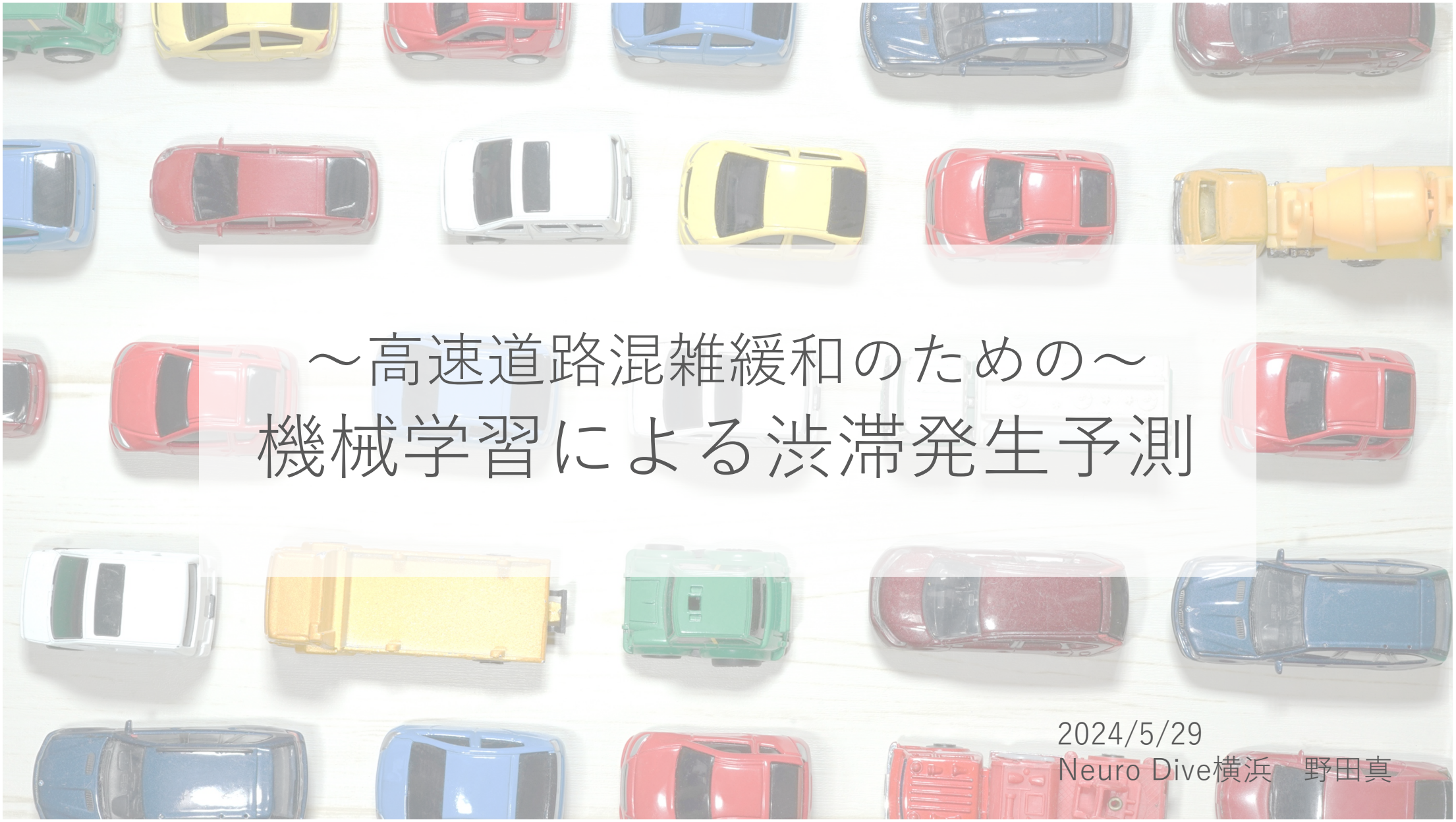


～高速道路混雑緩和のための～
機械学習による渋滞発生予測

2024/5/29

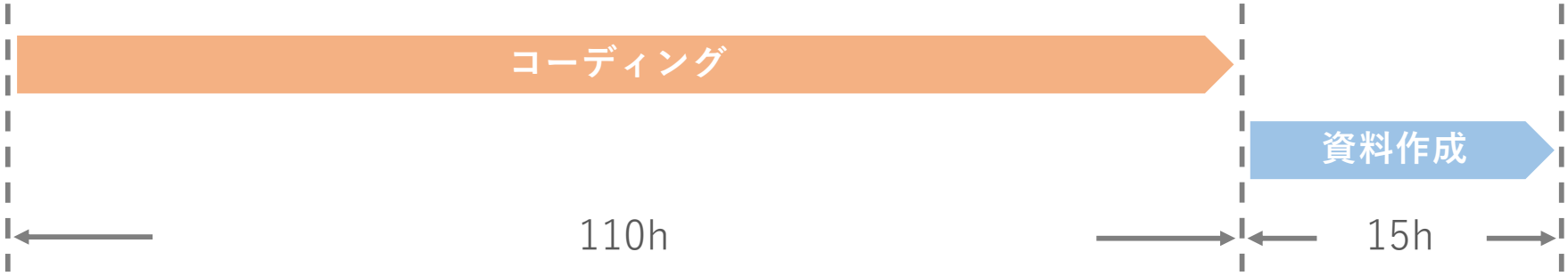
Neuro Dive横浜　野田真

目次

はじめに

- 発表者プロフィール：
  - 氏名：野田真
  - 経歴：高校卒業後、接客、介護、テレホンオペレーター、
    企業・省庁での事務職・IT事務職を経て、うつ病で休職。
  - 学習歴：2023年4月Neuro Dive横浜利用開始からPython、機械学習の学習を開始。

- 制作期間：
  - コーディング：2月19日 ~ 3月29日
  - 発表資料作成：5月8日 ~ 5月22日



課題の背景と目的

背景：

1. 国内の渋滞による労働時間損失は約50億時間、労働価値の損失は約11兆円と言われ、ドライバー労働力の不足が大きな問題になっている。
2. 国内全産業の、売上高に対する物流コストの比率の平均は、2019年の4.9%から2021年の5.7%に急騰している。

目的：

1. 高速道路渋滞状況の予測により渋滞回避を可能にし、輸送能力の逼迫を緩和するとともに、労働時間・価値の損失を低減する。
2. 輸送量あたりの燃料消費量を削減することで、物流の費用対効果を向上させ、あわせて環境負荷を低減する。

課題の概要（1）

予測の概要：

予測対象：

- 東北自動車道と関越自動車道の全925.8kmを268の区間に分け、各区間について1時間毎の渋滞予測を行う。

