2025

Informe técnico: EDA y Predicción de la variable “Churn” en Telco NN

ASIGNATURA: CIENCIA DE DATOS – I5571

ALUMNOS: FERNANDEZ BIANCO MARÍA PAZ Y PAIKOVSKY NICOLÁS.

1- Introducción y objetivos

Este trabajo práctico tiene como finalidad ayudar a la empresa de telecomunicaciones “Telco NN”, la cual plantea que existe una fuga de clientes que impacta de forma directa en los ingresos de la compañía.

Nuestros objetivos principales fueron realizar un análisis exploratorio de los datos (EDA) para así comprender el comportamiento de los clientes; e implementar un Pipeline de Machine Learning que ayude a predecir la variable Churn, es decir si el cliente cancelará el servicio, para así poder tomar acciones preventivas sobre los perfiles que demuestran intención de hacerlo.

Además, hemos entrenado y evaluado múltiples algoritmos de clasificación (Logistic Regression, Random Forest); y aplicado técnicas de reducción de la dimensionalidad (PCA) para luego comparar el rendimiento.

2- Descripción del dataset

El dataset provisto denominado “telco_churn_clusterai” consta originalmente de 7043 registros y 21 variables. Representando cada fila a cada uno de los clientes, y las columnas a los atributos descriptivos:

- Variable demográfica: gender
- Servicios contratados: OnlineSecurity, PhoneService, MultipleLines, InternetService, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies
- Información de la cuenta: Contract, Customer ID, SeniorCitizien, Partner, Dependents, tenure, TotalCharges, PaperlessBilling, PaymentMethod, MonthlyCharges
- Variable para predecir: Churn

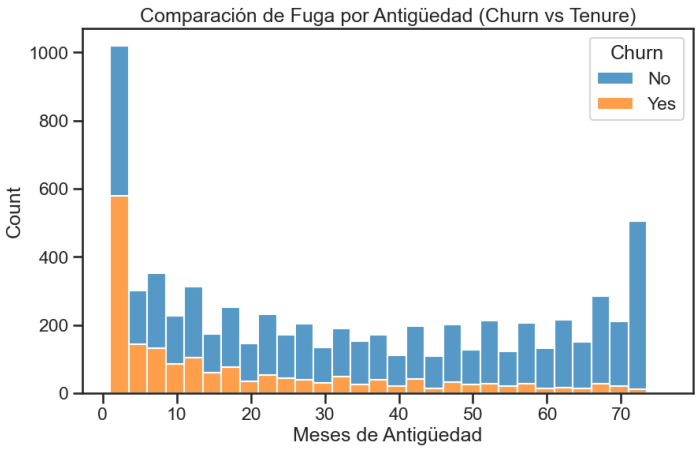
Durante la etapa de carga identificamos que la variable TotalCharges a pesar de ser numérica, contenía valores vacíos que eran interpretados como cadenas de texto. Por lo que realizamos una conversión forzada a numérico, detectándose 11 valores nulos que fueron eliminados del dataset, por lo que entonces para realizar el análisis contamos con un total de 7032 registros.

3- Análisis exploratorio de los datos (EDA)

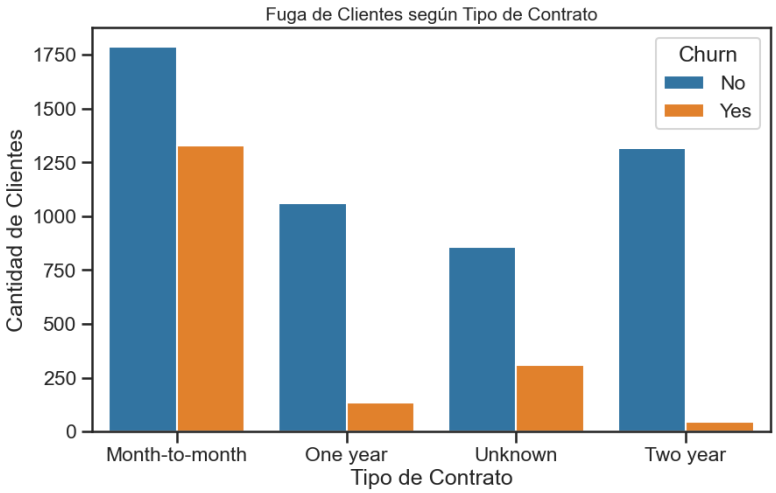
El análisis exploratorio nos permitió identificar patrones clave en el comportamiento de los clientes.

