UNIVERSIDADE DE SÃO PAULO
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO E CONTABILIDADE
DEPARTAMENTO DE ADMINISTRAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM MÉTODOS QUANTITATIVOS E INFORMÁTICA

MARCO ANTONIO LOPES

APLICAÇÃO DE APRENDIZADO DE MÁQUINA NA DETECÇÃO DE FRAUDES PÚBLICAS

São Paulo

2019

Prof. Dr. Vahan Agopyan
Reitor da Universidade de São Paulo

Prof. Dr. Adalberto Américo Fischmann
Diretor da Faculdade de Economia, Administração e Contabilidade

Prof. Dr. Moacir de Miranda Oliveira Júnior
Chefe do Departamento de Administração

Prof. Dr. Eduardo Kazuo Kayo
Coordenador do Programa de Pós-Graduação em Administração

MARCO ANTONIO LOPES

APLICAÇÃO DE APRENDIZADO DE MÁQUINA NA DETECÇÃO DE FRAUDES PÚBLICAS

Dissertação apresentada ao Programa de Pós-Graduação em Métodos Quantitativos e Informática da Faculdade de Economia, Administração e Contabilidade, da Universidade de São Paulo, como requisito parcial para a obtenção do título de Mestre em Ciências.

Orientadora: Prof.ª Dr.ª Alessandra de Ávila Montini

Versão corrigida

São Paulo

2019

Catalogação na Publicação (CIP)
Ficha Catalográfica com dados inseridos pelo autor

Lopes, Marco Antonio.
Aplicação de Aprendizado de Máquina na Detecção de Fraudes Públicas / Marco Antonio Lopes. - São Paulo, 2019.
68 p.

Dissertação (Mestrado) - Universidade de São Paulo, 2019.
Orientador: Alessandra de Ávila Montini.

1. inteligência artificial. 2. machine learning. 3. big data. 4. fraudes. 5. eventos raros. I. Universidade de São Paulo. Faculdade de Economia, Administração e Contabilidade. II. Título.

MARCO ANTONIO LOPES

APLICAÇÃO DE APRENDIZADO DE MÁQUINA NA DETECÇÃO DE FRAUDES PÚBLICAS

Dissertação apresentada ao Programa de Pós-Graduação em Métodos Quantitativos e Informática da Faculdade de Economia, Administração e Contabilidade, da Universidade de São Paulo, como requisito parcial para a obtenção do título de Mestre em Ciências.

Aprovado em: ____ de _____de 2019.

Banca Examinadora:

_____

Prof.ª Dr.ª Alessandra de Ávila Montini
Universidade de São Paulo

_____

Prof. Dr. Adolpho Walter Pimazoni Canton
Universidade de São Paulo

_____

Dr. Vitor Hugo Louzada Patrício
ETH Zürich

_____

Prof. ª Dr. ª Natália Cordeiro Zaniboni
Universidade de São Paulo

Dedico este trabalho aos meus pais, irmão, esposa e filhos.

Obrigado por sempre me incentivarem a buscar o meu melhor.

AGRADECIMENTOS

Aos meus pais, José Carlos Lopes e Vera Lucia Neri Lopes, que nunca mediram esforços para educar a mim e a meu irmão. Obrigado por não perderem a determinação mesmo nos momentos mais difíceis. Saibam que sempre me dedicarei para honrar todo o amor que me deram de forma tão generosa. Espero ser um pai tão bom aos meus filhos quanto vocês são para mim.

Ao meu irmão, Diego Vinicius Lopes, por sempre estar ao meu lado mesmo que a distância nos separe. Obrigado por sempre me pressionar a alcançar os meus sonhos e pela amizade de sempre.

À minha melhor amiga e esposa, Leila Dias Franco Lopes, por todo amor, cumplicidade e companheirismo dedicados. Enquanto eu escrevia este trabalho ela foi responsável por gerar três vidas dentro de si. Sempre serei grato por seu amor e sua dedicação, por me ajudar a criar uma família e por ter me transformado em pai. Guilherme, Daniel e Mariana, obrigado por terem esperado a conclusão deste trabalho para nascer.

À minha orientadora, Alessandra de Ávila Montini, pelas orientações, paciência e dedicação em ensinar. A maneira como ensina e lida com seus alunos, sem economizar energia e amor, será sempre uma inspiração para mim.

Aos amigos de São Paulo, Felipe Duarte, Karina Bindandi, Diego Silva, Karen Tanaka e Vitor Hugo. Aos amigos de Indaiatuba, Filipe Pirão, Thays Rodrigues, Carlos Ferro, Katia Mendes, Marco Candello, Andressa Armelim, Bruno Amstalden, Mariana Galvão, Sidnei Caus, Andréa Garcia e Gustavo Razzera, Fernando e Juliana. Por tornarem a vida mais leve com momentos que serão sempre recordados.

Aos colegas de trabalho que tornaram possível a tarefa de conduzir um mestrado atuando como professional. Agradeço especialmente aos líderes que tive. Eduardo Raul Hruschka, Diego Silva e Vitor Azeka, obrigado por terem investido em minha formação e minha carreira. Obrigado também ao meu time por entenderem quando precisei me ausentar.

À Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo, por tornar possível o sonho de me tornar Mestre. Obrigado a todos os professores por fazerem parte da minha formação acadêmica.

RESUMO

Nos últimos anos, os governos mundiais vêm participando de esforços conjuntos para aumentar a disponibilidade dos dados governamentais para seus cidadãos, e o resultado disso, no Brasil, foi a criação do Portal Brasileiro de Dados Abertos. Com mais dados disponíveis ao cidadão comum, várias análises que não são feitas pelo governo, em razão da falta de conhecimento ou de interesse, tornam-se possíveis, como, por exemplo, a identificação de fraudes em licitações públicas. Uma forma de identificar os padrões existentes nessas fraudes é o uso de aprendizado de máquina. Atualmente, existem *softwares* como R e Python que permitem o uso de diversas técnicas de aprendizado de máquina já implementadas. Esses *softwares*, devido à sua grande capacidade de processamento, também, podem auxiliar em problemas com dados desbalanceados, em que a ocorrência do evento que está sendo estudado é muito rara, como é o caso de fraudes. Assim, um exemplo desse tipo de problema e que é alvo do estudo desta dissertação é a detecção de fraudes em sistemas públicos por meio da descoberta de contratos que pertencem a empresas inidôneas. Tal desafio pode potencializar-se com grandes volumes de dados, visto que podem tornar o processamento dessas bases mais complexo. Assim, esta dissertação visa contribuir para a resolução desse problema propondo avaliar metodologias e técnicas de aprendizado de máquina que apresentam resultados satisfatórios nesse cenário.

Palavras-chave: *big data*, eventos raros, dados desbalanceados, *machine learning*, aprendizado de máquina, fraudes públicas, licitações, inteligência artificial.

ABSTRACT

In recent years, world governments have been participating in joint efforts to increase the availability of government data for their citizens, such as the creation of the Brazilian Open Data Portal. With more data available to the average citizen, several analyzes become possible, for example, the identification of fraud in public bids. One way to identify patterns in these scams is to use machine learning. These techniques can aid in the analysis of problems with unbalanced data, where the occurrence of the event being studied is very rare, as is the case of frauds. An example of this type of problem that is the subject of this dissertation is the detection of fraud in public systems discovering public contracts belonging to untrusted companies. Thus, this work uses public data for the identification of frauds to the public patrimony.

Palavras-chave: *big data*, rare events, imbalanced datasets, *machine learning*, public frauds, biddings, artificial intelligence.

LISTA DE FIGURAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

Com uma sociedade cada vez mais conectada e tecnológica, até mesmo tarefas simples, como realizar um pagamento no Internet Banking, podem gerar grande quantidade de dados. Nos processos de compras do governo essa também é uma realidade e, hoje, pode-se contar com bases de dados que não existiam num passado recente. E tais dados podem ser analisados a fim de se obter valor a partir dessas informações, como disponibilizar um serviço mais personalizado para seu cliente, reduzindo custos operacionais ou evitando perdas com fraudes, por exemplo.

O processo de compras governamentais, normalmente, utiliza-se de licitações para produtos e serviços de elevado valor. Licitação é o procedimento administrativo formal que estabelece a forma como as contratações de serviços e aquisições de produtos devem acontecer. No Brasil, para licitações por entidades que façam uso da verba pública, o processo é regulado pelas Leis Federais n.^{os} 8.666/93 e 10.520/02.

Apesar da regulamentação, o processo está sujeito a fraudes, tais como pagamento de propinas para manipular o processo licitatório. Entre as formas que isso pode acontecer estão: edição do edital de forma a privilegiar determinada empresa, processos licitatórios contendo apenas uma empresa ou informar determinada empresa sobre os lances das concorrentes através do vazamento de informações do processo licitatório por funcionários públicos.

Para Cressey (1953), estudioso do combate ao crime organizado, a fraude acontece quando existem três pilares: oportunidade, motivação e racionalização. Sendo que a motivação, também chamada de pressão, é a situação na qual o fraudador possui problemas financeiros não compartilhados que poderiam incentivar a prática de atos corruptos afim de solucionar tais problemas financeiros. A oportunidade vem da existência de uma relação de confiança da instituição com o fraudador, além da existência do conhecimento necessário para a realização da fraude. Já a racionalização é o processo em que o fraudador interpreta a ação da fraude como aceitável ou justificável (Machado & Gartner, 2017). Assim, por exemplo, o fraudador poderia ser uma pessoa que se encontra em uma situação de falta de dinheiro, conhece o sistema a ponto de saber suas fraquezas e racionaliza suas atitudes com o intuito de amenizar o ato ilícito que está cometendo.

Antes de se evitar uma fraude é necessário conhecer o padrão que determina se uma transação é ou não fraudulenta. Essa tarefa é difícil por diversos motivos, entre eles: as transações fraudulentas são muito semelhantes a transações não fraudulentas, a quantidade de

fraudes é muito pequena com relação ao total de transações observadas, o comportamento dos usuários é muito diversificado, o comportamento do fraudador se altera constantemente e, por fim, a necessidade de detectar a fraude em tempo hábil.

Dessa forma, este trabalho almeja estudar técnicas de aprendizado de máquina disponíveis que se mostrem viáveis na detecção de compras públicas fraudulentas ao analisar um grande volume de dados. Para isso, foram utilizados *softwares* de código aberto que são bastante difundidos no mercado, como R e Python. A escolha deu-se devido ao grande poder computacional desses *softwares* e por contarem com uma gama de algoritmos já implementada pela comunidade que mantém o produto. Assim, é possível testar diversas técnicas de classificação, dado que ambas as linguagens contam com um grande repositório de bibliotecas com diferentes algoritmos já implementados.

O escopo do trabalho abrangeu a utilização de técnicas de aprendizado de máquina com dados públicos, como os disponíveis no Portal Transparência (Brasil, 2019), de instituições públicas que realizaram compras através de processos licitatórios. No ano de 2011, a Lei de Acesso à Informação tornou os dados sobre os gastos públicos abertos à sociedade civil, com isso, possibilitando a realização do estudo.

A base de dados utilizada neste trabalho conta com dados disponibilizados por duas agências do governo: Ministério do Planejamento, Desenvolvimento e Gestão (MP) e Controladoria Geral da União (CGU). Criado em 1962, o MP é responsável por planejar a administração governamental, os custos, analisar a viabilidade de projetos, controlar orçamentos, liberar fundos para estados e projetos do governo. Por isso, detém os dados de todas as compras realizadas pelo governo federal, tanto compras feitas com licitações quanto compras diretas. A CGU foi criada, em 2003, com o objetivo de providenciar incremento da transparência da gestão, por meio das atividades de controle interno, auditoria pública, prevenção e combate à corrupção e ouvidoria. Dados sobre as empresas que tiveram afastamento ou proibição de participarem de processos de compras abertos pelo governo federal são mantidos e publicados sob responsabilidade da CGU.

Ambas as bases de dados foram obtidas por meio do portal dados.gov.br (Brasil, 2018). Esse portal foi construído para atender aos objetivos da Lei de Acesso à Informação Pública (Lei n.º 12.527/2011) e à Política de Dados Abertos, que foi consolidada pelo Decreto n.º 8.777, de 2016. Essas iniciativas integram um esforço global, do qual o Brasil faz parte, para que aumente a liberdade de acesso dos cidadãos aos dados públicos dirigida pela organização Open Government Partnership.

Dessa forma, este trabalho pretende responder à seguinte questão teórica: O aprendizado de máquina pode auxiliar na detecção de fraudes em contas públicas?

Para responder a essa questão o objetivo geral deste trabalho consistiu em avaliar se a utilização de técnicas de inteligência artificial (IA) apresenta resultados satisfatórios na classificação de gastos públicos como fraudulentos ou não. A técnica de IA escolhida foi o aprendizado de máquina.

Esse objetivo dividiu-se em objetivos específicos:

- Definir as variáveis explicativas do modelo.
- Elaborar modelos utilizando técnicas de *machine learning* (regressão logística, árvore de decisão, *random forest*, *gradient boosting* e *lightGBM*).
- Analisar qual modelo teve a melhor *performance* quanto à sua precisão e capacidade de detecção.

Com este estudo almejou-se propiciar uma ferramenta de combate a fraudes contra o patrimônio público que não exija pesados investimentos em processos de auditoria demorados e passíveis de corrupção. Esta ferramenta trata-se de um modelo capaz de realizar a classificação dos contratos entre o governo e a iniciativa privada em duas classes: “o contrato pertence à empresa inidônea” e “o contrato não pertence à empresa inidônea”. A determinação de uma empresa inidônea, ou seja, que não é confiável para se estabelecer um contrato com a administração pública, é feita pela Controladoria Geral da União. Neste trabalho também foram estudadas técnicas de aprendizado de máquina em dados que apresentam desbalanceamento entre as classes, como é o caso das fraudes em contratos públicos.

2 REVISÃO BIBLIOGRÁFICA

Como a ocorrência de fraudes públicas trata-se de um evento que acontece numa pequena parcela dos gastos públicos, este trabalho estudou a literatura sobre detecção de eventos raros. Nesta seção, são apresentados os principais conceitos e as definições sobre eventos raros e as dificuldades envolvidas em sua detecção, metodologias tradicionais para se trabalhar com dados com essas características, algoritmos de aprendizado de máquina presentes em *softwares* de análise de dados e como esses algoritmos podem ser utilizados na predição de eventos raros.

2.1 PROCESSO DE COMPRAS GOVERNAMENTAIS

As compras de produtos e serviços realizadas pela Administração Pública são feitas por meio de licitações, ou seja, um procedimento para a seleção da proposta comercial que é mais atrativo para o interesse público. Além disso, os administradores responsáveis por conduzir esse processo devem seguir alguns princípios, conforme a Lei n.º 8666/93:

> Art. 3 A licitação destina-se a garantir a observância do princípio constitucional da isonomia, a seleção da proposta mais vantajosa para a administração e a promoção do desenvolvimento nacional sustentável e será processada e julgada em estrita conformidade com os princípios básicos da legalidade, da impessoalidade, da moralidade, da igualdade, da publicidade, da probidade administrativa, da vinculação ao instrumento convocatório, do julgamento objetivo e dos que lhes são correlatos.

