Introdução a Ciência de Dados

-

Trabalho de Classificação no Weka

Dataset: GERMAN CREDIT DATA (DADOS DE CRÉDITO ALEMÃO)

Profª Tatiana Escovedo

Grupo B:

Daniel Almeida
Gabriel Botelho
Luiz Henrique
Rosilane Lessa
Vinicius Georges

Introdução a Ciência de Dados

-

Sumário:

Introdução a Ciência de Dados

Introdução:

O projeto de Data Science será ilustrado com base no dataset German Credit Data, seguindo as 7 etapas apresentada abaixo:

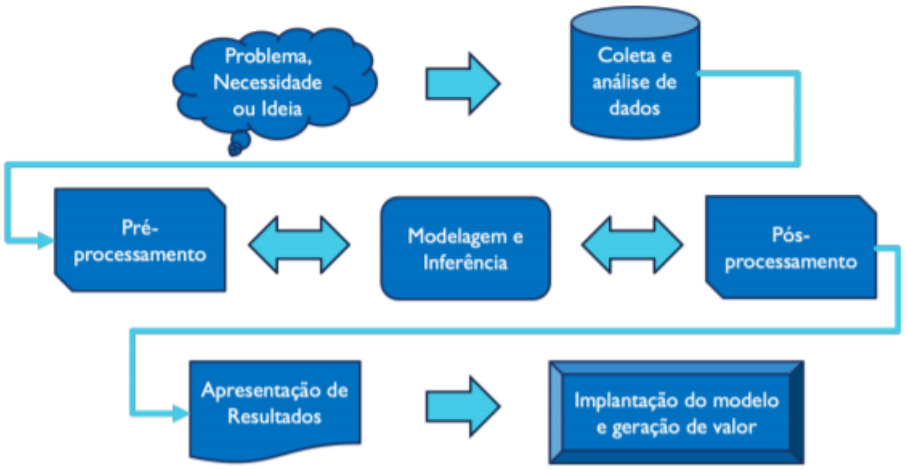


Figura 01: Etapas de Data Science

Seguindo as etapas, o projeto dataset German Credit Data inicia quando identificamos os problemas, as necessidades ou ideias. E com base nessas informações seguimos para etapa de coleta e análise dos dados aplicando o conhecimento adquirido para realizar o pré-processamento dos dados.

Na etapa de modelagem e inferência será relacionado os possíveis modelos para cada problema levantado, compondo as instâncias e as variáveis existentes no dataset. Já na etapa pós-processamento será apresentada as heurísticas de negócios. E na etapa apresentação de resultados será relatada as metodologias adotadas comparando os resultados do melhor modelo planejando os passos para a implantação da solução. Por fim, na etapa implantação do modelo e geração de valor será abordado a geração de valor ao empreendimento de forma qualitativa e quantitativa.

Introdução a Ciência de Dados

-

1. Apresentação do problema

Figura 02: Etapa 1

Analisando o dataset identificamos que trata-se de um problema preditivo de classificação, onde as pessoas solicitantes são avaliadas pelo seu risco de crédito, como bom (1) ou ruim (2).Essa análise ajudaria o banco credor na tomada de decisão na liberação de crédito, de acordo com o perfil das pessoas observadas, minimizando assim os riscos de inadimplências.

Foi observado no dataset que não há Missing (“valor ausente”) e aparentemente não há valores inconsistentes. Sendo assim, as etapas de limpeza e pré-processamento não serão implementadas neste momento, mas serão aprofundados as análises estatísticas com objetivo de identificar irregularidade no dataset.

Há uma variedade grande do tipo de variáveis encontradas no dataset - conforme aludido abaixo: há atributos que são variáveis qualitativas nominais: como propósito de obtenção do crédito, status marital, sexo, fiador, planos de prestações, status da habitação, telefone próprio e se a pessoa é um trabalhador estrangeiro; atributos que são variáveis qualitativas ordinais: status da conta, histórico de crédito, poupança/títulos, duração no emprego atual, propriedades, emprego.

Ademais, há os atributos numéricos; como as variáveis quantitativas discretas: duração (em meses), montante do crédito, tempo na moradia atual, idade em anos, número de créditos existentes no banco e número de dependentes. E por fim, há as variáveis quantitativas contínuas: taxa de prestação (em porcentagem da renda).

-

Nas tabelas abaixo, pode-se ver uma descrição mais aprofundada dos atributos contidos no presente *dataset*:

Descrição dos atributos:

Atributo 1	*checking_status*	**Tipo:**	Nominal
Descrição	Saldo da conta corrente, caso possua, por faixa de valor na unidade monetária alemã da época: Marco Alemão (M)		

Categorias	**Descrição das Categorias**
A11 : <0	saldo menor que 0 M
A12 : 0<=X<200	saldo maior ou igual a 0 M e menor que 200 M
A13 : X>= 200	saldo maior ou igual a 200 M
A14 : no checking	nenhuma conta corrente

Tabela 01: Descrição do atributo 1

Atributo 2	*duration*	**Tipo:**	Numérico
Descrição	Número de meses em que se financiou o empréstimo		

Tabela 02: Descrição do atributo 2

Atributo 3	*credit_history*	**Tipo:**	Nominal
Descrição	Classifica o comportamento de tomada de créditos		

Categorias	**Descrição das categorias**
A30 : no credits taken/ all credits paid back duly	nunca tomou crédito/quitou todos
A31 : all credits at this bank paid back duly	todos os créditos quitados com o banco
A32 : existing credits paid back duly till now	Créditos existentes sendo pagos corretamente
A33 : delay in paying off in the past	histórico de atraso no pagamento
A34 : critical account/ other credits existing (not at this bank)	conta crítica/existência de financiamento em outros bancos

-

Tabela 03: Descrição do atributo 3

Atributo 4	*Purpose*	**Tipo:**	Nominal
Descrição	Classifica de acordo com o propósito do financiamento		
Categorias	**Descrição das Categorias**		
A40: car (new)	carro novo		
A41: car (used)	carro usado		
A42: furniture/equipment	móveis/equipamentos		
A43: radio/television	rádio/televisão		
A44: domestic appliances	aparelhos de casa		
A45: repairs	consertos		
A46: education	educação		
A47: vacation	férias		
A48: retraining	requalificação		
A49: business	negócios		
A410: others	outros		

Tabela 04: Descrição do atributo 4

Atributo 5	*credit_amount*	**Tipo:**	Numeric
Descrição	Valor do crédito concedido na unidade monetária alemã da época: Marco Alemão (M)		

Tabela 05: Descrição do atributo 5

Atributo 6	*savings_status*	**Tipo:**	Nominal
Descrição	Saldo total das poupanças e títulos, caso possua, por faixa de valor na unidade monetária alemã da época: Marco Alemão (M)		
Categorias	**Descrição das Categorias**		

A61: < 100	menor que 100 M
A62: 100 <= X < 500	maior ou igual a 100M e menor que 500 M
A63: 500 <= X < 1000	maior ou igual a 500M e menor que 1000 M
A64: >= 1000	maior que 1000 M
A65: no known savings	nenhuma poupança ou título

Tabela 06: Descrição do atributo 6

Atributo 7	*employment*	**Tipo:**	Nominal
Descrição	Classifica a pessoa, conforme o tempo de emprego, caso esteja empregada		
Categorias	**Descrição das Categorias**		
A71: unemployed	desempregado		
A72: <1	Há menos de 1 ano		
A73: 1 <= X < 4	de 1 à 4 anos exclusive		
A74: 4 <= X < 7	de 4 à 7 anos exclusive		
A75: >= 7	7 ou mais anos empregado		

Tabela 07: Descrição do atributo 7

Atributo 8	*installment_commitment*	**Tipo:**	Numeric
Descrição	comprometimento da parcela do empréstimo em relação à renda		

Tabela 08: Descrição do atributo 8

Atributo 9	*personal_status*	**Tipo:**	Nominal
Descrição	Classifica de acordo com o gênero e o estado civil		
Categorias	**Descrição das Categorias**		
A91: male divorced / separated	homem divorciado / separado		
A92: female divorced / separated / married	mulher divorciada / separada / casada		

A93: male single	homem solteiro
A94: male married / widowed	homem casado / viúvo
A95: female single	mulher solteira

Tabela 09: Descrição do atributo 9

Atributo 10	*other_parties*	**Tipo:**	Nominal
Descrição	Classifica quanto à existência de garantidores/co-participantes		

Categorias	**Descrição das Categorias**
A101: none	sem garantia
A102: co-applicant	co-participante
A103: guarantor	garantidor

Tabela 10: Descrição do atributo 10

Atributo 11	*residence_since*	**Tipo:**	Numeric
Descrição	Tempo de permanência na residência atual em anos		

Tabela 11: Descrição do atributo 11

Atributo 12	*property_magnitude*	**Tipo:**	Nominal
Descrição	Assinala a existência de ativos e quais são esses		

Categorias	**Descrição das Categorias**
A121: real estate	residência
A122: if not A121 : building society savings agrément / life insurance	se não tem residência, tem seguro de vida ou outros serviços financeiros
A123: if not A121 / A122 : car or other, not in attribute 6	se não tem residência nem serviços financeiros, possui automóvel ou outros que não sejam poupanças ou títulos
A124: unknown / no property	não tem ativos ou sem informação.

Tabela 12: Descrição do atributo 12

Atributo 13	*age*	**Tipo:**	Numeric
Descrição	Idade dos tomadores de financiamento		

Tabela 13: Descrição do atributo 13

Atributo 14	*other_payment_plans*	**Tipo:**	Nominal
Descrição	Classifica de acordo com a existência de outros financiamentos		
Categorias	**Descrição das Categorias**		
A141: bank	bancos		
A142: stores	lojas		
A143: none	sem outros financiamentos		

Tabela 14: Descrição do atributo 14

Atributo 15	*housing*	**Tipo:**	Nominal
Descrição	Classifica de acordo com status da residência		
Categorias	**Descrição das Categorias**		
A151: rent	Aluguel		
A152: own	Residência própria		
A153: for free	doação		

Tabela 15: Descrição do atributo 15

Atributo 16	*existing_credits*	**Tipo:**	Numeric
Descrição	número de financiamentos existentes no banco		

Tabela 16: Descrição do atributo 16

Atributo 17	*job*	**Tipo:**	Nominal
Descrição	Classifica de acordo com a situação e tipo de trabalho dos tomadores de financiamento com o status de ser ou não residente		

-

Categorias	Descrição das Categorias
A171: unemployed/ unskilled - non-resident	desempregado/pouco qualificado-não residente
A172: unskilled - resident	pouco qualificado - residente
A173: skilled employee / official	bem qualificado
A174: Management / self-employed / highly qualified employee / officer	Gerente / autônomo / função bem qualificada / chefe

