Detecção de aderência em campanha de call center

Pré-processamento

- Tratamento de variáveis faltantes
  - Número total de registros: 41.188
  - Número de registros com alguma variável faltante: 10.700 (25,9%)

Pré-processamento

- Tratamento de variáveis faltantes
  - Número total de registros: 41.188
  - Número de registros com alguma variável faltante: 10.700 (25,9%)

- Formas de lidar com o problema
  - Remover registros ou algum atributo
  - Substituir por uma constante ou valor nulo
  - Tentar inferir: média, mediana, mais frequente

Pré-processamento

- Tratamento de variáveis faltantes
  - Número total de registros: 41.188
  - Número de registros com alguma variável faltante: 10.700 (25,9%)

- Formas de lidar com o problema
  - Remover registros ou algum atributo
  - Substituir por uma constante ou valor nulo
  - Tentar inferir: média, mediana, mais frequente

Pré-processamento - Tratamento de variáveis faltantes

- Analisando se algum atributo é responsável por parte significativa das variáveis faltantes.

Atributo	**Registros com variáveis Faltantes**
profissao	330
emprestimo_pessoal	990
emprestimo_moradia	990
inadimplente	8.597
estado_civil	80
educacao	1731

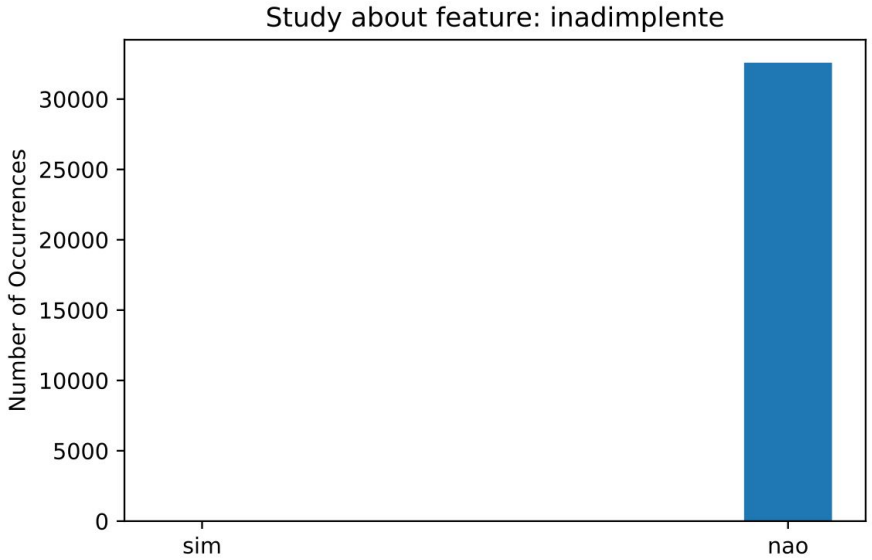
Pré-processamento - Tratamento de variáveis faltantes

- Analisando se algum atributo é responsável por parte significativa das variáveis faltantes.

Atributo	**Registros com variáveis Faltantes**
profissao	330
emprestimo_pessoal	990
emprestimo_moradia	990
inadimplente	**8.597 (20,9%)**
estado_civil	80
educacao	1731

