UNIVERSIDAD FRANCISCO MARROQUÍN
FACULTAD DE CIENCIAS ECONÓMICAS
DATA SCIENCE FOR FINANCE
MARIA ISABEL AVILA RIGALT

CASO PRÁCTICO NO. 4

PREDICCIÓN DE PÉRDIDA DE CLIENTES CON BOSQUE ALEATORIO

Instrucciones generales:

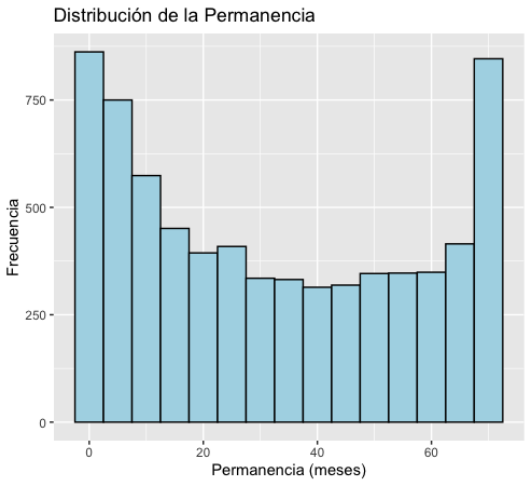
Usted trabaja para una empresa de telecomunicaciones que ha perdido muchos clientes en los últimos meses. Su tarea es analizar la cartera de clientes y determinar posibles razones por las cuáles se han perdido (datos en perdida_clientes.csv).

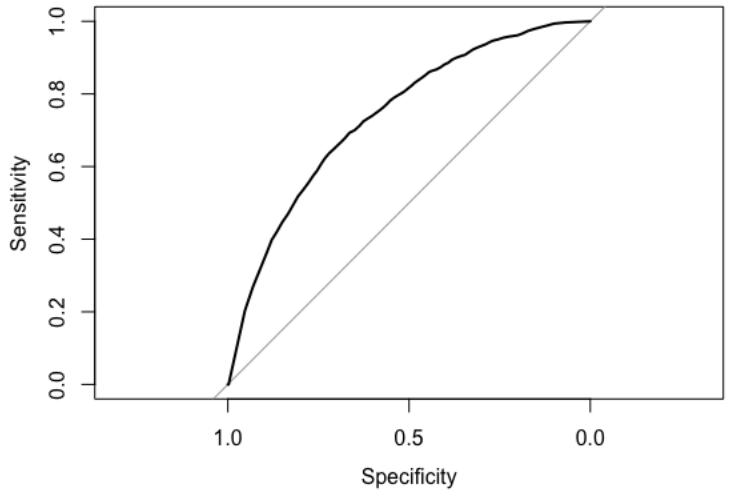
Entregables:

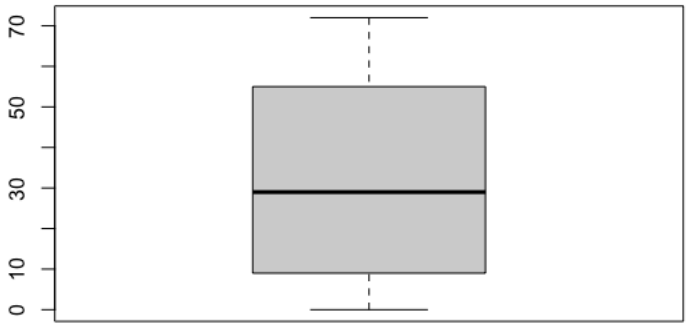
a) 1 código de R (o markdown) con su procedimiento
b) 1 Word (o markdown) con su trabajo escrito.

