Trabalho 2

Primeiramente foi realizada a importação da base de dados para o R Studio. Em seguida foi analisado os valores faltantes por coluna, onde foram encontradas algumas colunas com mais de 30% de valores faltantes. Foram retiradas as variáveis que possuíam mais de 30% de valores faltantes e que também não foram significativas para o modelo. As variáveis mantidas foram as seguintes: tp_camp, tot_env, tot_open, tot_clik, qtd_dias_ult_tran e a variável resposta fg_clik. Além disso, a variável tp_sexo e fg_clik foram transformadas para o tipo factor.

A base de dados é muito desbalanceada em relação a variável resposta fg_clik. Para resolver esse problema foi aplicado a função ovun_sample para aplicar o método de oversampling a fim de que a base de dados passe a ter 20% com valor 1 (positivo).
Em seguida, foi realizada a normalização dos dados por meio da função preProcess com método “range”. Após isso, foi aplicado a função cbind para juntar as variáveis normalizadas e a variável resposta. Posteriormente, foi feita a divisão da base de dados email em treino e teste. Depois foi aplicado a função glm com família binomial sobre a base de dados de treino e a função predict para aplicar o modelo sobre a base de teste. Após isso, foi possível calcular a ROC, AUC e fazer a matriz de confusão para as métricas de avaliação de modelos de classificação.

O modelo obteve uma AUC (medida agregada para todos os possíveis limites de classificação) de 0.9478 e a matriz de confusão retornou um sensibilidade de 0.8752 e uma especificidade de 0.9753. Então podemos concluir que o modelo captura com sucesso 87% dos resultados classificados como positivos (caso em que a pessoa realmente iria clicar na campanha) e 97% dos resultados classificados como negativos (caso em que a pessoa realmente não iria clicar na campanha).