Resumo Técnico da análise

Metodologia usada:
CRoss Industry Standard Process for Data Mining (CRISP-DM)

- **1 Entendimento de negócio:**

Precisamos diminuir ao máximo o número de inadimplentes e conceder o empréstimo a bons pagadores, algumas KPIs importantes:

- 1. prejuízo médio causado por um inadimplente.
- 2. lucro médio gerado por cliente que paga corretamente com os juros.
- 3. Quantidade de dinheiro do empréstimo dos não inadimplentes e dos inadimplentes.
- 4. custo médio de classificar um bom pagador como mal pagador e não conceder crédito, durante a modelagem.
- 5. Diferença entre não usar o modelo e usar o modelo no lucro final?

- **Até o final da análise queremos responder as perguntas de negócio da etapa 1, mas prosseguindo:**
- Iremos utilizar uma abordagem preditiva, pois neste caso a interpretação dos dados no momento da modelagem não é nosso foco, para validar os modelos vamos utilizar as métricas Kappa, MCC, G-mean, F1-score e KS(medir distinção das classes que o modelo faz) pois o Kappa, MCC e o G-mean considera as os verdadeiros positivos e negativos que é uma coisa importante nessa modelagem já que queremos reduzir gastos e maximizar os lucros, o f1-score para termos uma noção de como está a classificação dos verdadeiros positivos e a precisão do modelo.

- **2 Entendimento dos dados:**
  Temos dados passados de corte transversal que possuem as seguintes colunas(Temos 7 variáveis numéricas, 3 categóricas) :
  - person_age : Idade da pessoa pedindo o empréstimo. (numérica)
  - person_income : Renda da pessoa (numérica)
  - person_home_ownership : A casa em que a pessoa vive pertence a ela ou é alugada? (categórica)
  - person_emp_length : A quanto tempo a pessoa está empregada em anos. (inteiro)
  - loan_intent : Qual a intenção da pessoa ao pedir o empréstimo? existem algumas opções prontas, como por razoes medicas, educação entre outras. (categórica)
  - loan_grade : grau de empréstimo (categórica ordinal)
  - loan_amnt : qual a quantidade do empréstimo? (numérica)
  - loan_int_rate : Taxa de juro (ponto flutuante)
  - loan_percent_income : Renda percentual (ponto flutuante)
  - cb_person_default_on_file : se a pessoa já teve histórico de inadimplência antes ou não. (categórica dicotômica)
  - cb_preson_cred_hist_length : tamanho do histórico da pessoa, em anos. (inteiro)
  - loan_status : situação do empréstimo é a nossa variável alvo (categórica dicotomica) onde 1 indica inadimplência e 0 indica não inadimplência.

Variável resposta desbalanceada

- porcentagem de inadimplentes ~ 21.81%
- porcentagem de pagantes ~ 78.18%

Para tratar o desbalanceamento foi utilizada uma abordagem padrão onde a distribuição dos dados é mantida intocada.

total de linhas: 19548

total de colunas: 13

Consistência e qualidade dos dados

- Foi verificado que alguns valores para as idades eram irreais, igual a 144 anos, essas foram igualadas ao valor mais alto e plausível que era 94 anos.
- Dados faltantes: Havia dados faltantes nas variáveis:
- `person_emp_length = 552 ~ 2% das linhas`
- `loan_int_rate = 1883 ~ 9% das linhas`
- `Para entender melhor os dados essas linhas com dados faltantes foram excluídas da análise exploratória, com os insights da análise exploratória foi possível propor técnicas para substituir os dados faltantes.`

Descrição dados inadimplentes

	person_age	person_income	person_emp_length	loan_amnt	loan_int_rate	loan_percent_income	cb_person_cred_hist_length	l
count	4265.000000	4265.000000	4093.000000	4265.000000	3849.000000	4265.000000	4265.000000	
mean	27.445252	49678.865182	4.139995	10861.776084	13.032741	0.245341	5.703165	
std	6.221209	37801.329057	3.908621	7179.303678	3.301576	0.131263	4.137534	
min	20.000000	4000.000000	0.000000	1000.000000	5.420000	0.010000	2.000000	
25%	23.000000	30000.000000	1.000000	5000.000000	10.740000	0.140000	3.000000	
50%	26.000000	41682.000000	3.000000	9600.000000	13.480000	0.230000	4.000000	
75%	30.000000	60000.000000	6.000000	15000.000000	15.580000	0.340000	8.000000	
max	70.000000	604000.000000	34.000000	35000.000000	22.480000	0.770000	30.000000	

