PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA

Pós-graduação *Lato Sensu* em Inteligência Artificial e Aprendizado de Máquina

Leonardo Teles de Carvalho

DETECÇÃO DE ANOMALIAS NOS GASTOS DA COTA PARLAMENTAR

Belo Horizonte
2020

Leonardo Teles de Carvalho

DETECÇÃO DE ANOMALIAS NOS GASTOS DA COTA PARLAMENTAR

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Inteligência Artificial e Aprendizado de Máquina como requisito parcial à obtenção do título de especialista.

Belo Horizonte
2020

SUMÁRIO

1. Introdução

1.1. Contextualização

Os desafios do combate a fraudes financeiras sempre foram muito grandes. Há um enorme esforço de governos e empresas em busca da rápida e precisa detecção desses ilícitos. Cada aprimoramento nesses esforços, entretanto, é seguido, e muitas vezes antecipado, por mudanças nos comportamentos dos fraudadores. Estes, em um ambiente bastante dinâmico, ora aprendem com a fiscalização, ora inovam em busca do ganho ilegal.

Um exemplo da abrangência desse problema é evidenciado em pesquisa publicada em (CNDL,2019) em que a Confederação Nacional de Dirigentes Lojistas afirma que 2 em cada 10 brasileiros já foram vítimas de fraudes financeiras no ano de 2019. A CNDL aponta as transações em ambiente digital como principal foco para o cometimento desses crimes.

Num escopo mundial, o prejuízo contabilizado apenas em crimes relacionados ao mercado de cartões de crédito, segundo (THE NILSON REPORT, 2019), alcançou a marca de 27 bilhões de dólares no ano de 2018. Em um setor que movimentou 40 trilhões de dólares naquele ano, a *Nilson Report* estimou que 6.86 centavos foram perdidos em fraudes a cada 100 dólares gastos por meio de cartões de crédito e débito.

Em (CORREIO BRAZILIENSE, 2020), noticia-se que apenas um tipo de fraude relacionado a criação de poupanças virtuais para recebimento do Benefício Emergencial devido a pandemia de COVID-19 gerou um prejuízo de mais de 60 milhões de reais para os cofres públicos.

Soma-se a essa disputa uma pressão do mercado por uma maior praticidade e agilidade dos serviços financeiros além da justificada e desejada diminuição da burocracia. Vários são os setores econômicos nessa corrida, desde administradoras de cartão de crédito e comércios varejistas até órgãos governamentais, dada a pressão da sociedade por um ambiente com menor regulação. O surgimento de diversos serviços digitais ágeis por meio de aplicativos no mercado de transporte, de hospedaria, de *streaming* de áudio e vídeo e até mesmo a ascensão das criptomoedas são exemplos do vigor dessas novas tendências.

Ademais, pode-se afirmar que modelos tradicionais de detecção de fraudes são, tipicamente, baseados em regras fixas, construídas por especialistas. Essa

laboriosa e dispendiosa tarefa deve ser sempre atualizada, dada sua característica estática.

É nesse panorama que a utilização de tecnologias baseadas em Aprendizado de Máquina (AM) se torna bastante atrativa no combate a ilícitos fiscais e financeiros. Modelos dinâmicos e capazes de detectar padrões complexos e classificar fraudes em um volume crescente de dados poderiam, em tese, alavancar a precisão dos resultados dos especialistas.

Há, entretanto, algumas barreiras significativas a serem ultrapassadas. Na construção de tais modelos é desejável um rico histórico de casos rotulados. E esse histórico, na área em estudo, é bastante incomum. Usualmente, tem-se apenas um pequeno conjunto de casos históricos em que foi possível detectar fraude e uma enorme quantidade de casos não classificados. Dada a característica dinâmica do fraudador, esses poucos casos rotulados podem não ser suficientes para antever novas ações dolosas. O exemplo noticiado em (CORREIO BRAZILIENSE, 2020) é um exemplo dessa dinamicidade.

Ademais, há uma dificuldade em se explicar como os modelos de elevada precisão construídos por meio de AM geram seus resultados. E isso é essencial, uma vez que as auditorias que identificam fraudes devem sempre caracterizar com precisão o que aconteceu a fim de tipificar com clareza o ilícito.

Esta é uma área em que o simples resultado de um algoritmo classificador, ainda que com acurácia bastante elevada, é apenas uma fração da investigação de fraudes. Grande parte do trabalho dos analistas que lidam com esse problema consiste em bem caracterizar os indícios. Os pormenores de cada caso identificado são fundamentais para uma eventual ação judicial exitosa contra o fraudador. O resultado de um classificador que aponte para, por exemplo, 99,9% de chances de haver alguma anomalia é inútil se não embasar sua lógica em argumentos inteligíveis aos especialistas.

Por vezes, essa necessidade é suprida abrindo-se mão de um modelo mais sofisticado e utilizando-se modelos mais simples, bem menos precisos, mas que consigam expor os critérios que conduziram à identificação do fraudador. Há um aparente conflito entre a precisão (e acurácia) de modelos de Aprendizado de Máquina e sua explicabilidade ou interpretabilidade. E essa é uma questão primordial no uso de tais tecnologias na averiguação de atividades ilegais.

1.2. A Fraude como uma Anomalia no Conjunto de Dados

Os desafios que permeiam a detecção de fraudes utilizando-se AM, como visto anteriormente, são bastante variados. Não há, atualmente, ferramenta única que consiga responder às muitas questões demandadas por esta área.

Isto posto, uma abordagem bastante promissora é a visualização de eventos de fraude como anomalias em conjuntos de dados. As técnicas de detecção de anomalias consistem em um arcabouço de ferramentas capazes de encontrar padrões inconformes em relação ao comportamento esperado de um conjunto de dados.

Ora, é grande a interseção entre as características de fraudes e anomalias:

- **Desbalanceamento entre classes -** Pela própria definição das técnicas de detecção de anomalia, há grande desbalanceamento entre as classes normais e anômalas. Isso também acontece com as fraudes, que são, em sua maioria, eventos raros;
- **Ausência de rótulos –** É incomum que se tenha muita informação a respeito do processo gerador das anomalias. São, portanto, poucos os dados rotulados. E, muitas vezes, somente poucas instâncias de apenas uma das classes de comportamentos (normal ou fraude) é conhecida. Esse cenário também é usual no caso da detecção de fraudes, em que algumas poucas instâncias tem sua classe conhecida e há uma grande massa de dados com comportamento médio conforme, mas contaminado com amostras inconformes;
- **Comportamento atípico -** Além de escassos, os dados anômalos apresentam comportamento atípico, como se fossem originados de um outro processo gerador que não o dos dados em análise. Essa é, talvez, a principal característica que fundamenta a utilização dessas ferramentas em fraudes. Há uma certa expectativa de que o comportamento de fraudadores, embora possa se confundir com o comum, em alguns pontos será bastante distinto deste. E, supõe-se, o conjunto de dados analisados será capaz de evidenciar essas diferenças;

Embora seja óbvio que nem sempre as anomalias correspondam a fraudes, dadas as características comuns às duas, as técnicas de detecção de anomalias são consideradas boas candidatas na identificação inicial de fraudes. Destas, pode-se

destacar duas que tem obtido bons resultados em dados tabulares: *Isolation Forest* e redes *AutoEncoder*.

Isolation Forest (como em (LIU, 2008)) é uma técnica não supervisionada, baseada em Árvores de Decisão que conseguiu ótimos resultados em vários conjuntos de dados utilizados como referência na área como *Http and Smtp* da *KDD CUP 99*, *network intrusion data, Annthyroid Arrhythmia*, *Wisconsin Breast Cancer (Breastw)* entre outros.

