Universidade Federal da Bahia
ECD - Especialização em Ciência de Dados e Big Data

CHURN MODELLING: PREDIÇÃO DE CHURN PARA UMA BASE DE DADOS DE INSTITUIÇÃO FINANCEIRA

Nome: Ive de Oliveira Gavazza
Orientador: Prof. Dr. Ricardo Rocha

Introdução

Churn UK:ˈ/ˈtʃɜːrn/US:/tʃɜ-n/ ,(chûrn)

business (customers lost) (*anglicismo*)	taxa de churn *loc sf*	
		taxa de cancelamento *loc sf*
		taxa de evasão de clientes *loc sf*
		taxa de perda de clientes *loc sf*
	Prior churn had an adverse affect on this year's profits.	

UFBA

ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

Introdução

Por que analisar o churn

Custo do churn

Áreas de negócio que mais se beneficiam da análise do churn

UFBA

ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

Objetivo

Identificar numa base de dados de instituição financeira o perfil do cliente que poderá cancelar sua conta, saindo da base de clientes.

UFBA

ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

Dados utilizados

Dataset: “Churn_modelling.csv”

10.000 linhas e 13 variáveis

Quadro 1 - Descrição do Dataset

Nome da variável	Tipo de variável	Descrição da variável
CostumerId	quantitativa	Código numérico que identifica o cliente de forma única
Surname	Categórica	Sobrenome do cliente
CreditScore	quantitativa	Score creditício do usuário
Geography	Categórica	Localidade do usuário.
Age	quantitativa	Idade do usuário
Gender	Categórica	Identificação de gênero do usuário
Tenure	quantitativa	Período de fechamento de um empréstimo ou parcela do usuário.
Balance	quantitativa	Valor de saldo em conta
NumOfProducts	quantitativa	Número de produtos que o usuário tem na instituição
HasCrCard	quantitativa	Se possui cartão de crédito
IsActiveMember	quantitativa	Se é membro ativo da plataforma da instituição
EstimatedSalary	quantitativa	Salário estimado ao ano
Exited	quantitativa	Se saiu da base de clientes

Fonte: Elaboração própria

UFBA

ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

Dados utilizados

Necessidade de Normalização dos dados

Quadro 2 - Medidas básicas do dataset

	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Female	France	Germany
Count	10.000	10.000	10.000	10.000	10.000	10.000	10.000	10.000	10.000	10.000	10.000
Mean	650,5288	38,9218	5,0128	76.486	1,5302	0,7055	0,5151	100.090	0,4543	0,5014	0,2509
Std	96,6533	10,4878	2,8922	62.397	0,5817	0,4558	0,4998	57.510	0,4979	0,5000	0,4336
Min	350	18	0	0	1	0	0	11,58	0	0	0
25,00%	584	32	3	0	1	0	0	51.002	0	0	0
50,00%	652	37	5	97.199	1	1	1	100.194	0	1	0
75,00%	718	44	7	127.644	2	1	1	149.388	1	1	1
Max	850	92	10	250.898	4	1	1	199.992	1	1	1

