**EXAMEN FINAL FUNDAMENTOS DE APRENDIZAJE DE MÁQUINA**

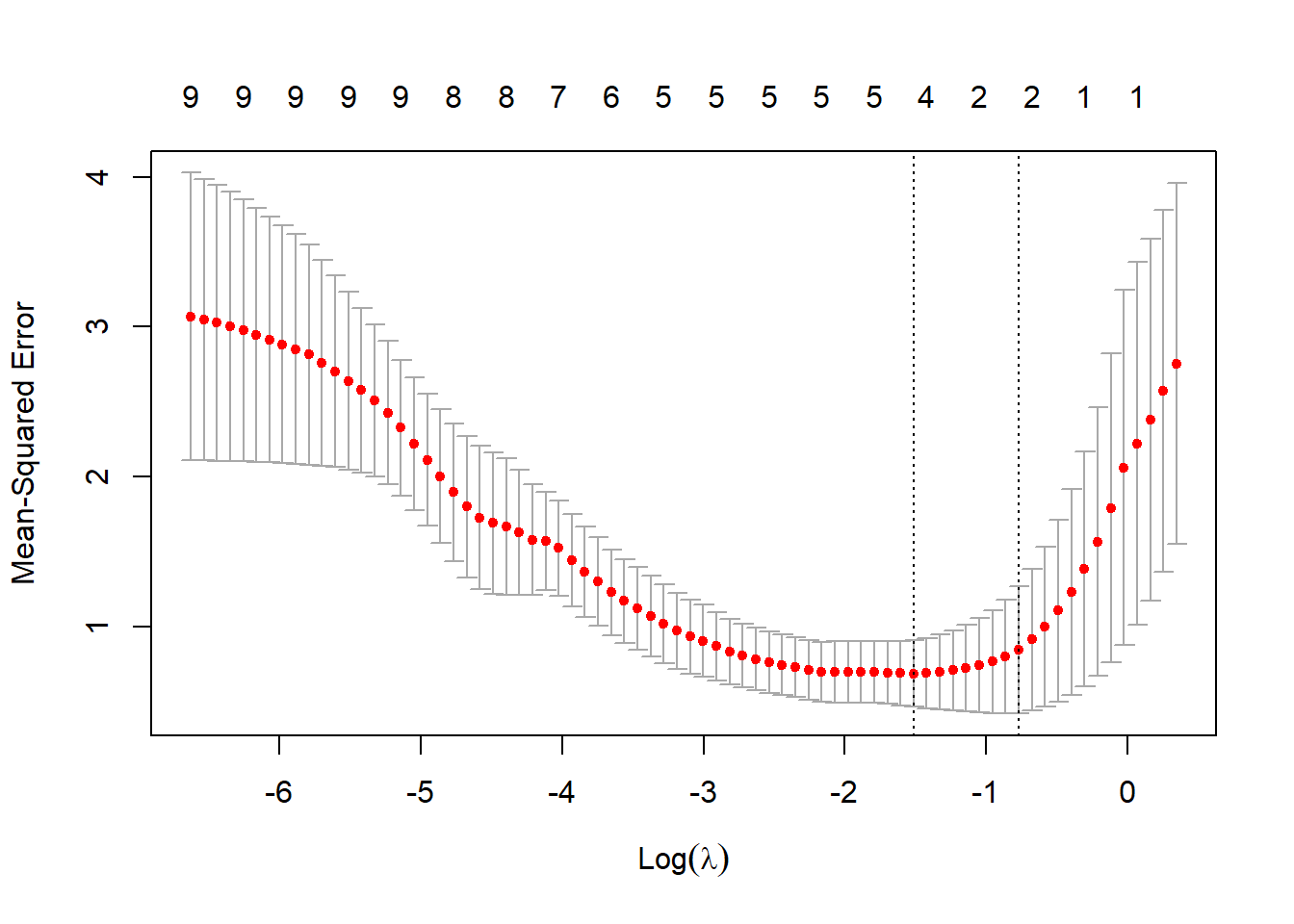
**Reglas del examen: Se pueden consultar materiales del curso. Está completamente prohibido el uso de Internet (más allá de Brightspace) y la comunicación entre estudiantes, con personas o entidades externas durante el examen, de forma presencial o virtual que le apoyen en tiempo real (incluidos foros e inteligencias artificiales basadas en LLM como chatGPT). Favor responder a continuación de la pregunta. Una vez diligenciado el cuestionario deben guardarlo en formato PDF y subirlo como respuesta a la asignación en Brightspace. La hora máxima de entrega es a las 10.00 a.m.**

