

ANÁLISE DO IMPACTO DO TEMPO DE ESPERA NA EVASÃO DE CLIENTES EM FILAS DE ATENDIMENTO: uma análise preditiva aplicada no modelo de regressão logística para atendimentos via chat

SANTOS, Vinicius Teixeira¹; BAIA, Joás Wesley ²



¹ Bacharel em Ciência da Computação - UNIFAGOC.
² Docente do curso de Ciência da Computação - UNIFAGOC.

vinicius14100@gmail.com
joas.baia@unifagoc.com.br

RESUMO

Os Sistemas de Informação (SI) surgem como ferramentas para o gerenciamento eficaz das organizações e, juntamente com eles, setores de suporte que oferecem auxílio para usuários desses SIs. O objetivo deste estudo foi o de analisar o impacto que o tempo de espera em filas de atendimento tem sobre a evasão nos setores Folha, Gestão e Fiscal de uma empresa que oferece suporte de atendimento a um sistema de informação na área Contábil, a fim de desenvolver um modelo capaz de estimar a evasão de clientes em espera por atendimento. Para alcançar o objetivo, métodos de tratamento de dados, correlação de variáveis, regressão logística e validação de modelo pela curva ROC foram empregados. A correlação obteve um resultado de relação negativa, simbolizando que quanto maior o tempo em espera na fila maior a chance de se manter no chat, enquanto a validação do modelo pela curva ROC encontrou valores de previsão para os 3 setores estudados de 70% para folha 63 % para gestão e 61% para fiscal. Concluiu-se que somente o tempo de espera não é um fator determinante para que os clientes abandonem as filas de espera.

Palavras-chave: Análise Preditiva. Regressão Logística. Atendimento ao Cliente. Sistemas de Informação. Evasão de Clientes, Abandono. Desistência. Tempo de Espera. Chat.

INTRODUÇÃO

No contexto empresarial, os Sistemas de Informação (SI) são ferramentas essenciais para o gerenciamento eficaz das organizações. Segundo Laudon e Laudon (2004), um Sistema de Informação é um conjunto de componentes inter-relacionados que coleta, processa, armazena e distribui informações para apoiar a tomada de decisões, a coordenação e o controle organizacional. Esses sistemas desempenham um papel crucial na automação de serviços, no aprimoramento da experiência do usuário e na facilitação das operações diárias, promovendo o alcance de metas e o cumprimento de obrigações.

Empresas de diversos setores adotam esses sistemas para otimizar o gerenciamento e aprimorar a tomada de decisões. Nesse contexto, o presente estudo se concentra no setor de Suporte de Atendimento ao Cliente de uma empresa que oferece soluções nas áreas de gestão e contabilidade, utilizando Sistemas de Informação. Esse setor de suporte utiliza um programa de comunicação via chat

como principal canal de interação com os clientes, com o objetivo de oferecer uma resolução rápida e eficaz para seus problemas. De acordo com Zendesk (2024), os chats surgiram como resposta à preferência dos clientes por uma comunicação direta e simplificada, evitando a necessidade de ligações e filas de espera prolongadas. Assim, o chat se tornou o principal meio de comunicação da empresa, utilizado para registrar problemas, solicitações e feedbacks detalhados dos usuários. Esse registro é essencial para identificar áreas que necessitam de melhorias e implementar soluções que atendam às expectativas dos clientes.

Para van Riel, Schepers e de Ruyter (2013), o feedback contínuo dos clientes oferece insights valiosos que auxiliam as empresas a entender melhor as necessidades e expectativas dos consumidores, contribuindo para o desenvolvimento de produtos mais alinhados com o mercado. Por meio da análise dos dados gerados no sistema de chat, é possível identificar padrões e comportamentos que ajudam a aprimorar o atendimento e reduzir a evasão dos chamados.

No setor de Suporte de Atendimento ao Cliente, a efetividade na resolução de problemas e a satisfação dos clientes podem ser prejudicadas pela falta de um sistema estruturado para analisar os dados das interações realizadas via chat. Embora o chat seja um canal de comunicação rápido e direto, a ausência de uma análise organizada pode resultar em tempos prolongados de espera e dificultar a captação de feedbacks relevantes sobre erros de software. Isso pode limitar a capacidade da empresa de promover melhorias contínuas e, eventualmente, aumentar a insatisfação e a evasão dos clientes.

Resolver essa questão é fundamental para aumentar a eficiência e a qualidade dos atendimentos, bem como a satisfação dos clientes. As interações via chat fornecem dados valiosos sobre os pontos de melhoria no software e no atendimento; contudo, sem uma análise adequada, esses dados permanecem subutilizados. A identificação de padrões nos dados de chat permite a antecipação de problemas, a melhoria dos processos de atendimento e o desenvolvimento de soluções mais alinhadas às necessidades dos clientes. Além disso, um sistema de análise estruturada pode reduzir a evasão nos atendimentos e melhorar a percepção dos clientes sobre o serviço prestado.

Com base nisso, este estudo busca responder à seguinte pergunta: qual o impacto do tempo de espera em filas de atendimento sobre o abandono dos chamados de suporte em diferentes setores?

Diante desse cenário, o objetivo principal deste estudo é analisar o impacto do tempo

e espera em fila sobre a evasão dos clientes em atendimentos via chat dos setores Folha de Pagamento, Fiscal e Gestão. Para isso, serão coletados e tratados dados de atendimento registrados no sistema de chat Contako, abrangendo o período de 2022 a 2023. Para alcançar o objetivo geral, os objetivos específicos são: coleta de dados; filtragem e limpeza, removendo erros e informações irrelevantes; inferir dados faltantes; preparar dados para aplicação de técnicas de correlação de variáveis e utilizar a Regressão Logística para analisar a relação entre o tempo de espera e a evasão dos clientes. A validação do modelo preditivo será realizada por meio da curva ROC, que permitirá avaliar a precisão do modelo em prever a evasão dos clientes.

A limpeza de dados foi realizada com o auxílio da ferramenta Planilhas Google, enquanto o algoritmo de inferência foi implementado em NodeJS. O modelo final foi desenvolvido na plataforma Google Colab com Python, e os resultados obtidos serão apresentados em relatórios e gráficos que contribuirão para uma análise e interpretação mais aprofundada.

REFERENCIAL TEÓRICO

Contako: sistema de atendimento ao cliente via Chat

O Contako (2024) define seu sistema como um serviço de atendimento multicanal via chat que permite estabelecer comunicação direta entre empresas e seus clientes. Esse sistema coleta informações essenciais para a análise, tais como Setor de Atendimento, Data de Entrada, Data de Primeira Mensagem e Data da Última Mensagem, que se configuram como variáveis-chave para a pesquisa sobre abandono de atendimento.

Segundo Creswell (2014), a coleta de dados é um pilar em estudos quantitativos, pois permite identificação de padrões, testagem de hipóteses e construção de modelos precisos. Esse processo é particularmente relevante em análises preditivas. Essas quatro variáveis representam os dados mais importantes para a pesquisa e serão analisadas no contexto de um banco de dados onde pode haver valores faltantes. Conforme Batista e Monard (2003), a presença de dados incompletos em uma base pode distorcer análises estatísticas, comprometendo a precisão de modelos preditivos como a regressão logística, pois dados ausentes podem reduzir a representatividade da amostra.

Os autores enfatizam que técnicas de imputação de dados são indispensáveis em pesquisas em que a completude dos dados é crítica:

A presença de dados ausentes em bases de dados pode levar a análises incompletas ou incorretas, comprometendo a qualidade e a confiabilidade dos resultados. Técnicas de inferência de dados faltantes, como a imputação, são eficazes para preservar a informação e reduzir o viés introduzido pela ausência de dados (Batista; Monard, 2003).

Esse processo de imputação, ao lidar com dados faltantes, contribui para a preservação da estrutura dos dados e integridade dos resultados, promovendo maior confiabilidade nas inferências e nos modelos preditivos desenvolvidos na pesquisa.

Correlação de Variáveis e Identificação de Padrões

A correlação de variáveis é uma medida estatística utilizada para avaliar a relação entre duas ou mais variáveis, ajudando a entender como mudanças em uma podem estar associadas a mudanças em outra. Esse método quantifica tanto a força quanto a direção da relação entre variáveis, com valores variando de -1 a 1, conforme explicado por Dancey e Reidy (2007) e detalhado no Quadro 2.

Quadro 2 - Interpretação dos valores de relação

Valor da Correlação	Interpretação
1	Correlação positiva perfeita (ambas as variáveis aumentam proporcionalmente)
0	Ausência de correlação linear (não há relação consistente entre as variáveis)
-1	Correlação negativa perfeita (à medida que uma variável aumenta, a outra diminui proporcionalmente)

Fonte: autoria própria.

No contexto de atendimento ao cliente, essas técnicas podem ser aplicadas para analisar a relação entre as variáveis tempo de espera e a evasão de clientes. Kumar e Chong (2018) destacam que a análise de correlação é uma ferramenta essencial na aprendizagem de máquina, permitindo identificar relações significativas entre atributos de um conjunto de dados e a variável alvo. Isso orienta a construção de modelos preditivos ao destacar os fatores mais relevantes para a análise. No projeto em questão, a variável quantitativa e contínua "tempo de espera" é analisada para verificar sua relação com a evasão de clientes como variável binária, auxiliando na identificação de fatores que influenciam a satisfação e a retenção no atendimento.

Análise Preditiva e Aprendizagem de Máquina

A análise preditiva utiliza dados históricos, algoritmos estatísticos e técnicas de aprendizagem de máquina para estimar a probabilidade de eventos futuros sendo utilizados em projetos que visam construir modelos capazes de replicar padrões através de fórmulas matemáticas. De acordo com Hyndman e Athanasopoulos (2018), essa abordagem aplicar modelos matemáticos para prever comportamentos e eventos, enquanto Kumar e Garg (2018) destacam que ela gera insights valiosos para a tomada de decisões, permitindo que organizações se antecipem e respondam a possíveis cenários futuros. No presente projeto, a análise preditiva é o tipo de análise eminente para estimar a probabilidade de evasão de clientes em função do tempo de espera.

A aprendizagem de máquina (AM), um campo da inteligência artificial, permite que sistemas identifiquem padrões em grandes volumes de dados e façam previsões ou tomem decisões baseadas nesses padrões. Nesse projeto, optou-se pela regressão logística como modelo de AM para previsão de evasão. Segundo Alpaydin (2020), a aprendizagem de máquina é especialmente eficaz em cenários onde é necessário analisar grandes conjuntos de dados para extrair insights estratégicos, como ocorre no setor de atendimento ao cliente, em que a previsão de evasão e a análise de satisfação pode ser utilizada para aumentar a retenção de clientes.

Regressão Logística

A regressão logística é uma técnica estatística utilizada para modelar a probabilidade de um evento binário ocorrer, baseada em uma ou mais variáveis independentes. Ela é amplamente utilizada em análise preditiva para prever a ocorrência de um evento e classificar os resultados. De acordo com Ménard (2002), a

regressão logística estima a probabilidade de uma determinada classe ou evento, como sucesso/falha, sim/não, através da utilização de uma função logística. No caso do projeto, será feita a verificação das informações referentes à chance de o cliente evadir após um certo tempo de espera no atendimento.

A regressão logística é particularmente útil em análise preditiva quantitativa, pois permite a modelagem da relação entre variáveis independentes e a probabilidade de um evento binário. Ao aplicar a regressão logística aos dados de atendimento ao cliente, é possível identificar padrões e fatores que influenciam o abandono do atendimento. Isso permite que as organizações desenvolvam estratégias preditivas e preventivas para minimizar a taxa de abandono, especialmente durante períodos de alto fluxo de atendimento.

MÉTODO DE DESENVOLVIMENTO

Para desenvolver este trabalho, foi utilizada uma pesquisa de natureza aplicada, que é caracterizada como tal por buscar a solução de problemas específicos e práticos. De acordo com Wieringa (2014), pesquisas aplicadas em tecnologia têm o objetivo de desenvolver soluções testáveis para problemas concretos, permitindo uma aplicação direta dos resultados em contextos reais. O método de análise utilizado neste estudo foi dividido em sete etapas, visando à análise preditiva dos dados de atendimento e à aplicação da regressão logística para prever a evasão dos clientes com base no tempo de espera. As etapas detalhadas são descritas a seguir.

Coleta de Dados

A primeira etapa envolve a coleta dos dados de atendimento ao cliente da plataforma Contako, entre o período de 2022 a 2023. Os dados incluem informações como Setor de Atendimento, Data de Entrada na Fila de Espera, Data de Primeira Mensagem e Data da Última Mensagem. Todos os dados irão constar em um arquivo CSV.

Filtragem e Limpeza dos Dados

Os dados coletados passaram por um filtro e serão depurados para eliminar erros e informações irrelevantes de atendimentos. Esta etapa garantirá que os dados utilizados na análise sejam precisos e consistentes. Dados inconsistentes ou irrelevantes são conhecidos por impactar negativamente a qualidade das análises, como observado por Rahm e Do (2000). Este processo visa assegurar que apenas dados úteis sejam utilizados para o aprendizado na modelagem.

Inferência sobre dados faltantes

Após a limpeza inicial dos dados, será realizada a inferência de valores ausentes, especialmente nas variáveis críticas para análise preditiva, como "Data da Primeira Mensagem" e "Avaliação do Atendimento". A inferência de dados ausentes visa reduzir o viés de análise que valores nulos podem introduzir e melhorar a representatividade dos dados para o modelo preditivo. Métodos como imputação de média ou modelos de aprendizado de máquina podem ser utilizados para inferir dados ausentes, evitando que lacunas interfiram na precisão da análise.

Segundo Batista e Monard (2003), o tratamento de dados faltantes é essencial para reduzir a incerteza e garantir uma análise preditiva mais robusta.

Preparação dos Dados para Análise Preditiva

Em seguida, os dados serão preparados para efetuar a correlação e análise preditiva, esse método inclui a transformação de variáveis, subamostragem, remoção dos limites inferiores e superiores e a divisão dos dados em conjuntos de treino e teste.

Aplicação da Regressão Logística

O método de Regressão Logística será utilizado para modelar a probabilidade de abandono do atendimento pelos clientes. A regressão logística é adequada para esse tipo de problema binário, cujo objetivo é prever se um cliente abandonará ou não o atendimento com base na variável independente Tempo de Espera.

Avaliação do Modelo

O modelo de regressão logística será avaliado utilizando métricas como acurácia, precisão, recall e a área sob a curva ROC (Receiver Operating Characteristic). Essas métricas ajudarão a determinar a eficácia do modelo em prever a evasão dos clientes tornando o estudo mais confiável.

Geração de Relatórios e Visualizações

Por fim, serão gerados gráficos e relatórios que mostrarão a probabilidade de evasão dos clientes e identificaram os padrões relevantes. Essas visualizações facilitarão a interpretação dos resultados.

RESULTADOS E DISCUSSÃO

Preparação dos dados

Para a análise preditiva do tempo de espera e sua relação com a evasão dos clientes em atendimentos via chat, partiu-se de um dataset original contendo 54.977 observações. Contudo, foram excluídos os dados do período de 01/12/2022 a 04/12/2022, pois não apresentavam registros válidos de "Data da Primeira Mensagem" e "Data da Última Mensagem", informações necessárias para calcular o tempo de espera e identificar atendimentos evadidos. Essa decisão foi baseada na importância da integridade dos dados para evitar vieses que possam distorcer os resultados (Rahm; Do, 2000).

Após essa filtragem, o dataset final incluiu 28.583 registros de atendimentos, divididos entre os setores de Folha (11.894), Gestão (9.996) e Fiscal (6.693). Para facilitar a modelagem preditiva, três novas colunas foram criadas: "Desistência" (variável binária indicando evasão ou não), "Data de Desistência" (data estimada para os casos de evasão) e "Tempo de Espera" (duração entre a entrada na fila e a resposta inicial ou o abandono).

Cálculo do Tempo de Espera para Desistentes e Não Desistentes

Para atendimentos não evadidos, o tempo de espera foi calculado como a diferença entre a "Data da Última Mensagem" e a "Data de Entrada na Fila", refletido no tempo que o cliente aguardou até ser atendido. Em contrapartida, para atendimentos evadidos, em que as datas de "Primeira Mensagem" e "Última Mensagem" estavam padronizadas como "01/01/2001", foi necessário adotar uma abordagem de inferência. Utilizando a técnica de inferência por mediana, a data de desistência foi aproximada pela mediana entre a "Data de Entrada" e a "Data do Primeiro Atendimento Bem-Sucedido" subsequente. A Figura 1 ilustra trecho de código responsável pela inferência.

Figura 1 - Trecho de código responsável por calcular a data de desistência

```
function calcularDataDesistencia(registroDesistencia: Registro, proximoAtendimentoSucedido: Registro): Registro {
  if (!proximoAtendimentoSucedido) {
    registroDesistencia.DataDesistencia = dayjs(ultimoAtendimentoSucedidoDoDia?.UltimaMensagem).toDate();
  } else {
    // Obtém os horários das mensagens
    const entradaDesistencia = dayjs(registroDesistencia.Data);
    const entradaProximoSucedido = dayjs(proximoAtendimentoSucedido.Data);

    // Calcula a diferença de tempo em milissegundos
    const diferencaMillis = entradaProximoSucedido.valueOf() - entradaDesistencia.valueOf();

    // Calcula o ponto médio do tempo
    const meioTempoMillis = diferencaMillis / 2;

    // Define a nova DataDesistencia como o ponto médio
    registroDesistencia.DataDesistencia = entradaDesistencia.add(meioTempoMillis, 'millisecond').toDate();
  }

  return registroDesistencia;
}
```

Fonte: autoria própria.

Esse método permitiu uma estimativa do horário de evasão, como sugerido por Batista e Monard (2003), sobre o uso de técnicas de imputação para dados ausentes. No entanto, observou-se um caso especial: atendimentos ocorridos ao final do dia. No entanto, foi identificada uma exceção no caso de atendimentos realizados ao final do expediente. Por se tratarem dos últimos registros do dia, esses atendimentos não apresentavam um sucessor bem-sucedido no mesmo dia, o que impediu a inferência precisa da Data de Desistência.

Ao final do processo, foi gerada uma tabela contendo as seguintes colunas: *Setor, Nome, Atendente, Primeira Mensagem, Última Mensagem, Desistência, Data de Desistência e Tempo de Espera* (Quadro 3). A estrutura dessas colunas foi projetada para garantir que as informações essenciais ao cálculo e a análise dos atendimentos evadidos fossem devidamente registradas.

Quadro 3 - Descrição de cada variável

Variável	Descrição
Setor de Atendimento	Identifica o setor solicitado pelo cliente, subdividido em Folha de Pagamento, Fiscal/Contábil e Gestão.
Data de Entrada	Representa o momento em que o cliente ingressou na fila de espera por atendimento.
Data de Primeira Mensagem	Marca a primeira interação do cliente com o sistema.
Data da Última Mensagem	Indica a última mensagem enviada pelo cliente.
Desistência	Classificação binária que indica se o atendimento foi finalizado com desistência (1) ou não (0).
Data de Desistência	Estimativa da data e horário de desistência do cliente, quando possível, usada para medir o tempo de espera até o momento de evasão
Tempo de Espera	Calculado com base no intervalo médio entre a Data de entrada e a Data de desistência.

Fonte: autoria própria.

Processo de Preparação dos dados e Treinamento do Modelo

O modelo preditivo para análise de desistência de clientes em atendimentos foi desenvolvido com um processo de preparação de dados, subamostragem e treinamento para garantir a maior precisão e relevância dos resultados. A seguir, descrevem-se cada etapa e motivo das escolhas.

Remoção de Dados com tempo de espera Nulo ou Igual a 0

A primeira etapa do processo envolveu a remoção de registros no qual a variável responsável por armazenar o tempo de espera era nula ou igual a zero. Essa variável é fundamental para o modelo, pois representa o tempo de espera que o cliente enfrentou antes de ser atendido ou desistir. Dados com tempo de espera igual a zero ou nulos indicam atendimentos inconsistentes, que poderiam distorcer os resultados da análise e reduzir a eficácia do modelo. Ao excluir esses dados, garante que o modelo trabalhe apenas com informações válidas e relevantes, eliminando possíveis vieses.

Análise Descritiva e Subamostragem

Antes de iniciar a análise preditiva, realizou-se uma avaliação inicial dos dados manipulados, que consistiam em três conjuntos de dados (datasets) representando os setores de atendimento da empresa. O maior deles era o setor de **Folha de Pagamento**, com um total de 11.893 observações, das quais 1.863 atendimentos foram classificados como "Desistência" (evasão) e 10.030 como "Não desistência". No setor de **Gestão**, foram registradas 9.995 observações, sendo 1.791 como "Desistência" e 8.204 como "Não desistência". O setor **Fiscal/Contábil** tinha 6.692 observações, com 1.186 classificadas como "Desistência" e 5.506 como "Não desistência".

Para balancear o conjunto de dados e evitar vieses, foi aplicada a técnica de subamostragem aleatória em cada um dos três setores, igualando as amostras de "Desistência" e "Não desistência". Abaixo, o Quadro 4 apresenta os resultados descritos acima para cada setor antes e depois da subamostragem.

Quadro 4 - Contagem de atendimentos antes e após a Subamostragem

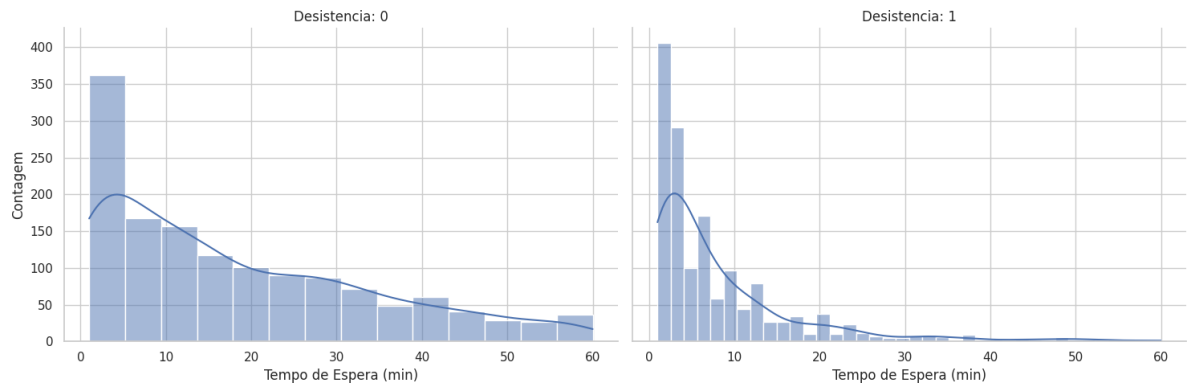
Setor	Contagem antes da Subamostragem	Contagem após a Subamostragem
Folha de Pagamento	Desistência: 1.863	Desistência: 1.863
	Não desistência: 10.030	Não desistência: 1.863
Gestão	Desistência: 1.791	Desistência: 1.791
	Não desistência: 8.204	Não desistência: 1.791
Fiscal/Contábil	Desistência: 1.186	Desistência: 1.186
	Não desistência: 5.506	Não desistência: 1.186

Fonte: autoria própria.

A aplicação desta técnica resultou em um conjunto de dados balanceado, garantindo que a análise preditiva fosse conduzida de maneira justa e sem tendência a um dos grupos.

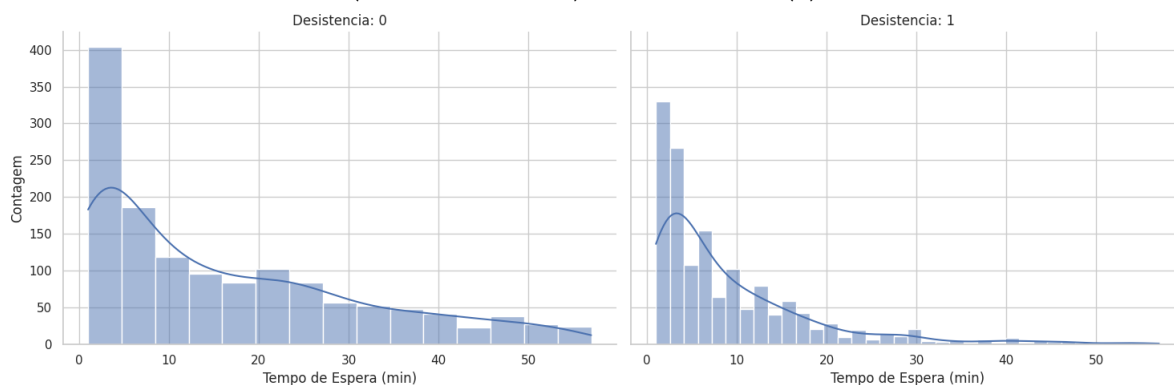
Após a remoção dos valores nulos e subamostragem randômica foi gerado gráficos de comparação de não desistentes e desistentes para cada um dos setores.

Figura 2 - Gráfico de comparação entre Desistentes: 0 (Não Desistentes) e Desistentes(1) do setor Folha



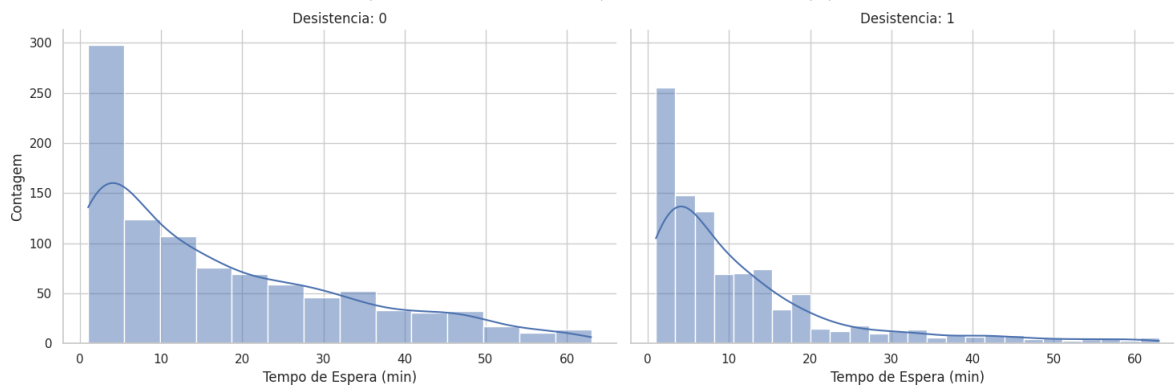
Fonte: autoria própria.

Figura 3 - Gráfico de comparação entre Desistentes: 0 (não desistentes) e Desistentes (1) do setor Gestão



Fonte: autoria própria.

Figura 4- Gráfico de comparação entre Desistentes: 0 (Não Desistentes) e Desistentes(1) do setor Fiscal



Fonte: autoria própria.

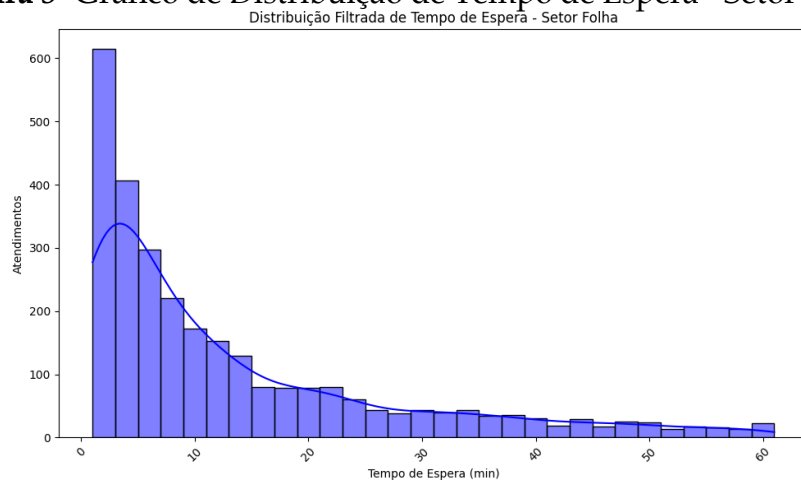
Observou-se na análise dos gráficos das Figuras 2, 3 e 4, para os setores Folha, Fiscal e Gestão respectivamente, que a distribuição dos atendimentos para Desistências igual a 0, ou seja, não desistentes possuíam valores que ultrapassaram os 50 minutos, essa distribuição já apresentava sinais de que quanto mais tempo se passava menor era a probabilidade de evasão dos atendimentos .

