Universidad Francisco Marroquín
Facultad de Ciencias Económicas
Curso: Machine Learning Models
Nombre: Juan Fernando González
Fecha: 02-04-2020

Reporte Bank Marketing

Introducción

El propósito de este reporte es poder predecir/clasificar si el cliente de un banco se suscribirá (sí/no) a un depósito a plazo. Esto será posible y realizado gracias a los algoritmos de árboles de decisiones y otros métodos de clasificación y selección de datos, los cuales permitirán analizar las variables, darnos *insights* valiosos de los datos. El *dataset* contiene datos e información sobre más de once mil datos de llamadas promocionales de bancos comerciales de Portugal, a sus clientes para suscribirlos a un depósito a plazo. Cada observación contiene el día, el mes, las veces que el cliente fue contactado, el balance del cliente en al banco, su edad, entre otras. La variable en la que nos vamos a centrar en este proyecto es en el depósito (*deposit*), el cual nos dice si el cliente acepto o no, valga la redundancia, el depósito a plazo.

La importancia de este reporte es el de poder analizar y sacar insights relevantes para los bancos comerciales, sobre que variables, eventos o acciones influyen más en la toma de decisión de un cliente a un depósito a plazo. Al poder clasificar a los clientes que se suscriben o no, también les será de importancia a estos bancos comerciales para segmentar de cierta manera a sus clientes, tomar decisiones acerca del tiempo y formas de comunicarse con sus clientes potenciales o clientes objetivo. Luego de aplicar los métodos de clasificación, predicción y selección de datos, se puedo observar que los bancos deben de enfocarse a hacer llamadas de forma breve, ya que, dado el árbol de decisiones, los clientes son más probables a aceptar un depósito a plazo cuando la llamada es corta. Más adelante en el reporte se explicará otros elementos a tomar en cuenta esta conclusión, así como recomendaciones de procesos, métodos y variables que pudieron ser valiosas para el desarrollo del proyecto.

Data

Algunas de las (*features* / columnas) que podremos encontrar en el *dataset* son las siguientes:

1. *Age:* Años del cliente contactado.
2. *Job:* El trabajo del cliente contactado en ese momento.
3. *Marital:* Estado civil de la persona con el cliente.

4. *Education:* Educación del cliente.
5. *Default:* Si el cliente tiene algún préstamo de crédito. (sí/no)
6. *Balance:* El balance del cliente en el banco.
7. *Housing:* ¿El cliente posee algún préstamos para pagar una casa?
8. *Loan:* ¿El cliente posee algún préstamo personal?
9. *Duration:* La duración de la llamada con el cliente.
10. *Previous:* Número de contactos realizados antes de esta campaña y para este cliente.
11. *Day:* Ultimo día en el que se contactó al cliente.
12. *Deposit:* Si el cliente decidió o no suscribirse al depósito a plazo.

Este *dataset* fue adquirido desde la página *kaggle.com,* el cual fue extraído de la página UCI. Los datos fueron obtenidos en el año 2012, y fueron procedentes de un banco situado en Portugal. Por otro lado, hay 11,162 filas (*observations*), que son los valores por cada columna en el dataset. En total hay 17 *features* y 18,297 *observations*, además, este dataset no presenta datos nulos, sin embargo, hay datos desconocidos (*Unknown*) o no relevantes para el análisis de los datos. Más adelante veremos cómo tratamos con estos valores.

Métodos

Para poder determinar que clientes acceden a suscribirse a un depósito a plazo, es necesario saber cuántos de ellos dijeron que si y cuales dijeron que no. Esto nos servirá para poder ver si es necesario realizar un balance de clases para nuestra variable a predecir, y así poder evitar que el algoritmo pierda generalización. Luego, será necesario saber cuáles podrían ser las variables candidatas para ser independientes, y a *priori,* las *features* qué apuntan a ser más relevantes es el estado civil, el balance, las veces contactadas, su educación y los años del cliente. Para verificar que mi hipótesis sobre las variables que podrían ser relevantes para mi análisis es correcta o no, utilizare una métodos estadísticos y métodos de selección de variables relevantes en base a nuestra variable independiente. Más adelante, ya teniendo las variables seleccionadas, empezaremos a implementar el algoritmo de árbol de decisión, y otros algoritmos que nos permita comparar rendimiento y el *score* de clasificación, y entonces así decidir cuál fue el mejor modelo. Con los árboles construidos, se empezará el *tuneo* de los parámetros de los métodos y algoritmos para buscar la solución óptima y en base a eso, poder entonces hacer el análisis de que hace que un cliente decida si suscribirse o no a un depósito a plazo.

