

# 오픈소스 소프트웨어 프로젝트 최종 발표

---

공공데이터를 활용하여  
성공적 창업을 돋는  
상권 분석 모델 개발

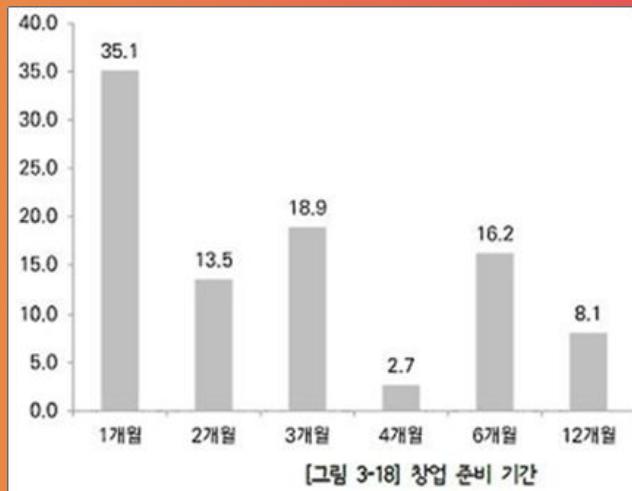
팀 3P

# 목차

1. 프로젝트 목표
2. 데이터 수집
3. 데이터 전처리
4. 모델링
  - XGBoost 모델을 이용한 데이터 분석
5. 분석 결과 예시 (자치구-상권-서비스업종 순)
  - a. 강남구 - 삼성동 코엑스 - 커피
  - b. 강남구 - 삼성동 코엑스 - 시계 및 귀금속
  - c. 중구 - 명동 - 한식음식점
6. 요약 및 의의

# 1. 프로젝트 목표

"공공 데이터를 활용하여 창업 예정자들의 성공적 창업을 돋는 상권 분석 모델 개발"



과거의 데이터

# 1. 프로젝트 목표

과거 1년 전

...

1분기 전

다음 분기

유동인구 수

추정 매출

주변 아파트

시세. 평수

주변 집객시설

상주인구 수

직장인구 수



폐업률 예측

창업 생존 전략 도출

## 2. 데이터 수집



No.	출처 : 서울 열린데이터 광장 ( <a href="http://data.seoul.go.kr/">http://data.seoul.go.kr/</a> )
1	서울시 우리마을가게 상권분석서비스(상권-추정유동인구).csv
2	서울시 우리마을가게 상권분석서비스(상권-추정매출).csv
3	상권 내 업종별로 점포 수, 개업·폐업률, 개업·폐업 점포 수
4	서울시 우리마을가게 상권분석서비스(상권-아파트).csv
5	서울시 우리마을가게 상권분석서비스(상권-집객시설)
6	서울시 우리마을가게 상권분석서비스(상권-점포)
7	서울시 우리마을가게 상권분석서비스(상권-직장인구)

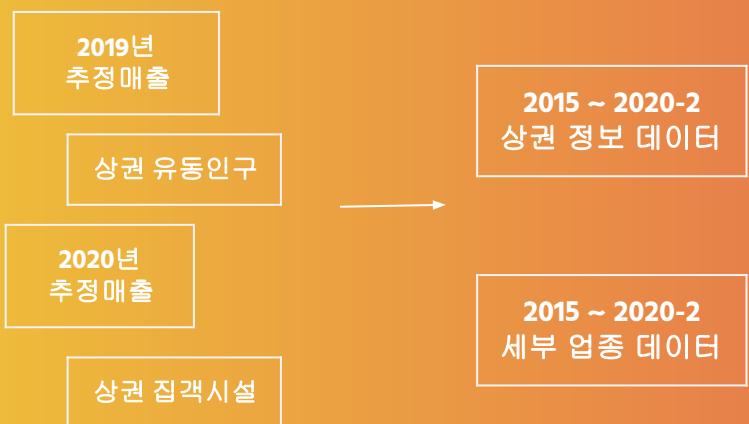
총 데이터 개수
( 32912 , 10 ) 약 33만개 데이터

### 'CC BY' 라이선스 보유 :

저작자가 자신의 저작물을 다른 이들이 자유롭게 쓸 수 있도록 미리 허락하는 라이선스로 출처 표기 후 자유로운 사용이 가능함

### 3. 데이터 전처리

#### (1) 데이터 통합



연도별, 항목별로 분리되어 있는 데이터들을  
상권 정보에 관한 데이터와 세부 업종에  
관한 데이터로 통합

#### (2) 결측치 처리

상권		상권	
1	→	[ 1,0,0 ]	
2	→	[ 0,1,0 ]	
3	→	[ 0,0,1 ]	
2018	2019	2020	
1	1	1	
2	?	2	
3	3	3	

##### \* One-Hot Encoding

상권 3은 상권 1 값의 3배라는 의미를 갖고  
있지 않다.

→ 이러한 관계성을 없애주기 위함

##### \* KNN Imputer

결측치를 해당 데이터와 유사한  
그룹의 관측값으로 대체함

### 3. 데이터 전처리

#### (3) 데이터 정규화

매출 건수	매출 건수
60,904	0.759007
59,855	0.739154
...	...