Definição dos limites percentuais: remoção dos 5% menores e maiores

Para refinar ainda mais o conjunto de dados e evitar que valores extremos interfiram na análise, foram definidos limites percentuais, excluindo os 5% menores e os 5% maiores valores de tempo de espera. Essa abordagem foi adotada para minimizar o impacto de outliers, que poderiam distorcer o modelo de regressão logística. Remover valores atípicos permite que o modelo se concentre em padrões mais representativos e gerais de desistência, garantindo uma previsão mais robusta.

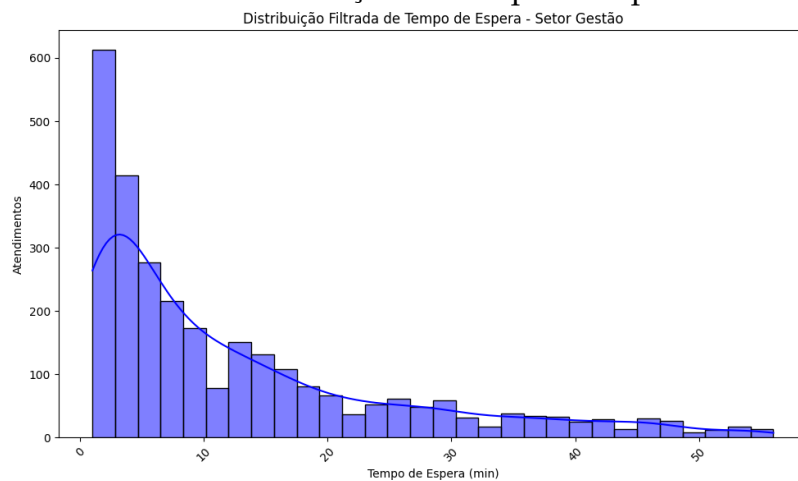
Após a remoção dos limites inferiores e superiores, os seguintes gráficos de Distribuição foram obtidos com relação a atendimentos por tempo de espera.

Figura 5- Gráfico de Distribuição de Tempo de Espera - Setor Folha



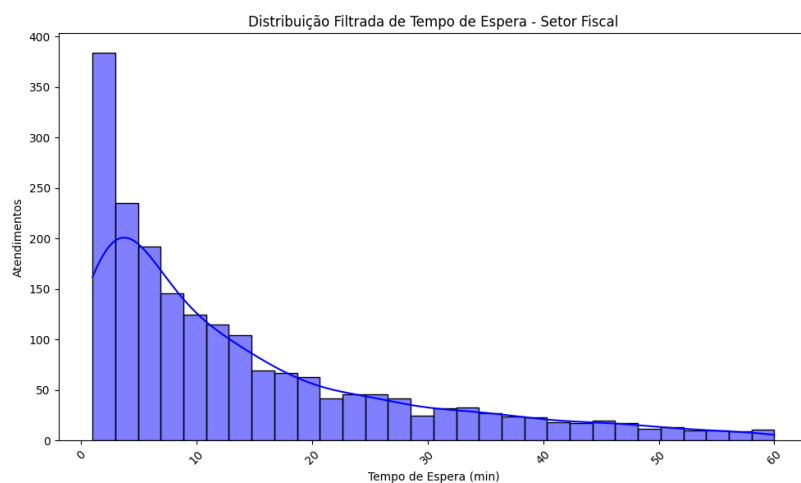
Fonte: autoria própria.

Figura 6 - Gráfico de Distribuição de Tempo de Espera - Setor Gestão



Fonte: autoria própria.

Figura 7 - Gráfico de Distribuição de Tempo de Espera - Setor Fiscal



Fonte: autoria própria.

Correlação de Point-Biserial

O coeficiente de correlação de Point-Biserial foi utilizado para medir a associação entre o tempo de espera e a desistência no atendimento. Os resultados indicaram uma correlação negativa para os três setores (Tabela 1), sugerindo que quanto maior o tempo de espera, menor a chance de desistência, comprovando de fato o que foi visto anteriormente nos gráficos das Figuras 2, 3 e 4.

Tabela 1 - Valor de correlação Point-Biserial para cada setor

Setor	Correlação Point-Biserial
Folha de Pagamento	-0,357
Gestão	-0,272
Fiscal/Contábil	-0,186

Fonte: autoria própria.

Vale ressaltar que o cálculo da correlação Point-Biserial foi realizada após a remoção dos outliers para verificar se o motivo da correlação inversa estava sendo causada pelos valores de limite superior e inferior, porém mesmo após a remoção dos valores ainda contava uma relação negativa para os três setores, reforçando ainda mais a ideia.

Divisão dos Dados em 60% para Treinamento e 40% para Teste

Para treinar e validar o modelo, os dados foram divididos em dois conjuntos: 60% para o treinamento e 40% para o teste. Essa divisão padrão em problemas de aprendizado de máquina ajuda a garantir que o modelo aprenda com uma amostra representativa, mantendo dados separados para avaliar sua eficácia. A proporção de 60/40 foi escolhida para oferecer um bom equilíbrio entre aprendizagem e validação, evitando o risco de overfitting.

Modelo de Regressão Logística

A regressão logística foi escolhida como o método estatístico principal para modelar a probabilidade de desistência com base no tempo de espera. A variável dependente "Desistência" foi tratada como uma variável binária (1 = desistência, 0 = não desistência), enquanto o "Tempo de Espera" foi a variável independente.

São apresentados nas tabelas abaixo os resultados gerados pela matriz de confusão para os 3 setores,

Tabela 2 - Matriz de confusão do setor de Folha de Pagamento

Previsão / Realidade	Desistência (Realidade)	Não Desistência (Realidade)
Desistência (Previsão)	519	289
Não Desistência (Previsão)	83	259

Fonte: autoria própria.

Tabela 3 - Matriz de confusão do setor de Gestão

Previsão / Realidade	Desistência (Realidade)	Não Desistência (Realidade)
Desistência (Previsão)	474	309
Não Desistência (Previsão)	122	255

Fonte: autoria própria.

