Relatório Competição 1

Isabella Beatriz da Silva, 201802727

Dezembro 2022

1 Análise exploratória dos dados

Na etapa exploratória dos dados, inicialmente a base de dados de treino foi aberta em uma planilha para facilitar a visualização dos valores disponíveis em cada coluna. Com isso, foi possível definir variáveis que do ponto de vista de negócio seriam interessantes, os tipos dos dados e se as variáveis eram discretas, categóricas ou se tinham uma ordem entre as categorias.

Em um segundo momento, com auxílio da biblioteca Pandas foi feita uma análise do tipo de dados de cada variável e a quantidade de valores não nulos como mostrado na Figura 1. Com isso, as variáveis que em sua maioria possuíam valores nulos foram excluídas do modelo de treinamento. Além disso, observando o tipo de cada informação foi decidido que tipo de pré-processamento aplicar para garantir que não informações erradas não seriam trazidas para o modelo.

Outra análise muito importante foi da quantidade de valores únicos contidos em cada variável. As que possuíam muitos valores distintos ou campos de texto livre exigiriam que o modelo e o pré-processamento dos dados fosse bem mais complexo. Por isso, esses atributos com muitas categorias foram excluídos do modelo. Como exemplificado na Figura 2, pode-se observar que a variável "DS_INDICACAO_CLINICA" por mais que pareça ter informações muito ricas para o problema é de texto livre, tornando seu uso muito difícil.

2 Pré-processamentos realizados

O primeiro pré-processamento realizado foi preenchimento de valores ausentes. Para as variáveis discretas esses valores foram preenchidos com o número zero (0) e para as categóricas com uma string de valor zero ('0').

As variáveis foram separadas entre categóricas e discretas. Para aquelas consideradas discretas foi realizado *Standard Scaler Encoding* o qual remove a média e dimensiona a variação dos valores. Para as categóricas foi realizado o processo de *One Hot Encoding*, o qual cria uma nova coluna para cada categoria em que o valor é um caso pertença à categoria e zero caso não pertença.

Para o enriquecimento dos dados, um dos pré-processamentos realizados foi a criação da variável idade utilizando o conceito de *Feature engineering*, que consiste na criação de informações a partir da combinação de outros atributos.

Essa nova variável foi chamada de 'idade' e obtida através da subtração da data da requisição da autorização pela data de nascimento do paciente.

Para o treinamento do modelo foram escolhidos alguns atributos do dataset, tanto variáveis discretas como categóricas. Cada uma delas tve um porquê para ser escolhida do ponto de vista do negócio.

Variáveis discretas

- IDADE: variável criada a partir de um pré-processamento que subtrai a 'DT_REQUISICAO' pela 'DT_NASCIMENTO' A idade é importante porque geralmente pessoas mais velhas tem uma necessidade maior que seus exames sejam autorizados.

Variáveis categóricas

- DS_TIPO_PREST_SOLICITANTE: caracteriza onde foi feita a solicitação do exame (hospital, clínica, laboratório...)
- DS_CBO: tipo de médico que atendeu o paciente;
- DS_INDICACAO_ACIDENTE: tipo do acidente que levou ao atendimento, inclusive mostra quando não houve acidente;
- DS_CARATER_ATENDIMENTO: se foi uma consulta de rotina ou emergência, geralmente os exames são autorizados em casos de emergência;
- DS_TIPO_INTERNACAO: o tipo da internação do paciente que levou a solicitação;
- DS_TIPO_ACOMODACAO: onde o paciente estava (enfermaria, UTI...);
- DS_TIPO_ATENDIMENTO: qual tipo de procedimento foi solicitado (exame, cirurgia, terapia).

3 Algoritmos utilizados

Para a etapa de classificação dos dados foram selecionados alguns algoritmos para teste e o que obteve o melhor resultado entre eles foi escolhido.

