**Instituto de Gestão e Tecnologia da Informação**

**Relatório do Projeto Aplicado**

Análise de Fraudes

Carlos Augusto Dias da Costa

Orientador(a): Máiron César Simoes Chaves

31/10/2021

****

**CARLOS AUGUSTO DIAS DA COSTA**

**INSTITUTO DE GESTÃO E TECNOLOGIA DA INFORMAÇÃO**RELATÓRIO DO PROJETO APLICADO

ANÁLISE DE FRAUDES

Relatório de Projeto Aplicado desenvolvido para fins de conclusão do curso MBA Engenheiro de Machine Learning.  
  
Orientador (a): Máiron César Simões Chaves

**Itaboraí  
31/10/2021**

**Sumário**

[1. CANVAS do Projeto Aplicado 5](#_Toc77190733)

[1.1 Desafio 6](#_Toc77190734)

[1.1.1 Análise de Contexto 6](#_Toc77190735)

[1.1.2 Personas 6](#_Toc77190736)

[1.1.3 Benefícios e Justificativas 9](#_Toc77190737)

[1.1.4 Hipóteses 11](#_Toc77190738)

[1.2 Solução 12](#_Toc77190739)

[1.2.1 Objetivo SMART 12](#_Toc77190740)

[1.2.2 Premissas e Restrições 13](#_Toc77190741)

[1.2.3 Backlog de Produto 14](#_Toc77190742)

[2. Área de Experimentação 15](#_Toc77190743)

[2.1 Sprint 1 15](#_Toc77190744)

[2.1.1 Solução 15](#_Toc77190745)

[● Evidência do planejamento: 15](#_Toc77190746)

[● Evidência da execução de cada requisito: 15](#_Toc77190747)

[● Evidência da solução: 16](#_Toc77190748)

[2.1.2 Lições aprendidas 17](#_Toc77190749)

[2.2 Sprint 2 18](#_Toc77190750)

[2.2.1 Solução 18](#_Toc77190751)

[● Evidência do planejamento: 18](#_Toc77190752)

[● Evidência da execução de cada requisito: 18](#_Toc77190753)

[● Evidência da solução: 21](#_Toc77190754)

[2.2.2 Lições aprendidas 22](#_Toc77190755)

[2.3 Sprint 3 23](#_Toc77190756)

[2.3.1 Solução 23](#_Toc77190757)

[● Evidência do planejamento: 23](#_Toc77190758)

[● Evidência da execução de cada requisito: 23](#_Toc77190759)

[● Evidência da solução: 23](#_Toc77190760)

[2.3.2 Lições aprendidas 24](#_Toc77190761)

[3. Considerações Finais 25](#_Toc77190762)

[3.1 Resultados Finais 25](#_Toc77190763)

[3.2 Contribuições 25](#_Toc77190764)

[3.3 Próximos passos 25](#_Toc77190765)

