Topic: Support Vector Machine (SVM)

Instructions

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable.

Please ensure you update all the details:

Name: Nukala Ayyappa Bharthwaj

Batch Id: DSWDMCOS 21012022

Topic: Support Vector Machines.

1. Business Problem

1.1. Objective

1.2. Constraints (if any)

2. Work on each feature of the dataset to create a data dictionary as displayed in the below image:

Name of Feature	Description	Type	Relevance
ID	Customer ID	Quantitative, Nominal	Irrelevant, ID does not provide useful information

2.1 Make a table as shown above and provide information about the features such as its Data type and its relevance to the model building, if not relevant provide reasons and provide description of the feature.

Using R and Python codes perform:

3. Data Pre-processing

3.1 Data Cleaning, Feature Engineering, etc.

3.2 Outlier Imputation

4. Exploratory Data Analysis (EDA):

4.1. Summary

4.2. Univariate analysis

4.3. Bivariate analysis

5. Model Building

5.1 Build the model on the scaled data (try multiple options)

5.2 Perform Support Vector Machines.

5.3 Train and Test the data and compare accuracies by Confusion Matrix and use

different Hyper Parameters

5.4 Briefly explain the model output in the documentation

6. Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided

Note:

The assignment should be submitted in the following format:

- **R code**
- **Python code**
- **Code Modularization should be maintained**
- **Documentation of the model building (elaborating on steps mentioned above)**

Problem Statement: -

A construction firm wants to develop a suburban locality with new infrastructure but they are faced with a challenge of incurring losses if they cannot sell the properties. To overcome this, they consult an analytics firm and would like to get insights on how densely the area is populated and different level of income group people reside. You as a Data Scientist perform Support Vector Machines Algorithm on the given dataset and bring out informative insights and also comment on if its viable for investment in that area.

Sol:

Business Objective: to predict the customers will buy the properties or not with respect to their salaries and other factors by using Support vector machine.

Constraints: Lack of analysis of the previous data related to the customers.

Data Types: The given data and its types are as follows:

Name of feature	Description	Data Type	Relevance
Age	Age of the customer	Ratio	Relevant
Workclass	Work Class of the customer	Nominal	Relevant
education	Education type of the customer	Nominal	Relevant
educationno	Rank given to the education of the customer	Ordinal	Relevant
maritalstatus	Whether the customer is married or not	Nominal	Relevant
occupation	Occupation of the customer	Nominal	Relevant
relationship	Relationship with the customer	Nominal	Relevant
Race	Racism of the customer	Nominal	Relevant

sex	Sex of the customer	Nominal	Relevant
capitalgain	Capital amount gained by the customer	Ratio	Relevant
capitalloss	Capital loss by the customer	Ratio	Relevant
hoursperweek	Number hours worked per week	Ratio	Relevant
native	Native country of the customer	Nominal	Relevant
Salary	Salary of the customer whether >50K or <=50K	Nominal	Relevant

Data Pre-Processing: All the variables of the given data is used for doing the analysis. Some of the variables in the given data is in categorical format so the same is converted into numeric data in order to do the analysis.

Support Vector Machine: After cleaning the data the same is used for using the analysis by taking the output variable as salary. SVM model is generated for the given data and the accuracy of the linear kernel model is 84.62% and the accuracy RBF kernel model is 85.43%.

	age	workclass	education	educationno	maritalstatus	occupation	relationship	race	sex	capitalg
1	39	State-gov	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174
2	50	Self-emp-not-inc	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0
3	38	Private	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0
4	53	Private	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0
5	28	Private	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0
6	37	Private	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0
7	49	Private	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0
8	52	Self-emp-not-inc	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0
9	31	Private	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084
10	42	Private	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178
11	37	Private	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0
12	30	State-gov	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0
13	23	Private	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0
14	32	Private	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0
15	34	Private	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0

Problem Statement: -

In California, annual forest fires can cause huge loss of wild life, human life and property damage can skyrocket in billions. Local officials would like to predict the size burned area in forest fires annually so that they can be better prepared in future calamities.

