UNIVERSIDADE FEDERAL DE SANTA CATARINA

CENTRO TECNOLÓGICO
DEPARTAMENTO DE AUTOMAÇÃO E SISTEMAS

Thiago Gouveia Rocha

Implementação e estruturação da área de Business Intelligence em uma PME de Marketing Digital

Florianópolis
2017

Thiago Gouveia Rocha

Implementação e estruturação da área de Business Intelligence em uma PME de Marketing Digital

Relatório submetido à Universidade Federal de Santa Catarina como requisito para a aprovação na disciplina **DAS 5511: Projeto de Fim de Curso** do curso de Graduação em Engenharia de Controle e Automação.

Orientador(a): Prof. Ph.D. José Ricardo Rabelo

Florianópolis
2017

Thiago Gouveia Rocha

Implementação e estruturação da área de Business Intelligence em uma PME de Marketing Digital

Esta monografia foi julgada no contexto da disciplina DAS5511: Projeto de Fim de Curso e aprovada na sua forma final pelo Curso de Engenharia de Controle e Automação.

Florianópolis, 29 de março de 2017

Banca Examinadora:

Eng. Gabriel Guerra Costa
Orientador na Empresa
Resultados Digitais

Prof. Ph.D. José Ricardo Rabelo
Orientador no Curso
Universidade Federal de Santa Catarina

Prof. Miguel Angel Chincaro Bernuy
Avaliador
Universidade Federal de Santa Catarina

Maurício Nunes de Oliveira
Debatedor
Universidade Federal de Santa Catarina

Leandro Hideki Shimanuki
Debatedor
Universidade Federal de Santa Catarina

AGRADECIMENTOS

Gostaria de, primeiramente, agradecer aos meus pais. Eles que sempre me guiaram, desde os primeiros passos, até onde estou hoje. Me ensinaram o valor do trabalho duro, a perseguir os meus sonhos, e que nada vem sem esforço. E ensinaram através do exemplo. Tenho o maior orgulho de ser filho de Marcelo Franco e Rocha e Kelly Cristiny Gouveia Rocha. Dois grandes vencedores na vida que, com bases extremamente fortes de amor, companheirismo, respeito e carinho, formaram a nossa família e tornaram possível esta e todas as minhas outras conquistas.

Gostaria de expressar também a minha profunda gratidão por todo o apoio, incentivo, carinho e amor que minha namorada Laísa Santos sempre dedicou a mim. Sem nunca esgotar as palavras de conforto, de confiança e de orgulho, foi minha grande companheira ao trilhar estas últimas etapas de uma fase tão marcante da vida. E seguirá sendo minha eterna companheira em todos os momentos que ainda estão por vir. Que continuemos a construir esta base sólida para um futuro repleto de conquistas e sonhos, compartilhando cada comemoração, cada vitória. É ao seu lado que saio agora para o mundo, e é para você que olharei de agora em diante antes de encarar cada próximo desafio, sabendo que sempre terei seu apoio. Muito obrigado!

Também gostaria de agradecer aos meus colegas de classe, pela amizade construída, que sempre ultrapassou a sala de aula, e pelos dias, noites e madrugadas de estudo que passamos juntos nas disciplinas mais complicadas, sempre compartilhando as dores e aprendizados trazidos pela graduação.

Devo também a minha gratidão ao grupo de estágio PET-MA do qual fiz parte durante 4 anos da minha graduação contribuindo com valores, aprendizados técnicos, gerenciais e pessoais sendo um fator muito importante para a minha formação profissional. Meus sinceros agradecimentos a todos!

Por fim, mas não menos importante, agradeço à Resultados Digitais pela disponibilização de todos os recursos necessários para desenvolver esse projeto além de me permitir evoluir como profissional gerando aprendizados a cada dai de trabalho. Gostaria de agradecer especialmente ao meu orientador, Gabriel Costa e aos colegas que contribuíram com esse projeto, Rafaela, Pedro, Bruno e Leandro. Obrigado a todos vocês pelo tempo dedicado neste projeto e por estarem sempre dispostos a compartilhar seus conhecimentos e experiências.

RESUMO

Em um contexto de globalização e com um cenário extremamente competitivo no qual as empresas estão inseridas, é fundamental ter dados de qualidade com o propósito de apoiar as decisões tomadas que buscam novas oportunidades ou formas de resolução dos seus problemas.

À medida que o número de softwares utilizado dentro da empresa e o volume de dados crescem, sem uma consistência na sua estrutura de dados e um objetivo claro nas análises, é demandado muito tempo para uma análise suficientemente boa resultando em perdas de oportunidades de negócio. Como solução, são utilizados sistemas de Inteligência de Negócios, os quais utilizam tecnologia e produtos para coletar os dados, transformá-los e fornecerem as informações necessárias para as tomadas de decisão estratégica.

Com a estruturação do fluxo de dados, surgiu a possibilidade de utilizar algoritmos de aprendizados de máquina para os processos de aquisição e retenção de clientes da Resultados Digitais. Com as metas agressivas e com um objetivo de aumentar a base de clientes em aproximadamente 150% no fim do ano, é necessário tanto aumentar a geração de leads [1]quanto a taxa de qualificação dentro do processo para que melhores oportunidades sejam entregues para os vendedores.

Diante disso, este trabalho de PFC trata da estruturação do fluxo de dados da Resultados Digitais, da implementação do sistema de BI para o fornecimento de recursos necessários para que os gestores realizem análises mais eficientes utilizando Dashboards mais confiáveis e, por fim, o trabalho aborda o desenvolvimento do algoritmo de Lead Scoring Preditivo – um algoritmo de aprendizado de máquina com o objetivo de melhorar a eficiência do funil de vendas.

Palavras-chave: Inteligência de negócio. Aprendizado de máquina. Estruturação de dados. Integração de dados.

1 É uma pessoa que tem, de alguma forma, interesse nos produtos ou serviços da sua empresa. No caso da Resultados Digitais, é toda pessoa que baixou algum material da empresa.

ABSTRACT

In a context of globalization and in an extremely competitive scenario where companies operate, it is essential to have quality data to support the decision making process.

As the number of softwares used within a company and the volume of data grows, without a consistency in the data structure and a clear objective in the analyses, it takes a lot of time for an efficient research to be carried out, which results in loss of business opportunities. As a solution, Business Intelligence systems have been developed, using technology and products to gather the data, transform it and provide the necessary information for strategic decision making.

Using a structuring of data flows, we have been able to use a Machine Learning algorithm to optimize the processes of acquisition and retention of clients at Resultados Digitais. Nonetheless, with aggressive goals to increase the customer base by approximately 150% by the end of the year, it was necessary to increase lead generation and the qualification rate within the funnel process in order to deliver better opportunities to the sales team to accomplish the goals.

Therefore, this work was developed with the purpose to structure the data flow and to implement the BI system in order to have managers making better analysis with the necessary resources like a reliable Dashboard. Lastly, this work also describes the development of Predictive Lead Scoring - a machine learning algorithm that aims to improve the efficiency of the sales funnel.

Key-words: Business Intelligence. Machine Learning. Data stream framework. Data integration

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BI – Business Intelligence (Inteligência de negócios)

ML – Machine Learning (Aprendizado de máquina)

CRM – Customer Relationship Management (Gerenciamento de relacionamento com o cliente)

DW – Data Warehouse (Armazém de dados)

FK – Foreign Key (Chave Estrangeira)

JS – JavaScript

PK – Primary Key (Chave Primária)

SQL – Structured Query Language (Linguagem estruturada de pesquisa)

ETL – Extract, Transform and Load (Extração, Transformação e carga)

API – Application Programming Interface (Interface de Programação de Aplicativos)

SUMÁRIO

1 INTRODUÇÃO

Os avanços tecnológicos dos últimos anos fizeram com que os modelos de negócios das empresas de software se transformassem completamente. Nesse contexto foi observado que a maioria destas empresas deixou de oferecer produtos físicos e começou a oferecer os mesmos produtos de uma maneira diferente: de forma digital. Estas empresas, conhecidas como SaaS (Service as a Software), se diferenciam dos modelos tradicionais, pois os seus clientes irão pagar somente pelos recursos utilizados do software, e não por toda a estrutura disponível que, não rara às vezes, não é desfrutada pelo cliente. [1]

Em relação a negócios B2B (*Business to Business*), uma das principais vantagens dessa mudança é a redução do custo fixo destas empresas, visto que elas não necessitam comprar inúmeras licenças de software ou possuir imensos servidores para fazerem com que seus produtos funcionem. Desta forma, elas podem se concentrar no desenvolvimento do seu principal negócio bem como utilizar serviços de outras empresas para prover recursos necessários para o desenvolvimento do seu produto fazendo com que esta evolução seja mais rápida e os custos sejam menores, uma vez que serão proporcionais a utilização do serviço. Com este crescimento e com a diminuição dos custos, será consequentemente gerada uma grande vantagem competitiva no mercado a essas empresas. [2]

Empresas como Netflix, Spotify, Sales Force, Cornerstone OnDemand e Workday são algumas entre tantas que crescem vertiginosamente devido a essa nova forma de estruturação de empresa e modelo de negócio. A sua receita é gerada sobre clientes que pagam mensalmente num modelo de assinatura por um serviço, o qual é apoiado por um produto digital (*software*) e uma equipe de gestão e suporte do mesmo.

Para o desenvolvimento do seu produto, da gestão e suporte para os clientes, a companhia utiliza os serviços de outras empresas onde o seu software também está hospedado na nuvem. Os fornecedores de serviços oferecem grandes quantidades de dados para que os clientes analisem e extraiam o maior valor possível do produto.

Conjuntamente, as empresas necessitam encontrar oportunidades dado as metas agressivas de crescimento e aquisição de mercado. Para algumas decisões mais complexas, são exigidas grandes quantidades de dados provenientes dessas

várias fontes de dados utilizados na empresa. No entanto, sem uma consistência na sua estrutura de dados e em um objetivo claro nas análises e em quais dados deveriam ser averiguados, é demandado muito tempo para uma análise satisfatória para embasar uma decisão, fazendo com que muitos direcionamentos dentro da companhia sejam feitos com base em intuição e por consequência, muitas oportunidades acabam sendo perdidas.

Além dos problemas supracitados, quando ao desenvolver um relatório, muitas vezes os dados não possuem garantia de confiabilidade e integridade devido as múltiplas origens dos dados e por falhas no processo ao alterar ou inserir os dados de forma manual resultando em incerteza sobre a fonte de dados mais confiável. [3]

1.1 Problemática

A Resultados Digitais cresce com taxas superiores a 100% desde a sua fundação em 2012. Em 2013, com o objetivo de melhorar a taxa de eficiência de aquisição de clientes e de diminuir a taxa de *churn*[2]*,* começaram a ser realizadas as primeiras análises sobre o perfil dos clientes.

Na época, a inserção dos dados nas planilhas e as correlações entre as tabelas eram realizadas de forma manual fazendo com que fosse gasto muito tempo com análises não complexas e inviabilizando àquelas com complexidade mais elevadas.

Além disso, esse método não se mostrava escalável ao passo que em 2013 a empresa possuía 300 clientes e agora, em 2017, já possui mais de 6000 clientes. No decorrer de 2014 e 2015 houve melhorias como por exemplo o uso de softwares robustos em cada área da empresa gerando milhares de informações de forma automática.

Porém, essas grandes quantidades de dados armazenados começaram a fazer com que os responsáveis por tomarem as decisões tivessem dificuldade de realizar as análises e se aprofundar no que realmente era importante.

2 Clientes que cancelam a assinatura do produto

Atualmente, apesar de a Resultados Digitais ter informações sobre milhares de *leads*[3], clientes e ex-clientes, estes dados não são muito utilizados para auxílio na tomada de decisão devido a falta de confiabilidade e interoperabilidade entre os sete sistemas utilizados nas áreas da empresa.

Com as informações ficando dispersas, elas se tornam ineficazes, pois faltam elementos que permitam validar sua integridade e fazer análises com maior complexidade.

Além disso, faltam informações de muitos clientes na base de dados da empresa visto que a mesma teve um crescimento acentuado em apenas quatro anos de existência e passou por várias mudanças processuais.

Esses tempos desprendidos em análises ineficientes e com dados não confiáveis são altamente prejudiciais para a empresa tanto do ponto de vista financeiro quanto de perdas de oportunidade e produtividade, principalmente se tratando da Resultados Digitais, que possui metas ambiciosas de aquisição de clientes no mercado brasileiro e latino americano nos próximos anos.

1.2 Objetivo Geral

O presente trabalho tem como objetivo geral fazer com que as tomadas de decisões dentro da empresa sejam realizadas com base em dados confiáveis garantindo uma análise fidedigna do cenário.

Para isso, o trabalho foi dividido em duas principais etapas: a primeira consistiu em implementar toda uma estrutura e fluxo dos dados para centralizá-los em um DW *(Data Warehouse)* e enriquecer as informações faltantes de clientes e ex-clientes de maneira automática.

A centralização das informações traz a facilitação do cruzamento de dados entre fontes diferentes. Além disso, há uma maior consistência na interpretação das informações e melhor correção de erros de dados visto que a definição e criação das métricas serão provenientes de somente um lugar. [4]

3 É uma pessoa que tem, de alguma forma, interesse nos produtos ou serviços da sua empresa. No caso da Resultados Digitais, é toda pessoa que baixou algum material da empresa.

A segunda etapa consistiu em implementar o *Predictive Lead Scoring* para melhorar a eficiência do funil de vendas e implementar uma ferramenta de BI (Business Intelligence) para desenvolver análises melhores e mais complexas.

O *Predictive Lead Scoring* é uma ferramenta de qualificação de leads que auxilia a decidir quais deverão ser abordados por um vendedor. O algoritmo de *Machine Learning* observa quais informações os leads que foram qualificados têm em comum, bem como quais informações os leads que não fecharam a venda têm em comum. A partir dessas observações, o algoritmo decide se um novo lead deverá ser abordado por um vendedor ou não.

Já o uso da ferramenta de BI irá habilitar os responsáveis pela tomada de decisões, de modo que o tempo investido seja gasto nas análises e não na limpeza dos dados ou verificação de integridade do mesmo. Além disso, o cruzamento de dados permitirá que sejam realizadas análises mais complexas seguindo metodologia que será estabelecida para garantir eficiência e o sucesso das análises.

1.3 Objetivos Específicos

- Centralizar os dados provenientes de todos os softwares utilizados dentro da empresa;
- Automatizar o processo de cadastro dos clientes para evitar perdas de informação;
- Modelar um Banco de Dados;
- Criar processo de atualização automática do Banco de Dados;
- Inserir informações faltantes de clientes e ex-clientes via scrapping;
- Desenvolver o processo de ETL (Extração, Transformação e Carga);
- Criar uma Dashboard para a visualização de dados;
- Desenvolver algoritmo de Machine Learning para a qualificação dos Leads;

1.4 Estrutura do Documento

O presente trabalho foi dividido em sete capítulos, de maneira que fossem abordados todos os aspectos teóricos necessários para o desenvolvimento do projeto, bem como as etapas da implementação do mesmo.

O primeiro capítulo trata sobre a introdução do trabalho, o problema a ser solucionado e os objetivos gerais e específicos.

No capítulo 2 será apresentado o local de desenvolvimento do presente trabalho, assim como o seu principal produto e estrutura organizacional.

No capítulo 3 será realizada uma análise bibliográfica, onde serão apresentados todos os conceitos necessários para a compreensão completa do trabalho.

O capítulo seguinte mostra a concepção da centralização dos dados e o enriquecimento das informações faltantes, desde os passos levados à compreensão do problema a ser solucionado até a especificação dos casos de uso, passando pelo levantamento dos requisitos. Logo após, é mostrado o desenvolvimento da solução, desde a expansão dos requisitos à escolha da plataforma de desenvolvimento, passando pela obtenção do modelo conceitual, e projetos das camadas de domínio e persistência.

De forma semelhante ao capítulo 4, no capítulo 5 é apresentada a concepção da implementação da ferramenta de BI e do algoritmo do *Predictive Lead Scoring,* além de todos os procedimentos realizados para a obtenção da solução.

No sexto capítulo discute-se detalhes da implantação e os impactos da utilização do sistema na empresa.

Por fim, no sétimo e último capítulo, será apresentada a conclusão sobre o trabalho e as perspectivas para trabalhos futuros.

2 EMPRESA RESULTADOS DIGITAIS

O presente trabalho foi desenvolvido na Resultados Digitais, uma empresa de tecnologia que desenvolve o RD Station, um software para automação de marketing. Fundada em 2011 por cinco ex-alunos da Universidade Federal de Santa Catarina, dois deles do curso de Engenharia de Controle e Automação.

Atualmente, a empresa é líder nacional no mercado de marketing digital possuindo cerca de 400 colaboradores e mais de 6000 clientes.

A empresa oferece uma solução SaaS para Marketing Digital, o RD Station, que permite que o cliente gerencie em uma única ferramenta todas as funcionalidades cruciais para desenvolver a melhor estratégia de marketing digital para aumentar os resultados de vendas de forma consistente.

Outrossim, ao assinar o serviço os clientes contam com o auxílio de gestores para a utilização de uma metodologia própria da Resultados Digitais, além de serviços de consultoria complementares com o objetivo de gerar mais resultados em tráfego, gerar mais *leads* e vendas para seus negócios, bem como construir um sólido ativo de Marketing Digital. Essas soluções são focadas em *Inbound Marketing*, método do qual a empresa utiliza desde a sua fundação comprovando a eficiência do método.

2.1 Inbound Marketing

O *Inbound Marketing* é um conjunto de estratégias que têm como objetivo atrair voluntariamente os consumidores para o site da empresa. É o contrário do marketing tradicional, e baseia-se no relacionamento com o consumidor ao invés de propagandas e interrupções. A principal diferença entre o marketing tradicional – que chamamos de *Outbound Marketing* – e o *Inbound* é que, no segundo, quem procura a empresa é o cliente e não o contrário. [5]

O termo começou a ser popularizado em 2009 [6] nos Estados Unidos, a metodologia segue uma sequência lógica de 5 etapas: Atrair, Converter, Relacionar, Vender e Analisar. A Figura 1 demonstra onde essas etapas se relacionam com o funil de vendas.

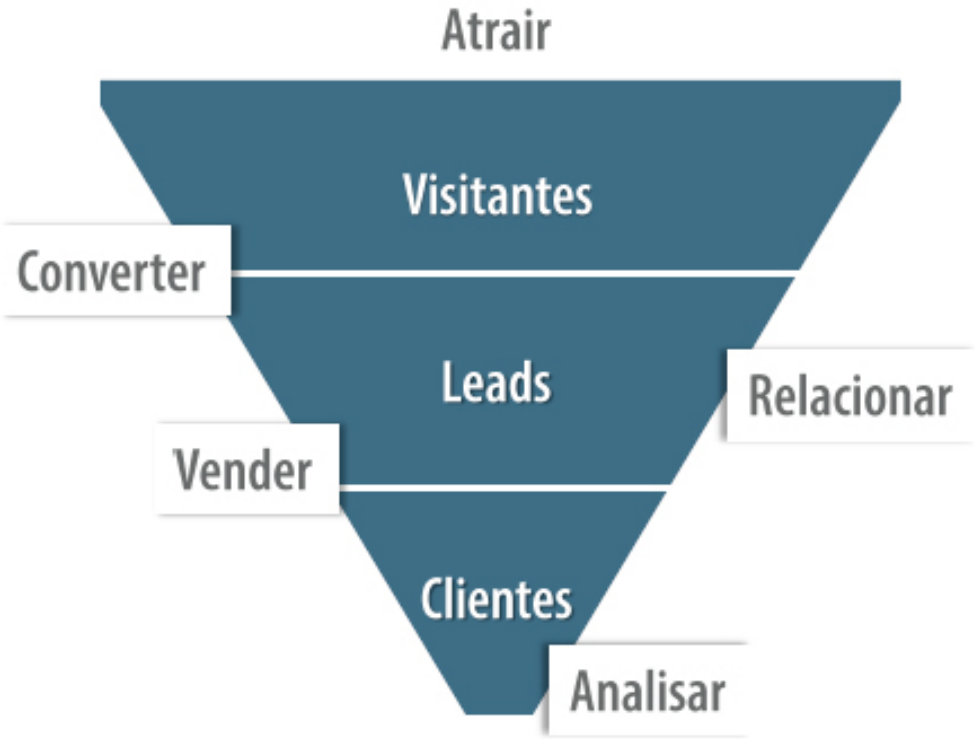

Figura 1 - Esquema demonstrativo da relação entre o Inbound Marketing e o Funil de Vendas

A grande vantagem do Inbound Marketing é ser mais barato (62%) que o Marketing convencional – ou Outbound Marketing – devido ao fato de você se relacionar com o público de forma customizada, otimizando a sua estratégia a partir da geração de inúmeros dados que as campanhas de marketing geram.

A partir do momento que o site da empresa começa a gerar milhares de Leads todo mês através de formulários via *Landing Pages*[4], há uma dificuldade de se relacionar com todos os *leads* para mostrar de forma eficaz a proposta de valor do produto. Além disso, se torna inviável abordar todos os leads pelo time de vendas, sendo necessário um processo decisório que planeje estrategicamente quais pessoas estão prontas para comprar o produto oferecido. [7]

A Resultados Digitais criou o RD Station com o objetivo de ser uma solução completa para a gestão e automação de marketing e vendas. Através dessa

4 São páginas criadas com foco na conversão. Ou seja, a pessoa se torna um Lead preenchendo um formulário e recebe um material em troca.

plataforma, empresas conseguem realizar todas as ações de Marketing Digital para atrair visitantes, gerar *leads*, relacionar-se com os *leads*, fazer o processo decisório de quais deverão ser abordados por vendedores, além de conseguir centralizar todos os dados para a otimização do processo, garantindo um menor custo e uma maior eficiência das ações de marketing.