Requerimientos específicos:

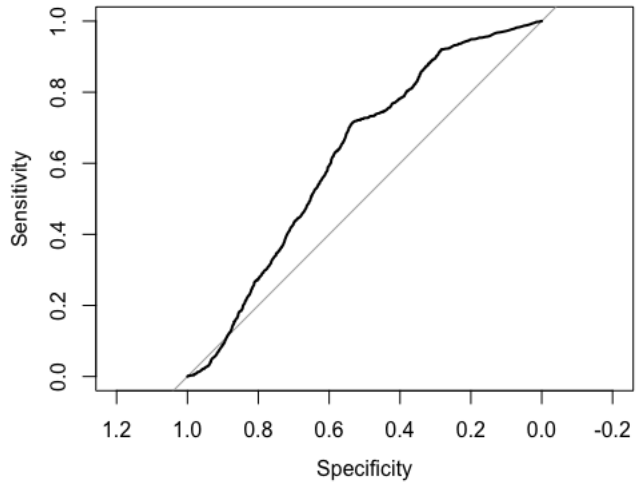
a. Explorar las variables.
   - Variables numericas:
     - Pemanencia: Tiene una media de 32 meses que un clientes ha estado con la empresa. No tiene outliers que afecten al modelo. Ahora bien su area bajo la curva esta entre 72%-75% por lo que decimos que tiene buena capaidad de prediccion.

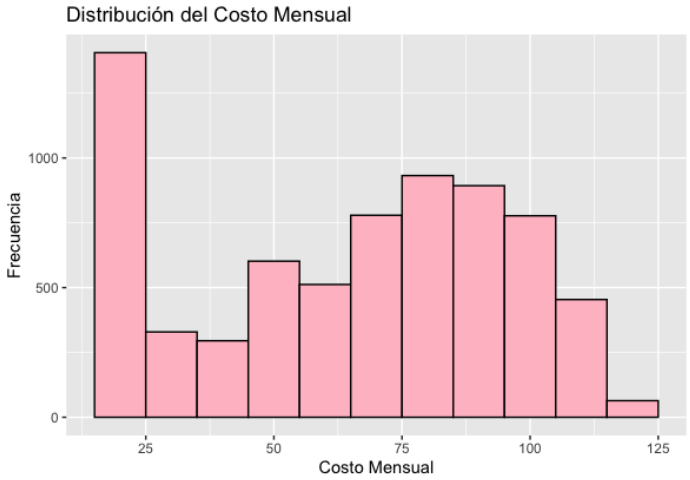


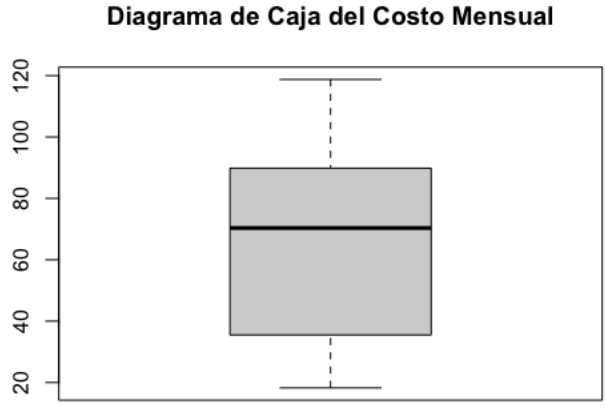




- Costo Mensual: Este costo tiene un promedio de 64.76, su auc es de un aproximado de 60%-64% por lo que tambien podria ser considerado un buen predictor, sin embargo no el mejor. En esta variable tampoco se encontro outliers que fueran a perjudicar el modelo.