予測の期間：

- 2023年7月1日~2023年7月31日

予測に用いるデータ：

- 予測対象日前日の1時間毎のデータ

課題の概要（２）

データの概要（１）：

道路構造データ：

- 区間の始点・終点となるインターチェンジの、コード、キロポストの値、緯度経度。
- 区間の始点・終点となるICに接続する他ICの数
- 方向（上り線・下り線）

トラフィックカウンターデータ：

- 日時
- トラフィックカウンターが位置するキロポストの値
- その時間・その区間の全車線の車両通過台数の合計（allCars）
- その時間・その区間の全車線の車両による占有率（OCC）
- その時間・その区間の全車線の車両平均速度（speed、**目的変数**）
- その時間・その区間の渋滞の有無（is_congestion、**目的変数**）

※ speedが40以下のときis_congestion=1

課題の概要（３）

データの概要（２）：

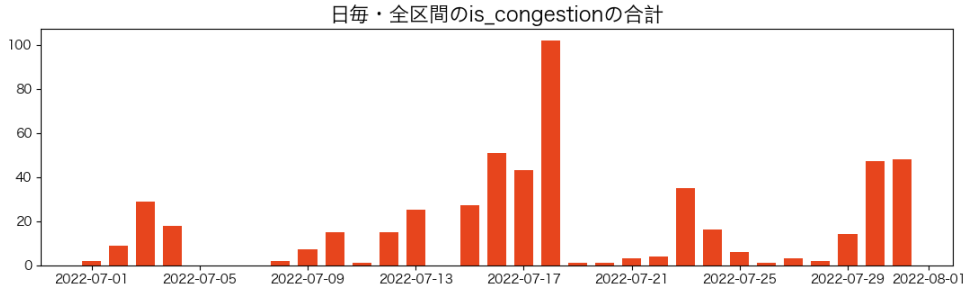
ドラぷら※検索ログ二次データ：

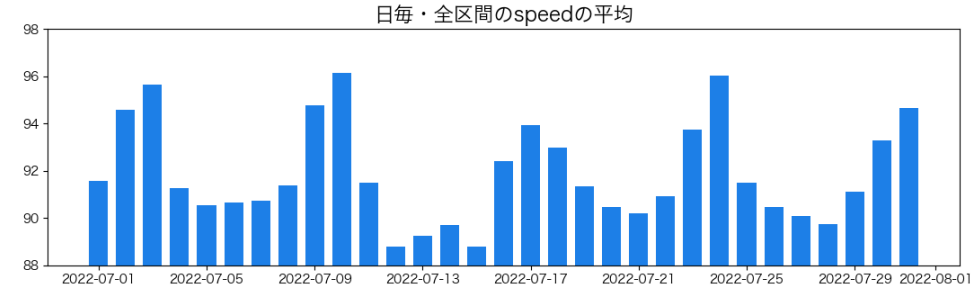
- その時間・その区間を対象にした時間指定ありドラぷらルート検索（search_specified）
- その区間を対象にした、
  前日24時間中に行われた時間指定なしドラぷらルート検索数。
  （search_unspecified）

