EndTerm Lab Exam

APPLIED MACHINE LEARNING

Harshit Singhai | E16CSE147 | 01 Nov 2019

EDA

- Found out what different types of data are present in the dataset i.e if the dataset contains any categorical, numerical values.
- Found the missing values from the dataset
- Found if there are any outliers.
- Feature engineering by making a correlation matrix and check how much independent variables depend on the dependent variable
- Plotted pairplot graph, heatmap and histogram to get valuable insights from the data
- Checked if the dataset is unbalanced.

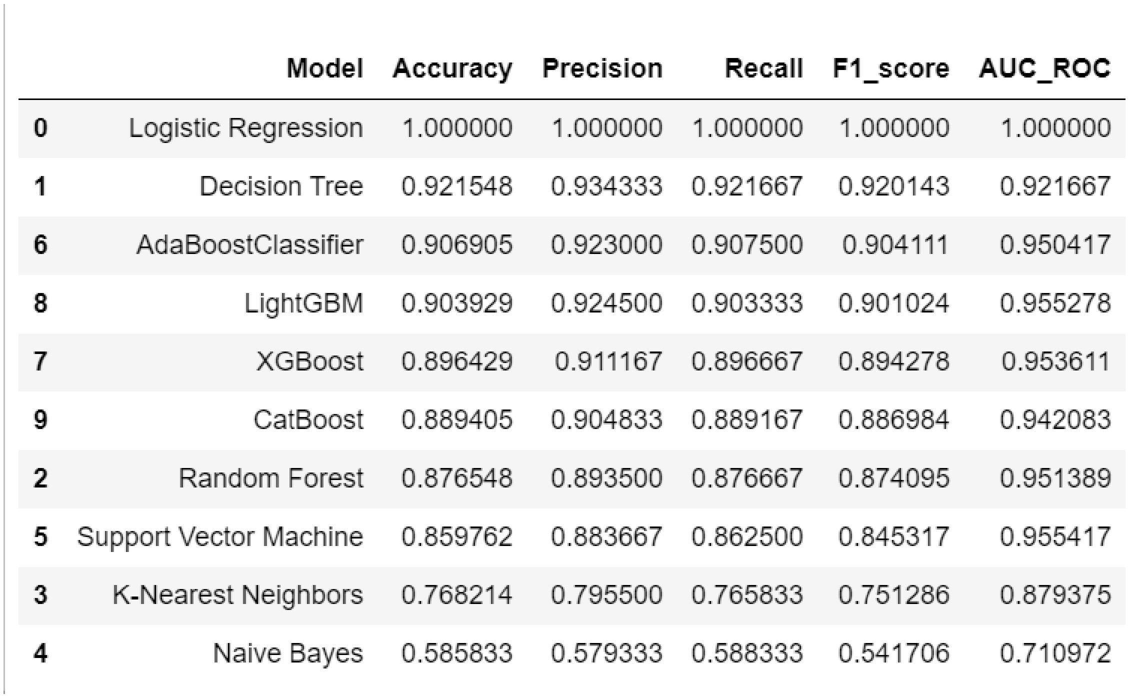
PRE-PROCESSING

- Found missing values
- Checked for categorical data
- Standardize the data
- Data splitting using train_test_split

FEATURE SELECTION

Used filter method for feature selection. In Filter method, the predictive power of each individual variable is evaluated. Correlation matrix was used to evaluate the impact of each feature with respect to the target variable.

ALGORITHMS USED



	Model	Accuracy	Precision	Recall	F1_score	AUC_ROC
0	Logistic Regression	1.000000	1.000000	1.000000	1.000000	1.000000
1	Decision Tree	0.921548	0.934333	0.921667	0.920143	0.921667
6	AdaBoostClassifier	0.906905	0.923000	0.907500	0.904111	0.950417
8	LightGBM	0.903929	0.924500	0.903333	0.901024	0.955278
7	XGBoost	0.896429	0.911167	0.896667	0.894278	0.953611
9	CatBoost	0.889405	0.904833	0.889167	0.886984	0.942083
2	Random Forest	0.876548	0.893500	0.876667	0.874095	0.951389
5	Support Vector Machine	0.859762	0.883667	0.862500	0.845317	0.955417
3	K-Nearest Neighbors	0.768214	0.795500	0.765833	0.751286	0.879375
4	Naive Bayes	0.585833	0.579333	0.588333	0.541706	0.710972

BEST RESULTS

Logistic Regression gave the best results for this problem. This is because logistic regression measures the relationship between the dependent variable (our label, what we want to predict) and the one or more independent variables (our features), by estimating probabilities using its underlying logistic function. Logistic regression performs best in binary classification problem. This is mainly because of the sigmoid function. The Sigmoid-Function is an S-shaped curve that can take any real-valued number and map it into a value between the range of 0 and 1.