

Tópicos em Engenharia de Software

ANÁLISE PREDITIVA APLICADA À IDENTIFICAÇÃO DE FRAUDES EM SEGUROS AUTOMOTIVOS

Gabriel Estevão, Leticia Lott, Luana Fleury, Rafael Ferraz e Yan Nalon

CONTEXTUALIZAÇÃO

INDÚSTRIA DE SEGUROS

- Mercado global de seguros cresceu **9,5% em 2024** nos serviços de proteção de propriedade pessoal e acidentes [1]
- No entanto, as **fraudes de pedidos de seguro** são um desafio intrínseco a essa área

Principais tipos de fraudes em seguros automotivos:

- Acidentes simulados
- Exagero de danos ou lesões
- Inclusão de “passageiros fantasma”



Impacto das fraudes:

- Altos custos operacionais e financeiros
- Sobrecarga nas investigações de sinistros
- Necessidade de métodos automatizados e escaláveis para mitigar prejuízos

PROBLEMA DE PESQUISA

A literatura carece de experimentos comparativos em Machine Learning específicos para previsão de fraudes automotivas

MOTIVAÇÃO

- Métodos tradicionais de auditoria e análise manual são ineficientes para grandes volumes de dados
- Uso de Machine Learning representa uma oportunidade nesse contexto



POTENCIAL DO APRENDIZADO DE MÁQUINA NO COMBATE À FRAUDE

- Machine Learning é uma ferramenta eficiente para combate à fraude, analisando grandes volumes de dados e identificando anomalias.
- Técnicas de aprendizado supervisionado permitem detecção com alta precisão.
- Alarfaj et al. (2022) utilizaram um algoritmo de aprendizado de máquina para detectar fraudes em cartões de crédito, alcançando resultados com alta acurácia, chegando a até 99,9%.



OBJETIVOS

OBJETIVO GERAL

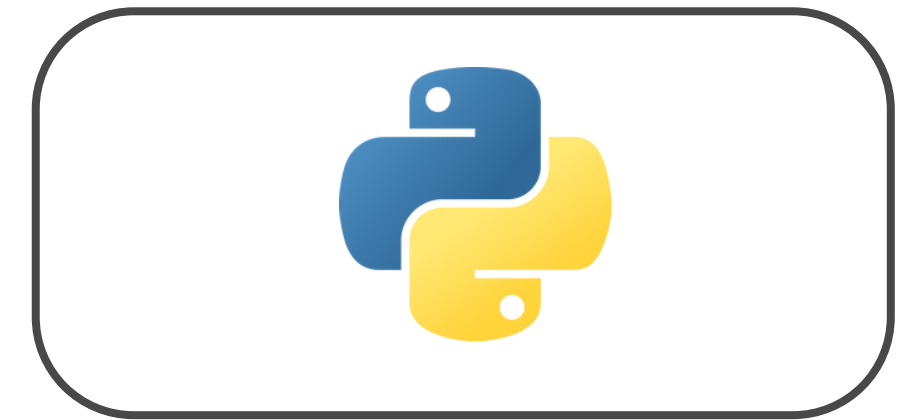
Avaliar o desempenho de técnicas de aprendizado de máquina na detecção de possíveis fraudes em seguros automotivos

OBJETIVOS ESPECÍFICOS

1. Investigar a performance de cada modelo (Regressão Logística, Naive Bayes, Floresta de Decisão e SVM) na identificação de fraude em seguros automotivos
2. Comparar métricas (acurácia, precisão, revocação e F1-Score) entre dados desbalanceados e balanceados (SMOTE)
3. Avaliar a consistência dos modelos diante de diferentes distribuições de dados, identificando a técnica mais adequada para aplicação prática.

PROJETO DA METODOLOGIA

INSTRUMENTOS



Bibliotecas: Pandas, Scikit-Learn, ImbLearn e Matplotlib

Justificativa: Adequação à manipulação de dados e implementação de algoritmos de inteligência artificial + uso em pesquisas na área de aprendizado de máquina.

