## Caso

No contexto recente, a diminuição das taxas de juros no mercado desencadeou um aumento significativo na demanda por solicitações de crédito. Os clientes veem uma oportunidade favorável para financiar compras importantes ou consolidar dívidas existentes, o que levou a uma afluência de solicitações de empréstimo no banco "Super Caja". A equipe de análise de crédito do banco está enfrentando uma carga de trabalho avassaladora devido à análise manual necessária para cada solicitação de empréstimo de clientes individuais. Essa metodologia manual resultou em um processo ineficiente e demorado, afetando negativamente a eficácia e a rapidez com que as solicitações de empréstimo são processadas. A situação se torna mais crítica devido à crescente preocupação com a taxa de inadimplência, um problema que está afetando cada vez mais a indústria financeira, aumentando a pressão sobre os bancos para identificar e mitigar os riscos associados ao crédito.

Com o objetivo de enfrentar esse desafio, a proposta é a automação do processo de análise utilizando técnicas avançadas de análise de dados, visando melhorar a eficiência, precisão e rapidez na avaliação das solicitações de crédito. Além disso, o banco já possui uma métrica para identificar clientes com pagamentos em atraso, o que poderia ser uma ferramenta valiosa para integrar na classificação de risco dentro do novo sistema automatizado.

O objetivo da análise é desenvolver um score de crédito a partir de uma análise de dados e avaliação do risco relativo que possa classificar os solicitantes em diferentes categorias de risco com base em sua probabilidade de inadimplência. Essa classificação permitirá ao banco tomar decisões informadas sobre quem conceder crédito, reduzindo assim o risco de empréstimos não reembolsáveis. Além disso, a integração da métrica existente de pagamentos em atraso fortalecerá a capacidade do modelo de identificar riscos, contribuindo assim para a solidez financeira e eficiência operacional do banco.

Alcance do projeto

Este projeto está dividido em 3 alcances, cada um marcado com um marco que determina o que você aprenderá e o que deve alcançar para considerá-lo concluído.

O marco 1 é sempre considerado indispensável para avaliar o projeto. Os demais são opcionais e destinam-se a permitir que você aprofunde alguma habilidade ou adiante um pouco do que será abordado em projetos futuros. Portanto, você pode completar apenas o marco 1 ou os marcos 1 e 2 ou os marcos 1 e 3 ou os marcos 1, 2 e 3.

ALCANCE DO PROJETO: Marcos 1/2/3

Processar e preparar a base de dados

Conectar/importar Dados A Ferramentas

Identificar E Gerenciar Valores Nulos

Identificar E Gerenciar Valores Duplicados

Identificar E Manejar Dados Fora Do Alcance Da Análise

Identificar E Gerir Dados Discrepantes Em Variáveis Categóricas

Identificar E Gerenciar Dados Discrepantes Em Variáveis Numéricas

Verificar E Alterar Tipo De Dado

Criar Novas Variáveis

Unir Tabelas

Construir Tabelas Auxiliares

Fazer uma análise exploratória

Agrupar Dados Segundo Variáveis Categóricas

Visualizar As Variáveis Categóricas

Aplicar Medidas De Tendência Central

Visualizar Distribuição

Aplicar Medidas De Dispersão

Calcular Quartis, Decis Ou Percentis

Calcular Correlação Entre Variáveis

Aplicar técnica de análise

Calcular Risco Relativo

Validar Hipóteses

Aplicar Segmentação

Regressão Logística

Resumir informações em um dashboard ou relatório

Representar Dados Através De Tabela Resumo Ou Scorecards

Representar Dados Através De Gráficos Simples

Representar Dados Através De Gráficos Ou Visuais Avançados

Aplicar Opções De Filtros Para Manejo E Interação

Apresentar resultados

Selecionar Gráficos E Informações Relevantes

Criar Uma Apresentação

Apresentar Resultados Com Conclusões E Recomendações

## **2. Mãos à obra! - Marco 1**

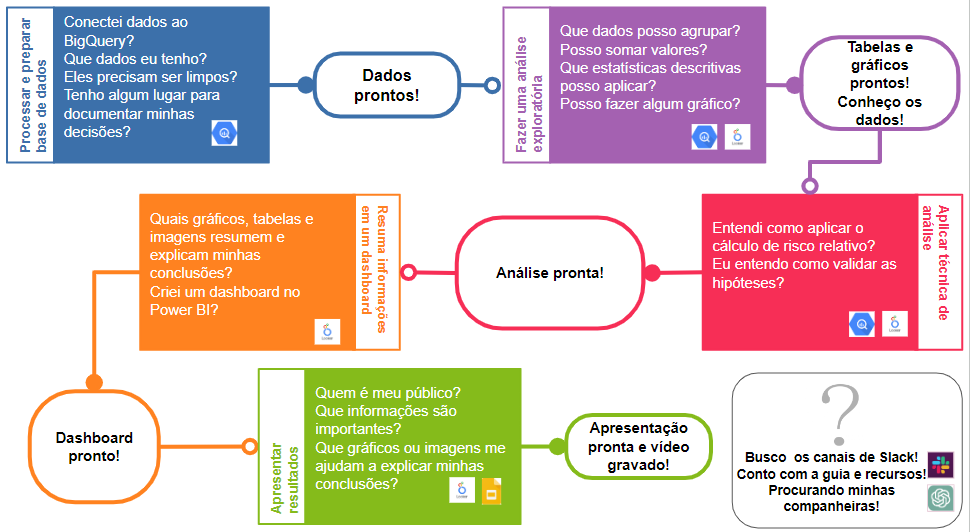
Em geral, cada marco deste projeto irá levá-la através de várias etapas da análise de dados, desde a compreensão do problema, através do processamento e limpeza de dados, até chegar à análise propriamente dita e à apresentação dos resultados, salvo algumas exceções. Neste primeiro marco, nos concentraremos na avaliação do risco relativo associado a cada uma das variáveis ​​fornecidas pela base de dados, que serão essenciais para posteriormente desenvolver uma pontuação/score de crédito robusta.

Você estará imersa no mundo financeiro, explorando os dados do banco com o objetivo de segmentar e avaliar a situação de crédito dos seus clientes. O contexto fornecido anteriormente já nos dá uma compreensão sólida do problema que enfrentamos: a necessidade de automatizar e otimizar o processo de análise de crédito para gerenciar e reduzir efetivamente o risco de não pagamento. As principais questões que procuraremos responder incluem: quais variáveis ​​mais influenciam o risco de não conformidade/inadimplência? Como essas variáveis ​​se correlacionam entre si e com o comportamento de pagamento do cliente?

