



igti

RELATÓRIO

PROJETO APLICADO

Instituto de Gestão e Tecnologia da Informação
Relatório do Projeto Aplicado

Melhoria do Serviço de Empréstimo de um Banco Utilizando Técnicas de Machine Learning

Renata Kelly Marcelino dos Santos

Orientador(a): Daniel Viana

20/03/2023



RENATA KELLY MARCELINO DOS SANTOS

INSTITUTO DE GESTÃO E TECNOLOGIA DA INFORMAÇÃO

RELATÓRIO DO PROJETO APLICADO

Melhoria do Serviço de Empréstimo de um Banco Utilizando Técnicas de Machine Learning

Relatório de Projeto Aplicado
desenvolvido para fins de conclusão do
curso MBA em Ciência de Dados.

Orientador (a): Daniel Viana

Recife

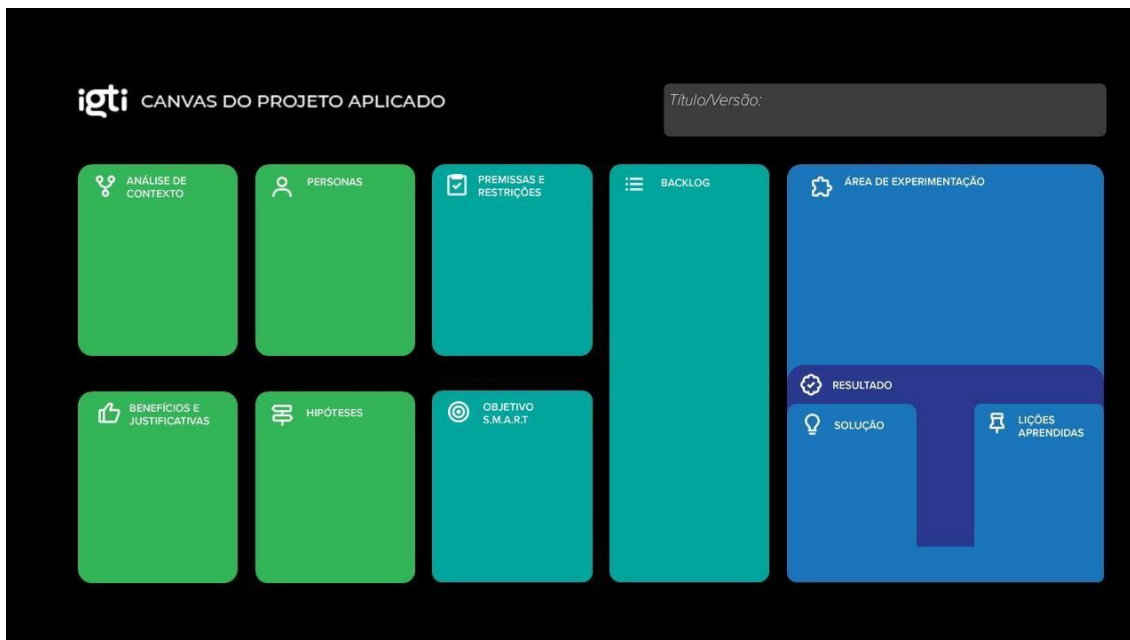
20/03/2023

Sumário

1. CANVAS do Projeto Aplicado	4
1.1 Desafio	5
1.1.1 Análise de Contexto	5
1.1.2 Personas	6
1.1.3 Benefícios e Justificativas	7
1.1.4 Hipóteses	8
1.2 Solução	9
1.2.1 Objetivo SMART	9
1.2.2 Premissas e Restrições	11
1.2.3 Backlog de Produto	13
2. Área de Experimentação	14
2.1 Sprint 1	16
2.1.1 Solução	16
• Evidência do planejamento:	16
• Evidência da execução de cada requisito:	16
• Evidência dos resultados:	16
2.1.2 Experiências vivenciadas	16
2.2 Sprint 2	17
2.2.1 Solução	17
• Evidência do planejamento:	17
• Evidência da execução de cada requisito:	17
• Evidência dos resultados:	17
2.2.2 Experiências vivenciadas	17
2.3 Sprint 3	18
2.3.1 Solução	18
• Evidência do planejamento:	18
• Evidência da execução de cada requisito:	18
• Evidência dos resultados:	18
2.3.2 Experiências vivenciadas	18
3. Considerações Finais	19
3.1 Resultados	19
3.2 Contribuições	19
3.3 Próximos passos	19

1. CANVAS do Projeto Aplicado

Figura conceitual, que representa todas as etapas do Projeto Aplicado.



