YOUBIKE 2.0

借車人數預測

シェアサイクル利用状況予測

科目：統計学習と深層学習

- GitHub Link

AGENDA


01
FRAME THE PROBLEM
02
DATA ACQUISITION
03
EXPLORE THE DATA
04
DATA PREPARING
05
MODEL SELECTION
06
MODEL ADJUSTMENT
07
CONCLUSION

01 FRAME THE PROBLEM

❏問題

　最近、学内の交流サイトでYouBikeのスポットに借りられる自転車がないか、または返せる場所がないという状況をよく見かけます。例えば、社会科学館から正門に向かう途中に9つのスポットがあっても、借りられる自転車が一台もないという報告があります。この状況は、シェアサイクルが設置された目的に反しています。従って、効率的な自転車調達の重要性がより際立っています。

01 FRAME THE PROBLEM

❏目標

　大学キャンパス内のシェアサイクルであるYoubike 2.0システムの効率化を支援するために、各スポットの未来の利用状況を予測し、運営者の車輌調達最適化という課題に役立てることを目指します。

01 FRAME THE PROBLEM

❑ 分析目標

私たちは、Youbike（台湾の公共シェアサイクルシステム）の利用状況に影響を与える要因が非常に複雑であると考えています。スポット内の残りの自転車の数、休祝日であるかどうか、当時の気温や降雨状況などが含まれ、ユーザーの利用意欲に影響する可能性があります。さらに、台湾大学内の不法駐輪撤去制度の存在は、キャンパス全体の自転車生態系において非常に重要な要素です。スポットの場所の近くに不正駐車された自転車が撤去されているかどうかも、Youbikeの貸し出し状況に影響を与える可能性があります。

02 DATA ACQUISITION

シェアサイクル・自転車撤去の関連情報

(1) 市政府の公開情報プラットフォームから大学エリアの各Youbike 2.0 スポットの運営データを15分ごとに収集します。

(2) 学内の自転車管理システムのから、15分ごとに撤去される自転車と自転車移置場のデータを収集します。

(3) 移置場の管理者に申請して、公開されていない期間（2019年から2021年10月末まで）のデータを取得します。

02 DATA ACQUISITION

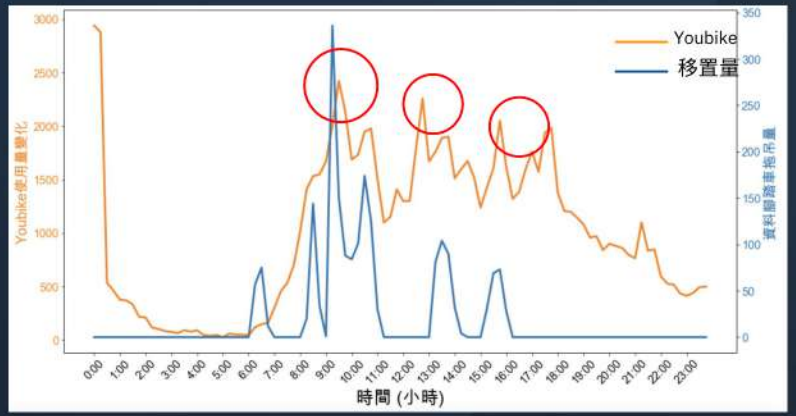
天気情報

　中央気象局のウェブサイトから、台湾大学の観測ステーションのデータを取得します。収集されたデータには気温や降水などの情報が含まれており、1時間ごとに1つのサンプルで記録されています。

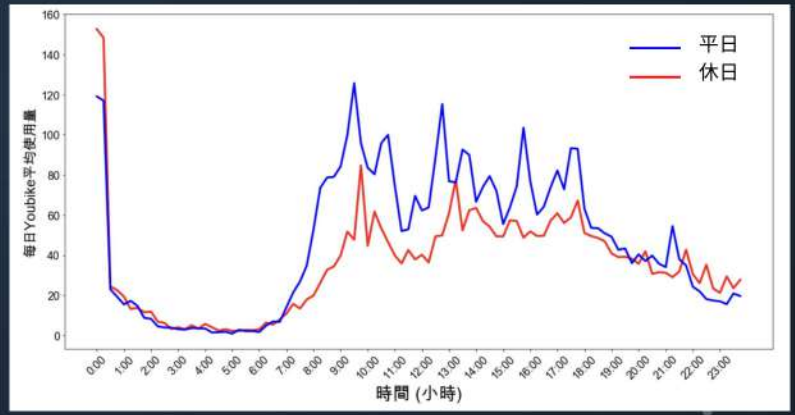
觀測時間	溫度 °C	天氣	風向	風力 (級)	陣風 (級)	能見度 (公里)	相對濕度 (%)	海平面氣壓 (百帕)	當日累積雨量(毫米)	日照時數
12/28 17:10	17.5		東北東	2	-	-	86	-	0.0	-
12/28 17:00	17.6		東北	2	3	-	86	-	0.0	-
12/28 16:50	17.7		東北	1	-	-	85	-	0.0	-
12/28 16:40	17.8		東北	2	-	-	84	-	0.0	-
12/28 16:30	17.9		北北東	1	-	-	84	-	0.0	-
12/28 16:20	18.0		東北	1	-	-	83	-	0.0	-
12/28 16:10	18.2		東北東	1	-	-	82	-	0.0	-
12/28 16:00	18.3		東北東	2	4	-	82	-	0.0	-