Redes Autoencoder obtiveram bons resultados como redutores de dimensionalidade na detecção de fraudes em cartões de crédito em (MISRA,2019). Além disso, em (AGGARWAL, 2017), o autor considera tal topologia como boa opção para se capturar relacionamentos não lineares na detecção de anomalias.

Ambos os métodos serão detalhados em seção própria.

1.2. O Aprendizado de Máquina e os gastos com a Cota Parlamentar

É crescente a quantidade de dados históricos, transacionais e bases correlatas para análise de irregularidades. Não somente a quantidade, mas também sua disponibilidade de maneira irrestrita tem visto grande crescimento. Um exemplo disso na legislação brasileira é o advento da Lei Complementar nº 131, de 27 de maio de 2009, que obriga União, Estados e Municípios a divulgar pormenorizadamente seus gastos na Internet.

Concomitantemente, o controle social dos gastos públicos ganha nova relevância e efetividade na medida em que enseja a participação de uma grande comunidade de cidadãos com interesse e conhecimento tecnológico para colaborar com o enfrentamento das fraudes no setor público.

Um exemplo dessa iniciativa foi noticiado em 2017 em (NOGUEIRA, 2017) em que um grupo de desenvolvedores de software analisou os gastos realizados por Deputados na chamada Cota Parlamentar. A cota é um valor que a União destina ao Poder Legislativo a fim de custear alguns gastos relacionados a atividade parlamentar como locomoção, alimentação e hospedagem, como descrito em (CÂMARA DOS DEPUTADOS, 2020a). Esse grupo conseguiu diagnosticar vários indícios de irregularidades. Essa ação autodenominada Operação Serenata de Amor logrou êxito ao motivar vários parlamentares a ressarcir alguns de seus gastos irregulares apontados pelo grupo.

Nessa esteira, passa-se a examinar novamente essa base de gastos parlamentares no intuito de testar algumas ferramentas promissoras em AM na área de detecção de anomalias.

1.3. Detecção e Explicação Local de Anomalias em Gastos na Cota Parlamentar

No presente trabalho, pretende-se abordar uma possível proposta para o exame de gastos da Cota Parlamentar. Foi visto nas seções anteriores que a análise de fraudes está profundamente relacionada com a detecção de anomalias.

Como o simples índice de anomalia mostra-se pouco informativo na identificação de indícios de fraude, buscou-se utilizar a técnica *Shapley Additive Explanation* (Valores SHAP) a fim de se obter a explicação local da instância. Essa técnica, como mostrado em (ANTWARG,2019) é bastante efetiva em explicar anomalias em modelos caixa preta, como *Autoencoder*. (LUNDBERG, 2017) descreve e implementa essa técnica em várias topologias, entre as quais as baseadas em árvores.

Objetiva-se, então, avaliar as técnicas de modelagem *Isolation Forest* e *Autoencoder* para detectar instâncias anômalas nas bases de dados relacionadas aos gastos com a Cota Parlamentar de deputados federais. A avaliação desse índice será feita comparando-se as instâncias anômalas com gastos já identificados como suspeitos. Esses consistem de instâncias relacionadas a diversos casos já noticiados na mídia como suspeitos de fraudes. Por fim, busca-se utilizar Valores SHAP para identificar quais as características daquelas foram relevantes para justificar o índice encontrado.

2 Descrição das Técnicas

Nesta Seção descreve-se brevemente os algoritmos *Isolation Forest*, *AutoEncoder* e Valores SHAP.

2.1 O algoritmo *Isolation Forest*

Muitos dos métodos para detecção de anomalias baseados em modelos se propõem a construir uma representação para dados normais. Assim, dados anômalos seriam identificados como incompatíveis com essa representação.

Em (LIU, 2008), o autor propõe um método bastante interessante para detecção de anomalias, baseado em Árvores de Decisão, denominado *Isolation Forest*. Baseia-se em duas premissas a respeito da anormalidade dos dados:

- São minoria, reduzem-se a apenas algumas instâncias;
- O valor de seus atributos é bem diferente daqueles encontrados em instâncias normais.

Nesse contexto, ao se construir uma Árvore de Decisão, em cada nó, pontos anômalos são mais facilmente separados do restante das instâncias. Assim, há uma maior probabilidade de que essas instâncias sejam isoladas ainda nos nós iniciais da árvore. Utilizando-se várias árvores, pode-se obter uma medida média da quantidade de nós pelos quais cada instância atravessa. E essa medida, segundo o autor, **seria uma boa métrica para indicar anomalias.**

O procedimento de construção da Árvore *Isolation Tree* pode ser resumido nos seguintes passos:

1. Escolhe-se, aleatoriamente, um atributo *q* e um valor de divisão *p* entre os valores máximo e mínimo deste atributo;
2. Classifica-se e faz-se a divisão do Espaço Amostral de acordo com p e, recursivamente, repete-se tal procedimento com novos *p* e *q*;

O método anterior continua até que: (i) a árvore atinja uma altura máxima determinada; (ii) haja apenas uma instância em cada nó filho ou; (iii) que todos os nós tenham a mesma classificação. Isso é feito em partições do Espaço, utilizando-se diversas *Isolation Trees*.

Finalmente, uma métrica de anomalia, denominada Escore de Anomalia *s(x,n)*, é definida para cada amostra baseada na quantidade média de nós necessária para seu isolamento (*E(h(x))*):

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}},$$

em que *x* é a instância, *n* é a quantidade de instâncias e a *c(n)* é uma normalização da função altura da árvore (os detalhes deste cálculo podem ser obtidos em (LIU, 2008)). Visto que as anomalias apresentam valores de *h(x)* tipicamente menores que dados normais, um valor de *s(x,n)* bem maior que a média das amostras apontará para instâncias anômalas.

2.2 Utilização de uma Rede Autoencoder na Detecção de Anomalias

Autoencoder é uma topologia auto associativa (como em (KRAMER, 1991)) de Rede Neural Artificial não supervisionada com algumas características específicas como:

- diferentemente de técnicas clássicas como Análise de Componentes Principais (PCA), é uma topologia que consegue representar relações não lineares entre as variáveis;
- sua estrutura interna é, geralmente, simétrica em relação às camadas escondidas, possuindo uma camada intermediária com quantidade de neurônios menor que o tamanho do vetor de entrada, conhecida como Espaço Latente;
- as camadas anteriores ao Espaço Latente são responsáveis por mapear o vetor de entrada ao espaço e são conhecidas como codificador. Analogamente, as camadas posteriores mapeiam o Espaço Latente ao vetor de saída e são nominadas decodificador da rede;
- a função de mapeamento da rede tem como objetivo produzir vetores de saída idênticos ao vetor de entrada.

Essa constrição do Espaço Latente em relação ao vetor de entrada impõe ao modelo de indução uma restrição na representação adequada entre entrada e saída. Quando presente, esse fato faz com que a rede *Autoencoder* possa ser utilizado como um método de redução de dimensionalidade. Em outros termos, a rede cria, via Espaço Latente, uma representação compacta da distribuição dos dados. E, diferentemente da PCA mencionada anteriormente, possibilita a incorporação de relações não lineares ao aprendizado.

