Predição de Irregularidade Fiscal dos Contribuintes do Tributo ISS

Glauco de Vasconcelos Soares[1], Rodrigo C. L. V. Cunha[1]

[1]Centro de Estudos e Sistemas Avançados do Recife (CESAR SCHOOL)
AV. Cais do Apolo, 77 – Recife, PE – Brasil

{gvs2,rclvc,femf}@cesar.school

Abstract. *The Service Tax (ISS) is a Brazilian municipal tax that is levied in cases where a service provision occurs. A common type of irregularity is to correctly declare the collection of services that have been provided, without, however, passing them on to the municipalities. In this context, this work aims to assess the use of Machine Learning techniques in order to obtain accurate forecasts that estimate the risks of companies becoming debtors in the next fiscal year. The models were developed and tested on a data set with approximately 60 thousand records, containing data from tax returns grouped by quarter for a period of 5 years. Preliminary results indicate that our model identifies this type of irregular behavior with 85.21% accuracy.*

Resumo. *O Imposto Sobre Serviço (ISS) é um tributo brasileiro municipal cuja incidência se dá nos casos em que ocorre uma prestação de serviço. Um tipo de irregularidade comum é declarar corretamente o recolhimento dos serviços que foram prestados, sem, entretanto, repassá-los aos municípios. Neste contexto, esse trabalho tem como objetivo avaliar o uso técnicas de Aprendizado de Máquina visando obter previsões precisas que estimem os riscos das empresas adimplentes ou inadimplentes se tornarem devedoras contumazes no próximo exercício fiscal. Os modelos foram desenvolvidos e testados sobre um conjunto de dados com aproximadamente 60 mil registros, contendo dados de declarações fiscais agrupadas por trimestre por um período de 5 anos. Os resultados preliminares indicam que nosso modelo identifica esse tipo de comportamento irregular com 85.21% de acurácia.*

1.Introdução

Sonegação ou evasão fiscal é um fenômeno global que afeta a sociedade como um todo. A legislação brasileira considera crime de sonegação o ato de prestar declarações falsas ou omitir, totalmente ou parcialmente, informações que deva ser produzida ao Fisco, com o intuito de não pagar ou pagar menos impostos que o realmente devido[1].

Com o objetivo de evitar a fiscalização, algumas empresas adotam a prática de declarar corretamente o recolhimento do imposto, sem, entretanto, repassá-los ao Estado. De fato, este capital sem destinação fiscal aumenta arbitrariamente o lucro das empresas, por meio de apropriação de valores que deveriam ter sido repassados ao

[1] http://www.planalto.gov.br/ccivil_03/leis/1950-1969/l4729.htm (acessado em 15/06/2020)

erário [Godoy and Basso 2015]. Em termos gerais, o inadimplemento contumaz pode ser entendido como a ação de deixar de recolher impostos de forma rotineira, sistemática e de caráter criminal[2], com o objetivo de obter lucro, prejudicando os cofres públicos, a concorrência e toda a sociedade. O tipo de irregularidade fiscal abordado neste trabalho é o inadimplemento contumaz.

Para lidar com esse tipo de problema, várias secretarias de finanças brasileiras estão desenvolvendo sistemas que auxiliam e aprimoram os processos de tomada de decisão [Matos et at. 2019]. Entre as várias decisões que uma administração tributária deve lidar, uma das mais importantes é, sem dúvida, saber quais contribuintes[3] inspecionar ou saber quem precisa de maior controle tributário. Neste sentido, o uso de Aprendizado de Máquina (AM) tem tido um significado prático nestas organizações, fornecendo um conjunto de ferramentas ideais para analisar os dados, aprender o comportamento, e fazer previsões a partir de dados históricos das bases de dados do governo.

Em virtude dos fatos que foram mencionados, este trabalho tem como objetivo propor um modelo de predição que calcule o risco do inadimplemento contumaz para os contribuintes do tributo Imposto Sobre Serviço (ISS).

O artigo está estruturado da seguinte forma. A Seção 2 apresenta os trabalhos relacionados ao tema deste artigo. Nas Seções 3 é apresentado o entendimento dos dados utilizados no modelo, bem como o pré-processamento dos mesmos. A Seção 4 explora os resultados dos classificadores *Logistic Regression* (RL), *Random Forest* (RF) e *LightGBM (LGBM)*. Por fim, a Sessão 5 apresenta as considerações finais do artigo.