1.1 Desafio

1.1.1 Análise de Contexto

É sempre um problema interessante e desafiador para os bancos prever a probabilidade de inadimplência de um cliente em um empréstimo, com apenas uma pequena informação sobre o cliente.

Atualmente, os cientistas de dados de um banco usam aprendizado de máquina para criar modelos preditivos. Os conjuntos de dados que eles usam podem ser da própria empresa e geralmente são coletados internamente por meio de suas operações diárias. Dessa forma, se quisermos trabalhar em tais projetos financeiros, não há muitos conjuntos de dados do mundo real que possamos usar. Felizmente, há uma exceção: o '*Berka Dataset*'. Este conjunto de dados nos proporciona realizar análises de crédito do *Czech Bank* (banco da República Tcheca).

Uma breve descrição do *Berka Dataset*, ou '*PKDD'99 Financial Dataset*', é uma coleção de informações financeiras anônimas reais de um banco tcheco, usado para o *PKDD'99 Discovery Challenge*.

O desafio deste trabalho é desenvolver uma solução em machine learning para prever a inadimplência do empréstimo como um problema de classificação binária, já que o banco deseja melhorar seus serviços a fim de diminuir o risco e aumentar o lucro — esse seria o cenário ideal.

A causa do problema de perda de receita com clientes inadimplentes em empréstimos, é que muitas vezes os gerentes do banco tem apenas uma vaga ideia de quem é um bom cliente (a quem oferecer alguns serviços adicionais ou vantagens) e quem é um cliente ruim (a quem observar cuidadosamente para minimizar as perdas do banco). Felizmente, o banco armazena dados sobre seus clientes, as contas (movimentos em vários meses), os empréstimos já concedidos, os cartões de crédito emitidos, etc.

- Matriz CSD - Certezas Suposições Dúvidas

MATRIZ CSD		Certezas	Suposições	Dúvidas
Diferentes Óticas de Análises	Atores	O banco deseja melhorar seu serviço de empréstimo para seus clientes .	O banco deseja diminuir os riscos e aumentar a receita. Os clientes dos bancos desejam empréstimos.	Existe alguma regra essencial que o banco impõe para que o cliente seja minimamente apto a ter um empréstimo?
	Cenários	Existem dados históricos com informações transacionais dos clientes no banco.	Os clientes podem atrasar o pagamento ao banco das parcelas dos empréstimos ou não atrasar.	O período histórico de informações dos clientes que o banco possui será suficiente para uma análise de crédito para conceder-lhes um empréstimo?
	Regras	Na base para análise, existem mais empréstimos que se enquadram na classe “cliente bom” do que os da classe “cliente ruim”.	Não se aplica	Não se aplica

- Observação do tipo POEMS.

PESSOAS	OBJETOS	AMBIENTE	MENSAGEM	SERVIÇOS
Quem está presente no contexto em análise?	Que objetos fazem parte do ambiente?	Quais são as características do ambiente?	Que mensagens são comunicadas?	Quais serviços são oferecidos?
Clientes do banco que desejam/estão aptos a um empréstimo.	Não se aplica	Ambiente para negociação de empréstimo pode ser no próprio banco.	Se o empréstimo foi concedido ou negado.	Além dos empréstimos, o banco tem serviço de contas bancárias e cartão de crédito.

1.1.2 Personas

As pessoas envolvidas diretamente nesse problema de concessão de empréstimo no Banco da República Tcheca, são os seus clientes que desejam ou estão aptos a um empréstimo. Mas quem são essas pessoas?

As informações reais dessas pessoas estão na base de dados, onde pode-se encontrar informações como gênero, data de aniversário (então posteriormente sua idade), informações demográficas, informações sobre sua conta bancária e o uso dela, entre outras informações.

Abaixo uma persona criada no site <https://www.geradordepersonas.com.br/>, para ilustrar um cliente do banco da República Tcheca.



JAN FRAIT

CONTADOR

ADULTO DE MEIA-IDADE (41-59)

Mini-bio

Jan Frait é um homem reservado, que gosta de viajar com sua família e agradá-los sempre que pode. Jan Frait tem um casal de gêmeos de 11 anos.



