Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Marcelo Bueno de Oliveira

MODELO PARA IDENTIFICAÇÃO DE CLIENTES COM PROPENSÃO AO DESENGAJAMENTO

Curitiba
2022

Marcelo Bueno de Oliveira

MODELO PARA IDENTIFICAÇÃO DE CLIENTES COM PROPENSÃO AO DESENGAJAMENTO

Trabalho de Conclusão de Curso apresentada ao Programa de Especialização em Data Science e Big Data da Universidade Federal do Paraná como requisito à obtenção do título de grau de especialista.

Orientador: Prof. Dr. Wagner Hugo Bonat

Curitiba
2022

Especialização em Data Science e Big Data
Universidade Federal do Paraná
dsbd.leg.ufpr.br

Modelo Para Identificação de Clientes com Propensão ao Desengajamento

Marcelo Bueno de Oliveira[1]
Prof. Dr. Wagner Hugo Bonat [2]

Resumo

O presente artigo tem por objetivo apresentar a métrica LF (*Life Time*) utilizada para monitorar o engajamento do cliente com o produto cartão de crédito em uma instituição financeira, a qual foi definida através da observação da frequência do uso do produto. Esta técnica atrelada à um modelo preditivo possibilita a avaliação dos fatores comportamentais que antecedem ao efeito do desengajamento. Desta forma a empresa pode agir antecipadamente e estabelecer estratégias de marketing, que visem manter o cliente usando o produto (engajado). Portanto monitorar o engajamento do cliente permite amadurecer o relacionamento com o mesmo, fortalecer a fidelização a marca, aumentar o faturamento da empresa e reduzir o cancelamento ou a substituição do cartão de crédito.

Palavras-chave: *Ciclo de Vida, Engajamento, Desengajamento, fidelização a marca, aumentar faturamento, reduzir cancelamento, cartão de crédito.*

Abstract

The article aims to present an objective LF metric (LIFE TIME) used to use the product presented customer engagement with a credit card in a financial institution, which was defined by observing the frequency. This technique linked to a predictive model enables the evaluation of behavioral behaviors that precede the effect of engagement. In this way the company can act and establish marketing strategies that aim to keep the customer using the product (engaged). Increase company credit and reduce credit card cancellation or replacement.

Keywords: *Lifecycle, Engagement, Disengagement, brand loyalty, increase revenue, reduce cancellation, credit card.*

I Introdução

Compreender o ciclo de vida do cliente e o produto faz parte do planejamento estratégico de marketing e vendas da maioria das empresas no cenário atual.

Esse trabalho se insere no contexto de uma instituição financeira que possui diversos produtos, entre eles o produto cartão de crédito que é um dos principais produtos da instituição.

Com base no conceito difundido pelo americano KOTLER, PHILIP (1960) a respeito da teoria dos 5´Ps (que são um conjunto de fatores que pode ser aproveitado para realizar estratégias que a ajudem a obter resultados positivos nas vendas), pode-se compreender a importância em avaliar este conceito para garantir o crescimento de uma carteira de clientes assim como o aumento do faturamento em vendas[1].

Observando dois dos itens apresentado na teoria do 5'Ps (Pessoas e Produtos) despertou o interesse em conhecer o perfil comportamental do cliente com o produto cartão de crédito ao longo do período de relacionamento entre ambos, o que foi possível através do modelo do ciclo de vida, que é um modelo que explica os padrões de consumo dos indivíduos[2].

Baseado nesse conceito foi possível elaborar o modelo de ciclo de vida para o produto cartão de crédito, somada às informações cadastrais e dados transacionais pode-se definir estágios do cliente no uso do produto, desde a venda, o engajamento, a maturidade, o desengajamento e o cancelamento.

O objetivo deste trabalho é criar uma métrica de engajamento para o produto cartão de crédito de uma instituição financeira.

Através dessa métrica utilizando-se de um modelo preditivo de regressão logística mapear os clientes com baixa, média e alta probabilidade de desengajamento.

Podendo então, oferecer subsídios para tomadas de decisões baseadas na probabilidade do desengajamento de um cliente com o produto cartão de crédito, e de forma antecipada elaborar estratégias de *marketing* para incentivar o cliente ao uso do produto.

O resultado desse trabalho é apresentado em sete capítulos podendo constatar que é possível através da métrica criada e somada a variáveis categóricas entender se há mudança no comportamento de uso do cliente em relação ao produto e então criar ações preditivas e não apenas reativas para que clientes com propensão ao desengajamento possam ser incentivados a usar o produto.

No segundo capítulo desse trabalho será apresentado como foi definido o conceito para desenvolvimento da métrica LF (*Life Time*) com a finalidade de identificar o cliente engajado assim como o cliente desengajado

[1]Aluno do programa de Especialização em Data Science & Big Data, marcelobueno09@hotmail.com.
[2]Professor do Departamento de Estatística da Universidade Federal do Paraná, wbonat@ufpr.br

O terceiro capítulo abordará informações pertinentes sobre a criação e escolha das variáveis categorizadas utilizadas no modelo preditivo. No quarto capitulo será apresentado a definição do modelo preditivo. No quinto capitulo será apresentado os resultados obtidos com a aplicação do modelo preditivo. No sexto capitulo será apresentado considerações finais.

2 Métrica LF (*Life Time*)

Ao avaliar o mercado é possível verificar que se tornou comum o uso de métricas para avaliar o desempenho da empresa assim como o relacionamento do cliente com seu produto.

