BLU365 @ FGV

Apresentação de Estudo de Caso - FGV

BLU 365

Agenda

Intro BLU365

- Breve contexto sobre a empresa
- Exemplo de como usamos Data Science no dia-a-dia

Estudo de caso

- Problema proposto
- Significado das variáveis
- Principais conceitos do dataset

Dicas de análise e modelagem

- Análise exploratória
- Tratamento das variáveis & feature engineering
- Modelagem e previsão

Plataforma BLU365

Plataforma BLU365

CLIQUE NO PARCEIRO ABAIXO
PARA NEGOCIAR SUA DÍVIDA ONLINE!

BLU em números

40MM

de CPFs com dívidas para negociação

150k

acordos por mês, em média

15MM

de SMSs enviados por mês e 20MM de e-mails

1,5MM

de acessos mensais à plataforma, em média

A BLU há 6 anos se dedica para transformar o cenário da gestão de dívidas através de competências únicas

USO DE

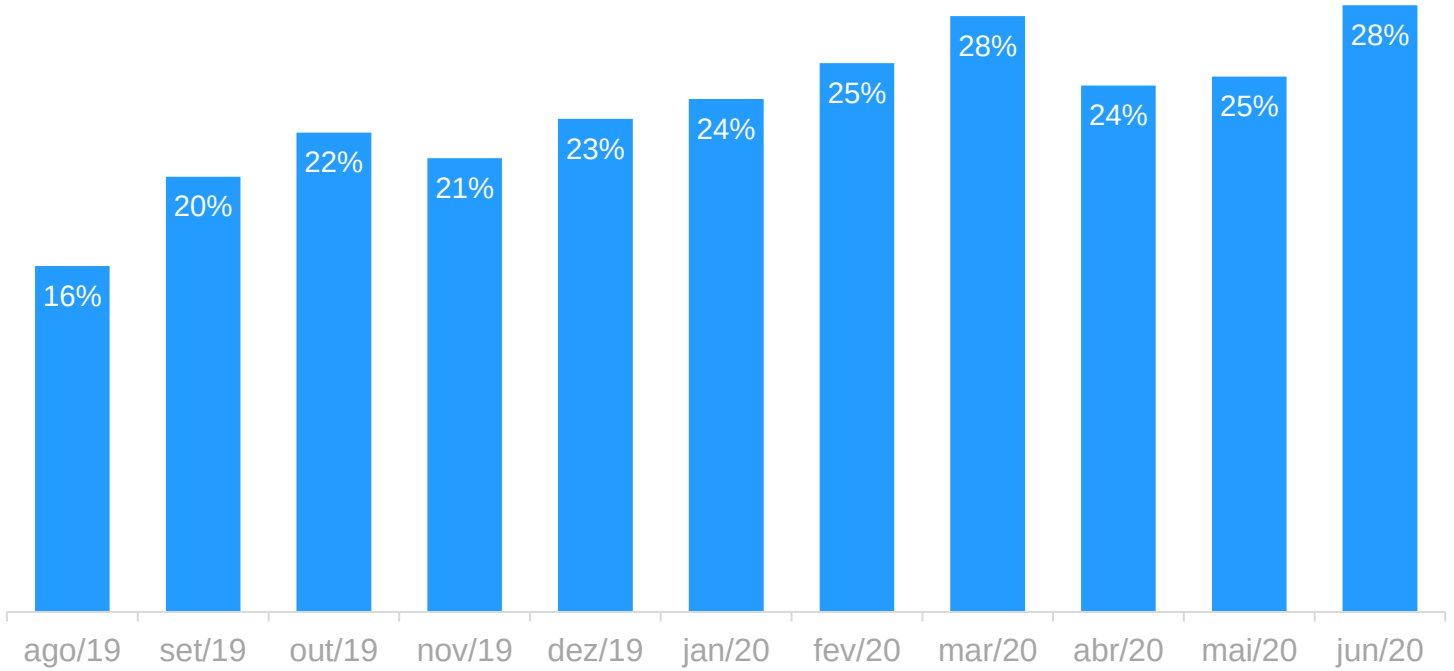
DATA-SCIENCE

NO NOSSO

DIA-A-DIA

Exemplo 1: Modelo de entrega de SMS

Percentual de **SMSs tarifados e não entregues** desde agosto/19.
A média histórica é 23,5%.