2.2 Estrutura organizacional e área de atuação

Na Resultados Digitais, abaixo do CEO existem 6 diretorias: Inbound Sales, Partners, Talent Management, Marketing, Financeiro e Customer Success.

Além disso, dentro de cada diretoria existem diversas áreas coordenadas por gestores. O presente trabalho foi realizado dentro de uma célula da empresa composta por um time de 5 pessoas sendo que somente o gestor e o autor deste trabalho estavam totalmente alocados no desenvolvimento da área.

A célula tem como objetivo encontrar a solução ideal de Business Intelligence para que sejam utilizados recursos humanos e financeiros de maneira coerente. Em uma visão a longo prazo, o objetivo da área de BI é estar posicionada junto às outras diretorias tendo como propósito principal fornecer a infraestrutura necessária para a análise de dados da empresa, realizar análises de negócio e treinar/prestar consultoria para os analistas responsáveis em cada área.

O fluxograma da Figura 2 mostra de forma simplificada a estrutura organizacional da empresa, destacando o local de atuação do presente trabalho.

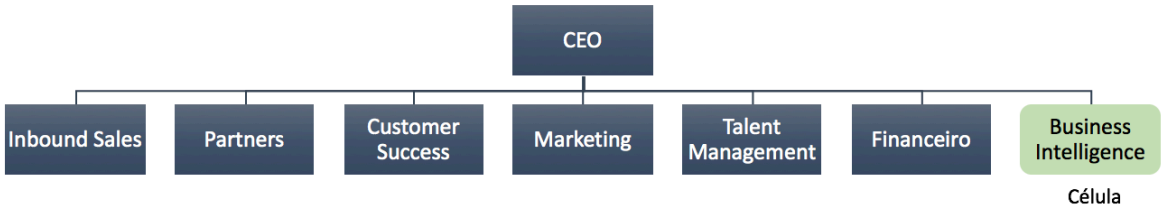

Figura 2 - Estrutura Organizacional da empresa

3 FUNDAMENTAÇÃO TEÓRICA

Primeiramente, para uma melhor compreensão de como foi desenvolvida a estrutura do fluxo de dados e as soluções de BI e *Machine Learning* da Resultados Digitais, é necessário apresentar os conceitos e teorias gerais usados no desenvolvimento do presente trabalho.

Este capítulo apresenta de forma sucinta os conceitos e ferramentas mais importantes em relação a Software as a Service, Plataforma de Integração SaaS, Banco de dados, *Business Intelligence e Machine Learning.*

3.1 Software as a Service

As transformações no mercado causadas pela expansão da Internet no mundo e até mesmo o conceito de Computação em Nuvem no qual é possível armazenar os dados e softwares em servidores de terceiros, tem permitido colocar em evidência a crescente tendência de adoção do modelo SaaS – Software as a Service [8].

Com modelo SaaS, o método de distribuição e comercialização de software é realizado remotamente. Ou seja, o cliente paga pelo serviço oferecido – sem a necessidade de adquirir licenças – e pode acessar o programa ou aplicativo de qualquer computador em qualquer lugar, através da internet.

Atualmente, as empresas de tecnologia, como a Resultados Digitais, além de comercializar um produto SaaS, também utilizam internamente estes modelos de softwares. Dessa forma, a grande maioria do armazenamento, processamento e *deploy* do software é realizado na nuvem evitando que a empresa tenha altos custos com servidores, espaço para armazená-los e refrigerá-los por exemplo.

Além disso, o modelo SaaS apresenta 6 principais benefícios para os negócios:

- Baixo custo de implementação;
- Acessível de qualquer lugar onde haja internet;
- Escalável – possível adaptar o plano assinado de acordo com que o negócio evolua com o produto;
- Acordo de nível de serviço dos líderes do setor garantindo boa performance e alto tempo disponível para acesso;
- Alta segurança devido à natureza compartilhada do serviço.

Em relação a integração de dados, a vantagem deste tipo de modelo é a facilidade de encontrar APIs – seja desenvolvida pela própria empresa SaaS ou por terceiros – ou até mesmo desenvolver internamente uma API para realizar as ações desejadas por meio de documentações que a empresa disponibiliza na internet.

API é um conjunto de rotinas e padrões de programação para acesso a um aplicativo de software ou plataforma baseado na Web. Ela facilita o desenvolvimento de um programa de computador, ao abstrair a implementação e fornecer as especificações para rotinas, as estruturas de dados, as classes de objeto, as variáveis e chamadas remotas que o desenvolvedor necessita.

No contexto de integração dos dados, a API facilita a conexão e comunicação entre os softwares permitindo a transferência de dados de forma confiável.

Nesse projeto, foi desenvolvido uma plataforma de integração SaaS para realizar requisições de dados aos softwares a cada intervalo de tempo definido e centralizar todos os dados de uso interno em um *Data Warehouse*.

3.2 Plataforma de Integração SaaS

Cada aplicação utilizada dentro de uma empresa pode gerar milhões de informações sobre os seus clientes e leads. A probabilidade de gerar insights significativos potencializa quando é possível cruzar todas essas informações.

No entanto, para a integração dos sistemas é necessário que haja interoperabilidade entre eles tendo processos de negócio bem sincronizados e padrões técnicos compatíveis entre si. O presente trabalho utilizada uma abordagem top-down com o foco em resolver o problema de interoperabilidade atual.

Além disso, apresenta-se 4 principais tipos de abordagem para resolver o problema de integração de sistemas:

- Transferência de Arquivos entre softwares - sistemas devem conhecer o nome do arquivo e local a ser lido ou registrado.
- Banco de Dados Compartilhado
- Invocação remota de procedimentos – sistemas compartilham funcionalidades para executar uma ação
- Baseado em Mensagens

Na maioria dos casos atuais, a integração via troca de mensagens resolve com grande eficiência a grande maioria dos problemas de integração podendo ser

aplicados a vários tipos de arquiteturas que basicamente são os principais padrões de sucesso para integração.

Antes da popularização da computação em nuvem, a boa prática de comunicação entre aplicações era a utilização da arquitetura de SOA (Arquitetura Orientada a Serviços) e ESB (Barramento de Serviços Corporativos) como um canal de comunicação.

Essa arquitetura tem como principais finalidades disponibilizar maior flexibilidade para mudanças, suportar serviços independentes de plataforma e protocolos e, principalmente, integrar as aplicações.

Os serviços são módulos de negócio que são invocadas via mensagens. Nessa arquitetura é tratado os requisitos de baixo acoplamento, computação distribuída independente de protocolo, sistemas legados, integração de aplicações e desenvolvimento baseado em padrões.

Um dos componentes mais importante em SOA é o ESB que oferece as funcionalidades para implementá-la. O barramento provê uma camada de abstração acima de um sistema de mensageria que permite a integração entre os aplicativos.

O ESB, ilustrado na Figura 3, é baseado em padrões de integração fornecendo uma infraestrutura para complexas arquiteturas de integração. Essa arquitetura funciona é controlada por eventos que disparam mensagens entre si através de um barramento logicamente comum a todos.

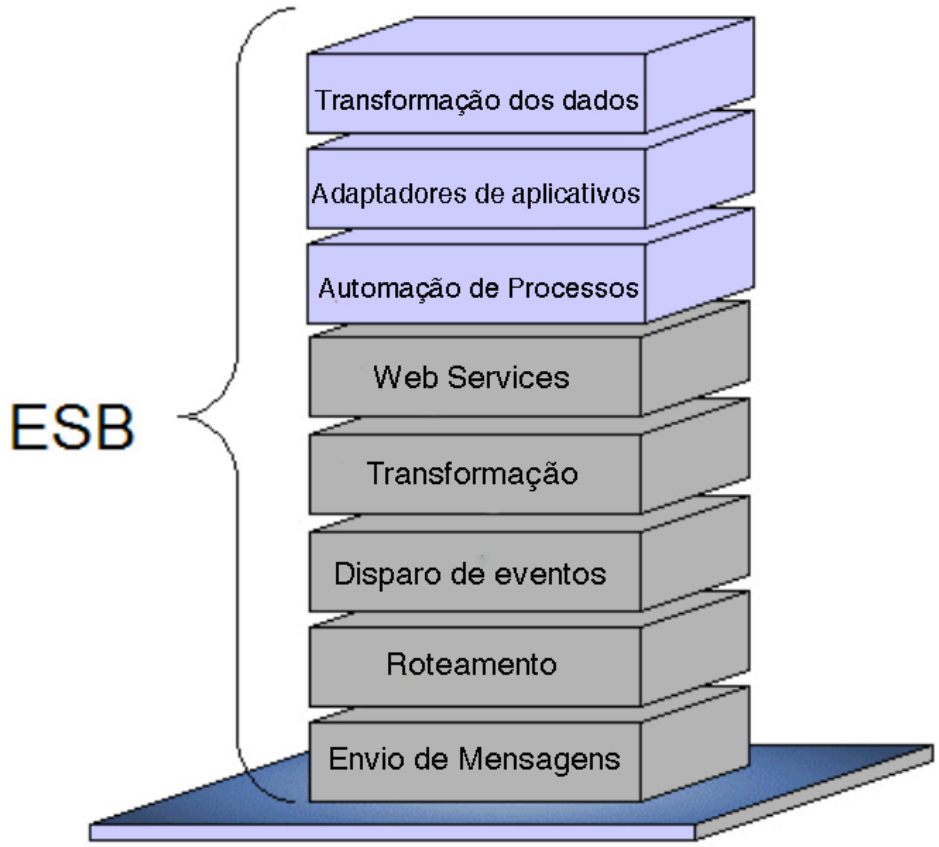

Figura 3 - Overview das funcionalidades de um ESB

Esses tipos de arquitetura tem o propósito de ter uma forma clara e escalável de integração das informações entre os softwares para que eles não se comuniquem diretamente entre si tornando difícil o gerenciamento do fluxo de informações.

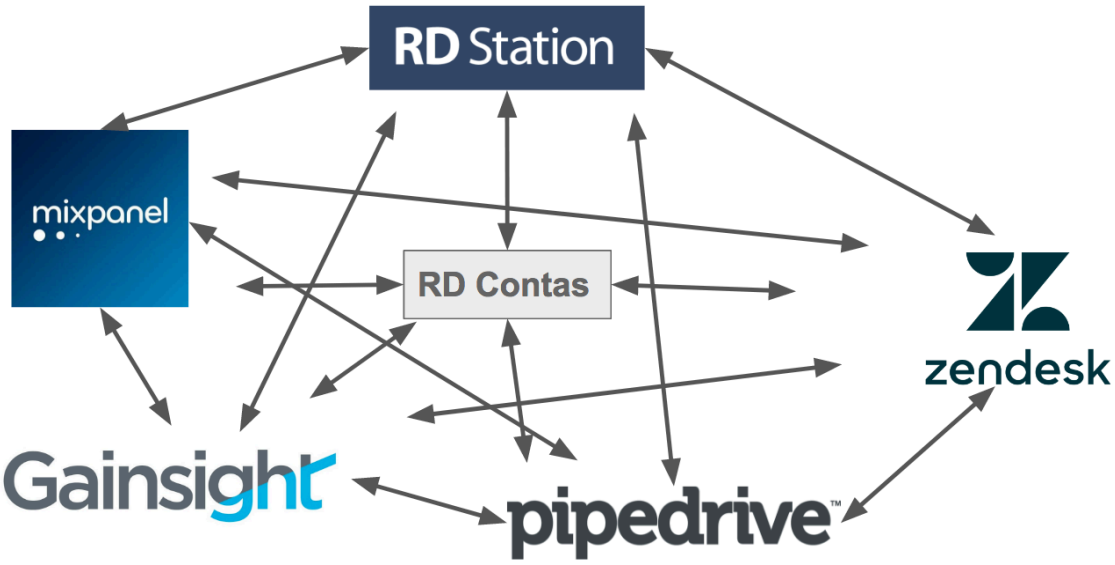

Figura 4 – Exemplo de como seria o fluxo de informações se os softwares na Resultados Digitais se comunicassem diretamente entre si

Ao longo dos anos, as empresas inseriram cada vez mais lógica de negócios no ESB como uma alta carga de sincronismo, roteamentos, transformações, cômputos intermediários e agregações. Isso fez com que o processo ficasse mais complexo e com mais problemas de desempenho, especialmente ao enviar com grandes volumes de dados através do barramento.

O ESB tradicional tem dificuldades em suprir as necessidades da natureza e volume dos dados modernos, pois se baseia principalmente em SOAP (Protocolo Simples de Acesso a Objetos) e XML (Linguagem Extensível de Marcação Genérica) para a representação de dados e eles utilizam uma grande quantidade de *tags* em torno dos dados.

Após o advento da computação em nuvem e com o uso de softwares SaaS, surgiram muitas soluções de plataforma de integração SaaS. O iPaaS, como são conhecidos, fornece uma plataforma com o propósito de facilitar a implementação de soluções de integração de aplicações, podendo essas aplicações serem locais (on-premisse), ou da nuvem.

Este serviço oferece uma plataforma para: modelar, executar, monitorar e gerenciar toda a solução de integração. Esta é a principal vantagem para as empresas que contratam este serviço, pois podem se concentrar em atividades como: criação de novas regras de negócio, projeção de fluxos de trabalho, dentre outras atividades.

Além disso, o iPaaS usa APIs (Interface de Programação de Aplicação) utilizando o protocolo REST (Transferência de Estado Representacional) com os dados codificados através do JSON para suprir ter uma boa performance mesmo com uma alta complexidade na lógica de negócios e com o grande volume de dados.

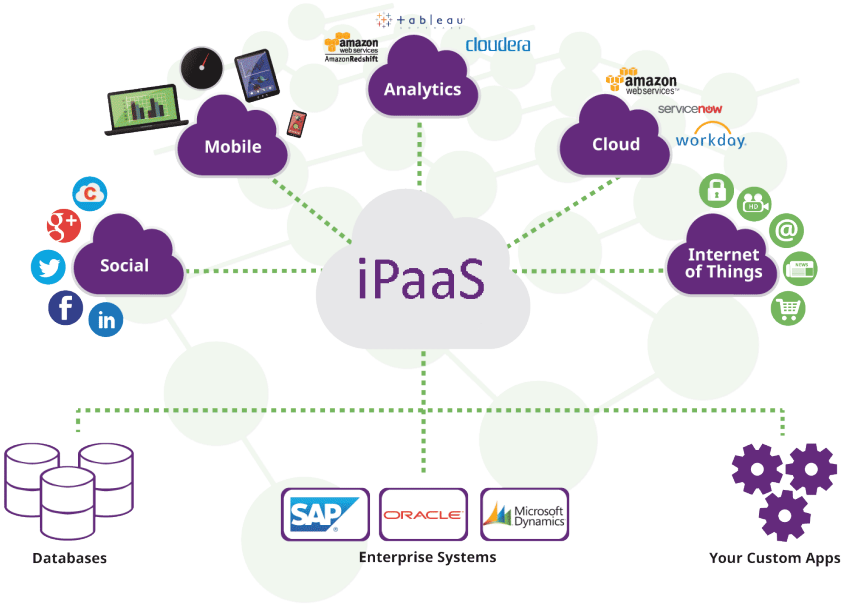

Figura 5 - Exemplo simplificado do funcionamento do Integration Plataform as a Service (iPaaS)[5]

Apesar de haver diversas soluções no mercado, para o presente trabalho, foi desenvolvido uma plataforma de integração em serviços de forma simplificada com o objetivo de suprir as necessidades atuais da empresa. Todos os dados serão centralizados em um único banco de dados também conhecido como *Data Warehouse.*

3.3 Banco de dados

Segundo Korth [9], um banco de dados “é uma coleção de dados inter-relacionados, representando informações sobre um domínio específico”, ou seja,

[5] Imagem adaptada e extraída do site: http://www.snaplogic.com/ipaas-integration-platform-as-a-service

sempre que for possível agrupar informações que se relacionam e tratam de um mesmo assunto, pode-se dizer que tem um banco de dados. Desde 1970, os bancos de dados relacionais se tornaram dominantes na grande maioria das aplicações comerciais, no entanto, com o crescimento do volume de dados houve o surgimento de modelos alternativos como o NoSQL.

3.3.1 Banco de dados relacional

A linguagem mais popular para se trabalhar com banco de dados relacional é conhecida como SQL (Linguagem de Consulta Estruturada) criada pela IBM em 1974.

Um banco de dados relacional organiza os dados em tabelas ou relações. Estas tabelas são compostas por linhas (registros) e colunas (atributos).

Para garantir a integridade dos dados, são utilizadas chaves primárias para identificar a unicidade dos registros e chaves estrangeiras que identificam como uma tabela se relaciona com a outra.

As chaves também são muito utilizadas em banco de dados relacionais para a criação de índices, com o objetivo de melhorar o desempenho de consultas no banco.

Por toda a facilidade de uso e simplicidade de expressão, o SQL se transformou na linguagem de consulta de dados mais usada no mundo, ajudando a consolidar o modelo relacional de banco de dados na maioria das aplicações atuais.

No entanto, a sua complexidade estrutural fez com que surgissem problemas, principalmente relacionados ao crescente volume de dados, muitas vezes não estruturados, que as empresas necessitam armazenar atualmente. Para resolver esse problema, nasceram os bancos de dados não relacionais.

No presente trabalho, todos os dados foram centralizados em um banco de dados relacional.

3.3.2 Banco de dados não relacional

Com a primeira versão sendo criada em 1998, os bancos de dados não relacionais, conhecidos como NoSQL, tem o principal objetivo de resolver o problema de escalabilidade dos bancos tradicionais.

A principal diferença entre os dois tipos de banco de dados é que em NoSQL não existe um esquema forte. A sua principal característica é possuir um armazenamento por chave-valor (*key-value stores*).

No presente trabalho, foi utilizado o MongoDB onde é trazido o conceito de Banco de Dados Orientado a Documentos. Eles têm como característica conter todas as informações importantes em um único documento, ser livre de esquemas, possuir identificadores únicos universais (UUID), possibilitar a consulta de documentos através de métodos avançados de agrupamento e filtragem conhecido como MapReduce e também permitir redundância e inconsistência. [10]

Figura 6 - Exemplo de estrutura em um documento no MongoDB

Os bancos de dados NoSQL apresentam vantagens sobre os bancos tradicionais quando precisamos de escalabilidade, flexibilidade, manipulação de quantidade massiva de dados, bom desempenho e facilidade para consultas.

O MongoDB possui consultas bastantes simples de serem realizadas, visto que não existem transações e *joins*. As consultas são mais fáceis de escrever e de ajustar. Neste projeto foram armazenadas todas as informações dos leads da empresa no MongoDB devido ao alto volume de dados e a grande quantidade de atualizações que são realizadas em cima de um mesmo lead.

Assim, os dados de marketing e vendas armazenados são primeiramente utilizados para três principais operações:

- Enriquecer dados dos leads
- Realizar o processo de qualificação de leads conhecido como *Lead Scoring*
- Enviar os leads qualificados para o software de vendas

A plataforma de integração de serviços extrai as informações de marketing e vendas diretamente do MongoDB e insere no DW para o desenvolvimento das soluções de BI e ML.

3.4 Business Intelligence

No cenário extremamente competitivo no qual estas firmas estão inseridas, é fundamental que as empresas tenham de forma rápida e eficiente acesso às informações para tomadas corretas de decisões. Para auxiliar nessa necessidade, são utilizados sistemas de *Business Intelligence* com 4 principais propósitos:

- Fazer com que os analistas utilizem mais tempo analisando gráficos a construindo-os;
- Acessar informações de modo rápido e fácil até para pessoas não técnicas;
- Ter decisões relevantes corretas;
- Rápido retorno sobre o investimento dado que as tomadas de decisões assertivas propiciam à companhia uma vantagem competitiva no mercado.

Segundo a A.T. Kearney [11], pode-se definir Business Intelligence como “campo particular de processamento de dados e consolidação para recuperar informações para tomada de decisão. O objetivo geral é fornecer - através de várias soluções - o conhecimento certo para as pessoas certas no momento certo”.

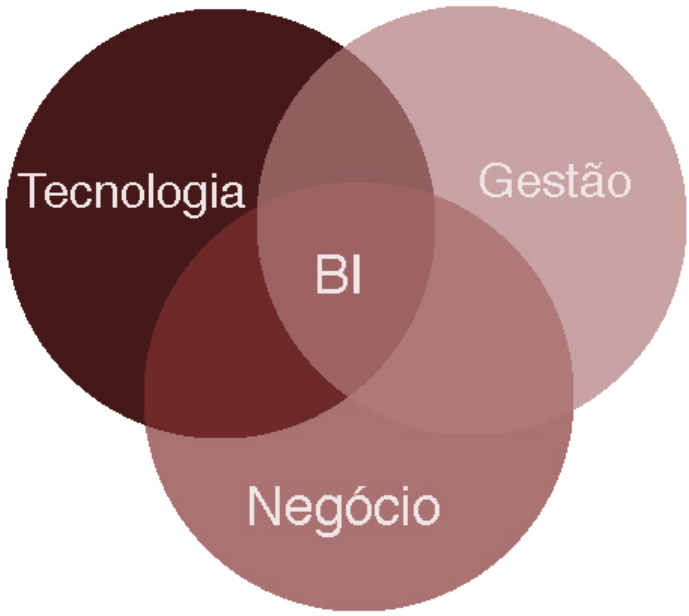

Figura 7 - Interssecções das principais entidades que compoem o Business Intelligence

Para que sejam gerados esses resultados, é necessária uma combinação de sistemas de TI, arquiteturas, estruturas de dados, processos de coleta de dados e planejamento para fornecer informações relevantes.