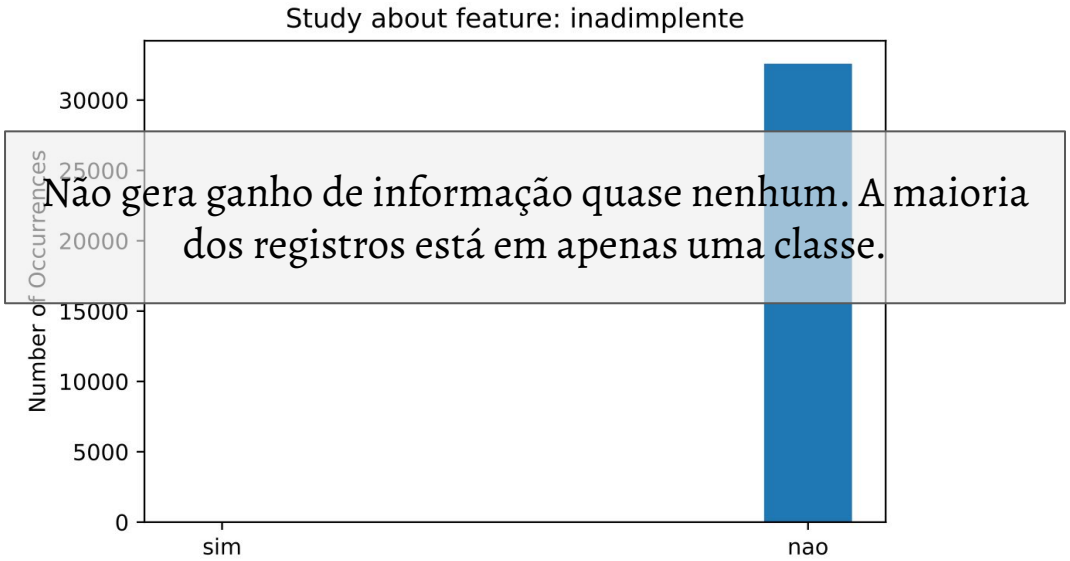
Pré-processamento - Tratamento de variáveis faltantes

- É necessário verificar o impacto da remoção do atributo “inadimplente”.



Pré-processamento - Tratamento de variáveis faltantes

- É necessário verificar o impacto da remoção do atributo “inadimplente”.



Não gera ganho de informação quase nenhum. A maioria dos registros está em apenas uma classe.

Pré-processamento - Tratamento de variáveis faltantes

- É feita a remoção do atributo “inadimplente” e verifica-se os registros com variáveis faltantes novamente:
  - Número total de registros: 41.188
  - Número de registros com alguma variável faltante: 2.943 (7,1%)
- Novamente pode-se lidar com o problema de três formas
  - Remover registros ou algum atributo
  - Substituir por uma constante ou valor nulo
  - Tentar inferir: média, mediana, mais frequente

Pré-processamento - Atributos categóricos

- A base de dados têm vários atributos categóricos. Por isso é preciso transformá-los em atributos numéricos.

- Implementamos três técnicas para lidar com esses atributos:
  - Label Encoding
  - One Hot Encoding
  - Find and Replace

Pré-processamento - Atributos categóricos

- Label Encoding

```
def labelEncoding(categories):
    id_count = 0
    category_mapping = dict()
    for category in categories:
        category_mapping[category] = id_count
        id_count += 1

    return category_mapping
```

Pré-processamento - Atributos categóricos

- One Hot Encoding

```
def oneHotEncoding(categories_list, register_category):

    category_encoding = []
    for category in categories_list:
        if category == register_category:
            category_encoding.append(1)
        else:
            category_encoding.append(0)

    return category_encoding
```

Pré-processamento - Atributos categóricos

- Find and Replace

```
def findAndReplace(replace_mapping, register_category):

    register_mapping = replace_mapping[register_category]

    return register_mapping
```

Esse método trata casos específicos, p.ex., para um atributo com dois valores, “sim” e “não”, podemos utilizar para substituir por 1 e 0, respectivamente.

Pré-processamento - Atributos categóricos

- Os 18 atributos foram divididos em quatro categorias:
  - Numéricos: os atributos que já são numéricos.
  - Categórico binário: são aqueles atributos que têm apenas dois valores e, portanto, sua transformação pode ser simplificada.
  - Categórico não-binário: todos os atributos categóricos com três ou mais valores.
  - Atributo alvo: “aderencia_campanha”

Pré-processamento - Atributos categóricos

- Transformação de atributos categóricos
  - Categórico binário: utilizamos o “Find and Replace” para mapear uma das classes para 0 e a outra para 1.
  - Categórico não-binário: utilizamos o “One Hot Encoding” para transformar os demais atributos categóricos em numéricos.
  - Atributo alvo: não é necessário fazer a transformação.