Segundo o manual de Licitações e Contratos do Tribunal de Contas da União (Brasil, 2010), os processos licitatórios brasileiros são regidos por algumas leis. A Lei n.º 8.666, de 21 de junho de 1993, Lei de Licitações e Contratos Administrativos, e a Lei n.º 10.520, de 17 de julho de 2002, Lei do Pregão, constituem a legislação básica sobre licitações e contratos para a Administração Pública. A Lei n.º 8666/93 define as seguintes modalidades de licitação: concorrência, tomada de preço, carta-convite, leilão e concurso. O pregão eletrônico ou presencial é definido pela Lei n.º 10520/02. Os Quadros 1 e 2, a seguir, apresentam os limites de valor exigidos para a utilização de cada modalidade.

Na modalidade carta-convite, a administração deve convidar pelo menos três fornecedores interessados, cadastrados ou não na Unidade Administrativa, que pertençam ao ramo de negócio do objeto a ser contratado. Caso não haja interesse de pelo menos três fornecedores ou o mercado seja limitado a um número menor de fornecedores, as circunstâncias devem ser explicadas no processo.

Durante a tomada de preços, é exigido que os fornecedores interessados estejam cadastrados até três dias antes do envio de propostas. Já a concorrência pode ser atendida por quaisquer interessados que possuam os requisitos mínimos apresentados em edital. É a modalidade exigida para compra de imóveis, para a alienação de imóveis públicos, para a concessão de direito real de uso, para as licitações internacionais, para a celebração de contratos de concessão de serviços públicos e para os contratos de parceria público-privada.

A lei também prevê a realização de concursos para a contratação de pessoas físicas e de leilões, principalmente, para a venda de imóveis que não sejam mais utilizados pelo governo. Aqui é importante diferenciar leilões de pregões, sendo que, em um leilão, participam diversos compradores e existe apenas um vendedor, no caso, o governo; já em um pregão, há diversos vendedores e apenas um comprador, também, o governo. O pregão também pode ser chamado, em alguns estudos, de leilão reverso, termo que foi utilizado neste trabalho.

Com exceção do pregão, as modalidades apresentadas anteriormente são explicadas para a população a partir de uma cartilha escrita pelo Sebrae (2014). O pregão é a modalidade de licitação utilizada para aquisição de bens e serviços comuns de qualquer valor em que a disputa pelo fornecimento é feita em sessão pública, por meio de propostas e lances, para classificação e habilitação do licitante com a proposta de menor preço, podendo ser realizada de maneira presencial ou eletrônica, usando sistemas de informações próprios para a atividade. Nessa modalidade é analisada apenas a documentação do fornecedor vencedor do pregão, simplificando o processo de compra. Por ser o processo de compra mais simples, é também o mais utilizado, como pode-se verificar mais adiante neste trabalho.

Quadro 1 - Valores para obras e serviços de engenharia

Modalidade	Valor
Carta-Convite	Até R$150.000,00
Tomada de Preço	Até R$1.500.000,00
Concorrência	Acima de R$1.500.000,00
Pregão	Sem limite de valor

Fonte: SEBRAE (2014)

Quadro 2 - Valores para compras e serviços

Modalidade	Valor
Carta-Convite	Até R$80.000,00
Tomada de Preço	Até R$650.000,00
Concorrência	Acima de R$650.000,00
Pregão	Sem limite de valor

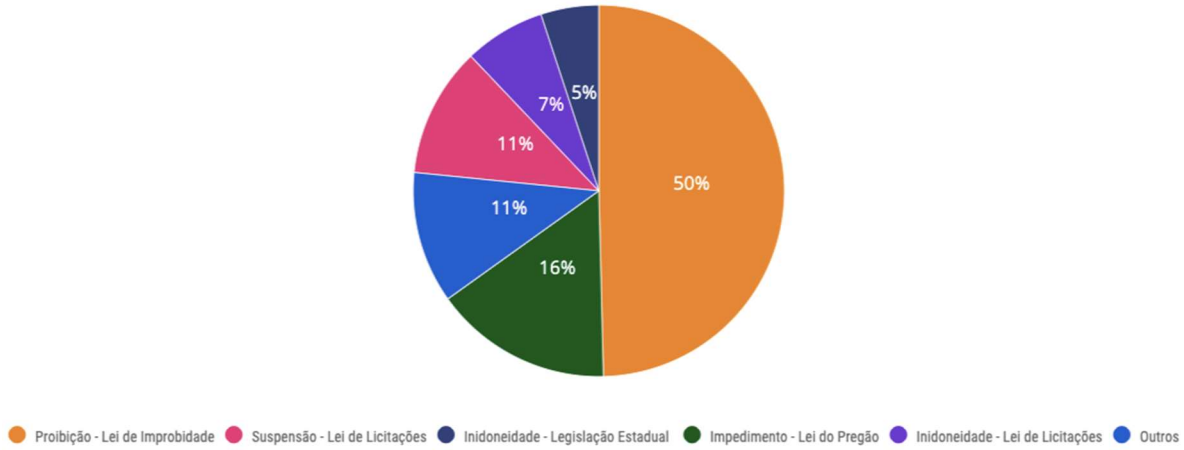
Fonte: SEBRAE (2014)

2.2 EMPRESAS INIDÔNEAS

A Controladoria Geral da União, mediante o Sistema Integrado de Registro do CEIS/CNEP (SIRCAD), mantém o cadastro de empresas que estão impedidas de realizar acordos comerciais com o governo ou foram punidas por exercerem condutas contrárias ao esperado. A inserção de empresas nesse cadastro é feita por entes públicos a fim de garantir a fidedignidade dos dados.

O SIRCAD apresenta, no Portal da Transparência, os dados do Cadastro Nacional de Empresas Inidôneas e Suspensas (CEIS) e do Cadastro Nacional das Empresas Punidas (CNEP), que estejam vigentes, atendendo às determinações da Lei n.º 12.846/2013 (Lei Anticorrupção). O CEIS consolida, em uma base de dados, as empresas e as pessoas físicas que sofreram sanções e tiveram restringido o direito de participar de licitações ou de celebrar contratos com a Administração Pública. Já o CNEP busca consolidar as sanções aplicadas a pessoas jurídicas pela prática de atos lesivos, com base na Lei n.º 12.846/2013.

Figura 1 - Proporção de empresas punidas pelas leis



Fonte: Brasil (2019a)

Segundo os dados do Portal da Transparência (Brasil, 2019a), 50% das empresas cadastradas no CEIS infringiram a Lei de Improbidade Administrativa (Lei n.º 8.429, 1992). A segunda maior responsável pelo cadastramento de empresas inidôneas é a Lei do Pregão (Lei n.º 10.520, 2002), sendo 16% das empresas do CEIS. A Lei de Licitações (Lei n.º 8666, 1993) foi quebrada por 18% das empresas, sendo que 11% foram suspensas e 7% foram consideradas inidôneas. Os 16% restantes infringiram leis estaduais ou outras leis.

Neste trabalho, foram consideradas apenas as empresas cadastradas no CEIS por ter a intenção de identificar contratos que pertençam a empresas que foram impedidas de negociar com o governo. As empresas cadastradas no CNEP são punidas mas não necessariamente impedidas de negociar com o governo, dessa forma, não foram considerados contratos com pessoas físicas ou empresas cadastradas no CNEP.

2.3 CORRUPÇÃO EM CONTRATOS PÚBLICOS

A prática de corrupção no Brasil é encarada como crime de improbidade administrativa e pode implicar na suspensão dos direitos políticos, na perda da função pública, na indisponibilidade de bens e no ressarcimento ao erário (Faria & Bianchi, 2018). Castro (2007) apresenta uma discussão sobre as condições econômicas e as práticas administrativas que aumentam a probabilidade de abrir oportunidade de se cometerem fraudes públicas em licitações. Seu trabalho propõe a redução de dificuldades para a participação de fornecedores em licitações evitando a ocorrência de dispensas de licitações, além de facilitar o monitoramento dos processos, assim, aumentando a transparência do processo licitatório.

Apesar de uma das atribuições dos processos licitatórios ser evitar a corrupção, algumas vezes, eles podem facilitar a seleção de determinados fornecedores com projetos não necessariamente semelhantes, desrespeitando critérios de isonomia que permitiriam a participação de todos os membros do mercado. Esse é o caso das contratações realizadas em Regime Diferenciado de Contratações Públicas (Ponte & Heringer, 2019)

Visto que o processo licitatório não é uma garantia da inexistência da corrupção e, em alguns casos, pode até aumentar a probabilidade de ocorrência, adotar a simplificação dos processos e a desburocratização pode trazer economia para o Estado e incentivar a economia como um todo, considerando que o governo é um grande consumidor de produtos e serviços, sem necessariamente aumentar a corrupção (de Albuquerque Nobrega & Brito, 2019).

Vasconcelos (2005) indica dados que corroboram essa visão, pois, analisando o uso da modalidade pregão, pôde verificar que o Estado de São Paulo conseguiu reduzir os gastos da Secretaria de Procuradoria em mais de 50% e, em 33,2%, na Secretaria da Educação. Essa eficiência é responsável por tornar tal modalidade a preferida da Administração Pública. Apesar de os pregões mostrarem-se ferramentas eficientes para a redução de custos governamentais, eles trazem preocupações pela quebra da hierarquia normativa e pelo aumento da facilidade de existência de conluio.

Na visão dos pregoeiros, os mecanismos de controle mais importantes para evitar a corrupção em pregões são a existência de um *check-list* e publicidade e propaganda. O *check-list* é uma lista de verificação para viabilizar a contratação por meio de diretrizes únicas e organizadas. O mecanismo de publicidade e propaganda é a obrigação da Administração Pública em divulgar, de maneira ampla, os editais de licitações na rede mundial de computadores. Como a responsabilidade da aplicação desses mecanismos é dos pregoeiros, o investimento em treinamento e educação para esses profissionais é crucial para condução íntegra dos processos licitatórios, assim, evitando a prática de corrupção (Aquino M., Sousa, Cavalcante, Duarte, & Aquino, F., 2018).

Nesse cenário, a existência de programas de *compliance* é de suma importância para evitar que funcionários da Administração Pública e representantes das empresas fornecedoras possam usar de manobras ilegais para garantir que determinada empresa se sagre vencedora da licitação de antemão. Quanto maior a transparência e a fiscalização entre pares, menores serão as chances de processos de corrupção efetivarem-se (Schramm, 2018).

A burocracia com processos rígidos nem sempre significa que uma fraude será evitada. Os processos de corrupção contam com a cumplicidade daqueles que deveriam zelar pelo

patrimônio público. Portanto, é preciso reduzir a burocracia para firmar um contrato de fornecimento entre a iniciativa privada e a Administração Pública, assim, aumentando a qualidade, eficiência e transparência dos mecanismos de controle e tornando possível que qualquer cidadão possa auditar de maneira autônoma os contratos firmados pelo governo (Dematté, 2009).

2.4 DIFICULDADES NA DETECÇÃO DE EVENTOS RAROS

Alguns problemas que apresentam a ocorrência de eventos raros são, por exemplo, a ocorrência de guerras entre países, epidemias, falhas de componentes eletrônicos ou mecânicos e de respostas a campanhas de *marketing*, cancelamentos de serviços ou uma transação fraudulenta. Por serem raros, esses eventos podem parecer completamente aleatórios.

Encontrar padrões em dados em que a ocorrência do evento que está sendo estudado é muitas vezes menor do que a não existência do evento é um problema de difícil solução. Como colocado por King e Zeng (2001), essa dificuldade pode acontecer por dois motivos: a probabilidade de o evento acontecer pode ser subestimada devido à quantidade muito menor de observações na base de dados, o segundo é que as estratégias mais comumente utilizadas são ineficientes em problemas como esse.

Conforme Van der Paal (2014), pode-se dividir a raridade dos eventos em dois tipos: raridade relativa e raridade absoluta. A raridade relativa, também, pode ser chamada de dados desbalanceados e acontece quando a classe que está sendo estudada ocorre em uma proporção muito pequena em relação à quantidade total de observações. Esse tipo de base de dados pode apresentar problemas com algoritmos de classificação, que tendem a diminuir sua importância. Já os eventos com raridade absoluta apresentam amostras de tamanho reduzido. Para saber se determinada amostra apresenta observações raras, Agresti (2013) aponta como sugestão analisar a quantidade de observações pela quantidade de variáveis existentes no problema. Allison (2012) defende que apenas este tipo de raridade influencia os resultados da regressão logística.

2.5 METODOLOGIA TRADICIONAL PARA DETECTAR EVENTOS RAROS

Prati, Batista e Monard (2009) publicaram seus estudos sobre mineração de dados em problemas com classes desbalanceadas, e uma das sugestões apresentadas é abordar o problema

fazendo amostragens de tal forma que as classes fiquem balanceadas. Há duas formas de conseguir esse resultado. Uma forma, chamada de sobreamostragem aleatória, consiste em fazer cópias aleatórias da classe com poucos casos, de forma que fiquem balanceadas. A outra é chamada de subamostragem aleatória. Neste caso, observações aleatórias da classe majoritária são excluídas da amostra até que a base fique balanceada.

As duas técnicas apresentam problemas. No caso da sobreamostragem aleatória, como são feitas cópias exatas de algumas observações, corre-se o risco de criar um modelo com *overfitting*, ou seja, o modelo funciona bem com os dados amostrados, mas não garante uma boa taxa de acerto com dados reais. Já a subamostragem aleatória pode eliminar observações que seriam importantes para que o modelo aprenda sobre os padrões dos dados.

Para diminuir esses problemas, algumas técnicas podem ser utilizadas. Um exemplo utilizado na sobreamostragem aleatória é a técnica chamada Synthetic Minority Over-sampling Technique (SMOTE), que consiste em gerar dados sintéticos ao fazer a interpolação dos dados da classe minoritária. Isso diminui o *overfitting* do modelo, visto que os dados que são adicionados à amostra não são cópias exatas das observações da classe estudada.

No caso da subamostragem aleatória, podem ser utilizados métodos para manter observações que estejam em regiões fronteiriças com a classe estudada a fim de melhorar a representação da região de borda. Outros métodos, pelo contrário, tendem a eliminar dados que possam causar confusão ao modelo, assim, são removidos aqueles que possam se parecer com dados da classe minoritária.

Outra forma de lidar com a modelagem de eventos raros foi proposta por Firth, em 1993, que descreve uma maneira de corrigir o viés causado pela presença muito maior de determinada classe. Esse método ficou conhecido como Máxima Verossimilhança Penalizada ou Correção de Viés de Firth.

2.6 APRENDIZADO DE MÁQUINA

Aprendizado de máquina é um ramo da inteligência artificial que estuda como os sistemas computacionais podem aprender a resolver determinados tipos de problema. De maneira muito simplificada, imagine-se que um algoritmo de aprendizado de máquina resolve um problema diversas vezes, sempre, utilizando uma medida de avaliação. Dessa forma, o algoritmo adapta-se para otimizar essa medida e aprender a resolver o problema. Monard e

Baranauskas (2003) apresentam os conceitos envolvidos em aprendizado de máquina de uma forma muito clara no capítulo quatro do livro "Sistemas Inteligentes para Engenharia", de 2003.

Os modelos de aprendizado de máquina podem ser divididos em dois tipos de aprendizados: supervisionado e não supervisionado. No aprendizado supervisionado, a base contém uma variável observada que indica qual a classificação de cada registro. No caso de um modelo para classificação de fraudes, por exemplo, haverá uma marcação em cada registro apontando se trata-se ou não de uma fraude. Essa marcação pode ser chamada de variável resposta, *target* ou variável dependente.

