Lucas Colucci
Github: https://github.com/lucascolucci/VivaReal/
June 17, 2016

Desafio Viva Real

Data Science Challenge

Este documento tem como objetivo explicar como as previsões salvas no arquivo result.csv foram estudadas e desenvolvidas.

Análise Gráfica

O primeiro passo para conhecer os dados que tinha em mão foi gerar gráficos que ilustrassem a relação entre cada feature e o target. Através disso consegui observar que alguns gráficos pareciam ter um mesmo tipo de distribuição, portanto fazia sentido estudar mais a fundo a correlação entre as features para que não gerássemos modelos com dados irrelevantes.

Análise de Correlação

Através do estudo de correlação gerado pelo pacote "mice" do R, identifiquei que a feature_8 tinha correlação maior que 0.99 com a feature_11. Sendo assim, fez sentido desconsiderar feature_8 do cálculo.

No entanto ainda não tinha convicção de que essa era a única feature que poderia ser removida ou modificada. Fazendo uma análise superficial, percebi que a feature_13 se mantinha igual em todas as linhas, portanto não impactaria no resultado final. Por isso ela foi removida também.

Outra situação observada foi que as features 12 e 14 tinham 97% de dados iguais, portanto decidi remover a feature_12 do dataset.

Análise dos Fatores

Após transformar o arquivo CSV em data frame, eu notei que as features 1, 2, 3 e 17 foram consideradas factors. No entanto estas fatures tinham milhares de leveis diferentes. Isso me fez pensar que, já que o conteúdo dessas colunas eram hexadecimais, eles poderiam conter números e não classes. Portanto transformei tais colunas de Hex para Numeric.

Tratamento de Valores NA

Outra análise mostrou que as features 0, 10 e 16 tinham muitas informações faltando, o que atrapalharia no cálculo de uma regressão. Com isso em mente decide testar 3 tipos de soluções para esse problema e escolher o que tivesse um melhor resultado. As três ideias eram:

- Ignorar colunas com NA
- Substituir NA com a media dos valores da coluna
- Substituir NA com zeros.

Modelos de Regressão

Os modelos de regressão escolhidos para serem comparados foram:

- Linear Regression
- SVM EPS Regression with Linear Kernel
- SVM EPS Regression with Polynomial Kernel
- SVM EPS Regression with Radial Kernel
- SVM EPS Regression with Sigmoid Kernel

Resultados

Para achar o melhor conjunto de substituição de valores faltantes e melhor modelo de regressão, desenvolvi minha própria função para fazer N-Fold Cross Validation, no qual passamos o modelo e como tratar os dados e ela performa a cross validation com esses atributos. Usei um cross validation com 5 fold com todas as combinações diferentes e cheguei no seguinte resultado:

- Melhor modelo: SVM EPS Regression with Radial Kernel
- Substituir NA com a média dos valores da coluna
- Erro = 0.0635724920681937

O erro foi calculado de acordo com as instruções recebidas do challenge:

"Error = median(abs(True - Pred) / True)"

Descrição dos Arquivos

Desafio_Analysis.R: Contém toda a análise dos dados, comparação de modelos e resultado do melhor modelo.

Desafio_Application.R: Contém a aplicação do melhor modelo e método escolhido durante a análise

Results.csv: Contém o resultado da regressão no formato CSV [id, target]