Tabela 17: Descrição do atributo 17

Atributo 18	*num_dependents*	**Tipo:**	Numérico
Descrição	Número de dependentes		

Tabela 18: Descrição do atributo 18

Atributo 19	*own_telephone*	**Tipo:**	Nominal
Descrição	Classifica de acordo com a posse de linha telefônica		
Categorias	**Descrição das Categorias**		
A191: none	sem linha		
A192: yes, registered under the customers name	com linha registrada em seu nome		

Tabela 19: Descrição do atributo 19

Atributo 20	*foreign_worker*	**Tipo:**	Nominal
Descrição	Classifica se é trabalhador estrangeiro ou não		
Categorias	**Descrição das Categorias**		
A201 : yes	É trabalhador estrangeiro		
A202 : no	Não é trabalhador estrangeiro		

Tabela 20: Descrição do atributo 20

Atributo 21	*class*	**Tipo:**	Nominal
Descrição	Variable class que classifica o tomador de empréstimo em risco de crédito bom ou ruim		

Categorias	**Descrição das Categorias**
A211 : good	Risco de crédito BOM
A212 : bad	Risco de crédito RUIM

Tabela 21: Descrição do atributo 21

2. Coleta e análise de dados

Figura 03: Etapa 2

O Dataset German Credit classifica pessoas descritas a partir de um conjunto de atributos como risco de crédito bom (classe good) ou como risco de crédito ruim (classe bad). É composto de 1.000 instâncias, 20 atributos preditores e uma class variable, onde 8 atributos são de variáveis qualitativas nominais, 6 atributos são de variáveis qualitativas ordinais, 6 atributos são de variáveis quantitativas discretas e 1 atributo de variável quantitativa contínua, ilustrando:

Relação de variáveis e atributos		
Variáveis qualitativas nominais	***Variáveis qualitativas ordinais***	***Variáveis quantitativas discretas***
❖ propósito de obtenção do crédito; ❖ status marital; ❖ sexo; ❖ fiador; ❖ planos de prestações; ❖ status da habitação; ❖ telefone próprio; ❖ se a pessoa é um trabalhador estrangeiro.	❖ status da conta; ❖ histórico de crédito; ❖ poupança/títulos; ❖ duração no emprego atual; ❖ propriedades; ❖ emprego. ❖ taxa de prestação (em porcentagem da renda).	❖ duração (em meses); ❖ montante do crédito; ❖ tempo na moradia atual; ❖ idade em anos; ❖ número de créditos existentes no banco; ❖ número de dependentes.

Tabela 22: Tipos de variáveis no Dataset

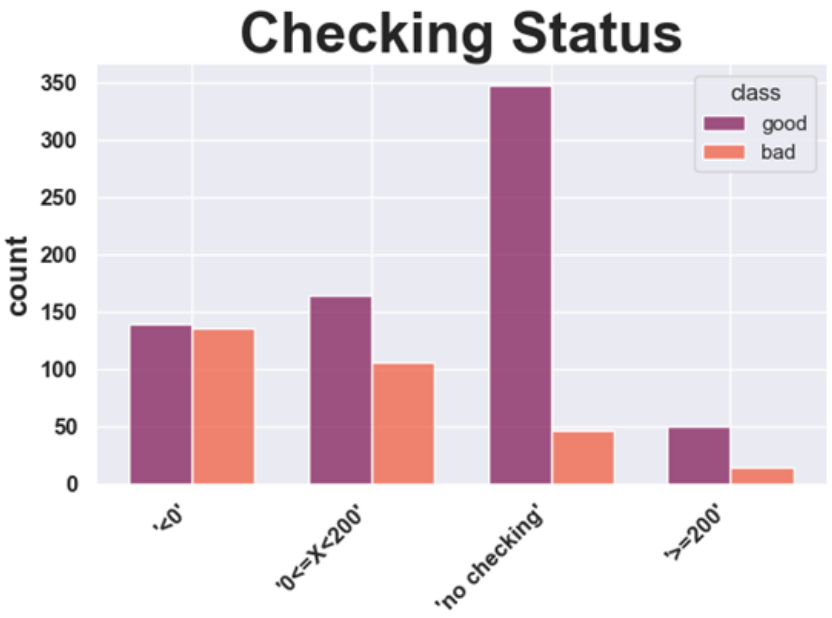
Variável	Mínimo	1º Quartil	Mediana	Média	Moda	3º Quartil	Máximo	Desvio Padrão	Amplitude
Duração	4	12	18	20.9	24	24	72	12.06	68
Montante do Crédito	250	1366	2320	3271	1393	3972	18424	2822.7	18174
Compromisso de Parcelamento	1	2	3	2.9	4	4	4	1.1	3
Tempo na Residência	1	2	3	2.8	4	4	4	1.1	3
Idade	19	27	33	35.55	27	42	75	11.3	56
Créditos Existentes	1	1	1	1.4	1	2	4	0.5	3
Número de Dependentes	1	1	1	1.1	1	1	2	0.3	1

Tabela 23: Estatísticas descritivas - Posição e Dispersão Nota: (i) As demais variáveis são qualitativas, dessa forma, não cabe inseri-las na presente tabela. (ii)Conforme supramencionado, não há valores faltantes no *dataset*

Há de se verificar, também, as distribuições, de forma ilustrada, dos atributos encontrados no *dataset*, relacionados com a variável de interesse, conforme exposto nos gráficos a seguir:

-

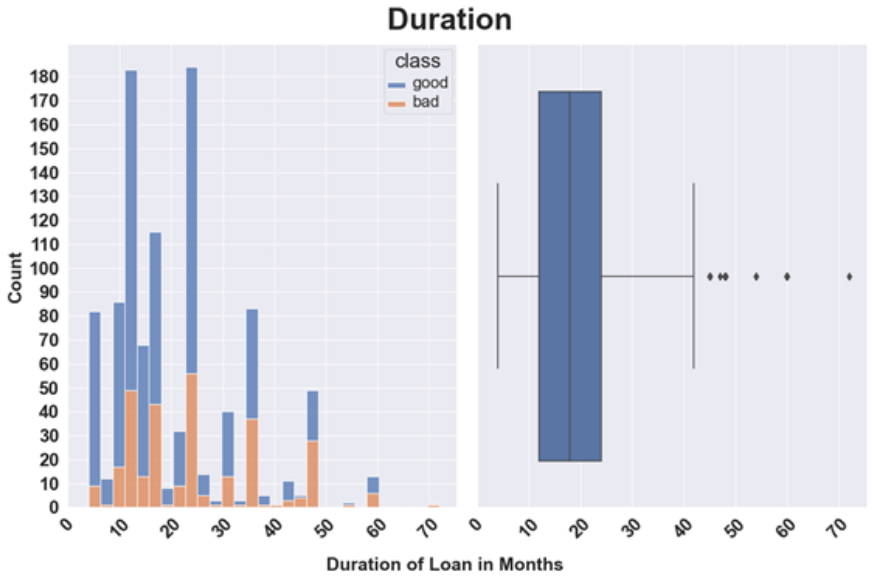
Atributo 1:



Checking status descreve o status da conta corrente do cliente. Pode ser observado que uma parcela bem pequena de clientes possuem mais de 200 mil marcos alemães na conta, de maneira que torna essa categoria um tanto desbalanceada em comparação com as outras 3. Também pode-se observar que para as outras duas categorias, a dispersão entre “good” e “bad” não são muito distintas (praticamente 50% de chance no item “<0”) e para o grupo que não possui conta corrente nesse banco, a taxa de concessão de empréstimo é bem alta. Isso pode ser por conta de uma estratégia de atração de novos clientes por parte do banco. Por conta disso, os atributos ‘<0’ e ‘>=200’ poderiam ser removidos no pré-processamento (assumindo que os atributos sejam implementados no modelo de ML usando “one hot encoding”).

-

Atributo 2:

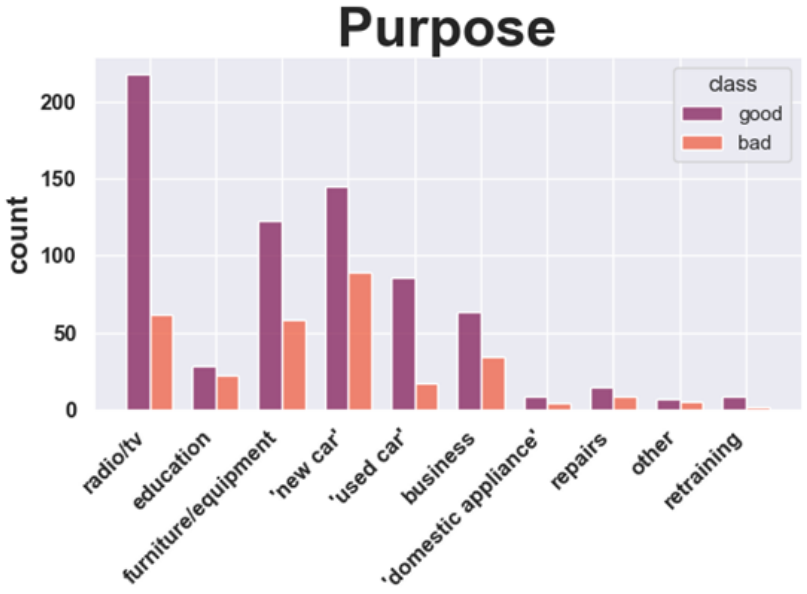


Esse atributo descreve a duração do empréstimo em meses. Considerando o boxplot, pode-se observar que empréstimos acima de 70 meses são outliers evidentes e que empréstimos próximos de 60 podem ser considerados como outliers. Observando a dispersão dos valores no barplot, pode-se afirmar que os dados seguem uma distribuição gama.

Atributo 3:

Credit History representa o histórico de crédito dentro do banco. Pode ser observado que a categoria “all paid” tem uma distribuição praticamente igual em relação a “class”. Essa classificação poderia ser removida do modelo em pré-processamento.

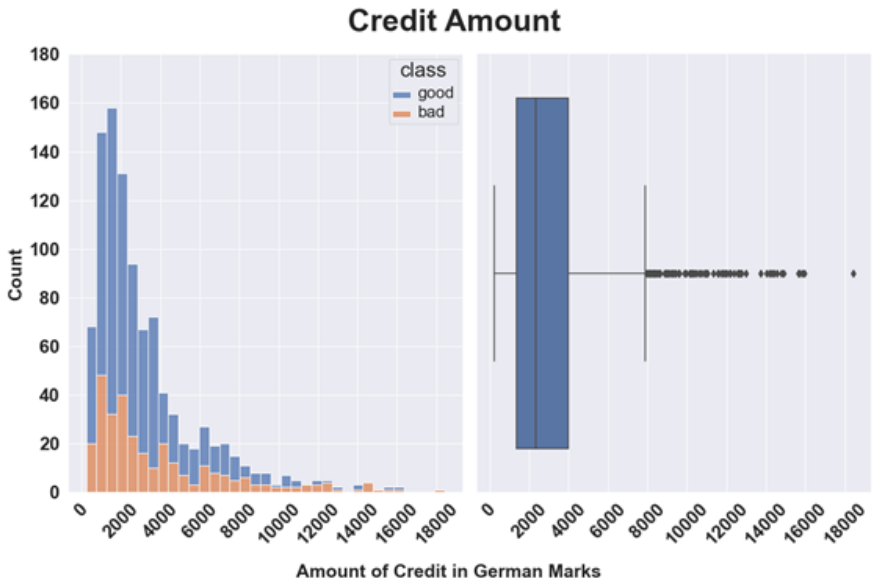
Atributo 4:



-

Purpose se refere ao propósito do empréstimo e nenhuma outra informação pode ser inferida dele usando KDD visual.

