Topic(s): Decision Tree & Random Forest

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: Decision Tree And Random Forest

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.

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 Decision Tree and Random Forest on the given datasets.

5.3 Train and Test the data and perform cross validation techniques, compare

accuracies, precision and recall and explain about them.

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 cloth manufacturing company is interested to know about the segment or attributes contributing to high sale. Approach - A decision tree & random forest model can be built with target variable 'Sales' (we will first convert it into categorical variable) & all other variables will be independent in the analysis.

Sol:

Business Objective: To know about the high sales of the clothes by using a decision tree.

Constraints: Lack of analysis of the sales data.

Data Types: given data and its data types are shown below

Name of feature	Description	Data type	Relevance
Sales	Sales of the company	Ratio	Relevant
CompPrice	Company Price	Ratio	Relevant
Income	Profit on the product	Ratio	Relevant
Advertising	Amount on advertisement	Ratio	Relevant
Population	Population of the place	Ratio	Relevant
Price	Price in the particular place	Ratio	Relevant
ShelveLoc	Rating for the sales	Nominal	Relevant
Age	Age of the buyer	Ratio	Relevant
Education	Education number of buyer	Ordinal	Relevant
Urban	Whether its urban or not	Nominal	Relevant
US	Whether its US or not	Nominal	Relevant

Data Pre-Processing: I have identified the non-numeric data in the given data set and then I have converted that into numeric using factor function in R and in python by using Label Encoding. For the sales column I have segregated into yes or no by taking the high sales greater than 9 and remaining all as no.

Decision tree: I have made the decision tree model for the given data set using Python. I have splited the data into training and testing as 70% and 30 % Respectively and My model have the accuracy of 74% for prediction of the data in the decision tree.

Random Forest: Since the accuracy is little bit low in the decision tree model I have done the random forest model for the same data sets of training and testing data and h I have got the accuracy as 82%. So this model can be used for predicting the future sales of the cloth manufacturing company.

	Sales	CompPrice	Income	Advertising	Population	Price	ShelveLoc	Age	Education	Urban	US
1	9.50	138	73	11	276	120	Bad	42	17	Yes	Yes
2	11.22	111	48	16	260	83	Good	65	10	Yes	Yes
3	10.06	113	35	10	269	80	Medium	59	12	Yes	Yes
4	7.40	117	100	4	466	97	Medium	55	14	Yes	Yes
5	4.15	141	64	3	340	128	Bad	38	13	Yes	No
6	10.81	124	113	13	501	72	Bad	78	16	No	Yes
7	6.63	115	105	0	45	108	Medium	71	15	Yes	No
8	11.85	136	81	15	425	120	Good	67	10	Yes	Yes
9	6.54	132	110	0	108	124	Medium	76	10	No	No
10	4.69	132	113	0	131	124	Medium	76	17	No	Yes
11	9.01	121	78	9	150	100	Bad	26	10	No	Yes
12	11.96	117	94	4	503	94	Good	50	13	Yes	Yes
13	3.98	122	35	2	393	136	Medium	62	18	Yes	No
14	10.96	115	28	11	29	86	Good	53	18	Yes	Yes
15	11.17	107	117	11	148	118	Good	52	18	Yes	Yes
16	8.71	149	95	5	400	144	Medium	76	18	No	No
17	7.58	118	32	0	284	110	Good	63	13	Yes	No

Problem Statement: -

Divide the data (Diabetes) into training and test datasets and create a Random Forest and Decision Tree Model to classify 'Class Variable' or “Outcome”

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0
7	3	78	50	32	88	31.0	0.248	26	1
8	10	115	0	0	0	35.3	0.134	29	0
9	2	197	70	45	543	30.5	0.158	53	1
10	8	125	96	0	0	0.0	0.232	54	1
11	4	110	92	0	0	37.6	0.191	30	0
12	10	168	74	0	0	38.0	0.537	34	1
13	10	139	80	0	0	27.1	1.441	57	0
14	1	189	60	23	846	30.1	0.398	59	1
15	5	166	72	19	175	25.8	0.587	51	1
16	7	100	0	0	0	30.0	0.484	32	1
17	0	118	84	47	230	45.8	0.551	31	1

Business Objective: To segregate the diabetes patients by using a decision tree and random forest.

Constraints: Lack of analysis of the patients data.

Data Types: given data and its data types are shown below

Name of feature	Description	Data type	Relevance
Number of times pregnant	Number of time the woman got pregnancy	Internal	Relevant
Plasma glucose concentration	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Ratio	Relevant
Diastolic blood pressure	Diastolic blood pressure (mm Hg)	Ratio	Relevant
Triceps skin fold thickness	Triceps skin fold thickness (mm)	Ratio	Relevant
2-Hour serum insulin	2-Hour serum insulin (mu U/ml)	Ratio	Relevant
Body mass index	Body mass index (weight in kg/(height in m)^2)	Ratio	Relevant
Diabetes pedigree function	Diabetes pedigree function	Ratio	Relevant
Age (years)	Age of patient	Ratio	Relevant
Classvariable	Variable for patiest having diabetes or not	Nominal	Relevant

Data Pre-Processing: the complete data can be used for preparing the model of decision tree and random forest.