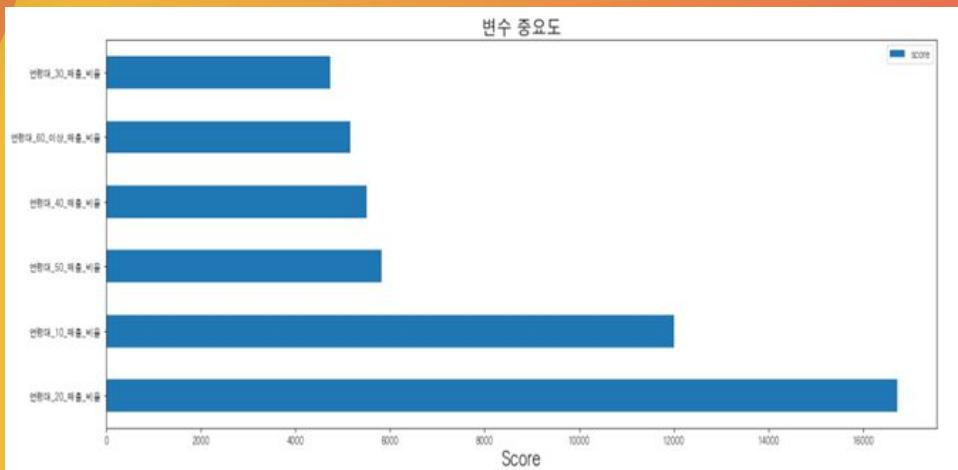
모델링 진행 시 데이터의 다양한 단위와 크기로 인해 모델의 성능이 좌우되지 않도록 '데이터 정규화'를 진행하여 각 항목의 단위와 크기의 영향 제거  
→ MIN-MAX Scaling 진행

#### (4) 시계열 데이터 생성



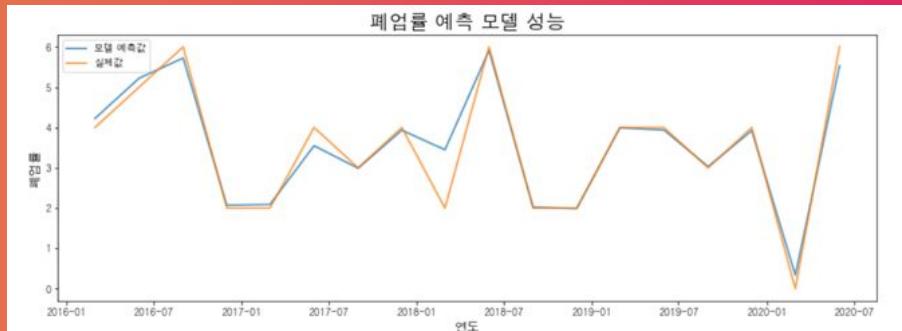
폐업률 예측 모델 구현을 위해 데이터를 시계열 형태로 변환  
→ 과거 1년 동안의 데이터로 다음 분기 폐업률 예측

## 4. 모델링 : Xgboost



〈XGBoost 모델 변수 중요도 시각화〉

- 주황선 : 실제 폐업률
- 파란선 : 과거 4분기 데이터로 예측한 미래 폐업률



〈XGBoost 모델 성능 시각화〉

- 후보로 선정된 3개의 모델 (RNN, 선형회귀, XGBoost) 중 XGBoost 모델이 실제 폐업률을 가장 잘 추적함
- XGBoost 모델의 경우 변수 중요도를 쉽게 도출하고 시각화할 수 있음
- XGBoost 모델을 최종 모델로 선정하고 이후의 모델링 진행

## 4. 모델링 : Xgboost

### (1) 폐업률 예측 모델

- 창업 예정자가 창업 위험도를 사전에 확인해볼 수 있도록 특정 업종의 상권 내 폐업률을 예측하는 모델 개발
- 특정 상권 내 업종의 과거 1년 동안의 데이터를 사용하여 해당 업종의 다음 분기 폐업률을 예측
- 전처리 과정에서 도출한 시계열 데이터 사용
- 모델 성능을 시각화하고 모델 예측 폐업률을 도출

