Trabalho 2

Aluno: Hugo Muniz Albuquerque

Disciplina: Análise Preditiva

Data: 07/07/2020

Primeiramente foi realizada a importação da base de dados para o R Studio. Em seguida foi analisado os valores faltantes por coluna, onde foi encontrado 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.

Foi notado a presença de de valores faltantes nas variáveis restantes. Para resolver esse problema foi aplicado a função mice, na qual é possível aplicar um método de imputação expecifico para cada caso. Para variáveis numéricas foi utilizado a "pmm" (imputação de correspondência media preditiva). Para variaveis do tipo factor que possuem mais de 2 levels foi aplicado o método "polyreg" (regularização polynomial). Para variáveis do tipo factor com mais de 2 levels foi aplicado o método "logreg" (regularização logística). Este processo possibilitou que o modelo passasse no teste de autocorrelação DW.

Pelo fato da base de dados ser muito desbalanceada em relação a variável resposta fg_clik foi aplicado a função ovun_sample para aplicar o método de oversampling a fim de que a base de dados passasse a ter 20% com valor 1 (positivo). Em seguida, foi realizado 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 feito 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. O modelo obtido passou somente no teste Durbin-Watson de autocorrelação.

Por último, foi aplicado 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 um

R^² aproximado de 0.51, a AUC foi de 0.9572 e a matriz de confusão retornou um sensibilidade de 0.76 e uma especificidade de 0.97. Então podemos concluir que o modelo funciona para prever a quantidade média demandada em um único dia.