PROJETO DA METODOLOGIA

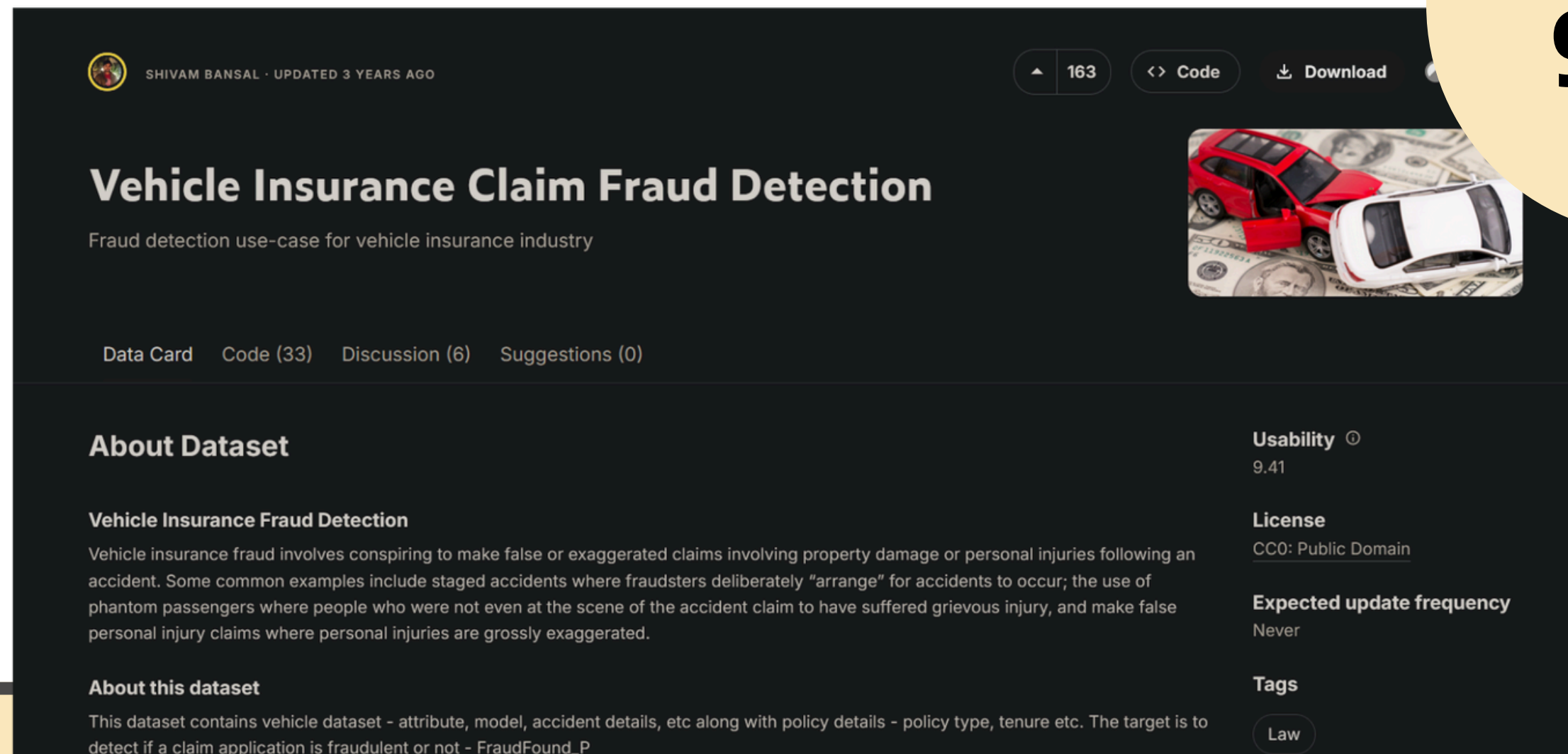
DATASET

Kaggle - *Vehicle Insurance Claim Fraud Detection*

- 15.420 registros, 33 atributos (colunas).
- Target: **FraudFound_P** (0 = não fraude, 1 = fraude)

