Proyecto Final Aprendizaje Automático de máquina, MACC

Brayan Steven Calderon Adames

Bogotá D.C

Noviembre de 2022

Introducción

1. Resumen del proyecto, descripción de la base y detalles generales
2. Exploratorios y descriptivos
3. Metodología
4. Resultados y conclusiones

Descripción de la base de datos

Los datos están relacionados con campañas de marketing directo de una entidad bancaria portuguesa. Las campañas de marketing se basaron en llamadas telefónicas. A menudo, se requería más de un contacto con el mismo cliente, para poder acceder si el producto (depósito bancario a plazo) estaría (sí) o no (no) suscrito.

1 - edad (numérico)
2 - trabajo: tipo de trabajo (categórico: administrador, cuello azul, empresario, criada, administración, jubilado, autónomo, servicios, estudiante), técnico, desempleado, desconocido)
3 - marital: estado civil (categórico: divorciado, casado, soltero, desconocido; nota: divorciado significa divorciado o viudo)
4 - educación: básico.4 años, básico.6 años, básico.9 años, bachillerato, analfabeto, curso. profesional, título. universitario, desconocido)
5- mora: ¿tiene crédito en mora? (categórico: no, sí, desconocido)
6 - vivienda: ¿tiene préstamo de vivienda? (categórico: no, sí, desconocido)
7 - préstamo: ¿tiene préstamo personal? (categórico: no, sí, desconocido)
8 - contacto: tipo de comunicación de contacto (categórico: celular, teléfono)
9 - mes: último mes de contacto del año (categóricos: ene, feb, mar, ..., nov, dec)
10 - day_of_week: último día de contacto de la semana (categórico: lun, mar, mié, jue, vie)
11 - duración: duración del último contacto, en segundos (numérico). Nota importante: este atributo afecta en gran medida el objetivo de salida (por ejemplo, si la duración = 0, entonces y = no). Sin embargo, la duración no se conoce antes de que se realice una llamada. Además, después del final de la llamada y es obviamente conocido. Por lo tanto, esta entrada solo debe incluirse con fines de referencia y debe descartarse si la intención es tener un modelo predictivo realista.
12 - campaña: número de contactos realizados durante esta campaña y para este cliente (numérico, incluye último contacto)
13 - pdays: número de días que transcurrieron desde la última vez que se contactó al cliente de una campaña anterior (numérico; 999 significa que el cliente no fue contactado previamente)
14 - anterior: número de contactos realizados antes de esta campaña y para este cliente (numérico)
15 - poutcome: resultado de la campaña de marketing anterior (categórico: fracaso, inexistente, éxito)
atributos del contexto social y económico
16 - emp.var.rate: tasa de variación del empleo - indicador trimestral (numérico)
17 - cons.price.idx: índice de precios al consumidor - indicador mensual (numérico)
18 - cons.conf.idx: índice de confianza del consumidor - indicador mensual (numérico)
19 - euribor3m: tasa euribor 3 meses - indicador diario (numérico)
20 - nr.employed: número de empleados - indicador trimestral (numérico)

Exploración y descripción de los datos:

La base de datos se encontró en un repositorio de estados unidos donde contienen muchas bases de datos que son utilizadas para Machine Learning, en esta ocasión se hace una

revisión preliminar de los datos con un resumen en la información, encontrando nulos y el tipo de dato de cada covariable, siguiente a esta revisión se hace un descriptivo de las variables numéricas observando los percentiles los cuales indican la posición o el rango en el que se mueven los datos de cada campo observado dentro de la tabla, posterior a la descripción de las variables cuantitativas se realizan tablas de frecuencia para poder de visualizar a primera vista como se distribuyen los datos, para verlo de mejor manera se decide hacer unos histogramas en donde se evidencia la distribución de los datos y para las categorías se hacen gráficos de barras.

Transformación de los datos:

Se decidió transformar las variables categóricas para poder tenerlas como input dentro del modelo de clasificación, cada variable fue convertida en números que pueden ser utilizados como evaluadores y métricas dentro de correlaciones.

Es importante resaltar que no se encontraron nulos dentro de la base, pero si categorías que contenían el valor de desconocido o no respondido, se pensó en quitar alguna de ellas, pero el porcentaje de datos con esta categoría es demasiado alto dentro de la tabla.

Correlaciones:

Se encontraron correlaciones demasiado altas en alguna de las covariables, para confirmar esto deberá hacerse una prueba de hipótesis para tener certeza, aun así, se decidió quitar dos covariables para evitar problemas de multicolinealidad.

Hallazgos:

Se encontró que la variable respuesta se encuentra desbalanceada haciendo que las métricas observadas para clasificación se vean afectadas, se realizaran pruebas para medir el rendimiento de los modelos con esta variación en la variable objetivo

Pruebas de modelos:

Se realizaron distintas pruebas con algunos modelos de clasificación variando los tipo de datos, estandarizaciones, balanceo y ajustando la cantidad de covariables ingresadas y los resultados dentro de los modelos fue muy buena a lo que se esperaba, los primeros modelos ocupados tuvieron como resultado un intervalo entre 0.61 y 0.68 de resultado en la curva roc sabiendo que entre mas cercano a uno el modelo es mejor, estos resultados fueron muy buenos, aun así se intento con unos árboles de clasificación con la misma variación en sus covariables se mejoraron las métricas de rendimiento teniendo como intervalo 0.73 a 0.78 de resultado en las métricas de rendimiento.

Conclusión:

Aunque se encontró muy buena cantidad de datos para el marketing del banco de Portugal y se encontraban muchas negativas de los clientes en la campaña realizada el modelo de clasificación puede ser útil para encontrar esos clientes con una buena respuesta y receptivos a las campañas que lance el banco

Referencias:

Moro (2014) Machine learning repository
http://archive.ics.uci.edu/ml/datasets/Bank+Marketing