## 1. CANVAS do Projeto Aplicado

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



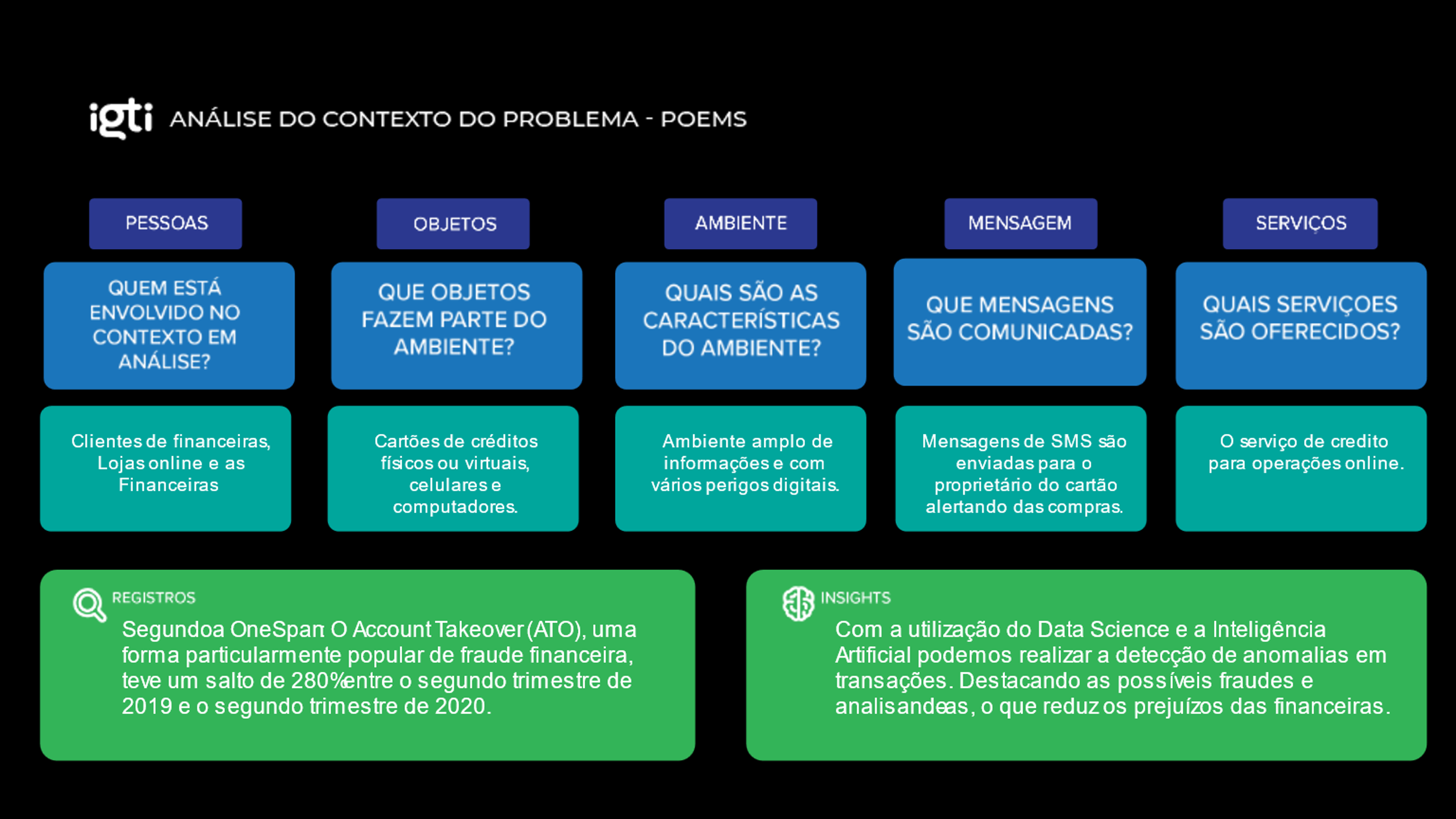
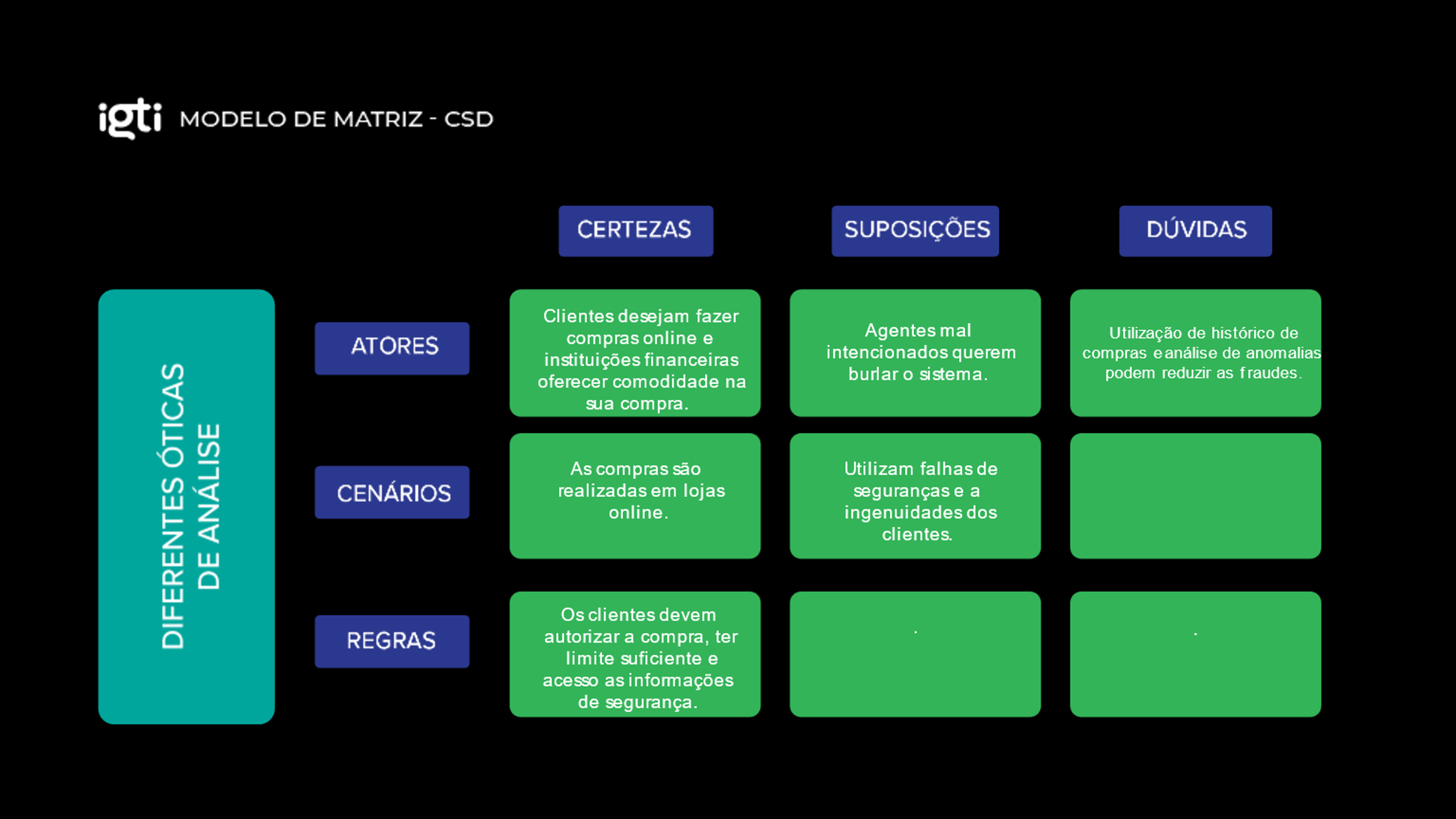
## Desafio

### 1.1.1 Análise de Contexto

As fraudes financeiras existem há anos, entretanto com o aumento da facilidade de compras online, elas veem se diversificando. Com a pandemia do COVID19 e a quarentena, muitos clientes tem usado mais os métodos digitais para efetuar seus pagamentos consequentemente nos últimos anos têm crescido as fraudes, segundo relatório da Mastercard SpendingPulse, um indicador de vendas no varejo, que inclui pagamentos de todos os tipos em uma serie de mercados globais, o e-commerce brasileiro apresentou 75% de crescimento em 2020 se comparado ao ano anterior. Segundo a Onespano Account Takeover (ATO), uma forma particularmente popular de fraude financeira, saltou mais de 280% entre o segundo trimestre de 2019 e o segundo trimestre de 2020.



