Informe Proyecto Final Programación

Por: Camilo Rojas Henao

RO

Descripción del problema

El problema escogido se trata sobre predecir si una persona va a adquirir un producto financiero, esto sin haber tenido ningún contacto con esta persona.

Cabe decir que cuanto se dice “contacto previo”, se refiere a haber llamado o directamente haber ofrecido a una persona dicho producto.

Para estos fines se encontró el siguiente dataset:

https://archive.ics.uci.edu/dataset/222/bank+marketing

Contiene 17 columnas y 45211 filas.

Este es sobre una campaña de marketing para CDTs.

En esta campaña un banco portugués se contactaba con clientes para ofrecerles un CDT.

El dataset tiene la siguiente información:

- Age: es la edad de la persona contactada
- Job: es el tipo de trabajo que tiene la persona contactada
- Marital: el estado marital de la persona
- Education: el nivel de educación
- Default: si ha fallado en pagar algún préstamo
- Balance: el dinero promedio en la cuenta
- Housing: si tiene préstamo para casas
- Loan: si tiene algún préstamo personal
- Contact: el modo de comunicación que se ha tenido con la persona
- Day_of_week: el último día de la semana en el que se tuvo contacto con la persona.
- Month: último mes en el que se tuvo contacto con la persona

- Duration: duración de la última llamada con la persona en segundos
- Campaign: número de contactos hechos durante una campaña
- Pdays: número de días que han pasado desde el contacto por alguna otra campaña
- Previous: número de contactos previos a esta campaña
- Poutcome: el resultado de la última campaña con este cliente
- Y: si el cliente abrió o no un CDT (esta es la variable que queremos predecir)

Forma de proceder

Análisis exploratorio

Tal como la descripción del trabajo indicaba, lo primero que se hizo fue un análisis exploratorio de los datos. Este análisis se hizo sin desechar las columnas que interesaban, por ende, en análisis se hizo sobre el dataset completo.

Estadísticas básicas

Para las columnas numéricas encontramos las siguientes estadísticas:

	age	balance	day	duration	campaign	pdays	previous
count	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000
mean	40.936210	1362.272058	15.806419	258.163080	2.763841	40.197828	0.580323
std	10.618762	3044.765829	8.322476	257.527812	3.098021	100.128746	2.303441
min	18.000000	-8019.000000	1.000000	0.000000	1.000000	-1.000000	0.000000
25%	33.000000	72.000000	8.000000	103.000000	1.000000	-1.000000	0.000000
50%	39.000000	448.000000	16.000000	180.000000	2.000000	-1.000000	0.000000
75%	48.000000	1428.000000	21.000000	319.000000	3.000000	-1.000000	0.000000
max	95.000000	102127.000000	31.000000	4918.000000	63.000000	871.000000	275.000000

De estas se puede ver que la mayoría de las personas tienen un balance menor que 1428 euros, pero debe de haber algunas personas con desproporcionadamente mucho más dinero, que llevan a que la media sea 1362, esto lo podemos confirmar sabiendo que el máximo es 102127, casi 228 veces más de lo que el tope del 50 porciento de las otras personas contactadas tiene.

Si bien esta es la columna más exagerada de todas, esto se puede evidenciar también en columnas como ‘duration’, ‘pdays’ y ‘previous’, como podemos ver gracias a los cuartiles, esto se debe a que la gran mayoría de las personas contactadas no habían sido contactadas antes (cosa que se vuelve a ver en las gráficas).

Numero de nulos

```
Data columns (total 17 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        45211 non-null  int64
 1   job        45211 non-null  object
 2   marital    45211 non-null  object
 3   education  45211 non-null  object
 4   default    45211 non-null  object
 5   balance    45211 non-null  int64
 6   housing    45211 non-null  object
 7   loan       45211 non-null  object
 8   contact    45211 non-null  object
 9   day        45211 non-null  int64
 10  month      45211 non-null  object
 11  duration   45211 non-null  int64
 12  campaign   45211 non-null  int64
 13  pdays      45211 non-null  int64
 14  previous   45211 non-null  int64
 15  poutcome   45211 non-null  object
 16  y          45211 non-null  object
dtypes: int64(7), object(10)
memory usage: 5.9+ MB
None
```

Como podemos ver en esta tabla, en nuestro dataset ninguna columna tiene valores nulos.

