知能情報実験 III（データマイニング班）
交通事故件数予測アプリケーションの作成

195726B 柴田大輝 195733D 花城可樹, 195753J 平良信長, 195764D 金城亘

2021 年 8 月 24 日

目次

概要

本文書は知能情報実験 III データマイニング班グループ 1 における実験の最終レポートとして作成したものである. この実験では, 交通事故件数予測アプリケーションの作成を目標に, 交通事故のデータを用いて機械学習を行い, 精度向上をはかり, 実験の結果と考察を述べた.

1 テーマ

本グループでは, 気象情報や行事イベント, 交通量などから, 交通事故数を予測することを対象問題として設定した. 年々減少傾向にはあるものの, 交通事故は昨年日本において年間 30 万件起きており [1], 交通事故数を予測することで交通事故予防に寄与する. また国土交通省によると, 「人生で交通事故にあう人は, 2 人に 1 人」という資料が社会資本整備審議会の会議の場で提供されており [2], 交通事故は身近なものであることがわかる. 私たちのグループでは具体的に交通事故をモデル化することでより解釈しやすくし, 対策を考えやすくなるのではないかと考えた.

2 実験方法

実験は PC を用いて行う. また, 実行環境として, Python が使える必要がある他, numpy, pandas, sklear などのライブラリが使用可能である必要がある. この実験の GitHub 上にあるリポジトリ (`https://github.com/Yoshiki-Hanashiro/group1_datamining`) をクローンし, 自分の PC 上にコードを引っ張ってくる. 次にデータセットを生成するため, group1_datamining/code/data_create.py を実行する. 数字の羅列がコンソールに出力されていれば無事に実行されている. データの生成が終了したら group1_datamining/code/Traffic_accident.py を実行し, モデルの精度や事故件数を出力する.

2.1 実験目的

この実験は, 気象情報や交通量から交通事故件数を予測することを目的としていて, 実際の交通事故データから, 交通事故にはどのような傾向があるのかを明らかにしたい.

2.2 データセット構築

警察庁のホームページから交通事故統計情報のオープンデータを利用した. (`https://www.npa.go.jp/publications/statistics/koutsuu/opendata/2019/opendata_2019.html`) 回帰分析を行うため 1 件の事故ごとの情報がまとめられてある交通事故統計情報のデータセットは天候などのデータセットと合わせて時間ごとの事故数として加工する必要があると判断した.

そのため, 気象庁のデータセットと組み合わせて天候の変化と路面形状を説明変数, 1 時間ごとの事故数を被説明変数として回帰分析を行なった. 気象庁のデータセットは各都道府県（県庁所在地もしくは中央付近）の天候として降水量データと気温のデータを採用し, 路面形状として, カーブのデータと坂道のデータを採用した. 路面形状についてはカテゴリデータを数値化したものになっていて, カテゴリに応じてカーブ, 坂が急になるような値であったため, 一時間に起きた事故の平均を取っている. 総サンプル数は 411720 個（365 日× 24 時間× 47 都道府県）となり, 総次元数は 4 とした.

2.3 モデル選定

モデルは SGD Regressor と SVR の両方を選定し比較した. まず，交通事故件数を予測するために回帰モデルを検討し，今回利用するデータセットが 40 万行近いレコード数であることから, sklearn の cheat sheet における回帰モデルの SGD Regressor を選定して実験を行った. はじめ, 事故一件ごとの事故データである 40 万件のデータセットを用いようとしていたが, 後ほどそのデータを 1 時間ごとにまとめ, 回帰学習用に加工したデータセットを作成する事となったため, データセットの総サンプル数に変更があった. SGDRegressor は，SGD(確率的勾配降下法) を用いた回帰モデルであり，SGD とは，バッチ学習の一つでパラメータを更新するために勾配を求める際に，データセットからデータのまとまりをランダムサンプリングして逐次的に学習するミニバッチ学習である. また，SVR はサポートベクトルを使った回帰学習であり, 非線形のカーネル関数を使うことで非線形の回帰分析を行うことが可能である. このように SVR では非線形な回帰学習を行い, 回帰曲面を予測するため, データセットの形にフィットした回帰が行えると考えたため SVR をアルゴリズムとして選択した.