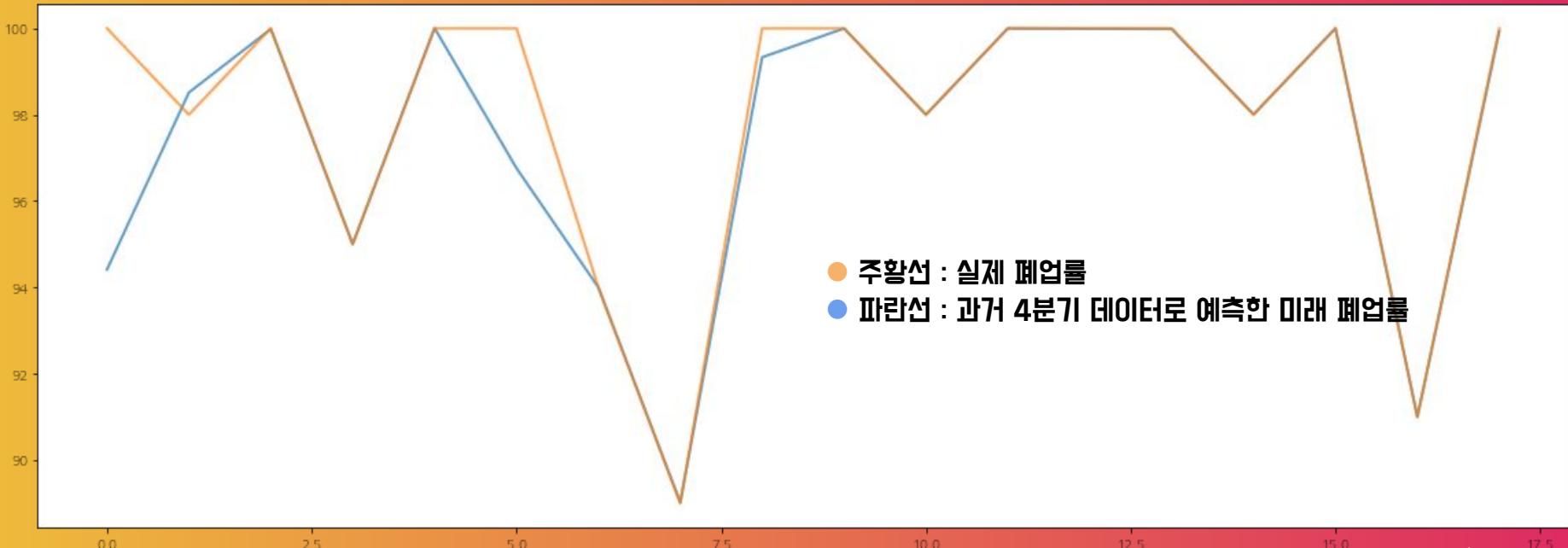
### (2) 창업 전략 분석 모델

- 창업 예정자에게 특정 업종이 상권 내에서 생존하는 데 필요한 정보를 제공하기 위해 모델링 진행
- '폐업률'과 반대되는 개념인 '생존율'을 기반으로 분석
- 직전 분기 특정 업종 데이터를 기반으로 다음 분기의 업종 생존율 분석
- 변수 카테고리별 중요도를 확인하기 위해 업종 매출 데이터를 '연령별 매출', '시간대별 매출', '성별 매출', '요일별 매출'로 구분하여 각각 모델링 진행
- 변수 중요도를 도출하고 시각화하여 창업 예정자로 하여금 창업 이후 생존에 필요한 요인이 무엇인지 파악할 수 있도록 함