Valores atípicos y extremos

En los gráficos de cajas y bigotes se notan que en general las columnas tienen varios valores extremos, como se mencionó previamente, los más notables son ‘balance’ y ‘previous’.

Como se mencionaba, en la columna ‘balance’ se puede ver que la gran mayoria de las personas tienen la misma cantidad de dinero, pero luego algunas personas tienen tanto (o tan poco) que aumentan desproporcionadamente la media de dinero.

En las otras columnas de ‘duration’, ‘pdays’ y ‘previous’ se puede evidenciar lo mencionado anteriormente, la gran mayoría de personas no habían sido contactadas antes; pero además, en la columna ‘previous’ se nota una anomalía, hay un valor demasiado extremo, este por su alto valor seguramente corresponde a un error de imputación, si no fuera porque no vamos a usar esta columna, ese valor tendría que ser eliminado.

Conteos:

```
Numero outliers de age es: 487
Numero outliers de balance es: 4729
Numero outliers de day es: 0
Numero outliers de duration es: 3235
Numero outliers de campaign es: 3064
Numero outliers de pdays es: 8257
Numero outliers de previous es: 8257
```


```
Numero outliers de age es: 487
Numero outliers de balance es: 4729
Numero outliers de day es: 0
Numero outliers de duration es: 3235
Numero outliers de campaign es: 3064
Numero outliers de pdays es: 8257
Numero outliers de previous es: 8257
```


En estas dos imágenes podemos ver el número de valores atípicos (y extremos) y solo extremos, respectivamente, en cada columna numérica

Histogramas

Realizamos histogramas tanto para las variables numéricas, como las categóricas.

Esto con el objetivo de revisar la distribución de las variables, y poder visualizar las cosas expuestas previamente.

Cosas notables acá son:

- a simple vista se puede ver que los trabajadores manuales, los gerentes y los técnicos representan poco más de la mitad de todos los contactados.

- Mas de la mitad de los contactados son casados.

- Mas o menos la mitad solo llegaron al bachillerato

- Casi ningún contactado ha incumplido en la deuda

- La mayoria de los contactados no tienen ningún préstamo

- Mucho más de la mitad de los contactados fueron por medio del celular

- En mayo se hacen más o menos el doble de llamadas (o contactos pues) que en el siguiente mes

- casi no se sabe ningún resultado de las campañas anteriores de marketing

- Solo el 12% (medido a ojo) de los contactos resultan en que los clientes abran un CDT en el banco

Relaciones entre las variables predictoras y el objetivo

Para las variables numéricas se hicieron histogramas para ver su relación con la variable objetivo, esto resulto en notar que las columnas ‘age’ y ‘balance’ (también ‘Day’, pero esta es en verdad más una variable categórica) que la distribución de estas en las personas que abrieron un CDT y las que no son muy similares, esto sugiere que estas dos columnas no van a ser muy buenas predictoras de si una persona va a abrir un CDT o no.

Por el contrario, las columnas ‘duration’ y ‘pdays’ si se diferencian más entre los que abrieron y los que no un CDT, indicando que estas columnas sí que serían unas buenas predictoras sobre si una persona va o no a abrir un CDT.

De igual manera para las variables numéricas también se hicieron histogramas, sin embargo, en estos no es tan fácil la interpretación como en los diagramas de cajas.

Para las columnas categóricas se hicieron histogramas sobrepuestos para cada columna, y encima de cada rectángulo se puso el porcentaje correspondiente a el número de personas con esa característica que contrataron un CDT.

Cosas notables acá son:

- los clientes que han accedido a un producto del banco anteriormente, son muchísimo más propensos a abrir un CDT.