Correlaciones detectadas:

a) Antigüedad (Tenure): podemos observar que la mayor tasa de fuga ocurre en los primeros meses de vida del cliente. A medida que aumenta su antigüedad, la probabilidad de que Churn sea positivo disminuye.

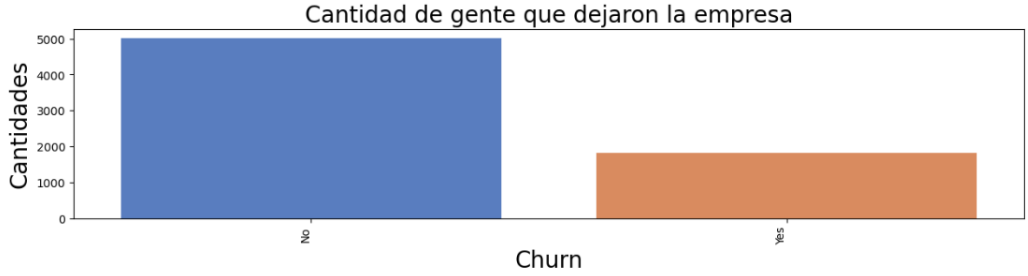


b) Contratos: Existe una fuerte relación entre el tipo de contrato del cliente y la fuga. Los clientes con contratos mensuales (“Month-to-month”) presentan una tasa de Churn positivo considerablemente superior a los que poseen contratos de uno o dos años.

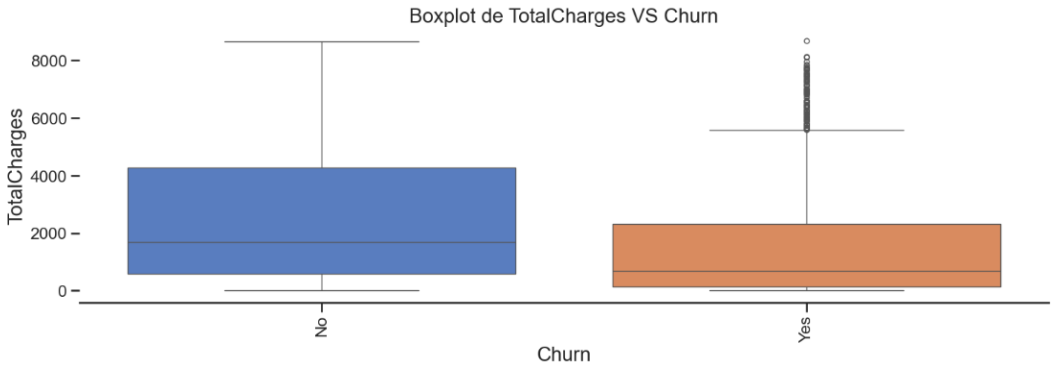


Churn	No (en %)	Yes (en %)
Contract		
Month-to-month	57.349166	42.650834
One year	88.638262	11.361738
Two year	96.769457	3.230543
Unknown	73.436161	26.563839

c) Observamos que la variable Churn nos informa que sobre los 6842 registros (no nulos) utilizados para el análisis, 5023 (73,41 %) corresponden a clientes que no abandonaron los servicios contratados, pero 1819 (26,59 %) corresponden a clientes que los han abandonado. Lo cual sugiere la fuerte necesidad de realizar mayores análisis y tomar medidas para evitar que futuros clientes abandonen la compañía.



d) Si bien se pueden observar presencia de valores outliers, los cargos totales de los clientes que abandonan la compañía suelen ser significativamente menores en comparación a los de los clientes que permanecen. Esto se debe a que, al abandonar de forma prematura, no llegan a acumular una suma considerable (en $) en su relación con la compañía; lo cual si sucede en clientes que no abandonan ya que mensualmente siguen acumulando pagos a lo largo del tiempo.