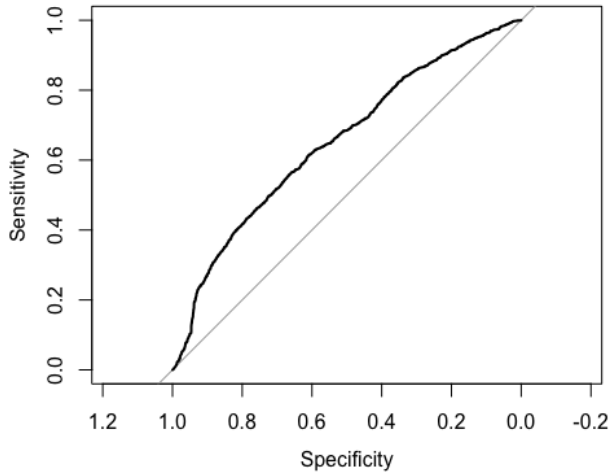


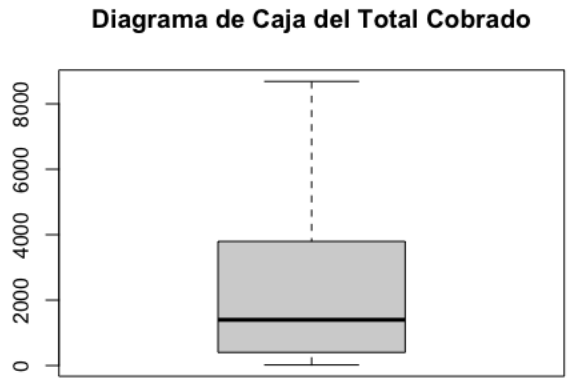


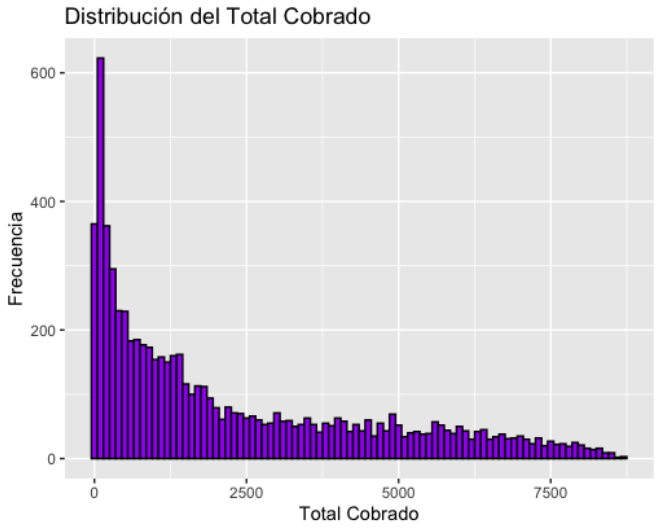


- Total cobrado: El promedio del total que se cobra durante su tiempo de permanencia es de 2,283.30, en esta variable se encontro que tiene un sesgo

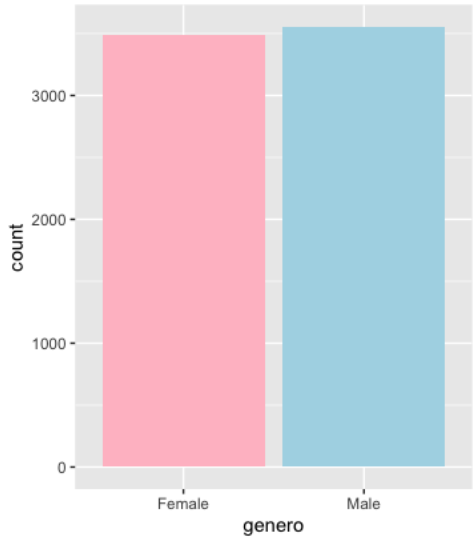
positivo. A pesar de al hacer un boxplot no se encontraron outliers. Su AUC fue de 64%-67% lo cual otra vez nos dice que es buena predictora.



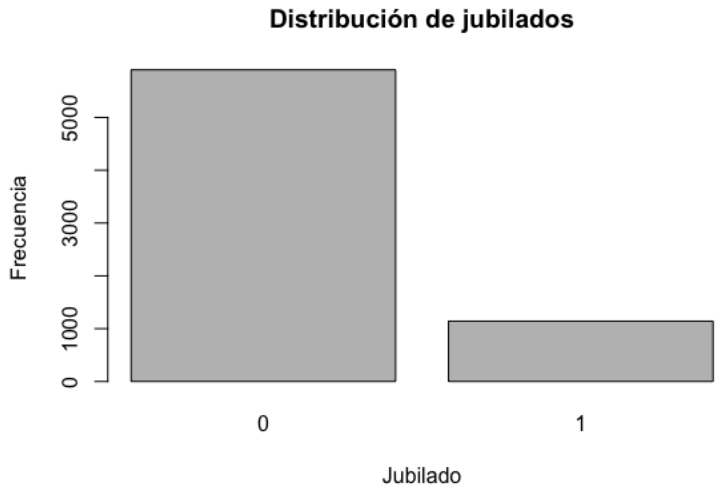




- Variables categoricas:
  - Genero: un 49.47% son mujeres y el 50.53% son hombres. Dentro de la categoria de mujeres hay un 73.1% que aun tienen el servicio mientras que un 26.9% ya no tienen el servicio. Por el otro lado, para los hombres son muy parecidas las cifras con un 73.7% que aun tienen el servicio y un 26.2% que ya no cuentan con dicho servicio. El valor p es de 0.4698, lo que nos dice que no hay evidencia para decir que el genero y perdida estan relacionadas entres si, por lo tanto no es buena predictora.



