ESTGF | POLITÉCNICO DO PORTO

LSIG LICENCIATURA SISTEMAS DE INFORMAÇÃO PARA A GESTÃO

Métodos e Técnicas de Análise de Dados

Docente:

Davide Carneiro

Aluno:

Francisco Ferreira nº 8170565

Índice

1. Introdução

Com base nos conteúdos lecionados na Unidade Curricular Métodos e Técnicas de Análise de Dados este trabalho tem como objetivo obter informações úteis na tomada de decisão de uma organização num determinado contexto, ambos fictícios.

Com os dados recolhidos é suposto utilizar a plataforma H2O para treino de modelos e posterior avaliação dos mesmos de forma a encontrar valor para suportar a tomada de decisão da organização através de uma aplicação web (Shiny Web App).

2.Análise e descrição do domínio

2.1 Apresentação e caracterização do Conjunto de Dados

Este relatório tem como objeto de estudo o do *Dataset* "*BankChurners.csv"* escolhido para realizar este trabalho prático, este foi retirado da plataforma *Kaggle*.
O Dataset tem1 10127 linhas e 20 colunas.
Este conjunto de dados reflete a base de dados de clientes de um banco onde são recolhidos dados pertinentes para o a gestão por parte da Entidade Financeira. Todas as variáveis recolhidas estão descritas na tabela que se segue.

Variáveis	**Significado**
CLIENTNUM	Número identificador do cliente
Attrition_Flag	Estado do cliente, ou seja, se se encontra como cliente existente ou se entrou em incumprimento
Customer_Age	Idade
Gender	Género
Dependent_count	Número dependentes
Education_Level	Categoria académica
Marital_Status	Situação matrimonial
Income_Category	Rendimento anual
Card_Category	Tipo de Cartão
Months_on_book	Tempo de relação com o banco em meses
Total_Relationship_Count	Nº produtos bancários detidos pelos clientes
Months_Inactive_12_mon	Nº de meses inativos no último ano
Contacts_Count_12_mon	Nº de contactos no último ano
Credit_Limit	Limite de crédito
Total_Revolving_Bal	Crédito "revolving" vai-se renovando consoante o pagamento da dívida é efetuado, assim como os respetivos juros associados.
Avg_Open_To_Buy	Saldo aquando abertura do crédito (média dos últimos meses)
Total_Amt_Chng_Q4_Q1	Variação do valor das transações (Q4 sobre Q1)
Total_Trans_Amt	Valor do total das transações (últimos 12 meses)
Total_Trans_Ct	Número total de transações (últimos 12 meses)
Total_Ct_Chng_Q4_Q1	Variação do número de transações (Q4 sobre Q1)
Avg_Utilization_Ratio	Rácio de utilização média do cartão

