Universidad Tecnológica Nacional
Facultad Regional Buenos Aires
Ciencia de Datos

Trabajo Práctico Final:
'ClusterAI Reporte - Análisis de Subscripciones Bancarias'.

Docentes: Mg. Martin Palazo
Ing. Nicolás Aguirre
Jefe de TP: Ing. Santiago Chas
Año: 2024
Curso: I5521

Grupo n° 2:

- Facundo Pérez - Legajo 168410-3
- Facundo Castro De Rosa - Legajo 161377-7

Introducción y Objetivos

En el ámbito bancario, entender el comportamiento de los clientes es clave para mejorar la eficiencia de las campañas de marketing. Este proyecto busca predecir qué clientes tienen más probabilidades de suscribirse a una campaña, utilizando un enfoque basado en datos. La predicción precisa permitirá al banco optimizar sus esfuerzos y recursos, dirigiendo las campañas a un público más segmentado y con mayores probabilidades de éxito.
El objetivo general es diseñar e implementar un pipeline de machine learning que analice el dataset, aplique técnicas avanzadas de análisis exploratorio y utilice métodos de reducción de dimensionalidad para mejorar la eficiencia del modelo. Este trabajo incluirá la evaluación de diferentes algoritmos y la comparación de su rendimiento.

Descripción del Dataset

El dataset consta de 45,211 registros y 17 variables que describen a los clientes del banco. Las variables incluyen características demográficas (edad, estado civil), información socioeconómica (nivel educativo, ocupación), y datos específicos de interacciones pasadas con el banco (tipo de contacto, duración de llamadas, número de contactos realizados durante la campaña, fecha de contacto, Performance de la campaña de marketing anterior para este cliente).
La variable objetivo es 'subscription', que toma el valor de 1 si el cliente se suscribió a la campaña y 0 en caso contrario.

```
#   Column                Non-Null Count  Dtype
---  ------                --------------  -----
0   Age                   40238 non-null  float64
1   Job                   40238 non-null  object
2   Marital Status        40238 non-null  object
3   Education             40238 non-null  object
4   Credit                40238 non-null  object
5   Balance (euros)       40238 non-null  float64
6   Housing Loan          37525 non-null  object
7   Personal Loan         37525 non-null  object
8   Contact               45211 non-null  object
9   Last Contact Day      45211 non-null  int64
10  Last Contact Month    45211 non-null  object
11  Last Contact Duration 37525 non-null  float64
12  Campaign              45211 non-null  int64
13  Pdays                 37525 non-null  float64
14  Previous              45211 non-null  int64
15  Poutcome              45211 non-null  object
16  Subscription          45211 non-null  int64
```

1. *Variables que conforman el dataset. Se puede observar la cantidad de nulos en cada Feature.*

Análisis Exploratorio de Datos

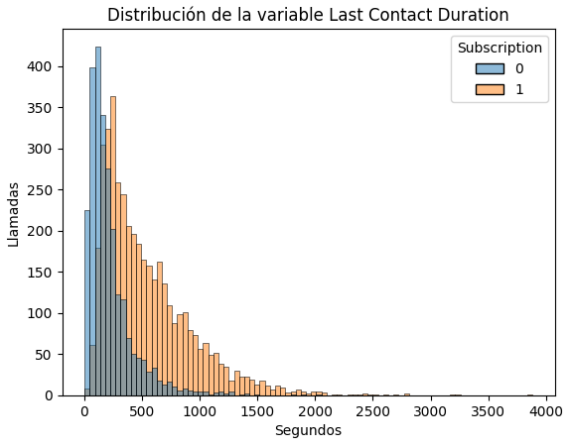
Esta etapa es crucial para comprender la información con la cual disponemos para explicar el problema, para ello nos valemos de herramientas como la estadística descriptiva, herramientas de visualización de datos, y el análisis de variables categóricas.
Primero identificamos la cantidad de valores nulos en el dataset, el cual era elevado. Resulta que los valores nulos son muchos en relación al tamaño del dataset, y no pertenecen a los mismos registros entre sí, por lo cual se buscó recuperar registros para el entrenamiento.