Métricas são essenciais para avaliar o desempenho de uma empresa e verificar a eficiência de suas estratégias.

Nem sempre o que foi planejado originalmente tem efeito positivo no mercado e, por isso, é tão importante acompanhar e mensurar os resultados[3].

Seguindo essa linha de raciocínio surgiu a necessidade da criação de uma métrica capaz de mapear o estágio do cliente no uso do produto cartão de credito.

Para criação da métrica LF (*Life Time*) foi necessário entender o comportamento do cliente engajado assim como o comportamento do cliente desengajado.

Então foi proposto exercícios para definição do conceito do público alvo de engajamento baseando-se na análise da variação do comportamento do faturamento do cliente em relação ao mês anterior, analisando o intervalo de 3 (três) meses.

Somado ao comportamento do cliente, foi considerado a data do início de relacionamento do cliente, a data de ativação do produto, a data do primeiro uso do produto, data de cancelamento e a variação do status de bloqueio do cartão no período analisado.

Para entender o exercício proposto imagine um cliente que para o primeiro mês da analise teve seu faturamento maior do que 15% (quinze por cento) em relação ao mês anterior, e no segundo mês da análise teve seu faturamento menor do que 15% (quinze por cento) em relação ao mês anterior e no terceiro mês da análise teve seu faturamento maior do 15% (quinze por cento) em relação ao mês anterior, esse cliente recebe a marcação (subiu-reduziu-subiu) e é considerado um cliente engajado.

Seguindo a linha do exercício acima foram sugeridas 220 (duzentos e vinte) combinações de gasto médio analisando o intervalo de 3 (três) períodos do faturamento do cliente.

Desses foram selecionadas 98 (noventa e oito) combinações de gasto médio para definir o estágio engajado de um cliente, a Figura 1 apresenta alguns exemplos desse exercício, os meses de analise serão chamados de MOB (*Month on Book* – frequência amostral em um intervalo de datas) e são representados na Figura 1 como Mob5 para o mês mais distante da data da análise e Mob3 para o mês próximo da data da análise.

Conceito Publico Alvo			
Mob5	Mob4	Mob3	Engajado?
↑	↓	↑	Sim
↓	↑	↓	Sim
↑	↑	↓	Sim
↓	↑	↑	Sim
↓	↓	↑	Sim
↓	↑	–	Sim
↑	↑	↑	Sim
↑	•	↑	Sim
–	↓	↑	Sim
–	↑	↓	Sim
Total de 98 combinações			

↑ Faturamento sobe >= 15% em relação ao mês anterior

↓ Faturamento desce >= 15% em relação ao mês anterior

– 15% em relação ao mês anterior

• Faturamento zerado no mês corrente

Figura 1: Conceito de combinações - Estágio engajado.

O motivo pelo qual utilizou-se a analise trimestral para definição do público alvo foi influenciada pelo prazo de duração de uma campanha de *marketing* dentro da instituição financeira.

Considerando que o estágio do cliente em Mob3 é engajado, podemos aplicar o modelo de desengajamento identificando assim a probabilidade do cliente desengajar em Mob0 e então de uma forma antecipada trabalhar com estratégias de *marketing* para incentivar o cliente ao uso do produto.

Uma vez que temos mapeado o estágio do cliente engajado, precisamos também mapear o estágio do cliente desengajado o qual é necessário para separar da base que será trabalhada no modelo de desengajamento, assim como aplicar como um método de análise de resultados para a base após ações de *marketing* de incentivo ao uso.

Seguindo o raciocínio para identificar o cliente engajado, das mesmas 220 (duzentos e vinte) combinações de gasto médio citadas anteriormente, foram selecionadas outras 55 (cinquenta e cinco) combinações para definir o estágio desengajado de um cliente conforme Figura 2.

Conceito Variavel Alvo			
Mob2	Mob1	Mob0	Engajado?
•	•	•	Não
•	↑	•	Não
↑	•	•	Não
↑	↓	•	Não
↓	↓	↓	Não
↓	•	•	Não
↓	–	↓	Não
↓	↓	•	Não
–	↓	•	Não
↓	–	•	Não
Total de 55 combinações			

↑ Faturamento sobe >= 15% em relação ao mês anterior

↓ Faturamento desce >= 15% em relação ao mês anterior

– 15% em relação ao mês anterior

• Faturamento zerado no mês corrente

Figura 2: Conceito de combinações - Estágio desengajado.

Do total de 220 (duzentos e vinte) combinações foi utilizado 98 para identificar o estágio engajado, 55 para

identificar o estágio desengajado, totalizando assim 153 combinações utilizadas na métrica LF (*Life Time*).

O restante das combinações foi descartado por entendermos que não refletia a realidade de um cliente engajado ou desengajado, sendo que nas demais combinações podemos encontrar os estágios como: pós-venda e cancelamento.

Para testar a métrica criada foi necessário aplicar o conceito em uma base contendo clientes aptos (clientes sem bloqueio de cancelamento) no período de análise.

Na base de dados onde a métrica LF (*Life Time*) foi aplicada pode-se observar que 33% (trinta e três por cento) dos clientes estavam engajados, e esses clientes foram observados em um intervalo de 60 (sessenta) dias e então constatado que 16% dos clientes engajados mudaram de estagio para desengajados após o período observado conforme Figura 3.

