COM6012 Assignment-2

Registration Number : **180128033**

1 Question1:

In the question 1 **part(a)**, **part(b)** and **part(c)**, we are using Decision Trees for Classification, Decision Trees for Regression and Logistic Regression by the help of pipelines and cross-validation for 25 percent and 100 percent of data to find the best configuration of parameters and their accuracy.In Decision Trees for Classification, we have given maxDepth(5,10,15), maxBins(30,31,32), impurity ("gini","entropy").In Decision Trees for Regression maxDepth(5, 10, 30),maxBins(20, 35, 40).In Logistic Regression maxIter (5, 10, 30) ,regParam (0.3, 0.6, 0.9) ,elasticNetParam (0.6,0.8 ,1.0).After running code for different parameter, we get the best of these, further on which we run are 100 percent of code.

1.1 Table of Accuracy,AUC and best parameter for 25 percent data

	Accuracy Core(10)	AUC Core(10)	Parameter Core(10)	Accuracy Core(20)	AUC Core(20)	Parameter Core(20)
DecisionTreeClassifier	0.703546	0.68900	maxBin 32 maxDepth 10 impurity gini	0.703546	0.68900	maxBin 32 maxDepth 10 impurity gini
DecisionTreeRegression	0703764	0.77595	maxBin 35 maxDepth 10	0703764	0.77595	maxBin 35 maxDepth 10
Logistic Regression	0.530278	0.5	elasticNetParam 0.6 maxIter 5 regParam 0.3	0.530278	0.5	elasticNetParam 0.6 maxIter 5 regParam 0.3

1.2 All Best Parameter

DecisionTreeClassifier Parameter	DecisionTreeRegression Parameter	Logistic Regression
cacheNodeIds False		aggregationDepth 2
checkpointInterval 10	cacheNodeIds False	elasticNetParam 0.6
featuresCol features	checkpointInterval 10	family auto
impurity gini	featuresCol features	featuresCol features
labelCol label	impurity variance	fitIntercept True
maxBins 32	labelCol label	labelCol label
maxDepth 10	maxBins 35	maxIter 5
maxMemoryInMB 256	maxDepth 10	predictionCol prediction
minInfoGain 0.0	maxMemoryInMB 256	probabilityCol probability
minInstancesPerNode 1	minInfoGain 0.0	rawPredictionCol rawPrediction
predictionCol prediction	minInstancesPerNode 1	regParam 0.3
probabilityCol probability	predictionCol prediction	standardization True
rawPredictionCol rawPrediction	seed -1407754390808368278	threshold 0.5
seed 956191873026065186		tol 1e-06

1.3 Table of Accuracy,AUC, Top 3 features and Time Taken for 100 percent data

	Accuracy Core(10)	AUC Core(10)	Top 3 Features	Time Taken	Accuracy Core(20)	AUC Core(20)	Top 3 Features	Time Taken
Decision Trees for Classification	0.70457	0.6740147160814 04	_c26 _c28 _c27	563 sec	0.70457	0.6740147160814 04	_c26 _c28 _c27	589 sec
Decision Trees for Regression	0.704763	0.7764754635451928	_c26 _c28 _c27	205 sec	704763	0.7764754635451928	_c26 _c28 _c27	194 sec
Logistic Regression	0.608702	0.5939483510118273	_c28 _c26 _c4	88sec	0.608702	0.5939483510118273	_c28 _c26 _c4	88 sec

1.4 Observation

- From the observation in the above table, we can say that time taken for 3 model is almost same for both the core.

- For both the core 10 and 20 value of accuracy,auc,parameter and features are same for their respective model.
- Difference in the accuracy obtained from Decision Trees for Classifier and Decision Trees for Regression are almost same in comparison to the logistic regression.

2 Question2:

2.1 Pre-processing of Data

- In this question, we are provided with 2.7 GB of train data.The data set has several fields with missing data(indicated by "?"). To balance data, we have drop those rows to have clean data.
- After cleaning data, we converted all the categorical values string and numeric values into "DoubleType" for suitable representation.
- After getting the desired scheme, we have taken "claim amount" as label and other column as feature and making them vector using vector assembler.
- we have dropped column "Row ID", "Household ID" and "Vehicle" because
- After that we have split data into training and testing by ratio 70:30 with seed = 50.
- we have then normalised the data and then found "rmse" over training data and testing data.
- There are various way to balance imbalanced data,by changing performance metric.We can use f1-score,Precision,Re-call.
- By Resampling data set.
- By using different algorithm
- We have used "rmse" here because it penalize data.

2.2 RMSE

Algorithum	RMSE(Core10)	RMSE(Core20)
Linear Regression on Training Model	0.003290	0.003290
Linear Regression on Test Model	0.00331967	0.00331967

After the observing result, we only find one difference that is time taken by core.Time taken by Core 10 was 66.72 sec where as time taken by Core 20 is 40.28