Proyecto final

Fabián Andrés Ortega Orduz, C.C.:1032507093

Diciembre 2020

1 Introducción

Un banco portugués hizo una campaña de marketing por vía telefónica para intentar convencer a las personas de suscribirse a un Bank term deposit (BKT). Durante el desarrollo de la campaña de marketing se recolectaron una serie de datos demográficos, laborales, financieros y de temporalidad de las distintas personas a las cuales se contactaron.
A partir del interés del banco de conseguir una mayor tasa de éxito en convencer a sus clientes surgen las siguientes preguntas; 1) ¿qué variables , de las recolectadas por el banco, consiguen predecir mejor la probabilidad de conseguir que un cliente dado se suscriba al (BKT), y 2) ¿qué modelo representa mejor la relación entre estas variables recolectadas por el banco y la probabilidad de éxito?
El informe esta divido en 4 partes. La primera es esta introducción donde se expuso el contexto y la pregunta de investigación. La segunda es la metodología, donde se expone brevemente la base de datos y el preprocesamiento que se realizo de los datos, además se describen las principales características de los modelos utilizados, como se seleccionaron sus hiperparametros y que criterios se usaron para la selección del mejor modelo. La tercera parte son los resultados, allí se expone cual fue el modelo seleccionado, que atributos les dio mayor importancia y una descripción de los atributos y de patrones encontrados en el análisis exploratorio. La cuarta parte son conclusiones y advertencias acerca de los resultados hallados por el modelo.

2 Metodología

La base de datos tiene en principio 45,211 observaciones y 17 atributos, 10 de los cuales corresponden a atributos categóricos y 7 a atributos numéricos, 1 atributo numérico llamado "balance" no estaba en el diccionario, por tanto se excluyo del análisis debido a que no se podía interpretar con certeza. También se excluyo del modelo una variable llamada "duration" por alta correlación con el resultado. Del total de observaciones se contaban con 39922 que correspondían con casos de fracaso de la campaña y 5289 de casos de éxito.
De los atributos categóricos fueron extraídas las distintas clases cada una en un

atributo diferente, luego de eliminar uno de los atributos que correspondía al "fracaso" de la variable principal se tienen 39 atributos. Dado el alto desequilibrio y el buen numero de observaciones de la base de datos se opto por una técnica de balanceo que se basa en la reducción de la clase dominante por medio de remuestreo, con este balanceo la base de datos quedo con 10,578 observaciones.
La selección de los hiperparametros se hizo tomando como criterio la media del score de distintos modelos variando los hiperparametros en cada iteración.
Debido a la cantidad de observaciones se decidio partir la muestra en un 80% para entrenamiento y un 20% para testeo.
Los modelos que se utilizaron, y sus características, aparecen en la siguiente tabla:

Modelo	**Características**	**Hiperparametros**
K neighbors nearest	Algoritmo de aprendizaje supervisado no parametrico basado en instancias. La clasificación se hace encontrando los k puntos más cercanos y viendo estos a que clase pertenecen.	n-neighbors $\equiv 6$
Árboles de decisión	Algoritmo de aprendizaje supervisado no parametrico que establece un esquema parecido a un juego en forma extendida, donde cada nodo representa un atributo. Los nodos son relacionados con los atributos de acuerdo a un criterio de ganancia de información dado.	criterion = "gini", max,depth = 8, max.features = 2
Random Forest	Algoritmo de aprendizaje supervisado no parametrico que se basa en la utilización de árboles de decisión. Para la clasificación de nuevos datos se toma en cuenta la decisión de clasificación de cada árbol y se realiza una votación.	criterion="gini", max,depth=10, max.features=2, n.estimators=16
Gradiant Boosting	Algoritmo de aprendizaje supervisado no parametrico se basa en una técnica secuencial para la construcción de arboles de decisión, en la que en cada rama buscara el subarbol que mayor exactitud arroje.	learning$_r ate$ = $0.0784, max.depth$ = $12, max, features$ = $3, n, estimators = 17$

Para la selección del mejor modelo se utilizo en primera instancia la curva de ROC. Luego, como segunda comprobación de los resultados se utilizo el área bajo la curva (AUC).