Fonte: Elaboração própria

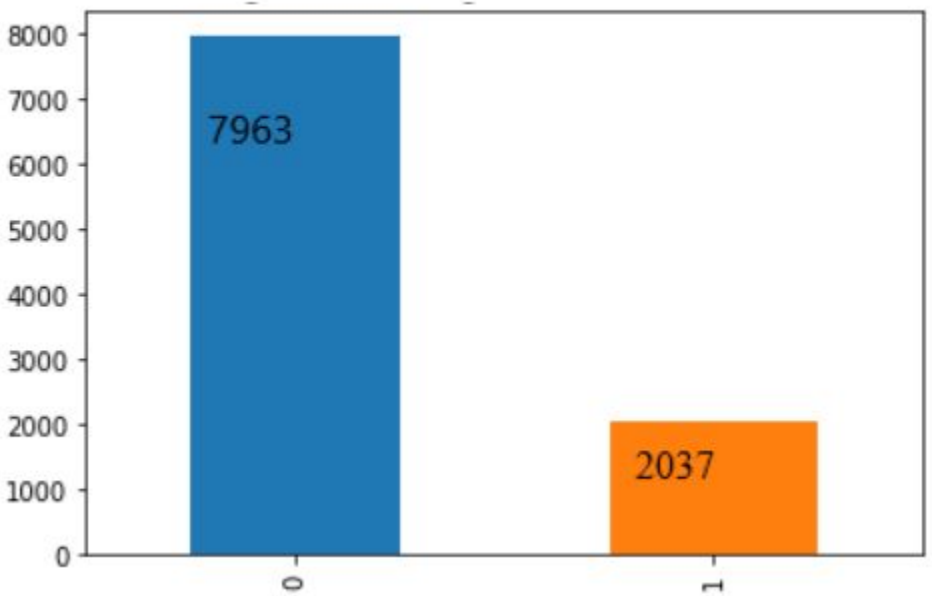
UFBA

ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

Análise exploratória

Análise da variável resposta

Figura 1 - Contagem variável "Exited"



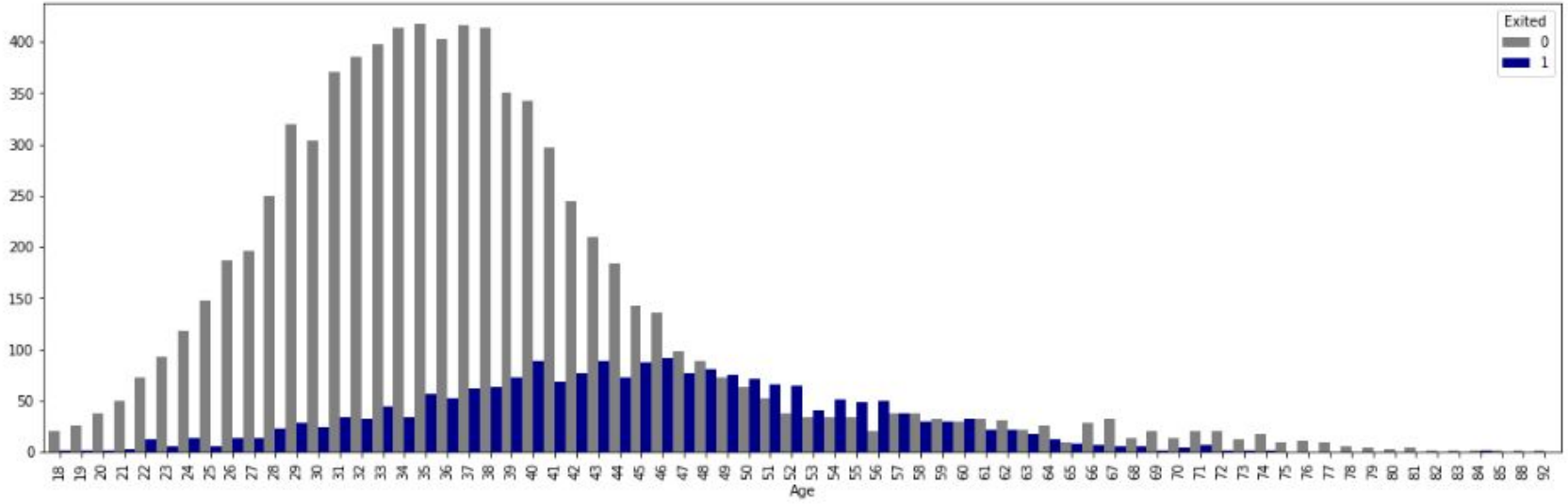
Fonte: Elaboração própria.