Descrição dados não inadimplentes

	person_age	person_income	person_emp_length	loan_amnt	loan_int_rate	loan_percent_income	cb_person_cred_hist_length
count	15283.000000	1.528300e+04	14903.000000	15283.000000	13816.000000	15283.000000	15283.000000
mean	27.788785	7.035327e+04	4.956049	9242.020546	10.466796	0.148885	5.821239
std	6.371859	5.198471e+04	4.052867	6011.873228	2.979545	0.086974	4.018375
min	20.000000	9.400000e+01	0.000000	500.000000	5.420000	0.000000	2.000000
25%	23.000000	4.200000e+04	2.000000	5000.000000	7.740000	0.080000	3.000000
50%	26.000000	6.000000e+04	4.000000	8000.000000	10.620000	0.130000	4.000000
75%	30.000000	8.452700e+04	7.000000	12000.000000	12.690000	0.200000	8.000000
max	144.000000	1.782000e+06	38.000000	35000.000000	22.060000	0.700000	30.000000

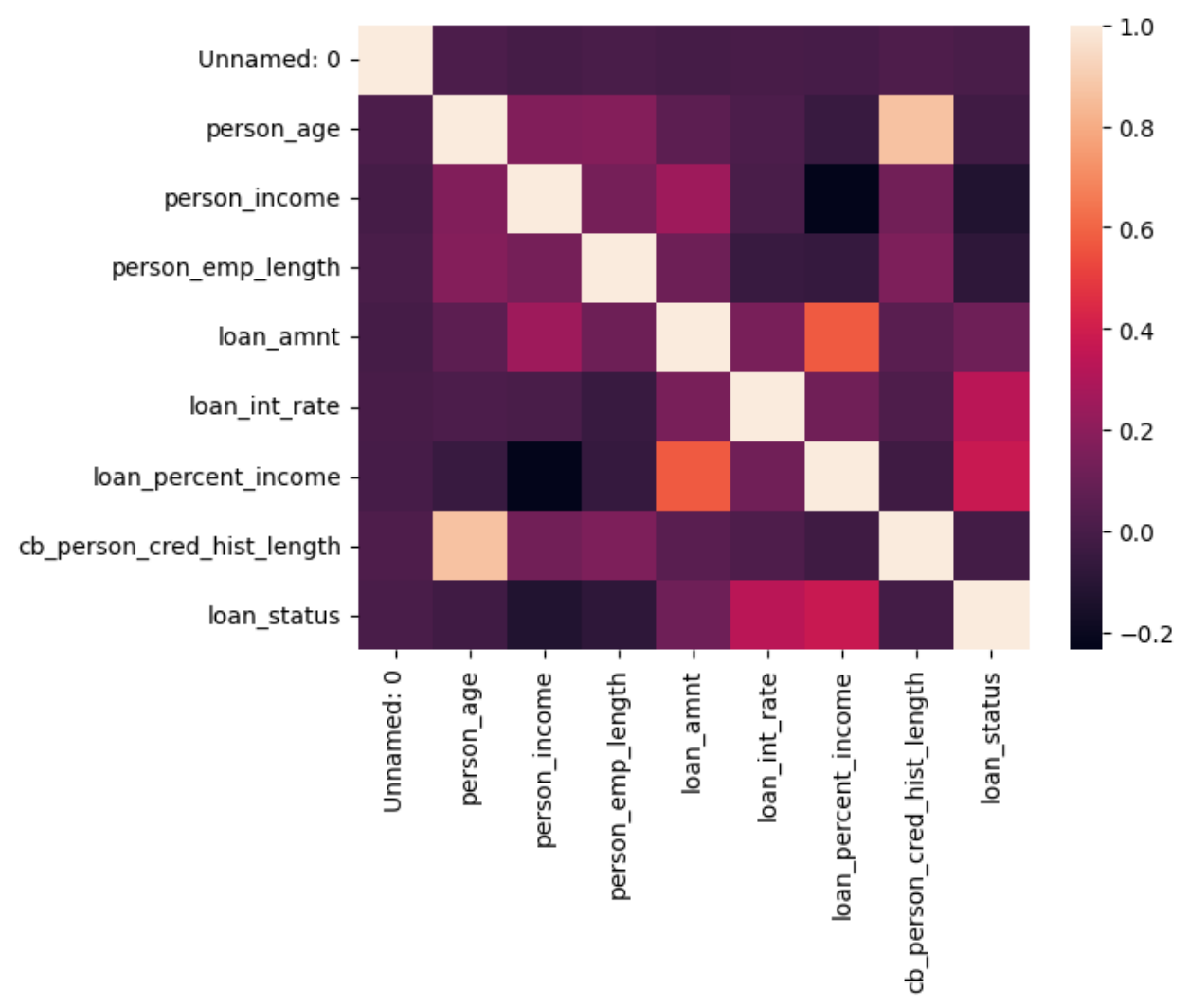
Já Podemos notar uma coisa

- Não seria legal substituir o loan_int_rate faltante por sua média, pois modifica a distribuição do loan_int_rate do grupo dos inadimplentes e dos não inadimplentes, e por consequência nos faz perder informação, precisamos de uma solução mais rebuscada para isso.
- O mesmo vale para o `person_emp_length`

