Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Mineração de dados na previsão de melhor canal de abordagem para próxima melhor ação no relacionamento hiper personalizado com o cliente bancário

Bruno Gomes Resende

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Orientador
Prof. Dr. Bruno César Ribas

Brasília
2024

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

Gm

Gomes Resende, Bruno
Mineração de dados na previsão de melhor canal de abordagem para próxima melhor ação no relacionamento hiper personalizado com o cliente bancário / Bruno Gomes Resende; orientador Bruno César Ribas. -- Brasília, 2024.
72 p.

Dissertação(Mestrado Profissional em Computação Aplicada) -- Universidade de Brasília, 2024.

1. Canais de interação. 2. Próxima melhor ação. 3. Hiper personalização. 4. Relacionamento com cliente. 5. Indústria financeira. I. César Ribas, Bruno, orient. II. Título.

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Mineração de dados na previsão de melhor canal de abordagem para próxima melhor ação no relacionamento hiper personalizado com o cliente bancário

Bruno Gomes Resende

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Prof. Dr. Bruno César Ribas (Orientador)
FGA/UnB

Prof. Dr. Fabiano Silva
Inf/UFPR

Prof. Dr. Glauco Vitor Pedrosa
FGA/UnB

Prof. Dr. Gladston Luiz da Silva
Coordenador do Programa de Pós-graduação em Computação Aplicada

Brasília, 13 de agosto de 2024

Dedicatória

Dedico este trabalho à minha família, com a esperança de que o empenho aqui empregado, possa servir de inspiração para o desenvolvimento do meu amado filho Miguel.

Agradecimentos

Meus sinceros agradecimentos à minha família pelo apoio e paciência ao longo dessa valorosa jornada e ao Banco do Brasil por investir continuamente na formação de seus colaboradores.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio do Acesso ao Portal de Periódicos.

Resumo

A existência de diversos canais de interação com o cliente gera um desafio na escolha do canal mais adequado nas diversas jornadas negociais nas organizações. Um banco brasileiro pretende prover as melhores experiências para os seus clientes. Para tal, deseja criar uma estratégia de abordagem hiper personalizada em suas jornadas negociais, de forma a antecipar desejos e necessidades dos consumidores de produtos e serviços oferecidos pela organização. Uma das alternativas de personalização de abordagens é a implementação de um modelo *Next Best Action (NBA)*, entregando a melhor mensagem, no momento mais adequado, pelo melhor canal de interação com o cliente. Este trabalho se propôs a modelar o comportamento de clientes bancários utilizando técnicas de Mineração de Dados a fim de elaborar um modelo de previsão para o melhor canal de interação no relacionamento negocial, de forma a atender a demanda da organização por maior eficiência operacional no envio de mensagens. Os dados utilizados incluem informações cadastrais, demográficas, comportamentais e históricos de interações, de forma a representar de forma única e individual cada cliente em seu relacionamento com a empresa, provendo assim cumprimentos à tomada de decisão do canal mais adequado para a abordagem. Foram comparadas abordagens probabilísticas, árvores de decisão, combinação de algoritmos e uso de redes neurais em modelos binários e multi classe, obtendo como resultado final uma acurácia de 0.9196, AUC de 0.9873, precisão de 0.9218 e F1 de 0.9197 em um modelo multi classe implementado com o algoritmo *Random Forest Classifier*.

Palavras-chave: Canais de Interação, Próxima Melhor Ação, hiper personalização, relacionamento com cliente, indústria financeira

Abstract

The existence of several channels for interacting with customers creates a challenge in choosing the most appropriate channel for the various business journeys in organizations. A Brazilian bank intends to provide the best experiences for its customers. To this end, it wants to create a hyper-personalized approach strategy for its business journeys, in order to anticipate the desires and needs of consumers of products and services offered by the organization. One of the alternatives for personalizing approaches is the implementation of a Next Best Action (NBA) model, delivering the best message, at the most appropriate time, through the best channel for interacting with the customer. This work proposed to model the behavior of bank customers using Data Mining techniques in order to develop a prediction model for the best interaction channel in the business relationship, in order to meet the organization's demand for greater operational efficiency in sending messages. The data used includes registration, demographic, behavioral and interaction history information, in order to uniquely and individually represent each customer in their relationship with the company, thus providing support for the decision-making process regarding the most appropriate channel for the approach. Probabilistic approaches, decision trees, combination of algorithms and use of neural networks in binary and multi-class models were compared, obtaining as a final result an accuracy of 0.9196, AUC of 0.9873, precision of 0.9218 and F1 of 0.9197 in a multi-class model implemented with the Random Forest Classifier algorithm.

Keywords: Interaction Channels, Next Best Action, hyper personalization, CRM, financial industry

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

AUC Area Under Curve.

B2B Business to Business.

CART Classification and Regression Tree.

CRISP-DM Cross Industry Standard Process for Data Mining.

CRN Coupled Recurrent Network.

CRU Coupled Recurrent Unit.

DCN Deep & Cross Network.

DNN Deep Neural Network.

FEBRABAN Federação Brasileira de Bancos.

FQI Fitted Q-iteration.

GRU Gated Recurrent Unit.

LightGBM Light Gradient Boosting Machine.

LSTM Long Short Term Memory.

LTV Lifetime Value.

MAE Mean Absolute Error.

MAPE Mean Absolute Percentage Error.

MLP Multi-layer Perceptron.

MSE Mean Squared Error.

MSLE Mean Squared Logarithmic Error.

NBA Next Best Action.

ReLU Rectified Linear Unit.

RF Randon Forest.

RL Reinforcement Learning.

RMSE Rooted Mean Squared Error.

RNN Recurrent Neural Network.

ROC Receiver Operating Characteristic.

SVM Support Vector Machine.

WDN Wide & Deep Network.

XGBoosting Extreme Gradient Boosting Machine.

Capítulo 1

Introdução

Proporcionar experiências positivas em jornadas negociais é um diferencial para as empresas em um mercado competitivo e global. Um banco brasileiro pretende ser a empresa que proporciona a melhor experiência para os seus clientes. Como estratégia para alcançar este objetivo, destaca-se a criação de abordagens hiper personalizadas em suas jornadas negociais, de forma a antecipar desejos e necessidades dos consumidores em seus produtos e serviços. Para personalizar a abordagem ao cliente, um sistema de recomendação *Next Best Action (NBA)* mostra-se aderente à expectativa da empresa, prevendo a próxima ação mais promissora na jornada negocial para cada cliente, fornecendo a mensagem certa, na hora e no canal mais adequado [1]. Contudo, a escolha do canal de interação mais adequado é uma tarefa desafiadora para a organização, que conta com diversas opções de canais. A partir dos dados cadastrais, demográficos, comportamentais e históricos de interações com seus clientes, pretende-se elaborar um modelo preditivo que forneça à organização um ranqueamento dos canais mais adequados para interação com um determinado cliente, contribuindo com a estratégia NBA.

1.1 Definição do Problema

Para a implementação de um modelo NBA, uma abordagem utilizada pelo mercado [2, 3] é a criação de uma arquitetura composta pela combinação de outros modelos que entreguem a mensagem mais adequada para o cliente, no instante e canal mais adequado, e isso envolve a combinação de diferentes modelos de propensão ao consumo de produtos e serviços, modelos de retenção e prevenção à evasão de clientes e modelos que tracem o momento de vida do cliente, propensão ao uso de canais e modelos de abordagem baseados em geolocalização. A Figura 1.1 representa uma abstração da arquitetura de um modelo NBA.

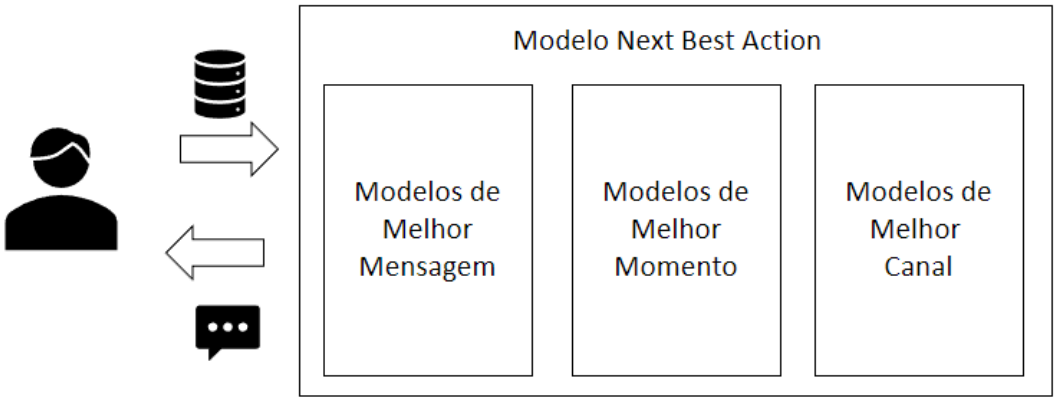

Figura 1.1: Arquitetura de Um Modelo NBA.

Segundo a Federação Brasileira de Bancos (FEBRABAN), os canais de atendimento podem ser classificados [1] em (1) Canais Digitais, (2) Canais Telefônicos e (3) Canais Presenciais. Nesta classificação estão canais como Internet Banking e APPS Bancários, Centrais de Atendimento e Agências, conforme se observa na Tabela 1.1.

Tabela 1.1: Canais de Atendimento

Tipo	**Canal**
Digital	Internet Banking
Digital	APPS Bancários
Digital	Redes Sociais
Digital	Caixas Eletrônicos
Telefônico	Central de Atendimento
Telefônico	Serviço de Atendimento ao Consumidor
Telefônico	Ouvidoria
Presencial	Agências
Presencial	Correspondentes

Na Pesquisa FEBRABAN de Tecnologia Bancária 2022 [4], o volume de transações em canais digitais é crescente nas instituições bancárias do Brasil, sendo que atualmente, 7 em cada 10 transações são realizadas por meio de canais como Mobile e Internet Banking.

O problema abordado neste trabalho trata, portanto, de melhorar o processo de comunicação com o cliente, racionalizando o uso de recursos tecnológicos para a entrega de mensagens em canais digitais.

[1]Disponível em: https://portal.febraban.org.br/pagina/3055/30/pt-br/canais-de-atendimento

1.2 Justificativa

A existência de diversos canais de interação no banco em questão é uma facilidade ofertada aos clientes, tornando a empresa presente na vida e nos meios de comunicação mais utilizados pelas pessoas. São cerca de 40 canais, entre físicos e digitais. Esta facilidade diminui o atrito nas interações mas pode gerar dificuldades em capturar a atenção do cliente em um determinado meio de comunicação. Portanto, definir uma ordem de prioridade no uso de canais para a entrega de comunicações, de forma personalizada ao interlocutor é um grande desafio para a empresa.

1.3 Hipótese

A partir dos dados cadastrais, demográficos, comportamentais e históricos de interações de clientes bancários, esta pesquisa pretende modelar um perfil de propensão ao uso dos canais de comunicação providos pela empresa, predizendo assim, de forma personalizada para cada cliente, a efetividade de uma ação de comunicação em um determinado canal de abordagem. Este modelo classificador de efetividade da abordagem em canais poderá ser utilizado como componente de um modelo de NBA para a organização. Assume-se neste trabalho portanto, a hipótese de que o uso de modelos personalizados permitem aprimorar a relação da empresa com o cliente e aumentar o sucesso das abordagens.

1.4 Objetivos

Esta pesquisa busca identificar o melhor canal de interação com um determinado cliente bancário, com base na predição do sucesso da abordagem, a partir de informações de perfil e históricos de interação para composição de um modelo final de NBA em um banco brasileiro. Este trabalho pretende utilizar uma abordagem de Mineração de Dados para inferir o melhor canal de interação com o cliente.

1.4.1 Objetivo Geral

Desenvolver um modelo hiper-personalizado para a verificação do sucesso de abordagens em um canal de interação a fim de prover a melhor experiência para os clientes utilizando aprendizado de máquinas.

1.4.2 Objetivos Específicos

De forma a conduzir este trabalho, objetiva-se especificamente a:

- Realização de *benchmark* entre diferentes abordagens de modelagem utilizando técnicas probabilísticas;
- Realização de *benchmark* entre diferentes abordagens de modelagem utilizando redes neurais;
- Comparação dos resultados obtidos dentre as abordagens de modelagem e técnicas utilizadas;

1.5 Contribuições

No aspecto tecnológico, a contribuição deste trabalho se destaca pela realização de *benchmarks* que comparam diferentes abordagens e modelos para atingir objetivos específicos de negócio. Essa prática não apenas oferece percepções sobre o desempenho de modelos sob diferentes cenários, como também possibilita a escolha das melhores soluções para demandas empresariais concretas. A inovação potencial surge com a implementação de um modelo de relacionamento hiper personalizado na indústria financeira. Ao personalizar a interação com os clientes com base em suas necessidades e comportamentos individuais, a tecnologia proposta pode aprimorar o modo como os serviços financeiros são oferecidos, elevando a experiência do cliente e promovendo maior eficiência operacional.

1.6 Organização do Trabalho

Este trabalho de pesquisa está dividido nos seguintes capítulos:

- Introdução, onde está descrita a justificativa, hipótese de pesquisa, objetivos geral e específicos do estudo;
- Fundamentação Teórica, onde são explicados conceitos e técnicas utilizadas no desenvolvimento do trabalho;
- Trabalhos Correlatos, onde são listados trabalhos relacionados, trazendo métodos e técnicas de destaque sobre o tema abordado nesta pesquisa;
- Modelo Preditivo: Melhor Canal de Interação, onde está relatado o experimento proposto no estudo;
- Conclusão, onde estão elencadas as lições aprendidas, resultados obtidos no trabalho e proposta de trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Neste capítulo serão abordadas técnicas e métodos, bem como conceitos-chave utilizados na elaboração deste trabalho.

2.1 Aprendizado de Máquinas

O aprendizado de máquinas, é uma disciplina que se insere no campo da inteligência artificial e tem se tornado um dos pilares fundamentais da computação moderna. Sua essência reside na capacidade de sistemas computacionais aprenderem padrões, estruturas e informações a partir de dados, sem uma programação explícita. Isso é possível por meio da aplicação de algoritmos e modelos estatísticos que permitem que as máquinas aprimorem seu desempenho à medida que mais dados são processados. O aprendizado de máquinas encontra aplicação em diversos domínios, desde reconhecimento de voz e visão computacional até previsão de mercado e diagnóstico médico.

Uma das vertentes mais relevantes do aprendizado de máquinas é o aprendizado supervisionado, em que os modelos são treinados com um conjunto de dados rotulados para realizar previsões ou classificações. Além disso, o aprendizado não supervisionado visa identificar padrões ocultos e estruturas nos dados sem a necessidade de rótulos. Outras abordagens incluem o aprendizado por reforço, no qual os agentes de aprendizado interagem com um ambiente para otimizar ações, e o aprendizado profundo, que se vale de redes neurais profundas para modelar representações complexas e de alto nível dos dados.

O campo do aprendizado de máquinas é dinâmico, com pesquisas em constante evolução, e suas aplicações continuam a se expandir, moldando significativamente a forma como as organizações processam informações e tomam decisões em um mundo cada vez mais orientado por dados.

2.2 Sistemas de Recomendação

Os sistemas de recomendação são uma classe de sistemas de filtragem de informações que desempenham um papel essencial na era da informação, auxiliando os usuários na descoberta de conteúdo relevante em uma ampla variedade de domínios, incluindo comércio eletrônico, entretenimento e mídias sociais. Esses sistemas são projetados para prever as preferências dos usuários, com base em seu histórico de interações, e fornecer recomendações personalizadas. Uma das abordagens fundamentais para a construção de sistemas de recomendação envolve a filtragem colaborativa, na qual as preferências dos usuários são inferidas a partir de padrões de comportamento semelhantes entre usuários ou itens. Além disso, os sistemas de recomendação podem se beneficiar do uso de técnicas de aprendizado profundo, como redes neurais, para capturar representações complexas e de alto nível dos dados, proporcionando recomendações mais precisas e personalizadas.

Uma dimensão importante dos sistemas de recomendação é a explicabilidade e a ética. A interpretabilidade dos sistemas de recomendação torna-se crucial, pois os usuários exigem cada vez mais compreender o raciocínio por trás das recomendações, garantindo transparência e confiança. Questões éticas relacionadas à privacidade e ao viés dos algoritmos de recomendação têm recebido atenção significativa, com a necessidade de abordar preocupações sobre discriminação e justiça algorítmica. Os sistemas de recomendação representam um campo de pesquisa em constante evolução, com desafios contínuos relacionados ao aprimoramento da qualidade das recomendações, ao equilíbrio entre precisão e explicabilidade, e à mitigação de riscos éticos, com o objetivo de atender às necessidades e expectativas dos usuários de forma responsável.

2.3 *Next Best Action (NBA)*

Os modelos preditivos *Next Best Action (NBA)* constituem uma abordagem sofisticada na área de aprendizado de máquina e análise de dados, direcionada à otimização de decisões em cenários de interação com o cliente. Esses modelos se baseiam em algoritmos de aprendizado de máquina, como árvores de decisão, redes neurais e algoritmos de reforço, para determinar a ação mais apropriada a ser tomada em tempo real com base no histórico de interações e nas características do cliente. A complexidade desses modelos reside na necessidade de equilibrar a exploração (aprendizado) e a exploração (ação) de maneira eficaz, a fim de maximizar a recompensa esperada. Além disso, a incorporação de técnicas de interpretabilidade e ética é fundamental para garantir a confiabilidade e a transparência dessas decisões, principalmente em setores sensíveis, como serviços financeiros e saúde. Em última análise, os modelos preditivos *NBA* desempenham um papel vital na personalização

de interações com o cliente e na otimização das estratégias de engajamento, contribuindo para o aprimoramento das relações e o alcance de metas empresariais.

2.4 *Ensemble*

As técnicas de *ensemble* [5] têm sido utilizadas em aprendizado de máquina e ciência de dados para melhorar o desempenho dos modelos. O *ensemble* objetiva combinar as previsões de vários modelos individuais para obter uma previsão mais precisa e confiável. Existem diferentes técnicas de *ensemble*, como *bagging*, *boosting*, *stacking* e *blending*. Cada técnica possui suas próprias características e métodos de combinação.

O *Bagging* (*bootstrap aggregating*) é uma técnica em que várias instâncias do mesmo modelo são treinadas em diferentes conjuntos de dados de treinamento, criados por amostragem com reposição dos dados originais. As previsões dos modelos individuais são combinadas usando média ou voto majoritário.

O *Boosting* é uma técnica em que vários modelos fracos são treinados sequencialmente, cada um corrigindo os erros do modelo anterior. A combinação das previsões é feita usando uma média ponderada.

O *Stacking* é uma técnica em que as previsões dos modelos individuais são usadas como entrada para um modelo meta, que é treinado para prever a saída final. Isso permite que o modelo meta capture padrões mais complexos que não são facilmente detectáveis pelos modelos individuais.

O *Blending* é uma técnica em que várias previsões são combinadas usando uma média ponderada ou voto majoritário. No entanto, ao contrário do *bagging*, os modelos individuais são treinados em todo o conjunto de dados de treinamento, em vez de conjuntos de dados amostrados.

Além disso, existem outras técnicas de *ensemble*, como *random forests*, *gradient boosting machines* e *deep ensemble*, que são variantes das técnicas básicas descritas acima.

Em resumo, as técnicas de *ensemble* são uma abordagem poderosa para melhorar o desempenho dos modelos de aprendizado de máquina e ciência de dados. A escolha da técnica de *ensemble* depende do problema em questão e dos dados disponíveis, e é importante avaliar cuidadosamente o desempenho do modelo para garantir que a abordagem escolhida seja a mais adequada para o problema em questão.

2.5 Redes Neurais

Uma rede neural é um modelo matemático que simula a forma como os neurônios no cérebro humano processam informações. Ela aprende a partir de dados, ajustando seus

parâmetros internos para realizar tarefas específicas. Essa capacidade de aprendizado e generalização faz das redes neurais uma técnica valiosa em uma variedade de aplicações, desde reconhecimento de voz e imagem até tomada de decisões complexas baseadas em dados.

Uma rede neural é composta por unidades interconectadas chamadas neurônios, organizadas em camadas. Cada conexão entre neurônios tem um peso que é ajustado durante o treinamento da rede. A camada de entrada recebe dados, que são propagados através das camadas ocultas até a camada de saída, onde a rede gera uma resposta. Durante o treinamento, a rede ajusta os pesos das conexões para minimizar a diferença entre suas previsões e os valores reais, utilizando algoritmos de otimização. Os modelos *Wide and Deep* e *Deep and Cross* a seguir são exemplos do uso de redes neurais no aprendizado de máquinas.

2.5.1 *Wide and Deep*

O modelo *Wide and Deep* proposto por Cheng et al. [6] em 2016 trás uma abordagem promissora para a recomendação de melhor canal de interação com o cliente, proposta neste trabalho. O modelo explora duas características principais: a memorização, no modelo *Wide* e a generalização no modelo *Deep*.

A memorização pode ser definida como o ato de aprender sobre a frequente ocorrência simultânea de elementos ou características e aproveitar a correlação disponível nos dados históricos. A generalização, por outro lado, é baseada na transitividade da correlação e explora novas combinações de recursos que nunca ou raramente ocorreram no passado.

Trata-se da combinação portanto de dois modelos, o primeiro, baseado em um modelo linear generalista, o segundo baseado em *Deep Learning*, conforme representado na Figura 2.1.

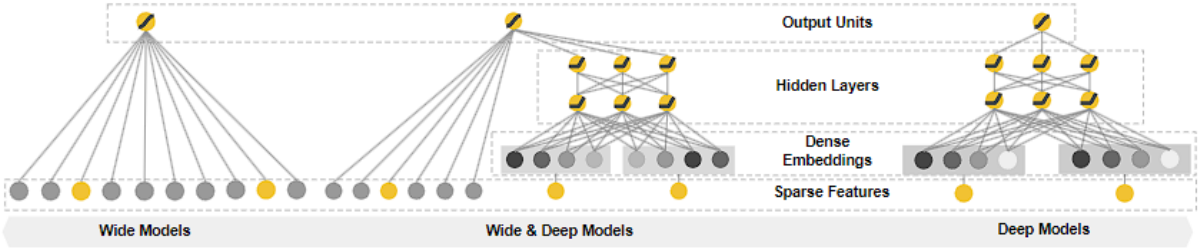

Figura 2.1: Especto dos Modelos Wide and Deep (Fonte: [6]).

2.5.2 *Deep and Cross*

O modelo *Deep and Cross* proposto por Wang et al. [7] em 2017 é uma arquitetura inovadora no campo de aprendizado de máquina, projetada para capturar relações não lineares de longo alcance entre diferentes características em conjuntos de dados. Desenvolvido para melhorar a expressividade das redes neurais, o modelo *Deep and Cross* incorpora uma operação de produto cruzado em suas camadas, permitindo que as características se influenciem mutuamente de maneira mais sofisticada. Essa operação promove uma exploração mais eficiente das interações entre variáveis, superando as limitações de modelos lineares e até mesmo de algumas redes neurais profundas tradicionais.

A arquitetura do *Deep and Cross* é composta por duas partes principais: a parte *Deep*, que representa as camadas convolucionais profundas, responsáveis por aprender representações complexas e hierárquicas das características, e a parte *Cross*, que incorpora as operações de produto cruzado para capturar interações entre essas características. O modelo é treinado de maneira fim a fim, ajustando automaticamente os pesos das conexões durante o processo de otimização. Esse design modular permite que o *Deep and Cross* seja aplicado a uma variedade de tarefas, desde recomendações personalizadas até sistemas de filtragem de informações, destacando sua versatilidade e capacidade de lidar com dados complexos e relacionamentos não lineares de maneira eficaz.

A Figura 2.2 representa o modelo *Deep and Cross* completo.

2.6 CRISP-DM

O Cross Industry Standard Process for Data Mining (CRISP-DM) [8] é um método padrão para mineração de dados que é frequentemente utilizada em projetos de ciência de dados. Desenvolvido em 1996 pelo grupo de trabalho da indústria de mineração de dados, trata-se de uma abordagem iterativa que guia os cientistas de dados através de seis fases distintas: compreensão do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e implementação do modelo. A abordagem estruturada ajuda a garantir que os cientistas de dados enderecem o problema de forma metódica e consistente, garantindo a qualidade e a eficácia dos resultados finais.

Na primeira fase do processo, a de compreensão do negócio, os objetivos do projeto são definidos e os requisitos do cliente são identificados. Isso envolve a compreensão do problema do negócio e a definição dos critérios de sucesso do projeto. Nesta fase, geralmente se estabelece um plano de projeto detalhado que inclui cronogramas, orçamentos e recursos necessários para a execução.

Na segunda fase, de compreensão dos dados, o objetivo é coletar e analisar os dados disponíveis para o projeto. Isso envolve a identificação de todas as fontes de dados rele-

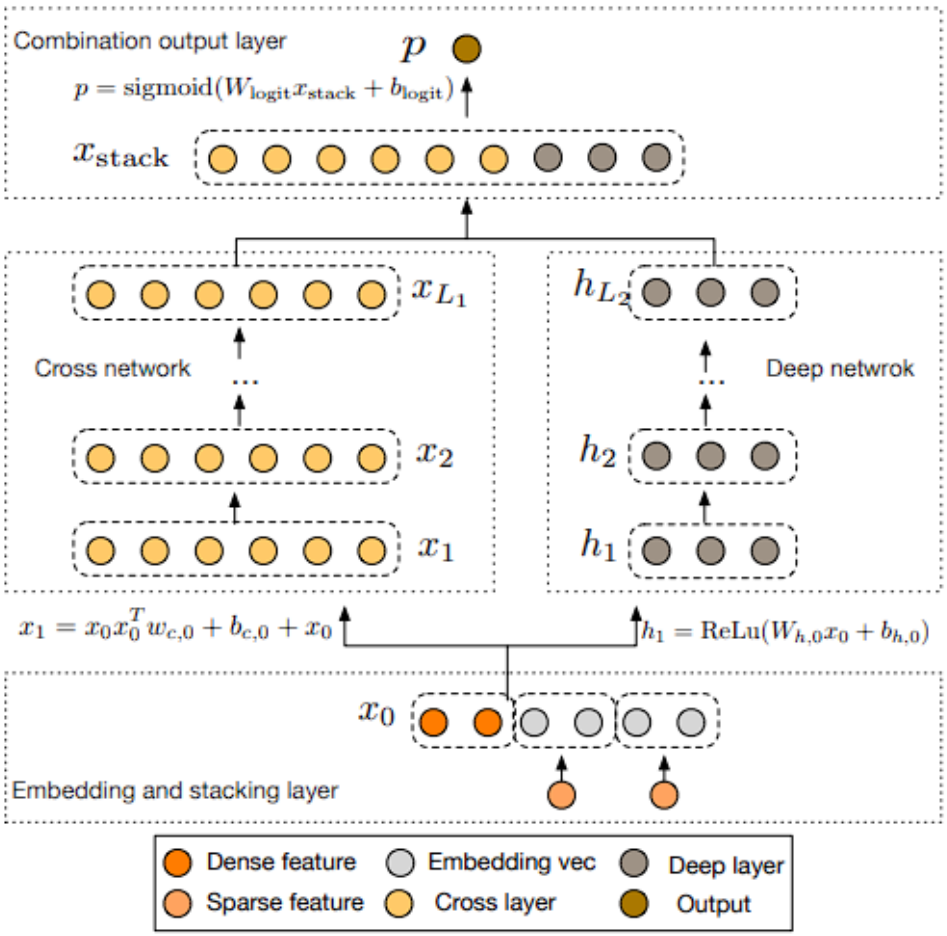

Figura 2.2: Deep and Cross Network (Fonte: [7]).

vantes, a coleta dos dados e a realização de uma análise exploratória inicial para entender a qualidade e a estrutura dos dados.

A terceira fase, de preparação dos dados é onde os dados são limpos, pré-processados e transformados em um formato adequado para análise. Isso envolve a seleção das variáveis relevantes, o tratamento de valores ausentes e a padronização dos dados.

Na quarta fase, de modelagem, são criados os modelos preditivos ou descritivos para resolver o problema de negócio identificado na fase de compreensão. Para isso é realizada a seleção de técnicas de modelagem apropriadas e a criação e avaliação de diversos modelos.

Na quinta fase, a de avaliação, os modelos criados na fase de modelagem são avaliados para determinar sua eficácia na resolução do problema de negócio. Para tal são avaliadas métricas de desempenho, como acurácia, precisão, *recall*, *F1 score* e *Area Under Curve (AUC)*, bem como a análise dos resultados para identificar possíveis melhorias.

Finalmente, na sexta fase, de implementação, o modelo escolhido na fase de avaliação é implementado no ambiente de produção organizacional e os resultados são monitorados

para garantir que o modelo continue a atender aos requisitos do negócio, retomando ao início do clico quando pertinente.

Uma representação gráfica das fases do CRISP-DM pode ser observada na Figura 2.3 a seguir.

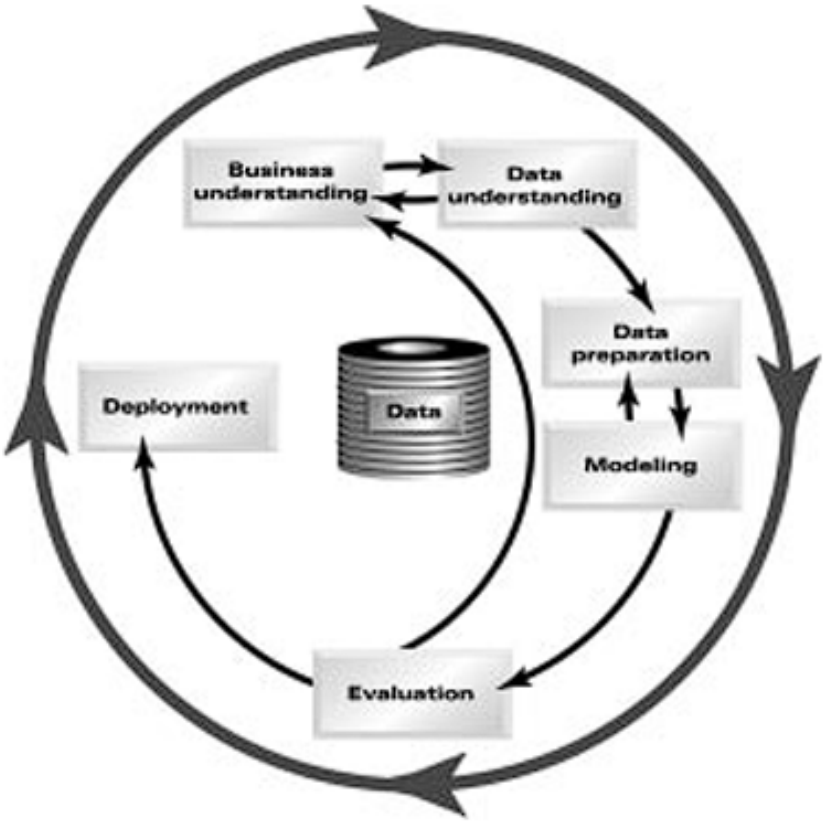

Figura 2.3: Fases do CRISP-DM (Fonte: [8]).

2.7 Stratified K-Fold

O Stratified K-Fold é uma técnica de validação cruzada frequentemente empregada em machine learning para avaliar o desempenho de um modelo de forma mais robusta. O método envolve a divisão do conjunto de dados em K subconjuntos (ou "folds") de forma estratificada, o que significa que a distribuição das classes no conjunto de dados é mantida em cada fold. Essa abordagem é particularmente útil quando se lida com conjuntos de dados desbalanceados, nos quais há uma discrepância significativa no número de exemplos entre diferentes classes. No Stratified K-Fold, cada fold contém uma proporção representativa de exemplos de cada classe, garantindo que o modelo seja avaliado de maneira equitativa em termos de todas as classes presentes no conjunto de dados.

Durante o processo de validação cruzada, o modelo é treinado K vezes, cada vez utilizando K-1 folds para treinamento e o fold restante para avaliação. Essa técnica oferece uma avaliação mais robusta do desempenho do modelo, uma vez que cada exemplo no conjunto de dados tem a oportunidade de ser utilizado tanto para treinamento quanto para validação, reduzindo a variabilidade nos resultados da avaliação do modelo. O Stratified K-Fold é particularmente útil quando a performance equitativa entre classes é crucial para a aplicação do modelo, tornando-se uma escolha valiosa em muitos cenários de aprendizado de máquina.

2.8 Métricas

As métricas são medidas numéricas utilizadas para avaliar o desempenho de um modelo de aprendizado de máquina em um conjunto de dados de teste, bem como para subsidiar os passos de análise de dados para a construção do modelo. Existem várias métricas de avaliação disponíveis [9], e a escolha da métrica a ser usada depende do tipo de problema que está sendo resolvido e do objetivo da análise. Dentre as métricas comumente utilizadas destacam-se:

- *Accuracy* (acurácia): é a proporção de predições corretas feitas pelo modelo em relação ao total de predições, conforme a Equação 2.1 a seguir. É a métrica mais simples e comum para avaliar o desempenho do modelo. No entanto, pode não ser adequada para dados desbalanceados, onde as classes possuem distribuições diferentes.

$$\text{Accuracy} = \frac{\text{Correct Classifications}}{\text{All Classifications}} \tag{2.1}$$

- *Precision* (precisão): é a proporção de verdadeiros positivos (TP) em relação à soma de TP e falsos positivos (FP). É uma medida da precisão do modelo em relação à previsão da classe positiva. É importante em problemas onde a classificação correta da classe positiva é fundamental. A representação matemática para a precisão pode ser observada na Equação 2.2 a seguir.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{2.2}$$

- *Recall* (revocação): é a proporção de TP em relação à soma de TP e falsos negativos (FN). É uma medida da capacidade do modelo em identificar todos os exemplos da classe positiva, conforme observamos na Equação 2.3. É importante em problemas onde a perda de um exemplo positivo é crítica.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{2.3}$$

- *F1 score*: é a média harmônica entre *Precision* e *Recall*. É uma métrica balanceada que considera tanto a precisão quanto a capacidade de identificação do modelo, observada na Equação 2.4 a seguir.

$$F1score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.4}$$

- *AUC* (área sob a curva Receiver Operating Characteristic (ROC)): é uma medida de desempenho que avalia a capacidade do modelo de distinguir entre as classes positiva e negativa. A curva ROC é uma curva que representa a taxa de verdadeiros positivos em relação à taxa de falsos positivos, para diferentes valores de limiar de decisão. A AUC é a área sob a curva ROC, e varia de 0 a 1, sendo que valores mais próximos de 1 indicam um melhor desempenho do modelo. A representação matemática para *AUC* pode ser verificada na Equação 2.5.

$$\text{AUC} = \frac{1}{M \cdot N^+} \sum_{i=1}^{M} \sum_{j=1}^{N^+} \text{rank}(i) \cdot \text{I}(y_i = 1, \hat{y}_j = 0) \tag{2.5}$$

  Onde:

  M é o número total de exemplos negativos.

  N^+ é o número total de exemplos positivos.

  rank(i) é o rank do i-ésimo exemplo negativo em relação aos exemplos positivos.

  I é a função indicadora, que é 1 se a condição dentro dos parênteses for verdadeira, e 0 caso contrário.

  y_i é a verdadeira classe do i-ésimo exemplo.

  $\hat{y}_j$ é a pontuação prevista para o j-ésimo exemplo.

- Coeficiente de Correlação de Pearson [10]: é uma medida estatística que avalia a relação linear entre duas variáveis quantitativas. Ele é calculado pela covariância das duas variáveis dividida pelo produto dos desvios padrão individuais. O valor do coeficiente varia entre -1 e 1, onde -1 indica uma correlação negativa perfeita, 1 indica uma correlação positiva perfeita e 0 representa a ausência de correlação linear. O coeficiente de Pearson é amplamente utilizado na análise estatística para quantificar a força e direção da relação entre variáveis, facilitando a compreensão de como as variáveis estão associadas e auxiliando na tomada de decisões informadas

em diversas áreas da ciência e pesquisa. A Equação 2.6 a seguir representa a fórmula matemática da métrica.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.6}$$

Onde:

n representa o número de observações.

x_i é o valor individual de x.

y_i é o valor individual de y.

$\bar{x}$ é a média de x.

$\bar{y}$ é a média de y.

Capítulo 3

Trabalhos Correlatos

Neste capítulo serão sumarizados os principais trabalhos estudados para a fundamentação teórica do modelo implementado neste projeto, com um resumo do objetivos dos trabalhos, abordagens e métricas utilizadas. Ao fim, uma tabela resumo dos trabalhos é apresentada.

No artigo *An ensemble based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in telecom industry*, Bilal et al. [11] apresenta uma abordagem para a previsão de evasão de clientes na indústria de telecomunicações baseada em *ensemble*, combinando técnicas de clusterização e classificação. Nesse sentido foram utilizados dois *datasets*, o primeiro composto de 5000 registros, contendo 707 registros de evasão e 4293 registros de não evasão. O segundo composto de 3333 registros com 21 atributos não especificados, contendo 483 registros de evasão e 2850 registros de não evasão. Foram utilizadas as seguintes técnicas para a elaboração dos modelos: *K-means clustering*, *K-medoids clustering*, *X-means clustering*, *Random clustering*, *K-nearest neighbor*, árvore de decisão, *Gradient boosted tree*, *Random forest*, *Deep learning classifier*, *Naïve Bayes*, *NB (K) (Naïve Bayes Kernel)* e *Ensemble classifiers (Voting, Bagging, AdaBoost, Stacking).*

Como método para o experimento, foram executadas as fases de pré processamento de dados, com limpeza de dados, seleção de variáveis e redução da dimensionalidade dos dados. De forma a mensurar o experimento foram utilizadas as medidas de acurácia, precisão, *recall* e *F1 score*. Neste trabalho, a combinação das técnicas *K-med clustering*, *Gradient boosted tree*, árvore de decisão e *Deep learning classifier* demonstrou uma maior acurácia quando comparado aos outros métodos.

Em *Proactive customer retention management in a non-contractual B2B setting based on churn prediction with random forests*, Gattermann-Itschert et al.[12] implementa um modelo de gestão para a retenção de clientes em relações Business to Business (B2B) não regidas por contrato. Foram utilizados 4952 registros de clientes, sendo desses 371 regis-

tros com indicativo de evasão e 4581 registros sem indicativo de evasão. O experimento seguiu as etapas de coleta de dados, seleção de variáveis, criação de rótulos de variáveis e elaboração de modelos onde as técnicas *Random forest*, Regressão Logística e *Support Vector Machine (SVM) linear* foram aplicadas. Como métricas neste trabalho foram utilizadas *AUC*, *Average precision* e *Top-decile lift*. Os resultados demonstraram um maior desempenho no uso de *Random forest* no modelo preditivo.

Cao e Zhu [13] elaborou um modelo de recomendação de próxima melhor ação personalizada baseada em aprendizado de máquina, utilizando a interação entre várias partes para a tomada de decisão automatizada. Os experimentos do modelo foram realizados utilizando dados de cobrança de dívidas em uma importante agência governamental australiana. Foram utilizados dados de 5 anos relacionados à dívida do contribuinte com o governo, que compreendem 61361 contribuintes, 10 ações de cobrança de dívidas selecionadas e 66126 sequências de ações governamentais de respostas de contribuintes em um total de 111514 operações de dívida. A base de dados é constituída de atributos sobre dados demográficos do contribuinte e circunstâncias, o valor e a duração da dívida em cada momento associado a um devedor, uma lista de ações opcionais de cobrança de dívidas e suas restrições de política de aplicação, uma sequência de ações históricas tomadas pelo governo em um devedor para recuperar a dívida em cada momento, o comportamento de resposta do contribuinte correspondente a cada ação de cobrança de dívidas e as informações de tempo informações associadas a casos de dívida, respostas e ações.

O modelo foi criado utilizando-se uma Coupled Recurrent Network (CRN) e confrontado com variações de três modelos estado da arte: o modelo Wide & Deep Network (WDN) do Google, Recurrent Neural Network (RNN) baseados em Long Short Term Memory (LSTM) e Gated Recurrent Unit (GRU) e a combinação do WDN com RNN. Uma Coupled Recurrent Unit (CRU) foi implementada para representação do estado atual do contribuinte. O modelo CRN representa os recursos demográficos de cada contribuinte (por exemplo, tipo de contribuinte, endereço, setor da indústria, etc.) para formar os estados iniciais da CRU. Isso resolve a partida a frio do problema na tomada de decisão, assumindo que contribuintes com características demográficas semelhantes provavelmente compartilham comportamentos semelhantes.

O modelo usa a função de ativação Rectified Linear Unit (ReLU) para mapeamento não linear e possui uma camada de normalização em lote após todas as camadas não lineares. As redes de perceptrons multicamadas Multi-layer Perceptron (MLP) no modelo têm três camadas. O modelo CRN foi treinado utilizando *Adam optimization algorithm* com um tamanho de lote de 128. Para mensurar a efetividade do modelo proposto foram usadas métricas de precisão (*Precision lift*), redução de erro (*Reward Mean Squared Error (MSE)*) e incremento na função de recompensa (*Average reward lift*). Os resultados

obtidos no experimento demonstraram que o modelo CRN proposto performa adequadamente e com melhores resultados frente aos *benchmarks* utilizados podendo ser estendido à outros domínios de informação e problemas de interação.

No trabalho *Multi-class imbalanced enterprise credit evaluation based on asymmetric bagging combined with light gradient boosting machine*, Sun et al. [14] desenvolvem um modelo de avaliação de crédito utilizando *bagging* assimétrico combinado com Light Gradient Boosting Machine (LightGBM) em classes desbalanceadas. O *dataset* utilizado contém uma amostra de 14487 registros de três classes distintas (1) baixo, (2) médio e (3) alto risco de inadimplência. Sendo que a quantidade de amostras das classes 1, 2 e 3 são 7967, 6314 e 206 registros respectivamente, o que se caracteriza por alto desequilíbrio de classes. Para avaliar a performance dos modelos foram elencados três indicadores: *recall* específico de classes, *macro recall* e a acurácia geral da predição. Os modelos foram elaborados utilizando LightGBM que é um algoritmo *gradient boosting* baseado em Classification and Regression Tree (CART). Como resultados obtidos, foram propostos dois novos modelos para a avaliação de crédito combinados em um LightGBM *ensemble classifier*, onde as métricas apuradas desse novo modelo foram melhores do que os *benchmarks* utilizados.

No artigo *Comparison of different ensemble methods in credit card default prediction*, Faraj et al. [15] comparam métodos de *ensemble* como *bagging*, *boosting* e *stacking* para a previsão de inadimplência em cartões de crédito. Para isso utilizou uma base de dados não balanceada composta de 30 mil registros de clientes, contendo 23 atributos, dentre eles um indicador de inadimplência como variável dependente e as variáveis independentes: quantidade de crédito estabelecido, gênero, escolaridade, estado civil, idade e variáveis contendo históricos de pagamento em diferentes períodos. Foram utilizadas as técnicas *XGBoosting*, Regressão Logística, Redes Neurais,*bagging*, *Ada Boost*, *voting ensemble* e *stacking*.

Como *benchmark* de desempenho do experimento utilizou-se o modelo de Rede Neural. O experimento foi realizado no primeiro momento com o *dataset* não balanceado e posteriormente balanceando a base de dados. Para mensurar resultados, o autor utilizou as métricas *F1 score* e *AUC*. Os resultados obtidos demonstraram um melhor desempenho no uso de *XGBoosting* em comparação com as demais técnicas de *ensemble* bem como com o *benchmark* em Rede Neural.

Em *A novel stacked generalization ensemble-based hybrid LGBM-XGB-MLP model for short-term load forecasting*, Massaoudi et al. [16] propõe um modelo para a previsão de carga elétrica de curto prazo, baseado em um *ensemble* de modelos LightGBM-XGBoosting-MLP. O modelo foi desenvolvido e validado utilizando dados de diferentes localizações, o primeiro *dataset* contendo 17519 amostras oriundas de uma base pública de uma indústria de fornecimento de energia localizada na Malásia e o segundo *dataset*

contendo 103775 registros oriundos da *ISO New England control area*, uma organização de transmissão regional de energia nos Estados Unidos. Destaca-se o uso de *Stacked Generalization (Stacking)* como abordagem de *ensemble*.

O trabalho foi realizado em uma sequência de fases de projeto iniciando pelo (1) estágio de engenharia de variáveis, (2) determinação do objetivo, (3) estágio de predição e por fim (4) estágio de avaliação. No estágio 3, destaca-se a aplicação do método de *Stacking* de forma a combinar os modelos LightGBM e XGBoosting onde os resultados são *input* para a camada do modelo MLP. De forma a avaliar a abordagem, foram utilizados os indicadores Rooted Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Squared Logarithmic Error (MSLE) e o coeficiente de determinação (R2). Os resultados obtidos demonstram uma maior *performance* do modelo *stacking* em relação aos modelos individualmente.

No trabalho *Deep and Cross network for ad click predictions*, Wang et al. [7] apresentou uma evolução da abordagem *Wide and Deep* proposta por [6]. Wang deu foco na engenharia de variáveis para o modelo, abordando uma deficiência das redes de aprendizado profundo DNN, que embora capazes de aprender automaticamente as interações entre as variáveis fazem isso de forma implícita, não sendo necessariamente eficientes no aprendizado.

O modelo *Deep & Cross Network (DCN)* introduz uma nova camada à DNN, uma rede cruzada, mais eficiente no aprendizado da interação entre as variáveis, aplicando de forma explícita o cruzamento entre as variáveis em cada camada da rede., sem que seja necessário o tratamento prévio dessas variáveis, adicionando uma complexidade extra insignificante ao modelo DNN. Os resultados experimentais demonstraram superioridade sobre os algoritmos de última geração no conjunto de dados, entre eles o *Wide and Deep* no conjunto de dados de classificação denso, em termos de precisão do modelo e uso de memória.

Em *Wide and Deep learning for recommender systems*, Cheng et al. [6] desenvolveu uma abordagem para sistemas de recomendação baseada em redes neurais chamada *Wide and Deep*. O modelo foi desenvolvido e testado no *Google Play*, uma loja comercial de aplicativos móveis com mais de um bilhão de usuários ativos e mais de um milhão de aplicativos. A abordagem consiste em uma composição de modelos. O modelo *Wide* é um modelo linear geral que aprende relações simples entre as variáveis de entrada. Tem a responsabilidade de capturar padrões simples em dados estruturados e categorizar eventos raros. O modelo *Deep* é um modelo não linear que aprende representações mais complexas das entradas. Ele é capaz de identificar padrões mais sutis e, portanto, é especialmente adequado para modelar dados não estruturados, como imagens e texto.

O modelo *Wide and Deep* combina os dois modelos de forma integrada, permitindo

que ambos compartilhem a entrada e sejam treinados em conjunto. A camada de saída do modelo *Wide and Deep* é uma combinação ponderada das saídas dos dois modelos, resultando em um modelo final que é capaz de capturar padrões simples e complexos. De forma a avaliar a abordagem proposta, foram utilizadas métricas de negócio *app acquisitions* e *serving performance*, além da métrica *AUC* de *performance* do modelo. O modelo *Wide and Deep* foi testado e comparado frente aos seus componentes *Wide-only* e *Deep-only* onde foi observada um importante ganho nos indicadores observados no uso da abordagem conjunta.

Em *Personalized ad recommendation systems for life-time value optimization with guarantees*, Theocharous et al. [17] propôs um modelo de recomendação personalizada de anúncios, de forma a otimizar o *Lifetime Value (LTV)* do cliente. Para tal foram utilizados duas bases de dados oriundas de sistemas de recomendação de dois bancos. Para a primeira amostra, foram coletados dados de uma determinada campanha de um banco durante um mês que teve 7 ofertas e aproximadamente 200.000 interações. Cerca de 20.000 das interações foram produzidas por uma estratégia aleatória. Para a segunda amostra, foram coletados dados de um banco diferente para uma campanha que teve 12 ofertas e 4.000.000 interações, das quais 250.000 foram produzidas por uma estratégia aleatória. Foram testadas duas abordagens, a primeira de curto prazo, chamada "gulosa"e a segunda de longo prazo, otimizando o *LTV*.

Para otimização gulosa, foi utilizado o algoritmo *Randon Forest (RF)* de forma a treinar um mapeamento de recursos para as ações. O sistema é treinado usando um modelo *RF* para cada uma das ofertas/ações para prever a recompensa imediata. Para a otimização do *LTV*, foi utilizado um algoritmo de *Reinforcement Learning (RL)* de última geração, chamado *Fitted Q-iteration (FQI)*, que permitiu lidar com variáveis contínuas e discretas de alta dimensionalidade. Como resultado, os autores relatam a adequabilidade da combinação das técnicas utilizadas, que otimizou o indicador *LTV* nos cenários testados.

Neste capítulo de trabalhos relacionados observa-se uma significante utilização de técnicas de combinação de modelos probabilísticos, gerando modelos *Ensemble* mais adequados e com melhor desempenho frente aos indicadores apurados através das técnicas de *Bagging*, *Stacking* e *Voting Ensemble*. Foram observados usos recorrentes das técnicas e algoritmos *Gradient Boosted Tree*, *Random Forest*, *Naïve Bayes*, *XGBoosting* e *LightGBM*. Quanto aos indicadores de avaliação de modelos, verificamos a utilização das métricas de Acurácia, *F1 score*, *Recall* e *AUC* de forma recorrente. A Tabela 3.1 apresenta um resumo dos trabalhos elencados neste capítulo.

Tabela 3.1: Estado da Arte

Autor	**Ano**	**Técnicas**	**Métricas**
Bilal et al. [11]	2022	*K-means clustering*, *K-medoids clustering*, *X-means clustering*, *Random clustering*, *K-nearest neighbor*, árvore de decisão, *Gradient boosted tree*, *Random forest*, *Deep learning classifier*, *Naïve Bayes*, *NB (K) (Naïve Bayes Kernel)* e *Ensemble classifiers (Voting, Bagging, AdaBoost e Stacking)*	Acurácia, Precisão, *Recall* e *F1 score*
Gattermann-Itschert et al.[12]	2022	*Random Forest*, Regressão Logística e *Support Vector Machine (SVM) Linear*	*AUC*, *Average Precision* e *Top-decile Lift*
Cao e Zhu [13]	2022	*Coupled Recurrent Network (CRN)*, *Wide & Deep Network (WDN)*, *Recurrent Neural Network (RNN)*, *Long Short Term Memory (LSTM)*, *Gated Recurrent Unit (GRU)*, *Rectified Linear Unit (ReLU)*, *Multi-layer Perceptron (MLP)* e *Adam Optimization Algorithm*	*Precision Lift*,*Reward Mean Squared Error (MSE)* e *Average Reward Lift*
Sun et al. [14]	2022	*Bagging*, *Light Gradient Boosting Machine (LightGBM)*, *Gradient Boosting*, *Classification and Regression Tree (CART)*	*Recall*, *Macro Recall* e Acurácia
Faraj et al. [15]	2021	*Ensemble*, *Bagging*, *Boosting*, *Stacking*, *XGBoosting*, Regressão Logística, *Ada Boost*, *Voting Ensemble* e Redes Neurais	*F1 score* e *AUC*

Massaoudi et al. [16]	2021	*Ensemble*, *Light Gradient Boosting Machine (LightGBM)*, *Extreme Gradient Boosting Machine (XGBoosting)* e *Multi-layer Perceptron (MLP)*	*Rooted Mean Squared Error (RMSE)*, *Mean Absolute Error (MAE)*, *Mean Absolute Percentage Error (MAPE)*, *Mean Squared Logarithmic Error (MSLE)* e Coeficiente de Determinação (R2)
Wang et al. [7]	2017	*Deep and Cross*	*Logloss*, *Accuracy* e *Memory Usage*
Cheng et al. [6]	2016	*Wide and Deep*	*APP Acquisitions*, *Serving Performance* e *AUC*
Theocharous et al. [17]	2015	*Randon Forest (RF)*, *Reinforcement Learning (RL)* e *Fitted Q-iteration (FQI)*	*Lifetime Value (LTV)*

Neste capítulo foram abordados trabalhos correlatos ao projeto elaborado, utilizando técnicas e métodos utilizados no desenvolvimento deste trabalho de pesquisa, bem como a sumarização de trabalhos que representam o estado da arte relacionado ao tema do projeto.

Capítulo 4

Modelo Preditivo: Melhor Canal de Interação

Neste capítulo serão relatados os procedimentos de planejamento e execução do estudo de caso de implementação do modelo preditivo de sucesso no melhor canal de interação com o cliente bancário. A condução do experimento será guiada pelo uso do CRISP-DM, exceto pela fase de implementação do modelo em produção, onde será gerada em substituição, uma proposta de implantação do modelo gerado. Iniciaremos descrevendo os aspectos de negócio envolvidos no caso a ser estudado. Posteriormente descreveremos as atividades relacionadas aos dados utilizados no processo de mineração, como as atividades de compreensão e preparação de dados. Posteriormente as atividades de modelagem serão descritas. Por fim serão descritos os procedimentos de avaliação dos modelos gerados.

Para a execução estruturada do processo de Mineração de Dados foi escolhido o método CRISP-DM devido a sua robustez e ampla aceitação no mercado e academia. Para a geração do modelo foram utilizadas técnicas de *ensemble* de forma a obter melhores resultados através da combinação de diferentes algoritmos e modelos preditivos. Foi considerado também a aplicação dos modelos *Wide and Deep* e *Deep and Cross* para a previsão do melhor canal de interação. Para a avaliação dos experimentos comparativos realizados, foram utilizadas como métrica final a acurácia dos modelos gerados.

4.1 Compreensão do Negócio

De forma a atender as necessidades negociais do banco alvo deste estudo de caso, foram realizadas reuniões de entendimento do negócio junto aos especialistas das áreas de atendimento, relacionamento e *marketing* da instituição. A partir dessas reuniões, o escopo de trabalho pôde ser definido, pelo tipo de cliente contactado e o período das interações realizadas.

Foram, portanto, definidos os escopos de tipo de cliente Pessoa Física da organização, bem como o período de interações relativo ao primeiro semestre do ano de 2022, para os respectivos clientes selecionados.

4.2 Análise dos Dados

Os dados obtidos para a etapa de Mineração de Dados deste estudo de caso são oriundos principalmente de duas das bases de dados de um grande banco comercial do país. Uma das bases contém informações cadastrais e de perfil do cliente, a outra base utilizada neste estudo contém dados históricos das interações realizadas com os clientes nos diversos canais de comunicação disponíveis na instituição.

4.2.1 Dicionário de Dados

Foram coletados 116208 dados cadastrais e de perfil de clientes pessoa física, contendo 41 variáveis por registro, e um total de 9091875 registros de interações, contendo 12 variáveis por registro de interação relacionados ao clientes amostrados. As Tabelas 4.1 e 4.2 descrevem as variáveis relacionadas a cada base de dados de perfil de cliente e interações, respectivamente.

Tabela 4.1: Variáveis de Perfil do Cliente

Nome	**Descrição**
CD_SEXO	Sexo
DT_NSC	Data de nascimento
CD_EST_CVL	Estado civil
IN_TEL_CADD	Indicador de telefone cadastrado
IN_ENVO_MSG_CELR	Indicador de envio de mensagem por celular
CD_NTZ_OCP	Código da natureza da ocupação
CD_OCP	Código da ocupação
VL_REND	Renda
CD_RSCO_CRD_CLI	Código de risco
DT_INC_PRMO_CCFX	Data do início de relacionamento
CD_EST_CC	Código de estado da conta corrente
CD_EST_CAD	Código de estado do cadastro
CD_EST_LMCR	Código de estado do limite de crédito
CD_SGM_CLI	Código do segmento do cliente
CD_MDLD_PKGE_SRVC	Código da modalidade do pacote de serviço

CD_RCBT_BNF_BB	Código de recebimento de benefício
VL_MGCT_OBSD	Margem de contribuição
IN_PSSE_CHQ_ESPL	Indicador de posse de cheque especial
IN_PSSE_POUP	Indicador de posse de poupança
IN_PSSE_DEPZ	Indicador de posse de depósito a prazo
IN_PSSE_FNDO_INVS	Indicador de posse de fundo de investimento
IN_PRVA_ABTO	Indicador de posse de previdência privada
IN_PSSE_SGRO_VCLO	Indicador de posse de seguro veículo
IN_PSSE_SGRO_PSS	Indicador de posse de seguro de pessoas
IN_PSSE_CPTZ_MSL	Indicador de posse de capitalização mensal
IN_PSSE_CPTZ_UNCO	Indicador de posse de capitalização pagamento único
IN_PSSE_CSR_VCLO	Indicador de posse de consórcio veículo
IN_PSSE_CSR_MOT	Indicador de posse de consórcio motocicleta
IN_PSSE_CSR_ETN	Indicador de posse de consórcio eletrônicos
IN_PSSE_CMZC_AGRP	Indicador de posse de comercialização agropecuária
IN_PSSE_CTE_AGRP	Indicador de posse de custeio agropecuário
IN_PSSE_INVS_AGRP	Indicador de posse de investimento agropecuário
IN_PSSE_PNF_CTE	Indicador de posse de PRONAF custeio
IN_PSSE_PNF_INVS	Indicador de posse de PRONAF investimento
IN_PSSE_CRT_CRD	Indicador de posse de cartão de crédito
IN_CRD_VCLO_PRPP	Indicador de posse de crédito veículo próprio
IN_PSSE_CRD_VCLO	Indicador de posse de crédito veículo
IN_PSSE_LCA	Indicador de posse de LCA
IN_PSSE_LCI	Indicador de posse de LCI
IN_PSSE_SGRO_PATR	Indicador de posse de seguro patrimônio
IN_PSSE_FNTO_ETDL	Indicador de posse de financiamento estudantil

Tabela 4.2: Variáveis de Interação com o Cliente

Nome	**Descrição**
CD_TIP_CNL	Código do tipo de canal
NM_TIP_CNL	Nome do tipo de canal
CD_ASNT_INRO	Código do assunto da interação
TX_ASNT_INRO	Texto do assunto da interação
CD_SUB_ASNT_INRO	Código do sub-assunto da interação

TX_SUB_ASNT_INRO	Texto do sub-assunto da interação
CD_TRAN_INRO_SIS	Código transação em sistema
TX_TRAN_INRO_SIS	Texto transação em sistema
CD_RSTD_INRO	Código do resultado da interação
TX_RSTD_INRO	Texto do resultado da interação
CD_SUB_RSTD_INRO	Código do sub-resultado da interação
TX_SUB_RSTD_INRO	Texto do sub-resultado da interação

4.2.2 Engenharia de Dados

Uma primeira etapa de engenharia e preparação de dados foi executada logo após a extração dos dados nas bases corporativas da instituição. Neste ponto torna-se fundamental salientar que, todos os processos de governança corporativa de dados foram seguidos, obtendo-se os acessos necessários por meio de papéis corporativos às bases da instituição, bem como respeitando regras de não identificação de dados referentes à informações pessoais de clientes.

A seguir, na Tabela 4.3, estão listadas as atividades referentes à engenharia e preparação de dados executadas na base de dados amostrada.

Tabela 4.3: Atividades de Engenharia de Dados

Atividade	**Descrição**
Exclusão de código identificador original do cliente	Exclusão do código identificador original, de forma a não ser possível identificar o cliente relacionado ao registro amostrado.
Remoção de valores nulos	Exclusão de linhas com valores nulos nas variáveis de interesse NM_TIP_CNL, CD_SEXO e CD_RSCO_CRD_CLI.

Substituição de indicadores por valores numéricos	Substituição dos valores textuais N e S por numéricos 0 e 1 nos indicadores de posse de produtos representado pelas variáveis IN_TEL_CADD, IN_ENVO_MSG_CELR, IN_PSSE_CHQ_ESPL, IN_PSSE_POUP, IN_PSSE_DEPZ, IN_PSSE_FNDO_INVS, IN_PRVA_ABTO, IN_PSSE_SGRO_VCLO, IN_PSSE_SGRO_PSS, IN_PSSE_CPTZ_MSL, IN_PSSE_CPTZ_UNCO, IN_PSSE_CSR_VCLO, IN_PSSE_CSR_MOT, IN_PSSE_CSR_ETN, IN_PSSE_CMZC_AGRP, IN_PSSE_CTE_AGRP, IN_PSSE_INVS_AGRP, IN_PSSE_PNF_CTE, IN_PSSE_PNF_INVS, IN_PSSE_CRT_CRD, IN_CRD_VCLO_PRPP, IN_PSSE_CRD_VCLO, IN_PSSE_LCA, IN_PSSE_LCI, IN_PSSE_SGRO_PATR e IN_PSSE_FNTO_ETDL
Substituição de indicadores por valores numéricos	Substituição dos valores textuais M e F por numéricos 0 e 1 na variável CD_SEXO
Substituição de indicadores por valores numéricos	Substituição dos valores textuais (A e A+), (B e B+), (C e C+), (D e D+), (E e E+), por numéricos 0, 1, 2, 3 e 4 respectivamente, na variável CD_RSCO_CRD_CLI, atribuindo -1 àqueles registros sem risco classificado.
Remoção de canais sem relevância	Canais internos de comunicação ou com indicativo de sistemas de controle foram removidos, a partir da deleção dos seguintes códigos em CD_TIP_CNL: (14, 25, 31, 44, 16, 23, 17).
Agrupamento de canais correlatos	Realizado o agrupamento de canais de forma a obter o seguinte grupo de interesse TAA, WEB, WAPP, SMS e FONE.
Remoção de datas irrelevantes	Exclusão de registros contendo 9999-12-31 nas variáveis DT_NSC e DT_INC_PRMO_CC.
Cálculo de idade atual e tempo de relacionamento atual	Transformação das variáveis DT_NSC e DT_INC_PRMO_CC em tempo em anos até a data atual, gerando as novas variáveis IDD_ATUAL e TEMPO_RLC_ATUAL e excluindo as datas originais.
Conversão de valores de pontos flutuantes	Conversão das variáveis VL_MGCT_OBSD e VL_RENDA representas por pontos flutuantes na base de dados para valores inteiros, truncando a parte decimal.

Obtenção dos indicativos de sucesso na interação	Implementação de filtro negocial a partir das variáveis `CD_RSTD_INRO` e `CD_SUB_RSTD_INRO`, gerando o indicador de sucesso na interação, na variável `IN_SUCC` definido pelos valores 0 e 1, para insucesso e sucesso na interação, respectivamente.
Separação de arquivos completo e amostra	Criação de dois arquivos, o primeiro contendo os 7613962 registros resultantes da etapa de engenharia e transformação dos dados, e um segundo arquivo contendo uma amostra de 5% do total, 380698 registros, ponderando essa amostragem pela quantidade de sucessos e insucessos na comunicação.

4.2.3 Análise Exploratória

Após a aquisição e preparação inicial dos dados, foi realizada uma etapa de entendimento dos dados coletados, analisando de forma exploratória a base de dados gerada na etapa de engenharia de dados. As subseções a seguir descrevem a análise exploratória sobre os dados de perfil dos clientes e das suas interações.

Perfil do Cliente

Para entender o perfil dos clientes da amostra realizada, variáveis demográficas e de perfil de negocial foram representadas visualmente. A Figura 4.1 exibe um histograma com as faixas etárias dos clientes amostrados, em que se percebe uma predominância de clientes entre 60 e 70 anos no conjunto de dados coletado.

Quanto à distribuição entre sexo dos clientes, se visualiza na Figura 4.2 uma certa igualdade na amostra, não havendo predominância.

No perfil de renda do cliente, percebe-se na Figura 4.3 uma maior concentração de clientes na faixa salarial de 0 a 10000 reais, tendendo à diminuição na medida que o valor de renda aumenta.

Na distribuição entre perfis de risco de crédito dos clientes no conjunto de dados amostrado percebe-se uma maior concentração no perfil de risco `C`, seguido por `B`, `E`, `D` e `A`. Os códigos de limites de crédito utilizados na Figura 4.4 representam uma abstração de risco atribuído ao cliente.

O mesmo conceito é utilizado na Figura 4.5 para representar os segmentos negociais dos clientes. São utilizados códigos representativos dos segmentos negociais. Percebemos uma maior concentração nos segmentos `PF C`, seguido por `PF B`, `PF E`, `PF D` e `PF A`.

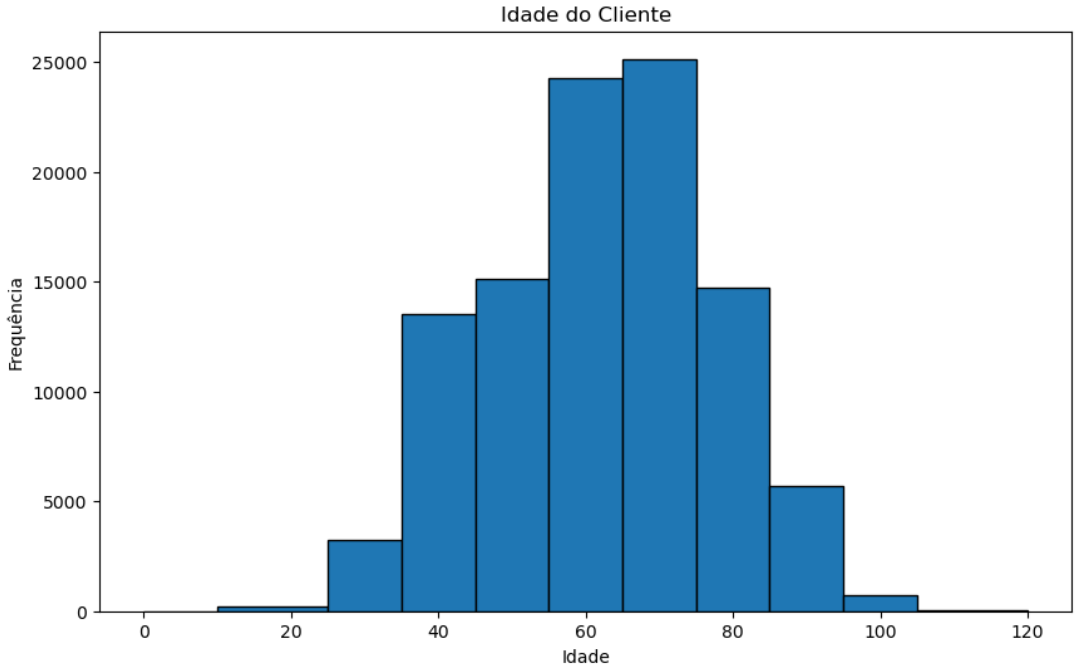

Figura 4.1: Faixas Etárias do Cliente.

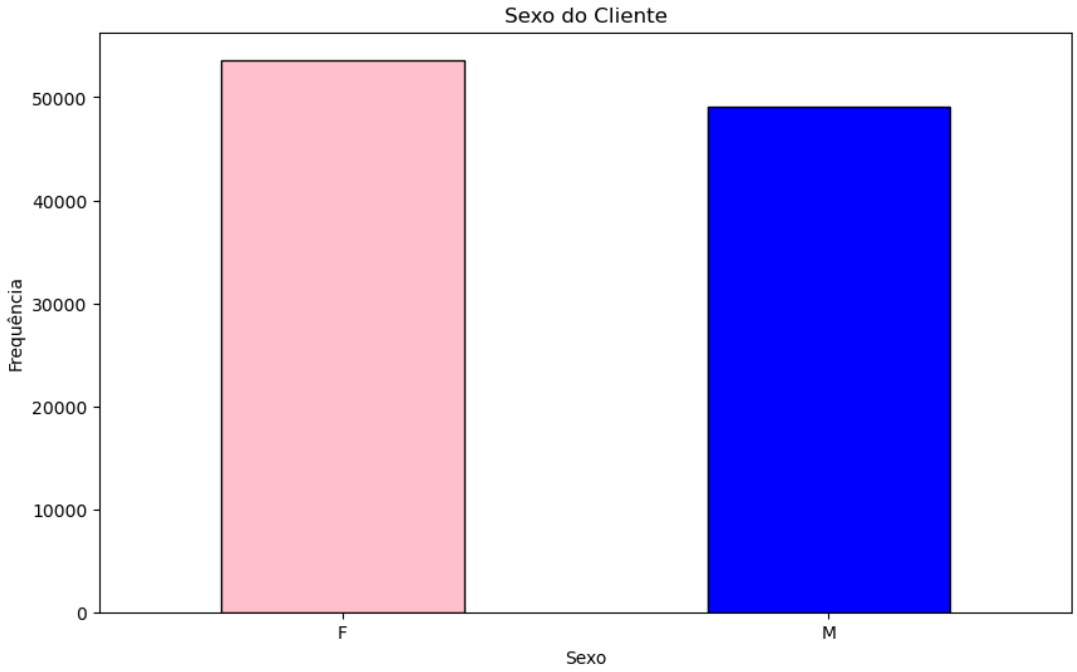

Figura 4.2: Sexo do Cliente.

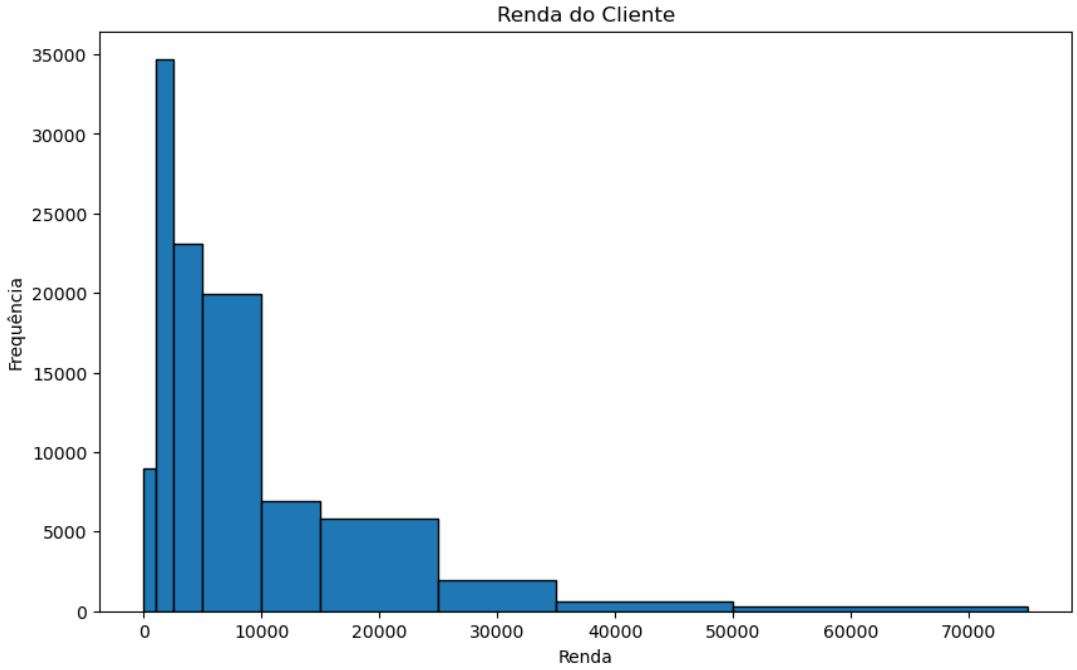

Figura 4.3: Faixas de Renda do Cliente.

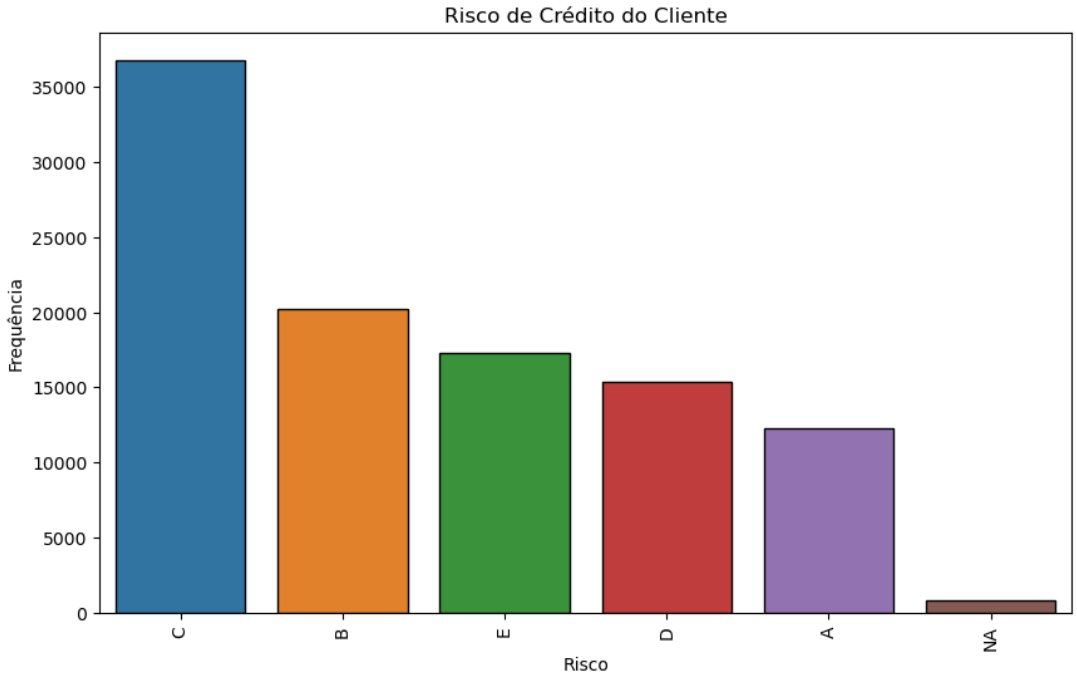

Figura 4.4: Risco de Crédito do Cliente.

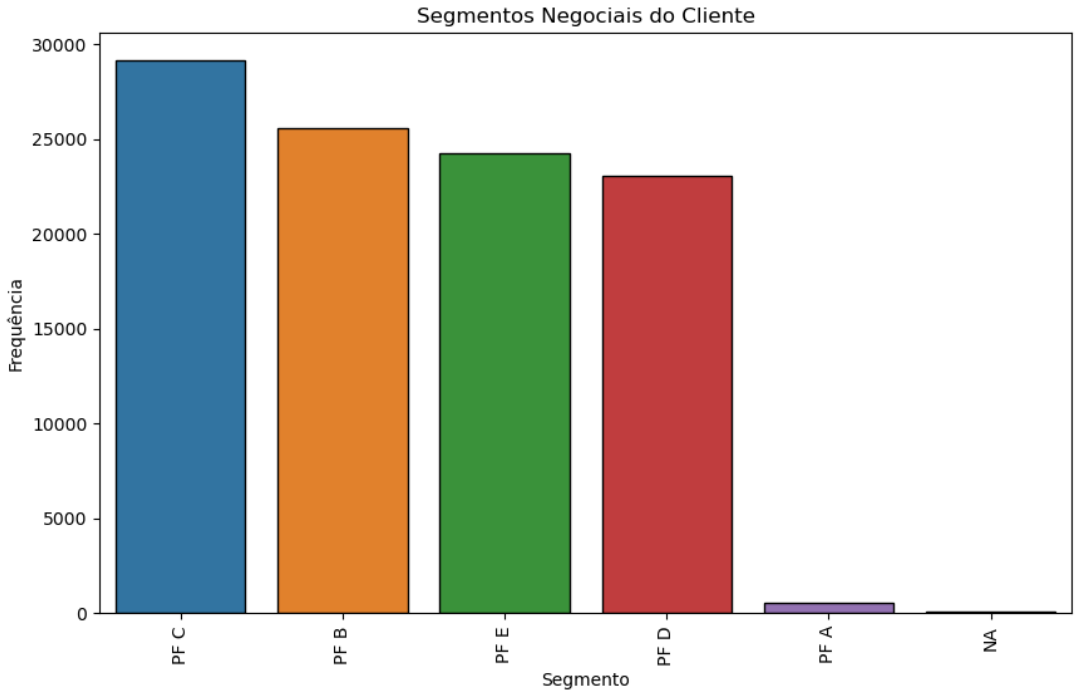

Figura 4.5: Segmentos Negociais.

O tempo de relacionamento negocial é representado na Figura 4.6 onde se observa uma preponderância de relacionamento entre 10 e 20 anos.

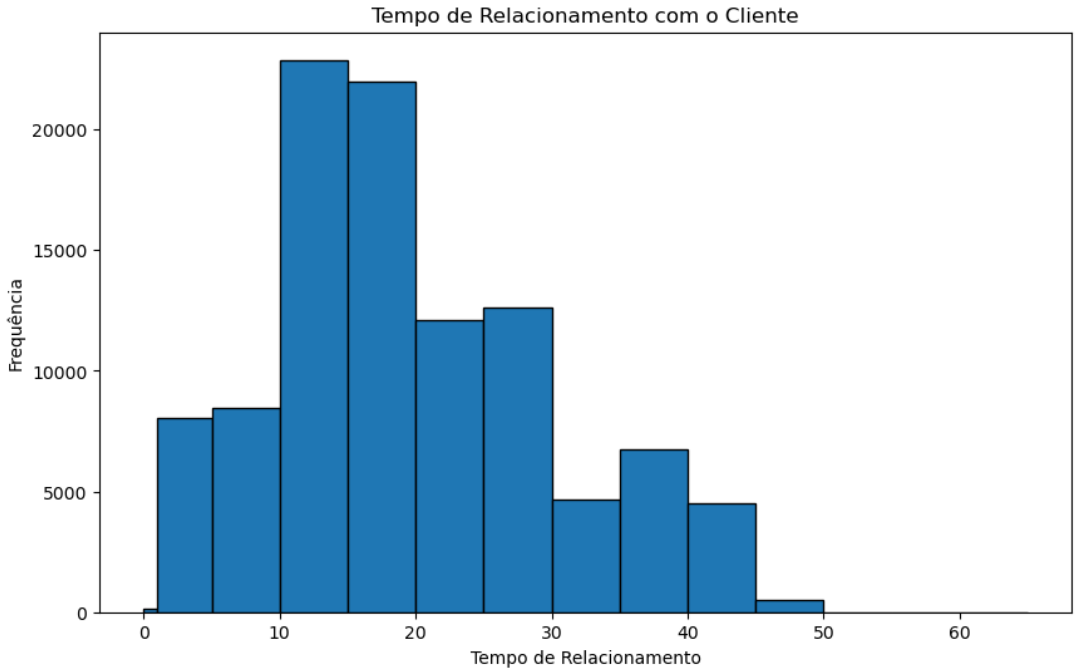

Figura 4.6: Tempo de Relaciomento Negocial.

Por fim, a Figura 4.7 demonstra a contagem de clientes em cada faixa de margem de contribuição. Esta é uma variável chave no relacionamento com o cliente, sendo esta uma variável dependente candidata em um modelo de maximização. Pode-se visualizar uma maior concentração entre 0 e 250 reais positivos em margem de contribuição.

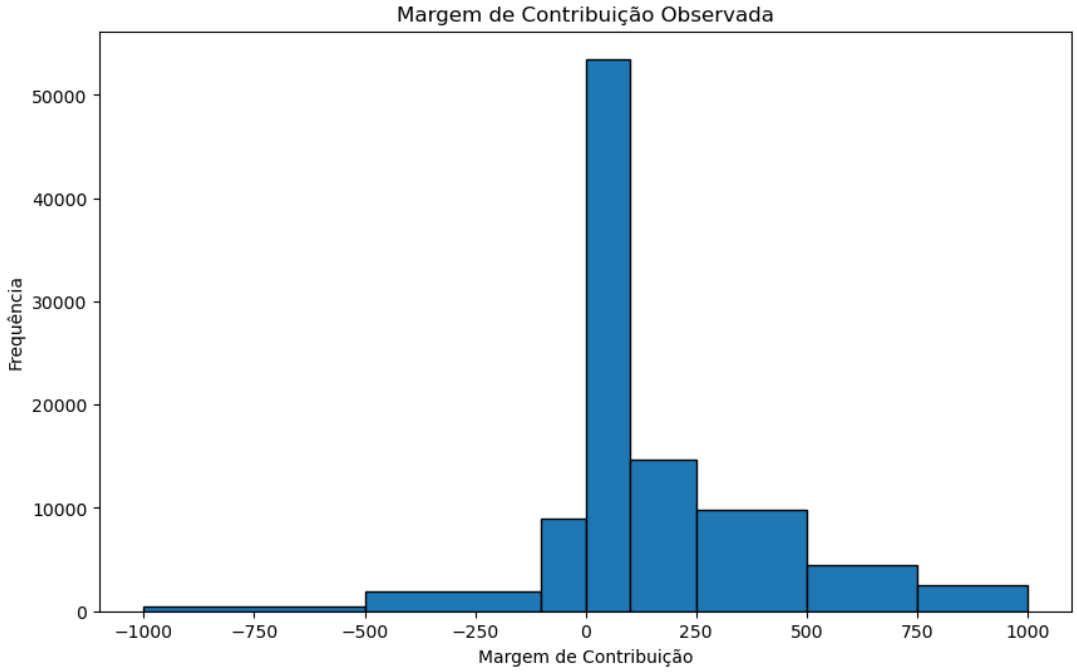

Figura 4.7: Margem de Contribuição do Cliente.

Interações do Cliente

No entendimento das informações de interação com o cliente, foram representadas visualmente as variáveis de: canais, assunto, sub assunto e o indicativo de sucessos nas interações.

Na Figura 4.8 visualizam-se os canais de interação listados no conjunto de dados. Esta é uma das principais variáveis dependentes candidatas para o modelo de melhor canal a ser gerado neste trabalho. Percebe-se uma predominância nas interações realizadas por `FONE`, seguido por `WAPP`, `TAA`, `WEB` e `SMS`.

Os assunto e sub assuntos de uma interação dizem respeito ao motivo daquele contato com o cliente. As Figura 4.9 e Figura 4.10 mostram uma predominância de assuntos relacionais e relativos a produtos como conta corrente, cartão de crédito e captação. Percebe-se na visualização de sub assuntos, uma predominância do registro `Outros`, o que pode vir a impactar negativamente a qualidade do modelo de melhor canal.

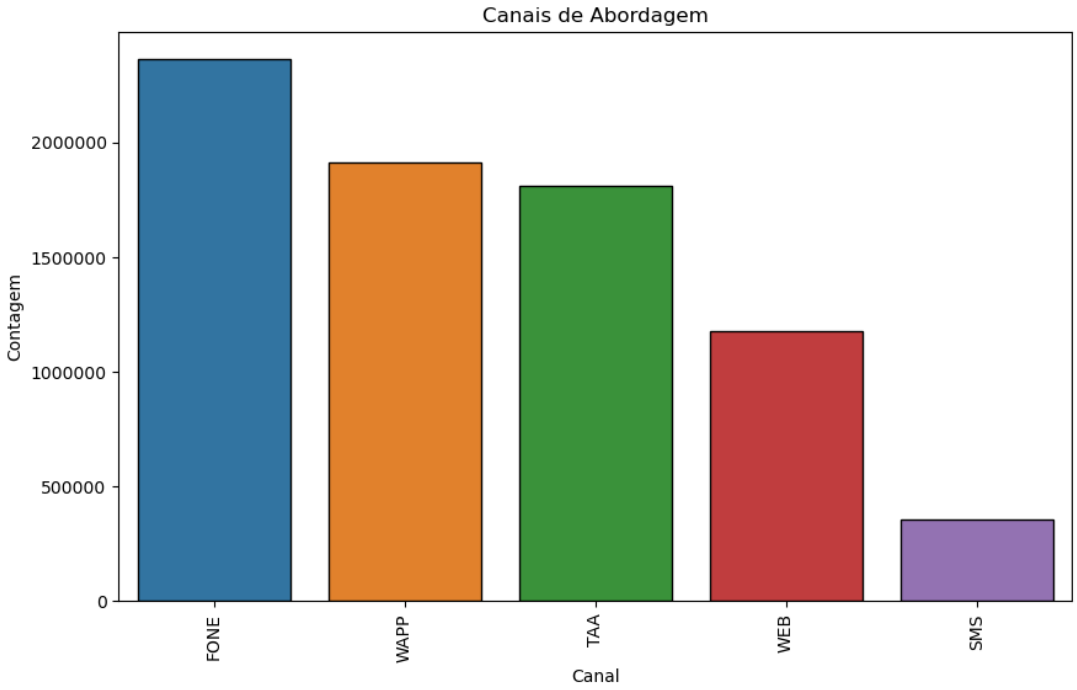

Figura 4.8: Canal da Interação com o Cliente.

A variável que indica o sucesso ou não da abordagem realizada ao cliente, observada na Figura 4.11, é uma das variáveis dependentes candidatas, principalmente numa abordagem de classificação binária quanto ao sucesso da utilização de um determinado canal de interação. Observa-se uma leve desbalanceamento na quantidade de interações com sucesso, podendo representar uma preocupação adicional na geração do modelo de melhor canal de interação.

Escolha de Variáveis

De forma a subsidiar a escolha das variáveis mais adequadas para a modelagem, foram verificados os coeficientes de correlação de Pearson [10] entre as variáveis do perfil do cliente e de interações, podendo ser observadas nas Figura 4.12 e Figura 4.13.

As correlações observadas variam entre -0.4 e 0.6 demonstrando uma correlação de fraca a moderada entre estas variáveis.

Conclui-se nesta análise que as variáveis em uso no conjunto de dados não possuem coeficientes de correlação que possam gerar impacto negativo na sua utilização para a modelagem do melhor canal de interação.

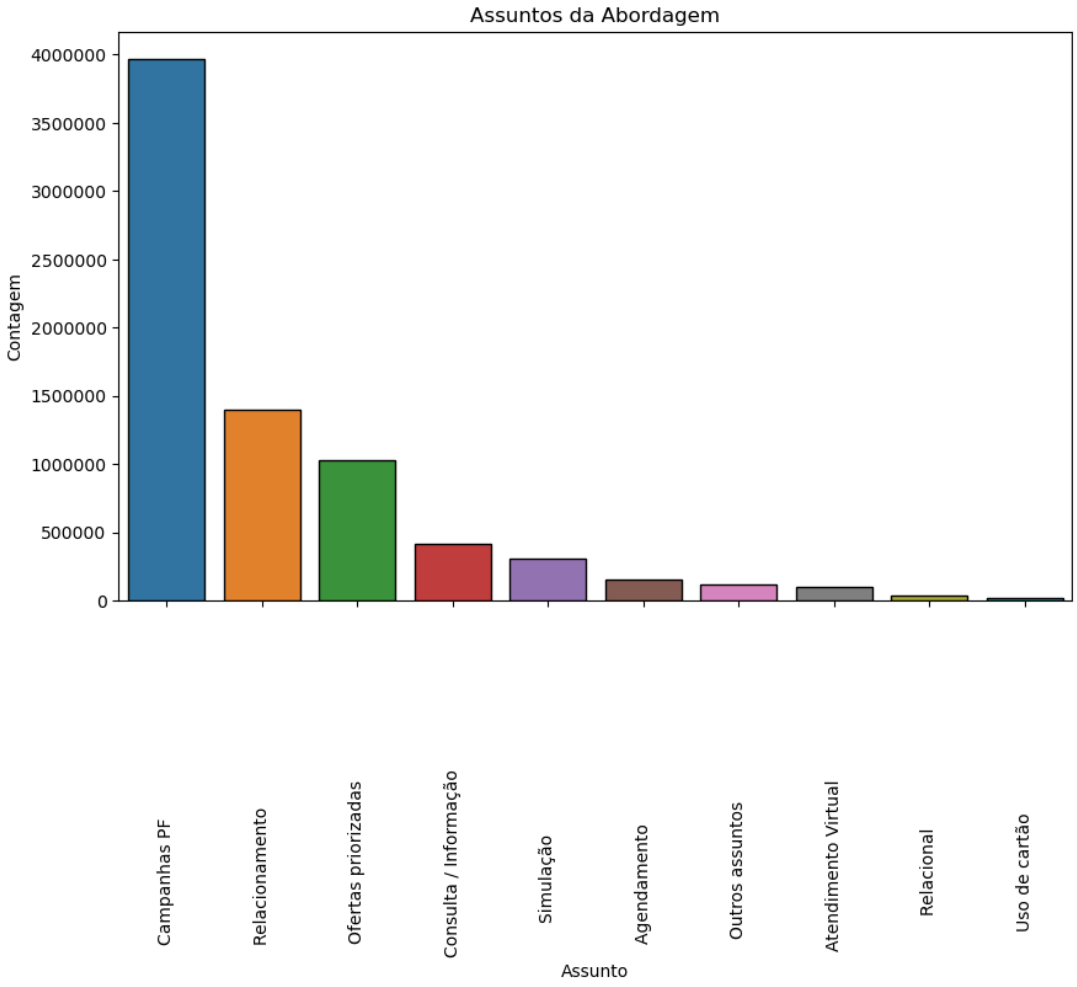

Figura 4.9: Assunto da Interação do Cliente.

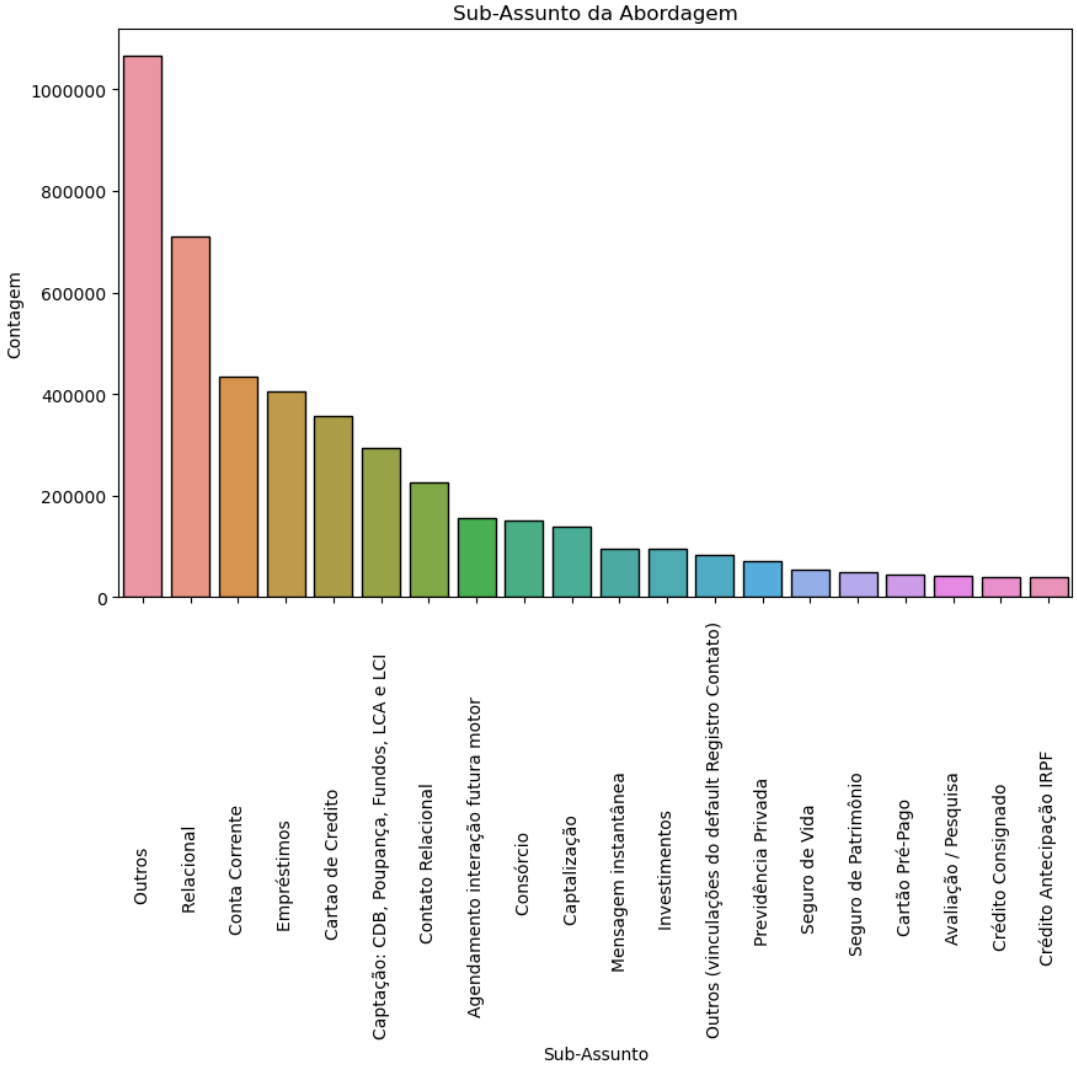

Figura 4.10: Sub-Assunto da Interação do Cliente.

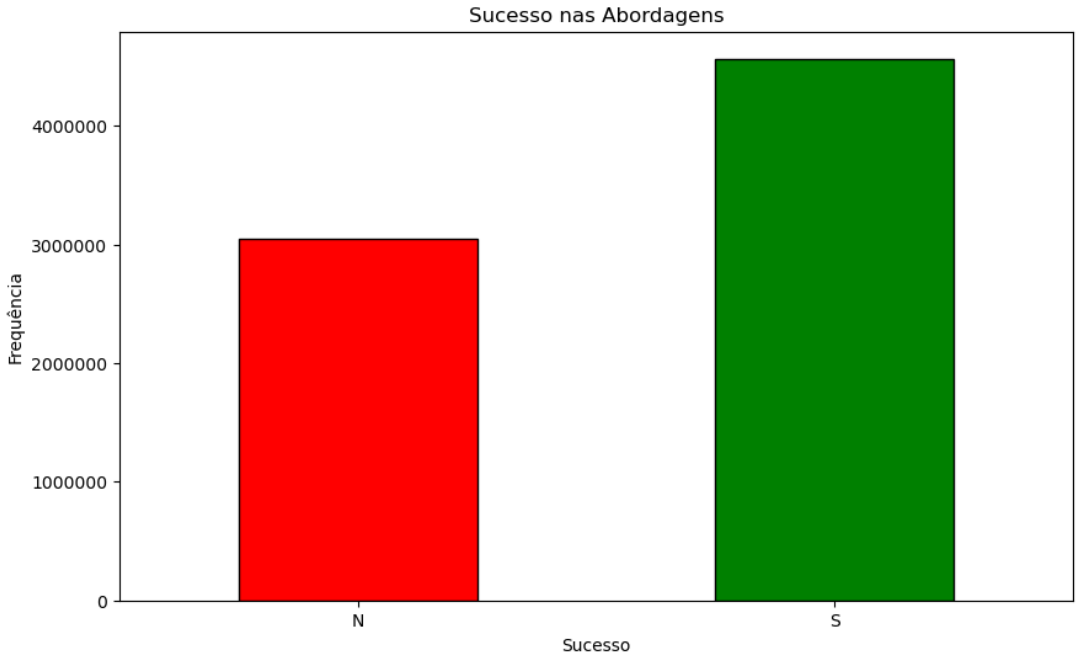

Figura 4.11: Sucesso na Interação com o Cliente.

4.3 Construção do Modelo de Melhor Canal de Interação

Na etapa de modelagem foram utilizadas duas abordagens: classificação binária e classificação multi classes. Em ambas as aplicações foram a utilizados uma série de algoritmos e técnicas de classificação, incluindo abordagens probabilísticas, *ensemble* de modelos e redes neurais. As sub-subseções a seguir descrevem as abordagens e resultados obtidos em cada experimento.

Especificações do Ambiente do Experimento

Para as etapas de engenharia de dados, preparação, análise exploratória, modelagem e treinamento foi utilizada uma máquina *Windows 10 Home Single Language* na versão *22H2*, equipada com processador *AMD Ryzen 5 3500X 6-Core Processor 3.95 GHz* e 36GB de RAM instalada.

Os experimentos foram realizados em ambiente *Python 3.9.13*, com a utilização principal das bibliotecas *Pycaret*, *TensorFlow* e *Keras* para modelagem e treinamento.

Os tempos de processamento, considerando o conjunto de dados total, podem ser verificados na Tabela 4.4 a seguir.

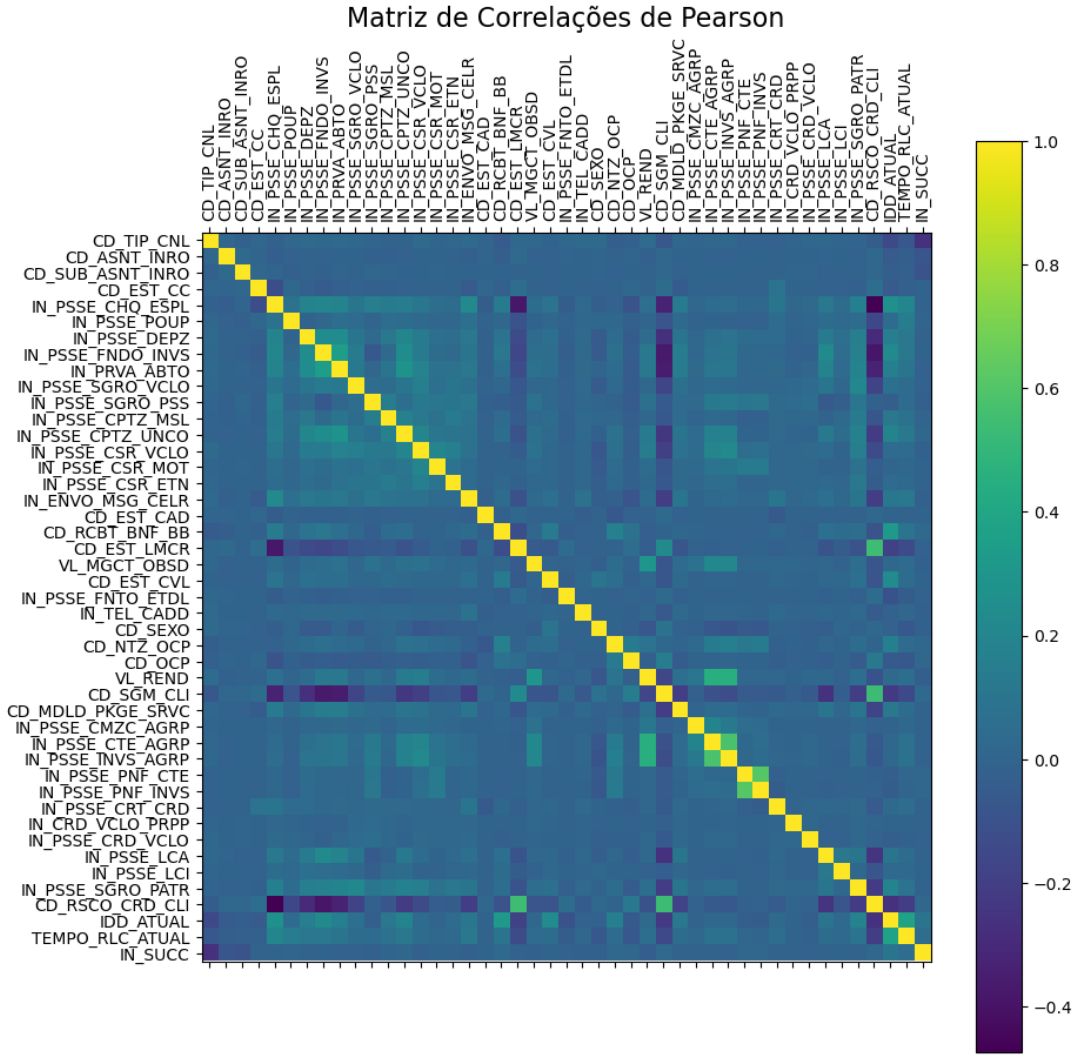

Figura 4.12: Matriz de Correlações de Pearson.

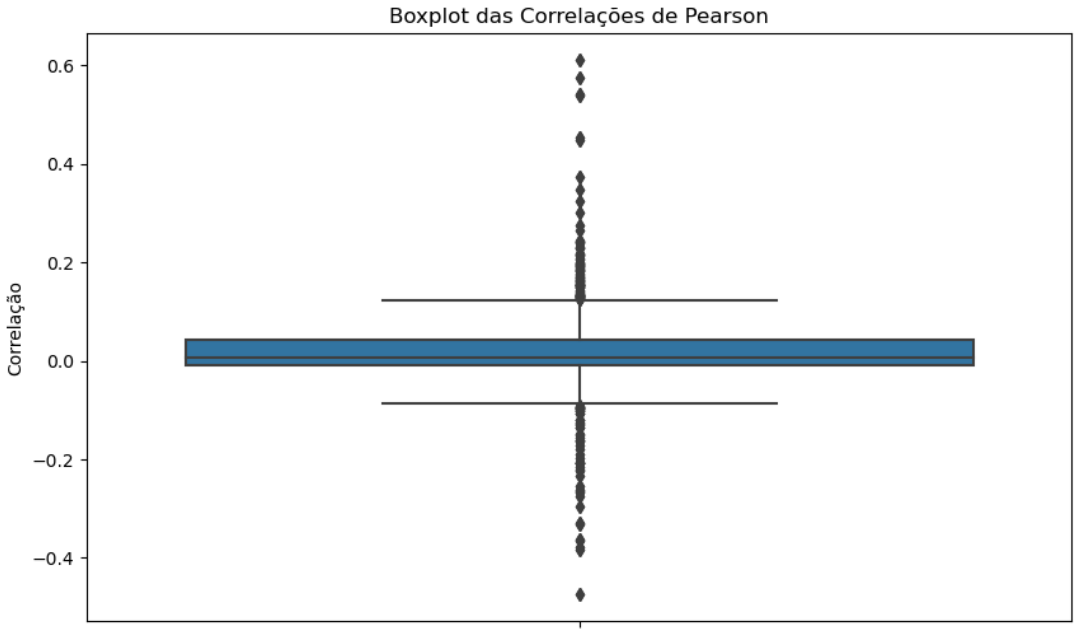

Figura 4.13: Boxplot de Correlações de Pearson.

Tabela 4.4: Tempo de Processamento

Atividade	**Tempo de Processamento**
Preparação de Dados	12.47 minutos
Análise Exploratória	1.66 minutos
Treinamento Classificador Binário Probabilístico	659.20 minutos
Treinamento Classificador Binário Redes Neurais	266.83 minutos
Treinamento Classificador Multi Classe Probabilístico	635.94 minutos
Treinamento Classificador Multi Classe Redes Neurais	98.37 minutos

Classificação Binária

Para a modelagem binária, foi considerado como variável dependente o indicador de sucesso da abordagem, representado por `IN_SUCC`. Sendo que, a partir das variáveis de perfil e interações realizadas, foi elaborado um modelo preditivo para antecipar o sucesso de uma determinada interação. Este modelo torna mais simples a modelagem, e permite a utilização de todo o conjunto de dados no aprendizado de máquina, em detrimento do multi classes, que se vale somente da porção de dados que indica sucesso na interação.

Os dados foram divididos em base de treinamento e teste, na proporção de 70% e 30% respectivamente, utilizando *Stratified K-Fold* para a validação cruzada, onde $K = 10$.

O modelo foi primeiramente treinado utilizando utilizando os algoritmos de classificação probabilísticos e de *ensemble* conforme a Tabela 4.5. A mesma tabela mostra um comparativo entre as principais métricas para avaliação do modelo de classificação. O objetivo nesta etapa foi de maximização da acurácia do modelo.

Tabela 4.5: Treinamento dos Modelos Probabilísticos: Abordagem Classificador Binário

Modelo	**Acurácia**	**AUC**	**Recall**	**Precision**	**F1**
Random Forest Classifier	0.8497	0.9431	0.9087	0.8509	0.8788
Decision Tree Classifier	0.8462	0.9358	0.8937	0.8562	0.8746
Extra Trees Classifier	0.8456	0.9366	0.8964	0.8536	0.8745
CatBoost Classifier	0.8456	0.9406	0.9566	0.8172	0.8814
Extreme Gradient Boosting	0.8376	0.9355	0.9592	0.8066	0.8763
Light Gradient Boosting Machine	0.8346	0.9337	0.9603	0.8027	0.8744
Gradient Boosting Classifier	0.8308	0.9294	0.9587	0.7993	0.8717
Ada Boost Classifier	0.8220	0.9217	0.9832	0.7788	0.8689
Quadratic Discriminant Analysis	0.6291	0.6997	0.9205	0.6307	0.7485
Naive Bayes	0.6119	0.6451	0.9533	0.6136	0.7466
Dummy Classifier	0.5998	0.5000	1.0000	0.5998	0.7499
Logistic Regression	0.5928	0.6253	0.8384	0.6186	0.7100
SVM - Linear Kernel	0.5830	0.0000	0.7564	0.6385	0.6730
Ridge Classifier	0.5657	0.0000	0.7342	0.6157	0.6698
Linear Discriminant Analysis	0.5652	0.6368	0.7261	0.6169	0.6670

Dentre os modelos treinados, o de melhores resultados iniciais foi o *Random Forest Classifier*, com acurácia inicial de 0.8497. Em seguida foi realizada uma etapa de otimização do modelo, utilizando *Bootstrap* com 100 estimadores, e critério de *Gini* no ajuste do modelo. Após a etapa de otimização, foi obtido o seguinte resultado ao aplicar o modelo na base completa, podendo ser verificado na Tabela 4.6.

Tabela 4.6: Treinamento dos Modelos Probabilísticos: Abordagem Classificador Binário

Modelo	**Acurácia**	**AUC**	**Recall**	**Precision**	**F1**
Random Forest Classifier	0.8659	0.9557	0.9229	0.8630	0.8920

Na Figura 4.14 se visualiza a matriz de confusão referente ao modelo otimizado. Nesta matriz verifica-se o números de falsos positivos, falsos negativos, verdadeiros positivos e verdadeiros negativos.

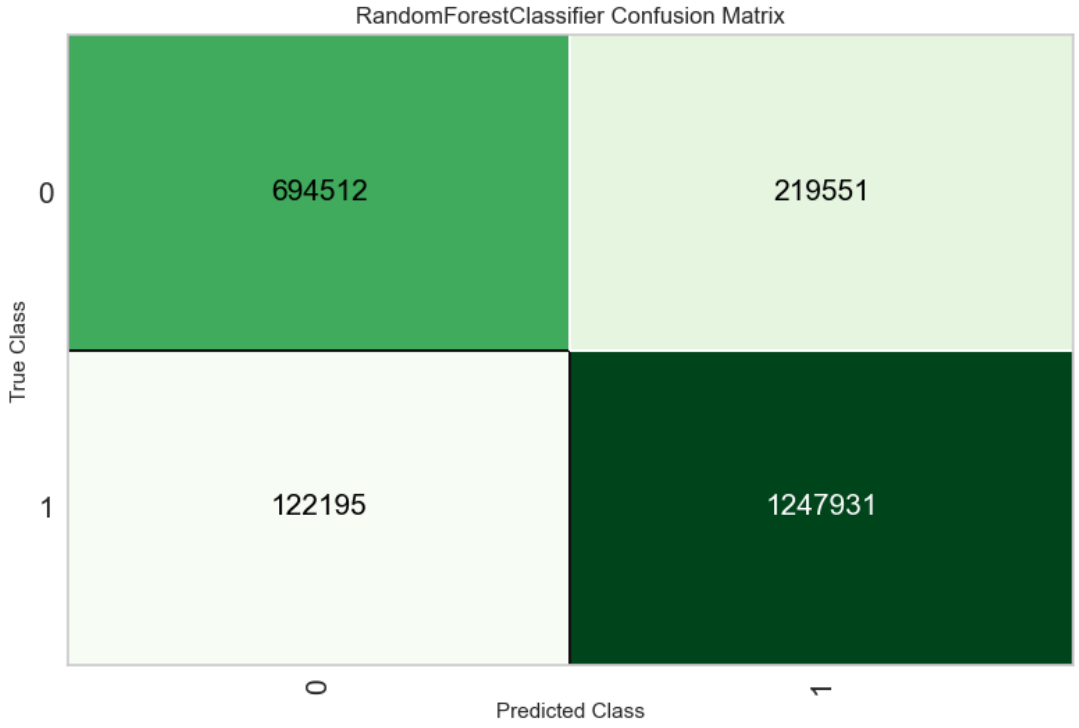

Figura 4.14: Classificador Binário: Matriz de Confusão.

Na Figura 4.15 se observa a curva ROC referente ao modelo otimizado. Nesta curva observa-se o desempenho do modelo em diferentes pontos de corte. A partir da curva ROC se obtém o indicador AUC. Quanto maior a área sob a curva (AUC), melhor é o desempenho do modelo.

Na Figura 4.16 se visualiza as principais variáveis utilizadas no modelo e seus níveis de importância.

Na Figura 4.17 se observa a curva de precisão versus recall do modelo otimizado, obtendo uma precisão média de 0.96.

Na Figura 4.18 se visualiza a proporção dos erros de previsão de classes no modelo otimizado.

Na Figura 4.19 observa-se a estatística KS (Kolmogorov-Smirnov), utilizada para avaliar a discrepância entre duas distribuições de probabilidade cumulativa, comumente utilizado em problemas de classificação binária para avaliar a eficácia discriminativa de um

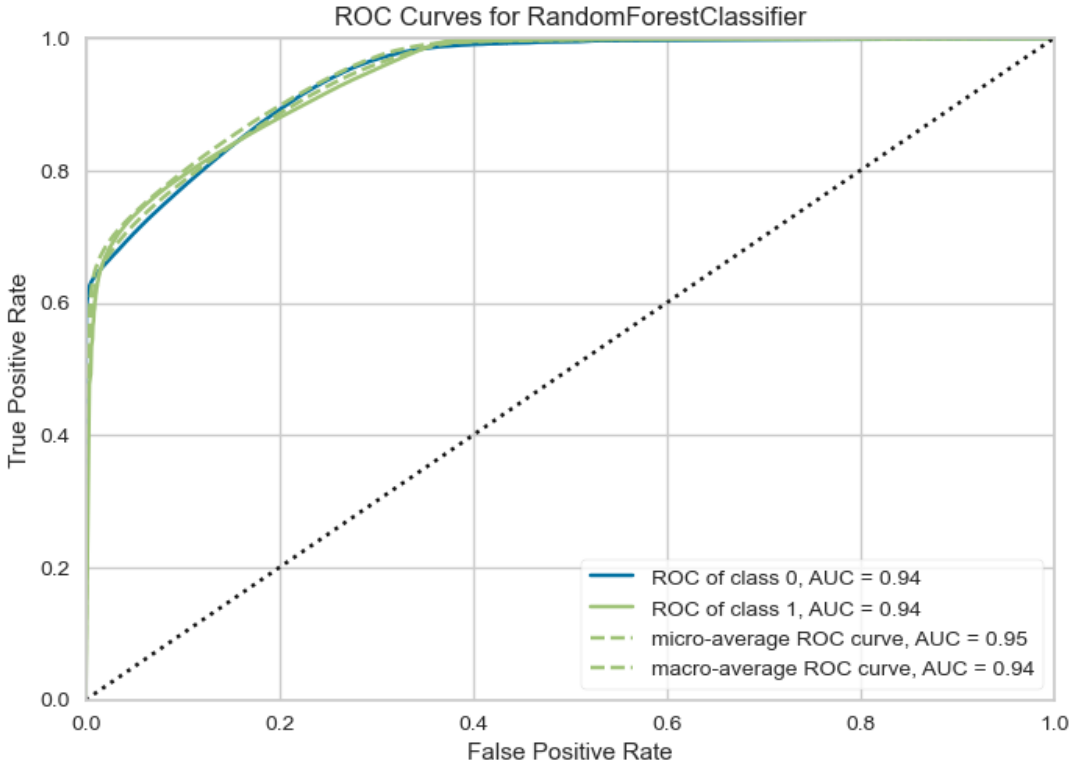

Figura 4.15: Classificador Binário: Curva ROC.

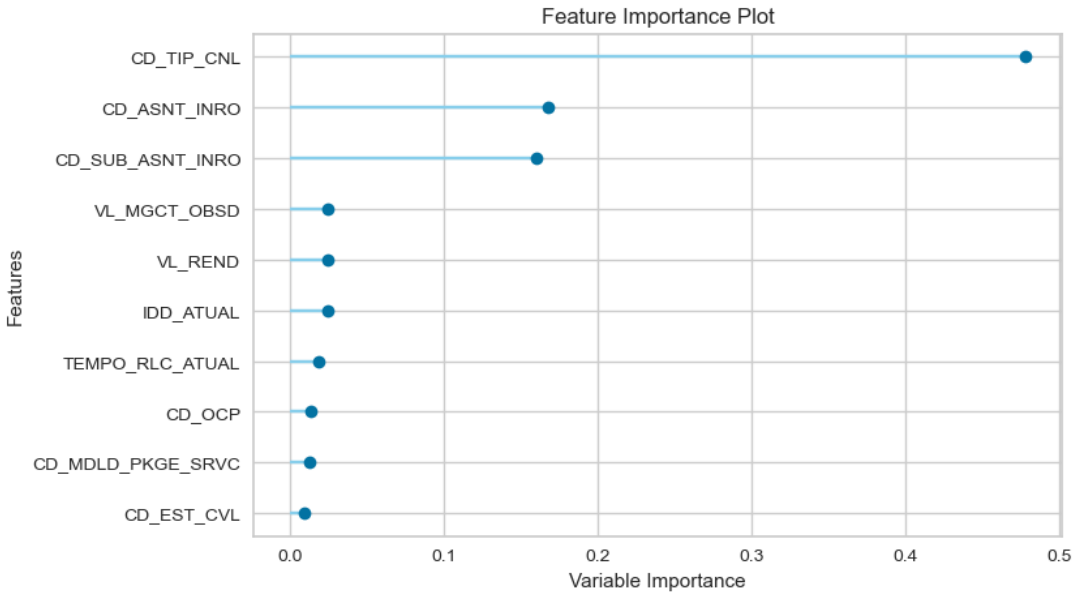

Figura 4.16: Classificador Binário: Variáveis de Maior Relevância.

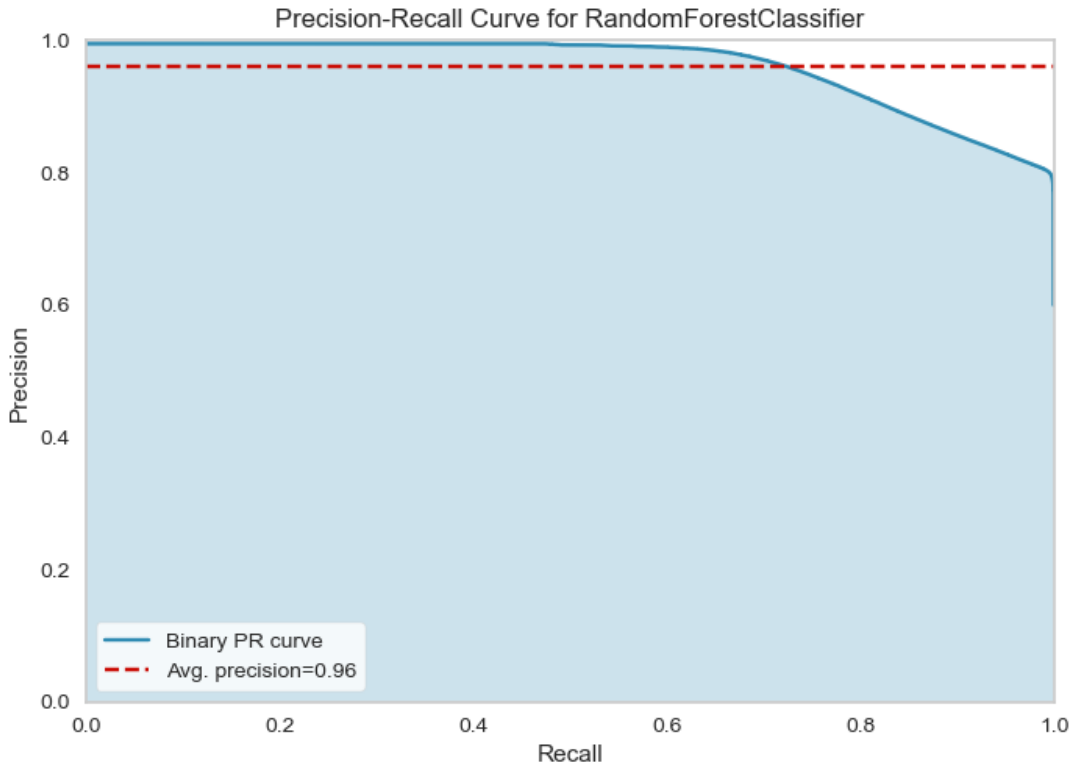

Figura 4.17: Classificador Binário: Precision-Recall.

modelo de classificação em relação à distribuição de probabilidade cumulativa dos eventos positivos e negativos.

Por fim, na Figura 4.20 se visualizam os indicadores de precisão, recall, F1 e suporte, por classes no modelo otimizado.

Em seguida, o modelo de classificação binária foi treinado utilizando redes neurais, especificamente as abordagens Deep & Cross Network (DCN) e Wide & Deep Network (WDN). Na etapa de *setup* do modelo utilizando redes neurais, foi utilizada a função de otimização *Adam*, uma taxa de aprendizado de 0.001 em 50 épocas de treinamento. A função de ativação utilizada foi *softmax*. De forma a estabelecer uma base comparativa para os algoritmos DCN e WDN, foi treinada uma primeira rede neural simples, sem a aplicação dos conceitos DCN e WDN, chamado de *Baseline Model*.

Na Figura 4.21 visualiza-se a rede neural gerada para o *Baseline Model*.

Na Figura 4.22 observa-se a rede neural gerada para o modelo WDN.

Na Figura 4.23 se visualiza a rede neural gerada para o modelo DCN.

Os resultados obtidos a partir da utilização das abordagens de redes neurais para a classificação binária podem ser observados na Tabela 4.7.

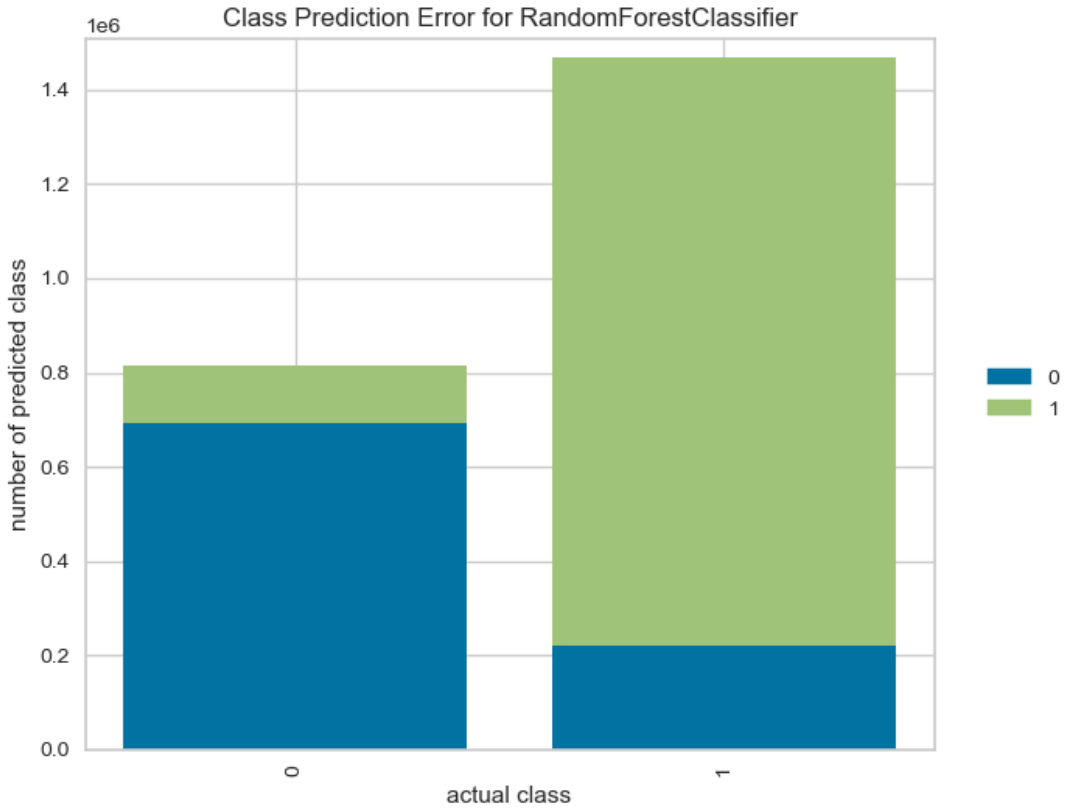

Figura 4.18: Classificador Binário: Erro.

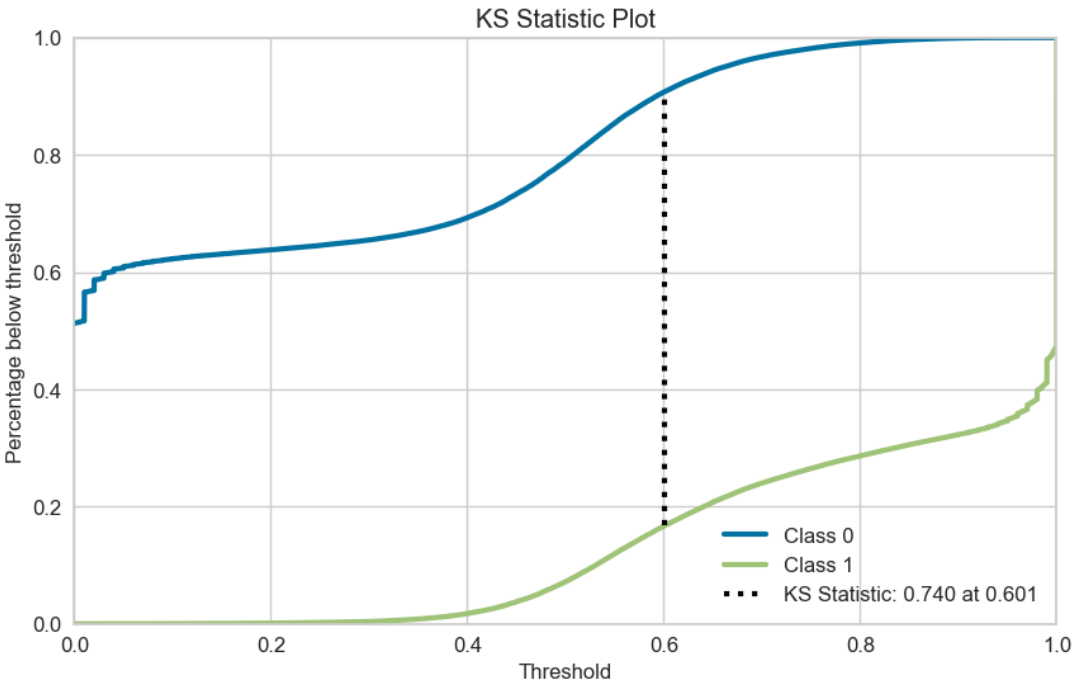

Figura 4.19: Classificador Binário: KS Statistic.

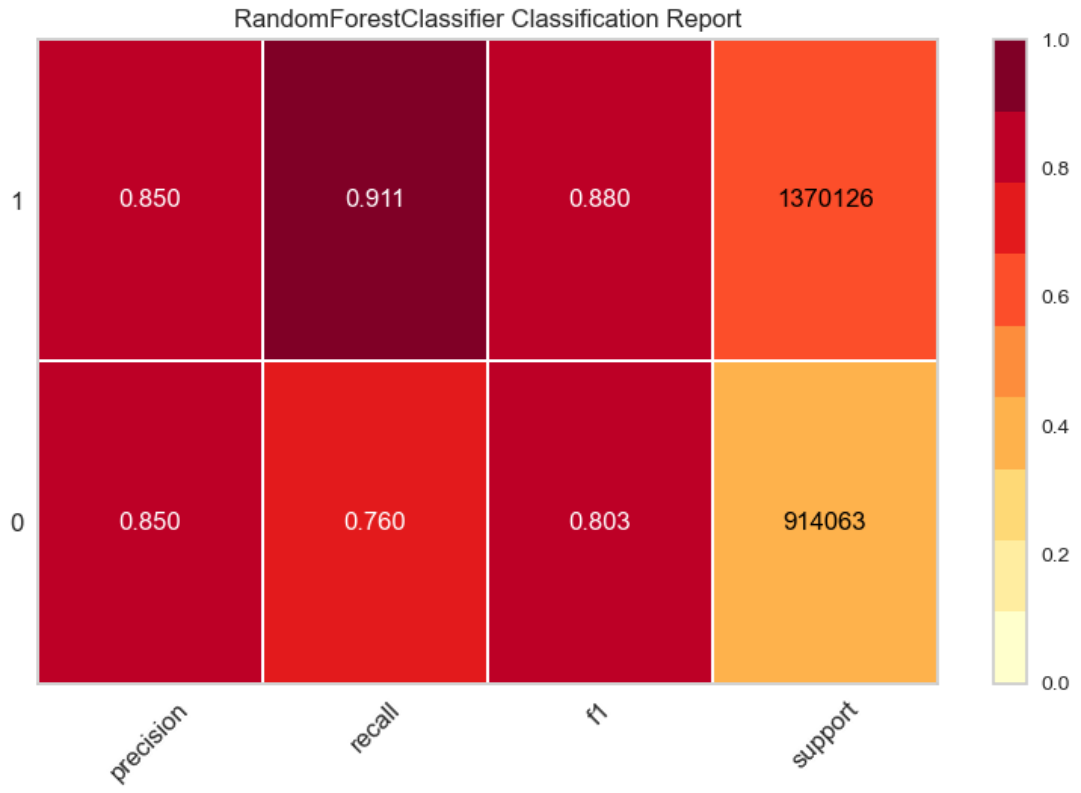

Figura 4.20: Classificador Binário: Resumo por Classe.

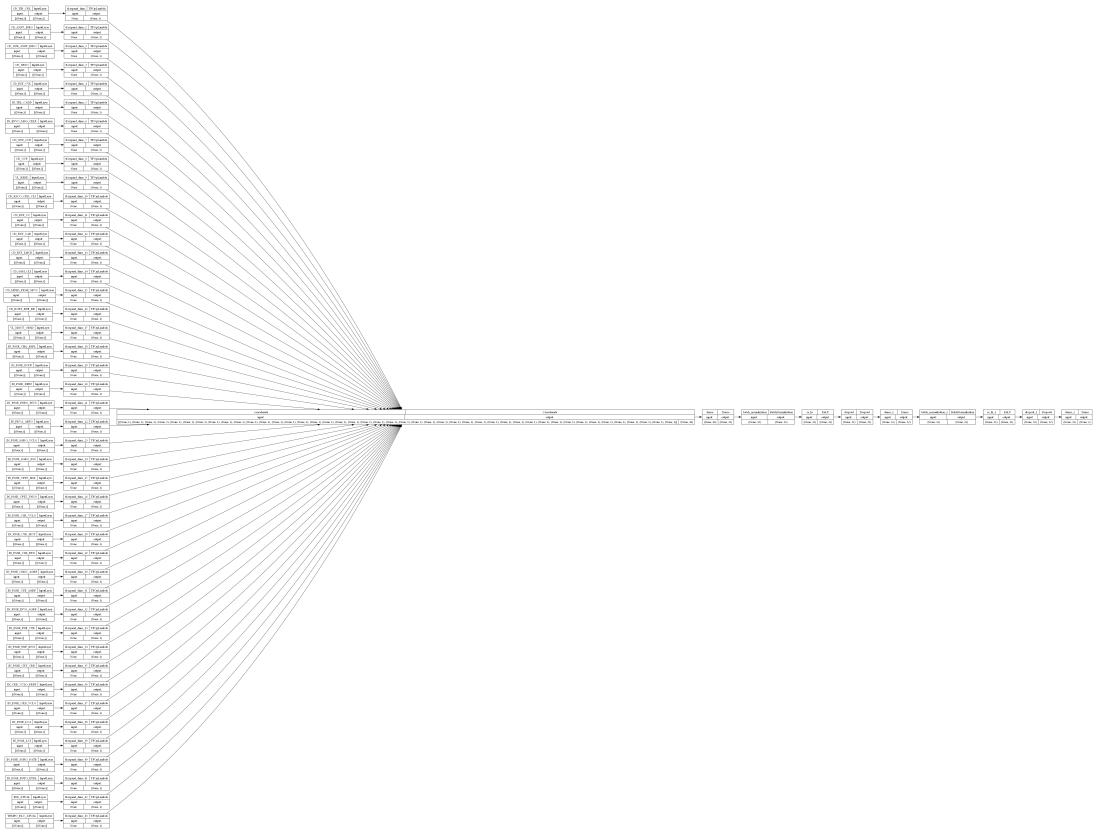

Figura 4.21: Rede Neural Binary Class: Baseline Model.

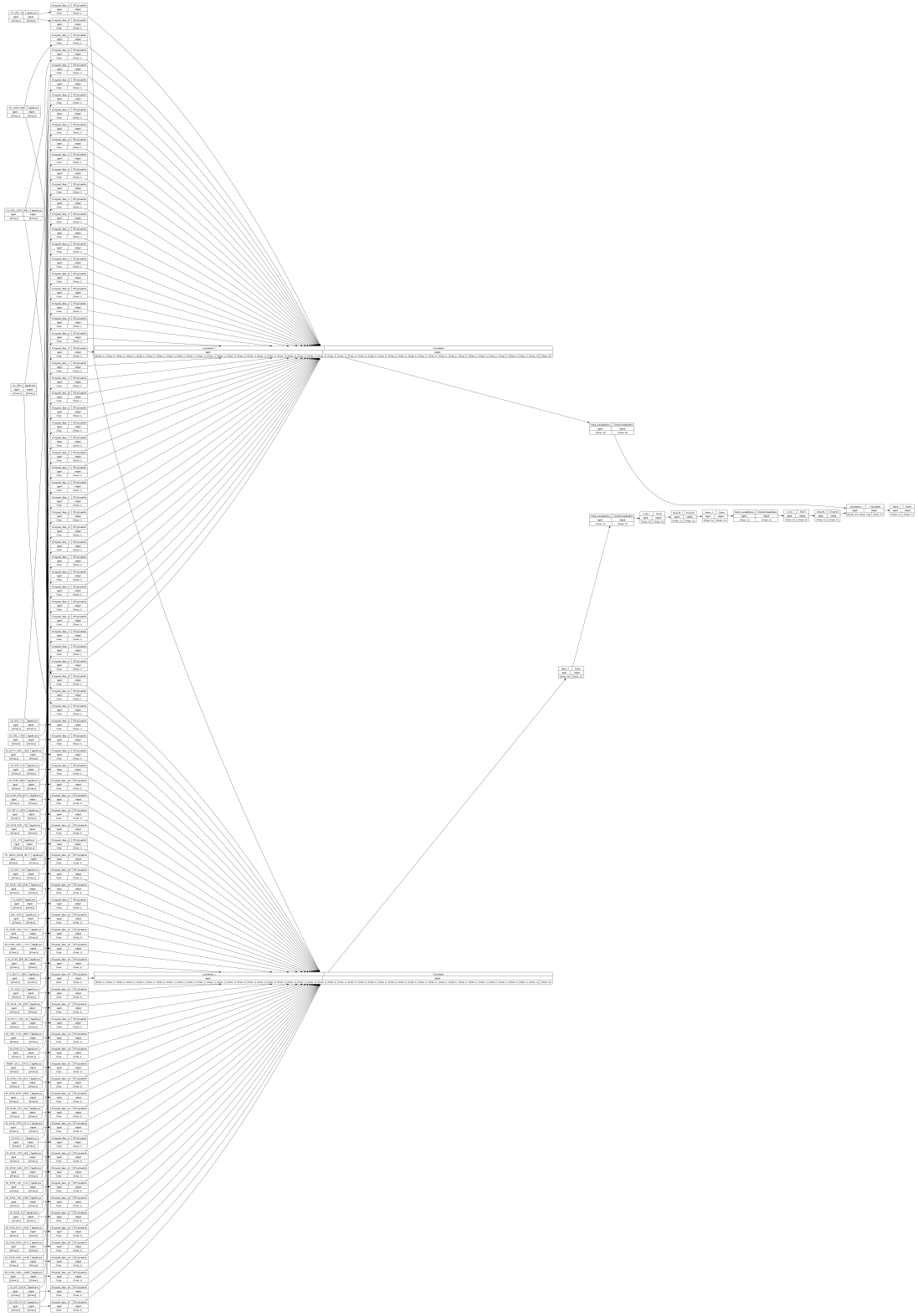

Figura 4.22: Rede Neural Binary Class: Wide and Deep Model.

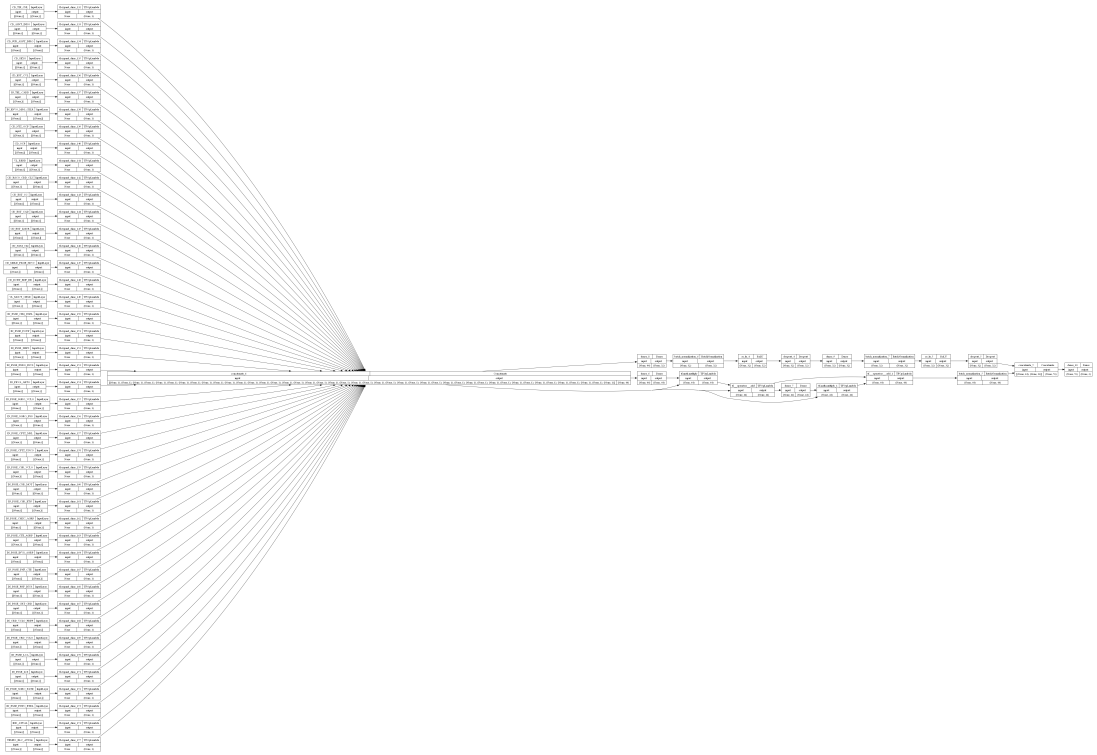

Figura 4.23: Rede Neural Binary Class: Deep and Cross Model.

Tabela 4.7: Treinamento dos Modelos em Redes Neurais: Abordagem Classificador Binário

Modelo	**Acurácia**
Baseline Model	0.8249
Wide & Deep Network (WDN)	0.8241
Deep & Cross Network (DCN)	0.8247

Nesta etapa do trabalho, o modelo que melhor performou para a abordagem de classificação binária foi o implementado com o uso de *Random Forest Classifier*.

Classificação Multi Classes

Para a modelagem multi classes, foi considerado como variável dependente o canal de interação, representado por `NM_TIP_CNL`. Sendo que, a partir das variáveis de perfil e interações realizadas, foi elaborado um modelo preditivo para a antecipação do canal adequado para o sucesso de uma determinada interação. Este modelo restringe a utilização de todo o conjunto de dados no aprendizado de máquina, uma vez que se vale somente da porção de dados que indica sucesso na interação.

Os dados foram divididos em base de treinamento e teste, na proporção de 70% e 30% respectivamente, utilizando *Stratified K-Fold* para a validação cruzada, onde $K = 10$.

O modelo foi primeiramente treinado utilizando utilizando os algoritmos de classificação probabilísticos e de *ensemble* conforme a Tabela 4.8. A mesma tabela mostra um comparativo entre as principais métricas para avaliação do modelo de classificação. O objetivo nesta etapa foi de maximização da acurácia do modelo.

Tabela 4.8: Treinamento dos Modelos Probabilísticos: Abordagem Classificador Multi Classe

Modelo	**Acurácia**	**AUC**	**Recall**	**Precision**	**F1**
Random Forest Classifier	0.8919	0.9756	0.8919	0.8942	0.8921
Decision Tree Classifier	0.8889	0.9533	0.8889	0.8926	0.8892
Extra Trees Classifier	0.8862	0.9636	0.8862	0.8891	0.8863
Extreme Gradient Boosting	0.8218	0.9577	0.8218	0.8341	0.8224
CatBoost Classifier	0.8216	0.9574	0.8216	0.8342	0.8222
Gradient Boosting Classifier	0.8003	0.9476	0.8003	0.8157	0.8003
Light Gradient Boosting Machine	0.7876	0.9241	0.7876	0.8042	0.7892
Ada Boost Classifier	0.6048	0.6651	0.6048	0.5951	0.5732
Logistic Regression	0.5326	0.7456	0.5326	0.4838	0.4750
Naive Bayes	0.5155	0.7215	0.5155	0.5168	0.4472
Ridge Classifier	0.4595	0.0000	0.4595	0.4146	0.4136
Linear Discriminant Analysis	0.4553	0.6642	0.4553	0.4315	0.4178
SVM - Linear Kernel	0.4162	0.0000	0.4162	0.4131	0.3256
Dummy Classifier	0.3898	0.5000	0.3898	0.1520	0.2187
Quadratic Discriminant Analysis	0.0435	0.5625	0.0435	0.4194	0.0740

Dentre os modelos treinados, o de melhores resultados iniciais foi o *Random Forest Classifier*, com acurácia inicial de 0.8919. Em seguida foi realizada uma etapa de otimização do modelo, utilizando *Bootstrap* com 100 estimadores, e critério de *Gini* no ajuste do modelo. Após a etapa de otimização, foi obtida uma pequena melhoria no resultado ao aplicar o modelo na base completa, podendo ser verificado na Tabela 4.9.

Tabela 4.9: Treinamento dos Modelos Probabilísticos: Abordagem Classificador Multi Classe

Modelo	**Acurácia**	**AUC**	**Recall**	**Precision**	**F1**
Random Forest Classifier	0.9196	0.9873	0.9196	0.9218	0.9197

Na Figura 4.24 se visualiza a matriz de confusão referente ao modelo otimizado. Nesta matriz verifica-se o número de falsos positivos, falsos negativos, verdadeiros positivos e verdadeiros negativos.

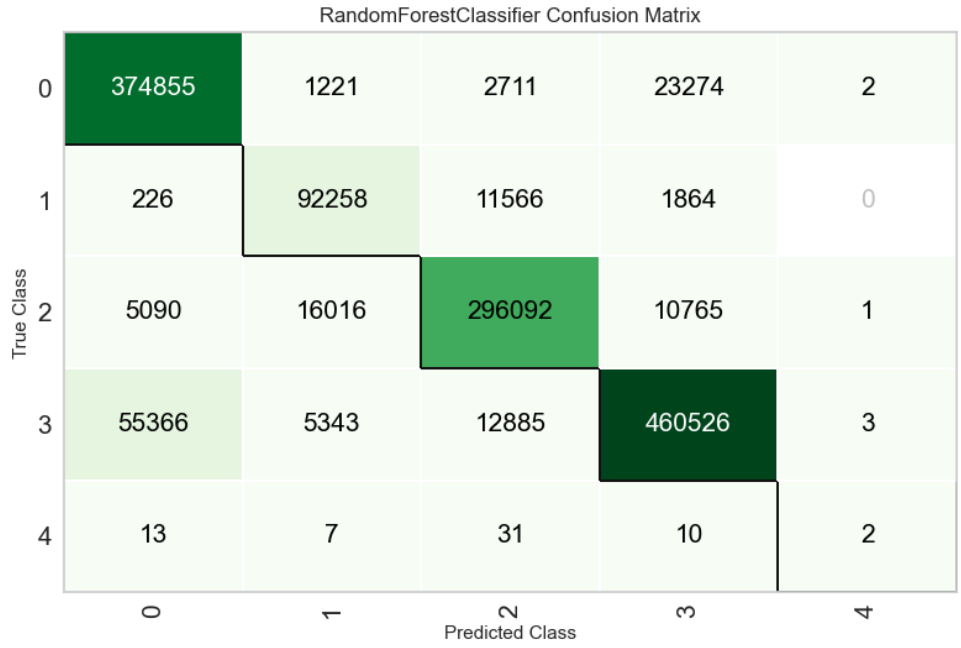

Figura 4.24: Classificador Multi Classe: Matriz de Confusão.

Na Figura 4.25 visualiza-se a curva ROC referente ao modelo otimizado. Nesta curva verifica-se o desempenho do modelo em diferentes pontos de corte. A partir da curva ROC se obtém o indicador AUC. Quanto maior a área sob a curva (AUC), melhor é o desempenho do modelo.

Na Figura 4.26 observam-se as principais variáveis utilizadas no modelo e seus níveis de importância.

Na Figura 4.27 se visualiza a curva de precisão versus recall do modelo otimizado, obtendo uma precisão média de 0.95.

Na Figura 4.28 se visualiza a proporção dos erros de previsão de classes no modelo otimizado.

Por fim, na Figura 4.29 se visualizam os indicadores de precisão, recall, F1 e suporte, por classes no modelo otimizado.

Em seguida, o modelo foi treinado utilizando redes neurais, especificamente as abordagens Deep & Cross Network (DCN) e Wide & Deep Network (WDN) de forma análoga à abordagem de classe única. Na etapa de *setup* do modelo utilizando redes neurais, foi utilizada a função de otimização *Adam*, uma taxa de aprendizado de 0.001 em 50 épocas

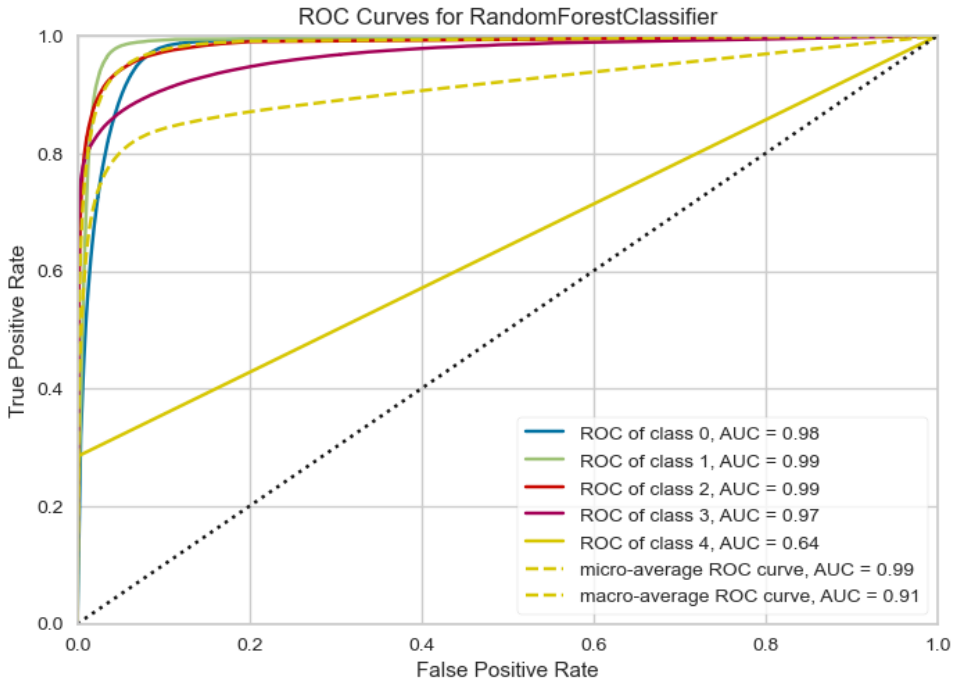

Figura 4.25: Classificador Multi Classe: Curva ROC.

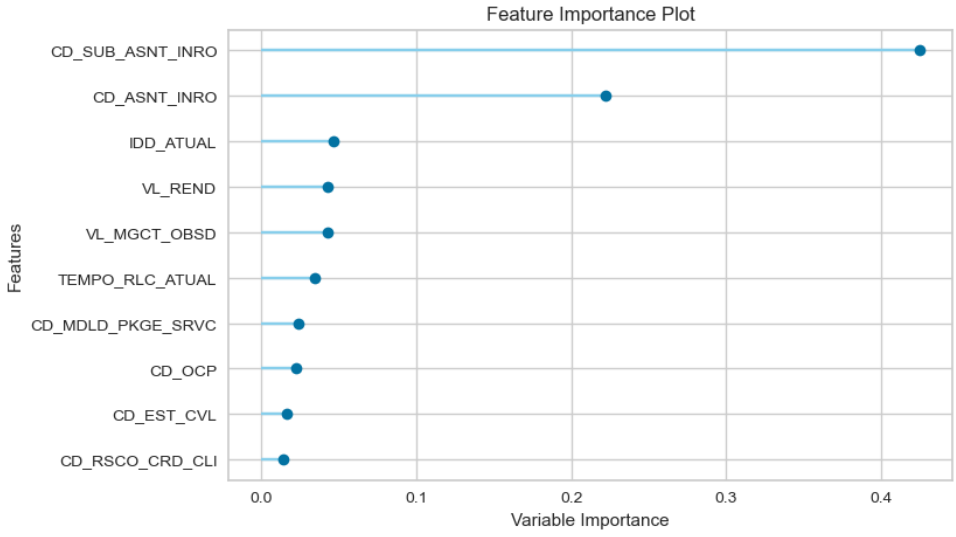

Figura 4.26: Classificador Multi Classe: Variáveis de Maior Relevância.

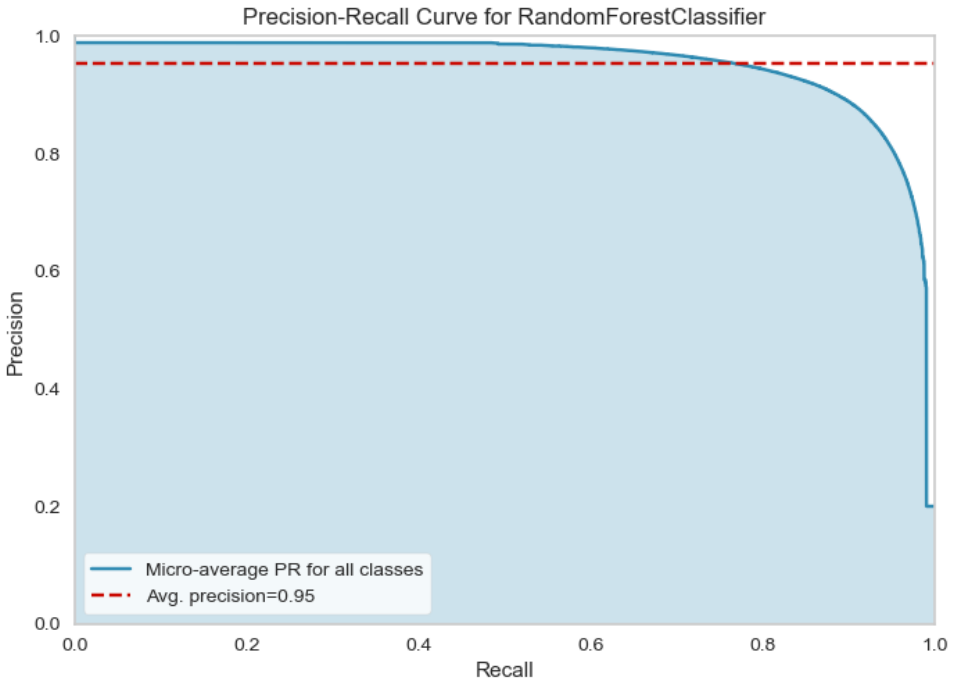

Figura 4.27: Classificador Multi Classe: Precision-Recall.

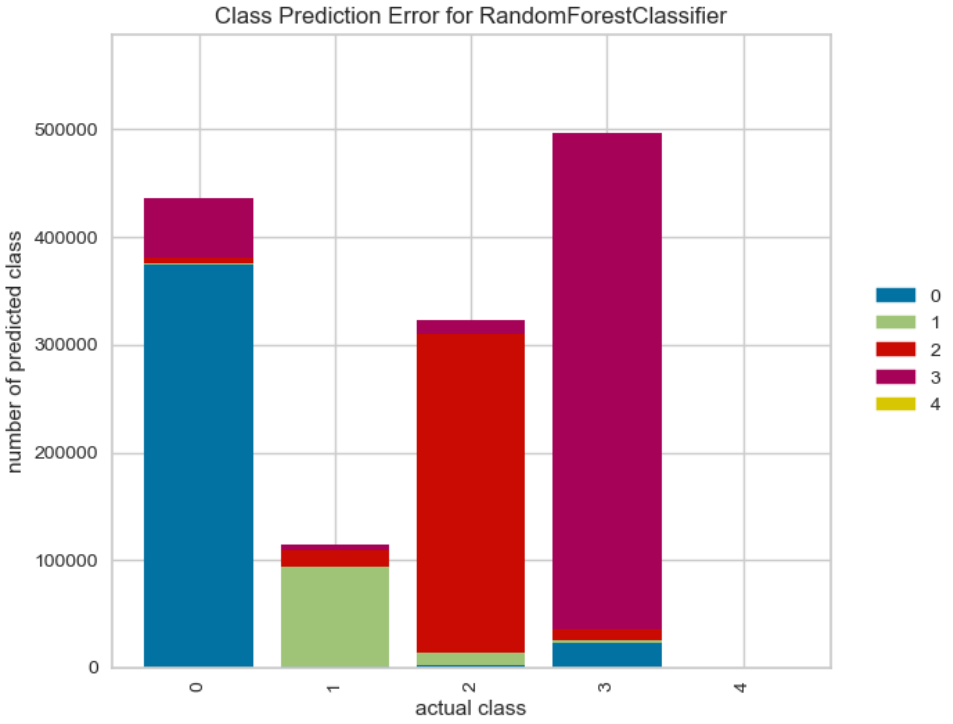

Figura 4.28: Classificador Multi Classe: Erro.

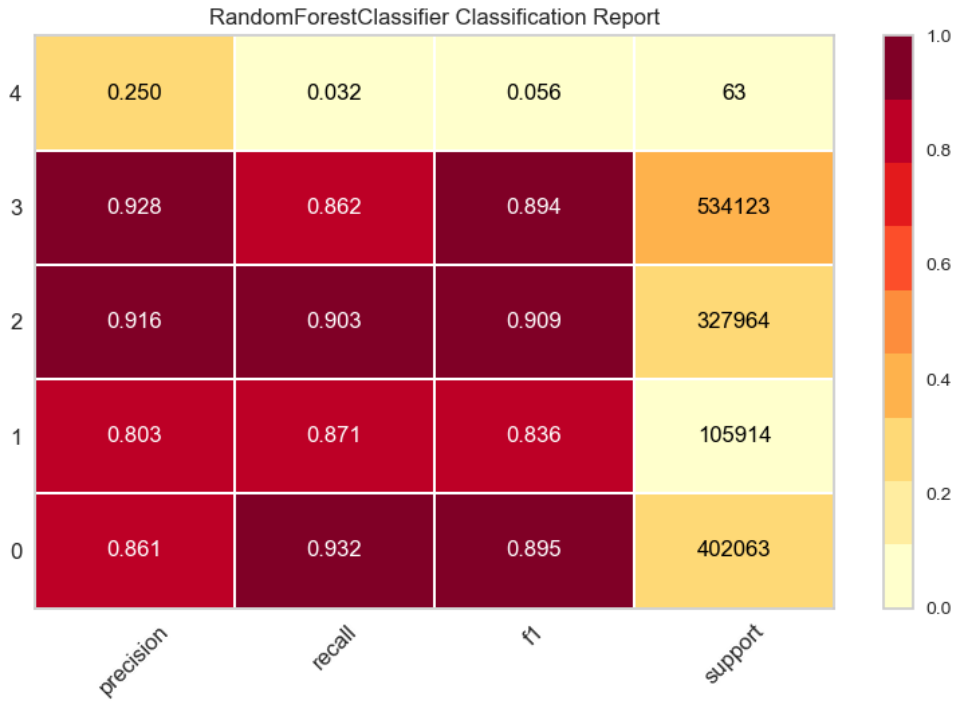

Figura 4.29: Classificador Multi Classe: Resumo por Classe.

de treinamento. A função de ativação utilizada foi *softmax*. De forma a estabelecer uma base comparativa para os algoritmos DCN e WDN, foi treinada uma primeira rede neural simples, sem a aplicação dos conceitos DCN e WDN, chamado de *Baseline Model*.

Na Figura 4.30 se observa a rede neural gerada para o *Baseline Model*.

Na Figura 4.31 se visualiza a rede neural gerada para o modelo WDN.

Na Figura 4.32 visualiza-se a rede neural gerada para o modelo DCN.

Os resultados obtidos a partir da utilização das abordagens de redes neurais para a classificação multi classe podem ser observados na Tabela 4.10.

Tabela 4.10: Treinamento dos Modelos em Redes Neurais: Abordagem Multi Classe

Modelo	**Acurácia**
Baseline Model	0.6119
Wide & Deep Network (WDN)	0.7543
Deep & Cross Network (DCN)	0.6819

Nesta etapa do trabalho, o modelo com o melhor desempenho para a abordagem de classificação multi classe foi o implementado com o uso do algoritmo *Random Forest Classifier*. Para fins de facilitação da reprodução integral do experimento, incluindo os

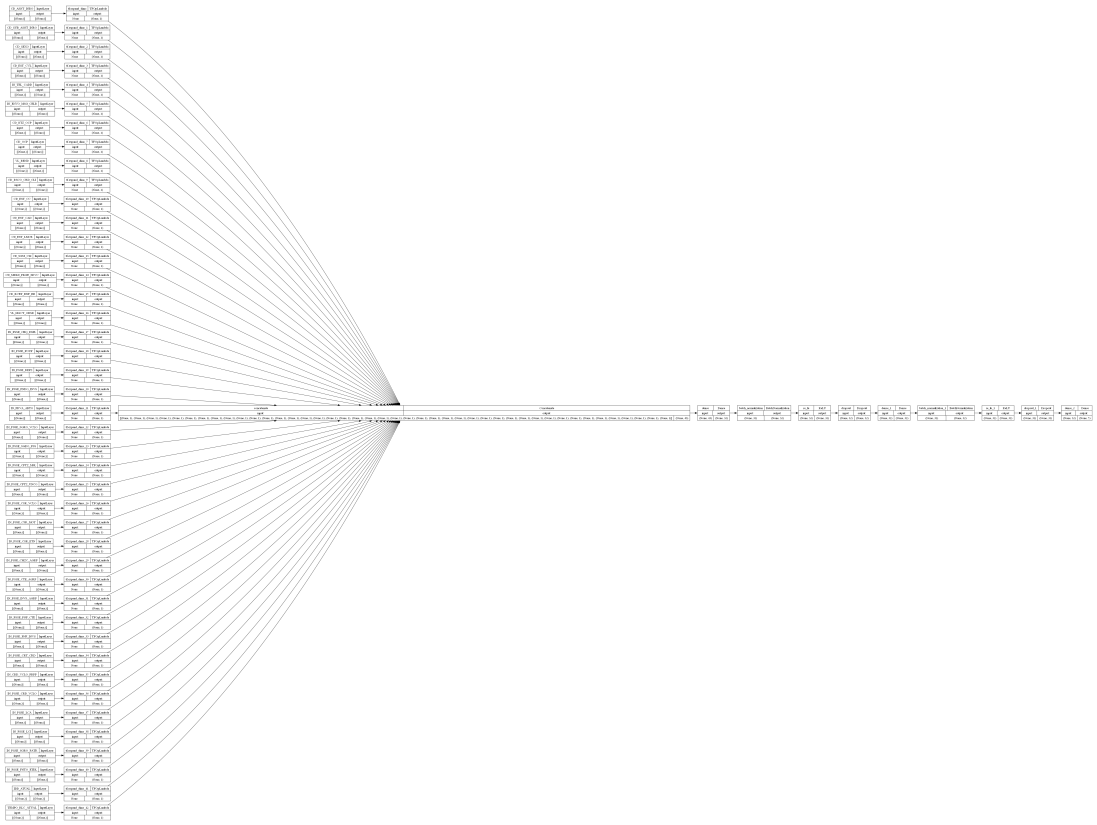

Figura 4.30: Rede Neural Multi Class: Baseline Model.

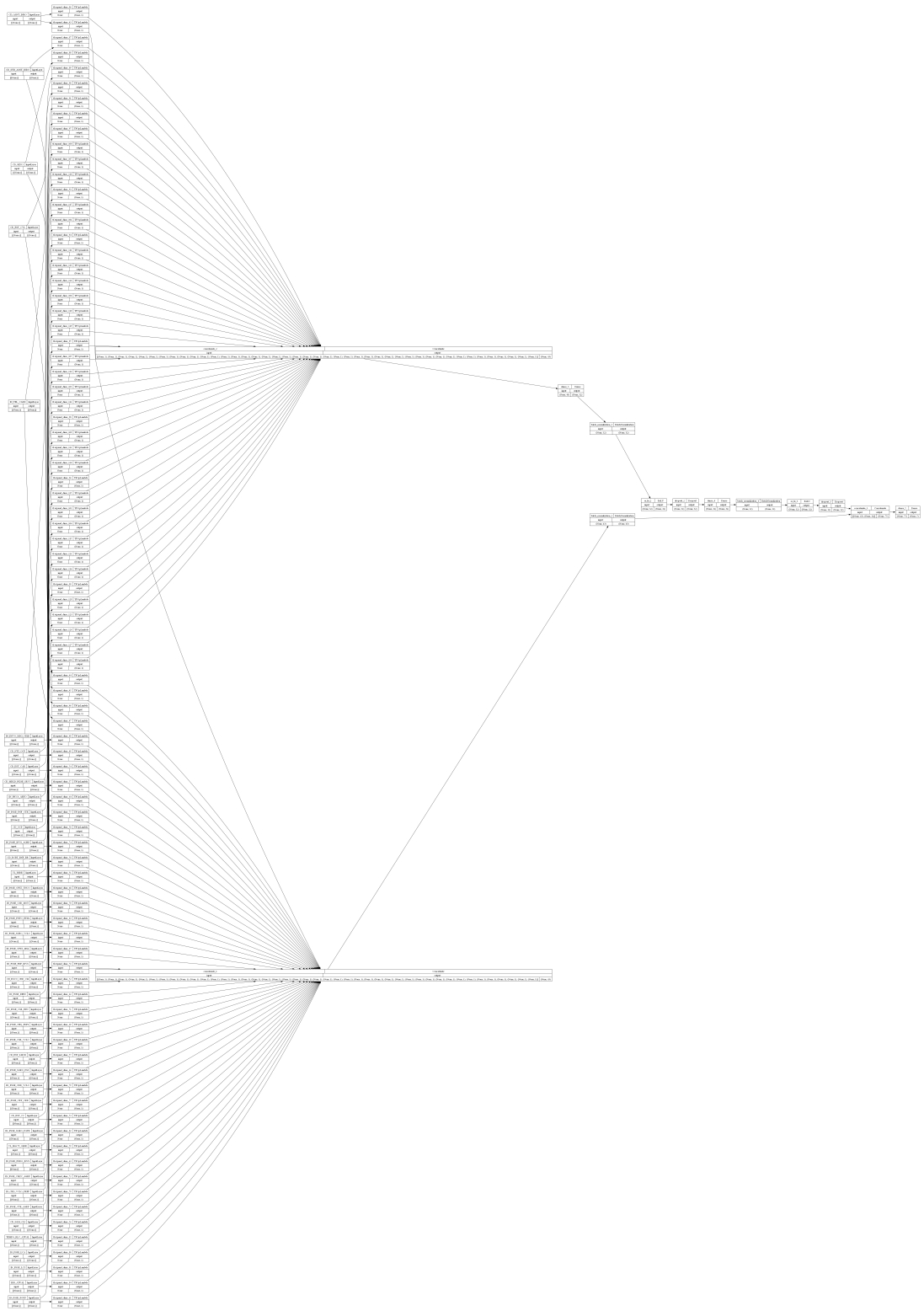

Figura 4.31: Rede Neural Multi Class: Wide and Deep Model.

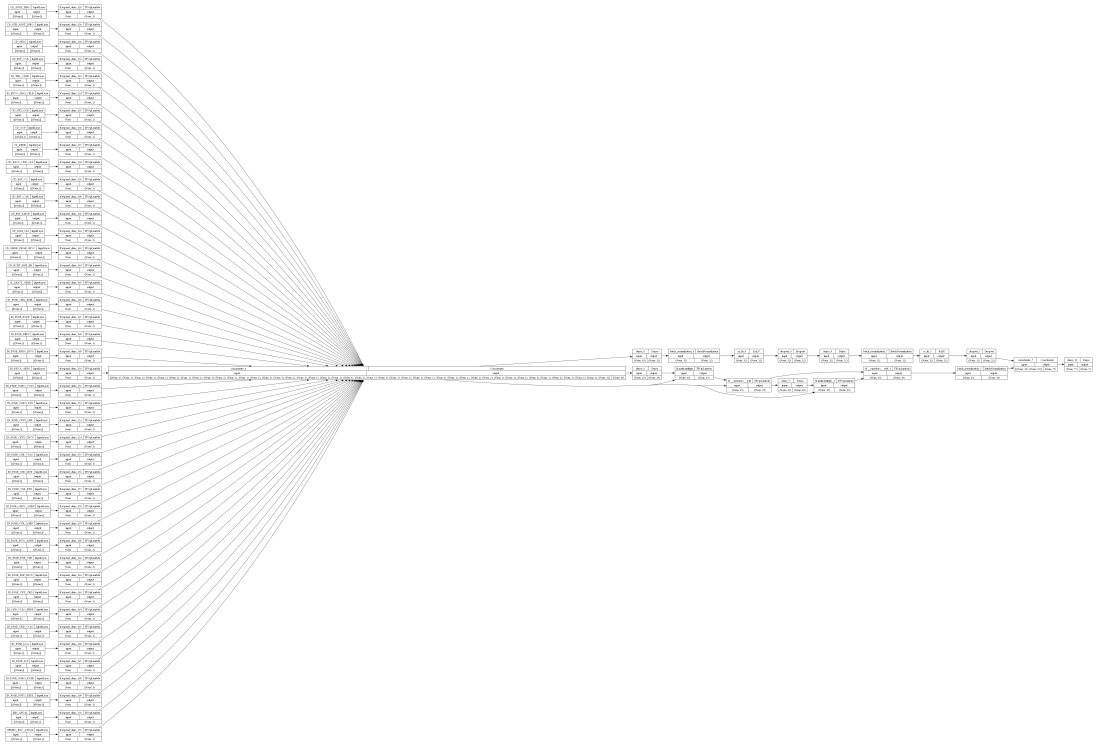

Figura 4.32: Rede Neural Multi Class: Deep and Cross Model.

processos de engenharia e análise de dados, modelagem e treinamento, o código fonte do trabalho pode ser consultado no repositório de projetos *Github.com* [18].

Capítulo 5

Conclusão

A hiper personalização é um desafio no relacionamento com clientes, sobretudo no segmento bancário, de alta competitividade, investimentos massivos em tecnologias, pesquisa e desenvolvimento de soluções inovadoras. Apesar das inúmeras fontes de dados de relacionamento com o cliente serem produzidos diariamente pelas organizações, extrair valor dessas massas de dados requer investigação de técnicas, experimentação e avaliação de conceitos da academia e da indústria.

De forma a implementar a hiper personalização, um banco brasileiro pretende implantar um modelo de Next Best Action (NBA) em seus processos de relacionamento. Para tal, investe no aprimoramento de abordagens com a mensagem mais adequada, no melhor momento, pelo melhor canal de interação com o cliente.

Este trabalho propôs uma abordagem comparativa de modelagem e técnicas para a previsão do melhor canal de interação com o cliente, de forma a compor a abordagem NBA pretendida na organização. A partir de dados demográficos que compõem o perfil do cliente e dados de interações realizadas, foram elaborados modelos para a indicação do sucesso na abordagem e do melhor canal de interação.

Foram testadas duas abordagens de modelagem, classificação binária e classificação multi classe. Para cada uma dessas abordagens foram utilizados algoritmos e técnicas probabilísticas como *Naive Bayes*, *Random Forest*, *SVM - Linear Kernel*, *Logisti Regression*, *Decision Tree*, *Gradient Boosting* e *CatBoost*, bem como algoritmos e técnicas de redes neurais como *Wide & Deep Network* e *Deep & Cross Network*.

Esses modelos foram comparados entre si, priorizando a acurácia em suas aplicações. Foram obtidos resultados satisfatórios ao negócio para ambas as abordagens de modelagem, sendo as acurácias máximas mensuradas no modelo de classificação binária de 0.8659 e de 0.9196 para o multi classe utilizando-se técnicas probabilísticas. Com a utilização de redes neurais, as acurácias máximas mensuradas no modelo de classificação binária foi de 0.8249 e de 0.7543 para o multi classe.

O melhor modelo obtido neste estudo utilizou-se da abordagem de classificação multi classe para a previsão do melhor canal de interação com o cliente, prevendo o sucesso de uma interação com acurácia de 0.9196, AUC de 0.9873, Precision de 0.9218 e F1 de 0.9197 implementado com o algoritmo classificador *Random Forest Classifier*.

Propõe-se como trabalhos futuros a evolução deste estudo, explorando técnicas de otimização e aplicando hiper parâmetros otimizados nos modelos, de forma a obter melhores resultados nos indicadores apurados. Uma segunda abordagem possível é o treinamento desses modelos em ambiente distribuído com maior capacidade computacional, viabilizando assim, uma maior utilização de dados para o treinamento, sobretudo das redes neurais.

Por fim, ao atender a expectativa negocial da instituição alvo do estudo, sugere-se a adoção das abordagens de modelagem e técnicas exploradas neste estudo na composição da abordagem NBA dessa organização, viabilizando a escolha do melhor canal de interação com o cliente em tempo de comunicação.

Referências

[1] *Next-best-action marketing: a customer-centric approach.* Customer relationship management (Malibu, Calif.), 16(5):42, 2012, ISSN 1529-8728. 1

[2] Deloitte: *Next best action - driving customer value through a rich and relevant multichannel experience in financial services.* `https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/consultancy/deloitte-uk-con-next-best-action.pdf`, acesso em 2023-10-10. 1

[3] Inc., Pegasystems: *Next-best-action marketing: A customer centric approach.* `https://www.pega.com/system/files/resources/pdf/NBA-Marketing-eBook-Jan2012.pdf`, acesso em 2023-10-10. 1

[4] FEBRABAN: *Pesquisa febraban de tecnologia bancária 2022*, 2022. `https://cmsarquivos.febraban.org.br/Arquivos/documentos/PDF/pesquisa-febraban-2022-vol-3.pdf`, acesso em 2023-04-17. 2

[5] Zhou, Zhi Hua: *Ensemble Methods: Foundations and Algorithms.* Chapman and Hall/CRC, 1st edição, 2012, ISBN 1439830037. 7

[6] Cheng, Heng Tze, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu e Hemal Shah: *Wide and deep learning for recommender systems.* Em *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, DLRS 2016, páginas 7–10, New York, NY, USA, 2016. Association for Computing Machinery, ISBN 9781450347952. `https://doi.org/10.1145/2988450.2988454`. 8, 18, 21

[7] Wang, Ruoxi, Bin Fu, Gang Fu e Mingliang Wang: *Deep & cross network for ad click predictions.* CoRR, abs/1708.05123, 2017. `http://arxiv.org/abs/1708.05123`. 9, 10, 18, 21

[8] Chapman, Peter, Janet Clinton, Randy Kerber, Tom Khabaza, Thomas P. Reinartz, Colin Shearer e Richard Wirth: *Crisp-dm 1.0: Step-by-step data mining guide.* 2000. 9, 11

[9] Hastie, T., R. Tibshirani e J.H. Friedman: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer series in statistics. Springer, 2009, ISBN 9780387848846. `https://books.google.com.br/books?id=eBSgoAEACAAJ`. 12

[10] Pearson, Karl: *Note on regression and inheritance in the case of two parents.* Proceedings of the Royal Society of London Series I, 58:240–242, janeiro 1895. 13, 32

[11] Fakhar Bilal, Syed, Abdulwahab Ali Almazroi, Saba Bashir, Farhan Hassan Khan e Abdulaleem Ali Almazroi: *An ensemble based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in telecom industry.* PeerJ. Computer science, 8:e854–e854, 2022, ISSN 2376-5992. 15, 20

[12] Gattermann-Itschert, Theresa e Ulrich W. Thonemann: *Proactive customer retention management in a non-contractual b2b setting based on churn prediction with random forests.* Industrial marketing management, 107:134–147, 2022, ISSN 0019-8501. 15, 20

[13] Cao, Longbing e Chengzhang Zhu: *Personalized next-best action recommendation with multi-party interaction learning for automated decision-making.* PLOS ONE, 17(1):1–22, janeiro 2022. `https://doi.org/10.1371/journal.pone.0263010`. 16, 20

[14] Sun, Jie, Jie Li e Hamido Fujita: *Multi-class imbalanced enterprise credit evaluation based on asymmetric bagging combined with light gradient boosting machine.* Applied soft computing, 130:109637, 2022, ISSN 1568-4946. 17, 20

[15] Faraj, Azhi Abdalmohammed, Didam Ahmed Mahmud e Bilal Najmaddin Rashid: *Comparison of different ensemble methods in credit card default prediction.* UHD Journal of Science and Technology, 5(2):20–25, 2021, ISSN 2521-4209. 17, 20

[16] Massaoudi, Mohamed, Shady S. Refaat, Ines Chihi, Mohamed Trabelsi, Fakhreddine S. Oueslati e Haitham Abu-Rub: *A novel stacked generalization ensemble-based hybrid lgbm-xgb-mlp model for short-term load forecasting.* Energy (Oxford), 214:118874, 2021, ISSN 0360-5442. 17, 21

[17] Theocharous, Georgios, Philip S. Thomas e Mohammad Ghavamzadeh: *Personalized ad recommendation systems for life-time value optimization with guarantees.* Em *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, páginas 1806–1812. AAAI Press, 2015, ISBN 9781577357384. 19, 21

[18] Resende, Bruno Gomes: *Repositório de fontes do trabalho.* `https://github.com/brunogresende/ppca-trabalho-final/tree/4f2b3062c4684381a93967c070f55770ffc34170/code`, acesso em 2024-09-06. 54