ifood_test_caio_lima

Case Ifood - Caio Lima

Introdução

Empresa

Considere uma empresa bem estabelecida que opera no setor varejista de alimentos. Atualmente possui centenas de milhares de clientes registrados e atendem quase um milhão de consumidores por ano. A empresa vende produtos de 5 categorias: vinhos, produtos de carnes raras, frutas exóticas, peixes especialmente preparados e produtos doces. Essas categorias podem ser subdivididas em produtos premium (Ouro) e produtos convencionais (Regulares). Os clientes podem fazer pedidos e adquirir produtos por meio de três canais de vendas: lojas físicas, catálogos e o site da empresa. Embora a empresa tenha registrado sólidos resultados financeiros globais e tenha mantido uma margem de lucro saudável nos últimos três anos, as perspectivas de crescimento dos lucros para os próximos três anos não são muito encorajadoras. Por esse motivo, estão sendo consideradas diversas iniciativas estratégicas para reverter essa situação, sendo uma delas a melhoria do desempenho das atividades de marketing, com um enfoque especial em campanhas de marketing.

Departamento de Marketing

O departamento de marketing enfrentou pressão para utilizar seu orçamento anual de forma mais eficaz. O Chief Marketing Officer (CMO) reconhece a importância de adotar uma abordagem mais orientada para dados ao tomar decisões. É por isso que uma pequena equipe de cientistas de dados foi recrutada com um objetivo claro em mente: desenvolver um modelo preditivo que apoie as iniciativas de marketing direto. Idealmente, o êxito dessas atividades demonstrará o valor dessa abordagem e persuadirá os indivíduos mais céticos dentro da empresa.

Objetivo

A equipe tem como meta desenvolver um modelo de previsão que otimize o lucro para a próxima campanha de marketing direto, agendada para o próximo mês, que é a sexta campanha da série. Esta campanha tem como objetivo principal a venda de um novo dispositivo para a base de clientes. Para criar o modelo, eles realizaram uma campanha piloto com 2.240 clientes selecionados aleatoriamente e contatados por telefone para adquirirem o dispositivo. Nos meses seguintes, identificaram os clientes que aderiram à oferta. O custo total dessa campanha de amostra foi de 6.720MU, enquanto a receita gerada pelos clientes que aceitaram a oferta atingiu 3.674MU. No geral, a campanha teve um prejuízo de -3.046MU, com uma taxa de sucesso de apenas 15%. A equipe tem como objetivo principal desenvolver um modelo capaz de prever o comportamento do cliente e aplicá-lo ao restante da base de clientes. Com sorte, o modelo permitirá que a empresa selecione os clientes mais propensos a comprar a oferta, excluindo os que não respondem, tornando a próxima campanha altamente lucrativa. Além de maximizar os lucros da campanha, o CMO também está interessado em estudar as características dos clientes dispostos a comprar o dispositivo.

Conjunto de dados

O conjunto de dados inclui informações sociodemográficas e firmográficas sobre 2.240 clientes que foram contatados. Além disso, ele contém uma marcação para identificar os clientes que participaram da campanha, adquirindo o produto. As informações coletadas estão abaixo:

- **ID**: Identificador único do cliente
- **Year_Birth**: Ano de nascimento
- **Education**: Nível de educação
- **Marital_Status**: Estado civíl
- **Income**: Renda
- **Kidhome**: Número de crianças na residência
- **Teenhome**: Número de adolescentes na residência
- **Dt_Customer**: Data de cadastro do cliente
- **Recency**: Número de dias desde a última compra
- **MntWines**: Valor gasto em vinhos nos últimos 2 anos
- **MntFruits**: Valor gasto em frutas exóticas nos últimos 2 anos
- **MntMeatProducts**: Valor gasto em carnes raras nos últimos 2 anos
- **MntFishProducts**: Valor gasto em peixes nos últimos 2 anos
- **MntSweetProducts**: Valor gasto em doces nos últimos 2 anos
- **MntGoldProds**: Valor gasto em produtos premium nos últimos 2 anos
- **NumDealsPurchases**: Número de compras feitas com desconto
- **NumWebPurchases**: Número de compras feitas no site
- **NumCatalogPurchases**: Número de compras feitas no catálogo
- **NumStorePurchases**: Número de compras feitas diretamente nas lojas
- **NumWebVisitsMonth**: NNúmero de visitas ao site no último mês
- **Complain**: 1 se o cliente reclamou nos últimos 2 anos, 0 c.c.
- **AcceptedCmp1**: 1 se o cliente aceitou a oferta da 1ª campanha, 0 c.c.
- **AcceptedCmp2**: 1 se o cliente aceitou a oferta da 2ª campanha, 0 c.c.
- **AcceptedCmp3**: 1 se o cliente aceitou a oferta da 3ª campanha, 0 c.c.
- **AcceptedCmp4**: 1 se o cliente aceitou a oferta da 4ª campanha, 0 c.c.
- **AcceptedCmp5**: 1 se o cliente aceitou a oferta da 5ª campanha, 0 c.c.
- **Response**: 1 se o cliente aceitou a oferta da última campanha, 0 c.c.

