ANÁLISE E PREDIÇÃO

Churn de clientes

Insights estratégicos & Machine Learning

- ☑ Machine Learning;
- ☑ Predição de churn;
- ☑ Análise de dados;
- ☑ Metodologia CRISP-DM;
- ☑ Análise exploratória de dados (EDA);
- ☑ Plano de ações futuras;
- ☑ Python;
- ☑ Análise preditiva;
- ☑ DataViz;

PRINCIPAIS TÓPICOS DO PROJETO

01

PROBLEMA DE NEGÓCIO

02

METODOLOGIAS E TÉCNICAS UTILIZADAS

03

ANÁLISE EXPLORATÓRIA
- Entendimento dos Dados;
- Preparação dos Dados;
- Modelagem dos dados;

04

ANÁLISE DE ASSOCIAÇÃO

CORRELAÇÕES E ASSOCIAÇÕES DE VARIÁVEIS

05

AVALIAÇÃO DOS RESULTADOS
- Machine Learning;

06

PLANOS DE AÇÃO (Futuras)

07

CONCLUSÃO

08

BASES E LINKS IMPORTANTES

Projeto **Prevenção de Churn** em um app de comida

Bem-vindos! Esta apresentação explorará um estudo detalhado sobre o churn de clientes no App ToComFome, com o objetivo de identificar os principais fatores de risco, segmentar clientes com alta probabilidade de churn, prever/atribuir probabilidade de churn dos clientes no próximos 4 meses e propor ações para aumentar a retenção de clientes.

Para isso, a área de CRM passou uma amostra de cerca de 10 mil clientes com suas respectivas informações de cadastro e transações nos próximos 4 meses a contar da data de extração usada como referência.

Metodologia
CRISP-DM/ESTATÍSTICA/ML

CRISP-DM (Cross Industry Standard Process for Data Mining) é uma metodologia padrão com uma estrutura flexível e iterativa que consiste em seis etapas principais:

- **Entendimento do Negócio:** Definir objetivos e requisitos do projeto.
- **Entendimento dos Dados:** Coletar e explorar dados relevantes.
- **Preparação dos Dados:** Limpar e transformar dados para análise.
- **Modelagem:** Selecionar e aplicar algoritmos de aprendizado de máquina.
- **Avaliação:** Avaliar a performance dos modelos.
- **Implementação:** Integrar e monitorar os modelos no ambiente de produção.

Técnicas Estatísticas Utilizadas:

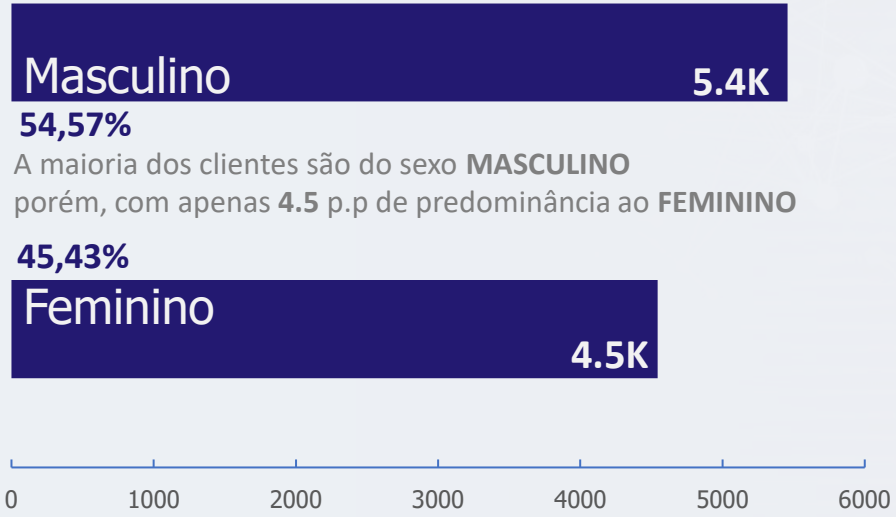
- **Análise Univariada:** Estudo de uma única variável para entender suas propriedades básicas.
- **Análise Bivariada:** Exploração da relação entre duas variáveis para identificar correlações e padrões.
- **Information Value (IV):** Métrica utilizada para determinar a importância de variáveis preditoras na modelagem de churn.
- **Random Forest:** Algoritmo de aprendizado de máquina utilizado para classificação de clientes com alta probabilidade de churn, combinando várias árvores de decisão para melhorar a precisão e a robustez do modelo.

Analise **UNIVARIADA** Qualitativa

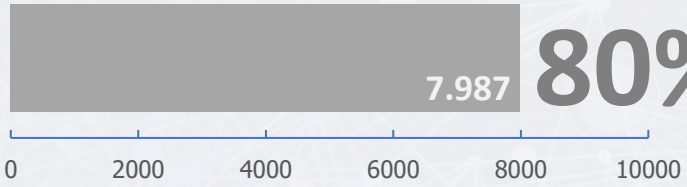
Churn

Um cliente é classificado como Churn nesse projeto quando nos próximos 4 meses em relação à data de referência, ele não transacionou por um período de 40 dias.

