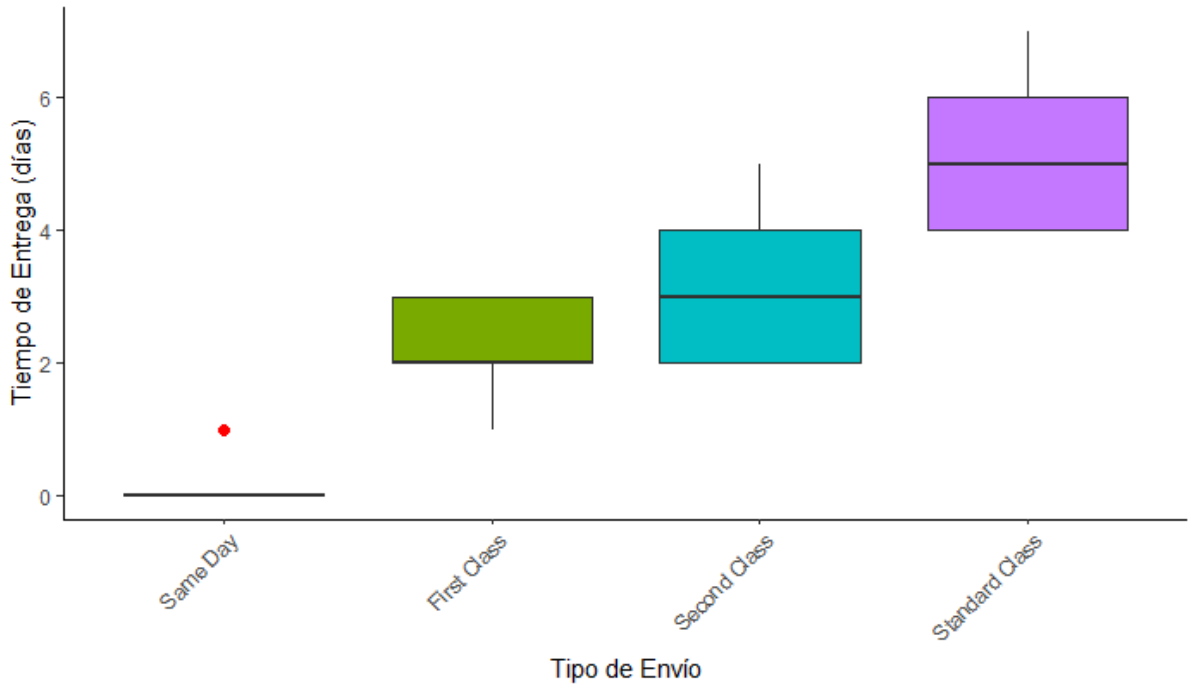

Figura 8: *Tiempos de entrega en función del tipo de entrega seleccionado para cada venta.*

En la Figura 9, se representa el tiempo de entrega respecto del tipo de cliente, que no cambia mucho. La mediana está en los 4 días, si bien el 25% de los clientes más afortunados en el tipo Home Office consiguen recibir su pedido antes de 2 días, cuando para el resto de clientes se han entregado antes de 3 días.

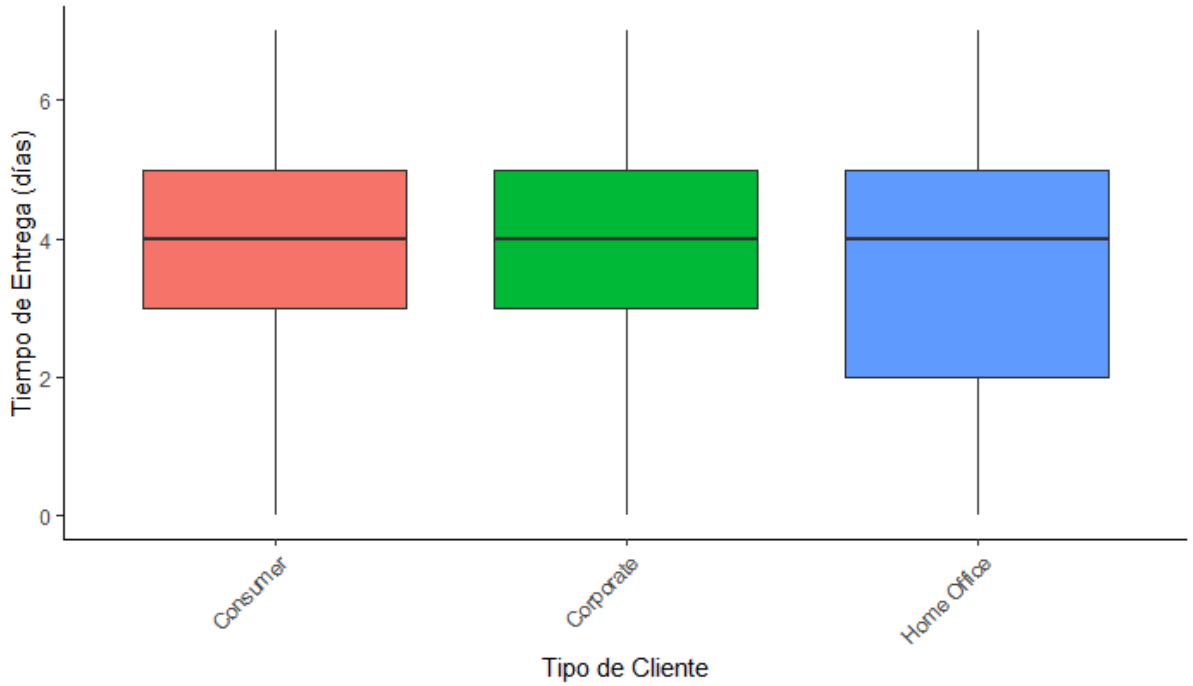

Figura 9: *Tiempos de entrega en función del tipo de cliente que ha realizado la compra..*

En la Figura 10 se muestra el tiempo de envío por meses a lo largo de los años. A simple vista observamos la estabilidad de las medianas y los diferentes rangos de variación a lo largo de los meses Por lo general agosto tiene más variabilidad en los dos últimos años (las cajas son más altas) y diciembre en los 3 primeros. Estas variaciones podrían venir condicionadas por los períodos vacacionales de verano y Navidad.

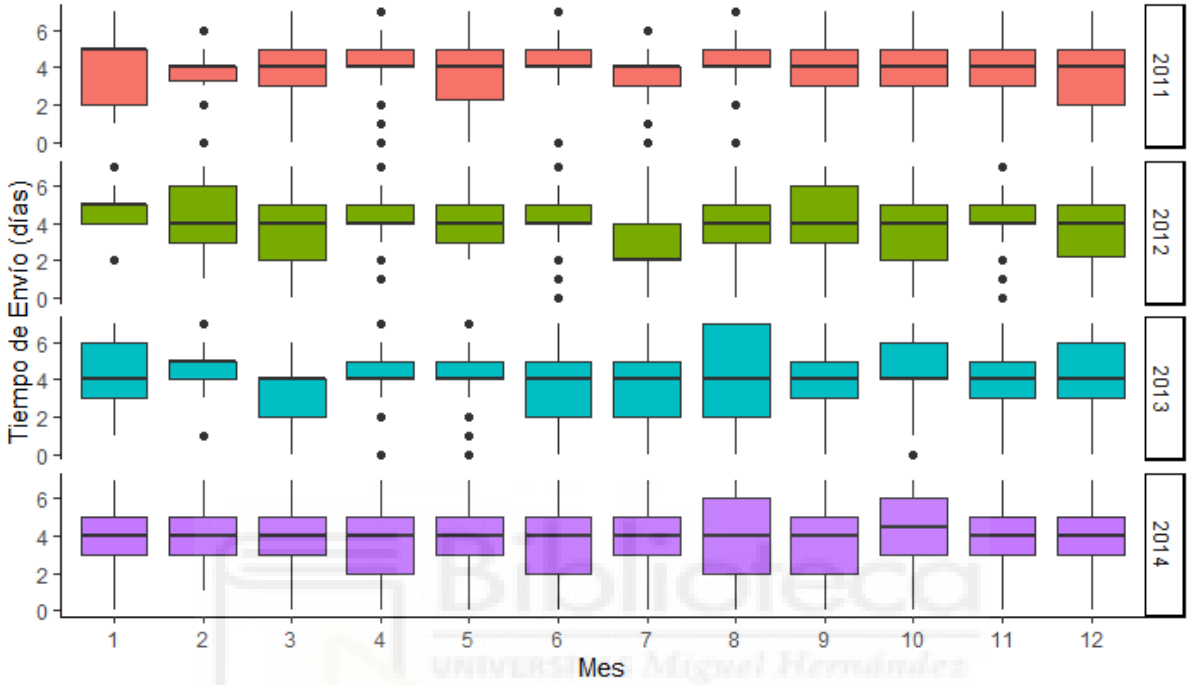

Figura 10: *Tiempos de envío en función del mes, diferenciado por años.*

Descuentos

En la Figura 11 se muestra el descuento aplicado a cada categoría de producto diferenciando por años. Podemos visualizar que Furniture (rojo) presenta una mayor dispersión en los descuentos en comparación con las otras categorías. También tiene valores atípicos más altos en cada año, lo que indica que algunos productos han recibido descuentos muy elevados.
Office Supplies (verde) y Technology (azul) constan de distribuciones más compactas, con menor variabilidad en los descuentos y menos valores atípicos en comparación con Furniture.