Análise da Base de Dados - Desbalanceamento

- Depois de realizar o pré-processamento, segue as informações básicas da base de dados:
  - Número total de registros: 38.245
  - Registros com a classe “SIM”: 4.258 (11,1%)
  - Registros com a classe “NÃO”: 33.987 (88,9%)

- Pontos importantes:
  - Se o classificador sempre escolher “NÃO”, acurácia já é 88%
  - Importante utilizar outras métricas: precision, recall, F1

Avaliação dos modelos de predição - Parte 1

- Escolha dos modelos de predição:
  - Naive Bayes (base)
  - Random Forest: tende a ter bom desempenho em base de dados desbalanceadas
  - SVM: tende a funcionar bem quando os dados são esparsos (sendo o caso depois da transformação dos atributos categóricos utilizando o “One Hot Encoding”)

Avaliação dos modelos de predição - Parte 1

- Parâmetros avaliados em cada um deles:
  - Naive Bayes: iremos utilizar os parâmetros padrões
  - Random Forest: vamos aumentar o número de estimadores de 100 para 1000 e iremos variar a profundidade das árvores
  - SVM: vamos variar o kernel e o parâmetro C

Avaliação dos modelos de predição - Parte 1

- Resultado base do Naive Bayes

Accuracy	Precision	Recall	F1
0,77	0,65	0,61	0,56

		Prediction	
		Não	Sim
Actual	Não	27.642	6.345
	Sim	2.558	1.700

Avaliação dos modelos de predição - Parte 1

- Resultado do SVM -- variando o kernel

Kernel	Accuracy	Precision	Recall	F1
linear	0,79	0,68	0,56	0,56
rbf	0,90	0,81	0,61	0,64
sigmoid	0,88	0,68	0,59	0,58

Avaliação dos modelos de predição - Parte 1

- Resultado do SVM -- variando o kernel

Kernel	Accuracy	Precision	Recall	F1
linear	0,79	0,68	0,56	0,56
rbf	0,90	0,81	0,61	0,64
sigmoid	0,88	0,68	0,59	0,58

Então agora vamos variar o parâmetro C fixando o kernel.

Avaliação dos modelos de predição - Parte 1

- Resultado do SVM -- kernel “rbf” e variando o parâmetro C

C	Accuracy	Precision	Recall	F1
0.01	0.90	0.76	0.60	0.61
0.1	0.90	0.81	0.62	0.64
1	0.90	0.81	0.61	0.64
10	0.90	0.80	0.62	0.65
100	0.87	0.78	0.62	0.63

Avaliação dos modelos de predição - Parte 1

- Resultado do SVM -- kernel “sigmoid” e variando o parâmetro C

C	Accuracy	Precision	Recall	F1
0.01	0.90	0.86	0.58	0.59
0.1	0.90	0.80	0.58	0.59
1	0.88	0.68	0.59	0.58
10	0.84	0.59	0.59	0.59
100	0.84	0.58	0.59	0.59

Avaliação dos modelos de predição - Parte 1

- Dentre todas as execuções, segue o melhor resultado com a respectiva matriz de confusão do SVM (kernel= “rbf” e C = 10)

C	Accuracy	Precision	Recall	F1
10	0.90	0.80	0.62	0.65

		Prediction	
		Não	Sim
Actual	Não	33.217	770
	Sim	3.113	1.145

Avaliação dos modelos de predição - Parte 1

- Resultado do Random Forest variando a profundidade máxima das árvores.