Se identificaron los distintos valores únicos de cada variable categórica:

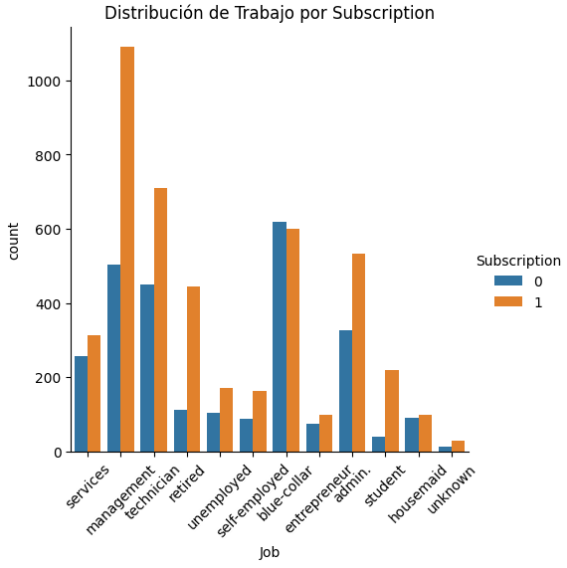
- Job: management, blue-collar, technician, admin., services, retired, unemployed, student, self-employed, housemaid, entrepreneur, unknown.
- Marital Status: married, single, divorced.
- Education: secondary, tertiary, primary, unknown.
- Credit: yes, no.
- Housing Loan: yes, no.

- Personal Loan: yes, no.
- Contact: Cellular, unknown, telephone.
- Poutcome: unknown, success, failure, other.

Se analizó por medio de histogramas las variables tipo float, segmentando los datos según la variable de salida subscription, y las categóricas por medio de gráficos y conteo de valores.
Se observa por ejemplo, en determinados trabajos cierta tendencia a suscribirse, así como en la duración de las llamadas algún grado de correlación.



2. *Histograma de la variable Last Contact Duration segmentado por Subscription.*



3. *Gráfico de categorías de trabajo por Salida de la variable Subscription.*

Con el objetivo de no perder muestras, se recurrió a asignar valores a los nulos:
Asignamos el promedio de edad en los valores nulos de la variable Age. A los valores nulos en Job,Education, Marital Status asignamos la categoría "Unknown". A los valores nulos en Credit asignamos "no". A los nulos en Balance (euros) le asignamos cero. A los nulos en Housing Loan le asignamos "yes". A los nulos en Personal Loan, podemos asignar "no". A los nulos en Last Contact Duration asignamos 0. A los nulos en Pdays asignamos -1 (se considera que no hubo contacto).
Continuamos reemplazando las variables categóricas yes no por 1 y 0 respectivamente, y generamos variables categóricas (dummies) para:

- Job, Marital status, Education, Contact, Poutcome.

Por último, eliminamos las variables Last Contact Day y Last Contact Month que no contienen información relevante a ser utilizada en el entrenamiento.

Materiales y Métodos

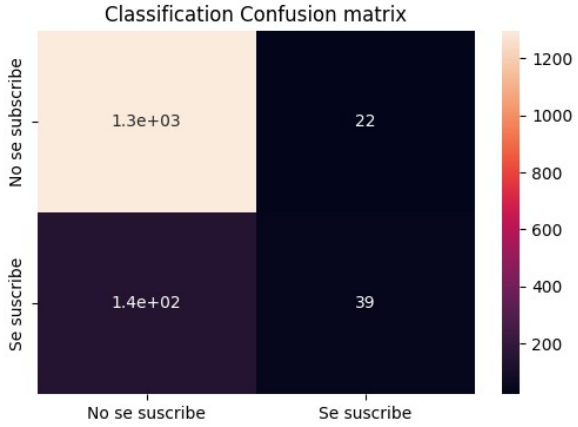
El pipeline se implementó en Python utilizando Pandas y Numpy para la manipulación y análisis de datos; Matplotlib y Seaborn para la generación de visualizaciones, Scikit-learn para la implementación de modelos de machine learning y la reducción de dimensionalidad, y Tensor Flow para la implementación de modelos de machine learning con redes neuronales.

Experimentos y Resultados

Se realizaron múltiples pruebas dividiendo el dataset en un conjunto de entrenamiento y uno de prueba (80/20), utilizando validación cruzada.

Las principales pruebas fueron utilizando LR, SVC y Neural networks.

En la evaluación de resultados, lo que parecían ser buenos valores de Accuracy cercanos a 0.90, terminó siendo una mala performance del modelo, ya que la muestra tomada para train y test poseían una gran cantidad de valores Negativos (Subscription = 0), por lo cual nos arrojaba una clasificación elevada de True Negative y de False Negative, sin cumplir el objetivo de identificar quienes se suscribieron a la campaña de marketing.

