🍙 한식당, 어디에 창업해야 성공할 수 있을까?

💡 Let’s think step by step!

팀원
권시은 기석광 김동현
김현지 이충원 원정인

목차

1. 주제

- **주제 : 서울시에서 한식당을 어디에 창업하면 성공 확률이 높을까?**

1) 사용한 데이터 셋

- '서울 열린 데이터 광장'의 서울시 상권분석 데이터
- 행정동 별 추정 매출, 소득 소비, 직장 인구, 상주 인구, 유동 인구, 점포 수, 임대료

2) 분석 목표

- 서울시의 다양한 데이터를 활용하여
  한식 전문점의 창업 성공 가능성을 예측하는 머신러닝 모델 구현

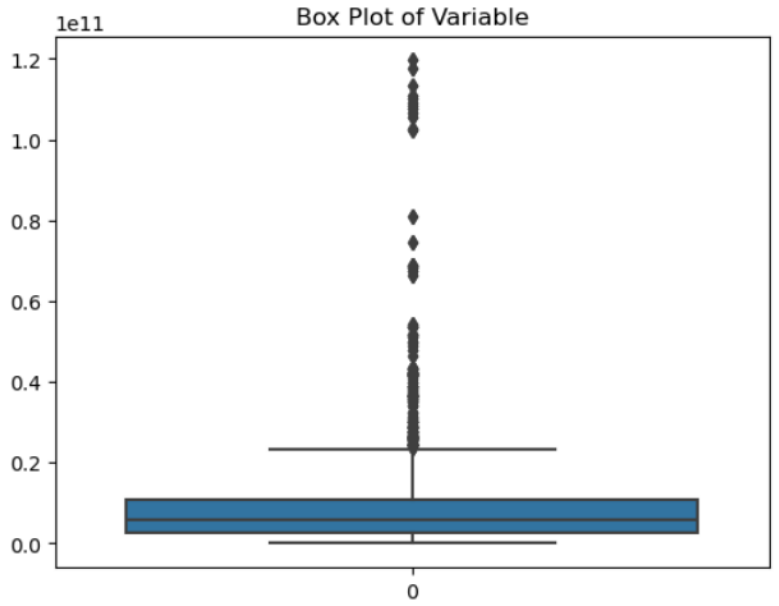
2. EDA

1. 이상치, 결측치 확인

a. Y값(당월 매출 금액)의 분산이 너무 크고 이상치가 많음

–〉 평균보다는 중간 값을 기준으로 하여 확인하기로 판단

b. y값이 정규 분포와 얼마나 가까운지 확인



▲ 매출액 값의 boxplot 그래프

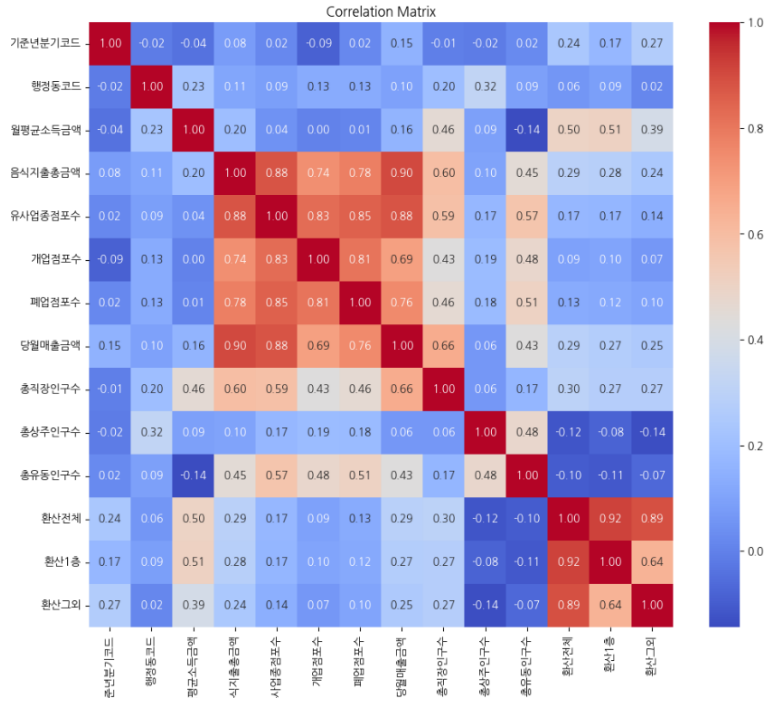
2. EDA

2. y값에 영향을 미치는 Feature Importance 확인

a. 변수 간 상관관계 분석

→ 매출 금액과 상관관계가 높은 feature: 유사 업종 점포 수, 개업 점포 수, 폐업 점포 수, 당월 매출 금액, 직장인구 수, 유동인구 수

b. y값과의 상관관계 분석 (랜덤 포레스트 모델을 통한 변수 중요도 확인)



▲ 변수 간 상관관계 분석 히트맵

```
당월매출금액      1.000000
음식지출총금액     0.903567
유사업종점포수     0.882786
폐업점포수        0.756771
개업점포수        0.691676
총직장인구수       0.656963
총유동인구수       0.429380
환산전체          0.286917
환산1층          0.268334
환산그외          0.249555
월평균소득금액      0.156896
기준년분기코드      0.154017
행정동코드         0.096096
총상주인구수        0.063227
Name: 당월매출금액, dtype: float64
```

▲ y값과의 상관관계 분석

3. 전처리

3. 분석에 사용할 컬럼 선택

a. 상관관계가 너무 작은 컬럼 제외

b. 유사한 컬럼은 묶어서 사용

c. 다중공선성 확인

데이터 Merging	행정동별 소득소비, 직장인구, 상주인구, 유동인구, 점포 수, 임대료 데이터셋 merge
불필요한 속성 제거	서비스 업종 코드 = 한식음식점
파생변수 추가	'점포당_당월평균판매건수' '점포당_당월평균매출금액(Predicted)' '점포당_당월평균매출금액' '행정동별_월평균소득금액' '행정동별_평당임대료'

4. 모델링

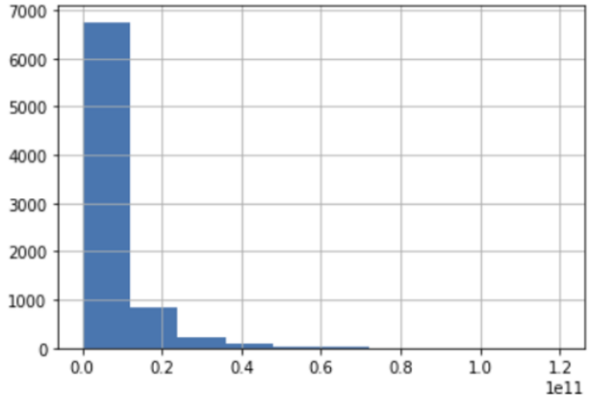
a. 단순 선형 회귀

- 최소제곱법(OLS)을 활용한 단순 선형 회귀 수행
  - 수치형 & 시계열 데이터를 예측하고자 하는 문제

OLS Regression Results

Dep. Variable:	y	R-squared (uncentered):	0.882
Model:	OLS	Adj. R-squared (uncentered):	0.882
Method:	Least Squares	F-statistic:	4568.
Date:	Tue, 28 May 2024	Prob (F-statistic):	0.00
Time:	14:43:16	Log-Likelihood:	-2786.3
No. Observations:	7955	AIC:	5599.
Df Residuals:	7942	BIC:	5689.
Df Model:	13		
Covariance Type:	nonrobust		