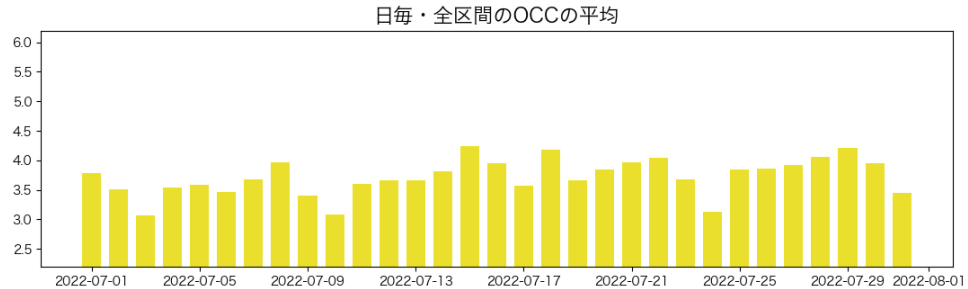
※NEXCO東日本が運営する、全国高速道路のルート・料金検索サイト。

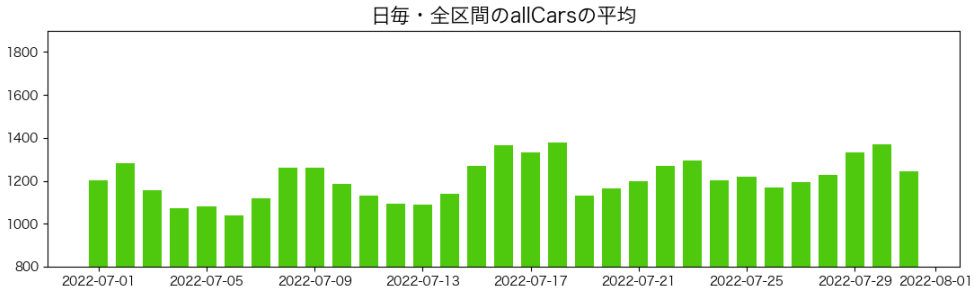
データの分析（1）

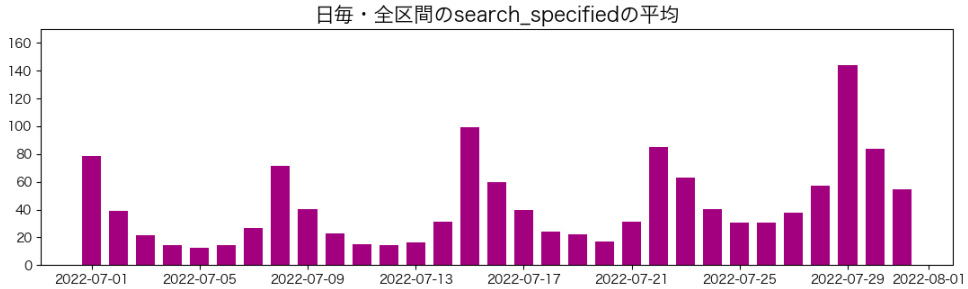
特徴量の分布（全区間）：

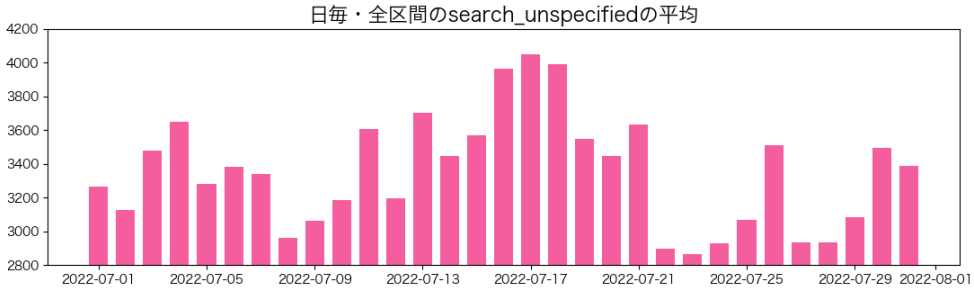








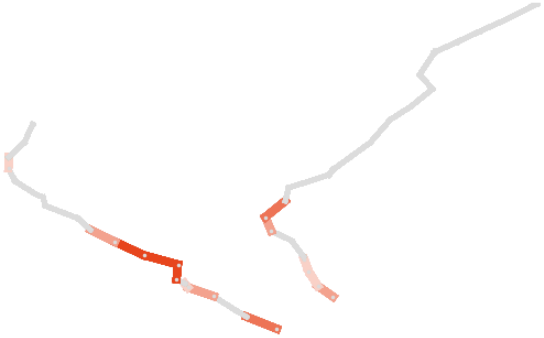




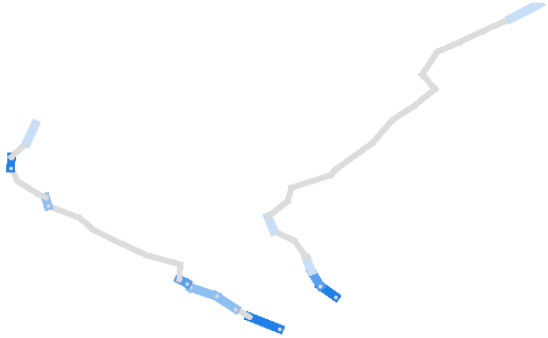
データの分析（２）

特徴量の分布（ 2022/7/1 ~ 7/31 の平均・上り線区間毎）：

