서울시 데이터 기반
일반음식점 최적 위치/업종/전략 추천

Data Mining Theory & Application
2019 Term Project

마스크사조

이태욱 2013147025

백상현 2013147051

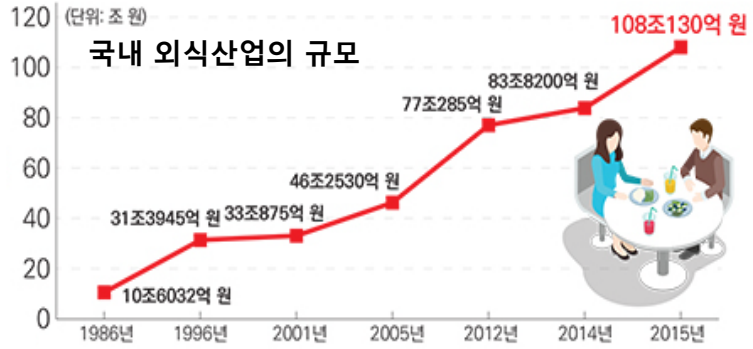
한재현 2014272026

김지환 2013147049

Abstract

Introduction

Project Goal



- ✓ 국내 외식업 시장, 양적 성장에 비해 저조한 수익 구조
- ✓ 자영업자의 체계적 분석 미비
- ✓ 자의적 창업결정방식의 한계를 극복하기 위한 체계적 상권 분석 필요

치열한 시장경쟁

높은 원가율 (매출액 70% 수준)