**NOMBRE: Victoria Sofía Chavarro Diaz**

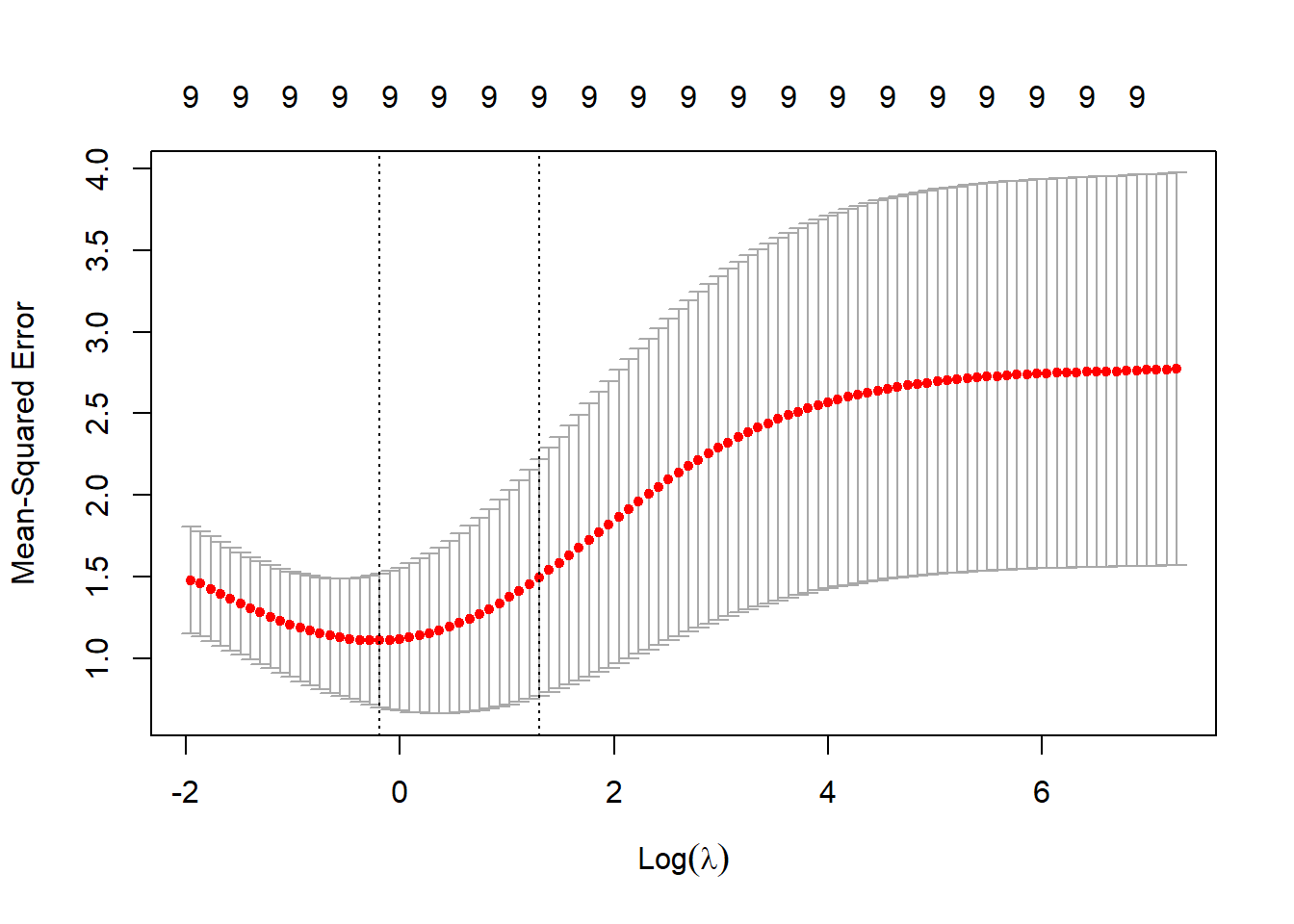
**Caso 1:** Usted hace parte de una organización que ha desarrollado una aplicación que permite predecir el precio de un bien raíz usado a partir de las características del inmueble. En el caso en particular, se desea predecir el precio de un apartamento en Cali (en millones de pesos) de estratos altos (4,5 y 6) a partir de 9 características del inmueble. Para realizar la predicción, primero se estandarizaron las variables (Z-score) y posteriormente se utilizó regresión lineal por mínimos cuadrados con stepwise y regularización con ridge y lasso. Se crearon bases de entrenamiento y prueba del 80% y 20% de los datos respectivamente. Se utilizó validación 5-folds para encontrar el parámetro lambda adecuado en la base de datos de entrenamiento en regularización. Los coeficientes de los tres modelos (Stepwise, Ridge y Lasso) obtenidos con el λ a una desviación estándar se presentan en la tabla presentada más abajo. Además, se incluyen los gráficos de Lambda (log ) vs. MSE para Ridge y Lasso en entrenamiento, a partir de la validación cruzada.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Stepwise | Lasso | Ridge |
| Intercept | 0,7832 | 0,0827 | 0,6556 |
| Numero de cuartos | 0,734 | 0 | 0,1478 |
| Area construída del inmueble (m2) | 2,56 | 1,4483 | 2,5672 |
| Número de baños | -0,771 | 0 | 0,3567 |
| Antigüedad | -1,8 | -0,2477 | -1,2945 |
| Ingreso promedio en la zona (estrato) | no incluido | 0 | 0,91 |
| Balcón | no incluido | 0 | 0,0035 |
| Distancia (en km) al transporte público | -0,5 | -0,0433 | -0,8956 |
| Parqueadero cubierto | no incluido | 0,0733 | 0,942 |
| Ascensor | no incluido |  | 0,058 |