Max Depth	Accuracy	Precision	Recall	F1
1	0.89	0.45	0.50	0.47
2	0.85	0.47	0.53	0.49
4	0.68	0.44	0.46	0.39
6	0.57	0.45	0.41	0.34
8	0.49	0.43	0.37	0.30
10	0.49	0.43	0.36	0.30

Avaliação dos modelos de predição - Parte 1

- Dentre todas as execuções, segue o melhor resultado com a respectiva matriz de confusão no Random Forest (Max Depth = 2)

Accuracy	Precision	Recall	F1
0.85	0.47	0.53	0.49

		Prediction	
		Não	Sim
Actual	Não	32.362	1.625
	Sim	4.036	222

Normalização dos atributos numéricos

- Normalização dos atributos:
  - Realizamos a normalização de forma que os valores de todos os atributos numéricos fiquem no intervalo [0,1].
- Caso anômalo:
  - Os valores do atributo “dias_ultimo_contato” ou estava entre 1 e 27 ou era 999. O valor 999, muito superior dos demais, acabava tornando a diferença entre registros com valores menores insignificante, perdendo o valor da informação. Assim, substituímos o valor 999 por 54 (dobro do segundo maior valor observado nesse atributo).

Avaliação dos modelos de predição - Parte 2

- Vamos reavaliar os modelos de predição para a base normalizada.
  - Utilizaremos os mesmos três classificadores e os mesmos padrões de variação de parâmetros.

Avaliação dos modelos de predição - Parte 2

- Resultado base do Naive Bayes

Accuracy	Precision	Recall	F1
0,77	0,65	0,61	0,56

		Prediction	
		Não	Sim
Actual	Não	27.781	6.206
	Sim	2.564	1.694

Avaliação dos modelos de predição - Parte 2

- Resultado do SVM variando o kernel

Kernel	Accuracy	Precision	Recall	F1
linear	0.79	0.73	0.56	0.54
rbf	0.74	0.59	0.48	0.45
sigmoid	0.77	0.57	0.52	0.5

Piora significativa dos resultados depois da normalização. Um atributo importante para o modelo possivelmente estava tendo um peso maior e melhorando o resultado.

Avaliação dos modelos de predição - Parte 2

- Resultado do SVM mantendo o kernel “rbf” e variando C

C	Accuracy	Precision	Recall	F1
0.01	0.89	0.47	0.52	0.49
0.1	0.8	0.74	0.55	0.52
1	0.74	0.59	0.48	0.45
10	0.72	0.56	0.48	0.45
100	0.71	0.55	0.48	0.46

Avaliação dos modelos de predição - Parte 2

- Resultado do SVM mantendo o kernel “rbf” e variando C

C	Accuracy	Precision	Recall	F1
0.01	0.89	0.47	0.52	0.49
0.1	0.8	0.74	0.55	0.52
1	0.74	0.59	0.48	0.45
10	0.72	0.56	0.48	0.45
100	0.71	0.55	0.48	0.46

Mesmo variando o parâmetro C e obtendo uma acurácia maior, as outras métricas continuam com valores muito baixos.

Avaliação dos modelos de predição - Parte 2

- Dentre todas as execuções, segue o melhor resultado com a respectiva matriz de confusão do SVM (kernel= “rbf” e C = 10)

C	Accuracy	Precision	Recall	F1
0.01	0.8	0.74	0.55	0.52

		Prediction	
		Não	Sim
Actual	Não	29.580	4.407
	Sim	3.318	940

Avaliação dos modelos de predição - Parte 2

- O resultado do Random Forest independe da normalização e por isso não vamos mostrar nessa etapa.

Seleção de features

- A piora dos resultados com a normalização é um outro indício da necessidade da seleção de features.
  - Possibilidade dos atributos antes destacados voltarem a contribuir para melhores resultados.
  - Utilizamos o método SelectKBest para fazer a seleção do conjunto de features.

Avaliação dos modelos de predição - Parte 3

- Resultado do Naive Bayes variando o número de **K** (atrb. selec.)

