Judge Panderson: Análise de leads com NLP e Detecção de Anomalias

Patrícia Pampanelli
Data Scientist
patricia.pampanelli@grupozap.com

Apresentação

Patrícia Pampanelli
Data Scientist @ Grupo ZAP
patricia.pampanelli@grupozap.com

Quem é o grupo ZAP ?

suahouse

O melhor e mais eficiente sistema de gestão, com CRM, plataforma de BI, controle de leads e muito mais.

data ZAP

Empresa de inteligência imobiliária que auxilia no processo de decisão para construção, investimento, financiamento e negociação.

fipe ZAP

Principal índice de preço de imóveis, de abrangência nacional. Parceria entre Fipe e o Grupo ZAP.

zap

O portal especialista em imóveis.

VivaReal

O maior portal de imóveis do Brasil.

conecta imobi

A plataforma que coloca o mercado imobiliário no centro de tudo.

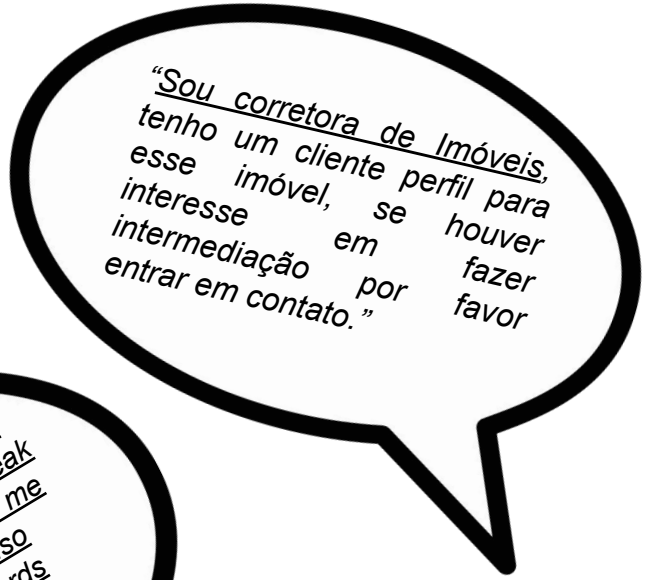
GEOIMOVEL

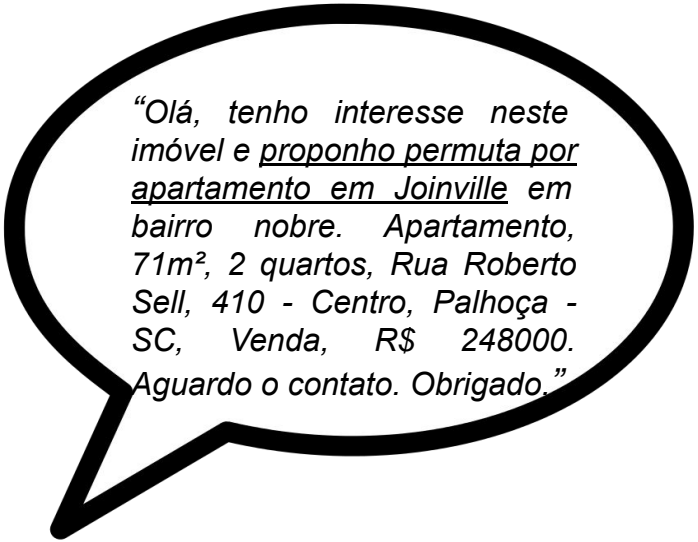
Plataforma de soluções e pesquisas imobiliárias com as mais diferenciadas ferramentas e instrumentos de análise mercadológica para as empresas que atuam no setor imobiliário

Motivação

Mais de 100.000 mensagens de texto por dia!


I read your advert and i am interested in the purchase. Can we discuss more on it? I can be reach on this ID: ***@outlook.com I speak only English Please get back to me with your Private E-Mail ID: Also your last Price. Thanks Regards Maria Olá, tenho interesse neste imóvel: Apartamento, 85m², 2 quartos, ...