No exemplo citado anteriormente, a variável resposta possui duas classes, fraude ou não fraude. Nesse caso, tem-se um problema de classificação cujo objetivo é acertar a quais classes pertencem os elementos da base. Caso a variável dependente fosse contínua, o problema seria de regressão e, por conseguinte, o objetivo seria o de acertar o valor que a variável resposta deveria apresentar.

O aprendizado não supervisionado, por sua vez, não apresenta uma variável dependente. Nesse caso, os algoritmos têm como objetivo comparar os elementos da base entre eles. A maioria dos algoritmos não supervisionados volta-se para agrupamento, ou seja, baseando-se em alguma medida de similaridade, o algoritmo marca os elementos que são mais parecidos entre si, formando grupos dentro da base.

Tanto os algoritmos supervisionados quanto os não supervisionados podem consumir muitos recursos computacionais, seja pela complexidade algorítmica envolvida seja pelo número de repetições necessárias para o aprendizado. Nesses casos, faz-se necessário o uso de ferramentas que ampliem o poder computacional, e uma forma de obter isso é usando computação distribuída ou paralela. Na computação distribuída, a execução de um mesmo algoritmo, ou diferentes partes de algoritmos, pode ser dividida em vários computadores diferentes. Na computação paralela, essa divisão acontece dentro dos processadores de um mesmo computador (McClelland & Rumelhart, 1987).

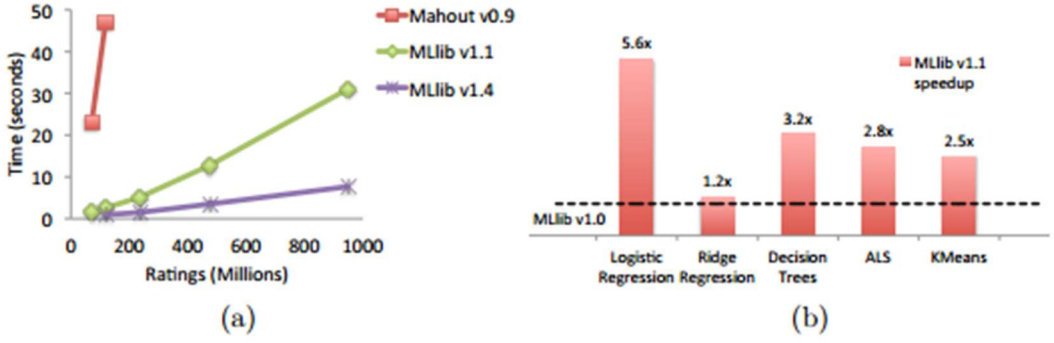
Em 2004, Dean e Ghemawat, funcionários do Google, publicaram um trabalho no qual descreviam uma forma de simplificar o uso de computação distribuída. Com a metodologia apresentada no trabalho, é possível utilizar computação distribuída sem a preocupação de implementar todos os controles operacionais que são esperados em arquiteturas paralelas. Esse artigo possibilitou o surgimento, em 2007, do *software* Hadoop, um arcabouço para processamento distribuído.

O Hadoop consegue dividir em partes grandes bases de dados e distribuí-las entre os diversos computadores de um *cluster*. Dessa forma, um algoritmo que pode ser processado de forma paralela, por exemplo, um contador de palavras, pode ser executado de forma simultânea para cada uma das partes.

Em 2010, Zaharia, Chowdhury, Franklin, Shenker e Stoica propuseram uma nova ferramenta baseada em processamento distribuído em memória, esta tipologia faz com que o processamento fique até cem vezes mais rápido se comparado com o mesmo processamento no Hadoop. Este ganho de desempenho está relacionado com a diminuição de utilização do disco rígido dos computadores do *cluster*. Etapas do processamento que eram consistidas em disco são realizadas na memória principal dos computadores, tornando todo o processo mais rápido.

Para fazer uso desse poder computacional para o aprendizado de máquina, surgiram bibliotecas para o desenvolvimento de aplicações dessa área. Uma delas é o Mahout (Owen, Anil, Dunning, & Friedman, 2011). Trata-se de uma biblioteca que reúne uma série de algoritmos de classificação, agrupamento e recomendação que são executados utilizando o mecanismo de processamento do Hadoop. Na mesma linha, existe a biblioteca MLlib (Meng et al., 2016), que também reúne diversos algoritmos para aprendizado de máquina além de trazer algumas funcionalidades que ajudam preparando os dados e automatizando processos. Outra vantagem da MLlib é o desempenho superior. Na Figura 2(a), há uma comparação entre as versões da MLlib e o Mahout durante a execução do algoritmo de recomendação ALS. A vantagem fica evidente ao observar que a MLlib consegue processar cinco vezes mais dados em um tempo menor que o Mahout. A Figura 2(b) mostra que a MLlib ainda está em evolução, apresentando desempenho diversas vezes superior com relação à versão anterior.

Figura 2 - (a) Resultados do benchmark para o algoritmo ALS. (b) Evolução na velocidade de processamento entre a versão 1.0 e a 1.1 da MLlib.



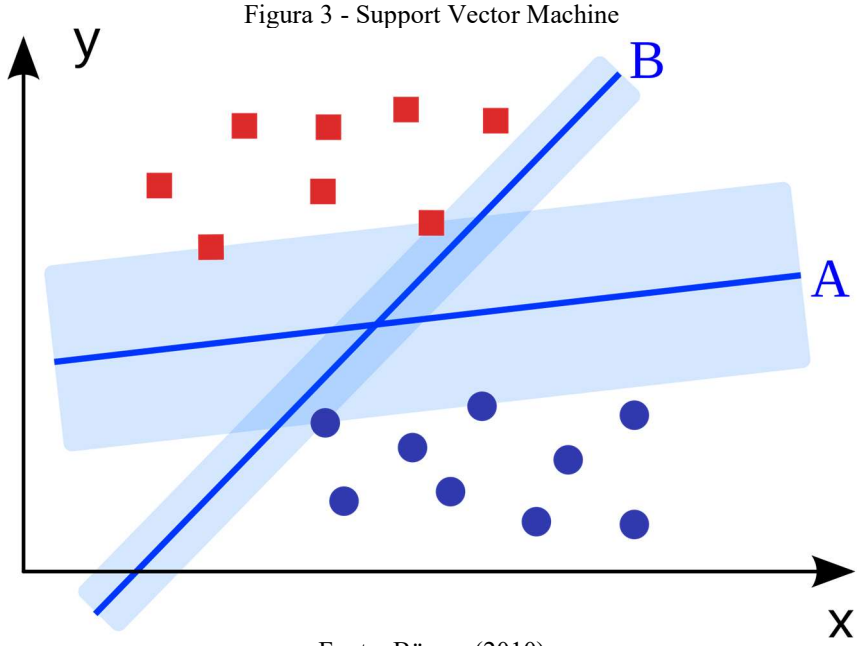
Fonte: Meng et al. (2016)

2.7 EVENTOS RAROS SENDO DETECTADOS COM ALGORITMOS DE APRENDIZADO DE MÁQUINA

Van der Paal (2014) testou a eficiência dos algoritmos de aprendizado de máquina (Support Vector Machine - SVM - e Random Forest) contra o desempenho de diversos algoritmos de regressão logística.

Como resultado da análise de Van der Paal (2014), infere-se que os algoritmos Random Forest e SVM têm um desempenho melhor do que os algoritmos de regressão logística estudados, sendo que o SVM apresentou resultados muito melhores do que Random Forest. Esse resultado, em parte, dá-se pela eficiência da técnica SVM em classificar dados não lineares. Quanto ao Random Forest, apresentou resultados equivalentes à melhor regressão logística em cada caso.

Figura 3 - Support Vector Machine



Fonte: Bürger (2010)

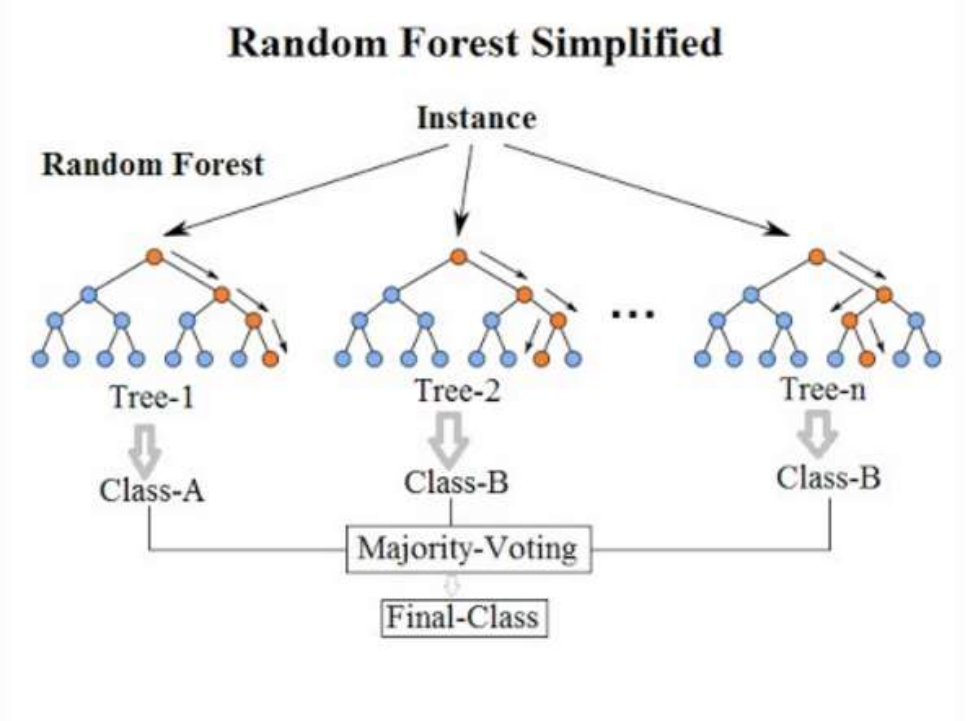
O Support Vector Machine (Cortes e Vapnik, 1995) é um algoritmo de classificação. Para entender seu funcionamento, imagine um problema com duas classes que devem ser separadas. Primeiro é aplicada uma transformação aos dados chamada de "*kernel trick*". Essa transformação faz uma seleção das variáveis relevantes para o problema e, então, realiza uma transformação geométrica em que as novas dimensões permitem a divisão das classes diferentes de maneira mais simples (Baudat & Anouar, 2001). A ideia é permitir que dados que não são naturalmente separáveis possam ser separados com a inclusão de uma nova dimensão. A partir

daí o algoritmo SVM traça um hiperplano entre as classes estudadas de tal forma que a margem de separação entre os indivíduos de classes diferentes seja maximizada (Simon, 2001). Na Figura 3, temos dois hiperplanos, A e B, que conseguem separar as classes apresentadas, porém percebe-se que a margem de separação é muito maior em A que em B. Na figura, a margem está representada por uma faixa azul em torno da linha que representa os hiperplanos.

O método Random Forest utiliza conjuntos de dados aleatórios extraídos da base de treinamento para criar N árvores de decisão. Cada árvore de decisão irá gerar um conjunto de variáveis que melhor generalize o problema. Dessa forma, há uma última etapa em que para cada observação imputada, cada árvore apresenta seus resultados, e a classe que com maior frequência for escolhida pelas árvores será eleita como a predição para determinada observação (Tan, 2018).

Na Figura 4, tem-se a representação de uma *random forest*. Em cada árvore, os nós que determinada observação percorreu foram pintados de laranja. Ao final, cada árvore realizou uma classificação independente, atribuindo uma determinada classe ao registro em questão. No caso representado, duas árvores escolheram a classe B e uma a classe A. A classe escolhida pelo modelo final é a que a maioria das árvores classificou, ou seja, a classe B.

Figura 4 - Diagrama de uma Random Forest.



Fonte: Jagannath (2017)

A vantagem de técnicas mais simples, como a regressão logística, está na facilidade de explicar o modelo. Sendo que cada variável receberá um coeficiente que tem explicação direta com a probabilidade atribuída à classe. Quanto aos outros algoritmos, não há uma forma simples de se explicar a contribuição de cada variável para o modelo.

Outra técnica testada neste trabalho será o Gradient Boosting Trees (Friedman, 1999). Nesta técnica uma árvore de classificação é treinada com a variável resposta do problema. Com base no erro de classificação apresentado por esta árvore, outras árvores são treinadas sequencialmente desta vez com o objetivo de reduzir o erro apresentado pela primeira estimativa. O número de árvores treinadas para reduzir o erro é um dos parâmetros do algoritmo. Esta técnica pode se sair bem com bases desbalanceadas por ter como comportamento tentar reduzir as imprecisões do algoritmo anterior.

A regressão logística é uma das técnicas estatísticas mais utilizadas para a classificação binária, ou seja, quando a variável resposta apresenta duas classes. A maior vantagem desse método é a facilidade de explicação do modelo obtido, que facilita a interpretação da importância e da relação que as variáveis independentes estabelecem com a variável resposta. (Hilbe, 2009).

Uma das maneiras que também foi utilizada neste trabalho foi a chamada *elasticnet* (Zou e Hastie, 2003). Neste caso a regressão logística é penalizada linearmente com os métodos Lasso (Tibshirani, 1996) e Ridge Regression (Hoerl e Kennard, 1988). Uma das vantagens aproveitadas do método Lasso é a seleção de variáveis. As duas formas de penalização servem para reduzir a importância das variáveis com o objetivo de evitar o *overfitting.*

2.8 VARIÁVEIS UTILIZADAS PARA DETECÇÃO DE FRAUDES PÚBLICAS

Durante a revisão bibliográfica inicial não foram encontrados trabalhos citando variáveis utilizadas na detecção de fraudes públicas. Por isso, optou-se por estudar os tipos de fraudes que são cometidos contra a Administração Pública para que fosse executado um levantamento de variáveis mais direcionadas. Estudos que relacionam o índice de percepção de corrupção utilizam uma série de variáveis que podem ser incorporadas neste trabalho. Elas são listadas a seguir.

Dentre as variáveis que podem ser observadas, destacam-se: a origem do sistema legal, o grau de fracionalização etnolinguístico, o nível de salários no setor público e o tipo de estrutura administrativa (Orth, 2012).

Além dessas variáveis, outras estão relacionadas com o grau de corrupção observado, como a produtividade do investimento público, o nível de investimentos estrangeiros diretos, o nível de gastos governamentais e o grau de informalidade da economia.

Outras variáveis que podem ser utilizadas, segundo a tese de doutorado de Lopes (2011): o grau de desigualdade econômico-social, o grau de predomínio religioso, o grau de descentralização, a persistência democrática, o PIB *per capita*, a taxa de inflação, o grau de intervenção estatal na economia, o tamanho do governo, a qualidade da burocracia, o grau de competição política, o grau de liberdade civil, o grau de liberdade de imprensa, o grau de mobilização da sociedade civil, a estrutura de mercado para a corrupção, além do grau de abertura econômica do país.