UFBA

ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

Análise exploratória

Figura 3 - Distribuição das idades com relação a variável "Exited"



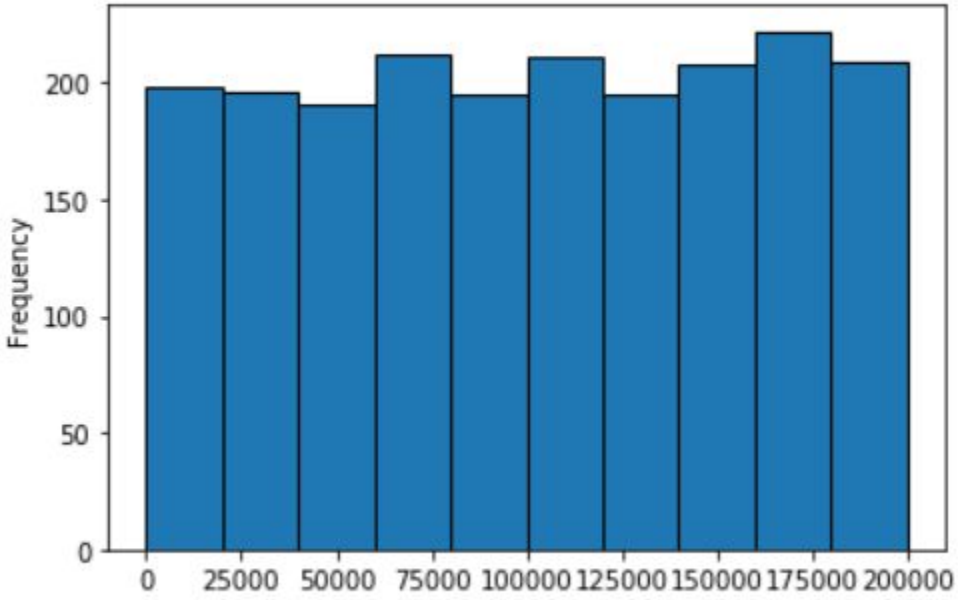
Fonte: Elaboração própria.

UFBA

ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

Análise exploratória

Figura 4 - Distribuição do salário anual estimado com relação a variável "Exited"



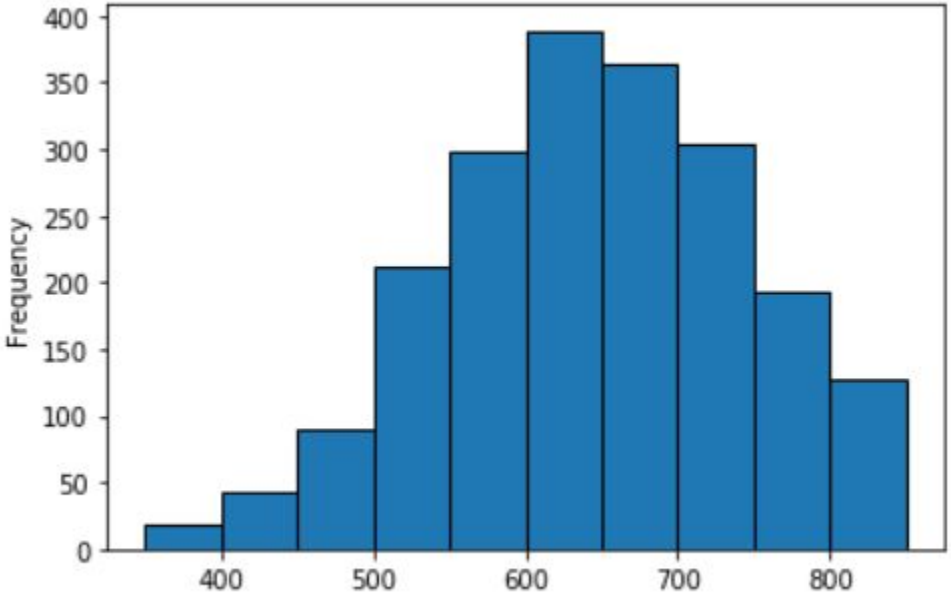
Fonte: Elaboração própria

UFBA

ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

Análise exploratória

Figura 5 - Distribuição do score de crédito com relação a variável "Exited"