"Sou corretora de Imóveis, tenho um cliente perfil para esse imóvel, se houver interesse em fazer intermediação por favor entrar em contato."


"Olá, tenho interesse neste imóvel e proponho permuta por apartamento em Joinville em bairro nobre. Apartamento, 71m², 2 quartos, Rua Roberto Sell, 410 - Centro, Palhoça - SC, Venda, R$ 248000. Aguardo o contato. Obrigado."

grupo ZAP

Análise de mensagens de texto

Pré-processamento

Pré-processamento

Leads de Mensagem

Extração de features: frequência que o usuário gera os leads (hora, dia, semana)

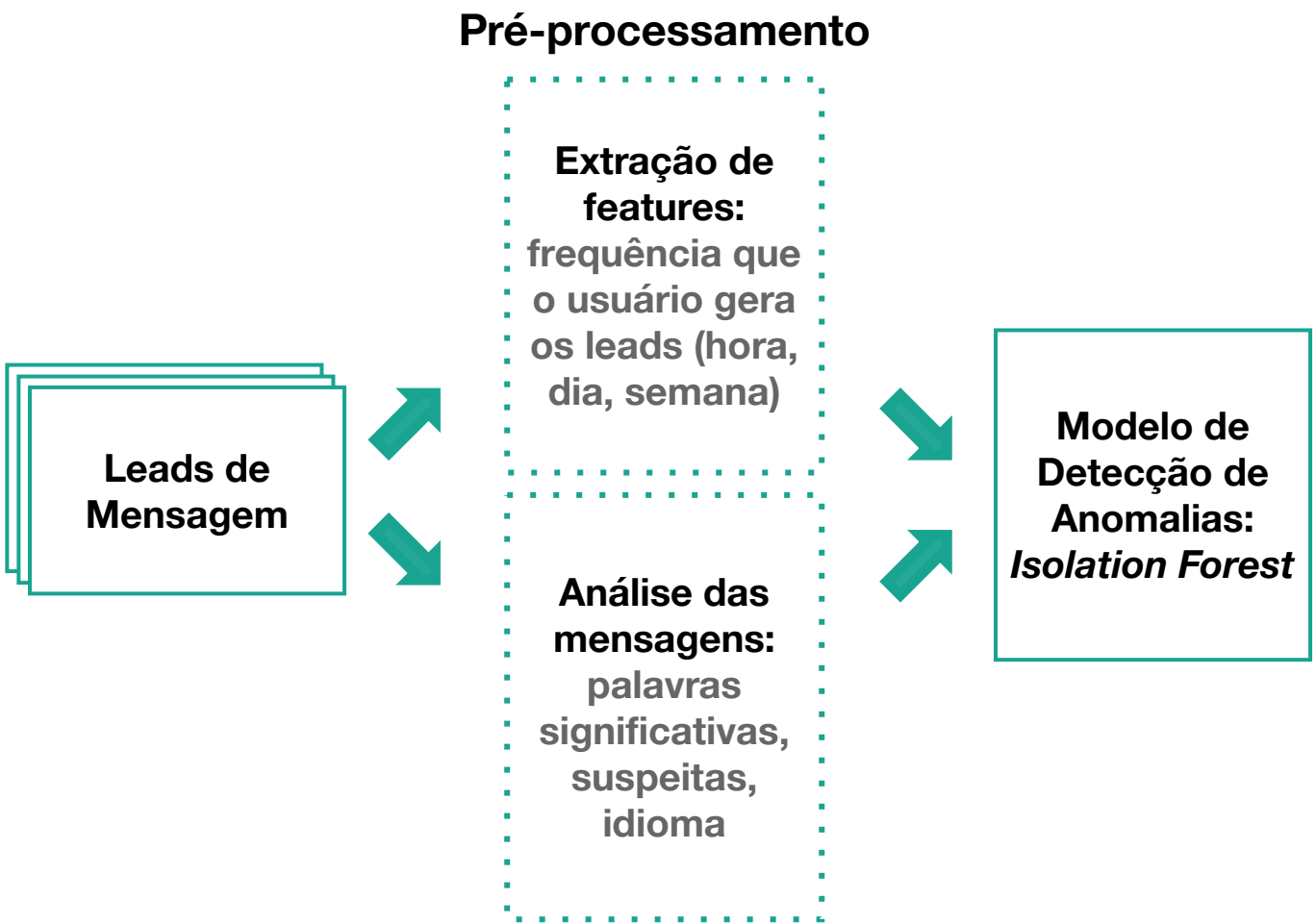
Pré-processamento

Leads de Mensagem

Extração de features: frequência que o usuário gera os leads (hora, dia, semana)

