Data Science

Nombre: Rodrigo Santa Maria

Fecha: 26/01/2024

Proyecto: "Bank-Marketing-Data"

Comisión: 42410

Profesor/a: Rebeca Figueroa

Tutor: Dario Ceballos

TABLA DE CONTENDIO

1. ABSTRACTO CON MOTIVACIÓN Y AUDIENCIA

Motivación:

La motivación detrás del análisis de los datos elegidos, basados en el conjunto de datos "Bank Marketing Data", radica en la necesidad de comprender y mejorar las estrategias de marketing en la industria bancaria. Este conjunto de datos proporciona una oportunidad única para explorar el comportamiento de los clientes en respuesta a campañas de marketing telefónico. La idea central es aprovechar la información contenida en estos datos para tomar decisiones informadas y diseñar estrategias de marketing más efectivas.

Audiencia:

La audiencia que podría beneficiarse de este análisis podría ser:

- **Gerentes de Sucursales Bancarias:**

Los gerentes de sucursales pueden aprovechar los insights para personalizar las ofertas y estrategias en función de las características de sus clientes locales, lo que podría aumentar las ventas en sus sucursales.

- **Profesionales del Marketing Bancario:**

Este análisis es relevante para los especialistas en marketing de la industria bancaria que buscan mejorar el rendimiento de sus campañas y aumentar las tasas de conversión de sus productos financieros.

- **Analistas Financieros:**

Los analistas financieros pueden utilizar los resultados de este análisis para comprender cómo las condiciones del mercado y las estrategias de marketing se relacionan con el éxito de los productos bancarios.

- **Economistas y Analistas del Mercado:**

Los economistas pueden utilizar los datos para evaluar el impacto de las tasas de interés y las tendencias macroeconómicas en las suscripciones a depósitos, lo que podría proporcionar información útil sobre la salud económica.

- **Reportes Financieros:**

Los periodistas especializados en finanzas pueden utilizar los insights para informar al público sobre las tendencias actuales en el sector bancario y su relación con las campañas de marketing.

- **Objetivo:**

El objetivo principal puede ser aumentar la tasa de suscripción a depósitos a plazo fijo entre los clientes.

Contexto Comercial:

El contexto comercial se refiere al entorno y las circunstancias en las que se llevan a cabo las actividades de marketing y negocios. El contexto comercial incluye la industria bancaria y financiera, donde el banco está buscando estrategias efectivas para promover sus productos financieros, como depósitos a plazo fijo.

Contexto Analítico:

El problema comercial podría ser la baja tasa de suscripción a depósitos a plazo fijo y la necesidad de aumentar esta tasa para mejorar la rentabilidad y la estabilidad financiera del banco.

Exploración de Datos (EDA):

Incluye la visualización de datos, la identificación de patrones, la detección de valores atípicos y la comprensión de las relaciones entre las variables para obtener información valiosa sobre el comportamiento de los clientes y su impacto en las suscripciones a depósitos a plazo fijo.

2. PREGUNTAS/HIPÓTESIS QUE QUEREMOS RESPONDER MEDIANTE EL ANÁLISIS DE DATOS

1. Como afecta la edad de los clientes a su suscripción a depósitos a plazo fijo?
2. Hay una relación entre el nivel educativo de los clientes y su disposición a suscribirse a depósitos a plazo fijo?
3. Cuál es el impacto de la situación laboral de los clientes en su decisión de suscribirse a depósitos a plazo fijo?
4. Existen diferencias en las tasas de suscripción entre los clientes con créditos en incumplimiento y los que no tienen?
5. Hay patrones estacionales en la suscripción a depósitos a plazo fijo a lo largo del año?

3. DESCRIPCIÓN DE LOS DATOS

El dataset es un conjunto de datos público el cual ha sido descargado del sitio Kaggle y se puede acceder en el siguiente link.

URL: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Tamaño del Conjunto de Datos:

- Nuestro conjunto de datos consta de 41,188 registros.

Valores Nulos:

- No se encontraron valores nulos en ninguna de las variables del conjunto de datos.

Duplicados:

- Realizamos un análisis exhaustivo y confirmamos que no hay registros duplicados en nuestro conjunto de datos.

Variable Objetivo "y":

- Hemos creado la variable objetivo "y" para utilizarla en nuestro modelo de machine learning. Esta variable es esencial para la clasificación y no presenta problemas de calidad de datos.

Esta información destaca la calidad del conjunto de datos y garantiza que estamos trabajando con datos limpios y completos.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 21 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             41188 non-null  int64
 1   job             41188 non-null  object
 2   marital         41188 non-null  object
 3   education       41188 non-null  object
 4   default         41188 non-null  object
 5   housing         41188 non-null  object
 6   loan            41188 non-null  object
 7   contact         41188 non-null  object
 8   month           41188 non-null  object
 9   day_of_week     41188 non-null  object
 10  duration        41188 non-null  int64
 11  campaign        41188 non-null  int64
 12  pdays           41188 non-null  int64
 13  previous        41188 non-null  int64
 14  poutcome        41188 non-null  object
 15  emp.var.rate    41188 non-null  float64
 16  cons.price.idx  41188 non-null  float64
 17  cons.conf.idx   41188 non-null  float64
 18  euribor3m       41188 non-null  float64
 19  nr.employed     41188 non-null  float64
 20  y               41188 non-null  object
dtypes: float64(5), int64(5), object(11)
```

Variables Numéricas:

* age: La edad del cliente
* duration: La duración de la última llamada en segundos.
* campaign: El número de contactos realizados durante esta campaña.
* pdays: El número de días que pasaron desde el último contacto con el
* previous: El número de contactos realizados antes de esta campaña.

* emp.var.rate: Tasa de variación del empleo.
* cons.price.idx (Índice de precios al consumidor - indicador mensual).
* cons.conf.idx (Índice de confianza del consumidor - indicador mensual).
* euribor3m: Tasa Euribor a 3 meses.
* nr.employed: Número de empleados.

Variables Categóricas:

* job: La profesión o empleo del cliente.
* marital: El estado civil del cliente.
* education: El nivel educativo del cliente.
* default: Si el cliente tiene crédito por defecto ("yes" o "no").
* housing: Si el cliente tiene un préstamo hipotecario ("yes" o "no").
* loan: Si el cliente tiene un préstamo personal ("yes" o "no").
* contact: El método de contacto utilizado para la campaña.
* month: El mes en que se contactó al cliente.
* day_of_week: El día de la semana en que se contactó al cliente.
* poutcome (Resultado de la campaña de marketing anterior).

Variable Objetivo Binaria (0 o 1)

* y: Indica si el cliente se suscribió ("yes") o no ("no") a un depósito a plazo fijo después de la campaña de marketing.

El dataset contiene 21 variables con 41.118 registros, entre los que se encuentra información de los clientes que se suscribieron a un plazo fijo, y los que no se suscribieron, la profesión y la educación de cada uno de ellos, si tienen préstamos hipotecarios, estado civil etc. También contiene información de la empresa, por ejemplo, la duración de la llamada con el cliente, el día de la semana que se lo contactó entre otros.

A continuación, se muestra una tabla resumen de todas las variables, con información de nulos, tipos de datos, cantidad de valores únicos.

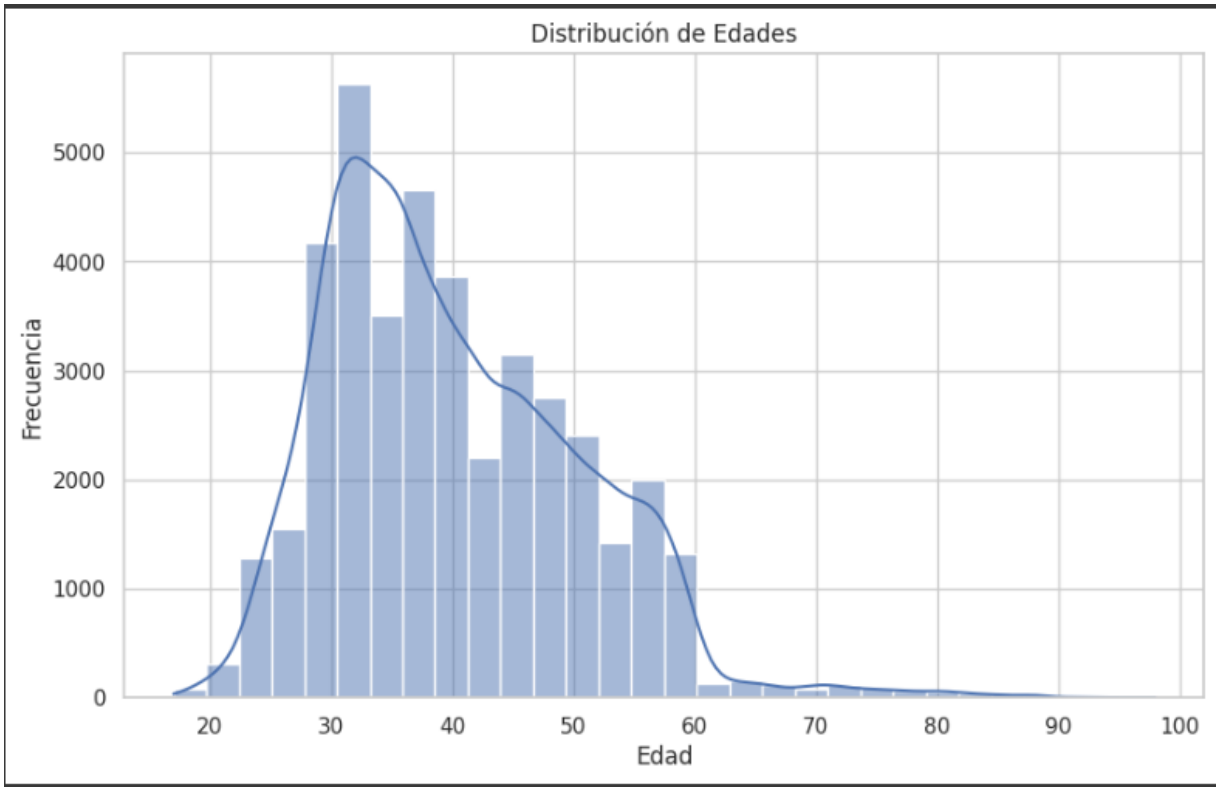
4. ANÁLISIS EXPLORATORIO DE DATOS (EDA)

En el análisis exploratorio se estudiaron las distribuciones de las variables numéricas y de las variables categóricas.

Entre los resultados que se destacan, se observó mediante un gráfico de barras, que de todos los datos registrados (41.188) solo %11.27 se suscribieron a un plazo fijo. El cual podemos encontrar información sesgada en nuestros datos para los clientes que "No" se suscribieron a un plazo fijo. Desde este punto podemos entender que, en nuestro entrenamiento de modelo deberíamos aplicar alguna técnica de Oversampling - Undersampling - Pesos de clase para más adelante.

1. **¿Cómo afecta la edad de los clientes a su suscripción a depósitos a plazo fijo?**