区間毎・全期間のis_congestionの合計(上り)

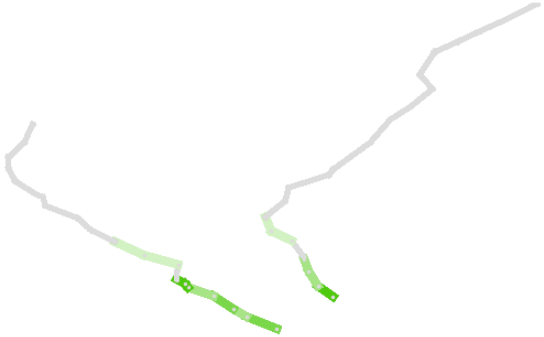


区間毎・全期間のspeedの平均(上り)

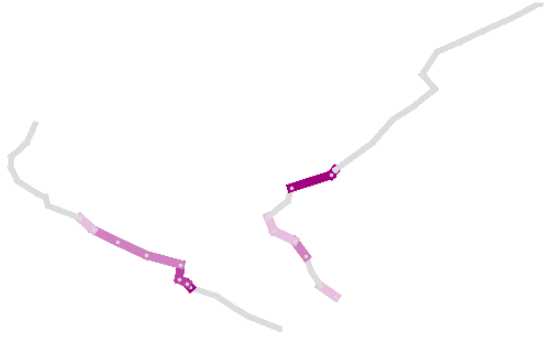


区間毎・全期間のOCCの平均(上り)

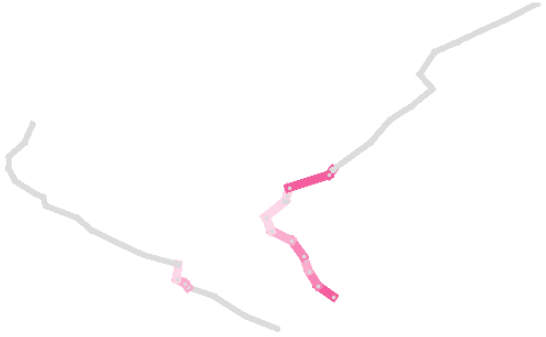
区間毎・全期間のallCarsの平均(上り)



区間毎・全期間のsearch_specifiedの平均(上り)



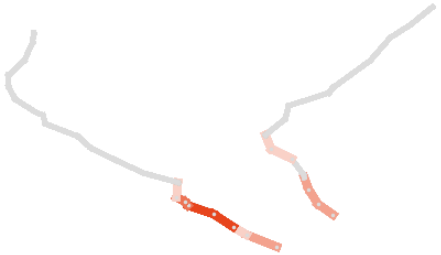
区間毎・全期間のsearch_unspecifiedの平均(上り)