2.3.1 SGD Regressor

パラメータ調整

SGD のハイパーパラメータとして，損失関数，学習率，early_stopping を選択してそれぞれの性能について評価した. 損失関数とは誤差の計算過程を表す関数であり，squared_loss, huber, epsilon_insensitive, squared_epsilon_insensitive の 4 つについて調査した. 学習率は勾配降下法におけるパラメータ更新時での変化量を表す指標であり，constant, optimal, invscaling, adaptive の 4 つについて調査した. early_stopping はブーリアン型のハイパーパラメータであり，テストデータに対して過学習が起こる前に学習を終了することで汎化性能を上げる手法である. また，標準化の有無における性能の変化についても調査した.

実験結果

実験はテストデータと教師データの誤差平均，RMSE，MAE の 3 つについて求めた. RMSE と MAE をそれぞれ評価した理由は，RMSE が外れ値の影響を強く受け，MAE が外れ値の影響を比較的受けない特性を持つことを考慮した. また，誤差はそれぞれ 5 回平均を求めた.

標準化

	誤差平均	RMSE	MAE
標準化あり	0.018	2.109	1.515
標準化なし	-0.004	2.180	1.548

表 1　標準化の有無による実行結果

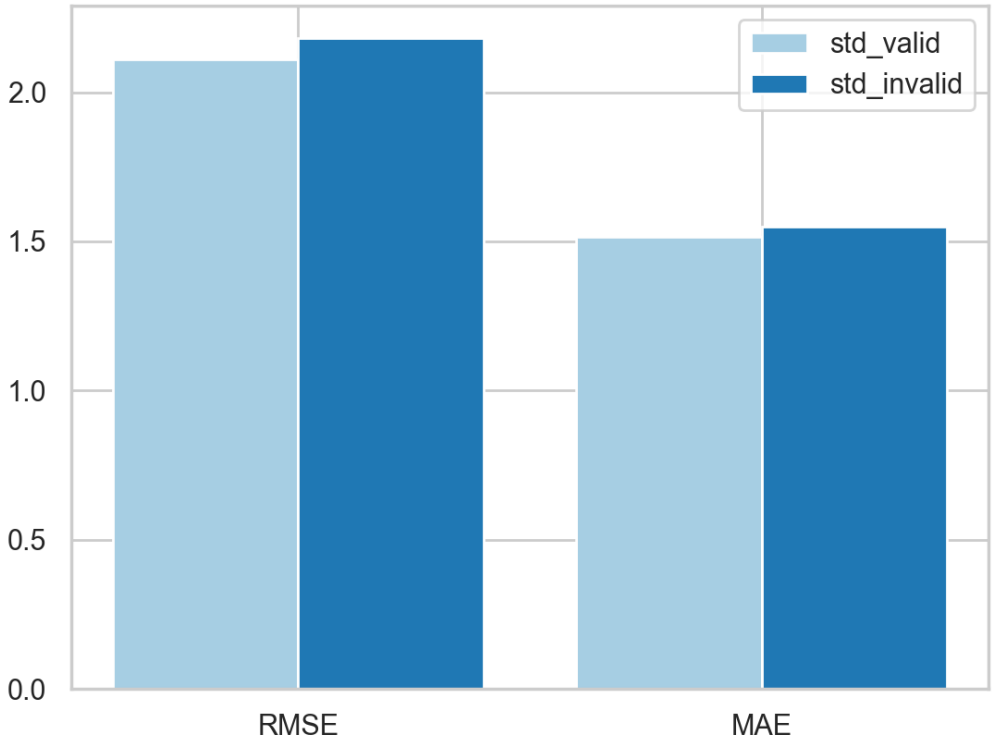

図 1　標準化の有無による比較

以下のハイパーパラメータでは，標準化ありでの実行結果

損失関数

	誤差平均	RMSE	MAE
squared_loss	0.028	2.106	1.510
huber	-1.086	2.367	1.354
epsilon_insensitive	-0.862	2.279	1.352
squared_epsilon_insensitive	0.009	2.104	1.515

