Profesor: Dr. Oldemar Rodríguez Rojas
Minería de Datos 1
Análisis Exploratorio y Métodos de Regresión
Fecha de Entrega: Jueves 23 de marzo - 8am

Tarea Número 2

- Las tareas serán revisadas en clase, no pueden ser enviadas por correo.
- Quienes no se presenten a la revisión de la tarea tendrán un cero de nota.
- Las tareas son estrictamente de carácter individual, tareas idénticas se les asignará cero puntos.
- Todas las tareas tienen el mismo valor en la nota final del curso, es decir, el promedio de las notas obtenidas en la tareas será la nota final del curso.
- Todos los ejercicios tienen el mismo valor.
- **Pregunta 1:** Suponga que trabajamos para un banco y se nos pide predecir el monto promedio de deuda en tarjeta de crédito de una cartera de clientes relativamente nuevos, basado en otra cartera de comportamiento y estructura similar de la cual sí se tiene información de deuda en tarjeta de crédito. En este ejercicio hacemos uso de la tabla de datos `DeudaCredito.csv` que contiene información de los clientes en una de las principales carteras de crédito del banco, e incluye variables que describen cada cliente tanto dentro del banco como fuera de éste.

  Esta tabla de datos contiene 400 clientes y 11 variables que los describen. Seguidamente se explican las variables que conforman la tabla.

  - `Ingreso`: Ingreso del cliente, en miles de dólares.
  - `Limite`: Límite de crédito global en tarjetas de crédito del cliente.
  - `CalifCredit`: Calificación crediticia del cliente.
  - `Tarjetas`: Cantidad de tarjetas de crédito del cliente.
  - `Edad`: Edad del cliente.
  - `Educacion`: Años de educación del cliente.
  - `Genero`: Género del cliente.
  - `Estudiante`: Indica si el cliente es estudiante o no.
  - `Casado`: Indica si el cliente es casado o no (1 = Sí, 0 = No).
  - `Etnicidad`: Indica si el cliente es caucásico, afroamericano o asiático.
  - `Balance`: Monto promedio de deuda en tarjeta de crédito del cliente, en dólares.

  Realice lo siguiente:

  1. Cargue la tabla de datos en `R` y asegúrese que las variables se están leyendo de forma correcta. ¿Es necesario recodificar variables? Seleccione la variable a predecir, y tome para entrenamiento un 80 % de la tabla de datos.

2. Realice un Análisis Exploratorio de los datos con todos los datos que incluya al menos:
   - `summary(...)`.
   - La matriz de correlaciones y alguno de sus gráficos. Interprete al menos dos correlaciones.
   - Encuentre al menos 3 datos atípicos, si es que existen en esta tabla de datos.
3. Basado en las estadísticas básicas explique cuál variable numérica parece ser la mejor para predecir la deuda en tarjeta de crédito.
4. Genere un modelo de regresión lineal múltiple incluyendo las todas las variables predictoras. ¿Cuáles coeficientes obtiene para los β? Dé una interpretación de 3 de los coeficientes que se obtienen en el modelo. ¿Cuál variable parece tener más impacto sobre la variable a predecir y por qué?
5. ¿Qué error se obtiene sobre la tabla de testing para el modelo de regresión lineal? Interprete las medidas de error obtenidas.
6. Si tuviera que eliminar alguna o algunas de las variables con la esperanza de que mejore la predicción ¿Cuál o cuáles de las variables eliminaría? ¿El nuevo modelo mejora la predicción?

- **Pregunta 2:** Un cliente nos contrata para estudiar una posible oportunidad de negocio, y para ver si le es rentable quiere una predicción de las ventas potenciales de asientos de niños para autos en su tienda. Para ello hacemos uso de los datos `AsientosNinno.csv` los cual contienen detalles de ventas de asientos de niños para auto en una serie de tiendas similares a las del cliente, y además los datos incluyen variables que definen características de la tienda y su localidad.

  La tabla de datos está formada por 400 filas y 11 columnas. Seguidamente se explican las variables que conforman la tabla.

  - `Ventas`: Ventas de asientos de niños para autos en cada localidad, en miles de unidades.
  - `PrecioCompt`: Precio promedio por asiento de niño cobrado por la competencia en cada localidad.
  - `Ingreso`: Nivel de ingreso promedio de los habitantes de la región, en miles de dólares.
  - `Publicidad`: Presupuesto que asigna cada tienda a publicidad, en miles de dólares.
  - `Poblacion`: Tamaño de la población en cada región, en miles.
  - `Precio`: Precio cobrado por la tienda por los asientos de niño para auto.
  - `CalidadEstant`: Indica la calidad de ubicación de los asientos de niño en los estantes de la tienda.
  - `Edad`: Edad promedio de los habitantes de la localidad.
  - `Educacion`: Años de educación promedio de los habitantes de cada región.
  - `Urbano`: Indica si la tienda está localizada en una zona urbana o no (1 = Sí, 0 = No).
  - `USA`: Indica si la tienda está ubicada en Estados Unidos o no (1 = Sí, 0 = No).

  Realice lo siguiente:

1. Cargue la tabla de datos en `R`. En caso de ser necesario, recodificar las variables de forma adecuada. Seleccione la variable a predecir, y para medir el error tome un 15 % de la tabla de datos.
2. Realice un Análisis Exploratorio de los datos con todos los datos que incluya al menos:
   - `summary(...)`.
   - La matriz de correlaciones y alguno de sus gráficos. Interprete al menos dos correlaciones.
   - Encuentre al menos 3 datos atípicos, si es que existen en esta tabla de datos.
3. Aplique el modelo de regresión lineal múltiple incluyendo todas las variables predictoras.
4. ¿Qué error se obtiene sobre la tabla de training para el modelo generado anteriormente?

- **Pregunta 3:** La Tabla de Datos `uscrime.csv` contiene el cálculo de índice de crímenes violentos por habitante en Estados Unidos, como son el asesinato, la violación, el robo y asalto. Las variables incluidas son, entre otras, el porcentaje de la población considerada urbana, la renta media de la familia, la participación de las fuerzas del orden, el número de policías per cápita, el porcentaje de los oficiales asignados a las unidades de la droga. La variable a predecir es `ViolentCrimesPerPop` (Per Capita Violent Crimes in US). Usando un 67 % de esta tabla para Tabla de Aprendizaje y el restante 33 % para Tabla de Testing efectue lo siguiente:

  1. Realice un Análisis Exploratorio de los datos con todos los datos.
  2. Construya un modelo predictivo para la variable `ViolentCrimesPerPop` usando una Regresión Lineal Múltiple con la función `lm(...)` en la Tabla de Aprendizaje y calcule *Error Estándar de los Residuos* para este modelo, además calcule el *Error Cuadrático Medio y el Error Relativo* para la Tabla de Testing.

- **Pregunta Optativa:** [25 puntos extra]

  1. Programe en **R** una función `lm2(...)` que recibe como parámetro una tabla de aprendizaje y retorna un modelo de Regresión Lineal, es decir, calcula y retorna $\beta = (X^t X)^{-1} X^t y$.
  2. Programe en **R** una función `predict2(...)` que recibe como parámetro el modelo construido en la pregunta anterior, una tabla de testing de modo tal que retorna la predicción para esta tabla de testing.
  3. Usando la tabla de datos `uscrime.csv` compare los resultados de las funciones `lm(...)`, `lm2(...)`, `predict(...)` y `predict2(...)`.
  4. Usando la tabla de datos `uscrime.csv` y la función de **R** denominada `system.time(...)` compare los tiempos de ejecución de las funciones `lm(...)`, `lm2(...)`, `predict(...)` y `predict2(...)`.

Entregables: Debe entregar un documento autreproducible HTML con todos los códigos y salidas, incluya pruebas de ejecución de las funciones programadas. No olvide poner un título para cada pregunta.