Nanodegree Engenheiro de Machine Learning

Projeto final

Rodrigo Marques Pessoa
07 de Outubro de 2018

I. Definição

As técnicas avançadas de análise de dados e aprendizagem de máquina trazem um ganho exponencial: permitem olhar adiante e auxiliar a medicina preventiva. Com base em dados de pacientes, é possível usar a análise de dados de uma maneira mais inteligente para sugerir cuidados preventivos em determinadas épocas do ano, ou até mesmo promover um estilo de vida mais saudável.

O repertório de informações que os cientistas precisam podem ser encontrados em softwares e equipamentos médicos que emitem dados.

Podemos minerar os registros de dados que os hospitais emitem sobre seus pacientes por meio de um CRM. A partir das informações selecionadas, é possível criar campanhas e ações preventivas.

Com algorítimos matemáticos cada vez mais precisos e baseado nesse histórico de dados, a máquina consegue alertar sobre possíveis doenças, identificar grupos de tendências, entre outros fatores.

Visão geral do projeto

Para tentar prever a necessidade de internação utilizaremos os dados das unidades de saúde com os dados dos pacientes e dados dos diagnósticos realizados pelos médicos. O modelo preditivo utilizará os dados públicos da prefeitura de Curitiba que está disponível para download através do seguinte endereço eletrônico: http://www.curitiba.pr.gov.br/dadosabertos/consulta/?grupo=1.

Descrição do problema

A falta de leito em hospitais é um problema constante encontrado em várias cidades brasileiras. Gerenciar essas vagas é cada vez mais essencial e ter essa informação o quanto antes pode ajudar a estabelecer prioridades de atendimento.

A medida que consultas são realizadas nas diversas unidades de saúde da cidade, essas informações poderiam ser enviadas a um servidor central que mediante a estes dados pode aplicar um modelo preditivo e contabilizar em tempo real a necessidade de utilização de leitos em hospitais, ou tomar ações preventivas, por exemplo.

Métricas

Espera-se que o modelo seja capaz de predizer com um índice alto de precisão as consultas que estão sendo realizada nas diversas unidades de saúde da cidade se necessitam ou não de internação.

Utilizaremos o indicador F-score com beta de 2 pois precisamos de alto recall, ou seja, realmente identificar as consultas que precisam de internação, pois se identificarmos com baixo recall vamos precisar de mais leitos do que estamos realmente prevendo. Assim, uma precisão baixa não é uma situação preocupante pois estamos prevendo leitos a mais do que realmente vamos precisar.

II. Análise

Exploração dos dados

Como o os dados apresentam as informações dos últimos 3 meses de atendimento das unidades de saúde, contendo quase 1.000.000 de registros utilizaremos uma amostragem para realizar o modelo preditivo.

Trabalharei com uma amostra de 664 registros, que foi calculado levando em consideração 5% de erro amostral e 99% de nível de confiança.

Como os dados não foram coletados com o intuito de utilização para o modelo de predição existe uma grande quantidade de dados ausentes e um trabalho de limpeza e padronização será necessário

Outra questão é fazer o balanceamento das classes já que a proporção de necessidade de internações comparada as da não necessidade são muito desproporcionais.

Segue abaixo um dicionário de dados com as variáveis disponíveis e as que foram escolhidas para serem preditoras do modelo. As variáveis em verde foram escolhidas pois podem apresentar alguma correlação com a variável classe(em azul), como por exemplo dados socioeconômicos dos pacientes (pacientes com menor poder aquisitivo ou que não tem infraestrutura como coleta de lixo, água tratada poderia necessitar de internações constantes). Já as variáveis em vermelhos foram descartadas pois não apresentam nenhuma correlação para o estudo deste projeto:

Dicionário de Dados

Nome do Campo	Descrição
Data do Atendimento	Data de Realização do Atendimento
Data de Nascimento	Data de Nascimento do Paciente
Sexo	Sexo do Paciente
Código do Tipo de Unidade	Código do Tipo de Unidade de Atendimento
Tipo de Unidade	Tipo de Unidade de Atendimento
Código da Unidade	Código da Unidade de Atendimento
Descrição da Unidade	Descrição da Unidade de Atendimento
Código do Procedimento	Código do Procedimento Realizado
Descrição do Procedimento	Descrição do Procedimento Realizado
Código do CBO	Código da Ocupação do Profissional
Descrição do CBO	Descrição da Ocupação do Profissional
Código do CID	Código do Diagnóstico
Descrição do CID	Descrição do Diagnóstico
Solicitação de Exames	Indica se ocorreu solicitação de Exames
Qtde Prescrita Farmácia Curitibana	Qtde de medicamentos prescritos na Farmácia
Qtde Dispensada Farmácia Curitibana	Qtde de medicamentos dispensados na Farmácia
Qtde de Medicamento Não Padronizado	Qtde de Medicamento Não Padronizado
Encaminhamento para Atendimento Especialista	Indica se houve encaminhamento para Atendimento de Especialista

Á rea de Atuação	Á rea de Atuação
Desencadeou Internamento	Indica se desencadeou Internamento
Data do Internamento	Data do Internamento do paciente
Estabelecimento Solicitante	Estabelecimento que solicitou o internamento
Estabelecimento Destino	Estabelecimento que houve a internação
CID do Internamento	Código do diagnóstico do internamento
Tratamento no Domicílio	Tipo de Tratamento de Água no domicílio
Abastecimento	Tipo de Abastecimento de Água no domicílio
Energia Elétrica	Indica se há energia elétrica no domicílio
Tipo de Habitação	Tipo de habitação no domicílio
Destino Lixo	Destino do lixo no domicílio
Fezes/Urina	Destino das fezes/urina no domicílio
Cômodos	Qtde de Cômodos no domicílio
Em Caso de Doença	Serviços procurados em caso de doença
Grupo Comunitário	Grupo Comunitário em que o paciente participa
Meio de Comunicação	Meios de Comunicação utilizados no domicílio
Meio de Transporte	Meios de Transporte utilizados no domicílio
Município	Município do paciente
Bairro	Bairro do paciente

Visualização exploratória

1 Dados antes do pré-processamento

Data nascimento	Sexo	Diagnóstico	Solicitação de exame	Encaminhado especialista	Tratamento de Água domicilio	energia	Tipo Habitação	Coleta Lixo	Coleta Fezes/Urina	Cômodos	Internamento
04/10/2012 00:00	F	EXAME MEDICO GERAL	Nao	Sim	SEM TRATAMENTO	Sim	TIJOLO/ALVENARIA COM REVESTIMENTO	COLETADO	SISTEMA DE ESGOTO	5	Nao

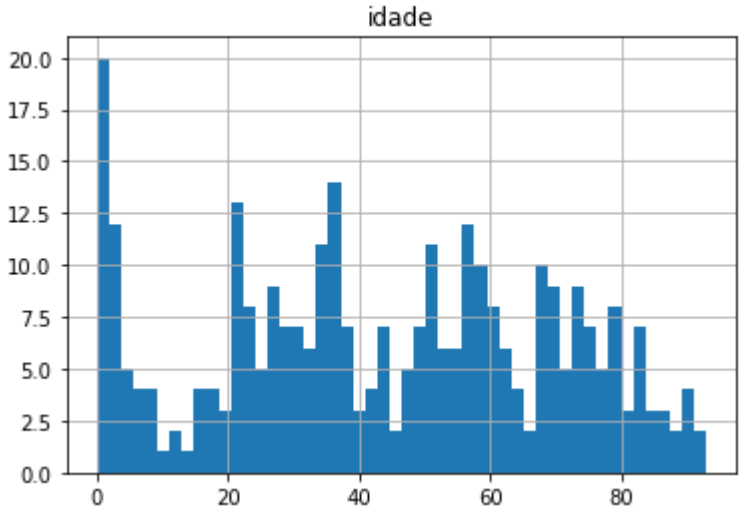
2 Dados de pacientes com necessidade de internamento

Analisando alguns registros que necessitaram de internação podemos ver que existe uma forte correlação entre internação e problemas relacionados ao tratamento de água. Ou seja, a falta de tratamento de água com cloro pode indicar uma necessidade maior de internação conforme indicado abaixo em vermelho.