Testes de hipótese sobre a distribuição das variáveis numéricas entre o grupo de inadimplentes e não inadimplentes foram feitos, afim de verificar se existia diferença na distribuição dos dois grupos

- Um detalhe importante é que para todos os testes era utilizada uma subamostra de cada grupo de tamanho 300 pois todos os testes utilizados são sensíveis ao tamanho da amostra, e eles eram realizados multiplas vezes para ver a distribuição do p-valor.
- Foi utilizado o teste de Shapiro wilk para verificar a normalidade, no caso de haver normalidade o teste-t de medias, caso não houvesse normalidade o teste não paramétrico de Mann Whitney era aplicado.

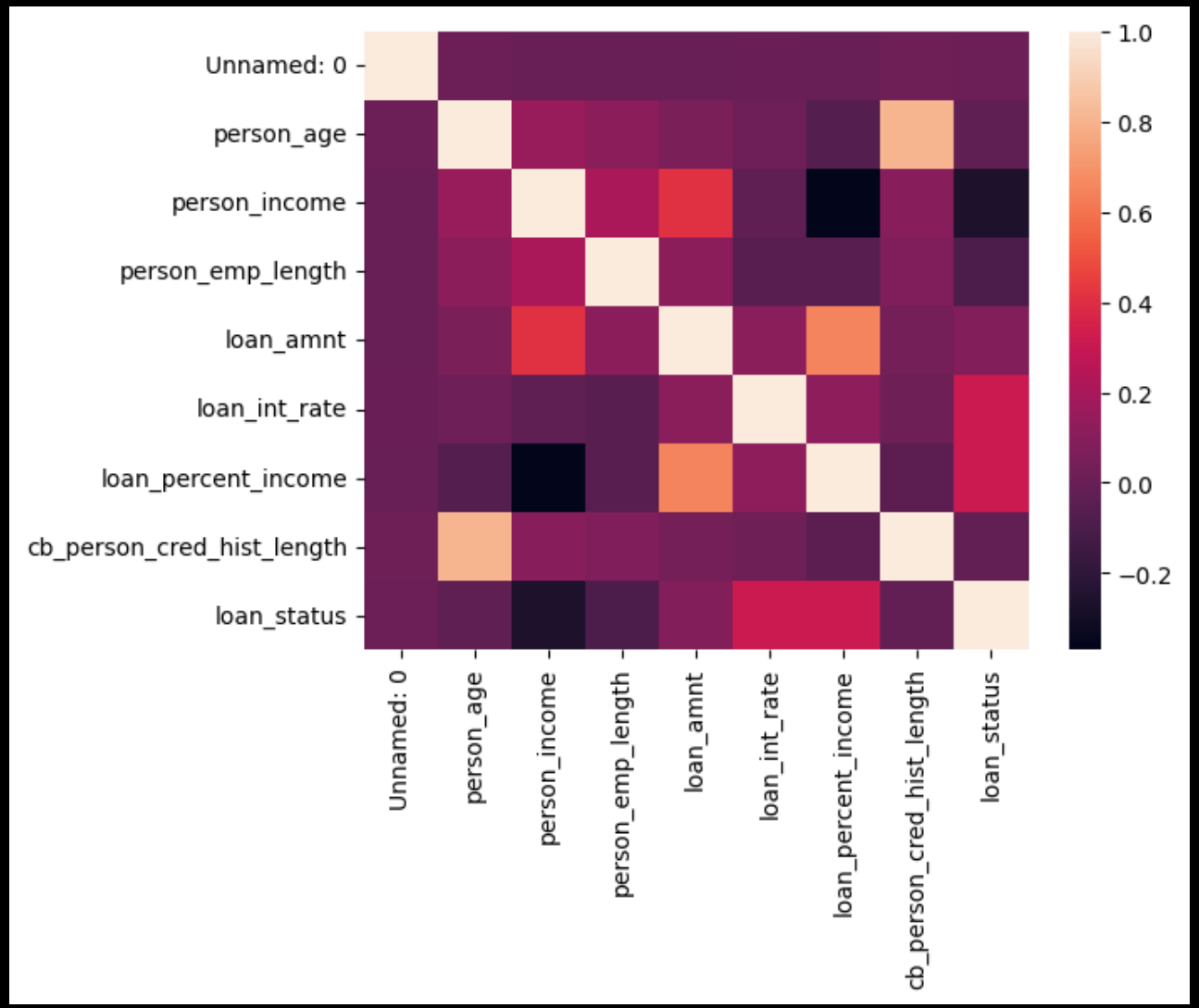
Verificação de correlação linear de pearson entre as variáveis numéricas.