Dos campos sitados, foram observadas algumas características determinantes para o não uso ou necessidade de transformação da informação, sendo eles:

- **ID**: Por ser um identificador único, não servirá como discrimador para a classificação dos clientes.
- **Year_Birth**: Pessoas nasceram entre 1940 e 1996, com exceção de 3 que "nasceram" antes de 1900 e dado o contexto teriam mais de 110 anos. Assim, podem se tratar de outliers.
- **Dt_Customer**: Algumas classes de modelos não aceitam o campo em formato de data, então seria interessante uma tranformação na informação para se adequar melhor aos cenários de classificação.
- **Marital_Status**: Há respostas incongruentes ('Absurd', 'Alone' e 'YOLO' - "You Only Live Once") e que serão interpretadas como a categoria 'Single'.
- **Income**: Há uma renda informada como sendo U$666666, um valor muito maior que a renda dos demais clientes e será considerado como Outlier. Além disso, houveram 16 clientes que não informaram a renda (resposta nula).
- **Complain**: Apenas 21 (0.9%) dos clientes fizeram reclamação. Como é pouco representativo, não servirá como discriminador para a classificação dos clientes.

Assim, os camops *ID* e *Complain* não serão considerados na classificação. Para *Marital_Status* foram feitos os ajustes necessarios. Foi criado um novo campo *DaysSinceEnrol*, uma transformação da variável *Dt_Customer*, que marca o número de dias desde o registro do cliente na empresa e agora passa a ser um campo númerico mais simples de ser interpretado nos modelos. Para *Year_Birth* pode-se desconsiderar os clientes Outliers ou substituir seus valores pela mediana/média dos demais clientes. O mesmo pode ser feito para o cliente Outlier em *Income*. Afim de não se perdar a informação destes clientes, optou-se por substituir os valores discrepantes pela mediana geral.

Análise Descritiva

O Gráfico 1 apresenta uma dispersão geral dos dados quantitativos (após as tratativas citadas) em relação à resposta, quanto maior e mais concentrada a núvem de pontos maior é quantidade de clientes e a concentração na parte superior indica um maior número de casos positivos (compra do produto) ao longo dos valores de cada variável de interesse. A linha azul indica uma tendência, se está crescente então conforme o valor da variável aumenta mais clientes compraram o produto após receber a campanha, sendo análogo quando está decrescente. Desta forma, é possível observar que conforma a renda aumenta a compra do produto também aumenta. O mesmo ocorre com os gastos em cada tipo de produto. Conforme aumenta o número de dias desde o último pedido, diminui as chances de o cliente ter comprado o produto.

