

Machine Learning Project

109550111 吳守元

Github

<https://github.com/billwu90325/ML-Final-Project>

Model

https://drive.google.com/drive/folders/1bfmlql13XmZxdolGSNjb_fbN4aluY2w-?usp=share_link

Brief introduction

這次的是要用label去預測boolean的output。主要需要處理或考慮的有以下幾個重點

- 如何解決含有NaN的row？
- 如何處理類別而非數字的column？
- 如何在原先的features中提取出更有用的訊息？
- 哪種model的表現最好？

Methodology

針對以上幾點，我應對的做法是

- 用Kuma_utils的LGBMImputer去impute
- 用sklearn的LabelEncoder去把類別轉換成數字
- (1) 用原先的data重新組合、計算出新的columns
(2) 用sklearn的mutual information score去選擇出幾個與label相關性較強的columns
- sklearn的Logistic Regression是表現最好的，而Model的Hyper parameter的部分我是用HW5的方法用kfold的方式找出一組最好的再去fit整個train

Summary

我覺得這次的project比較麻煩的地方是preprocessing的部分，像是如何填補NaN對data所帶來的影響，不同的方法都會對結果有影響，再者如何把原先的data拼湊、組合出更具有相關性的數據，也是一個需要花時間捉摸的地方。Model的部分我倒是覺得比較沒那麼複雜，主要是kfold比較怎樣的Hyper parameter會有最好的performance。

Comparisons of different approaches

* Preprocessing

- 直接drop掉有nan的row
- 用LGBMImputer賦值

我一開始是用a的方法，我原本的想法是missing的value不會很多，直接drop掉可以免去其他的過程，但發現處理完只剩一半的train，表現自然不會好。而後來採用的方法是impute，我就選了一個測試出來表換最好的Impute algorithm。

* Features Engineering

- Missing
- Group avg & std

原本是用這個columns是否missing(boolean)來造出一個新的column(column個數double)，表現有提升但一直沒辦法越過baseline，後來我就找其他engineering的方法，就找到我目前使用的這個組合，private score就有辦法突破0.59了。

* Model

- Logistic Regression
- NN(Keras)

原本預期NN的效果會比較好，但後來一直都是regression贏一點點，我後面就都採用sklearn的logistic regression。

Kaggle score

The screenshot shows the Kaggle competition page for 'Tabular Playground Series - Aug 2022'. The header includes the competition title, a description 'Practice your ML skills on this approachable dataset!', and statistics 'Kaggle · 1,888 teams · 4 months ago'. Below the header is a navigation bar with links: Overview, Data, Code, Discussion, Leaderboard (active), Rules, and Team. On the right of the navigation bar are links for Submissions, Late Submission, and a menu icon. The main section is titled 'Leaderboard' and includes buttons for 'Raw Data' and 'Refresh'. Under the 'YOUR RECENT SUBMISSION' section, a submission named 'submission_0109.csv' is listed, submitted by 'Wu Shou-Yuan' a few seconds ago. The submission has a score of 0.59041 and a public score of 0.58832. A button 'Jump to your leaderboard position' is also present.

Playground Prediction Competition

Tabular Playground Series - Aug 2022
Practice your ML skills on this approachable dataset!

Kaggle · 1,888 teams · 4 months ago

Overview Data Code Discussion **Leaderboard** Rules Team

Submissions **Late Submission** ...

Leaderboard [Raw Data](#) [Refresh](#)

YOUR RECENT SUBMISSION

submission_0109.csv
Submitted by Wu Shou-Yuan · Submitted a few seconds ago

Score: 0.59041
Public score: 0.58832

[Jump to your leaderboard position](#)

Reference

1. Address “NaN”

<https://www.kaggle.com/code/sfctrkl/tps-aug-2022/notebook#Modelling>

2. Feature Engineering:

<https://www.kaggle.com/code/themikejones/tps-aug-22-votingclassifier/notebook?scriptVersionId=102761965#4.1-Combine-features-and-create-new>

3. Kfold:

<https://www.kaggle.com/code/takanashihumbert/tps-aug22-9th-solution>