Sessão 08

WiDS Recife Live Coding, 08/02/2020

Como serão os live codings?

- Sessões ao-vivo todos os sábados das 14h às 15h
- Código e slides serão disponibilizados no nosso site
- O objetivo é treinar para participar do Datathon em 2020

Roteiro

- Estratégias de cross-validation para o modelo
- Feature engineering

O problema

- O objetivo é criar um modelo que preveja a probabilidade do paciente sobreviver a partir dos dados das primeiras 24 horas dele na UTI do hospital
- A métrica utilizada para avaliação é a ROC AUC

Etapas para resolver o problema

1. Importar os dados
2. Explorar os dados
3. Treinar o modelo com o conjunto de treinamento
4. Avaliar o modelo com o conjunto de testes
5. **Realizar interativamente as etapas 2, 3 e 4**

Estratégias de cross-validation

Estratégias de cross-validation

- Observando o ‘hospital_id’ dos dados que a gente precisa prever eu vi que só 9 hospitais estão no conjunto de treinamento e no conjunto de testes
- Então decidi que talvez seja uma boa ideia fazer a validação cruzada com **grupos de hospitais distintos** (em vez de ir direto pro StratifiedKfold)
- Para isso vamos usar o **GroupKFold**

GroupKFold

- Parâmetro **n_splits**: quantos grupos queremos
- Parâmetro grupos: ids do grupo para cada instância dos dados
- Com isso cada um dos grupos de dados (folds) tem grupos distintos (no nosso caso, ‘hospital_id’ distintos)

Feature engineering

Feature engineering

- Há diversas formas de codificar features usadas como entradas de um modelo, e existem representações que são **melhores ou piores do que outras**
- Feature engineering é o processo de criar representações que melhoram a eficiência de um modelo

Feature engineering

- Não existe uma fórmula para criar boas features, é um processo que acontece ao explorar os dados
- Você cria novas features e avalia se elas são úteis
- O scikit-learn possui vários métodos de seleção de features: https://scikit-learn.org/stable/modules/classes.html#module-sklearn.feature_selection

Novo algoritmo: catboost

Catboost

- Não querer muito trabalho quanto aos hiperparâmetros (ele já tem um bom desempenho com os parâmetros default)
- Algoritmo que consegue lidar com **variáveis categóricas**

Nossa estratégia

1. Treinar um modelo com catboost
2. Verificar a importância de cada feature
3. Criar novas features
4. Verificar a importância de cada feature novamente para ver se as novas features são úteis ou não

Obrigada!

E até semana que vem!

—

Referências

- [1] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GroupKFold.html
- [2] http://www.feat.engineering
- [3] https://catboost.ai/