Dados Utilizados

Todos os SMSs enviados pela BLU desde agosto/19

~60 MM de SMSs

Além disso, foram filtrados apenas números para os quais já tentamos algum contato no histórico (90% da base):

Base final: ~54 MM de SMSs

Variáveis explicativas

- Histórico de tentativas por número
- % Entrega e não-entrega
- Dia da semana
- Histórico de entrega dos 3 últimos envios por número

+ combinação entre as variáveis, totalizando 25 features

Variável dependente

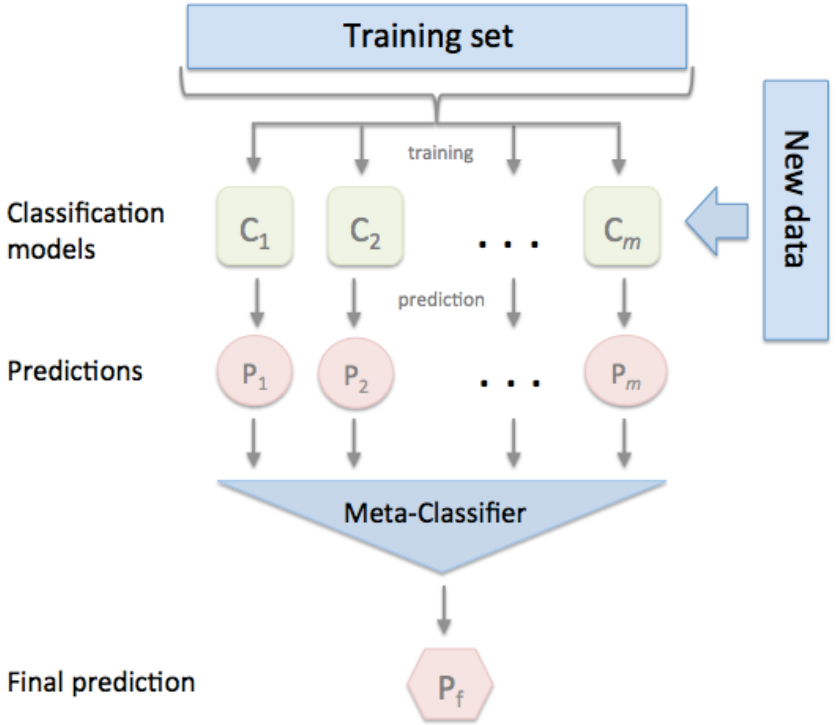
- Qual a probabilidade de entrega de um SMS após um disparo?

Modelagem

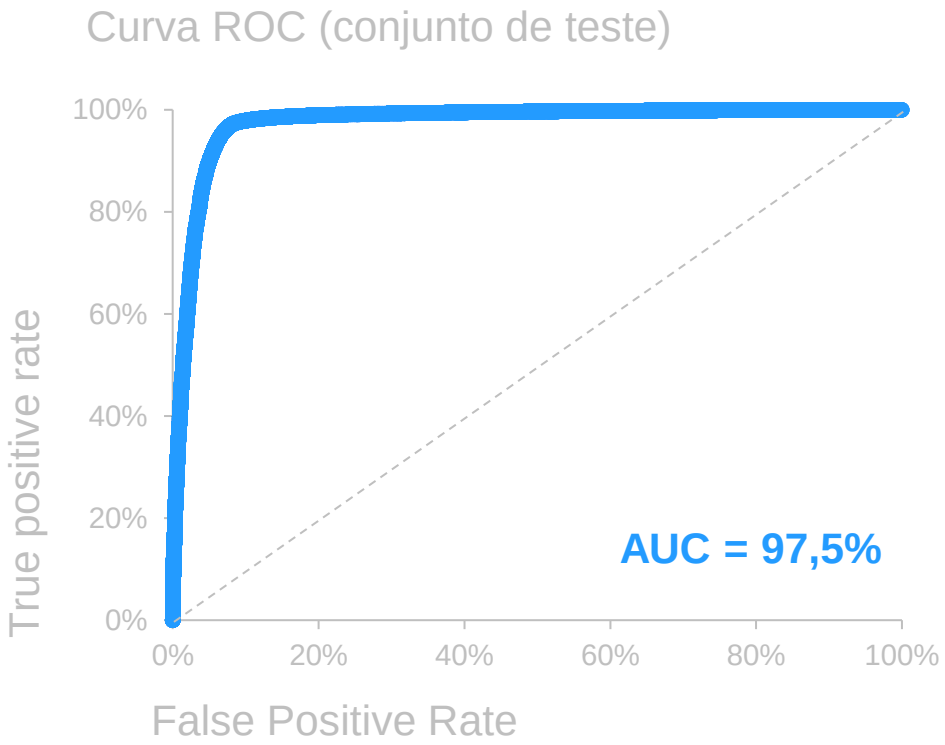
Foram treinados mais de 30 algoritmos diferentes em busca de performance.

