UTN.BA
UNIVERSIDAD TECNOLÓGICA NACIONAL
FACULTAD REGIONAL BUENOS AIRES

DATOS DEL TRABAJO PRÁCTICO FINAL - CIENCIA DE DATOS			
Año Inicio: 19/08/24	**2024**	**Curso: I5571**	
Tema:	**Telco Churn**		
Docentes:	Martin Palazzo		
	Nicolas Aguirre		
	Santiago Chas		

DATOS DEL GRUPO		
Apellido	**Nombre**	**Legajo**
Pacheco	Joaquin	160.892-7

Introducción y objetivo

El siguiente trabajo práctico se desarrolla en el marco de la materia Ciencia de datos, de la carrera Ingeniería Industrial de la Universidad Tecnológica Nacional.
El objetivo del trabajo fue aplicar técnicas de EDA y Machine Learning para predecir qué clientes dejaran la compañía. Para ello se presenta un dataset que contiene información de los clientes de la empresa.

Descripción del dataset

El dataset se compone de una base de 7.943 personas con 21 variables que muestran algunas características de los clientes en la empresa.

Variables contenidas por el dataset:

Diccionario dataset del Telco churn	
Variable	Significado
Customer ID	Valor identificador de clientes
gender	Género del cliente
SeniorCitizien	Si el cliente es un SeniorCitizien o no
Partner	Si el cliente tiene un socio o no
Dependents	Si el cliente tiene dependientes o no
tenure	Antigüedad del cliente
PhoneService	Si el cliente tiene un servicio de telefono o no
MultipleLines	Si el cliente tiene multiples lineas o no
InternetService	Tipo de servicio de internet que recibe. Si es que recibe
OnlineSecurity	Si el cliente tiene un servicio de seguridad online o no
OnlineBackup	Si el cliente tiene un servicio de backup o no.
DeviceProtection	Si el cliente tiene un seguro del dispositivo o no
TechSupport	Si el cliente tiene soporte de tecnología o no.
StreamingTV	Si el cliente tiene servicio de streaming o no
StreamingMovies	Si el cliente tiene servicios de streaming de peliculas o no
Contract	Tipo de contrato del cliente

PaperlessBilling	Si el cliente recibe la factura en papel o no.
PaymentMethod	Tipo de pago del cliente
MonthlyCharges	Costo mensual
TotalCharges	Cargos totales
Churn	Si el cliente se fue de la compañía o no

Análisis Exploratorio de Datos

Para entender mejor el análisis realizado, he dividido el proceso en dos partes: Preprocesamiento y Análisis Exploratorio de Datos (EDA).

Preprocesamiento

Estructuración del Dataset:

Reemplace el valor "No phone service" por "No" para facilitar la comprensión del conjunto de datos. Luego impute el valor "Sin valor" a las variables categóricas que contenían datos faltantes.

Conversión de Variables:

Converti las columnas 'TotalCharges' y 'MonthlyCharges' a formato numérico, gestionando los valores faltantes para permitir operaciones posteriores con datos numéricos. Luego elimine los valores faltantes de las variables numéricas y verifique que no quedaran valores NaN que pudieran afectar el análisis posterior. Finalmente elimine las variables "Unnamed: 0" y "customerID" ya que son variables key del dataset y no aportan información relevante.

Análisis Exploratorio de Datos (EDA)

Para realizar el EDA cree dos copias del dataset para realizar diferentes limpiezas y evaluar performa mejor en el modelo de entrenamiento posterior.

Opción 1:
Luego de copiar el dataset converti las variables binarias a 1 y 0 según sus valores Yes o No. A su vez, transforme las variables no binarias en variables dummy para obtener valores numéricos que permitan entrenar un modelo.

Opción 2:
Comencé aplicando el mismo procedimiento anterior:

Luego de copiar el dataset converti las variables binarias a 1 y 0 según sus valores Yes o No.. A su vez, transforme las variables no binarias en variables dummy para obtener valores numéricos que permitan entrenar un modelo.

Sin embargo, luego de analizar la correlación de las variables con la variable objetivo 'Churn' elimine aquellas con una correlación menor a 0.05 (umbral elegido por mí).

Este enfoque me permitió comparar la efectividad de ambas estrategias de limpieza de datos en el modelo de entrenamiento.

Modelo de entrenamiento

Materiales y métodos (algoritmos utilizados)

En esta sección hice un análisis comparativo de diferentes modelos de clasificación sobre los dos dataset procesados con el objetivo de evaluar cual era el mejor dataset y el modelo que mejor performaba.

Preprocesamiento y Modelado:

Preprocesamiento: Los datos numéricos se estandarizan utilizando StandardScaler.
Reducción de Dimensionalidad: Aplicación de Análisis de Componentes Principales (PCA) para reducir las características a 10 componentes.

Modelos Utilizados:

Regresión Logística (LogisticRegression)
Random Forest (RandomForestClassifier)
Máquina de Soporte Vectorial (SVC)

Evaluación:
Luego de dividir en conjunto de datos en entrenamiento y prueba utilice validación cruzada con 5 pliegues para calcular el score de precisión. Luego seleccione el modelo con mejor

rendimiento basado en la puntuación media de validación cruzada.Finalmente, evalúe el modelo seleccionado en el conjunto de prueba y visualización de la matriz de confusión

Experimentos y resultados

Se observa que el mejor dataset es telco_reduced con un score de 0.80 utilizando el modelo Logistic Regression.

Discusión y conclusiones

El análisis realizado sobre ambas opciones de dataset me permitió evaluar la efectividad aplicando diferentes técnicas de limpieza de datos. Por otro lado, usando técnicas de preprocesamiento, como la estandarización y la reducción de dimensionalidad con PCA, junto con la validación cruzada, pude determinar cuál de los modelos supervisados (Regresión Logística, Random Forest y Máquina de Soporte Vectorial) obtiene el mejor rendimiento en términos de precisión.

Los resultados indican que la combinación de un preprocesamiento adecuado y la selección del modelo adecuado es crucial para maximizar la precisión predictiva

Referencias (aunque sea 3 papers y/o libros)

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springe*r.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.*

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). *Scikit-learn: Machine Learning in Python*