MOTIV

레스토랑 수익 예측을 위한 기계학습 모델 연구: 터키의 데이터를 중심으로

ICT 공학부 202104168 김준석

MOTIV

목차 및 서론 설명

서론

자영업자 수 감소
음식업 종사자 수 증가
짧은 생존 기간의 음식업
코로나로 인한 영향

데이터 전처리

수익 영향 변수 확인
개장일로 나이 계산
변수간 관계성 확인
Train 셋 / Test 셋 분리
예측력 강한 변수 선정
StandardScaler & PCA

데이터 시각화

P-변수를 제외한 변수
데이터간 상관관계 파악

모델 연구

RMSE 성능 평가 지표
랜덤 포레스트
트리 개수에 따른 오차율

MOTIV

Id	Open Dat	City	City Grou	Type	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20
0	07/17/19	칼stanbul	Big Cities	IL	4	5	4	4	2	2	5	4	5	5	3	5	5	1	2	2	2	4	5	
1	02/14/20	Ankara	Big Cities	FC	4	5	4	4	1	2	5	5	5	5	1	5	5	0	0	0	0	0	3	
2	03/09/201	Diyarbak킈	Other	IL	2	4	2	5	2	3	5	5	5	5	2	5	5	0	0	0	0	0	1	
3	02/02/201	Tokat	Other	IL	6	4.5	6	6	4	4	10	8	10	10	8	10	7.5	6	4	9	3	12	20	
4	05/09/20	Gaziantep	Other	IL	3	4	3	4	2	2	5	5	5	5	2	5	5	2	1	2	1	4	2	
5	02/12/201	Ankara	Big Cities	FC	6	6	4.5	7.5	8	10	10	8	8	8	10	8	6	0	0	0	0	0	5	
6	10/11/201	칼stanbul	Big Cities	IL	2	3	4	4	1	5	5	5	5	5	2	5	5	3	4	4	3	4	2	
7	06/21/201	칼stanbul	Big Cities	IL	4	5	4	5	2	3	5	4	4	4	4	3	4	0	0	0	0	0	3	
8	08/28/201	Afyonkara	Other	IL	1	1	4	4	1	2	1	5	5	5	1	5	5	1	1	2	1	4	1	
9	11/16/201	Edirne	Other	IL	6	4.5	6	7.5	6	4	10	10	10	10	2	10	7.5	0	0	0	0	0	25	
10	08/09/201	Kocaeli	Other	FC	9	6	6	6	4	4	10	8	10	10	8	10	7.5	0	0	0	0	0	25	1
11	05/22/201	칼stanbul	Big Cities	IL	2	4	4	4	2															
12	02/28/201	Ankara	Big Cities	IL	2	2	4	4	2															
13	10/16/201	칼stanbul	Big Cities	FC	4	5	4	4	1															
14	12/29/201	Bursa	Other	FC	2	2	4	4	1															
15	02/07/201	칼stanbul	Big Cities	IL	12	7.5	6	6	2															
16	01/07/20	칼stanbul	Big Cities	FC	3	5	4	4	2															
17	11/08/20	칼stanbul	Big Cities	FC	2	4	4	5	1															
18	04/21/201	칼zmir	Big Cities	IL	4	5	4	3	1															
19	08/16/201	Sakarya	Other	IL	2	4	4	4	2															
20	08/25/201	Elaz캇헭	Other	IL	3	4	4	4	2															
21	01/25/201	칼stanbul	Big Cities	FC	5	5	4	4	2															
22	07/01/20	Kayseri	Other	FC	9	6	6	6	4															
23	06/03/20	Sakarya	Other	FC	2	4	4	4	2															
24	09/20/20	칼stanbul	Big Cities	IL	5	5	3	5	2															
25	12/23/201	Eski힐ehir	Other	FC	4	4	5	5	2															

AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ
P23	P24	P25	P26	P27	P28	P29	P30	P31	P32	P33	P34	P35	P36	P37	revenue
3	1	1	1	4	2	3	5	3	4	5	5	4	3	4	5653753
2	0	0	0	0	3	3	0	0	0	0	0	0	0	0	6923131
1	0	0	0	0	1	3	0	0	0	0	0	0	0	0	2055379
10	2	2	2.5	2.5	2.5	7.5	25	12	10	6	18	12	12	6	2675511
1	2	3	3	5	1	3	5	1	3	2	3	4	3	3	4316715
5	0	0	0	0	7.5	5	0	0	0	0	0	0	0	0	5017319
1	5	4	4	5	1	3	4	5	2	2	3	5	4	4	5166635
2	0	0	0	0	3	2	0	0	0	0	0	0	0	0	4491607
1	4	4	4	2	2	3	4	5	5	3	4	5	4	5	4952497
10	0	0	0	0	5	2.5	0	0	0	0	0	0	0	0	5444227
20	0	0	0	0	10	2.5	0	0	0	0	0	0	0	0	3745135
3	5	2	3	5	3	3	5	5	4	2	3	4	4	2	5161370
2	0	0	0	0	1	3	0	0	0	0	0	0	0	0	1734634
1	0	0	0	0	3	3	0	0	0	0	0	0	0	0	4807746
2	0	0	0	0	2	3	0	0	0	0	0	0	0	0	1999097
5	8	8	10	2.5	7.5	7.5	5	15	20	2	12	3	16	4	3218918
1	0	0	0	0	2	2	0	0	0	0	0	0	0	0	2E+07
1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	8213524
2	5	3	3	3	2	3	3	5	5	4	4	4	3	2	5337526
1	0	0	0	0	2	3	0	0	0	0	0	0	0	0	2021934
5	2	4	1	5	1	3	5	1	2	2	4	5	5	4	5525735
1	0	0	0	0	3	1	0	0	0	0	0	0	0	0	1149870
10	0	0	0	0	7.5	7.5	0	0	0	0	0	0	0	0	3956086
1	0	0	0	0	2	3	0	0	0	0	0	0	0	0	2999068
3	2	2	2	5	3	2	5	5	4	4	4	4	5	2	8904084
1	0	0	0	0	2	3	0	0	0	0	0	0	0	0	3778621

Train Data Set

MOTIV

데이터 전처리

Open Days

개장일로부터 나이 계산

변수의 데이터 타입을 변경

Type

레스토랑의 유형

범주형 데이터이므로 삭제

Big Cities & Other

수치형 데이터로 변환

더미로 만든 가변수로 변환

Random Forest Classifier

많은 변수 중 가장 예측력이 강한 변수가 무엇인지 파악

블랙박스 모형이기 대문에 설명변수와 반응변수의 설명력을 확보하기 어려움

MOTIV

데이터 전처리

Train 셋 / Test 셋 분리

Train Set
7

Test Set
3

예측을 위해 원데이터의 Train Set과
Test Set을 7:3으로 지정한다.

