Preguntas sobre Visualizaciones - Análisis de Customer Churn

Interpretación de Gráficas y Análisis Visual

Contents

Preguntas sobre Visualizaciones - Análisis de Customer Churn

Información del Proyecto

- **Dataset**: Telco Customer Churn (7,043 clientes)
- **Objetivo**: Predecir qué clientes abandonarán el servicio de telecomunicaciones
- **Mejor Modelo**: Logistic Regression Optimizado (ROC-AUC: 0.8503)
- **Técnica de Balanceo**: Undersampling (seleccionada automáticamente por mejor ROC-AUC)

CATEGORÍA 1: ANÁLISIS EXPLORATORIO DE DATOS (EDA)

Pregunta 1

¿Qué nos muestra la gráfica de distribución de Churn (barras y pastel) y por qué es importante visualizar el desbalance de clases?

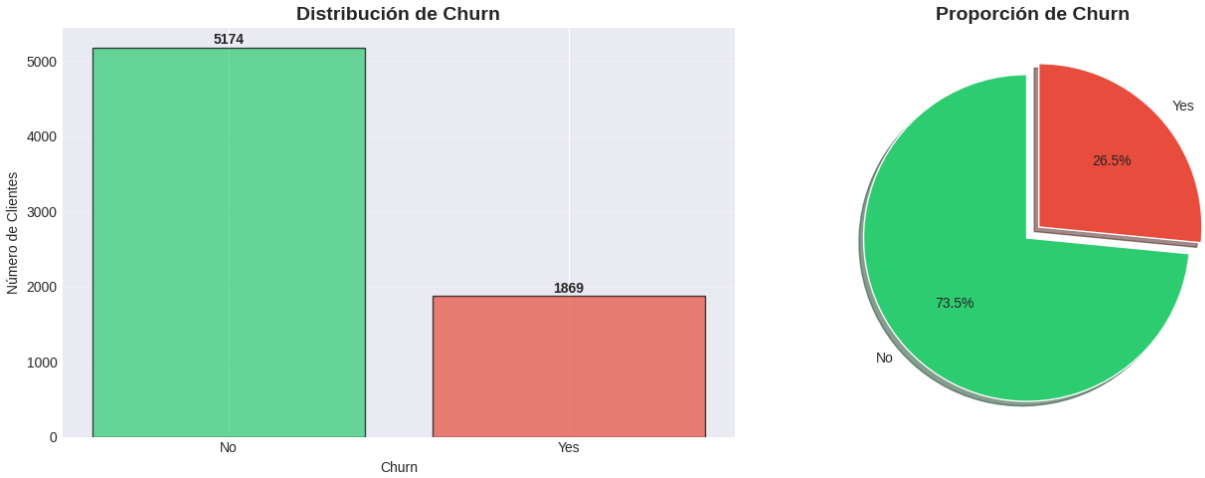


Figure 1: Distribución de Churn

Respuesta: La visualización muestra dos representaciones complementarias: un gráfico de barras que indica 5,174 clientes No Churn vs 1,869 clientes Churn, y un gráfico de pastel que muestra 73.5% vs 26.5%. Esta visualización es crucial porque revela inmediatamente el desbalance de clases (ratio 2.7:1), que es el principal desafío técnico del proyecto. Sin esta visualización, podríamos entrenar modelos que simplemente predicen "No Churn" para todos los casos y obtener 73% de accuracy, pero serían inútiles para el negocio.

Analogía: Es como un semáforo de advertencia en una carretera: te alerta inmediatamente que hay un problema (desbalance) que debes abordar antes de continuar el viaje (modelado).

Mini-glosario:

- **Desbalance de clases**: Cuando una categoría tiene significativamente más ejemplos que otra
- **Ratio de desbalance**: Proporción entre clase mayoritaria y minoritaria
- **Visualización dual**: Usar dos tipos de gráficos para mostrar la misma información desde diferentes perspectivas

---

Pregunta 2

¿Qué insights podemos extraer de las 6 gráficas de barras que muestran "Churn por Variable Categórica" (Contract, InternetService, PaymentMethod, TechSupport, OnlineSecurity, PaperlessBilling)?

Respuesta: Estas visualizaciones revelan los factores de riesgo más importantes:

1. **Contract**: Los contratos mes a mes tienen ~42% de churn vs ~3% en contratos de 2 años

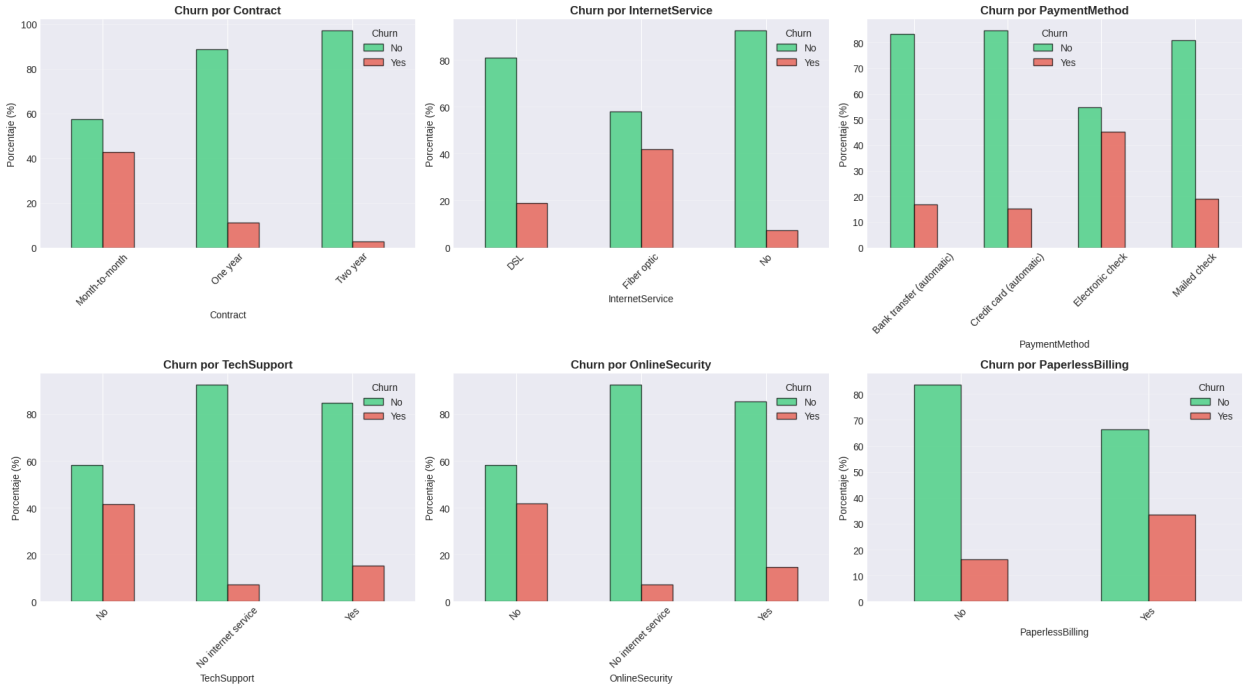


Figure 2: Churn por Variables Categóricas

2. **InternetService**: Fiber optic tiene mayor churn (~30%) que DSL (~19%)
3. **PaymentMethod**: Electronic check tiene el mayor churn (~45%)
4. **TechSupport**: Clientes sin soporte técnico tienen ~42% de churn vs ~15% con soporte
5. **OnlineSecurity**: Sin seguridad online ~42% churn vs ~15% con seguridad
6. **PaperlessBilling**: Facturación sin papel tiene mayor churn (~34% vs ~16%)

Estos insights son accionables: la empresa puede diseñar estrategias específicas para cada segmento de riesgo.

Analogía: Es como un médico que analiza diferentes síntomas (variables) para diagnosticar una enfermedad (churn). Cada síntoma aporta información valiosa para el diagnóstico final.

Mini-glosario:

- **Variable categórica**: Característica que representa grupos o categorías
- **Tasa de churn por segmento**: Porcentaje de abandono dentro de cada categoría
- **Insight accionable**: Descubrimiento que puede traducirse en acciones de negocio

---

Pregunta 3

¿Qué información proporcionan los histogramas superpuestos de las variables numéricas (tenure, MonthlyCharges, TotalCharges)?

Respuesta: Los histogramas superpuestos (verde para No Churn, rojo para Churn) revelan patrones de distribución:

1. **Tenure**: Los clientes con churn se concentran en los primeros meses (0-12 meses), mientras que los clientes leales tienen distribución más uniforme hasta 72 meses

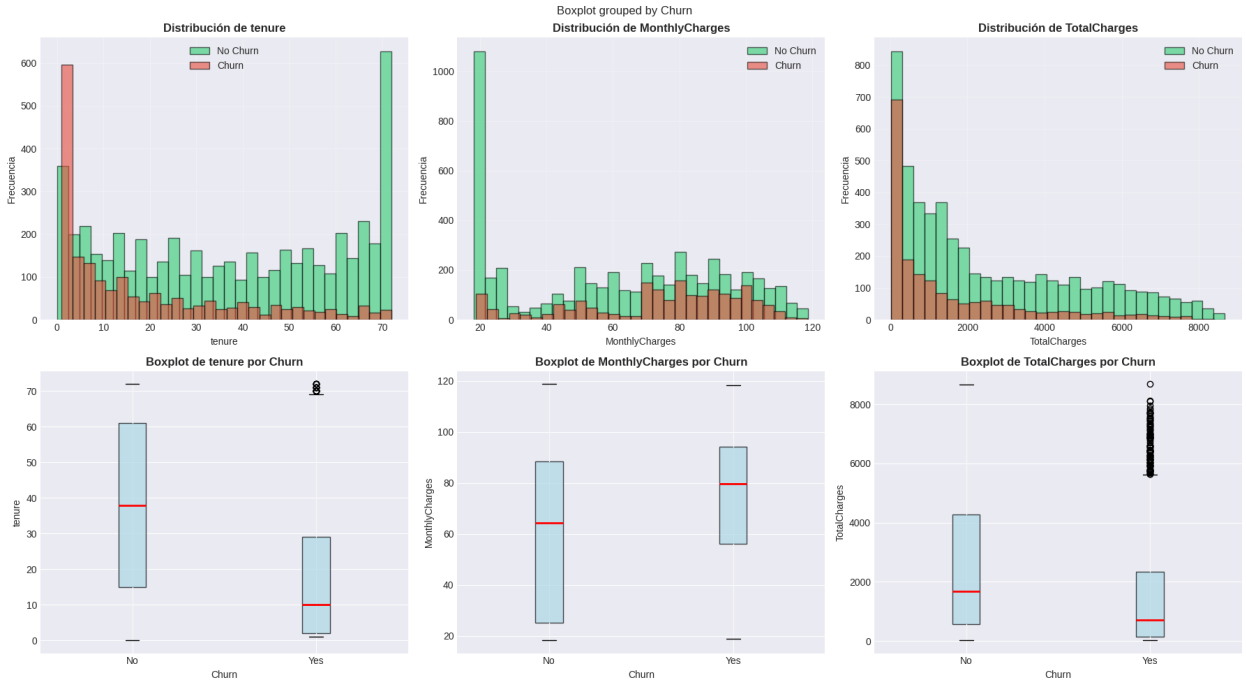


Figure 3: Distribuciones de Variables Numéricas

2. **MonthlyCharges**: Los clientes con churn tienden a tener cargos mensuales más altos ($70-$110), mientras que los leales tienen distribución más amplia
3. **TotalCharges**: Los clientes con churn tienen cargos totales bajos (concentrados cerca de $0-$2000), indicando que abandonan temprano

La superposición permite comparar directamente las distribuciones y identificar rangos de riesgo.

Analogía: Es como comparar dos poblaciones de plantas: una que sobrevive (verde) y otra que muere (roja). Al superponer las distribuciones de altura, agua recibida, etc., puedes identificar qué condiciones favorecen la supervivencia.

Mini-glosario:

- **Histograma**: Gráfico que muestra la distribución de frecuencias de una variable
- **Superposición**: Mostrar dos distribuciones en el mismo gráfico para facilitar comparación
- **Rango de riesgo**: Intervalo de valores donde se concentra el churn

---

Pregunta 4

¿Cómo interpretar los boxplots de variables numéricas por Churn y qué nos dicen las medianas?

Respuesta: Los boxplots complementan los histogramas mostrando estadísticas resumidas:

1. **Tenure**: La mediana de Churn está en ~10 meses vs ~38 meses para No Churn, confirmando que clientes nuevos tienen mayor riesgo
2. **MonthlyCharges**: La mediana de Churn es ~$80 vs ~$65 para No Churn, indicando que precios altos aumentan el riesgo

3. **TotalCharges**: La mediana de Churn es ~$1,400 vs ~$2,500 para No Churn, reflejando menor tiempo de permanencia

Los boxplots también muestran outliers (puntos fuera de los bigotes) que representan casos excepcionales.

Analogía: Es como comparar las notas de dos grupos de estudiantes: el boxplot te muestra rápidamente quién tiene mejor rendimiento promedio (mediana), qué tan dispersas están las notas (caja), y si hay casos extremos (outliers).

Mini-glosario:

- **Boxplot**: Gráfico que muestra mediana, cuartiles y outliers
- **Mediana**: Valor central que divide los datos en dos mitades iguales
- **Outlier**: Valor atípico que se aleja significativamente del resto

---

Pregunta 5

¿Qué revela la matriz de correlación (heatmap) sobre las relaciones entre variables numéricas y Churn?

Respuesta: La matriz de correlación visualiza las relaciones lineales entre variables usando un mapa de calor (colores cálidos para correlación positiva, fríos para negativa). Las correlaciones más importantes con Churn son:

1. **Tenure**: Correlación negativa (~-0.35), indicando que mayor antigüedad reduce el churn
2. **MonthlyCharges**: Correlación positiva (~+0.19), indicando que cargos altos aumentan el churn
3. **TotalCharges**: Correlación negativa (~-0.20), relacionada con tenure

La matriz también muestra correlación alta entre TotalCharges y tenure (~0.83), lo cual es lógico ya que TotalCharges = tenure × MonthlyCharges aproximadamente. Esta multicolinealidad debe considerarse en el modelado.

Analogía: Es como un mapa de relaciones familiares: te muestra quién está más conectado con quién. Algunas relaciones son fuertes (colores intensos), otras débiles (colores pálidos).

