Nombre del proyecto: Proyecto de ML para incrementar la efectivada de las campañas de marketing bancarias para la colocación del producto deposito a plazo.

Introducción

Los datos están relacionados con campañas de marketing directo (llamadas telefónicas) de una institución bancaria portuguesa. El objetivo de la clasificación es predecir si el cliente suscribirá un depósito a plazo (variable y).

Objetivos del Proyecto

1. **¿Cuáles son los objetivos del negocio?**

Determinar la probabilidad de comprar del producto depósito a plazo por parte de un cliente dadas sus características y de esta forma enfocar los esfuerzos del equipo comercial para incrementar la tasa de éxito de las campañas de marketing.

2. **¿Qué decisiones o procesos específicos se desean mejorar o automatizar con ML?**

A partir de los datos generados diariamente por los equipo comerciales mediante llamadas telefónicas, se pretender automatizar el procesamiento de esta información para generar conclusiones sobre la tasa de éxito de una venta y tomar decisiones estratégicas para el cumplimiento de los objetivos comerciales, al igual que personalizar la propuesta de valor del banco ofreciendo un portafolio de productos financieros que incluya el depósito a plazo a los clientes que cuentan con mayor probabilidad de requerirlo.

3. **¿Se podría resolver el problema de manera no automatizada?**

Dada la naturaleza de generación de datos diarios de una gran cantidad de agentes de ventas y clientes, las betas que ponderan el impacto de las variables independientes, constantemente se modifican por los gustos y preferencias de los consumidores, la situación económica que este atravesando la zona de influencia entre otros aspectos, este es un problema que requiere necesariamente una solución automatizada que sea capaz de capturar esos cambios a través del tiempo y brinde una solución alineada con la velocidad en la que se desarrolla específicamente el sector bancario.

Metodología Propuesta

Breve descripción del tipo y algoritmo de ML a utilizar.

El algoritmo más adecuado para resolver este problema es el ***Regresor de Bosques Aleatorios (Random Forest Regressor).*** Este algoritmo es eficaz para predecir valores numéricos, como el promedio en cuenta del ultimo año, la duración de las llamadas y los intentos de contacto realizados, porque maneja bien datos complejos y relaciones no lineales. Además, es robusto frente a valores atípicos y reduce el riesgo de sobreajuste (overfitting) al promediar los resultados de múltiples árboles de decisión y puede manejar variables categóricas como es el caso del job, housing, education, etc.

Las métricas de evaluación que se utilizarán son:

Error Cuadrático Medio (MSE | Mean squared error): Para medir la diferencia entre los valores predichos y los reales.

R^2 (Coeficiente de Determinación): Para evaluar qué tan bien el modelo explica la variabilidad del gasto anual.

Teniendo en cuenta que las features disponibles incluyen:

age, job, marital, education, default, balance, housing, loan, contact, day, month, duration, campaign, pdays, previous, poutcome, deposit.

La justificación de que el Regressor de Bosques Aleatorios (Random Forest Regressor) es una excelente elección recae en varias razones:

- **Capacidad de manejar tanto variables numéricas como categóricas:** Random Forest puede trabajar fácilmente con datos numéricos como el tiempo de interacción y el promedio de depósitos por año, así como con variables como el nivel educativo o trabajo. No requiere un extenso preprocesamiento o transformación de los datos categóricos, ya que los árboles de decisión dentro del bosque los manejan de manera natural.
- **Reducción del riesgo de sobreajuste (overfitting):** Dado que el modelo utiliza múltiples árboles de decisión (cada uno entrenado con diferentes subconjuntos de los datos), promedia los resultados, lo que reduce el riesgo de que un solo árbol aprenda demasiado los detalles específicos del conjunto de datos (sobreajuste) y pierda capacidad de generalización.
- **Tratamiento de relaciones no lineales:** Las interacciones entre el tiempo llamada y el el nivel educativo podrían tener efectos complejos en la probabilidad de que un cliente tome un depósito a término. Random Forest maneja bien estas relaciones no lineales, que serían difíciles de capturar con modelos más simples como la regresión lineal.
- **Robustez frente a datos faltantes o ruidosos:** Si hay datos incompletos o valores atípicos en las interacciones de los clientes o el historial de compras, Random Forest puede ser más tolerante y aun así ofrecer buenas predicciones sin necesidad de eliminar o imputar datos de manera intensiva.
- **Importancia de las características:** Random Forest ofrece una forma de medir la importancia de las características, lo que ayuda a identificar cuáles de las variables que tienen un mayor impacto en la probabilidad de éxito. Esta capacidad de análisis es muy útil para la empresa, ya que permite enfocar los esfuerzos en clientes con características especificas, y crear campañas a medida.