Mas não se limite a estas questões, durante sua exploração você descobrirá alguma informação ou padrão interessante que pode enriquecer a compreensão do risco de crédito, encorajamos que você busque e integre este insights a sua análise. Este projeto é tanto sobre resolução do problema apresentado quanto à exploração e descoberta de novas percepções que podem fortalecer sua abordagem para a gestão do risco de crédito.

Nesta fase inicial, o cálculo do risco relativo de cada variável é essencial para estabelecer uma base sólida sobre a qual construiremos a pontuação/score de crédito em marcos subsequentes. Sua análise neste marco estabelecerá as bases para uma avaliação de crédito mais precisa e automatizada que acabará contribuindo para a robustez financeira e operacional do banco.

No final deste marco, espera-se que você apresente suas descobertas e o cálculo do risco relativo associado a cada variável de forma clara e concisa, junto com quaisquer percepções adicionais que você descobriu. Este é um passo crucial na sua jornada rumo à criação de um sistema de análise de crédito automatizado e preciso.



O que é “crítico” ou o que há de novo?

Nesta seção, descreveremos algumas etapas ou processos que comumente podem gerar ansiedade/frustração/bloqueios ao abordar o problema de análise de risco de crédito. Contudo, é importante lembrar que essas reações não são a regra, e você pode passar por essas etapas sem nenhum problema. Este é apenas um pequeno conselho: se você chega a esses estágios e começa a se sentir bloqueada, isso é completamente natural!

Neste marco, existem três pontos que podem ser novos ou críticos para você:

* Tomada de decisão na etapa “Processar e preparar banco de dados”:  
  Você já viu isso durante todo o bootcamp e com certeza já internalizou que preparar dados é uma tarefa da qual você não pode escapar. O diferencial deste projeto é a necessidade de consultas um pouco mais complexas.
* Segmentação de risco na etapa “Aplicar técnica de análise”:  
  As questões que podem surgir aqui são: Como decidir quantos e que níveis de risco criar? Como determinar as regras para segmentar o risco com base nas variáveis ​​disponíveis?
* Como montar uma apresentação na etapa “Apresentar resultados”:  
  As perguntas que podem surgir aqui são: Que informações são realmente importantes? Como comunicar os riscos identificados de forma eficaz? Em que informação investir os minutos que tenho para apresentar?

Nestes passos o conselho é o mesmo: mantenha a calma, use todos os recursos disponíveis (canais Slack, sessões colaborativas, ferramentas IA, Google, etc.) para tomar decisões com confiança. É uma parte importante do seu aprendizado enfrentar e superar esses desafios. Se você não duvida nem erra, como você aprende? 😉 Além do mais, lembre-se de que cada obstáculo superado aproxima você um passo das habilidades de uma analista de dados.

### **🟦 2.1 Processar e preparar a base de dados**

⌛ Intervalo de tempo estimado: 8 a 12 horas

Como este é seu terceiro projeto e você já vem acumulando conhecimento sobre comandos e conceitos, incluiremos apenas recursos para tópicos apresentados pela primeira vez neste projeto. Use como recurso os conteúdos já disponíveis em outros projetos e também a documentação que você criou para projetos anteriores. E lembre-se, um novo projeto é sempre uma boa oportunidade para colocar em prática o que foi aprendido e procurar novas maneiras de atingir um objetivo.

| **Meta** | **Objetivo** | **Recursos** |
| --- | --- | --- |
| 🔵 Conectar/importar dados para ferramentas | Use a interface do Google BigQuery para criar as 4 tabelas (1 para cada arquivo) | - |
| 🔵 Identificar e tratar valores nulos | Identifique e trate valores nulos usando comandos SQL. | * 💡Observe que para a variável last*month*salary temos 7199 registros nulos, o que constitui aproximadamente 20% do nosso banco de dados de usuários. Portanto, talvez a melhor opção não seja simplesmente excluir esses registros. * Nosso objetivo nesta análise é encontrar o perfil dos clientes que pagam mal para gerar um mecanismo de regras de aprovação de crédito, portanto a variável last*month*salary é importante para nossa análise. Para tomar uma decisão de como tratar esses dados, podemos juntar os dados da tabela default utilizando a função LEFT JOIN, onde existe a variável default\_flag para entender se esses registros estão em maior proporção para clientes com flag 1 (mau pagador). |
| 🔵 Identificar e tratar valores duplicados | Identifique e trate dados duplicados usando comandos SQL. | - |
| 🔵 Identificar e gerenciar dados fora do escopo da análise | Use CORR para entender a correlação entre variáveis ​​ | * ℹ️ Identificar dados fora do escopo também nos faz considerar a seleção de variáveis. * 🚨 Altamente recomendado: * 📄 [Conceito de seleção de variáveis](https://docs.google.com/document/d/1q6UPnF3SMgHFcuAsy5DrRsiHGHBxb0qA1aut1Y9daEE/edit?usp=sharing) * 💡Podemos usar a função CORR para calcular a correlação entre duas variáveis. * Se calcularmos a correlação entre as variáveis ​​more\_ 90\_ days\_ overdue e number\_ times\_ delayed\_ payment\_ loan\_ 30\_ 59\_ days da tabela loan\_detail encontramos um valor de 0,98, indicando uma alta correlação entre essas duas variáveis, ou seja, quanto maior a correlação (quanto mais próximo de + 1 ou -1) mais correlacionadas são as variáveis. Consulte o documento de conceito para entender melhor este conceito e como selecionar a variável por meio de correlação. * 💡 Agora que você conhece a fórmula de correlação e como interpretá-la, verifique se não existem outras variáveis ​​com alta correlação. * 🧑‍💻 Se este conceito ainda não estiver claro, peça exemplos ao ChatGPT. |
| 🔵 Identificar e tratar dados inconsistentes em variáveis ​​categóricas | Identifique e resolva valores inconsistentes usando comandos SQL, como [UPPER](https://cloud.google.com/bigquery/docs/reference/standard-sql/string_functions#upper), [LOWER](https://nuvem.google.com/bigquery/docs/reference/standard-sql/string_functions#lower). | 💡Você também pode usar [CASE WHEN](https://cloud.google.com/bigquery/docs/reference/standard-sql/conditional_expressions#case_expr) ou IF para modificar o conteúdo de uma variável, por exemplo, substitua a palavra others por other e deixe os dados padronizados. |
| 🔵 Identificar e tratar dados discrepantes em variáveis ​​numéricas | Identificar e processar OUTLIERS | * 📄 [O que são outliers e como tratá-los](https://docs.google.com/document/d/1C5J2oyuATZLV-FnXUoE5DluNNnJaxRgd6ES2ORheRvg/edit?usp=sharing) * 💡 Uma das formas de identificar valores atípicos (outliers) são analisando os quartis e a distribuição dos dados. No Looker Studio temos duas opções de gráficos que podem ajudar nessa tarefa, o Histograma e o Boxplot. * 📹 [Vídeo oque são e como tratar valores extremos (outliers)](https://www.youtube.com/watch?v=_YWAb-RHL8g) * Você também pode se desafiar e buscar uma alternativa para identificar outliers no Bigquery ou avançar para a análise exploratória e realizar a análise através do Looker Studio. |
| 🔵 Verificar e alterar o tipo de dados | Em algum momento do projeto use CAST, SAFE\_CAST ou outro comando SQL para alterar o formato de uma variável. | - |
| 🔵Criar novas variáveis ​​ | Crie novas variáveis ​​e salve-as em uma nova tabela ou visualização. | * 📄 [Conceito feature engineering](https://docs.google.com/document/d/1AE0WZ4fWPtsSoWG92Wqtt8QQj4Yady94oqb6_XckqEw/edit?usp=sharing) * 💡Se revisarmos a base de dados de loans*outstanding, observamos que o user*id se repete diversas vezes, uma vez para cada empréstimo que o cliente faz, essa informação desta forma não nos ajuda muito, seria melhor se fosse agrupado por usuário, com uma linha para cada cliente, mostrando o tipo de empréstimo e o valor total. Para isso, você pode verificar quais tipos de empréstimos temos na base de dados, utilizando o comando DISTINCT. * 💡 Você criou novas variáveis, então pode ser interessante usar a função CORR para testar se também existe correlação entre as variáveis selecionadas e as novas variáveis. |
| 🔵 Unir tabelas | Entenda o comando JOIN e suas variedades para unir diferentes tabelas. | - |
| 🔵 Construir tabelas auxiliares | Use o comando WITH ou Subqueries para construir tabelas temporárias. | 💡Se olharmos o resultado final da tabela criada, podemos encontrar alguns valores nulos para real*estate*loan e total\_loan. Portanto, surge a questão de por que esses nulos sugiram se já revisamos todas as variáveis ​​que continham nulos 🤔 e também surge a questão se precisamos tratar este dados de alguma forma. |