최저임금 금리 인상

경기변동에 민감한 사업 특성

낮은 진입장벽

음식점 생존여부 분류모델 구성

음식점 관련 정보 변수 설정

음식점의 창업 후 2년 (Death Valley) 생존 여부 예측

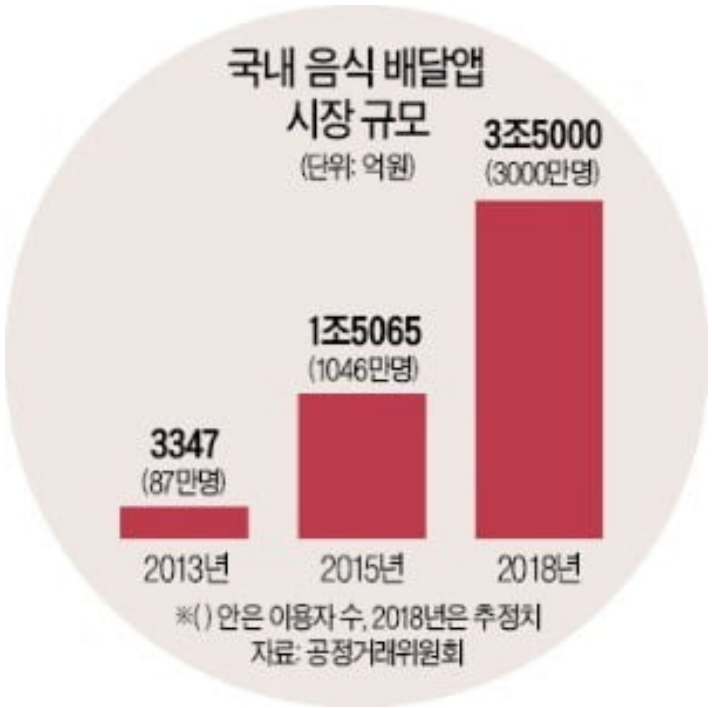
향상된 전략 및 솔루션 제공

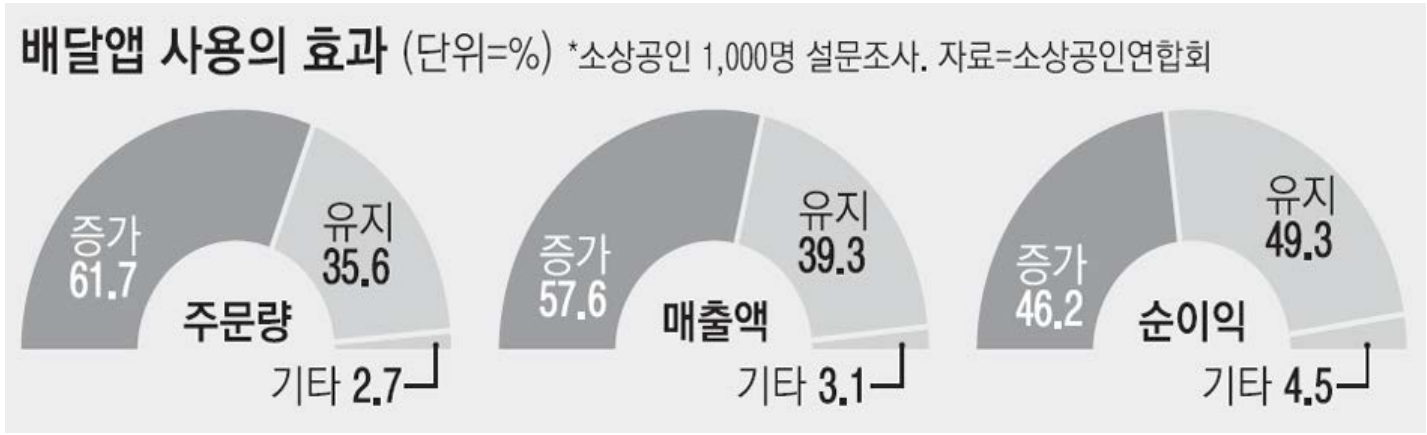
"타겟 지역과 업종 결정 시, 2년 뒤 생존확률 제공 및 최적화 전략 수립이 프로젝트 목적 !"

Introduction

Review Textmining

배달의민족 요기요 배달통





- ✓ 배달의 민족, 요기요, 배달통이 주도하고 있는 배달음식 어플 사용량이 꾸준히 증가하고 있으며, 그에 따른 음식 배달대행 서비스 시장 또한 확대운영 되는 추세.
- ✓ 최근 다양한 배달대행 서비스가 활성화됨에 따라 기존의 배달전문 음식점 뿐 아니라, 일반음식점 역시 배달서비스를 시행하고 있다.
- ✓ 이러한 트렌드를 반영해, 신규 개업하는 음식점 역시 배달 서비스 도입을 통해 시장에서 경쟁력을 보다 확보하고, 매출을 증대할 수 있으리라는 가설 수립.
- ✓ 각 지역별 사용자 리뷰에 어떠한 단어가 많이 등장하는지를 분석해 '맛', '양', '배달시간', '서비스 구성 ' 등의 요소 중 어느 부분을 중요시하는지 분석.

Literature Review

I. 텍스트마이닝을 이용한 고객 불만사항에 대한 사례연구

1. 개요
- 비구조화 데이터인 **고객의 불만사항** 분석 진행
- **단어-문서 행렬로 변환**
  - 지역적 가중치 함수 - 로그 사용
  - 전역적 가중치 함수 - 엔트로피 사용
- **군집 방법** - EM 군집
- **군집개수 선정** - 군집개수에 따른 평균값 정리

2. 결과
- 관련성이 높은 용어 간 관계 도식화 및 그룹화
- 연결 분석 기법 : 상호연관성의 시각적 표현을 통해 데이터 구조로의 접근성 향상

<텍스트마이닝을 이용한 고객 불만사항에 대한 사례연구>, 장지선 2012

▼

1) 다양한 변수를 포함한 자료의 분석을 통하여 데이터 그룹화

II. 뉴스 텍스트 마이닝과 시계열 분석을 이용한 주가예측

1. 개요
- 주식 방향 분류 모델 제시
- **BAG-OF-WORD기법**을 통해 변수 변환
  - 기존의 회계지표 경제지표
  - 타겟 기업과 관련된 뉴스 텍스트 마이닝