03 EXPLORE THE DATA

- Youbike利用量與移置量の比較



- 毎日平均利用量 (平日と休日の比較)



04 DATA PREPARING

❑データクリーニング、特徴選択

収集されたYoubikeデータから、各スポットの設置台数上限（tot）、残りの台数（sbi）、残りの駐輪スペース数（bemp）を取得できます。

各スポットの変動状況（リアルタイムの残り台数）を考慮し、以下の特徴を追加して各スポットの自転車台数の変化を予測します：

- スポットの残りの自転車数の遅延項（sbi_pastN）
- 残りの空きスペース数の遅延項（bemp_pastN）
- 未来N分後の残りの自転車数（sbi_N）
- 未来N分以内にそのスポットから借りられる自転車の淨差（rent_N）
-

N=15、30、45、60分など、4つのケースを考慮しています。

04 DATA PREPARING

❑データクリーニング、特徴選択

学内の移置システムのデータでのエリア分割はYoubikeの位置データと一致していないため、52つのスポット（place）とより広範囲のエリア（region）に人力で標記しました。これにより、各スポットの移置車両数（pCount_XXX）と各エリアの移置車両数（rCount_XXX）を計算できます。

また、極端な気温、降水状況はユーザーの利用意向に影響する可能性があるため、我々はその時間帯の気温（temp ）と降水量（precp）を特徴として使用しました。また、元のデータでは、降水量が0より大きく、かつ0.1mm未満の場合はTと記録されています。そこで、これらのデータを0.1mmとして標記して分析を行いました。

04 DATA PREPARING

❑データクリーニング、特徴選択

データを統合した後、次にカテゴリ変数をOne Hot Encodingで処理します。カテゴリ変数は時間帯、スポット名（sna）、曜日（weekday）、スポットが所在する移置地点（place）、スポットが所在する移置エリア（region）などが含まれます。

最後に、出現頻度が極めて低い変数を省略します。もし一つの変数が104回以下しか出現しない場合、これは2時間（52の駅に乗じて）にしか現れていないことを示していますので、これらの変数を削除します。

04 DATA PREPARING

❑ 処理したデータの例（カテゴリ変数はOne Hot Encoding前の形式で示す）

date	tot	sbi	countPlace	countRegion	bemp	sbi_pastN	bemp_pastN	pCount_地點	rCount_區域
2021/11/23	19	11	0.0	0.0	8	0.0	19.0	10	120
2021/11/23	10	7	0.0	0.0	3	7.0	3.0	15	140

temp	precp	time_時間	sna_地點	weekday	place_地點	region_區域	sbi_N	rent_N
15.5	0.1	16:00	大一女舍北側	1	研一、 大一女舍	A	10	0
15.5	0.1	12:15	社會系館南側	2	社會系館	D	15	5

サンプル数 **106,862**　特徴量 **267**　時間 **2021年11/23 ～2021年12/15**

04 DATA PREPARING

一、特徴量

sbi_15sbi_30、sbi_45、sbi_60、rent_15、rent_30、rent_45、rent_60、date 以外の全ての変数

二、目的変数

使用する8つの目的変数は、sbi_15、sbi_30、sbi_45、sbi_60、rent_15、rent_30、rent_45、rent_60 で、それぞれスポットの15、30、45、60分後の残り自転車数、および未来15、30、45、60分以内の借り出し自転車数を示しています。

三、Train / Test データセット

私たちはトレーニングセットとテストセットの比率を7:3と定義し、sklearn.model_selection の train_test_split を使用してデータを分割しました。その結果、74,803件のデータを含むトレーニングセットと32,059件のデータを含むテストセットが得られました。

四、モデル評価

モデルの予測結果を直感的に評価できるために、私たちはMAE（平均絶対誤差）を使用し、モデルの予測効果を示す。一方、モデルのトレーニング過程では、予測誤差が大きい場合には修正できるように、RMSE（平均二乗誤差の平方根）を訓練誤差の指標として使用します。