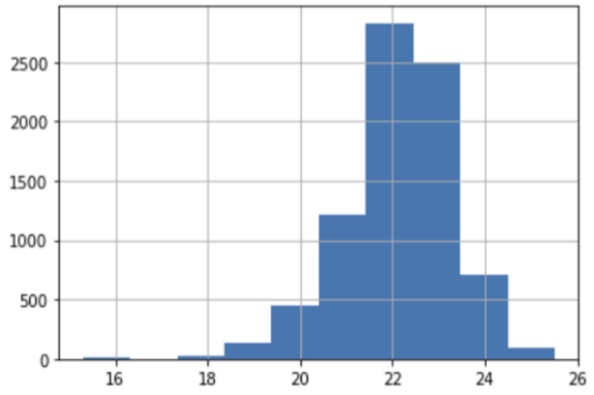
	coef	std err	t	P>|t|	[0.025	0.975]
기준년분기코드	0.0967	0.004	23.075	0.000	0.088	0.105
행정동코드	0.0073	0.004	1.729	0.084	-0.001	0.016
월평균소득금액	-0.0363	0.006	-6.533	0.000	-0.047	-0.025
음식지출총금액	0.4794	0.009	52.821	0.000	0.462	0.497
유사업종점포수	0.4368	0.012	36.771	0.000	0.413	0.460
개업점포수	-0.0866	0.008	-11.415	0.000	-0.101	-0.072
폐업점포수	0.0137	0.008	1.697	0.090	-0.002	0.029
총직장인구수	0.1612	0.006	27.632	0.000	0.150	0.173
총상주인구수	-0.0429	0.005	-8.813	0.000	-0.052	-0.033
총유동인구수	-0.0140	0.006	-2.473	0.013	-0.025	-0.003
환산전체	-747.4231	468.386	-1.596	0.111	-1665.584	170.737



RMSLE : 0.297

RMSE : 0.359

MAE : 0.199



RMSLE : 0.033

RMSE : 0.745

MAE : 0.522

▲ log 변환 전후 비교

4. 모델링

b. 릿지 회귀

- 릿지 회귀는 선형 회귀에 L2 규제를 추가한 모델
  L2 규제는 상대적으로 큰 회귀 계수 값의 예측 영향도를 감소시키기 위해서 회귀 계수값을 더 작게 만드는 규제 모델

-〉 실제 매출액보다 더 크게 예측을 하면 발생할 피해(손실)이 더 악영향을 준다고 판단
즉, 매출액을 실제보다 더 적게 예측하는 건 괜찮을 수 있지만, 더 크게 예측하는 것을 방지하고자 L2 규제를 사용

```
Mean Squared Error (MSE): 0.1290560580583785 5
Root Mean Squared Error (RMSE): 0.3592437307154831
Mean Absolute Error (MAE): 0.1988516412509639 7
R-squared (R2): 0.8787525835914736
```

4. 모델링

- 모델링 결과

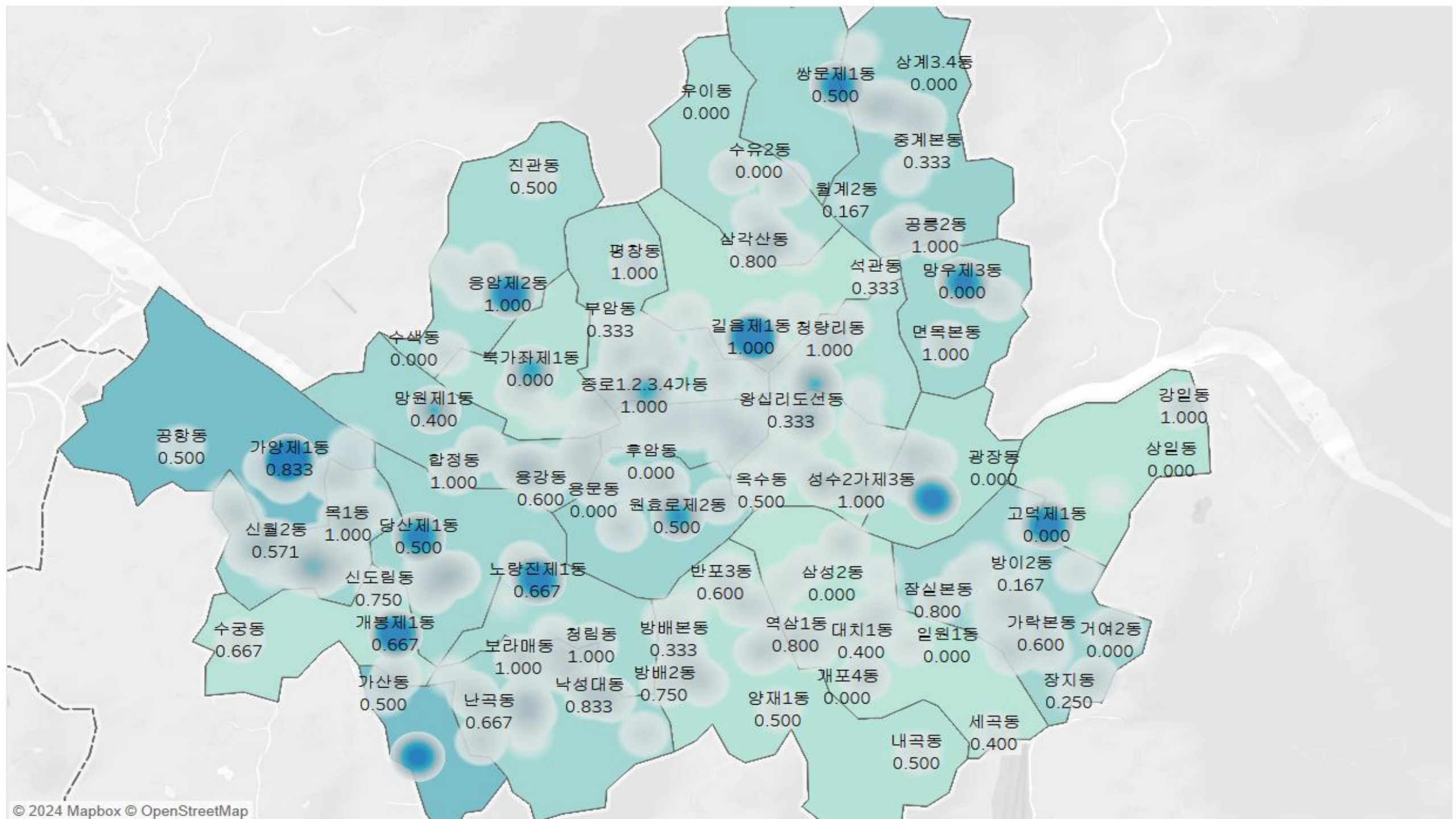
Model	Random_Forest (Standard Scaler)	Random_Forest (Min-Max Scaler)	Catboost	Xgboost	LightGBM
MSE (Mean Squared Error)	559229972.4899927	3446315098.2524085	1544859294.641698	1598069308.2946541	62361346210.091064
R-squared	0.97913	0.97943	0.97666	0.97271	0.97619