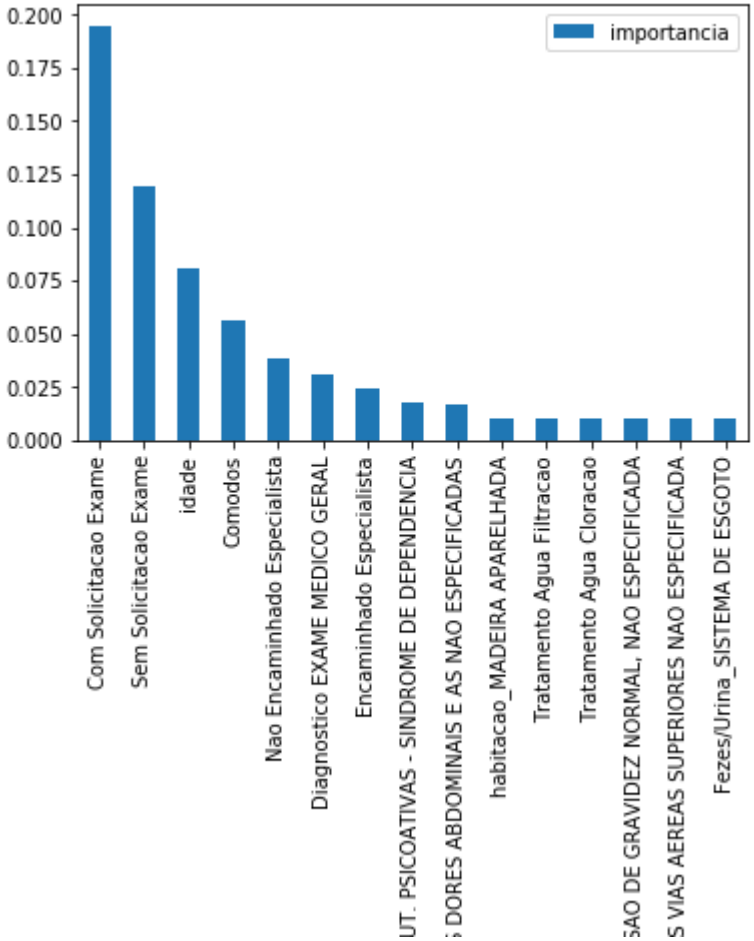
Diagnóstico	Solicitação de Exame	Tratamento De Agua domicilio	Internamento
ABDOME AGUDO	Sim	SEM TRATAMENTO	Sim
ACIDENTE VASCULAR CEREBRAL, NAO ESPECIFICADO C...	Sim	NAO ESPECIFICADO	Sim
OSTEOMIELITE NAO ESPECIFICADA	Sim	NAO ESPECIFICADO	Sim
AMEACA DE ABORTO	Sim	FILTRACAO	Sim
LARINGITE AGUDA	Sim	NAO ESPECIFICADO	Sim
PARESTESIAS CUTANEAS	Sim	NAO ESPECIFICADO	Sim
OUTRAS DORES ABDOMINAIS E AS NAO ESPECIFICADAS	Sim	SEM TRATAMENTO	Sim
EPISODIO DEPRESSIVO GRAVE SEM SINTOMAS PSICOTICOS	Nao	FILTRACAO	Sim
OUTRAS DORES ABDOMINAIS E AS NAO ESPECIFICADAS	Sim	FILTRACAO	Sim
TRANSTORNOS MENTAIS E COMPORT. DEV. USO COCAIN...	Nao	FILTRACAO	Sim

3 Histograma de Idade

Outra informação importante que podemos visualizar através do histograma abaixo é que existe uma necessidade maior de internação nos primeiros anos de vida.