Analizamos con un gráfico Histograma, que nos permite visualizar la distribución de una variable numérica, en este caso, "age" la edad de los clientes. El propósito principal de construir un histograma es entender la frecuencia con la que ocurren diferentes rangos de edades en el conjunto de datos. También facilita la identificación visual de valores atípicos o extremos en la distribución de edades, si los hubiera.



- Interpretación de Ejes:

Indica que el eje horizontal muestra los rangos de edades, mientras que el eje vertical representa la frecuencia de los clientes en cada rango.

- El %96,87 de los clientes suscriptos, tienen entre 20 y 60 años. El otro %3,13 se encuentra entre los primeros años de adulto, y mayores a 60 años de edad.
- Mediante el gráfico aplicado, observamos la distribución de datos en la edad de los clientes, donde tenemos los puntos más altos y bajos. Con la media como resultado en 40 años de edad, y los Outliers que comienzan a partir de los 60 años en adelante, lo cual podemos deducir que podría ser por planes sociales para gente mayor, como jubilaciones.

Variable Numérica vs Categórica (Boxplot)

- Los datos que obtuvimos con un gráfico de bigotes (Boxplot) en relación a la edad y la suscripción a depósito, es que los rangos intercuartílicos (IQR) tienen una pequeña diferencia, y la línea de la mediana es similar. Lo cual indica que las edades medianas son comparables entre los suscriptores y no suscriptores.
- Los Outliers nos muestran más información en los que no se suscribieron, debido a que podemos tener información sesgada por la cantidad de datos de clientes mayor a 60 años de edad, que no se suscribieron.

2. Hay una relación entre el nivel educativo de los clientes y su disposición a suscribirse a depósitos a plazo fijo?

3. Cuál es el impacto de la situación laboral de los clientes en su decisión de suscribirse a depósitos a plazo fijo?

- La observación de que clientes con mejor nivel educativo y profesiones laborales más altas son los que más se suscriben a un depósito puede llevar a varias interpretaciones y puede ser valiosa para la toma de decisiones estratégicas, como estrategias de Marketing más personalizadas a las necesidades y preferencias de estos segmentos claves.
- Los resultados sugieren que las estrategias de marketing podrían personalizarse para abordar las necesidades y preferencias de estos segmentos clave. Esto podría incluir mensajes publicitarios específicos, ofertas personalizadas o canales de comunicación preferidos.
- Podría ser útil realizar un análisis de rentabilidad para evaluar la contribución de estos segmentos a los ingresos generales. Si estos clientes son más rentables, podrías asignar recursos adicionales para satisfacer sus necesidades.
- La diferencia de clase no se observó como un factor determinante.
- Realizando una tabla de contingencia de nuestra variable "Y" y "Job" nos proporciona información de que los clientes Administrativos, estudiantes y jubilados tienen un número significativo a suscripciones. El resto es un poco más moderado.
- Realizamos un análisis comparativo entre los clientes suscritos a depósito a plazo fijo con estado civil "casado" y "soltero". No se observa una diferencia significativa en las suscripciones a depósito a plazo fijo entre clientes casados y solteros. Esta conclusión indica que el estado civil no parece ser un factor determinante en la decisión de suscribirse a depósitos a plazo fijo en nuestro conjunto de datos.

Tabla de contingencia entre el trabajo y la suscripción

y	no	yes
job		
admin.	9070	1352
blue-collar	8616	638
entrepreneur	1332	124
housemaid	954	106
management	2596	328
retired	1286	434
self-employed	1272	149
services	3646	323
student	600	275
technician	6013	730
unemployed	870	144
unknown	293	37

4. Existen diferencias en las tasas de suscripción entre los clientes con créditos en incumplimiento y los que no tienen?

- Observamos una notable diferencia en las tasas de suscripción entre clientes con créditos en incumplimiento y aquellos sin incumplimientos.