## 5. 분석 결과 예시

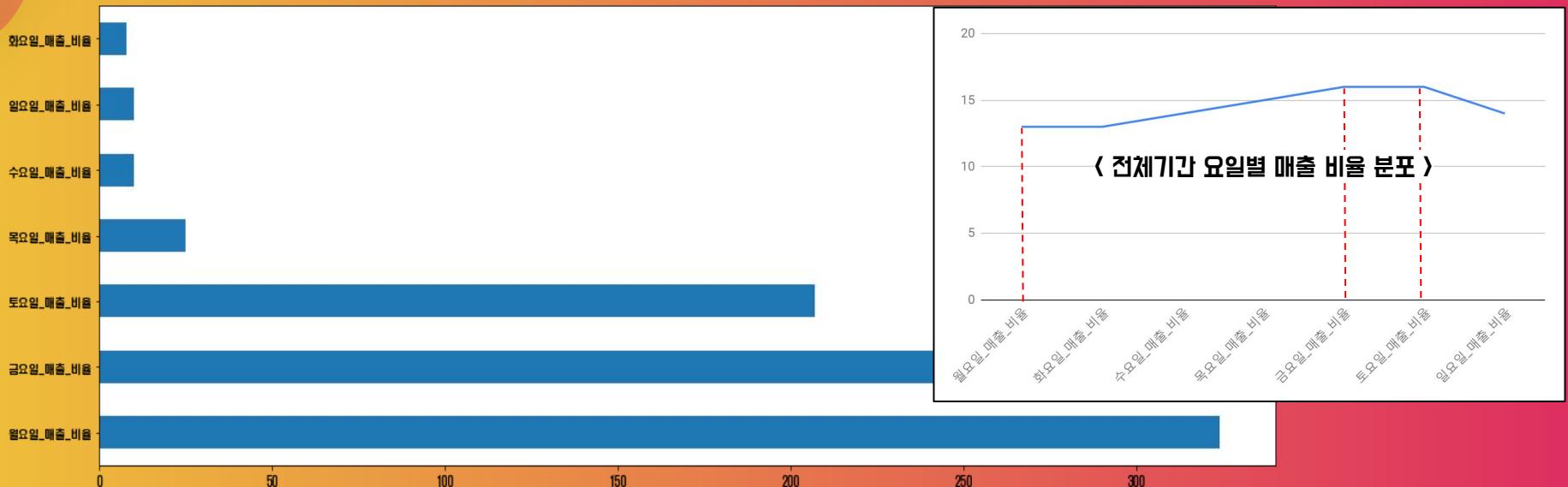
-삼성동 코엑스 상권의 시계&귀금속 업종-

폐업률 예측 모델 성능



## 5. 분석 결과 예시

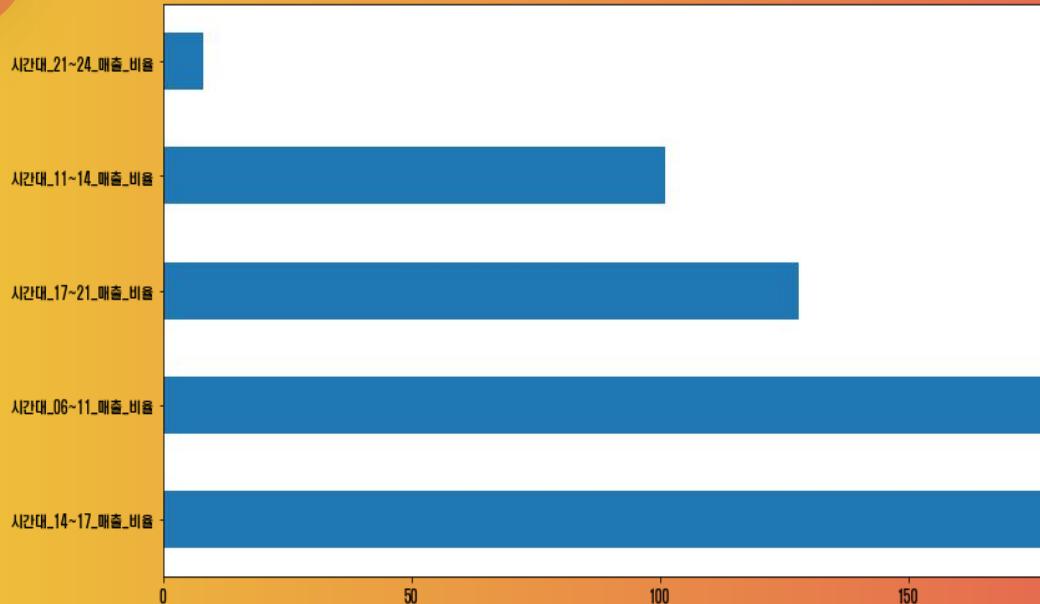
### -삼성동 코엑스 상권의 시계&귀금속 업종-



실제로, 시계&귀금속은 금요일, 토요일 가장 많이 팔린다.  
상대적으로 적게 팔리는 월요일에 많이 파는게 생존에 유리했고,  
그 다음으로 평소 많이 팔리는 요일인 금요일, 토요일에 많이 파는게 중요했다.

# 5. 분석 결과 예시

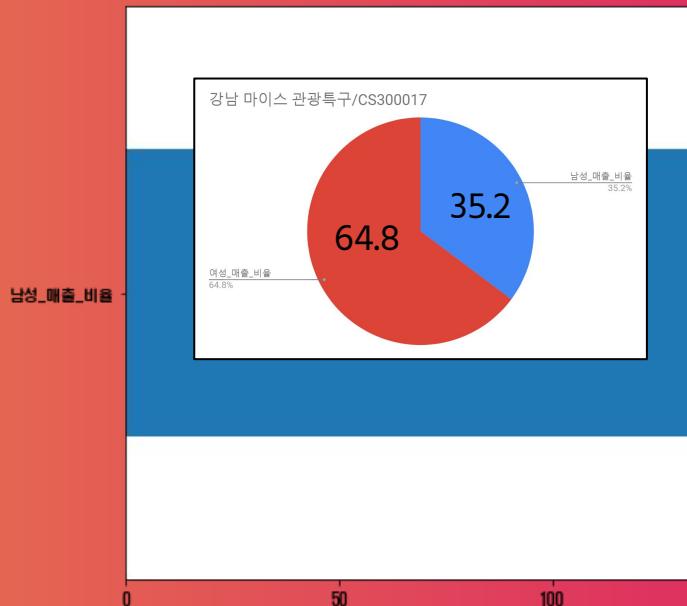
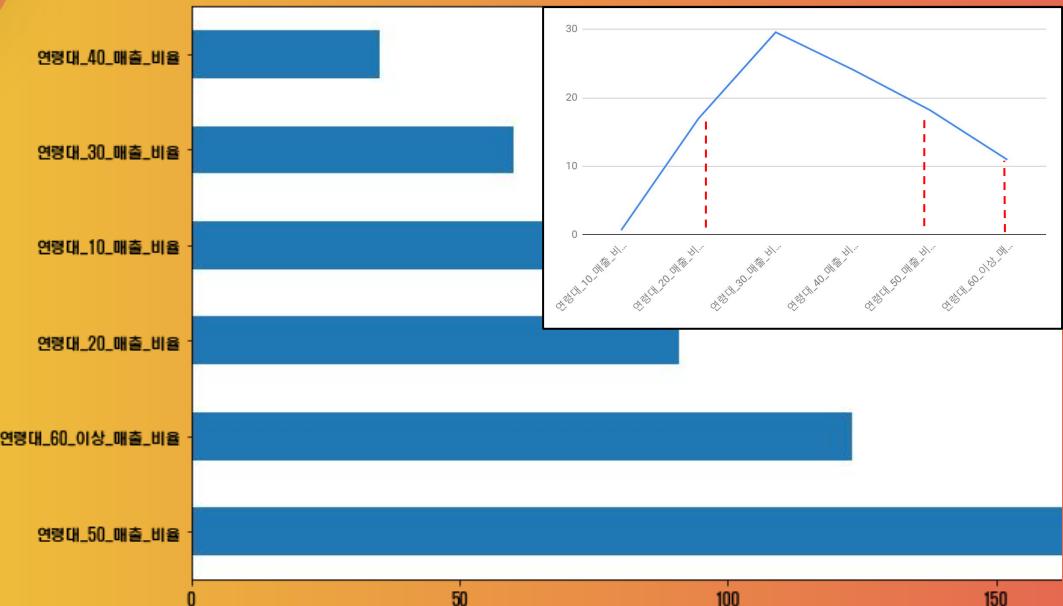
## -삼성동 코엑스 상권의 시계&귀금속 업종-



가장 잘 팔리는 오후 2시~5시에 잘 파는것이 생존에 중요했고,  
가장 덜 팔리는 오전 시간대에 잘 파는것이 생존에 유리한것으로 도출됨

# 5. 분석 결과 예시

## -삼성동 코엑스 상권의 시계&귀금속 업종-



매인 고객은 30대~40대이지만 생존에 중요한 연령층은 50~60대로 도출되었음.  
실제 고객의 65%는 여성이지만 남성 매출이 생존에 중요한 것으로 도출되었음.

## 5. 분석 결과 예시

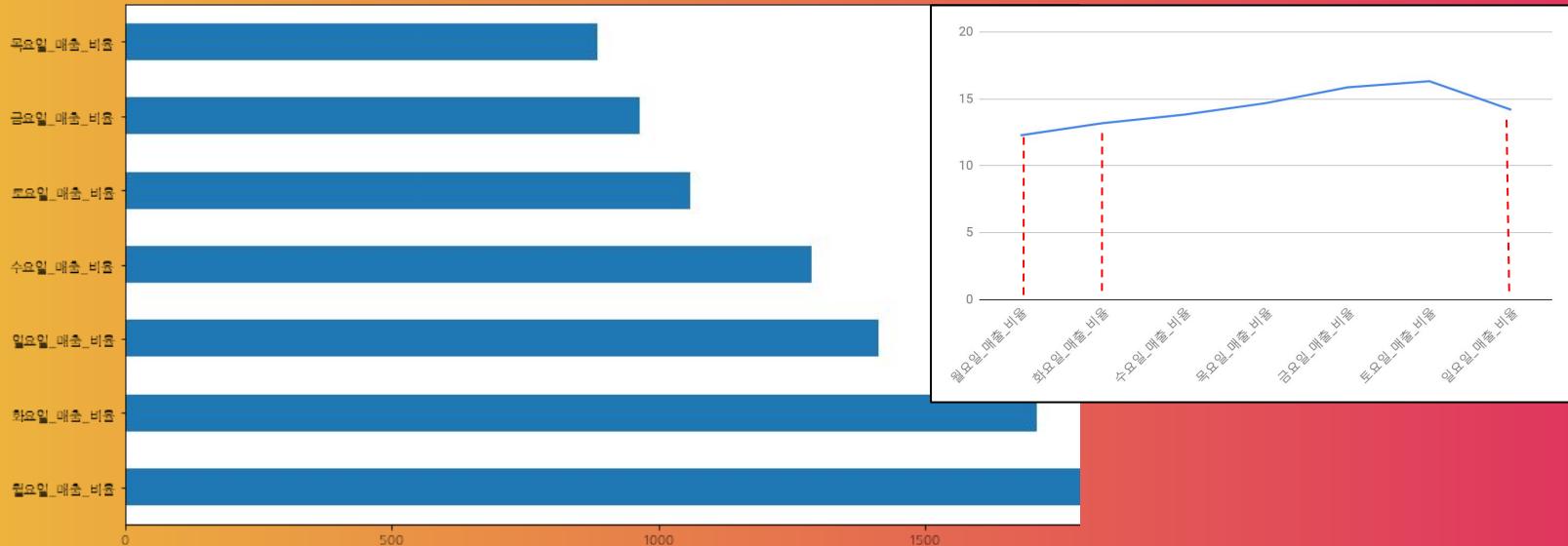
-삼성동 코엑스 상권의 커피 업종-

폐업률 예측 모델 성능



## 5. 분석 결과 예시

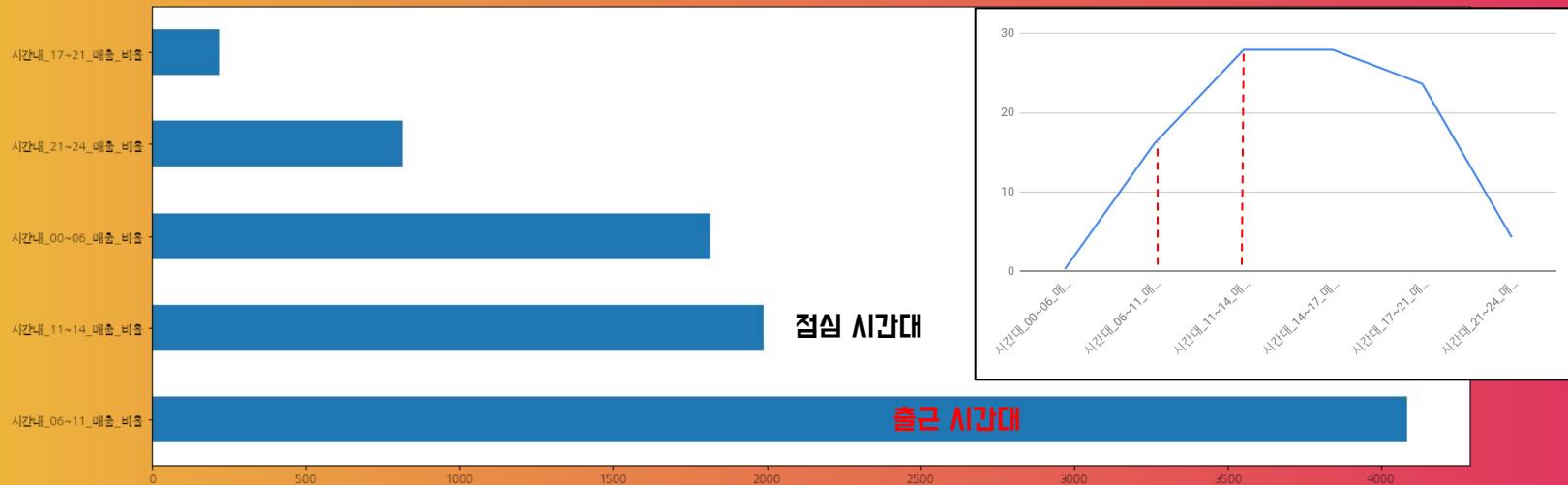
### -삼성동 코엑스 상권의 커피 업종-



커피는 요일별 매출량이 크게 차이가 나지 않지만 월요일, 화요일, 일요일이 상대적으로 더 적게 팔리는 경향이 있다. 이렇게 상대적으로 매출이 더 적은 요일에 많이 팔면 생존율이 올라갔다.

