Projeto Final

1. INTRODUÇÃO

Na análise preditiva queremos desenvolver modelos estatísticos que, com base em dados, conseguimos prever um determinado *outcome* de interesse. A partir desse modelo devemos garantir que com o aprendizado de comportamentos e experiências passadas conseguimos obter um modelo com um bom grau de generalização.

Nesse projeto, vocês irão desenvolver um modelo preditivo a partir de dados reais sobre um tema de interesse (dica: pense aqui nos diversos cases vistos em aula como finanças, CRM, seguros, saúde, *people analytics*, telecomunicações, *online shopping*, redes sociais, etc).

2. GUIDELINES

- Grupos: de até 4 pessoas;
- Peso: 30% da nota final;
- Data de entrega: até 2 semanas após a prova final do curso;
- Dados: a internet disponibiliza uma rica fonte de bases de dados reais com diversas aplicações (podem centralizar no kaggle);
- Cuidado com bases muito grandes e com muitas variáveis (+ trabalho e tempo de processamento);
- Apresentação: relatório em pdf + códigos .R
- Técnicas: escolham duas técnicas diferentes OU apenas uma técnica mas estressando a configuração de parâmetros

3. ROTEIRO GERAL

- Problema: breve descrição do problema de negócio que se deseja atuar;
- Fonte de dados: onde foi extraída a informação (link), tamanho da base, quantidade de variáveis;
- Descrição das variáveis, quais são qualitativas e quais são quantitativas (dica: tabela);
- O que quer dizer a variável resposta?
- Data prep.: existem variáveis com missing values, outliers? Foram tratadas, eliminadas?, descreva que tipo de tratamentos foram feitos (aqui podem usar qualquer pacote do R que vocês estiverem familiarizados)
- Como foi definida a amostra?
- Desenvolvimento do modelo:
  - no caso de dois algoritmos diferentes: construa cada modelo e faça a comparação de ambos baseados em algum argumento de performance (atentem-se para *overfitting*).
  - apenas um tipo de algoritmo: faça o *tuning* dos parâmetros e, em seguida, a comparação baseada em algum argumento de performance (atentem-se para *overfitting*). Se utilizar árvore atente-se para a poda.
- Com o modelo finalista apresente os parâmetros finais e as variáveis mais importantes. (Se for logístico apresente a equação final também). Plote a árvore caso tenha sido a finalista.
- Construa uma regra de negócio utilizando a matriz de confusão. Crie um racional para o ponto de corte para poder utilizar o modelo caso a sua empresa fictícia tenha optado para tal.
- Apêndice: para as variáveis finalistas do modelo, ou seja, apenas as que entraram no modelo final, apresente a análise uni e bivariada (seja criativo na construção dos plots: pense nas variáveis quali e quanti em qual forma seria melhor de mostrar a distribuição e a relação com a variável target).