No se observa una tendencia clara de aumento o disminución en los descuentos con el paso de los años. La distribución de los descuentos se mantiene relativamente constante en cada categoría.

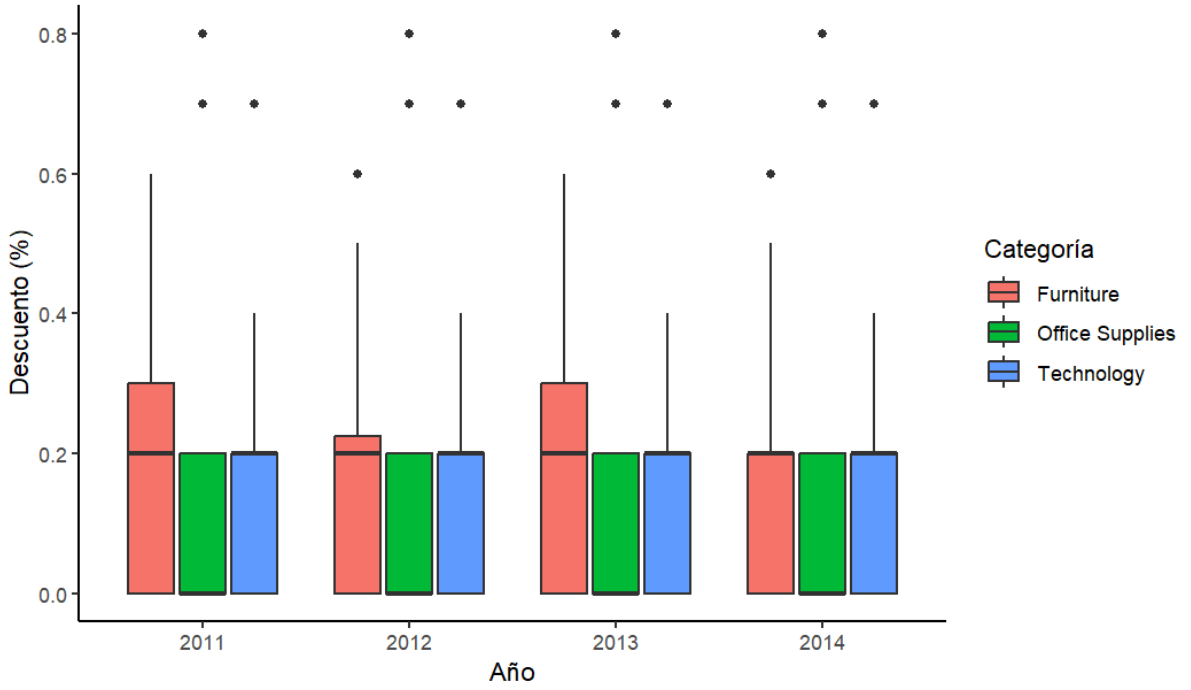

Figura 11: *Descuento por categoría de producto para cada año.*

Aunque no se observan diferencias notables entre categorías, en la Figura 12 se puede observar que en la mayoría de las subcategorías se aplican descuentos entre el 0% y el 20%. Sin embargo, Binders y Machines destacan por tener los descuentos más altos, llegando a alcanzar el 80%.

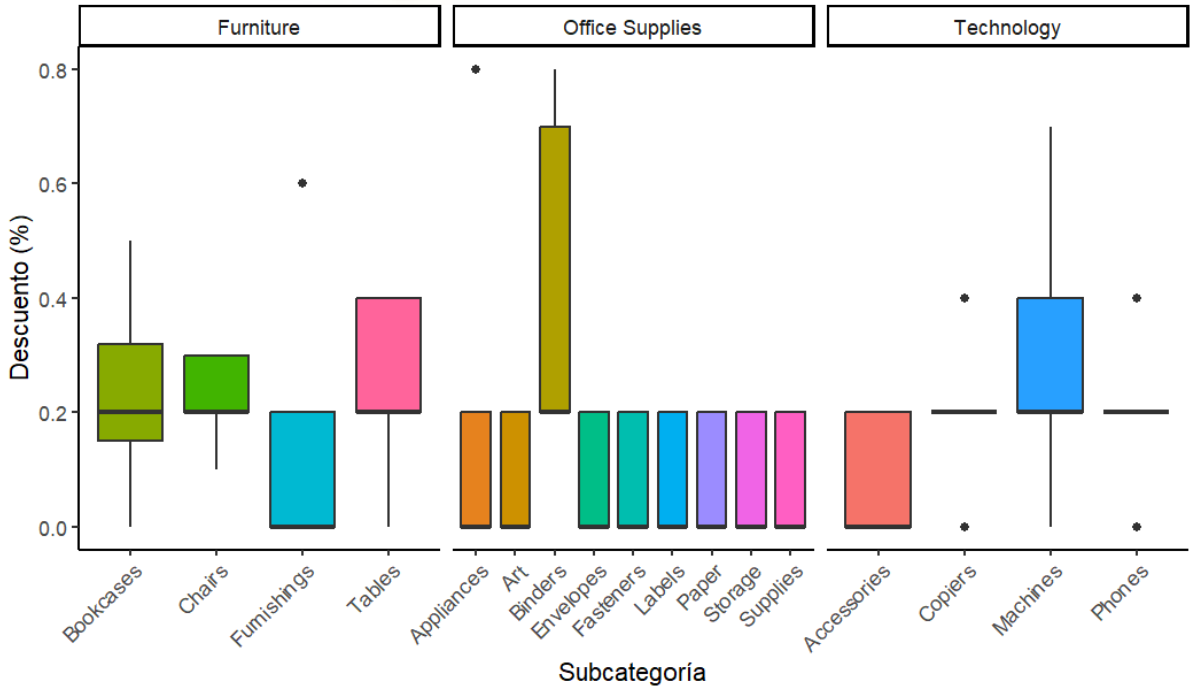

Figura 12: *Descuento por subcategoría de producto agrupadas por categoría.*

Ganancias y pérdidas

La figura 13 muestra el importe total de ganancias y pérdidas acumuladas durante los años 2011-2014, con la venta de productos en cada una de las categorías. Podemos observar que Technology es la categoría que más beneficios da (en verde) con respecto a las pérdidas (en rojo), que a diferencia de Furniture, que compensa los beneficios con las pérdidas.