**Lasso**



**Ridge**



***Resultados de RMSE en la base de datos de test:***

|  |  |  |  |
| --- | --- | --- | --- |
| **Modelo** | Stepwise | Lasso | Ridge |
| **RMSE** | 1,599647 | 1,309426 | 1,318787 |

1. Por favor, indique qué variables son las más influyentes para predecir el precio de un apartamento en Cali, y cómo se interpretan las relaciones entre esas variables y el precio del apartamento en el contexto del problema.

**Respuesta:** Las variables más influyentes en la predicción del precio de un apartamento en Cali, según los modelos de regresión utilizados, son:

1. **Área construída del inmueble (m2):** Esta variable consistentemente tiene un gran impacto positivo en el precio del apartamento en todos los modelos. Cuanto mayor sea el área construída, mayor será el precio.
2. **Antigüedad:** La antigüedad del inmueble tiene un impacto negativo en el precio en todos los modelos. A medida que aumenta la antigüedad, el precio disminuye.
3. **Distancia al transporte público:** Esta variable tiene un impacto negativo en el precio en todos los modelos. A medida que aumenta la distancia al transporte público, el precio disminuye.
4. **Parqueadero cubierto (principalmente en Ridge y Lasso):** Tener un parqueadero cubierto aumenta el precio del apartamento.
5. ¿Qué cambio esperaría en los coeficientes y en el sesgo del modelo Ridge si se utiliza el lambda mínimo en vez del lambda a una desviación estándar?

**Respuesta:** Cuando se utiliza el lambda mínimo en lugar del lambda a una desviación estándar en el modelo Ridge, se pueden esperar los siguientes cambios en los coeficientes y en el sesgo del modelo:

**Coeficientes del Modelo Ridge**

1. **Menor Regularización:**
   * Utilizar el lambda mínimo implica una menor penalización de los coeficientes. Esto se traduce en coeficientes de mayor magnitud en comparación con aquellos obtenidos utilizando el lambda a una desviación estándar.
   * Los coeficientes estarán más cerca de los obtenidos en una regresión lineal sin regularización, aunque no serán idénticos.
2. **Incremento en la Varianza:**
   * Con una menor regularización, el modelo se ajustará más a los datos de entrenamiento, lo que puede aumentar la varianza del modelo. Esto significa que los coeficientes podrían ser más sensibles a las fluctuaciones en los datos de entrenamiento.

**Sesgo del Modelo**

1. **Disminución del Sesgo:**
   * Menor regularización reduce el sesgo del modelo porque el ajuste es más fiel a los datos de entrenamiento. El modelo tendrá menor tendencia a simplificar excesivamente las relaciones entre las variables y el objetivo.

**Implicaciones**

* **Generalización:** Aunque el modelo con el lambda mínimo tendrá menor sesgo y podrá capturar mejor las relaciones en los datos de entrenamiento, esto podría venir a costa de un aumento en la varianza, afectando negativamente su capacidad de generalización a nuevos datos.
* **Overfitting:** Con el lambda mínimo, hay mayor riesgo de overfitting, donde el modelo se ajusta demasiado a los datos de entrenamiento y no se desempeña bien en datos de prueba o en aplicaciones del mundo real.

1. Un analista comenta que el número de cuartos no puede ser muy pequeño ni muy grande, porque en ambos casos no se adaptará a las necesidades actuales de los clientes y el precio bajará. ¿qué tipo de ajuste puede hacerse a las variables con el fin de que el modelo capture esta regla de negocio?

