Universidade Federal do Rio de Janeiro
Aluno: Danilo Filippo Reiszel Pereira
ADI CPF 160.636.807-99

COC786 - Inteligência Computacional
Prof. Alexandre Evsukoff

Relatório Parcial

1 Introdução

Com o grande aumento da quantidade de dados disponível nos tempos recentes, cada vez mais empresas tem utilizado a ciência de dados para a tomada de decisões. Seja apenas uma análise exploratória dos dados disponíveis ou a criação de um modelo, tal paradigma ficou conhecido no mercado como *Data Driven Decision Making* e, por mais que pareça intuitivo, tem revolucionado a forma como empresas alocam os seus recursos. Assim, este trabalho tem como objetivo estudar um conjunto de dados referentes à campanhas de marketing de um banco português, com a esperança de se criar um modelo capaz de otimizar os alvos (indivíduos) destas campanhas.

Considera-se este trabalho interessante pois, diferente de muitas aplicações de *machine learning*, na qual apenas uma das partes é beneficiada (como por exemplo análises de crédito ou currículo), a criação de um modelo de marketing efetivo é desejável para ambas as partes. Para a instituição bancária, permite a economia de recursos e um aumento na taxa de conversão ao concentrar esforços em candidatos mais propícios a contratarem o serviço. Para os indivíduos, minimiza o número de ligações de marketing indesejadas, e otimiza estas ligações para que possuam conteúdo de maior interesse.

Para o desenvolvimento do trabalho, será utilizada a linguagem de programação Python. Em especial, para a manipulação, exploração e visualização dos dados, serão utilizadas as bibliotecas *pandas* e *matplotlib*, enquanto os modelos serão desenvolvidos e treinados utilizando a biblioteca *scikit-learn*.

2 Caracterização e visualização

Os dados foram coletados entre Maio de 2008 e Novembro de 2010, totalizando 41188 registros de ligações, que contém dados do indivíduo, informações sobre a campanha e indicadores socioeconômicos gerais. O dataset contém um total de 21 variáveis, sendo 20 delas possíveis canditadas a entradas do modelo e uma o alvo desejado. Conforme podemos observar na Figura 1, o dataset não possui registros com dados faltantes.

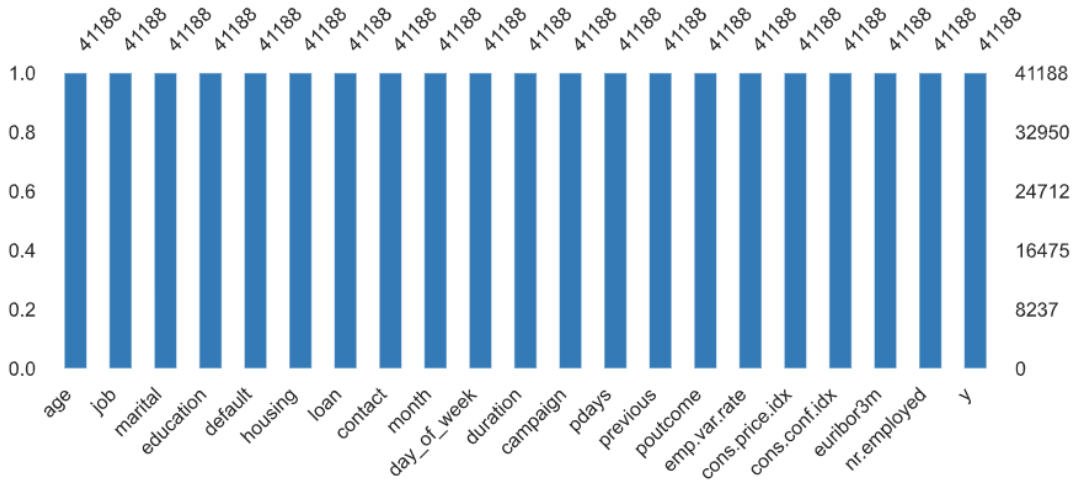

Figura 1: Contagem de valores não nulos por coluna

Uma breve descrição das variáveis pode ser vista a seguir:

2.1 Variáveis

- **y**:

  Variável alvo do modelo. Indica se o indivíduo contratou ou não o serviço ao final da chamada de marketing. Pode-se observar, na Figura 2, que o dataset é altamente desbalanceado, com apenas 11.3% dos registros contendo indivíduos que contratatam o serviço. Tal desbalanceamento era esperado em dados de marketing.