🤸Só por diversão:

* Você poderia juntar as tabelas usando diferentes tipos de JOIN e identificar a opção que melhor se adapta a este projeto?

### **🟪 2.2 Fazer uma análise exploratória**

⌛ Intervalo de tempo estimado: 8 a 12 horas

A Habilidade de realizar Análise Exploratória de Dados (EDA)é uma etapa crucial neste projeto. Como você já sabe, o EDA nos permite adquirir uma compreensão inicial e geral dos dados com os quais estamos trabalhando, antes de aplicar técnicas ou métricas específicas para calcular o risco relativo e, eventualmente, a pontuação de crédito. Esta prática, comum em análise de dados, ciência de dados e estatísticas, é uma ferramenta inestimável para descobrir padrões, tendências, relacionamentos e anomalias nos dados de crédito de forma exploratória e informal.

A análise exploratória em nosso contexto é realizada principalmente de forma visual e descritiva, sem chegar a conclusões definitivas. Algumas das técnicas e ferramentas usadas na análise exploratória de dados incluem:

Visualização de dados por meio de gráficos: essa técnica nos ajuda a visualizar a distribuição e as relações entre variáveis ​​como renda,idade, dívidas e comportamento de pagamento dos clientes, fornecendo percepções preliminares sobre como essas variáveis ​​podem influenciar o risco de crédito.

Resumo estatístico através de cálculos de medidas descritivas: calcular a média, mediana, desvio padrão, percentis, etc., para entender a tendência central, dispersão e distribuição dos dados.

Análise de tendências e padrões: observar tendências temporais, padrões sazonais ou outros padrões interessantes nos dados, como, por exemplo, comportamento de pagamento em diferentes grupos de renda ou o impacto do número de dependentes na capacidade de pagamento.

Segmentação de dados: dividir os dados em grupos ou segmentos, por exemplo, diferentes faixas de renda ou categorias de emprego, para analisá-los separadamente e encontrar diferenças significativas no risco crédito.

A análise exploratória é utilizada como uma fase iterativa (cíclica,que se repete) e criativa onde os analistas podem obter percepções ou hipóteses iniciais relacionadas ao risco de crédito (ou objetivo da análise), que posteriormente, passarão por testes e validações mais rigorosas. Este estágio inicial é essencial para entender os dados e tomar decisões sobre como proceder com análise e modelagem de crédito. Além disso, o EDA também é útil na detecção de problemas nos dados, como erros de entrada ou inconsistências, que não foram identificadas na fase de limpeza, permitindo a limpeza e preparação dos dados adequadamente antes de realizar uma análise mais aprofundada nos marcos subsequentes do projeto.

O uso de ferramentas de visualização na etapa de análise exploratória nos ajuda a entender os dados com mais facilidade e de forma visual. Neste projeto utilizaremos a ferramenta Looker Studio para realizar esta tarefa.

| **Meta** | **Objetivo** | **Recurso** |
| --- | --- | --- |
| 🟣 Agrupar dados de acordo com variáveis ​​categóricas | Use tabelas no Looker Studio para resumir dados em variáveis ​​categóricas | * [Link para acessar o Looker Studio](https://lookerstudio.google.com/navigation/reporting) * 📹 [Vídeo como começar no Looker Studio](https://www.youtube.com/watch?v=Gx5S8Ry54TU&pp=ygUjY29tbyBjcmlhciB1bSBwYWluZWwgbm8gZGF0YSBzdHVkaW8%3D) * 📹 [Vídeo Como conectar dados no Looker Studio](https://www.youtube.com/watch?v=dAZ0yfRSKjQ&pp=ygUiY29tbyBjb25lY3RhciBkYWRvcyBubyBkYXRhIHN0dWRpbw%3D%3D) * 📹 [Visão geral Looker Studio](https://www.youtube.com/watch?v=dfAHZ1C_ZQs&list=PLbnAsJ6zlidug4pEvq6tm0h7DPJGkP-aU) * 📹 [Vídeo como criar um painel/dashboard no Looker Studio](https://www.youtube.com/watch?v=qoKPW5Yn3Dk&pp=ygUjY29tbyBjcmlhciB1bSBwYWluZWwgbm8gZGF0YSBzdHVkaW8%3D) * 📄 [Tutorial do Google Como criar um relatório no Looker Studio](https://support.google.com/looker-studio/answer/06292570?0hl=ES#zippy=%2Csections-of-this-art%C3%ADculo%2Cneste-artigo) * ⚠️O Data Studio mudou recentemente seu nome para Looker Studio, então você pode encontrar referências ao Data Studio em alguns conteúdos na internet. Não se preocupe, é o mesmo produto. |
| 🟣 Ver variáveis ​​categóricas | Use gráficos de barras no Looker Studio para visualizar variáveis ​​categóricas. | 📹 [Video como criar gráficos de barras](https://www.youtube.com/watch?v=p4nNJWtStAs) |
| 🟣 Aplicar medidas de tendência central (moda, média, mediana) | Use as opções da tabela para calcular estatísticas descritivas para ajudar a compreender a distribuição dos dados. | * 📹 [Tabelas no Looker Studio](https://www.youtube.com/watch?v=WcxlQG1UdkY) * 💡 As tabelas dinâmicas no Looker Studio são semelhantes às tabelas dinâmica do Planilhas Google que você já conhece. |
| 🟣 Ver distribuição | Use histograma ou boxplot no LookerStudio para exibir variáveis ​​numéricas. | * 📹 [vídeo de conceito de boxplot (Diagrama de caixas)](https://www.youtube.com/watch?v=ebkzrOMjGXU) * 📹 [Interpretar boxplot](https://www.youtube.com/watch?v=qU2lANG4hYQ) * 📄 [Documentação do Google sobre Boxplot](https://cloud.google.com/looker/docs/boxplot-options?hl=pt) * 📹 [Como criar um boxplot no Looker Studio](https://www.loom.com/share/8dc217b506df40d9ad19cc2b698770c1?sid=7eeb2bd1-1460-4dce-918e-b1747c715d13) |
| 🟣 Aplicar medidas de dispersão (desvio padrão) | Use tabelas no Looker Studio para calcular o desvio padrão | 💡Assim como nas planilhas você pode calcular o desvio padrão diretamente por meio de tabelas dinâmicas. |
| 🟣 Calcular quartis, decis ou percentis | Calcular quartis para variáveis ​​de risco relativo no BigQuery | * 💡Para o cálculo do risco relativo precisamos definir um ponto de corte, pois o cálculo do risco relativo é baseado em grupos e nosso objetivo é encontrar o grupo com maior risco relativo de ser um mau pagador, por exemplo, queremos encontrar a faixa etária com maior risco de ser um mau pagador. * Podemos usar a estrutura da tabela temporária WITH, LEFT JOIN para unir as informações, a função [NTILE](https://cloud.google.com/bigquery/docs/reference/standard-sql/numbering_functions#ntile) para calcular quartis. |
| 🟣 Calcular correlação entre variáveis ​​numéricas | Compreender a relação que existe entre variáveis ​​numéricas através de correlações. Use gráficos de dispersão e linhas de tendência. Você também pode usar o comando CORR no BigQuery | * 📹 [Gráfico de dispersão](https://www.youtube.com/watch?v=ksRa12dmxNc) * 💡 Agora que você tem todos os dados em uma tabela , entender a relação entre variáveis ​​numéricas. Lembre-se de como usamos esse tipo de gráfico no projeto 2 para entender a correlação entre 2 variáveis ​​contínuas. |

### **🟥 2.3 Aplicar técnica de análise**

⌛ Intervalo de tempo estimado: 8 a 12 horas

As técnicas de análise na área financeira e de crédito podem ser muito variadas e a escolha por uma ou outra técnica, depende do problema específico em questão que você pretende resolver. Em alguns casos, diferentes técnicas combinadas podem ajudar a alcançar um resultado ou aprofundar as descobertas. Neste marco, vamos aplicar uma técnica de análise para avaliar o risco de crédito associado a cada cliente, através do que é conhecido como cálculo de risco relativo.

O cálculo do risco relativo é uma técnica estatística usada para estimar a probabilidade de que um determinado evento ocorra (como a inadimplência em um empréstimo) em um grupo específico, em comparação com outro grupo. No nosso caso, esta técnica nos permitirá analisar e comparar o risco associado a diferentes variáveis ​​e características dos clientes, com o objetivo de compreender melhor o seu comportamento de crédito e melhorar nossas estratégias de gestão de risco.

Ao contrário da análise RFM (Recência, Frequência, Monetária) que é comumente aplicada em marketing e análise de clientes para entender seu comportamento e melhorar as estratégias de negócios, o cálculo do risco relativo concentra-se na área financeira e de crédito, fornecendo uma medida quantitativa do risco associado a cada cliente do grupo de clientes.

No processo de cálculo do risco relativo, analisamos diversas variáveis, como o histórico de pagamentos, dívida total, renda, entre outros, para posteriormente determinar como essas variáveis ​​contribuem para o risco de crédito. Esta análise permitirá segmentar os clientes em diferentes categorias de risco, o que é essencial para criar, posteriormente, uma pontuação de crédito robusta e precisa.

O cálculo do risco relativo apresenta-se como uma técnica crucial neste projeto, o que nos permitirá abordar eficazmente o problema central de avaliação e gestão do risco de crédito de clientes do banco. Esta técnica estatística será importante na nossa tarefa de construir um modelo preditivo que possa identificar e classificar o risco com precisão, facilitando assim a tomada de decisão.

| **Meta** | **Objetivo** | **Recurso** |
| --- | --- | --- |
| 🔴 Calcular risco relativo | Calcule o risco relativo em cada grupo (Mau pagador, Bom pagador) de cada variável em relação à variável “inadimplência”. | 🚨 Altamente recomendado:   * 📄 [Conceito de Risco Relativo e como calculá-lo](https://docs.google.com/document/d/18r5xjCp15M3z7IoYlfnwxIMjreIM4w9hwL7sWm782Kk) * 📹 [Conceito de Risco Relativo e como calculá-lo](https://www.youtube.com/watch?v=TKVlS3bSV04) * 📹 [Exemplo de calculo de risco relativo no contexto de analise de credito](https://www.loom.com/share/673a975913b349b39979f7962268064c?sid=f42877f0-b90c-4d44-871e-72f2383c1cca) * 💡Você pode usar [CASE WHEN](https://cloud.google.com/bigquery/docs/reference/standard-sql/conditional_expressions?hl=es-419#case_expr) para categorizar os quartis de cada variável (idade, último*mês*salário , etc. ) de acordo com o resultado do risco relativo. * Use também o comando WITH para criar tabelas temporárias em Bigquery para auxilar na construção deste calculo. |
| 🔴 Validar hipótese | Nos grupos encontrados, valide a hipótese de quais apresentam risco relativo diferente. | * Hipótese: * Os mais jovens correm um risco maior de não pagamento. * Pessoas com mais empréstimos ativos correm maior risco de serem maus pagadores. * Pessoas que atrasaram seus pagamentos por mais de 90 dias correm maior risco de serem maus pagadores. * 💡 Após validar as hipóteses de acordo com o resultado do cálculo do risco relativo, construa uma tabela com os grupos de cada variável que apresenta maior risco de ser um mau pagador. |

