

Dados na Ciência Gestão e Sociedade

Relatório Final: Marketing Bancário

Trabalho realizado por:

Gonçalo Henriques, 123422

José Alberto, 121959

Jonasse Mbaki, 111900

ÍNDICE:

1.	INTRODUÇÃO.....	3
2.	CRISP-DM.....	5
2.1.	Business Understanding.....	6
2.1.1.	Marketing Bancário.....	6
2.2.	Data Understanding.....	7
2.3.	Data Preparation.....	10
2.4.	Modeling	15
3.	Bibliografia.....	20

1. INTRODUÇÃO

No âmbito da unidade curricular Dados na Ciência Gestão e Sociedade, durante o primeiro trimestre 2023/2024, foi proposto a realização de um Relatório referente a um projeto que incidiu no tema, “Marketing Bancário”. O mesmo pretende constatar todos os conhecimentos retidos durante o respetivo período de aulas, bem como, a metodologia CRISP-DM, e assim, obter uma melhor compreensão sobre a utilização e análise de um conjunto de dados.

Como tema, o Marketing Bancário, este projeto foi elaborado com a visão de otimizar as estratégias e ações neste campo que desempenha um papel tão crucial na atração e obtenção de clientes visto que se trata de um setor financeiro altamente competitivo.

A base de dados fornecida, compreende dados retirados do mundo real que estão publicamente disponíveis a fins de pesquisa (open data source), inclui dados demográficos e comportamentais, bem como dados relacionados a campanhas de marketing anteriores realizadas. Como tal, estas campanhas têm como foco a previsão da subscrição de depósitos a prazo por parte dos clientes, pelo que nos foi fácil definir qual seria o nosso target entre as variáveis que nos foram dadas.

De modo introdutório, este relatório pretende abordar os seguintes aspetos-chaves: exploração e análise dos dados para compreender a distribuição das variáveis e identificação de tendências, desenvolvimento e avaliação de modelos com base em métricas de desempenho definidas e explicar o processo para a construção do algoritmo de previsão.

Foram definidas previamente questões a qual o modelo preditivo pretende dar resposta, entre elas:

- Qual seria o perfil “ideal” para que um cliente subscrisse a um depósito a prazo?
- Qual das variáveis disponíveis no nosso conjunto de dados tem maior influência sobre o target definido?

- Qual será a melhor altura do ano para fazer uma campanha no marketing bancário?

Os aspetos-chaves anteriormente mencionados foram trabalhados na plataforma solicitada, Orange, que é uma ferramenta que permite a visualização de dados.

2. CRISP-DM

Business Intelligence (BI) é um termo abrangente que engloba, ferramentas, bases de dados e importantes metodologias com o objetivo de utilizar dados para apoiar uma decisão segura. Data-Mining (DM) é uma tecnologia de BI, que utiliza modelos com base em dados para extrair conhecimentos úteis. As técnicas referidas servem para aprimorar as campanhas deste tipo de marketing.

A CRISP-DM (Cross Industry Standard Process for Data Mining) é usada como procedimento padrão da Indústria para mineração de dados (Data-Mining). Inclui uma sequência de 6 fases não-rígidas, que visam a implementação de um modelo que representa o conhecimento apreendido, ajudando a toma de diversas decisões. Esta metodologia fornece uma abordagem estruturada para planejar o estudo e análise dos dados.

As respectivas fases são:

Business Understanding;

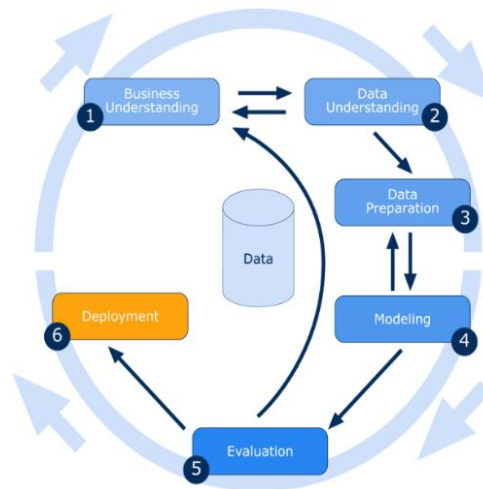
Data Understanding;

Data Preparation;

Modelling;

Evaluation;

Employment;



2.1. Business Understanding

A referida metodologia não implica o seu começo por esta fase, contudo é bastante importantíssima para definir os objetivos, a natureza do problema e dos requisitos do tema tratado, Marketing bancário.