Detalhes Pessoais

Localização

Praga

Renda Familiar

De R\$3.501,00 a R\$6.500,00

Nível Educacional

Ensino Superior

Status de Relacionamento

Casado(a)



Carreira

Empresa

Società Praga

Tamanho da empresa

Pequeno porte

Responsabilidades Profissionais

prestação de contas, escrituração fiscal e administração tributária

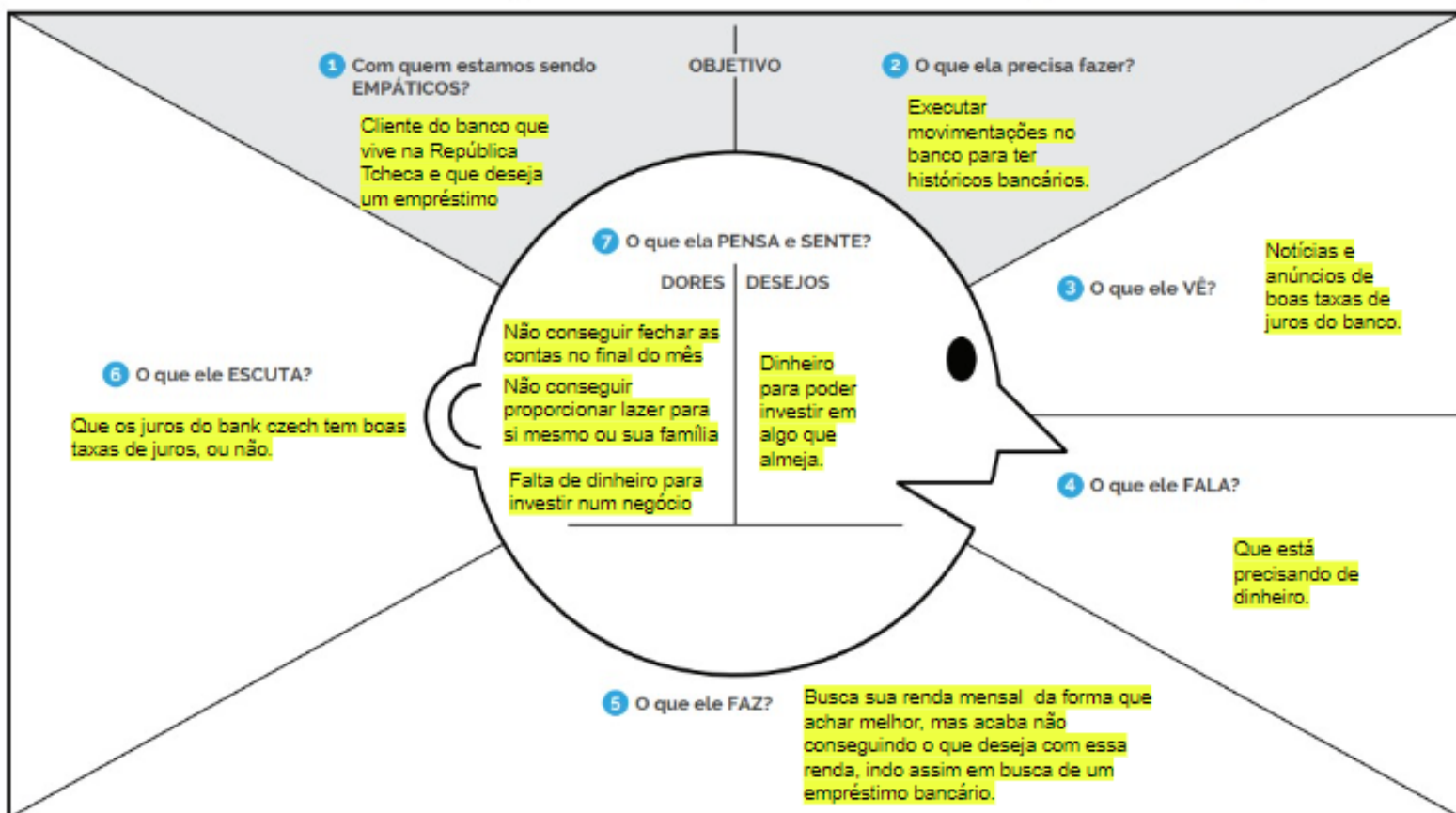
Objetivos

Sob demanda com objetivos de deixar o cliente satisfeito

Desafios

Nem sempre consegue deixar um dinheiro extra no final do mês para poder dar um lazer a sua família

- Mapa de Empatia



1.1.3 Benefícios e Justificativas

- Justificativa do projeto de melhoria do serviço de empréstimo do banco

Melhorar a assertividade do produto do serviço de empréstimo, a fim de diminuir a perda de receita que o banco tem com os clientes inadimplentes. O que será feito poderá dar um “Norte” para quem o banco deve dar uma atenção melhor baseado no seu score de risco de crédito (clientes com baixo score são os clientes com mais risco de se oferecer o empréstimo e clientes com alto score são aqueles com menor risco).

O investimento nesse projeto é justificado baseado no retorno financeiro que ele trará para o banco auxiliando na concessão de empréstimos, que são os benefícios futuros esperados.

