Avaliação de capacidade preditiva de um modelo para identificar preferência de anime por usuário

Bruno de Lima Santos
Instituto Federal do Espírito Santo

Contextualizando o problema

A indústria de anime utiliza de tentativa e erro: cada grande sucesso supera os gastos de uma gama de investimentos sem resultado.

Problema no futuro: com o streaming, muito material “ruim” fica disponível e os bons não se destacam.

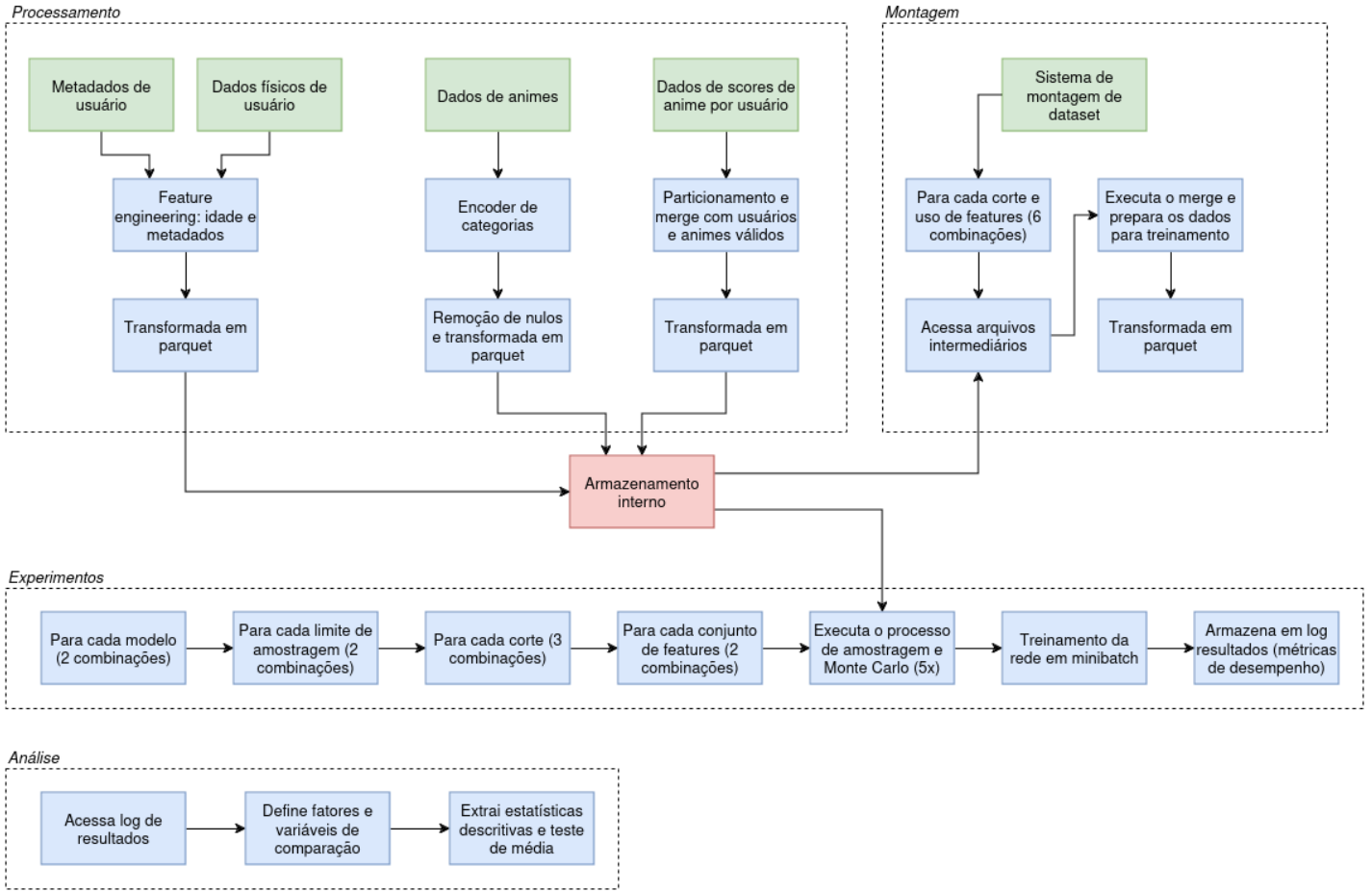
Saber se um anime vai ser bem recebido evita gasto com material ruim e reduz a quantidade de escolhas do usuário.

Desafios de modelagem

O disponível são os scores dados pelos usuários, logo é preciso decidir a regra de negócio mais adequada para converter a escala em binário.

Além dos dados físicos da animação (duração, gênero, material original e outros), os dados do usuário precisam ser considerados. Neste escopo, pode-se utilizar apenas dados físicos ou dados físicos acrescidos de metadados.