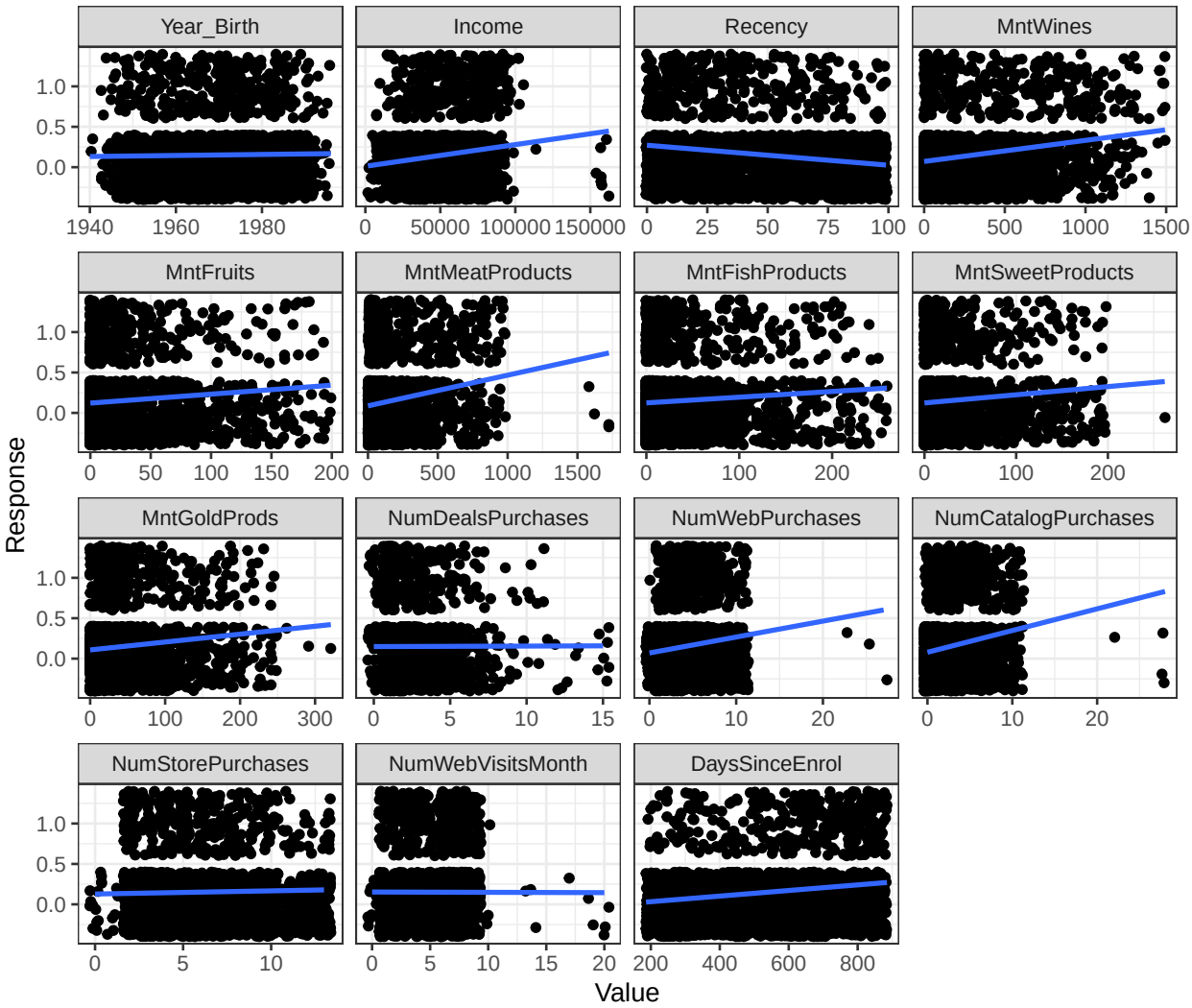

Figure 1: Disperção entre a Resposta e demais campos numéricos pós transformação

Já o Gráfico 2 apresenta uma dispersão geral dos dados qualitativos (após as tratativas citadas) em relação à resposta. Em educação é possível observar claramente que conforme o nível de edução aumenta maior é a proporção de clientes que comprarm o produto, um bom indicativo para o uso do campo. Em estado civíl há uma menor proporção de casos positivos para clientes que estão em relacionamento ("Together" ou "Married"). Conforme o número de crianças/adolescentes na residência aumenta, menos clientes compram o produto, proporcionalmente. Por fim, há um aumento muito grande nas chances do cliente comprar o produto

se aceitou alguma das últimas 5 campanhas, ou seja, quem já aceitou uma vez tem uma maior propabilidade de aceitar novamente.