Genêro



Churn



Tivemos **2.013** casos de churn, e esse será o foco da análise que será identificar as características desse cliente de churn, para projetar e encontrar futuros casos de churn.

Estados

50,14%
São Paulo

25,09%
Minas Gerais

24,77%
Rio de Janeiro

Os clientes estão centralizados na **região sudeste** com forte participação do estado de **São Paulo**

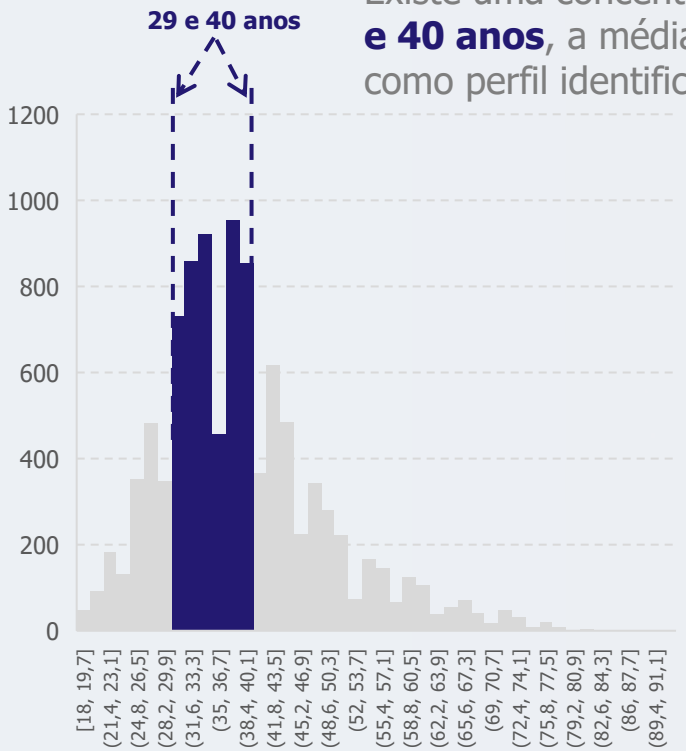
Analise **UNIVARIADA** Quantitativa

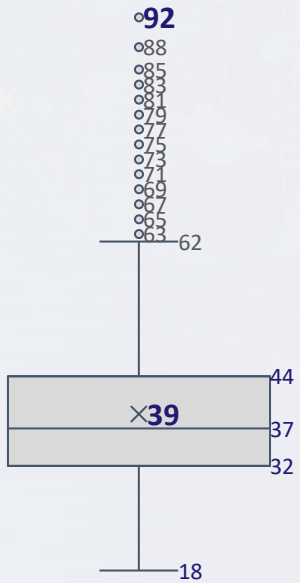
Variáveis Quantitativas

Nessa etapa, foi necessário avaliar cada uma das variáveis quantitativas de forma univariada, buscando, dessa vez, algum tipo de padrão e informação importante para a análise

Idade

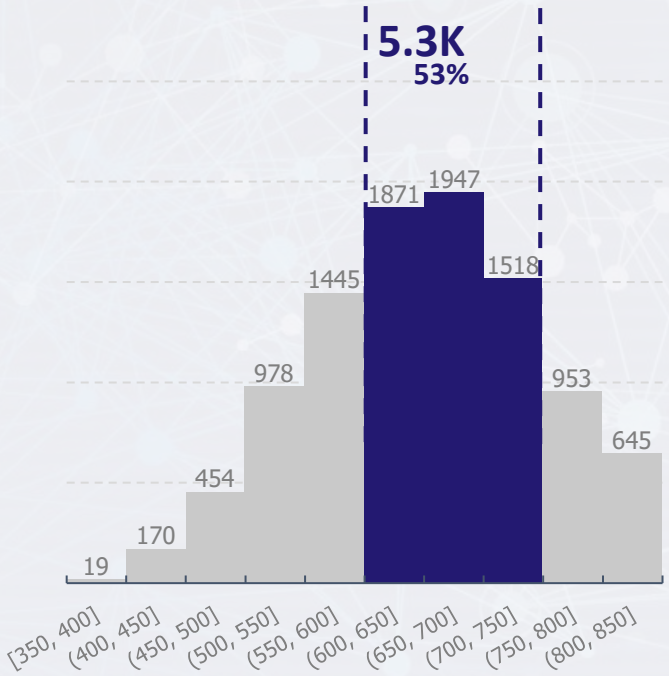
Existe uma concentração de cliente entre **29 e 40 anos**, a média de idade são 39 anos, como perfil identificado.