### **🟧 2.4 Resumir informações em um dashboard ou reporte**

⌛ Intervalo de tempo estimado: 8 a 12 horas

Um dashboard é uma ferramenta essencial de visualização de dados que fornece uma representação gráfica e resumida das informações relevantes, como, o risco de crédito dos clientes do banco. No mundo real, esta interface gráfica (assim como a base de dados) seria atualizada em tempo real para refletir alterações nas condições financeiras e de crédito, apresentando métricas de forma clara e concisa, como indicadores-chave de desempenho (KPIs), tendências de comportamento, detalhes de pagamento e outros dados importantes relacionados ao crédito, tudo em um lugar.

Como analista de dados, nosso papel é facilitar a tomada de decisões apresentando informações críticas sobre a concessão de crédito e risco associado de forma fácil de interpretar e com a capacidade de se adaptar às novas necessidades dos analistas e gestores de risco. Por exemplo, você pode incluir em seu dashboard, métricas como taxa de incumprimento e o risco relativo associado a diferentes variáveis.

| **Meta** | **Objetivo** | **Recurso** |
| --- | --- | --- |
| 🟠 Representar dados por meio de tabela resumo ou scorecards | Crie scorecards para números gerais do banco de dados | 📹 [Vídeo como criar scorecards no Looker Studio](https://www.youtube.com/watch?v=dEu4dDUT52o&pp=ygUjbG9va2VyIHN0dWRpbyBjb21vIGZhemVyIHNjb3JlY2FyZHM%3D) |
| 🟠 Representar dados através de gráficos simples | Inclua um gráfico univariado em seu dashboard. | * Univariado: São aqueles gráficos usados ​​para mostrar a distribuição de uma única variável. * 📹 [Conceito Gráfico Univariado](https://www.youtube.com/watch?v=tdF2PLa7Nfk) * 📹[Vídeo como criar gráficos de barras](https://www.youtube.com/watch?v=p4nNJWtStAs) |
| 🟠 Representar dados através de gráficos avançados | Inclua um gráfico bivariado/multivariado em seu painel. | * Bivariados São aqueles gráficos usados ​​para mostrar a associação entre duas variáveis. * 📹 [Conceito Gráfico Bivariado](https://www.youtube.com/watch?v=sox1gQRdL2w) * 📹 [Vídeo como criar um gráfico combinado](https://www.youtube.com/watch?v=rlcs7AdGfh0&t=1s) |
| 🟠 Aplicar opções de filtros (controle de dados) para gerenciamento e interação | Use filtros em seu painel para facilitar a exploração no painel. | * 📹 [Filtros de vídeo no Looker Studio](https://www.youtube.com/watch?v=zivYzP5GHTM) * 📹 [Dicas para criar um dashboard no Looker Studio](https://www.youtube.com/watch?v=9Lu9hHvwBFo) |

### **🟩 2.5 Apresentar resultados**

⌛ Intervalo de tempo estimado: 8 a 12 horas

Apresentar os resultados de uma análise de dados envolve comunicar de maneira efetiva as conclusões e resultados obtidos através dos dados processados ​​e analisados. Ou seja, não se trata apenas de mostrar números e gráficos, mas sim contar uma história coerente e significativa que ajude o público a compreender e tomar decisões informadas.

| **Meta** | **Objetivo** | **Recurso** |
| --- | --- | --- |
| 🟢 Selecionar gráficos e informações relevantes | Selecionar gráficos que respondam às questões apresentadas e ao objetivo da análise do marco 1 | 💡 Revise os recursos compartilhados em outros projetos e assista novamente aos seus vídeos de apresentação, faça uma análise e anote o que você acha que seria interessante melhorar e colocar em prática neste projeto. |
| 🟢Criar uma apresentação | Crie uma apresentação limpa e coerente que expresse a lógica usada no marco 1 | 💡 Você pode utilizar uma apresentação de slides em conjunto com o seu dashboard para apresentar os resultados deste marco |
| 🟢Apresentar resultados com conclusões e recomendações | Apresentar, além dos resultados, conclusões referentes aos achados encontrados no marco 1 | 💡Utilize suas colegas de bootcamp como audiência e peça a elas feedback sobre a sua apresentação 😉 |

ℹ️ Ao finalizar o projeto, compartilhe o link da sua ficha técnica ou repositório no Github (assista a este [video](https://www.loom.com/share/80eac7437e464e7c893d35e5b249a597?sid=56cfd171-57e5-437f-9570-759888eb35f7) sobre como criar uma conta no GitHub e criar repositórios de projetos para o seu portfólio). O repositório compartilhado deve conter as queries (consultas) usadas e o link público do seu dashboard em Looker Studio, onde seja possível ver claramente todos os passos que você seguiu e os resultados que você obteve.

### **2.6 Avaliação**

Um aspecto fundamental da sua preparação para uma carreira profissional é aprender a autoavaliar a sua aprendizagem e o resultado do seu trabalho. Cada guia contém uma diretriz muito específica e um objetivo para que você possa compreender os aspectos que determinam se você atingiu um determinado objetivo de aprendizagem ou se ainda não o fez.  
Independentemente de todos os 👍 que você tenha, um 👎(Veto) é suficiente para que você não alcance o objetivo.

🟦 Habilidade: Processar e preparar a base de dados

Metas de aprendizagem

🔵 Conectar/importar dados para outras ferramentas

* 👍 Você importou os dados como tabelas (uma para cada arquivo) dentro do ambiente de BigQuery?
* 👎Veto: Caso não tenha conseguido criar um projeto e importar as tabelas, você não atingiu essa meta.

🔵 Identificar e tratar valores nulos

* 👍Você identificou quais variáveis ​​​​possuem valores nulos usando comandos SQL?
* 👍Você tomou uma decisão sobre o que e como fazer com os valores nulos?
* 👎Veto: Se você não verificou valores nulos ou se não tomou uma decisão sobre quais valores nulos eliminar ou manter, você não atingiu essa meta.

🔵 Identificar e tratar valores duplicados

* 👍Você buscou valores duplicados utilizando comandos SQL?
* 👍Você removeu ou tratou os valores duplicados das suas tabelas?
* 👎Veto: Se você não procurou valores duplicados e não certificou a existência ou não de duplicados, você não atingiu essa meta.

🔵 Identificar e gerenciar dados fora do escopo da análise

* 👍Você selecionou as variáveis ​​​​por meio de correlação?
* 👍Você identificou e eliminou as variáveis ​​que não são permitidas (ética ou legalmente) em um modelo de crédito?
* 👎Veto: Se você não fez a seleção de variáveis, você não atingiu essa meta.

🔵 Identificar e tratar dados inconsistentes em variáveis ​​​​categóricas

* 👍Você padronizou as variáveis ​​​​categóricas corrigindo os dados do tipo string?
* 👎Veto: Se você não identificou e processou esses tipos de registros, você não alcançou essa meta.

🔵 Identificar e tratar dados inconsistentes em variáveis ​​numéricas

* 👍Você identificou e tratou os valores discrepantes?
* 👎Veto: Se você não identificou e processou esses tipos de registros, você não alcançou essa meta.

🔵 Criar novas variáveis

* 👍 Você criou novas variáveis ​​a partir das que possui no conjunto de dados utilizando comandos SQL?
* 👎Veto: Se você não criou nenhuma variável adicional, você não alcançou essa meta.

🔵 Unir tabelas

* 👍 Você reuniu as informações das diferentes tabelas com algum comando de JOIN?
* 👎Veto: Se você não conseguiu juntar os dados das diferentes tabelas, você não alcançou essa meta.

🔵 Construir tabelas auxiliares

* 👍 Você recorreu ao uso de tabelas temporárias usando o comando WITH?
* 👎Veto: Se você não criou tabelas temporárias ou novas tabelas permanentes, você não alcançou essa meta.

🟪Habilidade: Fazer uma análise exploratória

Metas de aprendizagem

🟣 Agrupar dados de acordo com variáveis ​​​​categóricas

* 👍Você entendeu o conceito de variáveis ​​​​categóricas?
* 👍Você usou uma ferramenta de visualização como o Looker Studio para realizar análise exploratória?
* 👎Veto: Se você não identificou quais são as variáveis ​​categóricas deste conjunto de dados ou não conseguiu resumir as variáveis ​​​​categóricas, você não alcançou essa meta.

🟣 Ver variáveis ​​categóricas

* 👍Você usou os resumos de variáveis ​​​​categóricas em alguma ferramenta de visualização como o Looker Studio?
* 👎Veto: Se você não conseguiu criar pelo menos um gráfico de barras, você não alcançou essa meta.

🟣 Aplicar medidas de tendência central

* 👍 Você resumiu variáveis ​​numéricas usando medidas de tendência central no Looker Studio?
* 👎Veto: Se você não conseguiu aplicar este cálculo em seu projeto, você não alcançou essa meta.

🟣 Ver distribuição

* 👍 Você usou histogramas e boxplot para visualizar a distribuição de variáveis ​​numéricas?
* 👎Veto: Se você não conseguiu aplicar esta visualização em seu projeto, você não alcançou essa meta.

🟣 Aplicar medidas de dispersão

* 👍 Você calculou a dispersão das variáveis ​​numéricas no Looker Estúdio?
* 👎Veto: Caso você não tenha conseguido aplicar esse cálculo através de tabelas em Looker Studio, você não alcançou essa meta.

🟣 Aplicar correlação entre variáveis ​​numéricas

* 👍 Você avaliou a relação entre as variáveis ​​numéricas do conjunto de dados?
* 👎Veto: Caso você não tenha conseguido aplicar esse cálculo através de fórmulas em SQL em seu projeto, você não alcançou essa meta.

🟣 Calcule quartis, decis ou percentis

* 👍 Você criou quartis para as variáveis a ​​analisar?
* 👎Veto: Caso você não tenha conseguido aplicar esse cálculo através de comandos SQL, você não alcançou essa meta.

🟥 Habilidade: Aplicar técnica de análise

Metas de aprendizagem

🔴 Calcular o risco relativo

* 👍 Você entendeu o conceito de risco relativo?
* 👍 Você calculou o risco relativo para cada segmento de clientes em cada variável?
* 👎Veto: Se você não entendeu ou não conseguiu aplicar esse cálculo através de fórmulas em SQL em seu projeto, você não alcançou essa meta.

🔴 Validar hipótese

* 👍 Você validou as hipóteses levantadas através do risco relativo?
* 👎Veto: Se você não entendeu ou não conseguiu aplicar esta análise em seu projeto, você não alcançou essa meta.

🟧 Habilidade: Resumir informações em um dashboard ou reporte

Metas de aprendizagem

🟠 Representar dados por meio de tabela resumo ou scorecards

* 👍Você criou um dashboard com scorecards (cartão de resultados) com números totais ou tabelas resumo que mostram os números totais do negócios?
* 👎Veto: Se você não criou um dashboard, você não alcançou essa meta.

🟠Representar os dados por meio de gráficos simples

* 👍Você criou gráficos de barras, gráficos de linhas ou gráficos de pizza para representar suas descobertas da análise exploratória?
* 👍Você criou gráficos para representar visualmente os resultados da análise?
* 👎Veto: Se você não conseguiu criar pelo menos um gráfico para representar os resultados obtidos, você não atingiu esse objetivo de aprendizagem.

🟠 Representar os dados por meio de gráficos ou recursos visuais avançados

* 👍¿Você criou gráficos bivariados e multivariados?
* 👎Veto: Se você não conseguiu criar pelo menos um gráfico para representar os resultados obtidos, você não alcançou essa meta.

🟠 Aplicar opções de filtros para gerenciamento e interação

* 👍Você aplicou filtros no dashboard para visualizar os dados de jeitos diferentes?
* 👎Veto: Se você não conseguiu criar pelo menos um filtro, você não alcançou essa meta.

🟩 Habilidade: Apresentar Resultados

Metas de aprendizagem

🟢 Selecionar gráficos e informações relevantes

* 👍Você selecionou para sua apresentação apenas os gráficos que explicam as informações que você deseja apresentar?
* 👎Veto: Se você não conseguiu selecionar e resumir seu projeto em apenas alguns gráficos, você não alcançou essa meta.

🟢 Criar uma apresentação

* 👍Você criou uma apresentação de slides coerente, clara e limpa?
* 👍Você treinou a apresentação sozinha ou com outras colegas para se sentir mais segura?
* 👎Veto: Se você não criou uma apresentação e vai apresentar diretamente em seu dashboard, você não alcançou essa meta.