Se destacam as correlações:

- correlação positiva entre idade e tempo de empregamento.
- correlação positive entre quantidade do empréstimo e porcentagem que o empréstimo representa da renda do cliente.
- correlação negativa entre renda do cliente e porcentagem que o empréstimo representa da renda do cliente.

Correlação não paramétrica de Spearman:



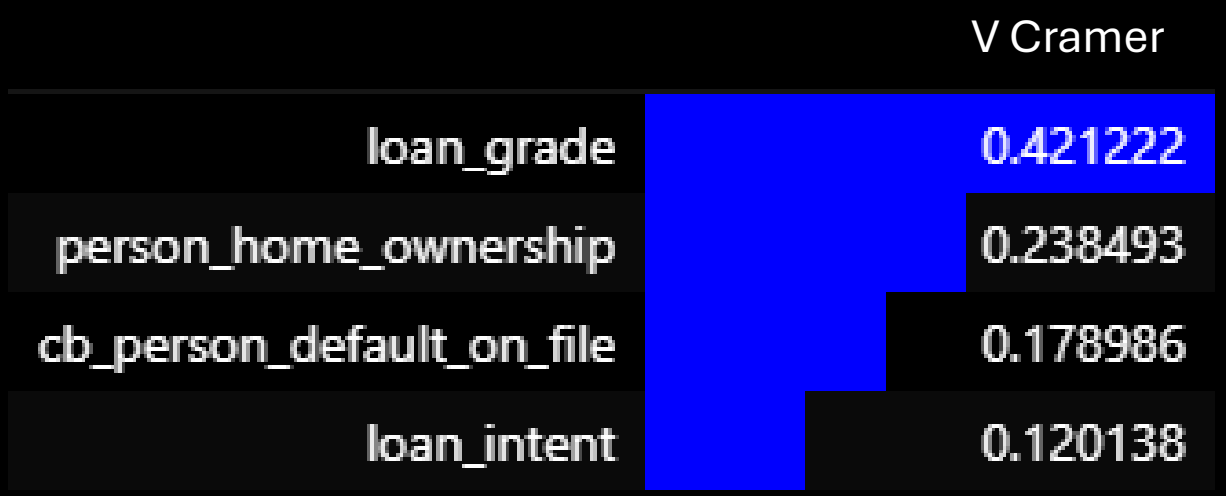
- Mesmas correlações, mas com o aumento em uma correlação:
- - correlação não linear positive entre quantidade do empréstimo e renda do cliente.

Teste de hipótese em relação as variáveis categóricas e a variável resposta.

- O teste qui-quadrado, que também é sensível ao tamanho da amostra, foi utilizado para testar relação de dependência entre as variáveis categóricas, considerando um tamanho de amostra igual a 300 foi testado 50 vezes os testes de dependência. No dicionário é visto a quantidade de vezes que a hipótese nula é rejeitada e temos que existe uma relação entre a variável resposta e a categórica:

```
{'cb_person_default_on_file': 37,
 'loan_grade': 50,
 'loan_intent': 16,
 'person_home_ownership': 48}
```

Além de verificar se existe relação entre as variáveis categóricas e a variável resposta é importante verificar a força dessa relação. Utilizando o coeficiente V de Cramer foi possível medir essa relação:



- Unindo valores dos testes e do coeficiente vemos que o grau do empréstimo possue forte relação com a variável resposta, logo em seguida o tipo de residencia, se a pessoa já foi inadimplente e por ultimo a justificative para o empréstimo.

Com essas visualizações e testes tiramos alguns insights dos dados.

- Todas as variáveis categóricas possuem relação significativa com a variável resposta.

- Entre as variávies categóricas a que possue maior correlação com a variável resposta é a variável loan_grade que representa o grau do empréstimo com V de cramer igual a 0.42, logo em seguida person_home_ownership que representa o tipo de residência do cliente com V de cramer igual a 0.23, a variável cb_person_default_on_file que indica se a pessoa já foi inadimplente ou não com v de cramer 0.178986 e por ultimo loan_intent que indica a justificativa para o empréstimo que tem V de cramer 0.12

- Existe diferença na distribuição das variávies numéricas para o grupode inadimplentes e não inadimplentes.