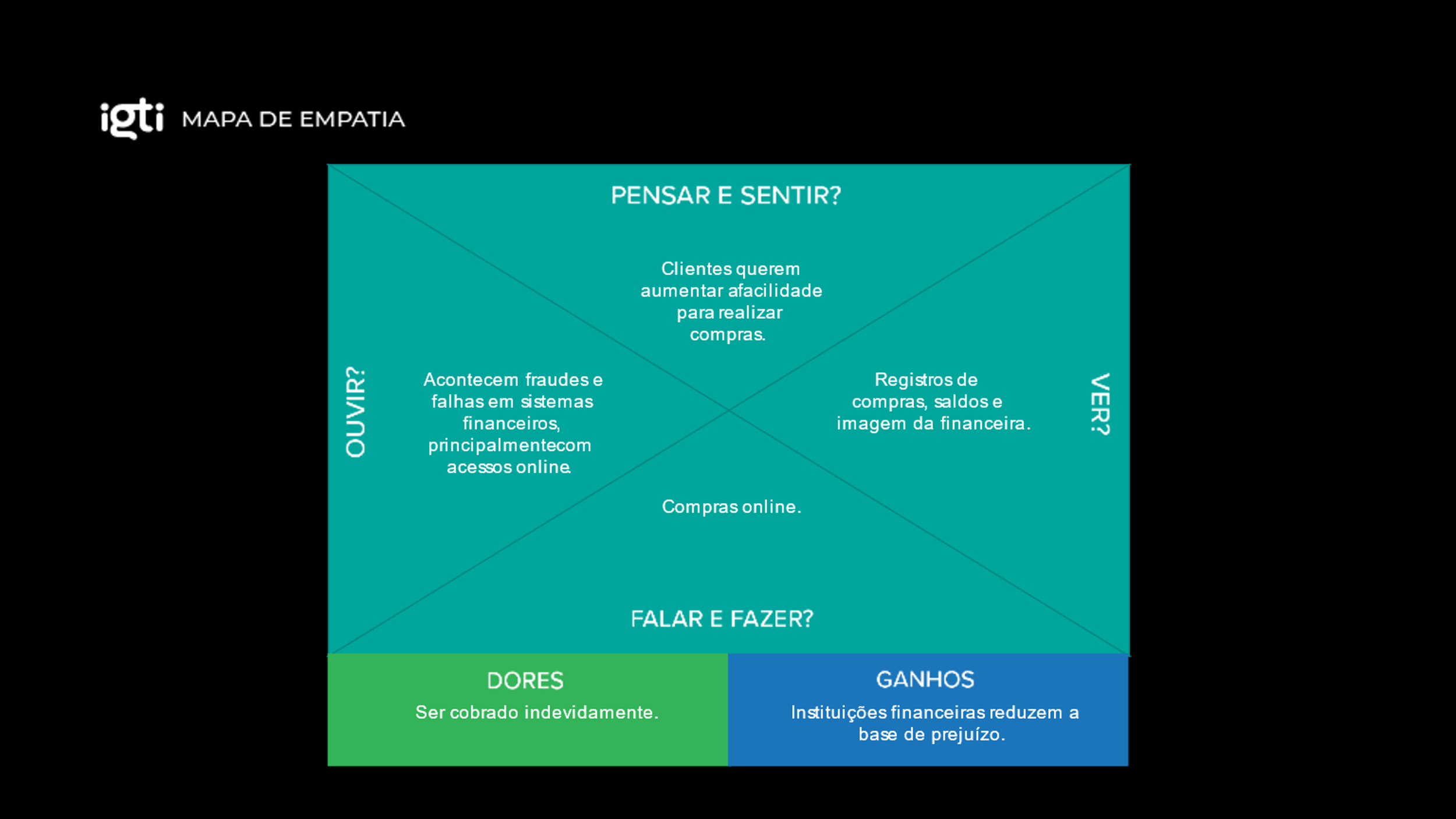
(censo de fraudes)



### 1.1.2 Personas

Claudio tem trinta e cinco anos, é um professor de História que possui duas matrículas no Estado. É extrovertido, gosta da praticidade das compras digitais e não gosta de filas de para pagá-las. Claudio não quer ser cobrado indevidamente e quer ter certeza ele está seguro das fraudes.

Otavio tem cinquenta anos, é dono de uma loja online. É exigente, perfeccionista, entretanto é bem alegre. Otavio quer minimizar qualquer perda tanto de clientes como de dinheiro para os fraudadores. Otavio quer um sistema melhor para mitigar suas perdas e um dia caminhar para um sistema perfeito contra fraudes. Otavio quase perdeu sua loja com ataques de fraudadores, teve um volume muito alto de transações em um único dia e quase teve um prejuízo imensurável. Otavio busca uma solução para não perder o seu negócio e não ter uma dívida na casa dos milhares de reais.