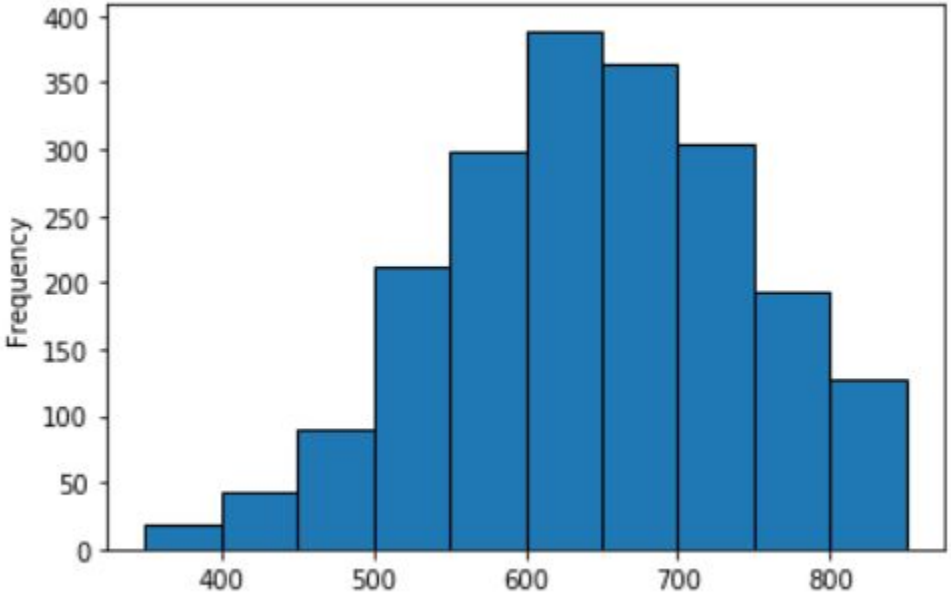
Fonte: Elaboração própria

UFBA

ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

Análise exploratória

Figura 5 - Distribuição do score de crédito com relação a variável "Exited"



Fonte: Elaboração própria

UFBA

ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

Feature Engineering

Principais problemas:

Desbalanceamento dos dados

Dados não normalizados

Dados com pouca relevância para a análise de churn

UFBA

ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

Feature Engineering

Setup 1

Retirada de valores nulos

Tratamento de variáveis não numéricas

Retirada de variáveis não relevantes para o presente estudo

UFBA

ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

Feature Engineering

Setup 2

Configuração do Setup 1

SMOTE - Synthetic Minority Over-sampling Technique

UFBA

ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

Feature Engineering

Setup 3

Análise bivariada

Quadro 4 - Análise bivariada

	RowNumber	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited	Female	France	Germany
RowNumber	1.0	0.0058	0.00078	-0.0065	-0.0091	0.0072	0.0006	0.012	-0.006	-0.017	-0.018	0.0086	-4.4e-05
CreditScore	0.0058	1.0	-0.004	0.00084	0.0063	0.012	-0.0055	0.026	-0.0014	-0.027	0.0029	-0.0089	0.0055
Age	0.00078	-0.004	1.0	-0.01	0.028	-0.031	-0.012	0.085	-0.0072	0.29	0.028	-0.039	0.047
Tenure	-0.0065	0.00084	-0.01	1.0	-0.012	0.013	0.023	-0.028	0.0078	-0.014	-0.015	-0.0028	-0.00057
Balance	-0.0091	0.0063	0.028	-0.012	1.0	-0.3	-0.015	-0.01	0.013	0.12	-0.012	-0.23	0.4
NumOfProducts	0.0072	0.012	-0.031	0.013	-0.3	1.0	0.0032	0.0096	0.014	-0.048	0.022	0.0012	-0.01
HasCrCard	0.0006	-0.0055	-0.012	0.023	-0.015	0.0032	1.0	-0.012	-0.0099	-0.0071	-0.0058	0.0025	0.011
IsActiveMember	0.012	0.026	0.085	-0.028	-0.01	0.0096	-0.012	1.0	-0.011	-0.16	-0.023	0.0033	-0.02
EstimatedSalary	-0.006	-0.0014	-0.0072	0.0078	0.013	0.014	-0.0099	-0.011	1.0	0.012	0.0081	-0.0033	0.01
Exited	-0.017	-0.027	0.29	-0.014	0.12	-0.048	-0.0071	-0.16	0.012	1.0	0.11	-0.1	0.17
Female	-0.018	0.0029	0.028	-0.015	-0.012	0.022	-0.0058	-0.023	0.0081	0.11	1.0	-0.0068	0.025
France	0.0086	-0.0089	-0.039	-0.0028	-0.23	0.0012	0.0025	0.0033	-0.0033	-0.1	-0.0068	1.0	-0.58
Germany	-4.4e-05	0.0055	0.047	-0.00057	0.4	-0.01	0.011	-0.02	0.01	0.17	0.025	-0.58	1.0