Os sistemas de BI centralizam todos os dados de software utilizados dentro da empresa em um *Data Warehouse* (DW) para ser a única fonte de dados. A principal vantagem de ter todos os dados centralizados é o auxílio na organização e na consistência dos dados, fazendo com que as análises sejam feitas sempre com as informações atualizadas e que as áreas se comuniquem de maneira mais eficiente retirando informações de uma única fonte de dados [12].

Com os dados armazenados no DW, é utilizada uma ferramenta de modelagem dos dados para a criação de métricas, colunas calculadas e extração somente dos dados necessários para análise ou desenvolvimento de *dashboards*.

A modelagem dos dados é feita pelo método tabular e as regras de pesquisa para extração dos dados são realizadas em linguagem SQL implicando em uma facilidade de implementação junto a uma boa performance comparando a modelagem pelo método de cubo. Ou seja, essa é a fase de consolidação dos dados da origem, com processos de transformação e tratamento de acordo com as características

intrínsecas, que o tornam consistente, estável, centralizado e à disposição das necessidades informacionais da organização.

Especificadamente, a ferramenta de modelagem permite o usuário clusterizar e agrupar dados, transformar campos de dados de TimeStamp para formato de data, corrigir erros de processo, remover linhas duplicadas e realizar cálculos mais complexos que sejam necessários para a análise. Grande parte dos cálculos realizados nessa etapa são a partir das funções de somar e contar acoplados a filtros e, em alguns casos, às operações matemáticas básicas

Dessa forma, a pesquisa realizada pela ferramenta de visualização de dados é mais eficiente, pois todos os pré-cálculos, estratégias de indexação e extração dos dados necessários já foram previamente realizados.

Por fim, o sistema de BI, através de uma ferramenta de visualização de dados, requisita as informações necessárias para a ferramenta de modelagem de dados. A ferramenta de visualização de dados requisita os dadosAs informações serão exibidas de forma clara e intuitiva através de painéis conhecidos como *Dashboards.*

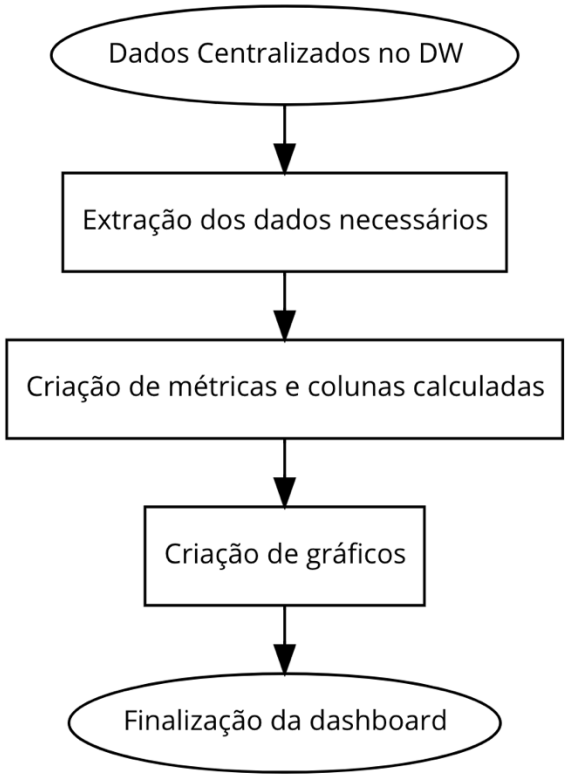

Figura 8 - Fluxograma do processo de BI

As *dashboards* contêm dados estratégicos agregados de alto nível, incluindo indicadores chave de desempenho - KPI's. Eles incluem relatórios interativos com dados traduzidos em gráficos, medidores e ilustrações para simplificar a comunicação de tópicos complexos. Além disso, elas permitem interações básicas - como as operações de *drill-down* e *slice-and-dice* - e fornecem vários níveis de detalhes para alcançar insights mais profundos.

Através de uma página web, o usuário final poderá interagir com a *Dashboard* para a extração de indicadores e insights. No próprio sistema da ferramenta de visualização de dados, é possível ter controle de acesso, restringindo as informações e dados para usuários específicos.

3.5 Machine Learning

Existem muitas formas de análise para a tomada de decisões de uma organização. É a combinação desses estilos que aumenta a maturidade dos recursos de análise de negócios dentro de uma empresa.

A Gartner, através da Figura 9, descreve o que eles chamam de Análise Contínua explicando os estilos analíticos, do descritivo ao preditivo. As análises descritivas são todas aquelas centenas, senão milhares, de relatórios que são gerados, mas nunca é feito nada com eles. Já a parte de análise prescritiva e de análise de diagnóstico são gerados através das *dashboards* utilizando sistemas de *Business Intelligence* como descritos no subcapítulo anterior.

Por fim, temos a análise preditiva que utiliza tecnologias como aprendizado de máquina (*Machine Learning*) e ter iniciativas como:

- Analisar tendências
- Identificar comportamento dos clientes e dos leads
- Tomar decisões de maneira proativa
- Melhorar desempenho nos negócios

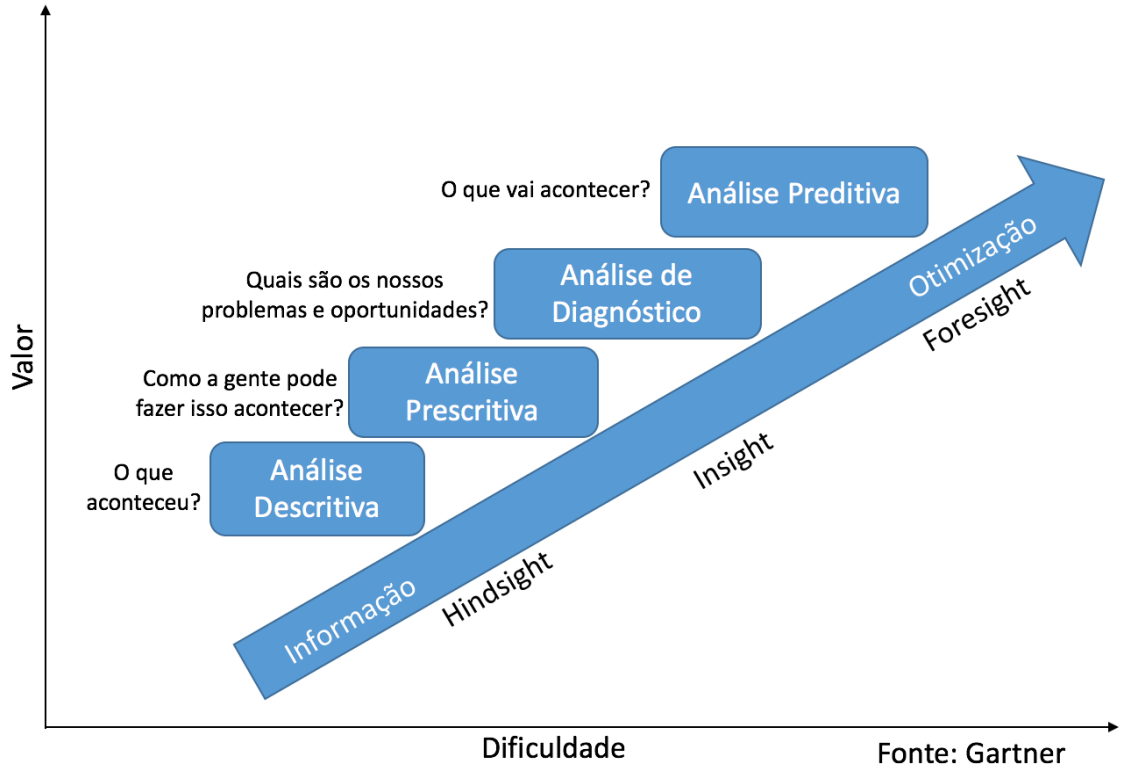

Figura 9 - 4 tipos de análises para melhorar a performance nos negócios

Fonte[Traduzida]:http://www.informationbuilders.es/sites/www.informationbuilders.com/files/intl/co.uk/presentations/four_types_of_analytics.pdf?redir=true

Machine Learning, ou ML, pode ser definido como um campo de Ciência dos dados no qual as máquinas podem aprender sem ser explicitamente programada por pessoas. Analisando dados históricos chamados de "dados de treinamento", o modelo de ML forma padrões e usa-os para aprender e fazer previsões futuras. [13]

No entanto, para um bom aprendizado da máquina, há a necessidade de seguir o processo descrito na Figura 10 principalmente na limpeza dos dados.

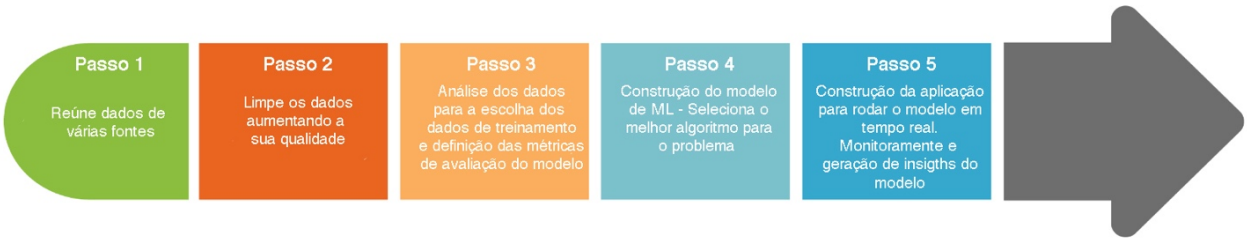

Figura 10 - Processo de desenvolvimento de um modelo de Machine Learning

Com a junção dos dados no primeiro passo, a limpeza dos dados consiste em popular os campos vazios. Para colunas numéricas, os campos podem preenchidos de acordo com várias abordagens tais como Naive Bayes, Prcesso Gaussiano, nearest neighbours e, principalmente, a média do conjunto de dados pelo fato de não alterar o desvio padrão da variável dependente.

Já para variáveis categóricas, uma abordagem comum é criar uma nova categoria de dados agrupando todos que não contem informação da variável em específico. Por fim, os últimos processos de limpezas são os tratamentos dos dados inseridos errados.

Dentre os vários tipos e técnicas de aprendizado em ML, pode-se categorizar em duas principais: supervisionado e não supervisiona. O tipo de aprendizado supervisionado, utilizado no presente trabalho, é deixado de forma explícita quais são as entradas de dados para a máquina prever uma saída específica. Já o tipo de aprendizado não supervisionado pode extrair inferências não explícitas a partir do conjunto de dados.

A técnica de classificação tem como objetivo predizer a qual de um conjunto de categorias uma nova observação pertence a partir do aprendizado com dados históricos conhecidos como dados de treinamento. Os principais algoritmos de classificação são regressão logísticas, árvore de decisão aleatória, classificadores de rede neural, adaboost, gradient boosting e KNN(*k-nearest neighbor*).

No presente trabalho, será utilizado um algoritmo de aprendizado de máquina para melhorar o processo de qualificação dos leads e entregar as melhores oportunidades para os vendedores.

Dado a fundamentação teórica do projeto, é ressaltado que o trabalho visa concentrar os dados utilizados da empresa em um *Data Warehouse* via uma plataforma de integração desenvolvida. Com os dados centralizados e processos bem definidos, foram desenvolvidas duas vertentes:

Primeiramente, foi implementado a solução de *Business Intelligence* com o objetivo de aumentar a eficiência das análises gerando *insights* valiosos para a empresa. Por fim, houve a implementação do *Predictive Lead Scoring* para melhorar o processo de qualificação dos *leads* entregando as melhorando oportunidades para os vendedores através de algoritmos de *Machine Learning.*

4 ESTRUTURAÇÃO DO FLUXO DE DADOS

As soluções de Business Intelligence (BI) visam auxiliar os processos de tomadas de decisões, fornecendo uma visão abrangente sobre os principais dados corporativos de uma empresa. Para isso, é necessário que os dados de análise sejam de qualidade, a fim de que não haja equívocos nos *insights* e decisões.

As informações erradas podem ser causadas por inúmeras complicações nos processos dentro da empresa tais como: inserção de dados de forma manual, mudanças de processos e erros de código.

Contudo, para que estas adversidades sejam minimizadas, é necessário que a empresa tenha uma estrutura de fluxo de dados definida e haja um armazenamento automatizado com todos os dados em um *Data Warehouse* para que, em todas as análises, haja somente um ponto de consulta.

Assim, este capítulo visa apresentar a estrutura de fluxo de dados definida para a Resultados Digitais e os componentes que foram desenvolvidos para a sua implementação.

4.1 Definição da Estrutura de dados

Antes do desenvolvimento deste projeto, a Resultados Digitais não possuía uma estrutura definida de fluxos de dados. As análises se limitavam a utilizar dados provenientes de um único software e, quando havia necessidade de cruzar diferentes fontes, era preciso exportar as informações em formato *csv*, integrá-los manualmente e após, realizar as análises através de ferramentas como o *Google Sheets.*

Ainda que funcione em situações menos complexas, a integração de dados utilizando planilhas apresenta dois principais pontos falhos ao longo prazo:

- As análises e *dashboards* em planilhas não levam em conta a experiência dos múltiplos usuários que podem acessar esses dados para a retirada de informações de forma simples e clara;
- Com muitos usuários tendo acesso às planilhas, as fórmulas de métricas podem ser editadas e alteradas não intencionalmente com extrema facilidade ao alterar uma fórmula ou quando se insere novas colunas. Caso esse erro não seja percebido, será gerado uma interpretação equivocada dos dados.

A fim de solucionar esses problemas, é necessária definir uma estrutura de fluxo de dados de modo que as informações tenham qualidade e que haja interoperabilidade entre os softwares utilizados dentro da empresa.

Atualmente, os softwares utilizados na Resultados Digitais são:

- Dados de clientes e financeiro – **Contas (software interno)**
- Suporte - **Zendesk**
- Marketing - **RD Station**
- Vendas - **Pipedrive**
- Dados de uso do RD Station pelos clientes – **Mixpanel e Segment**
- *Customer Success* - **Gainsight**
- Recursos Humanos - **Lever**

Dado o número de softwares utilizados pela empresa, foi visto que a centralização das informações facilita o cruzamento de dados entre fontes diferentes. Além disso, há uma maior consistência na interpretação das informações e melhor correção de erros de dados visto que a definição e criação das métricas serão provenientes de apenas um lugar. [4]

Desse modo, foi desenhado um fluxograma, representado na Figura 11, detalhando como serão centralizados e organizados os dados da empresa e quais interfaces farão o processamento, transferência e armazenamento dos dados.

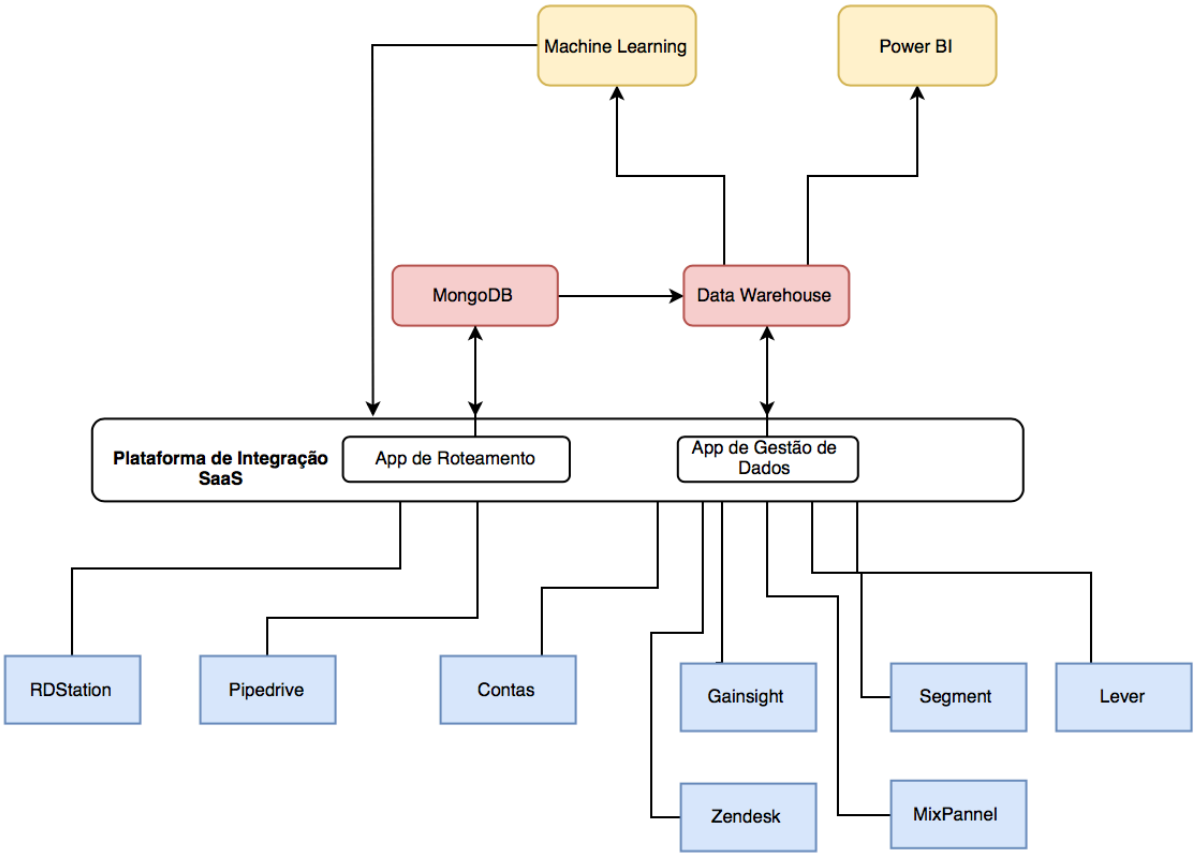

Figura 11 - Estrutura de Fluxo de Dados da Resultados Digitais

Para que a empresa tenha uma maior flexibilidade em eventuais substituições de ferramentas, é necessário que se opte por ter uma camada de barramento como uma plataforma de integração SaaS centralizando as informações em um *Data Warehouse.*

Para armazenar os dados de Marketing e Vendas, foi escolhido o MongoDB, já que um banco de dados não relacional possui alta flexibilidade para ajustes na estrutura das informações, além de que essas áreas alteram os seus processos em uma frequência relativamente alta comparada às outras áreas.

Já o aplicativo de roteamento será o responsável por inserir nesse banco todas as informações dos *leads* e *deals*. Antes da inserção dos dados, o aplicativo recebe as informações dos leads, realiza o enriquecimento deles, aplica o algoritmo de *Lead Scoring* e envia as melhores oportunidades para o Pipedrive para que sejam abordados pelos vendedores.

Para a comunicação com o software de BI escolhido, o Power BI, é necessário que os dados estejam em um banco de dados relacional, pois o software interpreta os

dados de maneira tabular. Assim, a escolha do Azure SQL para a centralização de todos os dados da empresa é devido a facilidade de comunicação com o Power BI, ambos desenvolvidos pela Microsoft, o preço ser acessível e ser hospedado em nuvem.

No entanto, para que a implementação da estrutura de dados definida fosse bem-sucedida, houve a necessidade de desenvolver variados componentes para melhorar o processo do fluxo de dados da empresa tais como:

- Desenvolver algoritmos de raspagem de dados para enriquecimento das informações faltantes sobre os clientes e enriquecimento dos dados dos leads para conhece-lo melhor;
- Desenvolver uma ferramenta auxiliar para a integração entre o CRM[6] (Pipedrive) e o Software Financeiro (RD Contas) para garantir que sempre as informações sobre o cliente sejam inseridas;
- Desenvolver a plataforma de integração dos dados para a centralização dos dados no *Dataware house* a partir de rotinas de atualização temporais*;*
- Desenvolver modelo lógico do Data Warehouse.

4.2 Enriquecimento de dados via scrapping

Ao analisar os dados provenientes do RD Contas, o software financeiro da empresa, foi constatada a falta de informações necessárias dos clientes, tais como: tamanho da empresa, cargo, segmento de atuação e localização da empresa.

Essa ausência de dados ocorre devido a erros de processos, acarretando a limitação dos números de amostras que contenham informações de qualidade e, assim, prejudicando as análises dos dados. Por consequência, a alteração dos processos de inserção das informações é fundamental.

Desta forma, foi desenvolvida uma ferramenta auxiliar na geração de contratos para os novos clientes. Contudo, a mudança desse processo não soluciona a falta de dados dos atuais clientes, sendo necessário o enriquecimento dos mesmos.

Há várias ferramentas disponíveis no mercado para o enriquecimento de Leads e Clientes como Clearbit, FullContact, Serasa Experian e Similar Tech. No

6 Customer Relationship Management - um sistema integrado de gestão com foco no cliente que reúne vários processos de uma forma organizada e integrada

entanto, essas ferramentas possuem um alto custo, inviabilizando a sua contratação para a situação atual da empresa e do projeto.