表 2　損失関数の違いによる実行結果

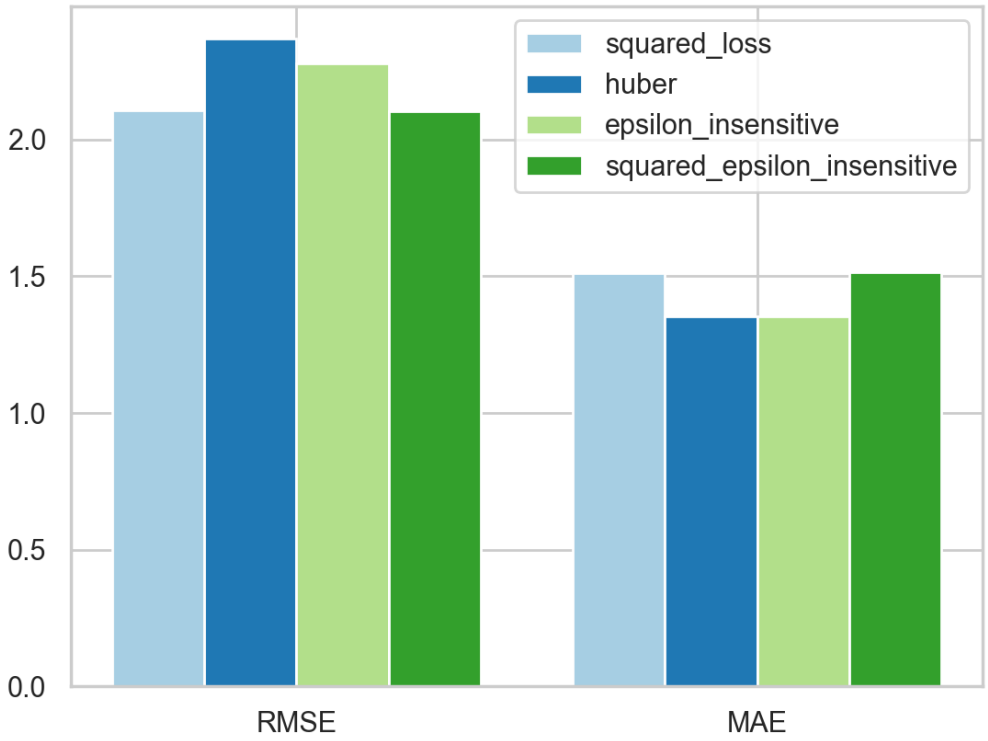

図 2　損失関数の違いによる比較

学習率

	誤差平均	RMSE	MAE
constant	0.020	2.127	1.530
optimal	-0.068	2.114	1.481
invscaling	0.004	2.104	1.510
adaptive	-0.006	2.113	1.514