**Respuesta:** Para capturar la regla de negocio mencionada por el analista, donde el número de cuartos no puede ser muy pequeño ni muy grande, ya que en ambos casos el precio del apartamento podría bajar, se puede realizar un ajuste no lineal a la variable "Número de cuartos".

Una forma de lograr esto es mediante la adición de términos polinómicos de orden superior para esta variable en el modelo de regresión lineal. Específicamente, se puede incluir un término cuadrático (número de cuartos)^2 junto con el término lineal original (número de cuartos).

El término cuadrático permitirá capturar la relación no lineal entre el número de cuartos y el precio del apartamento, donde el precio aumenta hasta un punto óptimo de número de cuartos y luego comienza a disminuir si el número de cuartos sigue aumentando. Esta transformación polinómica de la variable "Número de cuartos" se puede realizar antes de ajustar el modelo de regresión lineal. Una vez que se incluyen los términos polinómicos, los coeficientes estimados para estos términos capturarán la relación no lineal deseada.

Es importante tener en cuenta que la adición de términos polinómicos también puede introducir multicolinealidad en el modelo, lo que puede afectar la estabilidad de las estimaciones de los coeficientes. En ese caso, puede ser necesario aplicar técnicas adicionales, como la regularización o la selección de variables, para mitigar este problema. Además, es recomendable explorar gráficamente la relación entre el número de cuartos y el precio del apartamento para determinar si la transformación polinómica es adecuada o si se requiere otro tipo de transformación no lineal.

1. Basado en estos resultados, ¿cuál modelo escogería usted entre los tres y por qué?

**Respuesta**: Para escoger entre los tres modelos (Stepwise, Lasso y Ridge) en base a los resultados proporcionados, es importante considerar varios aspectos, incluyendo la precisión, la interpretación y la capacidad de generalización de los modelos.

El modelo con el **menor RMSE** en el conjunto de prueba es el Lasso (1.309426), seguido muy de cerca por el Ridge (1.318787), y el modelo Stepwise tiene el RMSE más alto (1.599647).

En base a **la Regularización y Robustez,** el modelo Lasso Realiza regularización L1, que no solo ayuda a prevenir el sobreajuste, sino que también puede reducir el número de características a considerar, estableciendo algunos coeficientes exactamente en cero. Esto conduce a modelos más simples y fáciles de interpretar.

Mientras que, el modelo Ridge Utiliza regularización L2, que tiende a reducir los coeficientes pero no los lleva a cero, permitiendo que todas las características contribuyan al modelo en mayor o menor medida. Esto puede ser ventajoso si se espera que todas las características tengan alguna importancia.

En base a la **Interpretabilidad**, el modelo Stepwise Selecciona características de forma iterativa y puede ser fácil de interpretar, pero no incluye ningún mecanismo para manejar multicolinealidad o prevenir el sobreajuste, lo cual es evidente dado su RMSE más alto.

El modelo Lasso, Además de ser preciso, facilita la interpretación al establecer coeficientes a cero, eliminando características menos relevantes. En el caso presentado, estableció a cero los coeficientes de varias características como "Número de cuartos" y "Número de baños".

En el caso del modelo Ridge, Aunque todos los coeficientes son menores en magnitud debido a la regularización, sigue considerando todas las características, lo que puede hacer el modelo más complejo de interpretar.

Con base en todo esto, yo escogería el **modelo Lasso** por las siguientes razones:

1. Precisión: El modelo Lasso tiene el RMSE más bajo (1.309426), lo que indica que tiene la mejor precisión en el conjunto de prueba.

2. Regularización y Robustez: La regularización L1 del modelo Lasso ayuda a evitar el sobreajuste, proporcionando un modelo que probablemente generalice mejor a datos nuevos.

3. Interpretabilidad: Lasso simplifica el modelo al eliminar características menos relevantes, lo que facilita la interpretación y la toma de decisiones basada en los resultados del modelo.

El modelo Lasso es preferido debido a su combinación óptima de precisión, capacidad de generalización y simplicidad. La capacidad del modelo para reducir los coeficientes de características irrelevantes a cero hace que sea más fácil de interpretar y utilizar para tomar decisiones informadas sobre los factores que realmente influyen en el precio de los apartamentos en Cali.

2) **Caso 2.** Usted va a asesorar una empresa de videojuegos online de suscripción mensual que piensa ofrecer una nueva consola de videojuegos personalizada con botones especiales para el uso del sitio y marcada con el logo del sitio web, ofreciendo un mecanismo de pago a través de un aumento en la mensualidad de suscripción mediante financiación sin intereses. A los clientes escogidos para esta promoción se les enviará un paquete promocional a sus casas que incluye material promocional del sitio web (esferos, códigos promocionales de acceso si compra la consola, camisetas). El costo de este envío sería de $60.000 por cliente, y las ganancias estimadas para la empresa son de $150.000 por venta.