2. Trabalhos Relacionados

A construção de modelos de prevenção e detecção automática de fraudes tem sido escolhida no mundo acadêmico como uma linha útil de pesquisa, recebendo interesse cada vez maior no contexto de evasão fiscal [Soares et al. 2020]. Na literatura diversos trabalhos têm sido propostos utilizando AM supervisionado através de dados históricos de auditoria, contudo, os dados de fiscalização em geral são bastante limitados, o que acaba refletindo na generalização e qualidade dos modelos [Vanhoeyveld et al. 2020]. No contexto de AM não supervisionado, os trabalhos [de Roux et al. 2018, Vanhoeyveld et al. 2020] propuseram modelos de detecção de fraude em declarações fiscais subfaturadas, ou seja, os dados das declarações são analisados em grupos com o objetivo de encontrar indícios de sonegação fiscal.

Além disso, o uso de AM também tem sido utilizado para combater a inadimplência, de modo a serem asseguradas a regularidade e a previsibilidade da arrecadação [Jupri and Sarno 2018]. No cenário brasileiro, [Dias and Becker 2017] apresentaram um modelo de classificação de fraude em notas fiscais eletrônicas de

[2] https://portal.stf.jus.br/noticias/verNoticiaDetalhe.asp?idConteudo=433114&ori=1 (acessado em 15/06/2020)

[3] Contribuinte é o sujeito passivo de uma obrigação tributária. Em outros termos, é aquele que se sujeita, por previsão legal, ao pagamento de tributos ao fisco.

serviço para o segmento da construção civil. A proposta é bastante promissora, contudo os dados rotulados como suspeitos também são bastante limitados.

Diferentemente dos estudos mencionados, estre trabalho utiliza técnicas de AM supervisionado para calcular o risco de um contribuinte adimplente ou inadimplente vir a se tornar um devedor contumaz.

3. Análise e Preparação dos Dados

3.1. Entendimento dos Dados

Os dados utilizados para treinamento dos modelos foram extraídos de um *Data Warehouse* (DW) de uma secretaria municipal de finanças situada em uma capital brasileira, que será identificada apenas como Secretaria Alfa. A mesma secretaria autorizou a divulgação das análises descritivas desde que fosse respeitado todo o sigilo fiscal.

Foram analisados dados de diversas declarações e escrituração fiscais de ISS, abrangendo o período de janeiro de 2015 a dezembro de 2019. Tanto a seleção dos dados para treinamento, como a escolha das variáveis foram realizadas por 2 especialistas, sendo um o administrador de dados da instituição, e o outro um auditor fiscal de carreira.

A variável alvo ou de resposta também estava disponível no DW, esta classificação é realizada analisando o comportamento da inadimplência em um determinado período, observando os critérios estabelecidos na legislação municipal.

- **Devedor contumaz** = 1, para todas as empresas sediadas no município que deixaram de recolher crédito tributário do ISS por três ou mais competências, consecutivas ou não, confessados por meio declarações fiscais.
- **Devedor contumaz** = 0, caso contrário, ou seja, apresenta-se adimplente com suas obrigações tributárias, pagando seus tributos conforme suas declarações, podendo ainda ter no máximo dois débitos em aberto.

Após a definição da classe alvo, a base de dados foi montada contendo 58.436 registros com dados fiscais de 20.737 contribuintes, sendo 5.796 devedores contumazes e 14.941 contribuintes regulares. É importante salientar que a granularidade da informação é baseada no contribuinte e que a data de corte utilizada é trimestral. A Tabela 1 apresenta as variáveis finais utilizadas para treinamento dos modelos.

ID	Descrição	ID	Nome
V1	Identificador da empresa	V10	Quantidade de débitos em aberto anterior ao trimestre
V2	Data de referência ou corte	V11	Proxy referente a atividade principal da empresa
V3	Valor do débito no trimestre	V12	Proxy referente ao bairro onde está localizada a empresa
V4	Valor do débito anterior ao trimestre	V13	Proxy referente ao segmento econômico de atuação da empresa
V5	Quantidade de escriturações abertas	V14	Idade da empresa em dia

	no trimestre		
V6	Quantidade de escriturações abertas anterior ao trimestre	V15	Indica se a empresa é optante do Simples Nacional
V7	Quantidade de notificações de débito no trimestre	V16	Valor faturado no trimestre
V8	Quantidade de notificações de débito anterior ao trimestre	V17	Valor faturado anterior ao trimestre
V9	Quantidade de débitos em aberto no trimestre	V18	Variável alvo ou de resposta, indica se o contribuinte está caracterizado como devedor contumaz um ano após a data de corte