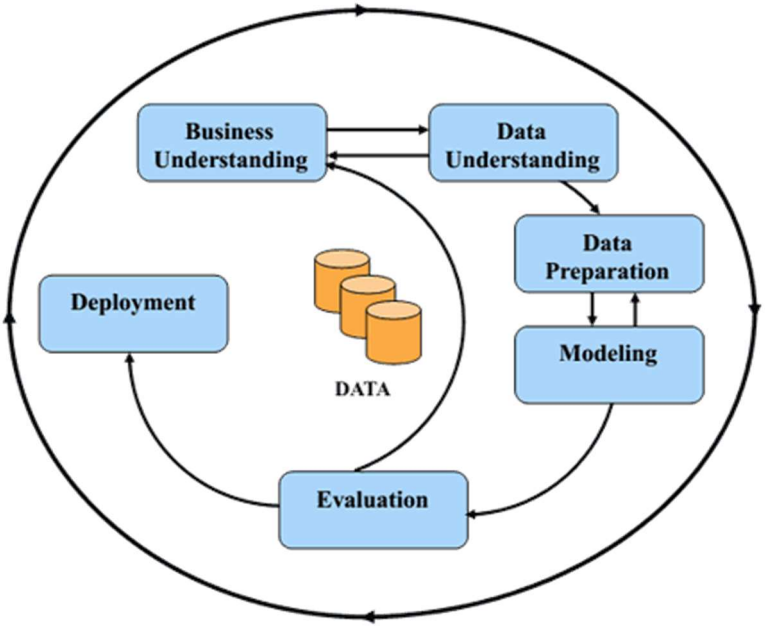
No caso deste estudo, muitas dessas variáveis não se aplicavam porque o objetivo era o de analisar os contratos estabelecidos pela União mediante licitações. Entretanto pôde-se utilizar os conceitos para a criação de novas variáveis. Por exemplo, a variável PIB *per capita* pode ser vista na visão estadual e, assim, foi possível criar variáveis sobre o PIB do Estado que solicitou a licitação, bem como o do Estado de origem da empresa que foi licitada.

2.9 METODOLOGIA PARA PROJETOS DE *MACHINE LEARNING*

A metodologia CRoss-Industry Standard Process for Data Mining (CRISP-DM) é um método popular que orienta projetos de análise de dados aumentando suas chances de sucesso (Chapman et al., 2000). Essa foi a metodologia escolhida porque há relatos na literatura sobre sua utilização com projetos que utilizam grandes bases de dados, como mencionado por Azevedo e Santos (2008).

Nessa perspectiva, são mencionadas seis etapas como fases necessárias para o sucesso de um projeto. Essas etapas permitem o início e a implantação de um modelo de aprendizado de máquina em um ambiente produtivo, ou seja, com dados reais. O objetivo final é que esses dados auxiliem na tomada de decisão gerando valor para o negócio.

Figura 5 - Visão das etapas da metodologia CRISP-DM



Fonte: Chapman et al. (2000).

Na Figura 5 são indicadas as etapas da metodologia, sendo possível perceber o caráter cíclico de um projeto de análise de dados. Essa característica está relacionada com as alterações que acontecem nos dados e nos negócios. Outro motivo é que, ao percorrer as etapas do processo, pode surgir um entendimento maior dos dados e possivelmente alguma etapa precise ser revisitada. Cada uma dessas etapas é descrita, a seguir, no contexto do presente.

2.9.1 Entendimento do Negócio

Nesta fase são entendidas as necessidades do negócio. Isso significa analisar o problema existente na administração do negócio e como os dados podem ajudar a resolvê-lo; são, então, decididos os objetivos do projeto. No caso deste trabalho, foram estudados os contratos firmados entre o governo e a iniciativa privada através de licitações afim de identificar aqueles que pertencem a empresas que foram identificadas como inidôneas.

2.9.2 Entendimento dos dados

Na fase de entendimento dos dados são feitas análises para encontrar problemas de qualidade nos dados, descobrir padrões através da visualização dos dados e construir novas variáveis ou subconjunto dos dados.

Nesta etapa é essencial garantir a qualidade dos dados que serão trabalhados para assegurar que eles representem de forma real o problema estudado. A falta de qualidade dos dados faz com que organizações precisem elevar seus custos com organização dos dados, percam oportunidades e tomem decisões erradas.

2.9.3 Preparação dos dados

Durante a preparação dos dados, o objetivo é criar uma base de dados que esteja pronta para ser processada pelos algoritmos de aprendizado de máquina. Esta etapa pode ser revisitada dependendo do tipo de algoritmo que será aplicado aos dados. As preparações podem conter seleção de atributos e observações, limpeza e transformações dos dados.

2.9.4 Modelagem

É na etapa de modelagem que são criados os modelos de aprendizado de máquina que irão atender às necessidades de negócio. Chapman (2000) cita 4 fases na etapa de modelagem: Seleção de técnicas de modelagem, geração dos testes, construção do modelo e avaliação dos modelos.

Neste trabalho a seleção de técnicas de modelagem foi amparada pela revisão bibliográfica. Para atender as necessidades de negócio, ou seja, classificar um contrato como fraude ou não, foram usadas técnicas de classificação. A geração dos testes, construção do modelo e a sua avaliação são apresentadas no próximo capítulo.

2.9.5 Avaliação

O processo de avaliação dos modelos utilizou técnicas de medida de desempenho clássicas como a área sob a curva ROC e medidas baseadas na matriz de confusão. Essas

métricas foram escolhidas por se comportarem melhor em cenários nos quais os dados são desbalanceados, como inferem Burez e Van Den Poel (2009).

É importante notar que as fases de avaliação e de modelagem são executadas concomitantemente para que os parâmetros das técnicas de modelagem possam ser otimizados para maximizar as métricas utilizadas para a avaliação.

2.9.6 Implantação

O processo de implantação do modelo em uma instituição governamental pode ser difícil de ser reproduzido por exigir o levantamento de quais são os ambientes informacionais disponíveis nos órgãos auditores dos contratos públicos.

Uma das formas que o modelo poderia ser utilizado seria executar a classificação dos contratos que foram firmados no último mês, por exemplo. Assim, as áreas responsáveis pela auditoria dos contratos poderiam priorizar os contratos com maior chance de pertencerem a empresas com um perfil semelhante ao de empresas que foram consideradas inidôneas. Para que isso funcione, será preciso que exista um conjunto de programas que atuem em conjunto para entregarem as etapas da metodologia explicada na seção 3. Isto é, um programa deverá fazer a captura dos dados, outro a sua preparação e um outro a classificação utilizando o modelo proposto.

Para simular esse desafio, o presente trabalho usou dados de meses que não estiveram presentes no processo de modelagem (*backtest*) para representar a qualidade preditiva do modelo esperada em uma situação real. Outra forma que pode ser aplicada é a criação de dados simulados baseados nas bases levantadas para a análise do problema.

3 METODOLOGIA

Este estudo tratou-se de uma pesquisa quantitativa baseada em dados secundários obtidos por meio de portais públicos que disponibilizam dados da Administração Pública no Brasil. Dos dados disponibilizados por esses portais foram extraídas as variáveis estudadas neste trabalho. Outras variáveis foram geradas com base na revisão bibliográfica. Além disso, este trabalho gerou novas variáveis a partir de técnicas de derivação estatística.

Após o levantamento dos dados fez-se o tratamento deles para que fossem atendidos alguns requisitos de algoritmos de *machine learning*, visto que alguns tratamentos podem melhorar o desempenho dos modelos. Assim, pode-se criar variáveis *dummies* para dados categóricos, ou seja, para cada categoria apresentada na variável foi criada outra variável indicando o valor 1 caso o registro pertencesse a esta categoria, ou 0 no caso contrário, realizar a imputação de valores faltantes ou remoção de valores *outliers*, isto é, valores maiores que uma vez e meia o intervalo interquartil acima do terceiro quartil e valores menores que uma vez e meia o intervalo interquartil abaixo do primeiro quartil. Com os dados preparados, foi realizada a fase de modelagem, na qual foram executadas diferentes técnicas com o objetivo de inferir o comportamento da variável dependente, no caso, indicando-se determinado contrato pertence ou não a uma empresa fraudulenta.

3.1 LEVANTAMENTO DOS DADOS

Os dados foram obtidos a partir do portal dados.gov.br utilizando duas fontes, o Ministério do Planejamento (MP) e a Controladoria Geral da União (CGU). Os dados do MP são referentes aos contratos públicos que foram originados a partir de licitações e são disponibilizados ao público através de uma *"Application Programming Interface"* (API). A API é uma interface com um sistema computacional onde é possível realizar consultas ao banco de dados do Ministério Público. Essa API só permite o retorno de 500 registros por consulta. Dessa forma, foi necessário o desenvolvimento de um programa para obter a base completa com os contratos licitados e salvá-los localmente no formato Comma Separated Values (CSV), resultando em um arquivo com em torno de 595 mil linhas, sendo que cada linha representa um contrato diferente. A execução do programa para a extração dos dados ocorreu no dia 20/3/2018, e esse programa é apresentado no Anexo A.

Os dados da CGU referem-se a empresas e a pessoas físicas que sofreram sanções tendo como efeito a restrição ao direito de participar de licitações ou de celebrar contratos com a Administração Pública. Essa base de dados é chamada de Cadastro Nacional de Empresas Inidôneas e Suspensas (CEIS) e está disponível para *download* já no formato CSV no portal dados.gov.br. O *download* foi feito no dia 8/9/2017 através do endereço <http://dados.gov.br/dataset/ceis>.

O quadro 3 apresenta as bases que foram utilizadas, quais são as variáveis que foram obtidas dessas bases e qual a origem desta base, Controladoria Geral da União ou Ministério do Planejamento.

Quadro 3 – Variáveis obtidas

Base	Variável	Fonte
Empresas	Tipo de Pessoa	Controladoria Geral da União
	CPF ou CNPJ do Sancionado	
	Nome Informado pelo Órgão Sancionador	
	Razão Social - Cadastro Receita	
	Nome Fantasia - Cadastro Receita	
	Número do processo	
	Tipo Sanção	
	Data Início Sanção	
	Data Final Sanção	
	Órgão Sancionador	
	UF Órgão Sancionador	
	Origem Informações	
	Data Origem Informações	
	Data Publicação	
	Publicação	
	Detalhamento	
	Abrangência definida em decisão judicial	
	Fundamentação Legal	
	Descrição da Fundamentação Legal	
	Data do Trânsito em Julgado	
	Complemento do Órgão	
	Contato da Origem da Informação	
Contratos	Identificador do Contrato	Ministério do Planejamento
	UASG	
	Modalidade da Licitação	
	Número do Aviso da Licitação	
	Código do Contrato	
	Licitação Associada	
	Origem da Licitação	
	Número	
	Objeto	
	Número de Aditivos	
	Número do Processo	
	CPF Contratada	
	CNPJ Contratada	
	Data de Assinatura	
	Fundamento Legal	
	Data de Início da Vigência	
	Data de Termino da Vigência	
	Valor inicial	
	Aditivos do Contrato	
	Apostilamentos do Contrato	
	Eventos do Contrato	
Fornecedores	Identificador do Fornecedor	Mnistério do Planejamento
	CNPJ	
	Razão Social	
	Nome Fantasia	
	Unidade Cadastradora	
	Natureza Jurídica	
	Ramo de Negócio	
	Porte da Empresa	
	CNAE	
	CNAE Secundário	
	Logradouro	
	Número do Logradouro	
	Complemento do Logradouro	
	Bairro	
	Município	
	CEP	
	Caixa Postal	
	Ativo	
	Recadastrado e Habilitado a Licitar	

Fonte: Elaboração própria

3.2 TRATAMENTO DOS DADOS

A base de contratos originou as variáveis explicativas utilizadas no modelo e a base de empresas inidôneas gerou a variável resposta. A forma como as variáveis foram criadas é descrita a seguir, assim como o processo de preparação de dados para a fase de modelagem. Nesta etapa, os arquivos CSV foram interpretados utilizando a linguagem Python e o pacote Pandas (Python Data Analysis Library), que permite a manipulação de bases de dados de maneira facilitada.

A variável resposta foi criada a partir da junção das duas bases, mantendo todos os dados da base de contratos e adicionando uma marcação para as linhas onde a empresa fornecedora do contrato estava presente na base de empresas inidôneas. Os contratos foram marcados apenas no período em que a empresa foi sancionada. Contratos assinados após o término das sanções impostas à empresa foram marcados com 0, ou seja, a empresa fornecedora foi considerada idônea por já ter cumprido sua punição, e isso aconteceu em apenas seis casos. Após a criação da variável resposta, a proporção de eventos observados na base foi de 7,5%, mostrando um desbalanceamento da classe “fraude” com a classe dos contratos bons.

Para obter as variáveis explicativas, a partir da base de contratos, foram excluídas as variáveis identificadoras, como nomes e chaves de identificação. Após essas exclusões, restaram 11 variáveis: Órgão Governamental, Natureza Jurídica, Ramo do Negócio, Porte da Empresa, CNAE, Município, Número de aditivos, Valor inicial do contrato, Extensão do contrato (em dias), Modalidade da Licitação e Origem da Licitação. As variáveis contínuas são as variáveis Número de aditivos, Valor inicial e Extensão do contrato. Todas as outras são categóricas.

Nas variáveis categóricas que apresentaram valores faltantes, a abordagem adotada para corrigir esses casos foi inserir um valor fora do domínio das variáveis para representar o valor ausente. No caso, foi inserido o valor -9999 onde antes não existia um valor. Isso faz com que os algoritmos interpretem de forma diferente esses casos e será possível distinguir caso a ausência de valor ajude a explicar o evento. A variável número de aditivos apresentou valores faltantes para os contratos que não tinham aditivos, neste caso foi inserido o valor zero. As outras variáveis contínuas não apresentaram valores faltantes.

Outras formas de inserir valores faltantes podem ser a inserção do valor médio, da mediana (Han, Pei, & Kamber, 2011), da moda ou de um valor determinado por um especialista. Também, pode-se fazer a imputação dos valores utilizando técnicas de *machine learning*

usando tanto algoritmos não supervisionados como supervisionados, como estudado por Jerez et al. (2010). Hunt e Jorgensen (2003) desenvolveram uma extensa revisão da literatura sobre o assunto em trabalho a respeito do preenchimento de valores faltantes.

Como parte do processamento das variáveis categóricas foram criadas variáveis *dummies*. Após esse tratamento, a variável Modalidade da Licitação originou 13 novas variáveis, Origem da Licitação originou outras 2, Natureza Jurídica ficou com 31, Ramo do Negócio com 86, Porte da Empresa com 4, CNAE apresentou 763 novas variáveis e Município 1.177 novas variáveis.

Esta grande quantidade de variáveis torna o processo de modelagem muito demorado, exigindo que seja feita uma seleção das variáveis que mais discriminam a variável resposta para que o processo de escolha do melhor modelo seja mais simples. A seleção de variáveis foi feita utilizando o pacote Boruta, disponível para R e Python, que foi desenvolvido seguindo o trabalho de Kursa e Rudnick (2010).

Este pacote usa como estratégia colocar variáveis aleatórias como sendo variáveis explicativas do modelo. A forma como o pacote cria as variáveis aleatórias é utilizar as variáveis explicativas, porém com os valores embaralhados entre os registros da base de modelagem, isso garante que as variáveis aleatórias tenham a mesma distribuição do que a variável real. O esperado é que toda explicação vinda destas variáveis aleatórias seja espúria, portanto variáveis com um desempenho pior do que as aleatórias poderiam ser descartadas por não melhorarem o modelo.