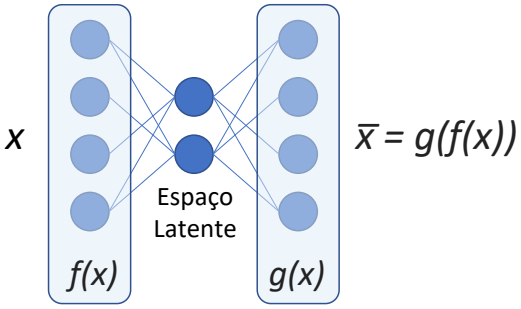
Em (GOODFELLOW, 2016), os autores fazem uma instrutiva revisão dessa topologia e denominam tal imposição à camada escondida de *Undercomplete Autoencoders*, citando várias aplicações como a redução de dimensionalidade e a filtragem de dados ruidosos.

Já a aplicação de *Autoencoders* como detector de anomalias não necessariamente implica em limitação da dimensão do Espaço Latente. O que se deseja é um mapeamento do comportamento normal, regular dos dados. Para isso, é necessário que os casos sabidamente anômalos sejam retirados do grupo de treinamento. Isso, porém, nem sempre é possível. É comum a situação em que os dados de treinamento estejam contaminados por anomalias.

Nessas condições, a manipulação da capacidade de representação do Espaço Latente é extremamente desejável. Isso permite que apenas as características mais representativas sejam modeladas e suas peculiaridades (possíveis *outliers*) sejam

relevadas. Ou seja, a redução de dimensionalidade alcançada pela rede favorece a modelagem do caso normal e a possível identificação de anomalias.

Durante o treinamento, considerando-se *f*, o codificador e *g* a função do Decodificador, tenta-se minimizar uma função perda (*L(x, g(f(x)*) de maneira a fazer com que a diferença média entre vetor de entrada *x* e o vetor de saída g(f(x)) seja a menor possível dada a restrição imposta pela quantidade de neurônios latentes. A Figura 1 exibe essa relação.



Supondo que uma rede *Autoencoder* já treinada tenha êxito em modelar a distribuição latente dos dados, a rede gerará um vetor de saída que, idealmente, seja próximo ao vetor de entrada oferecido. Entretanto, essa correspondência entre os vetores x e $\overline{x}$ só será efetiva se a amostra específica x_1 for gerada pela mesma função densidade de probabilidade modelada pela rede. Uma amostra x_2 incompatível com essa distribuição terá um valor g(f(x_2)) bastante distinta de x_2.

Pode-se, consequentemente, utilizar a função perda *L* baseada na diferença ($\overline{x}$ - *x)*, ou alguma medida derivada semelhante, como indicador de anomalia. Esta função L que, então, em tempo de treinamento é usada para se ajustar o modelo à identidade entre x e $\overline{x}$, pode ser utilizada, após a modelagem, para apontar dados anômalos. Neste trabalho, o Escore de Anomalia será dado pela soma do valor absoluto de ($\overline{x}$ - *x)* de todas as variáveis.

2.3 Shapley Additive Explanation (Valores SHAP) e a Explicabilidade de Modelos Caixa-Preta

A interpretabilidade de modelos de Inteligência Artificial (IA) é um tema de suma importância atualmente, com a crescente utilização de estruturas de Aprendizado Profundo. Redes de Aprendizado Profundo, que atualmente representam o estado da arte em detecção de padrões complexos, em regra, atingem seu alto desempenho ao custo de uma enorme quantidade de parâmetros. A estrutura interna dessas redes, em regra, guarda pouca ou nenhuma relação com a interpretação de seus resultados.

Essa necessidade de se examinar o raciocínio por trás de um simples número gerado por um classificador algorítmico não se dá simplesmente por particularidades das áreas, como é o caso das auditorias de fraude citadas anteriormente. Tal exame se faz necessário até mesmo para se descobrir falhas ou tendências não desejadas dos modelos e dos dados utilizados em seu treinamento. Em (RIBEIRO, 2016), um exemplo caricato explicita esse fato: um modelo de classificação de imagens utilizado para distinguir entre fotos de lobos e cachorros esquimós (*huskies*) apresenta sempre uma tendência a classificar como lobos os cachorros que estão em imagens com presença de neve ao fundo. Ao se analisar os dados de treinamento, percebe-se que todos os dados que apresentam lobos estão em imagens com fundos cobertos por neve. O autor, com esse exemplo, evidencia utilizando a técnica LIME (método introduzido no próprio artigo) que o conteúdo mais relevante utilizado *nessas imagens* para se distinguir as categorias é, para esse modelo Caixa-preta, a presença de neve.

Lundberg, em (LUNDBERG, 2017), a fim de tentar contornar a aparente dicotomia entre explicabilidade e desempenho, organiza a tarefa de Aprendizado de Máquina em duas partes: a criação de um modelo baseado em dados, nos moldes tradicionais, e; a criação de um Modelo de Interpretação (MI). A estratégia clara é desacoplar o problema da interpretabilidade com o de maximizar o desempenho de modelos complexos. Para tanto, se baseia em uma análise de jogos cooperativos advinda da Teoria dos Jogos, formulada por Lloyd Shapley.

O método proposto, denominado Valores SHAP (ou SHapley Additive exPlanation), gera valores para a importância de cada variável de instância, independentemente do modelo de previsão.

Considerando que o modelo utilizado é um classificador *f(z)* que gera uma probabilidade *E[f(z)]* de pertencimento à determinada classe para uma instância qualquer *z*, constrói-se um modelo de interpretação MI que tem como resultado a contribuição de cada variável para essa probabilidade. Essas contribuições se somam de maneira a totalizar o valor da probabilidade *E[f(z)]*.

Ou seja, o modelo MI gera valores ϕ para cada variável *i* de maneira que:

$$E[f(z)] = \phi_0 + \Sigma\phi_i$$

Tal MI modela a contribuição local de cada variável de entrada como o valor esperado da contribuição marginal dessa variável tomado em todas as combinações possíveis de variáveis. Em teoria de conjuntos isso coincide com a definição de Conjunto Potência. Considerando que cada variável é um elemento que contribui para

o resultado do modelo, deve-se, inicialmente, calcular a previsão do modelo para todos os subconjuntos possíveis em que estas variáveis estejam presentes. Como exemplo, se o vetor de entrada é constituído por regressores de valor *a, b* e *c*, para as variáveis A, B e C, obtém-se a previsão do modelo para as entradas *a; b; c; ab; ac; bc; abc* e conjunto vazio.

Então, para cada variável, deseja-se calcular o efeito médio que essa terá na previsão ao ser adicionada aos subconjuntos presentes no Conjunto Potência das demais variáveis. Seguindo o exemplo anterior, a contribuição da variável A tomando o valor de a é soma ponderada das seguintes contribuições marginais: *E[f(z|A=a)] - E[f(z)]; E[f(z|A=a,B=b)] - E[f(z|B=b)]; E[f(z|A=a,C=c)] - E[f(z|C=c)]; E[f(z|A=a,B=b,C=c)] - E[f(z|B=b,C=c)]*.

Em resumo, a contribuição de *A=a* é a variação na resposta do modelo ao se acrescentar a variável A na análise, para todas as combinações possíveis de *features*.

Vale ressaltar que essa abordagem é bastante distinta de análises do tipo *Feature Importance*, existentes em várias modelagens de IA. Estas são globais e analisam a influência de cada atributo no modelo como um todo e não em uma instância específica.

Os valores SHAP constroem a interpretabilidade local de certa observação. E essa propriedade vem exatamente suprir tal deficiência nas identificações de fraudes por modelos complexos.

Assim, pode-se utilizar modelos complexos bastante precisos e obter-se a explicação do porquê de seus resultados utilizando os valores SHAP.