2.2 Descrição do domínio e possíveis problemas do mesmo

Entidades como esta obtêm a sua principal fonte de rendimento em produtos financeiros, normalmente associados a pagamento de juros por parte dos seus clientes. Para que a gestão de clientes e produto seja feita da forma mais otimizada e personalizada é necessário que a Entidade Financeira conheça bem os seus clientes. Na tabela acima podemos aferir que o rendimento anual, a idade ou até a quantidade de dependentes podem ser fatores que deem mais ou menos credibilidade financeira, normalmente associado ao risco. A título de exemplo podemos considerar o crédito habitação, talvez o produto financeiro mais vendido por entidades bancárias, podemos considerar que um casal com 30 anos, sem dependentes e com um rendimento anual muito superior à prestação deste serviço terá bem menos risco de incumprimento da mesma.
O conjunto de dados que irei trabalhar está associado a Cartões de crédito e atribuição de um limite mensal, menor ou maior tendo em conta o tipo de cliente bem como as suas características.
É de extrema importância que a atribuição de crédito seja feita de forma consciente e com base numa avaliação exímia do cliente, caso contrário o incumprimento do mesmo irá se traduzir em perdas do banco.
Por fim, a imagem seguinte representa o sumário de todas as variáveis presentes no Dataset de estudo.

```
> summary(bank)
 Attrition_Flag  Customer_Age   Gender   Dependent_count Education_Level
 AC:1627        Min.   :26.00   1:5358   Min.   :0.000   3:1013
 EC:8500        1st Qu.:41.00   0:4769   1st Qu.:1.000   6: 451
                Median :46.00            Median :2.000   4:3128
                Mean   :46.33            Mean   :2.346   2:2013
                3rd Qu.:52.00            3rd Qu.:3.000   5: 516
                Max.   :73.00            Max.   :5.000   1:1487
                                                         0:1519
 Marital_Status     Income_Category Card_Category Months_on_book
 Length:10127       6: 727          0:9436        Min.   :13.00
 Class :character   2:1790          2: 116        1st Qu.:31.00
 Mode  :character   3:1402          3:  20        Median :36.00
                    4:1535          1: 555        Mean   :35.93
                    1:3561                        3rd Qu.:40.00
                    0:1112                        Max.   :56.00

 Total_Relationship_Count Months_Inactive_12_mon Contacts_Count_12_mon
 Min.   :1.000            Min.   :0.000          Min.   :0.000
 1st Qu.:3.000            1st Qu.:2.000          1st Qu.:2.000
 Median :4.000            Median :2.000          Median :2.000
 Mean   :3.813            Mean   :2.341          Mean   :2.455
 3rd Qu.:5.000            3rd Qu.:3.000          3rd Qu.:3.000
 Max.   :6.000            Max.   :6.000          Max.   :6.000

  Credit_Limit   Total_Revolving_Bal Avg_Open_To_Buy Total_Amt_Chng_Q4_Q1
 Min.   : 1438   Min.   :   0        Min.   :    3   Min.   :0.0000
 1st Qu.: 2555   1st Qu.: 359        1st Qu.: 1324   1st Qu.:0.6310
 Median : 4549   Median :1276        Median : 3474   Median :0.7360
 Mean   : 8632   Mean   :1163        Mean   : 7469   Mean   :0.7599
 3rd Qu.:11068   3rd Qu.:1784        3rd Qu.: 9859   3rd Qu.:0.8590
 Max.   :34516   Max.   :2517        Max.   :34516   Max.   :3.3970

 Total_Trans_Amt Total_Trans_Ct   Total_Ct_Chng_Q4_Q1 Avg_Utilization_Ratio
 Min.   :  510   Min.   : 10.00   Min.   :0.0000      Min.   :0.0000
 1st Qu.: 2156   1st Qu.: 45.00   1st Qu.:0.5820      1st Qu.:0.0230
 Median : 3899   Median : 67.00   Median :0.7020      Median :0.1760
 Mean   : 4404   Mean   : 64.86   Mean   :0.7122      Mean   :0.2749
 3rd Qu.: 4741   3rd Qu.: 81.00   3rd Qu.:0.8180      3rd Qu.:0.5030
 Max.   :18484   Max.   :139.00   Max.   :3.7140      Max.   :0.9990
```

1. Sumário dataset Bank

2.3 Descrição do Problema a abordar

Face ao domínio acima explicado, podemos aferir que com os dados recolhidos por esta ou outra entidade do mesmo meio é possível fazer análises e previsões que podem ser extremamente importantes.
Por exemplo tentar calcular o cumprimento financeiro de um cliente, não só com intuito de precaver a entidade financeira, mas também com intuito de oferecer o máximo de produtos financeiros ao cliente com vista em maiores lucros para a entidade financeira, mas de forma segura. Este equilíbrio pode ser extremamente vantajoso.
Neste relatório o problema que me propus tentar trabalhar foi prever com base em todos os dados recolhidos, se o cliente irá ou não entrar em incumprimento com o Banco. Tal como referido anteriormente esta análise irá permitir a obtenção de maiores receitas, mas sobretudo proteger o banco no sentido de ter cada vez menos clientes que entram em incumprimento.
Nos dados recolhidos dos clientes já existentes podemos aferir o estado de cliente na coluna "Attrition_Flag" que neste caso tem apenas dois possíveis resultados, "Attrited Customer" (AC) que representa os clientes que estão sinalizados por entrar em incumprimentos e Existing Customer (EC) que representa os clientes com obrigações cumpridas.