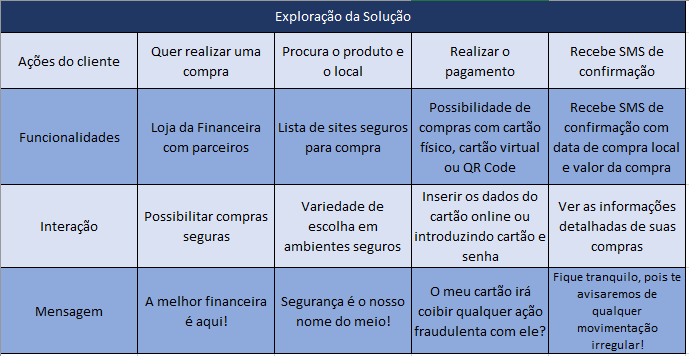
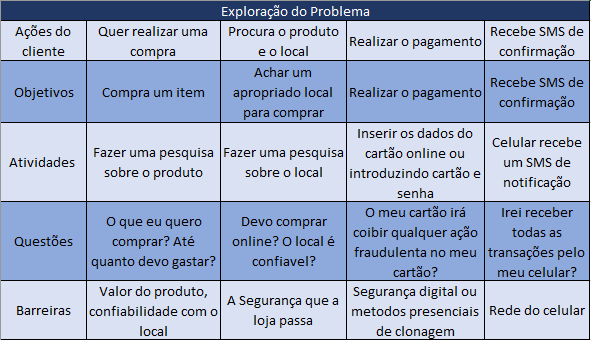
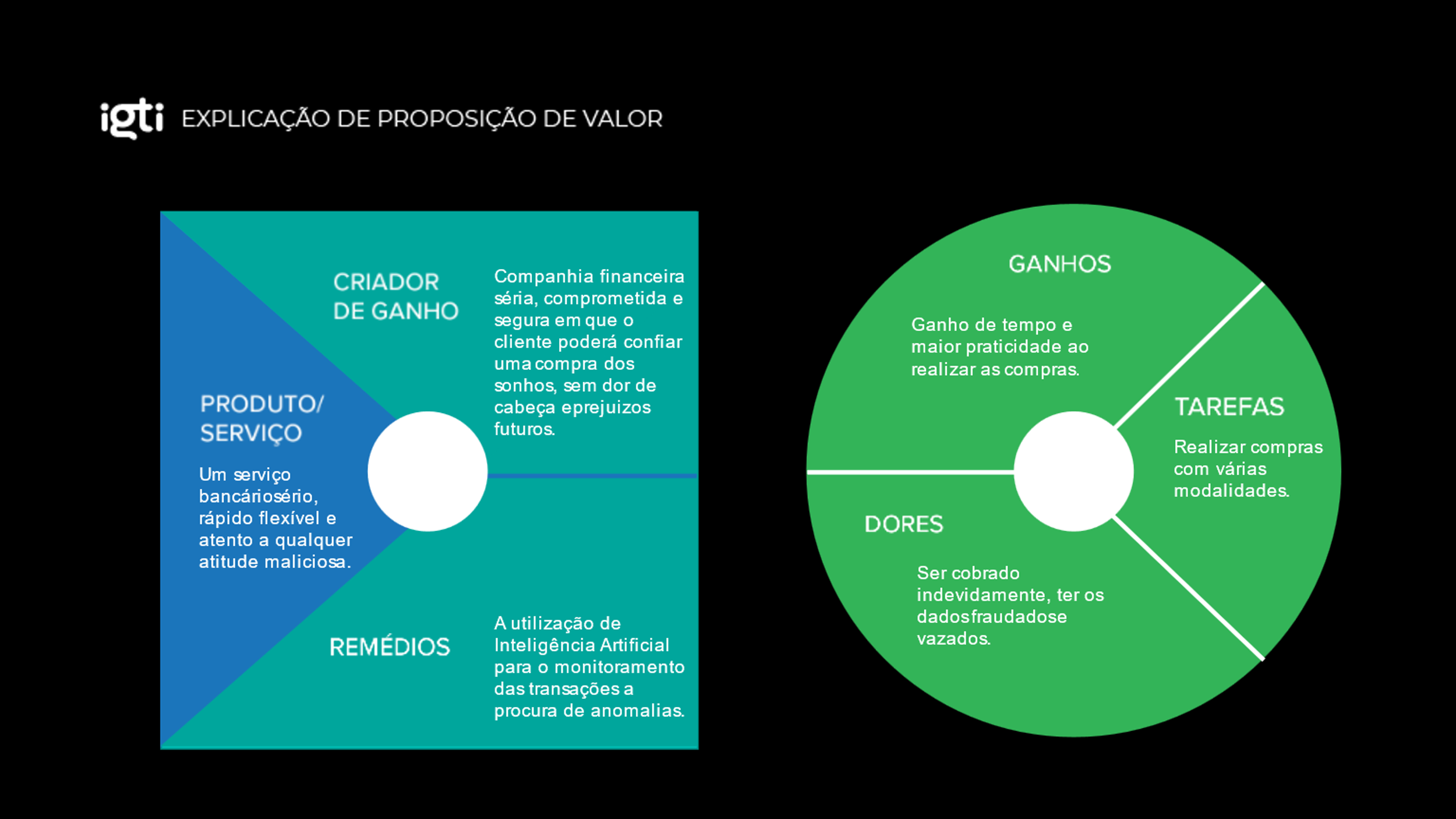


### 1.1.3 Benefícios e Justificativas

Os principais fatores para a explorações e aplicações das detecções de fraudes são, por uma necessidade na demanda tanto nas financeiras como dos clientes. Clientes que não tem cobranças indevidas são mais satisfeitos com o serviço e a financeira que coíbe aumenta sua margem de lucros, tem menos prejuízos. Uma pesquisa da Febraban (Federação Brasileira de Bancos), mostra que 86% dos brasileiros entrevistados têm medo de cair em golpes ou seus dados pessoais sejam violados.