Se dispone de una base de datos con los resultados de una oferta similar realizada en otro país en

Esta es la información de la que se dispone en base de datos:

|  |  |
| --- | --- |
| Variable | Etiqueta |
| ID\_Cliente | <ninguno> |
| C\_Postal | Codigo postal |
| Respuesta | ¿Interesado en compra de una nueva consola con descuento a través de la página? |
| Conocedor | Número de meses que lleva usando la suscripción |
| Cantidad | dinero gastado en videojuegos en miles (último año) |
| Frecuencia | Numero de utilizaciones (accesos) promedio por mes |
| Con\_niños | El afiliado tiene hijos menores de 15 años |
| Con\_Banda\_ancha | El afiliado disponde de banda ancha |
| Genero | Género del afiliado H=Hombre, M=mujer |
| Edad | Edad en años del afiliado |
| Grupo\_Ingresos | Variable ordinal de 1 a 9, de menores a mayores ingresos. |
| Grupo\_Edad | Edad categorizada en grupos de 1 a 4 |

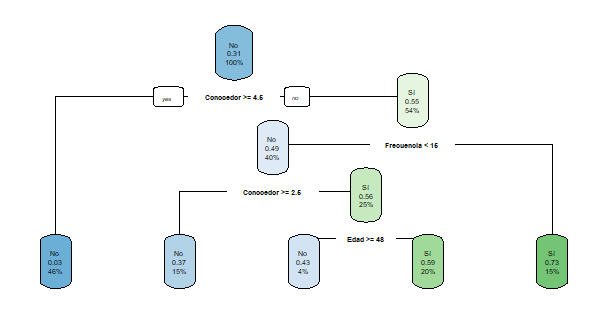
La variable Respuesta tiene la siguiente tabla de frecuencias:

No Sí

4104 1886

La base de datos se dividió en entrenamiento y prueba, en proporciones de 80% y 20% de forma completamente aleatoria.

El siguiente es el árbol ejecutado en la base de entrenamiento mediante validación cruzada con 10 grupos.



La siguiente tabla muestra los resultados de los valores de cp en las particiones usando el procedimiento printcp en R

CP nsplit rel error xerror xstd

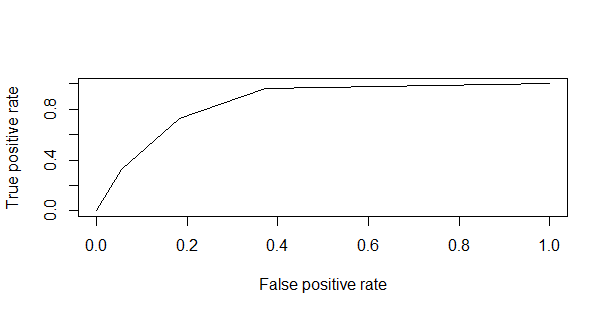
1 0.185578 0 1.00000 1.00000 0.019060

2 0.063097 1 0.81442 0.81442 0.017919

3 0.019088 3 0.68823 0.68929 0.016916

4 0.010000 4 0.66914 0.67550 0.016793

A continuación se presentan la curva COR resultante:



1. Indique en palabras del problema las características de el(los) grupo(s) que el árbol predice que sí comprará(n)

**Respuesta**: En este caso, las características como la cantidad de dinero gastado en videojuegos, la frecuencia de uso y la edad podrían ser determinantes. De acuerdo con el árbol de decisión mostrado en la imagen, las características del grupo que se predice que sí comprará la nueva consola son:

Análisis del Árbol de Decisión de Grupos Predichos que Sí Comprarán

1. Grupo 1:

- Clientes que llevan menos de 4.5 meses usando la suscripción y acceden al sitio con una frecuencia menor a 16 veces por mes.

- Probabilidad de compra: 54%

2. Grupo 2:

- Clientes que llevan más de 2.5 meses pero menos de 4.5 meses usando la suscripción y tienen una edad menor a 48 años.

- Probabilidad de compra: 25%

3. Grupo 3:

- Clientes que llevan más de 2.5 meses pero menos de 4.5 meses usando la suscripción y tienen una edad mayor o igual a 48 años.

- Probabilidad de compra: 20%

4. Grupo 4:

- Clientes que acceden al sitio con una frecuencia mayor o igual a 16 veces por mes.

