Trabajo Practico No.4
Final

Por:
Carlos Daniel Murillo
Andrés Ramírez

Docente:
Andrés Quintero Zea

Asignatura:
Programación

Universidad EIA
6/06/2024

1. Introducción

El presente informe documenta el análisis y la predicción del comportamiento de clientes de una institución bancaria portuguesa, con el fin de determinar si un cliente suscribirá un depósito a plazo. La base de datos utilizada proviene del UCI Machine Learning Repository y contiene información relacionada con las campañas de marketing directo de la entidad, realizadas a través de llamadas telefónicas. El objetivo principal es construir modelos de clasificación que puedan predecir con precisión la suscripción de depósitos a plazo por parte de los clientes.

2. Objetivo de la Práctica

El objetivo de esta práctica es aplicar técnicas de análisis y modelado de datos para predecir si un cliente suscribirá un depósito a plazo fijo. Esto implica:

- Realizar una exploración y preprocesamiento de los datos.
- Implementar y evaluar modelos de clasificación como Support Vector Classifier (SVC) y Random Forest.
- Comparar los modelos para identificar el más preciso y robusto.
- Presentar las conclusiones y posibles aplicaciones en un entorno real.

3. Metodología

3.1 Exploración de Datos

Primero, se realizó una exploración inicial de los datos para entender su estructura y características. Esto incluyó la generación de estadísticas descriptivas y la visualización de distribuciones de las variables numéricas y categóricas, así como la detección de valores atípicos mediante el método de 1.5 IQR.

Código de exploración de datos (ver Notebook 01 - exploración.ipynb)

3.2 Preprocesado de Datos

El preprocesamiento de datos es un paso crucial para preparar los datos para el modelado. Se eliminaron las filas con valores nulos y se estandarizaron las variables numéricas. Además, se codificaron las variables categóricas mediante OneHotEncoder. Esto se realizó utilizando pipelines de scikit-learn para asegurar un procesamiento eficiente y reproducible.

Código de preprocesado de datos (ver Notebook 02 - preprocesado.ipynb)

3.3 Modelado

Se construyeron dos modelos de clasificación: Support Vector Classifier (SVC) y Random Forest. Cada modelo se entrenó y evaluó utilizando técnicas de validación cruzada y GridSearchCV para encontrar los mejores hiperparámetros. Se evaluaron los modelos con métricas de precisión, matriz de confusión y curvas ROC.

Código del modelo SVC (ver Notebook 03 - modelo 1.ipynb)
Código del modelo Random Forest (ver Notebook 04 - modelo 2.ipynb)

3.4 Comparación de Modelos

Se compararon los modelos SVC y Random Forest en términos de precisión, matriz de confusión y área bajo la curva ROC (AUC). Esto permitió identificar cuál modelo proporciona una mejor capacidad predictiva y robustez.

Código de comparación de modelos (ver Notebook 05 - comparación.ipynb)

4. Resultados

4.1 Evaluación del Modelo SVC

- **Precisión**: El modelo SVC clasificó correctamente el 85% de los clientes no interesados y el 88% de los interesados.
- **Errores**: Se equivocó en un 10% al clasificar clientes no interesados como interesados y un 17% en clasificar interesados como no interesados.
- **Capacidad de Detección**: Tiene un área bajo la curva (AUC) de 0.91, indicando buena capacidad para diferenciar entre clientes interesados y no interesados.

4.2 Evaluación del Modelo RandomForest

- **Precisión**: El modelo RandomForest clasificó correctamente el 88% de los clientes no interesados y el 91% de los interesados.

- **Errores**: Se equivocó en un 8% al clasificar clientes no interesados como interesados y un 13% en clasificar interesados como no interesados.
- **Capacidad de Detección**: Tiene un AUC de 0.93, mostrando una capacidad ligeramente superior para diferenciar entre clientes interesados y no interesados.

4.3 Comparación de Modelos

- **Desempeño**: El modelo RandomForest tuvo un mejor desempeño general, siendo más preciso y con menos errores en la clasificación.

5. Conclusiones

Los resultados indican que ambos modelos tienen una alta capacidad predictiva, pero el modelo de Random Forest supera al SVC en términos de precisión y área bajo la curva ROC. El modelo de Random Forest también muestra una mejor capacidad para manejar la variabilidad de los datos, lo que lo hace más robusto para su uso en un entorno real.

Aplicación en un Entorno Real

En un entorno real, el modelo de Random Forest puede integrarse en el sistema de CRM (Customer Relationship Management) del banco para mejorar las campañas de marketing. Al predecir qué clientes tienen una mayor probabilidad de suscribir un depósito a plazo, el banco puede personalizar sus estrategias de contacto y optimizar recursos, incrementando la tasa de éxito de las campañas y mejorando la satisfacción del cliente.

Este enfoque puede ser ampliado a otras áreas del negocio bancario, como la detección de fraude y la evaluación de riesgos crediticios, donde la capacidad de manejar grandes volúmenes de datos y proporcionar predicciones precisas es crucial.