3. Critérios de Sucesso

Tendo como base os conhecimentos adquiridos na Unidade Curricular de Metodologias e Técnica de Análise de Dados considero que, face à discrepância no que diz respeito ao balanceamento dos dados, o indicador com mais preponderância seja o Recall com especial atenção no Recall na previsão Existing Customer (EC) , uma vez que prever EC e errar tem um impacto maior na organização do que prever AC e errar.

Com base nesta teoria espero obter um modelo com uma percentagem de Accuracy superior a 90%, um Recall superior a 85% e por fim uma e percentagem de Precision superior a 90%.

4. Machine Learning

4.1 Algoritmo de ML

Tendo em conta o conjunto de dados recolhidos para a execução deste trabalho foram utilizados algoritmos de ML de Aprendizagem Supervisionada uma vez que este tipo de Machine Learning (ML) permite treinar modelos de previsão baseados em dados input (variável independente) e de output (variável dependente)
O algoritmo de ML escolhido foi o Distributed Random Forest (DRF) ainda que também tenham sido treinados modelos com o algoritmo de Deep Learning (DP). A escolha de adotar DRF foi feita com base nas diferenças nestes dois algoritmos, uma vez que DP é mais exigente no que diz respeito a processamento e treino mas também modelo de dados mais complexos como imagens. Por sua vez o algoritmo DRF é menos complexo no que diz respeito à sua análise e em entender qual o comportamento e importância das variáveis do dataset.
Por fim o principal motivo para a escolho do algoritmo de DRF foi o facto de obter melhores modelos com este algoritmo do que com DP, esta decisão teve como base a análise das curvas de aprendizagem, e validação provenientes dos modelos. Para ambos algoritmos foi utilizado a metodologia de Cross Validation de modo a garantir divisão dos dados não dependessem do fator sorte ainda que esta metodologia se possa traduzir em mais tempo no treino do modelo.

4.2 Parâmetros de Configuração dos Modelos de ML

4.2.1 Deep Learning

No treino de modelos DL foram utilizados os parâmetros standard à exceção dos identificados na imagem seguinte.


```
buildModel 'deeplearning',
{"model_id":"dp_200_100_10e","training_frame":"bankag.train","validation_frame":"bankag.validation","nfolds":5,"response_column":"
Attrition_Flag","ignored_columns":[],"ignore_const_cols":true,"activation":"Rectifier","hidden":
[200,100],"epochs":10,"variable_importances":true,"fold_assignment":"AUTO","score_each_iteration":true,"balance_classes":true,"use
_all_factor_levels":true,"standardize":true,"train_samples_per_iteration":-2,"adaptive_rate":true,"input_dropout_ratio":0,"l1":0,"
l2":0,"loss":"Automatic","distribution":"AUTO","huber_alpha":0.9,"score_interval":5,"score_training_samples":10000,"score_validati
on_samples":0,"score_duty_cycle":0.1,"stopping_rounds":5,"stopping_metric":"AUTO","stopping_tolerance":0,"max_runtime_secs":0,"aut
oencoder":false,"categorical_encoding":"AUTO","auc_type":"AUTO","keep_cross_validation_models":true,"keep_cross_validation_predict
ions":false,"keep_cross_validation_fold_assignment":false,"class_sampling_factors":
[],"max_after_balance_size":5,"target_ratio_comm_to_comp":0.05,"seed":-1,"rho":0.99,"epsilon":1e-
8,"nesterov_accelerated_gradient":true,"max_w2":3.4028235e+38,"initial_weight_distribution":"UniformAdaptive","classification_stop
":0,"score_validation_sampling":"Uniform","diagnostics":true,"fast_mode":true,"force_load_balance":true,"single_node_mode":false,"
shuffle_training_data":false,"missing_values_handling":"MeanImputation","quiet_mode":false,"sparse":false,"col_major":false,"avera
ge_activation":0,"sparsity_beta":0,"max_categorical_features":2147483647,"reproducible":false,"export_weights_and_biases":false,"m
ini_batch_size":1,"elastic_averaging":false}
```


2. Parametros de Modelos DL

Com base na figura acima, podemos aferir que os parâmetros alterados foram : “nfolds” que representa o número de blocos em que os dados foram divididos no dataset de treino, atribuição da variável dependente no campo “response_column, escolha do numero de neurónios utilizados na camada de input para trabalhar os dados, “epochs” número de passagens pelo dataset no processo de aprendizagem ,score_each_iterarion : true” permite-nos marcar no gráfico de aprendizagem cada iteração do modelo época após época, foi ainda ativado o parâmetro de balanceamento dos dados uma vez que o dataset de estudo tem um grande discrepância no que ao número de resultados diz respeito.