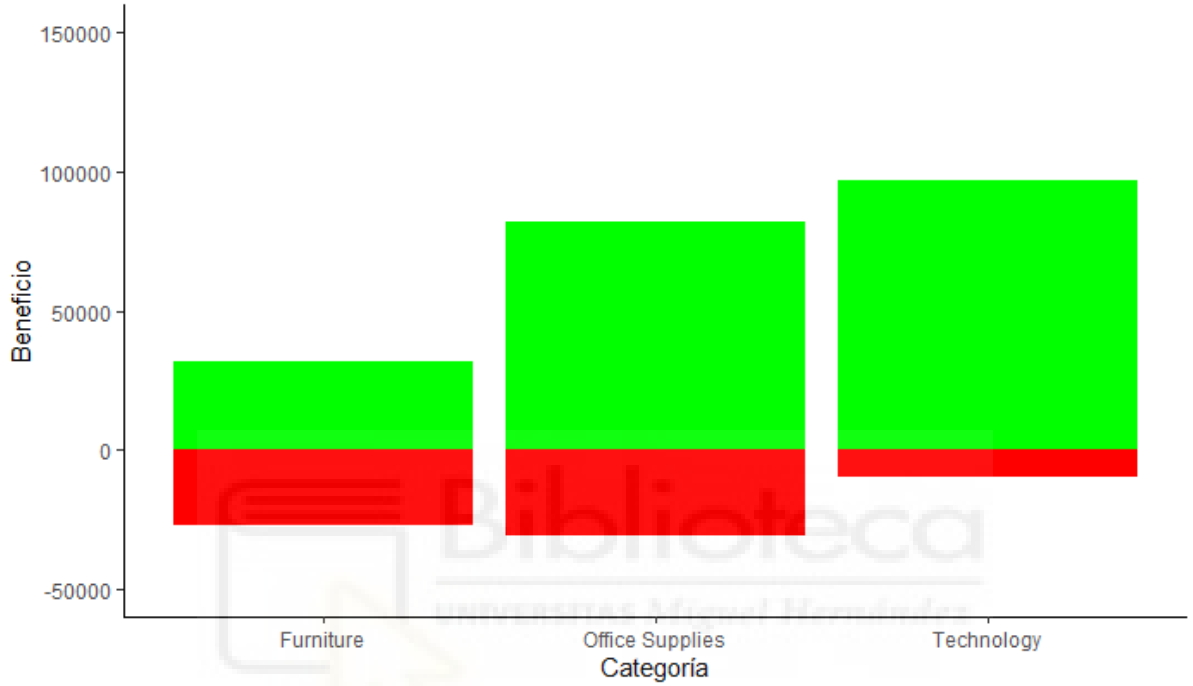

Figura 13: *Importe total de ganancias y pérdidas correspondientes a cada categoría de producto durante los años 2011-2014.*

La figura 14 muestra 3 gráficos comparativos en los que vemos el importe total de las ventas, el importe total y el beneficio obtenido con las ventas por cada categoría de producto. Technology, aunque es de la que menos ventas se han realizado, es la que más beneficios proporciona, seguida de Office Supplies y por último, Furniture. Office supplies es donde más productos se han vendido y más dinero se ha facturado, generando unos beneficios claramente inferiores a los de Technology.

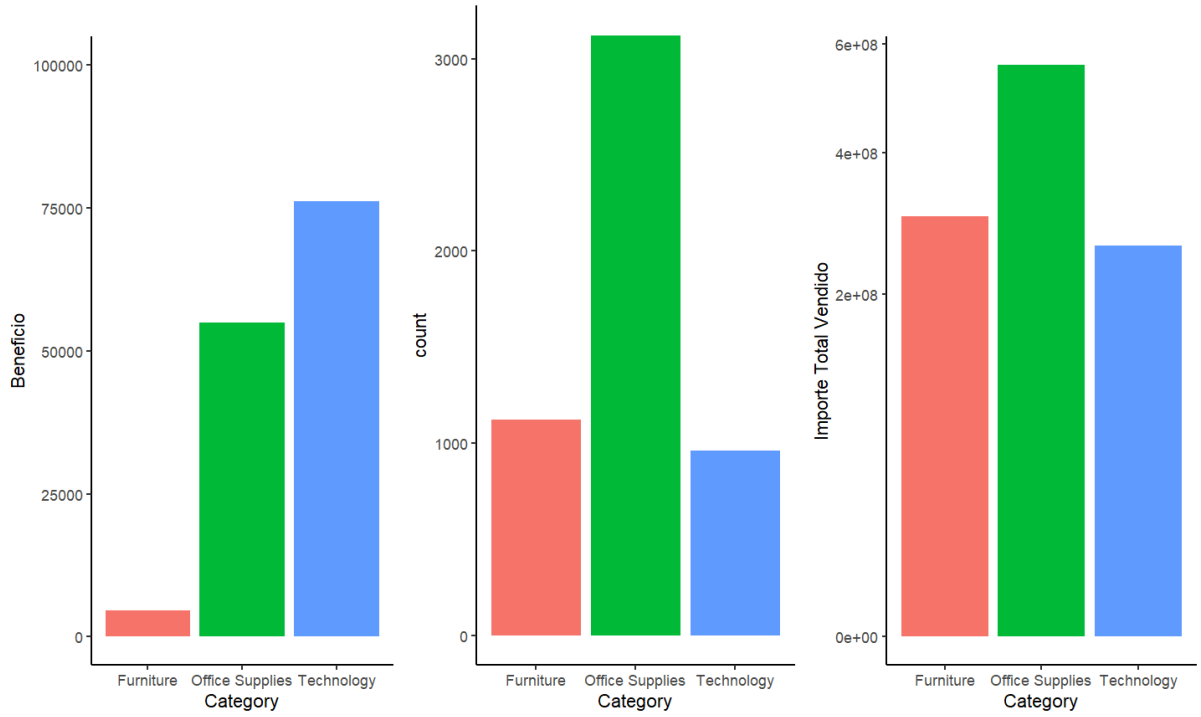

Figura 14: *Beneficios netos, importe total y ventas totales correspondientes a cada categoría de producto.*

En la Figura 15 se presenta el importe acumulado en el período de estudio, en cada una de las subcategorías de productos. A simple vista observamos que en general los beneficios superan a las pérdidas en todas las subcategorías, excepto en: Tables, Supplies y Bookcases, donde las pérdidas son mayores a las ganancias. Los productos que más ganancias producen son: Phones, Machines, Copiers, Binders y Accesories.

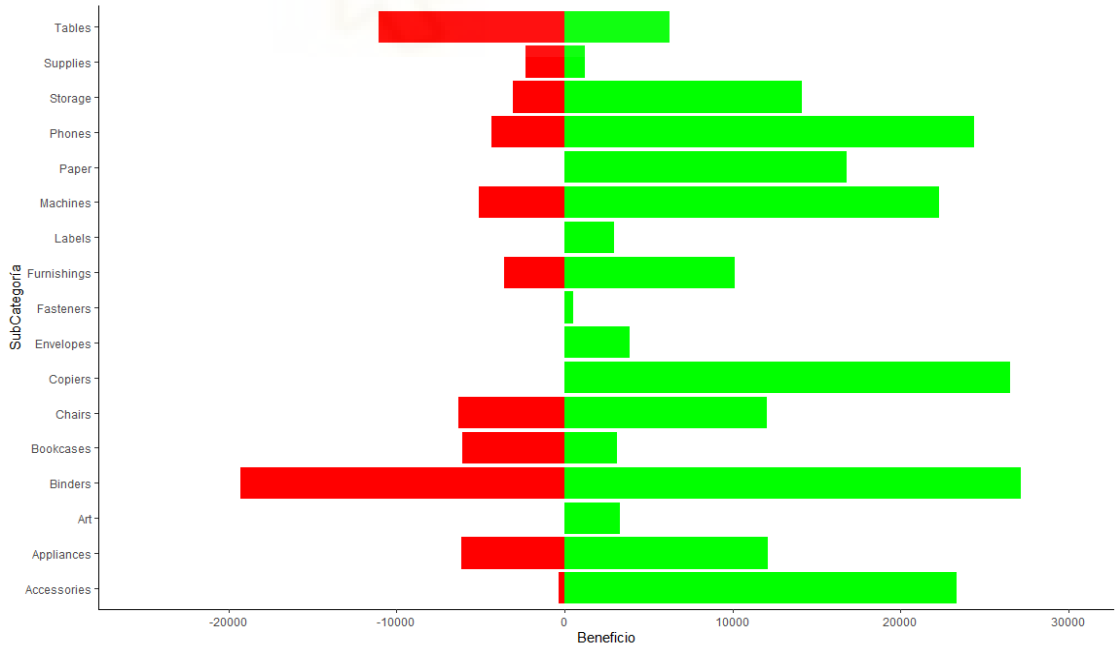

Figura 15: *Ganancias y pérdidas correspondientes a cada subcategoría de producto.*

La figura 16 muestra 3 gráficos en los que observamos el importe total de las ventas, el número de ventas y el beneficio obtenido con las ventas por cada categoría de producto. Podemos decir que Copiers, aunque es de las subcategorías con menos ventas, es la que más beneficio proporciona. En caso contrario encontramos Tables, Supplies y Bookases. Son subcategorías que se venden pero generan pérdidas.