No cenário atual de concessão de empréstimos, os gerentes do banco possuem apenas uma vaga ideia de quem é um bom cliente (a quem oferecer alguns serviços adicionais) e quem é um cliente ruim (a quem observar cuidadosamente para minimizar as perdas do banco). A proposta de valor desse projeto é direcionar melhor

esses bons e ruins clientes para os gerentes tomarem melhores decisões e assim diminuir as perdas do banco, mas também proporcionar uma concessão de crédito mais assertiva proporcionando que mais clientes bons sejam aprovados.

- Blueprint.

Exploração do problema do empréstimo no banco

Serviço de empréstimo oferecido pelo banco				
Ações dos clientes	Quer dinheiro emprestado	Procura banco	Faz simulação	Faz empréstimo
Objetivos	Comprar algo, investir em algo, etc..	Encontrar banco com menor taxa de juros	Se o empréstimo é aprovado ou não	Usar o dinheiro
Atividades	Procura indicações	Procura opções online, jornais, faz listas	Aguardar resultado que depende apenas do banco	Compra ou investe no que deseja, etc
Questões	Qual valor do empréstimo irá conseguir	Qual banco vai ter melhores taxas de juros? Qual banco tem melhor atendimento?	Quando terei o resultado da aprovação do empréstimo?	Quantas parcelas terei que pagar no empréstimo?
Barreiras	Taxa de juros alto	Banco muito longe de sua residência	Demora no resultado da aprovação	Taxas escondidas que não ficaram claras no momento do contrato

Exploração da solução proposta

Serviço de empréstimo oferecido pelo banco				
Ações dos clientes	Quer dinheiro emprestado	Procura banco	Faz simulação	Faz empréstimo
Funcionalidades	Não se aplica	Não se aplica	Melhoria do modelo de aprovação do empréstimo, tornando a concessão mais assertiva e com embasamentos estatísticos e probabilísticos	Não se aplica
Interação	Não se aplica	Não se aplica	Ter relacionamento com o banco, fazendo transações e deixando um bom histórico na base de dados para ter uma boa pontuação (score)	Não se aplica
Mensagem	Não se aplica	Não se aplica	Empréstimo aprovado ou não	Não se aplica

Tarefas e processos para alcançar os objetivos esperados

Serviço de empréstimo oferecido pelo banco				
Ações dos clientes	Quer dinheiro emprestado	Procura banco	Faz simulação	Faz empréstimo
Onde ocorre	Casa, escritório, bares, internet, etc	Internet, jornais, revistas, indicações	Direto no banco, por telefone, etc	Direto no banco
Tarefas aparentes	Não se aplica	Não se aplica	Não se aplica	Não se aplica
Tarefas escondidas	Não se aplica	Não se aplica	Modelagem de crédito para análise de aprovação ou não	Não se aplica
Processos de suporte	Não se aplica	Não se aplica	Resposta rápida para resposta de aprovação ou não do empréstimo	Não se aplica
Saída desejável	Não se aplica	Não se aplica	Que o empréstimo seja aprovado	Que o cliente usufrua do crédito concedido e que não se torne inadimplente

- CANVAS de Proposta de Valor

Proposta de Valor



Perfil do Cliente



1.1.4 Hipóteses

- Matriz de observações para hipóteses.

Exemplo de observação	Exemplo de hipótese
Modelos de predição em geral (não apenas os de concessão de crédito) tornam-se, em sua maioria, tendenciosos quando inseridas variáveis que relatam o sexo, raça ou escolaridade.	Os modelos de concessão de crédito não deveriam utilizar variáveis que envolvessem características físicas do cliente, como sexo e raça. A escolaridade também deveria ser ignorada, pois o quanto o cliente estudou não deveria influenciar em seu comportamento inadimplente.
As pessoas se sentem incomodadas quando sabem que possuem um bom comportamento bancário (não deixa de pagar suas dívidas, não realiza fraudes, etc) e mesmo assim seu pedido de empréstimo foi negado.	As variáveis incluídas num modelo de concessão de crédito devem fazer sentido e para isso o analista responsável pelo modelo precisa fazer uma boa análise exploratória bivariada com a variável resposta que possui.

- Priorização de Ideias.

SOLUÇÃO	B	A	S	I	C	O	TOTAL	PRIORIZAÇÃO
Criação de variáveis relevantes no modelo	5	5	5	5	2	2	24	1 ^a
Focar na assertividade do modelo	4	5	4	5	2	3	23	2 ^a
Validar se o modelo desenvolvido está coerente com o serviço de empréstimo do banco	4	3	4	5	3	2	21	3 ^a
Boa avaliação de desempenho para escolha do melhor modelo dentre os que foram desenvolvidos	3	4	3	5	2	2	19	4 ^a
Obtenção de dados históricos dos clientes do banco	5	5	2	2	2	2	18	5 ^a

1.2 Solução

1.2.1 Objetivo SMART

Objetivo geral: Melhorar o produto de análise de crédito para concessão de empréstimo aos clientes do banco.

