交通死亡事故に繋がる危険因子の特定と予防策に対する研究

青山学院大学理工学部
情報テクノロジー学科Dürst研究室

学籍番号：15822100

原田 裕大

目 次

第1章　はじめに

1.1　研究の背景

日本国内における交通事故件数は依然として高い水準にあり，喫緊の社会課題となっている．2024 年の統計によれば，交通事故発生件数は約 29 万件，負傷者数は 34 万人を超え，死者数は 2,663 名に達している [1]．この死傷者数は，2024 年の日本の総人口（約 1 億 2390 万人）[2] を基に換算すると，およそ 357 人に 1 人が 1 年間のうちに交通事故で死亡または負傷している計算となり，極めて高い確率であると言える．また，交通事故による社会的影響は人的被害にとどまらない．内閣府の調査（2022 年度推計）によれば，日本における交通事故起因の経済損失は年間約 10 兆 5,540 億円に上るとされており [3]，これは国民 1 人あたり年間約 8.5 万円の負担に相当する．したがって，交通事故の削減は人命を守るだけでなく，社会的な経済負担を軽減する上でも極めて重要である．

しかしながら，交通事故は単一の要因のみで引き起こされるものではなく，道路環境，気象条件，運転者の属性といった多様な要素が複雑に絡み合って発生する事象である [?]．そのため，真に有効な削減策を導出するためには，これらの多次元的な要因を包括的に捉え，その発生メカニズムを解明することが不可欠となる．こうした分析を可能にするリソースとして，近年，警察庁により「交通事故統計オープンデータ」が公開されている [4]．このデータセットは 2019 年から 2024 年までの記録を含み，発生日時，天候，道路線形，車両の種類など，事故状況に関する詳細な変数を網羅している．

1.2　関連研究と本研究の立ち位置

このオープンデータの公開に伴い，データ活用の試みは徐々に進められている．例えば，自転車関連事故に特化してアソシエーション分析を行った事例 [5] や，シチズンサイエンスの観点からデータの可視化や市民による分析を推進する試み [6] などが報告されている．また，一部では重大事故の要因探索に関する分析 [7] も行われているが，既存の研究の多くは，特定の車種や対象に限定されているか，あるいは要因の探索的な記述に留まる傾向にある．

したがって，全国規模の包括的なデータを用い，機械学習によって事故発生リスクを高精度に予測するモデルの構築や，それに基づいた定量的な予防策の提案に

至っている事例は未だ限定的である．人間が処理可能な範囲を超えた膨大なデータを活用し，多角的な要因が複合する事故メカニズムを解明する余地が残されている．

1.3　本研究の目的と構成

そこで本研究では, 警察庁のオープンデータを用い, 機械学習アルゴリズムを活用した死亡事故予測モデルの構築を行う．具体的には，勾配ブースティング決定木の一種である **LightGBM（Light Gradient Boosting Machine）**を採用し, ロジスティック回帰やランダムフォレストといった他の手法との比較検証を通じて，不均衡な事故データに対しても高精度な予測が可能なモデルを構築する．さらに本研究の特筆すべき目的は, 単なる予測にとどまらず, モデルの解釈性（Explainability）を担保することにある．**SHAP（SHapley Additive exPlanations）値**を用いた要因分析を行うことで，複雑に絡み合う事故要因を定量的に可視化し，「どのような条件下で死亡事故リスクが増大するか」を明らかにする．最終的には，これらの分析結果に基づき，交通事故死者数を減少させるためのデータ駆動型の客観的な予防策を提言することを目指す．

本論文の構成は以下の通りである．第 2 章では，本研究で使用する交通事故統計オープンデータの概要および前処理の手順について述べる．第 3 章では，予測モデルの構築に用いる機械学習アルゴリズムの詳細と評価指標について説明する．第 4 章では，構築したモデルの精度評価結果および，要因分析から得られた知見について考察を行う．最後に第 5 章で，本研究の結論と今後の課題についてまとめる．

第2章　使用データと前処理

本章では，本研究で使用したデータセットの概要と，機械学習モデルの学習に適した形式に変換するための前処理手法について述べる．特に，交通事故予測において致命的な問題となる「データリーク」の排除と，モデルの汎用性を高めるための特徴量エンジニアリングの過程について詳述する．

2.1　データセットの概要

本研究では，分析の基礎データとして，警察庁が公開している「交通事故統計オープンデータ」 [1] を採用した．このデータセットは，日本政府が推進する「世界最先端デジタル国家創造宣言・官民データ活用推進基本計画」に基づき，交通事故抑止や交通安全対策の高度化を目的として一般公開されているものである．分析対象期間は，2019 年から 2024 年までの 6 年間とした．本データセットは，日本国内で発生したすべての「人身事故（死亡または負傷を伴う事故）」を網羅しており，各事故について警察官が現場検分等を通じて記録した詳細な情報が含まれている．

データの形式は，都道府県別や月別にあらかじめ集計された統計表ではなく，発生した事故の 1 件 1 件が独立したデータ（レコード）として詳細に記録された形式で提供されている点が特徴である．

- **事故管理情報**：発生日時，発生場所（緯度・経度，都道府県，市区町村），天候，路面状態など
- **道路・環境情報**：道路形状，信号機の有無，道路幅員，規制速度，一時停止規制など
- **当事者情報**：当事者の年齢，性別，車両の形状，衝突部位，損傷程度など

オープンデータは年ごとに個別の CSV ファイルとして提供されており，各項目は「コードブック（データ定義書）」に基づいた数値コード（例：晴れ=1，曇り=2）で管理されている．本研究では，Python のデータ解析ライブラリである Pandas を用いてこれら 6 年分のデータを結合し，一つの統合データフレームを作成した．統合後の総レコード数は約 189 万件（1,895,275 件）に上る．このサンプルサイズは，単年度の分析では統計的な有意差を見出すことが困難な「死亡事故」のよう

な稀な事象の特性を学習するために十分な規模であり，かつ長期的なトレンドや季節性をモデルに取り込むことを可能にするものである．

2.2　データクリーニングとリークの排除

本研究におけるデータクリーニングの工程では，まずデータの欠損や形式の整合性を確認した．事前調査の結果，本データセットにおいて欠損値が含まれるレコードは全体の 0%であり，データの完全性が極めて高いことが確認された．したがって，欠損値に対する特別な処理（削除や補完）は不要であると判断し，全レコードを分析に使用した．

次に，予測モデルの構築にあたり最も注意を要する「データリーク（Data Leakage）」の排除を重点的に実施した．データリークとは，予測を行う時点では入手不可能な情報が学習データに含まれてしまう現象を指し，これが見過ごされると，モデルは実運用で機能しない虚偽の高精度を出力してしまう．本データセットに含まれる変数のうち，「事故内容」「人身損傷程度」「車両の損壊程度」などの項目は，事故が発生し，警察による検分が行われた後に確定する「事後情報」である．例えば，「事故内容」カラムには「死亡」や「負傷」といった結果が直接記述されており，これを入力変数として用いることは，答えを見ながら問題を解くことに等しい．したがって本研究では，これらの事後情報を含む計 14 カラムをデータリークの原因となる変数と見なし，学習用データから完全に除外した．これにより，事故発生直前の状況や環境要因のみからリスクを予測するという，実用的な問題設定を厳密に守っている．

2.3　特徴量エンジニアリング

元のデータセットに含まれる変数は，必ずしも機械学習モデルが解釈しやすい形式ではない．そこで，モデルの学習効率と予測精度を向上させるため，以下の 3 つの観点から特徴量エンジニアリングを実施した．

2.3.1　日時データの分解

元データの「発生日時」は日付と時刻を含むタイムスタンプ形式であるが，多くの機械学習アルゴリズムはこれを直接扱うことができない．また，交通事故には「季節による路面状況の変化」や「時間帯による視界の変化」など，周期的な傾向が存在する．そこで，「発生日時」カラムを「年（year）」「月（month）」「日（day）」「時間（hour）」の 4 つの数値データに分解した．これにより，モデルは年ごとのトレンド変化や，季節性・時間帯ごとのリスク変動を個別の特徴量として

学習することが可能となる．なお，情報が重複するため，元の「発生日時」カラムは削除した．

2.3.2　地理情報のクラスタリング (Geo-Clustering)

事故発生地点を示す「緯度」「経度」は，極めて詳細な位置情報を提供する一方で，そのまま連続値としてモデルに入力すると過学習（Overfitting）を引き起こすリスクがある．特定の交差点の座標そのものを記憶してしまうと，未知の地点での予測能力が低下するためである．この問題に対処するため，本研究では教師なし学習の一種である MiniBatchKMeans 法を用いて，日本全国の発生地点を 50 個のエリア（クラスタ）に分割する処理を行った．各事故データには，緯度・経度の代わりに `area_id`（0〜49 のカテゴリ変数）を付与した．これにより，詳細すぎる座標情報を「地域特性」という抽象的な概念に変換し，モデルの汎用性を高めている．