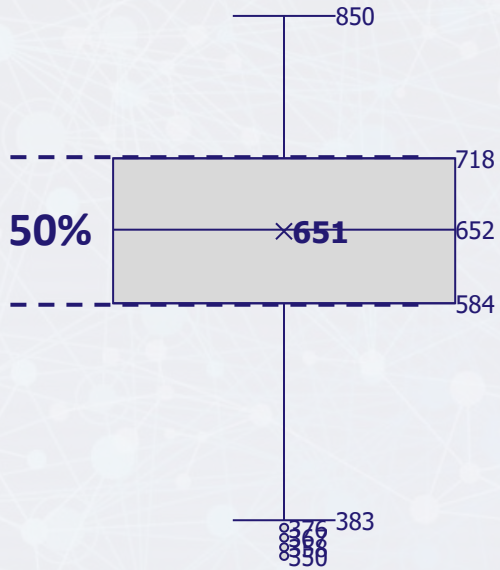




Score de Crédito

Cerca de **5.3K** de clientes possuem o **score de crédito** entre **600** e **750**, **53% dos clientes** da base utilizada para análise.

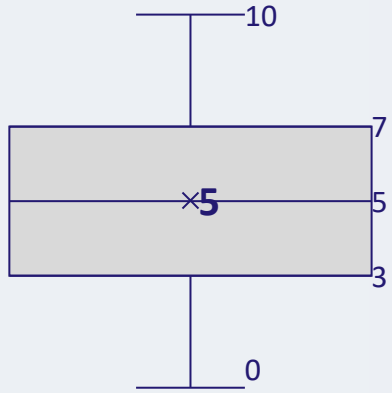




A **média de score** de crédito dos clientes são **651**, não sendo muito influenciada pelos outliers. Dessa forma o score possui **pouca variação** entre o maior volume de clientes.

Analise **UNIVARIADA** Quantitativa

Tempo como Cliente



A maior variação de tempo como cliente está em 3 e 7 meses. Sendo a **média 5** meses.

Tempo_cliente	Freq.	Fred Relativa	Freq. Relativa acc
0	413	4%	4%
1	1.035	10%	14%
2	1.048	10%	25%
3	1.009	10%	35%
4	989	10%	45%
5	1.012	10%	55%
6	967	10%	65%
7	1.028	10%	75%
8	1.025	10%	85%
9	984	10%	95%
10	490	5%	100%
Total Geral	**10.000**	**100%**	**100%**

Interessante entender que a média são de 5 meses, porém **91%** dos clientes possuem de **1 à 9 meses como clientes** com a mesma proporção de 10% da base de pesquisa.

Quant_Produtos

Quant. Categorias	Freq.	Fred Relativa	Fred Relativa
1	5.084	51%	51%
2	4.590	46%	97%
3	266	3%	99%
4	60	1%	100%
Total Geral	**10.000**	**100%**	

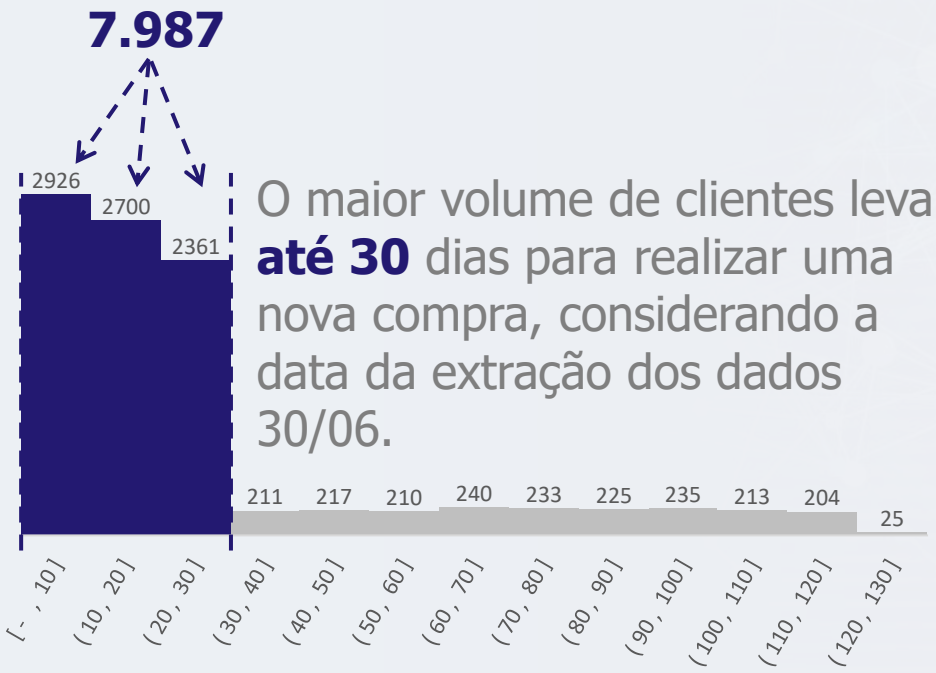
A grande maioria dos clientes **97%** compram até **duas categorias de produtos.**

*Como não temos os nomes dos produtos e categorias, essa variável pode ser inconclusiva, será necessário questionar e entender quais são as categorias e os produtos.