- **Específico:** Criar variáveis que façam sentido com o conceito de risco e empréstimo, a fim de usá-las para gerar um bom modelo;
- **Mensurável:** Obter um modelo com acurácia mínima de 70%;
- **Atingível:** Utilizar os dados disponibilizados pelo banco de dados do banco Tcheco;
- **Relevante:** Segmentar os clientes por faixa de score, ao final da modelagem, para ter uma melhor noção dos bons e maus clientes;
- **Temporal:** O modelo deve ser construído até o final dessa disciplina.

1.2.2 Premissas e Restrições

Premissas

Premissas	
Custo	Este projeto vai gerar um custo de infraestrutura (internet, computador, energia elétrica)
Escopo	negócio: este modelo será aplicado apenas para pessoas físicas
	base de dados: a base de dados utilizada é do ano de 1999
	técnico: modelo será desenvolvido em PySpark baseando-se na biblioteca MLlib
Qualidade	Para ser um modelo minimamente viável para ser usado em produção, seria ideal uma acurácia de 70%
Prazo	Concluir o modelo e suas análises até o final desta disciplina

- Matriz de Riscos.

Risco Identificado	Impacto Potencial	Ações preventivas	Ações corretivas
Falta de acesso às tecnologias necessárias para desenvolvimento (energia elétrica, WIFI, computador)	Pode gerar o impedimento do desenvolvimento do modelo e o atraso da entrega	Ter um ponto de apoio com wifi, energia elétrica e salvar arquivos importantes na nuvem.	Ter um ponto de apoio com wifi, energia elétrica e salvar arquivos importantes na nuvem.
Modelo com acurácia baixa (< 70%)	Não mapeamento de possíveis aprovações de bons clientes para empréstimo e caso de aprovar clientes ruins que tem mais chances de ser inadimplentes, o banco pode sair no prejuízo.	criar boas variáveis para que o modelo possa aprender de forma correta sobre o comportamento e perfil dos clientes.	caso o modelo não atinja o esperado, voltar e re-analisar (reamostrar, recalcular variáveis, procurar problemas na base.

1.2.3 Backlog de Produto

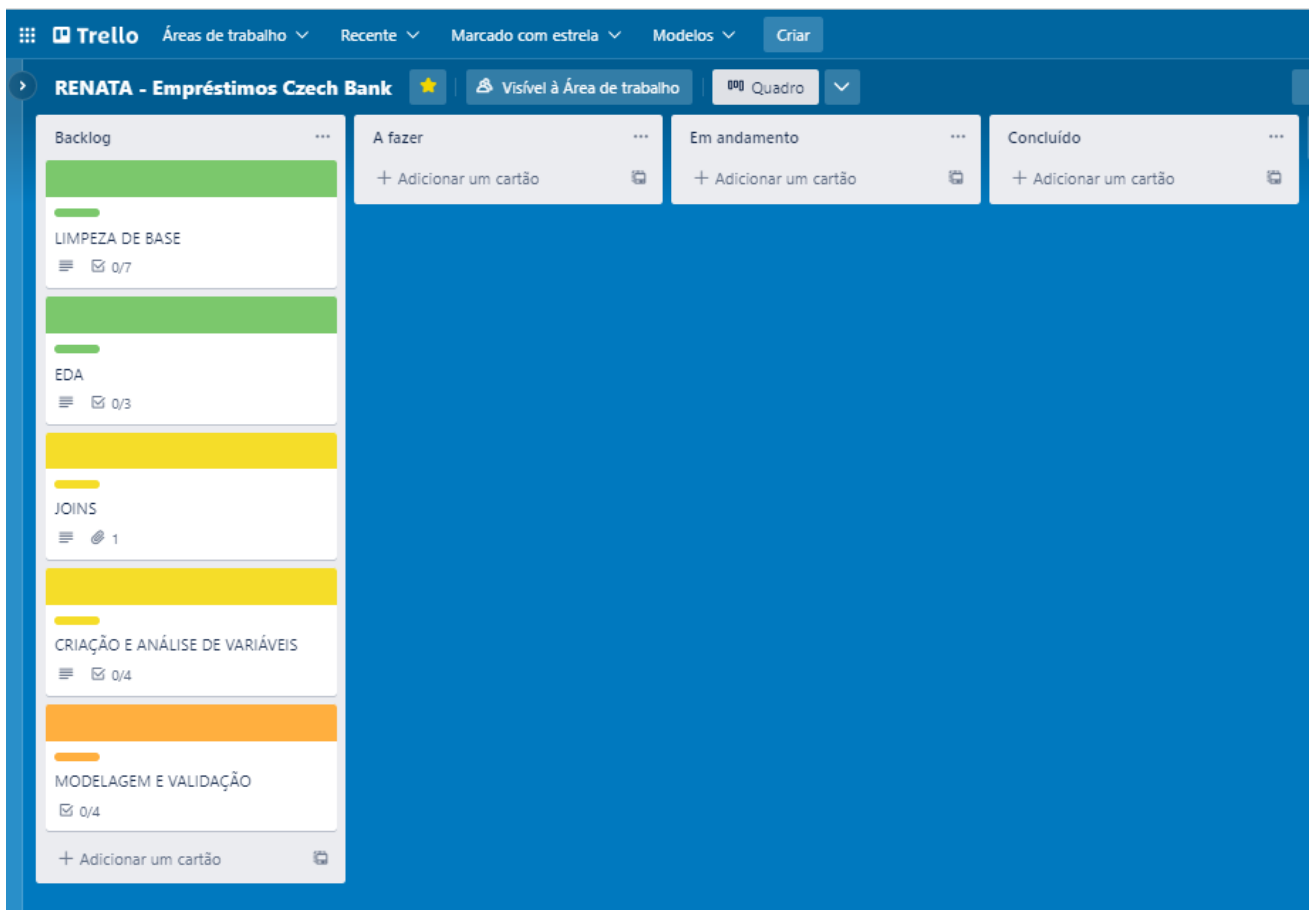
Link do trelo:

<https://trello.com/invite/b/JqGfVw86/ATTI2ae2eed9bbf003f9eab7702cd3bc699577231EEF/emprestimos-czech-bank>

VERDE: SPRINT 1

AMARELO: SPRINT 2

LARANJA: SPRINT 3



Capa

EDA

na lista [Backlog](#)

Etiquetas

SPRINT 1 +

Notificações

Seguir

Adicionar ao cartão

Membros

Etiquetas

Checklist

Datas

Anexo

Campos Personaliza...

Power-Ups

+ Adicionar power-...

Automação ⓘ

+ Adicionar botão

Ações

Descrição Editar

Análise exploratória dos dados para conhecer melhor as variáveis.

Estudar o significado das variáveis para cada tabela de dados no dicionário disponibilizado.

✓ Etapas EDA

Excluir

0%

☐ Análise univariada para todas as variáveis de todas as tabelas
 ☐ Análise bivariada com a variável de interesse na tabela LOAN
 ☐ Geração de gráficos para visualização dos dados

Adicionar um item

Capa

JOINS

na lista [Backlog](#)

Etiquetas

SPRINT 2 +

Notificações

Seguir

Adicionar ao cartão

Membros

Etiquetas

Checklist

Datas

Anexo

Campos Personaliza...

Power-Ups

+ Adicionar power-...

Automação ⓘ

+ Adicionar botão

Ações

Descrição Editar

Estudo do '**diagrama entidade relacional**' da base de dados para entender o relacionamento das tabelas e decidir quais atributos irei utilizar.

Você tem edições não salvas neste campo. [Exibir edições](#) • [Descartar](#)

Anexos

image.png ↗

Adicionado: 10 minutos atrás • [Comentário](#) • [Excluir](#) • [Editar](#)

Capa

Criar Capa

Adicionar um anexo

Atividade

Mostrar Detalhes

CRIAÇÃO E ANÁLISE DE VARIÁVEIS

na lista Backlog

Etiquetas

Notificações

SPRINT 2

+

Seguir

Descrição

Editar

Criação de variáveis que abrangem o contexto de empréstimos bancários.

Etapas de criação de variáveis

Excluir

0%

☐

Estudar quais variáveis criar para o contexto do projeto

☐

EDA das variáveis criadas

☐

Análise de correlação das variáveis

☐

Seleção de variáveis

Adicionar um item

Membros

Etiquetas

Checklist

Datas

Anexo

Campos Per

Power-Ups

+ Adicionar p

Automação

+ Adicionar b

Ações

Mostrar Detalhes

Capa

MODELAGEM E VALIDAÇÃO

na lista Backlog

Etiquetas

SPRINT 3

+

Notificações

+ Seguir

Descrição

Adicione uma descrição mais detalhada...

etapas modelagem

Excluir

0%

Amostragem

Seleção de variáveis

Testes de algoritmos

Validação com o algoritmo escolhido

Adicionar um item

Adicionar ao cartão

Membros

Etiquetas

Checklist

Datas

Anexo

Campos Personaliza...

Power-Ups

+ Adicionar power-...