Figura 2: Características da coluna de alvo

- **age**: Idade do cliente abordado. Variável numérica inteira.
- **job**: Variável categórica indicando o tipo de trabalho do indivíduo. Possíveis valores: ('admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')

- **marital**: Estado civil do cliente. Variável categórica. Possíveis valores: ('divorced','married','single','unknown'). Obs.: 'divorced' pode indicar divorciado ou viúvo
- **education**: Nível educacional, variável categórica. Possíveis valores: ('basic.4y','basic.6y','basic.9y
- **default**: Variável categórica. Indica se o cliente possui não pagas. Possíveis valores: ('yes','no','unknown')
- **housing**: Variável categórica. Indica se o cliente possui um empréstimo imobiliário. Possíveis valores: ('yes','no','unknown')
- **loan**: Variável categórica. Indica se o cliente possui um empréstimo pessoal. Possíveis valores: ('yes','no','unknown')
- **contact**: Variável categórica. Indica se a chamada foi feita atráves de telefone fixo ou celular. Possíveis valores: ('cellular','telephone')
- **month**: Mês da chamada
- **day_of_week**: Dia da semana da chamada
- **duration**: Duração da chamada
- **campaign**: Número de chamadas já feitas nesta campanha para o cliente
- **pdays**: Número de dias desde a última chamada para o cliente. 999 indica que o cliente não foi contactado anteriormente.
- **previous**: Número de chamadas anteriores à campanha atual já feitas para o cliente
- **poutcome**: Variável categórica. Indica o resultado da campanha de marketing anterior para este cliente. Possíveis valores: ('failure','nonexistent','success')
- **emp.var.rate**: Indicador socioeconômico trimestral (*employment variation rate*)
- **cons.price.idx**: Indicador socioeconômico mensal (*consumer price index*)
- **cons.conf.idx**: Indicador socioeconômico mensal (*consumer confidence index*)
- **euribor3m**: Indicador socioeconômico diário (*euribor 3 month rate*)
- **nr.employed**: Indicador socioeconômico trimestral (*number of employees*)

2.2 Correlações

Devido ao grande número de variáveis categóricas, optou-se por avaliar a correlação de Phik. Este método de correlação foi desenvolvido para ser consistente mesmo quando as variáveis são categóricas ou intervaladas, com um *fallback* para a correlação de pearson caso seja viável aplicá-la. Seu resultado pode ser observado na Figura 3.

Podemos observar uma forte correlação entre os indicadores econômicos presentes do dataset, o que é esperado, pois são todos medidas de características da economia geral. Também é possível observar uma forte correlação destes indicadores com a variável de mês, uma vez que muitos indicadores socioeconômicos apresentam sazonalidade anual. Já para a variável alvo, **y**, pode-se observar que o grupo de entradas com maior correlação é também o grupo de indicadores, o que sugere que momentos econômicos gerais são mais relevantes na taxa de conversão destas campanhas do que características individuais.

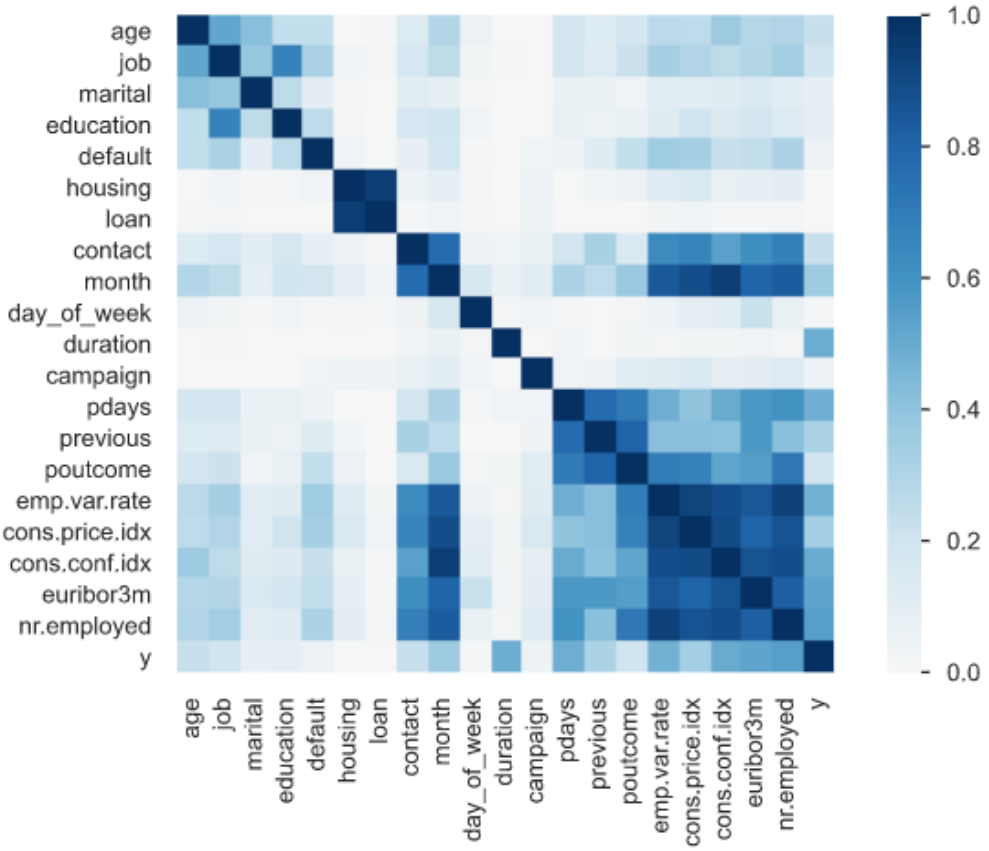

Figura 3: Correlação de Phik entre as variáveis