

BANK MARKETING

CASTRO IGNACIO JOAQUIN
CORREA BRAVO ALEXIS
DOTTO LUIS

TABLA DE CONTENIDOS

• <u>Descripción del negocio</u>	3
• <u>Objetivo del modelo</u>	5
• <u>Tabla de versionado</u>	6
• <u>Data Wrangling</u>	7
• <u>EDA</u>	9
• <u>Feature Selection</u>	13
• <u>Modelos</u>	14
• <u>Análisis de modelos</u>	15
• <u>Mejora de modelos</u>	16
• <u>Boosting models</u>	17
• <u>Conclusiones</u>	18
• <u>Mejoras / Futuras líneas</u>	19





DESCRIPCIÓN DEL NEGOCIO

Elegimos un **dataset** que contiene información relacionada con una campaña de marketing basada en llamadas telefónicas con el objetivo de ofrecer a los clientes la suscripción a un depósito a plazo.

Resulta interesante realizar un análisis para descubrir si existen variables y / o relaciones a las que debemos dedicar más o menos recursos, lo que permitirá a la compañía lograr sus objetivos de manera más eficiente.

OBJETIVO DEL MODELO

El objetivo de nuestro modelo es poder determinar qué tipo de clientes son los que suscriben a plazo fijo y estimar cuántos se suscribirán en la próxima campaña.

Averiguar cuántos se suscribieron en esta campaña y cuántos no, sus características comunes, tipo de correlaciones y otras medidas, nos permitirán descubrir si es necesario tomar decisiones que ayudarán a mejorar los resultados de ejercicios posteriores, manteniendo o hasta disminuyendo, si fuera posible, los costos.






A blurred background image of a financial chart with green bars and a red line graph, suggesting a market or investment context.

MOTIVACION

- ¿Será posible predecir cuántas personas suscribirán un depósito a plazo en la próxima campaña?
- ¿Qué pasa si nuestra clasificación no es precisa?
¿Qué pasa si estimamos de más?
- ¿Qué pasaría si hacemos una proyección inferior?
- ¿Cuál es la importancia de contar con un modelo sólido?

TABLA DE VERSIONADO

	SEMANA 16	SEMANA 19	SEMANA 22	SEMANA 23	SEMANA 25
LUN	Research individual 	Research grupal	EDA univariado	Boosting Models	
MAR			EDA multivariado 	Mejora de modelos	Agregado de Feature selection
MIE	Reunión grupal y puesta en común	Definición de caso de negocio 	Feature Selection		Mejora de ppt
JUE		Reunión con tutor para mostrar lo elegido		Conclusiones	Reunion de analisis
VIE		Investigación de Dataset	Models		
SAB	Reunión con nuevo tutor y elección del nuevo proyecto		Models y ppt	Mejora de markdown	Pulido del notebook y ppt 
DOM		Data Wrangling			

IMPORTANT

18/04 Primer entrega
11/05 Segunda Entrega
30/05 Tercer Entrega
03/07 Entrega Final

Herramientas utilizadas:

- Google calendar
- Google slides
- Zoom
- Google meet



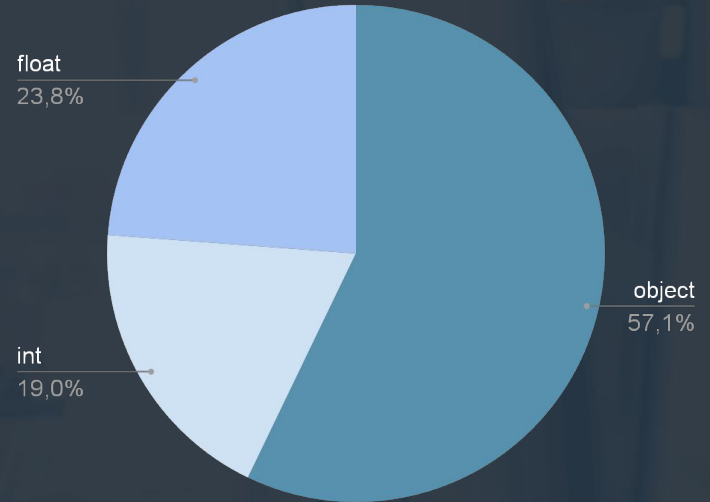


DATA WRANGLING

- En la etapa de limpieza, nos llamó la atención la presencia de una variable categórica “unknow”. En un inicio, realizamos todo el proyecto habiendo eliminado todas las filas en las que esta variable estuviese presente. Al final, luego de investigar y ver como trataban a este tipo de variable en otros proyectos, decidimos dejarla, ya que nos pareció que son variables que merecen análisis y forman parte de la predicción. Al hacer un análisis de los valores “unknow” solo en un feature representan el 20% del total.
- Revisamos cuántas filas duplicadas tenía el dataset y encontramos que eran 12, las cuales fueron eliminadas.

