**Lab 10**

|  |  |
| --- | --- |
| Name: | 劉邦彬 |
| Student ID: | R11631033 |
| Total Score: |  |

1. **Multiple Choice (35 points, 5 points each question)**

|  |  |  |  |
| --- | --- | --- | --- |
| # | Answer | Explanation (Optional) | Score |
| 1 | (a) |  |  |
| 2 | (c) |  |  |
| 3 | (b) |  |  |
| 4 | (b) |  |  |
| 5 | (c) |  |  |
| 6 | (c) |  |  |
| 7 | (e) |  |  |

1. **Simple Linear Regression (30 points, 10 points each question)**

|  |  |  |
| --- | --- | --- |
| # | Description | Score |
| 1 | Result: |  |
| 2 | Result: |  |
| 3 | Result: |  |

1. **Multiple and Polynomial Regression (35 points)**

|  |  |  |
| --- | --- | --- |
| # | Description | Score |
| 1 | import numpy as np  import matplotlib.pyplot as plt  import pandas as pd  from sklearn.linear\_model import LinearRegression  from sklearn.model\_selection import train\_test\_split  from sklearn.preprocessing import StandardScaler  df = pd.read\_csv("epldata\_final.csv")  df["age2"] = df["age"] \*\* 2  df["log\_page\_views"] = np.log2(df["page\_views"])  feature\_names = [      "fpl\_points",      "age",      "age2",      "log\_page\_views",      "new\_signing",      "big\_club",      "position\_cat",  ]  X = df[feature\_names].to\_numpy()  y = df["market\_value"].to\_numpy()  # df.fillna(0, inplace=True)  df.fillna(method="ffill", inplace=True)  X\_train, X\_test, y\_train, y\_test = train\_test\_split(      X, y, test\_size=0.2, random\_state=42, stratify=df["region"]  )  scaler = StandardScaler()  # X\_train\_scaled = scaler.fit\_transform(X\_train)  # X\_test\_scaled = scaler.transform(X\_test)  reg = LinearRegression(fit\_intercept=True)  reg.fit(X\_train, y\_train)  y\_pred = reg.predict(X\_test)  print("\nR^2: \t", reg.score(X\_test, y\_test))  from sklearn.metrics import mean\_squared\_error, mean\_absolute\_error  print("\n mean\_squared\_error : ", mean\_squared\_error(y\_test, y\_pred))  print("\n mean\_absolute\_error : ", mean\_absolute\_error(y\_test, y\_pred))  print("\ncoef: \n", reg.coef\_)  print("\ninter: \t", reg.intercept\_)  print(f"Coef of age and age^2 is {reg.coef\_[1]:.5f} and {reg.coef\_[2]:.5f}")  print(f"Coef of log2(page\_views) is {reg.coef\_[3]:.5f}")  print(f"Coef of big\_club is {reg.coef\_[5]:.5f}")  from sklearn.model\_selection import learning\_curve  train\_size\_abs, train\_scores, test\_scores = learning\_curve( reg, X, y,cv=10 ,train\_sizes=[0.1,0.25,0.5,0.75,1] ,scoring='neg\_mean\_squared\_error')  for train\_size, train\_scores, test\_scores in zip(train\_size\_abs, train\_scores, test\_scores):      print(f"{train\_size} samples were used to train the model")      print(f"The average train mse is {-train\_scores.mean():.2f}")      print(f"The average test mse is {-test\_scores.mean():.2f}") |  |
| 2 | 2-1    2-2  表示page\_view以及big\_club呈正相關，雖然age和market\_value呈正相關但age2卻呈現負相關表明年齡對marketvalue的影響可能是非線性的    2-3  球員應該加入big\_club等，參考係數和maeket\_value為正相關之行為去提升其市值。基於模型，球員應該提升其page\_view\*2^10倍以提升其市值10倍  2-4      Training\_data隨著訓練樣本增加，其MSE隨之下降，但是test\_data會因為訓練樣本增加而有波動，故模型可能有overfitting現象 |  |