Relatório Competição 1

Thiago Achcar Trevisan

November 2022

1 Análise exploratória dos dados

O conjunto de dados utilizado possui 227122 amostras no dataset de treino e 186144 amostras no dataset de teste, sendo que, para cada amostra há 31 colunas com informações dos tipos float64, int64 e object. Há também no dataset um total de 4034227 informações(células) perdidas/nulas e 68435 requisições únicas.

O dataset é composto por dados preenchidos por um colaborador da prestadora (hospital, clínica, laboratório ou consultório) requisitando cobertura das despesas de produtos e serviços prestados ao cliente (beneficiário do plano).

2 Pré-processamentos realizados

Foi feito um preenchimento dos espaços vazios(nulos) em algumas colunas do conjunto de dados, isso garantiu o aproveitamento de uma maior quantidade de parâmetros para elaboração do modelo.

Também foram usadas as ferramentas StandardScaler e OneHotEncoder, que pertencem à biblioteca Scikit-learn(sklearn.preprocessing), otimizando assim o uso dos dados e facilitando o trabalho do algoritmo na hora de processar os dados.

3 Configuração experimental

A linguagem utilizada foi Python versão 3.8.16.

Bibliotecas utilizadas foram Pandas, Numpy, Scikit-learn e Matplotlib.

4 Algoritmos utilizados

RandomForestClassifier(), biblioteca 'sklearn.ensemble'

Uma floresta aleatória é um estimador que ajusta vários classificadores de árvore de decisão em várias sub amostras do conjunto de dados e usa a média para melhorar a precisão preditiva e controlar o ajuste. É utilizada para gerar previsões razoáveis em uma ampla variedade de dados, exigindo pouca configuração.

5 Resultados

O modelo final teve score de aproximadamente 0.68, visto abaixo na matriz de confusão gerada.

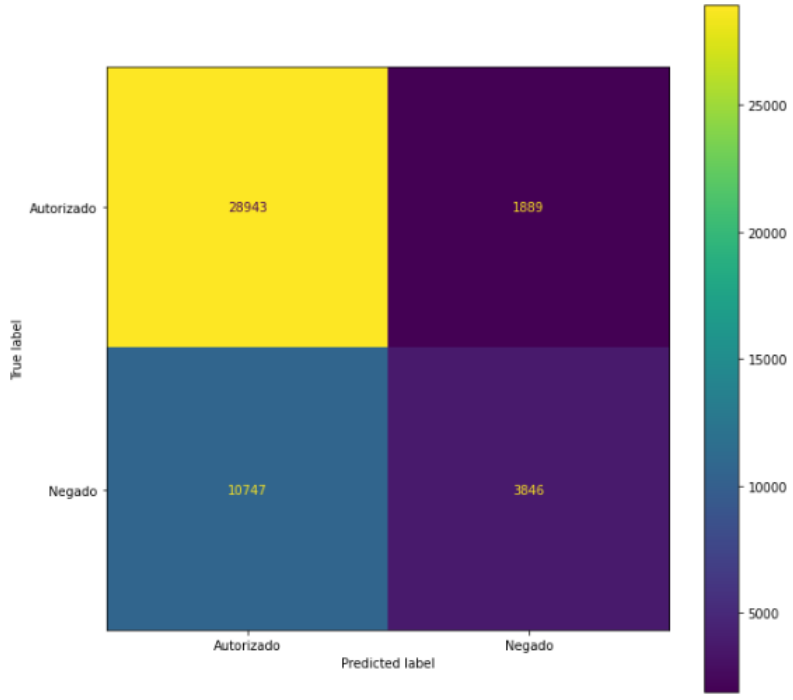
```
[[28943  1889]
 [10747  3846]]
              precision    recall  f1-score   support

  Autorizado       0.73      0.94      0.82     30832
      Negado       0.67      0.26      0.38     14593

    accuracy                           0.72     45425
   macro avg       0.70      0.60      0.60     45425
weighted avg       0.71      0.72      0.68     45425

0.7218271876719868
```

É possível visualizar também o gráfico da matriz de confusão, exibindo acertos e erros de predição do modelo.



6 Referências bibliográficas

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
https://www.kaggle.com/competitions/competicao-um-ic/data