Usabilidade

9.41



The screenshot shows the Kaggle dataset page for 'Vehicle Insurance Claim Fraud Detection' by Shivam Bansal. The page features a dark theme with a header bar containing the dataset title, author name, and update date. Below the header, there's a section for 'Vehicle Insurance Claim Fraud Detection' with a subtitle 'Fraud detection use-case for vehicle insurance industry'. A small image of a red car on a pile of money is visible. The page includes tabs for 'Data Card', 'Code (33)', 'Discussion (6)', and 'Suggestions (0)'. The 'About Dataset' section describes the dataset's purpose and content. On the right, there's a 'Usability' score of 9.41, a 'License' of CC0: Public Domain, and an 'Expected update frequency' of 'Never'. A 'Tags' section at the bottom shows the tag 'Law'.

SHIVAM BANSAL · UPDATED 3 YEARS AGO

163 Code Download

Vehicle Insurance Claim Fraud Detection

Fraud detection use-case for vehicle insurance industry

Data Card Code (33) Discussion (6) Suggestions (0)

About Dataset

Vehicle Insurance Fraud Detection

Vehicle insurance fraud involves conspiring to make false or exaggerated claims involving property damage or personal injuries following an accident. Some common examples include staged accidents where fraudsters deliberately "arrange" for accidents to occur; the use of phantom passengers where people who were not even at the scene of the accident claim to have suffered grievous injury, and make false personal injury claims where personal injuries are grossly exaggerated.

About this dataset

This dataset contains vehicle dataset - attribute, model, accident details, etc along with policy details - policy type, tenure etc. The target is to detect if a claim application is fraudulent or not - FraudFound_P

Usability 9.41

License CC0: Public Domain

Expected update frequency Never

Tags Law

SOLUÇÃO ADOTADA

PRÉ-PROCESSAMENTO

1. Limpeza de dados:

- Remoção de colunas irrelevantes
- Colunas que não agregam valor ao modelo, podendo prejudicar a capacidade de generalização dos modelos, tornando-os mais lentos e imprecisos.

2. Codificação de variáveis:

- Binária (LabelEncoder): aplicada em colunas com dois valores distintos (ex.: *Sex*, *AccidentArea*), exceto o alvo *FraudFround_P*.
 - Mapeia cada categoria para 0 ou 1.
 - Facilita a interpretação dos algoritmos, que conseguem lidar diretamente com variáveis binárias.
- One-Hot Encoding: aplicada em categorias sem ordem (ex.: *Make*, *MaritalStatus*, *PolicyType*).
 - Cria uma coluna nova para cada categoria de uma variável com múltiplos valores sem ordem.
 - Evita que os modelos interpretem incorretamente uma ordem inexistente (ex.: pensar que "Ford" < "Toyota").

SOLUÇÃO ADOTADA

PRÉ-PROCESSAMENTO

2. Codificação de variáveis

- Ordinal: aplicada em variáveis ordenadas (*VehiclePrice*, *AgeOfVehicle*).
 - Atribui valores inteiros baseados na ordem das categorias.
 - Ex.: “less than 20000” → 0; “20000 to 29000” → 1; ...; “more than 69000” → 5.
 - Permite que o algoritmo entenda diferenças proporcionais entre níveis e aprenda essa progressão natural.

3. Divisão Treino/Teste:

- 80% dos registros para treinamento, 20% para avaliação.
- Proporção escolhida por ter apresentado maior precisão dos modelos em comparação às proporções 90/10, 70/30 e 60/40.
- Nos testes, observou-se que o Naive Bayes obteve desempenho semelhante nas configurações 90/10 e 60/40.

