

Análisis PREDICTIVO PROYECTO FINAL





Table of contents



01

Introducción

Contexto, objetivo y público objetivo.

02

Preguntas

Hipótesis y preguntas de investigación.

03

Metadata

Características del dataset utilizado.

04

Análisis exploratorio

Exploración y visualización de datos

05

Modelado

Selección, entrenamiento y optimización del modelo.

06

Conclusiones

Resultados y recomendaciones



Introduccion

+ Contextualización y objetivo del proyecto +





Sobre este proyecto

El problema abordado en este proyecto es la tasa de abandono de clientes en una empresa de telecomunicaciones, también conocida como churn rate. Se busca analizar los factores que influyen en esta tasa con el objetivo de implementar estrategias efectivas para retener a los clientes actuales y atraer a nuevos.





Objetivos

Objetivo 1

Identificar los factores que influyen en la tasa de abandono de los clientes de la empresa de telecomunicaciones.



Objetivo 2

Desarrollar un modelo predictivo que permita predecir la probabilidad de abandono de los clientes y así ayudar a la empresa a retenerlos.





Audiencia

Este proyecto está destinado a personas interesadas en aprender técnicas de análisis de datos y modelado predictivo. Además de que es útil para empresas interesadas en identificar los factores que influyen en la tasa de abandono de los clientes.ra





02

Preguntas

+ Inquietudes que se intentan resolver con el análisis. +



Preguntas a resolver

Pregunta 1:	¿Cuáles son los factores que influyen en la tasa de abandono de los clientes?
Pregunta 2:	¿Hay alguna relación entre la duración del contrato y la tasa de abandono?
Pregunta 3:	¿Cómo afecta el tipo de servicio contratado a la tasa de abandono?
Pregunta 4:	¿Qué tan precisos son los modelos de predicción para identificar a los clientes con mayor probabilidad de abandonar?

El análisis exploratorio de datos y el modelado predictivo con técnicas de aprendizaje automático se utilizarán para responder las preguntas planteadas. El modelo se optimizará y se seleccionará el mejor para presentar recomendaciones y reducir la tasa de abandono de los clientes.





Metadata

+ Características y composición de los datos +





Descripción de los Datos

Los datos utilizados en este proyecto provienen del conjunto de datos "Telco Customer Churn", disponible en Kaggle. Este conjunto de datos incluye información sobre clientes de una empresa de telecomunicaciones, como información de contratos (tipo de contrato, duración, cargos mensuales), servicios contratados (internet, teléfono, seguridad en línea, soporte técnico, entre otros). También incluye información sobre si el cliente abandonó la empresa (churn) y la cantidad de cargos realizados por el cliente.

El conjunto de datos tiene un tamaño de 7,042 observaciones y 21 variables, incluyendo la variable de salida (churn).





Descripción de las Variables

- CustomerID: ID del cliente
- Demográficos: género y seniorcitizen (mayor de edad o no)
- Relaciones: partner (si tiene socio) y dependents (si tiene dependientes)
- Duración y servicios: tenure (número de meses en la compañía), phoneService (si tiene servicio telefónico) y multipleLines (si tiene múltiples líneas telefónicas)
- Servicios de internet: internetService, onlineSecurity, onlineBackup, deviceProtection, techSupport, streamingTV y streamingMovies
- Contrato y facturación: contract (tipo de contrato) y paperlessBilling (si tiene facturación sin papel)
- Pago y cargos: paymentMethod (método de pago), monthlyCharges (monto mensual cargado) y totalCharges (monto total cargado)
- Churn: indica si el cliente abandonó la compañía o no