Além disso, a escolha por desenvolver o próprio algoritmo de raspagem de dados resultou em enriquecer não só os dados dos clientes, como também os dados dos leads. Essas informações extras – como número de funcionários dedicados a marketing, proprietário de site, uso de ferramentas de marketing digital no site e níveis básicos de SEO[7] no site – ajudam em análises mais aprofundadas sobre o perfil de cliente ideal da Resultados Digitais e, principalmente, no processo de qualificação de leads, já que esses dados podem indicar a maturidade digital da empresa para trabalhar com uma ferramenta de automação de marketing como o RD Station.

Todos os dois algoritmos foram escritos em linguagem Java Script sendo que o primeiro código é o responsável por buscar as informações faltantes dos clientes e inseri-los na base de dados.

A Figura 12 mostra o fluxograma do algoritmo desse script. Nela, nota-se a simplicidade da operação de enriquecimento, exceto por dois pontos:

- Definir uma fonte confiável e legal para a extração de informações;
- Ser obrigado a utilizar proxies para não ter o IP bloqueado pelo Google devido ao alto número de requisições realizadas no sistema

7 É o conjunto de estratégias com o objetivo de potencializar e melhorar o posicionamento de um site nas páginas de resultados orgânicos nos sites de busca.

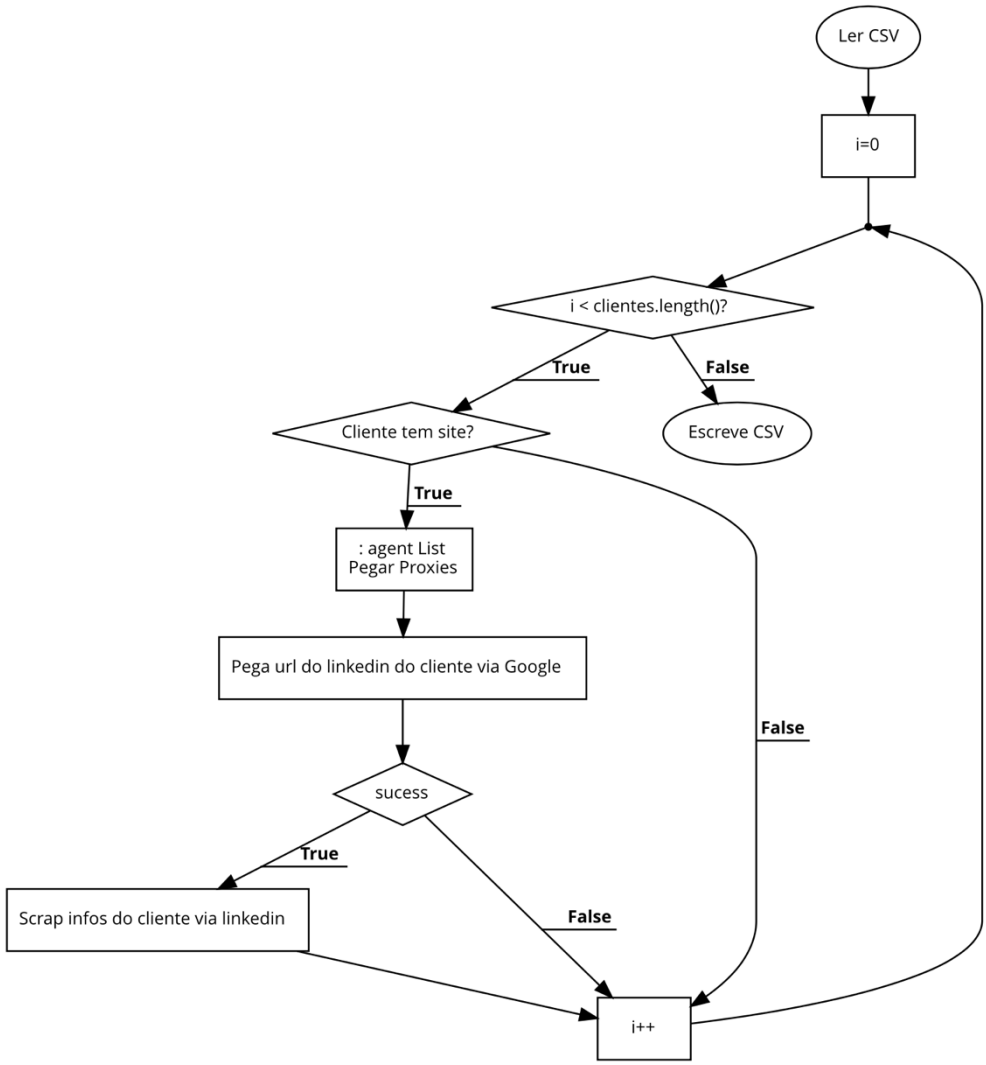

Figura 12 - Fluxograma do algoritmo de scrap

Já o segundo algoritmo é responsável pela obtenção das ferramentas de marketing e dos elementos de SEO (*Search Engine Optimization*) que o site do lead utiliza. Os dados buscados são:

- Elementos de SEO: Título, meta description, h1, h2, canonical tag e open graphics
- Ferramentas de Marketing: Wordpress, Pixel Adwords, Yoast SEO, Blog, Google Tag Manager, Google Analytics, RD Station, Facebook, Optimizely, Crazy Egg, Olark, Aweber, Hubspot e Hotjar.

Na Figura 13, é mostrado o fluxograma de funcionamento desse script. Como o site do lead já era obtido através dos formulários de conversão nas *Landing Pages*,

não é necessário utilizar proxies nesse algoritmo, pois não haverá requisições no Google.

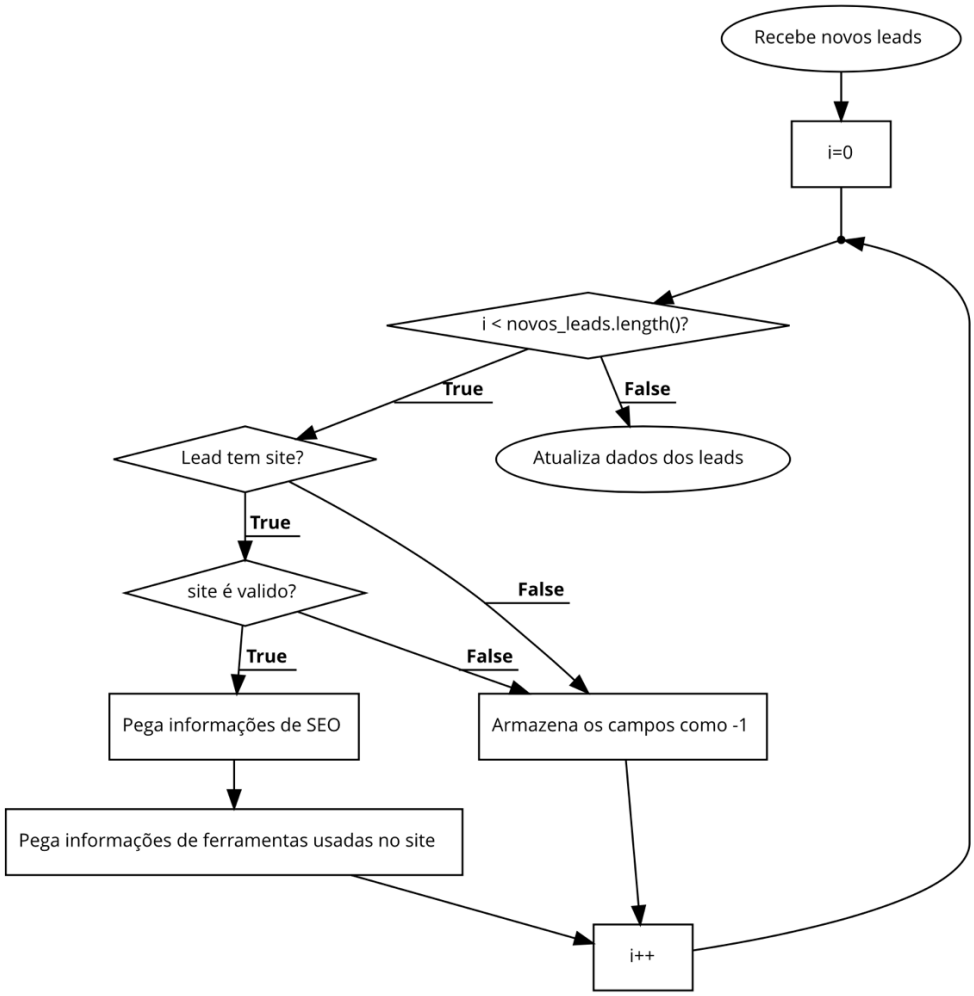

Figura 13 - Fluxograma do funcionamento do scrap de informações sobre o site do cliente

Esse algoritmo, além de fazer parte do aplicativo de roteamento que não foi desenvolvido pelo autor desse trabalho, também foi executado e atualizado para todas as bases de *leads* da Resultados Digitais com o objetivo que se atinja a qualidade desejável nos dados e que seja possível criar modelos preditivos precisos para a qualificação de *leads*.

4.3 Ferramenta auxiliar para a integração entre o CRM e Software Financeiro

Ao analisar a falta de informações dos clientes no banco de dados do software financeiro, foi constatado que os vendedores, no momento de cadastrar um novo cliente, não preenchiam todos os campos necessários, já que se tratava de um

processo extremamente burocrático, onde eles precisavam preencher as mesmas informações três vezes:

- No CRM
- Na geração do contrato – página web
- No RD Contas após o contrato assinado

Ademais, também foi constatado que não foi possível cruzar os dados dos clientes que estavam contidos no RD Contas com os dados que estavam no RD Station e no Pipedrive à época em que o cliente era apenas um lead.

Este problema ocorreu devido a falta de chave estrangeira que faria esta ligação, pois no sistema de automação de marketing, o RD Station, a chave primária do lead é o e-mail, sendo esta, a chave estrangeira no CRM. Assim, caso o lead seja qualificado, ele se torna um MQL[8] e é enviado ao CRM para que os vendedores façam a abordagem.

A ligação entre o lead e as informações do RD Contas acabam se perdendo quando a pessoa que assina o contrato não é a mesma que era o lead. Exemplificando: a abordagem e decisão da assinatura do serviço foi feita pelo gerente de marketing, contudo, o e-mail de assinatura colocado no contrato foi o do diretor financeiro.

A solução encontrada foi desenvolver uma ferramenta auxiliar para gerar o contrato e enviar o id do lead no CRM para o banco de dados do RD Contas e, com isso, o id será utilizado como chave estrangeira da tabela de clientes do RD Contas.

Além disso, para evitar a falta de informações sobre os futuros clientes, a ferramenta será o meio para a geração da url do contrato sendo que o vendedor insere os dados do cliente somente na ferramenta e, ao gerar a URL do contrato, será enviado, por meio do método GET, os dados do cliente para a página web de geração de contrato, para o banco de dados do RD Contas e, consequentemente, para o banco de dados onde estará centralizado todas as informações da empresa.

As classes do sistema desenvolvido podem ser representadas por meio de diagramas. A Figura 14 mostra o diagrama UML de diagrama de classes. Nesta

8 Acrônimo de Marketing Qualified Leads

imagem, cada retângulo representa uma classe, cada classe é composta por vários atributos e as setas representam como as classes se relacionam entre si.

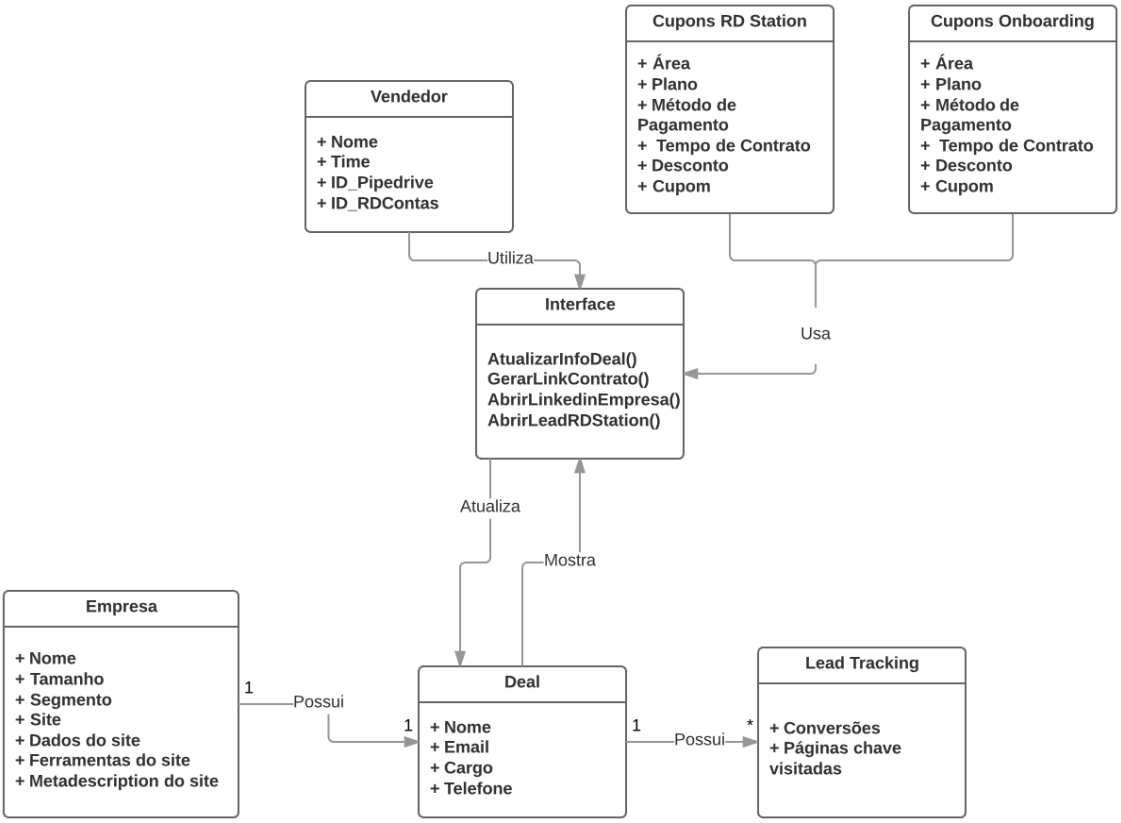

Figura 14 - Diagrama de classes

A fase de elaboração do processo unificado comporta também a determinação dos eventos e respostas do sistema, o que é feito mediante a criação de diagramas de sequência. O diagrama de sequência, representado na Figura 15, é uma forma de obter mais detalhes sobre o funcionamento do sistema [14].

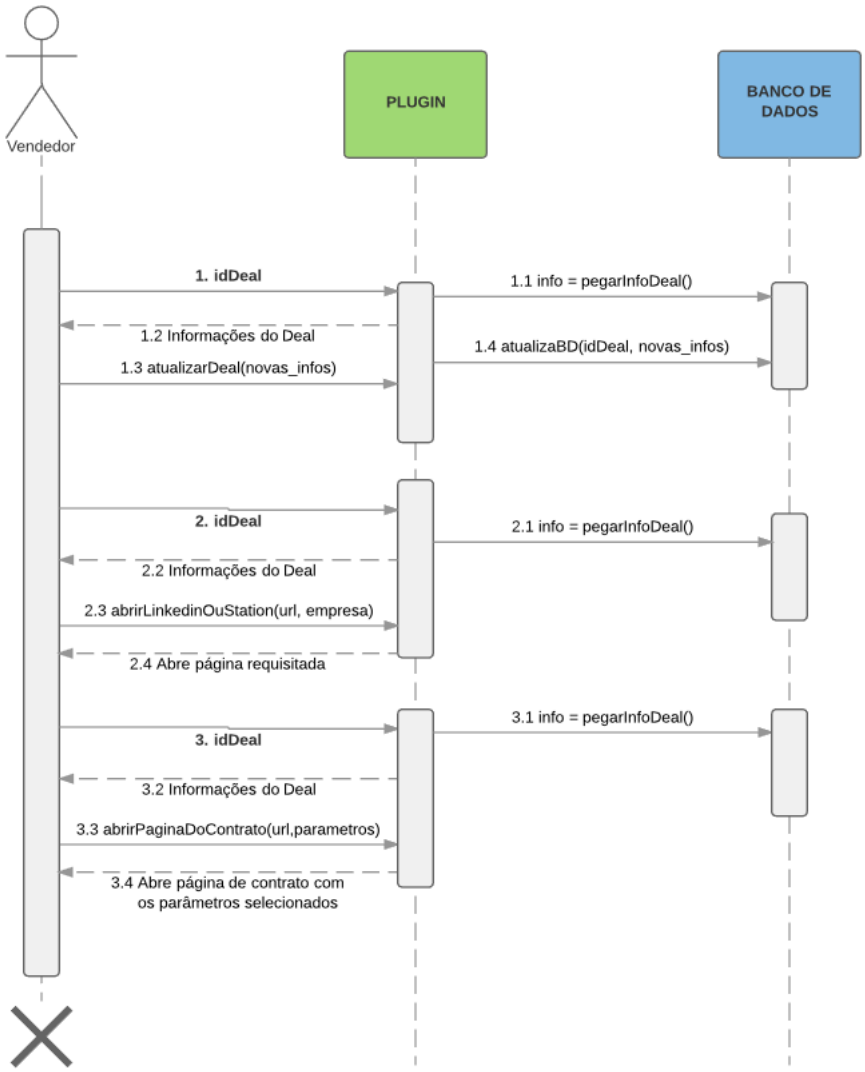

Figura 15 - Diagrama de Sequências

4.3.1 Implementação da Solução

A fase de construção é caracterizada pelo desenvolvimento físico do software, produção de códigos e testes preliminares.

Inicialmente foi escolhido desenvolver a ferramenta auxiliar como uma extensão do Google Chrome pela facilidade de instalação e atualização em todos os computadores, dado que a extensão fica hospedada no próprio *Market place* do Google, e estará disponível somente para pessoas que tenham o e-mail corporativo da empresa.

Outros fatores positivos são que: i) as linguagens de desenvolvimento são extremamente tradicionais (HTML, CSS e Javascript); ii) a fácil comunicação com as páginas ativas no browser; iii) a forma simples de fazer requisições externas.

A quantidade mínima de arquivos presentes em uma extensão do Google Chrome são 4:

- Arquivo Manifest – Contém metadados da extensão no formato JSON como propriedades como o nome da extensão, descrição, número de versão, etc;
- Arquivos HTML – Define toda a interface do usuário como informações, caixas de texto, botões, etc;
- Arquivo CSS – Define o estilo dos elementos que serão utilizados no arquivo HTML;
- Arquivo Javascript – Interage com o arquivo HTML e com a página ativa do browser podendo alterar as informações mostradas na interface.

Primeiramente, o código requisita as informações do *deal* (dados pessoais, dados da empresa e dados de uso conversões) utilizando o método GET para o MongoDB onde estão registrados todos os dados dos softwares de marketing e do CRM.

Em seguida, o mesmo método é utilizado para fazer a requisição das informações de vendedores, cupons de desconto do RD Station e cupons de desconto do *Onboarding* - treinamento pós-venda que é oferecido para os novos clientes visando um melhor uso da ferramenta.

Com todas as informações recebidas pelo sistema, as informações do *deal* serão mostradas na interface e, assim, o vendedor conseguirá abrir a ferramenta quando estiver na página específica do *deal* no Pipeprive.

Assim, ele poderá visualizar de forma objetiva as principais características do deal, atualizar as suas informações e gerar o link do contrato de assinatura de acordo com o plano contratado.

Para a geração do link do contrato de um novo cliente, o vendedor deve selecionar os campos:

- Plano: Basic, Pro e Enterprise;
- Período do contrato: 6 ou 12 meses;

- Número de contatos na base: 5, 10, 25, 50, 75, 100, 125, 150, 175, 200, 225, 250, 275, 300 mil contatos;
- Método de pagamento: Boleto ou cartão de crédito;
- Pagamento antecipado: Não, 6 meses, 12 meses;
- Desconto;
- Dia do vencimento.

Ao gerar o link do contrato, as informações necessárias do cliente serão transferidas de maneira automática para a página de contrato e para o sistema do RD Contas, diminuindo o retrabalho e a possibilidade de inconsistência nos dados. As Figura 16, Figura 17, Figura 18 e Figura 19 ilustram a interface do sistema desenvolvido.