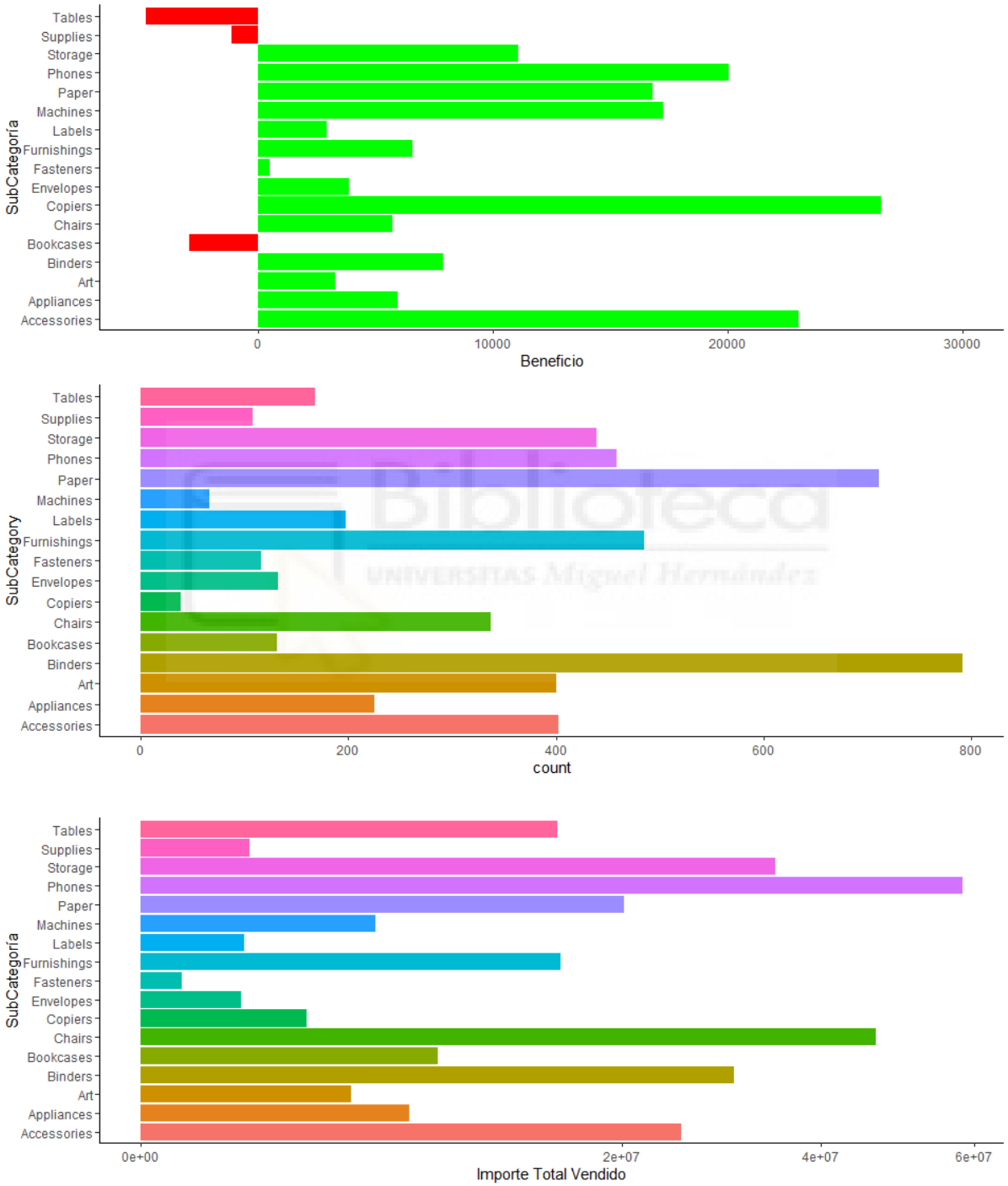

Figura 16 :*Beneficios netos correspondientes a cada subcategoría de producto.*

Por categoría, en la Figura 17 podemos ver que los productos tecnológicos (en azul), cuanto mayor es el descuento, menor es el beneficio. Además, el material de oficina (en verde) se distribuye preferentemente en los extremos, para descuentos del 0 al 20% y del 70 al 80%, generando pérdidas en las ventas para estos últimos dos tipos de descuento. Por otra parte, los descuentos aplicados para el mobiliario (en naranja) son como máximo del 60%, pero ya generan básicamente pérdidas con descuentos a partir del 30%. Vemos que, en general, cuanto mayor es el descuento, menores son los beneficios, es decir, se gana mucho más cuando no se aplica descuento o éste es muy bajo. Cuando los descuentos son relativamente altos, se empiezan a generar pérdidas. Descuentos muy bajos o inexistentes pueden llegar a producir beneficios de +3000 en algún producto y del mismo modo, con descuentos del 80% se generan pérdidas de - 3000. En este caso hemos eliminado estos valores para mejor visualización del gráfico.

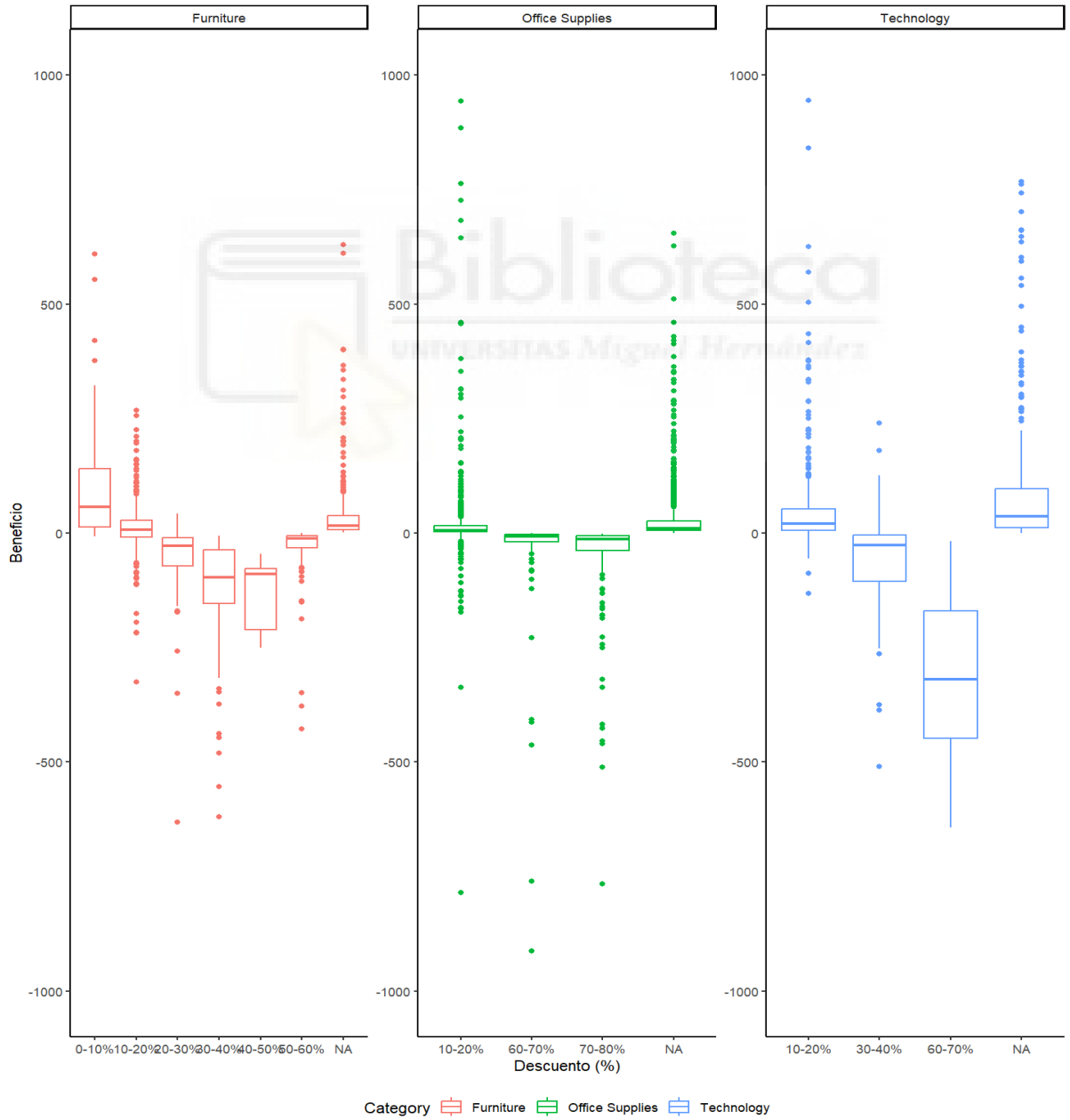

Figura 17: *Beneficio en función del descuento aplicado en cada categoría de producto.*

Conclusión de las estadísticas descriptivas

Como conclusión general podemos afirmar que la población de los estados es relevante a la hora de analizar el número de ventas, cuanta más población tenga el estado, éstas serán mayores. Además, a lo largo del tiempo los datos son constantes y no hay apenas diferencias entre años. En cuanto a tiempos de entrega, lo que más influye es el modo de envío asignado a la compra. Esta variable afecta de forma muy directa aumentando o disminuyendo los tiempos dependiendo del modo escogido. Mientras que “same day” hace que el pedido llegue en el mismo día al consumidor, como su nombre indica, las demás variables aumentan en uno o más días el envío.