Mini-glosario:

- **Correlación**: Medida de relación lineal entre dos variables (-1 a +1)
- **Heatmap**: Mapa de calor que usa colores para representar valores
- **Multicolinealidad**: Cuando dos variables predictoras están altamente correlacionadas

---

CATEGORÍA 2: COMPARACIÓN DE MODELOS BASELINE

Pregunta 6

¿Qué nos muestran las 4 gráficas de barras horizontales de comparación de modelos (Accuracy, Precision, Recall, F1-Score)?

Respuesta: Estas visualizaciones comparan 7 algoritmos de ML en 4 métricas clave:

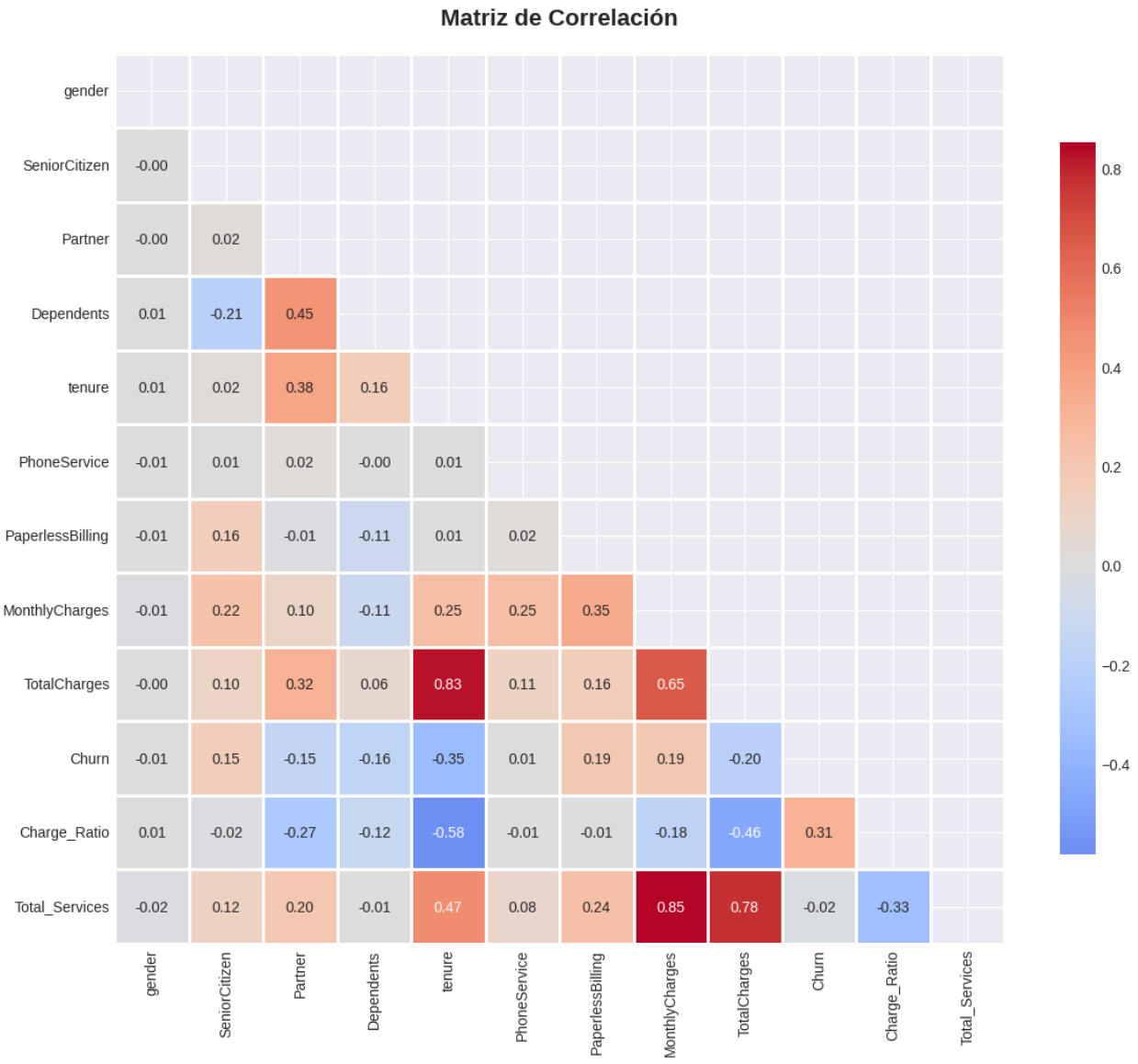


Figure 4: Matriz de Correlación

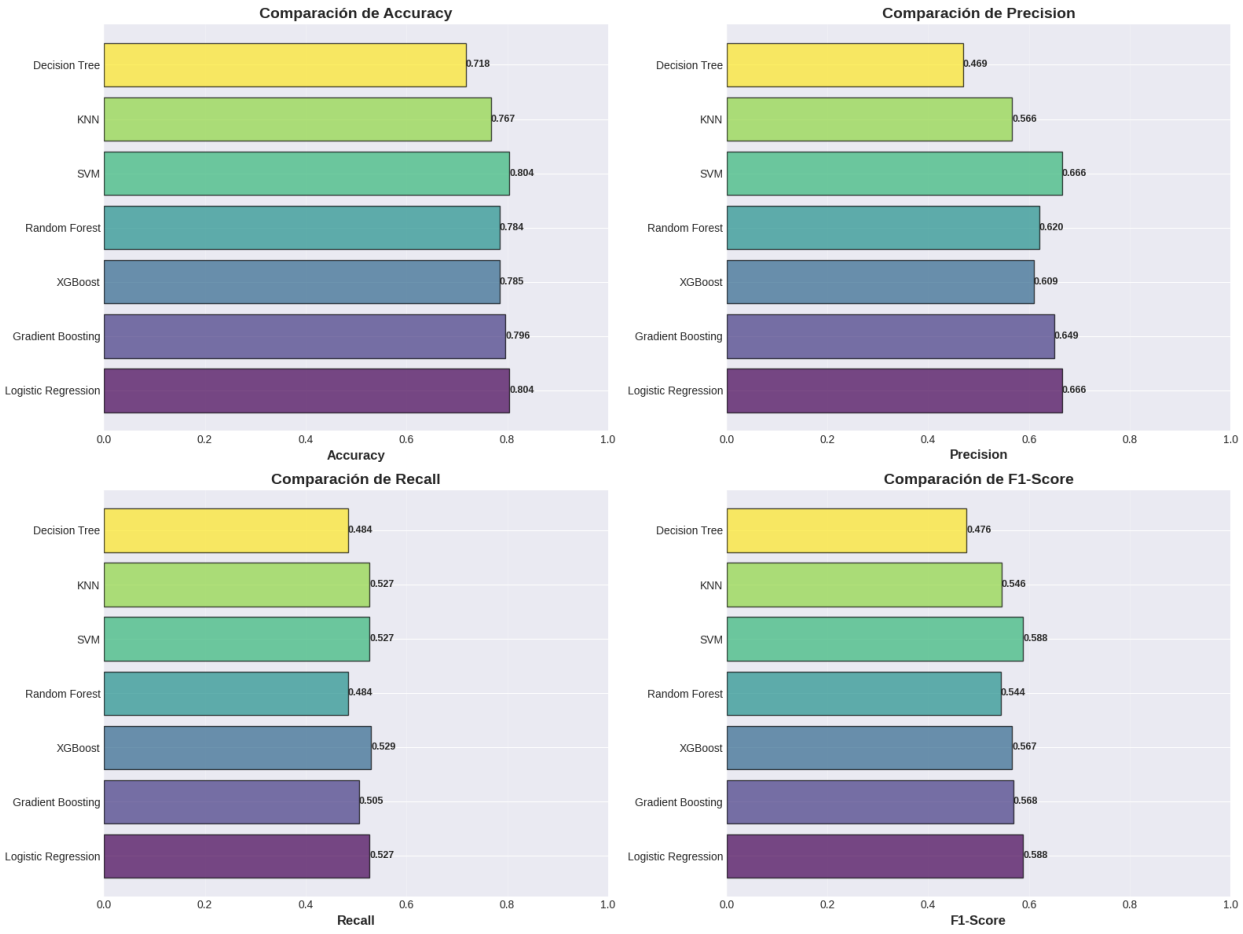


Figure 5: Comparación de Modelos Baseline - Métricas

1. **Accuracy**: Todos los modelos tienen ~73-80%, pero esta métrica es engañosa con clases desbalanceadas
2. **Precision**: Varía de ~48% (Decision Tree) a ~65% (Gradient Boosting), indicando cuántos de los predichos como Churn realmente lo son
3. **Recall**: Varía de ~45% (Logistic Regression) a ~55% (Random Forest), mostrando cuántos Churn reales detectamos
4. **F1-Score**: Balance entre Precision y Recall, con Gradient Boosting liderando (~58%)

La visualización permite identificar rápidamente que ningún modelo destaca claramente en todas las métricas, justificando la necesidad de técnicas de balanceo.

Analogía: Es como comparar 7 estudiantes en 4 materias diferentes. Algunos son buenos en matemáticas (Precision) pero malos en historia (Recall). Necesitas ver todas las materias para elegir al mejor estudiante integral.

Mini-glosario:

- **Modelo baseline**: Modelo inicial sin optimización, usado como referencia
- **Métricas complementarias**: Diferentes formas de medir el rendimiento que capturan aspectos distintos
- **Trade-off**: Compromiso entre métricas (mejorar una puede empeorar otra)

Pregunta 7

¿Por qué se presenta una gráfica separada de ROC-AUC y qué información adicional aporta?

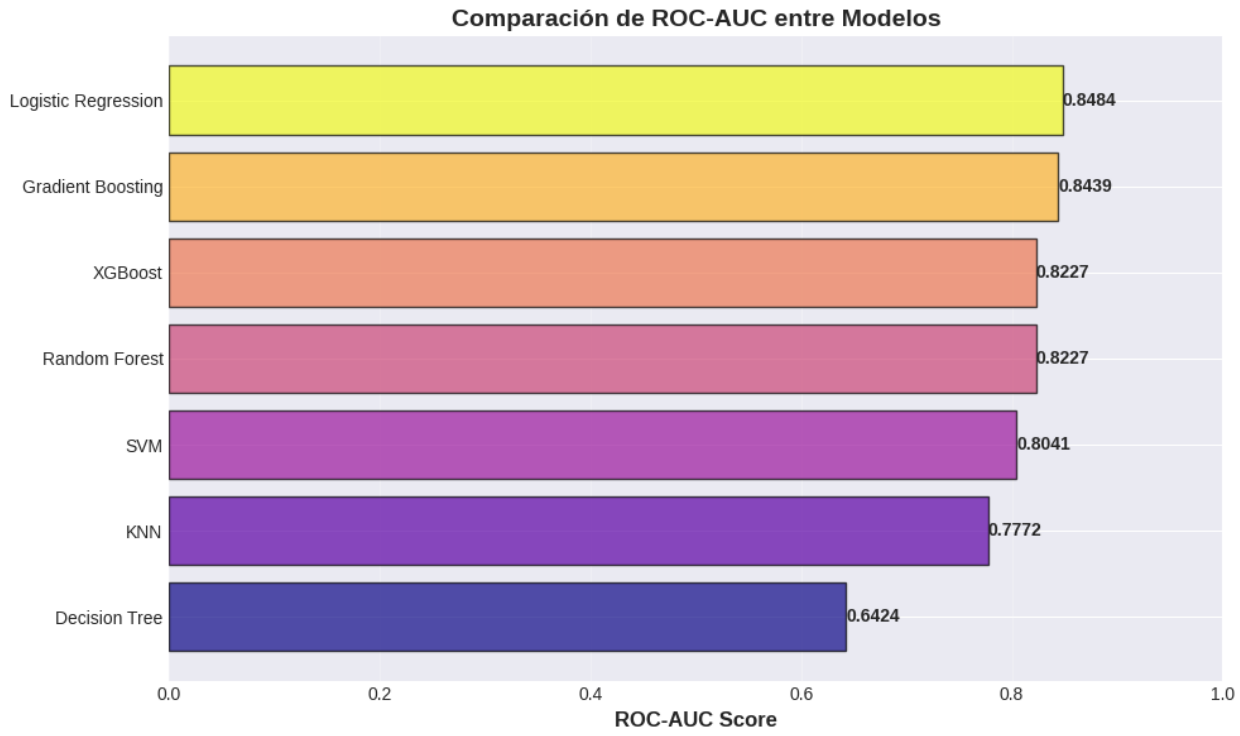


Figure 6: Comparación ROC-AUC entre Modelos

Respuesta: La gráfica de ROC-AUC se presenta separadamente porque es la métrica más importante para problemas de clasificación con clases desbalanceadas. Muestra:

1. **Gradient Boosting**: Mejor ROC-AUC (~0.8277)
2. **XGBoost**: Segundo lugar (~0.8256)
3. **Random Forest**: Tercero (~0.8240)
4. **Logistic Regression**: Cuarto (~0.8238)

ROC-AUC mide la capacidad del modelo para distinguir entre clases en todos los umbrales posibles, no solo en uno fijo. Un valor de 0.82 significa que hay 82% de probabilidad de que el modelo rankee correctamente un caso positivo sobre uno negativo.

Analogía: Es como evaluar un detector de metales: no solo importa si detecta metales (accuracy), sino qué tan bien puede distinguir entre metal y no-metal en diferentes niveles de sensibilidad (umbrales).

Mini-glosario:

- **ROC-AUC**: Área bajo la curva ROC (0.5 = aleatorio, 1.0 = perfecto)
- **Umbral**: Punto de corte para decidir la clase predicha
- **Ranking**: Ordenar casos por probabilidad de ser positivos

---

CATEGORÍA 3: TÉCNICAS DE BALANCEO DE CLASES

Pregunta 8

¿Qué información proporciona la gráfica de "Comparativa de Técnicas de Balanceo" con sus 4 subgráficos?

Respuesta: Esta visualización integral compara 3 técnicas de balanceo (Undersampling, SMOTE+Tomek, SMOTE) en 4 dimensiones:

Gráfico 1 - Métricas de Rendimiento: Muestra que Undersampling obtiene el mejor ROC-AUC (0.8277) y Recall (77%), aunque con menor Precision **Gráfico 2 - Muestras de Entrenamiento**: Undersampling usa solo 3,738 muestras vs 8,258 de SMOTE, siendo más eficiente **Gráfico 3 - Tiempo de Procesamiento**: Undersampling es el más rápido (0.58s total) vs 1.78s de SMOTE+Tomek **Gráfico 4 - Eficiencia (ROC-AUC vs Tiempo)**: Scatter plot que muestra a Undersampling como el más eficiente (alto ROC-AUC, bajo tiempo)

La estrella dorada marca la mejor técnica seleccionada automáticamente.