4.2.2 Distributed Random Forest

No treino de modelos DRF foram utilizados os parâmetros standard à exceção dos identificados na imagem seguinte.


```
buildModel 'drf',
{"model_id":"drf_50_20_0,3_bc","training_frame":"bankag.train","validation_frame":"bankag.validation","nfolds":5,"response_column"
:"Attrition_Flag","ignored_columns":
[],"ignore_const_cols":true,"ntrees":50,"max_depth":20,"min_rows":1,"nbins":20,"seed":-1,"mtries":-1,"sample_rate":0.3,"score_each
_iteration":true,"score_tree_interval":0,"fold_assignment":"AUTO","balance_classes":true,"nbins_top_level":1024,"nbins_cats":1024,
"r2_stopping":1.7976931348623157e+308,"stopping_rounds":0,"stopping_metric":"AUTO","stopping_tolerance":0.001,"max_runtime_secs":0
,"col_sample_rate_per_tree":1,"min_split_improvement":0.00001,"histogram_type":"AUTO","categorical_encoding":"AUTO","distribution"
:"AUTO","gainslift_bins":-1,"auc_type":"AUTO","keep_cross_validation_models":true,"keep_cross_validation_predictions":false,"keep_
cross_validation_fold_assignment":false,"class_sampling_factors":
[],"max_after_balance_size":5,"build_tree_one_node":false,"sample_rate_per_class":
[],"binomial_double_trees":false,"col_sample_rate_change_per_level":1,"calibrate_model":false,"calibration_method":"AUTO","check_c
onstant_response":true}
```

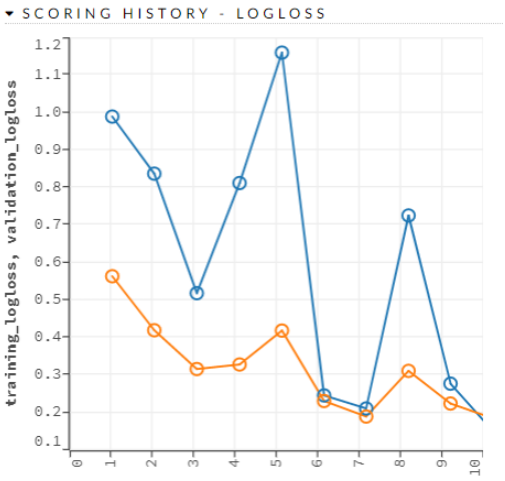

***3**. Parametros de Modelos DRF*

No treino de modelos DRF os parâmetros alterados foram os acima sinalizados: “ntrees” que define o número de árvores na camada de input, “maxdepth” que indica o número máximo de profundidade de cada árvore e por fim “sample_rate” que define a percentagem de amostra de dados face ao dataset de treino que cada árvore vai receber como input. Salientar ainda, ainda que neste modelo DRF todos os parâmetros que se repetem face ao modelo de DP foram igualmente alterados.