Figura 16 - Primeira tela da ferramenta de suporte para vendas

Figura 17 - Segunda tela da ferramenta de suporte para vendas

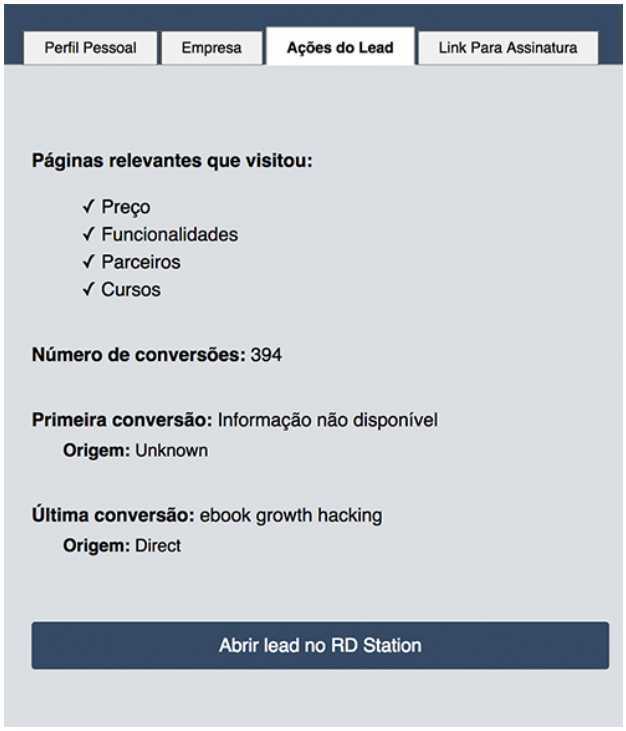

Figura 18 - Terceira tela da ferramenta de suporte para vendas

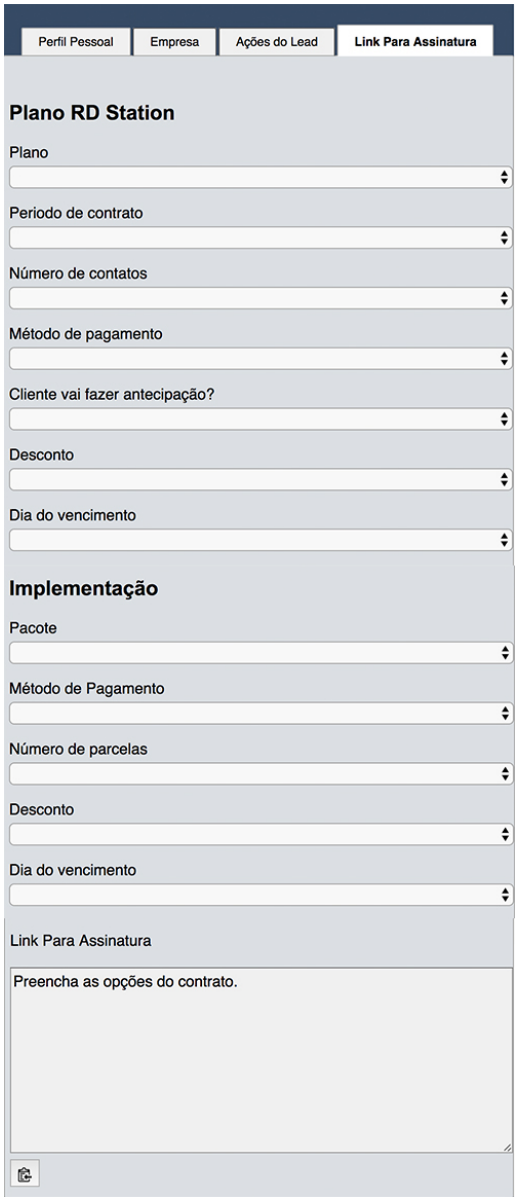

Figura 19 - Quarta tela da ferramenta de suporte para vendas

4.4 Plataforma de Integração de dados SaaS - iPaaS

Para obter *insights* mais valiosos, foi necessário centralizar todas as informações em um banco de dados. A centralização das informações traz a facilitação do cruzamento de dados entre fontes diferentes.

Além disso, há uma maior consistência na interpretação das informações e também uma melhor correção de erros de dados visto que a definição e criação das métricas serão provenientes de somente um lugar. [4]

Há várias plataformas iPaaS disponíveis no mercado: Informatica, Mulesoft e SnapLogic. No entanto, essas ferramentas oferecem mais opções do que a

necessidade atual da empresa, apresentando um alto custo de contratação com um baixo retorno sobre o investimento.

Desse modo, foi decidido internalizar o desenvolvimento do algoritmo que foi implementado em linguagem NodeJS e com a aplicação sendo hospedada no OpenShift, uma solução em nuvem *open source* conhecida como Plataforma como Serviço (PaaS).

Para o presente trabalho, foram desenvolvidas somente as rotinas de atualização do RD Station, do Pipedrive e do RD Contas como forma de MVP (mínimo produto viável) do iPaaS. Conceitualmente, um MVP é construir a versão mais simples e enxuta de um produto (ou parte dele), empregando o mínimo de recurso (tempo e dinheiro) e que entregará a proposta de valor principal da ideia.

Logo, foram implementadas as rotinas de atualização desses três softwares para a centralização dos dados, tendo assim, somente uma fonte de verdade para as próximas análises.

Após o desenvolvimento do presente projeto, o próximo passo será finalizar a centralização de todos os dados dos outros softwares para que possa ser feito análises mais complexas. Um exemplo deste tipo de análise é entender os motivos que levam os clientes a cancelarem a assinatura do software, analisando os dados de pós-venda, tickets do suporte e uso do software.

O fluxograma do algoritmo para cada comunicação entre o software e o *Data Warehouse* pode ser observado na Figura 20. As diferenças entre o algoritmo de atualização para cada software são somente algumas peculiaridades de comunicação com cada API e a frequência de atualização necessária já que, por exemplo, as informações de marketing mudam em uma velocidade maior do que informações de financeiras do cliente.

Nota-se também na Figura 20 que há uma redundância de 10 minutos entre a frequência de requisição e a janela de tempo de informações adquiridas em cada requisição. Isso acontece pela dificuldade de garantir com exatidão que o horário entre o Openshift e dos softwares sejam os mesmos.

Desta forma, essa estratégia insere somente no banco de dados as informações que não estão presentes na requisição anterior garantindo que todos os dados sejam inseridos no DW e que, assim, não haja falhas nas análises posteriores.

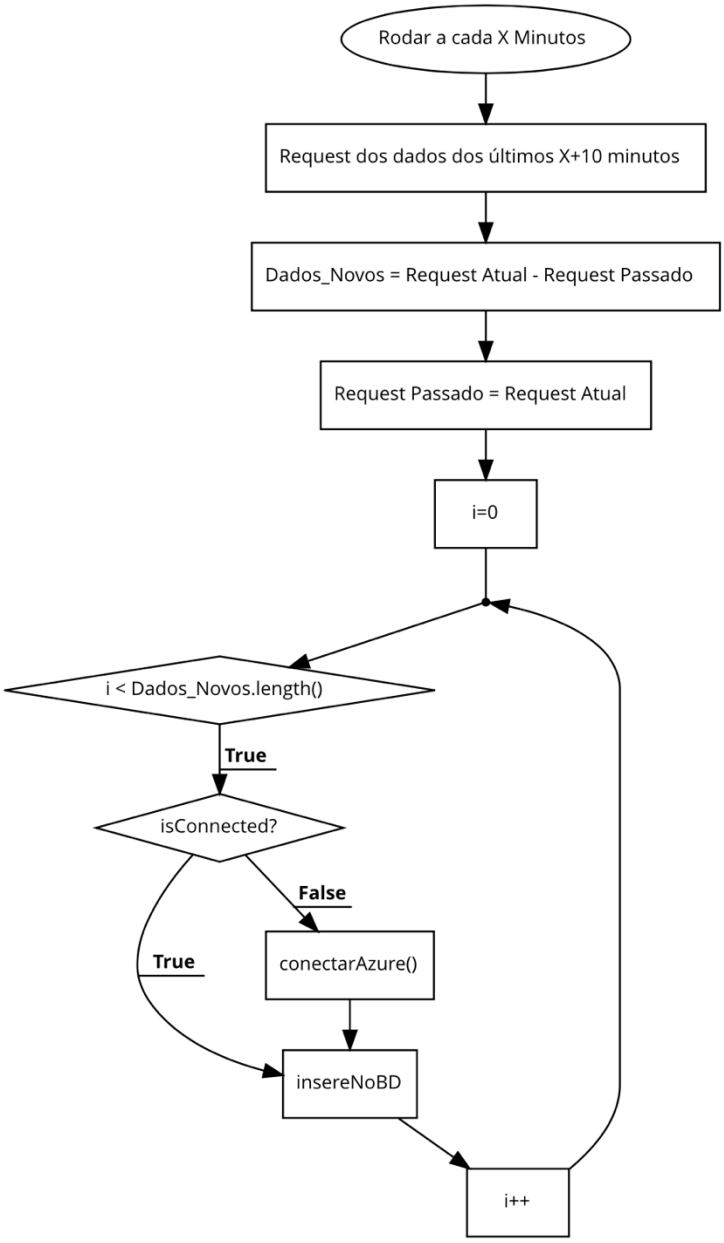

Figura 20 - Fluxograma padrão de comunicação entre o iPaaS e o software

4.5 Modelagem do Banco de dados

Com a definição da estrutura do fluxo de dados, enriquecimento das informações faltantes dos clientes e desenvolvimento da plataforma de integração dos dados, o próximo passo foi escolher o banco de dados, realizar a sua modelagem e implementá-lo.

O banco de dados escolhido foi um SQL Server hospedado no Azure, serviço em nuvem da Microsoft, por causa da facilidade de implementação e pelo custo ser proporcional ao uso do serviço.

A modelagem de dados é uma técnica usada para a especificação das regras e das estruturas de dados de um banco de dados. Ela faz parte do ciclo de desenvolvimento de um sistema de informação e é de vital importância para o bom resultado do projeto.

Construir um modelo lógico significa desenhar um esquema de dados consistente considerando algumas limitações do sistema de informações. Além de focar o esquema nas entidades lógicas e nas suas dependências, o modelo implementa recursos como adequação de padrão e nomenclatura, define as chaves primárias e estrangeiras, realiza normalização e implementa integridade referencial.

Deste modo, foi desenvolvido primeiramente o modelo lógico, representado na Figura 21, para o MVP desenvolvido com a centralização dos dados provenientes dos softwares de Marketing, Vendas e Financeiro.

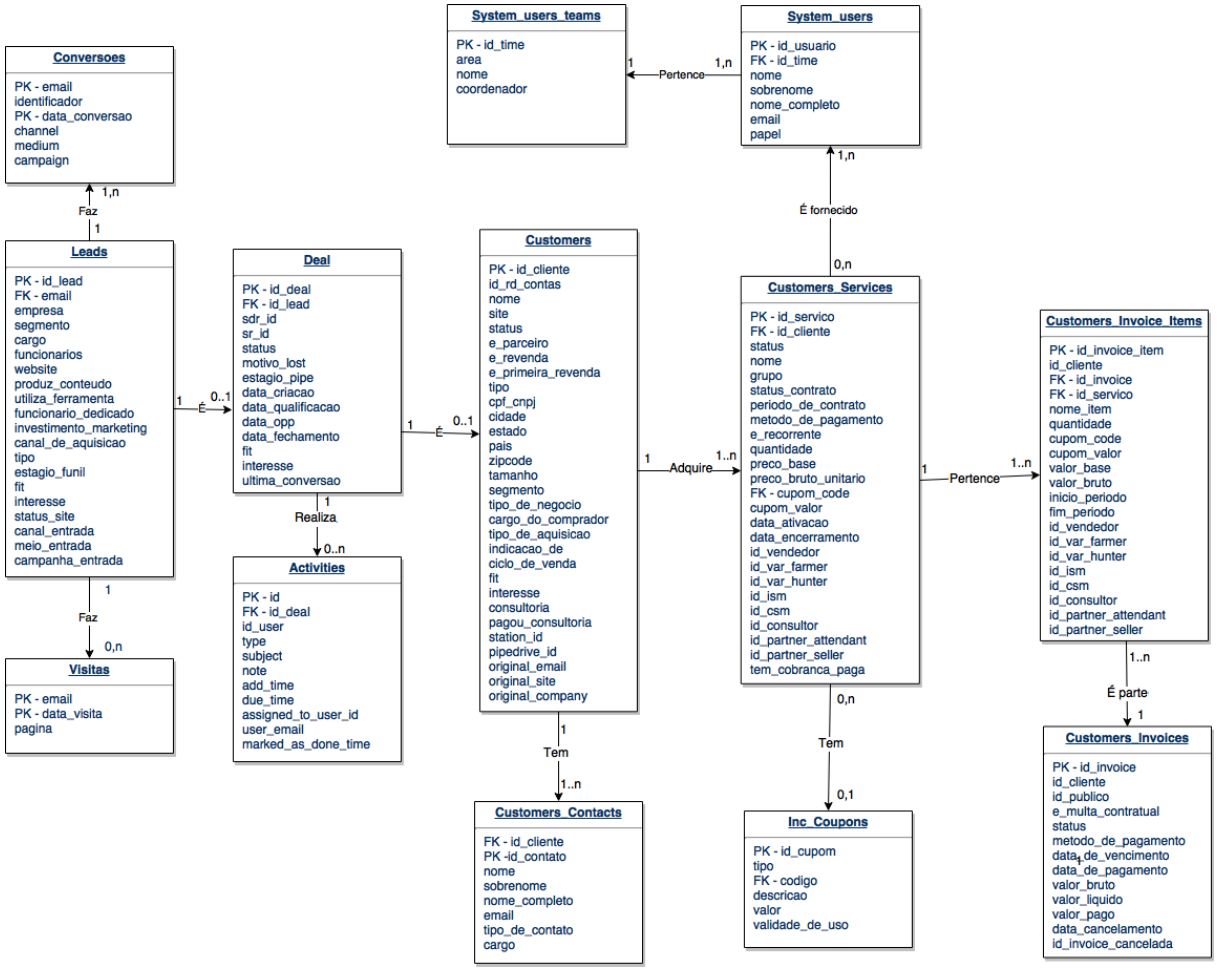

Figura 21 - Modelo Lógico do Data Warehouse com os dados dos softwares de vendas, marketing e financeiro

Por fim, o banco de dados foi criado de acordo com o modelo lógico definido e assim, aconteceram os primeiros testes para a inserção e atualização dos dados.

Com a validação de todo a estrutura de fluxo de dados desenvolvida, tem-se uma maior qualidade nos dados permitindo realizar análises por meio de ferramentas de *Business Intelligence* e implementar modelos preditivos com confiabilidade desejável.

5 IMPLANTAÇÃO DA SOLUÇÃO

A primeira parte do projeto visou construir uma estrutura de dados que ajudasse a gerar *insights* e melhorar os processos da empresa, a segunda parte do presente trabalho pretende desenvolver dois pontos principais:

- Finalizar a implementação do sistema de BI com a escolha da ferramenta e realizar modelagem dos dados;
- Estabelecer o primeiro modelo preditivo para a melhora de qualificação dos leads.

A implantação do sistema de *Business Intelligence* objetiva auxiliar os gestores e diretores para que organizem os dados necessários da análise de uma forma clara e intuitiva, através de painéis conhecidos como *dashboards.*

As *dashboards* contêm dados estratégicos agregados de alto nível, incluindo indicadores chave de desempenho - KPI's. Elas incluem relatórios interativos com dados traduzidos em gráficos, tabelas, medidores e ilustrações para simplificar a comunicação de tópicos complexos.

Já o desenvolvimento de um modelo preditivo com base em aprendizado de máquina visa melhorar o processo de decisão de quais leads devem ser abordados pelo time de vendas, conhecido como *Lead Scoring*. Com base nos dados históricos, o modelo preditivo forma padrões de quais características fazem um lead se tornar um cliente, além de usar os dados para aprender e fazer previsões sobre a possibilidade de um lead futuro ser qualificado.

Este capítulo abordará o desenvolvimento do sistema de *Business Intelligence* e do *Lead Scoring* preditivo onde a comunicação desses componentes com a estrutura dos dados explicada no capítulo 4 é ilustrada na Figura 11.

5.1 Business Intelligence

Os avanços tecnológicos nos últimos anos fizeram com que os modelos de negócios se transformassem completamente. Por consequência, as empresas necessitam encontrar diferentes oportunidades devido as metas agressivas de crescimento e a rápida mudança do mercado.

Para decisões mais complexas, são exigidas grandes quantidades de dados provenientes de diferentes fontes de dados utilizadas pela empresa. No entanto, sem haver uma estrutura de dados sólida e clara quanto aos seus objetivos, a análise das

informações demandará muito mais tempo do que o necessário, acarretando, por consequência, que a empresa perca oportunidades de negócio.

Como solução para isso, as empresas podem utilizar sistemas de Inteligência de Negócios (ou Business Intelligence), os quais utilizam tecnologia e produtos para buscarem os dados, transformá-los e fornecerem informações necessárias para as tomadas de decisões estratégicas, sem que haja gasto desnecessário de tempo.

No presente trabalho, a ferramenta de BI escolhida buscará as informações no DW desenvolvido no capítulo anterior, realizará os cálculos e a limpeza de dados através de um método conhecido como ETL (Extração, Transformação e Carga) disponibilizando somente os dados necessários para o desenvolvimento dos painéis de negócios e análises.

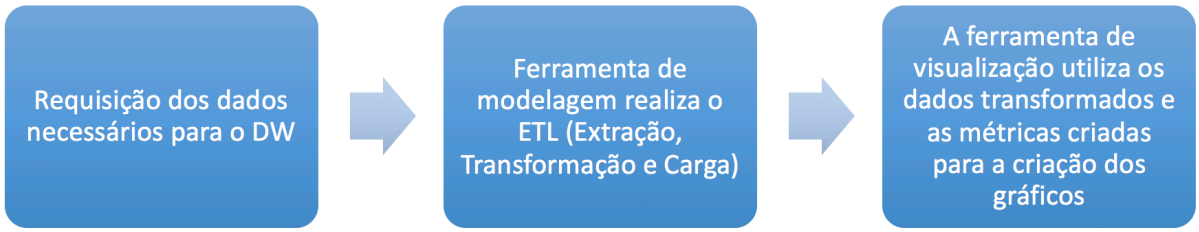

Figura 22 - Processo de desenvolvimento de uma dashboard

5.1.1 Escolha da ferramenta

Em paralelo ao desenvolvimento da estrutura do fluxo de dados, foram realizados vários testes com o propósito de selecionar a ferramenta mais adequada para atender as necessidades da empresa.

As 5 ferramentas escolhidas para a realização dos testes foram:

- Looker
- Microsoft Power BI
- Tableau
- Qlikview
- Good Data

A escolha dessas ferramentas se deu por diferentes fatores: i) liderança de mercado (Tableau e Qlikview); ii) novo produto de uma grande empresa no ramo de

tecnologia (Microsoft Power BI) iii) empresa SaaS com o produto totalmente em nuvem que pode entender melhor como calcular as métricas desse modelo (Looker); iv) atual parceira da Resultados Digitais no desenvolvimento do Marketing BI dentro do RD Station (GoodData).

Em função de todos os estudos realizados, foram descartados o Looker, Tableau e o Qlikview devido aos seus altos custos de implementação. Contudo, não se descarta que em um futuro próximo, quando a área de inteligência de negócios da empresa estiver mais consolidada, possam ser usados alguma dessas ferramentas.

Entre os outros dois softwares restantes, foi escolhido utilizar o GoodData devido ao conhecimento da existente da ferramenta pela área de Produto e por não haver custo adicional, já que a Resultados Digitais é cliente deles.

5.1.1.1 GoodData

Dentre as soluções avaliadas, o GoodData foi considerado o mais adequado para o contexto atual.

Foi levado em consideração o fato de que o time de produtos da empresa já possuía um conhecimento prévio da funcionalidade por terem desenvolvido a funcionalidade do marketing BI no RD Station, além de não ter custo.

Dada essa decisão, foi iniciado um desenvolvimento teste da estrutura e dos *dashboards* simples para a validação da ferramenta. Nesse processo, foram encontradas diversas dificuldades na utilização do software, desde a forma de configuração dos modelos de dados, que depende de um software para desktop, até a construção dos *dashboards*, que era bastante limitada.

De uma forma geral, o GoodData se mostrou uma solução com menos frequência de evolução do produto do que os seus concorrentes e voltada para a gestão de um grande volume de dados (bilhões de dados), que não é o caso da Resultados Digitais.

Após essa constatação, com um estudo mais aprofundado do Power BI, foi decidido que essa ferramenta traria um melhor desempenho para as nossas ações ao longo prazo. Isto pois, há um grande investimento da Microsoft nesse produto além do software ter várias funcionalidades modernas, assim como as líderes de mercado, mas por um preço mais acessível – U$10,00 por usuário. Por conta disso, optou-se em utilizar o Power BI como ferramenta.

5.1.1.2 Microsoft Power BI

O Power BI é um conjunto de ferramentas de análise de negócios que visa explorar melhor os dados, fornecendo uma visão holística para os usuários corporativos, já que possui suas métricas mais importantes em um só lugar, atualizadas em tempo real, e disponíveis tanto através do computador quanto pelo celular.

Em relação ao que já foi testado no GoodData, ele apresenta uma interface consideravelmente mais intuitiva e ágil, com bastante maturidade e entendimento dos pontos interessantes a serem levantados pelas ferramentas de visualização de dados.

A integração com o DW escolhido, Azure SQL, permite atualizar os dados periodicamente, facilitando o uso como uma *dashboard* de acompanhamento.

Apesar de ser um serviço novo, lançado em 2015, ele conta com uma comunidade ativa que visa esclarecer dúvidas e opinar sobre novas funcionalidades. Ademais, possui uma tendência de crescimento interessante, já que há novas atualizações a serem baixadas mensalmente, além da boa interação do time da Microsoft com a comunidade.

Além disso, o software, na sua versão desktop, engloba o Power Query utilizado para ETL. Em linhas gerais, essa ferramenta permite que o usuário efetue, dentro do próprio software de BI, a manipulação das informações obtidas, como selecionar somente as linhas e colunas necessárias, formatar o tipo das colunas e realizar a limpeza dos dados antes mesmo de transferi-las à interface de criação de gráficos para análise e geração das *dashboards* e relatórios.