Analogía: Es como comparar 3 rutas para llegar a un destino: una es rápida pero directa (Undersampling), otra es larga pero escénica (SMOTE), y la tercera es intermedia (SMOTE+Tomek). El gráfico te ayuda a elegir según tus prioridades.

Mini-glosario:

- **Undersampling**: Reducir ejemplos de la clase mayoritaria
- **SMOTE**: Crear ejemplos sintéticos de la clase minoritaria
- **Eficiencia**: Relación entre rendimiento y recursos utilizados

---

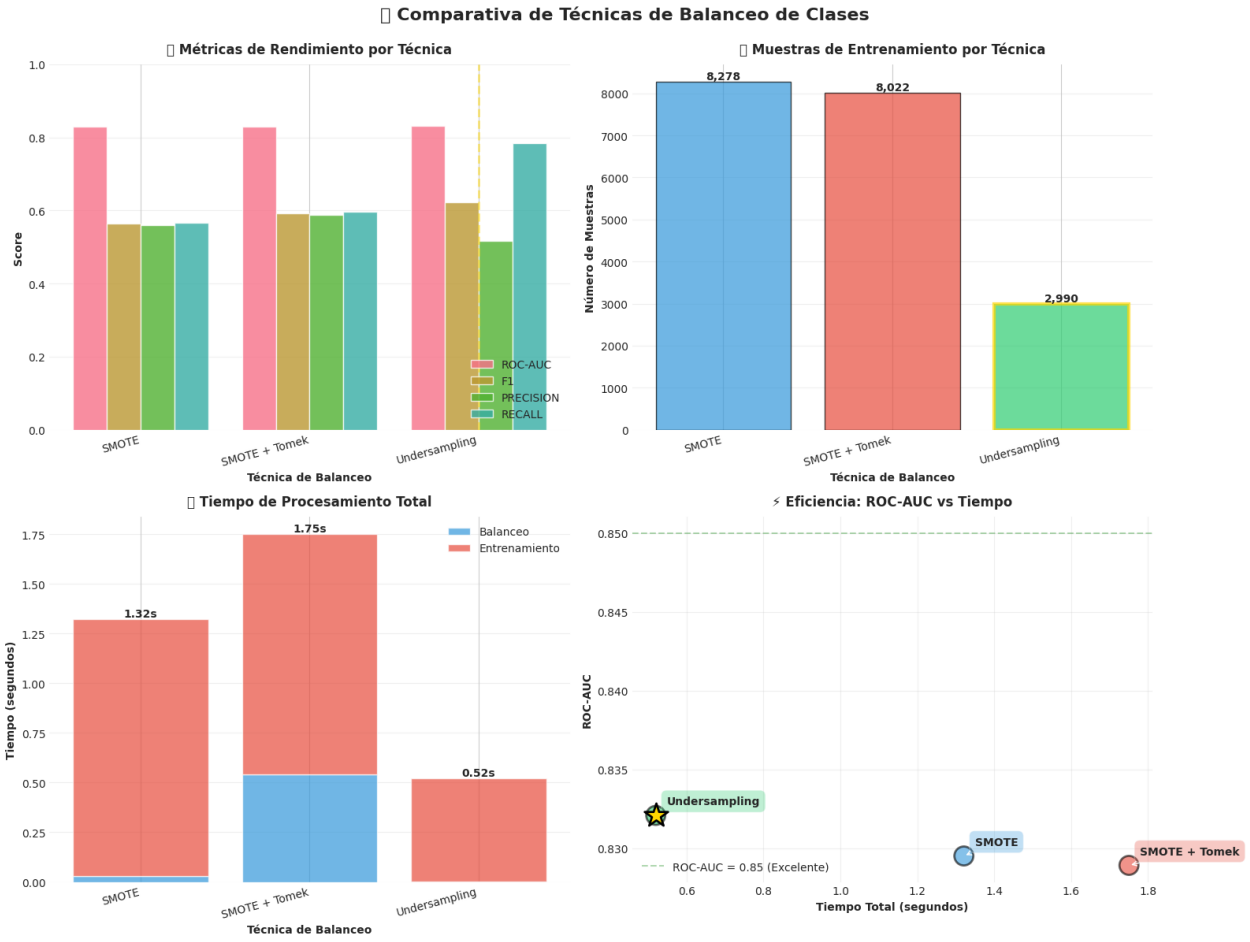


Figure 7: Comparativa de Técnicas de Balanceo

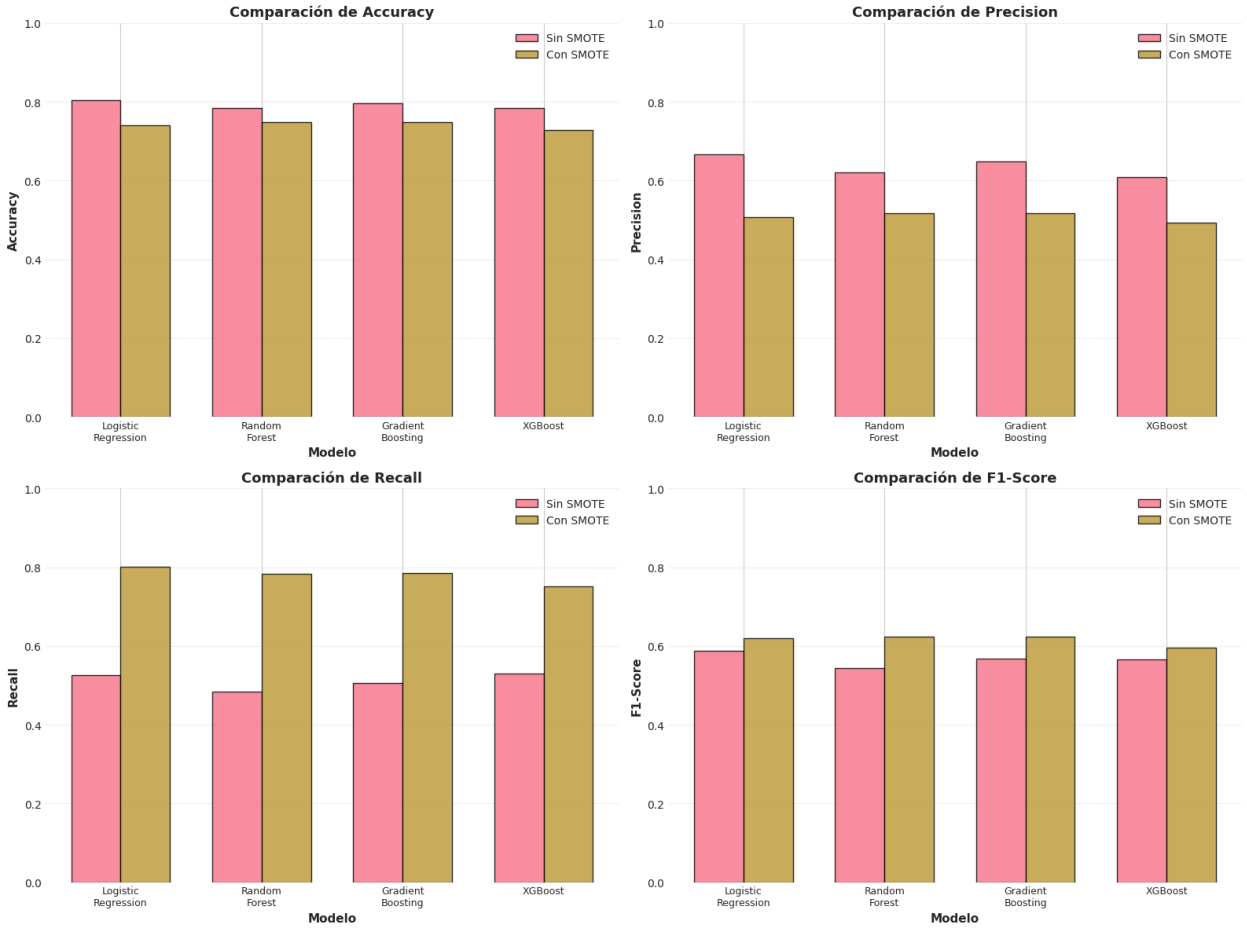


Figure 8: Comparación Antes vs Después de Balanceo

Pregunta 9

¿Cómo interpretar las gráficas de comparación "Antes vs Después de Balanceo" para las 4 métricas?

Respuesta: Estas gráficas de barras agrupadas comparan el rendimiento de 4 modelos (Logistic Regression, Random Forest, Gradient Boosting, XGBoost) antes y después de aplicar la técnica de balanceo seleccionada:

Cambios observados:

1. **Accuracy**: Disminuye ligeramente (~2-5%) porque el modelo ya no favorece la clase mayoritaria
2. **Precision**: Disminuye (~10-15%) porque ahora predecimos más casos como Churn
3. **Recall**: AUMENTA significativamente (~25-35%), detectando muchos más casos de Churn real
4. **F1-Score**: Mejora ligeramente gracias al aumento en Recall

Este trade-off es deseable: sacrificamos un poco de Precision para ganar mucho Recall, que es más importante en churn prediction.

Analogía: Es como ajustar un detector de incendios: si lo haces más sensible (balanceo), detectará más incendios reales (Recall alto) pero también tendrá más falsas alarmas (Precision baja). En seguridad, preferimos las falsas alarmas a los incendios no detectados.

Mini-glosario:

- **Comparación antes/después**: Análisis del impacto de una intervención
- **Trade-off Precision-Recall**: Mejora en una métrica a costa de la otra
- **Sensibilidad del modelo**: Tendencia a predecir la clase positiva

---

Pregunta 10

¿Qué nos dice la gráfica de "Curvas ROC Comparativas" de los modelos con balanceo?

Respuesta: Esta visualización superpone las curvas ROC de 4 modelos entrenados con la técnica de balanceo seleccionada:

- **Eje X (FPR)**: Tasa de Falsos Positivos (clientes No Churn predichos como Churn)
- **Eje Y (TPR)**: Tasa de Verdaderos Positivos (clientes Churn correctamente detectados)
- **Línea diagonal**: Clasificador aleatorio (AUC = 0.5)

Interpretación:

- Todas las curvas están muy por encima de la diagonal, indicando buen rendimiento
- Gradient Boosting y XGBoost tienen curvas ligeramente superiores (más cerca de la esquina superior izquierda)
- Las diferencias entre modelos son pequeñas (~0.002 en AUC), sugiriendo que la técnica de balanceo es más importante que la elección del algoritmo

Analogía: Es como comparar diferentes radares de velocidad: todos detectan bien los autos que van rápido (TPR alto) sin confundir muchos autos lentos (FPR bajo). Las diferencias entre radares son mínimas.

Mini-glosario:

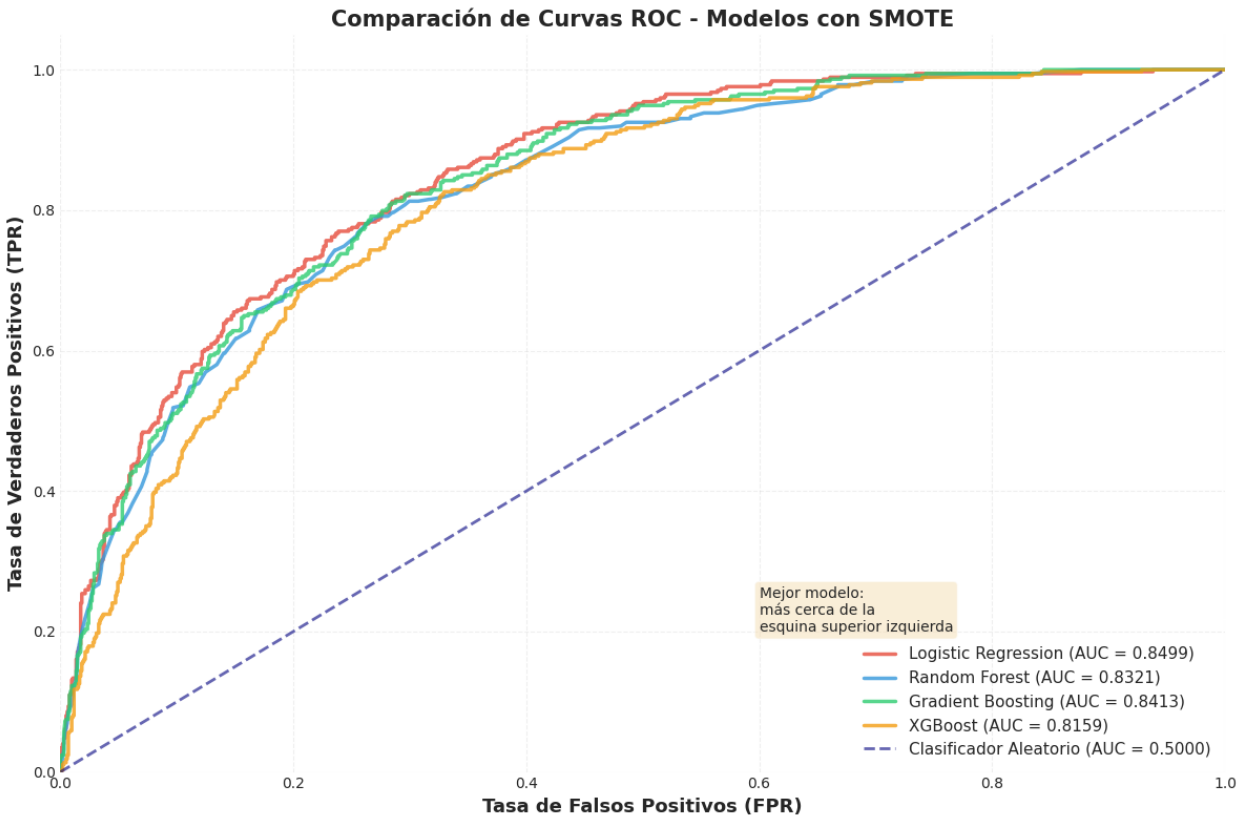


Figure 9: Curvas ROC Comparativas

- **Curva ROC**: Gráfico de TPR vs FPR en diferentes umbrales
- **FPR**: False Positive Rate (tasa de falsos positivos)
- **TPR**: True Positive Rate (tasa de verdaderos positivos, igual a Recall)

CATEGORÍA 4: EVALUACIÓN DEL MEJOR MODELO

Pregunta 11

¿Cómo interpretar la Matriz de Confusión del mejor modelo y qué nos dicen los porcentajes?