Melhor modelo: **stacked ensemble**

A figura ao lado detalha a intuição por trás do stacked ensemble, que funciona como uma camada final que agrega diversos modelos para chegar a uma previsão final.



Performance do modelo



True positive rate

"de todos que têm a doença, para quantos o exame deu positivo?"

Dentre todos os SMSs não-entregues, quantos números foram corretamente apontados como inválidos?

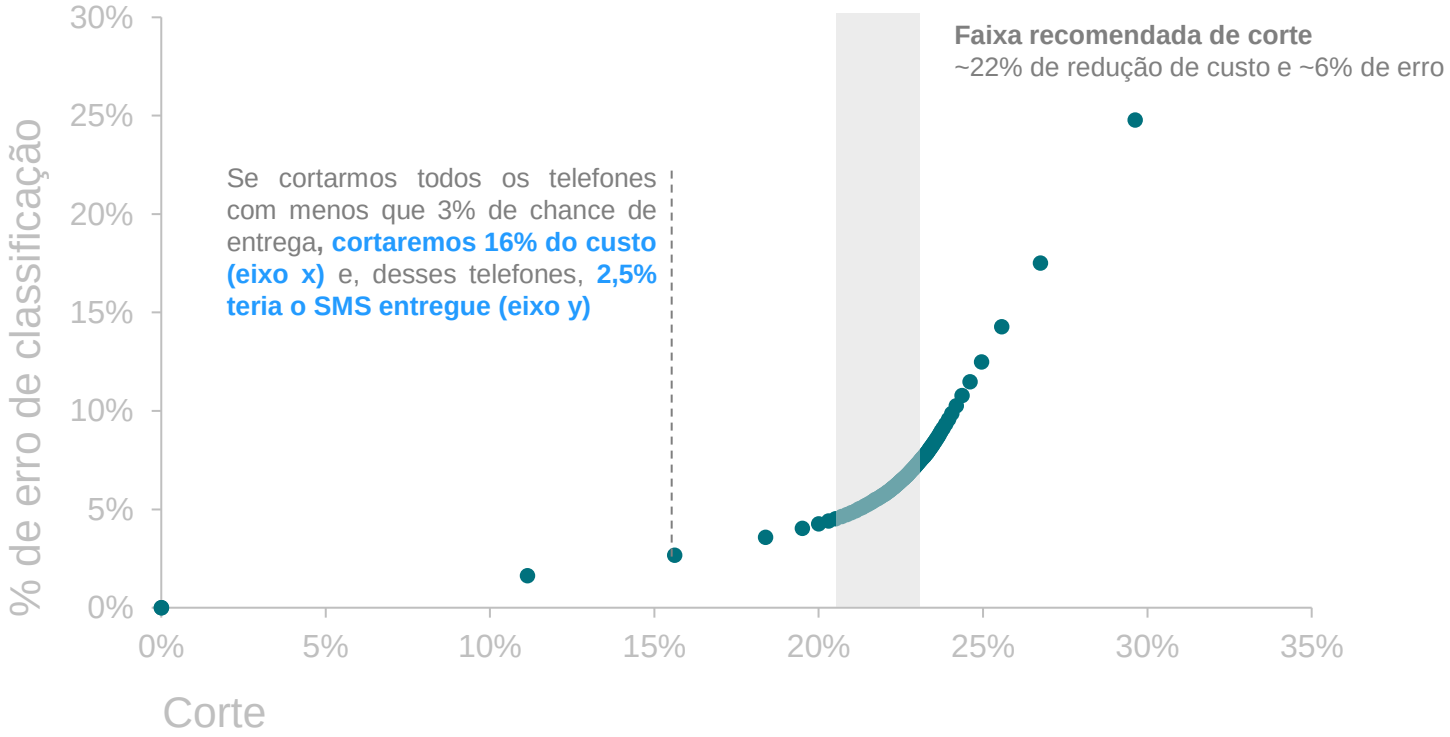
False positive rate

"de todos que não têm a doença, para quantos o exame deu positivo?"

Dentre todos os SMSs entregues, quantos números foram equivocadamente apontados como inválidos?

Trade-off

Trade off entre redução de custo e erro de classificação (SMSs não enviados que seriam entregues) para diferentes cortes de probabilidade



Exemplo 2: Modelo de Propensão de Acesso

Para uma lista de telefones aptos a negociar com a BLU em um dia, queremos saber: **qual é a probabilidade de conversão de cada pessoa?**

Com isso, poderemos ser muito mais eficientes se **enviarmos SMSs só para pessoas com maior chance de acessar** nosso site e fazer um acordo.

Dados Utilizados

SMSs enviados para diferentes pessoas em 4 dias específicos.

~530k SMSs

06/02/2020
40k envios

04/03/2020
240k envios

23/03/2020
213k envios
(primeiro dia da “crise”)

01/04/2020
40k envios

Variáveis explicativas

- Credor
- Data de envio
- Estado
- Valor da dívida & atraso
- Renda estimada
- Comportamento histórico:

Total de tentativas, último *step*, quantidade de acessos, % de entrega, etc.

Variável dependente

- Cliente acessou ou não o site nos 2 dias seguintes à campanha?