3 Resultados

Como se menciono, como primer criterio para la selección del mejor modelo se utilizo la curva de ROC, la cual aparece a continuación:

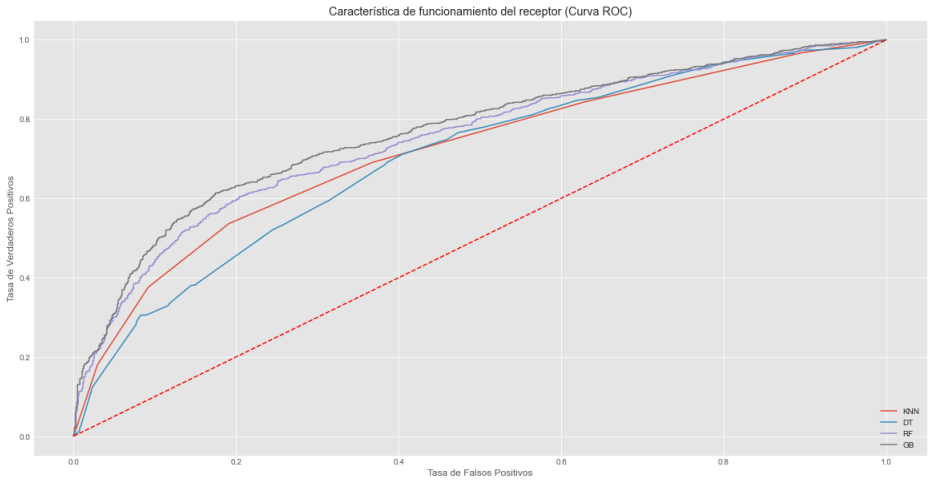

Figure 1: Curva de ROC para los modelos

Como se observa el modelo que que posee un mayor incremento en la tasa de verdaderos positivos a medida que aumenta la tasa de falsos positivos es el modelo que resulto de la aplicación del algoritmo de Gradiant Boosting. Sin embargo, como una medida de aseguramiento de que este si en el mejor modelo se presentan las medidas del AUC para los distintos modelos.

Modelo	**AUC**
K neighbors nearest	0.6726
Árboles de decisión	0.6502
Random Forest	0.6876
Gradiant Boosting	0.7042

Table 1: Valores AUC para los modelos

El AUC confirma aquello que se observo con la curva de ROC y permite ver que el poder predictivo de los modelos no están tan distantes entre si como puede hacer pensar la curva de ROC en algunos segmentos.

En base al modelo, los atributos que mayor poder predictivo se encuentran en la tabla 2.

A continuación se realiza una descripción de estos atributos y de patrones encontrados en el análisis exploratorio de los datos, realizado en el script del trabajo, que podrían explicar la importancia que el modelo de gradiant boosting les da.

- **month:** El atributo hace referencia al mes del año en el cual se hizo el ultimo contacto con el cliente, se observo en el análisis exploratorio

Atributo	Importancia
month	0.1313
age	0.1166
day	0.0935
campaign	0.0704
poutcome-success	0.0698
pdays	0.0533
contact-unknown	0.0466

Table 2: Importancia dada por el modelo a los atributos

una fuerte tendencia a tener exito en los meses de Marzo, Diciembre, Septiembre y Octubre, los cuales tuvieron mayor al 40% de éxito.

- **day:** Hace referencia al dia del mes en el cual se hizo el ultimo contacto, se observo que los dias 1 y 10 suelen tener una tasa de exito mayor al 20% .
- **age:** Hace referencia a la edad del cliente o potencial cliente, se observo que tanto los grupos de edad entre los 20 y los 30 y los mayores de 60 años tienen mayores tasas de éxito que la media.
- **campaign:** Hace referencia al numero de llamadas a este cliente durante la campaña de marketing, se observo que los clientes que fueron llamados menor cantidad de veces tuvieron mayor probabilidad de éxito.
- **pdays** Hace referencia a la cantidad de días que pasaron desde el ultimo contacto, se observo que llamadas muy frecuentes disminuian la tasa de exito pero que hay una franja que esta alrededor de los 15 a 100 dias en los cuales se tenian mayor tasa de exito.
- **poutcome-success** Hace referencia a que si se tuvo éxito con este cliente en campañas previas, se observa que los clientes que aceptaron en campañas predias tienen mayor tasa de éxito en esta campaña.

4 Conclusiones y advertencias

Los resultados exhibidos por el modelo deben ser tomados con cuidado debido principalmente a la gran perdida de información causada por el balanceo de la información, esto podría causar que algunas relaciones o todas se encuentren sesgadas y por tanto el modelo deje de ser efectivo. Dejando de lado lo anterior, el modelo puede llegar a ser útil para la toma de decisiones en el área de marketing al brindarle información acerca de que tipo de clientes deben ser su objetivo y de que manera programar la campaña de marketing.