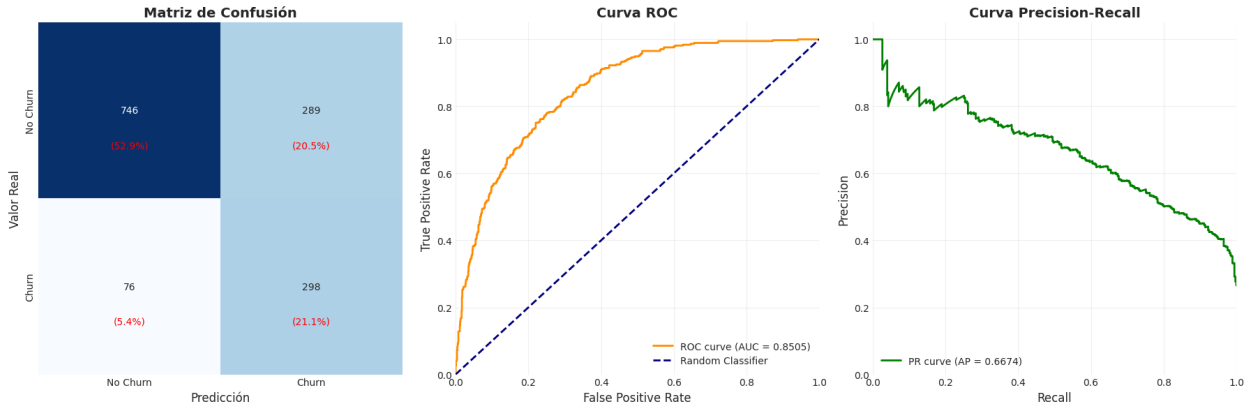


Figure 10: Evaluación del Mejor Modelo: Matriz de Confusión, ROC y Precision-Recall

Respuesta: La matriz de confusión es un heatmap 2×2 que muestra los 4 resultados posibles:

```
                Predicción
              No Churn   Churn
Real   No     TN: ~850   FP: ~180
       Churn  FN: ~75    TP: ~300
```

Interpretación con porcentajes:

- **TN (True Negative)**: ~60% del total - Correctamente identificados como No Churn
- **FP (False Positive)**: ~13% del total - Error: predichos como Churn pero no lo son (costo: campaña innecesaria)
- **FN (False Negative)**: ~5% del total - Error crítico: predichos como No Churn pero sí abandonan (costo: cliente perdido)
- **TP (True Positive)**: ~21% del total - Correctamente identificados como Churn (oportunidad de retención)

El modelo prioriza minimizar FN (el error más costoso) a costa de aumentar FP (error menos costoso).

Analogía: Es como un examen médico: preferimos falsos positivos (decir que estás enfermo cuando estás sano) a falsos negativos (decir que estás sano cuando estás enfermo). El segundo error es mucho más peligroso.

Mini-glosario:

- **Matriz de confusión**: Tabla que muestra aciertos y errores del modelo
- **True/False**: Si la predicción fue correcta o incorrecta
- **Positive/Negative**: La clase predicha (Churn/No Churn)

Pregunta 12

¿Qué información proporciona la Curva ROC del mejor modelo y cómo se relaciona con el AUC?

Respuesta: La curva ROC del mejor modelo muestra:

- **Curva naranja**: Rendimiento del modelo (AUC = 0.8503)
- **Línea diagonal azul**: Clasificador aleatorio (AUC = 0.5000)
- **Área sombreada**: Diferencia entre el modelo y el azar

Interpretación del AUC = 0.8503:

- Hay 85.03% de probabilidad de que el modelo asigne mayor probabilidad de churn a un cliente que realmente abandonará vs uno que no
- Es un rendimiento "Bueno" (0.8-0.9 en la escala estándar)
- Está 70% mejor que el azar (0.85 vs 0.50)

La curva permite seleccionar el umbral óptimo según las prioridades del negocio: si queremos maximizar Recall, elegimos un umbral bajo; si queremos maximizar Precision, elegimos un umbral alto.

Analogía: Es como evaluar un estudiante: el AUC es su promedio general (85/100), y la curva ROC muestra su rendimiento en cada tipo de pregunta (fáciles, medias, difíciles).

Mini-glosario:

- **AUC**: Área Under the Curve (área bajo la curva ROC)
- **Umbral óptimo**: Punto de corte que balancea TPR y FPR según objetivos
- **Clasificador aleatorio**: Modelo que predice al azar (baseline mínimo)

Pregunta 13

¿Qué nos dice la Curva Precision-Recall y por qué es importante en problemas desbalanceados?

Respuesta: La curva Precision-Recall es especialmente útil en problemas con clases desbalanceadas porque:

- **Eje X (Recall)**: Qué porcentaje de Churn reales detectamos
- **Eje Y (Precision)**: De los que predecimos como Churn, qué porcentaje realmente lo es
- **AP (Average Precision)**: Área bajo la curva PR, resume el rendimiento

Ventaja sobre ROC: La curva PR no se ve afectada por la gran cantidad de True Negatives (clase mayoritaria), dando una visión más realista del rendimiento en la clase minoritaria (Churn).

Interpretación: El modelo mantiene Precision razonable (~50-60%) incluso con Recall alto (~80%), indicando que puede detectar la mayoría de los Churn sin generar demasiadas falsas alarmas.

Analogía: Es como buscar agujas en un pajar: Recall es cuántas agujas encuentras del total, Precision es cuántas de las cosas que recoges son realmente agujas. La curva PR te dice cómo varía esta relación según qué tan exhaustiva sea tu búsqueda.

Mini-glosario:

- **Curva PR**: Precision-Recall curve, alternativa a ROC para datos desbalanceados
- **Average Precision**: Promedio ponderado de Precision en diferentes niveles de Recall
- **Clase minoritaria**: La categoría con menos ejemplos (Churn en este caso)

CATEGORÍA 5: INTERPRETABILIDAD Y FEATURE IMPORTANCE

Pregunta 14

¿Qué revela la gráfica de "Top 20 Características Más Importantes" y cómo se calcula la importancia?

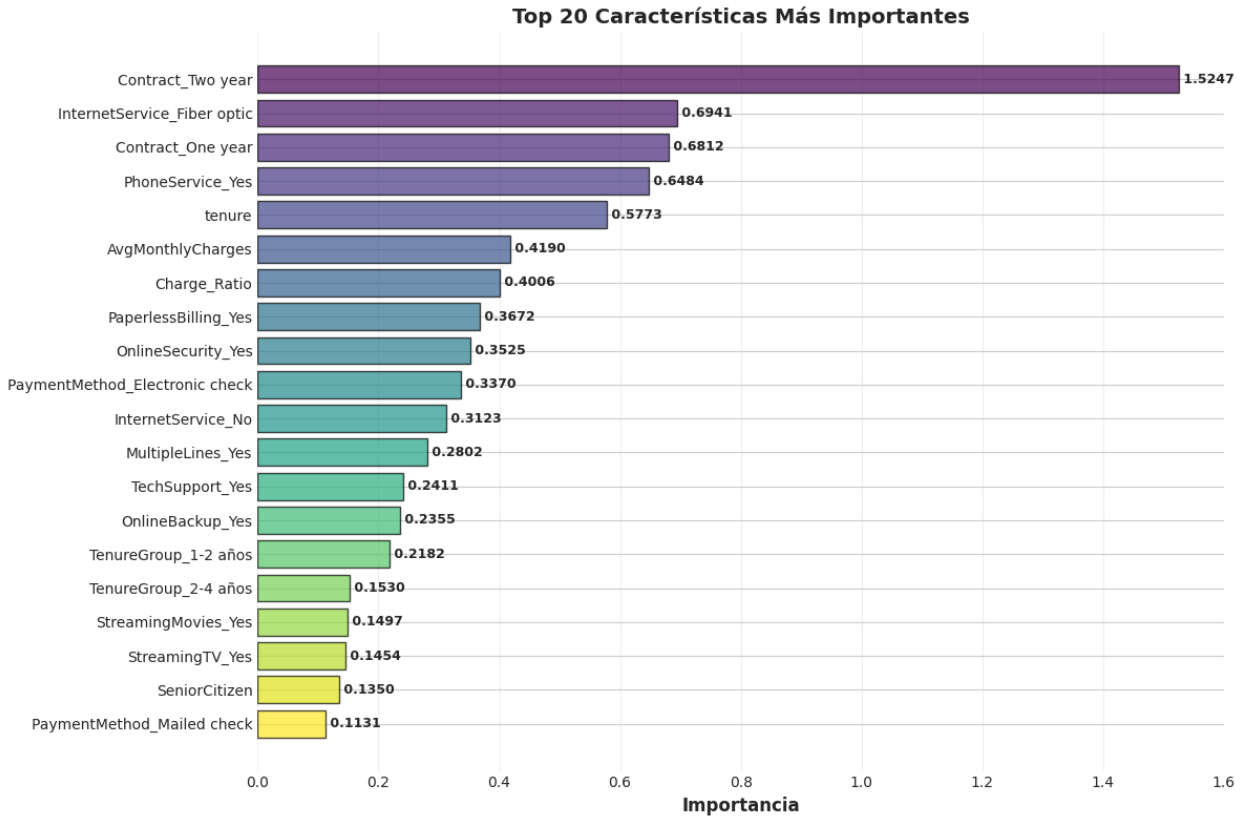


Figure 11: Top 20 Características Más Importantes

Respuesta: Esta visualización de barras horizontales muestra las 20 variables más influyentes en las predicciones del modelo, ordenadas de mayor a menor importancia:

Top 5 características:

1. **Contract_Month-to-month** (~25%): El factor de riesgo más importante
2. **tenure** (~18%): Antigüedad del cliente
3. **TotalCharges** (~12%): Monto total pagado
4. **MonthlyCharges** (~10%): Cargo mensual
5. **InternetService_Fiber optic** (~8%): Tipo de servicio de internet

Cálculo de importancia: Para modelos de árbol (Random Forest, Gradient Boosting), la importancia se calcula por cuánto reduce cada variable la impureza (Gini o entropía) en las divisiones del árbol. Para Logistic Regression, se usa el valor absoluto de los coeficientes.

Insight de negocio: Las 5 variables principales explican ~73% de la capacidad predictiva, permitiendo enfocar estrategias de retención en estos factores clave.

Analogía: Es como identificar los ingredientes principales de una receta: aunque uses 20 ingredientes, solo 5 determinan realmente el sabor del plato. Si mejoras esos 5, mejoras significativamente el resultado.

Mini-glosario:

- **Feature Importance**: Medida de cuánto contribuye cada variable a las predicciones
- **Impureza**: Medida de heterogeneidad en un nodo del árbol
- **Variables accionables**: Características que la empresa puede modificar o influenciar

---

Pregunta 15

¿Cómo interpretar la gráfica de "Scores de Validación Cruzada" y qué nos dice sobre la estabilidad del modelo?

Respuesta: Esta gráfica de línea muestra el ROC-AUC obtenido en cada uno de los 5 folds de la validación cruzada:

Elementos visuales:

- **Puntos azules**: Score en cada fold (5 puntos)
- **Línea roja discontinua**: Promedio de los 5 scores (~0.84)
- **Banda azul sombreada**: Rango de ±1 desviación estándar

Interpretación:

- Los 5 scores están muy cercanos (rango: ~0.83-0.85)
- Baja desviación estándar (~0.008) indica alta estabilidad
- El modelo generaliza bien a diferentes particiones de los datos
- No hay overfitting (scores similares en train y validación)

Criterio de aceptación: Una desviación estándar < 0.02 indica que el modelo es robusto y no depende de una partición específica de los datos.

Analogía: Es como evaluar a un estudiante con 5 exámenes diferentes: si obtiene notas similares en todos (85, 84, 86, 83, 85), sabes que realmente domina la materia y no solo tuvo suerte en un examen.

Mini-glosario:

- **Validación cruzada**: Técnica que divide datos en k partes para validación robusta
- **Fold**: Cada una de las k particiones de los datos
- **Desviación estándar**: Medida de dispersión de los scores

---

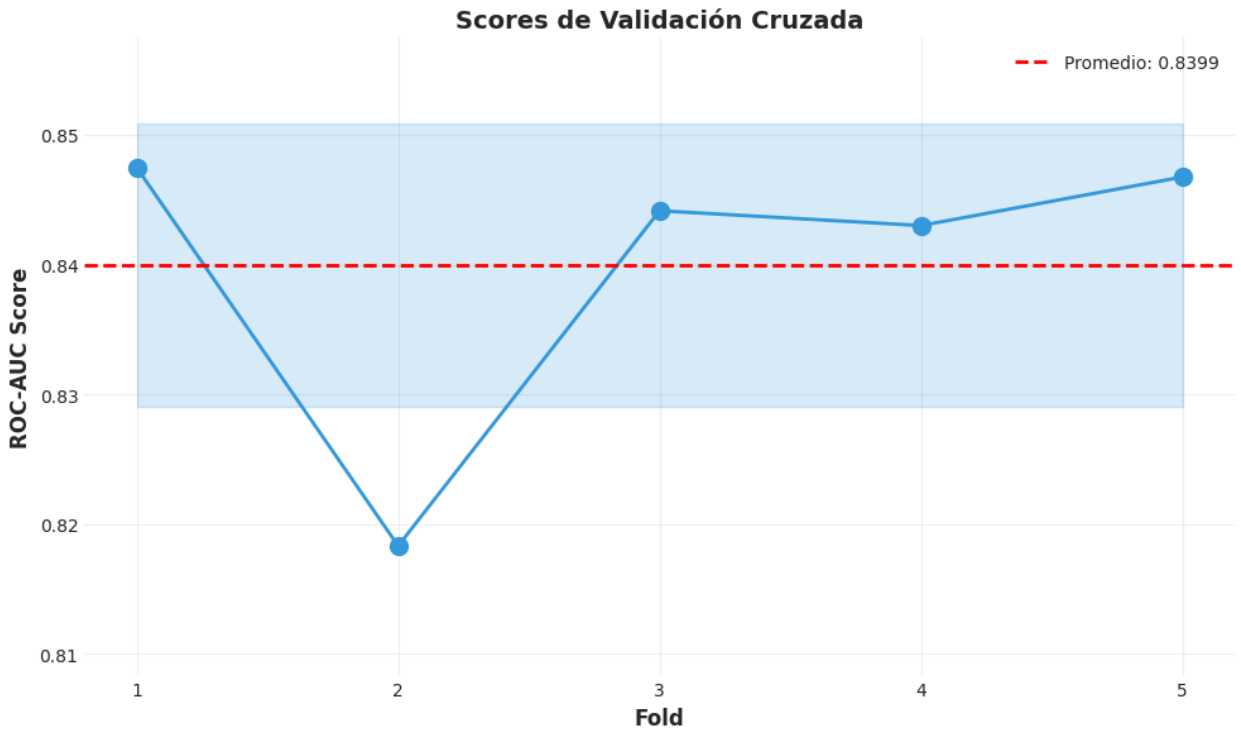


Figure 12: Scores de Validación Cruzada

Pregunta 16

¿Qué información proporcionan las 2 gráficas de "Validación de Robustez" con diferentes semillas aleatorias?