2. 결과
- 텍스트 마이닝과 분류기 **결합 방향성** 제시

<뉴스 텍스트 마이닝과 시계열 분석을 이용한 주가예측>, 안성원, 조성배

▼

2) 텍스트마이닝과 분류기 결합을 통한 연구 신뢰도 향상

III. Only the Bad Die Young : Restaurant Mortality in the Western US

1. 개요
- “새 식당의 90%는 첫 해 망한다?” 통념 확인 연구
- 데이터 : 미 연방통계국 고용임금조사
- 생존율 함수 : Kaplan-Meier Product Limit estimator
- **다른 산업과의 비교**

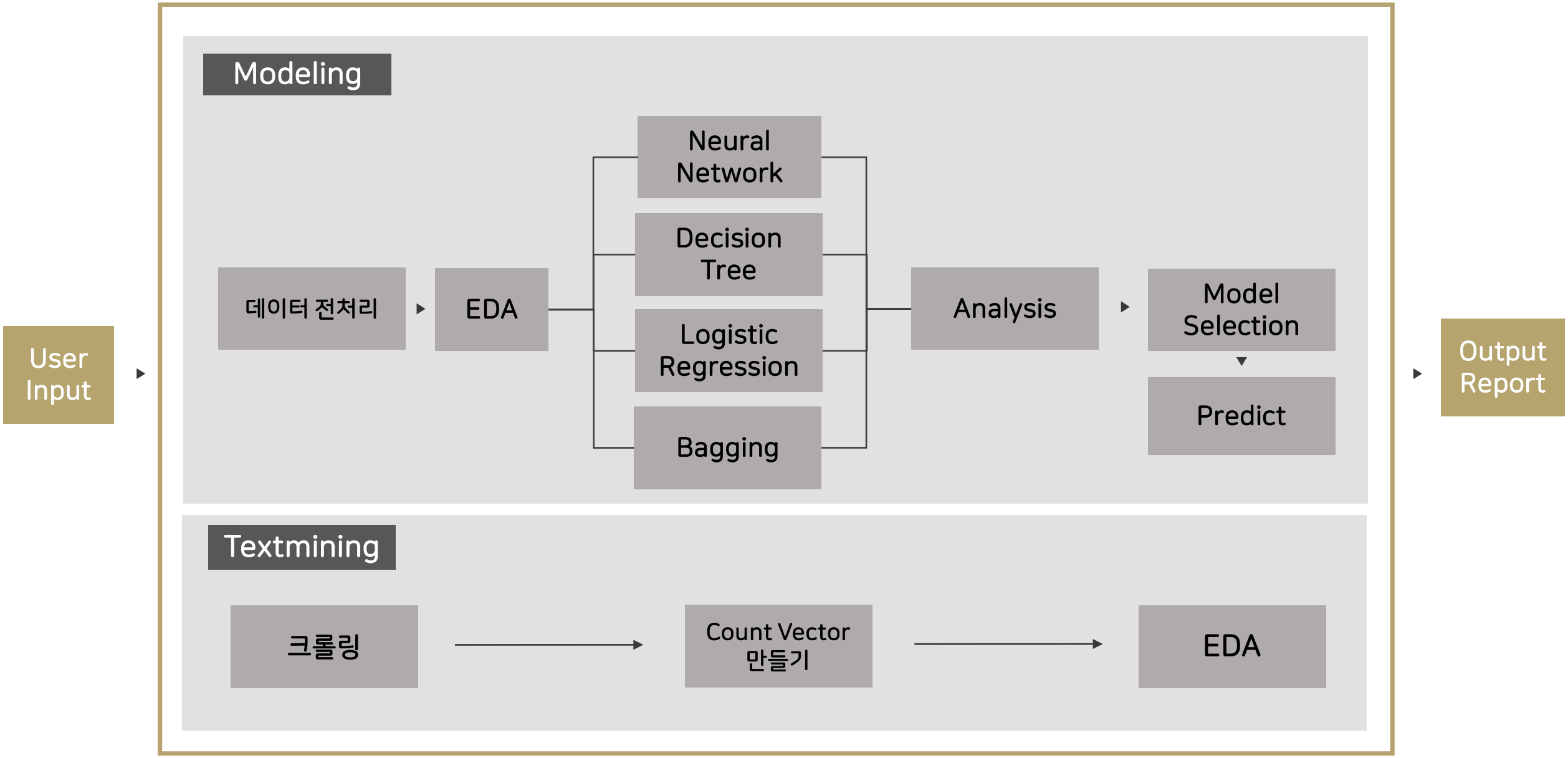
2. 활용
- 소득수준과 소비행태 등의 차이
- 지역별 특성 생존율에 유의미한 변화 발생