Tabela 4 - Matriz de confusão do setor Fiscal

Previsão / Realidade	Desistência (Realidade)	Não Desistência (Realidade)
Desistência (Previsão)	268	200
Não Desistência (Previsão)	122	201

Fonte: autoria própria.

Gráfico da Curva ROC para Cada Setor: Folha, Fiscal e Gestão

Por fim, o desempenho do modelo foi avaliado utilizando o gráfico da curva ROC (Receiver Operating Characteristic), que permite visualizar a taxa de verdadeiros positivos em relação aos falsos positivos para cada setor. A área sob a curva (AUC) representa a precisão do modelo – quanto mais próxima de 1, melhor a performance. Os gráficos ROC para os setores de Folha, Fiscal e Gestão permitem comparar a eficácia do modelo em cada setor específico, mostrando seu acerto em prever a desistência.

Figura 8 - Curva ROC - Setor Folha

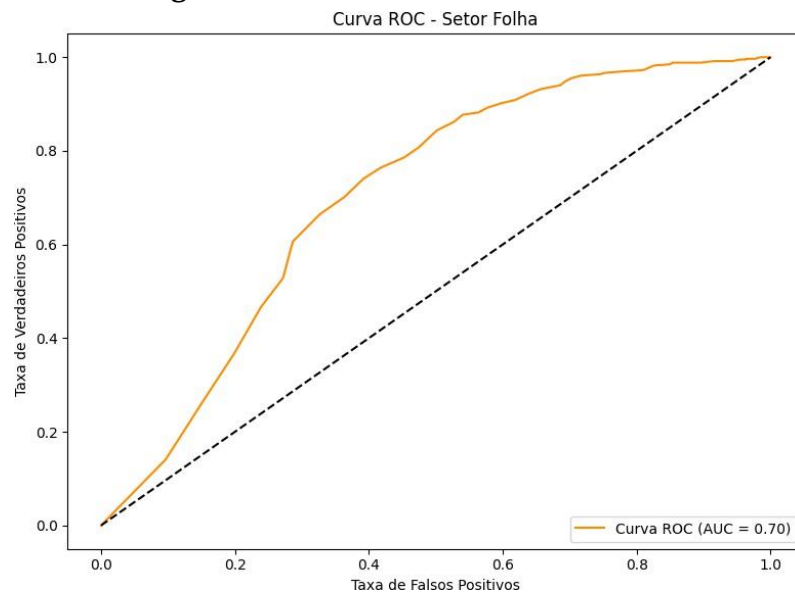
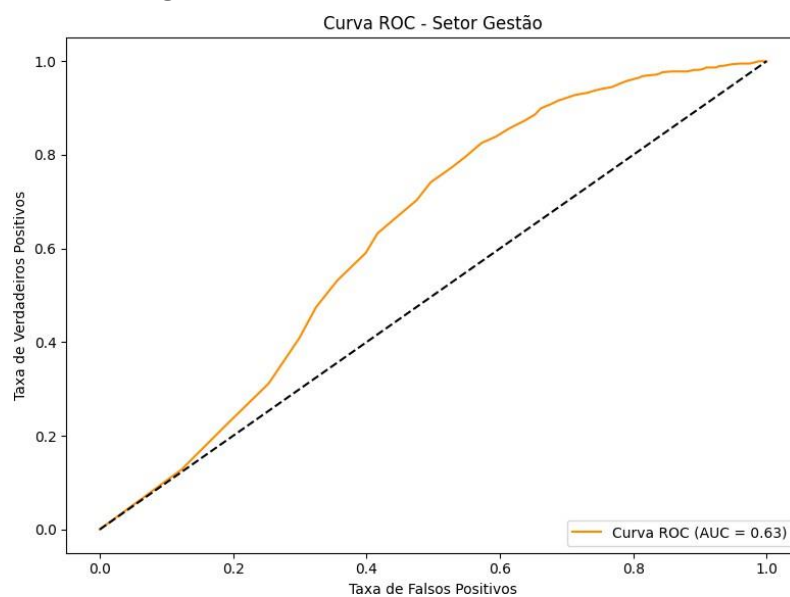
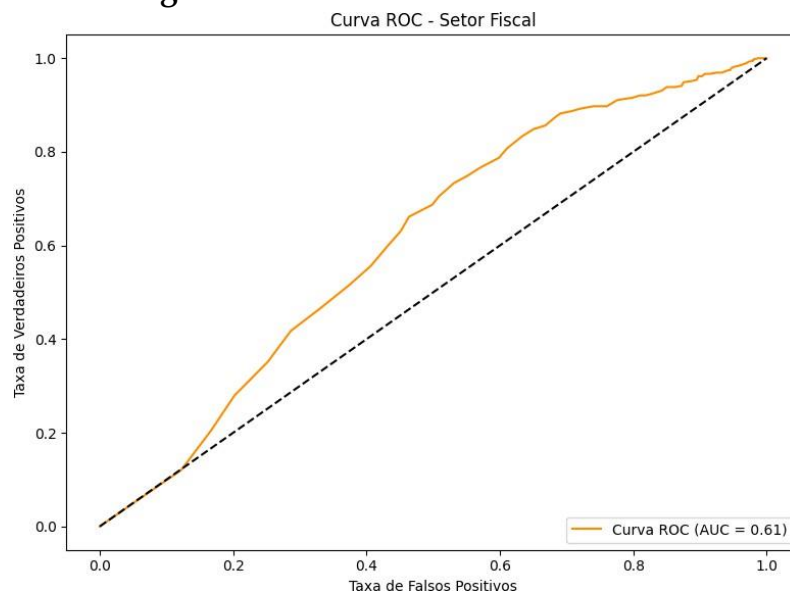


Figura 9 - Curva ROC - Setor Gestão



Fonte: autoria própria.