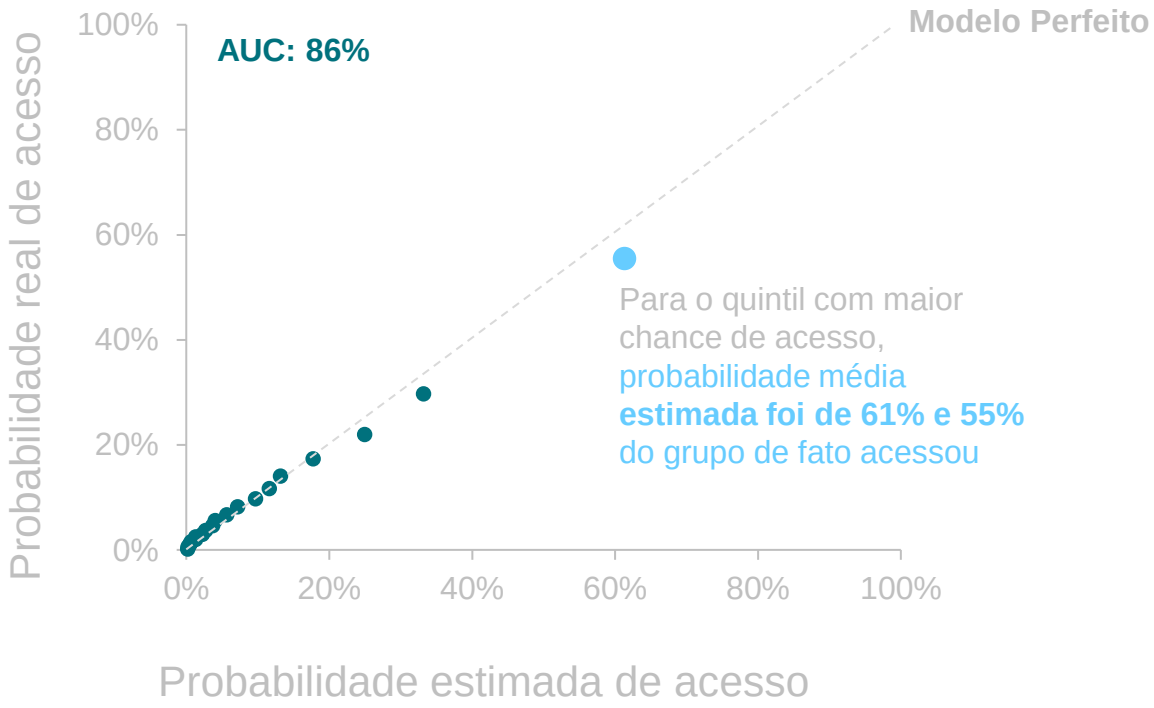
Modelagem

- Modelo de machine learning: Random Forest
- Divisão da base entre treino e teste. Conjunto de treino utilizado para estimar os parâmetros e teste para verificar performance.
- Output do modelo: probabilidade de acesso para cada indivíduo do conjunto de teste (que o modelo não viu para calcular os parâmetros).
- Tendo isso, ordena-se as pessoas em termos de chance de acesso e prioriza no “envio hipotético” aquelas com maior probabilidade de engajamento.

Resultados (conjunto de teste)

Probabilidade estimada vs real de acesso

média por quintil, conjunto de teste



Estudo de Caso

BLU 365

Problema proposto

Será que conseguimos **prever quais são os usuários com maior probabilidade acesso e acordo** após um SMS, com base em algumas variáveis comportamentais?

Variáveis explicativas

- DDD & Operadora
- Histórico de tentativas por número
- % Entrega e não-entrega
- Atraso, valor
- Histórico de acessos
- Variáveis demográficas (renda, idade, etc.)

Variável dependente

Qual a probabilidade de acesso e/ou acordo de cada pessoa após um disparo?

Base: histórico de ~700k SMSs enviados pela BLU entre abr e jun/20

- Cada linha do dataset representa um SMS enviado para um telefone diferente

Dicas & Perguntas-guia

Análise Exploratória

- Qual é a correlação entre as variáveis?
- Existe alguma variável que parece ser “a melhor” para prevermos acesso? E acordo?