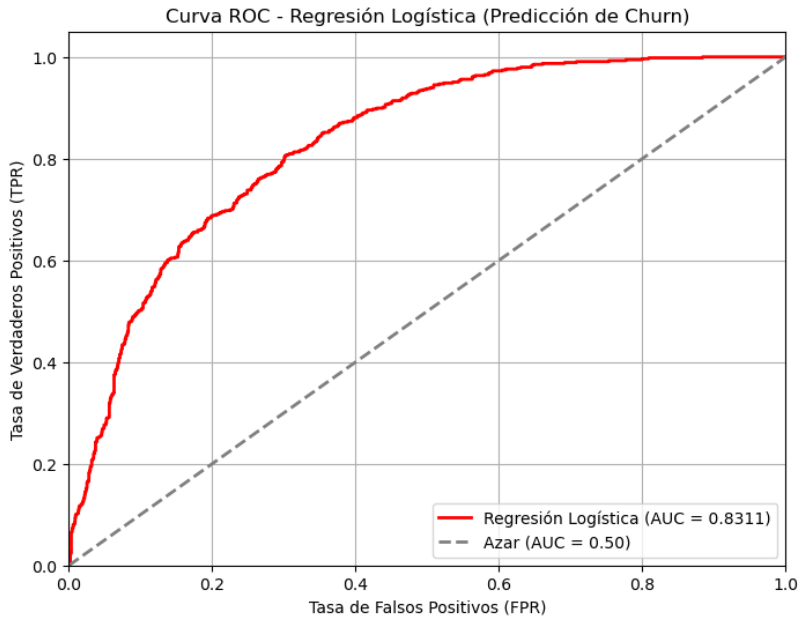


4- Materiales y Métodos (algoritmos utilizados)

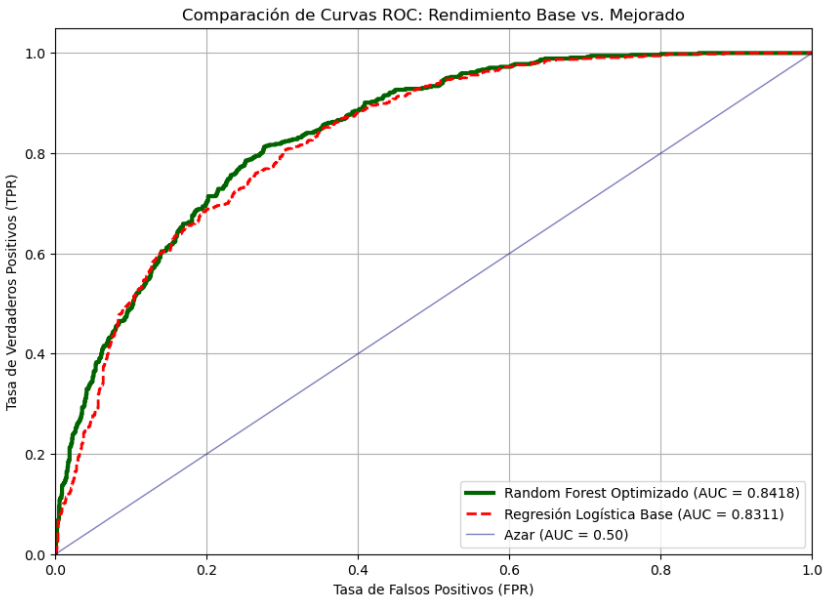
El proyecto fue desarrollado utilizando Python como lenguaje de programación, en el entorno de Jupyter Notebook. Las principales librerías que utilizamos fueron Pandas y Numpy (manipulación de los datos), Matplotlib y Seaborn (visualización) y Scikit-Learn (modelado).