Como primer punto, como ya se mencionó antes, el tratamiento del *dataset* fue más de imputación, normalización y manipulación de variables categóricas a variables

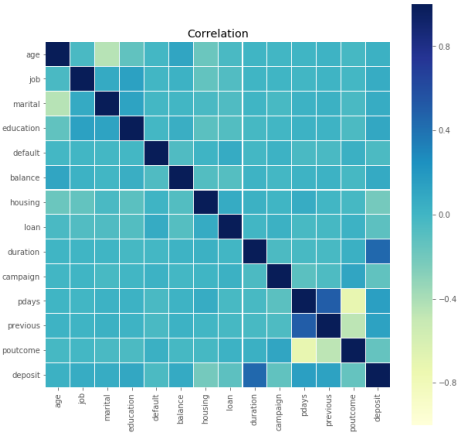
numéricas, debido a que no se presentaban datos nulos no se realizó ningún método de tratamiento de nulos. Se manipularon las variables categóricas para un mejor rendimiento en los algoritmos, luego estas fueron normalizadas y por último imputadas de modo que hubiera más interpretabilidad de los datos sin que se perdiera la esencia de los mismos. Por otro lado, se eliminaron columnas que eran irrelevantes en el análisis, y se cambiaron ciertas métricas en ciertas *features.* Luego de la limpieza de datos, se procedió a implementar los métodos y algoritmos necesarios para el análisis.

Se utilizo un diagrama de **correlación** como primer paso para identificar relación entre las variables independientes con nuestra variable dependiente. Luego de eso, se procedió a tomar todas esas variables que según su valor de correlación estaban fuertemente vinculadas con nuestra variable objetivo, estas variables fueron luego procesadas por el método de ***Feature Importance****,* lo que nos dio de *output* cuales de todas esas variables eran relevantes a partir de nuestra variable de *deposit.* Al ya tener las variables significativas, se comenzó con la implementación y la ejecución del algoritmo de **árboles de decisión**. Se probo primero con el criterio de *Gini*, con una profundidad límite de tres, luego de 4, de 6 y luego se le removió el límite. Luego, se realizó el mismo procedimiento solo que esta vez con el criterio de *Entropía*. Como ya se había mencionado antes, se utilizaron más algoritmos y métodos como el de ***Random Forest y Baggin***, para generar multiples arboles con un *subset* de features y observaciones aleatorias, para poder de cierta manera reducir la varianza y posible *overfitting* de nuestros algoritmos. Por último, se implementó ***AdaBoost Regression y Classification****,* para poder contrastar si se podía obtener mejores resultados a través del aprendizaje de los errores anteriores de predicción, y así crear un predictor final más fuerte. Los resultados que utilizaremos para comparar serán el *score* sobre los datos de entrenamiento y los datos de prueba, así como también el curva de ROC.

Resultados

A continuación, se presentarán los resultados obtenidos de los métodos aplicados al *dataset,* así como también el análisis y deducciones del mismo. Los primeros resultados que obtuve fueron sacados de la información que me proporcionaban el diagrama de correlación:

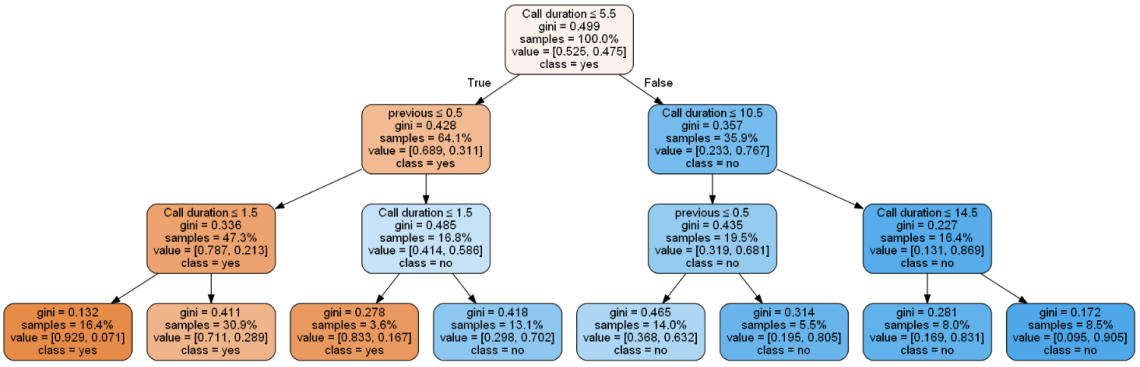
Correlación de la variables del dataset:



Como se puede observar, las variables que parecen estar fuertemente relacionas con la variable independiente son las siguientes: *Previous, pdays, duration, balance, education y marital*. Estas variables luego fueron utilizadas para ***Feature Importance***. El cual, dado al output, solo se escogieron las variables, *previous, duration, marital y education.*

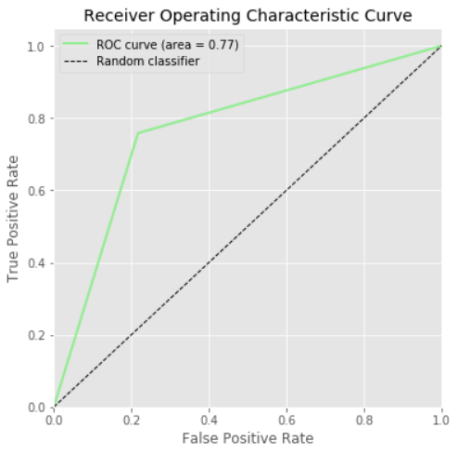
Luego, se procedió a realizar el algoritmo de árbol de decisión.