Figura 10 - Curva ROC - Setor Fiscal



Fonte: autoria própria.

Através da Matriz de confusão e sobre o gráfico da curva ROC, foram obtidos valores favoráveis para o setor de Folha com precisão de acerto sobre o modelo criado de 70% de precisão, enquanto para os setores Fiscal e Gestão os valores foram de 61% e 63%, respectivamente. Um dos possíveis motivos seria a baixa correlação entre as variáveis tempo de espera e desistência para esses setores e a quantidade de dados para treinamento inferior ao setor de folha.

CONCLUSÃO E SUGESTÕES

Para alcançar o objetivo principal de analisar o impacto do tempo de espera

na evasão de clientes em atendimentos via chat, o estudo realizou uma série de etapas. Primeiramente, os dados coletados da plataforma Contako, abrangendo o período de 2022 a 2023, foram cuidadosamente limpos, removendo erros e informações irrelevantes que poderiam comprometer a análise. Em seguida, procedeu-se à inferência de dados faltantes, garantindo a completude das informações cruciais para a pesquisa, como as datas de evasão. A fim de investigar a relação entre as variáveis, foram aplicadas técnicas de correlação de variáveis, buscando identificar padrões e tendências nos dados. Para aprofundar a análise e construir um modelo preditivo, utilizou-se a Regressão Logística, uma técnica estatística que permitiu modelar a probabilidade de evasão dos clientes com base no tempo de espera. Finalmente, para assegurar a confiabilidade do modelo preditivo, realizou-se a validação por meio da curva ROC, a qual permitiu avaliar a precisão do modelo na previsão da evasão dos clientes.

Os resultados da análise, utilizando a correlação de Point-Biserial, indicaram uma correlação negativa entre o tempo de espera e a evasão nos três setores estudados (Folha, Gestão e Fiscal). Isso significa que, ao contrário do esperado, quanto maior o tempo de espera, menor a chance de desistência. Essa relação foi mais forte no setor de Folha. A observação dos gráficos de distribuição de tempo de espera (Figuras 2, 3 e 4) já indicava essa tendência, com a maioria dos atendimentos não evadidos (Desistência = 0) concentrando-se em tempos de espera mais longos.

A Regressão Logística, utilizada para modelar a probabilidade de desistência, apresentou resultados favoráveis, principalmente para o setor de Folha, com precisão de 70%. Os setores Fiscal e Gestão obtiveram, respectivamente, precisões de 61% e 63%, possivelmente devido à menor correlação entre as variáveis e à menor quantidade de dados de treinamento. A curva ROC confirmou a maior confiabilidade do modelo no setor de Folha. Conclui-se que o tempo de espera, embora tenha impacto, não é o único fator determinante para a evasão, especialmente nos setores Fiscal e Gestão. Outros fatores, como o motivo da solicitação e a prioridade do atendimento, podem influenciar a decisão do cliente.

Sugestões para futuras pesquisas incluem a inclusão de variáveis explicativas adicionais, análise de horários de pico e padrões sazonais, e ampliação do conjunto de dados para aumentar a precisão do modelo preditivo. A investigação de outras métricas de atendimento, além do tempo de espera, também pode contribuir para uma compreensão mais abrangente do fenômeno de evasão.

Com essas medidas, a empresa poderá obter uma visão mais completa do comportamento de desistência dos clientes, otimizando seus atendimentos, aumentando a retenção e a satisfação.

REFERÊNCIAS

ALPAYDIN, E. **Introduction to machine learning**. MIT Press, 2020.

BATISTA, G. E.; MONARD, M. C. An analysis of four missing data treatment methods for supervised learning. **Applied Artificial Intelligence**, v. 17, n. 5-6), p. 519-533, 2023.

COHEN, J.; COHEN, P.; WEST, S. G.; AIKEN, L. S. **Applied multiple**

regression/correlation analysis for the behavioral sciences. 3. ed. Lawrence Erlbaum Associates, 2003.

CONTAKO. **Sistema de chat multicanal de atendimento**. 2024. Relatório técnico. Disponível em: <https://www.contako.com.br/>. Acesso em: 19 ago. 2024.

CRESWELL, J. W. **Research design: qualitative, quantitative, and mixed methods approaches**. SAGE Publications, 2014.

DANCEY, C. P.; REIDY, J. **Statistics without maths for psychology**. 4. ed. Prentice Hall, 2007.

HYNDMAN, R. J.; ATHANASOPOULOS, G. **Forecasting: principles and practice**. OTexts, 2018.

KUMAR, D.; GARG, P. K. Predictive analytics in human resource management. **Journal of Human Resource Management**, v. 6, n. 2, p. 28-33, 2018.

KUMAR, S.; CHONG, I. Correlation analysis to identify the effective data in machine learning: prediction of depressive disorder and emotion states. **International Journal of Environmental Research and Public Health**, v. 15, n. 12, p. 2907, 2018.

LAUDON, K. C.; LAUDON, J. P. **Sistemas de informação gerenciais**. 7. ed. São Paulo: Pearson Prentice Hall, 2004.

MENARD, S. **Applied logistic regression analysis**. SAGE Publications, 2002.

RAHM, E.; DO, H. H. Data cleaning: problems and current approaches. **IEEE Data Engineering Bulletin**, v. 23, n. 4, p. 3-13, 2000.

VAN RIEL, M. L.; SCHEPERS, J.; DE RUYTER, K. Customer feedback and innovation: the role of online customer communities in enhancing product development. **Journal of Service Research**, v. 16, n. 4, p. 514-528, 2013.

WIERINGA, R. J. **Design science methodology for information systems and software engineering**. Berlin: Springer, 2014.

ZENDESK. **Qual a importância das redes sociais para as empresas?** 2024. Disponível em: <https://www.zendesk.com.br/blog/qual-a-importancia-das-redes-sociais-para-as-empresas/>. Acesso em: 16 jun. 2024.