Como o tema incide num assunto bastante importante, foi feito inicialmente um estudo de pesquisa aprofundado, de modo, a compreender o assunto em questão, definir o objetivo, identificar as necessidades do negócio e considerar questões éticas.

2.1.1. Marketing Bancário

Como foi referido este setor apresenta um papel muito importante no que diz respeito à obtenção dos clientes. Deste modo, as empresas para promoverem os seus serviços/produtos utilizam 2 tipos de abordagens, através de campanhas em massa que englobam um público, em geral, indiscriminado ou através de marketing direto direcionado a um conjunto específico de contactos. Atualmente as respostas positivas as campanhas em massa são muito baixas, enquanto que, o marketing direto concentra-se em alvos que supostamente estarão mais interessados no respetivo serviço, tornando a campanha mais eficiente.

Esta segunda abordagem apresenta uma maior desvantagem, pois podem gerar respostas negativas por parte dos clientes, devido à invasão de privacidade a que estes estão sujeitos.

Devido à crise financeira existente nos últimos anos, os bancos europeus têm sofrido uma grande pressão em aumentar os seus ativos financeiros, como tal, a estratégia adotada é oferecer aplicações de depósito a longo prazo com boas taxas de juros por meio de campanhas de marketing direto.

O conjunto de dados fornecido foi obtido através de chamadas telefónicas, durante as quais foram oferecidas as aplicações de depósito a longo prazo.

O objetivo deste tipo de marketing, é melhorar a eficiência através da redução significativa do número de chamadas realizadas, obtendo uma maior

taxa de sucesso, ou seja, uma maior subscrição de depósitos a longo prazo por parte da população.

2.2. Data Understanding

Após um estudo aprofundado acerca do tema do trabalho os dados que nos foram fornecidos apresentam uma grande relevância para responder às questões elaboradas pelo grupo.

Esta fase permite familiarizarmo-nos com o conjunto de dados, perceber o significado e o valor de cada variável, analisar propriedades dos atributos, descrevê-los e explorá-los, analisar estatísticas, aferindo sempre as qualidades dos mesmos, resumindo, esta fase foca-se em alcançar um conhecimento aprofundado dos dados.

O nosso conjunto de dados é constituído, inicialmente, por 16 variáveis independentes, e 1 variável dependente(target). Deste modo, as variáveis independentes (features) são: Age, Job, Marital, Education, Default, Balance, Housing, Loan, Contact, Day, Month, Duration, Campaign, Pdays, Previous e Poutcome, o target é a variável y (define a resposta do cliente ao contacto).

A nossa base de dados tinha, inicialmente, um total de 7 variáveis **numéricas**, agora vamos abordar sobre cada uma delas:

Age: Representa a idade dos clientes que foram contactados durante as campanhas realizadas, que varia dos 18 aos 95 anos. O estudo desta variável mostrou que a distribuição de clientes que subscreveram/ não subscreveram não tem uma grande disparidade, pelo que, esta *feature* não tem grande impacto sobre o target considerado.

Balance: Esta variável representa o saldo médio anual de cada um dos clientes na altura em que foram contactados, em princípio, variavam entre -8019 até 102127 euros. Tivemos que considerar excluir certos intervalos desta variável, o que será tratado mais adiante.

Day: Esta variável representa o dia do mês em que o cliente foi contactado pela última vez. Esta variável toma o mesmo comportamento da variável “Age”,

visto que, a distribuição de taxa de sucesso/negação não varia significativamente.

Duration: Corresponde à duração da chamada realizada. No estudo desta variável, consoante o tema incidido, verificou-se que quanto mais durar a chamada maiores são as probabilidades de adesão.