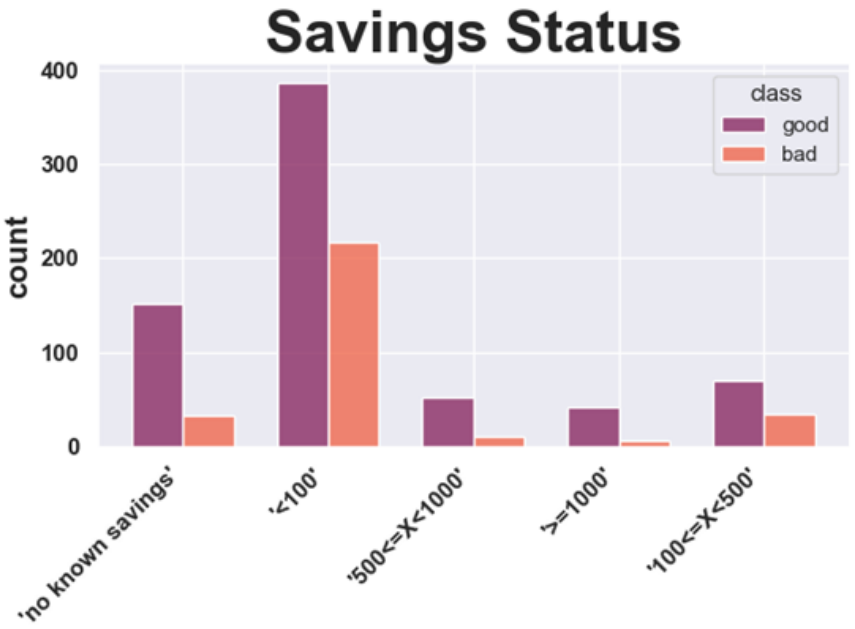
Atributo 5:



Credit amount se refere ao valor total do empréstimo. Credit amount apresenta uma distribuição gama com valores acima de 18 mil marcos podendo ser considerado um outlier.

-

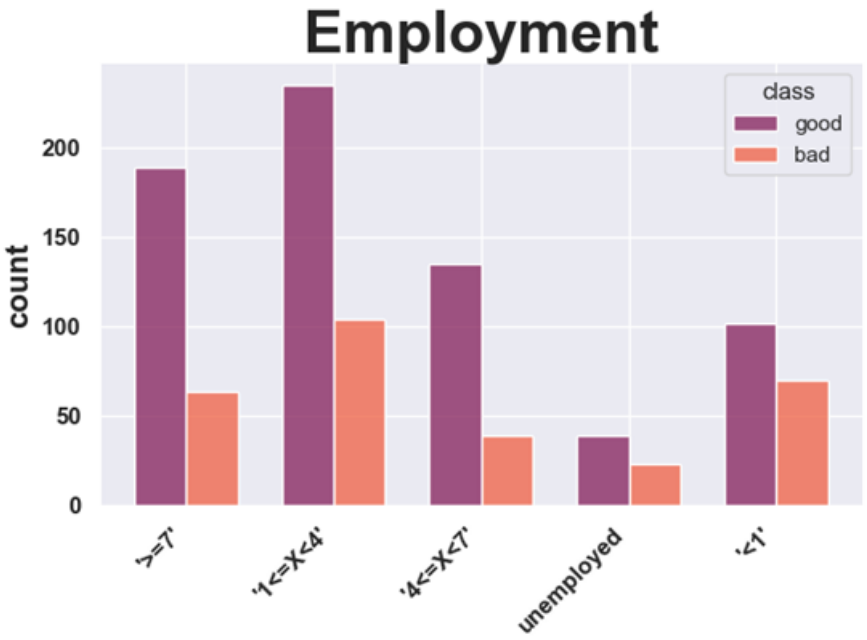
Atributo 6:



Saving status representa o quanto de dinheiro o cliente tem em economias / poupança. Esse gráfico demonstra que todas as categorias apresentam uma forte inclinação para classificação "good". Isso provavelmente não indica nada em que possa ser utilizado em pré-processamento.

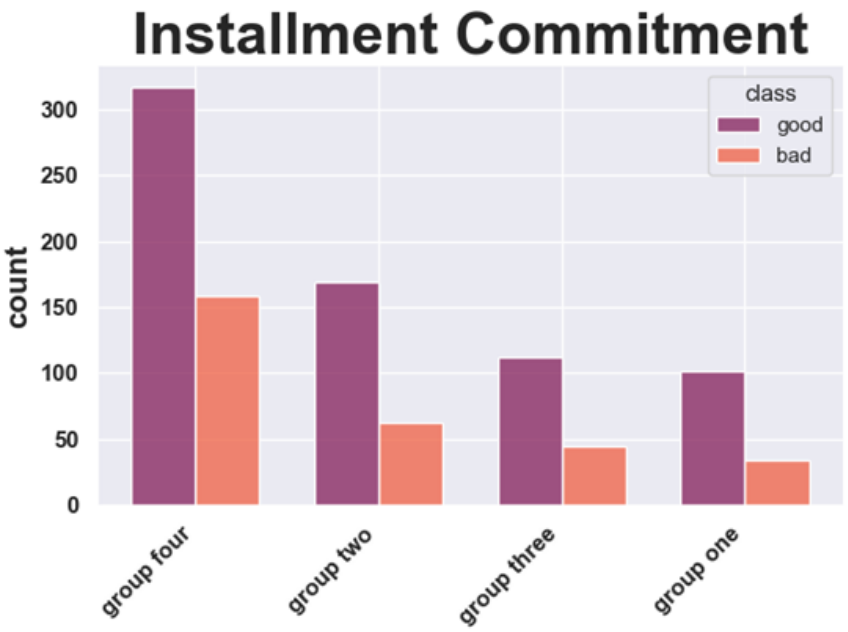
-

Atributo 7:



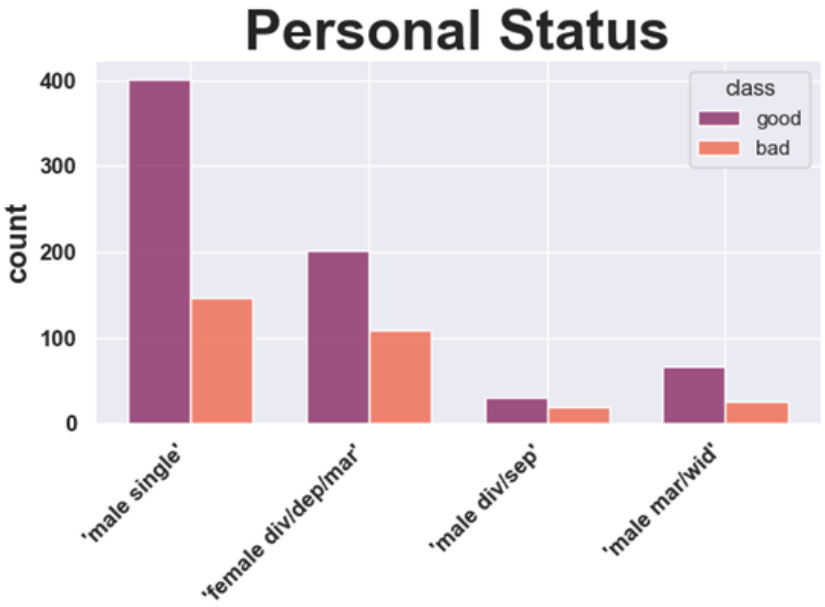
Employment representa a quantidade de anos que o cliente está empregado. Nenhuma informação adicional pode ser observada.

Atributo 8:



Installment commitment representa a parcela do contracheque que vai ser comprometida para arcar com o empréstimo. Assume-se que cada categoria representa um grupo em contrapartida a um valor percentual em si. Pode ser observado que todos os grupos apresentam tendência à classificação “good”.

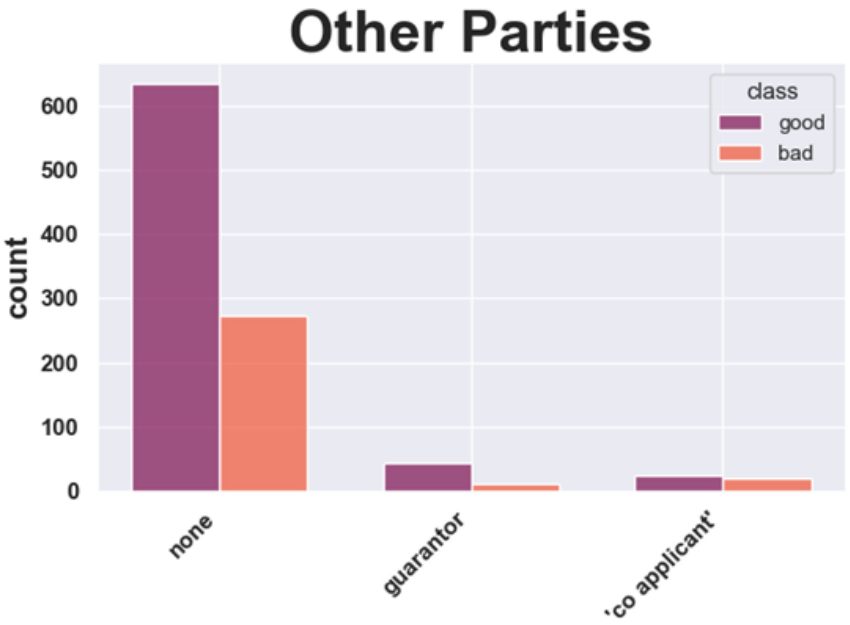
Atributo 9:



Personal status representa o status marital e sexo. Apenas para o status de homem divorciado/separado, há uma paridade entre as categorias bom ou mau pagador. No entanto, há poucas observações em tal status.