- las personas que no tienen prestamos de bienes raíces son más del doble de probables de abrir un CDT que las personas que cuentan con estos préstamos.

- el modo de contacto no aparenta ser un factor a tener en cuenta, esto debido a que las probabilidades de abrir un CDT son muy similares en ambos modos de contacto

Preprocesado

En el preprocesado solo se quitó las columnas las cuales dependían de algún contacto previo con el banco, las columnas numéricas se estandarizaron (ya que la regresión logística se beneficia de esto) y por último transformamos las columnas categóricas a valores los cuales puedan ser procesados por los modelos.

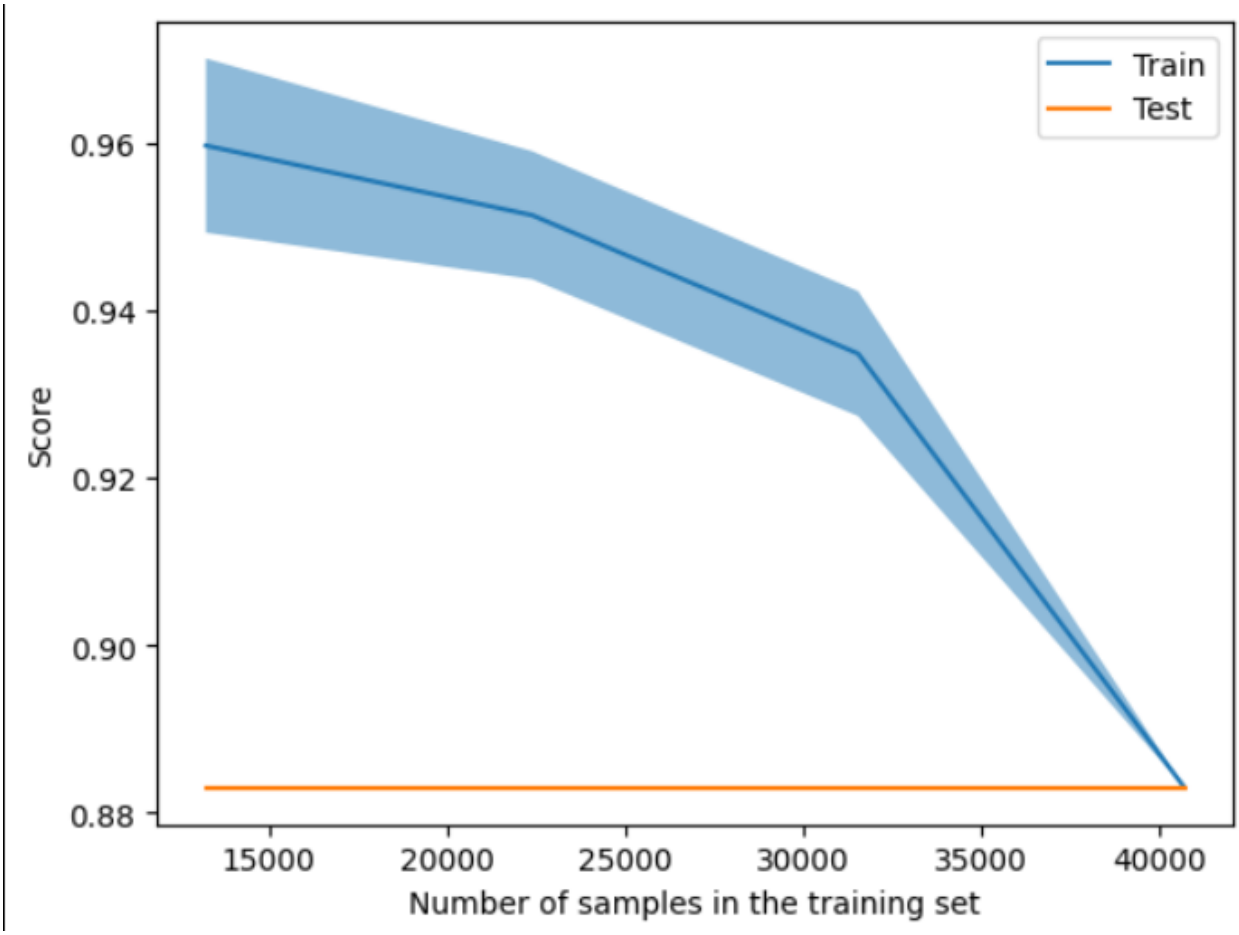
Modelos

Los modelos escogidos fueron:

- Regresión Logística
- Random Forest

A ambos modelos se les saco los parámetros óptimos mediante el uso de GridSearchCV.

El diagnostico para el modelo de regresion lineal fue el siguiente:



En la curva de aprendizaje, podemos notar que el modelo a medida que aumenta el tamaño de los datos de entrenamiento este poco a poco está generalizando más los datos, es decir, la distancia entre las curvas de aprendizaje y validación es cada vez menor, sin embargo, a su vez vemos que la curva de validación se quedó estancada en un valor, y esta no se mueve (o al menos no se puede apreciar el movimiento de esta).

Una de las primeras conclusiones que podemos sacar es que nuestro modelo no sufre de overfitting o underfitting, esto debido a que nuestras curvas de aprendizaje convergen a un valor y al final no se puede apreciar una diferencia entre estas.

El hecho de que la curva de validación no se mueva puede deberse a cosas como que el modelo es muy simple y no es capaz de aprender más sobre los datos, o que los datos que tenemos no son suficientes para poder predecir las variables adecuadamente.

Recomendaciones:

* Recolectar una mayor cantidad de datos, los cuales no dependan de contacto previo con los clientes, como puede ser, cantidad de productos financieros que ha tenido en su vida, estados financieros de personas allegadas a este, etc; esto debido a que debido a lo simple y generales que son nuestros datos finales que escogimos, estos no puedan estar dando un patrón (o tal vez puede ser el siguiente punto) muy claro, o no están dando información suficiente para poder realizar una predicción más acertada.

* Probar con otros modelos de clasificación más complejos, puede ser que el modelo de regresión logística sea demasiado simple para poder capturar los patrones que tienen nuestros datos, si esto fuera cierto, haría que no importe cuanta mayor cantidad de datos le metamos, el porcentaje de aciertos (de los datos de validación) no aumentaría.

En resumen, de las recomendaciones para este modelo, serian hacer una mayor recolección de características de los clientes e intentar usar otros modelos más complejos.

Nota: cabe recalcar que cuando se dice recolectar más características, se refiere a más columnas, no necesariamente aumentar la cantidad de filas, aunque por obvias razones, mientras más filas mejor.

El diagnostico para el modelo Random Forest fue el siguiente:

La curva de aprendizaje es prácticamente igual que la del modelo de regresión lineal, esto puede estar sugiriendo que el problema de que nuestra curva de aciertos en el set de validación este estancado, no se deba a un modelo más o menos complejo, si no que el problema se encuentre en los datos que estamos usando para entrenar nuestro modelo, datos los cuales, como ya fue mencionado previamente, son muy generales y muy pocos (en cantidad de atributos, no filas), esto puede estar resultando en que nuestros modelos en verdad no estén progresando.

Las recomendaciones para este modelo entonces se mantienen casi iguales a las que se le hicieron al modelo de regresión logística, el cual es:

* Recolectar una mayor cantidad de datos, los cuales no dependan de contacto previo con los clientes, como puede ser, cantidad de productos financieros que ha tenido en su vida, estados financieros de personas allegadas a este, etc; esto debido a que debido a lo simple y generales que son nuestros datos finales que escogimos, estos no puedan estar dando un patrón (o tal vez puede ser el siguiente punto) muy claro, o no están dando información suficiente para poder realizar una predicción más acertada.

