Module 4: Introduction to Machine Learning

Assignment

Module 4: Assignment

Linear Regression:

Analyze the information given in the Video_games dataset and predict the values using linear regression model.

The dataset can be loaded using this link:

Video_game_Data

Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
Wii Sports	Wii	2006	Sports	Nintendo	41.36	28.96	3.77	8.45	82.53
Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
Mario Kart Wii	Wii	2008	Racing	Nintendo	15.68	12.76	3.79	3.29	35.52
Wii Sports Resort	Wii	2009	Sports	Nintendo	15.61	10.93	3.28	2.95	32.77
Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37
Tetris	GB	1989	Puzzle	Nintendo	23.20	2.26	4.22	0.58	30.26
New Super Mario Bros.	DS	2006	Platform	Nintendo	11.28	9.14	6.50	2.88	29.80
Wii Play	Wii	2006	Misc	Nintendo	13.96	9.18	2.93	2.84	28.92
New Super Mario Bros. Wii	Wii	2009	Platform	Nintendo	14.44	6.94	4.70	2.24	28.32
Duck Hunt	NES	1984	Shooter	Nintendo	26.93	0.63	0.28	0.47	28.31
Nintendogs	DS	2005	Simulation	Nintendo	9.05	10.95	1.93	2.74	24.67
Mario Kart DS	DS	2005	Racing	Nintendo	9.71	7.47	4.13	1.90	23.21
Pokemon Gold/Pokemon Silver	GB	1999	Role-Playing	Nintendo	9.00	6.18	7.20	0.71	23.10
Wii Fit	Wii	2007	Sports	Nintendo	8.92	8.03	3.60	2.15	22.70
Kinect Adventures!	X360	2010	Misc	Microsoft Game Studios	15.00	4.89	0.24	1.69	21.81
Wii Fit Plus	Wii	2009	Sports	Nintendo	9.01	8.49	2.53	1.77	21.79
Grand Theft Auto V	PS3	2013	Action	Take-Two Interactive	7.02	9.09	0.98	3.96	21.04
Grand Theft Auto: San Andreas	PS2	2004	Action	Take-Two Interactive	9.43	0.40	0.41	10.57	20.81
Super Mario World	SNES	1990	Platform	Nintendo	12.78	3.75	3.54	0.55	20.61

The description of the attributes in the dataset are as follows:

- Name of the video game – text, each row depicts name of the video game
- Platform – platform on which game runs.
- NA_Sales – numeric variable, sales in North America for the respective games.
- EU_Sales – numeric variable, sales in Europe for the respective games.
- JP_Sales – numeric variable, sales in Japan for the respective games.
- Other_Sales – numeric variable, aggregate sum of sales in other parts of the world, for the respective games.

Perform the following tasks on the dataset

Task 1: Using the sales columns of our data

→ Create a test set and training set

→ Create a new linear regression model using train set.

→ Predict the values for other sales for the test

Task 2: Plot the values predicted by our model and the actual values to check deviation between them

→ Create a subset of 100 values so that you can see the plot clearly

→ Plot the actual values from the test set in black

→ And plot the predicted values in red so that we can differentiate between them clearly.

Logistic Regression:

Analyze the information given in the Employee_Data dataset and predict the values using logistic regression model.

The dataset can be loaded using this link:

https://www.edureka.co/medias/fdzs9gyrw1/download?media_file_id=189951110

srnumber	Edu_of_Emp	Edu_Cat	marital_Status	Occ_Of_Emp	Emp_rel_status	Emp_race_type	sex_of_emp	capital_gain	capital_loss	Work_hour_in_week	country_of_res	Emp_Sal
77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	India	>50K
122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	?	>50K
245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	0	45	Mexico	<=50K
176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0	40	United-States	<=50K
28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K

The description of the attributes in the dataset are as follows:

- Edu_of_Emp – gives the education level of the employees
- marital_satus – gives whether one is single, married, separated, divorced, or widowed.
- Occ_of_Emp – give the occupation of employees like manager, cleaner, professor etc.
- Emp_rel_status – gives the relation status of the employee like Husband, wife, not in family.
- Work_hour_in_week – gives the number of hours the employee worked in a week.
- Emp_sal – gives the salary range, >50k or<=50k

Perform the following tasks on the dataset

Task 1: Using the Emp_sal column of our data create a logistic regression model

→ Create a test set and training set

→ Build a model using all other columns and predict the values.

→ Create a confusion matrix with cut off value as 0.4