3. Coleta de Dados

O conjunto de dados a ser analisado no presente trabalho é de Despesas cobertas pela Cota para Exercício da Atividade Parlamentar (CEAP) desde o ano de 2008 até o dia 15 de setembro de 2020 utilizada pelos deputados federais.

De acordo com (CÂMARA DOS DEPUTADOS, 2020a), esta é uma cota destinada a custear os gastos dos deputados exclusivamente vinculados ao exercício da atividade parlamentar. São vários os tipos de despesas cobertas por essa cota, como passagens aéreas, alimentação, hospedagem, locação de veículos, combustíveis, contratação de consultorias e divulgação de atividade parlamentar. Há regras específicas para cada tipo de gasto, detalhado, atualmente, no Ato da Mesa nº 43, de 21 de maio de 2009, da Câmara dos Deputados.

Estes dados estão disponíveis em arquivos organizados por ano da despesa e podem ser obtidos em (CÂMARA DOS DEPUTADOS, 2020b).

Na data da extração, 17 de setembro de 2020, estes formavam uma base de 3.925.883 despesas descritas por 31 variáveis diferentes. Estas descrevem uma série de características como: nome e códigos identificadores do parlamentar que efetuou a despesa; valor, descrição e tipo da despesa; nome e CNPJ/CPF do estabelecimento fornecedor. A descrição desses dados está disponível em (CÂMARA DOS DEPUTADOS, 2020c).

Além disso, por meio do CNPJ da empresa identificada, pôde-se obter as informações de seu cadastro junto à Secretaria de Receita Federal do Brasil (SUARA, 2020). Variáveis como idade da empresa na data da Nota Fiscal, porte, capital social e sócios foram extraídos para cada despesa.

Dessas variáveis, foi construído um *dataset* com algumas informações que, à princípio, podem ser relevantes para identificar possíveis anomalias nos gastos. São elas:

Nome da coluna/campo	**Descrição**	**Tipo**
numMes	Mês de competência do gasto (1 a 12)	Numérica
numAno	Ano de competência do gasto	Numérica
idadeGasto	Quantidade de meses passados entre a competência do gasto e a data do documento fiscal	Numérica
sessão.legislativa	Ano do mandato do deputado (1 a 4)	Numérica
codTipo	Código que representa o tipo de gasto, como, por exemplo: locomoção, telefonia, divulgação. Na base há 26 tipos diferentes de gastos.	Categórica
vlrDocumento	Valor do documento fiscal apresentado	Numérica
capital social_empresa	Valor do Capital Social da empresa em que foi realizado o gasto, como obtido no CNPJ	Numérica
idade.empresa	Quantidade de dias passados entre a abertura da empresa relacionada ao gasto e a data do documento fiscal	Numérica
empresa.propria	Indicador que sinaliza se o deputado é sócio da empresa relacionada ao gasto	Categórica binária

Nome da coluna/campo	Descrição	Tipo
empresa.socio	Indicador que sinaliza se algum sócio da empresa relacionada ao gasto é sócio com o deputado em outra empresa	Categórica binária
porte_empresa	Código referente ao porte da empresa relacionada ao gasto constante no CNPJ (há 4 códigos de porte na base)	Categórica
opcao_pelo_mei	Indicador que sinaliza se a empresa relacionada ao gasto é Microempreendedor Individual	Categórica binária
Flag_CPF_CNPJ	Indicador que sinaliza se a pessoa relacionada ao gasto é Pessoa Física ou Pessoa Jurídica	Categórica binária

Algumas dessas métricas foram construídas de maneira indireta, a partir de outras e serão detalhadas na seção referente ao tratamento dos dados.

4. Processamento/Tratamento de Dados

Os dados extraídos dos sites da Câmara dos Deputados e da Receita Federal passaram por diversas transformações a fim de estarem adequados para o processo de modelagem.

A primeira e mais óbvia operação foi a junção dos dados cadastrais do CNPJ e quadro societário com a empresa informada na tabela de despesas. Despesas que possuem esse campo preenchido com CPF ou valores iguais a zero foram devidamente ignorados na junção. A partir dessa junção, pôde-se também criar duas variáveis indicadoras relacionadas ao quadro societário. A primeira indica se o deputado declarou despesas em empresas em que é sócio direto. A segunda variável refere-se a gastos em empresas cujos sócios também mantém sociedade com o deputado em questão em terceiras empresas. Essa é uma espécie de triangulação que poderia indicar algum tipo de irregularidade, dado o relacionamento entre as partes.

Das variáveis que indicam as datas “data de emissão da Nota Fiscal”, “número do mês”, “número do ano” (ambos referentes a data de competência da despesa) e “data de início de operação da empresa” pôde-se gerar variáveis numéricas que indicam o tempo decorrido em relação ao primeiro dia do mês de competência do

gasto. Assim, foram criadas as métricas “idade do gasto” (em número de meses) e “idade da empresa” (em quantidade de dias). Linhas que indiquem mais de 2 anos de “idade do gasto” foram considerados nulos dada a grande probabilidade de serem gerados por erros de digitação. Além disso, dado o ano de competência do gasto, foi criada a variável numérica “sessão legislativa”, com valores de 1 a 4, que indicam o ano do mandato do deputado.

Considerando-se o *dataset* formado pelas variáveis acima, foram necessários certos tratamentos a fim de obter vetores adequados ao treinamento dos modelos. Foram eles:

- Os dados constantes das notas fiscais com valores omissos ou nulos foram excluídos da base. Após a exclusão desses dados, restaram 3.112.508 instâncias para serem utilizadas;
- As variáveis categóricas foram transformadas em numéricas em um mapeamento dicotômico (zeros ou uns), criando-se novas variáveis indicadoras conforme a quantidade de categorias, como em um mapeamento *one hot encoding* ou *dummy*;
- Todas as variáveis, uma vez numéricas, foram transformadas (utilizando seus valores máximos e mínimos) a fim de que estivessem em um intervalo entre zero e um.

Após as operações anteriores, obteve-se uma base de dados com 3.112.508 linhas e 44 colunas a serem utilizadas para treinamento, validação e teste dos modelos.

4.1 A Variável Independente ‘Suspeito’

A fim de avaliar o desempenho dos modelos de detecção e explicação de anomalias, foi feita uma pesquisa a fim de buscar casos relevantes e fora do comum em despesas relacionadas à Cota Parlamentar. Desta busca, foi possível encontrar uma série de reportagens apontando para possíveis abusos ou irregularidades nesses gastos. Foram escolhidas como fontes de informação as seguintes reportagens: **(LUIZ, 2016)**, **(VILANOVA, 2017), (JORNAL DO COMÉRCIO, 2017), (ALMEIDA, 2013), (ESTADÃO CONTEÚDO, 2020), (REDAÇÃO PNOTÍCIAS, 2020), (AGÊNCIA ESTADO, 2016).**

Por meio dessas fontes, foi possível identificar nas bases quais foram os gastos aos quais elas se referiam e criar uma variável independente, denominada ‘suspeito’.