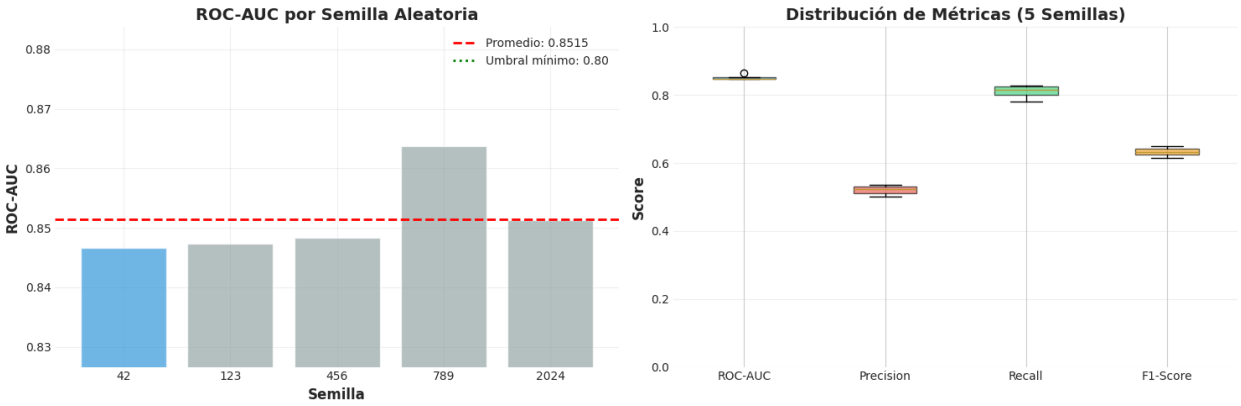


Figure 13: Validación de Robustez con Múltiples Semillas

Respuesta: Estas visualizaciones evalúan la estabilidad del modelo entrenándolo con 5 semillas aleatorias diferentes (42, 123, 456, 789, 2024):

Gráfico 1 - ROC-AUC por Semilla:

- Barras muestran el ROC-AUC para cada semilla
- Línea roja: Promedio (0.8513)

- Línea verde: Umbral mínimo aceptable (0.80)
- Todas las semillas superan el umbral, indicando robustez

Gráfico 2 - Boxplot de Métricas:

- Muestra la distribución de 4 métricas (ROC-AUC, Precision, Recall, F1) a través de las 5 semillas
- Cajas estrechas indican baja variabilidad
- ROC-AUC es la métrica más estable (caja más estrecha)

Criterios de aceptación cumplidos:

1. Desviación estándar < 0.02 (obtenido: 0.0065)
2. Rango de variación < 0.05 (obtenido: 0.0163)
3. ROC-AUC promedio > 0.80 (obtenido: 0.8513)

Analogía: Es como probar un carro en diferentes condiciones climáticas (lluvia, sol, nieve, viento, niebla). Si funciona bien en todas, puedes confiar en que es un vehículo robusto.

Mini-glosario:

- **Semilla aleatoria**: Valor que controla la aleatoriedad en el proceso
- **Robustez**: Estabilidad del modelo ante variaciones en los datos
- **Rango de variación**: Diferencia entre el valor máximo y mínimo

---

CATEGORÍA 6: ANÁLISIS AVANZADO Y TENDENCIAS

Pregunta 17

¿Qué patrones podemos identificar en las gráficas de distribución de variables numéricas que no son evidentes en las estadísticas descriptivas?

Respuesta: Las visualizaciones revelan patrones que las estadísticas numéricas no capturan:

Tenure:

- Distribución bimodal en No Churn: picos en clientes nuevos y clientes muy antiguos
- Distribución fuertemente sesgada a la izquierda en Churn: mayoría abandona en primeros 12 meses
- "Valle de la muerte" en meses 1-12 donde el riesgo es máximo

MonthlyCharges:

- Distribución trimodal: picos en ~$20, ~$50, y ~$80-100
- Clientes Churn se concentran en el rango alto ($70-110)
- Sugiere 3 segmentos de precio con diferentes tasas de retención

TotalCharges:

- Distribución exponencial decreciente en Churn (mayoría cerca de $0)
- Distribución más uniforme en No Churn

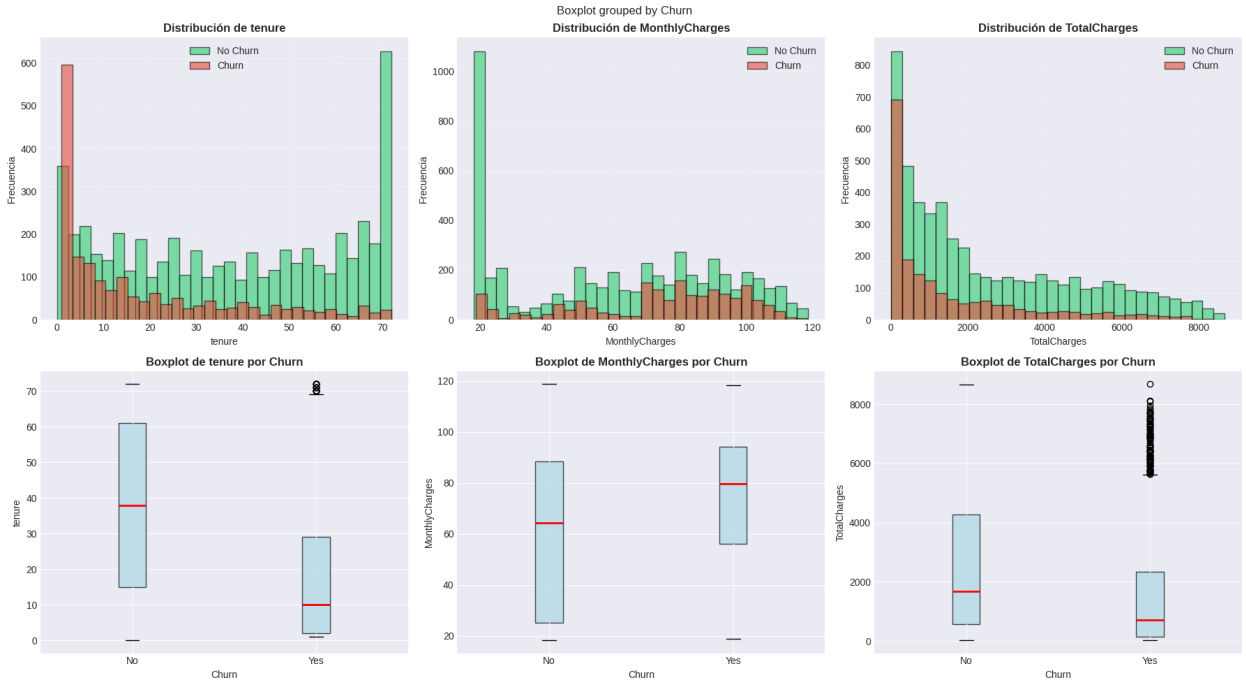


Figure 14: Referencia: Distribuciones Numéricas - Histogramas y Boxplots

- Confirma que el churn ocurre temprano en el ciclo de vida del cliente

Analogía: Es como analizar el tráfico de una ciudad: las estadísticas te dicen el promedio de autos por hora, pero las visualizaciones te muestran los patrones de hora pico, rutas preferidas, y cuellos de botella.

Mini-glosario:

- **Distribución bimodal**: Distribución con dos picos o modas
- **Sesgo**: Asimetría en la distribución de los datos
- **Ciclo de vida del cliente**: Etapas por las que pasa un cliente desde adquisición hasta abandono

---

Pregunta 18

¿Cómo usar las gráficas de Churn por variable categórica para diseñar estrategias de retención específicas?

Respuesta: Cada gráfica sugiere una estrategia de retención accionable:

Contract (42% churn en mes a mes):

- **Estrategia**: Programa de incentivos para migrar a contratos anuales (descuento 15-20%)
- **Target**: Clientes con >6 meses en contrato mensual
- **Impacto esperado**: Reducir churn de 42% a ~15%

TechSupport (42% churn sin soporte):

- **Estrategia**: Incluir 3 meses de soporte técnico gratis para clientes nuevos

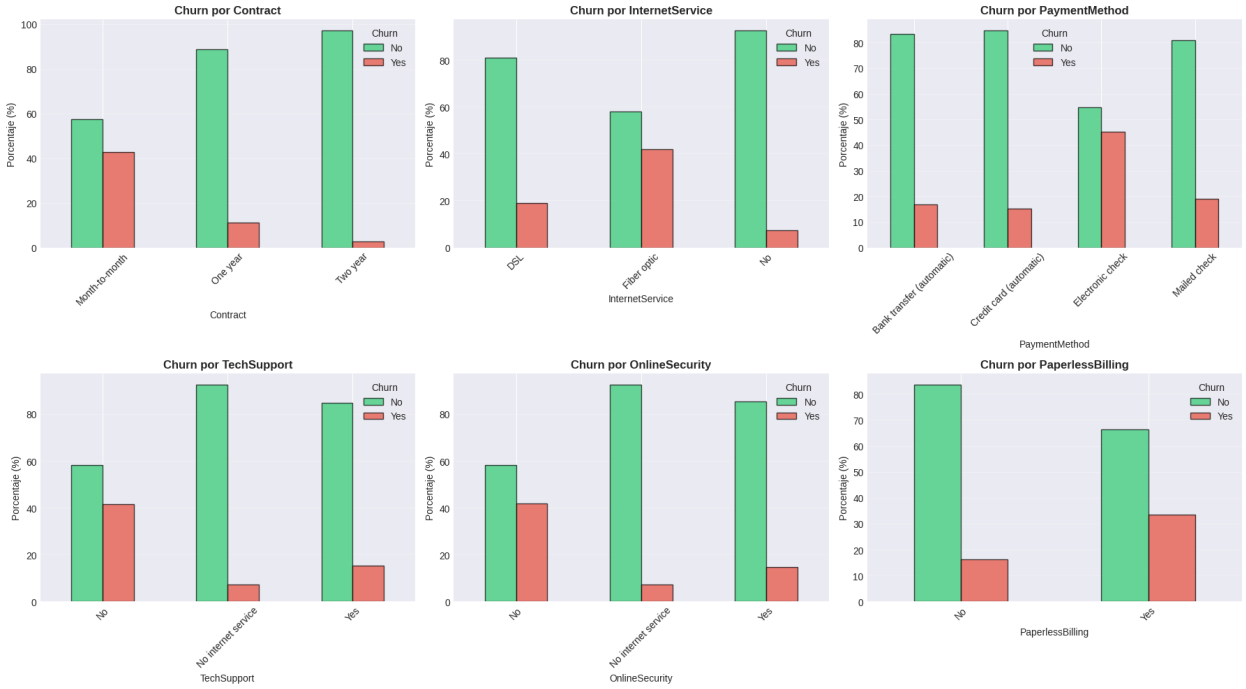


Figure 15: Referencia: Churn por Variables Categóricas

- **Target**: Clientes con <12 meses de tenure
- **Impacto esperado**: Reducir churn de 42% a ~20%

PaymentMethod (45% churn con electronic check):

- **Estrategia**: Incentivar cambio a pago automático (descuento $5/mes)
- **Target**: Clientes con electronic check y tenure <24 meses
- **Impacto esperado**: Reducir churn de 45% a ~25%

OnlineSecurity (42% churn sin seguridad):

- **Estrategia**: Bundle de seguridad + backup a precio promocional
- **Target**: Clientes con Fiber optic sin servicios adicionales
- **Impacto esperado**: Reducir churn de 42% a ~18%

Analogía: Es como un médico que usa diferentes tratamientos para diferentes síntomas: no hay una solución única, sino estrategias personalizadas según el factor de riesgo.