Após a seleção das variáveis restaram 15 variáveis: número de aditivos, valor inicial, tempo, 6 categorias de UASG, 5 categorias de modalidade da licitação e uma de ramo do negócio.

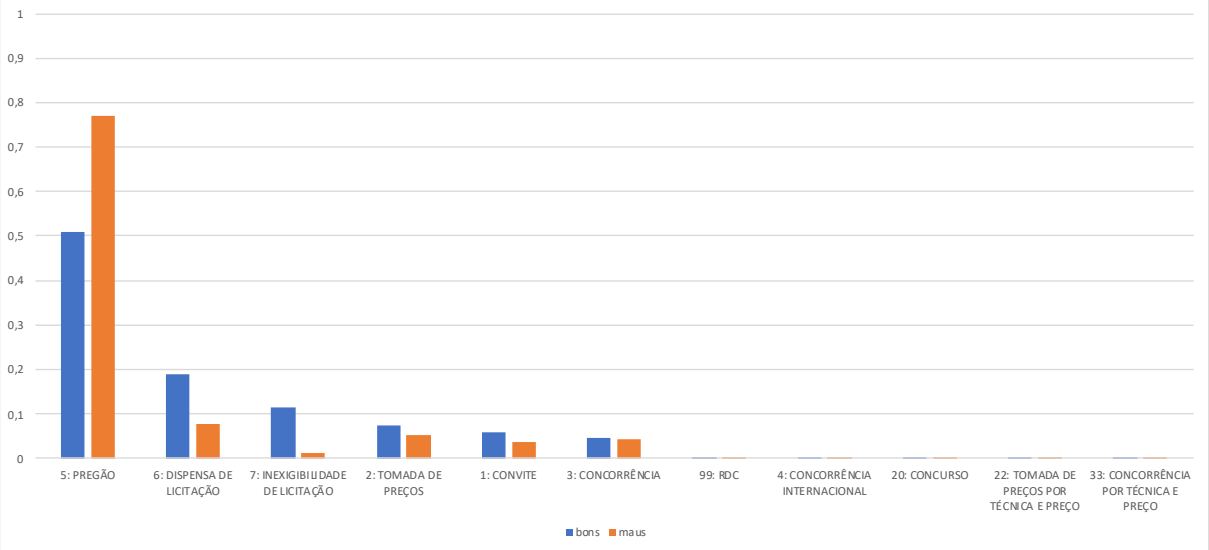
3.3 ANÁLISE DESCRITIVA

Ao usar a metodologia CRISP-DM, a análise descritiva é importante para o entendimento dos dados pertinentes ao negócio. Foram realizadas análises gráficas diferentes para variáveis categóricas e variáveis contínuas. Para todas as variáveis foram desenvolvidos gráficos para a população inteira e para apenas os contratos que pertencem a empresas inidôneas. O objetivo era identificar diferenças nas distribuições das variáveis para as diferentes classes.

Para as categóricas foram feitos gráficos de barras mostrando a proporção das categorias que aparecem em mais do que 0,8% da base. Além disso, foram apresentadas apenas as 15 categorias com maior número de contratos de cada variável. Essa análise permite conhecer as categorias mais representativas em uma variável, identificando categorias que concentram grande quantidade dos registros, por exemplo. Para as variáveis contínuas, os gráficos foram do tipo histograma e boxplot. Ambos mostram a distribuição dos dados, porém o boxplot dá destaque aos *outliers* existentes na variável.

Os gráficos de barra mostram a distribuição dos bons e dos maus nas categorias de cada variável. Assim, todas as barras de uma mesma cor somam 100% dos contratos da base estudada.

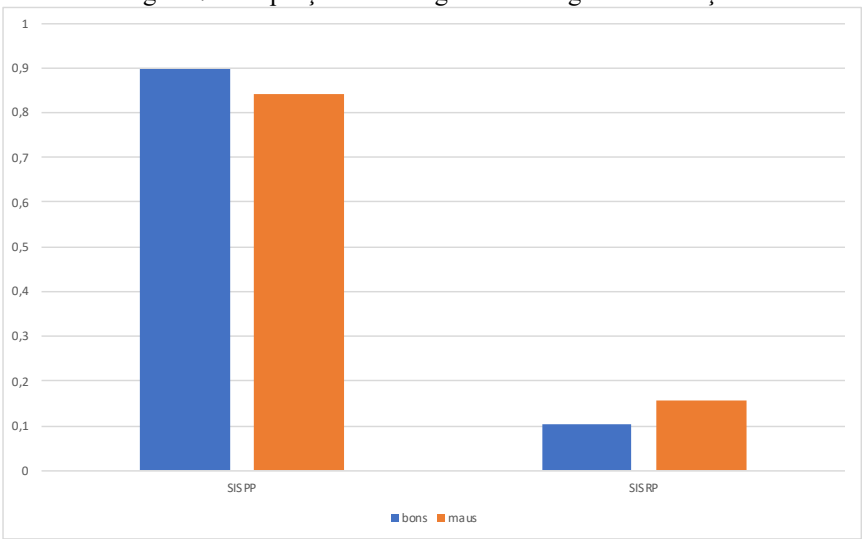
Figura 6 - Proporção das categorias de modalidade da licitação



Fonte: Elaboração própria

Para a variável modalidade da licitação, nota-se pela figura 6 que a modalidade pregão representa mais da metade das observações e mais de 70% dos contratos de empresas inidôneas. Comparando os dois gráficos, detecta-se uma alteração na ordem, sendo que a modalidade Inexigibilidade de Licitação apresenta a terceira maior frequência na população, sendo apenas a sexta maior quando se limita o público à classe estudada (maus).

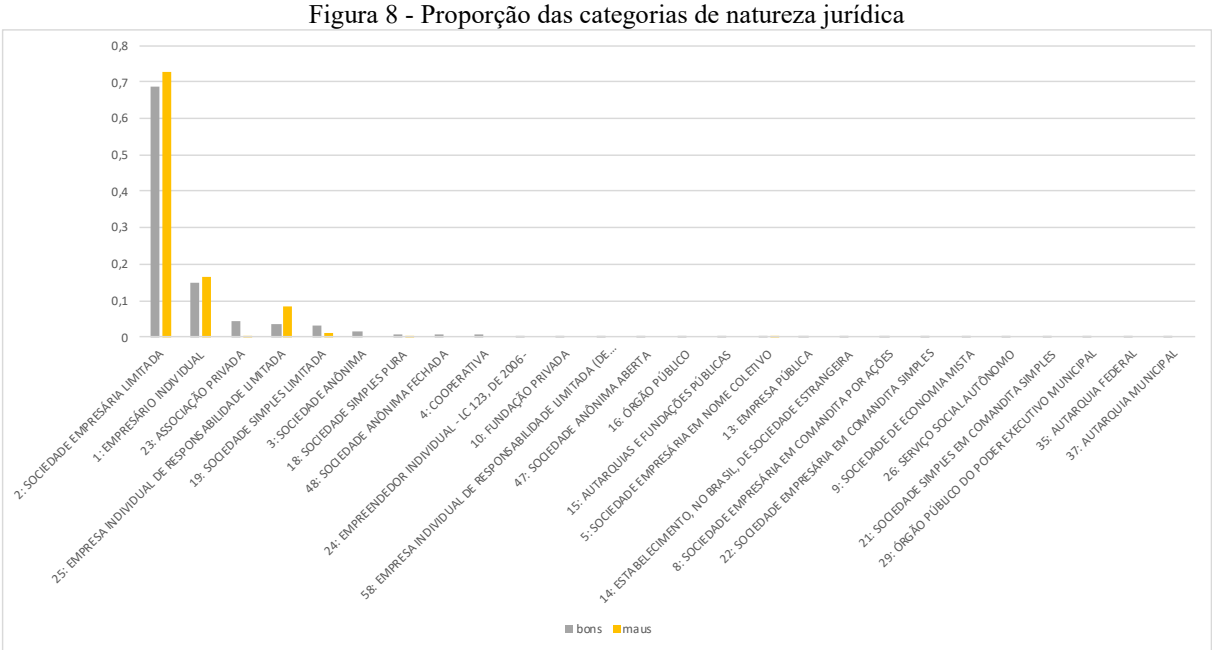
Figura 7 - Proporção das categorias de origem da licitação



Fonte: Elaboração própria

Quanto ao sistema que originou a licitação, figura 7, não existem diferenças significativas entre a população e a classe estudada. Há um leve aumento da proporção da categoria SSRP para os contratos de empresas impedidas.

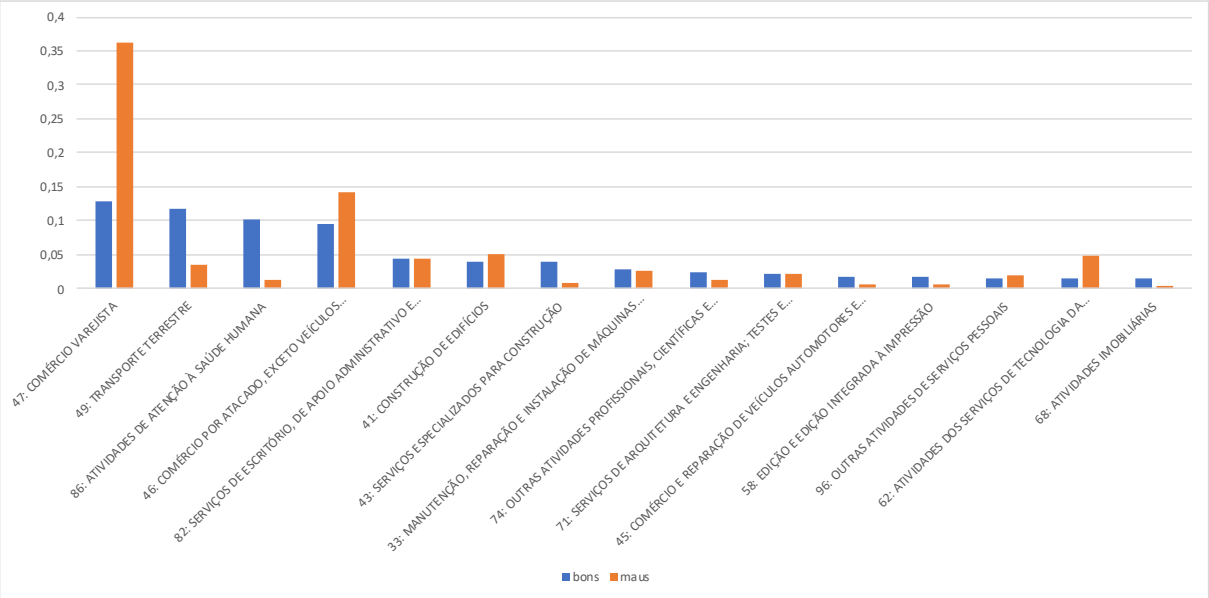
Figura 8 - Proporção das categorias de natureza jurídica



Fonte: Elaboração própria

Ao analisar a variável que apresenta a natureza jurídica, figura 8, é possível notar uma predominância da categoria Sociedade Empresária Limitada. Já quanto à diferença entre os dois gráficos, a categoria Associação Privada, em terceiro na população, desaparece nos contratos com eventos, pois apresenta um valor menor que 0,8%.

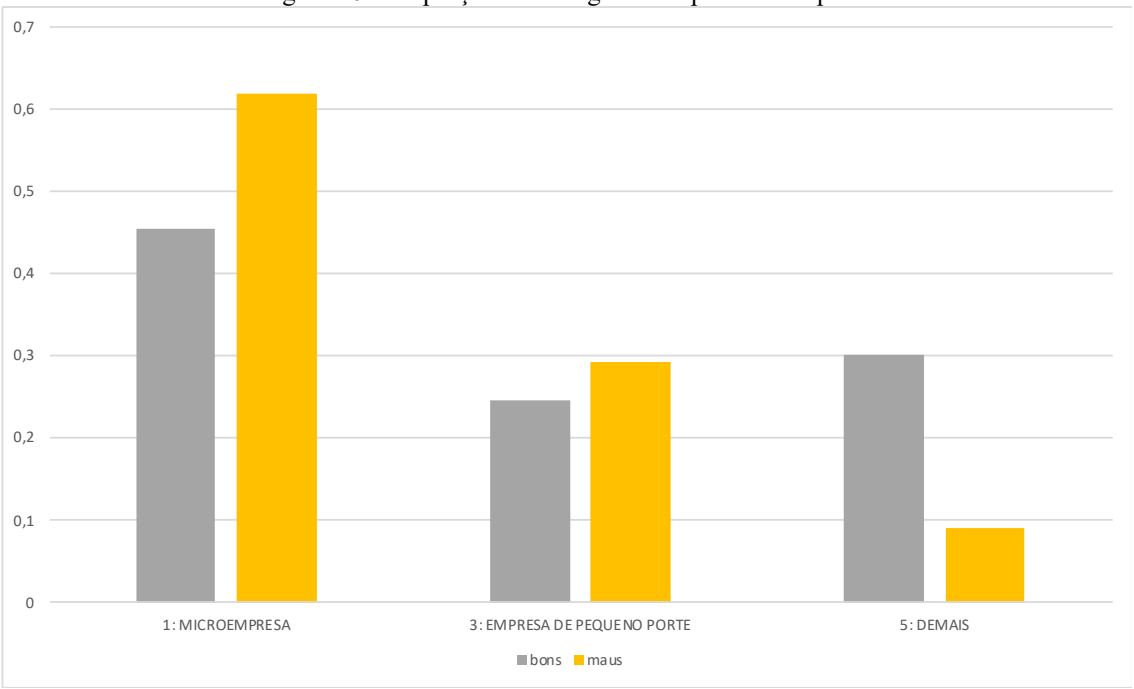
Figura 9 - Proporção das categorias de ramo de negócio



Fonte: Elaboração própria

A variável Ramo do Negócio, figura 9, mostra a categoria comércio varejista como a de maior frequência. Um ponto de observação interessante é que a ocorrência da categoria passa de 13% para 35% quando analisada a classe dos eventos. Há algumas alterações nas ordens entre os gráficos, porém sem um aumento de proporção significativo como no caso apresentado.