VARIABLES

- Age
- Job
- Marital
- Education
- Default
- Housing
- Loan
- Contact
- Month
- Day_of_week
- Duration
- Campaign
- Pdays
- Previous
- Poutcome
- Emp.var.rate
- Cons.price.idx
- Cons.conf.idx
- Euribor3m
- Nr.employed
- y



EXPLORATORY DATA ANALYSIS(EDA)

Para el EDA, nos pareció conveniente realizar un análisis dividiendo el dataset en secciones, analizando las variables de la siguiente manera:

- Relacionadas con el cliente
- Relacionadas con la campaña
- Relacionadas con el contexto económico y social.
- Otros atributos.

En los primeros dos casos nos pareció adecuado realizar un análisis exclusivamente univariado, y en los otros casos, por las características de las variables, realizar análisis bivariado o multivariados.

VARIABLES RELACIONADAS CON EL CLIENTE

AGE

En la distribución de edades observamos que los clientes tienen entre 17 y 98 años, y más allá de que la distribución tiene una tendencia marcada entre las edades que las personas son económicamente activas, no hallamos otra tendencia en particular.

JOBS

Aquí podemos observar que entre los clientes existen 12 categorías entre las cuales se destacan los trabajos de administrativos, blue-collar y técnicos. Las categorías con menor participación encontramos a los desempleados y estudiantes, como era esperado.

MARITAL

En la distribución de estado civil, vemos como la mayoría de los clientes son casados, seguidos por los solteros y los divorciados. También, encontramos otra vez la aparición de la categoría “unknow” es despreciable para el análisis de esta categoría.

EDUCATION

En esta categoría notamos que la mayoría de los clientes obtuvieron grados universitarios, o finalizaron sus estudios secundarios.. También, consideramos la categoría analfabeta como despreciable.

DEFAULT-HOUSING-LOAN

En este análisis vemos que la mayoría de los clientes no tienen préstamos en default. Que la cantidad de clientes con préstamos hipotecarios es mayor pero similar a la de los que no lo tienen. Por último, la cantidad de clientes sin préstamos es 5 veces superior a la cantidad de los que no tienen préstamos.

VARIABLES RELACIONADAS CON LA CAMPAÑA

DURATION

En el análisis de esta variable notamos que había 5 filas en las cuales tomaba valor 0, lo que significa que la llamada tuvo una duración de 0 segundos, por lo que creímos correcto eliminarlas ya que es obvio que si la llamada dura 0 segundos el cliente no contrató el servicio.

CONTACT

Aquí notamos que más de 25000 contactos se realizaron por celular y apenas 15000 por celular. Más adelante realizaremos un análisis con mayor profundidad.

MONTH

Analizando esta variable vimos que el mes con más registros es Mayo, y el que menos registros tiene es Diciembre.

DAY OF WEEK

Analizando los contactos según el día de la semana notamos que no hay una tendencia marcada de la empresa por contactarse en un día específico, pero los días que tienen mayor cantidad de registros son Lunes y Jueves.

ANÁLISIS BIVARIADO

En el análisis bivariado nos pareció relevante tomar en cuenta a las variables como nivel educativo, estado civil, ocupación, fechas y duración de las llamadas en relación a la variable target.

Luego de analizar las relaciones pudimos identificar que en general los clientes suscriben a un depósito a plazo, sin importar su estado civil, su ocupación o su edad. Notamos que en el caso de la edad, a partir de los 60 años los registros reducen muchísimo, como también la diferencias entre si contratan o no a un depósito a plazo, siendo prácticamente nulas.

Analizando fechas, notamos que los meses menos favorables fueron marzo, octubre, septiembre y diciembre.

En el caso de la duración de las llamadas, notamos que hay una leve relación entre duración de las llamadas y suscripción de depósitos a plazos.