<Only the Bad Die Young: Restaurant Mortality in the Western US >, Tian Luo, Philip B. Stark

▼

3) 신규 음식점의 생존 여부 측정을 넘어 예측을 통해 지역별, 업종별 추천을 통한 차별화 고려

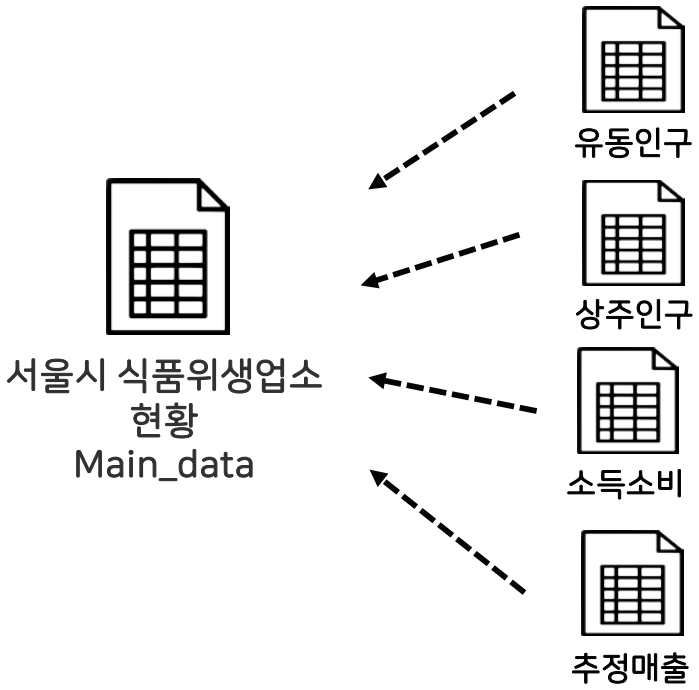
Methodology & Data



Methodology & Data

데이터 전처리_Data Set

전처리 및 Feature Engineering — EDA



- 필요없는 이상치 제거
- 필요없는 column 제거
- Feature engineering
- 결측치 제거

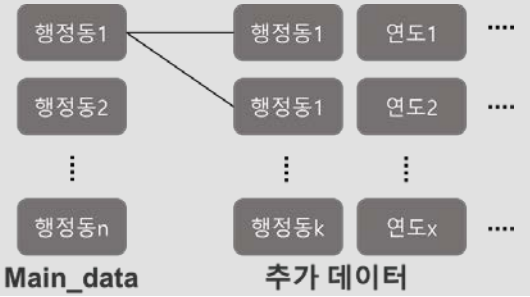
- **법정동-행정동 통일**

지역 단위 불일치 → 법정동으로 통일

1. 서울시 식품위생업소 현황 지역 단위: **행정동 → 법정동**
2. 서울시 우리마을가게 상권분석 서비스 지역 단위: **도로명 → 법정동**

- **통합전 row개수 통일**

1.Data 통합 시에 **key**가 되는 column **법정동**
2.추가할 데이터에는 기록 시점 등의 세부 정보에 의해 한 **법정동당 여러 개의 row 존재**
3.통합을 위해 법정동당 평균연산 **1row으로 통합**