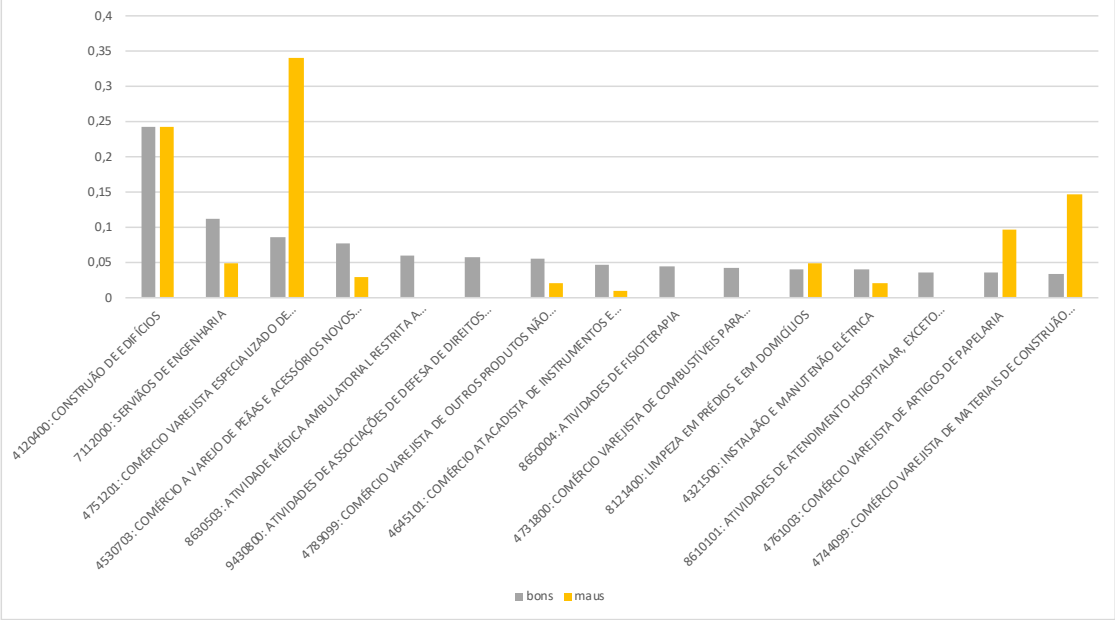
Figura 10 - Proporção das categorias de porte da empresa



Fonte: Elaboração própria

Ao analisar o porte da empresa fornecedora, figura 10, vê-se um aumento da proporção de microempresas e empresas de pequeno porte na classe estudada. A ocorrência de outros portes é reduzida para um terço da sua proporção na população.

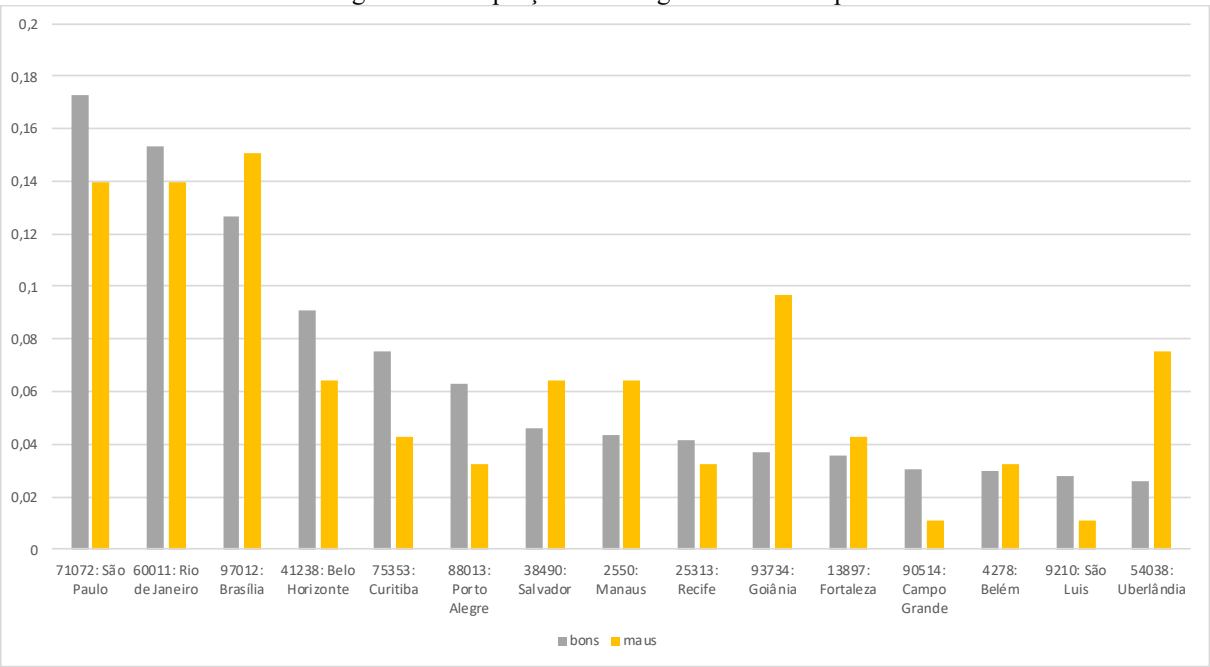
Figura 11 - Proporção das categorias de CNAE



Fonte: Elaboração própria

Ao observar o CNAE, figura 11, percebe-se que há uma distribuição dos contratos de empresas inidôneas em mais categorias.

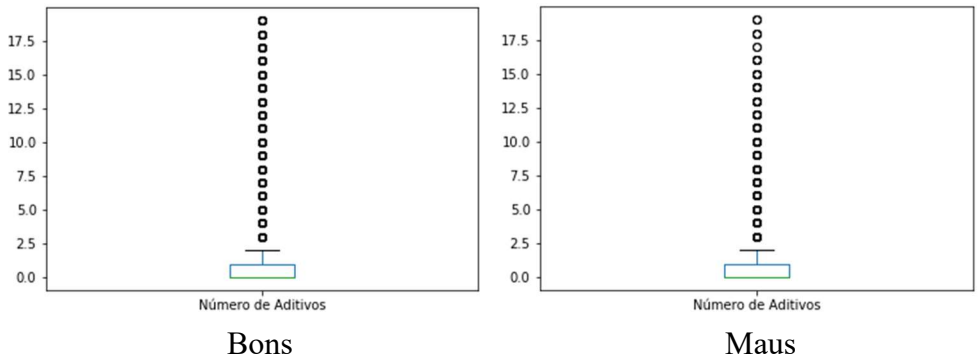
Figura 12 - Proporção das categorias de municípios



Fonte: Elaboração própria

Ao identificar a cidade de origem dos fornecedores, figura 12, individua-se uma ocorrência maior de eventos na cidade de Brasília. Também nota-se uma grande diferença na proporção dos contratos fraudulentos nas cidades de Goiânia e Uberlândia.

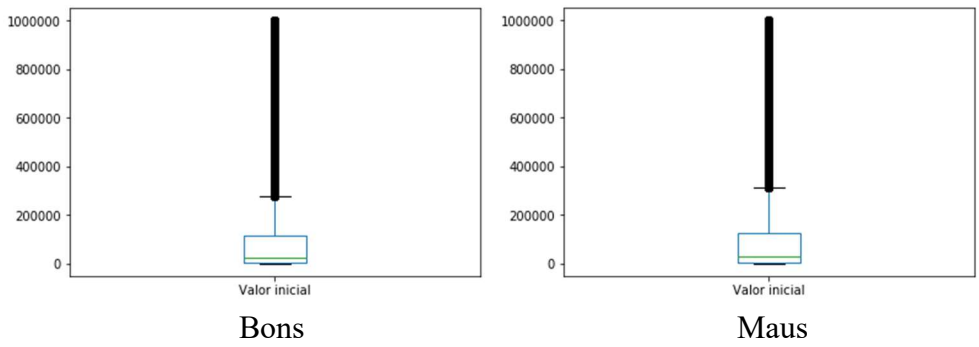
Figura 13 - Boxplot de número de aditivos



Fonte: Elaboração própria

Ao examinar a variável número de aditivos, figura 13, pode-se interpretar que mais de 50% dos contratos não possuem aditivos. Aproximadamente 75% dos contratos possuem até um aditivo. Alguns contratos indicam um número de aditivos muito superior, sendo que outros contratos apontam mais de 60 aditivos. A apresentação no boxplot foi limitada a 20 aditivos para facilitar a leitura do gráfico, isso significa um corte superior ao percentil 99 e não impacta a interpretação.

Figura 14 - Boxplot de valor inicial

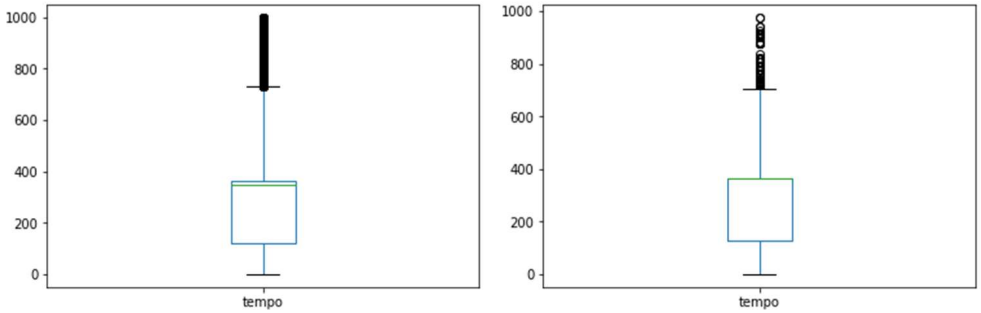


Fonte: Elaboração própria

A variável valor inicial indica um pequeno deslocamento da mediana e do terceiro quartil na subpopulação dos contratos pertencentes à classe estudada. Sendo que a mediana passa de R$34.652,00 para R$41.750,00. O mesmo acontece com o terceiro quartil, que vai de

R$172.800,00 para R$203.000,00. O valor máximo de um contrato é de 10 bilhões de reais e o de um contrato de empresa fraudadora é de 1.5 bilhão de reais. A apresentação no boxplot foi limitada a 1 milhão, o que assinala um corte no percentil 92.

Figura 15 - Boxplot de tempo



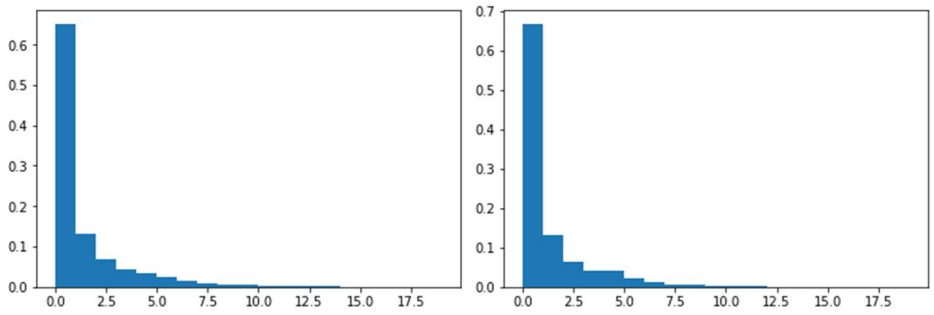
Bons Maus

Fonte: Elaboração própria

Na variável tempo, 75% dos contratos têm até 365 dias, o que mostra uma prevalência de contratos de até um ano. Observados os valores máximos para as duas populações, para todos os contratos existe um contrato vigente há 163 anos, que diz respeito a um contrato de fornecimento de esgoto tratado por uma companhia de saneamento básico. Na subpopulação de empresas inidôneas, um contrato de 85 anos trata da compra de um *scanner* por meio de uma empresa brasiliense de serviços de informática. Esses *outliers* são contratos com a data de término de sua vigência no futuro, sendo que o de 163 anos irá expirar em 2180 e o de 85 anos expirará em 2099. Como é uma parcela muito pequena dos contratos, esses foram mantidos nas bases.

No que tange aos boxplots das variáveis número de aditivos, valor inicial e tempo, não é possível identificar alterações significativas entre as distribuições para a população e para a classe minoritária. Apenas o boxplot do tempo apresenta uma redução no número de *outliers* e a de valor inicial um deslocamento da mediana para cima.

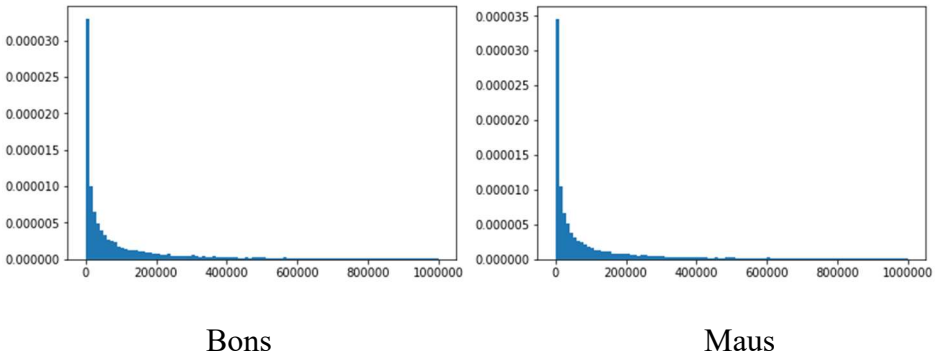
Figura 16 - Histogramas de número de aditivos



Fonte: Elaboração própria

Conforme o histograma da variável número de aditivos, nota-se que mais de 60% dos contratos não possuem aditivos. Assim, a distribuição apresenta uma cauda longa à direita e tem como moda e mediana o valor zero.

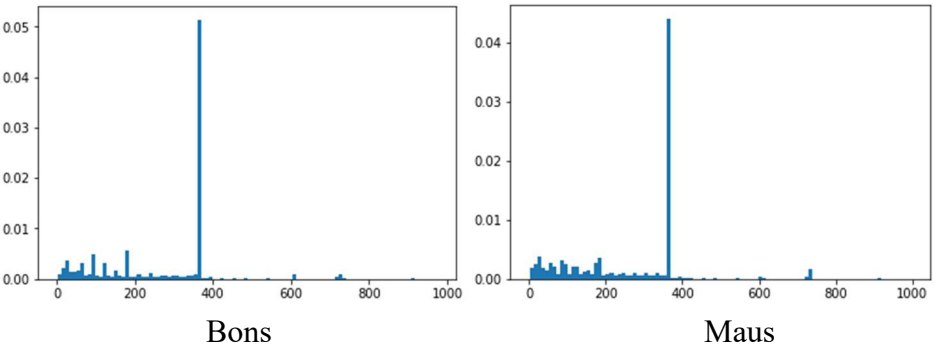
Figura 17 - Histogramas de valor inicial (R$)



Fonte: Elaboração própria

A variável que contém o valor inicial do contrato também apresenta inclinação à direita com cauda longa, 8.394 contratos assinalam seu valor inicial zerado, o que explica o pico no valor zero.

Figura 18 - Histogramas de tempo (dias)



Fonte: Elaboração própria

A variável de tempo ilustra a duração do contrato em dias. Percebe-se um pico no valor 365 dias, mostrando que vários contratos apresentam o prazo de 1 ano. Também é possível verificar que a grande maioria dos contratos não passa do prazo de 365 dias. Há outros picos, porém, bem menores, nos valores 180 dias e 720 dias, ou seja, nos prazos de 6 meses e 2 anos.

Ao comparar os histogramas da população completa com a seleção das empresas inidôneas, não se vê grandes diferenças nas distribuições. O que se pode perceber é um aumento na proporção de zeros na classe minoritária para o valor inicial e uma redução nos picos na variável tempo.

3.4 MODELAGEM

Com as variáveis preparadas, a base de dados está pronta para ser processada por algoritmos de aprendizado de máquina. Nesta etapa, separou-se um terço dos dados, de forma aleatória, para compor a base de teste, com 86.466 observações. Outras 246.313 observações foram reservadas para o treino dos modelos. Essa é uma técnica de validação cruzada que se chama holdout. O objetivo é separar uma amostra dos dados para que não seja considerada durante a etapa de treinamento do modelo. Essa amostra foi usada posteriormente ao treinamento para averiguar se o resultado obtido durante o aprendizado condizia com o que se esperava encontrar em uma situação real.

