**Pontificia Universidad Javeriana  
Octubre 27 de 2023**

**Reglas del examen: Se pueden consultar materiales e Internet, pero está completamente prohibida la comunicación entre estudiantes o con personas externas durante el examen, de forma presencial o virtual. Favor responder a continuación de la pregunta. Una vez diligenciado el cuestionario deben guardarlo en formato PDF y subirlo como respuesta a la asignación en Brightspace. La hora máxima de entrega es a las 5:00 pm.**

**NOMBRE:**

**Caso 1:** (15 puntos) La Oficina del Censo de EE. UU. ha publicado los datos del censo del estado de California donde existen mediciones acerca de las características y distribución de los diferentes hogares en el estado. La anterior información está segmentada por distritos los cuales poseen un volumen de habitantes de entre 600 a 3000 personas. Unos inversionistas están interesados en clasificar aquellos distritos en los cuales el precio promedio de la vivienda sea mayor a los US$200.000 para determinar posibles oportunidades de negocios. Se requiere hacer un ejercicio de clasificación para la variable binaria (greater200): 1 – el precio promedio de las viviendas por distrito es mayor o igual a los US$200.000 y 0 – en caso contrario. Para tal fin, se cuenta con las potenciales variables predictoras:

* housing\_median\_age: Edad promedio (en años) de la casa en el distrito.
* total\_rooms: Recuento del número total de habitaciones (excluyendo dormitorios) en todas las casas del distrito.
* total\_bedrooms: Recuento del número total de dormitorios en todas las casas del distrito.
* population: Recuento del número total de habitantes en el distrito.
* households: Recuento del número total de hogares en el distrito.
* median\_income: Mediana del ingreso familiar (en dólares) total de todas las casas en el distrito.

Una inspección inicial del conjunto de datos indica que la proporción de la variable de respuesta para 20433 distritos es

Texto

Descripción generada automáticamente con confianza baja

Para el ejercicio de clasificación, segmentamos el conjunto de datos en 80% entrenamiento y el restante para prueba. Se decide aplicar un modelo de regresión logística al conjunto de entrenamiento por medio de la función glm de R con la familia binomial, tomando como predictoras todas las variables antes mencionadas. Seguido, se aplica la metodología Stepwise al modelo inicial, cuyo resultado se observa a continuación

Tabla

Descripción generada automáticamente

1. (5) Tomando como probabilidad base la proporción de casas con precio mayor o igual a US$200.000, determine la probabilidad final que se obtiene al mostrar un aumento en 1000 al número de habitantes en el distrito suponiendo que las demás variables no tienen ningún cambio.

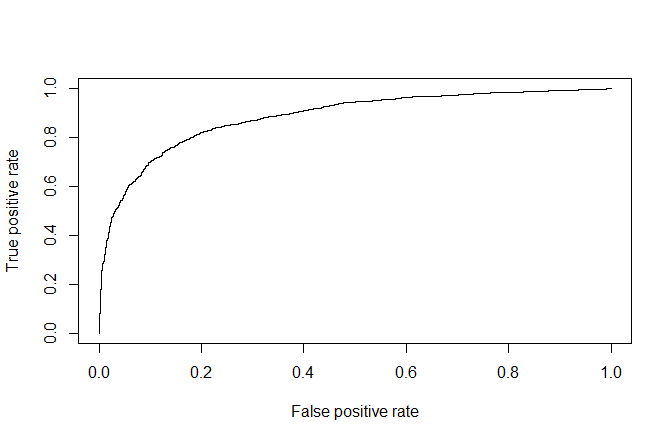
**Rta.**

Tomando un cutoff de 0.45, se obtiene la siguiente matriz de confusión en los datos de prueba

Imagen que contiene Texto

Descripción generada automáticamente

Finalmente, el AUC del modelo es de 0.884 y la curva ROC se muestra a continuación



1. (4) Basado en el cutoff de 0.45, señale con un círculo rojo de manera aproximada en la curva COR el resultado de dicho punto de corte e indique las coordenadas exactas de dicho punto. Analícelas.

**Rta.**

1. (6) Si el punto de corte se disminuye a 0.3,

¿qué ocurriría con la precisión y la exhaustividad, y cómo afectaría ello el uso del modelo según lo que se pretende en el negocio?

**Rta.**

¿afectaría el cambio de punto de corte el valor de AUC? Si lo afecta, ¿lo aumenta o disminuye? Si no lo afecta , ¿por qué?

**Rta.**

**Caso 2:** (15 puntos) Se quiere predecir la probabilidad de compra de un producto a partir de 11 predictores (10 escalares y 1 binario), para lo cual se tiene una base de datos de 7779 personas que ya decidieron si comprar o no el producto. El balance de la base de datos es 58% de sí (personas que compran) y 42% de no (personas que no compran).

La base de datos se partió en 80% para entrenamiento y 20% para prueba, balanceando el entrenamiento con undersampling.

Se ejecutó el algoritmo k-nn para realizar la predicción. Se escogió inicialmente y empleando un punto de corte de 0.5 para la probabilidad de compra, se obtuvo la siguiente matriz de confusión en la base de prueba

|  |  |  |
| --- | --- | --- |
|  | **Reference** | |
| **Prediction** | **0** | **1** |
| **0** | 502 | 672 |
| **1** | 149 | 1248 |

1. (7) Se desea realizar una campaña específica entre aquellos que el modelo predice como potenciales compradores, y el costo de la campaña se estima en $120.000 por potencial comprador. El potencial ingreso por cada éxito es de $150.000. Estime la ganancia (o pérdida) esperada si se piensa aplicar la campaña en una base de datos con 5000 personas.

**Rta.**

Tras efectuar un modelo de knn, con validación cruzada y 10 particiones (folds), y haciendo una búsqueda del parámetro k entre 1 y 50, se obtuvieron los valores anexos para la métrica de AUC (archivo aucknn\_2330.xlsx).

1. (8) Escoja un valor de k adecuado para el modelo, y mencione un valor de k con posible *overfitting* y un valor de k con posible *underfitting*. Apóyese para ello en un análisis de los datos entregados de la métrica de AUC.

**Rta.**