Clientes con incumplimiento de créditos:

- No presentan suscripciones a depósitos a plazo fijo.

Clientes sin incumplimiento de créditos:

- Constituyen la mayoría de las suscripciones (aproximadamente el 88.73% de los clientes suscriptos).
- El 3.83% de los clientes presenta un estado de crédito "unknown" en nuestros datos y no tiene suscripciones.

Estos resultados indican una asociación entre el estado de crédito y las suscripciones, sugiriendo que los clientes sin incumplimientos son más propensos a suscribirse a depósitos a plazo fijo.

5. Hay patrones estacionales en la suscripción a depósitos a plazo fijo a lo largo del año?

Realizamos un gráfico de bigotes (BoxPlot) para obtener un análisis más completo. Como resultado, se evidencia una diferencia notable en la duración de la última llamada entre los clientes que se suscribieron ("yes") y los que no ("no").

La mediana y el tamaño de la caja son mayores para los clientes suscriptos, indicando una tendencia hacia llamadas más largas en este grupo.
Ambos casos (suscripción y no suscripción) presentan numerosos outliers, sugiriendo variabilidad en la duración de la llamada para ambos grupos.

Observaciones Principales:

- Las llamadas que resultan en suscripciones tienden a tener una duración más larga en promedio.

- Se observan outliers tanto para llamadas con suscripciones como para aquellas sin suscripciones.

Interpretación de Resultados:

- La mayor duración en las llamadas podría indicar una mayor discusión y convencimiento, lo que podría influir en la decisión de suscribirse.

Recomendación:

- Dado que la duración de la llamada parece estar asociada con las suscripciones, podría ser beneficioso considerar este factor al planificar futuras interacciones telefónicas.
- Considerar una revisión más detallada de las llamadas con duraciones atípicas para comprender mejor el contexto y las razones detrás de estas variaciones.

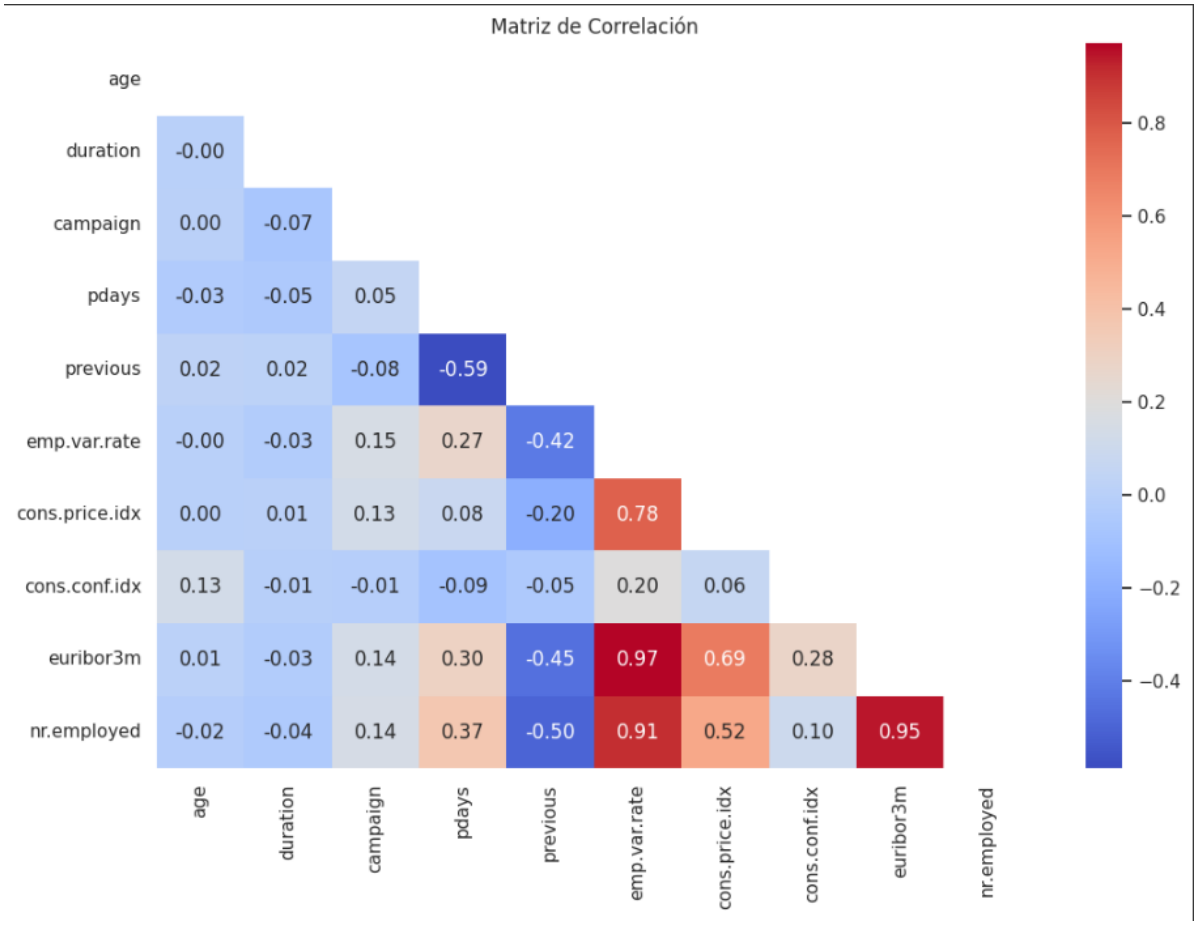
Estos hallazgos sugieren una relación entre la duración de la llamada y la decisión de suscripción, brindando información valiosa para estrategias de contacto y comunicación.

Estos hallazgos ofrecen una visión inicial de las relaciones en el conjunto de datos y establecen una base para análisis más detallados y la construcción de modelos predictivos. Próximos pasos incluirán una exploración más profunda de variables clave, la construcción y evaluación de modelos, así como un análisis de correlación para comprender las relaciones entre las variables. Con esta información, estamos listos para avanzar hacia una exploración más detallada y una comprensión más completa de nuestro conjunto de datos.

En nuestro proyecto, creamos un heatmap utilizando la matriz de correlación de las variables del conjunto de datos. Cada celda del mapa de calor representa el grado de correlación entre dos variables: cuanto más cercano sea el valor a 1 o -1, más fuerte será la correlación positiva o negativa, respectivamente. Los colores utilizados en el heatmap proporcionan una representación visual rápida de estas relaciones.

Al analizar el heatmap, pudimos identificar visualmente las variables que mostraban una correlación más fuerte o más débil. Esto nos ayudó a comprender mejor cómo diferentes características del conjunto de datos están interrelacionadas, lo que puede tener implicaciones importantes para nuestro modelo de machine learning.

Tabla de variables con índice de correlación mayor a 0,80.



Mediante nuestro HeatMap obtenemos la siguiente información más destacada.

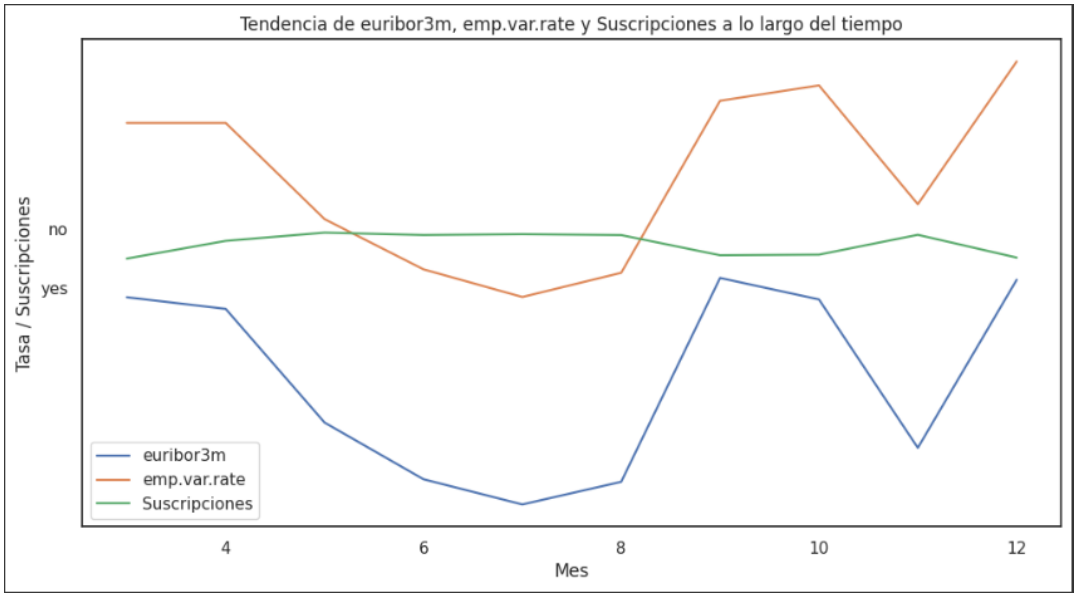
- **euribor3m y emp.var.rate:** Existe una fuerte correlación positiva de 0.97 entre la tasa Euribor a 3 meses (euribor3m) y la tasa de variación del empleo (emp.var.rate). Esto sugiere una relación cercana entre estos dos indicadores económicos.