Como modelo base utilizamos Logistic Regression, con él establecimos una línea de referencia del rendimiento, con AUC-ROC como métrica principal la cual nos dio una tasa de 0.8311 en test.

```
AUC (Regresión Logística): 0.8311
```



En busca de reducir el sobreajuste y mejorar el rendimiento obtenido anteriormente, aplicamos luego Random Forest.



El pipeline para predecir la variable Churn creado consistió en los siguientes pasos:

a) Aplicando One-Hot Encoding transformamos variables categóricas como por ejemplo Partner, PhoneService, InternetService, etc.
b) Dividimos el dataset en un conjunto de entrenamiento (70%) y de prueba (30%).

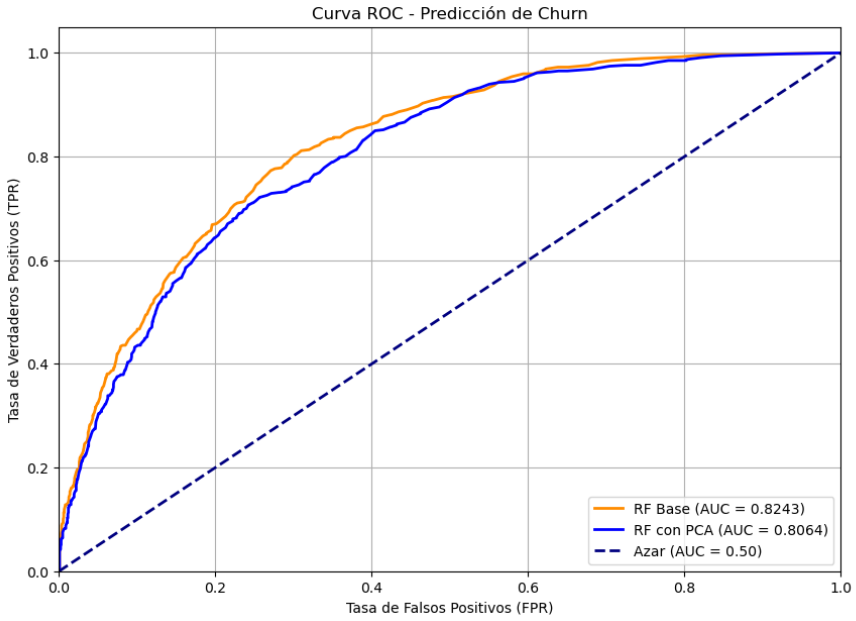
c) Escalamos las variables numéricas mediante el uso de StandardScaler, para normalizar las escalas.
d) Luego definimos y entrenamos los modelos, obteniéndose los resultados evidenciados aquí arriba.
e) Para el modelo de Random Forest, utilizamos la técnica de validación cruzada (3 cv) con GridSearchCV para así encontrar los mejores hiperparametros; Obteniéndose:

```
Mejores Hiperparámetros (RF):
{'max_depth': 10, 'min_samples_split': 10, 'n_estimators': 200}
AUC (Random Forest): 0.8418

Resultado Final: Random Forest (AUC=0.8418) > Regresión Logística (0.8311).
```

Para realizar una reducción de la dimensionalidad aplicamos PCA (con un RandomForestClassifier ya que como vimos se trata de un muy robusto modelo), reduciendo el número de características y evaluando si este modelo más compacto respecto a los anteriores puede mantener un buen rendimiento predictivo.

```
Comparación Final de Modelos (AUC-ROC):
1. Modelo Base (Sin PCA): 0.8243
2. Modelo con PCA (n=15): 0.8064
```



5- Experimentos y Resultados

Se entrenaron los modelos mencionados anteriormente y pudimos evaluarlos mediante las métricas de Accuracy y ROC-AUC.