データの分析（3）

特徴量の分布（ 2022/7/1 ~ 7/31 の平均・下り線区間毎） ：

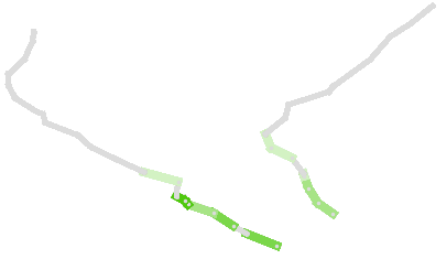
区間毎・全期間のis_congestionの合計(下り)



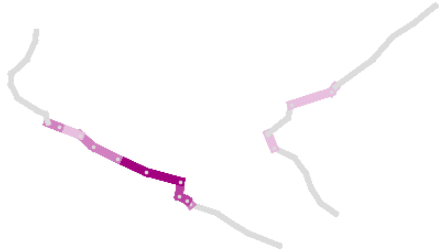
区間毎・全期間のspeedの平均(下り)

区間毎・全期間のOCCの平均(下り)

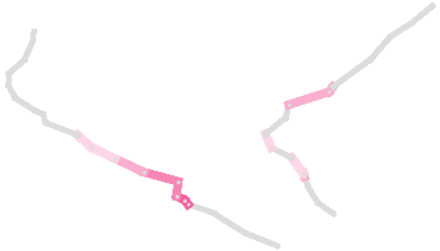
区間毎・全期間のallCarsの平均(下り)



区間毎・全期間のsearch_specifiedの平均(下り)



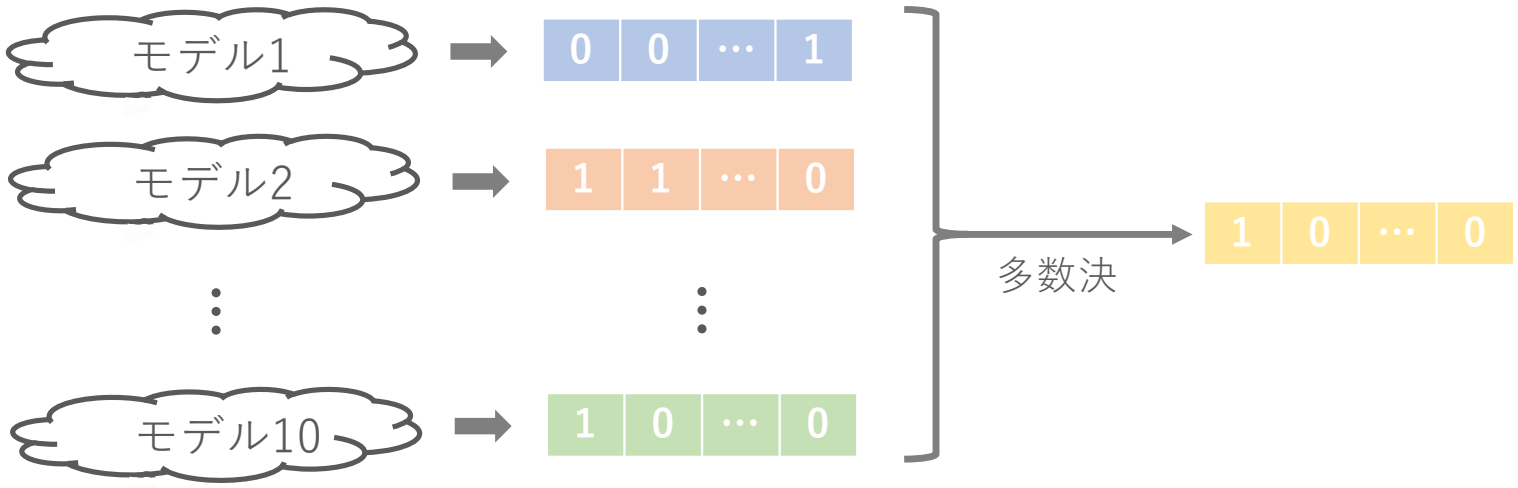
区間毎・全期間のsearch_unspecifiedの平均(下り)



予測のアプローチ（1）

アンサンブル：

- アルゴリズムにはlightGBMを用いた。
- speedを目的変数とする回帰モデルと
  is_congestionを目的変数とする二値分類モデルをつくり、それぞれについて、
  クロスバリデーションによって5個のモデルを作成した。
- 計10個のモデルの予測結果を、
  Votingによりアンサンブルして最終結果を得た。



予測のアプローチ（２）

特徴量生成：

回帰モデル・二値分類モデル共通：

- 曜日
- 同じ時刻のis_congestionの平均値
- 同じ日付・時刻のspeedの平均値
- 同じ日付・時刻のOCCの平均値
- 同じ時刻のspeedの7日間移動平均値