Automação

+ Adicionar botão

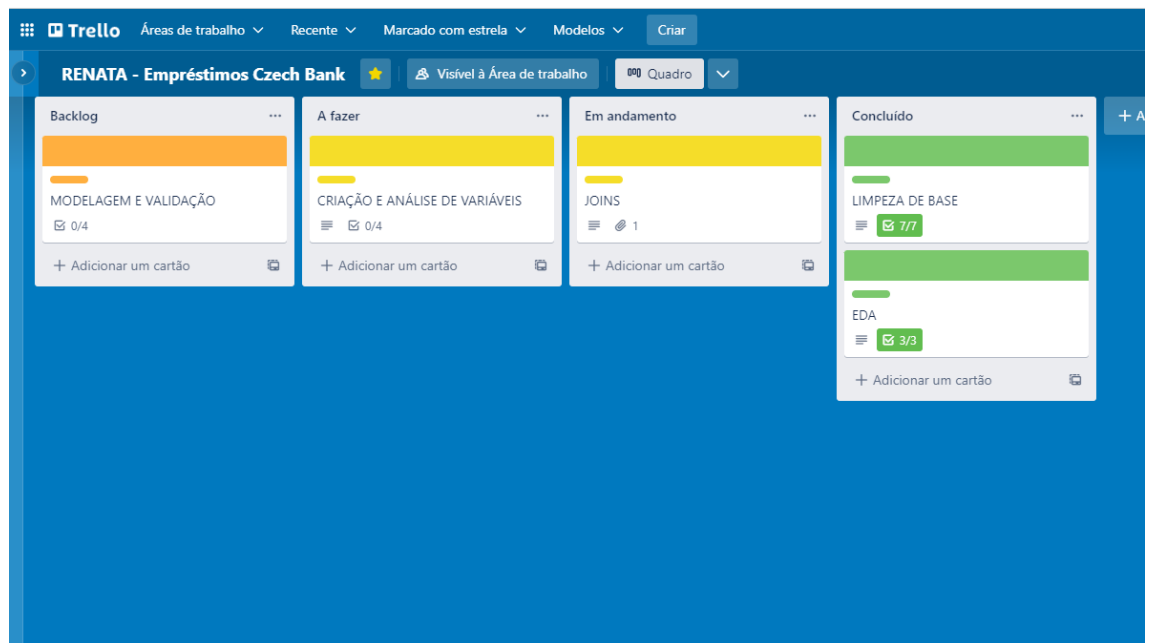
Acões

2. Área de Experimentação

2.1 Sprint 1

2.1.1 Solução

- Evidência do planejamento:



- Evidência da execução de cada requisito:

LIMPEZA DE BASE

A data veio em formato de EPOCH e a unix timestamp não estava resolvendo meu problema. Pois quando aplicada, retorna TODAS as datas iguais. Busquei outra forma de ajustar. A resolução está no notebook 'TRATAMENTO DATAS TRADUÇÕES'.

ANTES:

```
+-----+-----+-----+-----+-----+-----+
|loan_id|account_id|date|amount|duration|payments|status|
+-----+-----+-----+-----+-----+-----+
| 5314|    1787|930705|96396|    12| 8033.00|    B|
| 5316|    1801|930711|165960|    36| 4610.00|    A|
| 6863|    9188|930728|127080|    60| 2118.00|    A|
| 5325|    1843|930803|105804|    36| 2939.00|    A|
| 7240|   11013|930906|274740|    60| 4579.00|    A|
+-----+-----+-----+-----+-----+-----+
only showing top 5 rows
```

DEPOIS:

```
+-----+-----+-----+-----+-----+-----+
|loan_id|account_id|date_loan|amount|duration|payments|status|
+-----+-----+-----+-----+-----+-----+
|5314|1787|1993-07-05|96396|12|8033.0|B|
|5316|1801|1993-07-11|165960|36|4610.0|A|
|6863|9188|1993-07-28|127080|60|2118.0|A|
|5325|1843|1993-08-03|105804|36|2939.0|A|
|7240|11013|1993-09-06|274740|60|4579.0|A|
|6687|8261|1993-09-13|87840|24|3660.0|A|
|7284|11265|1993-09-15|52788|12|4399.0|A|
|6111|5428|1993-09-24|174744|24|7281.0|B|
|7235|10973|1993-10-13|154416|48|3217.0|A|
+-----+-----+-----+-----+-----+-----+
```

AINDA SOBRE LIMPEZA DA BASE:

antes:

```
+-----+-----+-----+-----+
|account_id|district_id|frequency|date|
+-----+-----+-----+-----+
|576|55|POPLATEK MESICNE|930101|
|3818|74|POPLATEK MESICNE|930101|
|704|55|POPLATEK MESICNE|930101|
|2378|16|POPLATEK MESICNE|930101|
|2632|24|POPLATEK MESICNE|930102|
+-----+-----+-----+-----+
only showing top 5 rows
```

