Nanodegree Engenheiro de Machine Learning

Proposta de Projeto Final

Analise e Predição em Consultas Médicas Marcadas

Aron Stall
02 de setembro de 2018

Proposta

Realizar uma analise e uma predição com os dados de milhares de consultas médicas, verificando se o paciente compareceu a consulta médica ou não. Esta analise e predição será realizada através do *dataset* disponibilizado no *Kaggle* chamado **Medical Appointment No Shows**, disponível neste link.

Será realizado uma analise utilizando os dados disponíveis, para verificar quais são os principais tipos de pessoas que mais acabam faltando nas consultas médicas, quais idades costumam mais faltar nas consultas médicas, se são os homens ou as mulheres que mais faltam nas consultas médicas, e mais inúmeras possibilidades de analises que serão realizadas através da analise do *dataset*.

Será realizado uma predição nos dados, predizendo se o paciente ira comparecer na consulta médica ou não, utilizando os conhecimento de Machine Learning realizados durante todas as aulas do curso de *Engenheiro de Machine Learning* da *Udacity*.

Histórico do Assunto

Este projeto esta relacionada a área de saúde, principalmente a área administrativa da saúde, pois é onde as consultas médicas são marcadas, e os dados do *dataset* incluem inúmeras informações relacionadas à saúde, como hipertensão, diabetes e alcoolismo. Um bom motivo para seja realizada uma analise e predição deste *dataset* é a economia de dinheiro que é possível se fazer através de escolher o melhor dia e horário para marcar a consulta para uma pessoa.

A escolha do *dataset* foi essa, pois gostaria que fosse na área da saúde e fosse utilizado uma predição dos dados, e este *dataset* contem inúmeras colunas interessantes para esse proposito e também existe uma grande quantidade de linhas disponível para a analise e a predição.

Descrição do Problema

Inúmeras consultas médicas são marcadas e o paciente acaba faltando a consulta médica. Será realizado uma analise no *dataset* para verificar os padrões nas pessoas

que marcam a consulta médica e vão a ela e as pessoas que marcam a consulta médica e acabam não indo a ela, e ao final será realizado uma predição nos dados para ver se é possível identificar se um paciente irá comparecer a consulta médica ou não.

Conjuntos de Dados e Entradas

O conjunto de dados que será utilizado neste projeto é o **Medical Appointment No Shows** que esta disponível no site do *Kaggle*, neste link. Este *dataset* é perfeito para o cenário descrito, pois ele foi feito para esse proposito, com inúmeras consultas médicas cadastradas, e o resultado, se o paciente compareceu a consulta médica, ou não compareceu a consulta médica.

As colunas do *dataset* são as seguintes:

- **PatientId** – A identificação do paciente.
- **AppointmentID** – A identificação da consulta médica.
- **Gender** – Sexo do paciente, sendo “M” para masculino e “F” para feminino.
- **ScheduledDay** – O dia em que a consulta foi marcada.
- **AppointmentDay** – O dia em que a consulta foi ou iria ser realizada.
- **Age** – Idade do paciente.
- **Neighbourhood** – Bairro onde a consulta foi ou iria ser realizada.
- **Scholarship** – Verdadeiro ou falso, referente se o paciente esta ingresso no programa social Bolsa Família.
- **Hipertension** – Verdadeiro ou falso, referente se o paciente é hipertenso.
- **Diabetes** – Verdadeiro ou falso, referente se o paciente é diabético.
- **Alcoholism** – Verdadeiro ou falso, referente se o paciente bebe bebidas alcoólicas.
- **Handcap** – Verdadeiro ou falso, referente se o paciente tem algum problema físico, mental ou social.
- **SMS_received** – Valor numérico com a quantidade de mensagem que o paciente recebeu para lembrar da consulta médica.

- **No-show** – Verdadeiro ou falso, referente se o paciente faltou na consulta médica ou não.

Descrição da Solução

A solução do projeto é dividido em duas partes, a primeira sendo uma analise dos dados, encontrando padrões nos dados e fazer representações com textos e gráficos para esses padrões, e após isso, será realizado uma predição nos dados, utilizando dados de treino e teste, para predizer com verdadeiro ou falso se o paciente ira comparecer a consulta médica, como demonstrado na coluna *No-show*.

Modelo de Referência (Benchmark)

O *dataset* **Medical Appointment No Shows** disponível no *Kaggle* disponibiliza a coluna *No-show* que contem o resultado, se o paciente compareceu a consulta medica ou não, retornando um valor de verdadeiro e falso.

A partir disso e analisando os *kernels* do *dataset* no *Kaggle*, tentarei chegar a *score* de ate 70%, utilizando alguns métodos de métrica para avaliação dos modelos, como *Train/Test Split*, *Cross Validation*, *ROC/AUC* e/ou outros.

Métricas de Avaliação

O projeto precisará predizer com um *score* de ate 70% quais pacientes podem faltar ou não nas consultas médicas, será utilizado os métodos de avaliação de métricas do *scikit-learn*, como por exemplo *accuracy_score* e *fbeta_score*, para *Train/Test Split*, *cross_val_score* para *Cross Validation*, *roc_auc_score* para *ROC/AUC* e/ou outros.

Como grande parte do *dataset*, os paciente foram na consulta médica (não faltaram), o aprendizado será realizado com dados positivos, melhorando ainda mais o *score*, tendo uma métrica de avaliação melhor.

Design do Projeto

Este projeto utilizara a técnica de aprendizado supervisionado, utilizando inúmeros algorítimos de predição, como por exemplo *GaussianNB*, *DecisionTreeClassifier*, *LogisticRegression* e outros, utilizando a biblioteca de Machine Learning para *Python scikit-learn*. A calibração do modelo será feito através do *scikit-learn*, utilizando algorítimos como *GridSearchCV*, *RandomizedSearchCV* e outros. Será utilizado a biblioteca *Matplotlib* para a geração dos gráficos apresentados no projeto final. Será utilizado o *Pandas* e *NumPy* para manipulação dos dados do *dataset*. E por fim, será utilizado o *Jupyter Notebook* como o ambiente de programação *Python* durante todo o desenvolvimento do projeto *Capstone*. Pode ser utilizado outras bibliotecas *Python* durante o projeto, se for necessário, e estas serão descritas no projeto.

Referências

Kaggle - **Medical Appointment No Shows**. Disponível em: <https://www.kaggle.com/joniarroba/noshowappointments>. Acesso em: 03 de setembro de 2018.

Paulo Vasconcellos — Cientista de Dados brasileiro - **Como saber se seu modelo de Machine Learning está funcionando mesmo**. Disponível em: <https://paulovasconcellos.com.br/como-saber-se-seu-modelo-de-machine-learning-est%C3%A1-funcionando-mesmo-a5892f6468b>. Acesso em: 03 de setembro de 2018.

Towards Data Science - **Train/Test Split and Cross Validation in Python**. Disponível em: <https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6>. Acesso em: 03 de setembro de 2018.