Proposta

Histórico do assunto

Ao longo desses anos o telemarketing sofreu uma série de mudanças, ou melhor, transformações que estão relacionadas diretamente com a evolução da tecnologia. Surgiram conceitos como Omnichannel (multicanal), Callback, Atendentes Virtuais, Chatbot, Portais de Autoatendimento e etc.

O cliente mudou, atualmente eles está no controle, não quer esperar 'pendurado' na linha e também não quer ser importunado por ligações oferecendo produtos que não são do seu interesse. Ele pode a qualquer momento deixar de comprar na sua loja, por exemplo, e escolher o mesmo produto em outra loja utilizando qualquer meio (ex: smartphone, tablet). Por outro lado, existe um potencial público de clientes que pode contratar o produto, desde que, seja realizado um contato e ofertado o produto adequado.

Nesse cenário é que a proposta desse projeto se enquadra, utilizando machine learning para selecionar o público que será ofertada a Campanha. O modelo de machine learning treinado e testado conseguirá prever quais são os clientes mais propensos a contratar um empréstimo bancário.

Encontrei dois estudos interessantes que abordam o tema:

1 – Feature Selection with Data Balancing for Prediction of Bank Telemarketing

Fonte: http://m-hikari.com/ams/ams-2014/ams-113-116-2014/vajiramedhinAMS113-116-2014.pdf

Escrito por: Chakarin Vajiramedhin e Anirut Suebsing

Esse estudo propõe a redução do conjunto de dados de entrada utilizando métodos para selecionar (Feature Selection) os recursos mais relevantes, uma vez que, o desempenho da previsão do modelo é afetado diretamente por isso. O estudo também aborta a seleção de subconjuntos baseado na correlação e balanceamento do conjunto (Data Balancing), visando aumentar a taxa de previsão do conjunto de dados.

2 – A Data-Driven Approach to Predict the Success of Bank Telemarketing

Fonte: https://repositorio.iscte-iul.pt/bitstream/10071/9499/1/post_print_dss_v3.pdf

Escrito por: Sérgio Moro, Paulo Cortez e Paulo Rita

Esse estudo aborda a importância das instituições financeiras trabalharem orientada a dados, utilizando modelos preditivos para selecionar o público das suas campanhas de Telemarketing. Os modelos de classificação estudados foram: logistic regression, decision trees, neural network, support vector machine (SVM).

Na seleção das features do conjunto, o estudo propõe uma seleção semi-automática. Na primeira etapa, o conhecimento do negócio é utilizado na seleção e na segunda etapa é feita uma seleção utilizando o método 'forward selection'.

Motivação da escolha do tema:

O que me motivou a escolher esse problema foi porque trabalho numa instituição financeira e há mais ou menos 1 ano estou atuando como engenheiro de dados. Nessa área de dados recentemente foi criado uma gerencia de machine learning e tenho interesse em migrar para essa gerencia.

Descrição do problema

O problema que estou propondo solucionar é selecionar de forma mais assertiva o público que será ofertada campanhas de marketing do produto empréstimo a prazo de uma instituição bancária. A seleção não assertiva implica um alto gasto (estudo e seleção do público, de tempo, em pessoas, em telefonia, etc) contatando clientes que não tem interesse em contratar o produto.

Dessa formar será dado um foco nos clientes que realmente desejam contratar o produto, com isso a instituição financeira poderá investir mais 'energia' em melhor atender e oferecer serviços mais atraentes aos seus clientes.

Conjuntos de dados e entradas

Esse conjunto de dados é referente a uma campanha de marketing realizada por uma institução bancária portuguesa. Essa campanha foi baseada em chamadas telefonicas. O produto ofertado foi um empréstimo a prazo.

Esse conjunto de dados é público, disponivel para pesquisa. Os detalhes estão descritos em [Moro et al., 2014]. O conjunto foi obtido no site da UCI (https://archive.ics.uci.edu/ml/datasets/bank+marketing)

[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

Features do Conjunto:

ID	Label	Descrição	Tipo	Dominio
1	Age	Idade	Numérica	
2	Job	Tipo de Emprego	Categórica	admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown'
3	Marital	Estado Civil	Categórica	divorced','married','single','unknown'; note: 'divorced' means divorced or widowed
4	Education	Escolaridade	Categórica	basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown'
5	Default	Tem Crédito Default	Categórica	no','yes','unknown'
6	Housing	Tem Empréstimo Habitacional	Categórica	no','yes','unknown'
7	Loan	Tem Empréstimo Pessoal	Categórica	no','yes','unknown'
8	Contact	Tipo de Contato	Categórica	cellular','telephone'
9	Month	Mês do último contato	Categórica	jan', 'feb', 'mar', ..., 'nov', 'dec'
10	day_of_week	Último dia de contato da semana	Categórica	'mon','tue','wed','thu','fri'
11	Duration	Duração do último contato em segundos	Numérica	Nota importante: este atributo afeta altamente a meta de saída (por exemplo, se a duração for = 0, então y = 'não'). No entanto, a duração não é conhecida antes de uma chamada ser executada. Além disso, após o término da chamada, é obviamente conhecido. Assim, essa entrada deve ser incluída apenas para fins de benchmark e deve ser descartada se a intenção for ter um modelo preditivo realista.
12	campaign	Número de contatos da última camapanha	Numérica	
13	Pdays	Número de dias que se passaram depois que o cliente foi contatado pela última vez de uma campanha anterior. 999 indica que cliente não foi contato em campanha anterior	Numérica	
14	Previous	Número de contatos realizados antes desta campanha	Numérica	