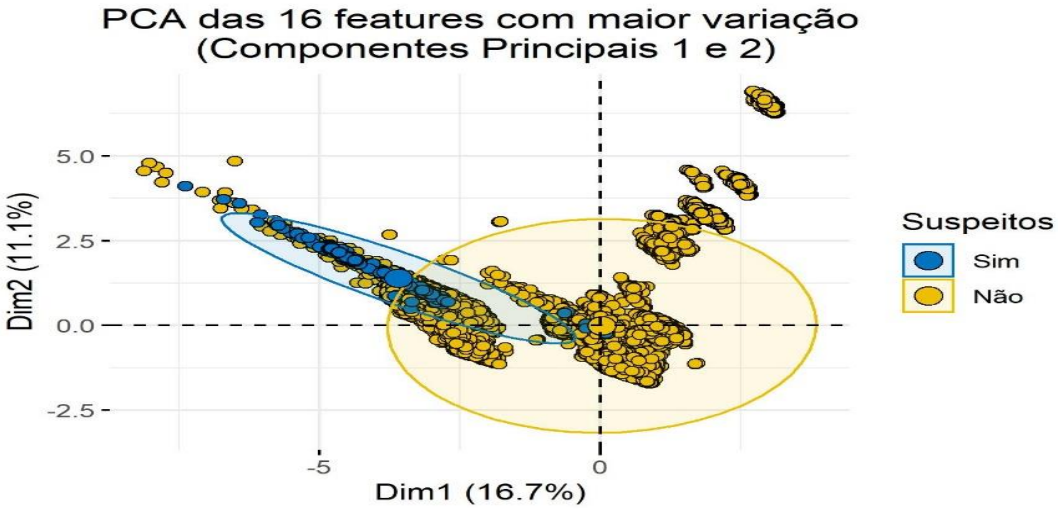
Frisa-se que não é possível saber se esses casos se referem realmente às fraudes ou qualquer ilícito. Não é, além disso, foco deste trabalho fazer alguma investigação ou aferição minuciosa dos gastos. Esses casos, por suas características distintas, foram utilizados para se testar ferramentas para anomalias.

O que faz desses casos interessantes para o presente trabalho é que todas as técnicas estudadas aqui são não supervisionadas. Portanto, um eventual alinhamento entre tais casos e os escores de anomalia vai ao encontro de tratar tais casos como realmente singulares, raros.

Por meio dessas referências, pôde-se identificar 323 instâncias de gastos, que serão denominadas ‘suspeitas’, em comparação com o restante das mais de 3 milhões de instâncias, denominadas, (e assim consideradas, a princípio) ‘comuns’.

5. Análise e Exploração dos Dados

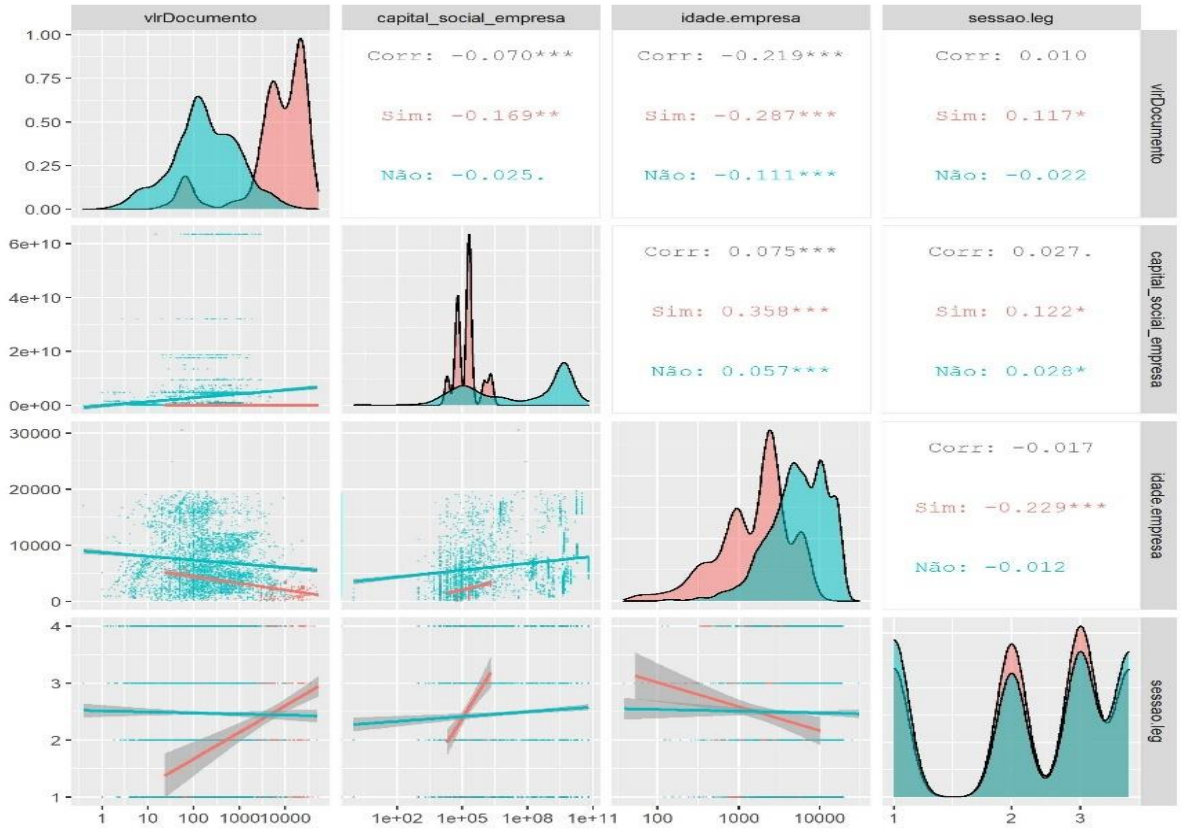
Partindo-se dos dados obtidos na Seção 3, percebe-se que eles são altamente desbalanceados. Como são 44 variáveis, inicia-se a análise exploratória fazendo-se uma breve Análise de Componentes Principais (PCA) nas variáveis que apresentaram alguma variação significativa. Optou-se por desprezar (apenas para a exploração dos dados) variáveis em que um único valor estava presente em mais de 95% dos dados. Após esse corte, para a PCA, restaram 16 variáveis. Em seguida à rotação efetuada pela PCA, gerou-se a visualização das duas primeiras Componentes Principais, que podem ser visualizadas abaixo:



Percebe-se que uma análise simplesmente linear como a PCA não consegue separar os casos suspeitos dos normais. Nesse contexto, as duas primeiras dimensões da PCA representam apenas 28% da variância dos dados. Ademais,

observa-se que esses gastos apresentam, em sua maioria, comportamentos homogêneos e distintos da grande massa amarela de dados comuns a direita.

A partir dos pesos dados a cada variável na Primeira Componente, foram escolhidas as 4 variáveis de maior peso a fim de se aprofundar na visualização da correlação entre elas e suas distribuições de valores. É o que pode ser visto na figura:



Por meio dessa figura, percebe-se que gastos considerados suspeitos têm valores mais elevados que os comuns. Além disso, empresas relacionadas a gastos suspeitos tem menos tempo em operação e menor Capital Social declarado. Ademais, pode-se observar uma discreta concentração desses gastos nas segunda e terceira seção legislativas.

6 Modelos de Detecção e Explicação de Anomalias

Relata-se nessa seção a construção dos modelos *Isolation Forest*, *Autoencoder* e Valores SHAP para a detecção e explicação de anomalias na base em estudo.

