UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

ESPECIALIZAÇÃO EM DATA SCIENCE E SUAS APLICAÇÕES

RAISSA PAVAN MACHADO

MODELAGEM PREDITIVA PARA IDENTIFICAÇÃO DA PROBABILIDADE DE REATIVAÇÃO DE REVENDEDORES EM MULTINACIONAL DO SETOR DE COSMÉTICOS

CURITIBA

2021

RAISSA PAVAN MACHADO

MODELAGEM PREDITIVA PARA IDENTIFICAÇÃO DA PROBABILIDADE DE REATIVAÇÃO DE REVENDEDORES EM MULTINACIONAL DO SETOR COSMÉTICO

Predictive Model To Identify The Probability Of Reactivation Of Door To Door Salesperson In Multinational Cosmetic Industry

Trabalho de conclusão de curso apresentado como requisito parcial à obtenção do título de pós-graduado em Ciência de Dados, do Departamento Acadêmico de Informática (DAINF), da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Msc Matheus Garibalde Soares de Lima

CURITIBA

2021

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Câmpus Ponta Grossa

RAISSA PAVAN MACHADO

MODELAGEM PREDITIVA PARA IDENTIFICAÇÃO DA PROBABILIDADE DE REATIVAÇÃO DE REVENDEDORES EM MULTINACIONAL DO SETOR COSMÉTICO

Natureza do trabalho: Trabalho de pesquisa apresentado como requisito para obtenção do título de especialista em Ciência de Dados da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Ciência da Computação.

Data de aprovação: 21/01/2021

Prof Matheus Garibalde Soares de Lima, Mestrado - Universidade Tecnológica Federal do Paraná

Profa. Rita Cristina G. Berardi, Doutorado - Pontifícia Universidade Católica do Rio de Janeiro

Profa. Marcelo de Oliveira Rosa, Doutorado - Universidade de São Paulo

Documento gerado pelo Sistema Acadêmico da UTFPR a partir dos dados da Ata de Defesa em 21/07/2021.

Dedico este trabalho à todos que acreditam na educação como forma de transformação social

AGRADECIMENTOS

Agradeço aos professores do curso que auxiliaram a solidificar boa parte dos conteúdos aqui utilizados e em especial ao meu orientador, Matheus, pelo apoio e paciência nas últimas semanas.

Agradeço ainda aos meus colegas que contribuíram muito para que essa fase da minha vida fosse agradável, mesmo com os desafios.

RESUMO

Os representantes comerciais de produtos cosméticos são uma das formas mais eficientes que as empresas do ramo encontraram para difundirem seus produtos, contudo, encontrar, reter e reativar revendedores para que continuem vendendo ainda é um desafio. Ao longo deste trabalho foram construídos modelos de *machine learning* em linguagem Python, baseado em três tipos de algoritmos, para se prever com maior segurança quem são os representantes que realmente tem chance de continuar revendendo produtos de cosméticos mesmo após algum tempo sem novos pedidos.

Palavras-chave: Predição. Python. Cosméticos. Ciência de dados. Venda Direta.

ABSTRACT

Door to door salesmen are one of the most efficient ways that cosmetic companies have found to spread their products, however, finding, retaining and reactivating salespeople so that they continue to sell is still a challenge. Throughout this work, machine learning models were built in Python language, based on three types of algorithms, in order to be able to predict with greater certainty who are the representatives who really have a chance to continue reselling cosmetic products even after some time without new orders.

Keywords: Predictions. Python. Cosmetics. Data Science. Door to door sales.

LISTA DE TABELAS

LISTA DE ILUSTRAÇÕES

SUMÁRIO

1 INTRODUÇÃO

Este trabalho foi construído com o objetivo de identificar oportunidades para reativação de revendedores que já fizeram parte da força de vendas de uma empresa multinacional do ramo de cosméticos como parte do canal de venda direta, através da construção de um modelo de aprendizado de máquina para previsão de revendedores com maior probabilidade de realização de novos pedidos.

1.1 CONTEXTUALIZAÇÃO

A revenda de produtos cosméticos é uma forma de trabalho bastante difundida no Brasil, o que permitiu que a empresa estudada conquistasse uma base significativa de pessoas dispostas a atuarem neste canal. Conforme os anos foram se passando, naturalmente uma parte desta base deixa de realizar novos pedidos e após algum tempo têm seus cadastros desativados, essas revendedoras passam a ser chamadas de cessadas, simultaneamente, novas pessoas se cadastraram ou revendedoras desativadas decidem retornar à base realizando novos pedidos. Essa rotatividade permite atuar junto à base de cessadas e tentar recuperar parte da base, mas para isso ser feito de forma otimizada, construiu-se um modelo preditivo que indica qual a probabilidade de uma pessoa da base inativa realizar um novo pedido.

A atuação com as pessoas que já estão na base, porém não realizam mais pedidos é essencial, pois é inviável para uma empresa sempre precisar de um volume muito grande de novos profissionais. O processo de convencimento pode ser caro e ineficiente, e um *turnover* alto, mesmo entre revendedores autônomos pode se tornar um indicativo de que melhorias no processo de manutenção dessas pessoas precisaria ser desenvolvido. Precisa-se considerar ainda outras situações em se tratando da manutenção da base de representantes, a saturação de uma região com mais vendedoras do que o necessário pode levar representantes a ficarem sem mercado e saírem da atividade, novas pessoas podem não se adaptar ao formato de trabalho e não obter os resultados esperados, entre tantas outras situações com potencial para dificultar a expansão da base de revendedoras. Assim, identificar oportunidades com pessoas que já conhecem e trabalharam com a marca tende a ser uma via cada vez mais importante para a manutenção e expansão dos canais de venda direta, não somente no setor cosmético.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

Criação de um modelo para previsão binaria da probabilidade de reativação de um revendedor cessado.