- Árvore de decisão: este algoritmo é utilizado para classificação e regressão. Uma árvore de decisão é uma estrutura de fluxograma em que o nó raiz é um dos atributos da base de dados e os nós-folha que são os resultados dos testes. A ligação entre esses nós se dá por regras do tipo "se-então", em que dependendo do valor do resultado o próximo nó ficará para esquerda ou direita.
- Random forest: as árvores de decisão têm uma alta variância. O algoritmo *Random forest* combina o resultado de várias árvores de decisão geradas em paralelo e para problemas de classificação o resultado com mais votos é escolhido. Por isso, É uma ótima opção para reduzir a variância da árvore de decisão e obter resultados melhores.

- SVM: Uma máquina de vetores de suporte (SVM)
- MLP: Um *Multilayer Perceptron* (MLP) é um tipo de rede neural artificial. Um MLP simples consiste em pelo menos 3 camadas de nós, sendo uma camada de input, uma camada oculta e uma de saída.

4 Resultados

O resultado dos algoritmos de Árvore de decisão e Random forest foram bastante próximos, como podem ser observados na Figura 3 e na Figura 4 apresentando um score um pouco melhor para a árvore sozinha. Com o conjunto de validação esse score foi de aproximadamente 0.743 já para o conjunto de testes esse número caiu para 0.68.

Já o algoritmo MLP apresentou um resultado um pouco inferior com o conjunto de validação, como apresentado na Figura 5 apresentou um score de 0,708 para os dados de validação. No entanto, quando validado com o conjunto de teste completo o score se manteve praticamente o mesmo com um valor de 0.705. Por isso foi o escolhido como submissão final.

Com o intuito de melhorar esses resultados, acredito que seria possível além de testar outros algoritmos reavaliar os atributos escolhidos para o modelo. Além disso pode-se melhorar a qualidade do conjunto de treinamento aplicando mais técnicas de pré-processamento como balancear o dataset.

5 Referências bibliográficas

SCIKIT learn. In: Sklearn.preprocessing.StandardScaler. [S. l.]. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html. Acesso em: 12 dez. 2022.

Get dummies vs one hot encoder qual método escolher. [S. l.], Agosto 2022. Disponível em: https://www.alura.com.br/artigos/get-dummies-vs-onehotencoder-qual-metodo-escolher. Acesso em: 12 dez. 2022.

OMNI parte 2: Feature engineering. [S. l.], abril 2022. Disponível em: https://www.kaggle.com/code/lusaugustodoprado/omni-parte-2?scriptVersionId=9386333. Acesso em: 19 dez. 2022.

GEEKS for Geeks. In: Decision tree. [S. l.], outubro 2022. Disponível em: https://www.geeksforgeeks.org/decision-tree/. Acesso em: 12 dez. 2022.

UNDERSTANDING Random Forests Classifiers in Python Tutorial. [S. l.], 1 maio 2018. Disponível em: https://www.datacamp.com/tutorial/random-forests-classifier-python. Acesso em: 11 dez. 2022.