回帰モデルのみ：

- speedと、speedの同時刻7時間移動平均の比。

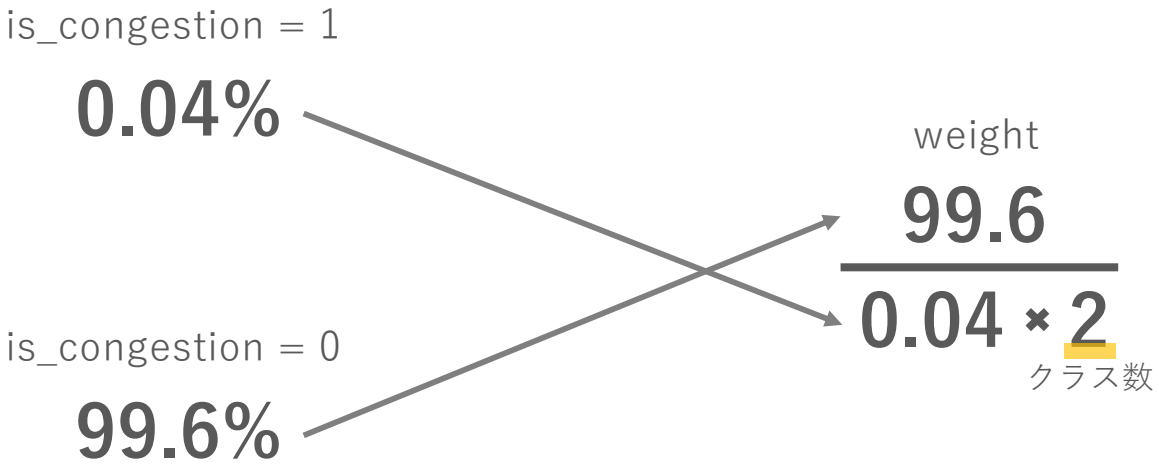
二値分類モデルのみ：

- 連休の何日目か
- 同じ日付・時刻のsearch_specifiedの平均値
- 同じ日付のsearch_unspecifiedの平均値
- 同じ時刻のOCCの7日間移動平均値

予測のアプローチ（3）

不均衡データへの対応：

is_congestionは99.6%の渋滞無しラベルに対して、
0.04%の渋滞有りラベルとなっており、極端な不均衡データである。
このため二値分類モデルではラベルの重み付けを行った。



モデルはweightの分だけ少数ラベルを重視して学習する

結果（1）

精度評価：

アンサンブルモデルによる、268区間、2023/7/1 0時~7/31 23時の予測結果。

TN (正解:渋滞していない) 197669	FP (誤り:渋滞していないが 渋滞していると予測) 529
FN (誤り:渋滞しているが 渋滞していないと予測) 523	TP (正解:渋滞している) 258