Fonte: Elaboração própria

UFBA

ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

Feature Engineering

Setup 3

Setup 1 e Setup 2

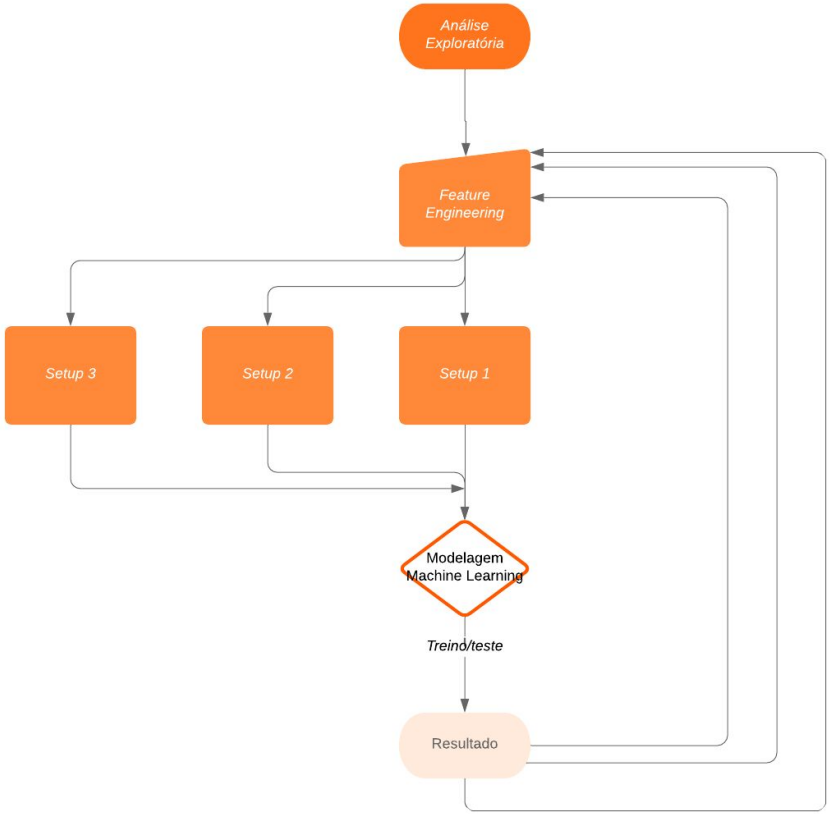
Selecionadas apenas as variáveis com maior correlação com “Exited”: “Age”, “IsActiveMember”, “Balance”, “NumberOfProducts”, “CreditScore” e nas dummies de localidade.

UFBA

ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

Modelagem Machine Learning

Pipeline de análise



UFBA

ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

Modelagem Machine Learning

Regressão Logística

KNN

SVM

Naive Bayes

Árvore de Decisão

Random Forest

Artificial Neural Network

UFBA

ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

Resultados

Visualização dos resultados: Matriz de confusão

Score de custo da predição: multiplicador de 6 para erro e multiplicador de 1 para acerto.

Métricas: Precision ou Precisão, Recall ou Revogabilidade e Accuracy ou Acurácia.

