Ciencia de Datos

Trabajo final grupal: Telco Churn

GRUPO 13

Integrantes: Villegas Avalos, Alan - 1509240

Raunich, Sol - 1756564

Curso: I5521

Introducción y objetivos

Contexto

En el mundo empresarial actual, la retención de clientes es crucial para el éxito a largo plazo de cualquier compañía. El churn, o la tasa de abandono de clientes, puede tener un impacto significativo en los ingresos y la sostenibilidad del negocio. Este informe presenta un análisis detallado utilizando técnicas de Machine Learning para predecir si un cliente abandonará la compañía.

Objetivo

El objetivo principal de este análisis es desarrollar un modelo que permita predecir la probabilidad de churn entre los clientes, utilizando diferentes algoritmos de Machine Learning y evaluando su rendimiento.

Descripción del dataset

El conjunto de datos utilizado en este análisis corresponde a una cartera de clientes de la empresa Telco NN, y contiene información sobre 7.043 clientes, con un total de 21 variables que describen diversas características relevantes. A continuación, se presenta un resumen de las variables incluidas en el dataset:

CustomerID: Un identificador único para cada cliente. Esta variable no es relevante para el análisis predictivo y se eliminará durante el preprocesamiento.

Gender: Género del cliente (masculino o femenino). Esta variable es categórica y puede influir en la decisión de abandono.

SeniorCitizen: Indica si el cliente es un ciudadano mayor (1) o no (0). Esta variable puede ser relevante para entender el comportamiento de los clientes mayores en relación con el churn.

Partner: Indica si el cliente tiene pareja (Yes/No). Esta variable puede influir en la estabilidad del cliente y su decisión de permanecer con la empresa.

Dependents: Indica si el cliente tiene dependientes (Yes/No). Similar a la variable de pareja, esta información puede afectar la permanencia del cliente.

Tenure: El número de meses que el cliente ha estado con la empresa. Esta variable numérica es crucial, ya que generalmente, los clientes que han estado más tiempo tienden a ser más leales.

PhoneService: Indica si el cliente tiene servicio telefónico (Yes/No). La disponibilidad de servicios puede influir en la decisión de continuar con la empresa.

MultipleLines: Indica si el cliente tiene múltiples líneas telefónicas (Yes/No). Esta variable puede estar relacionada con la satisfacción del cliente.

InternetService: Tipo de servicio de internet que tiene el cliente (DSL, Fiber optic, o No). Esta variable categórica puede ser un factor determinante en la satisfacción del cliente y su propensión a churn.

OnlineSecurity: Indica si el cliente tiene un servicio de seguridad en línea (Yes/No). Este servicio puede influir en la percepción del valor por parte del cliente.

OnlineBackup: Indica si el cliente tiene un servicio de respaldo en línea (Yes/No). Al igual que otros servicios, esto puede afectar la decisión del cliente.

DeviceProtection: Indica si el cliente tiene protección para dispositivos (Yes/No). Este servicio también puede influir en la satisfacción del cliente.

TechSupport: Indica si el cliente tiene acceso a soporte técnico (Yes/No). La calidad del soporte técnico es un factor importante para la retención del cliente.

StreamingTV: Indica si el cliente tiene acceso a servicios de streaming de televisión (Yes/No). Este servicio puede ser un factor atractivo para los clientes.

StreamingMovies: Indica si el cliente tiene acceso a servicios de streaming de películas (Yes/No). Similar a StreamingTV, esto puede afectar la lealtad del cliente.

Contract: Tipo de contrato del cliente (Month-to-month, One year, Two year). La duración del contrato puede influir significativamente en las decisiones de churn.

PaperlessBilling: Indica si el cliente opta por facturación sin papel (Yes/No). Esto puede estar relacionado con las preferencias personales y la comodidad del cliente.

PaymentMethod: Método de pago utilizado por el cliente (Electronic check, Mailed check, Bank transfer, Credit card). Las preferencias de pago pueden ofrecer información sobre la satisfacción y lealtad del cliente.