Análise das mensagens: palavras significativas, suspeitas, idioma

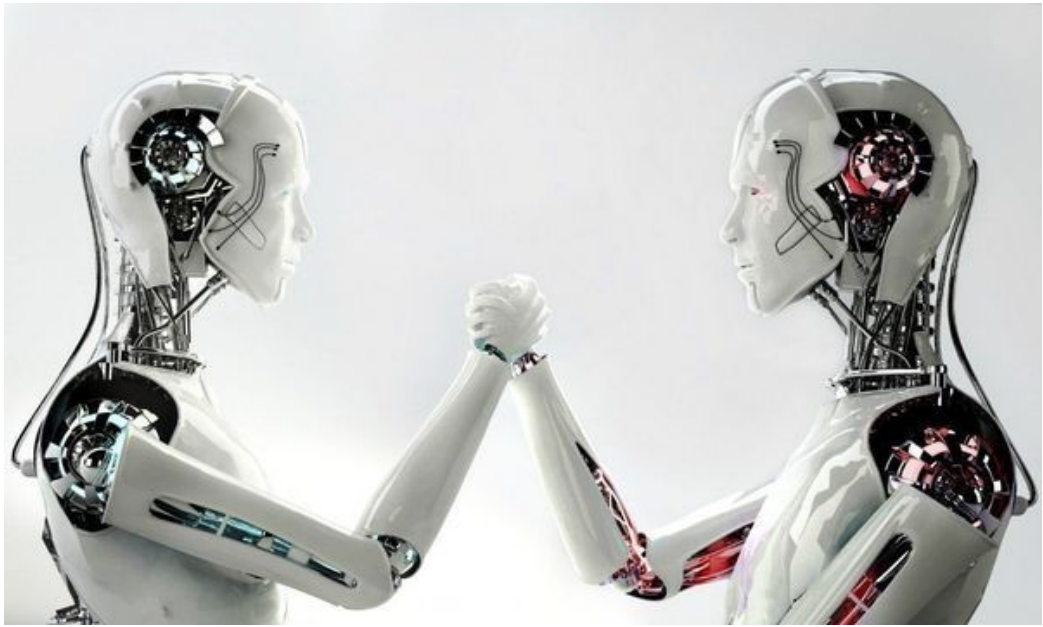

o
de
olá
imóvel
seu
mais
no
gost
inform
que
aguard
sobr
zap
sao
sp
ter
cas
client
retorn
contact
contat
email
aluguel
obrig
vil
apart
por
par
quitinet
encontr
paul
padrã
vend
este
obrigado
imóv
ser
telefon
condomíni
whatsapp
brooklin
marian
comercialsal
desej
pinheir
aceit
jardim
morumb
butantã
perdiz
santan
flat
itaim
valor
tatuap
trabalh
boa
vist
conjunt
liberdad
corretor
endereç
ipirang
leopoldin
locação
alto
madalen
mooc
comercial
rua
moem
consol
tenh
favor
paulist
está
precis
apto
bel
noit
alug
na
sã
do
tard
pod
visit
sant
agend
interess
uma
tem
possível
inter
aven
sab
cobertur
aclim
qual
interlag
estou
centr
fregues
com
da
ou
um
fund
cidad
rio
itaqu
marc
barr
fot
prudent
lojasalã
camp
jd
saúd
bom
higienópol
bib
meu
permut
cozinh
jabaqu
tucuruv
localiz
clementin
cambuc
autoriz
paraís
andar
verd
entrar
noss
mil
olímp
armári
dia
matild
imovel
sou
unidad
piritub
cerqueir
penh
faz
grat
vag
confirm
anunc
pagament
mobili