表 3　学習率の違いによる実行結果

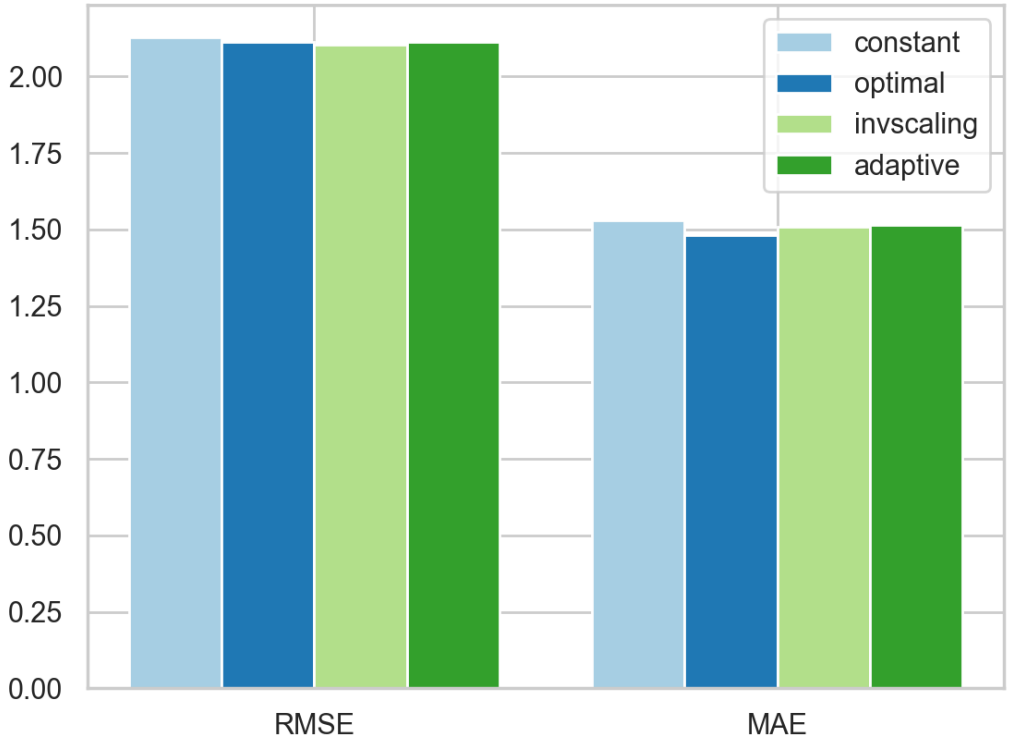

図 3　学習率の違いによる比較

early_stopping

	誤差平均	RMSE	MAE
early_stopping あり	0.017	2.107	1.521
early_stopping なし	-0.004	2.111	1.510

表 4　early stopping の有無による実行結果

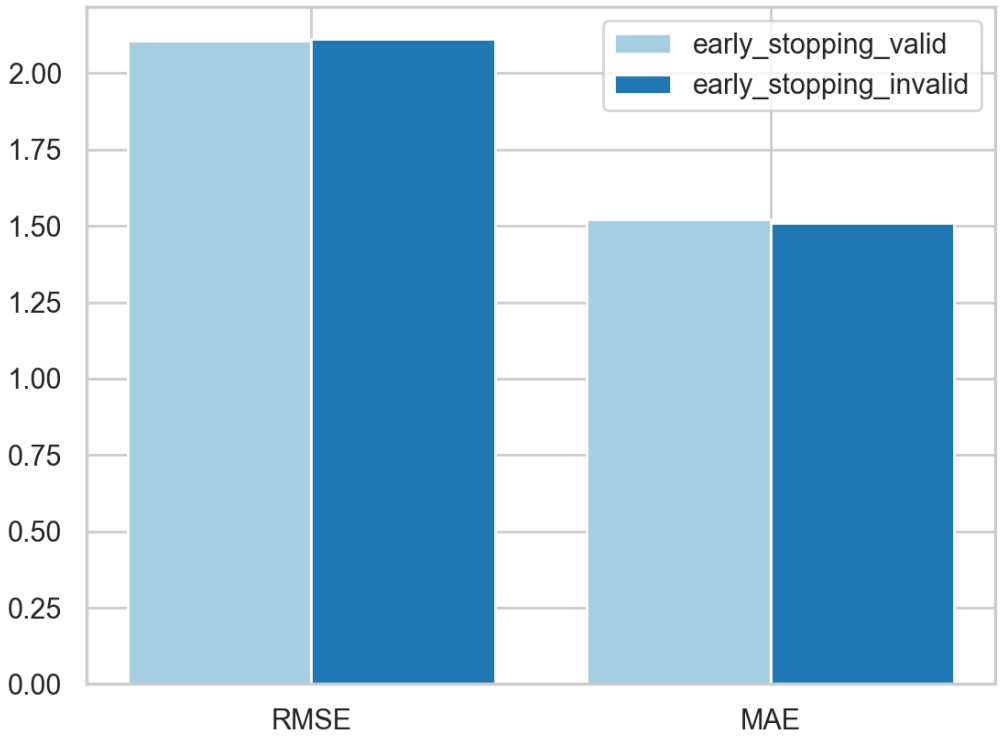

図 4　early stopping の有無による比較

考察

いずれの図からも読み取れるように RMSE と MAE を比較すると RMSE の値の方が大きくなっていることがわかる．このことから，データセットには外れ値のような教師データとの誤差の大きなデータが含まれていることが分かる．また，図 1 の標準化の有無による実行結果から標準化なし (std_invalid) の場合と標準化あり (std_valid) の場合を比較すると，標準化ありの方が棒グラフがわずかに小さくなっていることから誤差が小さくなっていることが分かる．よって，標準化をすることで精度が向上すると考えられる．図 2 の損失関数の実行結果から読み取れることとして，例えば RMSE で「squared_loss」と「squared_epsilon_insensitive」が「huber」と

「epsilon_insensitive」より低値であり，MAE ではその二つが相対的に高値となっているように，一方で RMSE と MAE の値が高くなっているものが他方では低くなっていることがわかる．この結果から「huber」と「epsilon_insensitive」は外れ値の影響を強く見積もるパラメータであることが分かる．最適であるパラメータを選択する場合外れ値の影響を考慮するかどうかにもよるが，MAE で低値を取り，且つ RMSE で「huber」と比較して低値を取っている「epsilon_insensitive」が最も良いと考えられる．図 3 の学習率と図 4 の early stopping の結果では，それぞれのパラメータを変更してもほとんど数値が変化していないことから，このモデルでは過学習が起こっていないことが読み取れる．