- **통합데이터**

x 120,013 obs

음식점 Index	SURVIVE	법정동	유동인구	...	Type
1	0	가산동	324004	...	한식음식점
2	1	가산동	324004	...	중국식 음식점
...	...	...	...	...	...

x 137 column

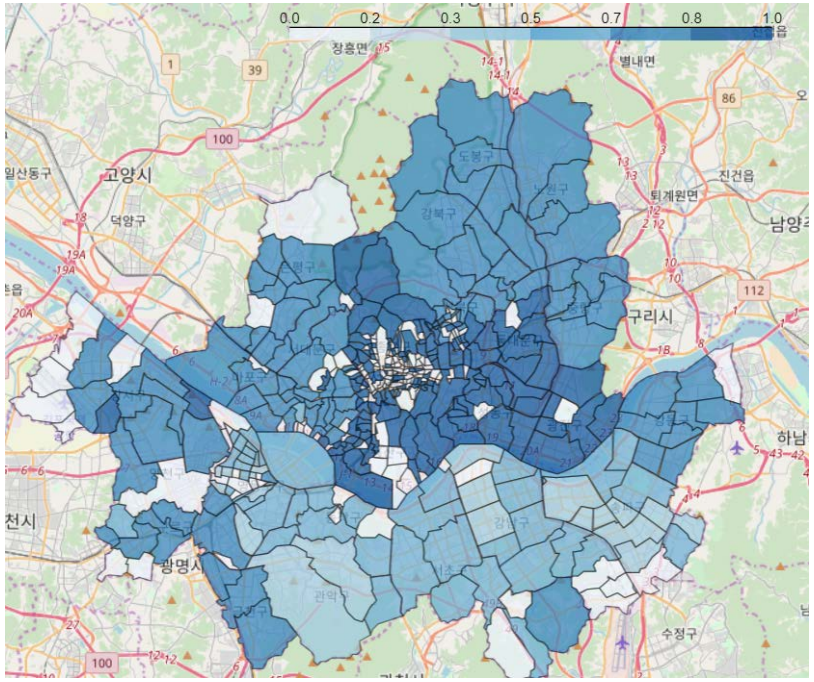
Methodology & Data

데이터 전처리_Data Set

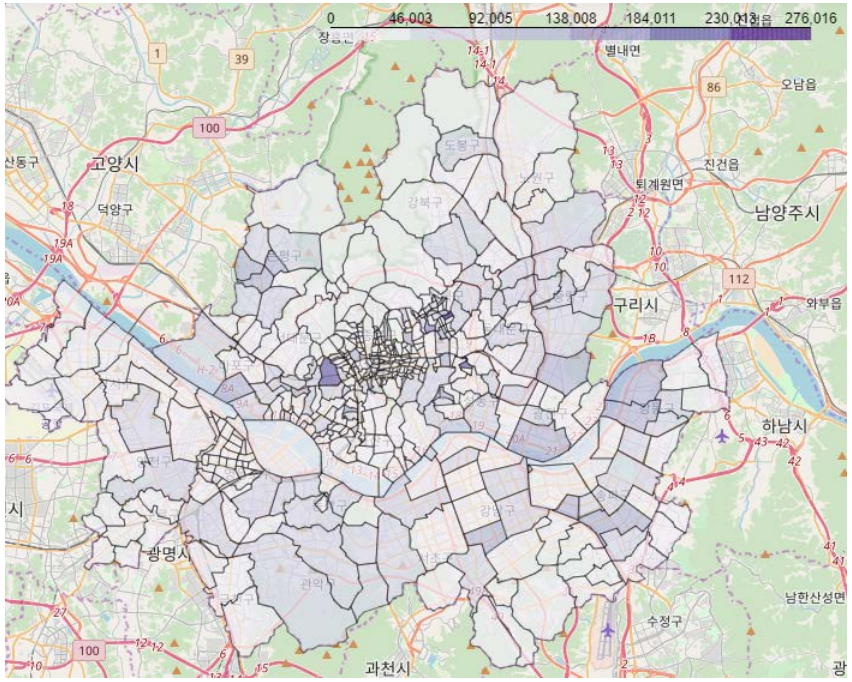
전처리 및 Feature Engineering

EDA

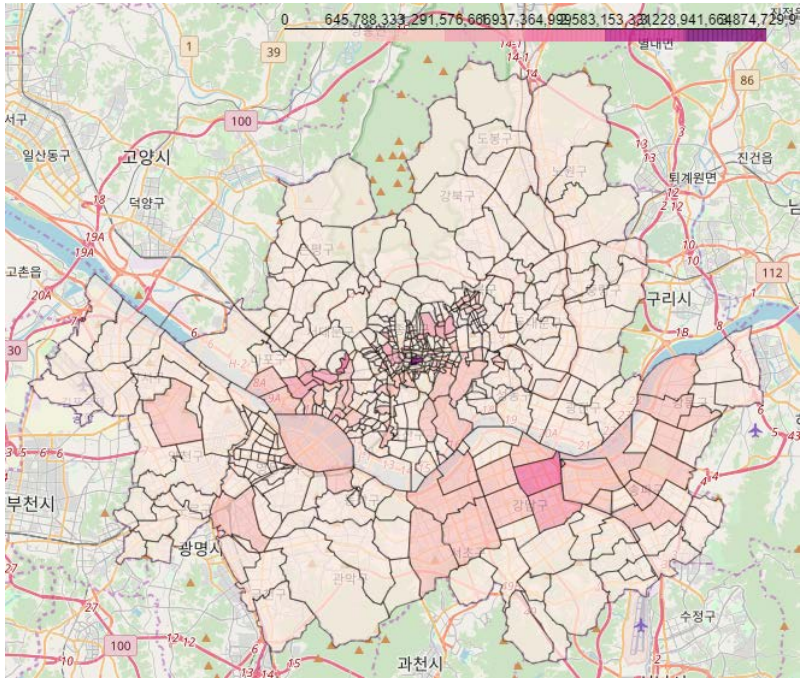
동별 생존율



동별유동인구



동별 추정매출



- ✓ 전처리 작업을 거친 Data로부터 생존율, 유동인구, 추정매출 등 주요 요인들의 분포를 살펴보았다.
- ✓ 전체 유동인구 중, **주말 유동인구의 비율**이 지역별 특성을 반영해 줄 것이라 판단, 새로운 변수를 생성하였다.

