Examen Parcial II
Econometría Avanzada
Profesora: Ana Laura Camacho
Estudiante: David Mora
Universidad de Costa Rica

I.

Página web del dataset elegido: [Bank Customer Churn Dataset | Kaggle]

a. Regresión Logística

a. ¿Cuál es la proporción de la variable respuesta en datos_train y datos_test?(Utilice una separación 0.75 train, 0.25 test al generar la partición con caret::createDataPartition)

Para datos_train se tiene:

```
        NO        SI
0.7962938 0.2037062
```

Para datos_test se tiene:

```
        NO        SI
0.7963185 0.2036815
```

b. ¿Cuál es el tamaño de cada uno de los sub sets mencionados en la pregunta anterior?

2499 observaciones para los datos de test.
7501 observaciones para los datos de train.

c. ¿Cuál es la métrica utilizada para elegir el modelo final que usted especificó en control Train?

La métrica a utilizar para este modelo de clasificación es “Accuracy”.

d. ¿Cuál es el valor del kappa y accuracy para cada una de las repeticiones de la primera partición (k=1 o Fold 1)?

Accuracy	Kappa	Resample
0.8160000	0.2552671	Fold01.Rep1
0.8064085	0.1778141	Fold01.Rep2
0.8066667	0.2047299	Fold01.Rep3
0.8160000	0.2563052	Fold01.Rep4
0.8149134	0.2581425	Fold01.Rep5

e. ¿Cuál es el accuracy y kappa del modelo final? (recuerde que esto no es lo mismo que el accuracy y kappa obtenidos al generar la matriz de confusión del siguiente paso)

Accuracy	Kappa
0.8075452	0.2120388

f. Utilice las predicciones raw para calcular y mostrar los resultados de una matriz de confusión con caret:confusionMatrix, utilizando como referencia el set de datos test que separó al inicio del ejercicio. Siga los ejemplos vistos en clase.

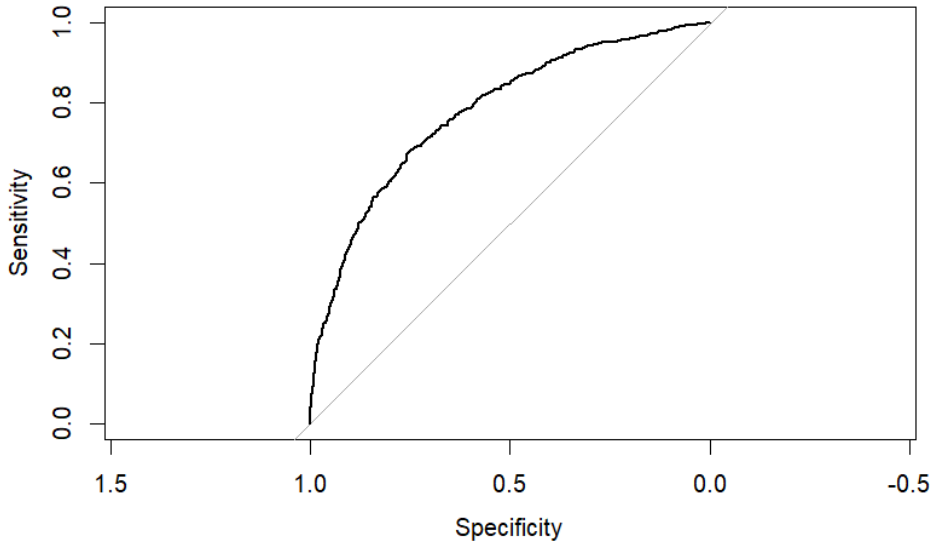
```
Confusion Matrix and Statistics

          Reference
Prediction   NO   SI
        NO 1936  397
        SI   54  112
```

Indique cuáles son las 2 métricas más importantes para su problema de clasificación y explique por qué, dentro del contexto de su problema de clasificación.

Para el contexto de este problema, cuyo interés es predecir la probabilidad de abandonar al Banco, es importante conocer la métrica de sensibilidad y la métrica de precisión. La primera me mostraría la capacidad de predecir a las personas que realmente abandonaron el banco, de manera que yo pueda conocer y clasificar las características de esas personas que salieron y usarlo como predictor para próximos clientes. La segunda asigna el valor de la precisión que tiene el modelo en predecir a esas personas que se esperan que iban a abandonar el Banco, de manera que uno se construya una expectativa de cuánto nivel de error soportar cuando se obtienen datos del futuro de los predichos por el algoritmo, así uno asigna más intención de generar una política empresarial enfocada a los que poseen mucha mayor probabilidad.

g. Utilice las predicciones tipo “prob” para generar una curva ROC y calcular el AUC.