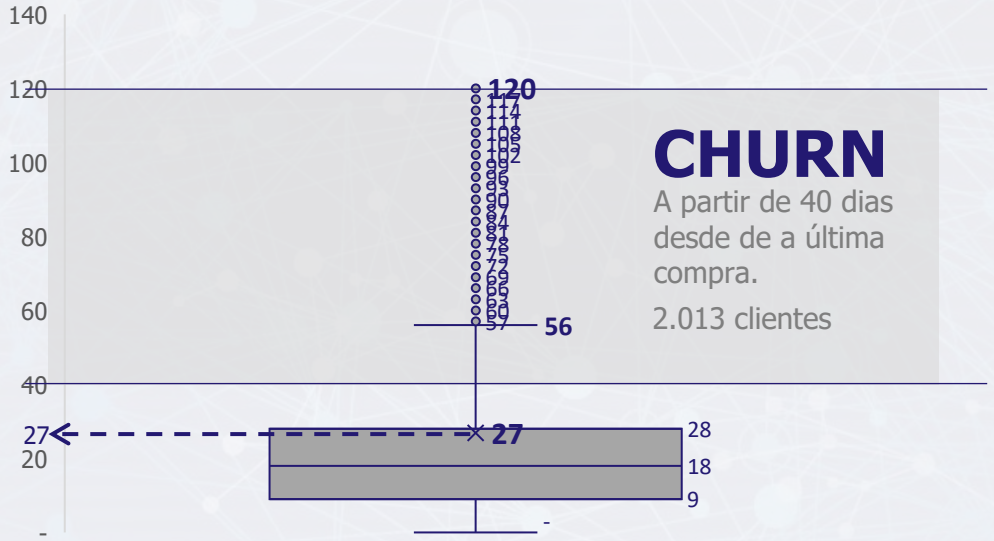
Analise **UNIVARIADA** Quantitativa

Tempo desde de a ultima compra

Considerando a data de extração do dia 30/06 e analisando a data da última compra, temos o resultado abaixo, que mostra quanto tempo cada cliente leva para realizar suas compras no aplicativo. Isso nos ajuda a entender melhor o comportamento dos clientes.



O maior volume de clientes leva **até 30** dias para realizar uma nova compra, considerando a data da extração dos dados 30/06.



A média de tempo para que o cliente efetue uma nova compra no aplicativo são de **27 dias**, influenciada fortemente por outliers de até 120 dias, que já são considerados casos de churn.

*Nesse caso podemos utilizar também a mediana de 18 dias, pois a média está sendo fortemente influenciada pelos outliers.

PERFIL DO CLIENTE

54,57%
MASCULINO

A maioria dos clientes são do sexo **MASCULINO**, porém existe um pequeno equilíbrio, aonde 45,43% são do sexo **FEMININO.**

IDADE MÉDIA
39 ANOS

50,14%
São Paulo

A **média de score** de crédito dos clientes são **651.**

02 CATEGORIAS DE PRODUTOS

Costumam comprar **02 categorias** de produtos por pedido.

Analise de **ASSOCIAÇÃO**

Analise de Associação

Information Value (IV)

Information Value (IV) é uma métrica estatística utilizada para avaliar a capacidade preditiva de variáveis em modelos de classificação, especialmente em modelos de risco e churn. A técnica ajuda a identificar quais variáveis possuem maior poder de discriminação entre diferentes classes. Nesse caso se o cliente permaneceu ou saiu da empresa.
Se uma variável tem poder forte de separação, significa que uma ou mais categorias da variável tem um alto nível de churn, sendo útil estuda-la com mais profundidade.

VARIÁVEIS	IV	Poder de separação
Quantidade de categorias	0,93	Muito Forte
Idade	0,85	Muito Forte
Estado	0,17	Médio
Programa fidelidade	0,15	Médio
Limite de Crédito	0,13	Médio
Genero	0,07	Médio
Tempo como cliente	0,01	Muito Fraco
Scores	0,01	Muito Fraco
Soma pedidos	0,00	Muito Fraco
usa cartão	0,00	Muito Fraco

IV TOTAL	PODER SEPARAÇÃO
<0,02	Muito Fraco
0,02 à 0,1	Fraco
0,1 à 0,3	Médio
0,3 à 0,5	Forte
>0,5	Muito bom

Dentro do nossa análise já conseguimos identificar as variáveis que poderão ser importantes para uma análise profunda e também uso para o machine learning.

As variáveis IDADE, QUANTIDADE DE CATEGORIAS, PROGRAMA DE FIDELIDADE E LIMITE DE CRÉDITO serão utilizadas nesse momento.

Qualquer variável com poder de separação MUITO FRACO, podemos desprezar, caso a empresa tenha um custo com a compra desses dados, podemos propor não utiliza-los mais, gerando economia.

Analise de **ASSOCIAÇÃO**

ESTADOS	SIM	NÃO	TOTAL	SIM	NÃO	FREQ. CHURN	IV
Minas Gerais	804	1.705	2.509	40%	21%	32%	0,12
Rio de Janeiro	407	2.070	2.477	20%	26%	16%	0,01
São Paulo	802	4.212	5.014	40%	53%	16%	0,04
Total Geral	**2.013**	**7.987**	**10.000**	**100%**	**100%**	**20%**	**0,17**

Clientes de **MINAS GERAIS**, possuem uma maior probabilidade de churn.

QUANT. CAT.	SIM	NÃO	TOTAL	SIM	NÃO	FREQ. CHURN	IV
1	1.395	3.689	5.084	69%	46%	27%	0,09
2	343	4.247	4.590	17%	53%	7%	0,41
3	217	49	266	11%	1%	82%	0,29
4	58	2	60	3%	0%	97%	0,14
Total Geral	**2.013**	**7.987**	**10.000**	**100%**	**100%**	**20%**	**0,93**

Clientes que por padrão **compram 3 ou 4 categorias de produtos**, possuem uma maior chance de **churn**, porém em nossa base não temos as explicação sobre essa variável, então nesse primeiro instante pode não trazer uma boa resposta, temos que questionar a empresa sobre isso.

PROG. FID.	SIM	NÃO	TOTAL	SIM	NÃO	FREQ. CHURN	IV
0	1.286	3.563	4.849	64%	45%	27%	0,07
1	727	4.424	5.151	36%	55%	14%	0,08
Total Geral	**2.013**	**7.987**	**10.000**	**100%**	**100%**	**20%**	**0,15**

Clientes sem o programa de fidelidade possuem uma **probabilidade** maior de churn com **27% de frequência de churn.**

IDADE	SIM	NÃO	TOTAL	SIM	NÃO	FREQ. CHURN	IV
18-22	20	206	226	1%	3%	9%	0,02
23-27	53	741	794	3%	9%	7%	0,08
28-32	138	1.632	1.770	7%	20%	8%	0,15
33-37	247	2.050	2.297	12%	26%	11%	0,10
38-42	366	1.653	2.019	18%	21%	18%	0,00
43-47	411	773	1.184	20%	10%	35%	0,08
48-52	346	324	670	17%	4%	52%	0,19
53-57	226	159	385	11%	2%	59%	0,16
58-62	133	163	296	7%	2%	45%	0,05
63-67	50	117	167	2%	1%	30%	0,01
68-72	20	87	107	1%	1%	19%	0,00
73-77	2	59	61	0%	1%	3%	0,01
78-82		17	17	0%	0%	0%	
83-87	1	3	4	0%	0%	25%	0,00
88-92		3	3	0%	0%	0%	
Total Geral	**2.013**	**7.987**	**10.000**	**100%**	**100%**	**20%**	**0,85**

A variável de IDADE, podemos identificar também que clientes **acima de 43 anos** possuem uma grande **probabilidade de churn.**

Podemos utilizar essas informações e variáveis para conseguir um melhor desempenho no algoritmo de machine learning para previsão de casos de churn de clientes.

Avaliação dos **RESULTADOS**

Machine Learning

É uma sub-área da inteligência artificial que permite que sistemas aprendam e façam previsões ou decisões baseadas em dados, sem serem explicitamente programados para tal. Utiliza algoritmos que identificam padrões e relações nos dados, melhorando a precisão das previsões à medida que novos dados são disponibilizados.

Regressão Logística

É um algoritmo de Machine Learning utilizado para modelar a probabilidade de um evento binário, como o churn de clientes. No nosso projeto, aplicamos a Regressão Logística para identificar os fatores que influenciam o churn e prever a probabilidade de um cliente cancelar o serviço. Esse modelo nos permite classificar os clientes como propensos ou não propensos a churn, auxiliando na criação de estratégias de retenção eficazes.

Algoritmo Python

Foi utilizado a linguagem python seguindo as seguintes etapas abaixo para construção do algoritmo de machine learning com o Random Forest.

1. **Coleta e Preparação de Dados**
   - Dados de cadastro e transações de clientes;
   - Tratamento de valores nulos e escalonamento; dos dados

2. **Divisão de Dados**
   - Separação em conjuntos de treinamento e teste (70/40);

3. **Modelagem**
   - Uso do algoritmo de Regressão Logística;
   - Treinamento do modelo com dados;

4. **Avaliação do Modelo**
   - Previsão de churn para dados de teste;
   - Cálculo de probabilidade de churn;
   - Geração de tabela de resultados;

Link para o código python: //github.com/Faustoalemos/Projeto_churn_machine_learning/blob/main/Machine%20Learning%20-%20Churn%20de%20aplicativo%20-%20Random%20Forest.v4.ipynb

Avaliação dos **RESULTADOS**

É importante entender alguns conceitos de custos e mais diretamente quais são os nossos objetivos com o projeto conforme abaixo.