- Probabilidad de compra: 15%

Según el árbol de decisión, los clientes que tienen una mayor probabilidad de comprar la consola con descuento a través de la página son aquellos que:

- Han usado la suscripción por menos de 4.5 meses y acceden al sitio con una frecuencia menor a 16 veces por mes.

- Han usado la suscripción por más de 2.5 meses y menos de 4.5 meses, independientemente de su edad.

- Acceden al sitio con una frecuencia mayor o igual a 16 veces por mes.

Estos grupos representan los clientes más propensos a comprar la consola según el modelo de decisión basado en los datos disponibles.

1. ¿Qué ocurrirá con el árbol y los resultados si cp se establece en 0.025? Y si se disminuye a 0?

**Respuesta**: Si el parámetro cp se establece en 0.025, el árbol se podará y se verá más pequeño y simple. Con un cp de 0.025, el árbol probablemente seguirá dividiendo el nodo raíz (Conocedor >= 4.5) y luego dividirá los dos nodos hijos (Frecuencia <= 15 y Frecuencia > 15), pero puede dejar de dividir algunos de los nodos más profundos. Esto haría que el árbol sea ligeramente más simple y menos ajustado a los datos de entrenamiento, pero podría mejorar su capacidad de generalización a nuevos datos.

Por otro lado, si cp se disminuye a 0, el árbol no se podará en absoluto y crecerá hasta su máxima extensión posible, ajustándose completamente a los datos de entrenamiento. Esto puede llevar a un sobreajuste (overfitting) del modelo, donde el árbol captura incluso el ruido y las peculiaridades específicas de los datos de entrenamiento, lo que puede afectar negativamente su capacidad de generalización a nuevos datos.

En términos generales, un cp más pequeño (cercano a 0) permitirá que el árbol crezca más y capture más detalles de los datos de entrenamiento, pero con un mayor riesgo de sobreajuste. Un cp más grande (cercano a 1) producirá un árbol más pequeño y simple, pero puede perder información importante.

1. Interprete en los términos del problema (es decir, no sólo de las definiciones sino explique que implican en el problema) el siguiente punto de la curva COR marcado en la gráfica: (0.20; 0.78).

**Respuesta**: El punto (0.20, 0.78) en la curva ROC representa un equilibrio específico entre la tasa de falsos positivos y la tasa de verdaderos positivos del modelo predictivo en el problema de determinar si un cliente comprará o no la nueva consola de videojuegos.

Una tasa de falsos positivos de 0.20 significa que el 20% de los clientes que no comprarían la consola son clasificados erróneamente por el modelo como compradores potenciales. En el contexto del problema, esto implica que 1 de cada 5 clientes que realmente no están interesados en la consola recibirían el paquete promocional y la oferta, lo cual representa un gasto innecesario para la empresa.

Por otro lado, una tasa de verdaderos positivos de 0.78 significa que el 78% de los clientes que sí comprarían la consola son correctamente identificados por el modelo como compradores potenciales. En el contexto del problema, esto implica que casi 4 de cada 5 clientes interesados en la consola recibirían la oferta promocional, lo cual es deseable para la empresa, ya que son compradores potenciales.

En general, este punto en la curva ROC representa un modelo que tiene una buena capacidad para identificar correctamente a los clientes interesados en la consola (alta tasa de verdaderos positivos), pero también genera un número considerable de falsos positivos, lo que implica enviar paquetes promocionales a clientes que no están interesados en la compra, incurriendo en gastos innecesarios.

**Caso 3:** Un banco desea realizar un modelo de regresión logística para la predicción de riesgo de caer en una mora superior a 1 año. Para ello tiene una base de 120.260 clientes, de los cuales 111.904 no han caído en mora superior a 1 año.

Se realizó un modelo de regresión logística con stepwise, las variables predictoras que quedaron finalmente en el modelo fueron:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Descripción** | **Coeficiente** |
| (Intercepto) | Intercepto de la regresión logística | 0,89 |
| Edad | Edad del cliente (años) | -0,028 |
| Créditos abiertos | Número de créditos abiertos | 0,013 |
| Retrasos previos | Número de veces que ha tenido retrasos inferiores a 90 días | 1,014 |
| Hipotecarios | Número de créditos hipotecarios que tiene | 0,044 |
| Dependientes | Número de personas a cargo | 0,066 |

Al testear en la base de datos de prueba (base cuyas proporciones de personas que han caído en mora de más de 1 año es similar a la de la base original) el modelo realizado, se obtuvieron los siguientes resultados:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Punto de corte actual | Precisión | Exhaustividad | F-score | AUC |
| 0,5 | 0,126 | 0,635 | 0,21 | 0,729 |

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Curva ROC - regresión logística

1. Especifique la partición train/test de la base de datos que realizaría para efectuar la creación y validación (prueba) del modelo sabiendo que el interés fundamental es predecir las personas que caerán en mora superior a 1 año y que se desea aprovechar al máximo la información disponible.