Campaign: A variável indica o número de contactos realizados durante a campanha decorrente. Segundo estudo da mesma, esta variável tem um grande impacto sobre o target em estudo.

Pdays: A variável diz respeito aos dias que se passaram desde que um referido cliente foi contactado pela última vez na campanha anterior a atual. Obtivemos o valor **-1** nos clientes que nunca tiveram sido contactados antes da presente campanha. Depois de estudos realizados verificamos que houve uma maior taxa de aceitação dos usuários que já tinham sido contactados em campanhas antecedentes.

Previous: Esta variável retrata o número de contactados realizados para um determinado cliente na campanha anterior. Análises apontaram uma forte correlação entre a respetiva variável e a variável “*Pdays*”, pelo que se considerou a exclusão de uma destas variáveis, o que será abordado

Com esta terminou a breve abordagem sobre as variáveis numéricas e abordaremos agora as variáveis **categóricas** onde estará também incluído o target em estudo.

Job: Corresponde a profissão do cliente contactado. Esta variável contém um total de 13 atributos onde temos, administrador (*admin.*), desconhecido (*unknown*), desempregado (*unemployed*), gestor (*management*), empregado/a doméstica (*housemaid*), empreendedor (*entrepreneur*), estudante (*student*), operário (*blue-collar*), trabalhadores independentes (*self-employed*), aposentado (*retired*), técnico (*technician*) e serviços (*services*). Depois de alguns estudos, pudemos identificar alguns atributos que se destacaram no que diz respeito a taxa de sucesso/insucesso.

Marital: Consiste no estado civil atual correspondente ao cliente contactado durante a campanha ocorrente. Entre os atributos temos, casados (*married*), solteiros (*singles*) e divorciados (*divorced*) que também corresponde aos viúvos. Tal como algumas variáveis antecedentes, esta também não tem grande impacto no nosso target, contudo, após um estudo da mesma verificou-se que um dos atributos tem uma maior adesão comparado com os outros.

Education: Esta variável corresponde ao nível de escolaridade atual do cliente contactado, entre eles, a primária (*primary*), o secundário (*secondary*), o terciário (*tertiary*) e os desconhecidos (*unknown*). Após o estudo da mesma, verificou-se que também não teria grande impacto com o target definido, contudo, existe um atributo com uma maior taxa de subscrição em comparação com os demais.

Default: Esta variável indica se o cliente contactado durante a campanha que decorre apresenta alguma dívida ou não, sendo assim, os únicos atributos são: não (*no*) e sim (*yes*). Após um estudo breve da mesma, considerou-se irrelevante para o tema em estudo.

Housing: Corresponde se o determinado cliente apresenta algum empréstimo habitacional, tal como os atributos da variável anterior referida, temos: sim (*yes*) e não (*no*). Depois de alguns estudos, verificou-se que esta variável

Loan: Semelhante a anterior, esta variável, representa se o cliente possui um empréstimo pessoal, como tal, os seus atributos são: não (*no*) e sim (*yes*). Um estudo da mesma permitiu-nos afirmar que existe uma maior adesão ao depósito a prazo por parte de um dos atributos.

Contact: Corresponde ao tipo de comunicação do contacto, os atributos correspondentes são: desconhecido (*unknown*), telefone (*telephone*) e celular (*cellular*). Verificou-se que esta variável é irrelevante para o objetivo do projeto.

Month: Coincide com o mês em que o contacto foi realizado para o respetivo cliente. Os atributos desta variável são: janeiro (*jan*), fevereiro (*feb*),

março (*mar*), abril (*apr*), maio (*may*), junho (*jun*), julho (*jul*), agosto (*aug*), setembro (*sep*), outubro (*oct*), novembro (*nov*) e dezembro (*dez*).

Poutcome: Por último esta variável corresponde ao resultado da campanha anterior realizada do respetivo contacto, entre eles, desconhecido (*unknown*), sucesso (*sucess*), falhado (*failure*) e outro (*other*). O estudo da mesma permitiu-nos aferir que um dos atributos desta variável se correlaciona com um dos atributos das variáveis numéricas, “*previous*” e “*pdays*”.