6.2 Modelización estadística

En este apartado, con el objetivo de predecir los tiempos de entrega y la tasa de beneficio en función de las variables que les afectan, vamos a realizar un modelado estadístico utilizando los cinco modelos explicados anteriormente: la regresión lineal múltiple, las regresiones ridge y lasso, los árboles de regresión y los bosques aleatorios. Cada uno de estos modelos ha sido ajustado a los datos históricos con el propósito de capturar las tendencias de las series de tiempo de las variables de interés.

A continuación, mostramos en detalle los resultados de cada modelo, y acabamos comparando en una tabla todos ellos y concluyendo sobre los mejores, tanto para los tiempos de entrega como las tasas de beneficio.

6.2.1. Tiempo de entrega

Primero presentaremos los modelos para la variable dependiente Tiempo de Entrega y entenderemos cuál se ajusta mejor a los datos fijándonos en las métricas elegidas.

6.2.1.1. Regresión lineal múltiple.

Hemos intentado ajustar un modelo en el que se prediga la variable dependiente TiempoEntrega en función de las variables independientes categóricas de las que disponemos.

TiempoEntrega ~ mes + SubCategory + anyo + State + Category + Segment + ShipMode.

Aplicando el modelo de Akaike, el modelo mejor ajustado queda:

TiempoEntrega ~ City + State + Segment + ShipMode

La tabla ANOVA (Figura 18) muestra que además todas las variables seleccionadas aportan diferencias significativas entre algunos de sus niveles:

Podemos decir que la ciudad, el estado, el segmento y el modo de envío influyen significativamente en el tiempo de entrega del producto.

```
Analysis of Variance Table

Response: TiempoEntrega
            Df Sum Sq Mean Sq   F value    Pr(>F)
City       172 1372.5     8.0    8.6135 < 2.2e-16 ***
State        4   14.1     3.5    3.8096  0.004269 **
Segment      2   61.3    30.6   33.0786 5.345e-15 ***
ShipMode     3 9702.8  3234.3 3491.2583 < 2.2e-16 ***
Residuals 5022 4652.3     0.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figura 18: *Análisis de la varianza relativa al modelo de regresión que explica el tiempo de entrega.*

6.2.1.2. Regresión Lasso.

Después de aplicar la penalización, observamos los coeficientes por cada variable en la Figura 19.
Al igual que en la regresión lineal múltiple, podemos comprobar que el modo de envío, “ShipMode” es la variable que más afecta al tiempo de entrega, con coeficientes superiores a 1. Por otra parte, las otras variables afectan, positiva o negativamente pero de forma leve sobre la variable dependiente. Observamos también que los coeficientes de las categorías de producto, las subcategorías “Furnishings”, “Copiers” o “Bindings”, además de algunos meses son 0, cosa que afirma que no tienen un efecto relevante en la predicción de la variable objetivo. Podemos fijarnos también en la ubicación desde la que se realiza el pedido fijándonos en las variables “city” o “state”. Observamos que el estado de Pennsylvania disminuye en 2 unidades la variable objetivo, mientras que no New York o Washington no afecta en la predicción.

Figura 19: *Gráficos de los coeficientes agrupado por variables al aplicar un modelo de regresión Lasso para explicar los tiempos de entrega*

6.2.1.3. Regresión Ridge.

En la Figura 20, vemos cómo algunos de los coeficientes de las variables, como las categorías de producto, aunque se aproximan mucho, no llegan a ser 0. Podemos analizar también que la variable que más afecta al tiempo de entrega es ShipMode. Además, llama la atención que las subcategorías Machines y Bookcases afectan disminuyendo el tiempo de entrega en casi 2 unidades.

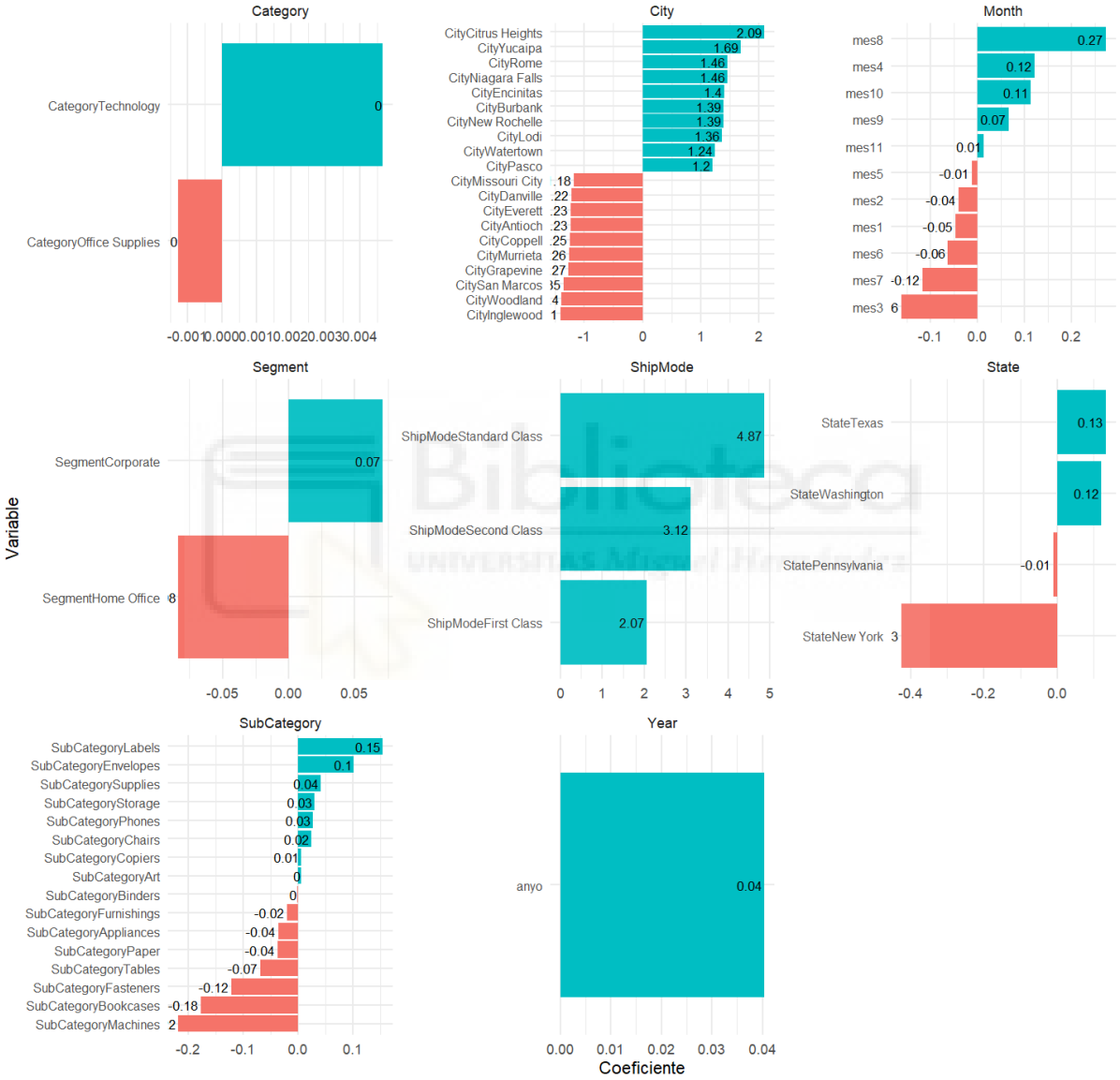

Figura 20: *Gráfico coeficientes de las variables del modelo de regresión Ridge para explicar la variable objetivo tasa de beneficio*

6.2.1.4. Árboles de regresión.

En la Figura 21 observamos que las divisiones del árbol se hacen con ShipMode, lo que indica que este factor tiene un impacto significativo en el tiempo de entrega. El árbol

muestra cómo los envíos realizados mediante Same Day, First Class o Second Class tienen un tiempo de entrega medio de 3.9 días. Sin embargo, los envíos que no utilizan estos métodos tardan en promedio 5 días, lo que indica que estas variables reducen el tiempo de entrega.

Por otra parte, dentro del grupo que usa Same Day, First Class o Second Class, se identifican diferencias importantes. Los envíos Same Day tienen el menor tiempo de entrega, con un promedio de 2.4 días, pero representan solo el 40% de los casos. Cuando el envío es First Class, el tiempo de entrega aumenta ligeramente a 2.7 días.

Añadir también que los envíos Same Day presentan tiempos de entrega extremadamente bajos, con un valor de 0.07 días en promedio, lo que significa que en algunos casos la entrega es prácticamente inmediata. Sin embargo, solo el 5% de los pedidos se procesan con esta rapidez.

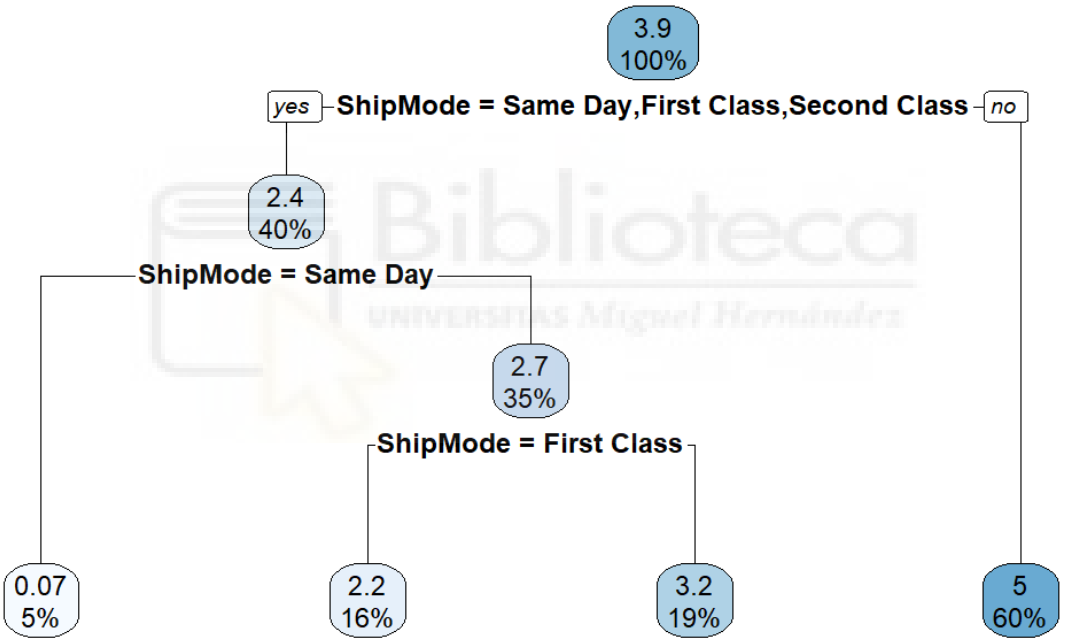

Figura 21: *Gráfico de árboles de regresión para explicar el tiempo de entrega*

6.2.1.5. Bosques aleatorios.

Observamos en la Figura 22, los dos gráficos que explican la importancia de las variables para la variable dependiente tiempo de entrega. En este caso, la variable más importante para el modelo es “ShipMode”, pues es la que consigue un valor mayor de importancia, y de incremento del MSE; del resto de variables, pueden considerarse algo relevantes, atendiendo a su contribución a una variación en el MSE, el mes, State, anyo y Segmento. El índice de importancia realmente no aporta mucha información sobre ellas.

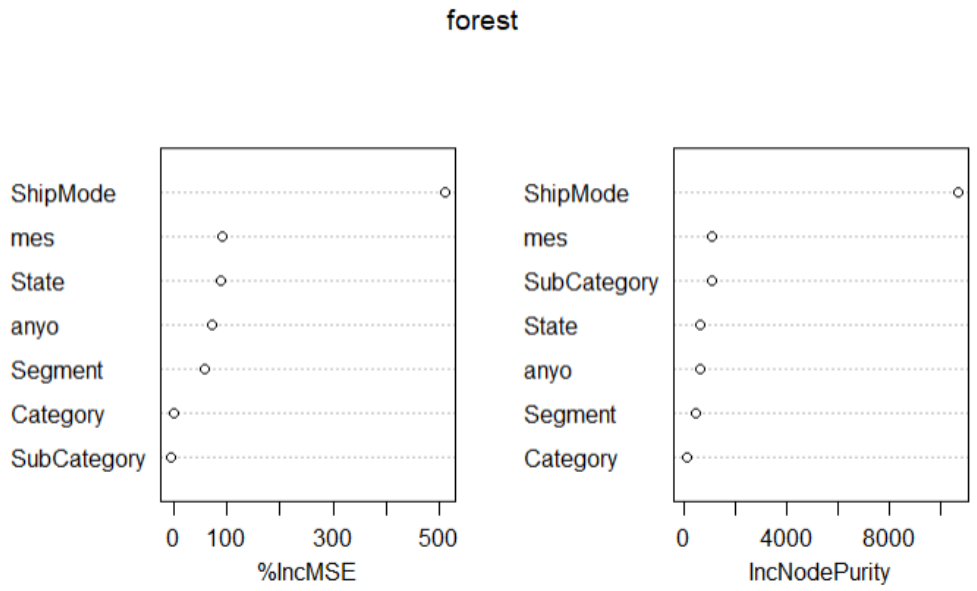

Figura 22: *Gráfico explicación de la importancia de las variables en un modelo de bosques aleatorios que explica el tiempo de entrega.*

Conclusiones sobre el tiempo de entrega

En la Figura 23, observamos esta tabla de comparación de resultados. Podemos afirmar que para predecir el tiempo de entrega del pedido, el modelo que mejor se ajusta a los datos son los Bosques aleatorios, pues tiene el menor MSE, 0.8764507. Esto quiere decir que es el modelo que menor error comete, y tiene el mayor R^2, lo que significa que es el que más variabilidad de los datos explica. Los demás modelos se ajustan de forma similar aunque no tan bien como los bosques.

	MSE	**R2**
Lineal Múltiple	0.8786479	0.6984000
Lasso	0.9366004	0.7075495
Ridge	0.9422042	0.7105763
Árboles de Regresión	0.9457538	0.6885589
Bosques Aleatorios	0.3751824	0.8764507

Fig 23: *Tabla comparativa de los modelos de regresión para predecir la variable Tiempo de entrega.*

Con estos datos, podemos concluir que que la variable que más influye sobre el tiempo de entrega es efectivamente “ShipMode”, pues ella en sí misma explica que siendo asignada a un pedido de venta, modificará los tiempos de envío.

6.2.2. Tasa de beneficio

Por otra parte, expondremos qué modelos son mejores para explicar la variable Tasa de Beneficio.

6.2.2.1. Regresión lineal múltiple.

Hemos intentado ajustar un modelo en el que se prediga la variable dependiente TasaBeneficio en función de las variables independientes categóricas de las que disponemos.

TasaBeneficio ~ mes + SubCategory + anyo + State + Category + Segment + ShipMode.

Mediante el modelo de Akaike, después de haber realizado los procedimientos "hacia delante" y "hacia atrás", llegamos al modelo mejor ajustado. Se nos quedaría tal que así:

TasaBeneficio ~ SubCategory + anyo + State

En este caso, la tabla ANOVA (Figura 24) quedaría de este modo:

```
Analysis of Variance Table

Response: TasaBeneficio
             Df  Sum Sq Mean Sq  F value  Pr(>F)
SubCategory  16 2468923  154308 128.9063 < 2e-16 ***
anyo          3   11659    3886   3.2467 0.02099 *
State         4 3295539  823885 688.2605 < 2e-16 ***
Residuals  5180 6200738    1197
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 24: *Análisis de la varianza relativa al modelo que explica la tasa de beneficio.*

Observamos que las variables "SubCategory" y "State" tienen un p-valor menor a 0'05, por lo que se rechaza la hipótesis nula y podemos decir que influyen muy significativamente en la tasa de beneficio. Sin embargo, la variable "anyo", aunque su p-valor también sea inferior a 0'05, es el mayor de los 3, por lo que concluimos que es la variable que menos influye.