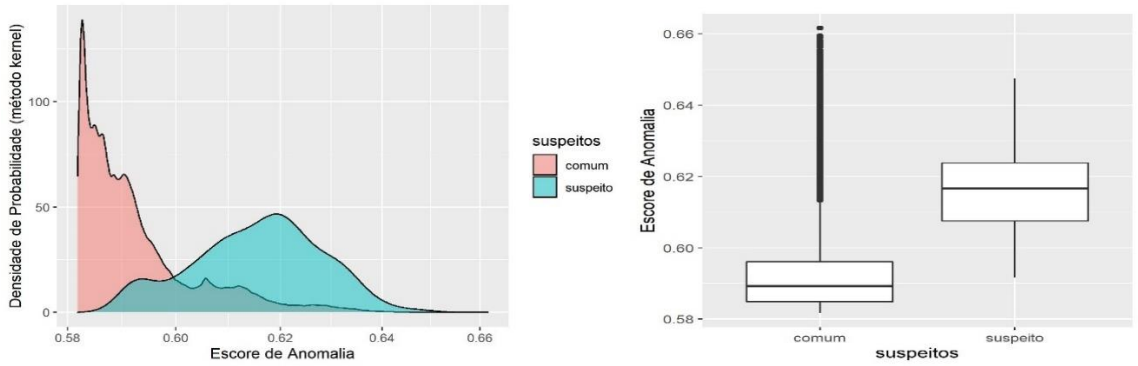
6.1 Criação do Modelo *Isolation Forest*

O algoritmo Isolation Forest, diferentemente da rede Autoencoder não modela o perfil comum dos dados. Com isso, seu treinamento pode incluir todos os dados (comuns e anômalos).

Realizou-se, então, uma amostragem aleatória em toda a base a fim de que os dados sejam oferecidos ao processo de treinamento de forma randômica.

O modelo foi treinado utilizando-se um total de 100 árvores. O parâmetro de sub-amostragem, como definido em (LUNDBERG, 2017) foi escolhido com valor 256, que internamente serve de base para a definição do tamanho de cada árvore.

Após o treinamento, o algoritmo já produziu os escores de anomalia para cada instância baseado na quantidade média de nós necessária para seu isolamento. Passou-se, então, a examinar a distribuição desses escores. De acordo com o algoritmo, instâncias com escore mais elevado têm comportamento anômalo. Essa distribuição pode ser vista nas figuras abaixo:

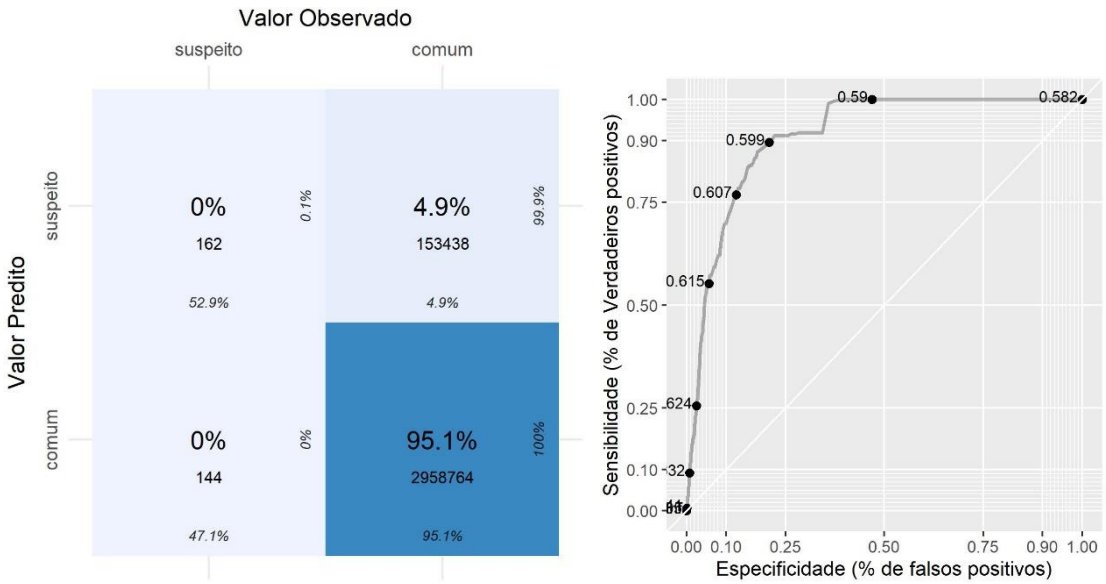


As figuras exibem, respectivamente, a densidade e o gráfico Boxplot dos escores para toda a base, segregados nas categorias comuns e suspeitos.

Percebe-se claramente que os casos suspeitos têm valor esperado de escore de anomalia maior que o dos normais. Ressalta-se que a quantidade de cada uma das categorias é bem distinta. A classe comum é, numericamente, muito maior. Contudo, em termos de quantidade de instâncias, ainda há bastante sobreposição. Ainda assim, é notável que um algoritmo completamente não-supervisionado consiga identificar tais casos suspeitos.

A fim de se calcular a matriz de confusão para uma possível classificação baseada no escore, escolheu-se como critério de corte o percentil 95, que corresponde ao escore de valor 0,616. Consequentemente, instâncias com tal indicador acima de

0,616 são consideradas anomalias. No intuito de melhor interpretar essa matriz, produziu-se também a Curva ROC. Ambas podem ser vistas na figura a seguir:



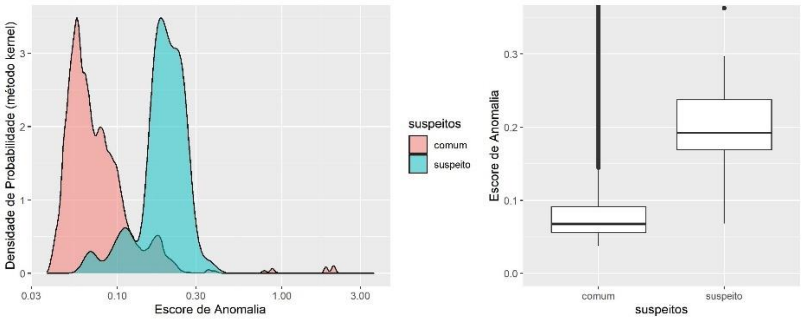
Ao se arbitrar esse valor para a classificação das instâncias em anômalas, conseguiu-se separar apenas 5% da base para uma futura investigação, sabendo-se que mais de 50% dos casos sabidamente suspeitos estão nesse grupo.

6.2 Criação do Modelo *AutoEncoder*

A detecção baseada em Autoencoder objetivou, inicialmente, a modelagem do caso comum. Ou seja, pretendeu-se a indução de um Espaço Latente que represente o comportamento regular dos dados.

Para isso, particionou-se o conjunto de dados em treinamento (60% das instâncias), validação (20%) e teste (20%), tendo-se o cuidado de separar as instâncias sabidamente suspeitas também de acordo com tal proporção.

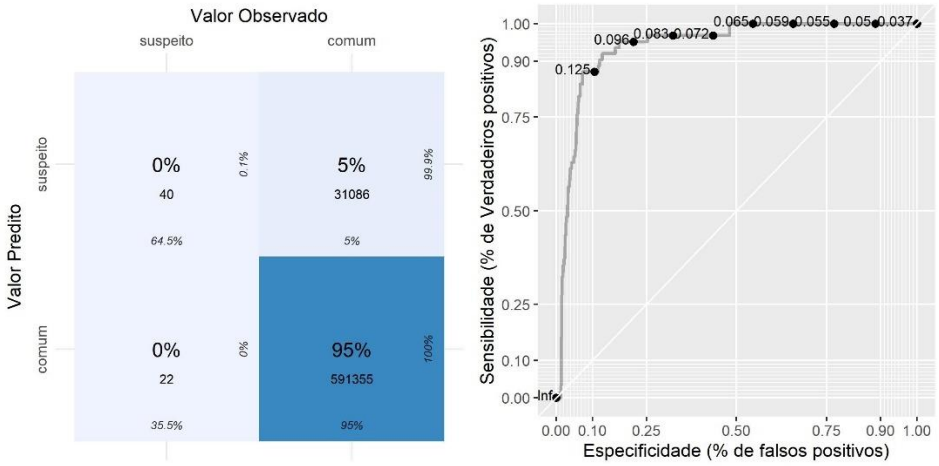
Após o treinamento, construiu-se o escore de anomalia obtendo-se o somatório ao longo de todas as variáveis do valor absoluto da diferença entre vetor de entrada e saída. Passou-se, então, a examinar a distribuição desses escores. De acordo com essa construção, instâncias com escore mais elevado têm comportamento anômalo. Essa distribuição pode ser vista nas figuras abaixo:



As figuras exibem a densidade e o gráfico Boxplot dos escores, segregados nas categorias: comuns e suspeitos.