El modelo de Regresión Logística demostró un desempeño estable, que permitió identificar patrones en los datos. Mientras que el Random Forest demostró ser el modelo superior, con mayor capacidad para capturar la complejidad del comportamiento del cliente.

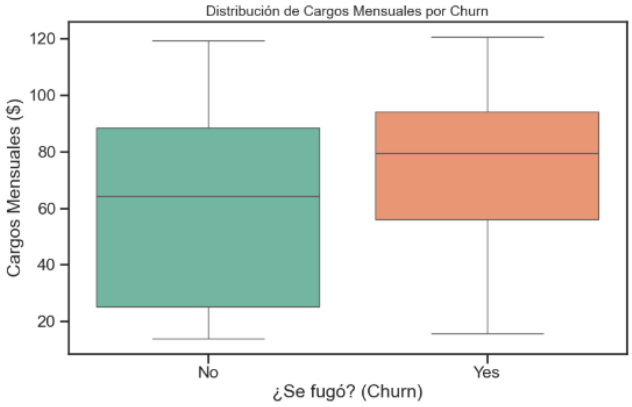
Respecto al experimento con reducción de la dimensionalidad (PCA), se observó que, si bien la complejidad se redujo, ello afectó al rendimiento del modelo. La Accuracy disminuyó levemente en comparación, lo cual sugiere que cada feature del dataset aporta valor significativo para la predicción de la fuga de clientes.

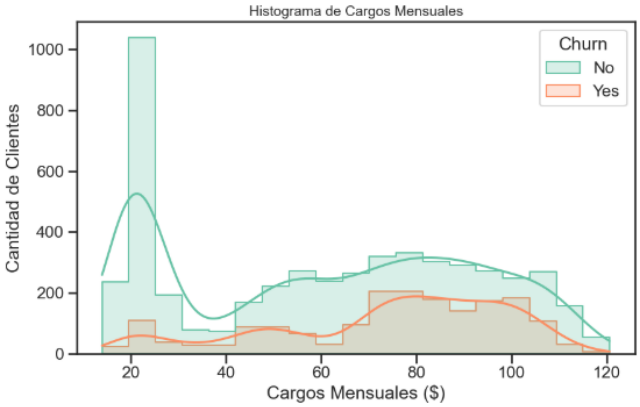
6- Discusión y Conclusiones

El análisis realizado nos confirma que es posible predecir la fuga de clientes con un nivel de confianza más que aceptable, superando el 78% de Accuracy en los mejores escenarios.

Los hallazgos principales fueron:

a) La influencia del contrato. El tipo de contrato es uno de los factores más determinantes, los clientes con contratos mensuales (“Month-to-month”) presentan una tasa de fuga drásticamente superior (42,65%) en comparación con quienes tienen contratos de uno (11,36%) o dos años (3,23%). La falta de compromiso a largo plazo facilita la migración hacia la competencia.
b) Antigüedad. La tasa de abandono se concentra en los clientes con baja antigüedad (“tenure”), la probabilidad de fidelización aumenta considerablemente a medida que transcurre el tiempo.
c) Sensibilidad al precio mensual. Los cargos mensuales elevados correlacionan positivamente con la fuga de clientes, quienes pagan un mayor monto son más exigentes y propensos a irse si no perciben valor.





7- Referencias

1. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. New York: Springer. Capítulo 8: Tree-Based Methods (Secciones 8.1 y 8.2)
2. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. New York: Springer. Capítulo 4: Classification (Sección 4.3: Logistic Regression).
3. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. New York: Springer. Capítulo 10: Unsupervised Learning (Sección 12.2: Principal Components Analysis).