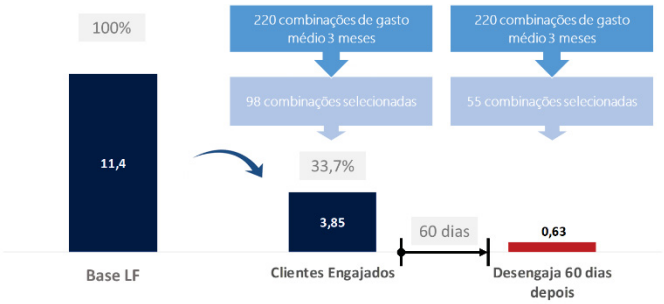

Figura 3. Aplicação da Métrica LF (*Life Time*)

Ao analisar o resultado apresentado após intervalo de 60 dias, pode-se verificar uma grande oportunidade de manter o faturamento, assim como impedir que esse venha diminuir devido ao desengajamento dos clientes com o produto e consequentemente com a marca.

A satisfação do cliente é o fator principal para uma empresa, e mensurar continuamente o engajamento do cliente possibilita o desenvolvimento de melhorias no produto assim como ações corretivas em relação ao relacionamento do cliente com o produto.

Segundo Marini (2018), é importantíssimo definir indicadores de engajamento e principalmente monitorar esses dados constantemente, até mesmo para saber se o cliente segue engajado ou está prestes a mudar o estágio no ciclo de vida, assim sendo, quanto mais rápido identificar este estágio, mais rápido é possível entrar com ações preventivas.

Métricas são indicadores que devem ser analisados constantemente e, portanto, avaliar a relação entre cliente e produto/serviço dá mostras do nível de engajamento do cliente.

No caso da Métrica LF (*Life Time*) o engajamento do cliente se dá através da utilização do cartão que pode ser medido através do gasto médio somado a dados cadastrais.

Uma vez que existe uma métrica para mapear o estágio engajado e desengajado do cliente com o produto, é necessário responder algumas perguntas: O que leva um cliente a mudar o seu comportamento em relação ao uso do produto contratado ou seja a desengajar? Quais fatores internos e externos podem influenciar nesse fenômeno? Que tipo de informações são necessárias para responder essas perguntas?

No próximo capítulo será abordado qual as informações utilizadas para definir as variáveis, assim como o modelo que melhor explicaria o fenômeno de desengajamento dos clientes.

3 Criação e escolha das variáveis

Nessa seção o foco principal é descrever como foi feito a seleção e construção de potenciais co-variáveis para que então na próxima seção possamos fazer o ajuste do modelo preditivo.

Iniciamos a seleção avaliando 3.000 (três mil) variáveis, as quais passaram por um processo de observação do dado e sua relevância ao evento, para então separarmos 264 (duzentos e sessenta e quatro). Na sequência realizamos ajuste de um modelo de regressão logística e ranqueamos conforme a importância de cada uma delas, para então separar 28 (vinte e oito) variáveis.

Essas vinte e oito passaram por um processo de categorização para então chegar ao número final de 11 (onze) variáveis selecionadas.

Respeitando as normas de sigilo da informação da instituição na qual o trabalho foi realizado, as variáveis utilizadas no artigo receberam o nome de VAR1, VAR2, VAR3 e assim sucessivamente.

Toda instituição financeira é portadora de uma elevada quantidade de dados oriundos de seus sistemas (autorizador de crédito, sistema cadastrais, sistemas de relacionamento, e APP´s, entre muitos outros) e também informações que são compradas do mercado (como BIROS de crédito, empresas reguladoras como o BACEN).

Para dar início ao processo de criação e seleção das variáveis foi enriquecido a base de dados com todo esse pool de informações apresentado na Figura 4, o que totalizou cerca de 3.000 (três mil variáveis).

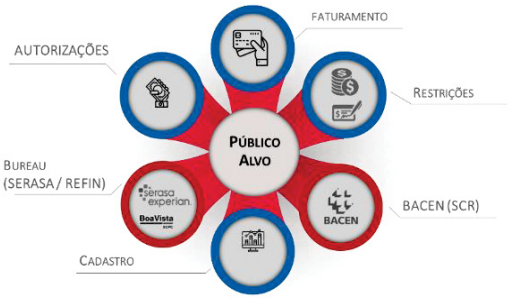

Figura 4. Enriquecimento com variáveis explicativas

Toda essa quantidade de informações torna a análise de dados algo complexo por se tratar de milhares de variáveis.

A seleção de variáveis é um componente muito importante no fluxo de trabalho de um cientista de dados.

Quando apresentados dados com altíssima dimensionalidade (quantidade excessiva de colunas), os modelos geralmente apresentam lentidão porque o tempo de treinamento aumenta exponencialmente.

Os modelos têm um risco crescente de se mostrar ineficaz quando a necessidade de prever novos resultados com o aumento do número de colunas, desta forma, perde-se eficácia ao tentar aplicar o modelo aos novos dados[4].

Geralmente utiliza-se a seleção de 10 a 20 variáveis para descrever o *Target* (variável alvo), assim sendo o processo de seleção de variáveis foi dividido em duas etapas.

Na primeira etapa foi aplicado duas metodologias na base de dados com todos os clientes que no MOB0 estavam com estagio de desengajamento.

- A primeira metodologia é a análise de correlação que é uma forma descritiva, que mede se há e qual o grau de dependência entre variáveis, ou seja, o quanto uma variável interfere em outra [5].