Tabela 1. Relação das variáveis após pré-processamento

3.2. Pré-processamento

Os dados extraídos do DW tiveram sua atividade de preparação de dados facilitada, uma vez que a limpeza, padronização e tratamento de *missings* (dados faltantes) já haviam sido realizadas. No entanto, para que os dados possam ser utilizados pelos classificadores eles devem sofrer algumas transformações. Para isso, utilizou-se diferentes tipos de técnicas de pré-processamento, estas foram: codificação de variável *proxy*, tratamento de *outliers* (pontos aberrantes) e normalização dos dados.

A técnica de codificação de variável proxy consiste em transformar uma variável que identifica alguma coisa no mundo real, como por exemplo um bairro, em uma ou mais variáveis que sejam análogas a coluna de origem, mas que o seu conteúdo armazenado seja um valor contínuo, no caso do bairro (V12) criamos uma variável com o somatório dos valores das edificações deste bairro.

A normalização dos dados é um processo que ajusta valores medidos em diferentes escalas para uma única escala, isso reduz a distância entre os valores e, no geral, tende a aumentar a precisão dos classificadores. Tanto a normalização quanto o tratamento dos *outliers* foram contornados com o uso de quartis, utilizando a função de pré-processamento *QuantileTransformer*[4]. Ao final, os valores das variáveis foram transformados para uma distribuição normal com valores entre 0 e 1.

4. Modelagem e Avaliação

Os modelos utilizados neste trabalho foram construídos usando a linguagem de programação *Python*[5] juntamente com o projeto *scikit-learn*[6]. Este projeto fornece a implementação dos principais algoritmos de aprendizado de máquina.

Para estimar os modelos de predição, a base de dados foi separada da seguinte maneira: 70% dos dados foram utilizados na etapa de treinamento e 30% na etapa teste. Assim, foram gerados 3 modelos distintos de classificação, *Logistic Regression*, *Random Forest* e *LightGBM*, utilizando o mesmo conjunto de dados e comparando seus

[4] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.QuantileTransformer.html

[5] https://www.python.org/

[6] https://scikit-learn.org/

resultados com as seguintes métricas: percentual de classificações corretas (Acurácia), área sob a curva ROC (AUC[7]), teste estatístico de *Kolmogorov-Smirnov* (KS2[8]) e matriz de confusão. O *Logistic Regression* e o *LigthGBM* foram inicializados com todos os parâmetros *default*, já o *Random Forest* foi instanciado com 200 árvores.

	LR	RF	LGBM
Total de acertos	13.979	14.938	14.351
Total de erros	3.552	2.593	3.180

Tabela 2. Comparação das predições

	LR %	RF %	LGBM %
Acurácia	79.74	85.21	81.86
AUC	76.64	90.18	84.36
KS2	48.30	68.20	53.90

Tabela 3. Comparação das métricas de avaliação

O resultado das matrizes de confusão apresentados na Tabela 2 mostram a quantidade de erros e acertos das predições sobre o conjunto independente de teste. Como podemos observar na tabela, o classificador *Random Forest* obteve um melhor desempenho que os demais classificadores. Além disso, conforme apresentado na Tabela 3, o mesmo classificador também alcançou melhores resultados nas métricas de acurácia, AUC e KS2, enquanto que, o algoritmo *Logistic Regression* teve o pior desempenho quando comparado com os outros.