StandardScaler & PCA

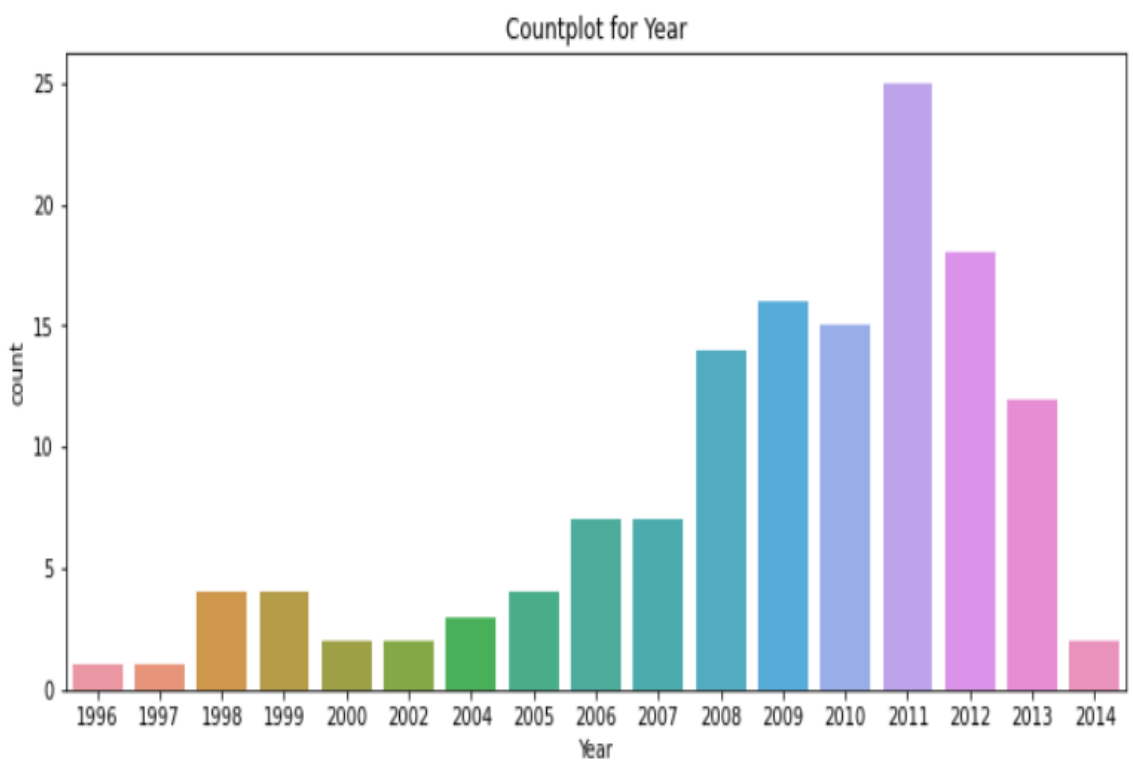
표준화를 통해
0의 평균
1의 표준편차

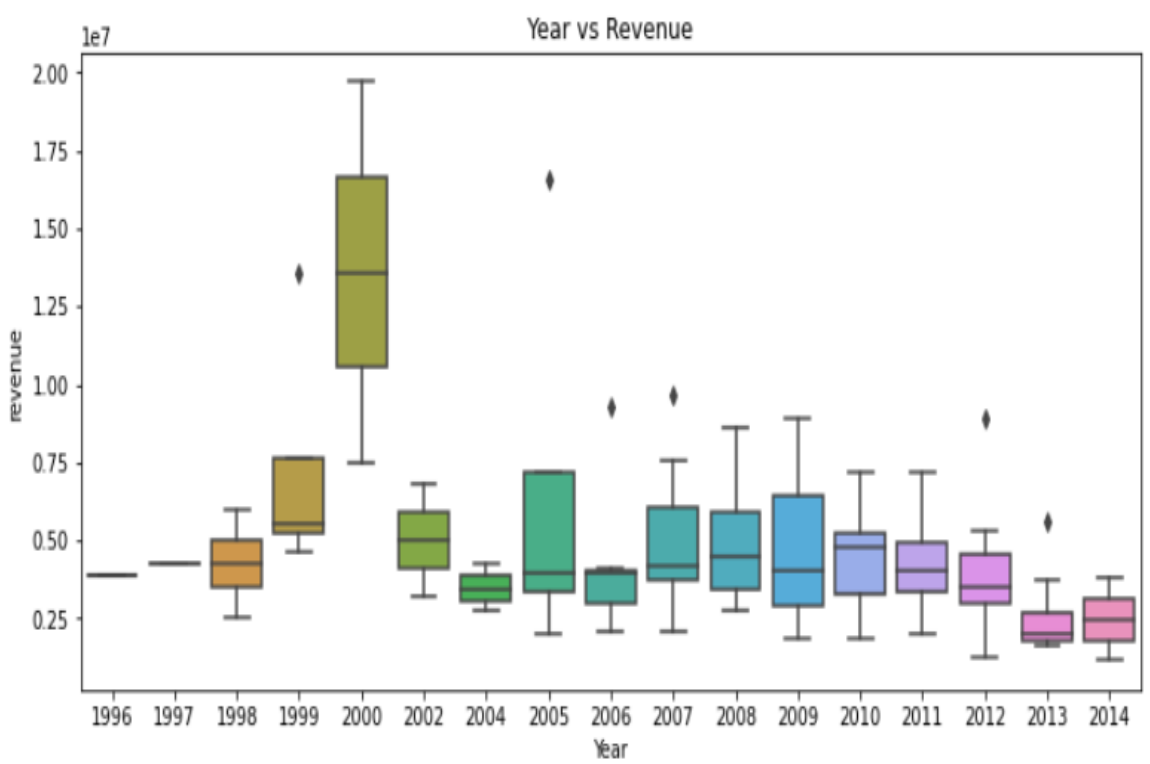
특성 행렬의
차원을 축소

선형적으로 구분되는 고차원 공간에서
주성분으로 투영된 결과를 반환하게 만든다.