  Somado a análise de correlação também foi aplicado analise descritiva, que segundo Reis, E.A., Reis I.A. (2002), é a fase inicial do processo de estudo dos dados coletados.

  Observando então a análise descritiva, com uso de gráficos, tabelas e dados organizados, foram excluídos da base, os dados indevidos, dados inconsistentes com a realidade da variável, os dados com volume *missing* (dados ausentes) acima do esperado.
- A segunda metodologia aplicada para eliminação de variáveis, foi levar em consideração a relação entre as variáveis e a variável alvo, ou seja, considerando a base de estudo quanto da variável analisada é presente na base contendo a variável alvo.

  O conceito de *Mutual Information* diz que a "Se X e Y são independentes, então nenhuma informação sobre Y pode ser obtida conhecendo X ou vice-versa. Se X é uma função determinística de Y, então podemos determinar X de Y e Y de X[4].

Após a aplicação das duas metodologias citadas, das 3.000 (três mil) variáveis iniciais foi selecionado 264 variáveis.

E então com as variáveis selecionadas entra-se na segunda etapa de seleção das variáveis.

Será usada a técnica *ODDS Ratio* (medida de associação entre uma exposição e um resultado)

A *ODDS Ratio* (OR) é uma estatística que quantifica a força da associação entre dois eventos, A e B[6].

É uma forma de se calcular a relevância que determinada variável tem dentro do cenário a ser estudado, então considerando o público total de clientes e observando determinada variável encontraremos os registros bons (quando a correlação acontece) e os registros maus (quando a correlação não acontece).

Na Tabela 1 podemos ver o resultado do cálculo realizado para a *ODDS* onde temos o total de maus sobre bons. *Odss Ratio* = Maus /Bons.

Tabela 1 – Calculo e ranqueamento de *ODDS* para seleção de variáveis.

Descrição	Classe	Maus	Bons	N_Total	ODDS
VAR1	001	513.669	256.835	770.504	2,00
VAR2	002	181.335	99.734	281.069	1,82
VAR3	003	181.947	109.168	291.115	1,67
VAR4	004	184.511	119.932	304.443	1,54
VAR5	005	224.214	156.950	381.164	1,43
VAR6	006	211.593	158.695	370.288	1,33
VAR7	007	274.939	219.951	494.890	1,25
VAR8	008	234.035	198.930	432.965	1,18
VAR9	009	221.449	199.304	420.753	1,11
VAR10	010	214.396	203.676	418.072	1,05
VAR11	011	206.172	206.172	412.344	1,00
VAR12	012	188.519	197.945	386.464	0,95
VAR13	013	188.388	207.227	395.615	0,91
VAR14	014	172.731	198.641	371.372	0,87
VAR15	015	188.615	226.338	414.953	0,83
VAR16	016	181.066	226.333	407.399	0,80
VAR17	017	191.617	249.102	440.719	0,77
VAR18	018	201.962	272.649	474.611	0,74
VAR19	019	183.202	256.483	439.685	0,71
VAR20	020	169.394	245.621	415.015	0,69
VAR21	021	169.960	254.940	424.900	0,67
VAR22	022	161.056	802.199	963.255	0,20
VAR23	023	147.402	778.430	925.832	0,19
VAR24	024	146.984	818.040	965.024	0,18
VAR25	025	132.559	775.690	908.249	0,17
VAR26	026	98.057	859.916	957.973	0,11
VAR27	027	90.501	851.541	942.042	0,11
VAR28	028	88.030	850.322	938.352	0,10

Analisando as 264 variáveis selecionadas na etapa anterior e após o cálculo da *ODDS* para cada uma das variáveis, foi ranqueado as variáveis e então separado as que continha o maior índice de relevância com o estudo, ou seja, aquelas que são mais representativas dentro do universo de desengajamento.

Após o ranqueamento das variáveis, chegou-se ao número de 28 variáveis relevantes para estudar o comportamento de um cliente que desengaja do seu relacionamento com o produto cartão de crédito.

A etapa seguinte, é o momento da categorização das variáveis.

Shimakura, Silvia (2012) diz que variáveis categóricas são as características que não possuem valores quantitativos, mas são definidas por várias categorias, ou seja, representa uma classificação do indivíduo.

Nessa etapa foi aplicado a *ODDS Ratio* para as categorias existentes dentro de uma variável.

Para definir as categorias existentes dentro de uma mesma variável, foram observados o máximo e o mínimo de abrangência de conteúdo da variável para definição

de intervalo percentual entre as categorias da variável.

Na Tabela 2 podemos verificar um exemplo de distribuição das categorias na variável VAR1 e o cálculo da *ODDS* para cada categoria antes do processo de categorização da variável.

Tabela 2 – Calculo de *ODDS* e definição das categorias para a variável VAR1.