Fluxograma de execução



Resultados: metadados

	factor	basic	full	ttest-basic-full
0	arch=1,sample=0.1,cut=cut6	0.888312	0.895944	1.124163e-07
1	arch=1,sample=0.1,cut=cut7	0.708368	0.762878	1.958774e-08
2	arch=1,sample=0.1,cut=cut8	0.311931	0.536893	4.948791e-09
3	arch=1,sample=0.2,cut=cut6	0.888384	0.896093	5.701522e-10
4	arch=1,sample=0.2,cut=cut7	0.705443	0.761860	1.809408e-09
5	arch=1,sample=0.2,cut=cut8	0.309045	0.555534	1.653350e-09
6	arch=2,sample=0.1,cut=cut6	0.887408	0.896105	9.693076e-08
7	arch=2,sample=0.1,cut=cut7	0.706057	0.762408	1.088779e-07
8	arch=2,sample=0.1,cut=cut8	0.280773	0.541882	1.329206e-08
9	arch=2,sample=0.2,cut=cut6	0.888257	0.896011	1.944084e-08
10	arch=2,sample=0.2,cut=cut7	0.705468	0.762367	3.437826e-08
11	arch=2,sample=0.2,cut=cut8	0.303647	0.550261	1.370115e-11

Resultados: local de corte

	factor	cut6	cut7	cut8	ttest-cut6-cut7	ttest-cut6-cut8	ttest-cut7-cut8
0	arch=1,sample=0.1,features=basic	0.888312	0.708368	0.311931	1.207986e-13	7.911281e-14	2.349218e-12
1	arch=1,sample=0.1,features=full	0.895944	0.762878	0.536893	9.855015e-13	1.485798e-11	7.582376e-10
2	arch=1,sample=0.2,features=basic	0.888384	0.705443	0.309045	1.720762e-14	1.496629e-12	3.513435e-11
3	arch=1,sample=0.2,features=full	0.896093	0.761860	0.555534	6.473266e-14	7.305112e-16	1.794684e-13
4	arch=2,sample=0.1,features=basic	0.887408	0.706057	0.280773	8.150079e-12	1.208087e-11	2.788426e-10
5	arch=2,sample=0.1,features=full	0.896105	0.762408	0.541882	4.742614e-15	3.858148e-14	2.026072e-12
6	arch=2,sample=0.2,features=basic	0.888257	0.705468	0.303647	1.802789e-14	2.340404e-15	8.447495e-14
7	arch=2,sample=0.2,features=full	0.896011	0.762367	0.550261	1.199389e-11	1.643081e-14	7.672061e-12

Resultados: arquitetura

	factor	1	2	ttest-1-2
0	sample=0.1,features=basic,cut=cut6	0.888312	0.887408	0.016724
1	sample=0.1,features=basic,cut=cut7	0.708368	0.706057	0.536916
2	sample=0.1,features=basic,cut=cut8	0.311931	0.280773	0.034157
3	sample=0.1,features=full,cut=cut6	0.895944	0.896105	0.787576
4	sample=0.1,features=full,cut=cut7	0.762878	0.762408	0.808731
5	sample=0.1,features=full,cut=cut8	0.536893	0.541882	0.513080
6	sample=0.2,features=basic,cut=cut6	0.888384	0.888257	0.553339
7	sample=0.2,features=basic,cut=cut7	0.705443	0.705468	0.990722
8	sample=0.2,features=basic,cut=cut8	0.309045	0.303647	0.556095
9	sample=0.2,features=full,cut=cut6	0.896093	0.896011	0.826883
10	sample=0.2,features=full,cut=cut7	0.761860	0.762367	0.854994
11	sample=0.2,features=full,cut=cut8	0.555534	0.550261	0.143383

Resultados: amostragem

	factor	0.1	0.2	ttest-0.1-0.2
0	arch=1,features=basic,cut=cut6	0.888312	0.888384	0.777809
1	arch=1,features=basic,cut=cut7	0.708368	0.705443	0.241177
2	arch=1,features=basic,cut=cut8	0.311931	0.309045	0.774331
3	arch=1,features=full,cut=cut6	0.895944	0.896093	0.733955
4	arch=1,features=full,cut=cut7	0.762878	0.761860	0.642415
5	arch=1,features=full,cut=cut8	0.536893	0.555534	0.026259
6	arch=2,features=basic,cut=cut6	0.887408	0.888257	0.013637
7	arch=2,features=basic,cut=cut7	0.706057	0.705468	0.867389
8	arch=2,features=basic,cut=cut8	0.280773	0.303647	0.081461
9	arch=2,features=full,cut=cut6	0.896105	0.896011	0.864736
10	arch=2,features=full,cut=cut7	0.762408	0.762367	0.987429
11	arch=2,features=full,cut=cut8	0.541882	0.550261	0.074701

Conclusões e Propostas Futuras

Nem o aumento da amostragem e nem o aumento da profundidade da rede demonstraram evidências de melhoria do modelo.

O foco no desenvolvimento deste tipo de projeto precisa ser nos metadados dos usuários, provavelmente quanto melhor os metadados, melhor a capacidade preditiva.

Desenvolver uma forma de agrupar os usuários a partir dos seus metadados pode fornecer importantes insumos para estratégias de marketing que utilizem o modelo proposto.

Explorar com mais detalhes o impacto do balanceamento das classes alvo ao se aplicar as regras de negócio para conversão da pontuação.

Repositório

GitHub: https://github.com/BrunoSantosPK/annproject