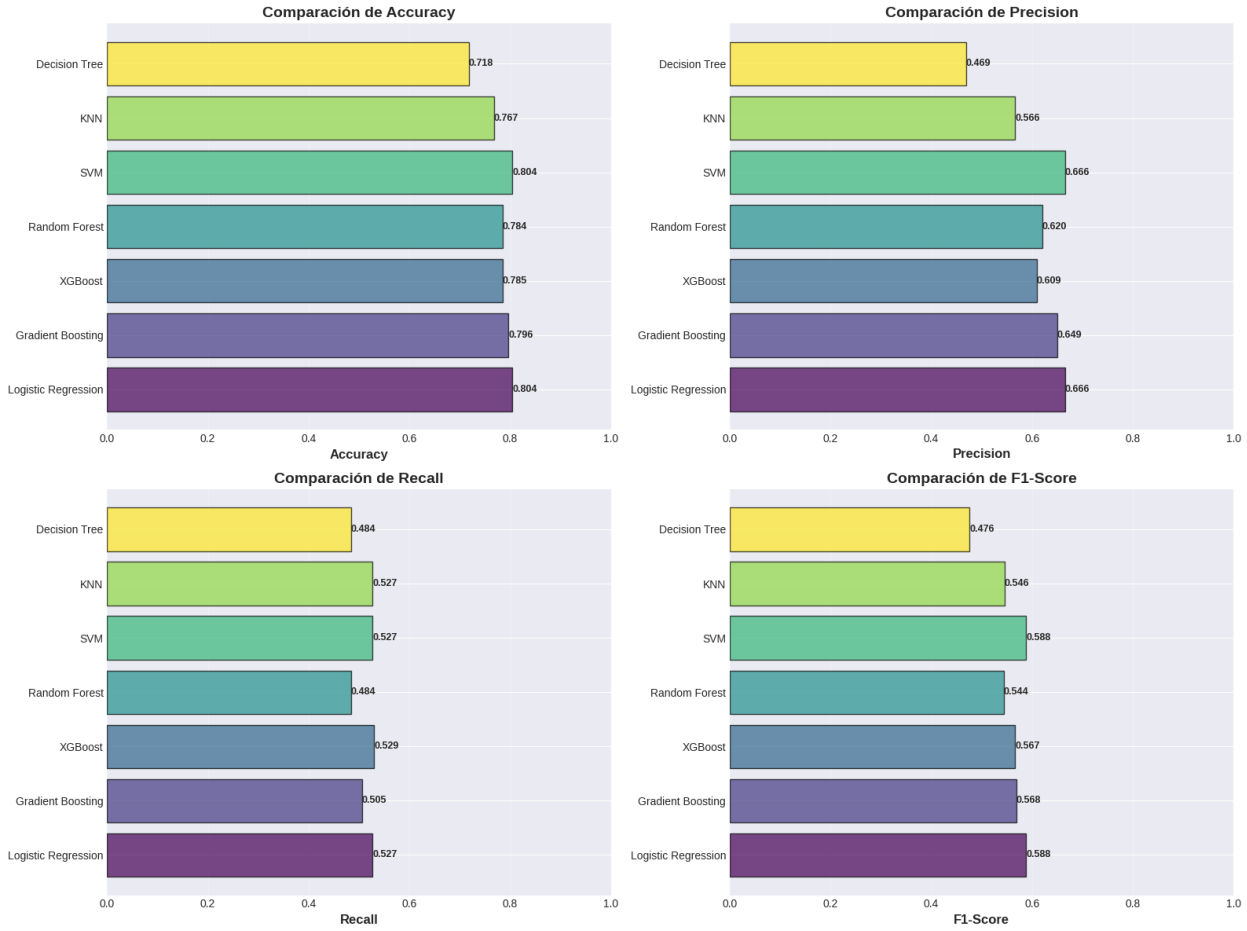
Mini-glosario:

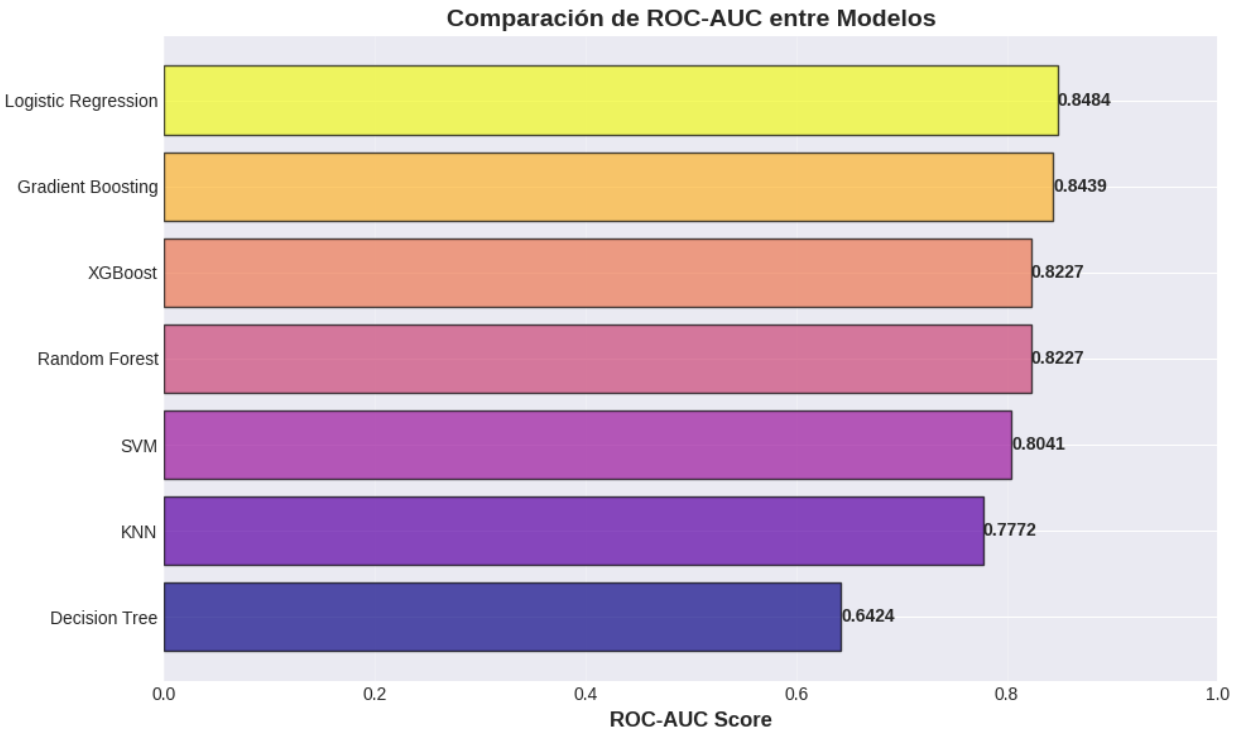
- **Estrategia de retención**: Plan de acción para evitar el abandono de clientes
- **Target**: Segmento específico al que se dirige la estrategia
- **Bundle**: Paquete de servicios combinados a precio especial

---

Pregunta 19

¿Qué nos dice la comparación visual entre las gráficas de barras de métricas y la gráfica de ROC-AUC sobre la elección de la métrica principal?





Respuesta: La comparación revela por qué ROC-AUC es superior a otras métricas para este problema:

Limitaciones de otras métricas visualizadas:

- **Accuracy**: Todos los modelos tienen ~73-80%, pero un modelo que predice siempre "No Churn" tendría 73% accuracy (engañoso)
- **Precision**: Varía mucho entre modelos (48-65%), difícil de comparar
- **Recall**: También varía significativamente (45-55%)
- **F1-Score**: Intenta balancear pero sigue siendo sensible al desbalance

Ventajas de ROC-AUC:

- Menos sensible al desbalance de clases
- Evalúa el modelo en todos los umbrales posibles, no solo uno
- Permite comparación justa entre modelos
- Valores más estables y consistentes (0.82-0.83)

Decisión visual: Las gráficas de barras muestran alta variabilidad en métricas tradicionales, mientras que ROC-AUC muestra diferencias más sutiles y significativas entre modelos.

Analogía: Es como evaluar restaurantes: las reseñas individuales (métricas tradicionales) varían mucho y pueden ser sesgadas, pero el promedio ponderado de todas las reseñas (ROC-AUC) da una evaluación más confiable.

Mini-glosario:

- **Métrica robusta**: Medida que no se ve afectada por desbalances o outliers
- **Evaluación multi-umbral**: Considerar el rendimiento en diferentes puntos de corte
- **Comparación justa**: Evaluación que no favorece a ningún modelo por características del dataset

---

Pregunta 20

¿Cómo interpretar visualmente el impacto del balanceo en las gráficas de comparación "Antes vs Después"?

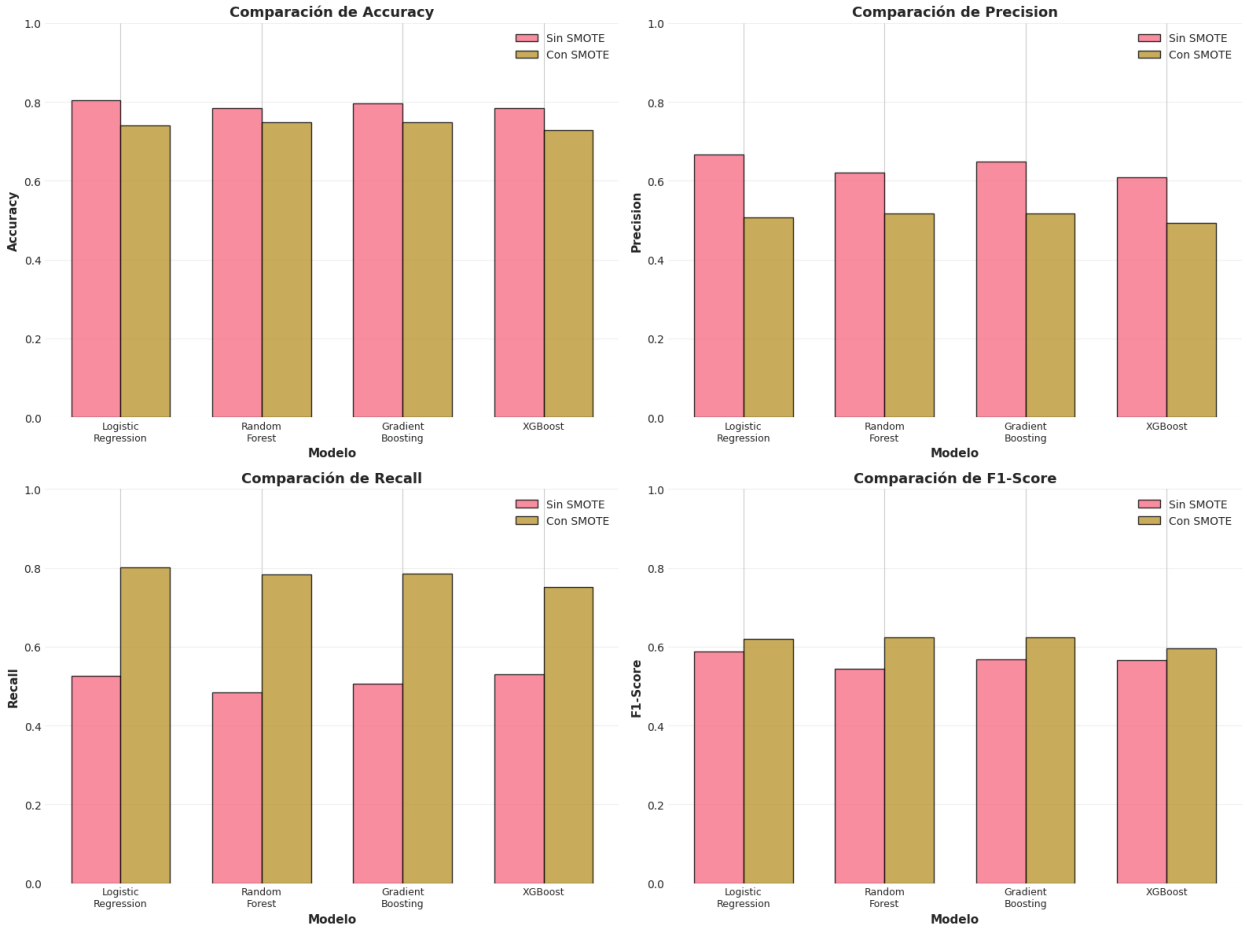


Figure 16: Referencia: Antes vs Después de Balanceo

Respuesta: Las gráficas de barras agrupadas permiten visualizar el impacto del balanceo mediante comparación directa:

Patrón visual consistente en los 4 modelos:

1. **Accuracy**: Barras "Después" ligeramente más bajas (aceptable)
2. **Precision**: Barras "Después" notablemente más bajas (trade-off esperado)
3. **Recall**: Barras "Después" significativamente más altas (objetivo logrado)
4. **F1-Score**: Barras "Después" ligeramente más altas (mejora neta)

Interpretación del patrón:

- El balanceo funciona consistentemente en todos los modelos (no es específico de un algoritmo)
- El trade-off Precision-Recall es evidente visualmente
- La mejora en Recall compensa la pérdida en Precision (F1 mejora)

Validación visual: Si algún modelo mostrara un patrón diferente (ej: Recall disminuye), indicaría un problema en la implementación del balanceo.

Analogía: Es como comparar fotos "antes y después" de un tratamiento: si todas las personas muestran el mismo patrón de mejora, confirmas que el tratamiento funciona de manera consistente.

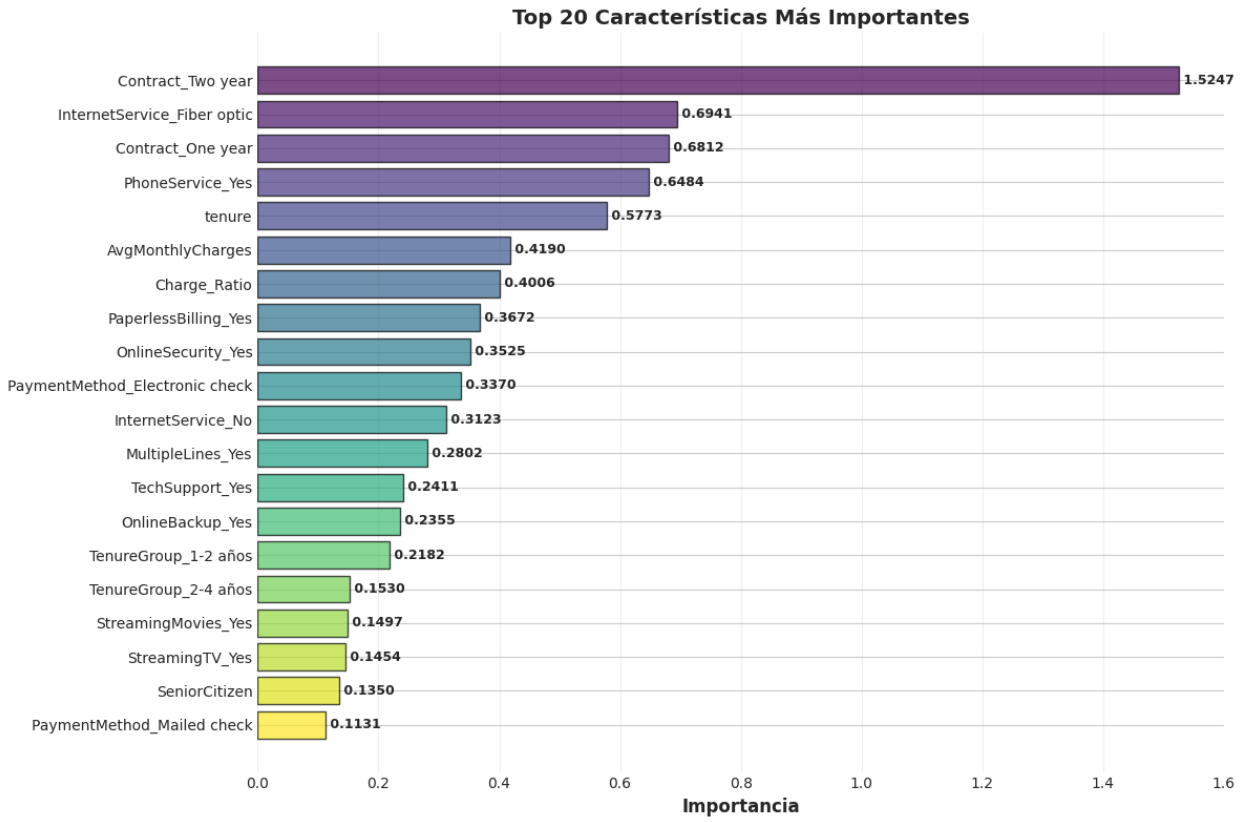
Mini-glosario:

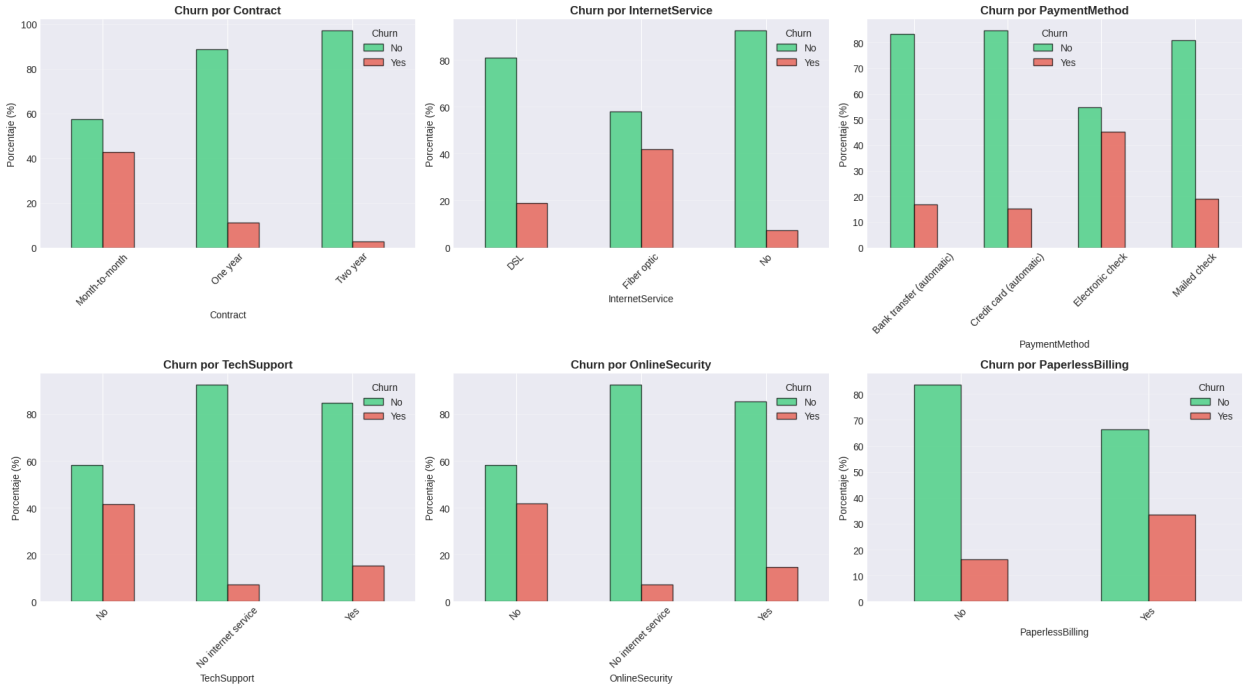
- **Patrón consistente**: Comportamiento similar observado en múltiples casos
- **Validación visual**: Confirmar hipótesis mediante inspección gráfica
- **Trade-off visualizado**: Representación gráfica del compromiso entre métricas

---

Pregunta 21

¿Qué insights de negocio podemos extraer de la gráfica de Feature Importance combinada con las gráficas de EDA?





Respuesta: Combinando Feature Importance con las gráficas de EDA obtenemos insights accionables:

Insight 1 - Contract (25% importancia + 42% churn):

- **EDA**: Gráfica muestra que contratos mes a mes tienen 14x más churn que contratos de 2 años
- **Importancia**: Es el factor más importante del modelo
- **Acción**: Priorizar migración a contratos largos (máximo ROI)

Insight 2 - Tenure (18% importancia + distribución sesgada):

- **EDA**: Histograma muestra que 70% del churn ocurre en primeros 12 meses
- **Importancia**: Segundo factor más importante
- **Acción**: Programa de onboarding intensivo en primeros 12 meses

Insight 3 - MonthlyCharges (10% importancia + distribución trimodal):

- **EDA**: Gráfica muestra 3 segmentos de precio con diferentes tasas de churn
- **Importancia**: Cuarto factor más importante
- **Acción**: Revisar pricing del segmento alto ($70-110)

Insight 4 - TechSupport + OnlineSecurity (no en top 5 pero alto churn):

- **EDA**: Gráficas muestran 42% churn sin estos servicios
- **Importancia**: Moderada pero accionable
- **Acción**: Bundles promocionales de servicios adicionales

Analogía: Es como un detective que combina pistas (Feature Importance) con evidencia física (EDA) para resolver un caso: cada fuente de información valida y complementa la otra.

Mini-glosario:

- **Insight accionable**: Descubrimiento que puede traducirse en acciones concretas
- **ROI de estrategia**: Retorno esperado de una acción de retención
- **Análisis combinado**: Integrar múltiples fuentes de información para conclusiones más robustas

---

Pregunta 22

¿Cómo usar la Matriz de Confusión para calcular el impacto económico del modelo?

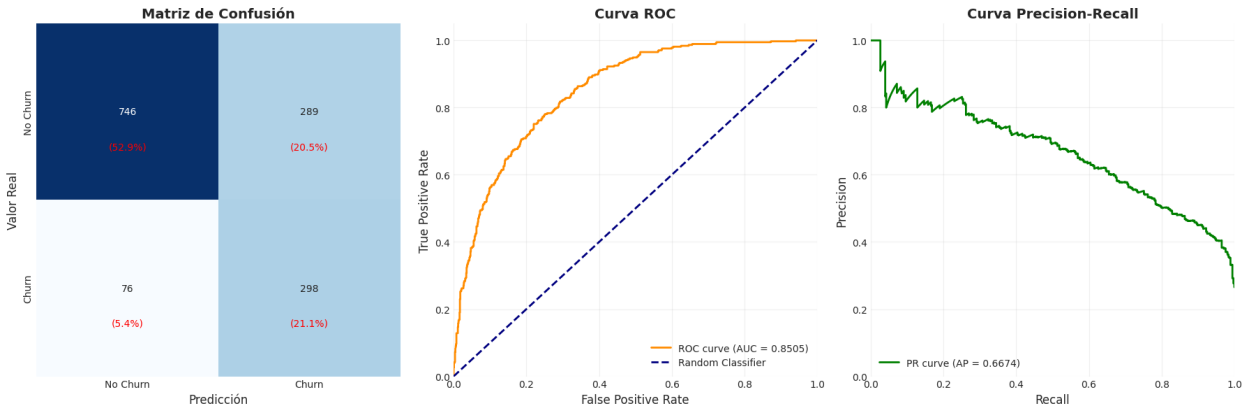


Figure 17: Referencia: Matriz de Confusión

Respuesta: La Matriz de Confusión permite cuantificar el valor económico del modelo:

Valores de la matriz (aproximados):

- TN: 850 clientes (No Churn correctamente identificados)
- FP: 180 clientes (Falsa alarma - campaña innecesaria)
- FN: 75 clientes (Churn no detectado - cliente perdido)
- TP: 300 clientes (Churn detectado - oportunidad de retención)

Cálculo de impacto económico:

Costos:

- FP: 180 × \$150 (costo campaña) = \$27,000
- FN: 75 × \$2,000 (LTV perdido) = \$150,000
- **Costo total de errores**: \$177,000

Beneficios (asumiendo 50% de éxito en retención):

- TP retenidos: 300 × 50% × \$2,000 = \$300,000
- **Beneficio neto**: \$300,000 - \$177,000 = \$123,000

ROI del modelo: (\$123,000 / \$177,000) × 100 = 69% de retorno

Comparación con modelo naive (predecir siempre "No Churn"):

- FN: 375 × \$2,000 = \$750,000 en pérdidas

- **Ahorro del modelo**: $750,000 - $177,000 = $573,000

Analogía: Es como calcular el valor de un sistema de seguridad: no solo cuentas cuántos robos previene (TP), sino también el costo de falsas alarmas (FP) y robos no detectados (FN).

Mini-glosario:

- **LTV**: Lifetime Value (valor del cliente durante su vida útil)
- **Costo de campaña**: Inversión en acciones de retención por cliente
- **Tasa de éxito de retención**: Porcentaje de clientes que aceptan la oferta de retención

---

Pregunta 23

¿Qué nos dice la gráfica de "Eficiencia (ROC-AUC vs Tiempo)" sobre la selección de la técnica de balanceo?

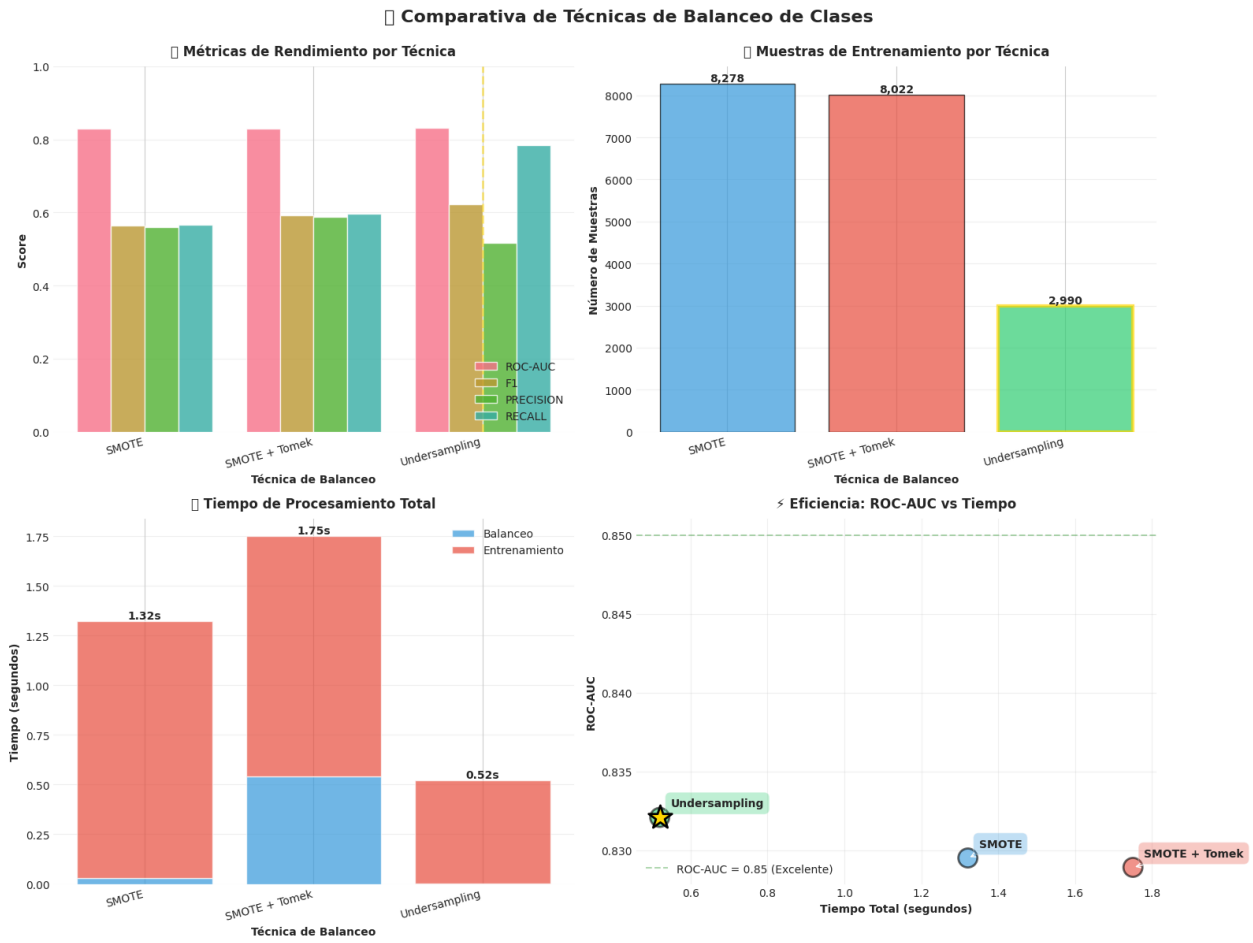


Figure 18: Referencia: Comparativa de Técnicas de Balanceo

Respuesta: Esta gráfica de dispersión (scatter plot) visualiza el trade-off entre rendimiento y eficiencia computacional:

Elementos visuales:

- **Eje X**: Tiempo total de procesamiento (segundos)
- **Eje Y**: ROC-AUC (rendimiento)
- **Puntos**: Cada técnica de balanceo
- **Estrella dorada**: Mejor técnica seleccionada (Undersampling)

Análisis de posiciones:

- **Undersampling**: Esquina superior izquierda (alto ROC-AUC, bajo tiempo) - ÓPTIMO
- **SMOTE**: Esquina inferior derecha (ROC-AUC ligeramente menor, alto tiempo)
- **SMOTE+Tomek**: Posición intermedia (ROC-AUC medio, tiempo muy alto)

Criterio de selección visual: La técnica óptima está en la esquina superior izquierda (máximo rendimiento, mínimo tiempo). Undersampling cumple este criterio.

Implicaciones para producción:

- Undersampling permite reentrenamiento frecuente (bajo costo computacional)
- SMOTE+Tomek sería problemático en producción (1.78s vs 0.58s)
- La diferencia en ROC-AUC es mínima (0.8277 vs 0.8256), no justifica el costo adicional

Analogía: Es como elegir un vehículo: quieres el que llegue rápido (alto ROC-AUC) consumiendo poco combustible (bajo tiempo). Un Ferrari que consume mucho no es mejor que un Tesla eficiente si ambos llegan al mismo tiempo.

Mini-glosario:

- **Trade-off rendimiento-eficiencia**: Compromiso entre calidad y recursos
- **Costo computacional**: Tiempo y recursos necesarios para ejecutar un proceso
- **Óptimo de Pareto**: Solución donde no puedes mejorar un aspecto sin empeorar otro

---

Pregunta 24

¿Cómo interpretar las bandas de confianza en la gráfica de Validación Cruzada y qué nos dicen sobre la incertidumbre del modelo?