En este análisis notamos que las variables que más influyen en el modelo son las variables como employment variation rate, consumer price index, consumer confidence index y otras variables más, que se relacionan con el contexto social y económico.

En un inicio nos llamó la atención, aunque tuviese cierta lógica. Luego, nos dimos cuenta que son variables que tienen mucha relación con la tasa de interés y que, casualmente, es una variable que no está en nuestro set de datos.

Entonces, podemos confirmar que las personas buscan mantener el valor de su dinero o, por qué no, obtener algún tipo de rendimiento, aprovechando la tasa de interés.

ANÁLISIS MULTIVARIADO

FEATURE SELECTION

PCA o Principal Component Analysis buscamos proyectar los puntos de varias dimensiones a un plano de menor dimensión

PCA

En esta etapa buscamos reducir la dimensionalidad de nuestro dataset. Analizamos ambos métodos pero terminamos quedandonos con el SFS, ya que el PCA generaba muchas variables sintéticas

SFS o Sequential Forward Selection va agregando cada feature secuencialmente hacia adelante y mostrando la performance

**WRAPPER METHOD -
SFS**



MODELOS



SUPPORT VECTOR MACHINE SVM

kernel='rbf'
C=0.1

n_estimator): 200
Class Weight: Balanced
Max Features: Log2

RANDOM FOREST



DECISION TREE

Altura máxima de 2

*Para todos los modelos se utiliza 30% test / 70% train

ANALISIS DE MODELOS

	EXACTITUD	PRECISION	F1 SCORE	ERRORES I Y II
SVM	89.6057%	12.9074%	9.0369%	1284
RANDOM FOREST	91.1033%	69.1003%	49.0968%	1099
DECISION TREE	90.3479%	60.3413%	49.6694%	1218

El modelo **SVM** regresa las peores métricas de Exactitud, Precisión y F1. También los errores tipo I y II son mayores que en los otros modelos. Por lo tanto, lo descartamos.

Ahora comparando los dos modelos restantes tenemos al **Decision Tree y Random Forest**.

Tiene mejores métricas en cuanto a Accuracy, Precisión y menor cantidad de errores, pero un F1 Score menor.

Si nos remontamos a las fórmulas de las métricas mencionadas anteriormente, vemos que el F1 incluye el Recall. Esta métrica de exhaustividad nos va a informar sobre la cantidad que el modelo de machine learning es capaz de identificar.

En el ejemplo, se refiere a que la exhaustividad (recall) es la respuesta a la pregunta: ¿Qué porcentaje de los clientes interesados somos capaces de identificar?

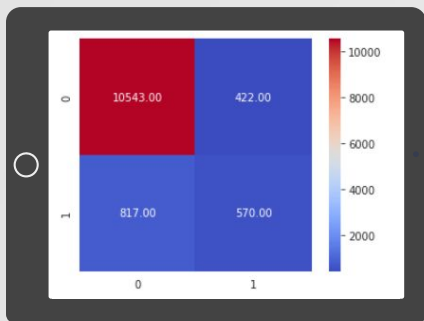
Por otro lado vemos que la variación de F1 del Decision Tree con respecto al Random Forest no es mucha.

Entonces, podríamos concluir que, al menos entre estos dos modelos, **podríamos elegir al Random Forest como el más adecuado**.

MEJORA DE MODELOS

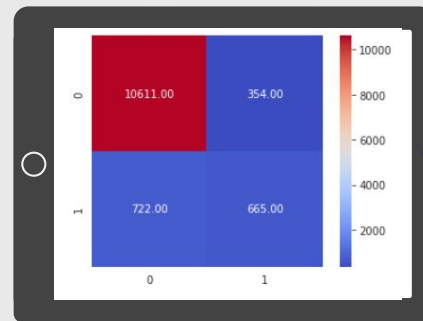
Aplicamos mejora de modelos al Decision Tree Classifier y al Random Forest a través de GridSearch y en ambos modelos las métricas de Exactitud como precisión mejoran significativamente

Decision Tree + GridSearch



Exactitud Decision Tree: **0.8572**
Exactitud Decision Tree + GridSearch: **0.8996**
Precisión Decision Tree: **0.4944**
Precisión Decision Tree + GridSearch: **0.7513**

