

Desenvolvimento de um Modelo Preditivo para Identificação de Fatores de Risco a partir de Dados Viários e de Acidentes de Trânsito

Development of a Predictive Model for Risk Factor Identification from Road and Traffic Accident Data

Diego Bernardo Artero – diego.bernardo.artero@gmail.com

Rafael Gutierrez Castanha – rafaelcastanha@unimar.br

INTRODUÇÃO

Os acidentes de trânsito figuram entre as principais causas de mortes evitáveis no Brasil, evidenciando a necessidade de antecipar fatores de risco. O problema da pesquisa consiste em identificar, a partir de dados públicos de sinistros viários do estado de São Paulo, os fatores mais associados à gravidade fatal de um acidente. Objetivou-se construir um pipeline de coleta e tratamento dos dados do Detran-SP, treinar um modelo em LightGBM e desenvolver uma rede neural em TensorFlow/Keras, permitindo comparar as duas abordagens. A originalidade do estudo reside na combinação entre geohash, features cíclicas de tempo e embeddings de entidade.

DESENVOLVIMENTO

Para a coleta dos dados do Infosiga – portal do Detran-SP – foi feito um web scraping, utilizando a biblioteca Selenium para Python. Enquanto o `etl_scraper.py` tem por função executar o web scraping na página do Infosiga obtendo o arquivo `dados_infosiga.zip`, o `etl_parser.py` tem a função de extrair arquivos e sintetizar os dados em 3 arquivos .csv: `Pessoas.csv`, `Sinistros.csv` e `Veículos.csv`.

A Análise Exploratória de Dados foi conduzida com as bibliotecas Pandas e Seaborn; foram descobertas inconsistências críticas da base de dados, com grande destaque à unicidade da gravidade dos acidentes anteriores a 2019 que eram classificados como Não Fatal. Vista a abundância de exemplares, todos os registros de 01/01/2015 até 31/12/2018 foram dispensados para a modelagem e as duplicidades foram tratadas via remoção dos exemplares idênticos, resultando em um dataset que abrange de 01/01/2019 até 31/08/2025.

Durante a Engenharia de Features, foram utilizados conceitos de seno e cosseno para categorizar dados cíclicos tais como dia, dia da semana, mês e hora. Além desse processo, foram aglutinadas ambas latitude e longitude em um geohash, por definição, um método de indexação espacial que transforma um ponto contínuo no espaço geográfico R^2 (latitude, longitude) em uma sequência discreta de símbolos, tal que prefixos comuns do código correspondem a regiões espaciais contíguas, permitindo consultas de proximidade e agregação espacial eficientes tratáveis como categoria, portanto, iteráveis pelo LightGBM.

Foi treinado um classificador LightGBM (Scikit-learn). Dado o desequilíbrio de classes ($\approx 25:1$), aplicou-se ajuste de threshold por maximização do F1, obtendo-se limiar ótimo de 0,1536 para a classe “Não Fatal”. Complementarmente, foi desenvolvida uma rede neural com tratamento híbrido das categóricas: embeddings de entidade (Guo; Berkahn, 2016) para as de alta cardinalidade (município, geohash, circunscrição, região administrativa) e one-hot para as de baixa cardinalidade (tipo_via, administração, conservação, turno, tipo_local); as numéricas são normalizadas internamente ao grafo; as representações são concatenadas em camadas densas com Batch Normalization e Dropout, ponderação de classe equivalente ao `scale_pos_weight` do LightGBM e o

mesmo critério de threshold por F1.

CONCLUSÃO

O projeto resultou em um pipeline funcional de coleta, tratamento, modelagem e visualização de dados de acidentes. A modelagem com LightGBM e uma rede neural em TensorFlow/Keras, tratando as categóricas via embeddings e one-hot conforme a cardinalidade, amplia o escopo comparativo do estudo e permite avaliar se a representação aprendida pelos embeddings supera a codificação nativa do LightGBM. Futuramente, recomenda-se comparar formalmente os dois modelos, ajustar hiperparâmetros e realizar novas validações cruzadas.

REFERÊNCIAS

GUO, C.; BERKHAHN, F. Entity embeddings of categorical variables. arXiv, 2016. Disponível em: <https://arxiv.org/abs/1604.06737>. Acesso em: 12 setembro 2026.

JINDAL, R. K.; AGARWAL, A. K.; SAHOO, A. K. Data analytics for analysing traffic accidents. Test Engineering and Management, v. 83, p. 14796-14803, mar./abr. 2020.

MORETTIN, P. A.; BUSSAB, W. de O. Estatística básica. São Paulo: Saraiva, 2017.

SANTOS, D. et al. Machine learning approaches to traffic accident analysis and hotspot prediction. Computers, v. 10, n. 12, p. 157, 2021.