Pré-processamento
Leads de Mensagem
Extração de features: frequência que o usuário gera os leads (hora, dia, semana)
Análise das mensagens: palavras significativas, suspeitas, idioma
Modelo de Detecção de Anomalias: Isolation Forest

Detecção de anomalias com Isolation Forest

Isolation Forest

- Método *ensemble* que funciona "isolando" as observações anômalas
- Métodos ensembles: conjunto de classificadores treinados para os quais os outputs são combinados

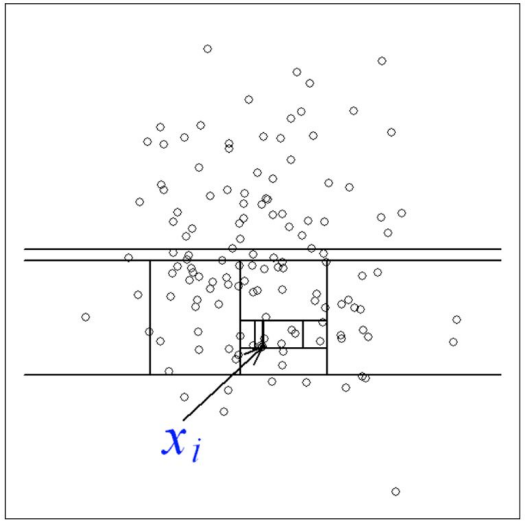


Isolation Forest

- Tem como base o método de árvores de decisão
- Define explicitamente quais são as amostras anômalas

Isolation Forest

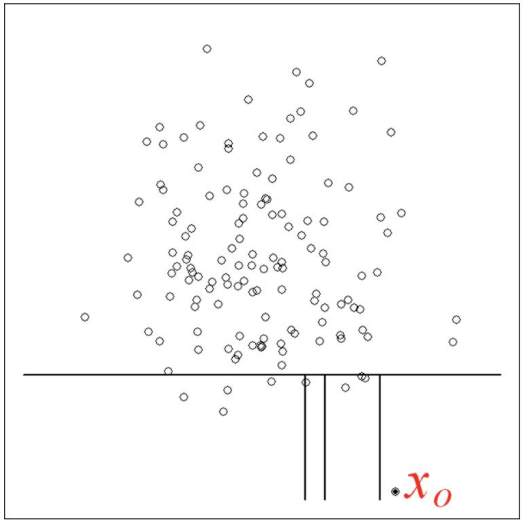
- **Amostras consideradas normais precisam de muitas partições para serem identificadas**

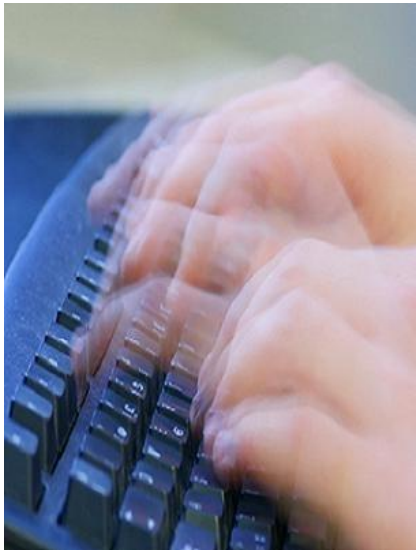


https://towardsdatascience.com/outlier-detection-with-isolation-forest-3d190448d45e

Isolation Forest

- **Amostras anômalas são identificadas por nós mais próximos da raiz da árvore**