# 5. 분석 결과 예시

## -삼성동 코엑스 상권의 커피 업종-

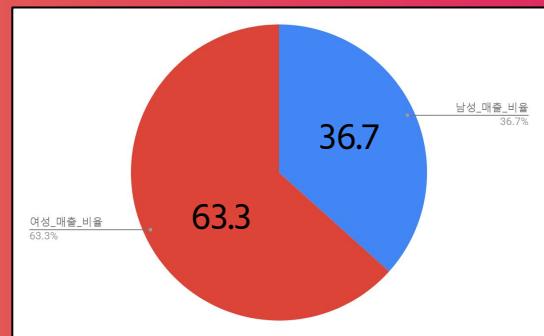
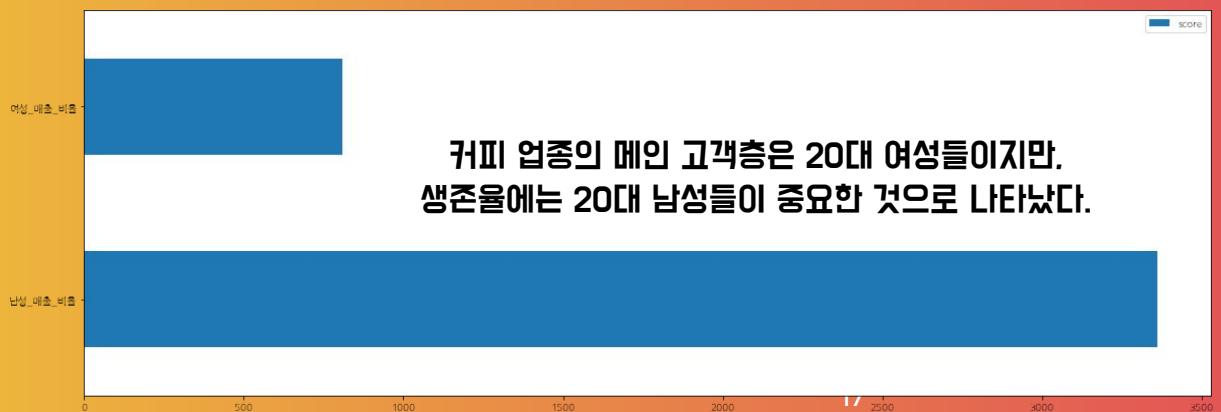
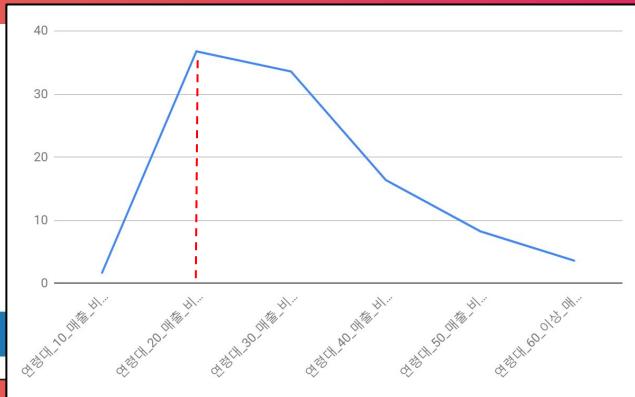
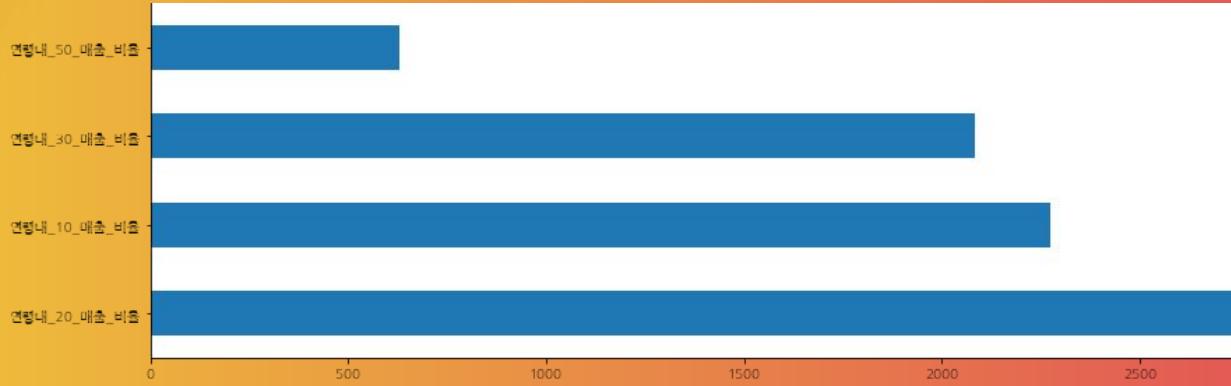


커피가 가장 많이 팔리는 시간대는 접심시간~오후 시간대 였음.

그러나 생존율에 가장 크게 영향을 미친 시간대는 출근시간대 였음.

# 5. 분석 결과 예시

## - 삼성동 코엑스 상권의 커피 업종-



# 5. 분석 결과 예시

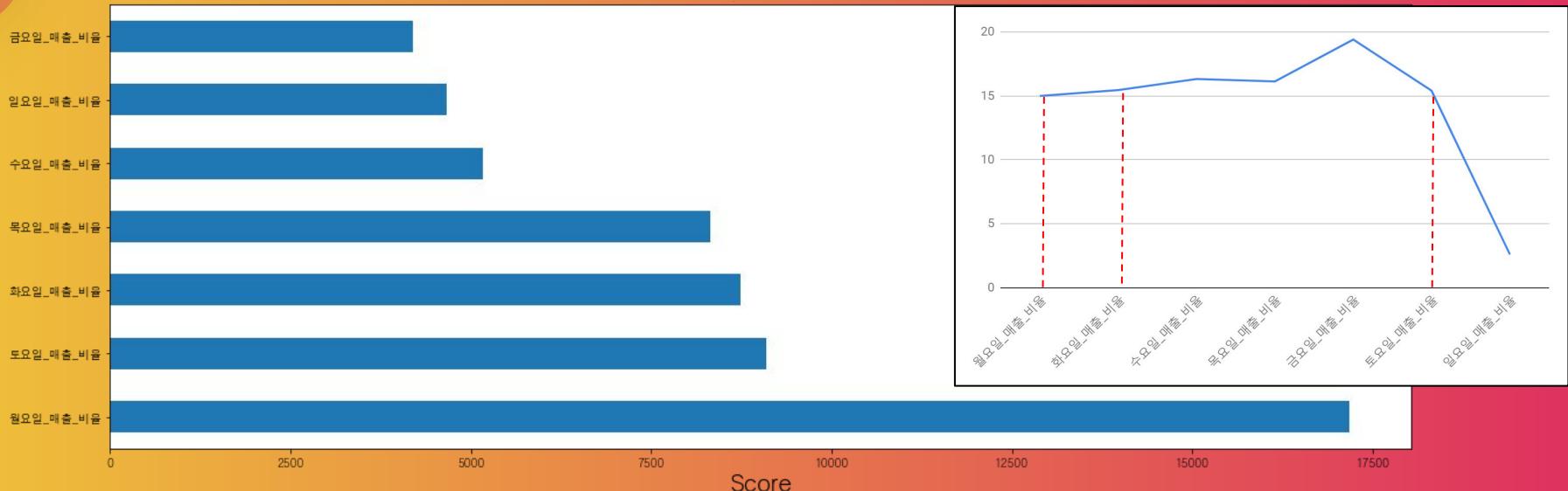
-명동의 한식음식점 업종-

폐업률 예측 모델 성능



## 5. 분석 결과 예시

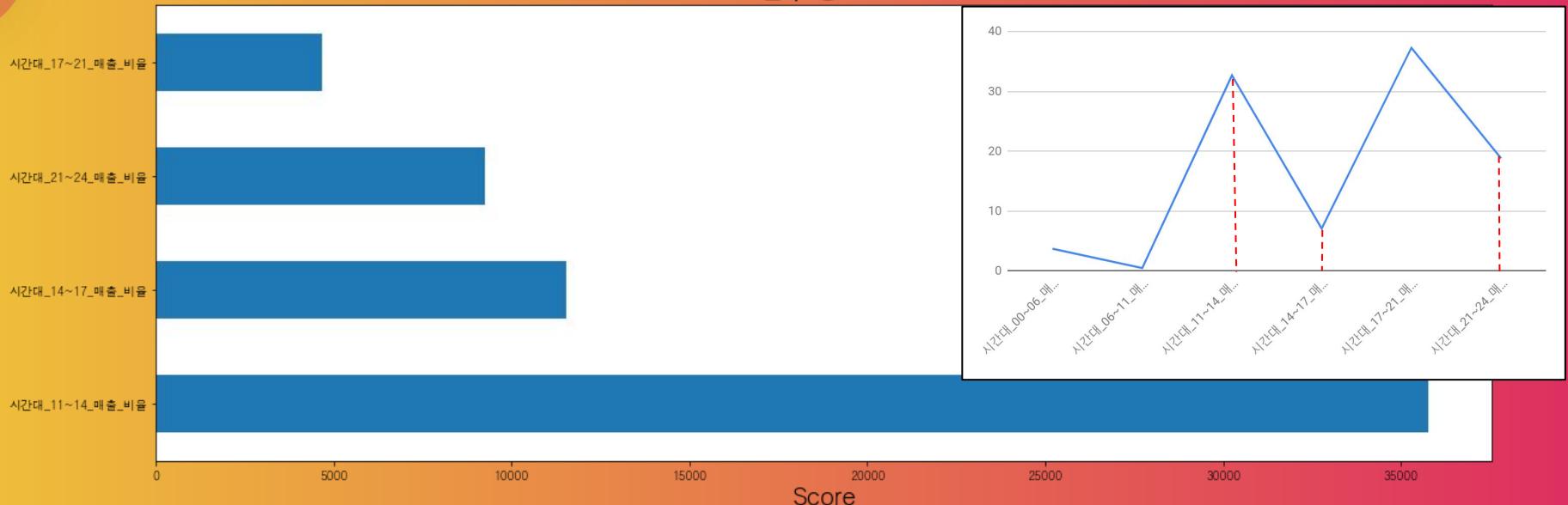
### -영동의 한식음식점 업종-



실제 한식음식점 매출은 금요일 가장 높은것으로 나타났지만,  
생존율에 있어서 월요일, 토요일, 화요일 매출이 중요한 것으로 도출되었음

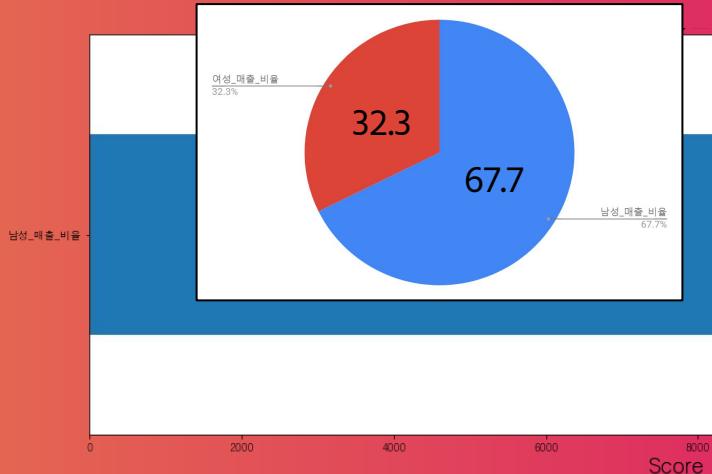
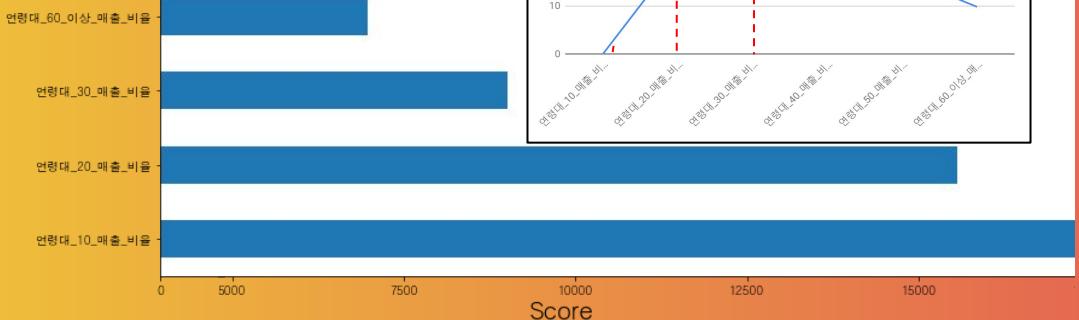
## 5. 분석 결과 예시

### -명동의 한식음식점 업종-



명동의 한식음식점은 점심시간대 매출과 저녁시간대 매출로 뚜렷히 나뉨.  
생존율에는 점심시간대 매출이 아주 중요한 요인으로 도출되었음

## 5. 분석 결과 예시



명동의 한식음식점은 20~30대가 가장 방문을 많이함.  
생존율에는 가장 적은 비중인 10대의 매출이 중요하게 작용하였음.

남성의 매출이 중요한 것으로 도출되었는데. 커플 접근성 좋은  
“분위기 좋은” 음식점이 생존율이 높을 것으로 예상됨