SOLUÇÃO ADOTADA

BALANCEAMENTO

- **SMOTE (oversampling)**
 - Utilizado somente no conjunto de treino, após divisão.
 - Criação de novas instâncias sintéticas da classe minoritária (*FraudFound_P = 1*).
 - Sem balanceamento, modelos tenderiam a classificar quase tudo como “não fraude”, mascarando detecção real.

MÉTRICAS

Elhussen et al. (2022) realizaram uma revisão sistemática sobre detecção de fraudes com aprendizado de máquina e utilizaram essas mesmas métricas

- **Acurácia:** proporção de previsões corretas
- **Precisão:** frações de fraudes preditas que realmente são fraudes
- **Revocação:** frações de fraudes reais que foram capturadas
- **F1-Score:** média harmônica entre precisão e revocação, essencial em classes desbalanceadas.

SOLUÇÃO ADOTADA

MODELOS

A escolha dos modelos é devido à predição binária, pois a variável alvo (*FraudFound_P*) é binária (0 e 1).

- **Regressão Logística**

- Estima a probabilidade de fraude a partir de variáveis preditoras (ex.: idade do veículo).
- Ajusta uma função logística que separa “fraude” e “não fraude”, retornando um valor entre 0 e 1.
- Configurado para realizar no máximo, 1000 iterações a fim de garantir convergência.

- **Naive Bayes**

- Combina probabilidades independentes de cada atributo para classificar um caso como fraude ou não.
- Calcula a probabilidade de fraude combinando separadamente as chances de cada característica. Depois, escolhe o resultado com maior probabilidade geral.

SOLUÇÃO ADOTADA

MODELOS

- **Support Vector Classifier (SVM)**

- Separa “fraude” de “não fraude” traçando uma fronteira complexa em espaço de alta dimensão.
- Devido ao tratamento dos dados, foi necessário utilizar o kernel RBF, pois o kernel linear não é capaz de representar fronteiras complexas, enquanto o RBF permite a criação de separações curvas entre as classes.
- O dataset pequeno exigiu um valor alto de C no SVC ($C = 10000$) para melhorar o desempenho, apesar do maior risco de overfitting.

- **Random Forest**

- Agrega várias árvores de decisão para decidir se um pedido é fraude.
- Cada árvore é treinada em subconjuntos de dados; as árvores “votam” e a maioria determina o resultado final.

RESULTADOS

TABLE I

DESEMPENHO DOS MODELOS NO CONJUNTO DE DADOS DESBALANCEADO

Modelo	Acurácia	Precisão	Revocação	F1-Score
Regressão Logística	0.9403	0.3750	0.0165	0.0316
Naïve Bayes	0.1560	0.0641	0.9780	0.1203
SVM (RBF Kernel)	0.9407	0.4800	0.0659	0.1159
Random Forest	0.9426	0.7778	0.0385	0.0733

- **Acurácia**

- Random Forest, SVM e Regressão Logística apresentaram desempenhos similares e altos. Naïve Bayes teve acurácia significativamente inferior.

- **Precisão**

- Naïve Bayes com desempenho inferior, indicando muitos falsos positivos.

- **Revocação**

- Naïve Bayes detectou quase todas as fraudes, mas com excesso de falsos positivos. Outros modelos tiveram baixa revocação.

- **F1-Score**

- Naïve Bayes teve o maior F1-Score devido à alta revocação. Random Forest foi o mais equilibrado entre os demais.