K	Accuracy	Precision	Recall	F1
2	0.9	0.78	0.61	0.62
4	0.89	0.78	0.63	0.63
6	0.89	0.78	0.65	0.64
8	0.89	0.80	0.70	0.70
10	0.88	0.78	0.69	0.69

Avaliação dos modelos de predição - Parte 3

- Melhor resultado do Naive Bayes - Top 8 atributos utilizados

Accuracy	Precision	Recall	F1
0.89	0.80	0.70	0.70

		Prediction	
		Não	Sim
Actual	Não	31.970	2.017
	Sim	2.334	1.924

Avaliação dos modelos de predição - Parte 3

- Resultado do Random Forest mantendo **K** (atrb. selec.) igual a 10

Max Depth	Accuracy	Precision	Recall	F1
1	0.89	0.50	0.54	0.52
2	0.74	0.52	0.50	0.41
4	0.72	0.59	0.51	0.42
6	0.58	0.45	0.44	0.34
8	0.58	0.46	0.44	0.34
10	0.55	0.55	0.42	0.33

Avaliação dos modelos de predição - Parte 3

- Resultado do Random Forest mantendo **K** (atrb. selec.) igual a 10

Max Depth	Accuracy	Precision	Recall	F1
1	0.89	0.50	0.54	0.52
2	0.84	0.58	0.46	0.41
4	0.72	0.59	0.51	0.42
6	0.58	0.45	0.44	0.34
8	0.58	0.46	0.44	0.34
10	0.55	0.55	0.42	0.33

Como os resultados ficaram ruins, vamos variar também o número **K** de features selecionadas.

Avaliação dos modelos de predição - Parte 3

- Resultado do Random Forest variando **K** (atrb. selec.)
  - max_depth = 2

K	Accuracy	Precision	Recall	F1
2	0.89	0.76	0.59	0.59
4	0.73	0.47	0.50	0.40
6	0.74	0.47	0.50	0.40
8	0.83	0.48	0.54	0.48
10	0.74	0.54	0.51	0.41

Avaliação dos modelos de predição - Parte 3

- Melhor resultado do Random Forest - Top 2 atributos utilizados

Accuracy	Precision	Recall	F1
0.89	0.76	0.59	0.59

		Prediction	
		Não	Sim
Actual	Não	33.314	673
	Sim	3.396	862

Avaliação dos modelos de predição - Parte 3

- Resultado do SVM mantendo **K** (atrb. selec.) igual a 10 e variando C
  - Kernel sendo utilizado: “rbf”

C	Accuracy	Precision	Recall	F1
0.01	0.88	0.73	0.59	0.59
0.1	0.87	0.72	0.58	0.57
1	0.85	0.62	0.57	0.54
10	0.85	0.61	0.56	0.53
100	0.85	0.6	0.56	0.52

Avaliação dos modelos de predição - Parte 3

- Resultado do SVM mantendo **K** (atrb. selec.) igual a 10 e variando C
  - Kernel sendo utilizado: “sigmoid”

C	Accuracy	Precision	Recall	F1
0.01	0.90	0.79	0.59	0.60
0.1	0.90	0.67	0.62	0.60
1	0.87	0.62	0.63	0.60
10	0.81	0.67	0.61	0.57
100	0.80	0.67	0.61	0.58

Avaliação dos modelos de predição - Parte 3

- Resultado do SVM mantendo **K** (atrb. selec.) igual a 10 e variando C
  - Kernel sendo utilizado: “linear”

C	Accuracy	Precision	Recall	F1
0.01	0.89	0.79	0.59	0.59
0.1	0.88	0.78	0.60	0.59
1	0.88	0.78	0.60	0.59
10	0.88	0.78	0.60	0.59
100	0.88	0.78	0.60	0.59

Avaliação dos modelos de predição - Parte 3

- Melhor resultado dentre as execuções do SVM:
  - kernel: 'sigmoid', C = 0.01
  - Vamos considerar esse cenário e variar o valor de **K**

K	Accuracy	Precision	Recall	F1
2	0.90	0.79	0.59	0.60
4	0.90	0.79	0.59	0.60
6	0.90	0.79	0.59	0.60
8	0.88	0.69	0.58	0.58
10	0.89	0.71	0.52	0.50