Descrição	Categoria	Mínimo	Maximo	Maus	Bons	Total	ODDS
VAR1	002	0	22	412.759	764.330	1.177.089	0,54
VAR1	003	23	24	331.674	794.784	1.126.458	0,42
VAR1	004	25	27	349.949	995.261	1.345.210	0,35
VAR1	005	28	29	242.739	819.977	1.062.716	0,30
VAR1	006	30	31	248.560	935.683	1.184.243	0,27
VAR1	007	32	33	212.187	938.740	1.150.927	0,23
VAR1	008	34	35	226.247	1.102.038	1.328.285	0,21
VAR1	009	36	36	148.777	830.555	979.332	0,18
VAR1	010	37	38	194.116	1.117.396	1.311.512	0,17
VAR1	011	39	40	143.312	879.831	1.023.143	0,16
VAR1	012	41	43	184.897	1.198.443	1.383.340	0,15
VAR1	013	44	45	132.376	873.751	1.006.127	0,15
VAR1	014	46	48	184.062	1.273.941	1.458.003	0,14
VAR1	015	49	50	104.034	744.903	848.937	0,14
VAR1	016	51	53	171.471	1.187.588	1.359.059	0,14
VAR1	017	54	57	165.445	1.184.260	1.349.705	0,14
VAR1	018	58	60	143.631	1.004.259	1.147.890	0,14
VAR1	019	61	65	135.961	916.506	1.052.467	0,15
VAR1	020	66	70	141.988	1.049.975	1.191.963	0,14
VAR1	021	71	120	95.860	1.049.595	1.145.455	0,09
VAR1	999			513.669	2.312.295	2.825.964	0,22

A Figura 5 apresenta a distribuição das classes baseada na *ODDS* antes da categorização para a variável VAR1 apresentada na Tabela 1.

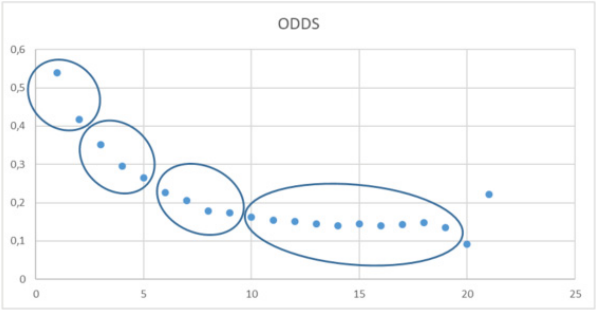

Figura 5 – Distribuição da variável VAR1 sem a categorização

Com o resultado da classificação anterior, tomou-se a decisão de realizar a categorização para a variável considerando a aproximação do índice da *ODDS* entre as categorias.

A ideia nesse exercício é desmembrar as categorias de uma variável ao máximo conforme a Tabela1 para então depois agrupa-las novamente da maneira que melhor descreve o efeito estudado conforme a Tabela 3.

Tabela 3 – Criação de novas classes para a variável VAR1.

Descrição	Categoria	Mínimo	Máximo	Maus	Bons	Total	ODDS
VAR1	1	0	24	744.433,00	1.559.114	2.303.547	0,48
VAR1	2	25	31	841.248	2.750.921	3.592.169	0,31
VAR1	3	32	38	781.327	3.988.729	4.770.056	0,20
VAR1	4	39	70	1.507.177	10.313.457	11.820.634	0,15
VAR1	5	71	120	95.860	1.049.595	1.145.455	0,09
VAR1	999	* classe descartada do estudo por conter dados nulos					

Podemos observar que para uma variável com mais de 20 categorias, após a aplicação da metodologia baseada na aproximação do índice da *ODDS*, temos apenas 5 categorias.

A Figura 6 apresenta a distribuição das classes após a categorização para a variável VAR1 apresentada na Tabela 2, onde foi realizado uma nova categorização considerando a aproximação dos índices *ODDS* entre as classes. A Classe 999 foi descartada do estudo por conter dados nulos.

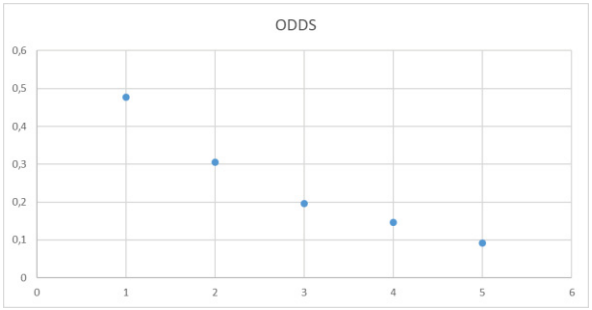

Figura 6 – Distribuição da variável VAR1 após a Categorização

Nesse momento foi levado em consideração alguns pontos importantes para a escolha das variáveis, exemplo: frequência da atualização da variável, nível de dificuldade de acesso a informação, nível de integridade do dado, e informações mais comuns para o mundo de cartão de credito.

Então após a categorização e as considerações citadas anteriormente, das 28 variáveis selecionadas na etapa anterior chegou-se ao número de 11 variáveis, que melhor traduzem e conseguem fazer a leitura do comportamento do cliente, possibilitando fazer uma previsão futura do desengajamento.

Na Tabela 4 podemos verificar os tipos de dados e categoria das variáveis selecionadas que serão utilizadas na modelagem.

Para os tipos de dados positivo considera que quanto maior é o índice da variável, maior é a chance de desengajamento, temos aqui relação diretamente proporcional.

Para o tipo de dados negativo quanto menor o índice da variável maior o risco de desengajamento.

Exemplo de tipo negativo: Variável que considera o perfil de uso do cliente, nesse caso quando o faturamento do cliente é menor, maior é a chance de desengajamento, e quando maior for o faturamento do cliente, menor a chance de desengajamento.

Tabela 4 – Tipo e categorias das 11 (onze) variáveis selecionadas para modelagem.