Os benefícios à longo prazo são inúmeros, a cada fraude descoberta há uma mitigação dos efeitos negativos. O sistema quanto sofre um ataque do tipo DoS de milhares de transações ao mesmo tempo, quando é detectada a anomalia poupa a empresa milhares de reais gastos para cobrir essas fraudes. Fraudes menores também tem o poder de causar prejuízo aos donos de comercio e financeiras. A sensação de segurança que leva ao negócio cria uma imagem positiva atraindo mais clientes que tem medo de fraudes. A Visa e Mastercard, têm programas que acompanham empresas que ultrapassam determinado número de contestações por mês, penalizando-as caso os índices se mantenham acima dos limites aceitáveis. O sistema baseado em inteligência artificial tem sua eficiência aperfeiçoada com o tempo o que justifica todos os investimentos iniciais.

O valor atrelado a atividade é grande em comparação aos investimentos iniciais, o que aumenta o custo benefício da atividade. A prevenção de fraudes ajuda a redução da taxa de contestação, evitando maiores complicações para a loja online. As ferramentas utilizadas serão: uma nuvem da AWS, Python, Numpy, Pandas, Matplotlib, Scikit Learn, Tensorflow.



### 1.1.4 Hipóteses

Por que há um aumento de compras online?

Pela maior praticidade e economia de tempo em comparação a métodos tradicionais.

Por que houve um aumento significativo de fraudes nessa era de compras online?

Com os sistemas online aumentaram as formas de aplicações de golpes e de obtenção de dados em sites de fachada ou phishing.

Por que as financeiras precisam cada vez mais da tecnologia para mitigar ou até acabar com as fraudes?

Com o auxílio da tecnologia podemos nos adequar as novas práticas fraudulentas ou até práticas que ainda serão criadas. O auxílio da tecnologia mantém a vigilância alta e fraudes que demorariam dias para serem descobertas podem ser descobertas instantaneamente. Os benefícios são muitos perante aos custos de implementação.

## 1.2 Solução

### 1.2.1 Objetivo SMART

* S (Specific – Específico), montar um modelo de Inteligência Artificial supervisionado.
* M (Mensurable - Mensurável), procurar um dataset de cartões de crédito no Kaggle e trabalhar 1 hora por dia. Utilizar de mensalidade para a manutenção e sustentação do projeto.
* A (Attainable – Atingível), projeto utilizando apenas dados públicos com sombra de vetores dos dados originais.
* R (Relevant – Relevante), com as realidades atuais uma financeira sem detecção de fraudes perderá muito dinheiro e provavelmente fracassará perante a um mercado tão competitivo. Utilização de técnica hibrida oversampling e undersampling.
* T (Time based – Temporal), até o final do curso terminarei meu projeto.

### 1.2.2 Premissas e Restrições

* Todos os dados pessoais são protegidos pela LGPD e assim há etapas de separações dos dados sensíveis e a utilização do PCA para reduzir a dimensionalidade dos dados também como de torná-los anônimos. A empresa sofrerá ações legais caso não cumpra essas determinações.
* As restrições do projeto envolvem a gama de operações financeiras utilizadas de primeiro momento. Apenas será coberto as transações de cartões de domínio público.

|  |  |  |  |
| --- | --- | --- | --- |
| Risco Identificado | Impacto Potencial | Ações Preventivas | Ações Corretivas |
| O modelo que não identifica uma fraude. | Fraude não detectada e sem prevenção. | Usar todo insight dado pelos usuários para melhorar o sistema. | Modelos treinados com realização de manutenção dos pesos. |
| Modelo para identificar uma compra normal por fraude. | Demora na liberação dos pagamentos. | Usar todo insight dado pelos usuários para melhorar o sistema. | Modelos treinados com realização de manutenção dos pesos. |

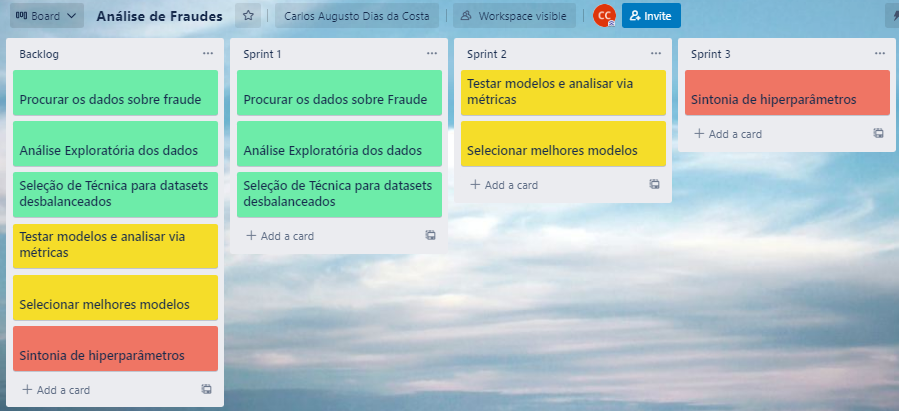
### 1.2.3 Backlog de Produto

O Backlog do produto foi separado em 3 etapas para 3 Sprints. O escopo do produto é montar um modelo Inteligência Artificial para testes de uma possível implementação futura.

Sprint 1 acontecerá a procura do dataset utilizado, como a análise exploratória dos dados para entender a sua natureza e assim finalizando com a escolha de técnica para datasets desbalanceados pois casos de fraudes são extremamente menores que transações normais

Sprint 2 será testado os modelos mais apropriados utilizando métricas para a avaliação, selecionando os melhores modelos para uma melhor sintonia na próxima sprint.

Sprint 3 haverá a sintonia dos Hiperparâmetros e assim a seleção do melhor modelo geral desta forma haverá o deploy com a Inteligência Artificial já treinada e com seus pesos já ajustados para a utilização de um protótipo.



# 2. Área de Experimentação

## 2.1 Sprint 1

### 2.1.1 Solução

#### Evidência do planejamento:

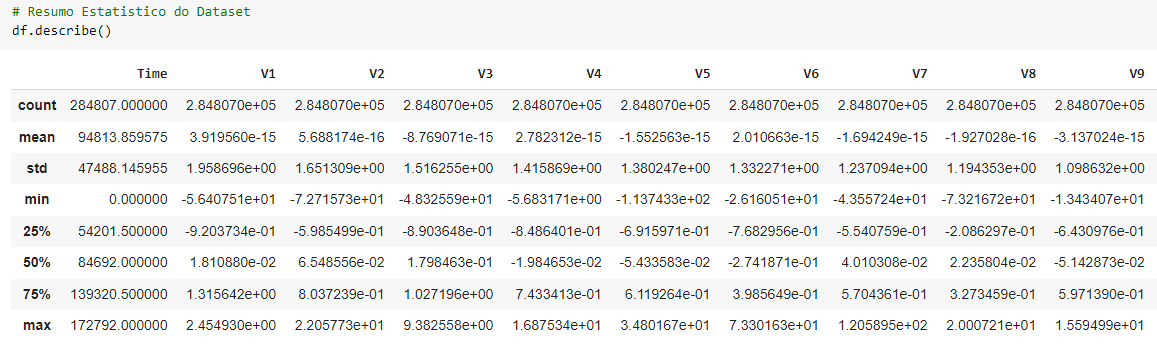


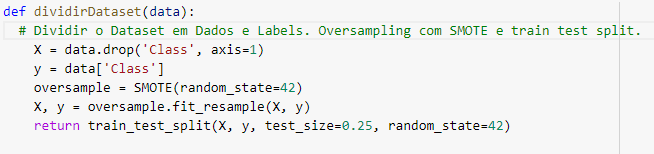
#### Evidência da execução de cada requisito:



Apresentação do dataset utilizado no projeto aplicado.

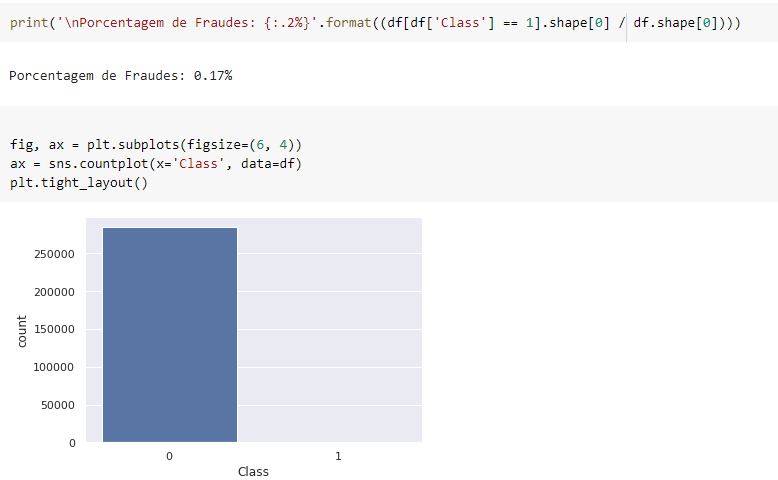
https://github.com/carlospc22/Projects/blob/main/Credit/Final.ipynb



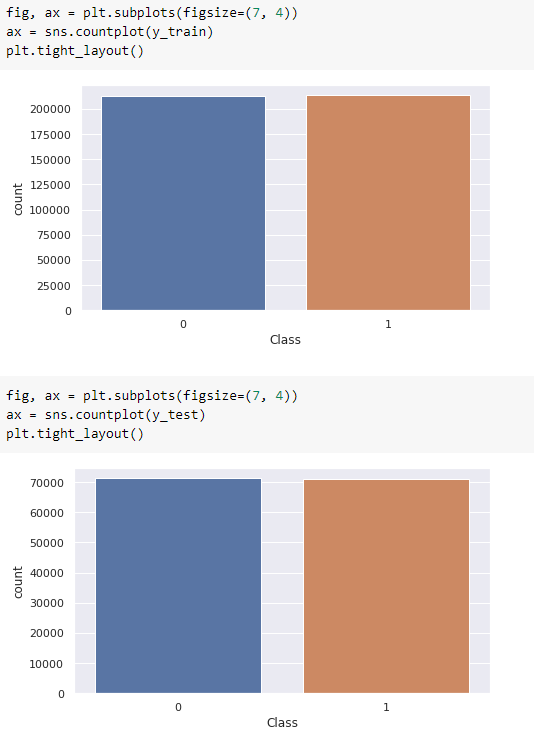


No exemplo acima temos o uso da técnica SMOTE (Tipo de modelo de Oversampling). Como eventos de fraudes são raros se faz necessário o uso de um algoritmo de réplica para aumentar a quantidade de amostras da classe inferior e ter um melhor treinamento. Segue artigo: https://datascience.eu/pt/programacao/smote/

#### Evidência dos resultados:



Evidência o desbalanceamento dos dados utilizados.



Após a aplicação dos algoritmos SMOTE o dataset está balanceado.

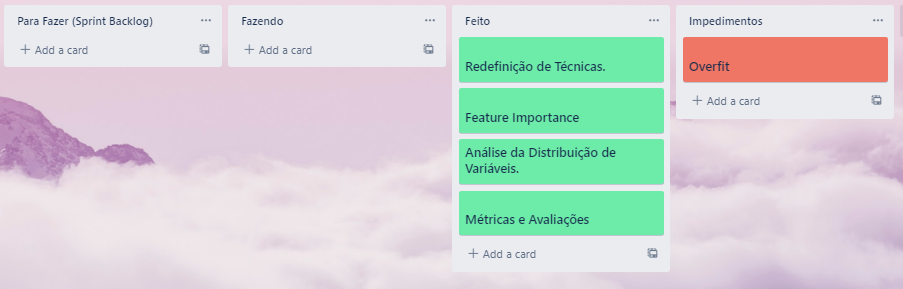
### 2.1.2 Lições aprendidas

As partes de pré-processamento mudam quando se trata de um dataset desbalanceado. Aprendi mais sobre o método de Oversampling e utilizei o SMOTE para isso.

## 2.2 Sprint 2

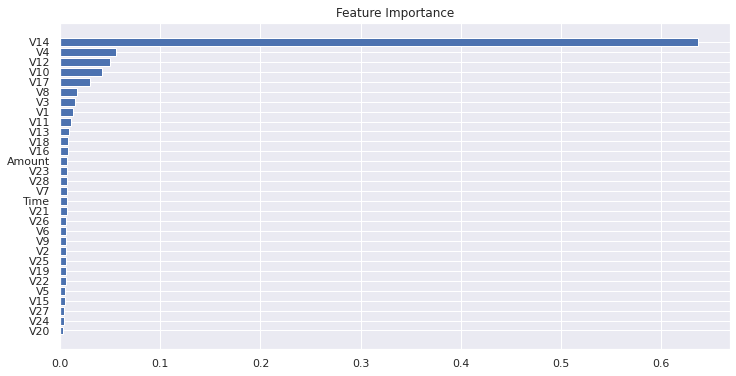
### 2.2.1 Solução

#### Evidência do planejamento:



#### Evidência da execução de cada requisito:

**XGBoost Importância de Variáveis.**

****

A análise de importância de variáveis é um método de alguns modelos de IA (Ensemble ou Árvore). Que ordenam as variáveis a nível de importância para a detecção da classe. As variáveis top 3 foram V14, V4, V12.

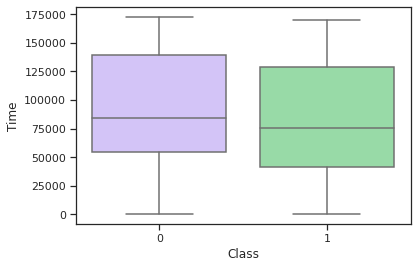
https://medium.com/big-data-blog/import%C3%A2ncia-de-vari%C3%A1veis-o-que-%C3%A9-para-que-serve-63843a35e68e

Métodos ensemble são métodos que unem vários outros modelos para criar um único bom modelo. Métodos do tipo Árvore trabalham com árvore de decisões, ou seja, uma série de regras para determinada avaliação.

Mais informações:

https://www.agatetepe.com.br/aprendizagem-ensemble-basica-floresta-aleatoria-adaboost-intensificacao-de-gradiente-explicacao-passo-a-passo/

**Análise da Distribuição de Variáveis.**

****

A análise da distribuição das variáveis tem como objetivo ver se há alguma variável em que o comportamento é diferençável e como mostrado acima nenhuma das classes apresentou tal diferenciação.

Para mais informações: https://edisciplinas.usp.br/pluginfile.php/3141014/mod\_resource/content/1/T%C3%B3pico\_08\_.pdf

**Métricas de Avaliação Oversampling.**

Em uma classificação binaria temos 2 classes, a classe positiva de maior interesse (Fraude) e a negativa (Legítima). Quando é acertado o valor computado verdadeiro e do contrário negativo.

## **Acurácia (Accuracy/Taxa de Acerto)**

É o número de acertos (positivos) divido pelo número total de exemplos, entretanto em datasets desbalanceados, levam a conclusões precipitadas e por isso usamos outras métricas auxiliares.

Acc= (Verdadeiro Positivo + Verdadeiro Negativo) / Número de elementos

## **F1 Score**

O F1 Score é uma média harmônica entre precisão e recall. É um método que leva em consideração os acertos da classe positiva e os erros das duas classes sendo assim um método bem equilibrado para análises.

F1 = 2 \* (precision \* recall) / (precision + recall)

## **Precisão (Precision)**

Precision = Verdadeiro Positivo / (Verdadeiro Positivo + Falso Positivo)

## **Sensibilidade (Recall)**

Recall = Verdadeiro Positivo/ (Verdadeiro Positivo + Falso Negativo)

**Especificidade (Specificity)**

Especificidade = Verdadeiro Negativo / (Falso Positivo + Verdadeiro Negativo)

Maiores informações

https://diegomariano.com/metricas-de-avaliacao-em-machine-learning/



Gradient Boosting Machine (para regressão e classificação) é um método de conjunto de aprendizagem. A heurística orientadora é que bons resultados preditivos podem ser obtidos por meio de aproximações cada vez mais refinadas. O GBM da H2O constrói sequencialmente árvores de regressão em todos os recursos do conjunto de dados de uma forma totalmente distribuída - cada árvore é construída em paralelo.

Maiores informações:

https://datarisk.io/gradient-boostings-parte-1-metodos-de-ensemble-learning/?cn-reloaded=1

Floresta Aleatória (Random Forest) é um algoritmo de aprendizagem de máquina flexível e fácil de usar que produz excelentes resultados a maioria das vezes, mesmo sem ajuste de hiperparâmetros. É também um dos algoritmos mais utilizados, devido à sua simplicidade e o fato de que pode ser utilizado para tarefas de classificação e também de regressão.