-

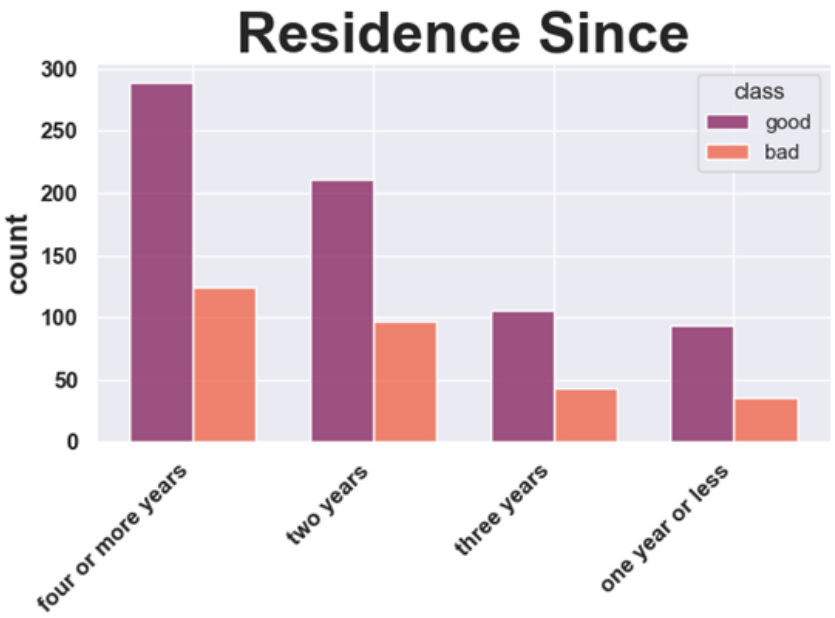
Atributo 10:



Other parties representa outras partes envolvidas em garantir o empréstimo tomado. Observa-se que quase todos os clientes se encontram na categoria “sem nenhum garantidor” por tanto é provável que esse atributo seja desconsiderado para a construção do modelo.

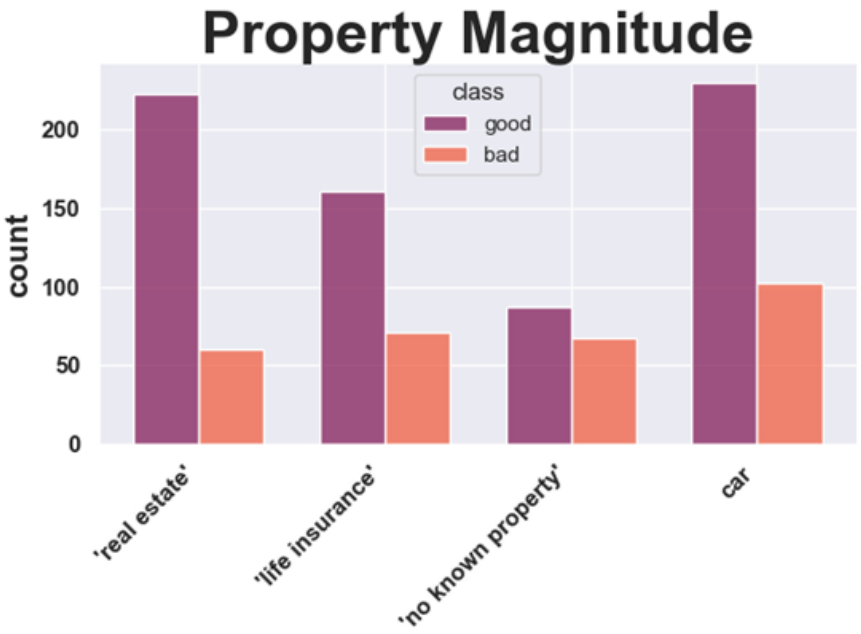
-

Atributo 11:



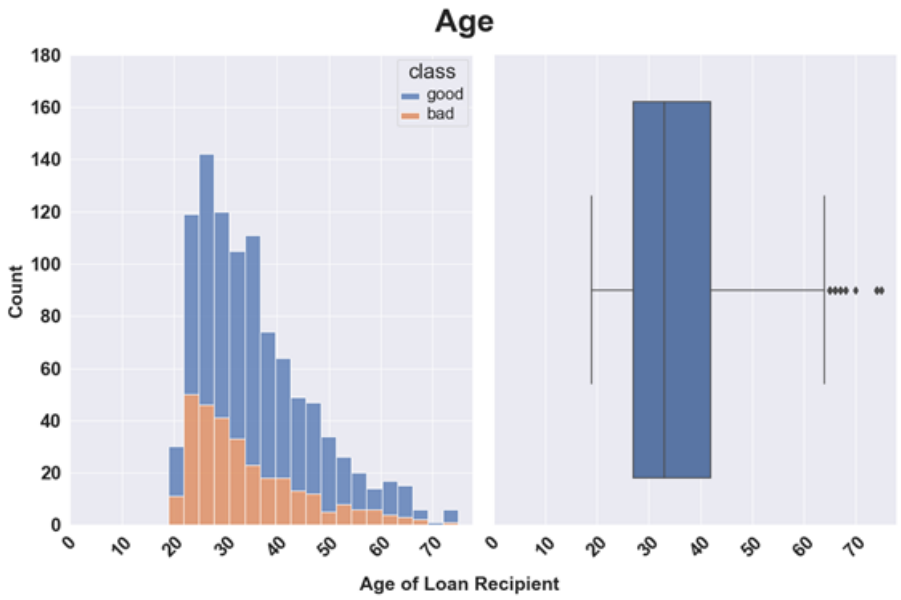
Residence since representa a quantidade de anos que um cliente reside no seu endereço atual. Aparentemente, não há relação entre as duas variáveis (Residence Since e a Categoria).

Atributo 12:



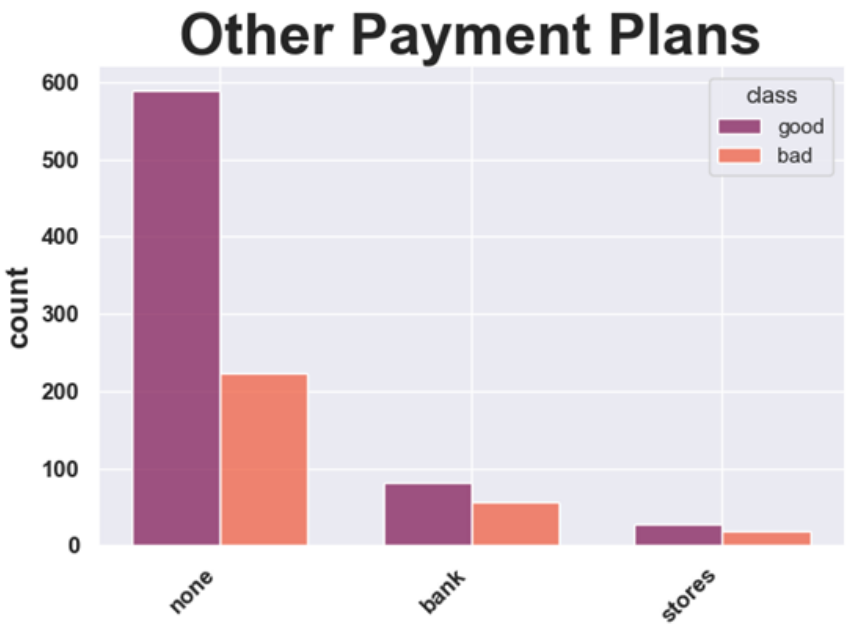
Property magnitude representa os tipos de propriedade que o cliente possui e que podem ser usados como colateral para o empréstimo.

Atributo 13:



Age representa a idade do cliente que procura tomar empréstimo. A distribuição segue padrão exponencial. Valores acima de 70 anos possivelmente podem ser considerados outliers, porém isso não é tão claro dado o boxplot.

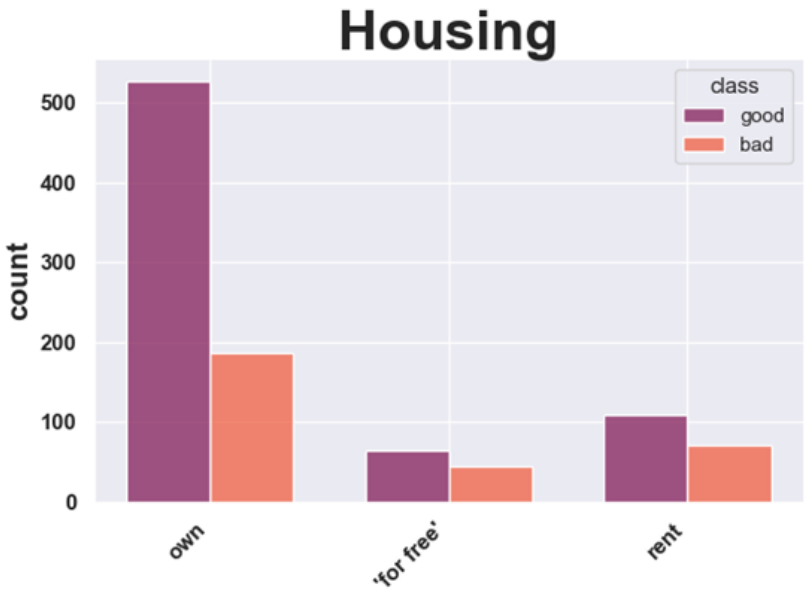
Atributo 14:



Other payment plans representa outras formas de pagamento.

-

Atributo 15:



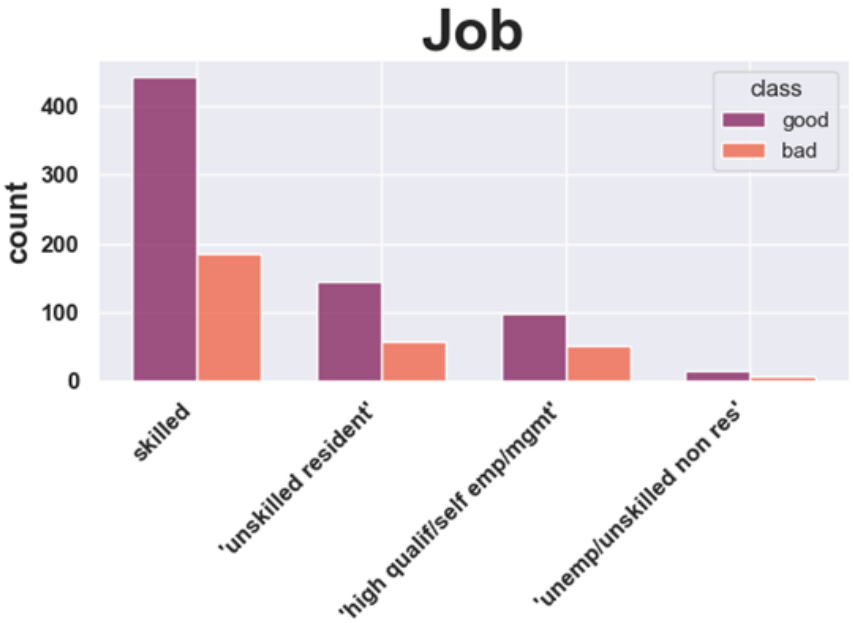
Housing representa que tipo de pagamento/propriedade que o cliente possui. O cliente pode ser proprietário, alugar ou morar de graça.