5.1.2 ETL

O processo de Extração, Transformação e Carga (ou em inglês *Extract, Transform and Load*) é a parte do sistema de BI que se encontra entre o *Data Warehouse* e a área de apresentação do BI.

A primeira etapa do processo é a extração que consiste em importar do *Data Warehouse* (DW) os dados necessários para a criação da *dashboard*. Em seguida, é começado a parte de transformação dos dados. Nela, serão realizadas todas as transformações necessárias para a representação do dado conforme a necessidade do usuário final. Alguns exemplos possíveis de ser realizados nessa etapa são:

- Excluir linhas com dados inconsistentes;
- Formatar as colunas para os seus respectivos tipos de dados;
- Criar colunas e tabelas calculadas;
- Filtrar valores específicos;
- Selecionar tabelas e colunas específicas.

Nessa etapa do processo, é fundamental que o desenvolvedor da *dashboard* tenha planejada e estruturado exatamente as informações que estarão disponíveis e como elas serão representadas. Esse processo de planejamento é importante tanto do ponto de vista do negócio com os estabelecimentos de gráficos acionáveis, simples e de fácil interpretação quanto no desenvolvimento, pois serão processados somente as linhas, colunas e tabelas indispensáveis.

A etapa final, de carga, é responsável por carregar as informações com suas devidas alterações para o sistema de BI.

Para o presente trabalho, foi utilizado a ferramenta Power Query para o desenvolvimento dos processos de ETL. Nela, os processos são geridos através de passos que podem ser escolhidos através da interface do software ou poderão ser escritos nas linguagens MDX ou R para a melhoria de performance.

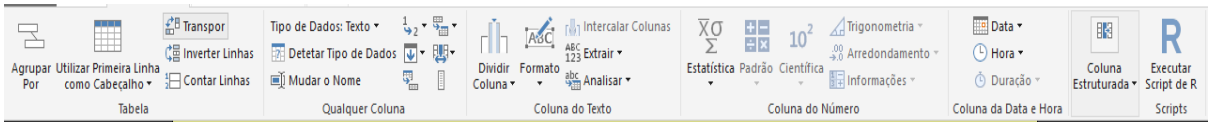

Figura 23 - Menu de transformação do Power Query

```
let
    Source = Sql.Database("master-rd.database.windows.net", "Master RD"),
    dbo_invoices = Source{[Schema="dbo",Item="invoices"]}[Data],
    #"Added Custom" = Table.AddColumn(dbo_invoices, "Preço_médio_mês = ", each [Preço_com_desconto]/(if [Número_de_parcelas] < 0
then [Número_de_parcelas]*-1
else if [Número_de_parcelas] = 0 then 1 else [Número_de_parcelas])),
    #"Renamed Columns" = Table.RenameColumns(#"Added Custom",{{"Preço_médio_mês = ", "Preço_médio_mês"}}),
    #"Grouped Rows" = Table.Group(#"Renamed Columns", {"ID_Contas"}, {{"Ticket Médio", each List.Average([Preço_médio_mês]), type number}})
in
    #"Grouped Rows"
```

Figura 24 - Editor avançado utilizando a linguagem MDX

Cada tabela, seja ela importada ou calculada, terá o seu próprio conjunto de passos a serem realizados, consistindo basicamente na sequencia de transformações que afetarão os dados importados para a criação da dashboard.

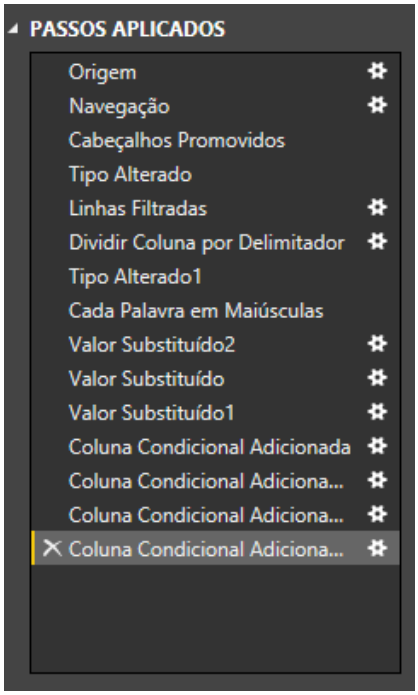

Figura 25 - Sequência de passos para transformação da tabela

5.1.3 Definição de relacionamentos entre tabelas

Após ETL, as tabelas podem ser relacionadas dentro do Power BI. Importante salientar que o software poderá fazer automaticamente a relação das tabelas, caso as chaves estejam com o mesmo nome.

A criação de relações entre tabelas é extremamente útil para a concepção de métricas, já que utilizam filtros de outras tabelas como, por exemplo, somar o número de conversões realizadas para um *lead* específico relacionando a tabela de conversões (conversions) e os dados do *Lead*(leads) representado na Figura 26.

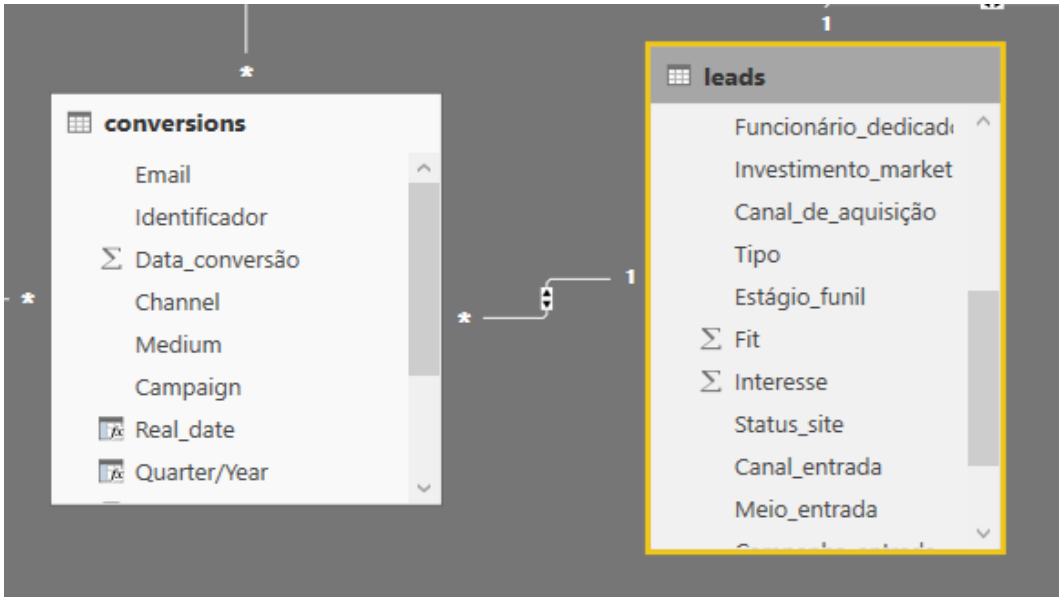

Figura 26 - Exemplo de relacionamento entre tabelas no Power BI

Essa funcionalidade do software se assemelha ao modelo lógico de um banco de dados relacional fazendo com que o modelo desenvolvido dentro do software fosse baseado na Figura 21. Além de selecionar as chaves estrangeiras que relacionam as tabelas, são definidas as direções da filtragem (unidirecional ou cruzada) e a cardinalidade (Muitos para um ou um para um).

A cardinalidade *Muitos para Um* é a relação do tipo fato para dimensão, por exemplo, uma tabela de vendas com várias linhas por produto com correspondência a uma tabela que lista os produtos em sua própria linha exclusiva. Já a cardinalidade *Um para Um* é usado para associar entradas únicas em tabelas de referência.

5.1.4 Criação das Dashboards

A criação dos painéis tem como principal objetivo ajudar os gestores a tomarem decisões assertivas de maneira célere, obtendo novas oportunidades de negócio e uma maior vantagem competitiva no mercado.

Apesar do Power BI ser uma ferramenta de fácil uso para a criação de métricas e gráficos, a parte mais importante do projeto de BI é o planejamento de quais dados devem ser informados e o por que, já que as empresas têm uma infinidade de dados que podem ser visualizados.

Essa importância é justificada quando 80% projetos de BI falham [15] principalmente por uma falta de objetivo de negócio claro ao desenvolver as *dashboards* e análises.

Dessa maneira, foi desenvolvido um processo para que o projeto tenha clareza do seu objetivo, de quem serão os usufruidores e quais ações serão tomadas por eles a partir de análises na *dashboard*. [16]

Além disso, esse processo foi testado e validado ao aplicar no desenvolvimento da *dashboard* de gestão de negócios que será apresentada nesse presente trabalho.

Os seguintes passos foram seguidos para o planejamento desse painel:

1) **Qual o objetivo da *dashboard*?**
2) **O que não é possível analisar hoje?**
3) **Quais ações seriam possíveis tomar com a nova *dashboard*?**
4) **Quais análises e informações seriam necessárias para que uma decisão fosse realizada de modo confortável?**
5) **Para cada análise, quais métricas ajudariam a tomar as ações definidas?**
6) **Resuma brevemente a história que a *dashboard* apresentará**

Desse modo, o objetivo da *dashboard* é fornecer aos gestores uma visão clara do perfil e distribuição da nossa base em relação as principais métricas de empresas SaaS como por exemplo MRR[9] e Taxa de *Churn*.

Antes do desenvolvimento desse trabalho, não era possível cruzar as informações entre os softwares com facilidade, além da complexidade de melhorar a qualidade dos dados sem um processo estruturado de ETL.

Com essa *dashboard*, os tomadores de decisão conseguem usufrui-la para variados projetos tais como alinhamento, priorização e segmentação das ações de marketing e vendas de acordo com os perfis de clientes mais interessantes segundo as análises de Perfil de Cliente Ideal. Para isso, é necessário que haja gráficos de distribuição de clientes (ativos e cancelados) em relação a alguns aspectos tais como:

9 Receita Recorrente por Mês

segmento, região, tamanho da empresa, tipo de negócio, ticket médio e plano adquirido.

Em resumo, o painel fornece uma visão holística da base de clientes da empresa considerando um cenário de multi-variáveis onde ainda é desconhecido o quanto cada fator influencia para aquisição e retenção dos clientes.

5.2 Aplicação de Machine Learning no processo comercial

É visto com frequência cada vez maior a evolução de algoritmos e aplicações de aprendizado de máquina no cotidiano das pessoas. Alguns exemplos clássicos dessas aplicações são os carros autônomos desenvolvidos pelo Google e o modo de recomendação de filmes que a Netflix utiliza de acordo com o perfil e filmes assistidos pelo assinante.

Com a estruturação do fluxo de dados, abriu-se a possibilidade de utilizar esse tipo de algoritmo para os processos de aquisição e retenção de clientes da Resultados Digitais. Com as metas agressivas e com um objetivo de aumentar a base de clientes em aproximadamente 150% ao final do ano, é necessário tanto aumentar a geração de leads quanto a taxa de qualificação dentro do processo para que melhores oportunidades sejam entregues para os vendedores fazendo com que essas metas sejam alcançadas.

Desse modo, essa etapa do projeto consistiu em implementar um algoritmo de aprendizado de máquina, conhecido como *Predictive Lead Scoring,* para melhorar a eficiência do funil de vendas.

Contudo, para melhor entender esta fase, será explicado brevemente o processo de vendas utilizado na empresa e como funciona a escolha de quais leads deverão ser abordados pelo time de vendas através do *Lead Scoring* tradicional.

5.2.1 Inside sales

A inovação tecnológica, principalmente as mudanças causadas pela internet, alteraram desde a forma de interação humana até o modo de aquisição de conhecimento. Sendo impactado da mesma maneira, houve mudanças no processo de vendas. Hoje, os clientes buscam diferentes informações em várias fontes sobre os produtos e sobre a empresa em que está negociando antes mesmo de entrar em contato com um vendedor. [17]

Com essa mudança por parte do cliente, o papel do vendedor, que antes era de convencer e informar, tornou-se de auxiliar na escolha e facilitar a compra.

Além disso, a empresa não necessita mais que os seus vendedores busquem novos clientes para uma comunicação pessoal, isto pois, o próprio interesse do cliente fará com que ele busque a empresa para uma comunicação virtual com o vendedor.

Com as vendas internas, tradução de *inside sales*, o time de vendas ganha em eficiência e produtividade, não tendo mais o risco de perder tempo em deslocamentos e reuniões canceladas.

Na Resultados Digitais, a grande maioria dos clientes vieram de *inside sales,* sendo que o processo na jornada de compra começa quando o lead vai atrás de um conteúdo para resolver algum problema da empresa. Ao encontrar o material, a pessoa deve preencher um formulário na *Landing Page* com informações pessoais para conseguir fazer o *download*.

Figura 27 - Exemplo de um formulário para download de material

Nesse processo, é importante que haja conteúdos ricos e de qualidade para um maior relacionamento com o *lead.* Este método acarretará no convencimento

contínuo da pessoa, até a percepção de retorno sobre o investimento e as facilidades que ela terá quando assinar o produto.

As etapas da jornada de compra na Resultados Digitais são:

- Lead – são pessoas que preenchem formulários demonstrando interesse no seu produto ou mercado, deixando suas informações de contato com a empresa;
- MQL *(Marketing Qualified Leads)* – dado as características do *lead* e o seu interesse no assunto, ele se torna um Lead Qualificado pelo time de Marketing;
- SAL *(Sales Accepted Leads)* – O time de pré-vendas qualifica os MQLs de acordo com condições necessárias para aquisição do produto que não eram possíveis de obter pelo preenchimento de formulários;
- Oportunidade – Os SALs que tem interesse e qualificação necessária são abordados pelo time de vendas para um maior entendimento do problema da pessoa e uma elaboração de proposta;
- Cliente – Lead que passou por todas as etapas do funil de vendas e assinou o produto.

O maior desafio do processo de vendas é ter previsibilidade acurada de quantos *leads* são precisos para gerar um determinado número de novos clientes. Além disso, com uma taxa de fechamento (Cliente/Lead) baixa, é preciso realizar grandes investimentos em mídia para a geração de leads, diminuindo assim, a rentabilidade do negócio.

Desta forma, ter taxas de qualificação a mais alta e previsível é importante, principalmente, a diminuição da discrepância entre o que a área de marketing considera como *lead* qualificado (MQL) e o que a área de vendas aceita como *lead* qualificado (SAL).

A decisão dos *leads* que são qualificados pelo marketing e enviados para vendas é realizado através de um sistema de pontuação conhecido como *Lead Scoring.*

5.2.2 Qualificação dos Leads através do Lead Scoring

Dentro do processo de vendas da empresa, há a qualificação dos leads realizado pela área de Marketing para eles serem repassados para abordagem pela

área de vendas. Esse processo de qualificação de leads tem o principal objetivo de melhorar a eficiência do time de vendas para que a empresa alcançar as metas estipuladas.

Em uma empresa que gera em média 2500 *leads* por dia, é necessário que haja um processo automático de qualificação e de encaminhamento desses *leads* para a área de vendas. Para isso, foi estabelecido um acordo entre Marketing e Vendas (SLA) definindo o que é um lead qualificado de acordo com o perfil e interesse no *Lead Scoring*, assim como, características ou condição do Lead que implica no bloqueio da entrega para vendas (e.g. Cargo ser estudante).

O *Lead Scoring* é um aplicativo onde são geradas duas notas de forma automática para os *Leads* uma de acordo com o seu perfil e a outra de acordo com o interesse do lead no negócio da empresa medidos através de interações que o lead tem com a empresa por meio de abertura de e-mails, conversões, visitas em páginas específicas, entre outras.

Figura 28 - Exemplo de pontuação por perfil

Fonte: http://ajuda.rdstation.com.br/hc/pt-br/articles/204365595-Lead-Scoring-A-qualifica%C3%A7%C3%A3o-de-um-Lead-a-partir-do-perfil-e-interesse

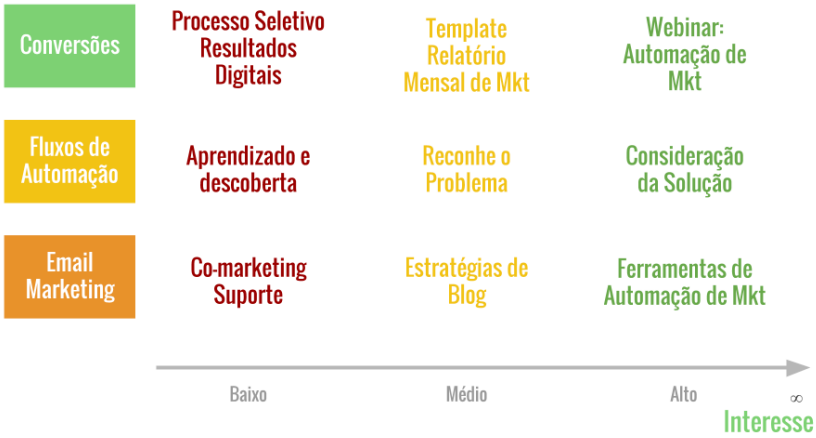

Figura 29 - Exemplo de pontuação por interesse

Fonte: http://ajuda.rdstation.com.br/hc/pt-br/articles/204365595-Lead-Scoring-A-qualifica%C3%A7%C3%A3o-de-um-Lead-a-partir-do-perfil-e-interesse

A partir disso, é realizada uma análise para a qualificação dos *leads* por meio desses dois indicadores. Os *leads* com um ótimo perfil tendem a necessitar de um menor interesse por estarem prontos para a compra. Por outro lado, os *leads* com perfil menor, em geral necessitam de maior tempo e interações com conteúdos ricos para chegarem ao momento da compra.

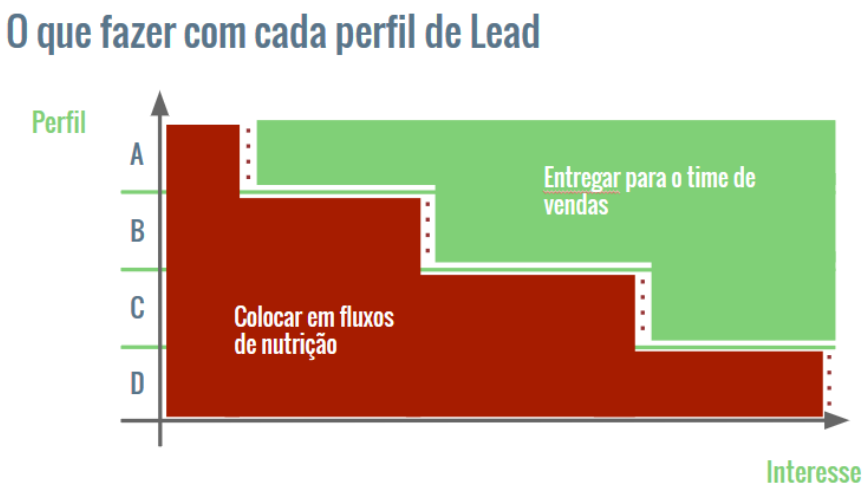

Figura 30 - Exemplo de gráfico (Interesse x Perfil) para qualificação do Lead

Fonte: http://resultadosdigitais.com.br/blog/lead-scoring-guia-completo/

Mesmo que seja um processo automatizado utilizando o RD Station, ele não é um processo ótimo, pois são dados pesos para as ações do Leads definindo de modo manual a pontuação do interesse e perfil.

Apesar de ser feito um estudo em cima da base de leads para saber quais ações e características tem maior peso, o comportamento dos leads é dinâmico ao longo do tempo e há um amadurecimento acelerado do mercado em relação a marketing digital.

Assim, essa pontuação estática em um contexto dinâmico pode acarretar tanto na discrepância entre MQL e SAL quanto no descarte de oportunidades que na época do estudo não eram interessantes.

Desse modo, o *Predictive Lead Scoring* atua com o objetivo de balancear de modo automático e dinâmico a importância de cada ação e característica do lead de acordo com o comportamento histórico dos clientes e dos *leads*.

5.2.3 Predictive Lead Scoring

O algoritmo do *Lead Scoring* Preditivo observa quais informações os leads que foram qualificados pela área de vendas têm em comum, bem como quais informações os leads que não fecharam a venda têm em comum. A partir dessas observações, o algoritmo decide se um novo lead deverá ser abordado por um vendedor ou não.

No entanto, há um processo para o desenvolvimento do modelo representado na Figura 10.

Para o desenvolvimento desse algoritmo, eram necessários os dados dos softwares de marketing e vendas com as informações sobre os *leads.* Como o processo de centralização dos dados já tinha sido desenvolvido, ficou mais fácil a exportação dessas informações para o formato *csv.*

A linguagem escolhida para o desenvolvimento do código foi Python devido a inúmeras bibliotecas disponíveis para o uso em desenvolvimento de modelos de *machine learning* como pandas, scikit-learn, seaborn, numpy e matplotlib. Além disso, há uma integração com o Jupyter, uma aplicação web que permite desenvolver o código de maneira visual auxiliando a limpeza, exploração e análise dos dados.

As informações dos leads que foram importados para a primeira análise foram:

- **Informações de perfil:** id_lead, e-mail, empresa, segmento, cargo, funcionarios, website, produz_conteudo, utiliza_ferramenta, funcionario_dedicado_mkt, investimento_mkt, canal_de_aquisicao, tipo, estagio_funil, status_site, website_tools, seo_elements, canal_entrada, meio_entrada, campanha_entrada
- **Informações de interesse:** total_conversions, bofu_conversion, mofu_conversion, tofu_conversion, mkt_conversion, cs_conversion, str_conversion, plan_conversion, autom_conversion, other_conversion, average_conversion_time, average_conversion_day, total_visits, visit_blog, visit_lp, visit_mat, visit_price, visit_home, average_time_visit

- **Informações sobre evolução no funil:** sent, quali, won.