-〉 최종 모델로 랜덤 포레스트 선택

5. 결론

1. 예상 매출액

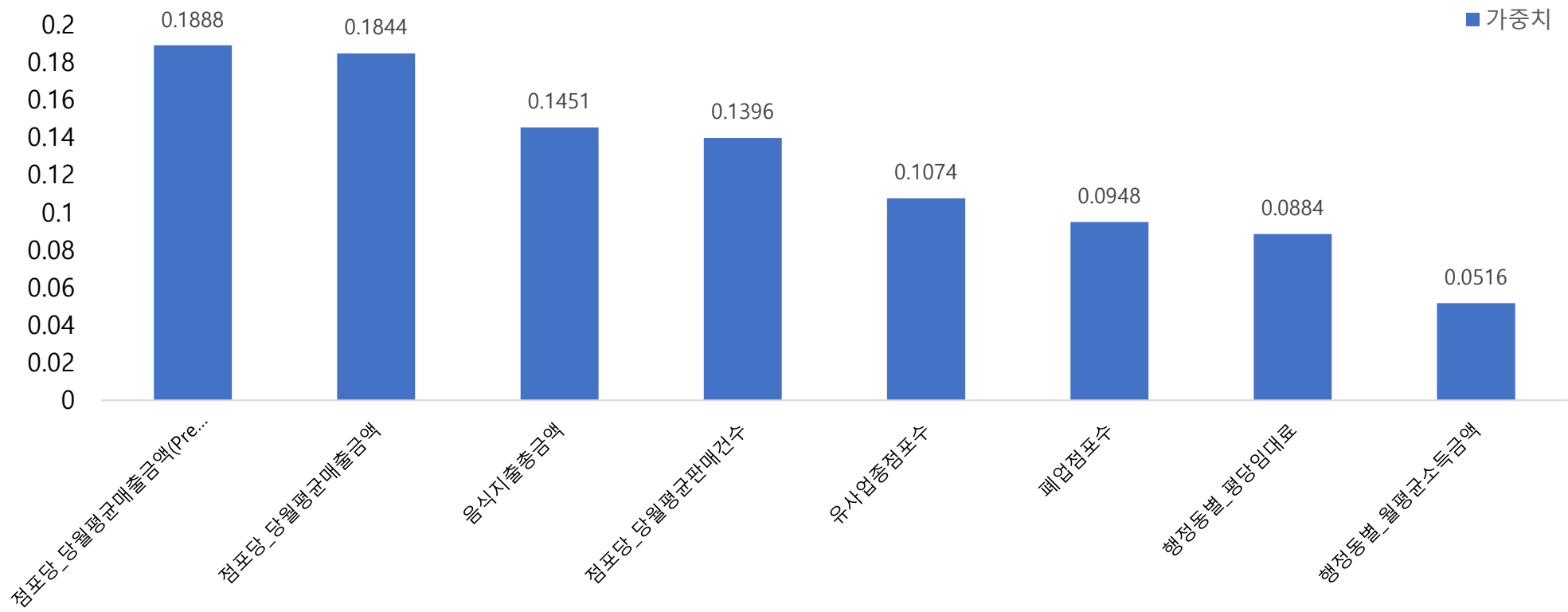
- 도출된 모델로 행정동 별 예상 매출액 예측
- Inverse_transform을 통해 스케일링 된 값을 원본 값으로 변환하여 예측 매출액 구하기
- 행정동 별 median값을 예측 매출액과 비교하여 창업 성공 여부를 성공(1)/실패(0) 로 처리
- 창업 성공 컬럼의 평균값 계산하여 각 동 별 창업 성공률 도출