Siglas e conceitos

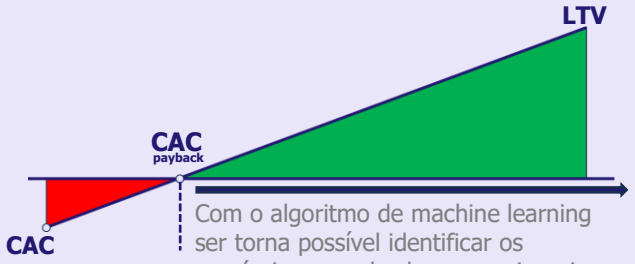
LTV (Life time Value)

Valor total previsto que um cliente trará para a empresa durante todo o período de seu relacionamento.

CAC (Customer Acquisition Cost)

Custo total gasto para adquirir um novo cliente, incluindo marketing, vendas e outros esforços relacionados.

Objetivo



Com o algoritmo de machine learning ser torna possível identificar os possíveis casos de churn e assim criar ações para que o cliente esteja sempre utilizando o aplicativo/produtos dentro desse Range positivo.

Custos Estimados

LTV
R$ 416,54

CAC
R$ 150,00

CUPOM DE DESCONTO
R$ 100,00

- Todos os valores são estimados, em um projeto real a área ou empresa poderia nos fornecer essas informações.

Confusion matrix

	CHURN	NÃO CHURN
CHURN	VP 2.061	FP 733
NÃO CHURN	FN 335	VN 1.664

Previsão correta
Previsão errada

Acurácia
78%

- Com a criação do algoritmo de machine learning chegamos a essa predição e acurácia a frente conseguiremos entender o retorno que a empresa terá com o projeto.

