Deusto

Predicción de ventas según el marketing

Aprendizaje automático basado en Regresión Lineal

Descripción del problema

- ¿Podemos predecir las ventas que conseguiremos según el dinero invertido en diferentes canales de publicidad?
  - Para ello, utilizaremos un dataset con inversiones de marketing y ventas, disponible en la plataforma **Kaggle** (origen).
  - El objetivo es determinar las ventas que podemos obtener a partir de una serie de inversiones en publicidad.

Descripción del Data Set

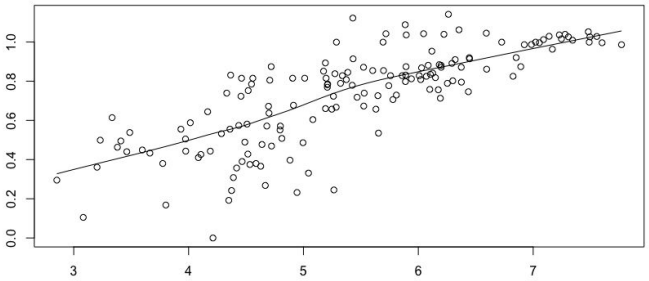
- id: identificador del registro
- TV: inversión de publicidad en TV (en miles de $)
- Radio: inversión de publicidad en radio (en miles de $)
- Newspaper: inversión de publicidad en periódicos (en miles de $)
- **Sales:** ventas conseguidas (en miles de $)

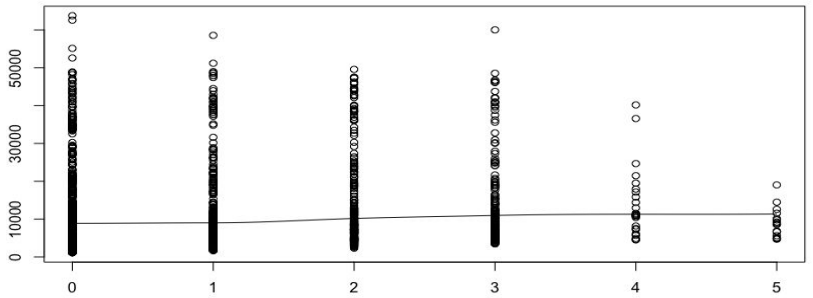
Se pide

- Elabora un script que:
  - Procese el dataset y lo adapte → Si algún valor es ‘null’ o vacío, sustituirlo por el valor más común en su columna;
  - Genere 10 modelos de regresión lineal, calcule los errores de cada uno (MAE, MSE y MAPE) e indique cuál es “el mejor”;
  - Muestre los coeficientes de cada atributo del mejor modelo;
  - Identifique la muestra con mayor error absoluto generado por el modelo;
  - Responda a las cuestiones planteadas.

Analiza los datos - Dependencia lineal

- Visualiza la dependencia lineal entre las variables
  - Usa la función scatter.smooth:
    - `scatter.smooth(x=data$x, y=data$y, main="X vs. Y")`





Analiza los datos - Correlación

- Calcula la correlación que existe entre cada atributo y el objetivo
  - Usa la función cor: `cor(data$y, data$x)`
    - Valores próximos a 1 o -1 indican la existencia de buena correlación
    - Baja correlación = -0.2 < x < 0.2

Crea el modelo de regresión lineal

- Divide el dataset en entrenamiento (75%) y test (25%)
  - Utiliza la función createDataPartition
  - `createDataPartition(y=data$Target, p=0.7, list=FALSE)`
- Entrena el modelo de regresión lineal
  - Utiliza la función lm // `lm(formula = ¿?, data = training_data)`
    - `formula=target~.` (usa todos los atributos del data.frame)
    - `formula=target~Att-1+Att-2+...+Att-N` (usa sólo algún atributo)
- Valida el modelo: `predict(model, test_data)`
- Calcula el error medio absoluto del modelo (Mean Absolute Error)
  - `mean(abs(prediction - test_data$Target))`

Analiza el modelo de regresión lineal

- Obtén los coeficientes del modelo de regresión lineal
  - Usa la función print: `print(model)`

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          686.852    527.290   1.303 0.193865
Batting_average    -2315.678   4078.164  -0.568 0.570647
On.base_percentage   100.318   4008.288   0.025 0.980052
Runs                  -1.191      7.640  -0.156 0.876229
Hits                  10.155      4.386   2.315 0.021392 *
Doubles               -6.892     11.883  -0.580 0.562407
Triples               -2.780     28.990  -0.096 0.923677
HomeRuns              61.650     17.682   3.487 0.000575 ***
Runs_batted_in         7.922      7.121   1.112 0.266956
Walks                 10.171      7.038   1.445 0.149645
Strike.Outs          -15.066      3.007  -5.010 1.01e-06 ***
Stolen_bases          13.091      6.489   2.017 0.044686 *
Errors               -19.927     10.020  -1.989 0.047791 *
```

El número de asteriscos que posee cada atributo (en la columna derecha de la vista de los coeficientes) indica la significatividad: cuantos más asteríticos, el atributo es más significativo.

Se consideran significativos valores inferiores a 0.05.

Analiza el modelo de regresión lineal

- Confirma si el modelo tiene significación estadística (p-value)
  - Usa la función print: `print(model)`

```
Residual standard error: 865.4 on 259 degrees of freedom
Multiple R-squared:  0.5647,    Adjusted R-squared:  0.5445
F-statistic:    28 on 12 and 259 DF,  p-value: < 2.2e-16
```

  - El valor debe ser inferior o igual a 0,05.
- Tras analizar el modelo
  - ¿Tiene sentido utilizar todas las variables del dataset?
  - Haz pruebas eliminando del modelo las variables que tienen menor relevancia

Responde a las siguientes preguntas

- Usando el mejor modelo, añade el código que permita responder a las siguientes cuestiones (justifica en la documentación cómo respondes a cada cuestión):
  - Identifica las 10 campañas con más beneficio: aquellas con mayor relación ventas/inversión.
  - Identifica las 10 campañas más ‘sorprendentes’:
    - Las 5 con mayor diferencia ‘positiva’ entre el valor real y la predicción
    - Las 5 con mayor diferencia ‘negativa’ entre el valor real y la predicción

Responde a las siguientes preguntas

- Usando el mejor modelo, añade el código que permita responder a las siguientes cuestiones (justifica en la documentación cómo respondes a cada cuestión):
  - Dado un presupuesto de 50.000$, y en pasos de 5.000$, usa el modelo para identificar la forma óptima de distribuir la inversión:
    - Cada canal debe tener un mínimo de 5.000$ en inversión. Por ejemplo:
      - Opción 1 → 5.000$ TV, 5.000$ Radio, 40.000$ Periódico
      - Opción 2 → 5.000$ TV, 10.000$ Radio, 35.000$ Periódico
      - …
      - Opción X → 40.000$ TV, 5.000$ Radio, 5.000$ Periódico
    - Recuerda que las cifras del dataset están en miles de $

Entrega

- Entrega el trabajo a través de la **tarea de ALUD**
  - Fecha de entrega: 12 de mayo
  - Completa el cuestionario de tiempos y dificultades
  - Formato: fichero .ZIP
- Evaluación - 10%
  - Ejecución correcta y libre de errores - 5%
  - Documentación de análisis de los resultados - 5%
- Esfuerzo individual
  - 20h. por persona