Y: É a nossa variável independente (o target definido) que tem como atributos, sim (*yes*) – para quando o cliente adere à subscrição ao depósito a longo prazo - e não (*no*) – para quando o cliente não adere à subscrição do depósito a longo prazo. O conjunto de dados é constituído por 45211 registos e as mencionadas 16 variáveis dependentes, e a variável independente. Neste número significativo de registos existem 5289 registos que aderiram ao depósito a prazo e os restantes 39922 não aderiram.

2.3. Data Preparation

Após a compreensão e exploração profunda das variáveis correspondentes à base de dados inicial, prosseguiu-se para a fase de processamento dos dados, Data Preparation, nela a ideia é realizar a seleção e qualquer tipo de processamento sobre os dados brutos, a fim de obter os dados tratados para a sua testagem em modelos preditivos.




O primeiro passo tomado foi atribuir à variável independente o seu respetivo “valor”, o target:

17	y	C	category	target	no, yes
----	---	---	----------	--------	---------

A seguir, estudou-se a variável “*Previous*” em relação à taxa de *outliers* no widget “*distributions*”, a partir do qual podemos aferir que a taxa de *outliers* que consideramos foi a partir dos 20 contactos realizados antes desta campanha, até ao máximo (275 dias). Isto porque a quantidade de registos nesse intervalo é insignificante para com o total de registos (63 para o total de 45211).

O mesmo critério foi usado em relação à variável “*Balance*”, contudo foi estudado no widget “*Scatter Plot*”, e assim, foi considerado a taxa de *inliers* no intervalo de –2000 a 60 000 euros, pois, tal como o critério usado na variável anterior, os registos não compreendidos nesse intervalo são insignificativos para o nosso estudo.

Posteriormente, foi estudada a variável “*Campaign*”, consideraram-se casos que pudessem ser precedentes, visto que esta variável faz referência ao número de contactos realizados na presente campanha a um único cliente, decidimos que seria útil para o nosso modelo que houvesse alguma paridade entre as grandes campanhas e as mínimas, ou seja, teve-se a ideia de que, para o caso de haver um total 5 clientes com 20 contactos realizados, teria que se obter para os clientes com apenas 19 contactados um número de clientes iguais ou superiores aos que tiveram 20 (5 clientes). Esta tendência confirmou-se apenas a partir dos clientes com o máximo de 24 contactos, o que nos motivou a eliminar os registos com o número de contactos maiores que este.

Conditions					
	previous	is below		20	×
	campaign	is below		25	×
	balance	is between	-2000	and 60000	×

Tratamento de *outliers*

Após o estudo referente aos *outliers* e a sua visualização na *data table*, foi atribuído o mesmo nome a atributos diferentes de determinadas variáveis, consoante o critério que iremos abordar.

Foi realizado um estudo analítico em relação à taxa de clientes que subscreveram e não subscreveram ao depósito a prazo para as seguintes features: *Job*, *Education* e *Month*, com a visão de atribuir o mesmo domínio a atributos que tivessem semelhanças, em relação, às percentagens de clientes que subscreveram e os que não subscreveram de acordo aos referidos atributos. Como por exemplo, na variável *Job*, os atributos, *Self-employed*, *Admin* e *Unknow*, apresentam uma relativa proporção semelhante, em relação à

percentagem, relativamente de “yes” e “no” (atributos do target), deste modo, atribuímos a estes atributos o domínio “*admin/self-employed*”, e assim, o atributo “*Unknow*” ficou inserido nos respetivos atributos, “*Admin*” e “*Self-employed*”. O mesmo foi aplicado ainda a determinados atributos desta variável, como também, às variáveis: *Education* e *Month*. Assim atribuímos os seguintes domínios nas consideradas features:

The image shows three screenshots of variable configuration widgets, likely from a data science tool like Orange3. Each widget is for a categorical variable and shows a list of values with their corresponding target labels.