## 6. 요약 및 의의

“공공 데이터를 활용하여 창업 예정자들의 성공적 창업을 돋는 상권 분석 모델 개발”

### 분석 목표

- 1) 현재로부터 과거 1년 간의 상권 데이터를 바탕으로 다음 분기의 폐업률을 예측
- 2) 창업 예정자에게 전략적 요인 제공

### 분석 결과

- 1) Xgboost 모델을 사용하여 폐업률 예측 및 시각화
- 2) 모델의 변수 중요도를 기반으로 상권의 생존에 중요한 요인 분석 및 시각화

## 6. 요약 및 의의

“공공 데이터를 활용하여 창업 예정자들의 성공적 창업을 돋는 상권 분석 모델 개발”

### 의의

상권의 생존을 결정하는 주요 요인과

폐업률이 낮을 것으로 예상되는 지역 정보 제공 가능

### 활용 가능성

창업 예정자들의 최적 창업 입지 선정 및 운영 전략 수립

## 6. 요약 및 의의

“공공 데이터를 활용하여 창업 예정자들의 성공적 창업을 돋는 상권 분석 모델 개발”

발전 가능성

서울시 공공 데이터

2015년 1분기~2020년 2분기까지 수집된 상태



상권 데이터 충분히 축적

더 강건하고 신뢰할 수 있는

폐업률 예측 모델을 구축할 수 있을 것

# 감사합니다

## 팀 3P

정순우 산업시스템공학과 (융합소프트웨어 복수전공)  
손민영 의생명공학과 (융합소프트웨어 복수전공)  
문지용 경영학부 (융합소프트웨어 복수전공)