**Respuesta**: Dado que el interés fundamental es predecir correctamente las personas que caerán en mora superior a 1 año y se desea aprovechar al máximo la información disponible, se recomienda realizar una partición estratificada de la base de datos en los conjuntos de entrenamiento y prueba.

La partición estratificada consiste en dividir la base de datos de manera que la distribución de la variable objetivo (en este caso, haber caído o no en mora superior a 1 año) sea aproximadamente la misma en los conjuntos de entrenamiento y prueba. Esto asegura que ambos conjuntos sean representativos de la población original y evita sesgos en el modelo debido a distribuciones desbalanceadas.

En este caso, dado que se tiene un total de 120.260 clientes, de los cuales 111.904 no han caído en mora superior a 1 año y 8.356 (120.260 - 111.904) sí han caído en mora superior a 1 año, se puede realizar la siguiente partición:

1. Conjunto de entrenamiento:

- Clientes que no han caído en mora superior a 1 año: 89.523 (80% de 111.904)

- Clientes que sí han caído en mora superior a 1 año: 6.684 (80% de 8.356)

2. Conjunto de prueba:

- Clientes que no han caído en mora superior a 1 año: 22.381 (20% de 111.904)

- Clientes que sí han caído en mora superior a 1 año: 1.672 (20% de 8.356)

De esta manera, tanto el conjunto de entrenamiento como el de prueba mantienen la misma proporción de clientes que han caído en mora superior a 1 año (aproximadamente 7%) que la base de datos original.Esta partición estratificada permitirá entrenar el modelo con una muestra representativa de la población y evaluar su desempeño en un conjunto de prueba que también refleje la distribución real de la variable objetivo. Esto ayudará a obtener una estimación más precisa del rendimiento del modelo en escenarios reales.

1. Un cliente cuyo riesgo de caer en mora estaba valorado en 30% ha tenido un retraso de 10 días que ya ha saldado. ¿Cuál es su riesgo ahora?

**Respuesta:** Para determinar el nuevo riesgo de caer en mora de un cliente que tenía un riesgo inicial del 30% y que ha tenido un retraso de 10 días (ya saldado), se necesita actualizar la probabilidad de acuerdo con el nuevo evento. En este caso, el evento relevante es el retraso de 10 días. La regresión logística, con la nueva información sobre el retraso, permitirá recalcular la probabilidad de mora.

−0.8473+1.014 = 0.1667

1/1+exp(-0.1667) = 0.5415

Por lo tanto, el nuevo riesgo del cliente de caer en mora, después de tener un retraso de 10 días, es aproximadamente 54.15%.

1. Un participante de la reunión de presentación del modelo opina que el F-score del modelo es muy bajo y que, por tanto, el modelo: a) no tiene utilidad para el banco y b) no supera una predicción realizada al azar. ¿Estaría usted de acuerdo o no? Justifique.

**Respuesta**: a) No estoy de acuerdo con la afirmación de que el modelo no tiene utilidad para el banco. A pesar de un F-score bajo, la capacidad del modelo para identificar a una parte significativa de los clientes de alto riesgo (con una exhaustividad del 63.5%) puede permitir al banco tomar medidas preventivas efectivas y reducir la tasa de mora.

b) No estoy de acuerdo con la afirmación de que el modelo no supera una predicción realizada al azar. El AUC de 0.729 demuestra que el modelo tiene una capacidad de discriminación sustancialmente mejor que el azar. Esto significa que el modelo es valioso para clasificar a los clientes en riesgo y tiene un rendimiento significativamente mejor que una predicción aleatoria.

**Caso 4.** Con el fin de retener a los clientes de cierto servicio de subscripción a canales premium de deportes en televisión, se decide usar *naive Bayes* para predecir los clientes que desertarán del servicio (*churn*). A continuación, se presenta una tabla cruzada de dos atributos que se usaron. El costo de realizar una llamada al cliente se ha estimado en 2 dólares y a cada cliente se le ofrecerá una promoción cuyo costo para la empresa es de 40 dólares. Se considera que el ingreso total que se obtiene por retener un cliente es de 200 dólares.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **¿Canceló el servicio?** | |  |  |  | **¿Canceló el servicio?** | |  |
|  |  | Sí | No |  |  |  | Sí | No |  |
| **¿Se retrasó en el Pago anterior?** | Sí | 89 | 250 |  | **Tipo de Servicio Contratado** | Full | 76 | 740 |  |
| No | 50 | 3502 |  | Básico | 63 | 3012 |  |