- **Nr.employed y emp.var.rate:** Hay una correlación fuerte de 0.91 entre el número de empleados (Nr.employed) y la tasa de variación del empleo (emp.var.rate). Esto indica una conexión significativa entre la cantidad de empleados y la variación en las tasas de empleo.
- **Nr.employed y euribor3m:** Se observa una fuerte correlación positiva de 0.95 entre el número de empleados (Nr.employed) y la tasa Euribor a 3 meses (euribor3m). Esto sugiere una relación estrecha entre la cantidad de empleados y las tasas de interés Euribor.
- **cons.price.idx y emp.var.rate:** Existe una correlación de 0.78 entre el Índice de Precios al Consumidor (cons.price.idx) y la tasa de variación del empleo (emp.var.rate). Esta relación señala una conexión significativa entre la inflación y la variación en las tasas de empleo.

En general, las correlaciones observadas proporcionan información valiosa sobre cómo ciertas variables del conjunto de datos están interrelacionadas, lo que puede ser crucial para comprender el comportamiento de los clientes y mejorar la precisión de los modelos de machine learning.

La baja correlación entre la mayoría de las variables sugiere que hay diversidad e independencia en los datos.

En la fase de análisis de variables, hemos explorado las relaciones entre los atributos destacando las correlaciones fuertes mediante estadísticas descriptivas. Este proceso nos proporcionó una comprensión más profunda de las conexiones lineales significativas presentes en el conjunto de datos.

Para preservar la integridad de nuestros datos originales y permitir un seguimiento detallado de la evolución de las variables, hemos creado una copia del DataFrame inicial. Esta copia nos brinda la flexibilidad de observar cómo se comportan las variables a lo largo del tiempo y cómo estas dinámicas afectan la variable objetivo.



Al comparar las versiones originales y copiadas de las variables, hemos incorporado una línea de tiempo que destaca los momentos de suscripciones. Esta adición visual facilita la identificación de patrones y tendencias asociadas con los resultados de suscripción a lo largo del tiempo.

Este enfoque meticuloso y sistemático nos proporciona una base sólida para la construcción de modelos de machine learning, permitiéndonos tomar decisiones informadas sobre cómo las variables afectan las suscripciones a depósitos a plazo fijo.

ALGORITMOS ELEGIDOS

1. Árbol de decisión

A continuación de nuestro análisis de las variables con fuerte correlación, hacemos una codificación One-Hot de las variables categóricas, e inicializamos un árbol de decisión y entrenamos el modelo para ver los resultados.

En esta fase, aplicamos el algoritmo de Árbol de Decisión para construir un modelo predictivo. Inicialmente, implementamos la codificación One-Hot a nuestras variables categóricas utilizando la función **pd.get_dummies**. Este paso nos permitió convertir las variables categóricas en representaciones numéricas adecuadas para el modelo.

A continuación, inicializamos nuestro clasificador de Árbol de Decisión utilizando la clase **DecisionTreeClassifier** de la biblioteca **scikit-learn**. Entrenamos el modelo con el conjunto de datos de entrenamiento, permitiendo que el árbol aprenda patrones y relaciones presentes en los datos.

Para garantizar la consistencia entre el conjunto de entrenamiento y prueba, nos aseguramos de que ambos conjuntos tuvieran las mismas columnas después de la codificación One-Hot. Esto es crucial para que el modelo pueda realizar predicciones en el conjunto de prueba.

Luego, aplicamos nuestro modelo entrenado para realizar predicciones en el conjunto de prueba y evaluamos su rendimiento mediante métricas como la matriz de confusión, la precisión, la recall y el F1-score y lo visualizamos. La matriz de confusión proporciona una visión detallada de las predicciones del modelo en comparación con las clases reales.

Experimentamos con la profundidad del árbol (**max_depth**) para evitar sobreajuste y mejorar la interpretabilidad.