- month widget:** Name: month, Type: Categorical. Values: apr → apr, aug → jan/jun/aug/nov (merged), dec → dec, feb → feb, jan → jan/jun/aug/nov (merged), jul → jul, jun → jan/jun/aug/nov (merged), mar → mar, may → may, nov → jan/jun/aug/nov (merged), oct → oct, sep → sep.
- education widget:** Name: education, Type: Categorical. Values: primary → primary, secondary → secondary, tertiary → tertiary (merged), unknown → tertiary (merged).
- job widget:** Name: job, Type: Categorical. Values: admin. → admin/self-employed (merged), blue-collar → blue-collar, entrepreneur → entrepreneur, housemaid → housemaid/services (merged), management → management, retired → retired, self-employed → admin/self-employed (merged), services → housemaid/services (merged), student → student, technician → technician, unemployed → unemployed, unknown → admin/self-employed (merged).

Eliminação de atributos desconhecidos

Em relação à variável *Poutcome*, verificou-se no widget “check_unknow” que o atributo *Unknow* correspondia aos clientes que tinham sido contactados pela primeira vez, ou seja, o atributo “0” na variável *Previous* e o atributo “-1” na feature “*Pdays*”, concluindo, estes 3 atributos destas features dizem o mesmo (estão correlacionadas), pelo que, se atribuiu o domínio de “*no_contact*” ao atributo *Unknow*.

The image shows a screenshot of the 'poutcome' variable configuration widget. It is a categorical variable with the following values and target labels: failure → failure (merged), other → failure (merged), success → success, and unknown → no_contact.

Eliminação de valores desconhecidos

Deste modo, os atributos desconhecidos estão quase todos tratados, restando apenas o atributo desconhecido na feature *contact*, pelo que foi feito o estudo do mesmo. Verificou-se que esta variável dependente é irrelevante pois é uma questão imprevisível pois raramente é possível saber se o tipo de comunicação vai ser realizado por telefone ou celular, pelo que foi excluída esta variável.

Por um critério semelhante excluiu-se a variável *Default*, e ele foi, existe uma maioria absoluta de clientes sem qualquer dívida, uma maioria de 98% (verificado no widget “Distributions Original”) e são estes clientes que aderiram massivamente ao depósito a longo prazo, e assim, eliminou-se esta variável, no widget “Features Redution”. Neste mesmo widget, exclui-se as variáveis *Pdays* e *Previous*, por consequente ao seguinte estudo. Tanto no widget “Correlations No” e “Correlation Yes” a correlação destas duas features era praticamente de 1, pelo que ficou considerado a exclusão de uma delas, e então, eliminou-se, primeiramente, a variável *Pdays* porque foi calculado o somatório de correlação destas duas variáveis com as demais (nos mesmos widgets), e verificou-se que *Previous* tinha um somatório de correlação superior em relação ao somatório da variável *Pdays*. Após a eliminação da referida variável, verificou-se no widget “Correlations Yes” que *Previous* e *Duration* apresentavam uma correlação negativa, relevante de eliminar uma delas, pelo que foi feito o mesmo procedimento antes referido, e excluiu-se *Previous*, no widget “Features Redution”.

Foi eliminada a variável *day* pois notou-se que ela acaba sendo uma variável casual, e como uma das questões consideradas pelo grupo referia-se à altura que seria melhor fazer este tipo de campanhas direcionadas, então por este motivo decidiu-se focar em meses.

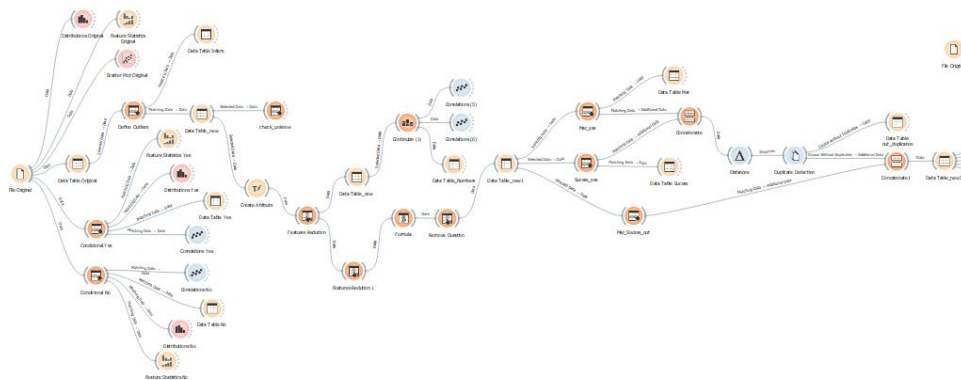
Posteriormente verificou-se a correlação das variáveis categóricas com as demais, pelo que foi transformado em ordinal todas as variáveis categóricas, e assim verificou-se a respetiva correlação. Foi então, que foi verificado a correlação negativa significativa entre a feature *Age* e *Marital*, e pelo procedimento anterior tomado para tratar de variáveis que apresentavam uma