유이동
0.000
쌍문제1동
0.500
상계3.4동
0.000
수유2동
0.000
중계본동
0.333
월계2동
0.167
진관동
0.500
삼각산동
0.800
공릉2동
1.000
평창동
1.000
석관동
0.333
망우제3동
0.000
응암제2동
1.000
부암동
0.333
길음제1동
1.000
청량리동
1.000
면목본동
1.000
수색동
0.000
북가좌제1동
0.000
종로1.2.3.4가동
1.000
왕십리도선동
0.333
망원제1동
0.400
강일동
1.000
공항동
0.500
가양제1동
0.833
합정동
1.000
후암동
0.000
광장동
0.000
상일동
0.000
용강동
0.600
용문동
0.000
옥수동
0.500
성수2가제3동
1.000
원효로제2동
0.500
고덕제1동
0.000
목1동
1.000
신월2동
0.571
당산제1동
0.500
방이2동
0.167
노량진제1동
0.667
반포3동
0.600
삼성2동
0.000
잠실본동
0.800
신도림동
0.750
수궁동
0.667
개봉제1동
0.667
청림동
1.000
방배본동
0.333
역삼1동
0.800
대치1동
0.400
일원1동
0.000
가락본동
0.600
거여2동
0.000
보라매동
1.000
방배2동
0.750
개포4동
0.000
장지동
0.250
가산동
0.500
낙성대동
0.833
난곡동
0.667
양재1동
0.500
세곡동
0.400
내곡동
0.500
© 2024 Mapbox © OpenStreetMap