  - Jubilados: Hay 5,901 que no son jubilados y 1,142 personas que son jubiladas. Dentro de los que no son jubiladas el 76.4% sigue con el servicio mientras que de las jubilidas solo un 58.3% sigue con el servicio. Ahora bien, su valor p es bajo siendo de: 9.478e-37 lo que nos dice que hay asosiciaion entre las variables por lo que es buena predictora.



  - Pareja: De los clientes, los que no tiene pareja, un 67% sigue con el servicio y un 33% ya no cuenta con el servicio. Por otra parte, los individuos con pareja

que siguen con el servicio es un aprox. de 80.3%. El valor p es de 1.519e-36 lo que es pequeño por lo que asumimos que es buen predictor.

- Contrato: Hay tres tipos de contrato, mes a mes, por un año y por dos años. De los de mes a mes, un 57.29% siguen con el servicio, de los de un año, un 88.73% siguen con el servicio, mientras que los de dos años tiene la mayor retención con un 97.17%. Esta variable tiene un valor p muy pequeño de: 5.863e-258 lo que indica que hay una asociación entre contrato y perdida. Esto nos dice que puede ser buen predictor.
- Protección: De los que no tenían protección y se quedaron es de 60.87% y el 39.13% si lo perdió. Esta variable también tiene un valor p muy chiquito de: 5.505e-122 por lo que si hay una relación entre las variables y esto nos puede ayudar al predecir.
- Internet: Según la prueba nos dice que el tipo de internet que tienen los clientes influencia la perdida de clientes, si tienen un DSL el 81.04% se queda, si es fibra óptica se queda solo el 58.11% y si no tienen internet se queda el 92.60%. Ahora bien, este también tiene un valor p bajo: 9.572e-160. Esto nos permite rechazar la hipótesis nula, y asumir que están relacionadas y que esta ayudara a la predicción.
- Método de pago: El valor p de esta es de: 3.682e-140, aunque no es tan bajo como las anteriores también sirve para predecir. Ahora bien, en base a los resultados de la tabla, los clientes que usan pago automático, ya sea bank transfer o credit card tienen una proporción más alta de quedarse.
- Backup online: los clientes que no tienen backup tienden a irse a diferencia con los que sí lo tienen. El valor p: 2.08e-131, hay relación entre las variables y también puede predecir.
- Soporte: Los que no tienen el servicio, el 58.36% se queda con el servicio. El valor p: 1.443e-180, el valor es bajo por lo que asumimos que es buen predictor.
- Streaming TV: los clientes que no tienen streaming TV tienden a irse a diferencia de los que sí tienen. El valor p: 5.529e-82. Hay una relación entre streaming TV y perdida. Es decir, esta variable es buena predictora.
- Streaming películas: Los que no tienen el servicio tienen una proporción más alta de personas que se van a diferencia de aquellos que sí lo tienen. Ahora bien, los que no tienen el servicio de streaming son más que aquellos que tienen el servicio. El valor p: 2.668e-82, se rechaza la hipótesis nula, lo que indica que puede servir para predecir.
- Servicio tel: Este tiene un valor p relativamente alto de 0.3162 por lo que puede que no esté relacionada con la perdida de clientes, por lo que no es buena predictora.
- Múltiples líneas: El valor p es bajo: 0.003464 por lo que es buena predictora. Los clientes que no tienen múltiples líneas tienden a irse más que los que sí tienen varias líneas.