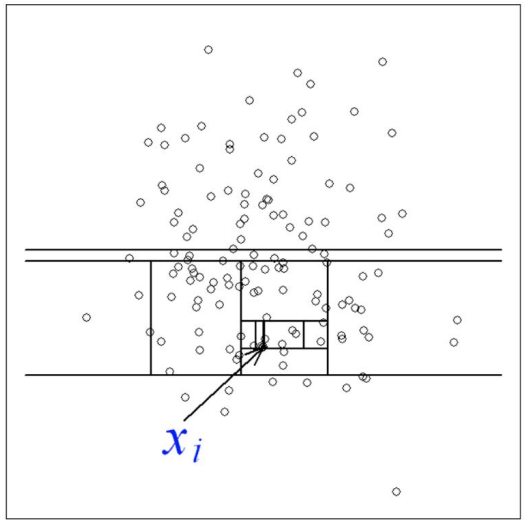




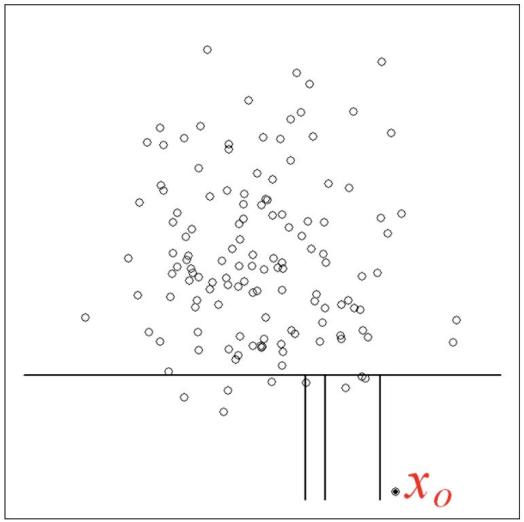
https://towardsdatascience.com/outlier-detection-with-isolation-forest-3d190448d45e

Isolation Forest

Score ≃ 0: normais



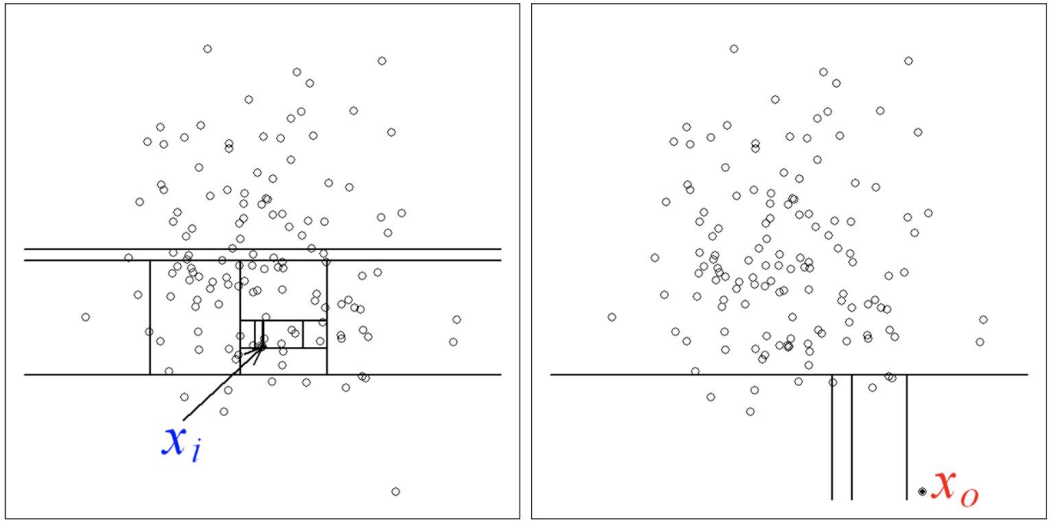
Score ≃ 1: anomalias



https://towardsdatascience.com/outlier-detection-with-isolation-forest-3d190448d45e

Isolation Forest

- Parâmetro de contaminação:
  - entre 0 e 0.5
  - representa a porcentagem de amostras que possivelmente são *outliers*