Como aconteceu na modelagem por *Isolation Forest,* casos suspeitos tiveram valor esperado de escore de anomalia maior que o dos normais. A classe comum ainda foi, numericamente, muito maior que a anômala. Todavia, em termos de quantidade de instâncias, ainda houve bastante sobreposição. Ademais, percebe-se que as anomalias identificadas pela rede guardaram grande correlação com os casos suspeitos.

A fim de se calcular a matriz de confusão para uma possível classificação baseada no escore, escolheu-se como critério de corte o percentil 95 dos escores, de valor 0,175. Instâncias com valor acima de 0,175 foram consideradas anômalas. Produziu-se também a Curva ROC a fim de melhor entender as consequências da escolha do ponto de corte. Ambas podem ser vistas na figura a seguir:



Percebe-se, pela matriz de confusão, que o valor escolhido para o ponto de corte, embora tenha separado apenas 5% da base como possíveis anomalias, conseguiu identificar corretamente mais de 64% das anomalias conhecidas.

6.3 Criação do Modelo *de Valores SHAP*

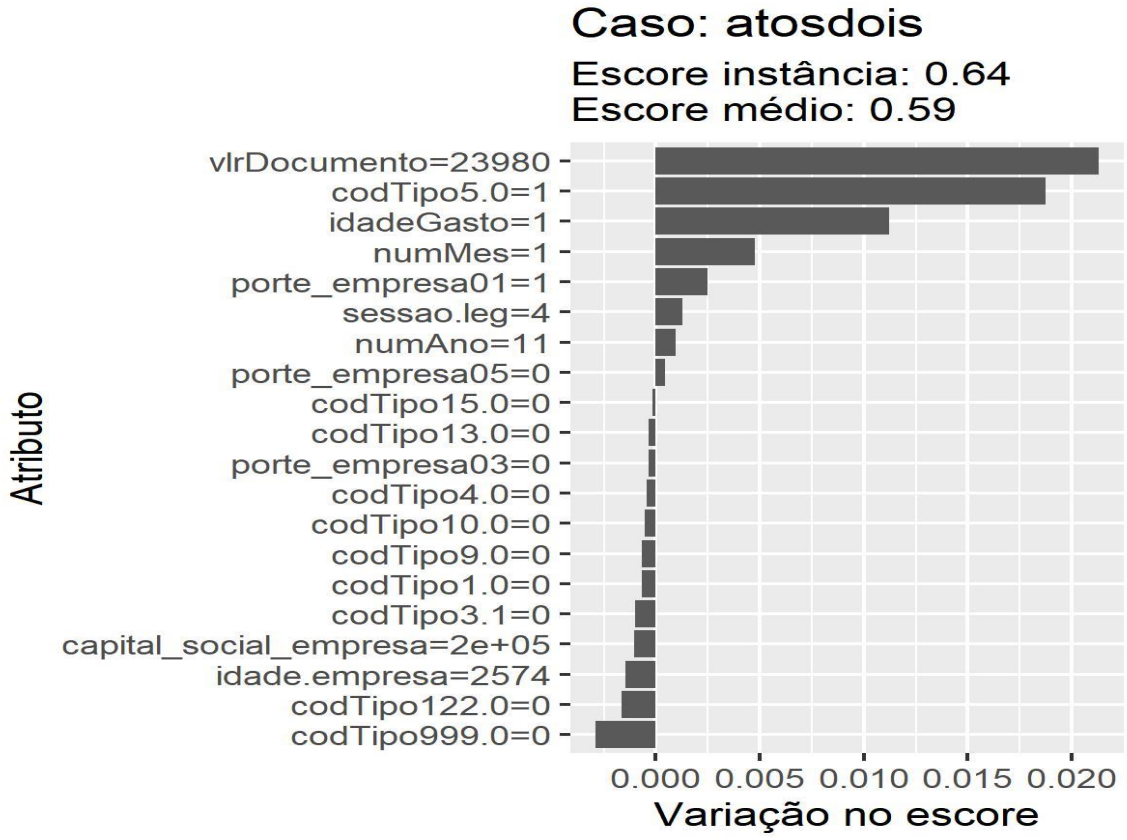
O modelo de explicação baseado em Valores SHAP tem como ponto de partida os resultados dos algoritmos *Isolation Forest* e *AutoEncoder*.

Criou-se, para tanto, uma estrutura de previsão local que inclui o modelo gerado (que deve ser capaz de gerar os escores a partir de um vetor de entrada) e a base de dados utilizada para a explicação. Nesse caso, utilizou-se toda a base (e não apenas a de teste) a fim de melhor explicar as instâncias classificadas.

A partir dessa estrutura de previsão e de uma determinada instância, o modelo de valores SHAP obtém a influência local de cada variável como explicação do escore de anomalia gerado.

Esse processo pôde ser melhor visualizado com um exemplo de instância corretamente classificada como anomalia: a de gasto suspeito mencionado em (ESTADAO CONTEUDO, 2020) com a empresa Atos Dois Propaganda e Publicidade Ltda.

Essa instância, em modelagem via Autoencoder possuiu escore igual a *0,27*, maior que a média da base (*0,09*). Além disso, foi maior que o valor limite (0,175) e, por isso, considerada anomalia. Esse gasto também foi considerado anomalia pelo modelo *Isolation Forest* com um escore de *0,64* em um limite é *0,616*. O gráfico a seguir exibe este resultado:



Pode-se perceber que algumas variáveis contribuem de maneira mais significante para que tal gasto seja anômalo. O principal é próprio valor do gasto, de R$23.980,00, consideravelmente mais elevado que a média. O tipo de gasto igual a 5.0 é código que indica ‘divulgação de atividade parlamentar’ e também tem um elevado peso na formação do escore de anomalia. A idade do gasto igual a unidade aponta para gasto efetuado em um intervalo de um mês da prestação de contas, e não no mesmo mês da prestação.

Outras variáveis que intuitivamente sinalizariam gastos suspeitos foram mostradas também no gráfico como: o gasto ter sido efetuado em janeiro (mês igual a unidade), o pequeno porte da empresa e o último ano do mandato (seção legislativa igual a 4).

Além disso, observou-se que há fatores que contribuíram negativamente como o capital social da empresa e sua idade, ambos valores comuns. Se olhados de maneira isolada, esses dois últimos fatores apontariam para um baixo escore de anomalia e, consequentemente, pouca probabilidade de ser um gasto suspeito.

Há, porém, uma soma de fatores que pesaram muito mais para que o gasto seja considerado anômalo.

7. Conclusão

Pôde-se estudar, neste trabalho, duas técnicas bastante eficientes para se identificar anomalias complexas em grandes bases de dados, a saber: *Isolation Forest* e redes *Autoencoder*.