correlação positiva significativa, aplicou-se o mesmo, e foi então excluída a variável *Age*, no widget “*Features Reduction 1*”.

Após isso, por motivo de escala, criou-se uma fórmula que converte os segundos (da variável *Duration*) em minutos, e assim, criou-se uma nova variável “*duration_in_min*” e excluiu-se a variável *Duration*.

Observando o target verificou-se que havia uma distribuição *imbalance*, o que significa que uma classe teve superioridade em relação à outra, a um nível que foi considerado grave. Estudos mostram que este tipo de situação é geralmente causado por um enviesamento dos dados, onde por maldade, ou por busca de benefícios particular é falsificado alguns registos atribuindo uma classe do target a registos que tinham antes uma outra classe.

Visto que não nos vimos qualificados para resolver este tipo de problema - por estarmos a utilizar uma plataforma com algumas limitações - e identificar quais foram os atributos enviesados decidiu-se indicar ao modelo uma nova tendência para a qual o modelo se devia focar para a fase de treino e teste. Para tal, destacou-se atributos das variáveis *month* e *outcome*, onde se verificou em alguns casos que a classe minoritária obteve alguma vantagem sobre a majoritária.



Visão geral da fase de Data Preparation

2.4. Modeling

Tendo sido feitas as análises e os tratamentos dos dados, passou-se para a fase da seleção dos modelos a serem utilizados para a resolução do problema em questão, com o foco em selecionar os que fossem de acordo a tendência dos dados. Tendo em consideração o target definido, foi possível verificar que se tratava de um problema de classificação e que estávamos diante de um modelo supervisionado, o que facilitou a delimitação dos tipos de modelos por pesquisar.

Pesquisando sobre os diferentes tipos de modelos, verificou-se, que o comportamento dos dados (visto pela análise dos gráficos), podiam ser resolvidos com uma grande variedade de modelos. Dos vários tipos de modelos selecionou-se apenas quatro que pelos estudos demonstrou-se que além de serem úteis para o tipo de problema em questão, também foi visto que os mesmos não exigiam muito da memória do computador, pois os dados utilizados não eram de enorme escala. Estes modelos foram, o SVM (Support Vetor Machine), Decision Tree, Naive Bayes e Logistic Regression.

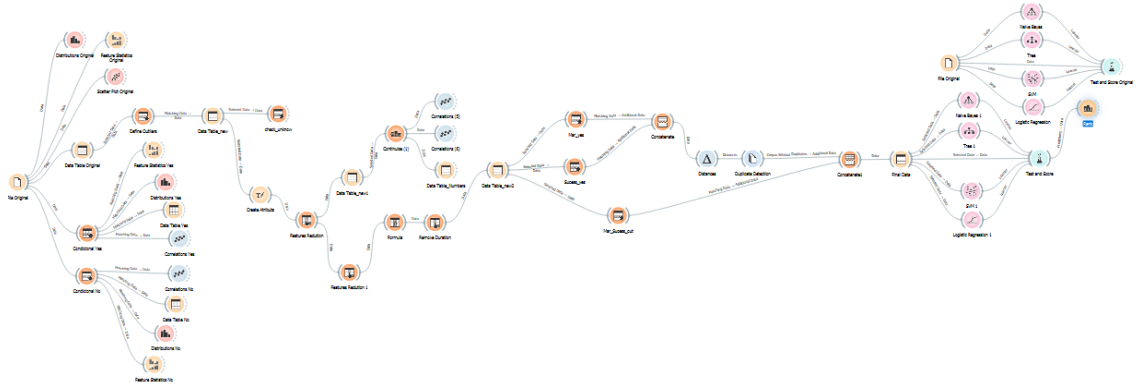
SVM: é um conceito na ciência da computação para um conjunto de métodos de aprendizado supervisionado que analisam os dados e reconhecem padrões, usado para classificação e análise de regressão.