Foram utilizadas duas amostras de treinamento para a modelagem. Uma delas mantendo a proporção de eventos original, com todas as 246.313 observações, e outra balanceando a amostra pela quantidade de eventos, ou seja, foram considerados todos os eventos presentes na base de treinamento e amostrado aleatoriamente da base o mesmo número de não eventos, totalizando em 36.214 observações. Assim, foi utilizada a técnica de subamostragem para o balanceamento dos dados. Esse processo resultou em duas bases de treinamento que foram igualmente tratadas para que os resultados pudessem ser comparados.

Na utilização das amostras de treinamento foram aplicadas cinco técnicas de aprendizado de máquina: árvore de decisões, regressão logística, SVM, *random forest* e *gradient boosting*. Essas técnicas foram explicadas na seção da Revisão Bibliográfica e a maneira como foram utilizadas é descrita a seguir. Foi usada a linguagem Python e o pacote Scikit-Learn, que contém a implementação de diversas técnicas de aprendizado de máquina. O pacote Scikit-Learn foi publicado por Pedregosa et al., (2011).

Os algoritmos de *machine learning* foram aplicados utilizando Grid Search, uma técnica para buscar os melhores parâmetros de configuração dos algoritmos. É, para tanto, criada uma lista de valores para cada parâmetro de cada algoritmo. Para cada combinação de parâmetros, ocorre uma execução do método em que é alterado um deles por um valor listado anteriormente. Esse processo repete-se até que todas as combinações entre os valores listados tenham sido executadas. Após esse processo ser executado, é escolhida a combinação de parâmetros que tenha apresentado o melhor resultado. O resultado é aferido por meio de métricas de avaliação de modelos, e, neste trabalho, foram utilizadas as medidas AUC (Area Under Curve), *precision* e *recall*.

A AUC pode ser interpretada como a porcentagem da variabilidade de uma variável resposta que pode ser representada por determinado modelo. Outra interpretação possível é a probabilidade de que o modelo acerte a classe de determinado elemento escolhido aleatoriamente. A métrica *precision* diz qual a probabilidade de um elemento realmente pertencer à classe estudada dentro de todos os elementos classificados como tal pelo modelo. A métrica *recall* mede qual a porcentagem dos elementos da classe estudada foram capturados pelo modelo.

Aplicar o Grid Search pode fazer com que se descubra os melhores parâmetros para a base de treino, mas não necessariamente os melhores parâmetros para a população inteira, causando *overfitting*. Para evitar este problema, aplicaram-se técnicas de validação cruzada. Foi utilizada a técnica 10-fold, em que a base de treino é dividida em 10 partes iguais e o mesmo modelo é executado 10 vezes, em cada uma das vezes, uma das partes é usada para validação e as outras para treinamento. Assim como as técnicas de modelagem, tanto o *grid search* quanto a validação cruzada foram realizadas utilizando o pacote Scikit-Learn.

4 RESULTADOS

Conforme descrito na seção anterior, foram treinados modelos em uma base balanceada e em outra mantendo a proporção original dos eventos. Os resultados mostram que balancear a amostra foi uma abordagem melhor em quase todas as técnicas e métricas observadas. No Quadro 3 é possível observar os resultados obtidos para cada modelo executado. Os valores em negrito mostram os melhores resultados para cada métrica, comparando a execução entre as bases e entre as técnicas testadas. A métrica *precision* apresentada pelo Gradient Boosting revelou-se melhor na base desbalanceada, porém a pequena diferença não é o suficiente para inferir que o modelo se sairia melhor em produção. Ressalta-se que todos os modelos tiveram a *performance* avaliada na mesma base reservada para testes, conforme descrito no item 3.3 deste trabalho. A coluna 3 do Quadro 3 indica as medidas de *performance* calculadas para as técnicas ajustadas na base original. A coluna 4 do Quadro 3 apresenta as medidas de *performance* calculadas para as técnicas ajustadas na base balanceada.

Quadro 4 - Medidas de *performance* dos modelos

Técnica	Medida	Base Original	Amostra Balanceada
Árvore de Decisão	AUC	0,51	0,58
	Precision	0,11	0,11
	Recall	0,49	0,49
Random Forest	AUC	0,50	0,60
	Precision	0,11	0,12
	Recall	0,02	0,47
Gradient Boosting	AUC	0,50	**0,61**
	Precision	**0,14**	0,12
	Recall	0,02	0,53
SVM	AUC	0,50	0,51
	Precision	0,11	0,11
	Recall	0,02	0,47
Elasticnet	AUC	0,52	0,61
	Precision	0,07	0,07
	Recall	0,63	**0,63**

Fonte: Elaboração própria

Conforme o modelo que apresentou o melhor desempenho: o Gradient Boosting, que foi treinado na base balanceada, pôde-se analisar as variáveis mais importantes para explicar a ocorrência de fraudes. O Quadro 4 elenca as variáveis mais importantes do modelo. A análise das variáveis importantes para o modelo pode ser feita por um especialista no processo de compras governamentais a fim de criar novas regras para a redução de fraudes.

Quadro 5 - Significado das categorias de modalidade da licitação.

Variável	**Importância**
Valor inicial	0,309577
Número de Aditivos	0,290803
Tempo	0,146157
Modalidade da Licitação_Pregão	0,05489
Modalidade da Licitação_Inexigibilidade_de_licitação	0,002476
Modalidade da Licitação_Concorrência Internacional	0,000441
UASG_Centro Federal de Educação	0,000417
UASG_Delegacia da Receita Federal 1	0,000202
UASG_Delegacia da Receita Federal 2	0,000195
UASG_Departamento de Engenharia e Construção	0,000086
Ramo do Negócio_Organismos Internacionais	0,000014
Modalidade da Licitação_Concurso	0,000014
UASG_Departamento da Policia Federal	0
Modalidade da Licitação_Concorrência Internacional por Técnica e Preço	0
UASG_Restaurante Universitário	0

Fonte: Elaboração própria

Analisando o quadro 6 com os coeficientes obtidos através do algoritmo regressão logística, um dos resultados obtidos é o sinal que aquela variável apresentaria caso fosse utilizada. Observar este sinal faz sentido pois as variáveis foram utilizadas por outras técnicas apresentando resultados melhores, assim, serve como uma explicação para os modelos mais complexos. Caso o sinal seja negativo, significa que quanto maior o valor daquela variável, maior a probabilidade de o contrato analisado não pertencer a uma empresa inidônea. Caso o sinal seja positivo, tem-se o comportamento oposto, aumentando a probabilidade do contrato ser de empresa inidônea.

É importante frisar que o resultado apresentado pela regressão linear ficou muito aquém do apresentado por outras técnicas, apresentando uma AUC de 0.51, que é muito próximo do aleatório. Além disso, os valores dos coeficientes apresentados são muito pequenos, o que permitiria mudanças de interpretação com a alteração de poucos registros estudados. Desta forma, a análise destes coeficientes é apenas uma sugestão de interpretação e precisa ser validada por profissionais do setor ou técnicas de explicação de modelos mais robustas que não serão abordadas neste trabalho.

Podemos identificar que a modalidade de licitação pregão aumenta a probabilidade de um contrato ser de empresa inidônea enquanto as outras modalidades ou não se mostraram variáveis importantes, ou apresentam redução na probabilidade de corrupção. Esse comportamento também pode ser identificado no gráfico da figura 6.

O valor inicial indica que quanto maiores os valores destas variáveis, maior será a probabilidade de fraude. Número de aditivos e tempo apresentaram uma relação inversa com a variável resposta, o que mostra um comportamento diferente do esperado. Uma unidade administrativa (UASG) apresentou sinais de que aumenta a probabilidade de corrupção, uma delegacia da Receita Federal. Todas as outras UASG apresentaram coeficientes negativos ou zerados.

Quadro 6 – Coeficientes das variáveis ajustadas na regressão logística.

Variável	Coeficiente
INTERCEPTO	0,000000053491
NÚMERO DE ADITIVOS	-0,000000133864
VALOR INICIAL	0,000000134717
TEMPO	-0,000104450001
UASG CENTRO FEDERAL DE EDUC TECNOLOGICA SAO PAULO	-0,000000002509
UASG RESTAURANTE UNIVERSITARIO	0,000000000000
UASG DEPARTAMENTO DE ENGENHARIA E CONSTRUÇÃO	-0,000000001546
UASG DELEGACIA DA RECEITA FEDERAL	0,000000000444
UASG DELEGACIA DA RECEITA FEDERAL 2	-0,000000000408
UASG_DPF	0,000000000000
CONCURSO	-0,000000003675
CONCORRÊNCIA INTERNACIONAL POR TÉCNICA E PREÇO	0,000000000000
CONCORRÊNCIA INTERNACIONAL	-0,000000004619
PREGÃO	0,000000899677
INEXIGIBILIDADE DE LICITAÇÃO	-0,000000300050
RAMO DO NEGÓCIO ORGANIZAÇÕES INTERNACIONAIS	-0,000000001131

Fonte: Elaboração própria

5 CONCLUSÃO

Com os estudos feitos neste trabalho, conclui-se que é viável obter dados públicos que permitam afirmar se determinada licitação pertence ou não a uma empresa fraudadora. Os dados utilizados foram publicados por órgãos do governo mediante o portal dados.gov.br, que faz parte de um esforço governamental para aumentar a transparência de seus processos aos cidadãos.

A partir dos dados coletados, foi possível realizar a criação de modelos de aprendizado de máquina que classificam, com um determinado grau de confiança, se uma licitação foi vencida por uma empresa inidônea. Isso permite que órgãos de governança pública como a Controladoria Geral da União possam priorizar a investigação de contratos com maior chance de estarem sendo fornecidos por empresas fraudulentas.

Com análises descritivas simples dos dados já é possível encontrar casos bastante suspeitos, como o contrato de compra de um scanner com uma duração de 85 anos. Também é possível ver que contratos de empresas inidôneas costumam não apresentar prazos com números “redondos”. Enquanto os contratos bons apresentam uma concentração maior em prazos de 6 meses ou 1 ano os contratos ruins tem prazos intermediários.

Nas variáveis categóricas, percebe-se uma concentração maior de fraudes em determinadas categorias. Um exemplo é o ramo de negócio “comércio varejista” que apresenta uma proporção de fraudes 3 vezes maior que a proporção de contratos bons. Também é possível encontrar municípios com uma concentração maior de fraudes. Uma forma de usar isso para evitar a corrupção seria utilizar mecanismos de auditoria mais rígidos e transparentes para licitações em categorias suspeitas.

Além disso, é possível interpretar as variáveis mais importantes do modelo para alterar determinados processos governamentais a fim de dificultar a ocorrência de novas fraudes. Por exemplo, é possível perceber que empresas que ganharam concorrências técnicas são menos prováveis de serem fraudadoras. Uma forma de reduzir a corrupção em compras públicas seria encontrar meios de incentivar esse tipo de processo de compra em detrimento a outros, nos quais ocorrem mais fraudes.

A modalidade de licitação mais utilizada é o pregão. Isso devido à simplificação do processo de compra e à comprovada eficiência do processo, reduzindo gastos em diversos órgãos governamentais. Os contratos firmados por meio dessa modalidade são os que apresentam maior probabilidade de pertencerem a empresas inidôneas. Isso acontece por

diversos motivos, como o conluio entre administradores públicos e empresas privadas ou especificações que favoreçam determinada empresa.

Apesar disso, a proibição da utilização da modalidade pregão poderia ser desastrosa causando atrasos nos processos licitatórios e trazendo consequências para a população. Nesse caso, a redução da corrupção passa pelo aumento da transparência dos processos, por conseguinte, facilitando o acesso de todos os cidadãos aos dados que permitem a auditoria autônoma das compras governamentais.

Este trabalho indica um modelo que torna possível a priorização de quais contratos deveriam ser auditados, dado que apresentam uma probabilidade maior de pertencerem a empresas corruptas ou empresas que tenham um comportamento semelhante ao de empresas corruptas e deveriam ser investigadas.

Nesse sentido, dada a questão teórica apresentada por este trabalho – O aprendizado de máquina pode auxiliar na detecção de fraudes em contas públicas? –, é possível respondê-la positivamente. As técnicas de aprendizado de máquina podem auxiliar na indicação de contratos que foram fornecidos por empresas inidôneas e, assim, auxiliar na detecção de fraudes contra o patrimônio público.

É importante notar que este trabalho se limitou a investigar dados de licitações públicas que foram divulgadas pelo Ministério do Planejamento, bem como analisou apenas as empresas cadastradas no CEIS, mantido pela Controladoria Geral da União. Dadas as variáveis mais importantes aqui mencionadas, trabalhos futuros podem aprofundar as análises buscando dados que melhor expliquem a relação do tempo de vigência do contrato com a ocorrência de fraudes; investigar as diferenças entre contratos com um grande número de aditivos e com valores iniciais muito altos. Outros estudos podem implementar uma segmentação do público a fim de investigar especificamente os ramos de negócios mais propensos a apresentar fraudes, como é o caso de contratos firmados com empresas de comércio varejista e construção civil.

REFERÊNCIAS BIBLIOGRÁFICAS

Agresti, A. (2013). *Categorical Data Analysis* (3rd. ed.). Wiley.

Albuquerque Nobrega, T. C., & Brito, M. F. L. (2019). A nova lei de licitações no brasil/a licitação diante das transições legislativas. *Revista de Estudos e Pesquisas Avançadas do Terceiro Setor, 5*(2), 68-98.

Allison, P. (2012). *Logistic Regression for Rare Events*. Recuperado de http://statisticalhorizons.com/logistic-regression-for-rare-events.

Azevedo, A. I. R. L., & Santos, M. F. (2008). *KDD, SEMMA and CRISP-DM: a parallel overview. IADS-DM*.

Baudat, G., & Anouar, F. (2001, July). Kernel-based methods and function approximation. In IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222) (Vol. 2, pp. 1244-1249). IEEE.

Brasil. (2013). Lei 12.846. Recuperado de http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2013/lei/l12846.htm

Brasil. (2002). Lei 10520. Recuperado de http://www.planalto.gov.br/ccivil_03/leis/2002/l10520.htm.

Brasil. (1992). Lei 8429. Recuperado de http://www.planalto.gov.br/ccivil_03/leis/l8429.htm.

Brasil. (1992). Lei 8443. Recuperado de http://www.planalto.gov.br/ccivil_03/leis/l8443.htm.

Brasil. (2019) Recuperado de http://www.portaltransparencia.gov.br/sancoes. Acessado em 20/07/2019a.

Brasil. Tribunal de contas da União. (2010). *Licitações e contratos: orientações e jurisprudência do TCU*. Recuperado de http://www.planalto.gov.br/ccivil_03/leis/l8666cons.htm.

Burez, J., & Van Den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications, 36*(Issue 3, 4626-4636.

Bürger, F. (2010. Recuperado de https://commons.wikimedia.org/wiki/File:Svm_intro.svg.

Castro, L. I. D. (2007). *Combate à corrupção em licitações públicas*. Recuperado de https://e-archivo.uc3m.es/handle/10016/719.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. *SPSS inc, 16*.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

Cressey, D. R. (1953). *Other people's money; a study of the social psychology of embezzlement.* Recuperado de http://dados.gov.br.

De Aquino, M. M. F., Sousa, R. G., Cavalcante, G. M., Duarte, F. M., & Aquino, F. C. D. O. B. (2018). Mecanismos de controle nos processos licitatórios: A percepção dos pregoeiros. *Revista Ambiente Contábil, 10*(2), 303-325.