Analise da importância dos atributos



Algoritmos e técnicas

Optamos por utilizar o algoritimo Random Forest ou floresta aleatória. Uma das vantagens da floresta aleatória é que ela pode ser usada para tarefas de regressão e classificação e que é fácil visualizar a importância relativa que atribui aos recursos de entrada.

O Random Forest também é considerado um algoritmo muito prático e fácil de usar, porque os hiperparâmetros padrão geralmente produzem um bom resultado de previsão. O número de hiperparâmetros também não é tão alto e eles são fáceis de entender.

Florestas Aleatórias também são muito difíceis de serem superadas em termos de desempenho. É claro que você pode sempre encontrar um modelo que possa ter um desempenho melhor, como uma rede neural, mas isso geralmente leva muito mais tempo no desenvolvimento. E além disso, eles podem lidar com vários tipos de recursos diferentes, como categórico e numérico que é o nosso caso.

Benchmark

Como não temos um modelo real utilizarei o modelo DummyClassifier do sklearn como modelo de referência para verificarmos e compararmos se o modelo escolhido realmente apresenta uma performance melhor que uma metologia aleatória.

DummyClassifier é um classificador que faz previsões usando regras simples. Esse classificador é útil como uma linha de base simples para comparar com outros classificadores (reais). Não deve ser usado para problemas reais conforme indicado na documentação do método: http://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html.

III. Metodologia

Pré-processamento de dados

Precisamos realizar uma normalização nos atributos numéricos e aplicar a técnica de one-hot encoding nos dados categóricos para que o modelo escolhido tenha uma melhor performance sobre os dados.

Implementação

Optamos por implementar uma função que pode ser chamada por vários modelos e à medida que ela faz o processo de predição ela também retorna as informações de acurácia, matriz de confusão e f-score para que possamos realizar a avaliação e a comparação entre os modelo selecionado e o de benchmark.

Refinamento

Por fim, melhoramos a performance testando alguns hiperparâmetros com o método gridsearch o que nos trouxe uma pequena melhora conforme os dados abaixo:

Melhor estimador:

RandomForestClassifier(bootstrap=False, class_weight=None,
criterion='gini', max_depth=3, max_features=20,
max_leaf_nodes=None, min_impurity_split=1e-07,
min_samples_leaf=1,min_samples_split=2,
min_weight_fraction_leaf=0.0,
n_estimators=15, n_jobs=1, oob_score=False,
random_state=999, verbose=0, warm_start=False)

Modelo não otimizado:

Acurácia dos dados de teste: 0.8421
F-score dos dados de teste: 0.8615

Modelo Otimizado:

Acurácia dos dados de teste: 0.8571
F-score dos dados de teste: 0.8869

IV. Resultados

Modelo de avaliação e validação

Um dos grandes problemas no aprendizado de máquina é o overfitting, mas na maioria das vezes isso não acontece tão fácil para um classificador de floresta aleatório. Isso porque, se houver árvores suficientes na floresta, o classificador não preparará o modelo. A principal limitação da Random Forest é que um grande número de árvores pode tornar o algoritmo lento e ineficaz para previsões em tempo real. Em geral, esses algoritmos são rápidos de treinar, mas muito lentos para criar previsões depois de treinados. Uma previsão mais precisa requer mais árvores, o que resulta em um modelo mais lento. Na maioria das aplicações do mundo real, o algoritmo de floresta aleatória é rápido o suficiente e uma alternativa é a utilização de vários processadores com o hiperparâmetro n_jobs.
A descrição acima nos dá muita segurança para afirmarmos que o nosso modelo é bem consistente e apresenta uma boa performance e atende as nossas necessidades.