MOTIV

데이터 시각화

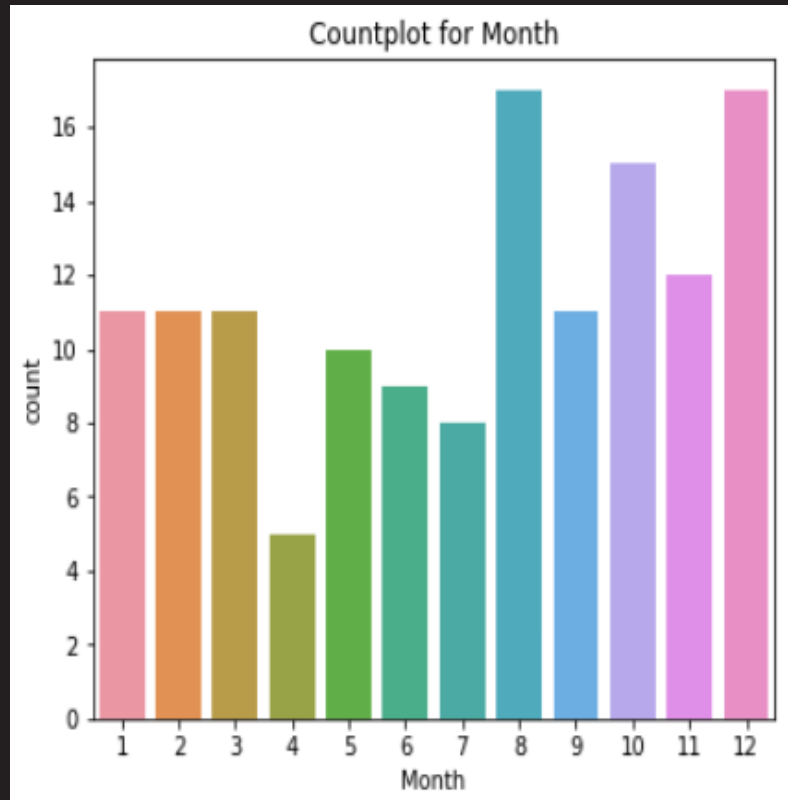


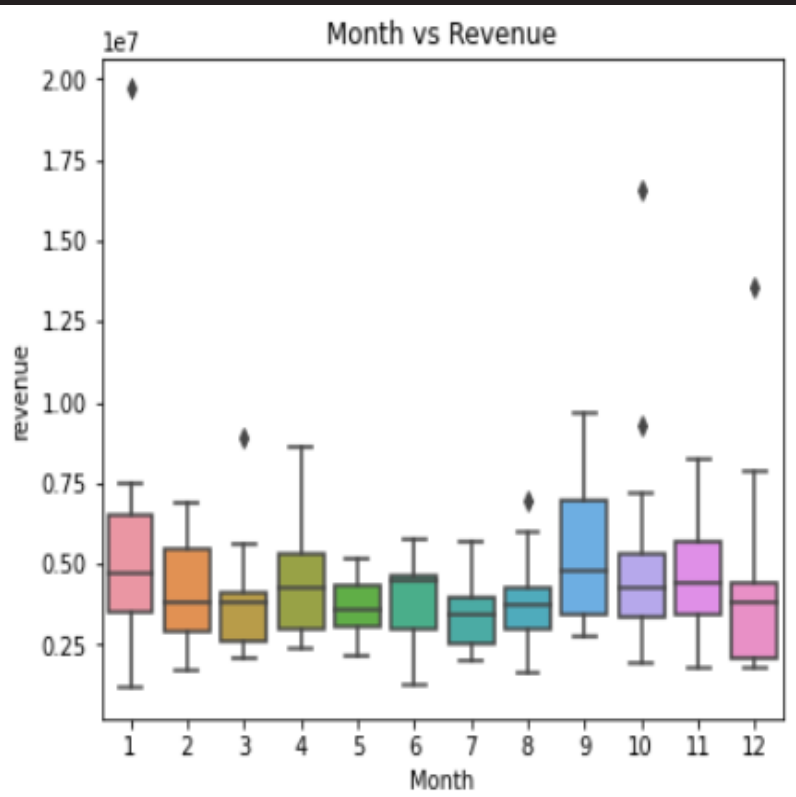


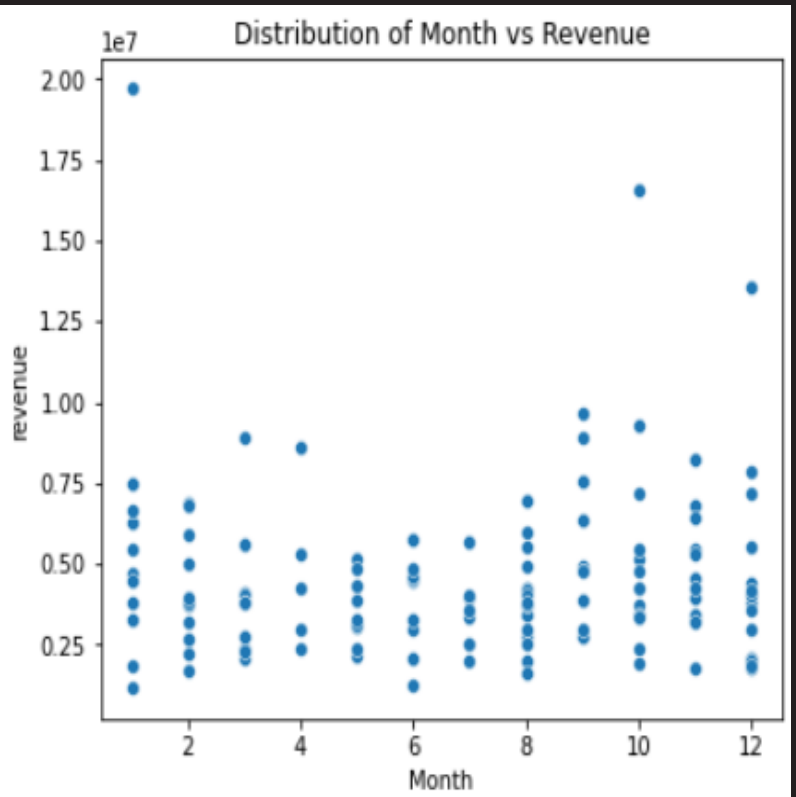
연별 레스토랑 오픈 횟수와 년도와 수익의 상관관계

MOTIV

데이터 시각화



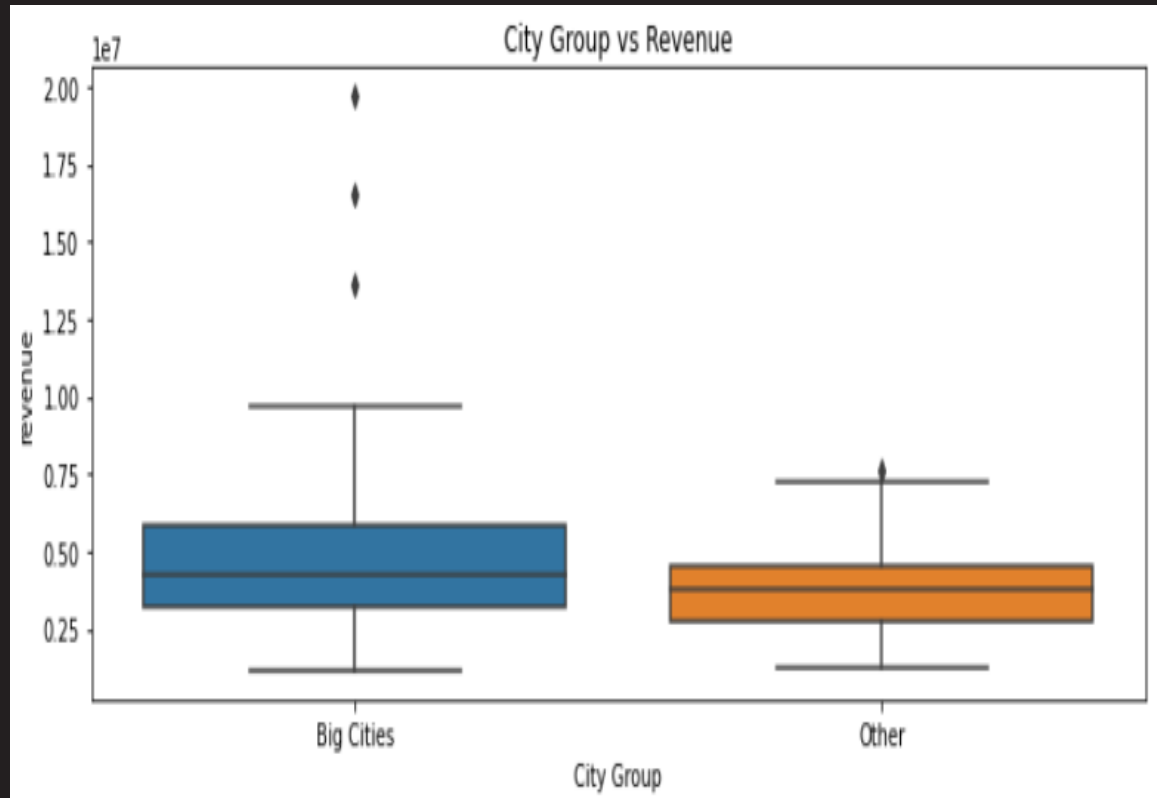


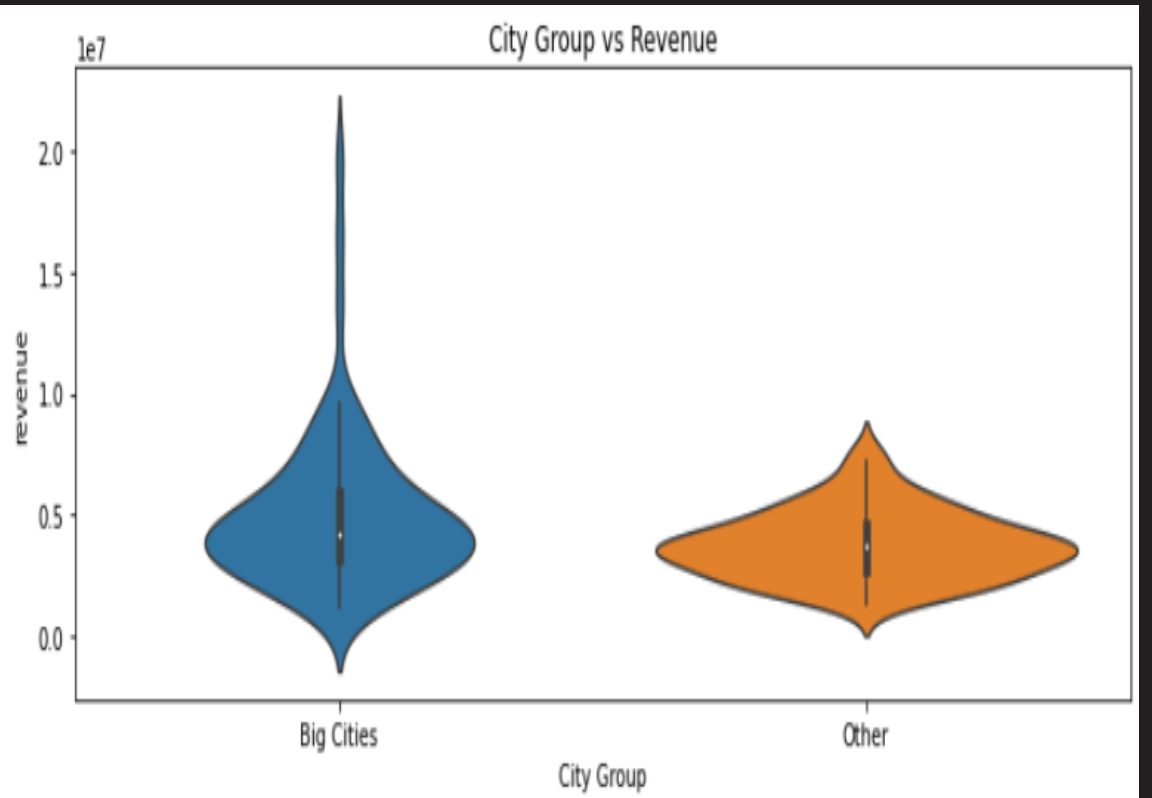


월별 레스토랑 오픈 횟수와 월과 수익의 상관관계

MOTIV

데이터 시각화





도시 유형과 수익의 상관관계

MOTIV

데이터 시각화

도시 유형과 년도, 수익의 상관관계

데이터 시각화

01	Open Days	0.087800	14	P29	0.030229	27	P33	0.013801
02	P19	0.049030	15	P12	0.027642	28	P31	0.013405