2.3.3　道路種別の集約

元データの「路線コード」は，国道や県道，バイパス区間などを区別するために約 6,800 種類のユニーク値を持っている．このようにカテゴリ数が膨大であると，データの希薄性を招き，モデルの学習が不安定になる．また，バイパスの区間番号のような微細な違いは，死亡事故リスクの予測において本質的な意味を持たない場合が多い．そこで，路線コードの上位桁に基づき，道路を「一般国道」「主要地方道」「高速自動車国道」「市町村道」など，意味のある 15 種類のカテゴリ（`road_type`）に集約した．この次元削減により，情報の粒度を適正化し，道路環境によるリスクの違いをモデルが捉えやすくなるよう設計した．

2.4　目的変数の設定と不均衡データ

本研究の目的は，ある事故が「死亡事故」に至るか否かを予測することである．前述の通り「事故内容」カラムはリーク変数として削除したが，目的変数（正解ラベル）を作成するために，「死者数」カラムを用いた．具体的には，死者数が 1 名以上の事故を「死亡事故（$y = 1$）」，0 名の事故を「非死亡事故（$y = 0$）」と定義した．ここで特筆すべきは，クラス間の極端な不均衡である．全データ約 189 万件のうち，死亡事故は約 1.6 万件に過ぎず，その比率はおよそ 118:1 となっている．このような不均衡データ（Imbalanced Data）をそのまま学習させると，モデルはすべてのデータを「非死亡事故」と予測するだけで 99%以上の正解率を出せてしまうため，死亡事故の検知に失敗する．この問題に対する対策（Class Weight および SMOTE の適用）については，次章の実験設定にて詳述する．

謝辞

本論文を執筆するにあたり，多くの方々から多大なる御指導及び御助言を賜りました．このような研究の契機と環境を与えてくださり暖かくご指導して頂いた Martin J. Dürst 教授に心から感謝致します．温かい研究室の仲間たちに感謝します．みんなのおかげでここまでたどりつくことができました．ありがとう．

2222 年 1 月

参考文献

[1] 警察庁. 交通事故統計. https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&toukei=00130002&tstat=000001027457&cycle=7&year=20240&month=0&stat_infid=000040249644&tclass1val=0, 2024. e-Stat 政府統計の総合窓口 (閲覧日: 2024 年 X 月 X 日).

[2] 総務省統計局. 人口推計（2024 年 10 月 1 日現在確定値）. https://www.stat.go.jp/data/jinsui/2024np/index.html#a05k01-a, 2024. 統計データ: 2024 年 10 月 1 日時点 (閲覧日: 2025 年 12 月 13 日).

[3] 内閣府. 令和 4 年度交通事故の被害・損失の経済的分析に関する調査研究（要旨）. https://www8.cao.go.jp/koutu/chou-ken/r04/pdf/youshi.pdf, 2023. 閲覧日: 2025 年 12 月 13 日.

[4] 警察庁. 交通事故統計情報のオープンデータ, 2024. （2019 年から 2024 年のデータを収録）.

[5] 坪井 志郎, 三村 康広, and 嶋田 善昭. アソシエーション分析を用いた自転車関連事故の特性分析. In 第 *44* 回交通工学研究発表会論文集（研究論文）, volume 44, pages 364–369, 2024. （自転車事故という特定の車種・形態に焦点を当てた研究）.

[6] 青木 和人. シチズンサイエンスによる交通事故統計情報オープンデータ分析. In *2022* 年度社会情報学会（*SSI*）学会大会予稿集：連携報告 *2*「オープンデータ・オープンガバメント」, volume 2022, pages 1–6, 2022. （オープンデータの可視化や市民参加型活用に関する研究）.

[7] 野口 直樹, 乾口 雅弘, 林 直樹, 関 宏理, 須山 崇仁, and 長谷川 潤. 交通事故データの解析による重大事故の要因探索. In 第 *37* 回ファジィシステムシンポジウム講演論文集 *(FSS2021* オンライン, volume 37, pages 675–678, 2021. （ファジィ理論等を用いた要因探索的研究）.

付録

A. 第123回情報処理学会全国大会発表論文
B. 質疑応答

付録 A
第123回情報処理学会全国大会発表論文

付録 B
質疑応答

発表後の質疑応答

X 先生の質問（要約）

ZZZ は他のものと比べてどのような違いは何か？

解答

Y 先生の質問（要約）

FFF はどのようなものでそのメリットは？

解答