Task Pure ML
from
Xenon Stack

by

Prateek Vikram

Data Exploration/Analysis

- The Data set has 1309 examples and 13 features + the target variable (survived)
- Need to convert a lot of features into numeric
- Features have widely different ranges, that we will need to convert into roughly the same scale.
- Spot some more features, that contain missing values (NaN = not a number), that we need to deal with
- Split the data into training and test set in which 80% data training set and 20% data test set ratio

Data Preprocessing

- **Missing Train Data:**
- We have to deal with Cabin(496), Body(496),Boat(495),Home.dest(273)
- A cabin number looks like ‘C452’ and the letter refers to the deck and missing values will be converted to zero
- Missing value of age is replace by its median

Data Preproccesing (Contd)

- **Converting Features:**
- ‘Fare’ is a float and we have to deal with 4 categorical features: Name, Sex, Ticket and Embarked
- Name feature to extract the Titles from the Name, so that we can build a new feature out of that.
- Convert ‘Sex’ feature into numeric.
- Drop the Ticket feature it will be a bit tricky to convert them into useful categories
- Convert ‘Embarked’ feature into numeric
- The ‘Fare’ feature first we will convert it from float into integer

Machine Learning Models

- **We will train several Machine Learning models**
- Random Forest
- Logistic Regression
- K Nearest Neighbor
- Linear Support Vector Machine
- Decision Tree

Best Model

- Case 1: Person Male Above 18

Model	Accuracy Score	Precision	Recall	F1Score
Random Forest	93.75	0.46	0.22	0.3055
Decision Tree	93.75	.20	.16	.18
KNN	86.35	0.29	0.18	0.22
Logistic Regression	82.40	0.27	0.027	.05
SVM	18.42	0.19	0.15	0.17

Best Model

- Case 2: Person Female Above 34

Model	Accuracy Score	Precision	Recall	F1Score
Random Forest	100	.88	.96	.92
Decision Tree	100	.88	.96	.89
KNN	91	0.87	0.91	0.89
Logistic Regression	89	0.88	0.94	0.91
SVM	86	.90	.87	.89

Best Model

- Case 1: Embarked C Age 55

Model	Accuracy Score	Precision	Recall	F1Score
Random Forest	93.09	0.48	0.23	0.31
Decision Tree	93.09	0.35	0.26	0.30
KNN	86.84	0.26	0.16	0.20
Logistic Regression	83.55	0.5	.019	0.037
SVM	83.22	0.37	0.29	0.33

Summary

- The data exploration where we got a feeling for the dataset, checked about missing data and learned which features are important.
- Seaborn and matplotlib to do the visualizations.
- Data preprocessing part, we computed missing values, converted features into numeric ones, grouped values into categories and created a few new features.
- 5 different machine learning models, applied **cross validation** on it.
- Lastly, we looked at it’s confusion matrix and computed the models precision, recall and f1-score.