Avaliação dos modelos de predição - Parte 3

- Melhor resultado dentre as execuções do SVM:
  - kernel: ‘sigmoid’, C = 0.01
  - Vamos considerar esse cenário e variar o valor de **K**

K	Accuracy	Precision	Recall	F1
2	0.90	0.79	0.59	0.60
4	0.90	0.79	0.59	0.60
6	0.90	0.79	0.59	0.60
8	0.88	0.69	0.58	0.58
10	0.89	0.71	0.52	0.50

Variar o valor de K não melhorou os resultados. A partir de determinado valor, os resultados começam a deteriorar.

Avaliação dos modelos de predição - Parte 3

- Melhor resultado dentre as execuções do SVM:
  - kernel: ‘sigmoid’, C = 0.01 e K=6

Accuracy	Precision	Recall	F1
0.90	0.79	0.59	0.60

		Prediction	
		Não	Sim
Actual	Não	33.474	467
	Sim	3.434	824

Redução de Dimensionalidade

- Vamos utilizar o PCA
  - Redução do tempo para modelos de predição (SVM principalmente)
  - O aumento de dimensões devido a utilização do “One Hot Encoding” possivelmente está prejudicando os resultados

Avaliação dos modelos de predição - Parte 4

- Melhor resultado obtido para cada modelo depois de variar o número de componentes do PCA

Modelo	Accuracy	Precision	Recall	F1
Naive Bayes	0.77	0.65	0.61	0.56
Random Forest	0.85	0.47	0.53	0.49
SVM	0.81	0.62	0.54	0.52

Avaliação dos modelos de predição - Parte 4

- Melhor resultado obtido para cada modelo depois de variar o número de componentes do PCA

Modelo	Accuracy	Precision	Recall	F1
Naive Bayes	0.77	0.65	0.61	0.56
Random Forest	0.83	0.77	0.68	0.69
SVM	0.81	0.62	0.54	0.52

Realizamos um número maior de experimentos variando o número de componente do PCA, mas como os resultados não foram satisfatórios, reportamos apenas estes. Os outros experimentos estão descritos no notebook e na planilha de resultados.

Melhores resultados

- Melhor resultado obtidos entre todas as avaliações

Id	Modelo	Parâmetro	Normalizada	Accuracy	Precision	Recall	F1
1	SVM	kernel=rbf C=10	Não	0.90	0.80	0.62	0.65
2	SVM	kernel=sigmoid C=0.01 Top 6 Features	Sim	0.89	0.79	0.59	0.59
3	Naive Bayes	Top 8 features	Sim	0.89	0.80	0.70	0.70

Melhores resultados

Id 1		Prediction	
		Não	Sim
Actual	Não	33.217	770
	Sim	3.113	1.145

Id 2		Prediction	
		Não	Sim
Actual	Não	33.474	467
	Sim	3.434	824

Id 3		Prediction	
		Não	Sim
Actual	Não	31.970	2.017
	Sim	2.334	1.924

Escolha do modelo a ser utilizado

- Depende das prioridades da empresa:
  - É importante conseguir identificar boa parte dos clientes que vão aderir à campanha mesmo correndo o risco de ter mais falsos positivos?
    - Modelo Id 3 (Naive Bayes)
  - É importante conseguir identificar boa parte dos clientes que vão aderir à campanha, mesmo que a geração de FP gere um custo muito alto para a empresa?
    - Modelo Id 1 (SVM)

Desenvolvimento de melhorias

- Pré-processamento:
  - Tratamento das variáveis faltantes
  - Utilizar outros métodos para *encoding* de atributos categóricos
- Seleção de features
  - Utilizar outros métodos para seleção de features
  - P.ex., remoção de features com pouca variância
- Utilização de outros classificadores e técnicas
  - Boosting (uma forma de dar mais peso para os FN)
  - GridSearch