```
Matriz de Confusión:
[[7147  156]
 [ 742  193]]

Reporte de Clasificación:
              precision    recall  f1-score   support

          no       0.91      0.98      0.94      7303
         yes       0.55      0.21      0.30       935

    accuracy                           0.89      8238
   macro avg       0.73      0.59      0.62      8238
weighted avg       0.87      0.89      0.87      8238
```

Exploramos cómo estas variables se correlacionan con la variable objetivo (suscripciones a depósitos a plazo fijo, 'y') y cómo afectan al rendimiento del modelo predictivo.

Pudimos exportar una imagen del árbol de decisión para tener mejor vista del trabajo realizado.

A continuación, realizamos un modelo "Random Forest" lo cual es lo más apropiado para nuestro Dataset por los datos obtenidos, ya que es una extensión de los árboles de decisión y generalmente ofrece un rendimiento robusto. También nos sirve para ver si obtenemos diferentes resultados probando otros modelos.

2. Random Forest

Inicialización del Modelo:

- Utilizamos la clase **RandomForestClassifier** de la biblioteca scikit-learn para inicializar nuestro clasificador Random Forest.
- Definimos un modelo básico sin ajustar hiperparámetros específicos en esta fase inicial.

Entrenamiento del Modelo:

- Entrenamos el modelo utilizando el conjunto de datos de entrenamiento. Durante este proceso, el Random Forest construye varios árboles de decisión independientes.

Realización de Predicciones:

- Aplicamos el modelo entrenado para realizar predicciones en el conjunto de prueba. Las predicciones se obtienen considerando las decisiones de múltiples árboles y agregando sus resultados.

Evaluación del Rendimiento:

- Evaluamos el rendimiento del modelo mediante métricas como la **matriz de confusión, precisión, recall y F1-score**. Estas métricas nos brindan información detallada sobre cómo el modelo clasifica las instancias positivas y negativas.

Importancia de Variables:

- Calculamos la importancia de cada variable en el modelo Random Forest. Esto nos proporciona información sobre qué características son más influyentes en las decisiones del modelo.

Ajuste de Hiperparámetros (Opcional):

- En algunos casos, podemos realizar un ajuste de hiperparámetros utilizando técnicas como **Grid Search** para encontrar la combinación óptima de hiperparámetros que maximice el rendimiento del modelo.

Visualización de la Importancia de Variables:

- Visualizamos la importancia de las variables mediante un gráfico de barras horizontal. Esto nos ayuda a identificar qué características son más relevantes para el modelo.

Este enfoque de Random Forest es particularmente beneficioso para mejorar la capacidad predictiva y reducir el sobreajuste en comparación con un solo árbol de decisión. La combinación de múltiples árboles proporciona una mayor robustez y generalización del modelo.

```
Accuracy del modelo Random Forest: 0.8909
Reporte de clasificación:
              precision    recall  f1-score   support

          no       0.91      0.98      0.94      7303
         yes       0.55      0.22      0.31       935

    accuracy                           0.89      8238
   macro avg       0.73      0.60      0.63      8238
weighted avg       0.87      0.89      0.87      8238


Importancia de las variables:
euribor3m: 0.7157
emp.var.rate: 0.2843
```

Como observaciones parece que el modelo Random Forest ha proporcionado resultados similares a los obtenidos con el árbol de decisión. Debido a los resultados, vamos a intentar ajustar los hiperparámetros, para evaluar una mejora del rendimiento del modelo utilizando técnica como la búsqueda de cuadrícula "**GridSearchCV**" para encontrar una combinación óptima.

Definición de Hiperparámetros:

- Especificamos una cuadrícula de hiperparámetros que queremos explorar. Esto incluye opciones para el número de árboles (**n_estimators**), la profundidad máxima de los árboles (**max_depth**), el número mínimo de muestras requeridas para dividir un nodo (**min_samples_split**), y el número mínimo de muestras requeridas en una hoja (**min_samples_leaf**).

Inicialización del Clasificador Random Forest:

- Inicializamos un clasificador Random Forest con la configuración base.

Grid Search:

- Utilizamos la clase **GridSearchCV** de scikit-learn para realizar una búsqueda de cuadrícula. Esta técnica evalúa exhaustivamente el rendimiento del modelo para cada combinación de hiperparámetros en la cuadrícula.

Entrenamiento con Grid Search:

- Entrenamos el modelo utilizando Grid Search en el conjunto de datos de entrenamiento. Durante este proceso, se evalúan diferentes combinaciones de hiperparámetros.

Mejores Hiperparámetros:

- Identificamos los mejores hiperparámetros encontrados por Grid Search. Estos hiperparámetros representan la combinación que maximiza la métrica de evaluación seleccionada.

Evaluación del Modelo Ajustado:

- Realizamos predicciones en el conjunto de prueba utilizando el modelo ajustado con los mejores hiperparámetros. Evaluamos el rendimiento del modelo utilizando métricas como la matriz de confusión, precisión, recall y F1-score.

Visualización de Importancia de Variables (Opcional):

- Opcionalmente, podemos volver a visualizar la importancia de las variables para ver si hay cambios significativos después del ajuste de hiperparámetros.

*Mejores hiperparámetros: {'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 100}

Una vez ya obtenido los mejores hiperparámetros, ajustamos el conjunto de prueba, mostrando una Matriz de Confusión y el reporte de clasificación.
Como mencionamos anteriormente, hay un desbalanceo o un sesgo de clases dado por nuestra variable objetivo "y", mayor cantidad de datos NO suscriptos que de suscriptos. Por eso decidimos aplicar el uso de "Pesos de clase en los modelos"

Dividimos el conjunto de datos en características (x) y variable objetivo (y), luego codificamos las variables categóricas con One-Hot encoding, dividimos en conjuntos de entrenamiento y prueba para inicializar el clasificador de **RandomForest** con pesos de clase. Entrenamos el modelo, realizamos las predicciones en el conjunto, y evaluamos el rendimiento del modelo con pesos de clase.