🟢 Apresentar resultados com conclusões e recomendações

* 👍Você gravou algum vídeo respeitando o limite de 5 minutos?
* 👍No seu vídeo, você falou claramente?
* 👍Você incluiu em sua apresentação conclusões e recomendações para o negócio?
* 👍Você definiu bem seu público e fluxo de pensamento?
* 👎Veto: Se você não respeitou o limite de 5 minutos na sua apresentação, você não alcançou essa meta.  
    
    
    
  COLUNAS:

| **Arquivo** | **Variável** | **Descrição** |
| --- | --- | --- |
| user\_info | user id | Número de identificação do cliente (único para cada cliente) |
|  | age | Idade do cliente |
|  | sex | Gênero do cliente |
|  | last month salary | Último salário mensal que o cliente informou ao banco |
|  | number dependents | Número de dependentes |
| loans\_outstanding | loan id | Número de identificação do empréstimo (único para cada empréstimo) |
|  | user id | Número de identificação do cliente |
|  | loan type | Tipo de empréstimo (real state = imóveis, others= outros) |
| loans\_detail | user id | Número de identificação do cliente |
|  | more 90 days overdue | Número de vezes que o cliente apresentou atraso superior a 90 dias |
|  | using lines not secured personal assets | Quanto o cliente está utilizando em relação ao seu limite de crédito, em linhas que não são garantidas por bens pessoais, como imóveis e automóveis |
|  | number times delayed payment loan 30 59 days | Número de vezes que o cliente atrasou o pagamento de um empréstimo (entre 30 e 59 dias) |
|  | debt ratio | Relação entre dívidas e ativos do cliente. Taxa de endividamento = Dívidas / Patrimonio |
|  | number times delayed payment loan 60 89 days | Número de vezes que o cliente atrasou o pagamento de um empréstimo (entre 60 e 89 dias) |
| default | user id | Número de identificação do cliente |
|  | default flag | Classificação dos clientes inadimplentes (1 para clientes já registrados alguma vez como inadimplentes, 0 para clientes sem histórico de inadimplência) |

**Etapa 1: Big Query: Estruturação e Modelagem**

Na consulta que executei, criei uma nova tabela chamada **projeto3-421317.Proj\_3.dataset\_risco\_relativo.**   
  
Passo a passo:

1- **Correção de informações do usuário** (user\_info\_corrected): Corrigi as informações dos usuários, preenchendo os valores ausentes na coluna last\_month\_salary com a média dos salários do último mês. Isso foi feito usando uma subconsulta para calcular a média dos salários e um CROSS JOIN para preencher os valores ausentes.

2- Correção de padrão de pagamento (default\_corrected): Corrigi o padrão de pagamento, atribuindo 1 para casos em que o default\_flag é igual a 1 e 0 para os outros casos. Isso foi feito usando uma estrutura CASE.

3- Correção de detalhes dos empréstimos (loans\_detail\_corrected): Corrigi os detalhes dos empréstimos, substituindo os valores nulos por 0 nas colunas relevantes.

Correção de empréstimos pendentes (loans\_outstanding\_corrected): Corrigi os empréstimos pendentes, substituindo os valores nulos por 0 nas colunas relevantes.

Correção de outliers (outliers\_corrected): Identifiquei outliers nos dados de empréstimos usando medidas estatísticas como média e desvio padrão. Os outliers foram marcados com 1 nas colunas relevantes, caso a diferença entre o valor observado e a média fosse maior que 3 vezes o desvio padrão.

4- Seleção e renomeação das colunas na nova tabela: Por fim, selecionei as colunas relevantes da tabela corrigida e renomeei-as de acordo com o idioma em português. Isso foi feito usando a estrutura SELECT ... AS ....

Resultado: Uma série de correções nos dados relacionados a informações de usuários, padrões de pagamento e detalhes de empréstimos, identificação de outliers nos dados de empréstimos e, por fim, criação de uma nova tabela com as informações corrigidas e renomeadas em português.

1. **Obtenção dos Dados no BigQuery:**
   * Iniciei o projeto realizando consultas no BigQuery para obter informações sobre os empréstimos e os usuários.
2. **Limpeza dos Dados:**
   * Criei uma tabela limpa chamada loans\_outstanding\_cleaned para armazenar os tipos de empréstimos padronizados.
3. **Sumarização dos Dados de Empréstimos por Usuário:**
   * Criei a tabela user\_loans\_summary para resumir o número total de empréstimos por usuário.
4. **Análise de Correlação entre Dados:**
   * Calculei a correlação entre o número total de empréstimos por usuário e várias características dos usuários, como idade, salário, etc.
5. **Correção e Preparação dos Dados:**
   * Criei tabelas para corrigir e padronizar os dados de usuários, empréstimos e padrões, garantindo consistência e completude dos dados.
6. **Criação do Conjunto de Dados Estruturado:**
   * Combinei todas as tabelas corrigidas em uma tabela estruturada chamada dataset\_risco\_relativo\_estruturado, que inclui informações detalhadas sobre os usuários e seus empréstimos, bem como a classificação de inadimplência, inclui o **Score de Crédito** que multiplica os dias em atrasos:
7. **Análise Exploratória dos Dados:**
   * Realizei uma análise exploratória dos dados, incluindo estatísticas descritivas e visualizações, para entender melhor a distribuição e as relações entre as variáveis.

Para criar o score de crédito:

* + Atraso superior a 90 dias: atribuído peso: 3
  + Atraso entre 60 a 89 dias atribuído peso: 2
  + Atraso entre 30 a 59 dias atribuído peso: 1

O resultado da multiplicação desses pesos é somada à taxa de endividamento e o resultado é o Score de Crédito.

1. **Exportação do Conjunto de Dados:**
   * Exportei o conjunto de dados estruturado para um arquivo CSV chamado dataset\_risco\_relativo.csv, pronto para ser utilizado em análises e modelagem de machine learning.

[Consulta no BigQuery  
- Projeto 3- Risco Relativo](https://console.cloud.google.com/bigquery?project=projeto3-421317&ws=!1m4!1m3!8m2!1s6616295630!2s956ebfdefd7d4460a980e8d69a3e1f19)

(Tempo de dedicação: 4 dias)

**Etapa 2- Cálculo de Risco Relativo- Google Colab (Python)**

Com o aumento da demanda por crédito devido à queda das taxas de juros, a equipe de análise de crédito do banco "Super Caja" enfrenta desafios significativos. Nosso objetivo é automatizar o processo de análise de crédito para melhorar a eficiência e precisão, utilizando técnicas avançadas de análise de dados.