-

Atributo 16:

Existing credits representa grupos segregados pelo número de outras obrigações financeiras já existentes. Considerando que não há quase nenhum dado para o grupo quatro quando comparado às outras categorias, é possível afirmar que essa categoria deve ser desconsiderada para a construção do modelo na fase de pré-processamento.

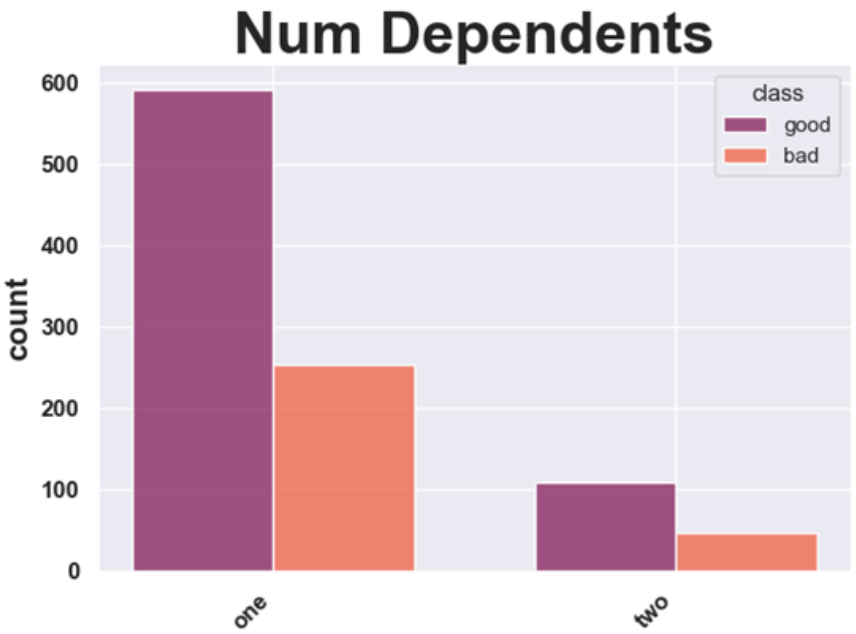
Atributo 17:



Job representa que categoria de emprego o cliente possui a categoria "unemp\unskilled non res" deve ser removida em pré-processamento considerando a falta de dados para tal categoria. Essa categoria agrupa não residentes desempregados / sem habilidades especiais

-

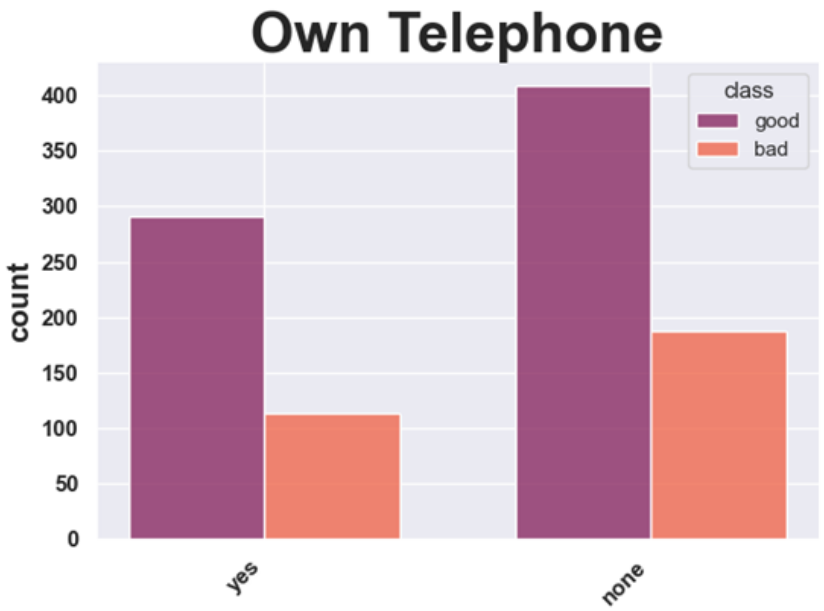
Atributo 18:



Num dependents representa o número de dependentes que o cliente possui.

-

Atributo 19:

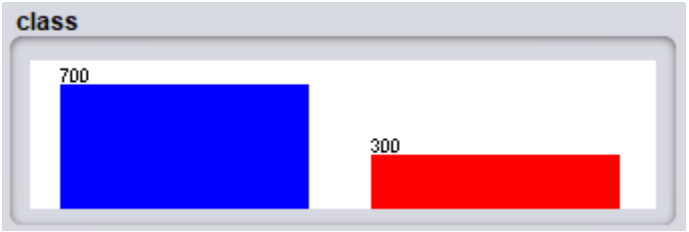


Own Telephone representa se o cliente possui um telefone ou não.

Atributo 20:

Foreign Worker representa se um trabalhador é estrangeiro ou não. Esse atributo pode ser desconsiderado para o modelo por conta da pequena quantidade de aplicantes que não são trabalhadores estrangeiros.

Distribuição de frequência do atributo classe:



Variável "class": é a variável classifica o tomador de empréstimo em risco de crédito bom ou ruim.
São 700 amostras classificadas em risco de crédito bom e 300 classificadas e risco de crédito ruim.

-

Distribuição dos atributos separados por classe:

Variável "checking status": Existência de marcos (M) na conta corrente, por faixa de valor

A11 : M <0 | A12 : 0<=M<200 | A13 : M>= 200 | A14 : sem conta bancária

No que se refere ao status da conta corrente, pode-se dizer que para pessoas que possuem menos de 200 marcos e ou conta negativa, o dataset tem uma distribuição relativamente simétrica entre risco de crédito bom e ruim. Já para quem tem valores maiores de 200 ou quem não tem conta, o dataset classifica sua maioria em crédito bom. Percebe-se também que quase 40% da base não possui conta corrente.

Variável "duration": mede quanto tempo se levou para pagar os empréstimos, em meses

O total de meses varia entre 4 e 72, com a maior concentração de valores até 38 meses.

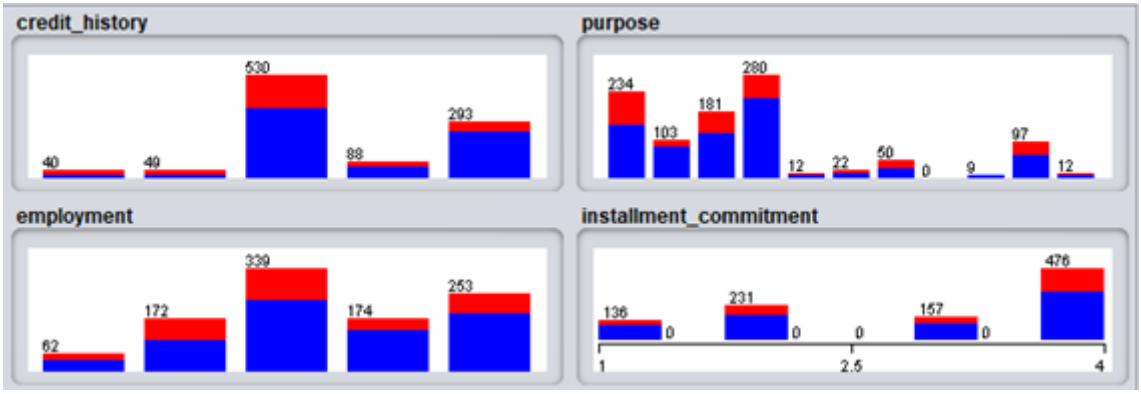
Variável "credit amount": valor do crédito concedido

Valores de 250 a 18424. Percebe-se uma incidência maior de risco de crédito ruim para montantes mais baixos e para montantes muito altos.

Variável "savings_status": faixas de valores de poupança/títulos

A61: M < 100 | A62: 100 <= M < 500 | A63: 500 <= M < 1000 | A64: M >= 1000 | A65: Desconhecido/Inexistente

A maior concentração de classificações em risco de crédito ruim é para valores abaixo de 100 marcos, enquanto poupança inexistente ou desconhecida recebem maior parte da classificação de risco bom de crédito.



Variável "credit_history": descreve o comportamento de tomada de créditos

A30: nunca tomou crédito/quitou todos | A31: todos os créditos quitados com o banco | A32: créditos existentes sendo pagos corretamente | A33: histórico de atraso no pagamento | A34: conta crítica/existência de financiamento em outros bancos.

A amostra está concentrada principalmente em quem tem financiamentos com o banco, sendo pagos corretamente ou com outros bancos e contas críticas. Percebe-se uma tendência maior em classificar créditos sendo pagos em risco de crédito ruim do que com quem tem histórico de atraso.

Variável "purpose": descreve o motivo do financiamento

A40: carro novo | A41: carro usado | A42: móveis/equipamentos | A43: rádio/televisão | A44: aparelhos de casa | A45: consertos | A46: educação | A47 : férias | A48: treinamento | A49: negócios | A410 : outros.

A amostra dos financiamentos se concentra na compra de bens de consumo em geral e também uma tendência de maior classificação de risco de crédito ruim nessa seção. Não existe valores para motivo de férias, o que pode indicar várias coisas, entre elas: como motivo "novo" na base e não foram coletados elementos suficientes com esse motivo ou simplesmente não se preenchia o motivo quando era férias.

Variável "employment": diz se a pessoa está empregada e há quantos anos

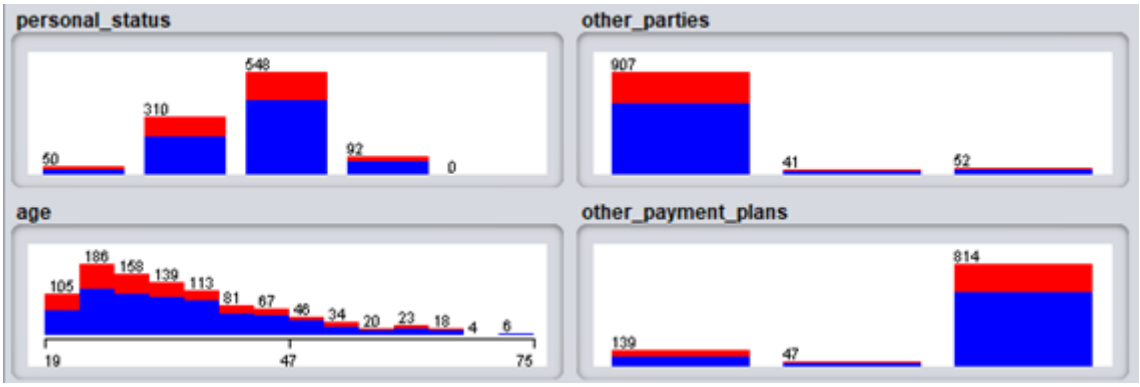
A71: desempregado | A72: Há menos de um ano | A73: entre 1 e 4 anos | A74: entre 4 e 7 anos | A75: mais de 7 anos empregado

Parece haver um bom balanceamento da amostra por faixa e também uma proporção parecida de risco de crédito bom/risco de crédito ruim entre as faixas.