```
Accuracy con pesos de clase: 0.9087157076960427
Matriz de Confusión con pesos de clase:
 [[7107  196]
 [ 556  379]]
Reporte de clasificación con pesos de clase:
              precision    recall  f1-score   support

          no       0.93      0.97      0.95      7303
         yes       0.66      0.41      0.50       935

    accuracy                           0.91      8238
   macro avg       0.79      0.69      0.73      8238
weighted avg       0.90      0.91      0.90      8238
```

- Parece que el modelo está mejorando en la clasificación de la clase positiva ("yes"), ya que los falsos negativos (556) disminuyeron en comparación con el modelo sin pesos de clase.
- El informe de clasificación proporciona métricas detalladas para cada clase. Podemos observar mejoras en la precisión, recuperación y puntuación F1 para la clase "yes" en comparación con el modelo sin pesos de clase.

Utilizaremos la validación cruzada para evaluar el rendimiento del modelo en diferentes divisiones de los datos y realizaremos una búsqueda de hiperparámetros para encontrar la combinación óptima.

Definimos los hiperparámetros a ajustar

```
param_grid = {
    'n_estimators': [50, 100, 150],
    'max_depth': [5, 10, 15],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
}
```

Inicializamos el clasificador RandomForest con la búsqueda de hiperparámetros con validación cruzada en los datos de entrenamiento, y obtenemos los mejores hiperparámetros encontrados iniciar el modelo.

Al ajustar los hiperparámetros y utilizar pesos de clase, observamos una disminución en la precisión global del modelo, pero esto podría estar relacionado con el equilibrio que se logra al mejorar el rendimiento para la clase minoritaria "yes". Aunque la precisión global disminuyó, nos acercamos a nuestro objetivo.

En resumen, el objetivo era reducir los falsos negativos e identificar las suscripciones a plazo fijo de la clase "yes", por ende, los ajustes del modelo fueron beneficiosos.

Realizamos un gráfico para visualizar la Curva AUC-ROC para el modelo RandomForest, convirtiendo las etiquetas de clase a valores binarios con Label-Encoder. Obteniendo las probabilidades predichas en lugar de las etiquetas de clase, calculamos la curva ROC y la graficamos.

- Eje X (Tasa de Falsos Positivos):

Explica que el eje X representa la tasa de falsos positivos. Proporción de instancias negativas incorrectamente clasificadas como positivas.

- Eje Y (Tasa de Verdaderos Positivos):

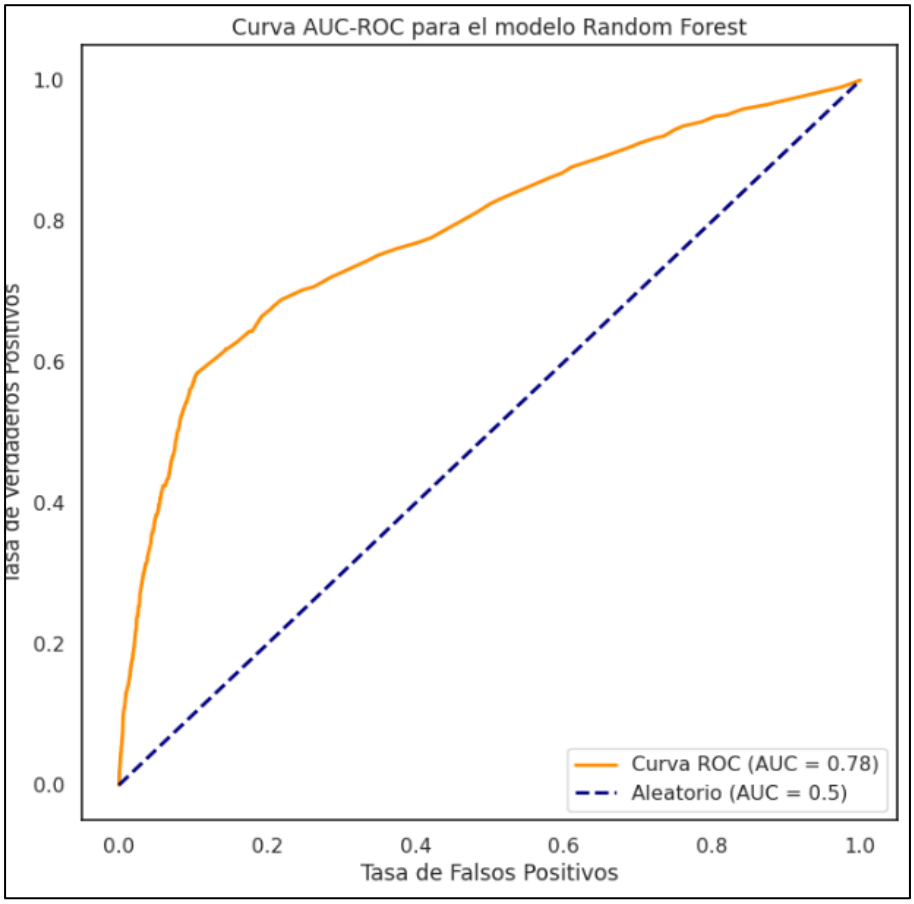
Describe que el eje Y representa la tasa de verdaderos positivos. Proporción de instancias positivas correctamente clasificadas como positivas.

- Área bajo la Curva (AUC):

Menciona el valor específico del AUC (en este caso, 0.78) y explica que indica la capacidad del modelo para distinguir entre clases positivas y negativas.

- Línea de Referencia Aleatoria:

Indica que la línea diagonal desde (0,0) hasta (1,1) representa una predicción aleatoria, y el AUC debería estar por encima de esta línea para considerarse un buen modelo.



A continuación, realizamos una prueba de Oversampling, para disminuir los datos, ya que obtenemos información sesgada por el desbalance de nuestro dataset para el modelo. Para eso primero creamos una copia para no poner en riesgo nuestros datos e importamos la biblioteca **imbalanced-learn.** La técnica utilizada para abordar el desequilibrio de clases es SMOTE (**Synthetic Minority Over-sampling Technique**) y probamos el modelo.

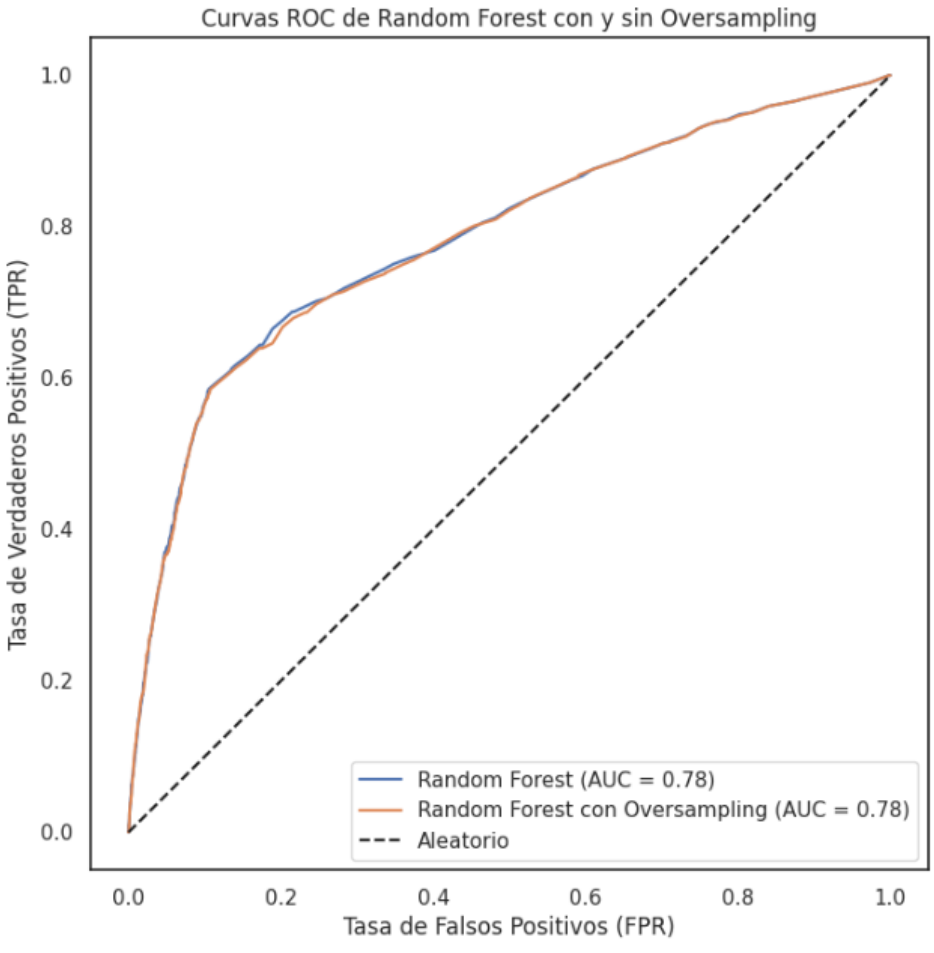
```
Accuracy del modelo Random Forest con oversampling: 0.8350
Reporte de clasificación:
              precision    recall  f1-score   support