- Seguridad online: Los clientes que no tienen seguridad tienden a irse más que los que sí la tienen. El valor p es de: 2.661e-185, es buena predictora.

b. Utilizar uno o más modelos de bosque aleatorio para predecir la variable de pérdida.
   - Se utilizaron 4 random forest:
     - En el primero se utilzaron 70 arboles, con un node size de 10 y todas las variables.
     - En el segundo se uilizaron 30 arboles, con un node size de 4 y se le quito el servicio tel y el genero.
     - En el tercero se utilizaron 50 arboles y un node size de 10 y se le quitaron las variables de servicio tel, genero, dependientes, pareja, multiples lineas, proteccion y metodo de pago.
     - En el cuarto se utilizaron 60 arboles con un nodesize de 6 y se le quito el servicio tel y el genero.

c. Interpretar sus resultados y métricas.
   - Modelo rf:
     - Recall: 0.64
     - Precision: 0.52
     - Accuracy: 0.79
     - F1 Score: 0.58
   - Modelo rf2:
     - Recall: 0.63
     - Precision: 0.54
     - Accuracy: 0.79
     - F1 Score: 0.58
   - Modelo rf3:
     - Recall: 0.63
     - Precision: 0.53
     - Accuracy: 0.79
     - F1 Score: 0.58
   - Modelo rf4:
     - Recall: 0.63
     - Precision: 0.54
     - Accuracy: 0.79
     - F1 Score: 0.58
   - Todos tienen mediciones parecidas, pero tecnicamente el mejor fue el primer bosque con todas las variables. Luego estan iguales el cuarto y segundo bosque. En el primer arbol como se menciono arriba utilizo 70 arboles, con una selección de 5 variables al azar en cada división. La tasa de error del modelo es de 20.54% en los datos que no se utilizaron durante el entrenamiento. Luego en la matriz de confusion muestra que se clasificaron correctamente 3252 clientes como no y 661 clientes como si, mientras que se equivocaron al clasificar 381 clientes No que era Si y 630

clientes como Si que eran No. Por lo tanto, las tasa de error en la clasificaion total es del 48.80%.

- Ahora bien, el recall es la proporcion de casos positivos que fueron correctamente indentificados, precision es los casos positivos correctamente clasificados entre todos los casos clasificados como positivos por el modelo, accuracy es la proporcion de casos clasificados correctamente en general y el F1 es el equilibrio entre precision y recall. Ahora bien, en este caso donde se quieren prevenir las perdads entonces es clave el recall. Es decir, es importante identificar y retener los clientes que estan en riesgo de irse. Sin embargo, estas metricas puede que estan bajas porque los bosques tienden a ser más sensibles al overfitting.

d. Dar al menos tres recomendaciones para disminuir la pérdida de clientes.
    - Mejorar la seguridad y calidad del servicio de internet: ya que se vio que esta variable era importante, ver de mejorar la conexión, es decir que sea rapida, estable y confiable, siempre y cuando cumpla con medidas de seguirdad para proteger la privacidad de los clientes.
    - Ofrecer precios competitivos y que aporten valor, ya que el costo mensual es clave. Tal vez se puede dar ofertas, paquetes o sorteos.
    - Mejorar el servicio de soporte porque esta tambien es siginifictiva, entonces ver de cómo mejorar el soporte tecnico y atencion al cliente, que este sea eficiente y receptivo. Que no solo sean capacitados sino que sean amables para resolver las consultas.

Diccionario de variables

Genero – categórica

Jubilado – categórica, si el cliente está jubilado

Pareja – categórica, si el cliente tiene pareja

Dependientes – categórica, si el cliente tiene dependientes

Permanencia – numérica, meses que el cliente ha estado con la empresa

Servicio_tel – categórica, si el cliente tiene contratado servicio telefónico

Múltiples_lineas – categórica, si el cliente tiene múltiples líneas contratadas

Internet – categórica, tipo de internet que el cliente tiene contratado

Seguridad_online - categórica, si el cliente tiene contratado el servicio de seguridad en línea

Backup_online – categórica, si el cliente tiene contratado el servicio de backup en línea

Protección – categórica, si el cliente tiene contratado el servicio de protección de dispositivos

Soporte – categórica, si el cliente tiene contratado el servicio de soporte técnico

Streaming_TV – categórica, si el cliente tiene contratado el servicio de streaming tv

Streaming_movies – categórica, si el cliente tiene contratado el servicio de películas en streaming

Contrato – categórica, tipo de contrato

Factura_electronica – categórica, si el cliente pide sus facturas en línea

Método_pago – categórica

Costo_mensual – numérica, costo mensual del servicio

Total_cobrado – numérica, total cobrado al cliente durante su permanencia con la empresa

Perdida – categórica, si el cliente se perdió o no