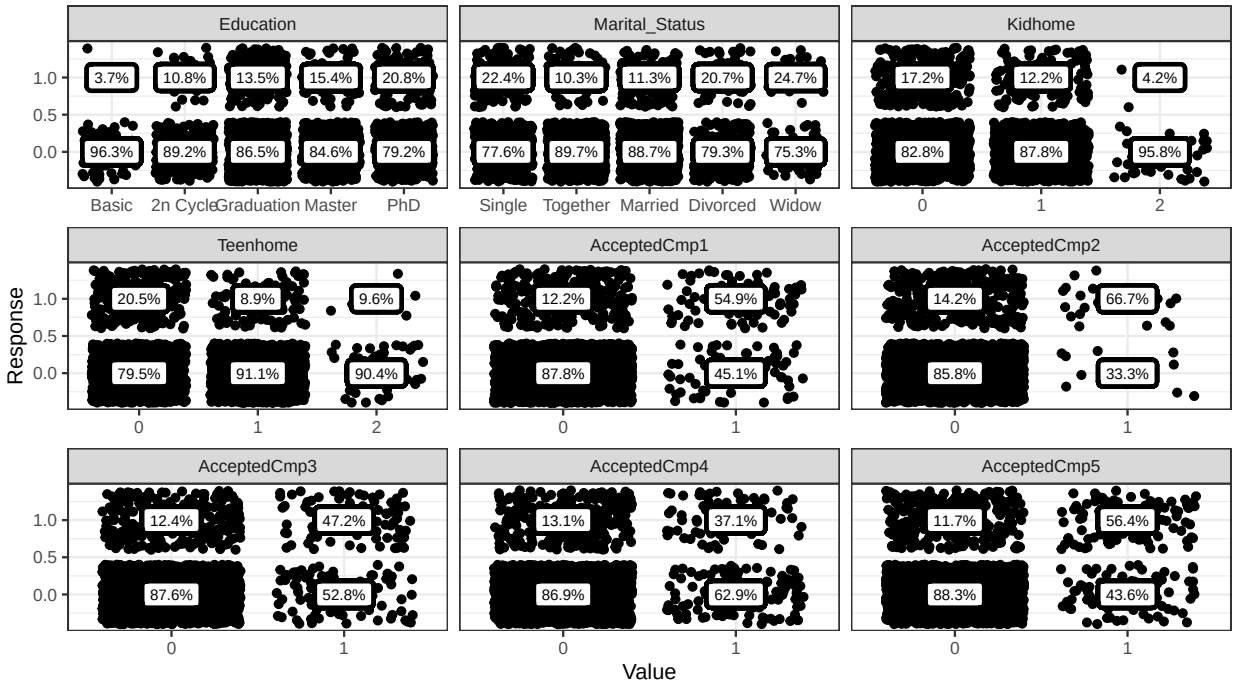

Figure 2: Disperção entre a Resposta e demais campos categóricos pós transformação

A Figura 3 apresenta de forma mais concreta as relações entre as variáveis. É possível notar que não há informações que são extremamente correlacionadas com a resposta. O grupo de variáveis das outras campanhas são as mais relacionadas com o resultado da última e vale ressaltar que elas apresentam uma correlação entre si. Além disso, o número de pedidos entre os diferentes canais e os gastos em diferentes produtos apresentaram uma correlação positiva. Ano de nascimento tem uma correlação quase nula, ou seja, a idade do cliente não aparenta ter influência.