Maiores informações:  
https://medium.com/machina-sapiens/o-algoritmo-da-floresta-aleat%C3%B3ria-3545f6babdf8

O XGBoost Classifier é um algoritmo de aprendizado de máquina escalável e feito para melhorar algoritmos baseados em árvores. Esse algoritmo vem sendo muito utilizado na área para criar soluções estado-da-arte em ciência de dados. O XGBoost é uma implementação do Gradient Boosting Machines criado por Chen.

Maiores informações:

https://1library.org/article/xgboost-classifier-algoritmos-de-aprendizado-de-m%C3%A1quina.yrk5jjvz

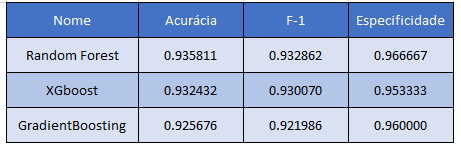
Após os treinamentos usando as técnicas de oversampling foi observado um possível sobre ajuste. A métrica especificidade foi adicionada assim como testes em undersampling e método hibrido (Oversampling e Undersampling).

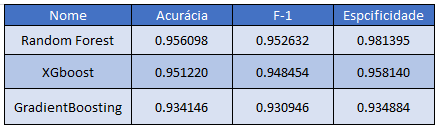
Undersampling é uma técnica de retirada de amostras da classe majoritária.

Segue Artigo: https://blog.letscode.com.br/problemas-de-vies-e-desbalanceamento-de-amostras-na-ciencia-de-dados/

#### Evidência dos resultados:

**Métricas de Avaliação Undersampling.**



**Métricas de Avaliação Modelo Hibrido.**

### 2.2.2 Lições aprendidas

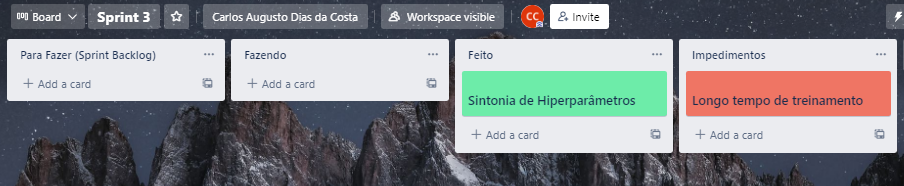
### A utilização da importância de variáveis e da distribuição de variáveis houveram algumas lições recolhidas. As variáveis tem distribuições parecidas havendo bastante interseção. O que mostra a dificuldade de se trabalhar com esse tipo de problema, já a importância de variáveis apontou uma hierarquia parecida em todos os modelos o que aponta que há algumas variáveis críticas para a diferenciação das classes.

### Utilizei modelos híbridos de undersampling e oversampling para um melhor equilíbrio e menor propensão de sobre ajuste.

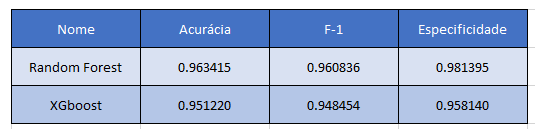
## 2.3 Sprint 3

### 2.3.1 Solução

#### Evidência do planejamento:

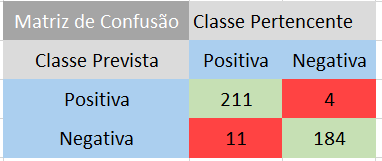


#### Evidência da execução de cada requisito:



Com a sintonia de hiperparâmetros utilizando o método de grades (Testa inúmeras configurações por varredura) o modelo Random forest conseguiu acurácia acima de 96% com F1 de também 96%, o que indica o acerto nas duas classes.

#### Evidência dos resultados:



O modelo Random Forest teve essa matriz de confusão que é uma representação dos acertos e erros na classificação. 11 classes foram etiquetadas como negativas e eram positivas assim como 4 foram etiquetadas positivas e eram negativas.

### 2.3.2 Lições aprendidas

Uma boa sintonia de Hiperparâmetros pode auxiliar a um modelo performar da melhor. E na etapa de deploy (Colocar em funcionamento prático) a linha de corte poderia ser ajustada a um valor inferior 50% para um maior aproveitamento na captura de fraudes.

# 3. Considerações Finais

## 3.1 Resultados Finais

A detecção de fraudes não é uma tarefa fácil, há inúmeros desafios. A quantidade de ocorrências, os seus formatos, padrões e sua constante inovação. Neste estudo foi utilizado um dataset publico do Kaggle em que os dados estão anonimizados e vetores dos dados reais.

Com os modelos estudados para a importância de variáveis e alcançar melhor entendimento da base de dados foi constatado que os vetores V14, V4 e V12 são os mais importantes. Métodos do tipo ensemble ou tipo árvore tem uma melhor performance nesse tipo de classificação.

Os datasets desbalanceados contribuíram em um desafio extremamente proveitoso. Para um dataset tão desbalanceado o melhor método encontrado foi o híbrido oversampling e undersampling.

O Random Forest performou com acurácia acima de 96% após a sintonia de hiperparâmetros.

## 3.2 Contribuições

O modelo proposto em condições parecidas, com dados anonimizados tem total condição de performar bem. Como é um assunto bem complexo que requer cautela e inúmeras estratégias, entretanto os resultados foram bem com uma acurácia e F1 de 96% sendo assim um bom resultado.  Logo o gerenciamento de transações conseguiria pegar várias fraudes.

## 3.3 Próximos passos

Os próximos passos são colocar em deploy e testar com dados externos de formato similar. Testar a performance do modelo com exclusão de variáveis almejando melhores performances.