Build a Support Vector Machines algorithm on the dataset and share your insights on it in the documentation.

Note: - Size_ category is the output variable.

Business Objective: To predict the forest area with other factors using support vector machine model.

Constraints: Lack of analysis of the previous forest fire data.

Data Types: Given data and its types are as follows:

Name of feature	Description	Data Type	Relevance
Month	Month when fire occurred	Nominal	Relevant
Day	Day when fire occured	Nominal	Relevant
FFMC	Fine Fuel Moisture Code for moisture content in fire data	Ratio	Relevant
DMC	Duff Moisture Code for moisture content in fire data	Ratio	Relevant
DC	third moisture index for moisture content in fire data	Ratio	Relevant
ISI	Initial Spread Index of the fire data	Ratio	Relevant
Temp	Temperature of the area	Ratio	Relevant
RH	Relative humidity of the fire area	Ratio	Relevant
Wind	Wind speed of the fire area	Ratio	Relevant
Rain	Rain forecast of the fire area	Ratio	Relevant
Area	Area of the fired region	Ratio	Relevant
Day type	All the days are converted into 1 or 0 based on type of day	Nominal	Relevant
Month type	All the months are converted into 1 or 0 based on type of months	Nominal	Relevant
size_category	Size of the area whether small or bir	Nominal	Relevant

Data Pre-Processing: All the given data is used for doing the analysis. Some of the data are in categorical format so the same is converted into numeric data to do the analysis on the data.

Support Vector Machine: After cleaning the data the same is used for using the analysis by taking the output variable as size category. SVM model is generated for the given data and the accuracy of the linear kernel model is 96% and the accuracy RBF kernel model is 82%.

	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area	dayfri	daymon	daysat	daysun
1	mar	fri	86.2	26.2	94.3	5.1	8.2	51	6.7	0.0	0.00	1	0	0	0
2	oct	tue	90.6	35.4	669.1	6.7	18.0	33	0.9	0.0	0.00	0	0	0	0
3	oct	sat	90.6	43.7	686.9	6.7	14.6	33	1.3	0.0	0.00	0	0	1	0
4	mar	fri	91.7	33.3	77.5	9.0	8.3	97	4.0	0.2	0.00	1	0	0	0
5	mar	sun	89.3	51.3	102.2	9.6	11.4	99	1.8	0.0	0.00	0	0	0	1
6	aug	sun	92.3	85.3	488.0	14.7	22.2	29	5.4	0.0	0.00	0	0	0	1
7	aug	mon	92.3	88.9	495.6	8.5	24.1	27	3.1	0.0	0.00	0	1	0	0
8	aug	mon	91.5	145.4	608.2	10.7	8.0	86	2.2	0.0	0.00	0	1	0	0
9	sep	tue	91.0	129.5	692.6	7.0	13.1	63	5.4	0.0	0.00	0	0	0	0
10	sep	sat	92.5	88.0	698.6	7.1	22.8	40	4.0	0.0	0.00	0	0	1	0
11	sep	sat	92.5	88.0	698.6	7.1	17.8	51	7.2	0.0	0.00	0	0	1	0
12	sep	sat	92.8	73.2	713.0	22.6	19.3	38	4.0	0.0	0.00	0	0	1	0
13	aug	fri	63.5	70.8	665.3	0.8	17.0	72	6.7	0.0	0.00	1	0	0	0
14	sep	mon	90.9	126.5	686.5	7.0	21.3	42	2.2	0.0	0.00	0	1	0	0
15	sep	wed	92.9	133.3	699.6	9.2	26.4	21	4.5	0.0	0.00	0	0	0	0
16	sep	fri	93.3	141.2	713.9	13.9	22.9	44	5.4	0.0	0.00	1	0	0	0
17	mar	sat	91.7	35.8	80.8	7.8	15.1	27	5.4	0.0	0.00	0	0	1	0