Árbol de decisión: *Gini, profundidad 3*



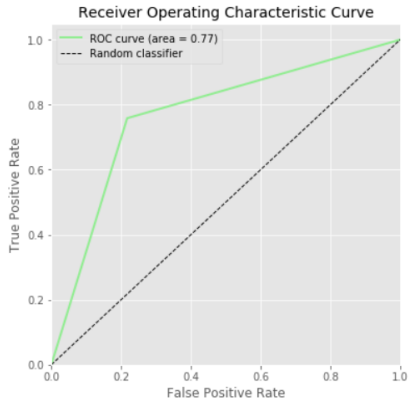
Dado el primer árbol de decisión, pude concluir que entre más breve sea la llamada del agente bancario con el cliente final, es más probable que el cliente decida suscribirse al depósito a plazo. Por otro lado, se obtuvieron unos resultados de precisión de *0.77* en los datos de entrenamiento y un *0.76* en los datos de entrenamiento. Luego se realizaron pruebas con diferentes profundidades de árbol, pero este solo mejoraba en los datos de entrenamiento, por lo que al seguir aumentando la profundidad podía caer en *overfitting*.

Baggin: *Bootstrap, 500 estimadores*



En este caso, se utilizó *Baggin* con *Bootstrap* para generalizar mejor el modelo y utilizar las *features* más fuertes, sin embargo, el *ROC curve* nos dio u resultado casi igual a el árbol de decisión normal, *0.77*.

Random Forest: *Bootstrap, 50 muestras máximo, 3 hojas nodo máximo*



El resultado que se obtuvo del *Random Forest* no fue distinto a los demás métodos y algoritmos. Se obtuvo un *ROC curve area* de *0.77,* y al aumentar el número de hojas y muestras, básicamente se quedaba constante.

Conclusiones

Luego del proceso de exploración y análisis de los datos ya antes mencionados, puedo concluir y responder a la pregunta de la introducción. Los bancos comerciales tendrán una mayor probabilidad de que los clientes se suscriban a depósitos a plazo entre más breve sea la llamada, o bien, que las llamadas fueron breves y que el cliente haya sido contactado al menos más de una vez. Sin embargo, dado a que los *scores* de los algoritmos no fueron los mejores, también llegue a la conclusión que pudo haber habido variables más significativas no agregadas al *dataset* que posiblemente nos darían una mejor idea y visualización de que es lo que realmente afecta la decisión del cliente, es

decir, como la tasa de interés ofrecida, la hora de la llamada, o el largo del plazo ofrecido. Y por último, también note una de las variables seleccionadas pudo haber afectado nuestro análisis y predicciones, como lo es la duración de la llamada, ya que a pesar de que la mayoría de clientes que aceptaron el depósito a plazo fueron llamadas breves, hubieron persona que aceptaron el depósito a plazo con una duración de llamada bastante más tardada, y esto nos da la pauta que puede que el cliente pudo haber estado ya enterado de la campaña acerca del depósito, y que por eso no fue tan extensa la llamada, y que hay clientes que también estaban interesados, pero que tal vez necesitaban más información o incluso, podrían haber estado negociando, como lo mencione antes, una tasa preferencial, plazos, etc.
(S. Moro, 2012) (Brid, 2018) (Chan, 2020) (Hershy, 2019)

Apéndice

Feature Importance:

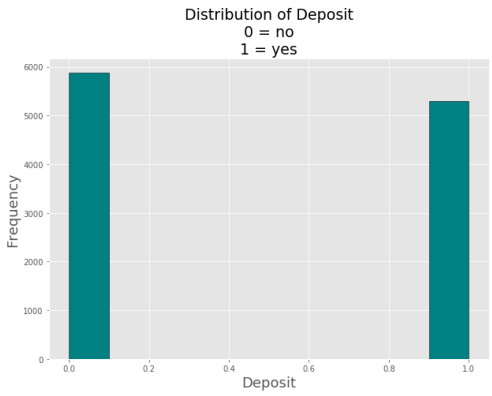
Variables ingresadas:

	previous	duration	job	marital	pdays	education	balance
0	0	17	0	1	-1	1	2343
1	0	24	0	1	-1	1	45
2	0	23	9	1	-1	1	1270
3	0	9	7	1	-1	1	2476
4	0	11	0	1	-1	2	184