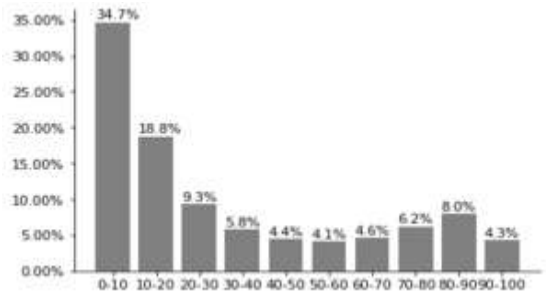

Figura 1. Frequência das empresas por faixa de escore

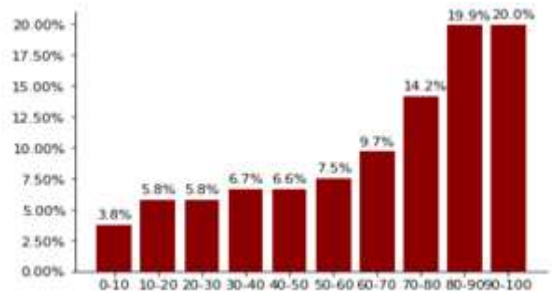

Figura 2. Devedores contumazes por faixa de escore

Após estimar os modelos, foi feita uma análise dos escores das predições, pois para que o sistema possa ser utilizado em produção é necessário definir um limiar (ponto de corte). O objetivo é separar as empresas irregulares das regulares, dando um maior poder de análise para os auditores. Pela Figura 1, nota-se que existe uma pequena quantidade de empresas nas últimas faixas de escore, entre 70-100, que correspondem às empresas de maior risco. Já na Figura 2 pode-se observar que quanto maior o escore, maior o risco da empresa ser contumaz. Isso reforça que o modelo ajudará os auditores na prática a monitorar as empresas de maior risco. Em conversa com os auditores ficou definido o ponto de corte de 70 (escores a partir de 70) para serem as empresas

[7] A métrica AUC permite avaliar um modelo comparando a proporção de falsos positivos à medida em que é aumentada a taxa de verdadeiros positivos.

[8] O KS2 é um teste não paramétrico sobre a igualdade de distribuições de probabilidade contínuas, ele mede a distância máxima entre duas distribuições e pondera os erros pela probabilidade de ocorrência de cada classe.

monitoradas. A escolha do ponto de corte foi uma relação entre a capacidade de investigação dos auditores e o risco estimado.

Espera-se que o modelo construído venha auxiliar o planejamento e a pauta fiscal durante o ano, sugerindo quais contribuintes tem a maior probabilidade de torna-se um devedor contumaz no próximo exercício.

5. Conclusão

Este trabalho apresentou uma aplicação real de AM no contexto de irregularidade fiscal, apresentando resultados estatisticamente expressivo quanto ao desempenho das predições. A abordagem proposta possibilita uma melhoria no processo de tomada de decisão dos auditores fiscais, uma vez que sugere quais contribuintes são mais propensos a cometer o inadimplemento contumaz no próximo exercício. Adicionalmente, os resultados obtidos pelo modelo também foram considerados relevantes por um auditor especialista da secretaria Alfa. Além disso, também foi possível constatar que para este problema de predição o classificador *Random Forest* obteve melhores resultados que o *Logistic Regression* e *LightGBM*. Os próximos passos consistirão na aplicação de outros algoritmos de AM, além do ajuste fino dos modelos através das configurações dos hiperparâmetros para melhorar a performance dos classificadores.

Referências

de Roux, D.; Pérez, B.; Moreno, A.; Villamil, P. and Figueroa, C. (2018). "Tax Fraud Detection for Under-Reporting Declarations Using an Unsupervised Machine Learning Approach". In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18), Londor, UK, pp. 2015-222.

Dias, M. and Becker, K. (2017). "Identificação de Candidatos à Fiscalização por Evasão do Tributo ISS". In: 5th Symposium on Knowledge Discovery, Mining and Learning, Uberlândia, MG, 2017.

Godoy, L. S., Basso, J. P. (2015) "Sonegação e inadimplência contumaz: prejuízo à concorrência empresarial". Revista Digital ESAPERGS, p. 04 -10.

Jupri, M. and Sarno R. (2018) "Taxpayer compliance classification using C4.5, SVM, KNN, Naive Bayes and MLP". 2018 International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, pp. 297-303.

Matos, T., Macedo, J. A., Lettich, F., Monteiro, J. M., Renso, C., Perego and R., Nardini, F. M. (2019). "Leveraging Feature Selection To Detect Potential Tax Fraudsters". Expert Systems with Applications. vol. 145, 113128.

Soares, G. V., Cunha, R. C. L. V., Filho, F. E. M. (2020) "O Uso de Inteligência Artificial no Combate à Evasão Fiscal: Uma Revisão Sistemática da Literatura". In: 8th Workshop de Computação Aplicada em Governo Eletrônico (WCGE), Cuiabá, MT, 2020. No Prelo

Vanhoeyveld, J., Martens, D., Peeters, B. (2020). "Value-added tax fraud detection with scalable anomaly detection techniques". Applied Soft Computing, vol. 86, 105895.