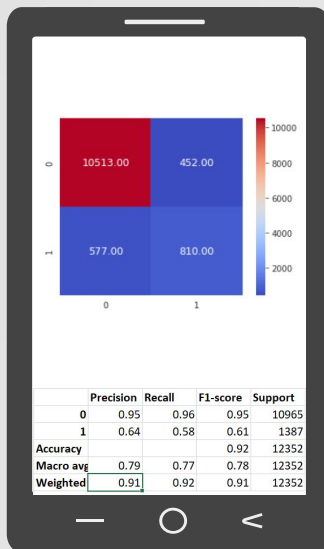
Random Forest + GridSearch



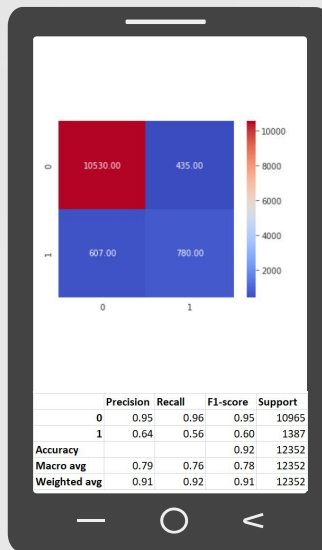
Exactitud Random Forest: **0.8261**
Exactitud Random Forest + GridSearch: **0.9128**
Precisión Random Forest: **0.4981**
Precisión Random Forest + GridSearch: **0.7944**

BOOSTING MODELS

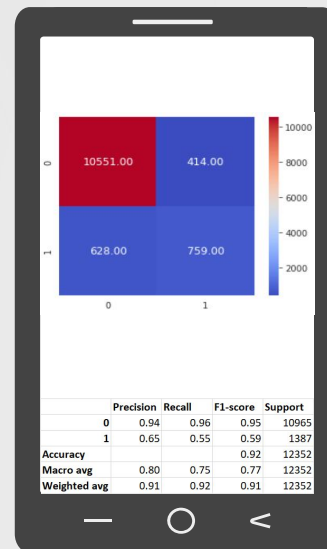
LIGHTGBM



XGBOOST



CATBOOST



Se obtienen mejores resultados utilizando los modelos de boosting.

*Para todos los modelos se utiliza 30% test / 70% train

CONCLUSIONES

	Precision	Recall	F1 score	Exactitud	Score
SVM	0.79	0.60	0.64	0.90	3.57
D. Tree	0.74	0.59	0.61	0.89	3.44
R. Forest	0.80	0.72	0.75	0.91	3.93
XGBoost	0.79	0.76	0.78	0.92	4.03
LIGHTGBM	0.79	0.77	0.78	0.92	4.04
CatBoost	0.80	0.75	0.77	0.92	4.01

Calculando los score vemos que XGBoost y Lightgbm son los que ofrecen los mejores resultados. Elegimos LightGBM porque tiene el score más alto, brindándonos una mejor predicción en las personas que se suscribirán a un plazo.

Podemos concluir también que al implementar la solución ofrecida, el banco no solo se beneficiaría en una reducción de costos por campañas que no se orientan al tipo de usuario adecuado, sino que también sería redituable a nivel de ingresos, ya que mejorar la eficiencia de las campañas puede orientarse también a captar nuevos clientes. Ambos casos sirven para mejorar las campañas de marketing enfocándose en los distintos perfiles de usuarios.



MEJORAS / FUTURAS LÍNEAS

- Para seguir profundizando en el proyecto se pueden usar otros métodos en Feature selection, un método exhaustivo por ejemplo. Para comparar la relevancia de los features y de allí elegir los mejores según los métodos empleados.
- También se podría generar categorización de perfiles de forma automática para clientes existentes o nuevos.
- Para que el modelo sea más robusto y no esté sesgado solo a un caso de negocio, podemos utilizar datasets de marketing de otros bancos para ver cómo features similares, o no, impactaron en los ingresos.
- En base a nuestras observaciones, podemos retroalimentar las futuras encuestas modificandolas quitando features sin aportes relevantes a la modelización, para recortar tiempos de adquisición de datos y su procesamiento.
- Podrían buscarse patrones y relaciones con otros dataset que no necesariamente sean provenientes de llamadas telefónicas, por ejemplo encuestas online para tener mayor conocimiento de los usuarios.