Dean, J., & Ghemawat, S. (2004). *Mapreduce: simplified data processing on large clusters. OSDI'04. Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation.* Berkeley, CA: USENIX Association.

Dematté, F. R. (2009). Punição de empresas por corrupção em licitações e contratos com o governo.

Faria, L., & Bianchi, B. G. (2018). Improbidade administrativa e dano ao Erário presumido por dispensa indevida de licitação: uma crítica à jurisprudência do Superior Tribunal de Justiça. *A&C-Revista de Direito Administrativo & Constitucional, 18*(73), 163-187.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*. p. 27-38.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hilbe, J. M. (2009). *Logistic regression models*. Chapman and hall/CRC.

Hoerl, A. and Kennard, R. (1988). *Ridge regression*. In Encyclopedia of Statistical Sciences, vol. 8, pp. 129–136. New York: Wiley

Hunt, L.; & Jorgensen, M. (2003). Mixture model clustering for mixed data with missing information. *Computational Statistics & Data Analysis*, *41*(Issues 3-4, 429-440. ISSN 0167-9473.

Jagannath, V. (2018) Recuperado de https://commons.wikimedia.org/wiki/File:Random_forest_diagram_complete.

Jerez, J., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine, 50* (Issue 2), 105-115. ISSN 0933-3657.

King, G., & Zeng L. (2001). Logistic regression in rare events data. *Political Analysis*, 137-163.

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. J Stat Softw, 36(11), 1-13.

Machado, M. R. R., & Gartner, I. R. (2017, agos.). Triângulo de fraudes de Cressey (1953) e teoria da agência: estudo aplicado a instituições bancárias brasileiras. *Revista Contemporânea de Contabilidade, 14*(32), 108-140. ISSN 2175-8069.

McClelland, J. L., Rumelhart, D. E., & PDP Research Group. (1987). *Parallel distributed processing* (vol. 2). Cambridge, MA: MIT press.

Meng, et al. (2016). MLlib: Machine Learning in Apache Spark. *Journal of Machine Learning Research*, *17*, 1-7.

Monard, M. C., & Baranauskas, J.A. (2003). *Conceitos sobre Aprendizado de Máquina. Sistemas Inteligentes-Fundamentos e Aplicações*.

Orth, C. F. (2012). *Indicadores socioeconômicos como determinantes do nível de corrupção nos municípios brasileiros: uma análise a partir de regressão espacial.*

Owen, S., Anil, R., Dunning T, & Friedman E. (2011). *Mahout in Action*. Manning.

Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python. *JMLR 12*, 2825-2830.

Ponte, L. R., & Heringer, M. (2015) Autor da Lei de Licitações: "RDC propicia a corrupção e a injustiça". Recuperado de https://www.causp.gov.br/autor-da-lei-de-licitacoes-rdc-propicia-a-corrupcao-e-a-injustica/.

Prati, R. C., Batista, G. E., & Monard, M. C. (2009). Data mining with imbalanced class distributions: concepts and methods (pp. 359-376). In 4th Indian International Conference on Artificial Intelligence. Tumkur, Karnataka, India.

Schramm, F. S. (2018). *O compliance como instrumento de combate à corrupção no âmbito das contratações pública.*

Serviço Brasileiro de Apoio às Micro e Pequenas Empresas (Sebrae*). Compras Públicas: um bom negócio para a sua empresa*. Brasília: Sebrae, 2014.

Tan, P. N. (2018). Introduction to data mining. Pearson Education India.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288.

Van Der Paal, B., & Benoit, D. (2014). *A Comparison of Methods for Modelling Rare Events Data*.

Vasconcelos, F. (2005). Licitação pública: análise dos aspectos relevantes do Pregão. Prim@ facie: *Revista da Pós-Graduação em Ciências Jurídicas, 4*(7), 151-163.

Zaharia, M., Chowdhury, M. J., Franklin, S., & Shenker I. (2010). Stoica. Spark: cluster computing with working sets. *In Proc. HotCloud, 10*.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, *67*(2), 301-320.

ANEXO A – CRIAÇÃO DAS BASES

```
# coding: utf-8



from urllib.request import urlopen
import urllib


ff = open("contratos_todos2.csv", "wb")
reg_num=0
while reg_num<594512:
   link                                                    =
"http://compras.dados.gov.br/contratos/v1/contratos.csv?offset
="+str(reg_num)
   try:
       f = urlopen(link)
       ff.write(f.read())
       f.close()
       ff.flush()
       print(str(reg_num)+" gravado")
   except:
       print(str(reg_num)+" errado")
       pass
   reg_num=reg_num+500
ff.close()
```

ANEXO B – MODELAGEM

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import Imputer, LabelBinarizer, StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import GridSearchCV
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
from sklearn.svm import SVC
import matplotlib.pyplot as plt

import os
import pickle

df = pd.read_csv('contratos_todos2.csv')

df.columns

cnpj = df['CNPJ Contratada']

cnpjs = []

for c in list(cnpj):
    if c == 'CNPJ Contratada':
        cnpjs.append(np.nan)
    else:
        cnpjs.append(int(c.split(':  ')[0].replace('Fornecedor  ', '').replace('.','').replace('-
','').replace('/','')) if type(c) is str else c)

df['cnpj'] = cnpjs

excluidos = pd.read_csv('20180301_CEIS.csv', sep=";", encoding='latin-1')

excluidos.columns

df_model = pd.merge(df, excluidos, left_on=['cnpj'], right_on=['CPF ou CNPJ do Sancionado'],
how='left', indicator='Exist')
df_model['Exist'] = np.where(df_model.Exist == 'both', 1, 0)

df_model.columns
```

```
df_model = df_model[df_model['UASG'] != 'UASG']

df.shape

df_model.shape

df_model[df_model['Exist']==1].shape

df_model['Data de Início da Vigência']

import datetime
ass = list(df_model['Data de Assinatura'])
fin = list(df_model['Data Final Sanção'])
datas = []
dataf = []
deleta = []
for n, i in enumerate(ass):
    if type(i) is str and type(fin[n]) is str:
        if datetime.datetime.strptime(i, '%d/%m/%Y')>datetime.datetime.strptime(fin[n],
'%d/%m/%Y'):
            deleta.append(1)
        else:
            deleta.append(0)
    else:
        deleta.append(0)
len(datas)

len(deleta)

len(ass)

df_model['deleta'] = deleta

df_model = df_model[df_model['deleta']==0]

import datetime
ini = list(df_model['Data de Início da Vigência'])
ini_t = []
fin = list(df_model['Data de Termino da Vigência'])
fin_t = []
for i in ini:
    ini_t.append(datetime.datetime.strptime(i if type(i) is str else '12/05/2018', '%d/%m/%Y'))
for f in fin:
    fin_t.append(datetime.datetime.strptime(f if type(f) is str else '12/05/2018', '%d/%m/%Y'))
```

```
dif = []
for i in range(len(ini_t)):
   d = fin_t[i]-ini_t[i]
   dif.append(d.days)

df_model['tempo'] = dif

df_model[df_model['Exist']==1][['Data de Assinatura','Data de Início da Vigência', 'Data de Termino da Vigência','Data Início Sanção', 'Data Final Sanção']]

44727/595701

df_model_clean = df_model.drop(['Identificador do Contrato',
      'Número do Aviso da Licitação', 'Código do Contrato',
      'Licitação associada', 'Número', 'Objeto',
      'Número do Processo', 'CPF Contratada',
      'CNPJ Contratada', 'Data de Assinatura', 'Fundamento Legal',
      'Data de Início da Vigência', 'Data de Termino da Vigência',
      'Aditivos do Contrato > uri',
      'Apostilamentos do Contrato > uri', 'Eventos do Contrato > uri', 'cnpj',
      'Tipo de Pessoa', 'CPF ou CNPJ do Sancionado',
      'Nome Informado pelo Órgão Sancionador',
      'Razão Social - Cadastro Receita', 'Nome Fantasia - Cadastro Receita',
      'Número do processo', 'Tipo Sanção', 'Órgão Sancionador', 'UF Órgão Sancionador',
      'Origem Informações', 'Data Origem Informações', 'Data Publicação',
      'Publicação', 'Detalhamento',
      'Abrangência definida em decisão judicial', 'Fundamentação Legal',
      'Descrição da Fundamentação Legal', 'Data do Trânsito em Julgado',
      'Complemento do Órgão', 'Contato da Origem da Informação', 'Data Início Sanção',
      'Data Final Sanção', 'UASG', 'deleta'
], axis=1)

X_0 = df_model_clean[df_model_clean['Exist']==0].sample(n=df_model_clean[df_model_clean['Exist']==1].shape[0])

X_1 = df_model_clean[df_model_clean['Exist']==1]

X = pd.concat([X_0, X_1])

y = X['Exist']

X = X.drop(['Exist'], axis=1)

X.columns
```

```
X = X[X['Número de Aditivos'] != 'Número de Aditivos']

X.columns

Mod = []
for c in list(X['Modalidade da Licitação']):
   Mod.append(int(c.split(': ')[0]) if type(c) is str else c)

Valor = []
for c in list(X['Valor inicial']):
   Valor.append(float(c.replace(',', '').replace('R$ ', '')) if type(c) is str else c)

Origem = []
for c in list(X['Origem da Licitação']):
   Origem.append(0 if c == 'SISPP' else 1)

X['Modalidade da Licitação'] = Mod

X['Valor inicial'] = Valor

X['Origem da Licitação'] = Origem

X['Número de Aditivos'] = X['Número de Aditivos'].fillna(0)

adt = list(X['Número de Aditivos'])
for n, i in enumerate(adt):
   adt[n]=int(i)

X['Número de Aditivos'] = adt

X = X.fillna(-9999)

X = pd.get_dummies(X, drop_first=True,
                   columns=['Modalidade da Licitação', 'Origem da Licitação'])

list(X.columns)

X_train, X_test, y_train, y_test = train_test_split(
   X, y, test_size=0.33, random_state=42)

X_test.shape

sum(y_train)/y_train.shape[0]
```

```
sum(y_test)/y_test.shape[0]

X_train.shape

from sklearn import tree
clf = tree.DecisionTreeClassifier()
clf = clf.fit(X_train, y_train)

clf

vars = pd.DataFrame(list(zip(list(clf.feature_importances_), list(X.columns))))

vars.sort_values(by=[0], ascending=False)

set(df['Modalidade da Licitação'])

clf.tree_.max_depth

pred = clf.predict(X_test)

metrics.roc_auc_score(y_test,pred)

print(metrics.classification_report(y_test,pred))

param_grid = {
    'n_estimators': [3000],
    #'min_impurity_decrease': [0],
    'max_depth' : [20],
    #'criterion' :['gini', 'entropy']
}

rfc=RandomForestClassifier(random_state=42)
rf = GridSearchCV(estimator=rfc, param_grid=param_grid, cv=10, n_jobs=-1, verbose=3)
rf.fit(X_train, y_train)

rf.best_score_

pred_rf = rf.predict(X_train)

metrics.roc_auc_score(y_train,pred_rf)

pred_rf = rf.predict(X_test)

metrics.roc_auc_score(y_test,pred_rf)
```

```
param_grid = {
    'n_estimators': [100, 200, 300, 500],
    #'min_impurity_decrease': [0],
    'max_depth' : [10, 20, 30, 72],
    #'criterion' :['gini', 'entropy']
}

rfc=RandomForestClassifier(random_state=42)
CV_rfc = GridSearchCV(estimator=rfc, param_grid=param_grid, cv=10, n_jobs=-1,
verbose=3)
CV_rfc.fit(X_train, y_train)

CV_rfc.best_score_

CV_rfc.best_estimator_

pred_cvrf = CV_rfc.predict(X_test)

p = CV_rfc.predict_proba(X_test)

metrics.roc_auc_score(y_test,pred_cvrf)

metrics.accuracy_score(y_test,pred_cvrf)

metrics.average_precision_score(y_test,pred_cvrf)

metrics.f1_score(y_test,pred_cvrf)

print(metrics.classification_report(y_test,pred_cvrf))

%matplotlib inline
from sklearn.metrics import precision_recall_curve
import matplotlib.pyplot as plt

precision, recall, _ = precision_recall_curve(y_test,p[:,1])

plt.step(recall, precision, color='b', alpha=0.2,
         where='post')
plt.fill_between(recall, precision, step='post', alpha=0.2,
                 color='b')

plt.xlabel('Recall')
plt.ylabel('Precision')
plt.ylim([0.0, 1.05])
plt.xlim([0.0, 1.0])
```

```
plt.title('2-class Precision-Recall curve: AP={0:0.2f}'.format(metrics.average_precision_score(y_test,pred_cvrf)))

fpr, tpr, t = metrics.roc_curve(y_test,pred_cvrf)

plt.figure()
lw = 2
plt.plot(fpr, tpr, color='darkorange',
        lw=lw, label='ROC curve (area = %0.2f)' % metrics.roc_auc_score(y_test,p[:,1]))
plt.plot([0, 1], [0, 1], color='navy', lw=lw, linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic example')
plt.legend(loc="lower right")
plt.show()

from sklearn.ensemble import GradientBoostingClassifier

gb = GradientBoostingClassifier(max_depth=15, n_estimators=300, random_state=0)
gb.fit(X_train, y_train)

vars = pd.DataFrame(list(zip(list(gb.feature_importances_), list(X.columns))))

vars.sort_values(by=[0], ascending=False)

pred_gb = gb.predict(X_test)

proba_gb = gb.predict_proba(X_test)

metrics.roc_auc_score(y_test,pred_gb)

metrics.accuracy_score(y_test,pred_gb)

metrics.average_precision_score(y_test,pred_gb)

metrics.f1_score(y_test,pred_gb)

print(metrics.classification_report(y_test,pred_gb))

%matplotlib inline
from sklearn.metrics import precision_recall_curve
import matplotlib.pyplot as plt
```

```
precision, recall, _ = precision_recall_curve(y_test,proba_gb[:,1])

plt.step(recall, precision, color='b', alpha=0.2,
        where='post')
plt.fill_between(recall, precision, step='post', alpha=0.2,
                color='b')

plt.xlabel('Recall')
plt.ylabel('Precision')
plt.ylim([0.0, 1.05])
plt.xlim([0.0, 1.0])
plt.title('2-class Precision-Recall curve: AP={0:0.2f}'.format(metrics.average_precision_score(y_test,pred_gb)))

fpr, tpr, _ = metrics.roc_curve(y_test,pred_gb)
plt.figure()
lw = 2
plt.plot(fpr, tpr, color='darkorange',
        lw=lw, label='ROC curve (area = %0.2f)' % metrics.roc_auc_score(y_test,proba_gb[:,1], average='micro'))
plt.plot([0, 1], [0, 1], color='navy', lw=lw, linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic example')
plt.legend(loc="lower right")
plt.show()

X.describe()

X[['Valor inicial', 'tempo', 'Número de Aditivos']].describe()

X[['Valor inicial', 'tempo', 'Número de Aditivos']].hist(bins=25)

X[X['Valor inicial']<400000][['Valor inicial']].boxplot()

X[X['tempo']<1000][['tempo']].boxplot()

X[X['Número de Aditivos']<10][['Número de Aditivos']].boxplot()
```