Comparación de los modelos

Por el objetivo y contexto de los datos, debemos es de minimizar los Falsos Negativos, esto debido a que será para un banco peor el perder posibles ganancias por no haber llamado a una persona, que tener un costo marginal más bajo de contactar a una persona que el modelo incorrectamente predijo que si iba a contratar (esto porque existe la posibilidad, por más remota de que sea, que efectivamente lo haga), es decir, es mejor pecar por intentar contactar de más, que perder por no contactar.

Es decir, la métrica más importante es el `recall`.

Además, no se utilizó la curva ROC debido a que como vimos en la exploración de datos, nuestro dataset no está balanceado, esto debido a que apenas hay personas que contrataron el CDT en comparación con las que lo rechazaron.

Para hacer la comparación se utilizó la prueba de Kohen-Cappa, esto con el objetivo de ver que tan similares eran los resultados arrojados por ambos modelos, sin embargo, esta métrica resulto ser un indicativo de que los modelos no estaban funcionando; esto debido a que el resultado erar Nan, es decir, la prueba estaba teniendo problemas, y leyendo algunas advertencias arrojadas por Python, se encuentra que todas las columnas solo eran de una sola categoría, esto estaba indicando que los modelos solo estaban dando una misma respuesta (False).

Confirmando estas sospechas, se hizo un reporte de la clasificación, esto con el objetivo de ver el valor de la variable ‘recall’, la cual como fue indicado antes, es la que más nos

importa, sin embargo, se confirmó que los modelos estaban dando un único resultado, el cual es ‘False’.

Inicialmente se pensó que podría ser debido a que se estaba repartiendo mal la información de entrenamiento y de validación, sin embargo, luego de intentarlo varias veces resulto ser que simplemente los datos que se tienen no son suficientes, o más bien, no convienen la suficiente información de un individuo para poder hacer una predicción adecuadamente.

De igual manera, con el ánimo de mostrar cómo se procedería de manera normal, se realizó una matriz de confusión para cada modelo, en donde se confirmó por tercera vez que los modelos solo estaban dando un único resultado.

De esto se sacó la siguiente conclusión:

Realmente ninguno de los dos modelos está funcionando bien. Solo están diciendo que `no` a todo, lo cual no solo es erróneo, si no que hace que el nivel de Falsos Negativos sea muy alto.

Lo que se debería de hacer es mejorar la calidad de los datos, por medio de recoger datos los cuales sean más específicos, y sí que en verdad puedan representar a diferentes poblaciones con diferentes características, y de esa manera ahí si se pudiera hacer un modelo realmente útil.

Pero tal como están los datos, no se puede realizar un modelo de predicción con el cual el banco, sin ningún contacto previo con el posible cliente, pueda predecir de manera acertada (o al menos aceptable) si dicho posible cliente va a abrir un CDT o no.

Algunos de los datos que se podrían recolectar serian:

- Cuantos productos financieros a tenido la persona anteriormente (independientemente si es con el mismo banco o no)

- Ingresos de la persona (debido a que la columna `balance` se refiere es a cuanto tiene en el banco, no cuento gana)

- Lugar de residencia de la persona (ya que el contexto social puede afectar decisiones financieras).

- Flujo mensual de dinero.

- Establecimientos en los cuales gasta el dinero.

- Relacionar datos de familiares (ya que el contexto familiar influye mucho en las finanzas)

Etc.

Conclusiones

SI bien con los datos que se tienen se puede sacar un par de conclusiones sobre estos, los datos resultan ser insuficientes, esto debido a que son muy pocos y que son demasiado amplios como para poder crear un modelo de predicción efectivo (o si quiera funcional) para poder resolver el problema escogido.

Solución planteada:

Recolección de más datos, y más específicos para cada persona, es decir, recolectar datos que puedan más claramente discriminar las personas en segmentos mucho más específicos; ya que los modelos utilizados no aparentan ser el problema.

Se reitera que con más datos se refiere a más columnas, no filas, ya que, en verdad, 45211 filas es muy buena cantidad, aunque claro, mientras más mejor.