2.3.2 SVR

パラメータ調整

手動調整を行うパラメータとして, 今回は, kernel, epsilon を選択した．まず kernel について, kernel とは, 使用するカーネル関数を指定するためのパラメータで, 適切なカーネル関数はデータの形状によって異なるため, 変更の必要があると考えた．epsilon は, 不感度関数と呼ばれ, イプシロンの役割は訓練時に真値と予測値の誤差が小さいサンプルを無視して, 頑健なモデルを学習すること．ε が大きいほど, 無視するサンプルの数が増加し, 学習に影響するサンプルであるサポートベクトルの数が減少する．頑健なモデルはロバストなモデルとも言われ, 多様なデータに対して, 特に学習データ内に多く存在しない傾向のデータに対して, 適切に予測を行うことができるのことモデルのことを指す.

実験結果

カーネル関数を線形カーネル, RBF カーネル, シグモイドカーネルに各々指定すると, 表 5 を見ると分かるように精度には大きな違いが出なかった．以下の結果は複数回実行した際の平均をとっている．誤差の平均は, テストデータの実際のラベルと予測したラベルの誤差を平均したもの.

	線形カーネル	RBF カーネル	シグモイドカーネル
誤差平均	-0.3380	-0.3075	-0.3416

表 5　選択したカーネルでの誤差平均の比較

データと回帰曲面をプロットしてみると図 5 のようになり, データ配置に対して沿った曲面になっているシグモイドカーネルを選択することにした.

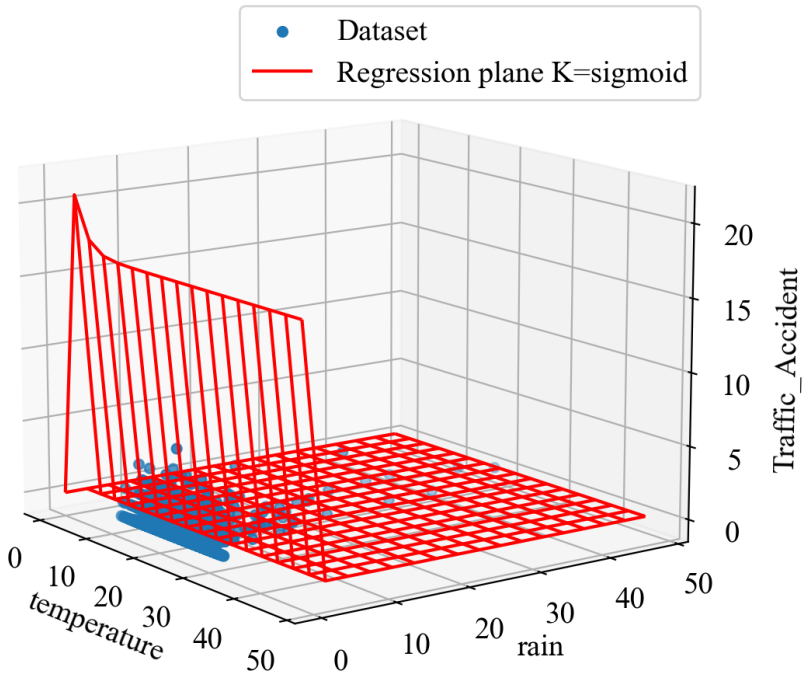

図 5　sigmoid カーネルを使用した際のデータ配置

シグモイドカーネルを選択した上で, イプシロンを 0.1, 0.5, 1 と変化させ, 誤差の平均を見てみると, 表 6 の通りイプシロン を 0.5 にした際に, 誤差平均が 0.0518 に減少し, 精度の改善がみられた.