          no       0.94      0.87      0.90      7303
         yes       0.36      0.58      0.44       935

    accuracy                           0.84      8238
   macro avg       0.65      0.72      0.67      8238
weighted avg       0.88      0.84      0.85      8238
```

El resultado muestra que el modelo Random Forest entrenado con oversampling tiene un rendimiento similar al modelo sin oversampling. La precisión para la clase 'yes' ha mejorado, pero el recall ha disminuido, lo que indica que el modelo puede estar clasificando menos casos positivos correctamente.

Es importante considerar que el oversampling puede no siempre mejorar el rendimiento del modelo, y en algunos casos, incluso puede empeorar.



CONCLUSIONES:

La variable objetivo, "y" (Si o No a la suscripción a un depósito a plazo fijo), está desbalanceada; tenemos un sesgo en nuestros datos, teniendo un %11,27 para los que "Si" se suscribieron mientras que el %88,73 restante no se suscribieron. Llegando a esta información, utilizamos técnicas de Oversampling para obtener mejores resultados del modelo.

De esta manera, concluimos que el mejor modelo fue RandomForest, junto con la técnica aplicada de pesos de clases para disminuir la información sesgada que teníamos en nuestro proyecto. A pesar de probar con Oversampling, no logramos mejorar el resultado. Optamos por dejar la optimización del modelo, ya que es el que mejor puede predecir si un cliente se suscribe a un plazo fijo o no.

El modelo alcanzó una exactitud (**accuracy**) del 91%, una precisión (precisión) del 66% para aquellos que sí (yes) se suscribieron a un depósito a plazo fijo, una tasa de verdaderos positivos (**Recall**) del 40.64%, y un área bajo la curva **ROC** igual a 0.78.