Métrica de éxito del proyecto

Incrementar el porcentaje de efectividad de una campaña de marketing en un 30%: Medida a través de la cantidad de depósitos a termino colocados en un periodo de tiempo determinado.

Responsabilidades Éticas y Sociales

Privacidad y protección de datos: El modelo utilizará datos personales e identificativos de los clientes.

Transparencia en las decisiones automatizadas: Es importante que las predicciones los clientes no sean utilizadas de manera discriminatoria o para limitar las oportunidades de algunos clientes. El modelo debe ser transparente y auditable, de forma que se pueda explicar cómo se toman las decisiones y evitar sesgos que perjudiquen a ciertos grupos de clientes.

Evitar el sesgo algorítmico: Asegurarse de que el modelo no favorezca a un tipo específico de cliente. El sesgo en el modelo podría generar desigualdad en el trato y experiencia de compra de los clientes.

Impacto social: El acceso a los productos financieros es un derecho de los ciudadanos, se debe ser cauteloso con las conclusiones ya que el tratamiento de este tipo de instrumentos, impacta en el bienestar social.

Diego Lagos
Business Intelligence Analyst
Linkedin
+34 632 67 31 73
diegoandres0905@gmail.com

Segundo envío actividad 3.2 - Sprint 3

Comentario en Moodle Instructora It Academy: Verónica Figueroa

"El enfoque del trabajo sería adecuado si la variable resultado fuera continua. Sin embargo, dado que el objetivo es predecir una variable binaria relacionada con la contratación de depósitos, el algoritmo que has seleccionado no es el más apropiado. Te invito a realizar un segundo intento para que puedas reformular tu propuesta y ajustar la elección del algoritmo a las necesidades del proyecto."

Respuesta estudiante Diego Lagos:

Después de estudiar nuevamente la necesidad del proyecto y teniendo en cuenta el dataset proporcionado por el ejercicio y su explicación:

"Bank Marketing

Donated on 2/13/2012

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y)."

Variables Table						
Variable Name	***Role***	***Type***	***Demographic***	***Description***	***Units***	***Missing Values***
age	Feature	Integer	Age			no
job	Feature	Categorical	Occupation	type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')		no
marital	Feature	Categorical	Marital Status	marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)		no
education	Feature	Categorical	Education Level	(categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')		no
default	Feature	Binary		has credit in default?		no
balance	Feature	Integer		average yearly balance	euros	no
housing	Feature	Binary		has housing loan?		no
loan	Feature	Binary		has personal loan?		no
contact	Feature	Categorical		contact communication type (categorical: 'cellular','telephone')		yes
day_of_week	Feature	Date		last contact day of the week		
month	Feature	Date		last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')		no
duration	Feature	Integer		last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.		no
campaign	Feature	Integer		number of contacts performed during this campaign and for this client (numeric, includes last contact)		no
pdays	Feature	Integer		number of days that passed by after the client was last contacted from a previous campaign (numeric; -1 means client was not previously contacted)		yes
previous	Feature	Integer		number of contacts performed before this campaign and for this client		no
poutcome	Feature	Categorical		outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')		yes
y	Target	Binary		has the client subscribed a term deposit?		no

Reitero mi justificación del uso del algoritmo Random Forest Regression para este proyecto y, respetuosamente, solicito una explicación más detallada por parte de la instructora sobre por qué no considera apropiado este algoritmo para un problema de clasificación binaria.

El algoritmo Random Forest Regression es el más apropiado para cumplir con los objetivos de este proyecto debido a su capacidad para manejar diferentes tipos de datos y su flexibilidad en situaciones donde las relaciones entre las características y la variable dependiente no son necesariamente lineales. Por su parte, aunque la regresión logística es uno de los algoritmos más utilizados para problemas de clasificación binaria, modelando la probabilidad de que una instancia pertenezca a una clase (0 o 1) mediante una función sigmoide, presenta limitaciones que lo hacen menos adecuado para este caso.