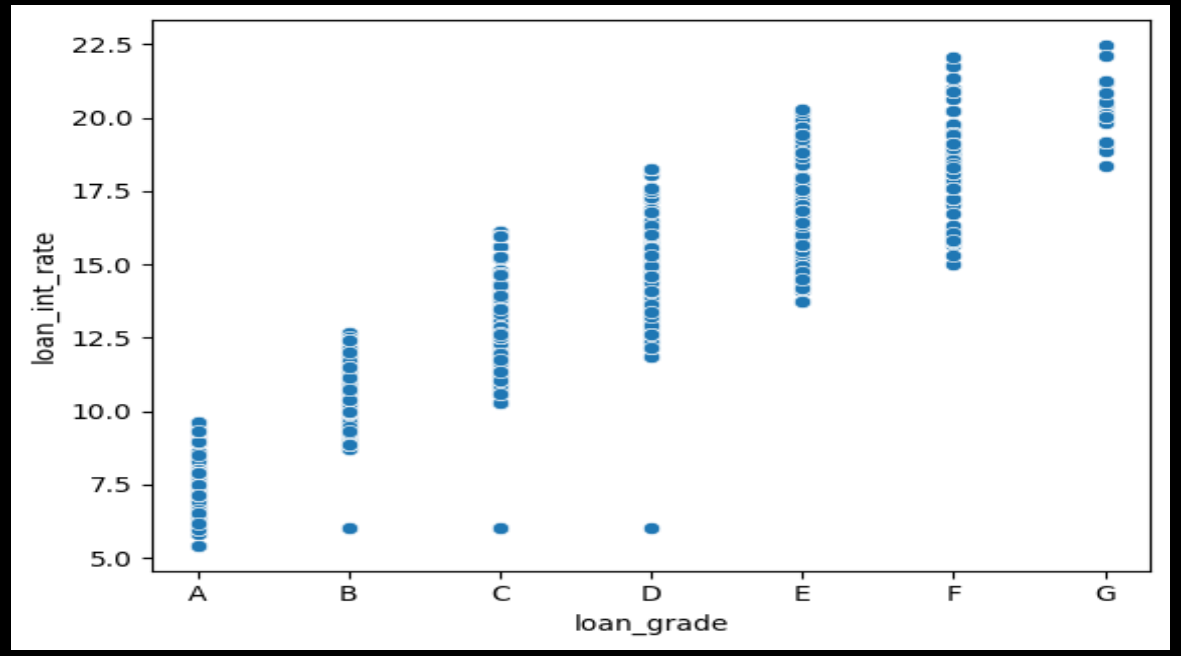
- Existe uma forte correlação positiva linear e não linear entre a idade da pessoa, e o tempo de empregamento, que é um dos dados faltantes e podemos utilizar essa correlação ao nosso favor para prever o dado faltante.

- Existe um forte correlação positiva entre a renda do cliente e a quantidade de dinheiro no empréstimo

- Existe uma forte correlação negativa linear e não linear entre a quantidade de dinheiro no empréstimo e o quanto a empréstimo representa da renda do cliente, indicando que quanto maior a renda do cliente, menos o empréstimo representa da renda dele.

Dados faltantes:

- Utilizando a correlação de spearman e Kendal foi possível verificar entre as variáveis categóricas qual possuia maior correlação com o loan_int_rate que era a taxa do empréstimo e foi verificado que o grau do empréstimo possue maior correlação com a taxa do empréstimo.



Foi calculada a média do loan_int_rate por grau de empréstimo afim de substituir os dados faltantes dessa variável, causando mínimo impacto na distribuição dessa variável numérica entre o grupo de inadimplentes e não inadimplentes, já que há diferença na distribuição dessa variável numérica entre esses dois grupos.

Dados faltantes (`person_emp_length`):

Como havia uma correlação linear forte entre a idade e o tempo de empregamento foi ajustado um modelo de regressão linear simples baseado na idade para prever o tempo de empregamento (`person_emp_length`).

- Após substituir os dados faltantes Podemos observar a distribuição dos dados com a substituição e perceber que não houve mudanças drásticas na distruição nos dois grupos quando havíamos desconsiderado os dados faltantes, evitando que a gente perca informação.
- Um detalhe importante é que precisamos substituir esses dados faltantes depois de fazer a divisão de treino, validação e teste para que não soframos de Data Leakage.

Grupo inadimplentes

	person_age	person_income	person_emp_length	loan_amnt	loan_int_rate	loan_percent_income	cb_person_cred_hist_length
count	19548.000000	1.954800e+04	19548.000000	19548.000000	19548.000000	19548.000000	19548.000000
mean	27.713833	6.617245e+04	4.779211	9595.420248	11.025592	0.169930	5.795478
std	6.340725	6.558040e+04	3.980277	6320.414975	3.212845	0.106111	4.044860
min	20.000000	4.000000e+03	0.000000	500.000000	5.420000	0.000000	2.000000
25%	23.000000	3.850000e+04	2.000000	5000.000000	7.900000	0.090000	3.000000
50%	26.000000	5.500000e+04	4.000000	8000.000000	11.009722	0.150000	4.000000
75%	30.000000	8.000000e+04	7.000000	12025.000000	13.458466	0.230000	8.000000
max	144.000000	6.000000e+06	38.000000	35000.000000	22.480000	0.770000	30.000000