F1 = 0.329

Precision(適合率) = 0.328

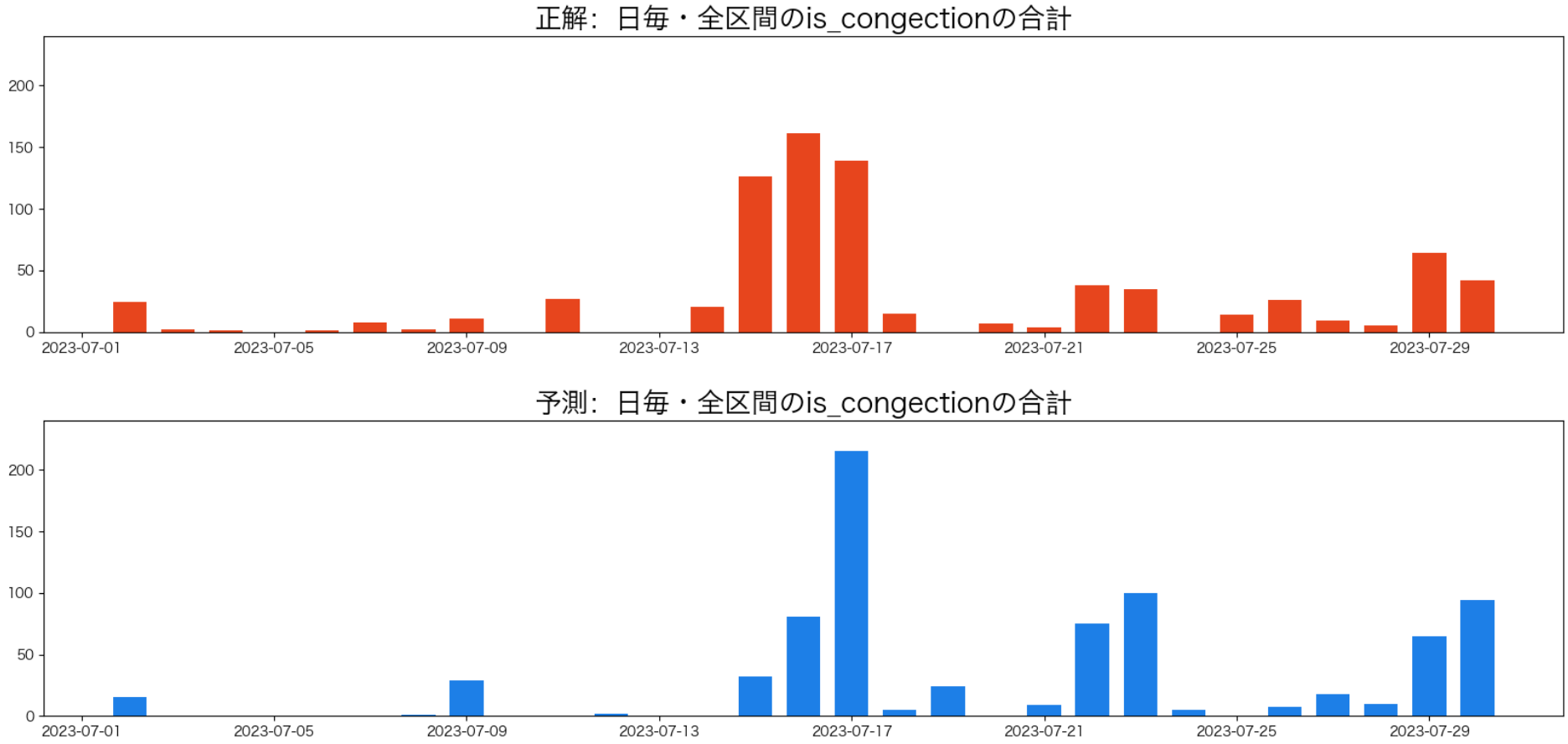
Recall(再現率) = 0.330

Accuracy(正解率) = 0.995

コンペティションサイト上の順位は、71位 / 177エントリー（上位40%）であった。

結果（2）

is_congectionの正解・予測比較（全区間）：

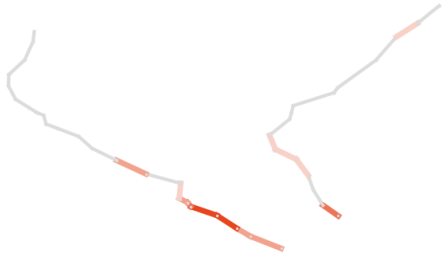


結果（3）

is_congectionの正解・予測比較（2023/7/1~2023/7/31の合計）：

正解：区間毎・全期間のis_congestionの合計(上り)

正解：区間毎・全期間のis_congestionの合計(下り)



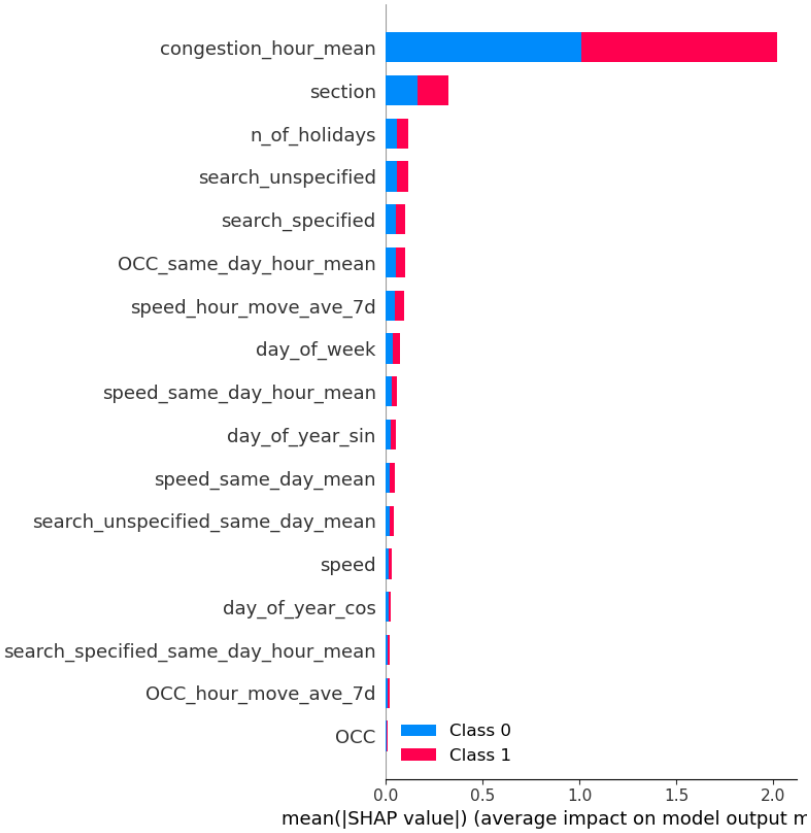
予測：区間毎・全期間のis_congestionの合計(上り)