6 Apêndices

```
RangeIndex: 227122 entries, 0 to 227121
Data columns (total 32 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   Unnamed: 0                  227122 non-null  int64
 1   NR_SEQ_REQUISICAO           227122 non-null  int64
 2   NR_SEQ_ITEM                 227122 non-null  int64
 3   DT_REQUISICAO               227122 non-null  int64
 4   DS_TIPO_GUIA                227122 non-null  object
 5   DT_NASCIMENTO               227112 non-null  float64
 6   NR_PRODUTO                  227122 non-null  int64
 7   DS_TIPO_PREST_SOLICITANTE   227122 non-null  object
 8   DS_CBO                      227122 non-null  object
 9   DS_TIPO_CONSULTA            10511 non-null   object
 10  QT_TEMPO_DOENCA             266 non-null     float64
 11  DS_UNIDADE_TEMPO_DOENCA     266 non-null     object
 12  DS_TIPO_DOENCA              531 non-null     object
 13  DS_INDICACAO_ACIDENTE       209539 non-null  object
 14  DS_TIPO_SAIDA               0 non-null       float64
 15  DS_TIPO_INTERNACAO          59863 non-null   object
 16  DS_REGIME_INTERNACAO        59863 non-null   object
 17  DS_CARATER_ATENDIMENTO      227122 non-null  object
 18  DS_TIPO_ACOMODACAO          59781 non-null   object
 19  QT_DIA_SOLICITADO           58995 non-null   float64
 20  CD_GUIA_REFERENCIA          37463 non-null   float64
 21  DS_TIPO_ATENDIMENTO         168045 non-null  object
 22  CD_CID                      131250 non-null  object
 23  DS_INDICACAO_CLINICA        179944 non-null  object
 24  DS_TIPO_ITEM                227122 non-null  object
 25  CD_ITEM                     227122 non-null  int64
 26  DS_ITEM                     227122 non-null  object
 27  DS_CLASSE                   227122 non-null  object
 28  DS_SUBGRUPO                 227122 non-null  object
 29  DS_GRUPO                    227122 non-null  object
 30  QT_SOLICITADA               227122 non-null  float64
 31  DS_STATUS_ITEM              227122 non-null  object
dtypes: float64(6), int64(6), object(20)
```


Figure 1: Valores únicos disponíveis em cada variável.

Unnamed: 0	227122
NR_SEQ_REQUISICAO	80699
NR_SEQ_ITEM	227122
DT_REQUISICAO	357
DS_TIPO_GUIA	3
DT_NASCIMENTO	16557
NR_PRODUTO	1
DS_TIPO_PREST_SOLICITANTE	12
DS_CBO	59
DS_TIPO_CONSULTA	4
QT_TEMPO_DOENCA	17
DS_UNIDADE_TEMPO_DOENCA	3
DS_TIPO_DOENCA	2
DS_INDICACAO_ACIDENTE	4
DS_TIPO_SAIDA	0
DS_TIPO_INTERNACAO	6
DS_REGIME_INTERNACAO	3
DS_CARATER_ATENDIMENTO	2
DS_TIPO_ACOMODACAO	8
QT_DIA_SOLICITADO	34
CD_GUIA_REFERENCIA	4610
DS_TIPO_ATENDIMENTO	13
CD_CID	1626
DS_INDICACAO_CLINICA	40428
DS_TIPO_ITEM	2
CD_ITEM	6220
DS_ITEM	6146
DS_CLASSE	460
DS_SUBGRUPO	72
DS_GRUPO	9
QT_SOLICITADA	270
DS_STATUS_ITEM	2
dtype: int64	

Figure 2: Valores únicos disponíveis em cada variável.

```
[[27245  3551]
 [ 8083  6546]]
              precision    recall  f1-score   support

   Autorizado       0.77      0.88      0.82     30796
       Negado       0.65      0.45      0.53     14629

    accuracy                           0.74     45425
   macro avg       0.71      0.67      0.68     45425
weighted avg       0.73      0.74      0.73     45425

0.7438855255916346
```

Figure 3: Métricas para o resultado do algoritmo de Árvore de decisão.

```
[[26852  3944]
 [ 7749  6880]]
              precision    recall  f1-score   support

   Autorizado       0.78      0.87      0.82     30796
       Negado       0.64      0.47      0.54     14629

    accuracy                           0.74     45425
   macro avg       0.71      0.67      0.68     45425
weighted avg       0.73      0.74      0.73     45425

0.7425866813428729
```

Figure 4: Métricas para o resultado do algoritmo Random Forest.

```
[[27261  3535]
 [ 9727  4902]]
              precision    recall  f1-score   support

  Autorizado       0.74      0.89      0.80     30796
      Negado       0.58      0.34      0.43     14629

    accuracy                           0.71     45425
   macro avg       0.66      0.61      0.61     45425
weighted avg       0.69      0.71      0.68     45425

0.7080462300495322
```

Figure 5: Métricas para o resultado do algoritmo MLP.