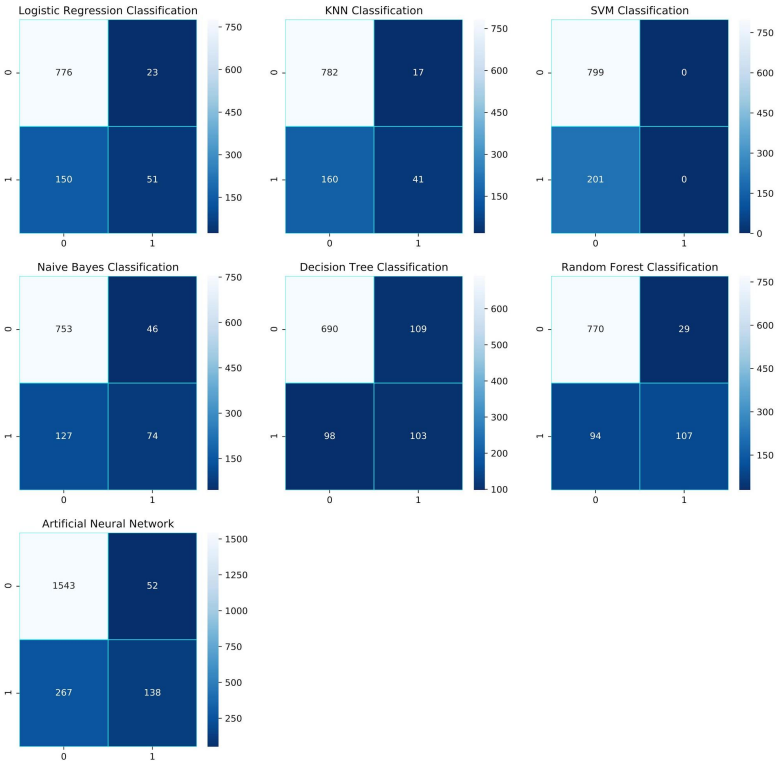
UFBA

ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

Resultados

Setup 1

Figura 9 - Matriz de confusão - Modelo 1



Fonte: Elaboração própria.

Resultados

Precision e recall altos valores para "Exited" = 0 e baixos para "Exited" = 1, que é o padrão buscado.

Quadro 5 - Métricas - Modelo 1

LR	precision	recall	f1-score	accuracy
	0 0.97	0.84	0.90	0.83
	1 0.25	0.69	0.37	
KNN:	precision	recall	f1-score	accuracy
	0 0.98	0.83	0.90	0.82
	1 0.20	0.71	0.32	
SVM:	precision	recall	f1-score	accuracy
	0 1.00	0.80	0.89	0.80
	1 0.00	0.00	0.00	
NB	precision	recall	f1-score	accuracy
	0 0.94	0.86	0.90	0.83
	1 0.37	0.62	0.46	
DT	precision	recall	f1-score	accuracy
	0 0.86	0.87	0.87	0.79
	1 0.48	0.47	0.48	
RF	precision	recall	f1-score	accuracy
	0 0.96	0.89	0.93	0.88
	1 0.53	0.79	0.64	
ANN	precision	recall	f1-score	accuracy
	0.796537	0.454321		0.80

Fonte: Elaboração própria.

UFBA

ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

Resultados

Melhor modelo: ANN

Base de dados desbalanceada para casos de usuários que não deram churn faz com que ao detectar os altos números de casos verdadeiros negativos (VN) o modelo acerte muitas vezes.

Quadro 6 - Score do Setup 1

Setup	Classificador	FP	FN	Custo do Erro	VN	VP	Custo do Acerto	Score
1	LR	150	23	288	776	51	1064	776
1	KNN	160	17	262	782	41	1044	782
1	SVM	201	0	201	799	0	1000	799
1	NB	127	46	403	753	74	1156	753
1	DT	94	109	748	690	103	1438	690
1	RF	94	29	268	770	107	1038	770
1	ANN	267	52	579	1543	138	2122	1543

Fonte: Elaboração própria.