Decision Tree: é uma das abordagens de modelagem preditiva usadas em estatística, mineração de dados e aprendizado de máquina. Os modelos de árvore podem ser de regressão e de classificação. Nos casos em que a variável de destino pode assumir um conjunto de discreto de valores está-se diante de uma árvore de classificação.

Naive Bayes: é uma família de classificadores probabilísticos que se baseiam na aplicação da inferência bayesiana com fortes suposições de independência entre as variáveis. Classificadores de Naive Bayes são altamente escaláveis, exigindo um grande número de variáveis para o aprendizado.

Logistic Regression: é um modelo frequentemente usado para classificação e análise preditiva. Este modelo estima a probabilidade de

ocorrência de um evento, como um voto, com base em um determinado conjunto de dados de variáveis independentes.



Visão Geral do workflow

2.5. Evaluation

Com base no maior objetivo da campanha de marketing, que é prover a maior aceitação da parte dos clientes para que os mesmos possam aderir ao depósito a prazo, usamos os devidos classificadores considerando as seguintes métricas, Recall, Precision e F1-Score.

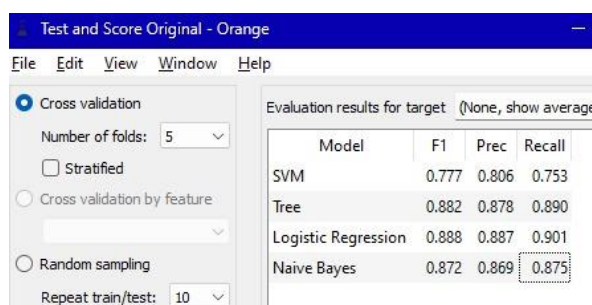
Para avaliar o desempenho do modelo foi utilizado o widget *Test and Score* que permite avaliar o desempenho de modelos de aprendizado de máquina, é usado para medir a forma como um modelo se ajusta aos dados de treino e como ele se generaliza para novos dados.

O **Recall** é uma métrica usada em problemas de classificação, avalia a capacidade de um modelo encontrar possíveis situações positivas e neste modelo, o Recall tinha como função identificar o maior número possível de clientes que irão aderir ao depósito a prazo, pois pretende-se minimizar o número de falsos negativos.

O **Precision** geralmente usado em estatísticas quando os falsos positivos têm um impacto muito negativo relativamente aos falsos negativos, irá ajudar o modelo a evitar falsos clientes nesse caso clientes que realmente não tem interesse no depósito a prazo, quanto maior for a precisão o modelo estará a minimizar os falsos positivos.

O **F1-Scores** usado para uma avaliação geral do desempenho do modelo em relação aos clientes que aceitaram os depósitos a prazo, Isso permitira o alcance de um equilíbrio entre a identificação correta de aceitantes (recall) e evitar falsos positivos (precisão), mantendo um equilíbrio entre os dois quando o F1-Scores tem um resultado baixo é porque ou recall ou precision é baixo.

Primeiramente foi feito um teste com a base de dados original, com os modelos de classificação mencionados acima, svm, tree, logistic regression e naive byes obtendo os seguintes resultados relevantes do modelo sem alterações.