Dado que são técnicas completamente não supervisionadas e que os casos rotulados foram utilizados apenas em ambiente de teste, pareceu promissora sua utilização em um ambiente preliminar de análise de fraudes.

É notável a capacidade de tais algoritmos em apontar anomalias que guardaram bastante correlação com casos suspeitos em gastos da Cota para Exercício da Atividade Parlamentar.

Por fim, utilizou-se a técnica de Valores SHAP para explicar os escores de anomalia gerado por tais modelos. Essa técnica trouxe uma clareza e intuição ao resultado do escore e possibilitou uma melhor interpretação desse escore como indício de fraude.

REFERÊNCIAS

CÂMARA DOS DEPUTADOS. **Assessoria de Imprensa – Cota Parlamentar**, 2020a. Disponível em: <https://www2.camara.leg.br/comunicacao/assessoria-de-imprensa/guia-para-jornalistas/cota-parlamentar> Acesso em 29/9/2020

CÂMARA DOS DEPUTADOS. **Dados Abertos – Despesas pela Cota para Exercício da Atividade Parlamentar**, 2020b. Disponível em: <https://dadosabertos.camara.leg.br/swagger/api.html#staticfile> Acesso em 28/9/2020

CÂMARA DOS DEPUTADOS. **Cota para Exercício da Atividade Parlamentar – Explicações sobre o formato dos arquivos**, 2020c. Disponível em: <https://www2.camara.leg.br/transparencia/cota-para-exercicio-da-atividade-parlamentar/explicacoes-sobre-o-formato-dos-arquivos-xml > Acesso em: 28/9/2020

SUARA. RECEITA FEDERAL - **Dados públicos CNPJ**, 2020. Disponível em: <https://receita.economia.gov.br/orientacao/tributaria/cadastros/cadastro-nacional-de-pessoas-juridicas-cnpj/dados-publicos-cnpj> Acesso em: 29/9/2020

CORREIO BRAZILIENSE. **Fraudes no auxílio emergencial já dão prejuízo de mais de R$ 60 milhões**, 2020. Disponível em: <https://www.correiobraziliense.com.br/app/noticia/economia/2020/06/27/internas_economia,867374/fraudes-no-auxilio-emergencial-ja-dao-prejuizo-de-mais-de-r-60-milhoe.shtml> Acesso em: 6/10/2020

CNDL BRASIL. **Em cada dez brasileiros, dois foram vítimas de fraudes nos últimos 12 meses, mostra levantamento CNDL/SPC Brasil**, 2019. Disponível em: <https://site.cndl.org.br/em-cada-dez-brasileiros-dois-foram-vitimas-de-fraudes-nos-ultimos-12-meses-mostra-levantamento-cndlspc-brasil/> Acesso em: 6/10/2020

THE NILSON REPORT. **Card Fraud Losses Reach $27.85 Billion**, 2019. Disponível em: < https://nilsonreport.com/mention/407/1link/> Acesso em: 6/10/2020

NOGUEIRA Italo - FOLHA DE SÃO PAULO. **Jovens criam robô que monitora despesas de deputados federais**, 2017. Disponível em: < https://www1.folha.uol.com.br/poder/2017/01/1852180-jovens-criam-robo-que-monitora-despesas-de-deputados-federais.shtml> Acesso em: 6/10/2020

MISRA Sumit, THAKUR Soumyadeep, GHOSH Manosij, SAHA Sanjoy Kumar. **An Autoencoder Based Model for Detecting Fraudulent Credit CardTransaction.** International Conference on Computational Intelligence and Data Science (ICCIDS 2019), 2019.

AGGARWAL Charu C. **Outlier analysis, Second Edition**. pg 102–105. Springer, 2017

LIU Fei Tony, TING Kai Ming, ZHOU Zhi-Hua. **Isolation Forest.** Eighth IEEE Internation Conference on Data Mining, 2008.

LUNDBERG Scott M., LEE, Su-In. **A Unified Approach to Interpreting Model Predictions.** Long Beach: 31 Conference on Neural Information Processing Systems (NIPS 2017), 2017.

RIBEIRO Marco Tulio, SINGH Sameer, GUESTRIN Carlos. **"Why Should I Trust You?" Explaining the Predictions of Any Classifier.** San Francisco: KDD 2016, 2016.

KRAMER Mark A. **Nonlinear Principal Component Analysis using autoassociative Networks**. AIChE Journal 37 (2): 233-243, 1991.

GOODFELLOW Ian, BENGIO Yoshua, COURVILLE Aaron. **DeepLearning**. MIT Press, 2016

ANTWARG Liat, MILLER Ronnie Mindlin, SHAPIRA Bracha, ROKACH Lior – Deep AI. **Explaining Anomalies Detected by Autoencoders Using SHAP**, 2019. Disponível

em: <https://deepai.org/publication/explaining-anomalies-detected-by-autoencoders-using-shap>, Acesso em 6/10/2020.
LUIZ, Gabriel. **Após ser flagrado por app, deputado devolve à Câmara R$ 727 por 13 refeições no mesmo dia.** 6/12/2016. Disponível em: <https://g1.globo.com/distrito-federal/noticia/apos-ser-flagrado-por-app-deputado-devolve-a-camara-r-727-por-13-refeicoes-no-mesmo-dia.ghtml>, Acesso em 9/10/2020
VILANOVA, Pedro. **O que a resposta do Dep. Marcon à Rosie tem a nos dizer sobre o trabalho da Serenata de Amor**. 9/5/2017. Disponível em: <https://medium.com/data-science-brigade/o-que-a-resposta-do-dep-marcon-%C3%A0-rosie-t%C3%AAm-a-nos-dizer-sobre-o-trabalho-da-serenata-de-amor-c7f898a4655f>, Acesso em 6/10/2020
JORNAL DO COMÉRCIO. **Parlamentares federais gaúchos esclarecem despesas fiscalizadas por levantamento eletrônico**. 17/5/2017. Disponível em: <https://www.jornaldocomercio.com/_conteudo/2017/05/politica/562699-parlamentares-federais-gauchos-esclarecem-despesas-fiscalizadas-por-levantamento-eletronico.html>
ALMEIDA, Amanda. **Deputados utilizam verba de custeio para pagar gastos de campanha**. 13/1/2013. Disponível em: <https://www.em.com.br/app/noticia/politica/2013/01/13/interna_politica,343003/deputados-utilizam-verba-de-custeio-para-pagar-gastos-de-campanha.shtml>, Acesso em: 9/10/2020
ESTADÃO CONTEÚDO, **Rosa Weber autoriza inquérito contra utilização irregular de cota parlamentar**. 1/9/2020. Disponível em: <https://istoe.com.br/rosa-weber-autoriza-inquerito-contra-utilizacao-irregular-de-cota-parlamentar/>, Acesso em: 9/10/2020
REDAÇÃO PNOTÍCIAS, **Deputados usam cota parlamentar para alugar carros de empresas investigadas por fraude**. 29/7/2020 Disponível em: <https://pnoticias.com.br/noticia/politica/238862-deputados-usam-cota-parlamentar-para-alugar-carros-de-empresas-investigadas-por-fraude>, Acesso em 9/10/2020
AGÊNCIA ESTADO, **André Moura e outros 29 deputados são investigados por fraudes**. 23/10/2016. Disponível em: <https://www.otempo.com.br/politica/andre-moura-e-outros-29-deputados-sao-investigados-por-fraudes-1.1389605>, Acesso em 9/10/2020