6.2.2.2. Regresión Lasso

En la Figura 25, vemos que al realizar la validación cruzada con 10 particiones para encontrar el valor óptimo de λ y la penalización, los coeficientes se distribuyen según el gráfico que se muestra a continuación. Comprobamos mirando los coeficientes de las variables que hay algunas variables, como la ciudad Tyler o el estado de Texas que afectan significativamente a la tasa de beneficio, pues los coeficientes son bastante grandes, ya sea en positivo o en negativo. Algo que llama bastante la atención es que gran parte de las variables tienen el coeficiente nulo, como el año, el mes o el segmento.

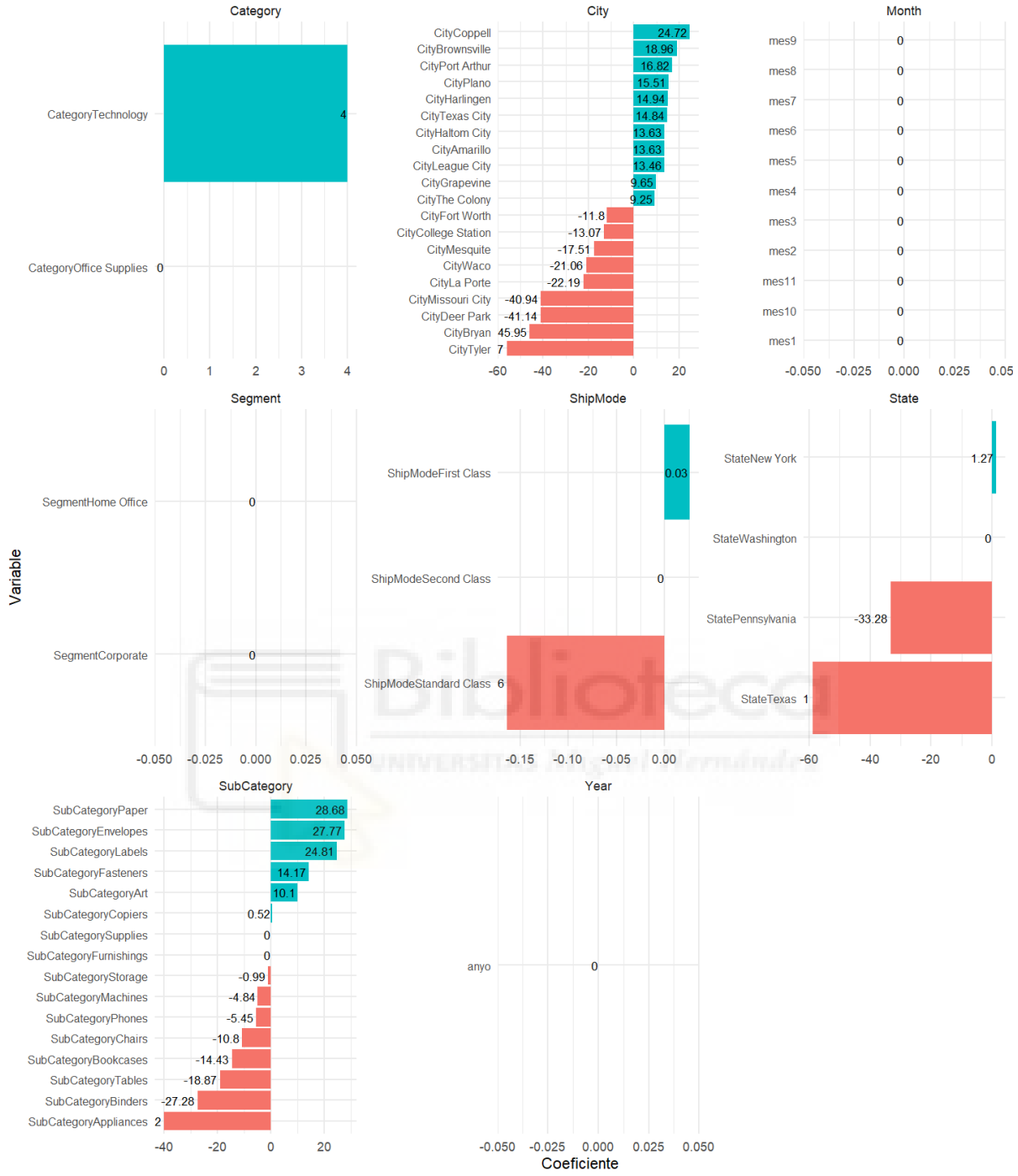

Figura 25: *Gráfico coeficientes de las variables del modelo de regresión Lasso para explicar la variable objetivo tasa de beneficio*

6.2.2.3. Regresión Ridge

En la Figura 26, visualizamos que este modelo sugiere que el modo de envío, la ubicación geográfica, con las variables State y City y la categoría del producto son factores determinantes en los tiempos de entrega. Los productos tecnológicos y ciertas subcategorías, como Appliances o Binders, presentan mayores retrasos, mientras que los pedidos enviados mediante "Standard Class" tienden a llegar más rápido.

Figura 26: *Gráfico coeficientes de las variables del modelo de regresión Ridge para explicar la variable objetivo tasa de beneficio*

6.2.2.4. Árboles de Regresión

En la Figura 27, vemos que el árbol de regresión que explica la tasa de beneficio, se han abreviado las variables para que cupiesen todas en un mismo árbol. El modelo ha dividido los datos en función de variables como State, City, SubCategory, y ShipMode, lo que indica que afectan significativamente a la variabilidad de la TasaBeneficio.

También se observa que la variable SubCategory juega un papel importante en la segmentación del beneficio. Algunas subcategorías como Appliances y Binders parecen tener valores más bajos de TasaBeneficio, lo que sugiere poca rentabilidad. Por otro lado, subcategorías como Bookcases, Chairs, Phones, Storage, Supplies y Tables muestran una

mayor contribución a la rentabilidad. La segmentación basada en City sugiere que hay diferencias significativas en la rentabilidad según la región.
Algo relevante que observamos es que algunos nodos terminales muestran valores negativos de tasa de beneficio, es decir, algunos productos en ciertas regiones pueden estar generando pérdidas.
Por otro lado, hay ramas que muestran valores positivos elevados, lo que indica que ciertas combinaciones de productos y ubicaciones son más rentables que otras.