Decision tree: I have made the decision tree model for the given data set using Python. I have splited the data into training and testing as 70% and 30 % Respectively and My model have the accuracy of 72% for prediction of the data in the decision tree.

Random Forest: Since the accuracy is little bit low in the decision tree model I have done the random forest model for the same data sets of training and testing data and h I have got the accuracy as 80%. So this model can be used for predicting the whether the patient will get diabetes or not.

Problem Statement: -

Use decision trees & random forest algorithm to prepare a model on fraud datatreating those who have taxable_income <= 30000 as "Risky" and others are "Good".

	Undergrad	Marital.Status	Taxable.Income	City.Population	Work.Experience	Urban
1	NO	Single	68833	50047	10	YES
2	YES	Divorced	33700	134075	18	YES
3	NO	Married	36925	160205	30	YES
4	YES	Single	50190	193264	15	YES
5	NO	Married	81002	27533	28	NO
6	NO	Divorced	33329	116382	0	NO
7	NO	Divorced	83357	80890	8	YES
8	YES	Single	62774	131253	3	YES
9	NO	Single	83519	102481	12	YES
10	YES	Divorced	98152	155482	4	YES
11	NO	Single	29732	102602	19	YES
12	NO	Single	61063	94875	6	YES
13	NO	Divorced	11794	148033	14	YES
14	NO	Married	61830	86649	16	YES
15	NO	Married	64070	57529	13	YES
16	NO	Divorced	69869	107764	29	NO
17	YES	Divorced	24987	34551	29	NO

Sol:

Business Objective: To know about the check is risky or good using a decision tree and random forest models.

Constraints: Lack of analysis of the fraud checks data.

Data Types: given data and its data types are shown below

Name of feature	Description	Data type	Relevance
Undergrad	Whether the applicant is undergraduate or not	Nominal	Relevant
Marital.Status	Marital status of applicant	Nominal	Relevant
Taxable.Income	Taxable income of the applicant	Ratio	Relevant
City.Population	Population of the city	Ratio	Relevant
Work.Experience	Work experience of the applicant	Ratio	Relevant
Urban	Whether the state is urban or not	Nominal	Relevant

Data Pre-Processing: I have identified the non-numeric data in the given data set and then I have converted that into numeric using factor function in python by using Label Encoding. For the taxable income column I have segregated into Risky or Good by taking the taxable income greater than 30000 as Risky and remaining all as Good.

Decision tree: I have made the decision tree model for the given data set using both R and Python. I have splitted the data into training and testing as 70% and 30 % Respectively and My model have the accuracy of 70% for prediction of the data in the decision tree.

Random Forest: Since the accuracy is little bit low in the decision tree model I have done the random forest model for the same data sets of training and testing data and I have got the accuracy as 78%. So this model can be used for predicting the whether the check is risky or not.

Problem Statement: -

In Recruitment domain, HR faces with the challenge of predicting if the candidate is faking his salary or the candidate is genuine. In order to do it manually, let us use our Machine Learning algorithm to correctly classify using Decision Tree and Random Forest. We have a scenario where, a candidate claims to have 5 years of experience and earning 70000 per month working as regional manager and the candidate is expecting more than his previous CTC. A sample data has been collected, find out the candidate claims are genuine or fake.

Sol:

Business Objective: To know whether the candidate is expediting the correct CTC or not by using decision tree and random forest models.

Constraints: Lack of analysis of the previous employee data.

Data Types: given data and its data types are shown below

Name of feature	Description	Data type	Relevance
Position of the employee	Employee position in the company.	Nominal	Relevant
no of Years of Experience of employee	Experience of the employee	Ratio	Relevant
Monthly income of employee	Salary of the employee	Ratio	Relevant

Data Pre-Processing: I have identified the non-numeric data in the given data set and then I have converted that into numeric using factor function in python by using Label Encoding. For the Monthly income of employee column I have segregated into High or Medium & Low by taking the Monthly income of employee greater than 70000 as High and remaining all as Medium & Low.

Decision tree: I have made the decision tree model for the given data set using both R and Python. I have splitted the data into training and testing as 70% and 30 % Respectively and My model have the accuracy of 75% for prediction of the data in the decision tree.

Random Forest: Since the accuracy is little bit low in the decision tree model I have done the random forest model for the same data sets of training and testing data and I have got the accuracy as 84%. So this model can be used for whether the new employee is demanding correct CTC or Not.

A	B	C
Position of the employee	no of Years of Experience of e	monthly income of employee
Business Analyst	1.1	39343
Junior Consultant	1.3	46205
Senior Consultant	1.5	37731
Manager	2	43525
Country Manager	2.2	39891
Region Manager	2.9	56642
Partner	3	60150
Senior Partner	3.2	54445
C-level	3.2	64445
CEO	3.7	57189
Junior Consultant	3.9	63218
Senior Consultant	4	55794
Manager	4	56957
Region Manager	4.1	57081
Business Analyst	4.5	61111
Junior Consultant	4.9	67938
Senior Consultant	5.1	66029
Manager	5.3	83088
Country Manager	5.9	81363
Region Manager	6	93940
Partner	6.8	91738
Senior Partner	7.1	98273
C-level	7.9	101302

truth_or_bluffy