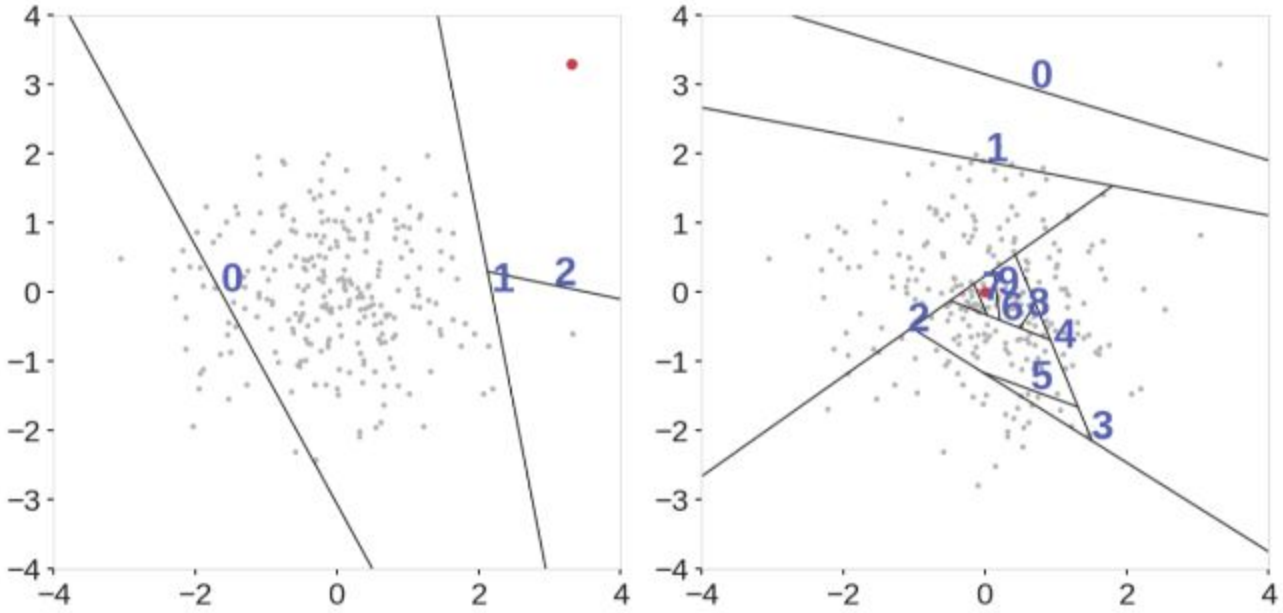
https://towardsdatascience.com/outlier-detection-with-isolation-forest-3d190448d45e

```
from sklearn.ensemble import IsolationForest
```