15	poutcome	Resultado da campanha de marketing anterior	Categorica	failure','nonexistent','success'
16	emp.var.rate	Taxa de variação de emprego	Numérica	
17	cons.price.idx	Indice de preços ao consumidor	Numérica	
18	cons.conf.idx	Indice de confiança do consumidor	Numérica	
19	euribor3m	Euro	Numérica	
20	nr.employed	Número de empregos	Numérica	
21	Target	Cliente contratou produto bancário	Binaria	yes','no'

Descrição da solução

A solução proposta para esse problema é a execução de um modelo de aprendizagem supervisionado com o dataset real. Será necessário efetuar uma preparação dos dados do conjunto para execução do modelo. Esse modelo será treinado com esse conjunto (já preparado), gerando no final do seu processamento, um modelo matemático que consegue prever (a partir dos dados do cliente) se o cliente irá contratar o produto/serviço bancário.

Farei um split do arquivo, usando 80% do dataset para aprendizagem (treinamento) e 20% para testes. Os resultados dos testes serão comparados com os dados reais e dessa forma conseguiremos medir a efetividade do modelo gerado.

A partir desse momento, nas próximas campanhas, a seleção do público será feita apresentando o novo conjunto (público) para o modelo treinado. Ao término da execução do modelo teremos a informação de quais clientes provavelmente contratarão o produto.

Modelo de referência (benchmark)

Encontrei alguns trabalhos no GitHub, eles serão como modelo de referencia para o modelo que estou propondo.

Modelo de Referencia 1:

Fonte: https://github.com/krishtanwani/bank-additional

Modelo: LogisticRegression

- ⇨ Acurácia: 0.896981467994
- ⇨ Tamanho do Teste: 30%

Modelo de Referencia 2 (essa fonte possui 3 modelos):

Fonte: https://github.com/juliencohensolal/BankMarketing

Modelo: SVM

- ⇨ Acurácia: 0.899193939983
- ⇨ Tamanho do teste: 25%

Modelo: RandomForestClassifier

- ⇨ Acurácia: 0.88798679227
- ⇨ Tamanho do teste: 25%

Modelo: LogisticRegression

- ⇨ Acurácia: 0.900320481694
- ⇨ Tamanho do teste: 25%

Métricas de avaliação

As métricas utilizadas serão (do Pacote Scikit.metrics):

accuracy_score (acurácia): Proporção de casos que foram corretamente previstos, sejam eles verdadeiro positivo ou verdadeiro negativo. Com isso, mede a frequência que o modelo faz a previsão correta.

f1_score (score): Média ponderada dos scores de precisão (precision) e sensibilidade (recall). Varia entre 0 e 1, sendo 1 a melhor pontuação possível.

precision_score (precisão): Proporção de casos positivos que o modelo classificou como positivo e eram positivas, ou seja, é a capacidade de não rotular como positiva uma amostra que é negativa (Verdadeiro Positivo / (Verdadeiro Positivo + Falso Positivo))

recall_score (sensibilidade): Proporção de casos realmente positivos e que foram classificados pelo modelo como positivo, ou seja, é a capacidade do modelo de

encontrar todas as amostras positivas. ((Verdadeiro Positivo / (Verdadeiro Positivo + Falso Negativo))

- confusion_matrix (Matriz de confusão): Essa métrica calcula a quantidade de falso positivo e falso negativo, e de verdadeiro positivo e verdadeiro negativo

Design do projeto

O design desse projeto consistirá de:

1 – Analise Exploratória

- Identificar correlações (para auxiliar na eliminação de features)

- Selecionar features mais preditivas

2 – Feature Enginnering:

    - Features Numéricas:

        - Inputar dados nulos (se houver)

        - Ajustar a scala dos valores

    - Features Categoricas

        - Transformar em códigos para execução do modelo

3 – Treinar o Modelo (com 80% do conjunto)

4 – Testar o Modelo (com 20% do conjunto)

5 – Avaliar Modelo (acurácia e generalização)

6 – Otimizar modelo

- Ajustando variáveis do modelo

- Ajustando as escalas das features e/ou adicionando, eliminado ou trocando features

7 – Treinar novamente o Modelo

8 – Testar novamente o Modelo

9 – Avaliar Modelo (acurácia e generalização)

10 – Otimizar novamente modelo até chegar no ponto que o ajuste não reflita mais na melhoria da acurácia e generalização