03	P11	0.048616	16	P21	0.027468	29	P14	0.013398
04	P28	0.048073	17	Big Cities	0.021086	30	P34	0.012988
05	P6	0.047423	18	Others	0.019282	31	P16	0.012948
06	P2	0.046847	19	P13	0.018407	32	P15	0.012927
07	P20	0.044270	20	P9	0.018130	33	P25	0.012733
08	P22	0.041101	21	P7	0.017888	34	P37	0.012634
09	P5	0.038940	22	P10	0.017070	35	P24	0.011999
10	P23	0.035039	23	P32	0.015280	36	P30	0.011465
11	P4	0.033633	24	P27	0.015114	37	P36	0.010757
12	P8	0.032433	25	P26	0.014458	38	P35	0.010173
13	P3	0.031882	26	P17	0.014213	39	P18	0.009416

Random Forest Classifier

모델 연구

RMSE(Root Mean Square Error)

RMSE

RMSE는 회귀 모델 성능 평가 지표

오차를 제곱해서 평균한 값의 제곱근

값이 작을수록 정밀도가 높은 평가 지표를 사용한다.

모델별 RMSE Value

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

RMSE가 작을수록 성능 향상

Model	RMSE Value
Lasso Regression	1710660.3807717038
Ridge Regression	1710666.6082206795
Random Forest	1644100.1807500264

모델 연구

예측값과 실제값 그리고 오차율

예측값	실제값	오차율(%)	예측값	실제값	오차율(%)
3714896.756	2267425.0	63.837691	4705920.552	7495092.0	37.213305
4486442.688	4952497.0	9.410492	5180054.496	3956086.0	30.938875
4347229.640	3354383.0	29.714498	5105501.692	2156098.0	136.793582
4504207.520	3871344.0	16.347385	5220805.288	3752885.0	39.114449
4118482.116	2732645.0	50.714129	4712343.156	8904084.0	47.076609

결정 트리 개수에 따른 평균 오차율

결정 트리 개수	평균오차율(%)
1	58.984111557494074
10	62.5592519485054
100	43.33222656084231
200	39.91403430274294
300	40.439834917879175

MOTIV

레스토랑 수익 예측을 위한 기계학습 모델 연구: 터키의 데이터를 중심으로

ICT 공학부 202104168 김준석

감사합니다 :)