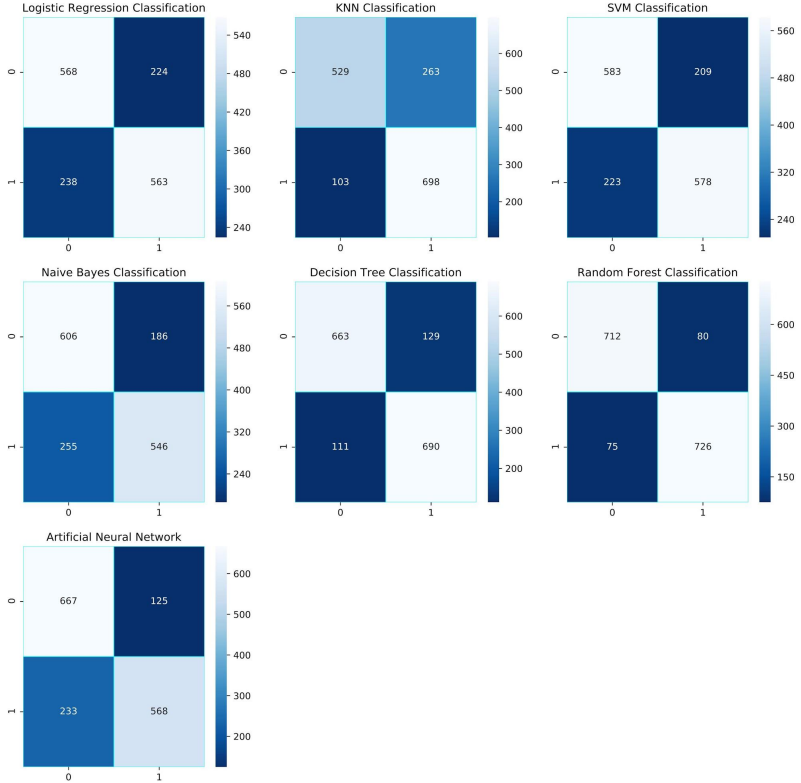
UFBA

ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

Resultados

Setup 2

Figura 10 - Matriz de confusão - Modelo 2



Fonte: Elaboração própria.

Resultados

Melhora geral das métricas para a detecção de respostas iguais a 1, cumprindo o objetivo de melhorar o conhecimento dos modelos quanto ao churn na base de clientes do banco.

Quadro 7 - Métricas - Modelo 2

LR	precision	recall	f1-score	accuracy
	0 0.72	0.72	0.72	0.72
	1 0.72	0.72	0.72	
KNN:	precision	recall	f1-score	accuracy
	0 0.68	0.86	0.76	0.79
	1 0.89	0.74	0.81	
SVM:	precision	recall	f1-score	accuracy
	0 0.74	0.74	0.74	0.74
	1 0.75	0.74	0.75	
NB	precision	recall	f1-score	accuracy
	0 0.77	0.72	0.74	0.74
	1 0.71	0.75	0.73	
DT	precision	recall	f1-score	accuracy
	0 0.82	0.85	0.83	0.84
	1 0.86	0.83	0.84	
RF	precision	recall	f1-score	accuracy
	0 0.90	0.90	0.90	0.90
	1 0.90	0.90	0.90	
ANN	precision	recall	f1-score	accuracy
	0.779026	0.779026		0.50

Fonte: Elaboração própria.

UFBA

ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

Resultados

Melhor modelo: Random Forrest

Conseguiu reduzir o erro mantendo o mesmo nível de acerto, enquanto outros modelos apesar de terem aumentado significativamente seus acertos fizeram o mesmo com os erros, gerando um impacto quase nulo no score, ficando no mesmo patamar anterior ou até mesmo reduzindo seu score.

Quadro 8 - Score do Setup 2

Setup	Classificador	FP	FN	Custo do Erro	VN	VP	Custo do Acerto	Score
2	LR	238	224	1.582	568	563	2150	568
2	KNN	103	263	1.681	529	698	2210	529
2	SVM	223	209	1.477	583	578	2060	583
2	NB	255	186	1.371	606	546	1977	606
2	DT	111	129	885	663	690	1548	663
2	RF	75	80	555	712	726	1267	712
2	ANN	233	125	983	667	568	1650	667

Fonte: Elaboração própria.