05 MODEL SELECTION

- 使用したモデル


OLS


Cat-boost


Decision Tree


Random Forest


MLP

❑全てのモデルの最善の **testing MAE**

	sbi15	sbi30	sbi45	sbi60	rent15	rent30	rent45	rent60
OLS	1.05	1.50	1.81	2.07	1.05	1.50	1.81	2.07
catboost	1.01	1.41	1.67	1.88	1.01	1.40	1.67	1.88
Decision Tree	1.05	1.51	1.82	2.05	1.04	1.46	1.8	2.03
Random Forest	1.04	1.48	1.79	1.99	1.02	1.44	1.75	1.98
MLP	1.04	1.38	1.62	1.78	1.06	1.43	1.62	1.78

05 MODEL SELECTION

❑予測結果に対して 最も重要な１０つの特徴量（MLP以外の４つのモデル）

	OLS	CatBoost	Decision Tree	Random Forest
1	sbi	sbi	sbi	sbi
2	bemp	temp	sbi_past15	sbi_past15
3	sbi_past30	bemp	tot	sbi_past30
4	bemp_past30	sbi_past60	sbi_past30	bemp
5	sbi_past60	sbi_past15	bemp_past15	tot
6	bemp_past60	sbi_past30	bemp	sbi_past60
7	sbi_past15	sbi_past45	sbi_past45	temp
8	bemp_past15	sna_臺大大一女北側	bemp_past60	sbi_past45
9	sna_臺大檔案展示館	bemp_past60	bemp_past30	bemp_past45
10	place_土木系館	region_B	temp	bemp_past60

06 MODEL ADJUSTMENT

❑ MLP Model

「Linear MLP」は「CatBoost」よりも優れた予測効果がありますので、私たちはこのモデルを選択して Fine Tuning を行いました。このステップでは、以下の方法を試しました。

- モデルの隠れ層の数を調整
- モデルのノードの数を調整
- 異なるOptimizerの使用
- ミニバッチの数を調整
- 学習率の調整

06 MODEL ADJUSTMENT

- **Random Forest Model**

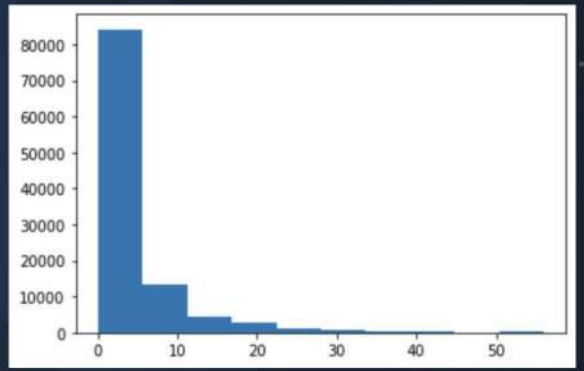
元々は、 grid search を使用して、複数のパラメータセットでモデルを調整し、最良の予測効果を得るための最適なパラメータを見つけるようにしていました。しかし、この方法では、効果の悪いパラメータを試すために多くの時間がかかる可能性があり、また試行するパラメータが不足しているため、最適な結果を見つけることができないことがありました。

この問題を改善するために、fine-tuningの段階では random searchを採用しました。

06 MODEL ADJUSTMENT

❑ データ操作

データの分布と予測結果を見ると、スポットの利用可能な自転車数は主に0〜5台です。そのため、モデルが予測誤差を抑えるため、全ての予測が0の状況がありました。



この問題を解決するために、私たちはSMOTE（Synthetic Minority Over-sampling Technique）を使用して、データの不均衡をより均等にしようと試みました。

この方法を使用すると、訓練データセットのデータ数が1,676,370に増加しました。右の表のように、SMOTEを使用するとMAE（平均絶対誤差）が以前よりも高くなります。しかし、データの分布がより均等になり、モデルが全部のデータを0と予測しなくなったためです。

	sbi15	sbi30	sbi45	sbi60	rent15	rent30	rent45	rent60
MLP	1.03	1.42	1.66	1.81	1.05	1.40	1.64	1.78
Random Forest	1.01	1.39	1.63	1.87	0.98	1.34	1.58	1.82

07 CONCLUSION

- モデルの予測結果により、未来30分間のYoubikeの利用状況を予測し、MAEを1.5以下に抑えました。
- この予測結果を活用することで、Youbikeの運営者は効果的に将来の駐輪場の利用状況を予測し、現在の自転車の調達に関する課題を改善できると考えます。