- Para cada uma das variáveis, qual é o percentual de dados “missing”?

Dicas & Perguntas-guia

Tratamento das variáveis

- Qual a melhor forma de tratar os dados faltantes? Excluir linhas? Preencher com base em alguma estatística (média, mediana, moda, etc.)? Isso muda se a variável é categórica ou numérica?

- Para as variáveis categóricas, como fazer para que elas possam ser ingeridas pelo modelo?

- Existe alguma inconsistência entre as variáveis (por exemplo, data de atraso maior que a data de disparo)? O que fazer nesse caso?

Dicas & Perguntas-guia

Feature Engineering

- Como criar variáveis de dias de atraso e sazonalidade (ex: dia-da-semana, quinzena do mês, 5º dia útil, etc.)?
- Será que podemos usar o DDD para criar outras variáveis (cruzando com outras fontes)?
- Para quais variáveis faz sentido criar interações/não-linearidades (ex: valor^2)

Dicas & Perguntas-guia

Modelagem & Previsão

- Como fazer para escolher quais variáveis vão entrar no modelo?
- Qual é o percentual da base que deve ser separado entre treino e teste?
- Para esse problema específico, qual é a melhor métrica de avaliação para saber se temos um modelo bom ou não?
- As probabilidades que estão saindo do modelo estão “calibradas”?
- Considerando que o percentual de acessos e acordos é apenas uma pequena fração do total de envios, será que se “balancearmos” a amostra para treino conseguimos resultados melhores?

Obrigado!

BLU365

João Henrique Netto
joao.netto@blu365.com.br