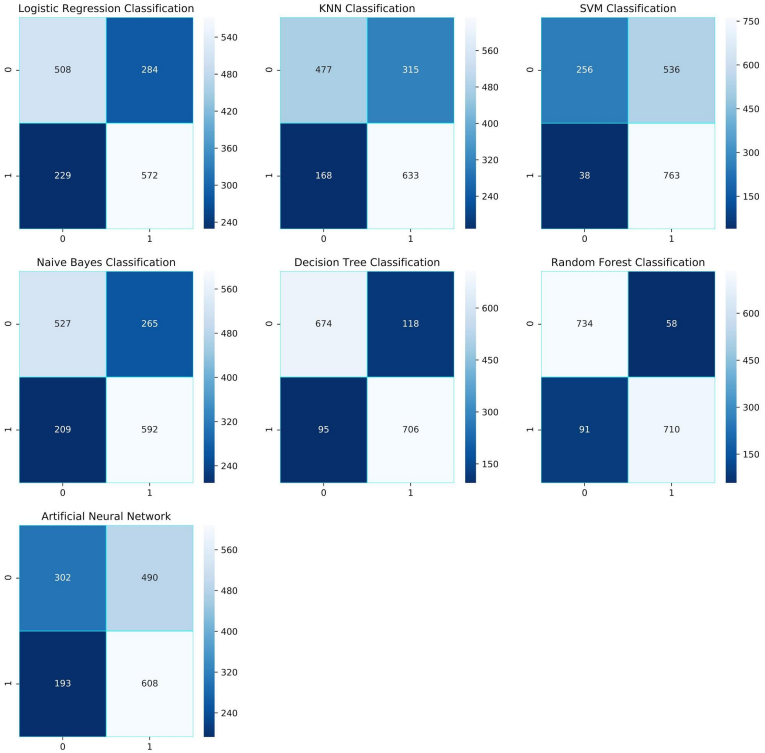
UFBA

ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

Resultados

Setup 3

Figura 11 - Matriz de confusão - Modelo 3



Fonte: Elaboração própria.

Resultados

Piora nas métricas dos modelos, denotando que apesar de selecionadas apenas as variáveis de correlação mais fortes com a variável resposta os modelos se beneficiam mais da inclusão de todas as variáveis do dataset.

Quadro 9 - Métricas - Modelo 3

LR	precision	recall	f1-score	accuracy
	0 0.64	0.68	0.66	0.67
	1 0.71	0.66	0.68	
KNN:	precision	recall	f1-score	accuracy
	0 0.60	0.73	0.66	0.69
	1 0.78	0.66	0.72	
SVM:	precision	recall	f1-score	accuracy
	0 0.32	0.90	0.48	0.65
	1 0.96	0.59	0.73	
NB	precision	recall	f1-score	accuracy
	0 0.66	0.71	0.69	0.70
	1 0.74	0.69	0.71	
DT	precision	recall	f1-score	accuracy
	0 0.86	0.88	0.87	0.87
	1 0.88	0.86	0.87	
RF	precision	recall	f1-score	accuracy
	0 0.92	0.89	0.91	0.91
	1 0.89	0.92	0.90	
ANN	precision	recall	f1-score	accuracy
	0.566916	0.851436		0.50

Fonte: Elaboração própria.

UFBA

ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

Resultados

Melhor modelo: Random Forrest

consegue aumentar seu nível de acerto mantendo o nível de erro do setup anterior, diferente dos outros modelos que aumentaram erro e acerto em proporção semelhante não impactando o resultado final acumulado.

Importante notar que o Setup 3 rebaixa os escores de todos os modelos.

Quadro 10 - Score do Setup 3

Setup	Classificador	FP	FN	Custo do Erro	VN	VP	Custo do Acerto	Score
3	LR	229	284	1.933	508	572	2441	508
3	KNN	168	315	2.058	477	633	2535	477
3	SVM	38	536	3.254	256	763	3510	256
3	NB	209	265	1.799	527	592	2326	527
3	DT	95	118	803	674	706	1477	674
3	RF	91	58	439	734	710	1173	734
3	ANN	193	490	3.133	302	608	3435	302

Fonte: Elaboração própria.

UFBA

ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

Considerações Finais

Para o problema de churn é importante que o algoritmo consiga balancear um falso positivo baixo, mantendo a média de acerto para que os custos de não prever a saída de um cliente da base seja minimizado, mantendo a média de verdadeiro positivos. No caso dos modelos estudados o modelo Random Forest foi o mais balanceado em entregar uma boa predição de verdadeiros positivos com o menor custo de erro.

Para a completar a análise do churn recomenda-se a implementação do modelo com o objetivo de se tornar uma ferramenta de predição para novos dados.

UFBA

ECD
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E BIG DATA