As bases vieram com o idioma da República Tcheca, então traduzi para inglês.

depois:

```
+-----+-----+-----+-----+
|account_id|district_id|stmt_frq|date_account|
+-----+-----+-----+-----+
|576       |55         |monthly |1993-01-01 |
|3818      |74         |monthly |1993-01-01 |
|704       |55         |monthly |1993-01-01 |
|2378      |16         |monthly |1993-01-01 |
|2632      |24         |monthly |1993-01-02 |
|1972      |77         |monthly |1993-01-02 |
|1539      |1          |after_tr|1993-01-03 |
|793       |47         |monthly |1993-01-03 |
|2484      |74         |monthly |1993-01-03 |
```

- Evidência dos resultados:

EDA:

Análise univariada para entender como as variáveis de primeiro interesse estão distribuídas:

summary	count	mean	stddev	min	25%	50%	75%	max
account_id	682	5824.162756598241	3283.512680895359	10001	2962.0	5735.0	8688.0	993
district_id	682	37.489736070381234	25.184325712717584	1	13.0	39.0	60.0	9
stmt_frq	682	None	None	after_tr	None	None	None	weekly
date_account	682	None	None	1993-01-13	None	None	None	1997-12-22
loan_id	682	6172.466275659824	682.5792792264359	4959	5576.0	6175.0	6753.0	7308
date_loan	682	None	None	1993-07-05	None	None	None	1998-12-08
amount	682	151410.1759530792	113372.4063095917	100080	66696.0	116832.0	210744.0	99936
duration	682	36.49266862170088	17.075218552321882	12	24.0	36.0	48.0	60
payments	682	4190.6642228739	2215.8303442941256	1000.0	2477.0	3931.0	5814.0	9910.0
status	682	None	None	A	None	None	None	D
days_between	682	398.24046920821115	164.6113589799098	102	261	395	529	697

Aqui irão apenas alguns recortes das análises gráficas que foram feitas comparando o público bom e público mal definido pelo banco.

O gráfico abaixo é apenas uma amostra dos demais que foram feitos para comparar o comportamento do público BOM e MAU.

Ele mostra os dias entre a abertura da conta bancária até a emissão do

empréstimo. Ou seja, graficamente, podemos observar que o público bom demora mais a emitir empréstimos após a abertura da sua conta bancária quando comparado com o público mau.

```
# days_between
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(16, 6))
df_good_pd.days_between.hist(bins=20, ax=ax1, label='good', color='green', alpha=0.6)
df_bad_pd.days_between.hist(bins=20, ax=ax2, label='bad', color='red', alpha=0.6)
ax1.set_title('Days Between Account Creation and Loan Issuance')
ax2.set_title('Days Between Account Creation and Loan Issuance')
ax1.legend()
ax2.legend()
plt.show()
```



2.1.2 Experiências vivenciadas

A experiência vivenciada nessa sprint foi a questão de manipular dados em formato DATETIME. Quando me deparei com as datas em formato de EPOCH e utilizando funções já prontas de bibliotecas - tanto de pandas quanto de Pyspark - para tentar reverter esse quadro e não obtive sucesso, me vi quase sem saída. Para os problemas de concessão de crédito as variáveis de DATETIME são muito importantes para construção de variáveis e análises temporais do comportamento dos usuários.

Percebi que nem todas as soluções estão implementadas em bibliotecas e que muitas vezes devemos pesquisar coisas mais específicas e usar da própria criatividade para resolver alguns problemas.

2.2 Sprint 2

2.2.1 Solução

- Evidência do planejamento:
- Evidência da execução de cada requisito:
- Evidência dos resultados:

2.2.2 Experiências vivenciadas

2.3 Sprint 3

2.3.1 Solução

- Evidência do planejamento:
- Evidência da execução de cada requisito:
- Evidência dos resultados:

2.3.2 Experiências vivenciadas

3. Considerações Finais

3.1 Resultados

Por meio de um texto detalhado, apresente os principais resultados alcançados pelo seu Projeto Aplicado.

Cite os pontos positivos e negativos, as dificuldades enfrentadas e as experiências vivenciadas durante todo o processo.

3.2 Contribuições

Apresente quais foram as contribuições que o seu Projeto Aplicado trouxe para que o Desafio proposto fosse solucionado.

Cite, por exemplo, as inovações, as vantagens sobre os similares, as melhorias alcançadas, entre outros.

3.3 Próximos passos

Descreva quais são os próximos passos que poderão contribuir com o aprimoramento da solução apresentada pelo seu Projeto Aplicado.