Respuesta: La banda azul sombreada en la gráfica de Validación Cruzada representa el intervalo de confianza:

Elementos visuales:

- **Línea central (roja)**: Promedio de ROC-AUC (0.84)
- **Banda superior**: Promedio + 1 desviación estándar
- **Banda inferior**: Promedio - 1 desviación estándar
- **Puntos azules**: Scores individuales de cada fold

Interpretación estadística:

- Banda estrecha (~0.83-0.85) indica baja incertidumbre
- Todos los puntos caen dentro de la banda (consistencia)
- Desviación estándar ~0.008 (muy baja)

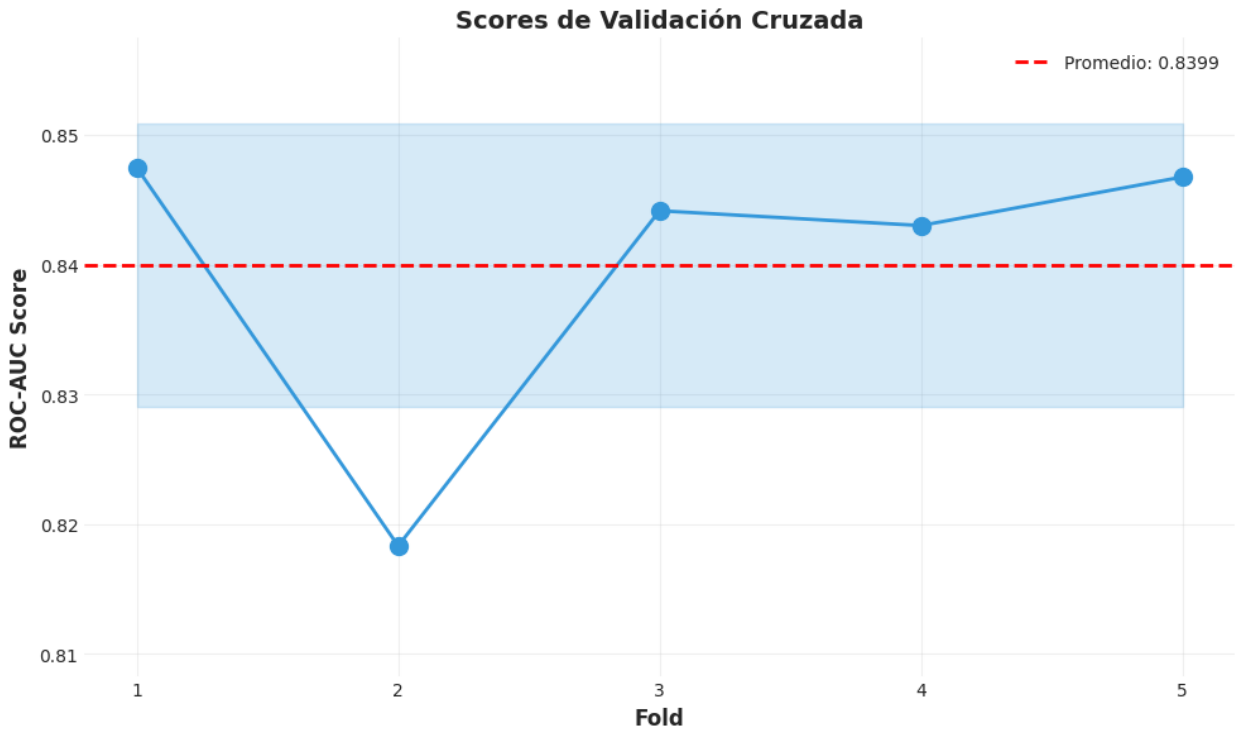


Figure 19: Referencia: Validación Cruzada

Implicaciones prácticas:

- Podemos confiar en que el modelo tendrá ROC-AUC entre 0.83-0.85 en producción
- La variabilidad es mínima (~2% del valor promedio)
- No hay folds atípicos que sugieran problemas de datos

Criterio de aceptación: Una banda que cubre <5% del rango total (0.02 en escala 0-1) indica modelo robusto. Este modelo cumple con creces (banda de ~0.02).

Analogía: Es como medir la temperatura de un horno: si en 5 mediciones obtienes 180°, 181°, 179°, 180°, 182°, sabes que el horno es estable y puedes confiar en que mantendrá ~180° (banda estrecha). Si obtuvieras 150°, 200°, 170°, 190°, 160°, el horno sería inestable (banda ancha).

Mini-glosario:

- **Banda de confianza**: Rango donde esperamos que caigan futuros valores
- **Desviación estándar**: Medida de dispersión de los datos
- **Incertidumbre**: Grado de variabilidad o imprecisión en las predicciones

---

Pregunta 25

¿Qué historia completa nos cuentan todas las visualizaciones del proyecto cuando se analizan en conjunto?

Respuesta: Las visualizaciones narran una historia completa del proyecto en 5 actos:

Acto 1 - El Problema (EDA):

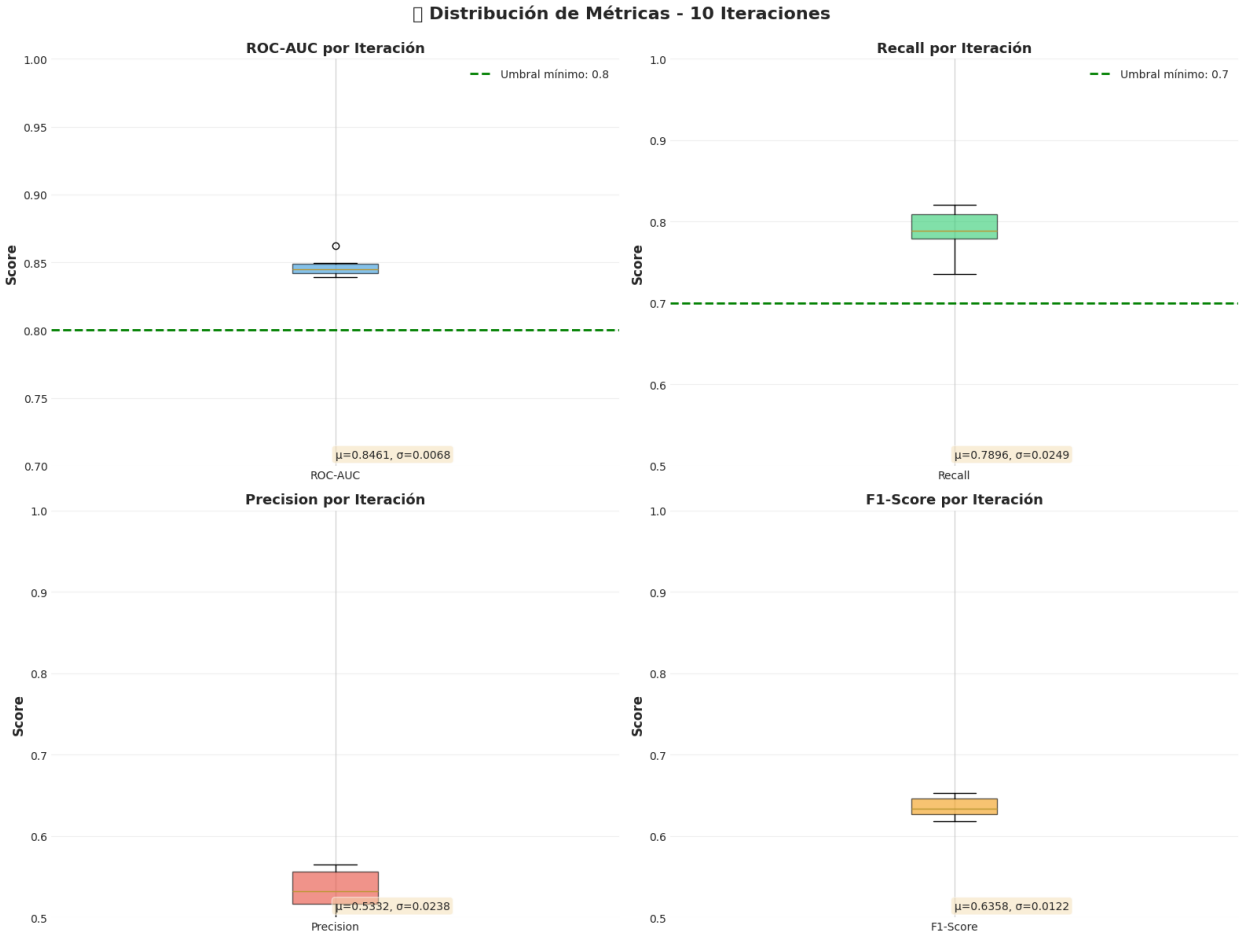


Figure 20: Visualización Integrada: Todas las Gráficas del Proyecto

- Gráficas de distribución revelan desbalance 73/27
- Histogramas y boxplots identifican factores de riesgo (tenure bajo, cargos altos)
- Matriz de correlación confirma relaciones esperadas

Acto 2 - El Desafío (Modelos Baseline):

- Gráficas de comparación muestran que modelos sin balanceo tienen bajo Recall (~45-55%)
- ROC-AUC razonable (~0.82) pero insuficiente para detectar Churn

Acto 3 - La Solución (Balanceo):

- Gráfica comparativa de técnicas identifica Undersampling como óptima
- Gráficas antes/después demuestran mejora significativa en Recall (+30%)
- Curvas ROC confirman que el balanceo funciona en todos los modelos

Acto 4 - La Validación (Evaluación):

- Matriz de confusión muestra balance adecuado entre TP y FP
- Curvas ROC y PR confirman rendimiento robusto (AUC 0.85)
- Feature Importance identifica palancas de acción

Acto 5 - La Confianza (Robustez):

- Validación cruzada demuestra estabilidad (bandas estrechas)
- Validación con múltiples semillas confirma reproducibilidad
- Boxplots muestran métricas consistentes

Conclusión visual: El proyecto pasa de un problema (desbalance) a una solución validada (modelo robusto con ROC-AUC 0.85) mediante un proceso sistemático y transparente, donde cada visualización aporta evidencia que sustenta las decisiones tomadas.

Analogía: Es como un documental que te lleva desde el descubrimiento de un problema (escena 1), pasando por los intentos fallidos (escena 2), la solución innovadora (escena 3), las pruebas rigurosas (escena 4), hasta el éxito final (escena 5). Cada escena (visualización) es necesaria para entender la historia completa.

Mini-glosario:

- **Narrativa de datos**: Contar una historia coherente usando visualizaciones
- **Evidencia visual**: Gráficas que sustentan decisiones y conclusiones
- **Proceso sistemático**: Metodología estructurada y reproducible

---

GLOSARIO GENERAL DE VISUALIZACIONES

Término	Definición
Gráfico de barras	Visualización que compara categorías usando barras horizontales o verticales
Gráfico de pastel	Visualización circular que muestra proporciones de un todo

Término	Definición
Histograma	Gráfico que muestra la distribución de frecuencias de una variable continua
Boxplot	Gráfico que muestra mediana, cuartiles, rango y outliers
Heatmap	Mapa de calor que usa colores para representar valores en una matriz
Scatter plot	Gráfico de dispersión que muestra relación entre dos variables
Curva ROC	Gráfico de TPR vs FPR en diferentes umbrales
Curva PR	Gráfico de Precision vs Recall en diferentes umbrales
Matriz de confusión	Tabla 2×2 que muestra TP, TN, FP, FN
Gráfico de línea	Visualización que muestra tendencias o evolución temporal
Barras agrupadas	Gráfico que compara múltiples categorías en diferentes grupos
Banda de confianza	Área sombreada que representa incertidumbre o variabilidad

---

RESUMEN VISUAL DE TODAS LAS GRÁFICAS

Gráficas de Análisis Exploratorio (EDA)

1. **Distribución de Churn** - Barras y pastel mostrando desbalance 73/27
2. **Churn por Variables Categóricas** - 6 gráficas de factores de riesgo
3. **Distribuciones Numéricas** - Histogramas y boxplots de tenure, charges
4. **Matriz de Correlación** - Heatmap de relaciones entre variables

Gráficas de Comparación de Modelos

5. **Métricas Baseline** - 4 gráficas comparando 7 algoritmos
6. **ROC-AUC Baseline** - Ranking de modelos por capacidad discriminativa

Gráficas de Técnicas de Balanceo

7. **Comparativa de Balanceo** - 4 subgráficos evaluando 3 técnicas
8. **Antes vs Después** - Impacto del balanceo en 4 métricas
9. **Curvas ROC Comparativas** - Superposición de 4 modelos balanceados

Gráficas de Evaluación Final

10. **Evaluación Completa** - Matriz confusión + ROC + Precision-Recall
11. **Feature Importance** - Top 20 características más influyentes
12. **Validación Cruzada** - Estabilidad en 5 folds
13. **Validación de Robustez** - Consistencia con 5 semillas

Gráficas Adicionales (Multi-iteración)

14-15. **Análisis de Múltiples Iteraciones** - Boxplots y tendencias

---

RECOMENDACIONES PARA IMPLEMENTACIÓN

Herramientas Recomendadas

Opción 1: Quarto (RECOMENDADO)

Ventajas:

- Integración nativa con Python, R y Julia
- Genera documentos HTML, PDF y presentaciones interactivas
- Soporte para código ejecutable y visualizaciones dinámicas
- Sintaxis Markdown familiar
- Ideal para reportes técnicos y presentaciones ejecutivas

Uso sugerido:

```
# Instalar Quarto
# Descargar de https://quarto.org/docs/get-started/

# Renderizar documento
quarto render preguntas-graficas-cliente-insight.md --to html
quarto render preguntas-graficas-cliente-insight.md --to pdf
quarto render preguntas-graficas-cliente-insight.md --to revealjs  # Presentación
```

Opción 2: Shiny (Para dashboards interactivos)

Ventajas:

- Dashboards interactivos con filtros y controles
- Actualización en tiempo real
- Ideal para exploración de datos por usuarios no técnicos
- Puede integrarse con Quarto

Uso sugerido: - Crear dashboard interactivo donde usuarios puedan:

- Filtrar por segmentos de clientes
- Explorar diferentes umbrales de predicción
- Simular escenarios de retención
- Visualizar ROI de diferentes estrategias

Recomendación Final

Usar ambas herramientas de forma complementaria:

1. **Quarto**: Para documentación estática, reportes y presentaciones
2. **Shiny**: Para dashboard interactivo de monitoreo en producción

---

Documento generado para la sustentación del proyecto de predicción de Customer Churn Bootcamp de IA - Cliente Insight Fecha: 2025-11-28