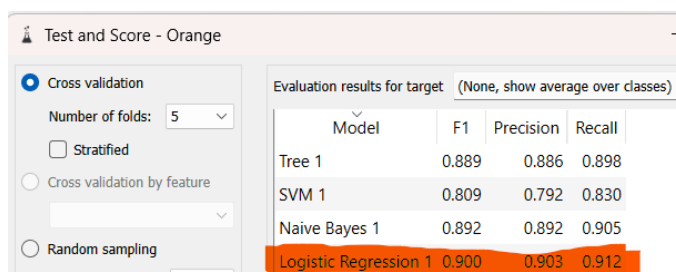


The screenshot shows the 'Test and Score Original' window in Orange Data Mining. The 'Cross validation' method is selected with 5 folds. The evaluation results table is as follows:

Model	F1	Prec	Recall
SVM	0.777	0.806	0.753
Tree	0.882	0.878	0.890
Logistic Regression	0.888	0.887	0.901
Naive Bayes	0.872	0.869	0.875

Métricas dos dados originais

Após as etapas do CRISP-DM implementadas ao modelo na sequência de passos mencionados acima, tendo em conta os modelos de classificação usados, destacando-se nas métricas Recall, Precision, F1-Score. Obteve-se os melhores resultados possíveis com base no objetivo a alcançar “prever se o cliente irá subscrever um depósito a prazo”.



The screenshot shows the 'Test and Score' window in Orange Data Mining. The 'Cross validation' method is selected with 5 folds. The evaluation results table is as follows:

Model	F1	Precision	Recall
Tree 1	0.889	0.886	0.898
SVM 1	0.809	0.792	0.830
Naive Bayes 1	0.892	0.892	0.905
Logistic Regression 1	0.900	0.903	0.912

Métricas dos dados tratados

Chegado aqui, seria oportuno responder as questões colocadas ao início do trabalho, após alcançar o objetivo deste trabalho e os resultados obtidos.

Qual seria o perfil “ideal” para que um cliente subscrevesse a um depósito a prazo?

Com base nas análises feitas neste trabalho para classificação de um cliente ideal, não há relevância com a idade do mesmo, deve apresentar os seguintes padrões, ter uma profissão de estudante, retirado ou desempregado, com estado civil solteiro pois estes não tem muitas responsabilidades relativamente aos outros, deve ter uma educação do nível terciário pois estes tem uma noção mais ampla referente a tomada de decisões que implicam o seu amanhã em investimentos do género, com um saldo não

inferior a –2.000 mil euros e não superior a 60.000 mil euros , e sem empréstimos de habitação e pessoal .

Qual das variáveis disponíveis no nosso conjunto de dados tem maior influência sobre o target definido?

Com base no widget *rank*, que permite classificar o grau de importância das variáveis dependentes no nosso conjunto de dados, foi assim identificado a variável com mais relevância em relação ao target esta que foi a *feature Job*.

		#	ReliefF
1	C month	9	0.020
2	C job	9	0.114
3	C poutcome	4	0.066
4	C marital	3	0.022
5	C education	3	0.030
6	C loan	2	0.034
7	C housing	2	0.014
8	N duration_in_min		0.039
9	N balance		0.005
10	N campaign		0.024

Demonstração do rank

Qual será a melhor altura do ano para fazer uma campanha no marketing bancário?

Segundo as análises obtidas, a melhor altura do ano para fazer a campanha do marketing bancário, é o mês de março pois foi o período na qual houve maior aceitação ao depósito a prazo, a taxa de aceitação referente ao target (S/N) foi positiva referente aos outros meses que a campanha decorreu.

3. BIBLIOGRAFIA

Scikit-Learn. (s.d.). Obtido em Outubro de 2023, de 1.4. Support Vector Machines: <https://scikit-learn.org/stable/modules/svm.html>

Scikit-Learn. (s.d.). *Naive Bayes*. Obtido em Outubro de 2023, de Scikit-Learn: https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html

Wikipédia. (2014). *Máquina de Vetores de Suporte*. Obtido de https://pt.wikipedia.org/wiki/M%C3%A1quina_de_vetores_de_suporte

Wikipédia. (s.d.). *Aprendizagem de árvore de decisão*. Obtido em Outubro de 2023, de Wikipédia.

Wikipédia. (s.d.). *Naive Bayes*. Obtido em Outubro de 2023, de Wikipédia: https://pt.wikipedia.org/wiki/Naive_Bayes