Variável "installment_commitment": comprometimento da parcela do empréstimo em relação à renda

É um índice que mede a proporção da parcela do financiamento e na renda disponível, variando de 1 a 4, sendo 1, a parcela não tem potencial de comprometer a renda e 4 sugere que a parcela compromete significativamente a capacidade de pagamento, podendo resultar em inadimplência.

Percebe-se um balanceamento razoável da amostra entre as faixas e uma tendência esperada em uma proporção maior em classificar um comprometimento maior da renda com a parcela em risco de crédito ruim.



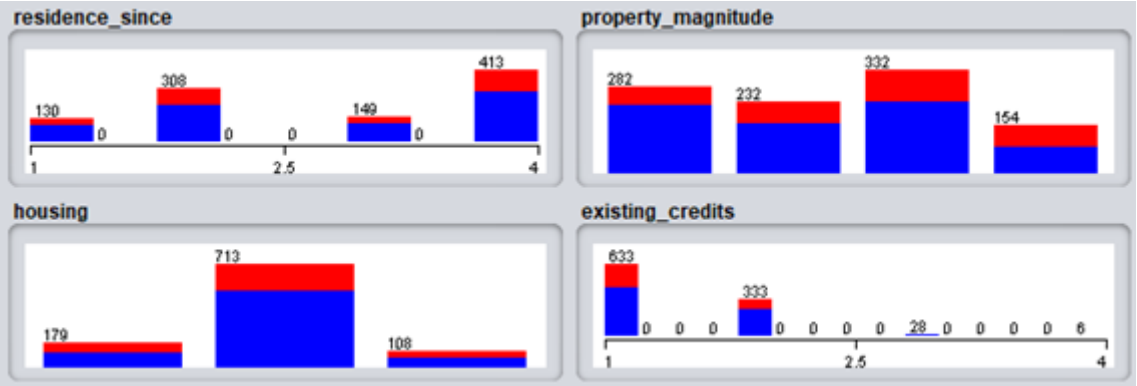
Variável "personal_status": cruza o gênero com o estado civil
A91 : homem divorciado/separado | A92: mulher divorciada/separada/casada | A93 : homem solteiro | A94 : homem casado/viúvo | A95: mulher solteira
A amostra se concentra em homens solteiros e mulheres divorciadas/separadas/casadas, com a maior incidência de classificação de risco de crédito ruim nesses grupos.
Outro fator interessante é a pouca informação sobre mulheres, não tendo informações sobre mulheres solteiras. Além disso, para mulheres, o estado civil não parece ser relevante para a classificação de risco de crédito.

Variável "other_parties": existência de garantidores/co-participantes
A101 : sem garantia | A102: co-participante | A103 : garantidor
A amostra sugere que praticamente todos os empréstimos com risco de crédito ruim pertencem ao grupo dos financiamentos sem um garantidor/co-participante.

Variável "age": relaciona a idade dos tomadores de financiamento
Valores de 19 a 75 anos. Amostra aparentemente balanceada para toda a faixa de idade e uma proporção relativamente estável entre risco de crédito bom e ruim ao longo da faixa de idade.

Variável "other_payment_plans": existência de outros financiamentos
A14: bancos | A142: lojas | A143: sem outros financiamentos
A amostra pode sugerir que ter financiamento em outros bancos é um fator relevante para se classificar em um risco de crédito ruim.

-



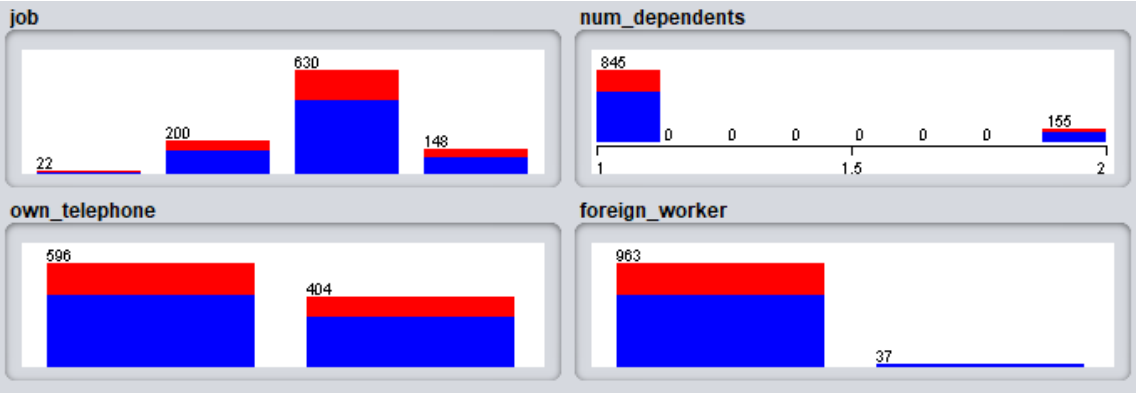
Variável "residence_since": tempo de permanência na residência atual em anos
É uma variável numérica discreta, com valores 1, 2, 3, 4. Existe um bom balanceamento de amostra entre os anos e as classificações de crédito.

Variável "property_magnitude": assinala a existência de ativos e quais são esses
A121 : residência | A122 :se não tem residência, tem seguro de vida ou outros serviços financeiros | A123 : se não tem residência nem serviços financeiros, possui automóvel ou outros que não sejam poupanças ou títulos | A124: não tem ativos ou sem informação.
Bom balanceamento entre as faixas e as classificações de risco de crédito.

Variável "housing": status da residência
A151: aluguel | A152 : residência própria | A153: doação
A maior concentração de classificação em risco de crédito ruim fica na faixa de quem possui residência própria. Vários fatores podem explicar isso, entre eles o não pagamento correto dos financiamentos da residência.

Variável "existing_credits": número de financiamentos existentes no banco
Valores discretos (1,2,3,4). A maior parte da amostra possui 1 ou 2 financiamentos com o banco e o risco de crédito ruim se concentra nessa parte.

-



Variável "job": cruza a situação e tipo de trabalho dos tomadores de financiamento com o status de ser ou não residente

A171 : desempregado/pouco qualificado-não residente | A172 : pouco qualificado - residente | A173: bem qualificado | A174 : gerente/ autônomo/ função bem qualificada/ chefe

A maior concentração de amostra com risco de crédito ruim se encontra nos trabalhadores bem qualificados.

Variável "num_dependents": número de dependentes

Possui valores de 1 e 2. A maior concentração de amostra com risco de crédito ruim se encontra com quem tem 1 dependente.

Variável "own_telephone": diz se possui linha telefônica ou não

A191 : sem linha | A192 : com linha registrada em seu nome

Nos anos 90, a linha de telefone possuía um valor de ativo como um bem, e isso poderia resultar em condições positivas reais de tomada de financiamento. Nos dias atuais, essa variável não é relevante para explicar comportamento de crédito e provavelmente será descartada na fase de ETL.

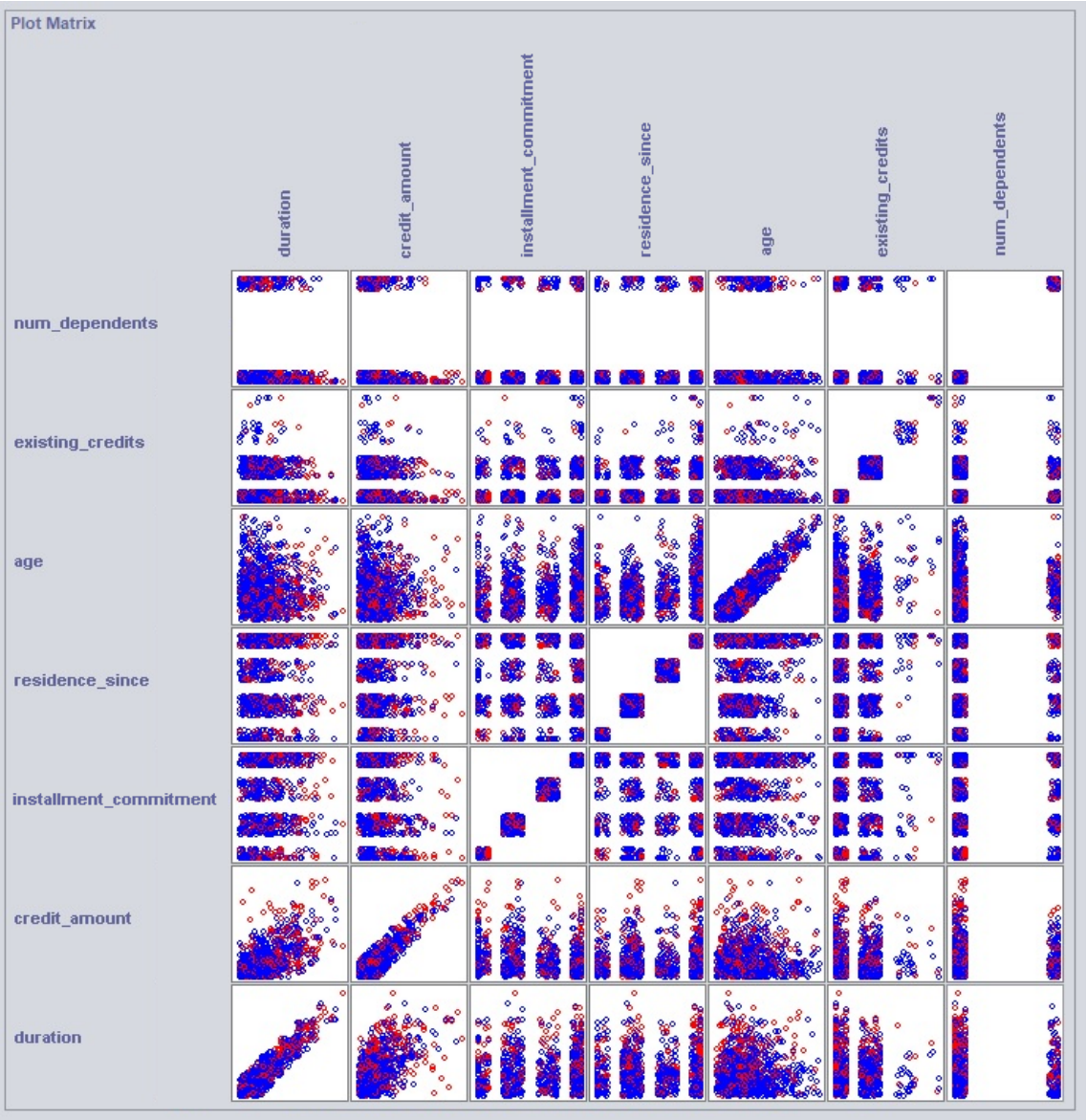
Variável "foreign_worker": diz se o tomador de financiamento é um trabalhador estrangeiro ou não.