Análisis exploratorio →

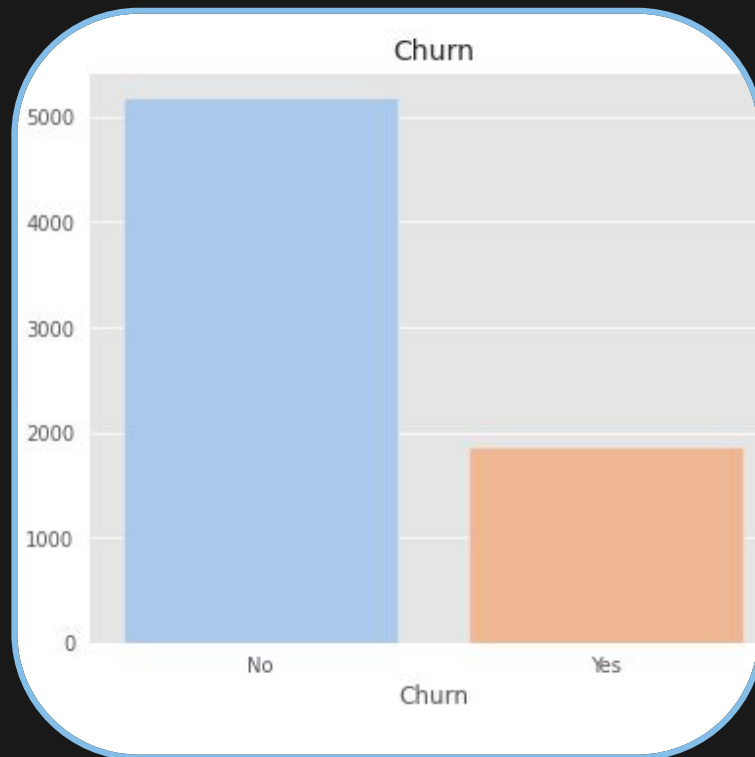
+ Exploración y comprensión de los datos +





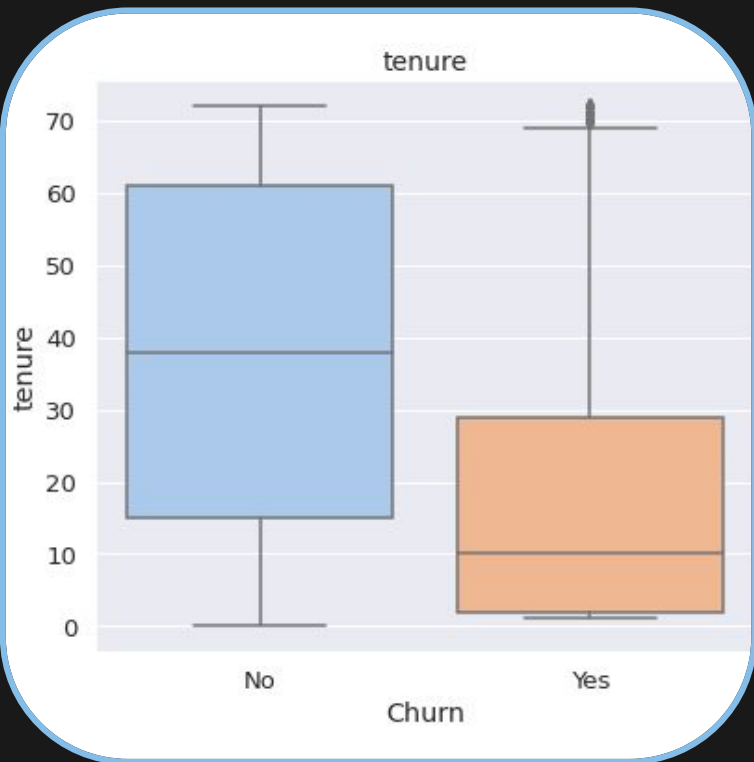
Distribución Tasa de abandono

Se observa una distribución desbalanceada, ya que la mayoría de los clientes no han abandonado la compañía.





Relación Churn y Tenure



Se observa que los clientes que abandonan la compañía tienen una mediana menor en la variable Tenure, lo que indica que son clientes más nuevos.





Relación contrato y churn

Se observa que los clientes que tienen un contrato mensual tienen una tasa de abandono mucho mayor que los clientes que tienen contratos anuales o de dos años.





