

CS564 – DATA SCIENCE METHODOLOGY

TERM PROJECT REPORT

NGUYEN BA VINH QUANG – 20190710

1. Overview

- Using popular frameworks: pandas, numpy, scikit-learn, catboost...
- Read and process datasets as data frames with pandas.
- Pre-process for both train and test datasets.

2. Data cleaning

- Find out which columns are numerical and categorical. Save the categorical column indexes into a list.
- Calculate the percentage of missing values in each column then try to fill them in. The number of records that have the missing values in each column is not much, so this is the possible solutions
 - o For the numerical column: replace the missing with the average value.
 - o For categorical columns: replace the missing with the mode of all values.
 - o Note that: for the name column, do not fill it, this column will be dropped in later steps.

	% nulls
PassengerId	0.0%
HomePlanet	0.02312%
CryoSleep	0.02496%
Cabin	0.02289%
Destination	0.02094%
Age	0.02059%
VIP	0.02335%
RoomService	0.02082%
FoodCourt	0.02105%
ShoppingMall	0.02393%
Spa	0.02105%
VRDeck	0.02163%
Name	0.02301%
Transported	0.0%






- Split the “Cabin” column into three columns “Deck”, “Num”, and “Side”.

3. Modeling

- Drop the column “Name”, and “Cabin”.
- Add two new columns “NetSpent”, and “AgeGroup”. “NetSpent” is the sum of all spending types of passengers. “AgeGroup” is based on the “Age” of passengers. Besides that, the passenger ID is unique for every customer, so set it as the index for our data frames.
- Convert all columns of categorical into integer numbers with LabelEncoder. Now the data frame is ready to feed into the model.

PassengerId	HomePlanet	CryoSleep	Destination	Age	VIP	RoomService	FoodCourt	ShoppingMall	Spa	VRDeck	Transported	Deck	Num	Side	AgeGroup	NetSpent
0001_01	1	0	2	39.0	0	0.0	0.0	0.0	0.0	0.0	0	1	0	0	3	0.0
0002_01	0	0	2	24.0	0	109.0	9.0	25.0	549.0	44.0	1	5	0	1	2	736.0
0003_01	1	0	2	58.0	1	43.0	3576.0	0.0	6715.0	49.0	0	0	0	1	5	10383.0
0003_02	1	0	2	33.0	0	0.0	1283.0	371.0	3329.0	193.0	0	0	0	1	3	5176.0
0004_01	0	0	2	16.0	0	303.0	70.0	151.0	565.0	2.0	1	5	1	1	1	1091.0

- Using 5 methods to build the model:
 - o The catboost model.
 - o The random forest.
 - o The support vector machine.
 - o The grid search CV.
 - o The voting classifier is a combination of the above models.
- I did not divide the given train dataset into train and test datasets, so all the given train dataset is fed into the model. After that, I use each model to predict the given test dataset and submit it to the Kaggle.
- The result for all methods is as follows

Submission and Description		Public Score 📄
<div> All Successful Errors </div> <div>Recent ▾</div>		
 submission.csv Complete · 13m ago · rf1	0.78933	
 submission.csv Complete · 13m ago · sv1	0.79027	
 submission.csv Complete · 14m ago · cb2	0.80476	
 submission.csv Complete · 15m ago · grid1	0.80453	
 submission.csv Complete · 16m ago · vote1	0.80289	

- The best performance is the catboost model which is more than 0.8 scores.
- The lowest scores belong to the random forest model.
- Do not have much difference between the catboost model and the voting model.

4. References

1. https://www.youtube.com/watch?v=pUSi5xexT4Q&ab_channel=AladdinPersson
2. https://www.youtube.com/watch?v=HReBjpi9dCY&t=969s&ab_channel=AutomataLearningLab

Submission capture



submission.csv

Complete · 14m ago · cb2

0.80476