Variables Seleccionadas por el método:

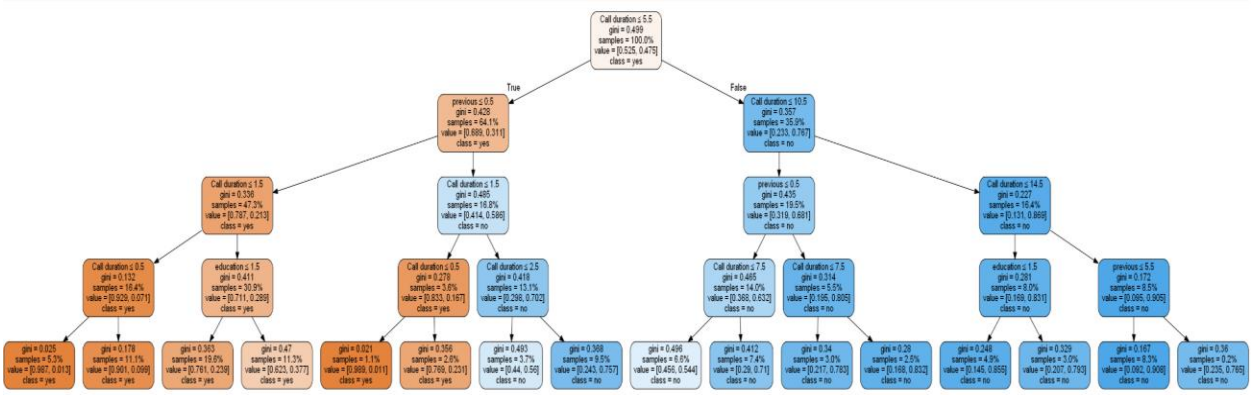
```
[1 1 2 1 3 1 4]
array([ True,  True, False,  True, False,  True, False])
```

Distribución de variable dependiente (deposit):



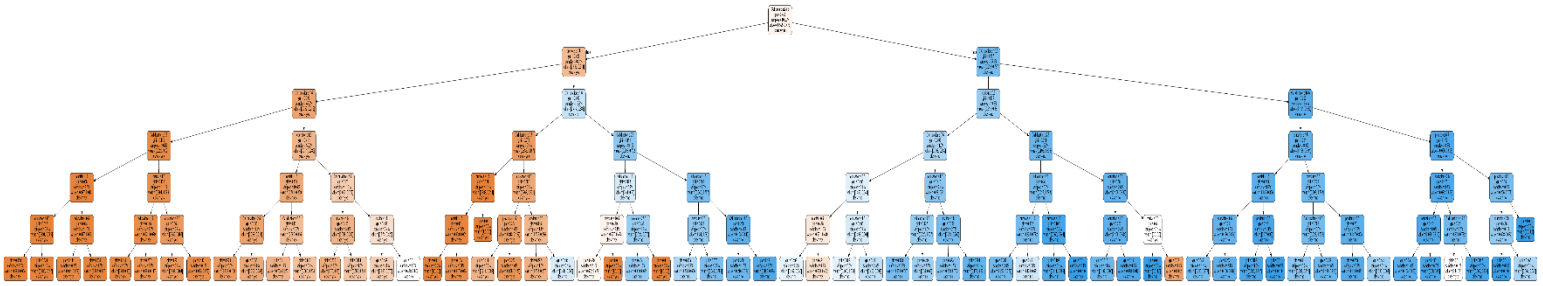
Clientes no suscritos: 52.6%
Clientes suscritos: 47.4%

Árbol de decisión: Profundidad 4



Training Score: 0.77
Testing Score: 0.76

Árbol de decisión: Profundidad 6



Training Score: 0.77
Testing Score: 0.77

Referencias

Brid, R. S. (25 de 10 de 2018). *Medium*. Obtenido de https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb

Brownlee, J. (11 de 05 de 2018). *Machine Learning Maestry*. Obtenido de https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/

Chan, C. (2020). *DisplayR*. Obtenido de https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb

Creech, S. (2019). *Statistically SIgnificant* . Obtenido de https://www.statisticallysignificantconsulting.com/RegressionAnalysis.htm

Hershy, A. (10 de 07 de 2019). *Medium*. Obtenido de Gini Index vs Information Entropy: https://towardsdatascience.com/gini-index-vs-information-entropy-7a7e4fed3fcb

Magiya, J. (16 de 06 de 2019). *Medium*. Obtenido de Medium: https://towardsdatascience.com/kendall-rank-correlation-explained-dee01d99c535

S. Moro, P. C. (14 de 02 de 2012). *Bank Marketing Data Set*. Obtenido de Bank Marketing Data Set: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing