**Enlace Repositorio:** <https://github.com/luna-as18/Proyecto-2-Analitica>

1. Análisis de negocio + Tablero de datos. (Laura Hurtado)

2. Ingeniería de datos + Ciencia de datos (María Camila Luna)

3. Análisis de datos + Despliegue (Gabriela Correa)

**Prototipo de Analítica de Datos para la Predicción de Compra de Certificados de Depósito a Término (CDT)**

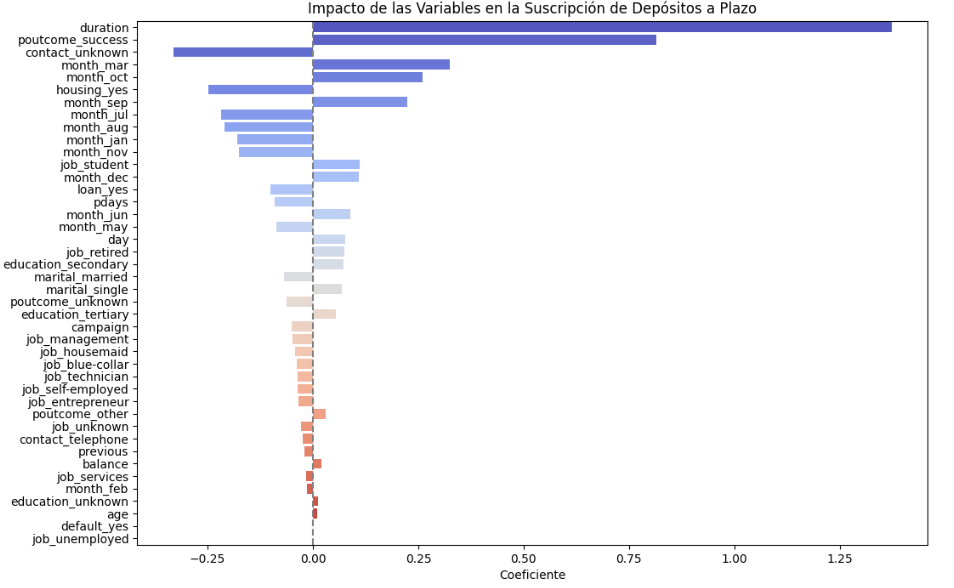
Este proyecto tiene como objetivo el desarrollo de un prototipo funcional de analítica de datos para la predicción de compra de Certificados de Depósito a Término (CDT) en una entidad bancaria. El prototipo está diseñado para proporcionar información clave al área comercial, facilitando la identificación de clientes con alta probabilidad de adquirir un CDT. La herramienta estará orientada a maximizar las ventas de productos financieros mediante recomendaciones basadas en datos, apoyando la toma de decisiones informada en las estrategias comerciales.

**Preguntas de negocio:** A partir de este enfoque, se pretende estudiar las siguientes preguntas de negocio:

1. **¿Cuáles son los factores más influyentes en la suscripción de depósitos a plazo?**

El análisis del modelo de clasificación ha identificado las siguientes variables como las más influyentes en la decisión de suscripción a depósitos a plazo:

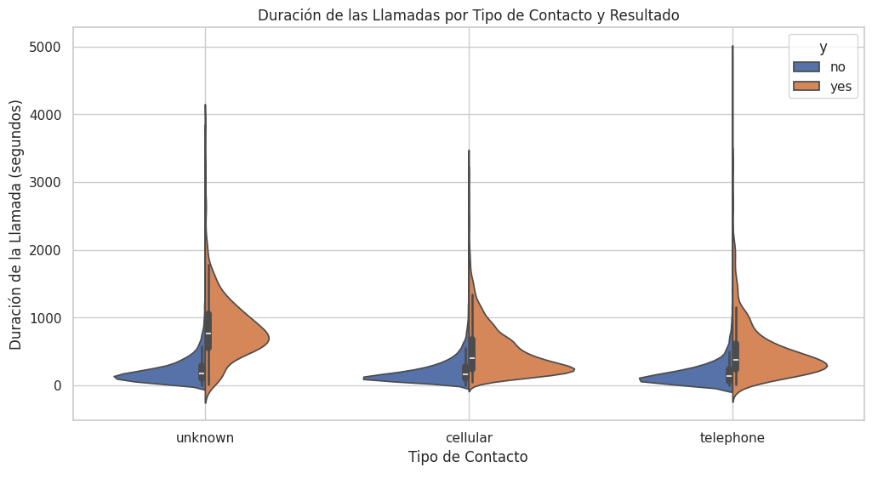
* **Duración de la llamada**: Coeficiente de **1.37**. Lo que quiere decir que a mayor duración, mayor probabilidad de suscripción.
* **Éxito de la campaña anterior (poutcome\_success)**: Coeficiente de  **0.82**. Lo que quiere decir que un resultado exitoso en campañas previas incrementa la probabilidad de suscripción.
* **Mes de contacto**: Los meses de **marzo** (0.33) y **octubre** (0.26) son los más favorables, mientras que **julio** (-0.22) y **noviembre** (-0.18) muestran un impacto negativo.

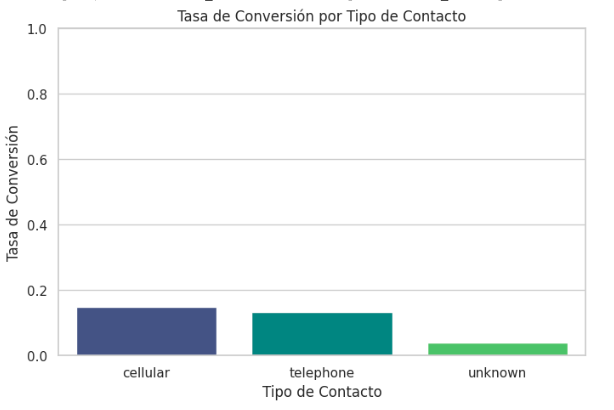


Estos hallazgos sugieren que optimizar la duración de las llamadas y enfocarse en clientes con un historial positivo en campañas anteriores podría aumentar las tasas de suscripción.

1. **¿Cómo varía la duración de las llamadas exitosas según el tipo de contacto?**

* **Duración y Probabilidad de Éxito:** La duración de la llamada es un factor determinante en el éxito, ya que las llamadas que resultan en una respuesta afirmativa tienden a durar más que las negativas. Esto sugiere que mantener la atención del cliente durante más tiempo podría aumentar la probabilidad de éxito.
* **Eficiencia por Tipo de Contacto:** Los tres tipos de contacto presentan patrones similares, con una tendencia a que las llamadas a teléfonos móviles sean más efectivas en términos de duración y éxito, lo cual podría implicar un enfoque estratégico en este canal.
* **Variabilidad en Duraciones:** La categoría unknown muestra mayor variabilidad en la duración, lo cual podría indicar una falta de consistencia en este tipo de contacto. Optimizar los protocolos de estas llamadas podría ser una oportunidad de mejora en la campaña.

****

****

Estas observaciones permiten orientar la estrategia de contacto y mejorar las probabilidades de éxito en futuras campañas.

**Limpieza de Datos:**

El proceso que se realizó para la limpieza de datos fue el siguiente:

1. Se valido que no hubieran duplicados, para el cual no se encontraron duplicados.
2. Se valido que el tipo de dato de cada columna hubiera cargado correctamente.
3. Se valido que no hubiera nulos, a lo cual no se hallaron nulos.
4. Se uso el comando. unique() para validar que los datos registrados fueran acorde al diccionario. Para la cual se obtuvo un resultado satisfactorio.
5. Se hizo un análisis para ver que columnas tenían mas del 90% con el mismo valor. Y se encontró que “Default” tenia el 98.2% en “NO”. Se considero eliminar esta columna, ya que, al tener tanta unicidad podría no aportarle mucho al modelo. Sin embargo, se vio que esta variable tenía una correlación de 0.25 con la variable de interés. Por lo que, se mantuvo y se dejó esa decisión para el modelo.
6. Se revisaron las correlaciones entre las variables de manera general y dieron correlaciones bajas en general:

Imagen que contiene Gráfico

Descripción generada automáticamente

1. Se buscaron valores anómalos con Boxplots, sin embargo, a pesar de que habían aparentes valores “anómalos”, bajo el contexto de un banco eran valores aceptables. Como, por ejemplo, un balance negativo.

Diagrama, Gráfico de cajas y bigotes

Descripción generada automáticamente

1. Se reviso que la edad mínima en los datos fueran personas mayores de edad, y que el máximo fuera una edad razonable. Y los datos se comportaban acorde.
2. Una vez hecho este análisis, se procedió a transformar las categorías a dummies, eliminando una de las categorías en las nuevas columnas generadas, por cada categoría. Y estos datos fueron los se usaron para entrenar el modelo, por lo cual el dataframe termino siendo de 43 columnas, incluida la variable de interés.

**Modelo de predictivo que indica si se adquirió el CDT o no:**

En este caso, se realizaron varios modelos para evaluar el que tuviera un mejor rendimiento, así fue el procedimiento:

1. Se realizo una regresión logística, con un test size del 20% de los datos. En general, los resultados sugieren que el modelo tiene un ajuste razonable, explicando alrededor del 33.74% de la variabilidad y mostrando una mejora significativa en comparación con un modelo sin predictores. Sin embargo, el Pseudo R-cuadrado de 0.3374 también sugiere que hay una cantidad considerable de variabilidad en los datos que el modelo no logra explica.
2. Por lo que se decidió hacer una búsqueda de hiper parámetros y además escalar los datos.

Texto

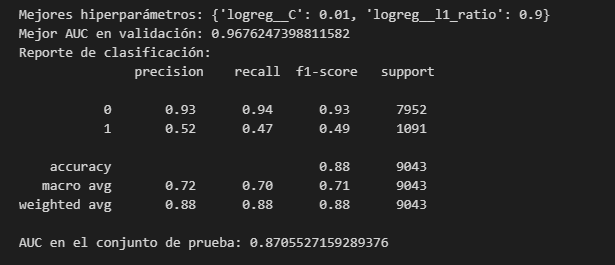
Descripción generada automáticamente

Texto

Descripción generada automáticamente

Se uso la penalización elastic net. La razón por la que se uso, fue basado en (Zou, 2005) quien estableció que Elastic Net combina las penalizaciones L1 y L2 de las regularizaciones Lasso y Ridge, respectivamente. Este método es especialmente útil cuando hay correlación entre las variables predictoras, ya que permite seleccionar grupos de variables correlacionadas, manteniendo al mismo tiempo el efecto de regularización para reducir la varianza del modelo.

Y para este modelo se obtuvieron las siguientes métricas:



Las métricas obtenidas muestran el rendimiento del modelo en términos de precisión, recall, y F1-score para cada clase. Para la clase "0" (clientes que no adquieren el producto), el modelo tiene una alta precisión (0.93), recall (0.94) y F1-score (0.93), lo cual indica que es muy efectivo para identificar correctamente los casos negativos. Sin embargo, para la clase "1" (clientes que adquieren el producto), las métricas son significativamente más bajas, con una precisión de 0.52, un recall de 0.47 y un F1-score de 0.49. Esta diferencia indica que el modelo tiene dificultades para identificar correctamente a los clientes que sí adquieren el producto, posiblemente debido a un desbalance en las clases (donde los casos de la clase "0" son mucho más comunes). Aunque el modelo tiene un AUC alto en validación (0.9676) y en el conjunto de prueba (0.8706), lo que sugiere una buena capacidad general para distinguir entre las clases, su rendimiento en la clase "1" es limitado, afectando la utilidad práctica para predecir casos positivos.

1. Por lo que se decidió hacer un modelo LDA preprocesado con PSL ya que, como menciono (S. Wold, 1978), PLS ayuda a reducir la dimensionalidad del conjunto de datos al extraer componentes que capturan la mayor parte de la varianza relevante, abordando el problema de la multicolinealidad al seleccionar solo las combinaciones de variables más informativas. Esto permite que LDA opere en un espacio de características optimizado, mejorando la separabilidad de las clases y haciendo que el modelo sea más interpretable y menos propenso al sobreajuste. En contraste, la regresión logística, aunque es un modelo robusto, puede tener dificultades en este tipo de escenarios, ya que la multicolinealidad entre las variables afecta la estabilidad de los coeficientes y, sin una reducción de dimensionalidad previa, el modelo podría no capturar de manera efectiva las relaciones relevantes en los datos.

En la siguiente imagen se muestra que se realizaron intervalos de confianza para ver la significancia de las variables, y ver si estas incluían el 0:

Texto

Descripción generada automáticamente

Sin embargo, algo que sucedía es que en las dummies, una de ellas no era significativa. Pero al ser parte de una categoría completa, para no complejizar el modelo estas no fueron eliminadas. Algo importante a denotar, es que se evaluo la significancia de las variables, y pesar de que “Age” y “Default” no parecían ser relevantes para el modelo, estas se removieron y se re-entreno el modelo, se obtuvieron las mismas métricas de rendimiento sin estas variables. Por lo que, se decidió incluirlas de igual forma, ya que, al presentar el mismo resultado es mejor que el modelo tenga la mayor cantidad de información posible.

Además de que se trató el desbalanceo de clases, ya que había un 88% aprox en clase tipo 1, y esto podía afectar el modelo se obtuvo lo siguiente:

Calendario

Descripción generada automáticamente

Aunque se obtuvieron pequeñas mejoras, se concluye que el puntaje medio de clase uno se debe al desbalanceo de clases pero que el rendimiento general del modelo es bastante bueno. Por lo que se procede a hacer el archivo keras, para serializar el modelo. Todo esto implementando en una **red neuronal** para usar keras.

**CONCLUSIONES**

* La duración de la llamada (duration) es un factor clave, ya que las llamadas más largas están fuertemente asociadas con una mayor probabilidad de suscripción al depósito. El balance del cliente, aunque presenta valores extremos, también resulta relevante, ya que aquellos con balances más altos parecen más propensos a suscribirse. La edad muestra ciertas diferencias en la distribución, sugiriendo que algunos rangos de edad podrían estar más dispuestos a suscribirse entre 20 y 40 años. La variable de días previos (pdays), que indica el tiempo desde el último contacto en una campaña anterior, influye en la probabilidad de suscripción, especialmente si el contacto fue reciente. Además, el número de contactos previos (previous) refleja una ligera tendencia de mayor suscripción entre aquellos clientes que han sido contactados varias veces en el pasado.
* Los clientes en ocupaciones de alta calificación, con educación terciaria o secundaria, sin historial de impago ni préstamos personales, y sin préstamos de vivienda son más propensos a suscribirse. Además, el contacto por celular es más efectivo, y la probabilidad de suscripción varía según el mes, siendo mayor en marzo y octubre. Finalmente, aquellos que han tenido éxito en campañas previas muestran una tendencia significativamente mayor a suscribirse, mientras que los que tuvieron fracasos previos o desconocen el resultado de campañas pasadas tienden a no suscribirse. No obstante, los resultados de campañas previas en su mayoría son desconocidas.

**Bibliografía**

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301-320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>

Wold, H., Ruhe, A., Wold, S., & Dunn, III, W. J. (1984). The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3), 735–743. <https://epubs.siam.org/doi/10.1137/0905052>