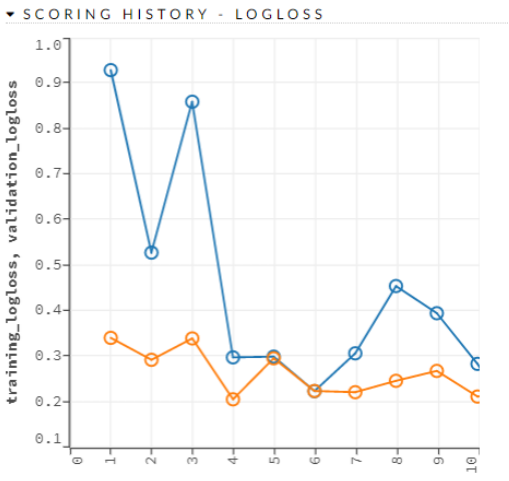
4.3 Avaliação dos Modelos

4.3.1 Deep Learning

Na elaboração de Modelos com o algoritmo DL não foi conseguido obter modelos cujas curvas de aprendizagem nos indicassem viabilidade. Tal como é possível aferir nas imagens 4 e 5 podemos concluir que ambos os modelos não são representativos, muito provavelmente porque não existem dados suficientes para representar padrões. Ainda assim na como medida de tentativa de diminuir ruído nos modelos e que a configuração do mesmo fosse muito complexa e fizesse com que o modelo ao invés de aprender decorasse, defini que houvesse menos neurónios na camada de input no segundo modelo, imagem 5.



4. Modelo DL 200_100



5. Modelo DL 100_50

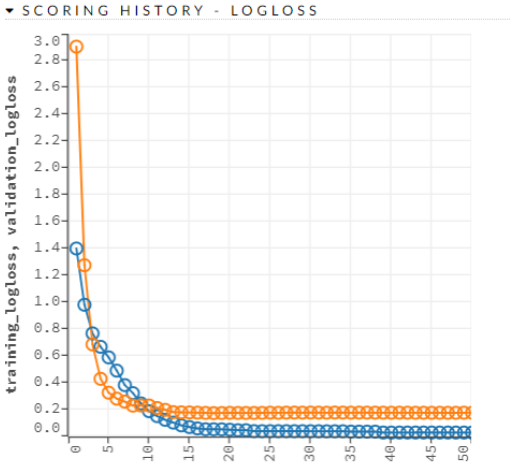
4.3.2 Distributed Random Forest

Tendo em conta os dois tipos de modelos abordados no relatório, bem como os resultados dos modelos DL foi no modelo DRF que obtive melhores resultados. Foi por isso também que a busca na otimização do modelo também foi mais enfatizada nos modelos DRF.

Tal como mencionado acima no relatório existe uma grande discrepância no dataset no que à variável dependente diz respeito. Existem 2 resultados possíveis Attrited Customer (AC) e Existing Costumer (EC), sendo que cerca de 80% da amostra é de cliente EC e 20% (AC). Ainda assim no primeiro modelo não foi ativado o parâmetro de balanceamento dos dados e o resultado das curvas está representado na figura 7. Podemos aferir que ainda que o gap nas curvas (treino e validação) é pequeno o que é positivo ainda assim gostaria de ver a sobreposição da curva de validação face à de treino nas primeiras árvores.

Parâmetros:

Ntrees = 50
Max_depth = 20
Sample_rate = 0,6
Balance_classes = False



7. Modelo DRF_50_20_0,6

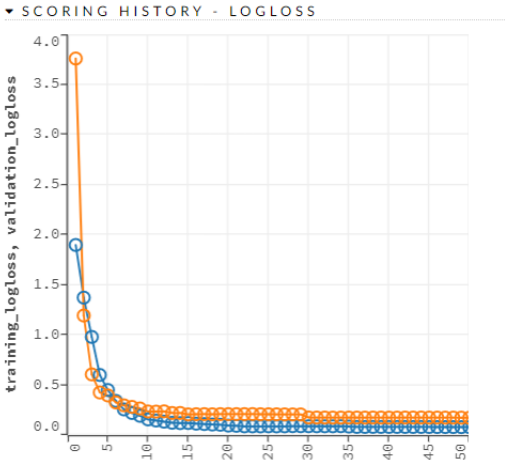
VALIDATION METRICS

	AC	EC	Error	Rate	Precision
AC	253	54	0.1759	54 / 307	0.95
EC	14	1701	0.0082	14 / 1.715	0.97
Total	267	1755	0.0336	68 / 2.022	
Recall	0.82	0.99			

6. VM DRF_50_20_0,6

Com intuito de resolver a sobreposição acima mencionada, e tentar cada vez mais diminuir o gap entre curvas resolvi ativar o balanceamento de dados e diminuir a amostragem de dados que cada árvore recebe para que exista ainda menos probabilidade de o modelo “decorar” padrões.