1.2.2 Objetivos Específicos

O modelo apresentado foi construído com os seguintes objetivos específicos:

- Aplicação de um modelo de *machine learning* para previsão binária da probabilidade de reativação de um revendedor cessado;
- A solução precisa ser aplicável há um pipeline de produtivo em ambiente *cloud*;
- Compreender os padrões fundamentais sobre a base de dados estudada, de forma a organizar as primeiras informações conceituais sobre revendedores que não atuam mais ativamente com revenda de produtos da empresa.

2 REVISÃO BIBLIOGRÁFICA

As técnicas de *Machine Learning* estão cada vez mais difundidas e sendo amplamente utilizadas nos serviços ao consumidor. Os sites cada vez mais buscam personalizar a experiência dos usuários através de recomendações, páginas e condições exclusivas, adaptação de conteúdos, entre muitas outras.

Algumas das técnicas empregadas para garantir estas personalizações são complexas, como as redes neurais profundas e os algoritmos *transformers*, porém em muitos casos essa complexidade não é necessária. Abhilasha (2018) mostra isso no seu paper sobre análise de sentimentos, onde ele utiliza *Naive Bayes* e *Support Vector Machine* para classificar sentimentos em comentários e depois analisa os mesmos com uma regressão logística.

Neste trabalho o foco é em um problema que não é visível aos clientes, mas que afeta muito as empresas, o *churn* de revendedores, porém partindo do ponto de que muitos deles são *churns* "temporários", retornando a fazer pedidos algum tempo após a inativação ou ao menos têm maior probabilidade de realizar novos

pedidos, considerando o histórico com a empresa, o que é um tema ainda pouco difundido no contexto de Ciência de Dados por ser específico de algumas indústrias, quando consideramos trabalhos sobre o previsão de *churn* especificamente, encontrasse alguns conteúdos de interesse que podem ser fornecer uma base inicial. Ullah (2019) construiu um modelo focado em *churn* de clientes no setor telecom utilizando o modelo de Random Forest e conseguiu um F1-*score* de mais de 80% nas previsões, dentro da indústria estudada um modelo semelhante é aplicado e os resultados da previsão ficam próximas aos 70% de acurácia, o que é um problema diretamente relacionado ao estudado neste caso. Podemos ainda visualizar o caso proposto por Spanoudes (2017), no qual ele conseguiu construir uma arquitetura toda para um modelo de rede neural para previsão de *churn* de forma generalista, indicando que mesmo sob diferentes condições, o problema pode ser resolvido de formas similares e apresentar bons resultados.

Christodoulou (2019) nos trás ainda uma revisão sistemática sobre trabalhos de predição na área da saúde, que envolvem previsão de diagnósticos e prognósticos, ele compara o uso de regressão logística versus outros modelos de *machine learning*, como árvores de decisão, *support vector machines* e redes neurais. Os resultados da revisão indicam que não há evidência de melhor performance dos demais modelos sobre a regressão logística, porém também indica que parte dos trabalhos sofre com problemas de validação das previsões ou possuem base de dados limitada, tornando a comparação entre os trabalhos algo ainda complicado.

Dados os estudos acima, podemos entender a necessidade de se expandir os estudos em previsões de áreas ainda não desvendadas da empresa também como a aplicação de algoritmos tidos como clássicos podem agregar valor para o problema colocado, especialmente quando envolvem dados tabulares e bases relacionais.

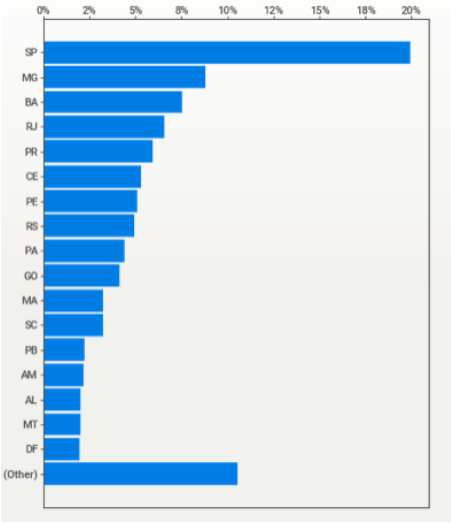
3 ANÁLISE DAS BASES

O ciclo de vida das revendedoras atuando com a venda de cosméticos pode seguir diversos rumos. Para entender a base e definir o problema que o modelo se propõem a resolver utilizamos alguns conceitos chave:

- Revendedor ativo: Realizou ao menos um pedido nas últimas 21 semanas;
- Revendedor cessado: Cadastro está inativo pois ela passou mais de 21 semanas em realizar um novo pedido;
- Revendedor reativado: A pessoa passou pelo processo de inativação do cadastro, mas retornou para fazer um novo pedido após esse tempo.