Los resultados de las predicciones, evaluados en la base de datos de prueba, se presentan a continuación:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Real** | |
|  |  | **¿Canceló el servicio?** | |
|  |  | Sí | No |
| **Predicción** | Sí | 250 | 150 |
|  | No | 100 | 500 |
|  |  |  |  |
|  |  |  |  |

1. Según el esquema de clasificación de *naive Bayes*, ¿cuál de las dos variables tiene mayor probabilidad para predecir un cliente que canceló el servicio?

**Respuesta**: Para determinar cuál de las dos variables tiene mayor probabilidad para predecir un cliente que canceló el servicio según el esquema de clasificación de naive Bayes, debemos calcular las probabilidades condicionales de cancelar el servicio para cada valor de las variables "¿Se retrasó en el Pago anterior?" y "Tipo de Servicio Contratado".

1. Variable: ¿Se retrasó en el Pago anterior?

Probabilidad de cancelar el servicio dado que se retrasó en el pago anterior (Sí):

- (Canceló = Sí | Retraso = Sí):

- Número de clientes que cancelaron y se retrasaron: 89

- Número total de clientes que se retrasaron: 89 + 250 = 339

- 89/339 =

|  |
| --- |
| [1] 0.2625369 |
|  |
| |  | | --- | |  | |

Probabilidad de cancelar el servicio dado que no se retrasó en el pago anterior (No):

- (Canceló = Sí | Retraso = No):

- Número de clientes que cancelaron y no se retrasaron: 50

- Número total de clientes que no se retrasaron: 50 + 3502 = 3552

- 50/3552 =

|  |
| --- |
| [1] 0.01407658 |
|  |
| |  | | --- | |  | |

2. Variable: Tipo de Servicio Contratado

Probabilidad de cancelar el servicio dado que tiene el servicio Full:

- (Canceló = Sí | Servicio = Full):

- Número de clientes que cancelaron y tienen servicio Full: 76

- Número total de clientes con servicio Full: 76 + 740 = 816

- 76/816 = 0.09313725

Probabilidad de cancelar el servicio dado que tiene el servicio Básico:

- (Canceló = Sí | Servicio = Básico):

- Número de clientes que cancelaron y tienen servicio Básico: 63

- Número total de clientes con servicio Básico: 63 + 3012 = 3075

- 63/3075 = 0.0204878

La variable "¿Se retrasó en el Pago anterior?" es más significativa para predecir que un cliente cancelará el servicio, dado que la probabilidad de cancelación es mucho mayor para los clientes que se retrasaron en el pago anterior (26.2%) en comparación con cualquier otra combinación de variables y valores.

(12) ¿Cuál sería la ganancia o pérdida asociada en la base de prueba si se aplica a la estrategia de retención mencionada?

**Respuesta**: Costo de realizar una llamada al cliente: $2

Costo de la promoción ofrecida al cliente: $40

Ingreso total por retener un cliente: $200

Cálculo de Costos y Beneficios

1. Clientes Predichos que Cancelarán (Predicción: Sí)

- Clientes que realmente cancelarán (TP): 250

- Clientes que no cancelarán pero fueron predichos como que sí (FP): 150

2. Clientes Predichos que No Cancelarán (Predicción: No)

- Clientes que realmente no cancelarán (TN): 500

- Clientes que cancelarán pero fueron predichos como que no (FN): 100

Costos Asociados a la Estrategia de Retención

Para todos los clientes que fueron predichos como que cancelarán, se realizará una llamada y se ofrecerá una promoción:

- Número total de llamadas: 250 (TP) + 150 (FP) = 400

- Costo total de llamadas: 400 \* $2 = $800

- Número total de promociones: $400

- Costo total de promociones: 400 \* $40 = $16,000

- Costo total de la estrategia: $800 (llamadas) + $16,000 (promociones) = $16,800

Beneficios Asociados a la Retención

Los beneficios provienen de los clientes que realmente iban a cancelar pero fueron retenidos debido a la estrategia:

- Número de clientes realmente retenidos (TP): 250

- Ingreso total por retención: 250 \* $200 = $50,000

Ganancia Neta

La ganancia o pérdida neta se calcula restando los costos de la estrategia de los ingresos obtenidos:

- Ganancia Neta: $50,000 (ingreso total) - $16,800 (costo total) = $33,200

La ganancia neta asociada a la estrategia de retención aplicada a la base de prueba es de $33,200.