Area under the curve: 0.7789



b. K vecinos más cercanos

a. ¿Cuál es la proporción de la variable respuesta en datos_train y datos_test?(Utilice una separación 0.75 train, 0.25 test al generar la partición con caret::createDataPartition)

Para datos_train se tiene:

```
        NO        SI
0.7962938 0.2037062
```

Para datos_test se tiene:

```
        NO        SI
0.7963185 0.2036815
```

b. ¿Cuál es el tamaño de cada uno de los sub sets mencionados en la pregunta anterior?

2499 observaciones para los datos de test.
7501 observaciones para los datos de train.

c. ¿Cuál es la métrica utilizada para elegir el modelo final que usted especificó en control Train?

La métrica a utilizar para este modelo de clasificación es “Accuracy”.

d. ¿Cuál es el valor del kappa y accuracy para cada una de las repeticiones de la primera partición (k=1 o Fold 1)?

Accuracy	Kappa	Resample
0.8266667	0.3374648	Fold01.Rep1
0.8160000	0.3050800	Fold01.Rep2
0.8266667	0.3492451	Fold01.Rep3
0.8437917	0.4222241	Fold01.Rep4
0.8226667	0.3589249	Fold01.Rep5

e. ¿Cuál es el accuracy y kappa del modelo final? (recuerde que esto no es lo mismo que el accuracy y kappa obtenidos al generar la matriz de confusión del siguiente paso)

K*	Accuracy	Kappa
10	0.8075452	0.2120388

f. Utilice las predicciones raw para calcular y mostrar los resultados de una matriz de confusión con caret:confusionMatrix, utilizando como referencia el set de datos test que separó al inicio del ejercicio. Siga los ejemplos vistos en clase.

```
Confusion Matrix and Statistics

          Reference
Prediction   NO   SI
        NO 1920  359
        SI   70  150
```

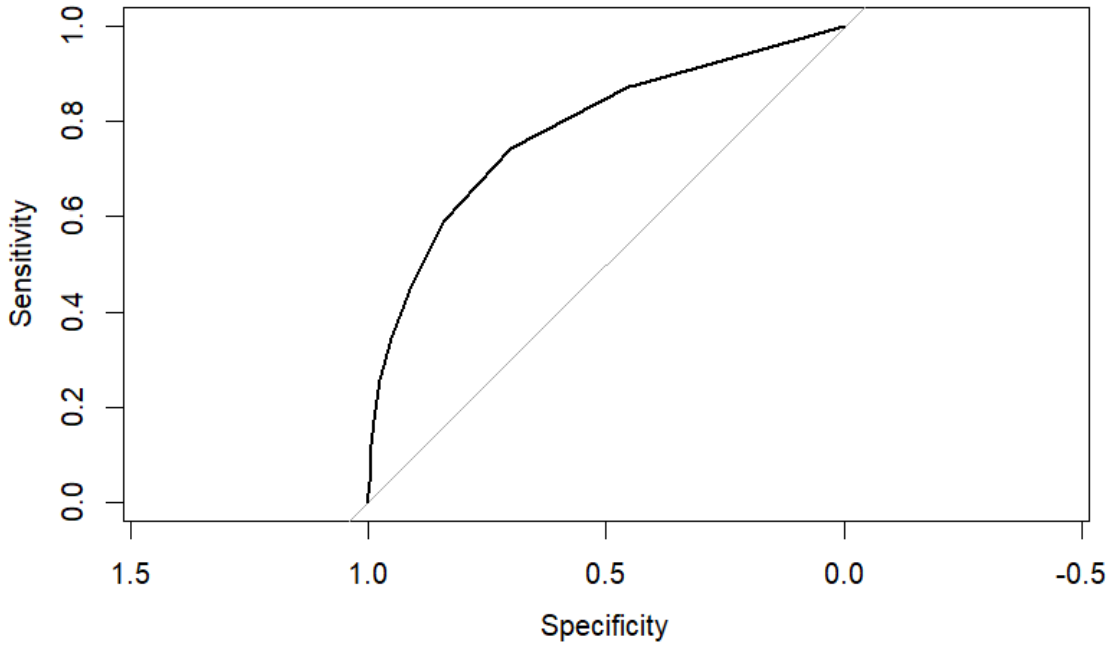
Indique cuáles son las 2 métricas más importantes para su problema de clasificación y explique por qué, dentro del contexto de su problema de clasificación.