予測：区間毎・全期間のis_congestionの合計(下り)

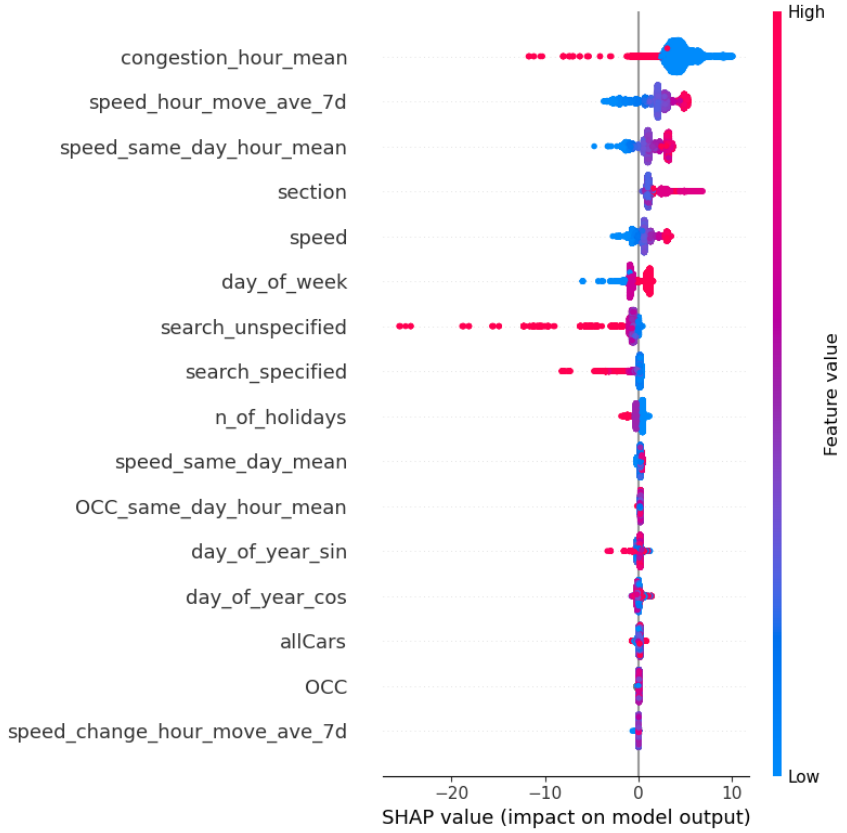
結果（4）

SHAPによる特徴量の解析：

二値分類モデル：



回帰モデル：

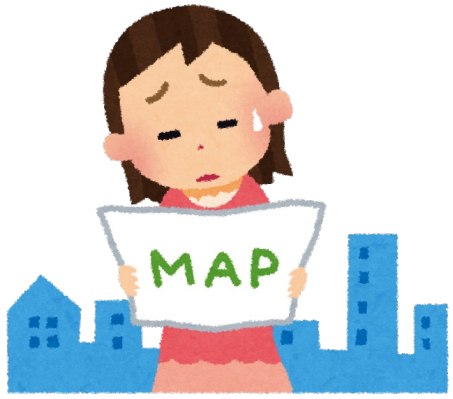


結果（4）

結果のまとめ：

- 作成したモデルでは、渋滞発生のタイミングについては、
  おおむね傾向を捉えることができた。
  一方渋滞が発生する場所については、
  その分布をうまく捉えることができなかった。

- 予測に用いた特徴量では、
  is_congestionの同時刻平均値、speedの同時刻7日平均値、区間
  などが予測に有効であった。



精度向上のための施策案

- 地理情報オープンデータから緯度・経度に紐づいた高度情報を得て、
  区間の勾配を説明変数に加える。

  → 速度変化をより正確に予測できると考えられる。

- 過去の交通規制情報を参照することで、
  学習データから事故・悪天候・天災などによる渋滞の情報を取り除く。

  → 突発的な要因で発生した渋滞をモデルに学習させないことで、
  より汎化性能が高い予測モデルが構築できると考えられる。

EOF