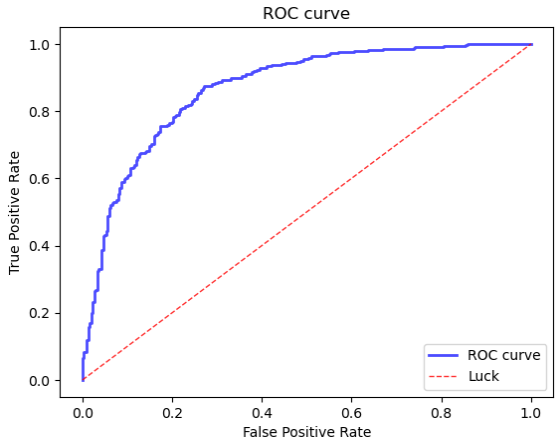


4. *Matriz de confusión de resultados con alto Accuracy pero con elevado False Negative.*

Este desbalance de muestras se fue corrigiendo probando distintas relaciones entre cantidad de subscription = 0 y subscription = 1, buscando mejorar el aprendizaje, buscando observar la mejora de resultados en la matriz de confusión.

Luego de exhaustivas pruebas, los mejores resultados obtenidos fueron con Logistic Regression, tomando una mayor cantidad de Subscription = 1 respecto a Subscription = 0 (62,5 % y 37,5% respectivamente):

- Accuracy score: 0,815

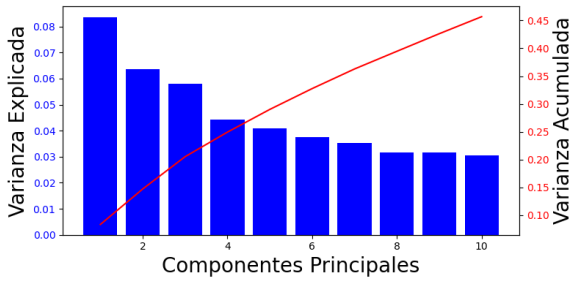


5. *Área bajo la curva ROC (LR). AUC = 0.86*

6. *Matriz de confusión con mejor resultado obtenido (LR)*

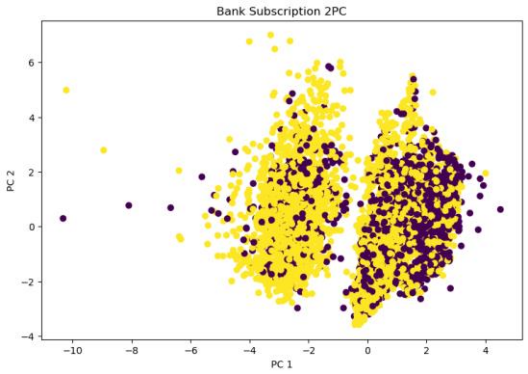
Resultados de PCA:

Se aplicó el Principal Component Analysis (PCA), buscando reducir el número de variables del dataset, eliminando redundancias y mejorando la eficiencia computacional.



7. *Varianza explicada y porcentaje por cada componente principal (PC).*

En el pca vemos que no hay una gran concentración de información en las primeras componentes (PC), sino que se distribuye en muchas componentes.



8. *Scatter plot de las 2 PC.*

Luego de entrenar usando la proyección vemos que los resultados no se ven mejorados al concentrarnos en pocas componentes principales, si no que empeoran, por lo cual se necesita más información para mejorar el entrenamiento.

El mejor resultado en este caso se obtuvo con SVC: score: 0.71

9. *Matriz de confusión de SVC luego de reducir a las primeras 10 PC.*

Discusión y Conclusiones

El análisis realizado destacó la importancia de un adecuado tratamiento de los datos, como la imputación de valores nulos y la generación de variables categóricas, para garantizar la calidad y estabilidad de los modelos. Sin embargo, el hecho de que para recuperar datos hacemos reemplazo de valores nulos con valores predefinidos puede ser la causa de que nunca llegamos a un modelo más confiable, lo que sugiere la necesidad de explorar métodos alternativos de imputación en futuros estudios.

Además, se abordó el desbalance en la variable objetivo 'Subscription', cuya redistribución (62,5 % para suscripciones positivas y 37,5 % para negativas) mejoró significativamente la sensibilidad y redujo los falsos negativos. Dado el objetivo inicial de detectar clientes que se suscribirían, nos resulta más importante mantener una tasa baja de falsos negativos que de falsos positivos, priorizando la identificación correcta de clientes potenciales.

Entre los modelos evaluados, Logistic Regression fue el más efectivo, logrando un accuracy de 0,815 y un AUC de 0,86, destacándose por su equilibrio entre simplicidad y rendimiento. Si bien el uso de SVC y PCA mostró cierto potencial, los datos no concentraban suficiente información en pocas componentes principales, limitando su efectividad.

Referencias

- VanderPlas, J. (n.d.). *Python Data Science Handbook*. Recuperado de https://jakevdp.github.io/PythonDataScienceHandbook/
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. Recuperado de https://www.statlearning.com/
- Murphy, K. P. (2022). *Probabilistic Machine Learning: An Introduction*. Recuperado de https://probml.github.io/pml-book/book1.html
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Recuperado de https://www.deeplearningbook.org/
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). New York: Springer.
- Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.