Variáveis	Tipo	Categoria
VAR01	Categorica	Metrica LF
VAR02	Positiva	Avaliação de Risco
VAR03	Negativa	Perfil de Uso Interno
VAR04	Negativa	Perfil de Uso Mercado
VAR05	Positiva	Avaliação de Risco
VAR06	Categorica	Cadastral
VAR07	Negativa	Perfil de Investidor
VAR08	Categorica	Cadastral
VAR09	Negativa	Cadastral
VAR10	Negativa	Perfil de Uso Interno
VAR11	Positiva	Perfil de Uso Mercado

4 Definição do modelo preditivo

Nessa seção iremos utilizar as 11 variáveis que foram construídas na seção anterior para construir um modelo preditivo para prever os clientes com propensão ao desengajamento.

Porque uma análise preditiva? Organizações do mundo todo estão realizando analises preditivas para resolver problemas difíceis e descobrir novas oportunidades.

Os modelos preditivos ajudam as empresas a atrair, reter e expandir seus clientes mais valiosos [6].

Os modelos preditivos usam resultados conhecidos para desenvolver (ou treinar) um modelo que pode ser usado para prever valores para dados diferentes ou novos.

A modelagem fornece resultados na forma de previsões que representam uma probabilidade da variável de destino (no caso desse artigo o desengajamento) com base na importância estimada de um conjunto de variáveis de entrada [6].

Existem vários modelos preditivos, o modelo escolhido foi o de regressão pelo motivo desse modelo observar informações históricas para projetar o futuro.

Foi usado do tipo logístico, porque parte das variáveis são categóricas, além de que se trata de um modelo comumente usado na instituição onde o estudo foi realizado.

Setores como CRM, Crédito já tem a cultura de uso desse tipo de modelo, então optamos por aplica-lo na modelagem desses dados.

O Modelo de Desengajamento foi desenvolvido utilizando a plataforma do *SAS Enterprise Guide* versão 8.0 que é uma das ferramentas do SAS[7], devido ser a plataforma escolhida pela instituição como ferramenta de análise de dados.

Todo o modelo de regressão logística gera o intercepto e também gera o coeficiente de regressão, conhecido como beta, para cada uma das categorias das variáveis.

A Tabela 5 mostra o valor do intercepto encontrado que foi igual a 131,71 e o β para cada uma das categorias das 11 variáveis selecionadas.

Tabela 5 – Distribuição do β para cada categoria das variáveis e valor do α.

Intercepto (α) = 131,71		
Variável	Categoria	Beta (**β**)
VAR1	A	48,17
	B	27,83
	C	- 9,79
	D	7,79
VAR2	A	8,46
VAR3	A	13,37
	B	27,07
	C	16,50
VAR4	A	37,03
	B	30,22
	C	55,06
VAR5	A	- 1,93
	B	10,69
	C	3,08
	D	9,35
	E	6,39
	F	0,45
VAR6	A	- 24,08
	B	7,56
	C	24,33
	D	19,22
	E	8,58
VAR7	A	- 32,83
	B	- 13,69
	C	5,51
	D	14,84
VAR8	A	- 90,24
	B	- 51,74
	C	- 4,75
	D	6,94
VAR9	A	18,02
	B	2,87
	C	- 32,06
VAR10	A	2,11
	B	2,68
	C	11,76
VAR11	A	3,67
	B	2,17

A formula utilizada para calcular a escoragem do modelo preditivo de desengajamento é:

$$\hat{N} = \alpha + \beta(VAR1^{(c)}) + \beta(VAR2^{(c)}) + \beta(VAR3^{(c)}) + \beta(VAR4^{(c)}) + \beta(VAR5^{(c)}) + \beta(VAR6^{(c)}) + \beta(VAR7^{(c)}) + \beta(VAR8^{(c)}) + \beta(VAR9^{(c)}) + \beta(VAR10^{(c)}) + \beta(VAR11^{(c)}),$$

onde (c) refere-se à categoria que o cliente se encaixou.

Uma vez encontrada o valor de escoragem através da formula acima, entendemos ser mais compreendido pelo usuário final, a transformação do valor de escoragem para porcentagem da probabilidade de desengajamento (PD).

Então para encontrar a PD foi verificada a mediana de desengajados por faixa escore, carimbando assim a porcentagem no gráfico.

A Figura 7 apresenta um exemplo do cálculo do modelo preditivo de desengajamento considerando um determinado cliente e as suas respectivas categorias dentro das variáveis analisadas, também visualizamos o valor do predito linear e a probabilidade de desengajamento.

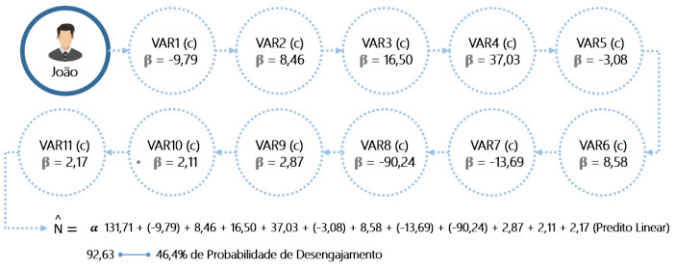

Figura 7 – Cálculo do modelo preditivo de desengajamento

Uma vez construído o modelo a partir da base de estudo, em seguida foi realizado testes na base de controle para avaliar a performance e identificar o KS (definido como o máximo entre as distâncias das curvas).

Essa métrica é utilizada para medir o quão bem o modelo separa as classes da variável resposta[10], de cada base onde o modelo foi testado.

Na sequencia o modelo foi pilotado usando a base *OUT OF TIME* com intuito de validar o modelo com dados novos e também mensurar o *KS* da base modelada.

Na Figura 8 vemos a performance obtida em cada um dos grupos usados para construir, testar e pilotar o modelo.

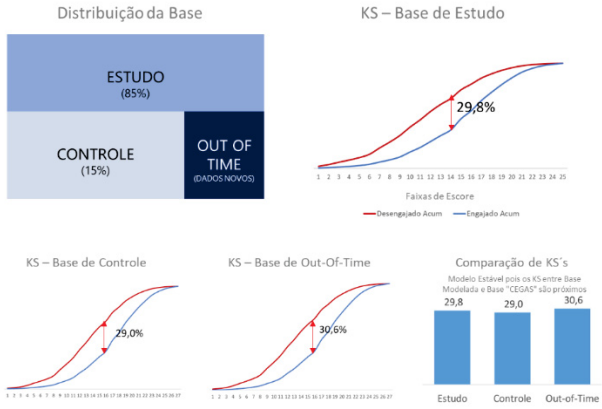

Figura 8 – Distribuição da base e resultado do *KS* para construção, teste e piloto do modelo

Conforme avaliação do *KS* podemos verificar que o modelo é estável pois o índice entre a base modelada e a base de piloto são muito próximos.

Após o piloto do modelo foi realizado distribuição dos 3.8MM clientes em 27 faixas observando a PD, pois acredita-se que o *capacity* disponível para tratamento da base como um todo não atenderia a demanda, portanto as 27 faixas foram segmentadas em 5 grandes faixas de abordagem, sendo altíssima propensão, alta propensão, media propensão, baixa propensão e baixíssima propensão ao desengajamento, conforme apresentado na Figura 9.

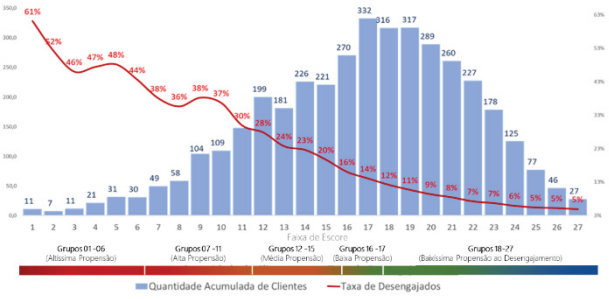

Figura 9 – Distribuição das faixas conforme propensão ao desengajamento

A Figura 9 nos mostra que podemos dividir as ações considerando o nível de propensão ao desengajamento associado ao número de clientes alocadas na faixa da PD.

O modelo de desengajamento tem a periodicidade de processamento mensal devido a entrega dos dados serem mensais, e tem sido treinado com a periodicidade semestral, devido a quantidade de clientes entrantes e também os clientes que saem da base nesse período.

De toda a base de estudo, decidimos por marcar os registros com o escore e a probabilidade do desengajamento classificado por faixas de propensão e entregar a base como um todo para as áreas de campanha, relacionamento e produtos.

5 Resultados

Após construído o modelo a partir da base de estudo, realizado testes na base de controle e pilotado o modelo usando a base *OUT OF TIME* pode-se então avaliar os resultados do modelo preditivo de desengajamento quando aplicado em uma base contendo cliente aptos em uma determinada data prevendo então desengajamento 90 dias após.

No exemplo da Tabela 6 consideramos um grupo de clientes que foram selecionados pela métrica LF (*Life Time*) como sendo clientes engajados e então aplicado o modelo para prever o estágio do cliente no mês de fevereiro subsequente ao analisado.

Tabela 6 – Faixa de desengajamento e clientes por faixa – Previsão – *OUT OF DATE* - FEV

Faixa de Desengajamento dos Clientes	Faixa de Probabilidade de Desengajamento	QTD de Clientes em cada faixa - Mês NOV	Qtd Clientes na PD Mínima	Qtd Clientes na PD Máxima
Altíssima	39-61%	31.956	12.463	19.493
Alta	29-38%	189.785	55.038	72.118
Média	20-28%	502.471	100.494	140.692
Baixa	14-19%	505.284	70.740	96.004
Baixíssima	5-13%	1.366.478	68.324	177.642
Total		2.595.974	307.059	505.949

Para validar a eficiência do modelo, aplica-se novamente a métrica LF (*Life Time*) na mesma base modelada e avalia a mudança de comportamento do cliente 60 dias após.

Podemos então verificar na Tabela 7 que o índice de desengajamento ficou dentro do previsto para as faixas altíssima, alta, média e baixa propensão.

Para a faixa baixíssima propensão o índice de desengajamento foi maior do que previsto.

Tabela 7 – Faixa de desengajamento e clientes por faixa - Efetuado

Faixa de Desengajamento dos Clientes	Faixa de Probabilidade de Desengajamento	QTD de Clientes em cada faixa - Mês FEV	Qtd Clientes desengajados FEV	PD Realizada em FEV
Altíssima	39-61%	31.956	15.208	48%
Alta	29-38%	189.785	68.125	36%
Média	20-28%	502.471	132.703	26%
Baixa	14-19%	505.284	82.530	16%
Baixíssima	5-13%	1.366.478	202.450	15%
Total		2.595.974	501.016	19%

6 Considerações Finais

O objetivo deste trabalho foi criar e aplicar a métrica LF (*Life Time*) com intuito de mapear o estágio do cliente no uso do produto cartão de credito em uma instituição financeira foi alcançado, inclusive apresentando resultados satisfatórios.