A201 : sim | A202 : não

A amostra sugere um desbalanceamento visto que quase o total da amostra se concentra em trabalhadores estrangeiros e todas as classificações em risco de crédito ruim se encontram neste grupo. Temos que pensar em como tratar esse problema para que não exista uma tendência do modelo de Machine Learning em classificar qualquer tomador de empréstimo estrangeiro como alto risco de inadimplência.

-

Atributos em partes, com scatter plots:



	duration	credit_amount	installment_commitment	residence_since	age	existing_credits	num_dependents
duration	1.000	0.016	0.004	0.002	-0.001	0.000	0.000

credit_amount	0.016	1.000	-0.012	0.001	0.001	0.000	0.000
installment_commitment	0.004	-0.012	1.000	0.004	0.002	0.001	-0.004
residence_since	0.002	0.001	0.004	1.000	0.011	0.004	0.002
age	-0.001	0.001	0.002	0.011	1.000	0.003	0.003
existing_credits	0.000	0.000	0.001	0.004	0.003	1.000	0.003
num_dependents	-0.001	0.001	-0.004	0.002	0.003	0.003	1.000

Tabela XX: Matriz de Correlação (Pearson)

Observando os gráficos de dispersão e a matriz de correlação (coeficiente de Pearson), concluímos que não há correlação entre os atributos preditores. Também não foi possível identificar nos gráficos um par de atributos, que seja importante na identificação de um padrão para separação das classes.

3. Pré-processamento

Nessa etapa preparamos os dados com base nos conhecimentos adquiridos ao analisar o comportamento dos dados.

Na etapa 2 identificamos a necessidades de limpar as seguintes informações:

- ➔ *No Atributo Duration foi eliminado os empréstimos de 60 meses para cima, porque são outliers;*
- ➔ *No Atributo Checking Status foi alterado categoria>=200 por zero, porque são empréstimo se ter conta corrente no banco;*
- ➔ *No Atributo Credit Amount foi alterado o valor total do empréstimo acima de 18 mil por zero, porque são outlier;*
- ➔ *O Atributo Other Parties foi removido do dataset, observamos que todos os cliente se encontram na categoria "sem nenhum garantidor";*

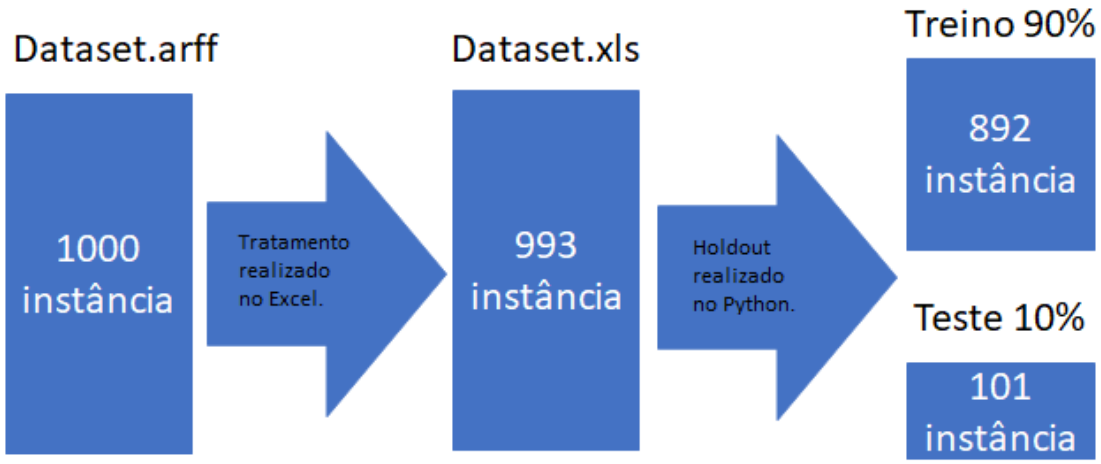
- ➔ *No Atributo Age foi eliminado os clientes com a idade acima de 70 anos, porque consideramos outliers;*

- ➔ *No Atributo Existing Credits foi eliminado as categorias four and three, porque não é relevante para o negócio;*
- ➔ *No Atributo Job foi eliminado a categoria unemp\unskilled non res por dados inconsistentes. Essa categoria agrupa os "não residentes desempregados / sem habilidades especiais";*
- ➔ *No Atributo Foreign Worker foi eliminado do dataset, porque não é pertinente ao negócio;*

Após a limpeza de dados o dataset ficou com 993 instâncias para o experimento.

Utilizamos o excel para realizar o tratamento de limpeza e a ferramenta python para criar o holdout.

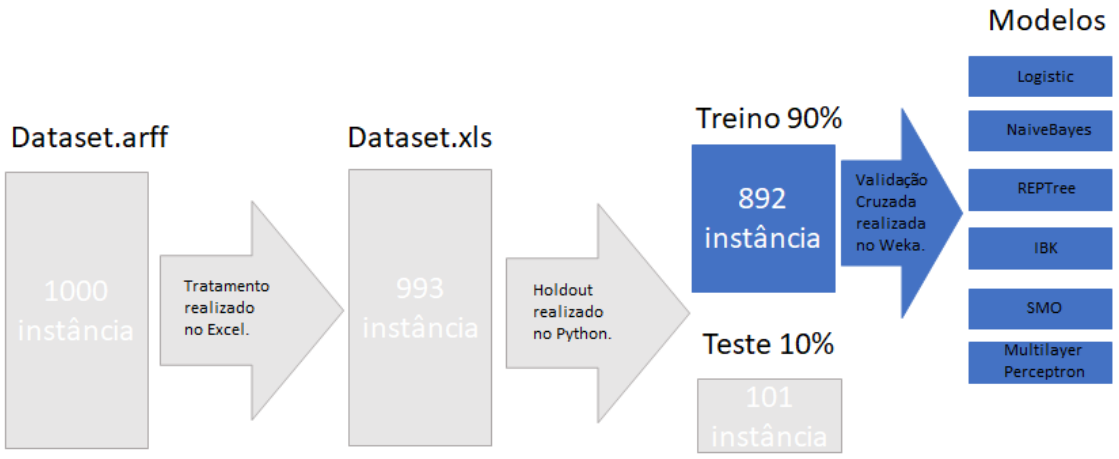
Decidimos utilizar o python para criar o holdout, porque a ferramenta weka não aceita a extensão xls.



O dataset foi dividido em 90% para treino e 10% para teste. Resultando 892 instâncias para treino e 101 instâncias para teste.

4. Modelagem e inferência

Nessa etapa criamos a validação cruzada, utilizando a massa de dados para treino. Foi utilizado 6 modelos como experimento: Logistic, Naive Bayes, REPTree, IBK,SMO e Multilayer Perceptron.



Início dos experimentos utilizando Dataset treino:

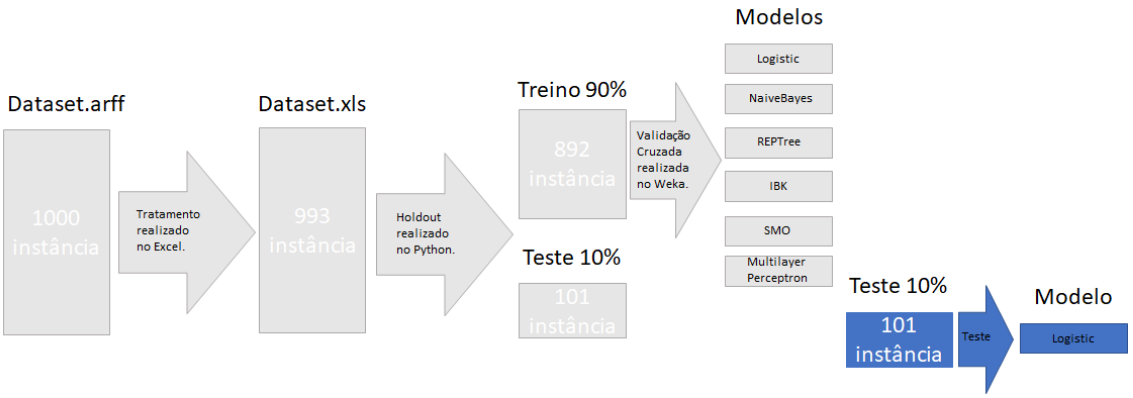
Resumo-Modelos de Classificação no Weka		
Modelo	**Biblioteca Weka**	**Acurácia**
Regressão Logística - Logistic	**functions.Logistic**	**78.59%**
NaiveBayes - NaiveBayes	bayes.NaiveBayes	75%
Classification and RegressionTrees (CART) - REPTree	trees.REPTree	72.53%
k-Nearest Neighbors (KNN) - IBk	lazy.IBK	72.08%
Support Vector Machines (SVM) – SMO	functions.SMO	75.89%
Redes Neurais multicamadas	functions.MultilayerPerceptron	71.63%

A tabela acima representa o desempenho da acurácia de cada modelo treinado.

5. Pós-processamento

-

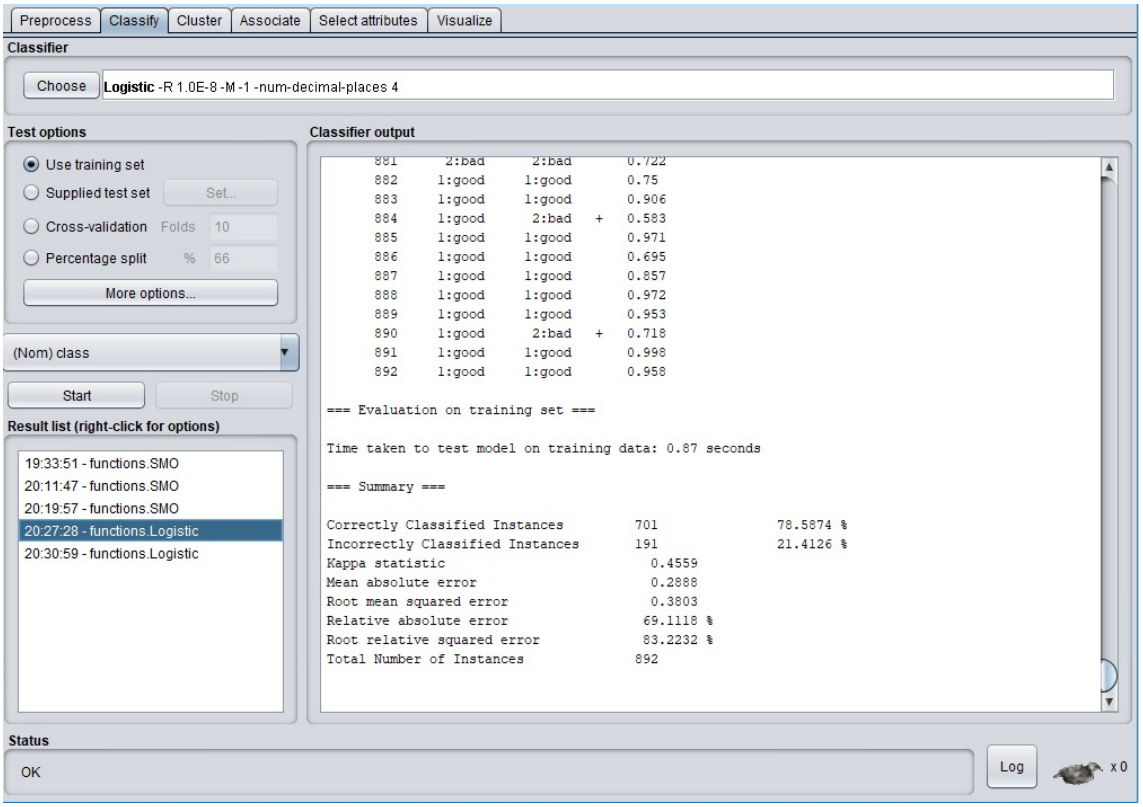
Nessa etapa é realizado o experimento do modelo escolhido, utilizando as instâncias de teste. O modelo escolhido foi Logistic, porque a acurácia foi de 78.59% teve o melhor desempenho com relação aos demais modelos utilizados.