Avaliação dos **RESULTADOS**

MATRIZ DE IMPACTO POTENCIAL

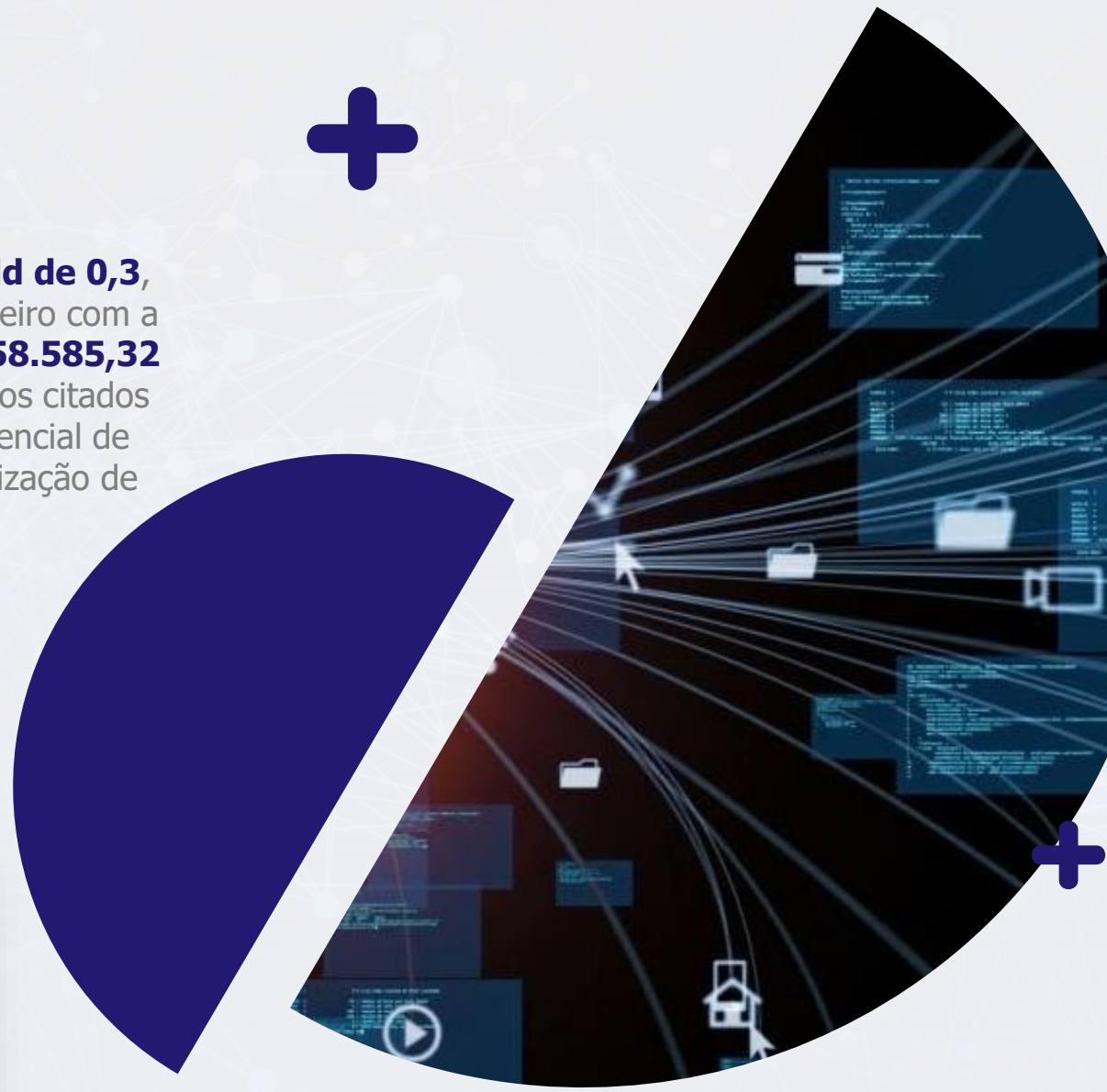
Resultado Final

	CHURN	NÃO CHURN
CHURN	R$ 343.239	-R$ 50.250
NÃO CHURN	R$ 122.073	R$ 443.522

Previsão correta Previsão errada

* Custos Estimados | LTV = R$ 416,54 | CAC = R$ 150,00 | CUPOM DE DESCONTO = R$ 100,00

Com a **acurácia de 78%** e o **thershold de 0,3**, conseguimos chegar ao resultado financeiro com a aplicação do machine learning de **R$ 858.585,32** positivo considerando os custos estimados citados anteriormente, trazendo um retorno potencial de **R$ 60.359,10** se comparado a não utilização de machine learning.



PROJEÇÃO DE RESULTADOS

CENÁRIOS

MACHINE LEARNING

R$ 798.226,22

Cupom para todos clientes

MACHINE LEARNING

R$ 858.585,32

Cupom apenas para clientes com potencial churn

Retorno potencial

R$ 60.359,10

*Retorno para cada 4.793 clientes.