Ntrees = 50
Max_depth = 20
Sample_rate = 0,4
Balance_classes = True



9. Modelo DRF_50_20_0,4_bc

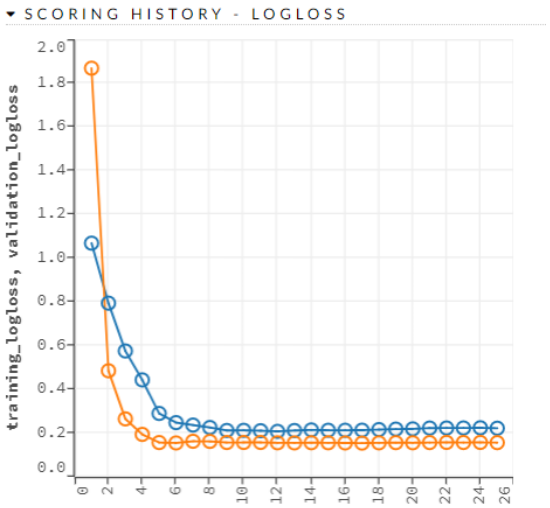
VALIDATION METRICS -

	AC	EC	Error	Rate	Precision
AC	271	36	0.1173	36 / 307	0.88
EC	38	1677	0.0222	38 / 1.715	0.98
Total	309	1713	0.0366	74 / 2.022	
Recall	0.88	0.98			

8. VM_50_20_0,4_bc

Com base nas imagens 8 e 9 é válido afirmar que os objetivos anteriores foram concretizados. No modelo a seguir foi definido como objetivo diminuir a complexidade do modelo, ou seja, reduzir o número de árvores e a profundidade das mesmas.

Ntrees = 25
Max_depth = 10
Sample_rate = 0,4
Balance_classes = True



11. Modelo DRF_25_10_0,4_bc

VALIDATION METRICS

	AC	EC	Error	Rate	Precision
AC	280	27	0.0879	27 / 307	0.82
EC	62	1653	0.0362	62 / 1.715	0.98
Total	342	1680	0.0440	89 / 2.022	
Recall	0.91	0.96			

10. VM_DRF_25_10_0,4_bc

É possível afirmar que a diminuição da complexidade em nada ajudou na qualidade do modelo, uma vez que o gap entre as curvas aumentou e a curva de treino sobrepôs-se à curva de validação.

4.4 Escolha do Modelo

Com base em todos os modelos treinados no ponto anterior foi escolhido para "produção" o modelo da figura 8. Ainda que os critérios se tenham baseado essencialmente na interpretação das curvas de treino e validação outras métricas também devem ser analisadas.

Antes de abordar métricas como Precision e Recall importa referir que no contexto estudado a importância de Falsos AC e Falsos EC, ou seja, se o modelo prever que o cliente irá incumprir no pagamento AC, mas na verdade ele até cumpre EC não é tao mau como o modelo prever que o cliente irá cumprir, EC, e vir-se a verificar que este incumpriu AC.

Posto isto, neste contexto procuro uma percentagem de Recall elevada na previsão de EC ainda que a percentagem de Precision naturalmente diminua. Assumindo a introdução acima é mais prejudicial prever EC e o cliente vir a ser AC. Salientar por fim que a Accuracy deste modelo é dado pela fórmula TAC + TEC / TAC + TEC + FAC + FEC, ou seja 307 + 1715 / 307 + 1715 + 36 + 38 = 2022/2096 = 0,964. Este indicador o a taxa de acerto do modelo.

5. Conclusão

Em conclusão este estudo permitiu através da plataforma H2O e através de R testar vários algoritmos de Machine Learning e obter um modelo que nos permitisse, através dos conceitos lecionados na Unidade Curricular, perceber se tem qualidade para implementação e posteriormente que conhecimento e mais valias traria para o nosso cliente com o objetivo de obter mais e melhores ferramentas para a tomada de decisão. Por fim e dada a complexidade do problema, foi-me proposto que de forma mais intuitiva e prática, através de uma aplicação Shiny conseguisse mostrar algumas destas ferramentas e conhecimento ao cliente.