Descrição do Conjunto de Dados Utilizamos um conjunto de dados com 36.000 amostras e 14 variáveis, incluindo variáveis numéricas como idade, último salário informado, e variáveis categóricas como gênero e estado civil.

Análise exploratória detalhada, incluindo análise descritiva, visualização de dados e análise de correlações. Identificamos padrões, tendências e correlações entre as variáveis.

3. Cálculo do Risco Relativo

Após o cálculo do risco relativo, encontramos os seguintes resultados:

Atraso superior a 90 dias (5.2160):

Clientes com atraso superior a 90 dias possuem um risco de inadimplência 5,22 vezes maior que a média. Essa variável também demonstra um impacto significativo no risco.

Atraso entre 60 e 89 dias (4.7570):

Clientes com atraso entre 60 e 89 dias apresentam um risco de inadimplência 4,76 vezes maior que a média.

Atraso entre 30 e 59 dias (7.1770):

Clientes com histórico de atraso entre 30 e 59 dias apresentam um risco de inadimplência 7,18 vezes maior que a média da população. Esse é o fator de maior impacto no risco de inadimplência.

Taxa de endividamento (3.1630):

Clientes com uma alta taxa de endividamento têm um risco de inadimplência 3,16 vezes maior que a média.

Total de empréstimo por ID (2.0840):

Clientes com um valor total de empréstimos mais elevado apresentam um risco de inadimplência 2,08 vezes maior que a média.

Score segmentado (1.0900):

Essa variável, possivelmente proveniente de um modelo de score de crédito pré-existente, contribui para o risco de inadimplência com um impacto de 1,09 vezes a média.

Demais variáveis que não são tão importantes:

Idade (0.9841), número de dependentes (0.8274) e último salário informado (0.7459) apresentaram impacto relativamente baixo no risco de inadimplência.

4. Treinamento e Avaliação do Modelo

Foram utilizados diferentes algoritmos de aprendizado de máquina, incluindo Regressão Logística, Random Forest, XGBoost e LightGBM. O modelo com melhor desempenho foi o XGBoost, com acurácia de 84,05% no conjunto de treinamento. No entanto, após ajustes, o modelo Random Forest apresentou uma acurácia impressionante de 99,99%.

O desempenho do modelo foi avaliado em diferentes conjuntos de dados, utilizando técnicas como validação cruzada e holdout. Os resultados comprovam a robustez do modelo e sua capacidade de generalização para novos dados.

Análise Detalhada das Métricas Acurácia: A acurácia do modelo Random Forest ajustado foi de 99,99%, indicando sua capacidade de classificar corretamente as amostras do conjunto de dados.

Matriz de Confusão Ajustada (Random Forest).  
  
Já quando olhamos para curva ROC AUC, identificamos que nos modelos ajustados, as previsões para bons pagadores são feitas corretamente para 79% dos casos, enquanto maus pagadores são identificados corretamente 89% dos casos.

Conclusão

O modelo de predição de inadimplência desenvolvido para o banco "Super Caja" é robusto e eficaz, fornecendo informações valiosas para a tomada de decisão. Sua implementação permitirá ao banco melhorar a eficiência e rapidez na análise de crédito, reduzindo o risco de empréstimos não reembolsáveis e fortalecendo sua posição no mercado financeiro.

Os gráficos demonstram que em diversas modelagens de previsão as variáveis que contabilizam os números de dia em atraso e a taxa de endividamento são variáveis importantes para definir o perfil de risco dos clientes.

[Google Colab- Projeto 3- Risco Relativo](https://colab.research.google.com/drive/1Q--zEXfLQFZKjjaH-mmcpXYvqf1pumgE#scrollTo=xZ84GCUeA48o) [Trilha de Aprendizagem em Python- Projeto 3](https://docs.google.com/document/u/0/d/1ieL6AbhanGT9ylwEWjI4ChgM8vu0cR1W3C45GZVZXSA/edit)

**Etapa 3- Criação de Dashboard no Looker**

Para criar o dashboard no Looker, dois arquivos foram importados: dataset\_risco\_relativo e dataset\_risco\_relativo\_rr, um com os dados manipulados anteriormente via BigQuery e outro com dados manipulados via Python no Google Colab.   
  
A apresentação é dividida por tópicos:  
1- Análise exploratória  
2- Score de Crédito  
3- Buscador Super Caja  
5- Hipóteses  
6- Modelos de Previsão  
7- Pontos de Atenção  
  
Na análise exploratória aspectos gerais foram levantados para ilustrar o aspecto do bando de dados antes e após intervenções e melhorias, ilustrando em gráficos a amostra da população de acordo com os aspectos encontrados, incluindo um gráfico que avalia as variáveis com maior risco relativo.  
  
Em Score de Crédito é explicado a maneira de criação do mesmo e as regras utilizadas, além de ilustrar em gráfico de barras e linha a proporção da população de acordo com a faixa etária e o Score desses clientes.   
  
O Buscador Super Caja entretanto, vem como a ferramenta de solução encontrada após a análise, testes e avaliação de hipóteses, uma ferramenta simples com um buscador de id e visualizações dos empréstimos ativos, dias que este cliente atrasou, taxa de endividamento e um apontador de score de crédito.   
  
Em hipóteses nos aprofundamos nos gráficos de dispersão para compreender a distribuição desses clientes e avaliar seus comportamentos de acordo com cada variável, mostrando quais delas são fundamentais e quais não têm impacto.   
  
Já na página do modelo de previsão, a ideia é mostrar a curva ROC AUC, que define a acurácia geral do modelo de acordo com alguns aspectos, dando enfase então no recall ou seja, a taxa de previsões corretas para bons ou maus pagadores, finaliza explicando a aplicabilidade e a eficiência do modelo de classificação de inadimplência.   
  
Em pontos de atenção é levantado então o apontamento a descoberta feita na análise exploratória, indicando a diferença salarial informada por homens e mulheres e a taxa de endividamento maior para mulheres, indicando alguns insights como a criação de um modelo de crédito só para mulheres, diferentes taxas de juros, facilitação no pagamento e até a prática de menores parcelas.   
  
Ainda em pontos de atenção, finalizamos indicando alguns aspectos que não foram encontrados no banco de dados como data de efetivação e finalização dos empréstimos, valor dos empréstimos e valor das dívidas; também é apontado um insight referente a classificação dos clientes, ao invés de maus ou bons pagadores, a ideia seria atribuir níveis de inadimplência.   
  
[Dashboard- Super Caja](https://lookerstudio.google.com/u/0/reporting/88e9a323-6cc6-410b-b19f-715310512c30/page/p_tztbukukhd)

[Apresentação](https://www.youtube.com/watch?v=pJruKkJTj9I)