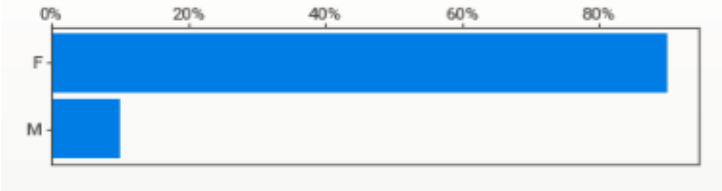
A base utilizada para a construção desse modelo foi composta por informações cadastrais e transacionais do conjunto de revendedores que compuseram a base de venda direta da empresa de cosméticos abordada neste estudo nas categorias cessada e reativadas. O perfil das pessoas ainda ativas foi consultado apenas para identificação de padrões em relação ao alvo do modelo. Os dados utilizados para avaliar o histórico compõem um período de seis anos, onde temos um perfil de revendedores majoritariamente feminino, distribuídas por todo o Brasil de forma correspondente ao tamanho dos núcleos populacionais e com destaque para mulheres adultas entre 25 e 50 anos.

Figura 1 - Percentual de revendedores em cada estado



Fonte: autoria própria

Figura 2 - Percentual das revendedoras por gênero



Fonte: autoria própria

Dentre as premissas utilizadas para dar seguimento a análise podemos destacar dois pontos:

- Revendedoras que cessaram há menos tempo vão apresentar perfis de dados mais confiáveis e devem ter maior chance de responder a estímulos para retorno.
- Perfis mais engajados com a marca durante o período ativo tendem a ser bons candidatos a reativação.

A partir desses pontos conseguimos identificar informações que poderiam ser relevantes para o modelo. Além de dados cadastrais, organizou-se dados referentes a quantidade e variedade das vendas, valores dos pedidos realizados (considerando que um pedido pode incluir vários produtos), tempo de cessamento, período em que o último pedido foi realizado, tempo como ativas e informações de crédito.

Comparando os três grupos selecionados para a análise exploratória, os revendedores cessados, reativados e ativos, pode se verificar que o perfil de compras dos revendedores reativados se assemelha mais ao perfil dos ainda ativos do que daqueles que cessaram e nunca retornaram, isso considerando valores e quantidades. Comportamentos particulares em relação ao período de compras desses revendedores também puderam ser identificados, demonstrando tendências sazonais de uma parte da base, o que desafia o conceito de o que é um revendedor ativo, visto que a frequência de compras é o único fator considerado. Os dados específicos dessa parte da análise foram usados no treinamento do modelo, porém não foram detalhados pois apresentaram informações que podem influenciar as estratégias de campanhas da empresa, e por isso os padrões não serão demonstrados.

3.1 CONSTRUÇÃO DAS FEATURES

O problema proposto neste trabalho ainda não havia sido analisado pela área de dados dentro da empresa, o que significa que não havia bases organizadas sobre este tema específico. Tomando como base dados relevantes no contexto geral de Venda Direta ou VD, utilizou-se principalmente bases com informações cadastrais, detalhes dos pedidos realizados e produtos.

3.2 DESCRIÇÃO DAS FEATURES

O processo de construção das *features* finais ocorreu diretamente no banco de dados. Todas as tabelas tiveram por base o agrupamento por código de revendedor, sendo calculados diferentes variáveis em cada contexto necessário. Inicialmente, mais de 50 variáveis foram analisadas, porém a lista final testada nos modelos resultou em 30 variáveis, listadas abaixo:

Tabela 1 - Variáveis construídas para o modelo de previsão

(continua)

Nome Variável	Tipo	Descrição
COD_REVENDEDOR	STRING	Código único de identificação do revendedor
IDADE_ATUAL	INT	Idade atual da revendedora
ANO_CADASTRO	STRING	Ano em que foi realizado o cadastro como revendedor
DES_GENERO_REVENDEDOR	STRING	Identificação fornecida pelo revendedor como sendo do sexo feminino ou masculino
COD_LOGRADOURO_UF_REVENDEDOR	STRING	Estado da federação do revendedor
DES_BLOQUEIO_CADASTRO_REVENDEDOR	STRING	Informação sobre eventuais problemas no cadastro
DES_ORIGEM_CADASTRO_REVENDEDOR	STRING	Forma como o revendedor se cadastrou para trabalhar com venda direta
DES_CLUSTER_SEGMENTACAO	STRING	Segmentação de revendedoras considerando o valor acumulado de vendas
QT_LIMITE_CREDITO	FLOAT	Limite de crédito disponibilizado para a revendedora
QT_SALDO_CREDITO	FLOAT	Saldo de crédito disponível no momento da consulta das informações
PEDIDOS	INT	Total de pedidos de produtos realizados durante a vida útil do revendedor
CICLO_PRIMEIRA_COMPRA	INT	Período do ano em que foi feita o primeiro pedido considerando a segmentação do ano realizada pela própria empresa
CICLO_ULTIMA_COMPRA	INT	Período do ano em que foi feita o último pedido considerando a segmentação do ano realizada pela própria empresa
MES_PRIMEIRO_PEDIDO	INT	Mês do ano em que foi realizado o primeiro pedido
MES_ULTIMO_PEDIDO	INT	Mês do ano em que foi realizado o último pedido
SEMANA_ULTIMO_PEDIDO	INT	Semana do ano em que foi realizado o último pedido
DIAS_ULT_COMPRA	INT	Dias entre a última compra e a data de quebra da base em 31/12/2019
MESES_ULT_COMPRA	INT	Meses entre a última compra e a data de quebra da base em 31/12/2019
DISTINCT_MATERIALS	INT	Quantidade de produtos diferentes comprados ao longo da vida útil do revendedor
SUM_QTD_VENDA	INT	Soma da quantidade de produtos pedidos
VLR_PRATICADO_MAX_PEDIDO	FLOAT	Valor mais alto registrado em um único pedido

VLR_PRATICADO_MIN_PEDIDO	FLOAT	Valor mais baixo registrado em um único pedido
VLR_PRATICADO_MEDIO_PEDIDOS	FLOAT	Média dos valores de pedidos realizados ao longo da vida útil do revendedor antes de cessar
ANO_PRIMEIRO_PEDIDO	INT	Ano em que foi realizado o primeiro pedido
ANO_ULTIMO_PEDIDO	INT	Ano em que foi realizado o último pedido
ANOS_ATIVA	INT	Tempo em anos em que o revendedor realizou pedidos sem ultrapassar o intervalo de 21 semanas entre os pedidos
MESES_ATIVA	INT	Tempo em meses em que o revendedor realizou pedidos sem ultrapassar o intervalo de 21 semanas entre os pedidos
SEMANAS_ATIVA	INT	Tempo em semanas em que o revendedor realizou pedidos sem ultrapassar o intervalo de 21 semanas entre os pedidos
VLR_LIQUIDO_ULT_VENDA	FLOAT	Valor líquido em reais do último pedido realizado pelo revendedor
VLR_PRATICADO_ULT_VENDA	FLOAT	Valor praticado em reais do último pedido realizado pelo revendedor

4 MODELO

4.1 TRATAMENTO DOS DADOS

As bases de dados utilizadas para a construção das *features* exigiram tratamentos prévios antes de entrarem como data frames no python, para que fossem removidas vendas inválidas e revendedores com muitas informações faltantes. Também foi estabelecido um limite temporal para garantir a integridade do histórico utilizado.

Os tratamentos dentro do python tiveram foco sobretudo em corrigir preenchimentos incorretos, campos que por padrão seriam vazios e tipo dos dados após a importação da base de dados.

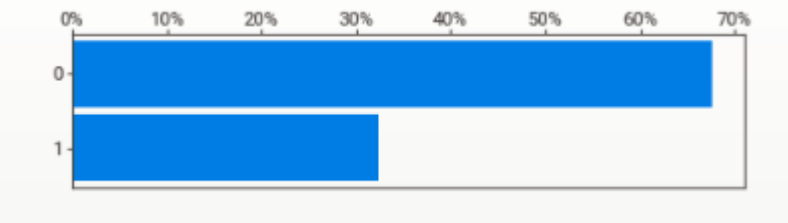
4.2 MODELOS TESTADOS

O objetivo desta modelagem é realizar uma classificação binária entre as representantes que cessaram as atividades como revendedoras de cosméticos autônomas e identificar quais vão retornar dentro de um período máximo de doze meses da última compra. Para atingir este objetivo foram testados três modelos de machine learning utilizando a linguagem Python: a regressão logística como baseline, um modelo de árvores de decisão e uma rede neural sequencial.

Os resultados dos modelos tradicionais, a regressão e a árvore, conseguiram apresentar taxas maiores de acerto com um menor esforço, a rede neural inicialmente teve resultados satisfatórios, porém convergiu para o *overffiting*, o que não pode ser corrigido mesmo com rebalanceamento e *layers* de de ruído ou *dropout*.

O target construído apresentou um percentual médio de desbalanceamento, ficando em aproximadamente 70% de cessadas e 30% de reativadas, considerando dois anos de dados. Para validar os modelos, os dados foram divididos em dois períodos, um ano e meio para treino e teste e no formato 70/30 e seis meses para previsão. As duas parcelas foram isoladas e tratadas separadamente, da mesma forma que um modelo produtivo seria tratado.

Figura 3 - Percentual da base de cessadas que reativou



Fonte: autoria própria

Os modelos foram testados com o conjunto de *features* descrito anteriormente neste trabalho e numa segunda rodada com o conjunto das vinte variáveis mais relevantes encontradas pelo modelo *Randon Forest,* e que explicavam quase 80% do ganho de *performance* com a árvore. Porém, o resultado com a redução de variáveis não apresentou melhoria nas métricas, piorando algumas, por isso a versão mantida contempla todas as *features* filtradas inicialmente.

O modelo foi construído e processado em máquina local com processador de Intel i5 Quad-Core, 16Gb de memória RAM e 500Gb de armazenamento, e sistema operacional macOS, as seguintes bibliotecas foram utilizadas durante o processamento:

- Keras
    - *Sequential*
    - *Dense*
- Sklearn

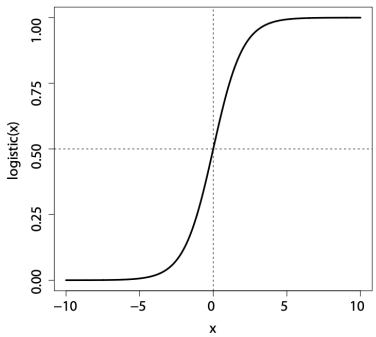
    - *StandardScale*
    - *LogisticRegression*
    - *RandonForestClassifier*
    - *Metrics*
- Pandas
- Matplotlib
- Seaborn

4.2.1 Regressão Logística

A regressão logística é um dos métodos mais comuns para a construção de modelos binários de classificação (Hilbe, 2011), tendo como base os modelos lineares generalizados de previsão. Ela calcula a probabilidade de que uma variável esteja entre 0 e 1 considerando um uma curva exponencial dada pela seguinte função e curva:

Figura 4 - Função logística

$$logistic(x) = \frac{1}{1 + e^{-x}}$$



Fonte: Kelleher, 2020 (pg 343)

A aplicação da regressão logística neste problema tinha como objetivo inicial definir valores *baseline* para comparar com os demais modelos. Por ser um algoritmo simples e de rápido processamento, ele permite identificar rapidamente, por comparação, se os demais modelos estão performando dentro de uma faixa desejável. O modelo foi construído com os parâmetros padrão definidos no pacote do *sklearn*.

Os resultados encontrados na previsão para um período de seis meses estão detalhados na tabela 2. Podemos ver que a taxa de acertos nas previsões

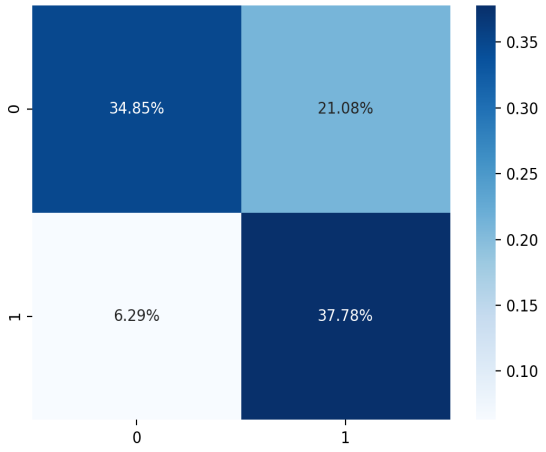
ficou acima de 70%, porém como nosso alvo não possui uma distribuição balanceada, é necessário verificar também as demais métricas.

Tabela 2 - Métricas para o modelo de regressão logística

Métricas - Regressão Logística	
accuracy_score	0,72
balanced_accuracy	0,74
roc_auc	0,74
precision	0,61
recall	0,85
f1_score	0,73

Observando o recall e também a matriz de confusão podemos ver que o modelo apresentou bons resultados encontrando os positivos verdadeiros, com uma taxa de acertos de 85%, o que em termos de negócio representa um percentual muito bom para aprimorar a gestão de campanhas focadas no público cessado da base.

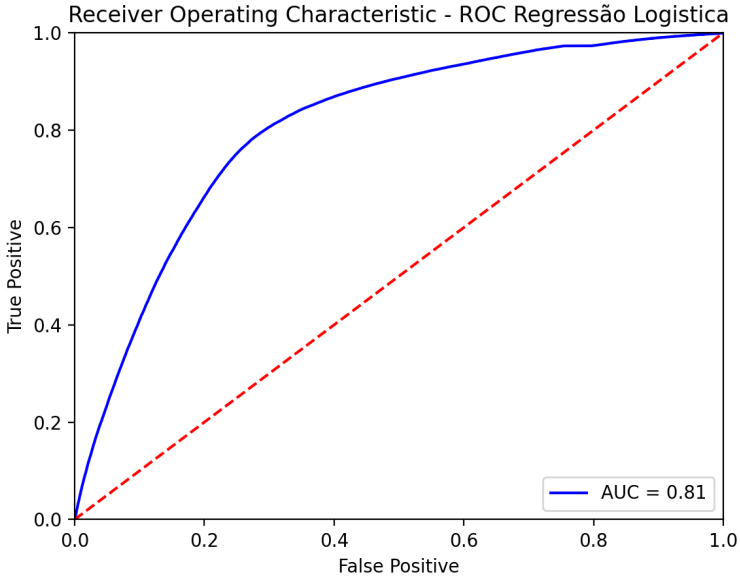
Figura 5 - Matriz de confusão para a Regressão Logística



Fonte: autoria própria

Observando o *auc score* (score geral) de 0.81 podemos ver que o modelo conseguiu alcançar uma medida de separação entre as classes bastante satisfatório, o que avaliado juntamente com as demais métricas, torna o modelo uma opção satisfatória para a construção de uma solução produtiva ao problema.

Figura 6 - Curva ROC para o modelo de Regressão Logística



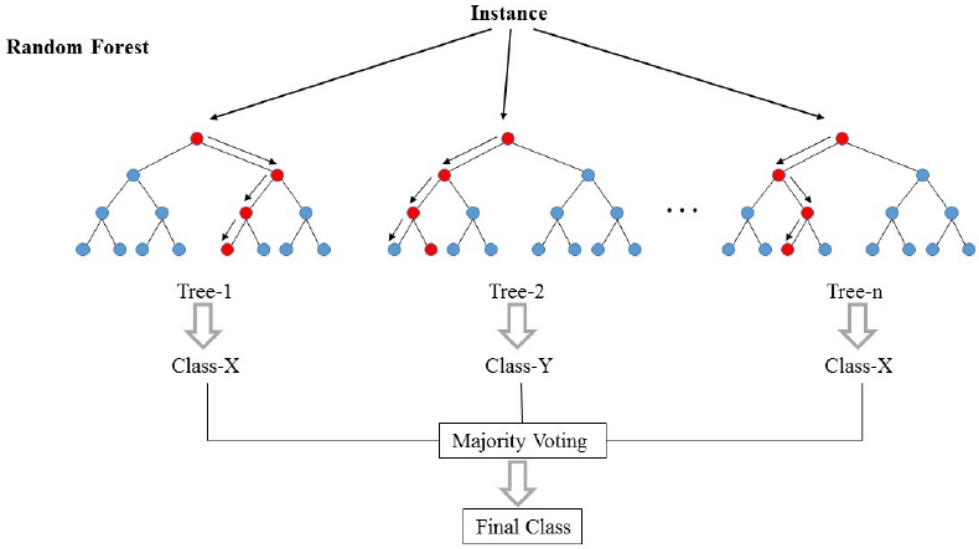
Fonte: autoria própria

4.2.2 Random Forest Classifier

As árvores de decisão são modelos clássicos do *machine learning*, partindo de uma lógica explicável para a compreensão humana. A base de uma árvore de decisão é a busca pela pureza nos diferentes níveis que vão sendo criados a partir de condições aplicadas as *features* criadas. A mesma lógica que se aplica por exemplo, para dividir adultos de crianças, a partir da faixa etária de cada pessoa, porém as árvores permitem fazer esse processo com uma quantidade muito maior de dados.

O modelo chamado de *Random Forest* usa as árvores como ferramenta para resolver problemas mais complexos, pois ele cria não apenas uma, mas diversas árvores de decisão, dando preferência para árvores com baixa correlação entre si e fazendo a previsão com base no resultado de todas as árvores. Algumas vantagens das florestas consistem em ser um algoritmo mais rápido do que outros como boosting e bagging, apresentam resultados robustos mesmo com outliers ou ruído e pode ser visualizado de forma simples (Breiman, L., 2001).

Figura 7 - Assemble de árvores no algoritmo Random Forest



Fonte: https://www.analyticsvidhya.com/blog/2020/12/lets-open-the-black-box-of-random-forests

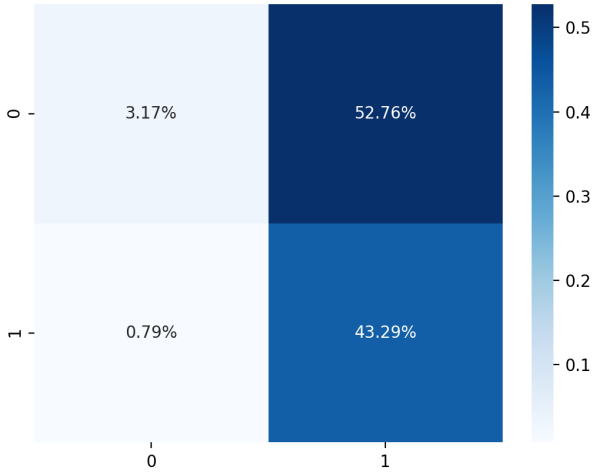
O modelo foi construído utilizando os parâmetros padrão do sklearn inicialmente e também foi feito um tunning com random search para tentar otimizar o modelo. Foram testadas cem combinações aleatórias de parâmetros e os resultados utilizados foram: profundidade máxima de 20, mínimo de amostras por nó de 2 e no máximo 1000 árvores antes do resultado. Contudo, o *ensemble* de árvores teve problemas para generalizar os resultados, tanto com os parâmetros padrão quanto com os otimizados, apresentando bons números para o treino, porém com *overffiting* nas previsões, trazendo resultados que indicariam que a quase totalidade dos revendedores cessados retornam num período de doze meses, o que sabemos que não ocorre. Visualizando as métricas podemos ver que a taxa de acerto se deu na proporção de pessoas que efetivamente retornaram no período previsto e que o recall alto indica a tendência do modelo em prever apenas a classe positiva.

Tabela 3 - Métricas para o modelo *Random Forest*

Métricas - Random Forest	
accuracy_score	0,46
balanced_accuracy	0,51
roc_auc	0,74
precision	0,45
recall	0,98
f1_score	0,62

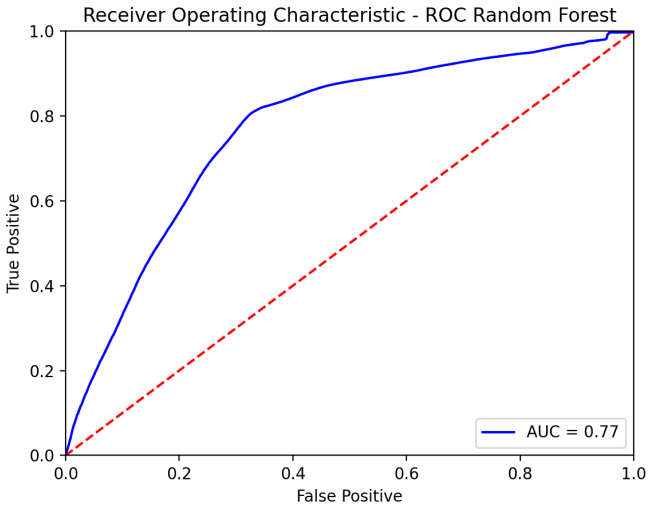
O AUC score ficou próximo do apresentado pelo modelo de regressão, um dos indicativos do perigo de se considerar métricas isoladamente. Ensembles tendem ao *overfitting*, um dos problemas identificados neste caso foi o desbalanceamento extra da base de previsão, o que indica que previsões de mais curto prazo podem exigir um volume de dados de treino maior e correções para a sazonalidade do problema.

Figura 8 - Matriz de confusão para o modelo *Random Forest*



Fonte: autoria própria

Figura 9 - Curva ROC para o modelo Random Forest



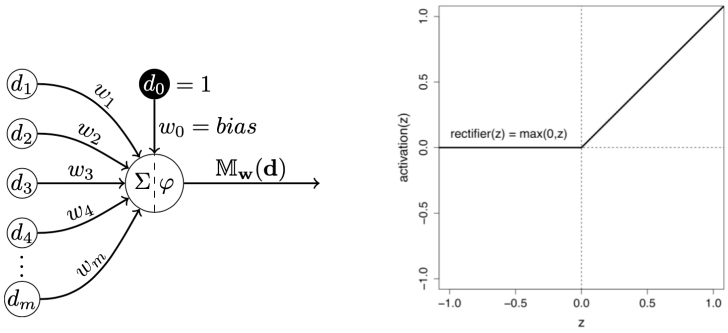
Fonte: autoria própria

4.2.3 Rede Neural

As redes neurais são modelos muito eficientes, especialmente quando utilizadas com um grande volume de dados ou quando a complexidade do

problema é elevada. Constituídas por neurônios artificiais, que visam imitar o cérebro humano, eles são preparados para receber diversos inputs e transmitir apenas um resultado para a próxima camada de neurônios. O resultado desse processo são múltiplas operações matriciais nas quais os inputs recebem pesos distintos, conforme a informação é transferida entre camadas os valores são recalculados. Nesse processo entram também as chamadas funções de ativação, cujo objetivo básico é introduzir não linearidade no processo, possibilitando que problemas mais complexos sejam resolvidos (Kelleher, 2020).

Figura 10 - Esquema resumido de uma rede neural (esquerda) e função de ativação ReLU (direita)



Fonte: Kelleher, 2020 (pg 387)

A rede neural construída para os testes neste trabalho é simples, e caracteriza-se por uma rede sequencial com um *layer* de entrada com 32 neurônios que utilizou a função de ativação ReLU, uma camada interna com a mesma função de ativação mas com 4 neurônios e uma camada de saída que utiliza a função sigmoide para fazer a previsão binária. Foram testadas ainda a adição de padrões de *dropout* e ruído, porém nenhuma delas melhorou o desempenho da rede e por isso não foram consideradas na arquitetura final. O modelo foi compilado com perda por cross entropia binária e otimizador *rmsprop*. O resumo da rede processada é apresentado abaixo:

Figura 11 - Resumo da rede neural aplicada

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_3 (Dense)              (None, 32)                2592
_________________________________________________________________
dense_4 (Dense)              (None, 4)                 132
_________________________________________________________________
dense_5 (Dense)              (None, 1)                 5
=================================================================
Total params: 2,729
Trainable params: 2,729
Non-trainable params: 0
```

Fonte: autoria própria

Os resultados apresentados durante o treino do modelo adiantam o que as métricas mostram logo em seguida, a alta acurácia indica que a rede construiu uma função de previsão muito muito ajustada aos dados de treino.

Figura 12 - Treinamento da rede neural com 3 epochs

```
val_loss: 0.1773 - val_accuracy: 0.9171 - val_auc: 0.9676

val_loss: 0.1711 - val_accuracy: 0.9195 - val_auc: 0.9692

val_loss: 0.1701 - val_accuracy: 0.9207 - val_auc: 0.9697
```

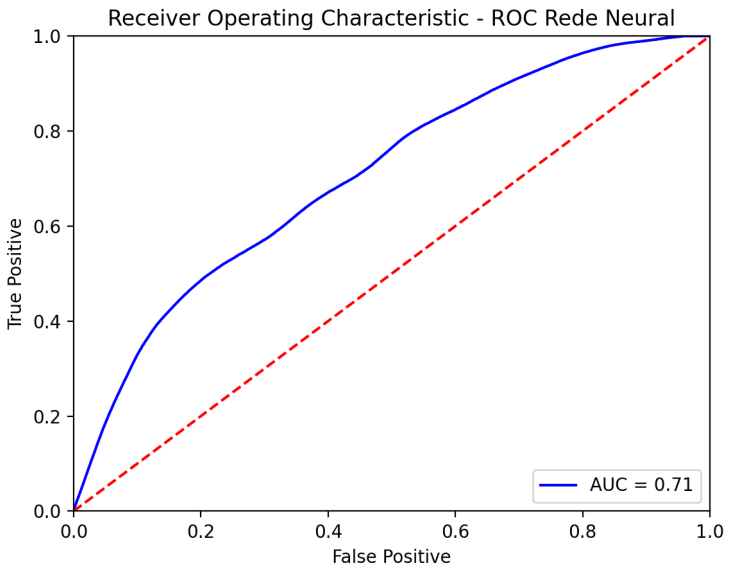
Fonte: autoria própria

O problema de generalização mostrado durante o treino ficou evidente durante e previsão: o score de acurácia reflete o percentual de classes negativas na base de previsão, visto que ao contrário das árvores, a rede previu apenas a classe com volume maior de dados, negativa, perdendo as nuances da classe positiva, o que condiz com o desbalanceamento da base teste. Testes de rebalanceamento foram realizados, porém não conseguimos alcançar melhoras expressivas na rede.

Tabela 4 - Métricas - Rede Neural

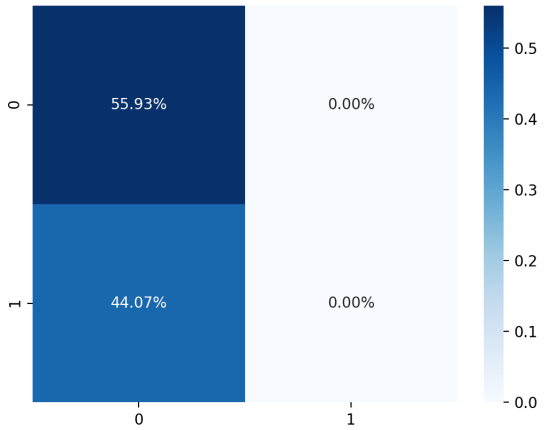
Métricas - Rede Neural	
accuracy_score	0,56
balanced_accuracy	0,5
roc_auc	0,5
precision	0,44
recall	0
f1_score	0

Figura 13 - Curva ROC para Rede Neural



Fonte: autoria própria

Figura 14 - Matriz de confusão para Rede Neural



Fonte: autoria própria

4.3 FEATURE IMPORTANCE

A construção de diferentes modelos permitiu identificar as principais variáveis utilizadas pelos modelos para realizar as previsões, o que mostrou que os modelos parecem estar de acordo com o conhecimento de negócio sobre o tema. Podemos destacar aqui que os principais fatores que explicam as reativações são relacionados a performance de vendas e o tempo que permaneceram atuando como revendedores.

Tabela 5 - Features selecionadas e importância relativa

FEATURE	IMPORTANCE
SEMANAS_ATIVA	0,085
DIAS_ULT_COMPRA	0,068
MESES_ATIVA	0,059
CICLO_ULTIMA_COMPRA	0,056
MESES_ULT_COMPRA	0,048
PEDIDOS	0,048
SUM_QTD_VENDA	0,045
DISTINCT_MATERIALS	0,038
VLR_PRATICADO_MAX_PEDIDO	0,034
VLR_PRATICADO_MEDIO_PEDIDOS	0,031
SEMANA_ULTIMO_PEDIDO	0,030
IDADE_ATUAL	0,029
VLR_PRATICADO_ULT_VENDA	0,055
ANO_ULTIMO_PEDIDO	0,029
VLR_LIQUIDO_ULT_VENDA	0,028
ANOS_ATIVA	0,027
VLR_PRATICADO_MIN_PEDIDO	0,025
DES_BLOQUEIO_CADASTRO_REVENDEDOR (sem bloqueio)	0,018
COD_MES_ULTIMO_PEDIDO	0,017

5 CONSIDERAÇÕES FINAIS

O modelo que conseguiu melhor identificar os padrões de reativação de revendedores foi também o mais simples, a regressão logística utilizada permitiu identificar cerca de 40% dos revendedores que efetivamente retornaram para a base e outros 35% que não efetivamente não retornaram, considerando uma base completamente nova para o modelo. Essa informação é muito relevante para a base de cessadas e agrega aos demais detalhes identificados como padrões de comportamento deste grupo, permitindo identificar com mais precisão com quem é possível trabalhar em termos de campanhas publicitárias e incentivos, o que em combinação com modelos de recomendação por exemplo, permitem personalizadas campanhas e trazer mais valor para um representante que deixou de comprar há alguns meses.

O modelo de árvores e a rede neural tiveram problemas generalizando as informações. Um dos maiores desafios nesse caso será construir uma janela temporal que consiga trazer exemplos adequados para a aplicação destes modelos mais complexos sem o problema do *overfitting*, como a base é naturalmente

desbalanceada e ainda existe uma tendência sazonal no para a realização de novos pedidos por cessadas, tratamentos adicionais serão necessários.

O problema estudado neste trabalho foi construído com uma base ainda incipiente de variáveis. Uma das formas mais adequadas de aprofundar as análises e modelos é aprofundar os conceitos de negócio por trás da base de *features*, o que é essencial para dar seguimento com a produtização do projeto.

REFERÊNCIAS

Kelleher, J. D., Namee, B.M., Fundamentals of Machine Learning for Predictive Data Analytics. 2 ed. The MIT Press: Cambridge, EUA, 2020.

Logistic Regression. JM Hilbe. International encyclopedia of statistical science 2011. Generalized linear models and extensions, Arizona State University, 2011.

Breiman, L., Random Forests. Machine Learning, 45, 5–32. Kluwer Academic Publishers, 2001.

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. Acessado em 14 de junho de 2021.

Sklearn Linear Model Logistic Regression - https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html. Acessado em 21 de junho de 2021.

Tyagi A., Sharma N., Sentiment Analysis using Logistic Regression and Effective Word Score Heuristic. International Journal of Engineering & Technology, 2018.

Spanoudes P., Nguyen T., Deep Learning in Customer Churn Prediction: Unsupervised Feature Learning on Abstract Company Independent Feature Vectors. Cornell University, 2017

Ullah I., Raza B, et al. A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector. IEEE Access, V7, 2019.

Christodouloua E., Ma J., et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology. Vol 110. 2019.