MACHINE LEARNING

R$ 464.570,76

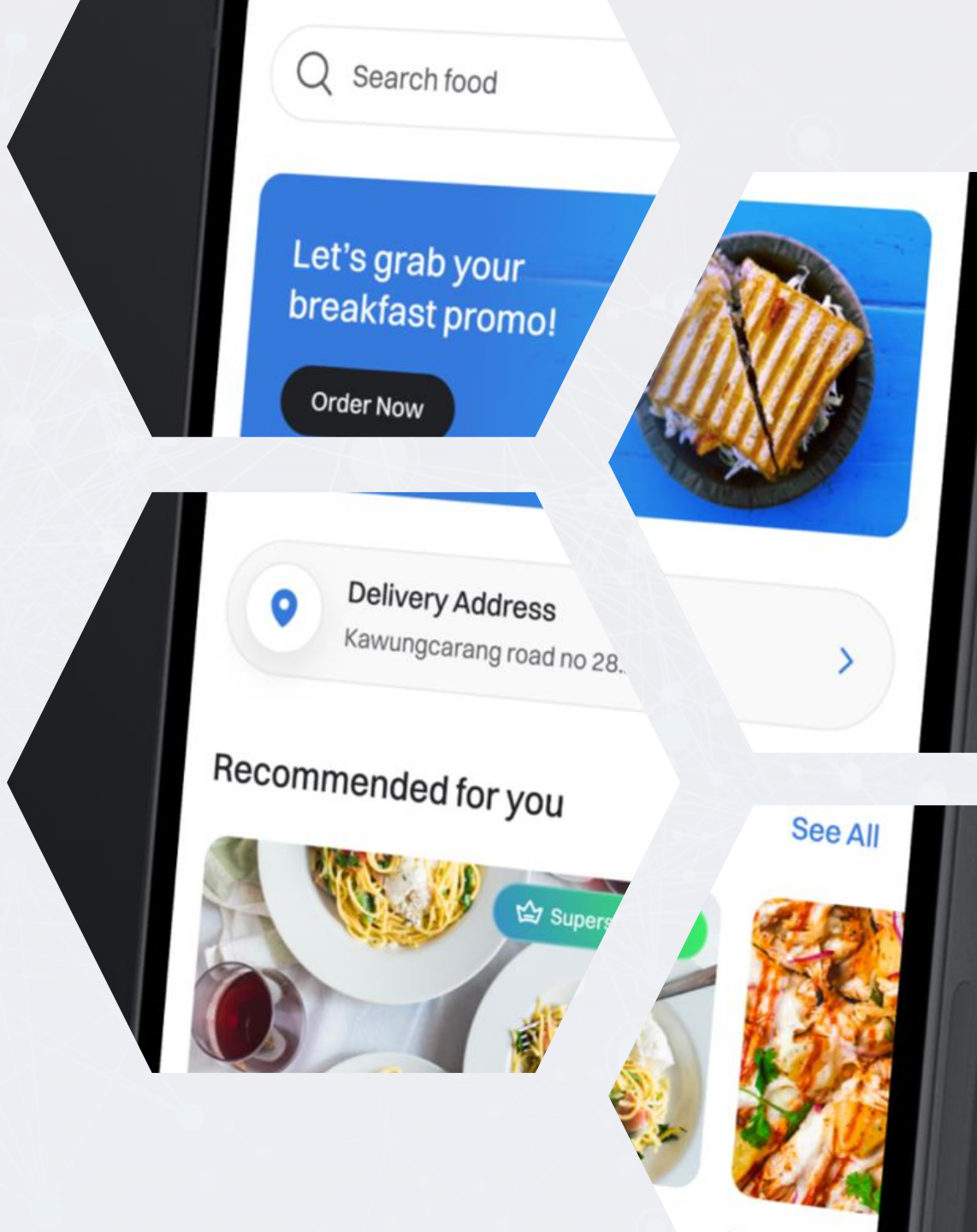
Não utiliza cupom

Plano de **AÇÕES**

Ação	O que	Por quê	Quem	Quando	Onde	Como	Quanto Custa
Desenvolvimento de Ofertas Personalizadas	Criar campanhas de marketing direcionadas e descontos exclusivos	Incentivar a renovação e reduzir o churn	Equipe de Marketing e CRM	Início imediato, com revisões mensais	Aplicativo e plataformas de comunicação	Análise de dados de clientes para criar ofertas direcionadas	Investimento inicial em análise de dados e campanhas de marketing
Melhoria na Experiência do Cliente	Personalizar a jornada do cliente e oferecer atendimento dedicado	Aumentar a satisfação e lealdade dos clientes	Equipe de Atendimento ao Cliente e UX	Início imediato, com avaliações contínuas	Aplicativo e canais de suporte	Implementar personalização baseada em dados e treinamento da equipe	Investimento em sistemas de personalização e capacitação da equipe
Implementação de Ações Proativas	Enviar alertas personalizados e realizar verificações de satisfação	Antecipar problemas e aumentar a satisfação do cliente	Equipe de Atendimento ao Cliente e Marketing	Início imediato, com revisões trimestrais	Aplicativo e canais de comunicação	Estabelecer mecanismos de notificação e pesquisas de satisfação	Investimento em sistemas de notificação e coleta de feedback
Análise Avançada de Dados	Utilizar machine learning para previsões e análise de sentimentos	Melhorar a precisão das previsões e identificar sinais de insatisfação	Equipe de Data Science e BI	Início após a implementação das outras ações	Sistemas de análise de dados	Implementar modelos avançados e análise de feedbacks	Custos associados a ferramentas de análise avançada e tempo da equipe
Estratégias de Engagement	Oferecer conteúdo exclusivo e gamificação	Aumentar o engajamento e lealdade dos clientes	Equipe de Marketing e Desenvolvimento de Produto	Início após a implementação das outras ações	Aplicativo e plataformas de comunicação	Desenvolver conteúdo exclusivo e implementar elementos de gamificação	Investimento em criação de conteúdo e desenvolvimento de gamificação
Personalização	Adaptar recomendações e ofertas com base no histórico de compras	Aumentar a relevância das ofertas e melhorar a experiência do usuário	Equipe de Data Science e Marketing	Início imediato, com atualizações contínuas	No aplicativo e nas comunicações com os clientes	Utilizar análise de dados para personalizar recomendações e ofertas	Investimento em ferramentas de análise de dados e desenvolvimento
Atendimento ao Cliente	Proporcionar suporte rápido e eficiente para resolver problemas	Melhorar a satisfação e fidelidade dos clientes	Equipe de Atendimento ao Cliente	Início imediato, com monitoramento contínuo	Canais de suporte (telefone, chat, e-mail)	Treinamento contínuo da equipe e implementação de sistemas de atendimento eficientes	Investimento em treinamento e ferramentas de atendimento
Comunicação Regular	Manter os clientes informados sobre novidades, promoções e atualizações	Manter os clientes engajados e informados	Equipe de Marketing e Comunicação	Comunicações regulares (semanal, mensal)	E-mails, notificações no aplicativo, redes sociais	Criar um calendário de comunicação e conteúdo relevante para os clientes	Investimento em criação de conteúdo e ferramentas de comunicação

CONCLUSÃO

Com a implementação das estratégias de retenção baseadas em Machine Learning e a personalização de ofertas e comunicações, conseguimos identificar e atuar proativamente sobre clientes em risco de churn. A automação e a análise contínua de dados não só irão aumentaram a eficiência operacional, como também irão melhorar a precisão das nossas previsões, garantindo uma experiência superior para nossos clientes. O ideal é continuar aprimorando as abordagens para manter e fortalecer a lealdade dos nossos clientes.



BASES E LINKS

Referências de Bases de Dados

- Base Churn Tratada: https://github.com/Faustoalemos/Projeto_churn_machine_learning

Links do código versão python:

- Repositório GitHub: https://github.com/Faustoalemos/Projeto_churn_machine_learning/blob/main/Machine%20Learning%20-%20Churn%20de%20aplicativo%20-%20Random%20Forest.v4.ipynb