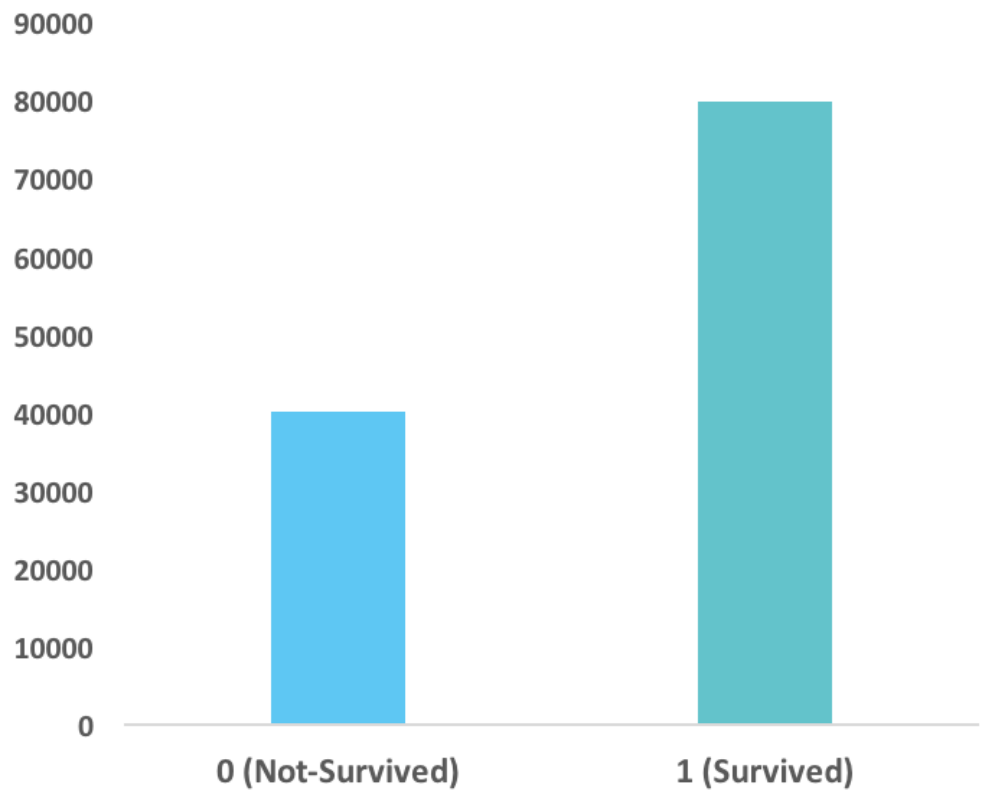
Methodology & Data

데이터 전처리_Data Set

전처리 및 Feature Engineering

EDA

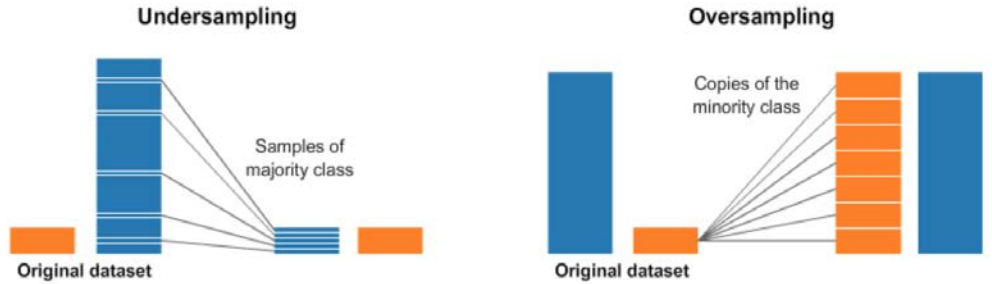
TARGET



TARGET이 굉장히 unbalanced임을 확인 할 수 있다.
(1: 음식점 2년이상 생존, 0: 음식점 2년 이상 생존 못함)

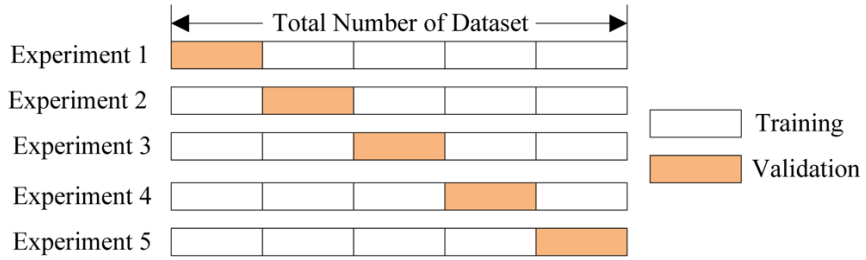
Sampling

불균형 데이터셋 클래스 편향 보정 언더/오버 샘플링:



Training Data에만 적용

과적합(Overfitting) 방지를 위한 교차 검증 (CV)