Para el contexto de este problema, cuyo interés es predecir la probabilidad de abandonar al Banco, es importante conocer la métrica de sensibilidad y la métrica de precisión. La primera me mostraría la capacidad de predecir a las personas que realmente abandonaron el banco, de manera que yo pueda conocer y clasificar las características de esas personas que salieron y usarlo como predictor para próximos clientes. La segunda asigna el valor de la precisión que tiene el

modelo en predecir a esas personas que se esperan que iban a abandonar el Banco, de manera que uno se construya una expectativa de cuánto nivel de error soportar cuando se obtienen datos del futuro de los predichos por el algoritmo, así uno asigna más intención de generar una política empresarial enfocada a los que poseen mucha mayor probabilidad.

g. Utilice las predicciones tipo “prob” para generar una curva ROC y calcular el AUC.

Area under the curve: 0.7822



II.

a. Haga una comparación entre los resultados de los modelos que generó, tomando como base la matriz de confusión y el AUC. Su respuesta debe ser de al menos 5 líneas y máximo 15 líneas. (4 puntos)

El proceso de clasificación entre ambos modelos parte de criterios teóricos diferentes, el primero mediante el uso de una estimación, el segundo realiza un análisis espacial de probabilidad. Bajo esta lógica, el proceso de predicción, bajo los mismos parámetros de validación cruzada, obtiene la capacidad del modelo para ajustar a los nuevos datos, y estos se digitan en la matriz de confusión. Para el interés de este investigador, es importante notar que el modelo RNN mejora significativamente la predicción de sí, es, decir, la sensibilidad, de manera que pasó de predecir 112 a 150, por lo que también redujo su error de falsos negativos, lo cual, es un mejor indicio. La importancia de la predicción de los verdaderos negativos y falsos positivos es menor para esta investigación.

Finalmente, el AUC muestra un mejor ajuste para el modelo espacial KNN, esto es especial para elegir como método predictor el KNN, pues este índice se puede interpretar como la probabilidad de que un clasificador ordene o puntúe dos observaciones con dos clasificaciones diferentes de forma correcta.

b. Responda brevemente: (8 puntos)

i. ¿Cuáles son los vectores de soporte de un clasificador de soporte vectorial?

En un clasificador de soporte vectorial, solamente las observaciones que se posicionan justamente sobre el margen o que se encuentran en el lado incorrecto del margen afectarán al hiperplano, y, por tanto, al clasificador obtenido. A estas observaciones se les llama “vectores de soporte” y estos vectores si afectan al clasificador de soporte vectorial.

Cuando C es muy grande, el margen es grande y, por tanto, habrá más vectores de soporte. En contraste, si C es pequeño, habrá menores vectores de soporte y, por tanto, se esperará menores sesgos y mayores varianzas.

ii. ¿Qué significa el C en una máquina de soporte vectorial?

Partiendo del problema que intenta resolver los clasificadores de soporte vectorial en la que no queda más remedio que admitir errores en la clasificación de algunos datos de entrenamiento que van a estar en el lado equivocado del hiperplano, aparece el parámetro C.

C es un parámetro de ajuste que podemos elegir a través de la validación cruzada, este valor limita la cantidad de los errores que podemos mantener en el lado equivocado del hiperplano. Cc controla el conocido trade-off de sesgo-variancia. C determina el número y severidad de las violaciones al margen y al hiper plano que se van a tolerar.

En términos del balance entre el sesgo y la varianza: valores pequeños de C van a dar lugar a modelos muy complejos, con mucha varianza y poco sesgo (con el consiguiente riesgo de sobreajuste); y valores grandes a modelos con mucho sesgo y poca varianza.

iii. En un modelo knn, ¿Cuál de estos dos valores (5 o 20) para K (k vecinos, no particiones) obtenemos una frontera de decisión más flexible?

Con un k de 5 obtenemos una frontera más flexible, de manera que se vea más como un camino de muchas vueltas o más “quebrado” que un camino lineal. Mientras que un valor de 20 si obliga a obtener menor flexibilidad.

iv. Explique, en palabras sencillas, cuál es la diferencia entre odd y probabilidad. (Puede ilustrarlo con un ejemplo como: odd vs probabilidad de que una persona de 100 gane el primer lugar)

Odds es el ratio (razón, fracción, proporción) de que ocurra un escenario respecto a que no ocurra un escenario. Si Odds es mayor a 1, entonces existen más ocurrencias que no ocurrencias, si es menor a 1, imperan las no ocurrencias.

La probabilidad indica el grado de que ocurra un escenario respecto a todos los escenarios.

En un ejemplo: imagine que usted realizó un examen de matemáticas de entre 100 personas, y el profesor califica el examen entre quienes hicieron buena calificación vs los que no. El profesor desea obtener el odd, entonces toma la cantidad de personas que hicieron buena calificación y lo divide entre los que no, de manera que posea por cuánto se multiplica una persona que saca buena calificación respecto a quien no.

La probabilidad muestra cuánto representa de los estudiantes totales, los estudiantes con buena calificación, en otras palabras, cuántos estudiantes en promedio del total de la población, va a clasificarse en la clase de buena calificación.