Relación Churn y Internet



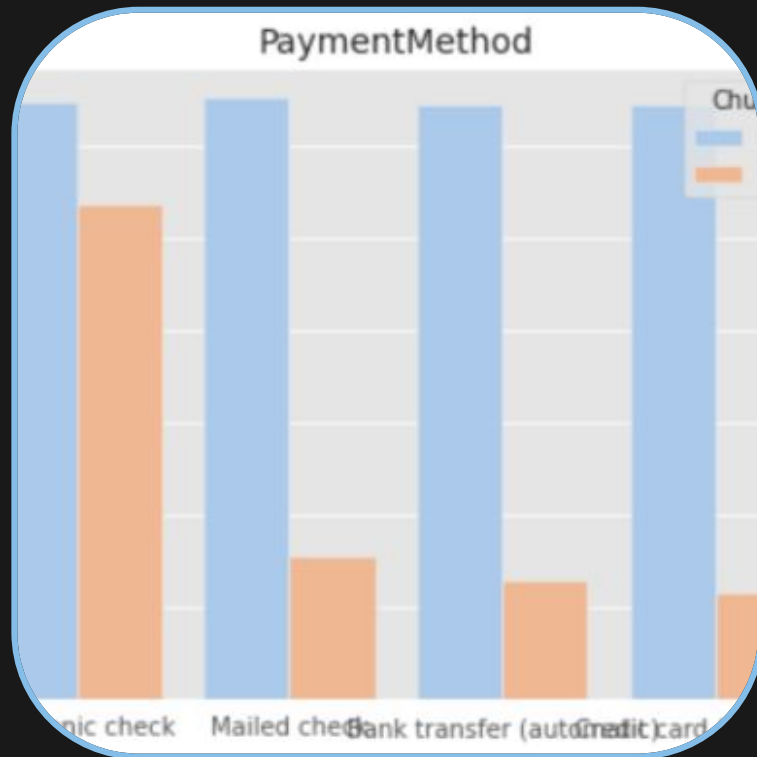
Los clientes que contratan fibra óptica son los que tienen mayor tasa de abandono, seguidos de los clientes que tienen internet DSL. Por otro lado, los clientes que tienen internet por marcado no tienen una tasa significativamente alta de abandono.





Relación tipo de pago y churn

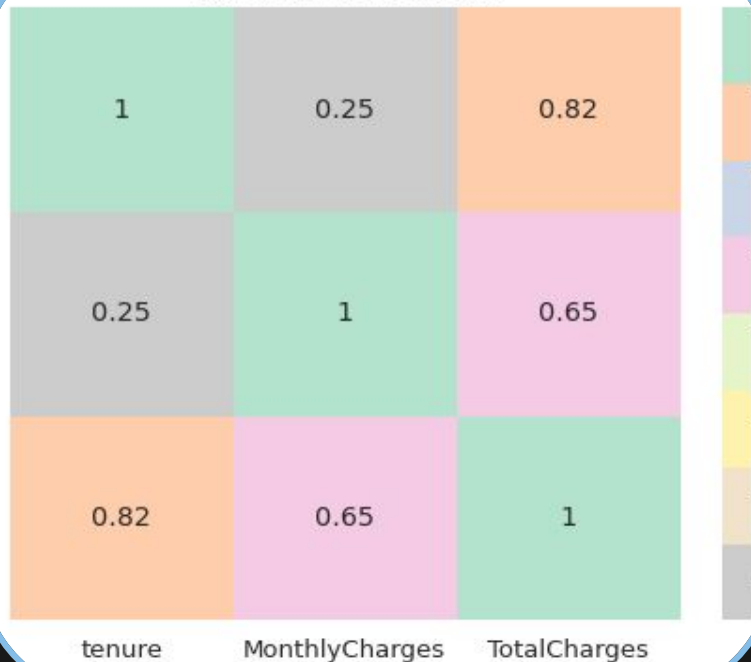
Se observa que los clientes que utilizan cheques electrónicos tienen una tasa de abandono mayor que los clientes que utilizan otros métodos de pago.





Correlación variables numéricas

Matriz de correlación



Se observa que la variable TotalCharges está altamente correlacionada con la variable MonthlyCharges y la variable Tenure.





Conclusiones

Las variables más relevantes para predecir la tasa de abandono de los clientes son la variable Contract, PaymentMethod, MonthlyCharges, InternetService, Tenure, y TotalCharges. Además, se observa una distribución desbalanceada de la variable Churn y una fuerte correlación entre las variables numéricas.





Modelado

+ Optimización y selección del mejor modelo +





El objetivo del modelado es crear un modelo predictivo que pueda predecir si un cliente abandonará o no la compañía de telecomunicaciones. Este modelo se entrenará utilizando técnicas de aprendizaje automático. El modelo final deberá tener una alta precisión para poder ser utilizado por la compañía para implementar estrategias efectivas de retención de clientes.