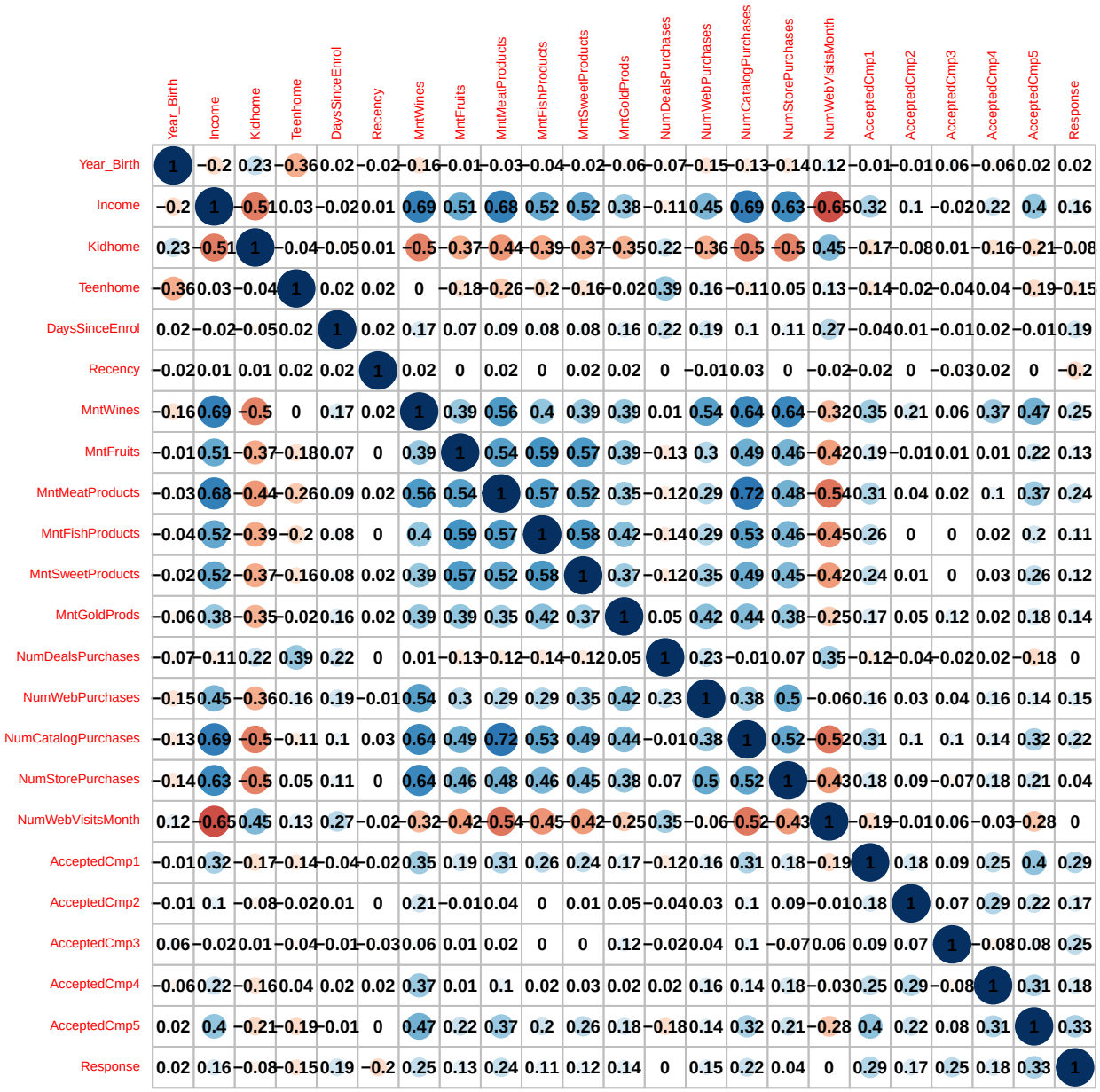

Figure 3: Correlação entre todas as variáveis

Modelo preditivo (Classificação)

Divisão dos dados

É de extrema importância começar pela etapa de divisão dos dados em treino e teste, assim será possível comparar os resultados ajustados ao treino com a aplicação do modelo em dados não treinados, assim conseguimos avaliar a performance do modelo (o quão generalista é) e evitar vieses da amostra inicial. Optou-se por separar um 70% para dados de treino e 30% para dados de teste. Além disso, os dados foram configurados em 6 conjuntos de variáveis diferentes, fazendo algumas alterações no formatado de algumas variáveis ou removendo outras afim de chegar em diferentes formatos, que podem se adequar melhores para

alguns modelos do que outros (há modelos que só aceitam variáveis numéricas, então foi criado uma versão em que é retirado as variáveis categóricas). Segue as diferentes versões:

1. Conjunto original.
2. Retirado os resultados das outras campanhas. O intuito é gerar modelos que não dependam dos resultados destas campanhas (pensando na performance a longo prazo).
3. Transformação das variáveis categóricas em "Dummys", com um campo para cada opção com a marcação de 1 se positivo e 0 c.c..
4. Retirado os resultados das outras campanhas e Transformação das variáveis categóricas em "Dummys".
5. Retirado de Ano de nascimento e Renda, por serem variáveis quantitativas de alta escala.
6. Transformação de Ano de nascimento e Renda para o formato de faixas. Ano de nascimento dividido em 5 e 5 anos e Renda para U$10mil em U$10mil.