Justificativa

Como podemos ver nos dados abaixo o modelo conseguiu uma boa performance se comparada ao modelo de benchmark o que mostra que o modelo escolhido atende as nossas necessidades.

Modelo Benchmark

```
DummyClassifier registros treinados = 531
DummyClassifier acuracia teste= 0.481203007519
DummyClassifier f-score(2) teste= 0.536912751678

Modelo Escolhido

RandomForestClassifier registros treinados = 531
RandomForestClassifier acuracia teste= 0.842105263158
RandomForestClassifier f-score(2) teste= 0.823529411765
```

V. Conclusão

Forma livre de visualização

Com Solicitação de Exame	0.195
Sem solicitação de Exame	0.119
Idade	0.081
Número de Cômodos	0.056
Não Encaminhado ao Especialista	0.038
Diagnóstico EXAME MEDICO GERAL	0.031
Encaminhado ao Especialista	0.024
Diagnóstico TRANST. MENTAIS/COMPORT. DEV. USO MULT. DROGAS/OUT. PSICOATIVAS - SINDROME DE DEPENDENCIA	0.018
Diagnóstico OUTRAS DORES ABDOMINAIS E AS NAO ESPECIFICADAS	0.017
Tipo de habitação - MADEIRA APARELHADA	0.010
Tipo de tratamento de Água - FILTRACAO	0.010
Tipo de tratamento de Água - CLORACAO	0.010
Diagnóstico - SUPERVISAO DE GRAVIDEZ NORMAL, NAO ESPECIFICADA	0.010
Diagnóstico - INFECCAO AGUDA DAS VIAS AEREAS SUPERIORES NAO ESPECIFICADA	0.010
Coleta Fezes/Urina - SISTEMA DE ESGOTO	0.010

Um dado interessante é que a partir da análise das características importantes retornados pelo modelo Random Forest foi possível ver que o tipo de tratamento de água e o sistema de saneamento básico(destaca em vermelho acima) tem uma importância na determinação da necessidade de internação o que indica que as prefeituras deveriam investir mais em saneamento básico para tentar diminuir o número de internação e liberação das vagas dos hospitais para casos realmente que não podem ser evitados com saneamento.

Reflexão

Neste trabalho iniciamos com o processo de levantamento dos dados. Como os dados eram muito numerosos foi necessário realizar uma amostragem e também balancear as classes para melhoria do processo do modelo escolhido. Uma vez que os dados foram importados e pré-processados o modelo foi aplicado de uma forma geral. Foi realizado uma comparação com um modelo aleatório para certificarmos que realmente houve uma melhora na performance se comparado ao modelo aleatório. Além disso um processo de melhoria do modelo foi definido para escolhermos os melhores hiperparâmetros.

O processo mais complicado foi o de encontrar dados para analise, e como foi mencionado os dados não foram planejados para utilização de modelo preditivo e por isso uma análise mais complexa não foi necessária, mas como o objetivo era puramente educacional acredito que isso não teve impacto direto na resolução do problema proposto mas mostra claramente a importância do planejamento de coleta de dados. Em um projeto real poderíamos trabalhar junto a prefeitura para melhorarmos os dados coletados com o resultado de exames de sangue, fezes, urina, eletrocardiogramas e até talvez de radiogramas para aplicarmos técnicas de deep learning para detecção de doenças mais graves.

Melhorias

Utilizamos os dados disponíveis para realizar a previsão. Estes dados não foram coletados com uma visão de utilização para predição e, portanto, podem não ser os mais indicados, mas como nosso intuito e apenas educacional isso não será um fator crítico. No entanto em um ambiente real poderíamos realizar um trabalho junto a prefeitura para obtermos outros dados que não estão disponíveis atualmente, como por exemplo, os dados dos resultados de exames de sangue, fezes, urina e eletrocardiograma que poderiam ser feitos nestas unidades básicas de atendimento a fim de melhorarmos a precisão das previsões.