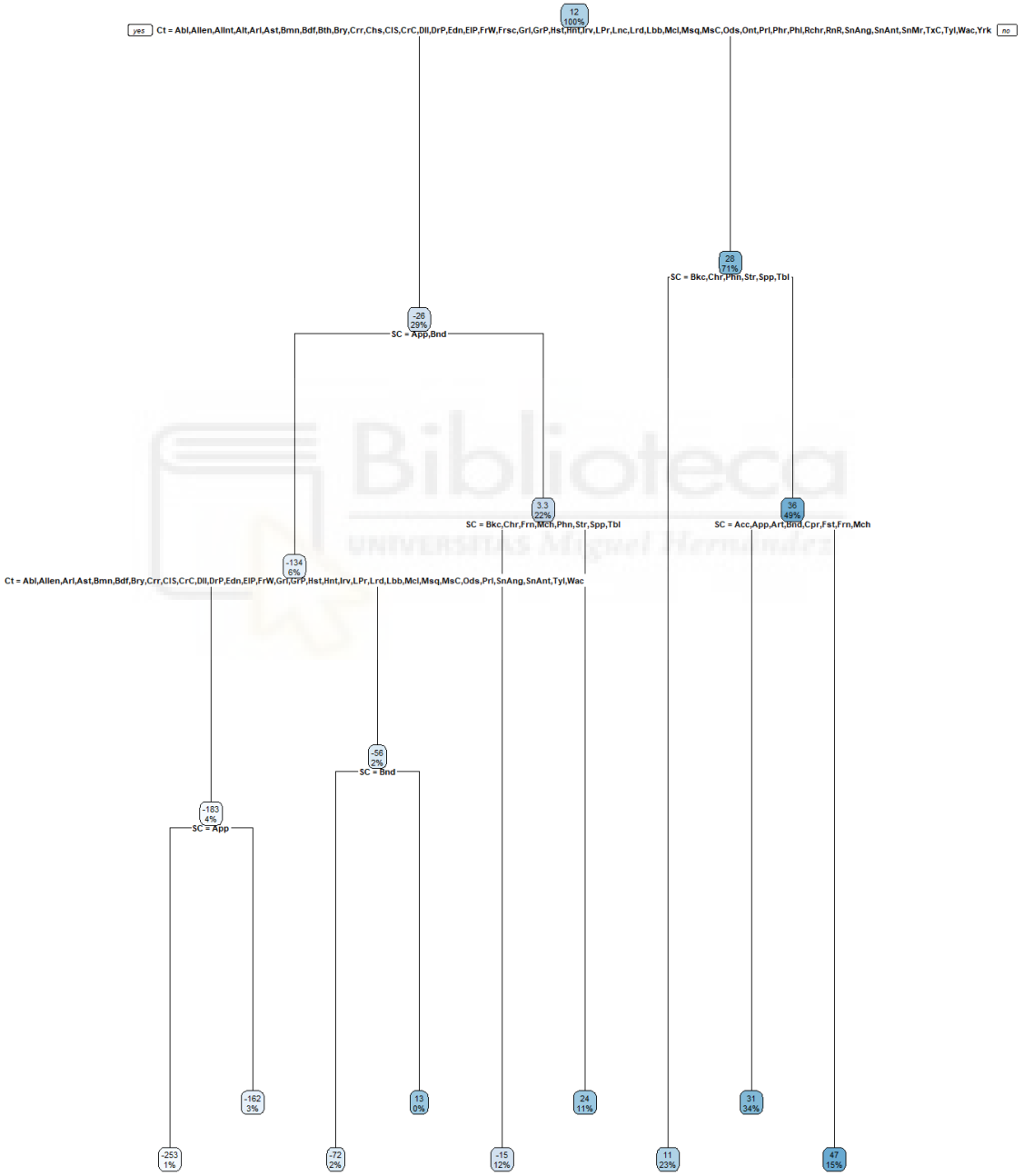

Figura 27: *Gráfico del árbol de regresión que explica la variable objetivo Tasa de Beneficio.*

6.2.2.5. Bosques aleatorios

En la Figura 28 observamos que en el modelo que explica la tasa de beneficio, las variables más relevantes son "State" y "SubCategory", ya que en ambos gráficos aparecen en las primeras posiciones con una diferencia significativa respecto a las demás variables. Esto indica que el estado en el que se realiza la venta y la subcategoría del producto tienen un fuerte impacto en la variabilidad de la tasa de beneficio. La variable "Category" también muestra cierta relevancia, pero con menor impacto relativo en comparación con las dos primeras. Otras variables como "ShipMode", "anyo", "mes" y "Segment" parecen tener una menor contribución en la predicción del modelo, aunque siguen aportando información. Eliminar las variables más influyentes, como "State" y "SubCategory", podría reducir la precisión del modelo, lo que sugiere que cualquier modelo que se utilice para explicar esta variable debería utilizarlas.

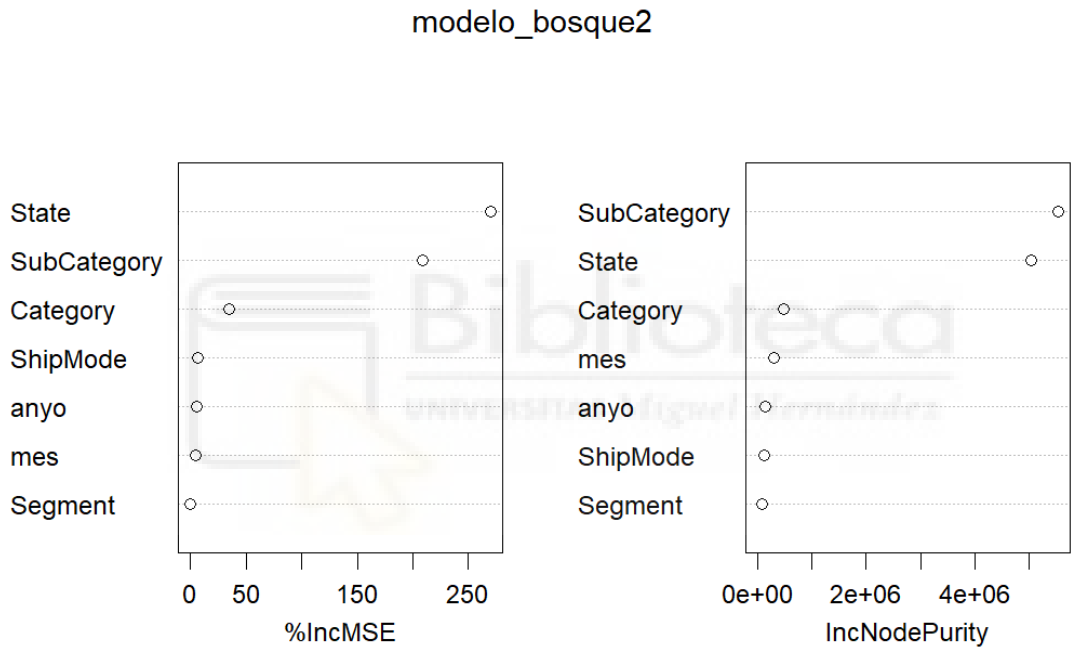

Figura 28: *Gráfico sobre el análisis de la importancia de las variables en un modelo de bosques aleatorios que explica la tasa de beneficio*

Conclusiones sobre la tasa de beneficio

Por otra parte, para predecir la tasa de beneficio, tenemos más variabilidad entre modelos de regresión y nos lo muestra la tabla de la Figura 29. El mayor R2 lo encontramos en los bosques aleatorios, lo que indica que es el modelo que mejor ajusta los datos, pues explica casi el 98% de la variabilidad de la variable dependiente

	MSE	R2
Lineal Múltiple	0.8786479	0.4823000
Lasso	1222.991000	0.4898353
Ridge	0.5030789	0.5030789
Árboles de Regresión	264.5997000	0.8850302
Bosques Aleatorios	51.7600412	0.9775100

Fig 29: *Tabla comparativa de los modelos de regresión para predecir la variable Tasa de Beneficio.*

Conclusiones

Aunque no hemos utilizado todas las variables de la base de datos y nos hemos quedado únicamente con las más repetidas (lo que implica perder parte de la información disponible), podemos concluir que el tiempo de entrega depende casi en su totalidad del modo de envío. No es de extrañar, pues la variable en sí misma explica que, dependiendo de la modalidad seleccionada, se puede tardar menos de un día en entregar el pedido o, por el contrario, extenderse más allá de ese plazo.

En lo que respecta a la tasa de beneficio, ésta se ve afectada tanto por el estado como por la ciudad en la cual se realice la compra, lo que sugiere que factores geográficos, demográficos y logísticos pueden influir significativamente en la rentabilidad. Convendría, por tanto, llevar a cabo un análisis más exhaustivo por parte de la empresa para implementar una estrategia de mercado y estudiar la viabilidad de abrir nuevos establecimientos en zonas prometedoras.

Por último, pensar en la inclusión de más variables explicativas en futuros modelos permitiría obtener una visión más completa de los factores que determinan el éxito del negocio, tanto en términos de rapidez de entrega como de rentabilidad.

Referencias

Amat, J. (n.d.). *Introducción a la Regresión Lineal Múltiple*. Cienciadedatos.net. Retrieved January 28, 2025, from https://cienciadedatos.net/documentos/25_regresion_lineal_multiple

Tellez, C. F., & Morales, M. A. (2016). *Modelos Estadísticos lineales con aplicaciones en R*. Ediciones de la U.

Random Forest: Bosque aleatorio. Definición y funcionamiento. (n.d.). DataScientest.com. Retrieved February 10, 2025, from https://datascientest.com/es/random-forest-bosque-aleatorio-definicion-y-funcionamiento

Tibshirani, R., & Breiman, L. (n.d.). *LASSO (estadística)*. Wikipedia. Retrieved February 10, 2025, from https://es.wikipedia.org/wiki/LASSO_(estad%C3%ADstica)

Machine learning: Fundamentos, algoritmos y aplicaciones para los negocios, industria y finanzas. (2024). Ediciones Díaz de Santos.

ggplot2 - Essentials - Easy Guides - Wiki. (n.d.). STHDA. Retrieved February 12, 2025, from https://www.sthda.com/english/wiki/ggplot2-essentials

Zhu, H. (2024, January 23). Create Awesome HTML Table with knitr::kable and kableExtra. Retrieved February 1, 2025, from https://cran.r-project.org/web/packages/kableExtra/vignettes/awesome_table_in_html.html