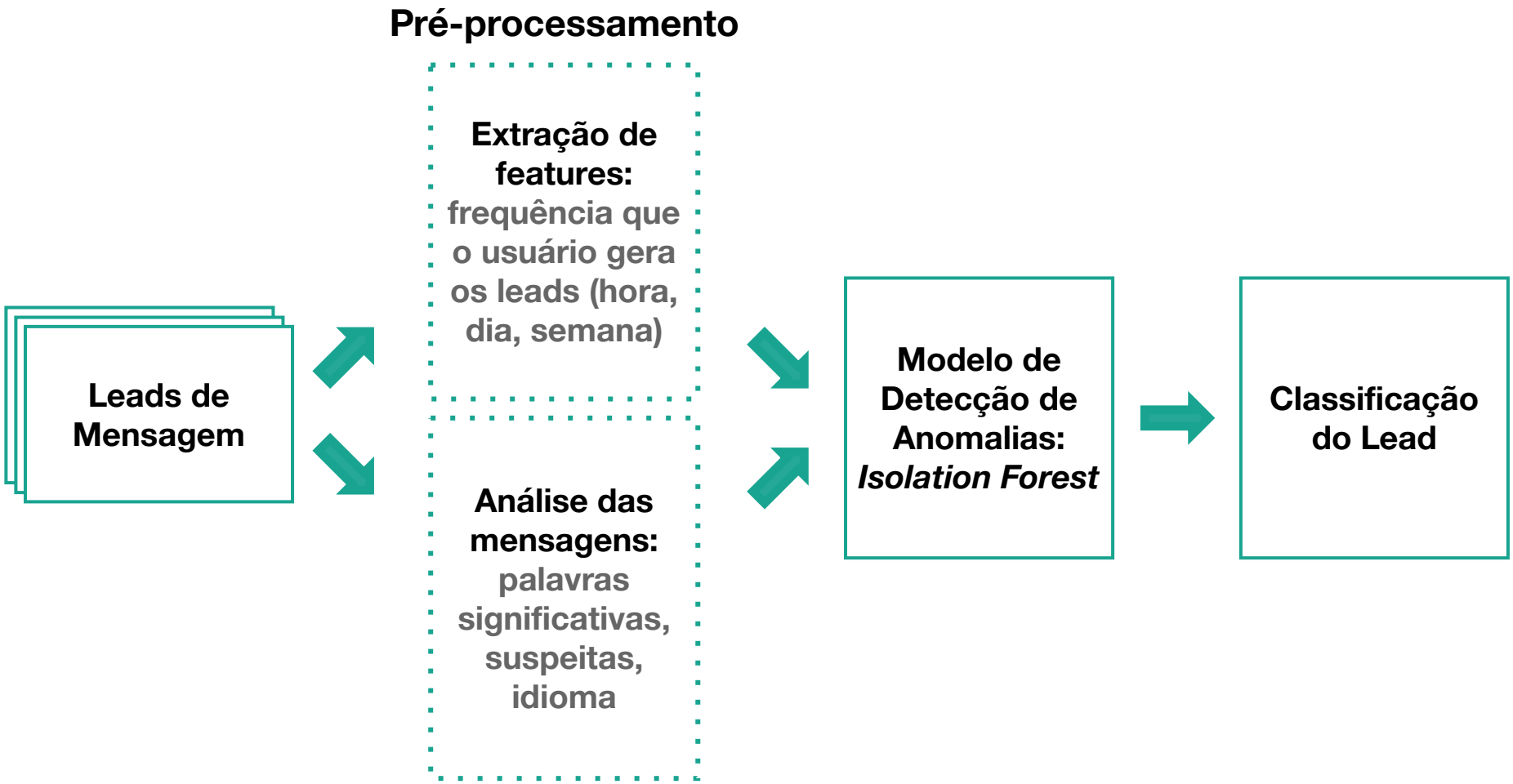
```
from sklearn.ensemble import IsolationForest

model = IsolationForest(n_estimat=300, max_features=0.7, contamination=0.015)
```

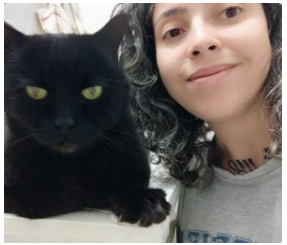
Extended Isolation Forest



https://towardsdatascience.com/outlier-detection-with-isolation-forest-3d190448d45e


Pré-processamento
Leads de Mensagem
Extração de features: frequência que o usuário gera os leads (hora, dia, semana)
Análise das mensagens: palavras significativas, suspeitas, idioma
Modelo de Detecção de Anomalias: Isolation Forest
Classificação do Lead

Ninguém faz nada sozinho...



Ana Carolina

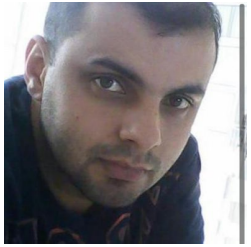
Audrey Ferreira

Daniel Vainsencher

Vinícius Silva

Fábio Zacarias

Guilherme Bertola



Renan Barbioni

Classificação de leads

- **BLOCKED:** leads que contém palavras ou links proibidos
- **AGENT:** leads de corretor
- **NOT_IN_PORTUGUESE:** leads que não estão em português
- **UNDEFINED:** menos de X leads do mesmo usuário e em menos de X minutos
- **SUSPECT:** leads gerados em uma frequência que o modelo identifica como sendo bot;
- **EXCHANGE:** leads gerados com a intenção de fazer troca/permuta de imóveis
- **NORMAL:** leads considerados normais
- **ETC**...

Obrigada!

Estamos contratando!
jobs.kenoby.com/grupozap