Modelos Ajustados

Foram testados uma série de modelos de classificação para averiguar qual se adequava melhor ao problema, os quais foram:

- **Regressão Logística**: Modelar a probabilidade de um ponto de dados pertencer a uma das classes, utilizando uma função logística que varia de 0 a 1.
- **Logística com Lasso**: Regressão logística com o método de regularização Lasso, que adiciona um custo a cada coeficiente e que leva muitos deles a se tornarem exatamente 0.
- **KNN**: Classifica um novo ponto de dados com base nas classes dos K pontos mais próximos no conjunto de treinamento.
- **Naive Bayes**: Calcula a probabilidade de uma observação pertencer a uma classe específica com base nas características observadas.
- **Árvore de Decisão**: Estrutura de árvore que divide os dados em classes ou valores de destino com base nas características dos dados. A árvore é construída recursivamente, começando com o atributo mais informativo na raiz e dividindo os dados em ramos com base nos valores desse atributo.
- **Árvore de Decisão com Bagging**: Cria várias árvores de decisão durante o treinamento, cada uma usando um subconjunto aleatório dos dados de treinamento (bootstrapping), para ajuda a criar modelos diversificados, o que geralmente leva a um melhor desempenho.
- **Random Forest**: Similar ao Bagging na árvore de decisão, com a diferença de que para árvore criada em um subconjunto, também só é selecionado um subconjunto aleatório das variáveis, com o objetivo de retirar o efeito de características muito dominantes e que pode gerar uma melhora no desempenho.

Métrica de Avaliação

Para definir qual modelo obteve a melhor performance é preciso decidir qual critério de avaliação. As principais métricas em problemas de classificação estão a seguir:

- **Acurácia**: Percentual de acertos total. (VP + VN)/N
- **Sensibilidade**: Percentual de casos positivos classificados corretamente. VP/(VP + FN)

- **Especificidade**: Percentual de casos negativos classificados corretamente. VN/(FP + VN)
- **Precisão**: Percentual de verdadeiros positivos dentre todos os classificados como positivo. VP/(VP + FP)
- VP: verdadeiros positivos; FN: falsos negativos; FP: falsos positivos; VN: verdadeiros negativos; P: precisão; S: sensibilidade; N: total de elementos
- **F-score**: Uma média harmônica calculada com base na precisão e na revocação. 2x(PxS)/(P+S)

Como o objetivo é maximizar o lucro, é relevante avaliar a sensibilidade, para alcançar o máximo de clientes potenciais com a companha, e a precisão, para obter uma taxa de sucesso da campanha maior e consequentemente um maior lucro. Desta forma, é interessante a utlização de ambas, analogo ao "F-score". Porém, pode-se também utilzar os resultados da campanha piloto para melhorar a métrica de avaliação. De um total de 2240 cliente, 334 aderiram a campnha (15%). Os custos da campanha foi 6720MU e a receita gerada foi de 3674MU, resultando em um prejuízo -3046MU.

Afim de criar um racício lógico, suponhamos que 100 seja quantidade de clientes potenciais (que vão comprar o produto se estiverem na campanha). Seja também o Modelo X o escolhido para auxiliar na campanha e ele tenha uma sensibilidade de 40% e uma precisão de 80%. Pode-se então chegar a relação de que:

- O custo médio de campanha por cliente foi de 3MU (6720MU / 2240 clientes)
- A receita média gerada por cliente foi de 11MU (3674MU / 334 clientes)
- A Sensibilidade indica que a probabilidade do cliente estar corretamente classificado para aderir a campanha é de 40%.
- Dos 100 clientes potenciais, em média 40 receberiam a campanha. (100 * S)
- A Precisão indica que há 80% de chances de o cliente classificado como positivo ser de fato um cliente que irá comprar o produto.
- Assim, a cada 40 clientes (80%) que recebem a campanha corretamente, também há 10 clientes (20%) que recebem incorretamente e não iram comprar o produto (prejuízo).
- Desta forma, pode-se chegar ao cálculo que serão 80 clientes que receberam a campanha no total. (100*S/P)
- O custo da campanha seria de 240MU (80 clientes * 3MU/cliente) (100 * S * 3)
- A receita da campanha seria de 440MU (40 clientes * 11MU/cliente) (100 * S/P *11)
- E então, haveria um lucro do modelo de 200MU (440MU - 240MU)
- Em um cenário ideal, só seria mandado campanha para os 100 clientes potenciais, logo haveria um custo de 300MU e uma receita de 1100MU, ou seja, um lucro potencial de 800MU.
- Como o objetivo é maximizar o lucro, pode-se focar em aproximar lucro do modelo ao lucro potencial.
- Logo, teriamos a seguinte métrica final de porcentagem do lucro potencial: (S * 11 - S/P * 3) / (11 - 3)