Entende-se que como todo modelo é necessário treina-lo, com ajustes das variáveis e inclusive das faixas da probabilidade de desengajamento.

Podemos ver que as faixas de probabilidade dão a opção de separar bases com números menores de clientes, porém com maior assertividade na probabilidade.

Dependendo do *capacity* da empresa e das áreas de campanhas, de relacionamento e produtos, essa pode ser uma ótima estratégia de atuação nas campanhas de marketing.

Uma vez mapeado o cliente com propensão de desengajamento cabe as instituições trabalhar de forma antecipada e elaborar estratégias de *marketing* para incentivar o cliente ao uso do produto, evitando assim aumento de cancelamento do produto e consequentemente garantindo o engajamento do cliente com a marca.

Agradecimentos

Este trabalho só pode ser concluído com a ajuda de algumas pessoas que não posso deixar de mencionar.

Quero agradecer ao Professor Orientador Dr. Wagner Hugo Bonat por dar o suporte necessário para que as ideias fossem organizadas e apresentadas nesse documento, agradeço por todo o conhecimento em relação a Analitics que foi compartilhado durante todo o curso.

Também agradeço a todo corpo docente pelo empenho em nos capacitar para que o Data Science se torne um diferencial em nossas carreiras profissionais.

Quero agradecer aos meus colegas de trabalho, Daniel Jesus Soares, e Daniel Alves Gato, por compartilhar o conhecimento e a paciência direcionada com o meu aprendizado, e que de forma solicita sempre estiveram disponíveis para responder inúmeras duvidas que surgiram durante todo o período de aprendizado.

Agradeço ao Marcos Vinicius Alvarenga Ramos da Silva, por acreditar e investir em mim, obrigado pela oportunidade e por todo conhecimento que foi compartilhado comigo durante todo esse processo.

Agradeço a minha esposa por todo empenho, paciência e dedicação prestada em todo esse período acadêmico.

E mais importante agradeço a Deus por tudo que me é possível viver.

Referências

[1] 5 Ps do Marketing: trabalhe as pessoas em sua estratégia – Disponível em: https://www.agencianovofoco.com.br/5-ps-do-marketing-estrategia/

[2] Teoria do Ciclo de Vida – Disponível em: https://pt.wikipedia.org/wiki/Teoria_do_ciclo_de_vida#cite_note-1

[3] Métricas de Customer Sucess – Disponível em: https://www.iugu.com/blog/metricas-de-customer-success-para-voce-utilizar

[4] Métodos para selecionar as melhores variáveis do dataset em Python – Disponível em: https://medium.com/@alegeorgelustosa/m%C3%A9todos-para-selecionar-as-melhores-vari%C3%A1veis-do-dataset-em-python-2c374b2e9df2

[5] Análise de correlação usando Python e R – Disponível em: https://medium.com/omixdata/estat%C3%ADstica-an%C3%A1lise-de-correla%C3%A7%C3%A3o-usando-python-e-r-d68611511b5a

[6] Szumilas, Madalena (agosto de 2010). "Explicando o odds ratio" ISSN 1719-8429 . PMC 2938757 . PMID 20842279 . - https://en.wikipedia.org/wiki/Odds_ratio#cite_note-1

[7] SAS ENTERPRISE GUIDE – Disponível em: https://support.sas.com/en/software/enterprise-guide-support.html

[8] Definição INTERCEPT – Disponível em: https://blog.minitab.com/pt/analise-de-regressao-como-interpretar-a-constante-intercepto-y#:~:text=O%20termo%20constante%20na%20an%C3%A1lise,rela%C3%A7%C3%A3o%20%C3%A0%20interpreta%C3%A7%C3%A3o%20da%20constante.

[9] A lógica da regressão logística – Disponível em: https://www.scielo.br/j/rsocp/a/RWjPthhKDYbFQYydbDr3MgH/?lang=pt#:~:text=O%20coeficiente%20de%20regress%C3%A3o%2C%20(%CE%B2,Y)%20e%20diferentes%20vari%C3%A1veis%20independentes.

[10] Kolmogorov-Smirnov - Disponível em: ttps://pt.wikipedia.org/wiki/Teste_Kolmogorov-Smirnov#:~:text=A%20estat%C3%ADstica%20de%20Kolmogorov%E2%80%93Smirnov,distribui%C3%A7%C3%A3o%20emp%C3%ADrica%20de%20duas%20amostras.

[11] Analise Preditiva - Disponível em: https://www.sas.com/pt_br/insights/analytics/predictive-analytics.html

[12] MARINI, Carolina. Os 3 pilares do Customer Success: engajamento; churn; métricas. Disponível em: http://useronboarding.com.br/pilares-do-customersuccess/. Acesso em 16/11/2018.

[13] Shimakura, Silvia (2012) - Tipos de variáveis https://leg.ufpr.br/~silvia/CE055/node8.html

[14] Reis, E.A., Reis I.A. (2002) Análise Descritiva de Dados. Relatório Técnico do Departamento de Estatística da UFMG. Disponível em: www.est.ufmg.br