MonthlyCharges: Cargos mensuales que paga el cliente. Esta variable numérica es importante para entender la relación entre costos y churn.
TotalCharges: Cargos totales acumulados por el cliente desde su inicio con la empresa. Al igual que MonthlyCharges, esta variable es crucial para evaluar la relación entre costos y abandono.
Churn: Variable objetivo que indica si el cliente ha abandonado la empresa (1) o no (0). Esta es la variable que se busca predecir mediante técnicas de Machine Learning.

Análisis exploratorio de datos

Durante el análisis exploratorio de datos (EDA), se llevó a cabo un proceso exhaustivo de limpieza y transformación de los datos, con especial énfasis en la columna churn, que indica si un cliente ha abandonado la compañía. Este proceso fue fundamental para garantizar la calidad y la integridad de los datos antes de proceder con el modelado.
En primer lugar, se identificaron y gestionaron los valores nulos en el conjunto de datos. Se observó que la columna churn contenía algunos valores faltantes, lo que podría afectar negativamente el rendimiento del modelo. Para abordar este problema, se decidió eliminar las filas que contenían valores nulos en esta columna, ya que su porcentaje era relativamente bajo y su eliminación no comprometería la representatividad del conjunto de datos.
Además, se realizó una transformación de la variable churn para convertirla en una forma adecuada para el análisis. Originalmente, esta columna podría estar representada como categórica, utilizando etiquetas como "sí" o "no". Para facilitar el trabajo con algoritmos de Machine Learning, se transformó esta variable a una representación numérica: se asignó un valor de 1 para los clientes que habían abandonado la compañía y un valor de 0 para aquellos que habían permanecido. Esta codificación permite que los modelos interpreten correctamente la variable como un problema de clasificación binaria.
Una vez completadas estas transformaciones, se llevó a cabo un análisis descriptivo para explorar la distribución de la variable churn. Se generaron visualizaciones que mostraban la cantidad de clientes que habían abandonado frente a aquellos que permanecieron, lo que proporcionó una visión clara sobre el problema del churn en el conjunto de datos.

Materiales y métodos

En este análisis, se implementaron diversos algoritmos de Machine Learning con el objetivo de predecir la probabilidad de churn entre los clientes. Cada uno de estos métodos fue seleccionado por su capacidad para abordar problemas de clasificación y su efectividad en conjuntos de datos desbalanceados.
Uno de los algoritmos utilizados fue Random Forest, que es un método de ensamble que combina múltiples árboles de decisión para mejorar la precisión y reducir el riesgo de sobreajuste. Este enfoque permite capturar interacciones no lineales entre las variables y proporciona estimaciones robustas del churn, lo que lo convierte en una opción adecuada para este tipo de análisis.
Otro algoritmo implementado fue el Support Vector Machine (SVM), un clasificador potente que busca encontrar el hiperplano óptimo que separa las diferentes clases en el espacio de características. Aunque SVM puede ser menos efectivo en conjuntos de datos grandes, se

utilizó en este estudio para comparar su rendimiento con otros modelos, especialmente en términos de precisión y recall.
El Gradient Boosting también formó parte del análisis. Este algoritmo se basa en la técnica de boosting, donde se construyen secuencialmente modelos débiles (generalmente árboles de decisión) que corrigen los errores del modelo anterior. Gradient Boosting es conocido por su capacidad para lograr un alto rendimiento en tareas de clasificación y se destacó como uno de los modelos más efectivos en este análisis.
Además, se utilizó XGBoost, que es una implementación optimizada del algoritmo de boosting. Este modelo es ampliamente reconocido por su velocidad y rendimiento, siendo una elección popular en competiciones de Machine Learning. Su capacidad para manejar grandes volúmenes de datos y su eficacia en la clasificación binaria lo hicieron particularmente adecuado para este estudio.
Por último, se implementó un modelo de redes neuronales utilizando Keras, con varias capas densas y funciones de activación ReLU. Este enfoque permite capturar relaciones complejas entre las características del cliente y la probabilidad de churn. Aunque las redes neuronales pueden requerir más datos y ajustes, su flexibilidad las convierte en una opción interesante para este tipo de análisis.

Cada modelo fue evaluado utilizando métricas clave como la matriz de confusión, precisión, recall, F1-score y AUC-ROC. Estas métricas proporcionaron una visión integral del rendimiento del modelo al clasificar correctamente a los clientes que abandonan frente a aquellos que permanecen.

Experimentos y resultados

Los experimentos comenzaron con la preparación del conjunto de datos, donde se realizó una limpieza exhaustiva. Esto incluyó el manejo de valores nulos y la transformación de la columna churn para convertirla en una variable numérica adecuada para el análisis. Se eliminaron las filas con valores nulos en la columna churn y se codificó la variable como 0 para clientes que permanecen y 1 para aquellos que abandonan.
Una vez que los datos fueron preparados, se dividieron en conjuntos de entrenamiento y prueba. Cada modelo fue ajustado a los datos de entrenamiento y luego se realizaron predicciones sobre el conjunto de prueba. Para evaluar el rendimiento de los modelos, se utilizaron métricas como la matriz de confusión, precisión, recall, F1-score y AUC-ROC.
Los resultados mostraron que el modelo de Gradient Boosting tuvo el mejor rendimiento, con un AUC-ROC de aproximadamente 0.8561, lo que indica una excelente capacidad para discriminar entre clientes que abandonan y aquellos que permanecen. Random Forest también mostró un rendimiento sólido con un AUC-ROC de 0.8554. En contraste, las redes neuronales tuvieron un rendimiento inferior, con un AUC-ROC de 0.7827.
Además de las métricas estándar, se realizaron análisis adicionales para entender mejor cómo cada modelo se comportaba en relación con diferentes características del cliente. Esto incluyó la identificación de patrones en los datos que podrían indicar un mayor riesgo de churn.

Discusión y conclusiones

Modelo	AUC-ROC	Exactitud	Precisión Clase 1	Recall Clase 1	F1-Score Clase 1
Gradient Boosting	0.8561	0.81	0.65	0.6	0.62
Random Forest	0.8554	0.8	0.63	0.64	0.64
XGBoost	0.8273	0.78	0.61	0.51	0.56
Redes Neuronales	0.7827	0.74	0.52	0.45	0.48
SVM	0.7676	0.76	0.55	0.51	0.53

Los resultados obtenidos mostraron que el modelo de Gradient Boosting fue el más efectivo, con un AUC-ROC de aproximadamente 0.8561. Este resultado indica que el modelo tiene una excelente capacidad para diferenciar entre clientes que abandonan y aquellos que permanecen. La robustez del Gradient Boosting se puede atribuir a su enfoque en corregir los errores de modelos anteriores, lo que le permite capturar relaciones complejas en los datos.
Por otro lado, Random Forest también mostró un rendimiento sólido con un AUC-ROC de 0.8554, lo que refuerza la idea de que los métodos basados en árboles son particularmente eficaces para este tipo de problemas. Sin embargo, los resultados para XGBoost fueron ligeramente inferiores, con un AUC-ROC de 0.8273, aunque sigue siendo un modelo competitivo.
Las redes neuronales, a pesar de su potencial para modelar relaciones complejas, presentaron un rendimiento inferior con un AUC-ROC de 0.7827. Esto puede deberse a la necesidad de más datos o ajustes más finos en la arquitectura del modelo para lograr un rendimiento óptimo. Este hallazgo sugiere que, aunque las redes neuronales son herramientas poderosas, no siempre son la mejor opción para todos los conjuntos de datos o problemas.
La discusión también destacó la importancia del preprocesamiento adecuado de los datos. La transformación y limpieza de la columna churn fueron cruciales para garantizar que los modelos pudieran interpretar correctamente la variable objetivo. La eliminación de valores nulos y la codificación adecuada permitieron establecer una base sólida para el análisis.

Los resultados obtenidos no sólo proporcionan información valiosa sobre las tasas de churn, sino que también ofrecen una base para desarrollar estrategias efectivas de retención de clientes. Futuros trabajos podrían explorar enfoques adicionales, como el uso de técnicas avanzadas de ensamblado o la incorporación de variables adicionales que puedan influir en el churn, con el fin de mejorar aún más la precisión predictiva.

Referencias

Burkov, A. (2020). The hundred-page machine learning book. Andriy Burkov.

Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems (2nd ed.). O'Reilly Media.

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies. MIT Press.