5. 결론

2. 창업 성공 예상 점수 계산

- 점포당_당월평균매출금액(Predicted)에 대한 파생변수들의 상관관계를 분석하여 가중치 설정
- 창업 성공 예상 점수 컬럼 추가



5. 결론

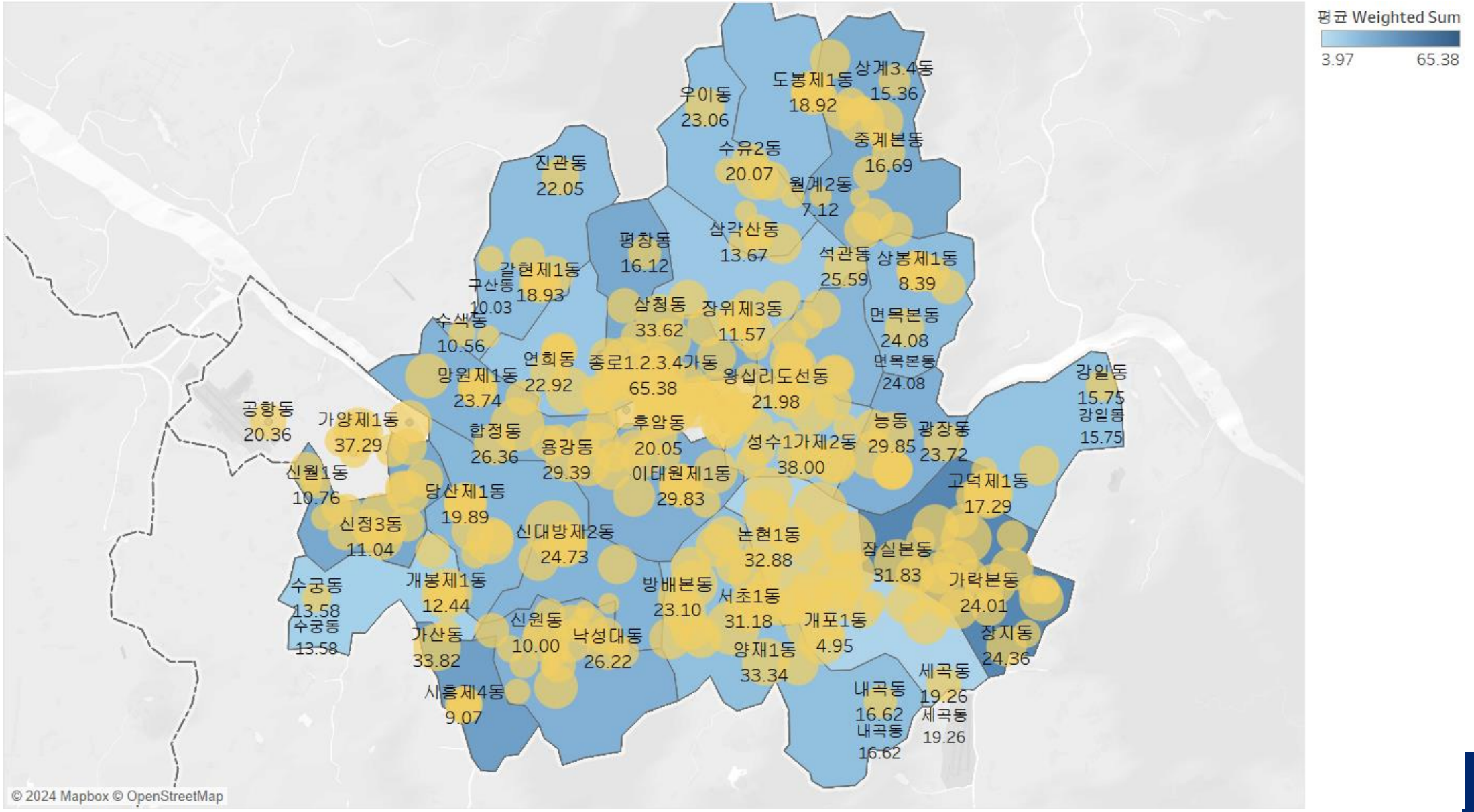
2. 창업 성공 예상 점수 계산

```python
def calculate_pred_proba(df):
    scaler = MinMaxScaler()
    columns = list(weights.keys())
    df_scaled = df[columns].copy()
    df_scaled[columns] = scaler.fit_transform(df_scaled[columns])
    weighted_columns = []

    for column in columns:
        weighted_column = df_scaled[column] * weights[column]
        weighted_columns.append(weighted_column)

    pred_proba = sum(weighted_columns) * 100
    return pred_proba.tolist()
```

5. 결론 – 동 별 창업 성공 점수 분포



6. 개선방향

리뷰 데이터 분석

고객 리뷰 데이터를 수집하고 감성 분석을 수행하여
보다 심층적인 인사이트를 도출할 계획

웹 서비스 구현

예비 창업자들이 직접 입력한 정보를 바탕으로
예상 매출액, 추천 객단가, 창업 점수 등을 제공하는
웹 서비스를 구축

지속적 개선

이번 프로젝트를 통해 얻은 경험과 피드백을 바탕으로
모델을 지속적으로 개선하고 발전시켜 나갈 것

금융 데이터 연계

향후에는 금융 데이터와의 연계를 통해 보다 종합적인
창업 지원 서비스를 제공할 수 있을 것으로 기대

감사합니다