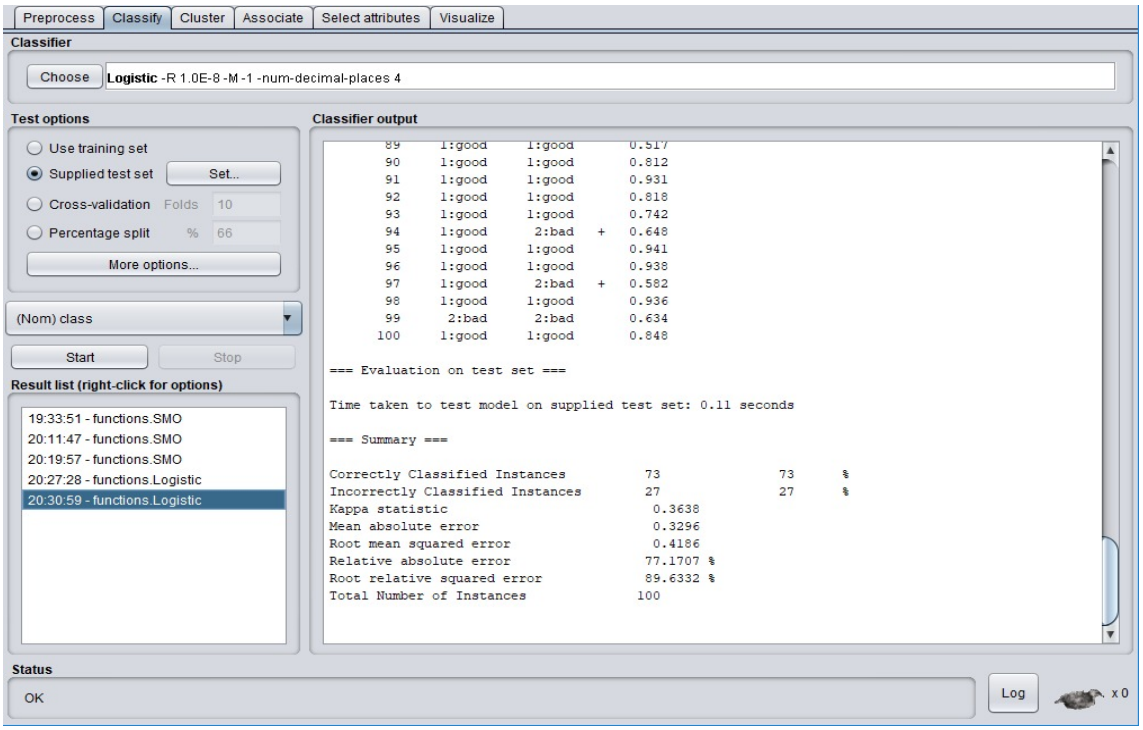


Início dos experimentos utilizando Dataset teste:

Introdução a Ciência de Dados

-





-

O modelo foi treinado com êxito. Considerando que aproximadamente 70% dos aplicantes são aprovados para empréstimo, a métrica escolhida foi acurácia e o benchmark escolhido foi 70% considerando a distribuição de "class". A acurácia de treino foi de 78.6% usando 10 fold cross validation e a de teste foi de 73%. A acurácia de teste poderia ser mais alta . O resultado, embora aceitável, poderia ser melhorado com ajustes feitos ao modelo, sua função de custo e aos hiperparâmetros. Ele também poderia ser melhorado com novos parâmetros por nenhum parâmetro individual aparenta ter uma correlação muito forte com a classificação. Recomendamos novos dados coletados.

6. Apresentação de resultados

Nessa etapa o melhor modelo deve ser apresentado da forma mais clara possível para os gestores, explicando a metodologia adotada apresentado a solução do problema e pontuando a descobertas na massa de dados.

Dados de crédito Alemão.(fig.Ilustrativa)

- *Qual era o problema?*

Classificar as pessoas descritas por um conjunto de atributos como riscos de crédito bons ou ruins

- *Qual foi a solução proposta?*

Elaborar uma análise para obter uma visão preliminar sobre a credibilidade de um solicitante de crédito ao analisar uma demanda de crédito. Como seus modelos preditivos consomem muitos recursos e podem ser executados por um longo tempo, essa visualização preliminar pode ser útil na eliminação de demandas insatisfatórias antes de executar modelos no restante.

Em segundo lugar, pelos próprios solicitantes de crédito se mostra útil para estimar sua credibilidade antes de fazer uma solicitação oficial aos seus bancos. Isso os ajudará a adequar sua demanda de crédito de acordo ou adiá-la até o momento em que atinjam uma pontuação de credibilidade favorável.

- *Quais foram as principais descobertas?*

Carro novo e televisão / rádio são os principais motivos para a maioria dos candidatos buscarem um crédito. Infelizmente ou felizmente, muito poucos candidatos estão buscando um crédito para educação ou reciclagem. Isso pode significar que a educação e o retreinamento não valem um crédito ou que seus custos foram totalmente cobertos pelas escolas, universidades, governo ou de alguma outra forma que parece muito improvável.
Os empregados qualificados e os homens casados / viúvos também são responsáveis por quase dois terços da soma do crédito. As mulheres tomadoras de crédito nos dados estão na extremidade inferior, o que supera a lógica, considerando o fato de que ambos os sexos enfrentam as mesmas dificuldades financeiras.

- *Quais foram as limitações e dificuldades encontradas?*

- *Quais são as principais conclusões?*

Quando um banco recebe um pedido de empréstimo, com base no perfil do solicitante, o banco deve tomar uma decisão quanto a prosseguir ou não com a aprovação do empréstimo. Dois tipos de riscos estão associados à decisão do banco.

- *Se o candidato apresenta um bom risco de crédito, ou seja, é provável que pague o empréstimo, a não aprovação do empréstimo para a pessoa resulta em perda de negócios para o banco*

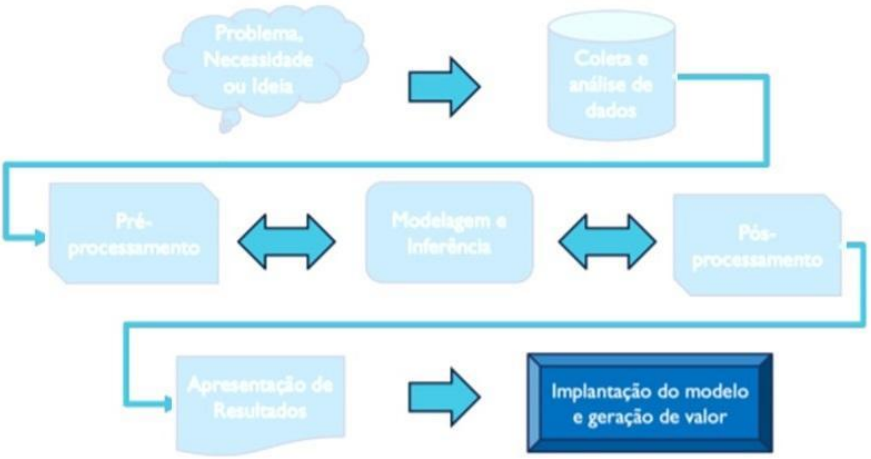
- *Se o requerente apresenta um risco de crédito ruim, ou seja, não é provável que pague o empréstimo, a aprovação do empréstimo para a pessoa resulta em perda financeira para o banco.*

Nosso objetivo foi:

Para minimizar a perda do ponto de vista do banco, o banco precisa de uma regra de decisão sobre quem deve dar a aprovação do empréstimo e quem não deve. Os perfis demográficos e socioeconômicos de um candidato são considerados pelos gerentes de empréstimos antes de uma decisão ser tomada em relação ao seu pedido de empréstimo.

O software Weka pode sim trabalhar com grandes conjuntos de dados, porém, fica limitado nas aplicações com interface gráfica especialmente na versão Explorer do aplicativo por trabalharem somente com dados armazenados na memória principal do computador e podendo ocorrer sobrecargas para visualização destes dados, limitando a quantidade de dados que pode ser processada.

7. Implantação do modelo



Para a predição de novos dados:

Para a predição de novos dados faríamos análises futuras com alguns ajustes no modelo **Logistic** Regression no próprio Weka para tentar melhorar ainda mais as predições e monitorar novas entradas de massas de dados .

Conclusão de Nossos Algoritmos de Machine Learning via WEKA:

Utilizamos em nosso experimento 6 algoritmos de classificação e realizamos outros experimentos e comparativos a parte , no qual apenas dois apresentaram melhores resultados: SVM e **Logistic** regression. Escolhemos Regressão Logística (functions.Logistic) pela melhor Acurácia no conjunto de treino (892) instâncias de 1000 , atingindo o resultado de 701 instâncias classificadas corretamente , representando 78.58% do treinamento.O algoritmo **Logistic** , por ser um algoritmo de classificação binária se mostrou eficaz , pois neste dataset escolhido ele assumiu que as entradas são numéricas, e aprendeu um coeficiente para cada característica de entrada , linearmente combinados em uma função sigmóide (em forma de 's') que mapeia as saídas em valor zero e um , que no nosso problema de negócio seria 1 = concede crédito (bom pagador) , e 0(zero) = não conceder crédito (mal pagador).

Para o conjunto de teste usamos uma massa de Dados de 100 instâncias , no qual nosso modelo previu uma acurácia de 73% , ou seja atingiu o resultado de 73 instâncias classificadas corretamente.

Concluímos a partir do estudo realizado e dos resultados obtidos, que através da aplicação de técnicas de Machine Learning (Aprendizado de Máquina), a classificação de risco de créditos pode ser feita por instituições financeiras de forma autônoma e assertiva.