Algoritmos a utilizar

Regresion
Logistica



Arbol de
Decision



SVM



Red
Neuronal





Evaluación de Modelos

Para evaluar el rendimiento de los modelos predictivos, se utilizó la técnica de validación cruzada. Se calculó la precisión y la desviación estándar para presentar en cada modelo

Metricas

Regresion Logistica: 0.8021 (0.00743)

Arbol de Decision: 0.72631 (0.01051)

SVM: 0.78825 (0.011)

Red Neuronal: 0.7682 (0.01408)





Evaluación de hiperparámetros

Para mejorar aún más la precisión del modelo, se utilizaron técnicas de optimización de hiperparámetros para encontrar los valores óptimos de los parámetros del modelo.

```
Logistica:
parametros: {'Regresion Logistica__Cs': 10, 'Regresion Logistica__penalty': 'l2', 'Regresion Logistica__solver': 'lib
n del entrenamiento: 0.80334
n del test: 0.81192
-----

Decision:
parametros: {'Arbol de Decision__max_depth': 6, 'Arbol de Decision__min_samples_split': 2}
n del entrenamiento: 0.78577
n del test: 0.79915
-----

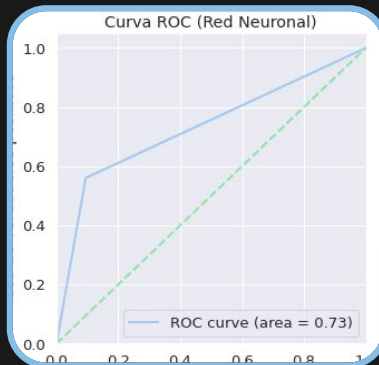
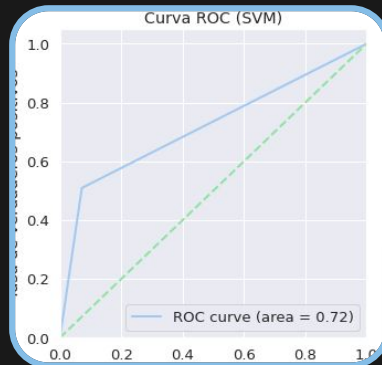
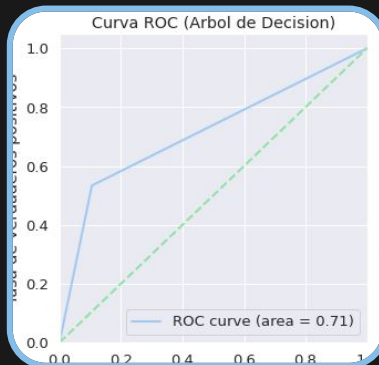
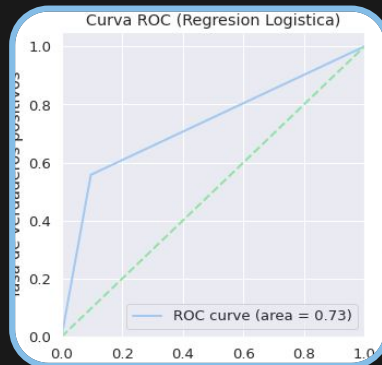
parametros: {'SVM__C': 10, 'SVM__kernel': 'linear'}
n del entrenamiento: 0.80334
n del test: 0.8176
-----

onal:
parametros: {'Red Neuronal__activation': 'logistic', 'Red Neuronal__alpha': 0.01, 'Red Neuronal__hidden_layer_sizes':
n del entrenamiento: 0.80422
del test: 0.81405
```





Curvas ROC





Conclusiones

+ Resultados y recomendaciones del proyecto +





Resultados

Se identificaron varios factores que influyen en la tasa de abandono de los clientes, como la duración del contrato, el cargo mensual y el tipo de servicio contratado.

Se encontró una relación entre el tipo de servicio y la tasa de abandono, siendo los clientes con servicios de fibra óptica los que tienen una tasa de abandono más alta en comparación con aquellos que solo tienen servicios de teléfono.

El modelo SVM tuvo el mejor desempeño en términos de precisión de prueba, seguido por Regresión Logística y Red Neuronal.





Recomendaciones

La compañía debería prestar especial atención a los clientes que tienen servicios de fibra óptica, ya que tienen una tasa de abandono más alta.

La compañía debería tomar medidas preventivas para mejorar la satisfacción del cliente y retenerlos, como ofrecer promociones, descuentos o servicios adicionales.