	$\varepsilon = 0.1$	$\varepsilon = 0.5$	$\varepsilon = 1$
誤差平均	0.3408	0.0518	0.5613

表 6　イプシロンを変化させた際の誤差平均の比較

カーネル関数をシグモイドカーネルと選択し，イプシロンを 0.5 にした際のデータ配置が図 6 の通りとなる．

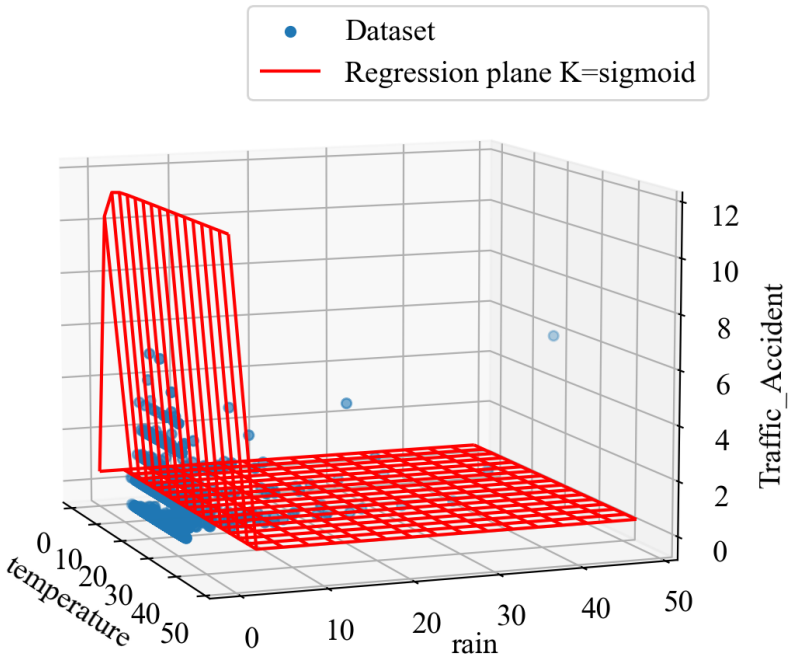

図 6　sigmoid カーネルを使用し，ε を 0.5 とした際のデータ配置

考察

実験結果として，イプシロンを 0.5 に変更した際に精度の改善がみられた．このデータセットでは，降水量を扱っており，どうしても雨の日のデータが少なくなってしまう．そのためイプシロンを調整することで，学習データ内に多く存在しない傾向のデータに対して，より適切に予測を行うことができるような頑健なモデルになったのではないかと考える．

3　意図していた実験計画との違い

当初の目標はアプリケーションまで作成することだったが，データセットの構築やモデルとパラメータの検証に多くの時間を費やしたため，アプリの制作を断念して精度を向上させることに尽力した．機械学習の経験が未だ浅いことやアプリを制作したことのないメンバーもいたため，中間報告等のタイミングに目標の見直しをすれば，グループ全体としてよりアクティブに活動できたのではないかと思う．また，最小限のモデルを完成させて，それを徐々にアップデートしていく開発をする予定であったが，データセットの情報収集や特徴量の選択基準が分からなかったりと手探りの状態であったため，モデルの開発までに時間がかかってしまったことが挙げられる．使用した生

データが 30 万件を超える大きなものだったため，データ数を極端に減らしたり特徴量を小さくすることで最小限の情報量に留めてデータセットを制作するようにすれば，より効率的に進められたと思われる．データセットが交通事故一件ごとのデータであったため, データセットを自前で生成する必要があった．生成できたはいいものの, 後から追加したい説明変数が出てきた際に追加するのが難しく断念した箇所があった．

4　まとめ

github を利用したグループ開発を行い，データセットの構築を通して，生データからモデルに適用するデータセットを作成する方法や予測モデルに最適な説明変数の選択基準，方法について理解することができた．選択したモデルにより得られた予測結果をグラフを利用して最適なパラメータの比較を分かりやすくしたり，3 次元空間にサンプルと回帰平面を出力してテストデータと教師データとの誤差精度の変化について比較することができた．

参考文献

[1] 交通事故総合分析センター 交通事故発生状況, https://www.itarda.or.jp/situation_accidents, 2021/06/03

[2] レスポンス 興味深いデータ「一生のうち交通事故にあう確率は何 % でしょうか?」,https://response.jp/article/2002/03/05/15433.html, 2002/03/05

[3] 警察庁 オープンデータ, https://www.npa.go.jp/publications/statistics/koutsuu/opendata/2019/opendata_2019.html, 2021/06/03

[4] 授業資料 レポートテンプレート, https://github.com/naltoma/info3dm-report-template/blob/master/template.pdf, 2020/07/02

[5] sklearn SGDRegressor https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html#sklearn.linear_model.SGDRegressor