Resultados

Após treinar os modelos para todas as versões da amostra, chegou-se ao resultado da tabela abaixo que apresenta os 10 modelos com melhor performance no amostra de teste, de acorodo com a métrica de lucro potencial. Como pode ser visto, o modelo de Regressão Logística na amostra 5 apresentou uma acurácia de 89.43% um Lucro Potencial de 43.62%, um lucro que nos testes chega a 3.49MU/cliente. Além disso, temos uma margem de lucro 141.35%, ou seja, a cada MU gasto tem-se de lucro 1.41MU (Isto seguindo as premissa do problema, com um gasto de 3MU/cliente e uma receita de 11MU/cliente).

Vale ressaltar também o Modelo Lasso da amostra 1 tem a melhor precisão. Neste modelo teria a melhor taxa de sucesso da campanha, chegando uma margem de lucro de 218.42%, i.e, a cada MU gasto teria 2.18MU de lucro. Porém, este é o de menor Sensibilidade e estariamos deixando de alcançar pela campanha cerca 65% dos clientes potenciais, em termos de lucro nominal total o modelo de Regressão Logística na amostra 5 continua sendo melhor.

Table 1: Resultados dos 10 modelos mais performaticos nas amostras de teste

Modelo	Versão Amostra	Acurácia	Precisão	Sensibilidade	Especificidade	Gastos	Receita	Lucro	Mergem de Lucro	Lucro_Potencial
Logistic	5	89.43%	65.82%	54.17%	95.31%	2.47MU/cl	5.96MU/cl	3.49MU/cl	141.35%	43.62%
Logistic	3	88.54%	67.50%	51.43%	95.41%	2.29MU/cl	5.66MU/cl	3.37MU/cl	147.50%	42.14%
Logistic	1	89.14%	62.34%	52.17%	95.00%	2.51MU/cl	5.74MU/cl	3.23MU/cl	128.57%	40.35%
Random Forest	1	89.29%	67.86%	41.30%	96.90%	1.83MU/cl	4.54MU/cl	2.72MU/cl	148.81%	33.97%
Bagging Tree	3	86.61%	59.26%	45.71%	94.18%	2.31MU/cl	5.03MU/cl	2.71MU/cl	117.28%	33.93%
Lasso	1	90.48%	86.84%	35.87%	99.14%	1.24MU/cl	3.95MU/cl	2.71MU/cl	218.42%	33.83%
Random Forest	3	87.95%	70.00%	40.00%	96.83%	1.71MU/cl	4.40MU/cl	2.69MU/cl	156.67%	33.57%
Logistic	6	88.54%	65.00%	41.05%	96.36%	1.89MU/cl	4.52MU/cl	2.62MU/cl	138.33%	32.76%
Bagging Tree	1	88.54%	61.90%	42.39%	95.86%	2.05MU/cl	4.66MU/cl	2.61MU/cl	126.98%	32.61%
Naive Bayes	5	84.38%	46.09%	55.21%	89.24%	3.59MU/cl	6.07MU/cl	2.48MU/cl	68.99%	30.99%

Observações

Também poderia ser testado modelos mais complexos, explorado melhor a iterações entre as variáveis, testado "tunnings" para melhorar a performance, redução de dimensão e entre outras abordagens, mas dado o problema e a solução apresentada, os resultados foram positivos e atenderam os requisitos pra um lucro maior para a campanha.

Como próximos passos, seria interessante aplicar um ou mais pilotos, utilizando agora a recomendação do modelo, para verificar se a performance com dados novos continua suficiente. Dado que o modelo esteja validado, o colocamos em produção para se "auto-alimentar" conforme a informação de novos clientes vão sendo atualizadas.