Una de las principales ventajas de la regresión logística es su interpretabilidad; los coeficientes obtenidos indican la magnitud del impacto de cada variable independiente sobre la probabilidad del resultado, lo que no solo permite determinar si el cliente tomará o no el producto financiero, sino también qué variables tienen más o menos peso en su decisión final a través de los parámetros estimados. Además, su simplicidad permite un entrenamiento más rápido en comparación con Random Forest. Sin embargo, esta técnica asume relaciones lineales entre las características y la variable dependiente, lo cual puede disminuir la precisión del modelo drásticamente si los datos no se ajustan a esta suposición. Por lo tanto, la regresión logística puede resultar inadecuada para las necesidades de este proyecto.

Una de las fortalezas de los Random Forest es su capacidad para manejar una amplia variedad de tipos de datos. En el caso de la clasificación binaria, es fundamental aplicar técnicas de preprocesamiento adecuadas debido a la diversidad de las variables (numéricas, categóricas, binarias y de fechas) para obtener como resultado una clasificación binaria, 1 o 0. Este preprocesamiento garantiza que el algoritmo Random Forest pueda interpretar el conjunto de datos de manera efectiva.

Preprocesamiento requerido para el manejo de diferentes tipos de variables:

- **Variables numéricas:** Datos como edad, ingresos y saldo pueden ser utilizados directamente por Random Forest, ya que los árboles de decisión que lo componen realizan divisiones basadas en estos valores para asignar resultados de 0 o 1 según pertenencias a rangos determinados.
- **Variables booleanas:** Random Forest puede manejar características booleanas, como si un cliente ya posee un producto financiero.
- **Variables categóricas:** Se puede utilizar la técnica de codificación One-Hot para transformar categorías en variables binarias. Por ejemplo, para el estado civil: casado (1) o no casado (0), lo que permite que Random Forest procese adecuadamente esta información.
- **Fechas:** Aunque las fechas no se pueden utilizar en su forma original, se pueden extraer componentes valiosos, como el día, el mes y el año. Además, se pueden calcular diferencias entre fechas, lo que resulta útil para medir la antigüedad del cliente o incluir variables estacionales. Por ejemplo, agrupar los meses en estaciones del año (primavera, verano, otoño, invierno y asignarles valores de 1 y 0 al igual que en las categóricas por medio de One-Hot) puede ser relevante para campañas de marketing y desempeño comercial.

Resumen:

El algoritmo Random Forest construye múltiples árboles de decisión utilizando una combinación de características que requieren ser preprocesadas (numéricas, categóricas transformadas, variables binarias y fechas transformadas). Cada árbol realiza una predicción (0 o 1) y el resultado final se obtiene mediante el voto mayoritario de estos árboles, lo que proporciona una clasificación binaria. Además, Random Forest puede ofrecer probabilidades si se requiere una predicción probabilística, aunque para fines de clasificación, el resultado será 0 o 1.

Conclusión:

El algoritmo Random Forest es adecuado para problemas de clasificación binaria, como el que se plantea en este proyecto: predecir si un cliente tomará un producto financiero (resultado 1) o no (resultado 0). Este algoritmo es capaz

de manejar tanto problemas de regresión (variables continuas) como problemas de clasificación (variables discretas o categóricas), teniendo en cuenta la necesidad de preprocesar los datos como se explicó anteriormente.

Justificación de Random Forest para clasificación binaria:

- Clasificación binaria: Random Forest es eficaz para problemas donde la variable objetivo es binaria, ya que construye múltiples árboles de decisión y vota por la clase mayoritaria (0 o 1).
- Manejo de datos heterogéneos: Random Forest trabaja bien con variables numéricas y categóricas simultáneamente, lo cual es esencial para este proyecto, dado que se incluyen características mixtas (ingresos, estado civil, edad, etc.).
- Capacidad para manejar ruido y evitar sobreajuste: Al promediar múltiples árboles, Random Forest es menos propenso al sobreajuste en comparación con un solo árbol de decisión, lo que es especialmente útil en contextos con datos ruidosos o complejos.

Diego Lagos
Business Intelligence Analyst
Linkedin
+34 632 67 31 73
diegoandres0905@gmail.com