5.2.3.1 ETL

O pré-processamento e a limpeza de dados são tarefas importantes e devem ser realizadas antes que o conjunto de dados possa ser usado com eficiência para o aprendizado de máquina.

Isto pois, os dados brutos não são confiáveis já que costumam conter informações fora de padrão, falhas de preenchimento e valores ausentes. Caso eles sejam utilizados, podem produzir predições erradas no modelo, prejudicando todo o processo de qualificação da empresa.

Primeiramente, foram excluídos os *leads* que tinham e-mail da "@resultadosdigitais" e os que não possuíam informações no campo total_conversions, pois um *lead* deve ter no mínimo uma conversão. Foi realizado o preenchimento dos valores vazios por 0 para os campos quantitativos e "None" para os campos qualitativos.

Nos dados importados, havia uma inconsistência na coluna de status do site e que poderia ser tratada de acordo com as ferramentas presentes no site (coluna website_tool). O método de tratamento para esta coluna é:

Se status vazio e website_tool = "site inválido", status = "inválido"
Se status vazio e website_tool = "None", status = "sem website"
Se status vazio e website_tool = Não nulo, status = "válido"

Os dados de ferramentas no site e dos elementos de SEO eram adquiridos de forma qualitativa, sendo necessário transformar para a quantidade de ferramentas e elementos que o lead possui. Isso era importante, pois é um indício de maturidade digital que faria um cliente gerar valor com o RD Station, ajudando na retenção do mesmo.

A última etapa de correção de dados imputados foi feita através da padronização dos campos fechados de tamanho de empresa e cargo, que foram colocados de maneiras distintas em diferentes formulários.

Para finalizar a etapa de transformação dos dados, foi necessário converter os campos categóricos para campos numéricos, pois os algoritmos de aprendizado de máquina interpretam somente dados numéricos. Além disso, foi realizada a normalização dos campos numéricos os transformando em valores entre 0 e 1 de forma linear a fim de evitar ruídos.

5.2.3.2 Escolha do tipo de algoritmo

Há dois métodos de aprendizado de máquina: o supervisionado e o não supervisionado. O método supervisionado, escolhido para o presente trabalho, é comumente usado em aplicações nas quais os dados históricos preveem prováveis acontecimentos futuros. Nele, o modelo é treinado usando exemplos rotulados como entrada e a saída desejada já conhecida.

Dentro do método supervisionado, há vários tipos específicos de aprendizados de máquina como sendo os principais: regressão, classificação, detecção de anomalias e clusterização.

No presente trabalho, o tipo escolhido foi o supervisionado por classificação sendo que as entradas serão os campos de perfil e interesse do *lead*, e a saída é a classificação binária se o *lead* é qualificado.

5.2.3.3 Resolvendo problemas de classes desbalanceadas

Em algoritmos de aprendizado de máquina do tipo classificação, uma base de dados é definida desbalanceada quando existem muito menos casos de algumas classes do que de outras. [18]

Os algoritmos de classificação desenvolvidos com métodos tradicionais são sensíveis a este tipo de desbalanceamento e tendem a valorizar classes predominantes e a ignorar classes de menor representação – conhecidas também como classes raras.

Os classificadores gerados a partir de bases de treinamento desbalanceadas apresentam altas taxas de falsos negativos para as classes raras, o que é problemático quando a classe de interesse é classe rara. A fim de evitar esse viés, são utilizadas técnicas de amostragem para um rebalanceamento das classes.

Há três principais métodos de amostragem para a resolução dessa questão:

- Subamostragem - eliminação de casos da classe majoritária;
- Superamostragem - replicação de casos da classe minoritária dos dados de treinamento visando obter classificadores melhores do que os obtidos a partir da distribuição original;
- Superamostragem por geração sintética – replicação da classe minoritária sendo que as novas amostras são baseadas na interpolação de instâncias das classes minoritárias.

A remoção de observações com a subamostragem pode fazer com que os dados de treinamento percam informações importantes pertencentes à classe majoritária.

Já a simples replicação de casos positivos (instancias de dados associados à classe de interesse) pode produzir classificadores muito específicos para os casos replicados e com baixo poder de generalização para outros casos positivos.

Por fim, a superamostragem por geração sintética evita o baixo poder de generalização. Contudo pode apresentar o efeito indesejável de criação de casos positivos que invadem o espaço de decisão da classe negativa. Essa característica, denominada sobreposição de classes, tende a degradar o desempenho de classificadores obtidos a partir de tais dados.

Dessa forma, a definição do melhor método de amostragem depende do tamanho da base de dados e o comportamento das classes no espaço amostral. No caso do processo de qualificação de *leads,* somente 10% dos leads gerados são enviados para o time de vendas e somente a metade deles são aceitos por eles se tornando SAL.

Para a escolha do melhor tipo, foram desenvolvidos modelos de Regressão Logística com variados tipos de amostragem. As quatros métricas de avaliação sobre o modelo foram:

- AUC - Área sob a curva ROC – A curva ROC fornece a medida para comparar performances de classificadores. Quanto mais próximo de 1 a área AUC, melhor a performance global do classificador. Sendo que o AUC igual a 0.5 é definido como classificador randômico;
- Cohen-kappa Score – Acurácia de classificação normalizada pelo desequilíbrio das classes nos dados;
- Precisão da classe majoritária – define qual é a percentagem das informações rotuladas como positiva são realmente positivas;
- Recall da classe minoritária – define qual é a percentagem das informações positivas que foram realmente rotuladas como positivas.

A Tabela 1 mostra os resultados de cada modelo amostrado. O Random Undersampler foi escolhido dado a sua melhor performance segundo as métricas descritas acima.

Tabela 1 - Resultados de cada tipo de amostragem

Método	**AUC**	**Cohen-Kappa Score**	**Precisão Classe Majoritária**	**Recall Classe Minoritária**
Unbalanced	0.511263	0.038569	0.934226	0.027964
Balanced Weight	0.780849	0.242023	0.980285	0.782438
Random Under Sampler	0.841801	0.306438	0.989591	0.883110
Near Miss	0.742753	0.170340	0.978577	0.788031
Near Miss 2	0.709199	0.336492	0.962023	0.492729
Near Miss 3	0.548458	0.136276	0.939038	0.115772
Tomek Links	0.513311	0.044363	0.934489	0.034116
Random Oversampling	0.735945	0.308761	0.967347	0.583893

SMOTE	0.736143	0.315119	0.967204	0.579978
SMOTE + Tomek Link	0.739171	0.308891	0.967939	0.592841
SMOTE + ENN	0.773241	0.275659	0.976135	0.719239
EasyEnsemble	0.823634	0.259754	0.989409	0.887025
Balance Cascade	0.786406	0.182570	0.990412	0.911074

5.2.3.4 Seleção das *features* e desenvolvimento do modelo

Na maioria dos conjuntos de dados, nem todos os atributos contribuem para a definição ou determinação dos rótulos de classe. Em teoria, espera-se que o aumento do vetor de característica proporcione mais confiabilidade na predição da classificação.

Na prática, no entanto, grandes vetores de características grandes reduzem significativamente o processo de aprendizagem, bem como fazem com que o classificador se prenda aos dados de treinamento, conhecido como *overfitting,* comprometendo a generalização do modelo. [19]

Em geral, as características são identificadas como:

- Relevante: características que influenciam o produto e seu papel não pode ser assumido pelos demais;
- Irrelevante: características que não influenciam o produto;
- Redundante: Um recurso pode assumir o papel de outro.

O objetivo da seleção de recursos é encontrar o conjunto de características que fornecem a maior parte das informações relevantes por um menor custo computacional.

No presente trabalho, foi desenvolvido um primeiro modelo utilizando o algoritmo *Random Forest* e amostrado com o *Random Under Sampler* para a utilização da função *features_importances,* disponíveis em todos os algoritmos com base em árvore. A Tabela 2 expõe as 10 características mais importantes para a qualificação de um lead segundo o algoritmo, sendo essas características escolhidas para o desenvolvimento do modelo dentre as 49 disponíveis.

Tabela 2 - 10 características mais relevantes para a determinação da saída do modelo

Ranking	***Features***	**Importância**
1	Conversões Total	0.23134
2	Segmento	0.21748
3	Número de Funcionários	0.18724
4	Status site	0.17363
5	Visitas Total	0.14728
6	Cargo	0.14010
7	Dia da conversão	0.13869
8	Outras conversões	0.08237
9	Estágio funil	0.04726
10	Campanha de entrada	0.03125

A última etapa do desenvolvimento do modelo é a escolha do melhor algoritmo de predição. Com a definição das características a serem analisadas e o tipo de amostragem, foram criados 8 modelos com algoritmos diferentes sendo eles avaliados pelo AUC [10], matriz de confusão e o número de *leads* que tornaram vendas e o modelo não aprovou. Cada modelo é criado e é feito a predição em cima de uma base de dados histórica na qual não fora usada para o treinamento do modelo.

10 Acrônimo de Area Under Curve – Área abaixo da curva ROC

- Regressão Logística

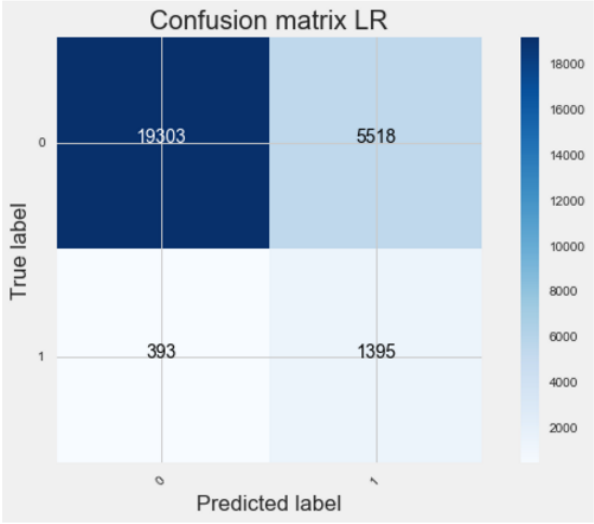

Figura 31 - Matriz de confusão do modelo por Regressão Logística

AUC: 0,778945

Leads não qualificados pelo modelo que foram vendas: 40

- Gradient Boosting

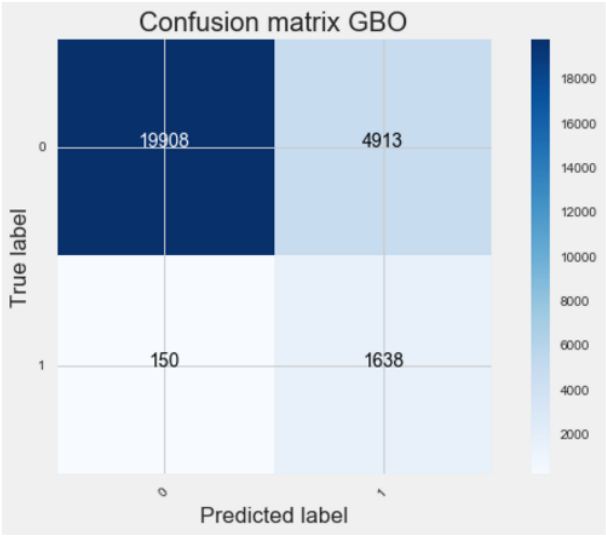

Figura 32 - Matriz de confusão do modelo por Gradient Boosting

AUC: 0,859085

Leads não qualificados pelo modelo que foram vendas: 0

- MLP – Rede Neural

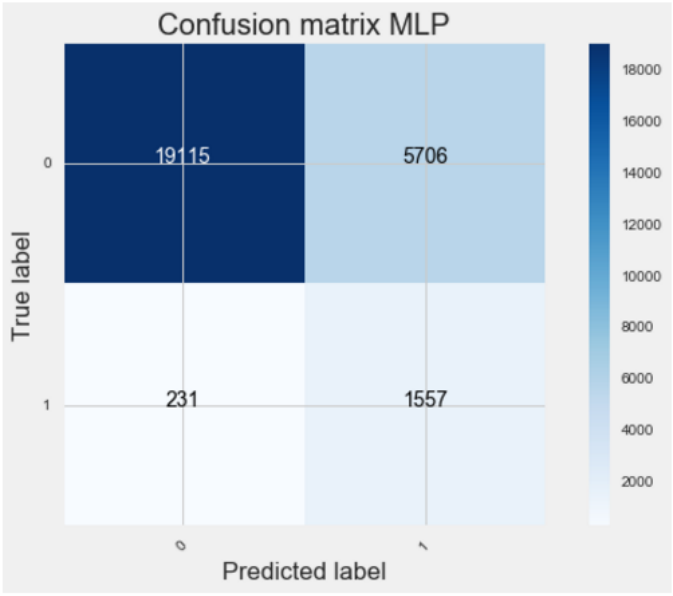

Figura 33 - Matriz de confusão do modelo por MLP – Rede Neural

AUC: 0,820460

Leads não qualificados pelo modelo que foram vendas: 7

- Random Forest

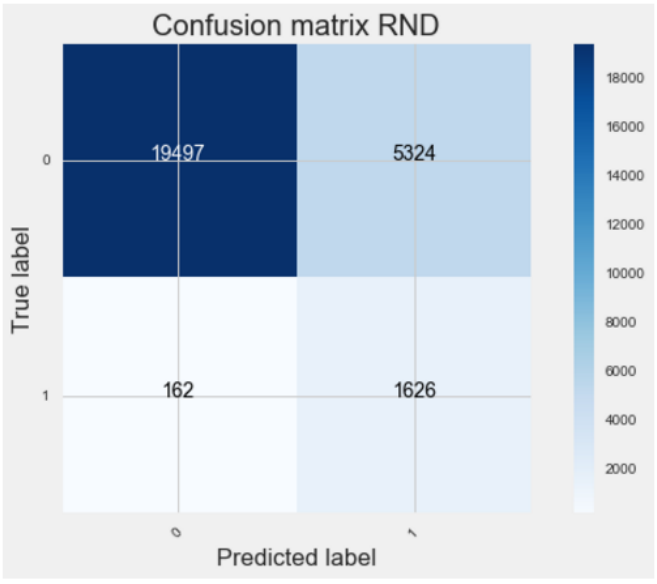

Figura 34 - Matriz de confusão do modelo por Random Forest

AUC: 0,847450

Leads não qualificados pelo modelo que foram vendas: 1

- Blagging

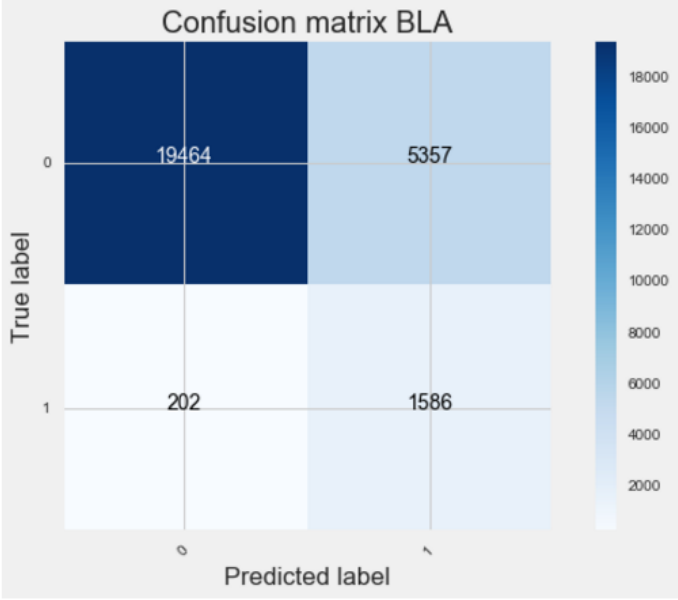

Figura 35 - Matriz de confusão do modelo por Blagging

AUC: 0,835600

Leads não qualificados pelo modelo que foram vendas: 5

- AdaBoost

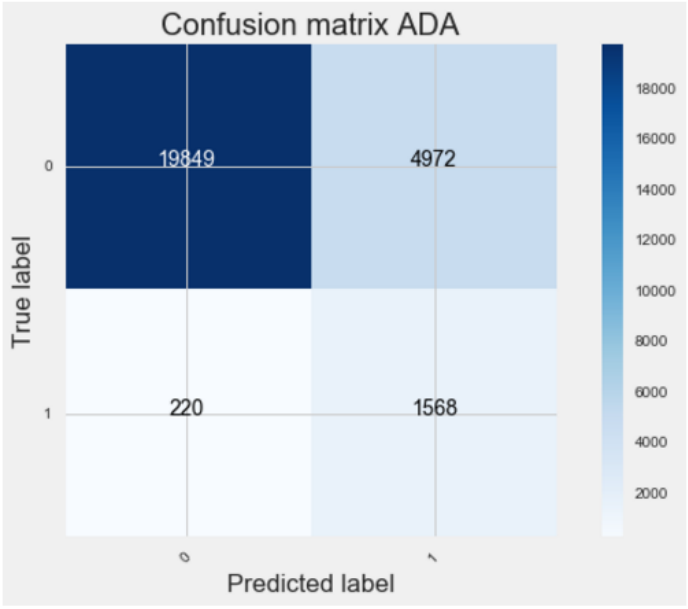

Figura 36 - Matriz de confusão do modelo por AdaBoost

AUC: 0,838322

Leads não qualificados pelo modelo que foram vendas: 8

- KNN

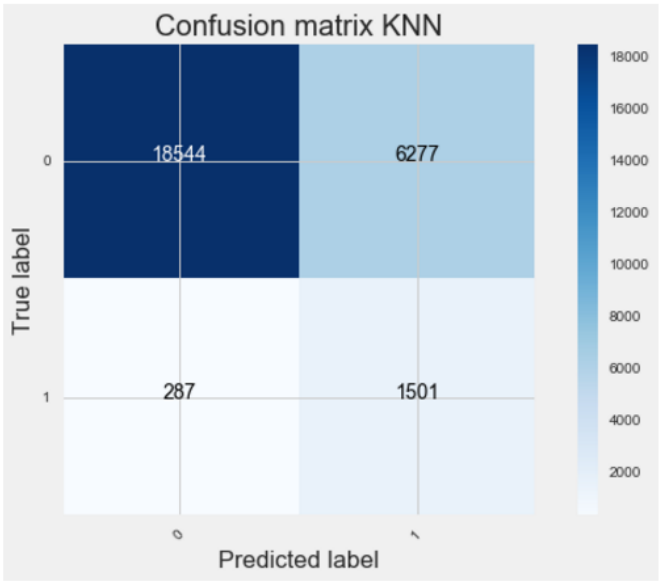

Figura 37 - Matriz de confusão do modelo por KNN

AUC: 0,793297

Leads não qualificados pelo modelo que foram vendas: 12

- Classificador por árvore de decisão

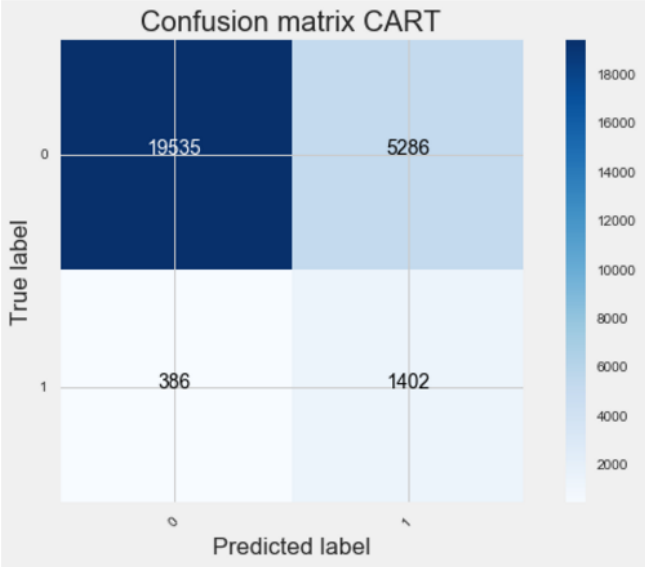

Figura 38 - Matriz de confusão do modelo por Classificador por árvore de decisão

AUC: 0,785576

Leads não qualificados pelo modelo que foram vendas: 15

Por ser o algoritmo que apresenta o maior AUC, o *Gradient Boosting* foi o escolhido. Isto pois, ele apresentou um baixo Falso Positivo e ainda qualificou todos os leads de testes que viraram vendas.

Esse classificador é baseado no princípio de que se pode construir um estimador forte a partir da junção de vários estimadores fracos, havendo minimização de uma função de custo que penaliza a diferença entre os valores obtidos pelo modelo preditivo e os valores medidos.

Esta minimização é aplicada para cada variável selecionada obtendo-se um estimador por cada uma destas variáveis. Após este processo, os estimadores são combinados resultando em uma função de estimadores, a qual será representativa do modelo. [20]

5.2.3.5 Implantação do Lead Scoring Preditivo na operação de vendas

Com o modelo desenvolvido, faz-se necessário implementá-lo no processo de qualificação de leads em tempo real. O código foi inserido no aplicativo de roteamento, onde era feito o cálculo do *lead scoring* tradicional.