Grupo não inadimplentes

	person_age	person_income	person_emp_length	loan_amnt	loan_int_rate	loan_percent_income	cb_person_cred_hist_length
count	19548.000000	1.954800e+04	18996.000000	19548.000000	17665.000000	19548.000000	19548.000000
mean	27.713833	6.617245e+04	4.780217	9595.420248	11.025886	0.169930	5.795478
std	6.340725	6.558040e+04	4.036093	6320.414975	3.231083	0.106111	4.044860
min	20.000000	4.000000e+03	0.000000	500.000000	5.420000	0.000000	2.000000
25%	23.000000	3.850000e+04	2.000000	5000.000000	7.900000	0.090000	3.000000
50%	26.000000	5.500000e+04	4.000000	8000.000000	10.990000	0.150000	4.000000
75%	30.000000	8.000000e+04	7.000000	12025.000000	13.480000	0.230000	8.000000
max	144.000000	6.000000e+06	38.000000	35000.000000	22.480000	0.770000	30.000000

Etapa de preprocessamento, modelagem e validação

- Irei explicar essas três etapas juntas pois acho dificil separá-las.
- Para a validação foi usado o k-fold estratificado, onde os dados foram divididos em três folds com a mesma proporção da variável resposta em cada fold, para ter estimações sobre a performance dos modelos mais confiáveis e evitar overfitting, foram criados 3 folds.
- Cada fold foi dividido em três subconjuntos estratificados pela variável resposta, treino (60% do fold), validação(20% do fold), teste(20% do fold). O conjunto de validação é usado para ajustar hiperparâmetros dos modelos.

Preprocessamento nas variáveis categóricas

- Nas variáveis categóricas foi aplicado o ordinal encoding nas variáveis com maior correlação V de Cramer, a `'loan_grade'`, `'person_home_ownership'`,`'cb_person_default_on_file'` pois de acordo com o teste e o coeficiente as categorias dessas variáveis possuem relação com a variável resposta.
- No restante foi aplicado o One Hot Encoding, no caso de ser dicotomica foi mantida uma coluna com 0,1 .

Seleção de variáveis

- Foi utilizado o critério mutual information para verificar a importância de cada covariável para prever a variável resposta.
- Após selecionar as variáveis com menor importância por este critério, foi ajustado os modelos e observado o impacto da retirada de cada variável na performance dos modelos, e pelo princípio da parsimonia escolhido o menor conjunto de variáveis que gerava a melhor performance para os modelos.

Para modelagem foi utilizado os modelos:

- Regressão binária clássica com função de ligação assimétrica
- Regressão binária bayesiana com função de ligação assimétrica.
- Random Forest
- Gradient boosted trees (XGBoost).

A regressão binária com função de ligação assimétrica é mais apropriada para dados desbalanceados, pois ela representa a função de distribuição acumulada (f.d.a.) da probabilidade da classe resposta, como os dados são desbalanceados faz mais sentido ela ser assimétrica, diferente da regressão logística que possue função de ligação simétrica.

Para mais detalhes tenho meu artigo onde explico essa diferença em mais detalhes:

Link:

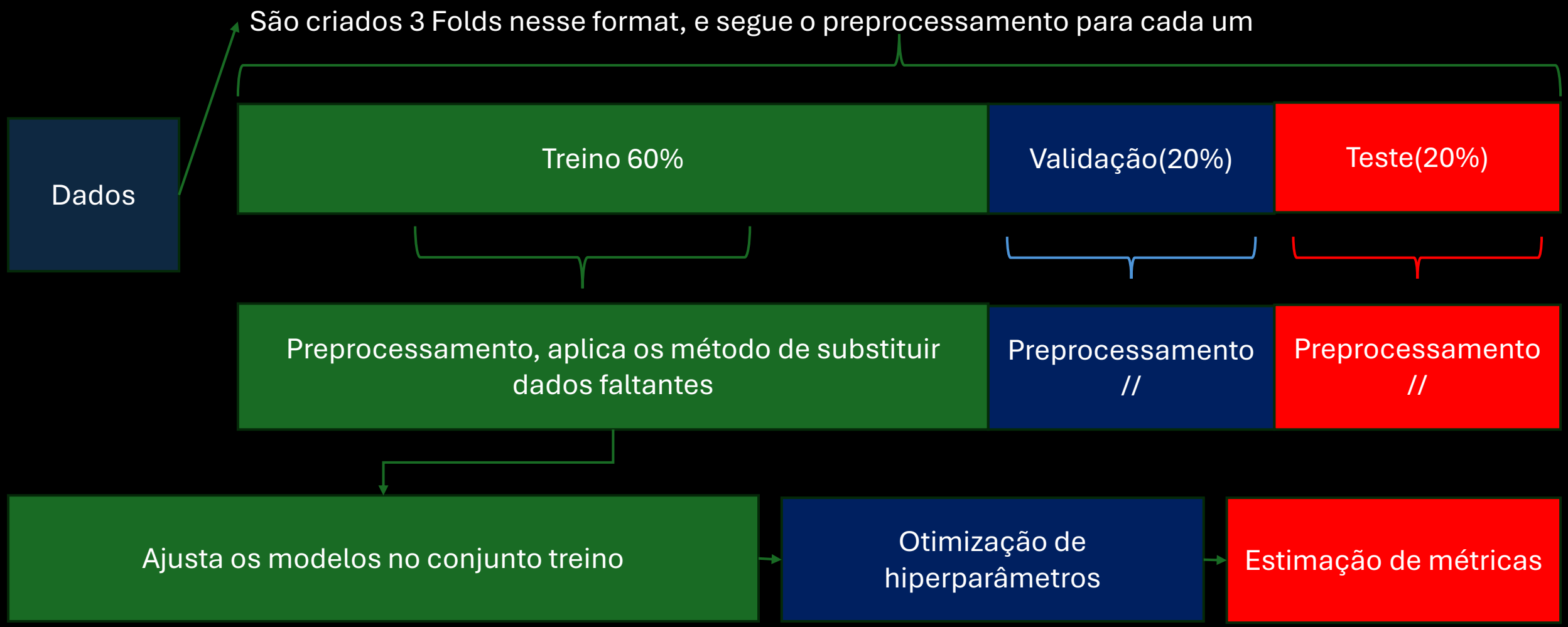
Regressão clássica e bayesiana

- As funções de ligação clássica loglog e cloglog estão disponíveis no python, elas possuem um formato fixo que não se modifica de acordo com a assimétria dos dados. Já a ligação bayesiana cauchito que irei utilizar possui um parâmetro (lambda) responsável em controlar o nível de assimetria na função de ligação.
- Além de potencialmente ter um boa capacidade preditiva as técncias bayesianas possuem uma interpretação considerada mais fácil.

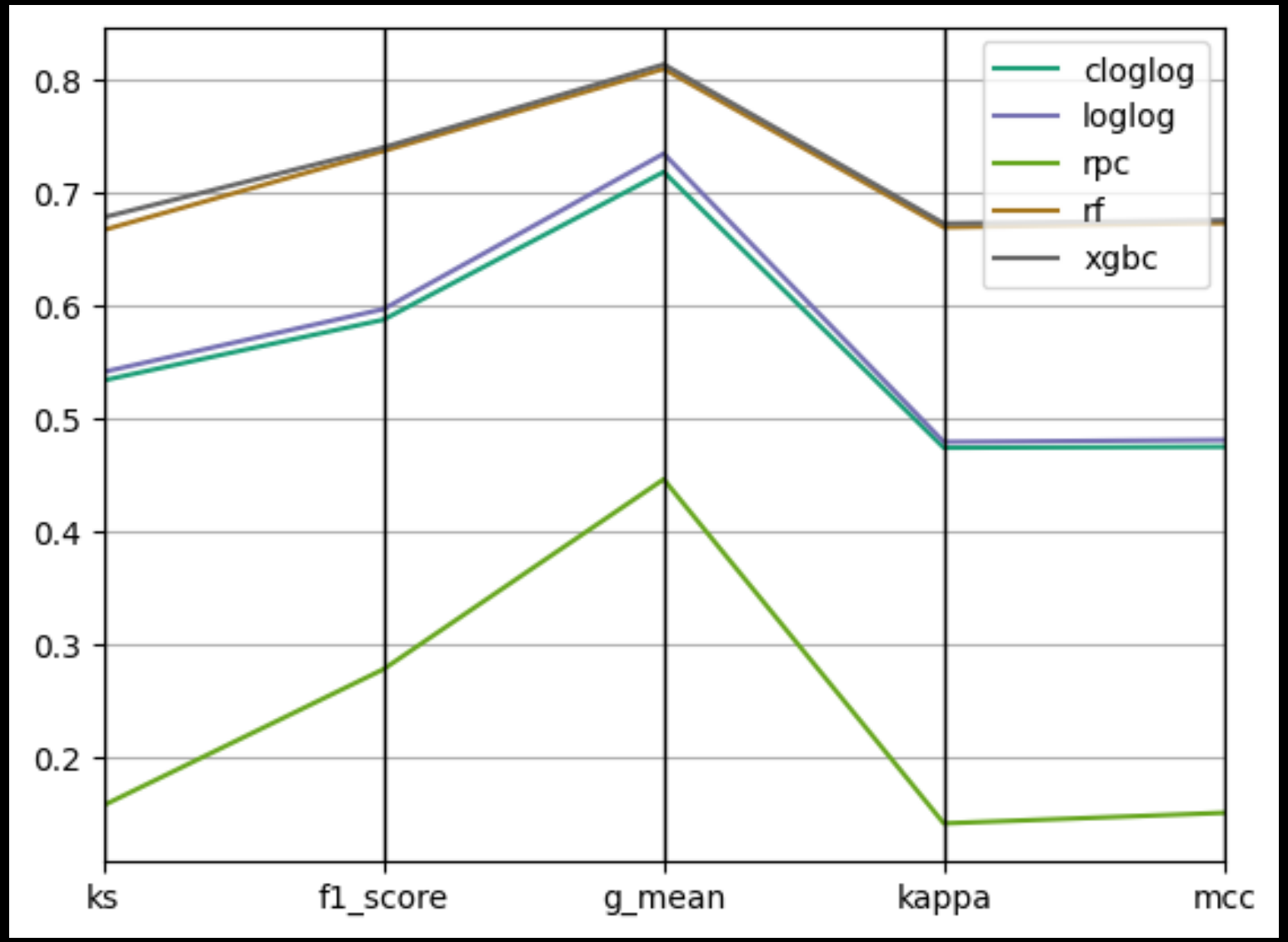
Na etapa de validação:

- A regra de corte de todos os modelos são maximizadas de acordo com a métrica Kappa que considera a distribução original dos dados para calcular o nível de concordância entre os valores preditos e verdadeiro dos dados, além disso ele considera as classes positivas e negativas.
- Nos modelos ensemble, random forest e xgboost, o método GridSearchCV foi utilizado estimar os hiperparâmetros, no caso do random forest foi utilizado o mesmo número de árvores em todos os testes pois esse hiperparâmetro não faz diferença porque o random forest não sofre de overtraining[1].
- [1] - Link : https://www.ibm.com/topics/random-forest#:~:text=However%2C%20when%20there's%20a%20robust,overall%20variance%20and%20prediction%20error.

Resumo pipeline dos dados até os modelos



Performance dos modelos



Cloglog, loglog = regressão com função de ligação cloglog e loglog.
RPC = é a regressão bayesiana com função de ligação cauchito.
rf = random forest
Xgbc = grandient boosted trees classifier.
Podemos observer que o gradient boosted trees e o random forest possuem melhores performances.

Sobre a regressão bayesiana

- Na regressão bayesiana podemos ter uma melhor performance com alguns ajustes nos hiperparâmetros do modelo, mas infelizmente esse modelo demora muito tempo para ser treinado e testar os seus hiperparâmetros, e como os modelos ensemble tiveram uma boa performance ele foi deixado de lado nesta análise.

Utilizando a matriz de confusão foi possível estimar:

- TPR = taxa de verdadeiros positivos.
- TNR = taxa de verdadeiros negativos.
- FPR = taxa de falsos positivos.
- FNR = taxa de falsos negativos.

No momento da classificação vamos considerar que: Falso positivo = gasto igual a média de lucro por cliente igual a R$989, pois perdemos a oportunidade.

Falso negativo = gasto de um inadimplente, pois é um inadimplente R$1436

Verdadeiro negativo = lucro de R$989 que é a media de lucro de um não inadimplente.

Verdadeiro positivo = não causa gasto nenhum nem lucro nenhum.

O modelo com melhor performance foi o random forest.

- Com faturamento aproximadamente igual a R$28.695.206
- Gasto total de R$2.606.842
- Lucro = faturamento total – gasto total = R$**26.088.364**

Lembrando que o gasto total é igual a quantidade media de uma empréstimo vezes sua taxa de juros, mais a quantidade media de um empréstimo de um não inadimplente(pois já havia esse dinheiro em caixa).

Sem nenhum modelo o lucro final era negativo, igual a –R$37.327.036, ou seja, tivemos um aumento no lucro de 169% utilizando o modelo.

Outras métricas importantes para o negócio que foi calculada:

- Lucro médio gerado em um empréstimo dado a um não inadimplente: R$989
- Prejuízo médio causado por um inadimplente é igual a quantidade média de um empréstimo + o suposto lucro médio em cima do empréstimo é igual a R$1436.
- Custo médio de classificar um bom pagador como mal pagador é igual a R$989 pois perdemos a oportunidade.
- Quantidade média de um empréstimo de um não inadimplente: R$9242
- Quantidade média de um empréstimo de um inadimplente: R$10861

Outras métricas importantes de negócio:

- Porcentagem de inadimplentes: 21.81%
- Porcentagem de não inadimplentes: 78.18%

Sugestões:

- Poderíamos fazer mais experimentos como por exemplo:
- - Modificar a transformação que é feita nas variável categóricas, utilizando outras técnicas como por exemplo o WOE(Weight of Evidence), entre outras já que existem várias, e verificar o impacto que isso tem no modelo.
- - Adicionar mais variáveis ao modelo.
- - Ajustar outros modelos de classificação.
- ...