RESULTADOS

TABLE II
DESEMPENHO DOS MODELOS NO CONJUNTO DE DADOS BALANCEADO

Modelo	Acurácia	Precisão	Revocação	F1-Score
Regressão Logística	0.9403	0.3750	0.0165	0.0316
Naïve Bayes	0.2007	0.0655	0.9451	0.1225
SVM (RBF Kernel)	0.9400	0.4444	0.0659	0.1148
Random Forest	0.9410	0.5000	0.0495	0.0900

- **Acurácia**
 - Regressão Logística, SVM, Random Forest mantiveram desempenhos semelhantes e altos. Naïve Bayes continuou com acurácia baixa.
- **Precisão**
 - Naïve Bayes manteve desempenho inferior.
- **Revocação**
 - Naïve Bayes novamente com desempenho superior. Demais modelos com revocações baixas, continuando a priorizar a classe majoritária.
- **F1-Score**
 - Naïve Bayes apresentou o maior valor.

RESULTADOS

- **Naïve Bayes:**

- Maior variação positiva em acurácia (+28,69%), embora permanecendo com desempenho geral insatisfatório.

- **Random Forest:**

- Queda acentuada na precisão (-35,71%), mas foi o que mais se beneficiou em revocação (+28,57%) e F1-score (+22,78%). SMOTE ampliou sua capacidade de identificar fraudes, mas sacrificou o equilíbrio entre classes (aumentando falsos positivos).

- **Observação Geral**

- Efeitos do balanceamento são dependentes do modelo e da métrica.

TABLE III
VARIAÇÃO PERCENTUAL APÓS BALANCEAMENTO DOS DADOS

Modelo	Acurácia (%)	Precisão (%)
Regressão Logística	0.00	0.00
Naïve Bayes	+28.69	+2.18
SVM (RBF Kernel)	-0.07	-7.42
Random Forest	-0.17	-35.71
Média	+9.48	-13.65

Modelo	Revocação (%)	F1-Score (%)
Regressão Logística	0.00	0.00
Naïve Bayes	-3.36	+1.83
SVM (RBF Kernel)	0.00	-0.95
Random Forest	+28.57	+22.78
Média	+8.40	+7.88

CONCLUSÃO

Analizando os resultados obtidos, as principais conclusões e recomendações práticas deste estudo são:

- **Estabilidade da Acurácia**
 - Modelos Regressão Logística, SVM e Random Forest apresentaram acurácias elevadas (~94%) e relativamente estáveis em ambos os cenários (desbalanceado e balanceado).
- **Random Forest como Opção Equilibrada**
 - Apesar de não atingir revocações elevadas, manteve-se como o mais equilibrado em termos de F1-score no conjunto balanceado (0.0900), sendo uma opção para buscar compromisso entre detecção e controle de falsos positivos.

CONCLUSÃO

- **Problema do Naïve Bayes:**

- Alta capacidade de identificar fraudes por conta da alta revocação. Porém, acompanhado de baixa precisão.
- Na prática, pouco recomendado para a indústria devido aos potenciais custos operacionais elevados com investigações de falsos positivos.
- Alta revocação não deve ser analisada isoladamente quando há grande disparidade de precisão.

- **Recomendação Final**

- Modelos baseados em Random Forest ou SVM oferecem maior eficácia operacional para detecção de fraudes em seguros automotivos em comparação ao Naïve Bayes, especialmente sob a ótica dos custos de falsos positivos.

REFERENCIAS

- [1] McKinsey & Company, Global Insurance Report 2025: The pursuit of growth. 2024. Disponível em: <https://www.mckinsey.com/industries/financial-services/our-insights/global-insurance-report>. Acesso em 13 de maio de 2025
- F. K. Alarfaj, I. Malik, H. U. Khan, N. Almusallam, M. Ramzan, and M. Ahmed, “Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms”, IEEE Access, vol. 10, pp. 39700–39715, 2022, doi: 10.1109/ACCESS.2022.3166891.
- N. S. Elhusseny, S. M. Ouf, and A. M. Idrees, “Credit Card Fraud Detection Using Machine Learning Techniques”, Future Computing and Informatics Journal, vol. 7, no. 1, Article 2, 2022. doi: 10.54623/fue.fcij.7.1.2.



OBRIGADO