Jeydson Lopes da Silva
jeydsonl@gmail.com

Desafio Técnico IEL/FORD

Possíveis soluções para problemas de Data Scarcity, Unbalanced e Not Validated.

- **Data Scarcity**

Os modelos de Machine Learning no geral necessitam de uma grande quantidade de dados para apresentarem um desempenho adequado. No entanto, nota-se que é comum ocorrer dificuldades na coleta de um conjunto de dados de treinamento grandes o suficiente.

De forma a contornar tal situação pode-se optar por algumas soluções, destacando-se:

i. **Data Generation** – Busca-se realizar manipulações com a base de dados existente de forma a criar novos dados. Exemplo: a partir de uma única imagem de um cachorro, pode-se gerar outras apenas invertendo, recortando, diminuindo o tamanho, ou dando um zoom.
ii. **Transfer Learning** – Busca-se utilizar modelos já treinados com outro conjunto de dados. Nesse caso, deve-se considerar o fato de não haver dados de treinamento suficientes e os domínios de origem e destino têm algumas semelhanças, mas não necessariamente idênticos.

- **Data Unbalanced**

Os modelos de Machine Learning no geral necessitam de uma grande quantidade de dados para apresentarem um desempenho adequado. De fato, a maioria dos modelos de Machine Learning não apresentam um bom funcionamento quando relacionados a conjuntos de dados desequilibrados.

De forma a contornar tal situação pode-se optar por algumas soluções, destacando-se:

1) **Resample do conjunto de treinamento**
    i. **Under-sampling –** Equilibra o conjunto de dados reduzindo o tamanho da classe abundante.
    ii. **Over-sampling -** Equilibra o conjunto de dados aumentando o tamanho de amostras raras.
2) **K-fold Cross-Validation de forma adequada** – É importante salientar que a validação cruzada deve ser aplicada adequadamente ao usar o método de **Over-sampling** para resolver os problemas de desequilíbrio dos dados. A validação cruzada deve sempre ser feita antes do s **Over-sampling** dos dados.
3) **Clusterização da classe abundante -** Em vez de apenas depender de amostras aleatórias para cobrir a variedade das amostras de treinamento, sugere-se clusterizar a classe abundante em R grupos, com R sendo o número de casos em R. Para cada grupo, apenas o Medoid (centro do cluster) é mantido. O modelo é então treinado com a classe rara e apenas os Medoids.
4) **Combinar diferentes conjuntos de dados resampled** - Geralmente alguns classificadores como a logistic regression ou random forest tendem a generalizar "jogando fora" a classe rara. Assim sendo, pode-se construir N modelos que usem todas as amostras da classe rara e N amostras diferentes da classe abundante. Por exemplo, deseja-se reunir 10 modelos, manter-se-ia os 1.000 casos da classe rara e amostram aleatoriamente 10.000 casos da classe abundante. Assim sendo, divide-se as 10.000 caixas em 10 blocos e treina 10 modelos diferentes.

- **Data Not validated**

A validação de dados é muito importante essencial de qualquer tarefa de tratamento de dados. Nesse sentido, caso os dados não sejam precisos desde o início, provavelmente os resultados também não serão precisos. Dessa forma, faz-se necessário verificar e validar os dados antes de serem usados.

Jeydson Lopes da Silva
jeydsonl@gmail.com

De forma a contornar tal situação pode-se optar por algumas soluções, destacando-se:

iii. **k-Fold Cross-Validation**– Cross-validation é um procedimento de resampling usado para avaliar modelos em uma amostra de dados limitada. Este método popular porque é simples de entender, geralmente resulta em uma estimativa menos tendenciosa ou menos otimista da habilidade do modelo do que outros métodos, como uma divisão simples de train/test.

iv. **Leave-one-out cross-validation–** É um caso especial da cross-validation em que o número de folds é igual ao número de instâncias no dataset. Assim, o algoritmo de aprendizado é aplicado uma vez para cada instância, usando todas as outras instâncias como um conjunto de treinamento e usando a instância selecionada como um conjunto de teste de item único.