O fluxograma representado na Figura 39 expõe a atividade de qualificação e o encaminhamento de um *lead* a partir do RD Station até o Pipedrive.

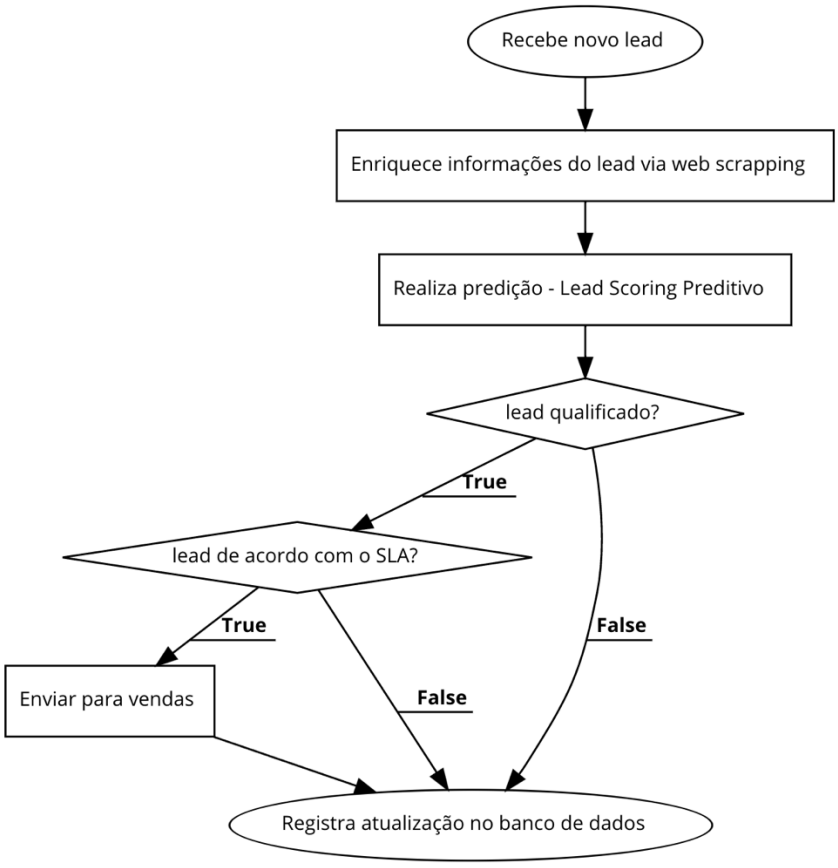

Figura 39 - Fluxograma do aplicativo de roteamento com a implantação do Lead Scoring Preditivo

Durante esse processo, é realizada uma filtragem dos *leads* qualificados de acordo com o SLA[11] da empresa. O SLA é um acordo entre marketing e vendas que define quais os leads que devem e quais não devem ser qualificados como tais. Isso acontece, pois, um lead pode apresentar características que o impossibilitam de

11 Acrônimo de Service Level Agreement – Acordo de nível de serviço

usufruir do produto, como por exemplo microempreendedor individual, empresa sem site e estudantes.

Já o fluxograma da Figura 40 ilustra o funcionamento em específico do Lead Scoring Preditivo.

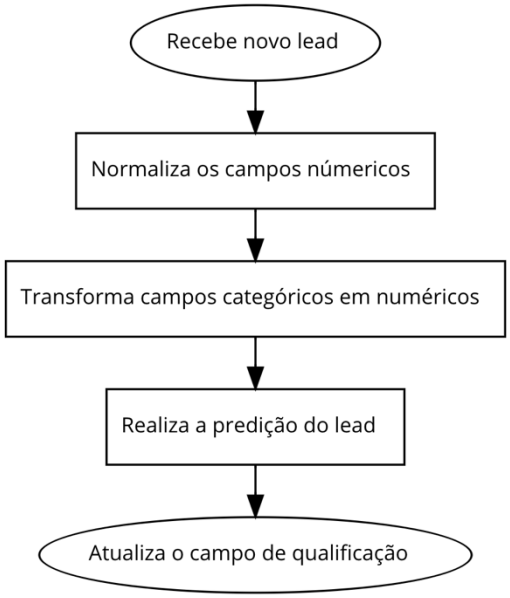

Figura 40 - Fluxograma representando o processo do Lead Scoring Preditivo

6 RESULTADOS

Como resultado final do projeto alcançou-se a implantação de um sistema de BI que visa auxiliar os funcionários da Resultados Digitais, em especial os gestores, a tomar importantes decisões de negócios.

O desenvolvimento do sistema de BI incluiu a estruturação do fluxo de dados com o enriquecimento de informações, bem como o desenvolvimento do aplicativo para gerar a URL de contrato do produto.

Após isso, foi desenvolvido toda a modelagem do *Data Warehouse* através dos processos de Extração, Transformação e Carga (ETL) das informações, da definição do relacionamento entre as tabelas e do desenvolvimento da *dashboard* dentro do Power BI.

Com a estrutura de dados definidas, também foi desenvolvido um algoritmo com base em aprendizado de máquina para a qualificação dos *leads*, auxiliando o time de vendas a alcançar as metas pretendidas.

Esse capítulo apresentará todos os impactos resultantes do presente trabalho, tanto no contexto de *Business Intelligence* quanto em *Machine Learning* aplicado na operação de vendas.

6.1 Business Intelligence

Do ponto de vista prático, a estruturação do fluxo de dados e a utilização da ferramenta permitiu com que gráficos e relatórios mais detalhados e com uma maior confiabilidade pudessem ser gerados. Por consequência, os dados puderam ser melhor analisados e discutidos durante a reunião de acompanhamento de negócios.

Além do desenvolvimento do sistema de BI, a estruturação do fluxo de dados permitiu com que fossem planejados *dashboards* para o time de vendas, marketing e recursos humanos.

A Figura 41 mostra um exemplo com valores fictícios de uma das *dashboard* criadas:

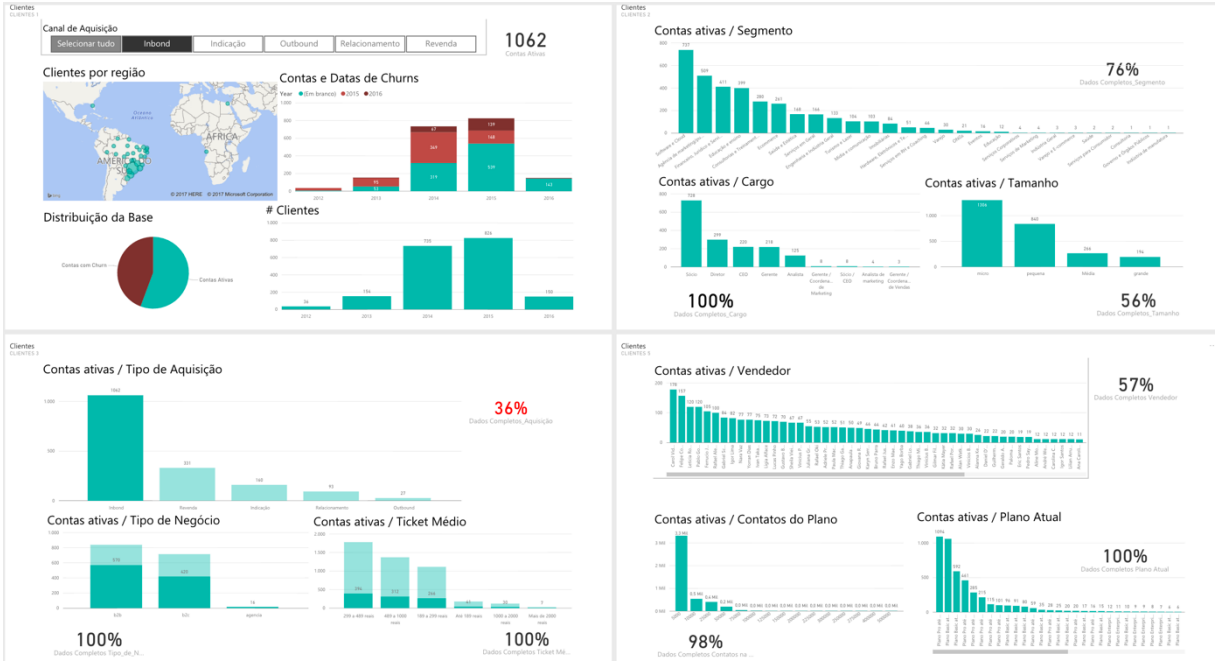

Figura 41 - Exemplo com valores fictícios da dashboard de Gestão de Negócios

No canto superior esquerdo do painel apresenta-se as regiões onde estão os nossos clientes e a distribuição da base de clientes ativos e inativos no decorrer dos anos.

Já no canto superior direito, trata-se de uma análise da base de clientes ativa em relação a três principais características: segmento, cargo e tamanho.

Logo abaixo, no canto inferior direito, é observada a parte financeira. O primeiro gráfico é o número de contas vendidas por vendedor, podendo elas serem filtradas pelos outros dois gráficos, com a distribuição das contas ativas por contatos do plano e plano atual assinado.

Por fim, no canto esquerdo inferior, há a distribuição dos clientes em relação ao tipo de aquisição, tipo de negócio e ticket médio. Nesse mesmo quadrante, na Figura 41, nota-se a funcionalidade de filtragem disponibilizada pelo Power BI, colocando em destaque como é a distribuição do tipo de negócio e ticket médio de acordo com o filtro selecionado no gráfico de tipo de aquisição.

Com um formato de apresentação diferente, também é possível visualizar a *dashboard* em uma versão mobile (tablet ou celular) através do aplicativo do Power BI como é apresentado na Figura 42:

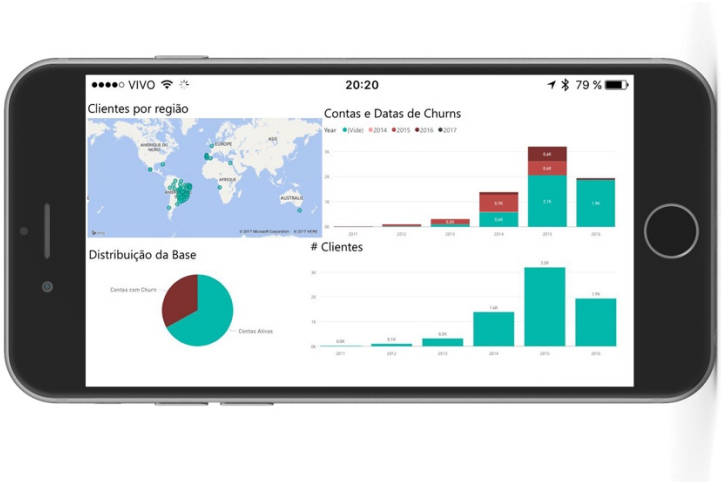

Figura 42 - Exemplo da dashboard na versão mobile

A última etapa do processo foi validar com os gestores a facilidade de entendimento do gráfico para extração de *insights significativos* de maneira rápida. Além disso, foi possível programar o horário e período de atualização dos dados e os usuários que terão acesso ao painel.

Como um dos objetivos do presente trabalho foi implementar uma ferramenta de BI (Business Intelligente) para poder desenvolver melhores e mais complexas análises de dados, a estruturação e a implementação deste sistema deu maior poder e autonomia para os responsáveis de cada área, principalmente no tocando a tomada de decisões, de modo que o tempo investido em cada decisão se tornasse menor e mais eficiente mesmo que não fosse possível quantificar o quão mais eficiente uma pessoa seria mais eficiente em um processo de análise pela complexidade desse tipo de quantificação. Além disso, os processos de elaboração de relatórios mensais se tornaram 50% mais rápidos devido a facilidade da extração das métricas necessárias e da confiabilidade dos dados.

Desta maneira, as análises feitas passaram a trazer resultados mais rápidos e significativos a empresa. Pelo fato de haver apenas uma fonte de dados diminuiu as discussões existentes sobre as formas de extração de dados e, por consequência, este tempo dispendido passou a ser focado em novas oportunidades.

6.2 Aplicação de Machine Learning no processo comercial

A implantação do modelo dentro do processo comercial foi simples, uma vez que o Lead Scoring já existia. Foi necessário apenas substituir o modelo tradicional pelo preditivo.

Apesar de haver resultados satisfatórios com a base de dados de teste, realizar uma mudança substancial no processo de qualificação de *leads* pode acarretar em um grande prejuízo financeiro, caso o modelo não funcione conforme o planejado.

Como solução, foi planejada uma substituição de forma gradual onde, no primeiro mês, o Lead Scoring tradicional e o preditivo funcionaram em paralelo sendo que cada um gerou aproximadamente metade do volume de MQL.

Dessa forma, teve-se a possibilidade de analisar a performance de cada modelo e avaliar como o modelo preditivo funciona na operação em tempo real.

A Tabela 3 ilustra os resultados coletados no mês de dezembro, sendo que os valores absolutos apresentados são fictícios e as taxas são reais:

Tabela 3 - Comparativo dos resultados gerados pelo Lead Scoring já existente e pelo Lead Scoring Preditivo

Lead Scoring Tradicional		**Taxa de eficiência**
MQL	3274	37,5%
SAL	1227	33,3%
Oportunidades	409	42,1%
Vendas	172	

Lead Scoring Preditivo		**Taxa de eficiência**
MQL	3205	46,9%
SAL	1502	43,3%
Oportunidades	650	48,0%
Vendas	312	

Nota-se o aumento da taxa de qualificação de 37,5% com o Lead Scoring Tradicional para 46,9% com o novo modelo devido ao fato que o algoritmo de aprendizado de máquina correlacionar dados de maneira complexa para prever se o

lead é qualificado e, por outro lado, o modelo tradicional de lead scoring é baseado em um estudo estático de como uma variável sozinha impacta na qualificação do lead.

Além disso, com aproximadamente o mesmo número de MQL gerados, foram realizadas 81,3% mais vendas provenientes do Lead Scoring Preditivo, certificando a eficácia do modelo desenvolvido e comprovando o seu aprendizado com a base de dados histórica.

Outro fator importante foi que o modelo preditivo qualificou 164 das 172 vendas provenientes do Lead Scoring Tradicional. Por outro lado, ao avaliarmos a situação oposta, o modelo tradicional qualificou somente 229 das 312 vendas provenientes do Lead Scoring Preditivo. Em uma frequência mensal, o classificador aprende novamente com a base de dados maior e atualizada. Por ser um processo off-line, a velocidade de aprendizado não é um ponto crítico do presente trabalho.

7 CONSIDERAÇÕES FINAIS E PERSPECTIVAS

A partir do trabalho desenvolvido, foi possível atingir o resultado esperado do projeto, ou seja, foi estruturado um fluxo de dados, estes dados foram centralizados e enriquecidos e, após isso, houve a construção de um painel de indicadores, e o desenvolvimento do Lead Scoring Preditivo.

Por trás do desenvolvimento de uma *dashboard* e a comprovação de sua eficiência, há sempre uma base de dados qualificada que visa propiciar novas análises. E é através desta análise qualificada que, no próximo trimestre, mais duas dashboards entraram em fase de concepção e desenvolvimento na empresa: a Dashboard de Resultados de Negócios de Vendas e a Dashboard Operacional de Recrutamento.

Além disso, como forma de um mínimo produto viável, foram desenvolvidas somente as rotinas de atualização do RD Station, do Pipedrive e do RD Contas. Para as próximas atualizações, haverá a centralização de outros softwares utilizados pela empresa, para melhor compreender, por exemplo, a razão dos clientes cancelarem a assinatura do RD Station, por meio de análises com os dados de pós venda, tickets do suporte e uso das ferramentas.

Ao inserir no processo lógicos de negócio maiores e mais complexas, será necessário dar robustez na plataforma de integração dos dados adquirindo um serviço de iPaaS ou tornando a aplicação desenvolvida em um produto. Além disso, com uma maior quantidade de dados, a performance de realizar o processo de ETL e visualização de dados no mesmo software será deteriorado sendo que uma solução no futuro será adotar uma ferramenta para ETL e manter o Power BI somente para a visualização dos dados.

O presente projeto trouxe, a cada etapa, um conhecimento muito grandioso para àqueles que ajudaram a construí-lo. Foram diversas conversas com empresas do Vale do Silício, que geraram boas práticas e novos olhares. Além disso, as participações em eventos e interações com a comunidade de *Business Intelligence e Machine Learning* que, apesar de ainda pequena no País, ajudaram a resolver problemas particulares do mercado brasileiro.

A partir do desenvolvimento deste trabalho, pode-se analisar que os resultados e as aplicações de Machine Learning no processo comercial trazem resultados muito

satisfatórios. Apesar de um bom resultado, foi verificado que muitos leads não possuíam informações consideradas importantes para a qualificação. Para uma melhor aprendizagem do classificador, pode-se buscar formas de enriquecimento desses dados contratando serviços como Datanyze, Clearbit, entre outros.

Com o impacto positivo do modelo, abre-se a possibilidade da criação de um modelo de *Churn* Preditivo, que seria responsável por informar a possibilidade do cliente cancelar a assinatura segundo vários parâmetros como perfil, dados financeiros, dados do suporte e dados de uso do software.

Por fim, somando as soluções implementadas com um grande sucesso, a empresa visa aumentar o número de empregados alocados na área para a evolução cada vez mais ascendente, fornecendo mais inteligência nas decisões tomadas pelos gestores.

Bibliografia

1] F. Sales, “The 7 Secrets of SaaS Startup Success,” [Online]. Available: https://www.salesforce.com/assets/pdf/misc/WP_7Secrets_0408.pdf.

2] NTT Data, “Cloud Computing transforming the Enterprise,” [Online]. Available: http://americas.nttdata.com/services/services/cloud-lifecycle-services/~/media/Documents/white-papers/Cloud-Computing-Transforming-the-Enterprise.pdf.

3] Gartner, “Nine Fatal Flaws in Business Intelligence Implementations,” 10 Outubro 2008. [Online]. Available: http://www.gartner.com/newsroom/id/774912.

4] Looker, “Is It Worth Centralizing My Data?,” 2 Maio 2016. [Online]. Available: https://looker.com/blog/centralizing-your-data.

5] Resultados Digitais, “Inbound Marketing,” [Online]. Available: http://resultadosdigitais.com.br/inbound-marketing/.

6] B. Halligan e D. Shah, Inbound Marketing: seja encontrado usando o Google, a mídia social e os blogs, Alta Books, 2009.

7] G. Costa, “http://resultadosdigitais.com.br/blog/lead-scoring-guia-completo/,” [Online].

8] V. P. F. d. Pinho, “SaaS: Análise de impacto na transformação da investigação e desenvolvimento de produto para serviço,” 20 07 2007. [Online]. Available: https://repositorio-aberto.up.pt/bitstream/10216/59412/1/000134894.pdf.

9] H. e. S. A. KORTH, Sistemas de Bancos de Dados, 2a. edição revisada ed., Makron Books, 1994.

10] B. Kyle., MongoDB in Action, Manning, 2011.

11] A. Kearney, “Better Decision Making with Proper Business Intelligence,” [Online]. Available: https://www.google.com.br/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwiI38HzvYnTAhWCgJAKHRcvCHoQFggcMAA&url=https%3A%2F%2Fwww.atkearney.com%2Fdocuments%2F10192%2F247903%2F

Better_Decision_Making_with_Proper_Business_Intelligence.pdf%2Fe55e6880-ed1b-4b25-a0b6-33b94c0cc641&usg=AFQjCNH4Pvks81gHrKPkuZLwaW0DvnDelg&sig2=LN0e-ce1uSxSWNKf27-lNA.

[12] Looker, “Is It Worth Centralizing My Data?,” [Online]. Available: https://looker.com/blog/centralizing-your-data. [Acesso em 18 10 2016].

[13] P. Harrington, Machine Learning in Action, Shelter Island, New York: Manning Publications Co., 2012.

[14] R. S. Wazlawick, Object-Oriented Analysis and Design for Information Systems. Modeling with UML, OCL, and IFML, Morgan Kaufmann, 2014.

[15] T. Elliot, “Why BI Projects Fail,” 1 11 2011. [Online]. Available: http://assets.timoelliott.com/docs/why_bi_projects_fail_2012.pdf. [Acesso em 4 1 2017].

[16] Juice Analytics, “A Guide to Creating Dashboards People Love to Use,” 1 11 2009. [Online]. Available: http://www.cpoc.org/assets/Data/guide_to_dashboard_design1.pdf. [Acesso em 19 01 2017].

[17] Resultados Digitais, “Guia Completo de Inside Sales,” 01 12 2015. [Online]. Available: https://d335luupugsy2.cloudfront.net/cms%2Ffiles%2F2%2F1436190926O+guia+completo+de+Inside+Sales+%281%29.pdf. [Acesso em 25 12 2016].

[18] N. Chawla, N. Japkowicz e A. Kotcz, “Editorial: Special Issue on Learning from Imbalanced Data Sets,” 1 1 2014. [Online]. Available: https://www3.nd.edu/~dial/publications/chawla2004editorial.pdf. [Acesso em 18 12 2016].

[19] G. H. M.A. Hall, “Benchmarking attribute selection techniques for discrete class data set mining,” *IEEE Transactions on Knowledge and Data Engineering,* pp. 1437 - 1447, 17 11 2003.

[20] J. H. Friedman, “Sthochastic Gradient Boosting,” 26 03 1999. [Online]. Available: http://statweb.stanford.edu/~jhf/ftp/stobst.pdf. [Acesso em 30 01 2017].