Projetos em Ciência de Dados

Gianlucca Devigili e Maisa de O. Fraiz

Redução do uso de memória

- Transformar cada coluna no "menor" datatype possível. E.g float64 -> float32
- Dados do tipo object -> category

Redução do uso de memória:
Datasets Base

Df Targets

```
df_targets = reduce_mem_usage(df_targets);
```

✓ 0.2s

```
Memory usage of dataframe is 102.84 MB
Memory usage after optimization is: 71.99 MB
Decreased by 30.0%
```

Player Box Scores

```
df_pbs = reduce_mem_usage(df_pbs);
```

✓ 0.8s

```
Memory usage of dataframe is 92.20 MB
Memory usage after optimization is: 29.97 MB
Decreased by 67.5%
```

Redução do uso de memória:
Datasets Principal

Reduce memory usage

```
df_train = reduce_mem_usage(df_train);
```

✓ 12.7s

```
Memory usage of dataframe is 1331.04 MB
Memory usage after optimization is: 450.51 MB
Decreased by 66.2%
```

- Redução de ≈880.5 MB (66.2%)
- Redução do tempo de treinamento dos modelos:
  - Gradient Boosting: 2h 48m 6s -> 1h 46m 58s

Modelos de Multitask

- Motivação: resultado do MultitaskLASSO

3	multitask LASSO	0.748527	1.503057	0.672377	0.789836	0.928449

- Uso do *sklearn.MultiOutputRegressor* ao invés do código programado pelo próprio grupo
- Possibilidade de usar *feature Chaining*

Modelos de Multitask

	model	target1	target2	target3	target4	average
0	Média (sem PBS)	1.126844	2.739029	1.068968	1.477766	1.603152
1	Média por Jogador (sem PBS)	0.939999	2.251019	0.954300	1.025011	1.292582
2	Mediana (sem PBS)	0.712801	1.651943	0.498075	1.139852	1.000668
3	Mediana por Jogador (sem PBS)	0.702606	1.560620	0.493126	0.925954	0.920577
4	Naive (sem PBS)	1.168903	1.808041	0.761283	1.520494	1.314680

	model	target1	target2	target3	target4	average
0	Lasso | MultiOutput	0.780746	1.344433	0.703082	0.746220	0.893620
1	Ridge | MultiOutput	0.784643	1.331393	0.729669	0.756929	0.900659
2	ElasticNet | MultiOutput	0.782295	1.342960	0.710303	0.747843	0.895850
3	DecisionTreeRegressor | MultiOutput	2.361119	2.124396	1.964909	1.574683	2.006277
4	GradientBoostingRegressor | MultiOutput	0.731110	1.066651	0.666981	0.726966	0.797927

Cross Validation com LASSO

- Procurar o valor ótimo para o parâmetro de regularização. Mas ótimo para que?
  - α_{pred} -> otimiza as predições
  - α_{ms} -> dá o valor do modelo verdadeiro
- Em geral, α_{pred} != α_{ms} e $\alpha_{pred} < \alpha_{ms}$
- Em resumo, overfitting porque estávamos procurando o α_{ms} ao usar o LassoCV, o que levava a um erro grande.

Em progresso

- Uso da coluna de data:
  - A data como variável ordinal
  - Decomposição em ano, mês, dia, dia da semana, dia do ano, semana do ano...
  - Dummies vs. valor numérico

Projetos em Ciência de Dados

Gianlucca Devigili e Maisa de O. Fraiz