Empirical Analysis

Model 1

Modeling — Assess — Predict

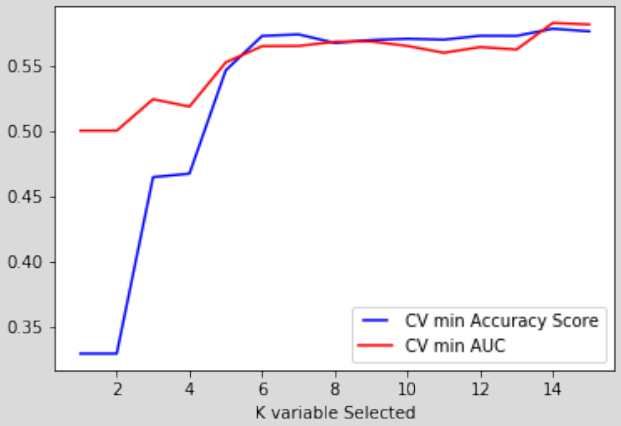
Logistic regression

1. 변수 선택

카이제곱 검정: 1~15개 변수 선택: n=1,2,…,15

n 개 변수가 채택된 데이터에 대하여 10 Fold Cross Validation 진행: Train/Test Data 10세트

Training Data에 대하여 Undersampling & Oversampling 진행 후 모델 피팅 실시.



10Fold CV 스코어 중 가장 낮은 점수를 채택하여 시각화.

변수가 6개 이상 선택되면 Score의 증가 대폭 감소.

최적 변수 개수: 6개

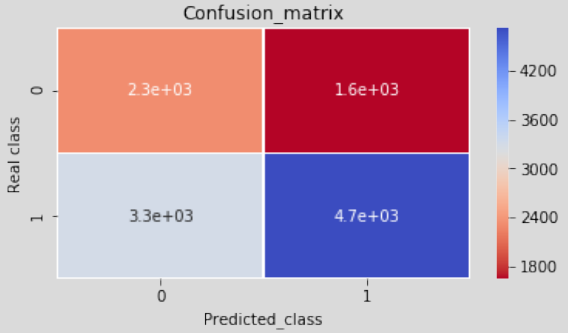
2. C 파라미터 튜닝

C값: 0.001~100 탐색: c=0.001, 0.01,…,100

6개 변수가 채택된 데이터에 대하여 10 Fold Cross Validation 진행: Train/Test Data 10세트

Training Data에 대하여 Undersampling & Oversampling 진행 후 모델 피팅 실시.

3. 모델 성능



Accuracy: 0.5881176568619282

---------Classification Report----------

	precision	recall	f1-score	support
0	0.41	0.59	0.48	3972
1	0.74	0.59	0.66	8029
micro avg	0.59	0.59	0.59	12001
macro avg	0.58	0.59	0.57	12001
weighted avg	0.63	0.59	0.60	12001

예측 정확도: 0.58

Empirical Analysis

Model 2

Modeling — Assess — Predict

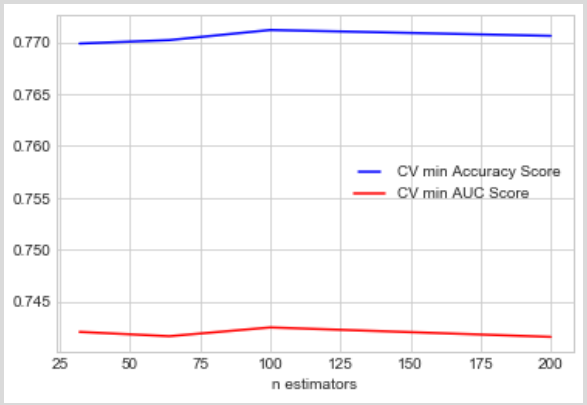
Random Forest

1. N estimator 튜닝

n_estimator: 32, 64, 100, 200 시도

데이터에 대하여 10 Fold Cross Validation 진행:
Train/Test Data 10세트

Training Data에 대하여 Undersampling & Oversampling 진행 후 모델 피팅 실시.



n_estimator 값에 따른 유의한 CV 스코어의 변화 없음.

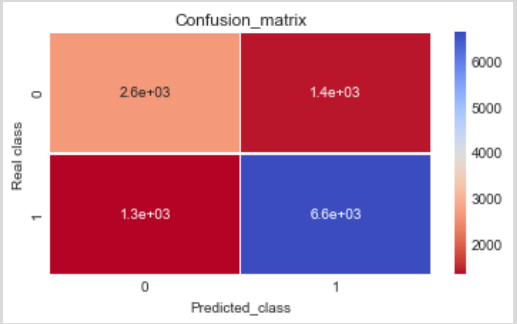
32개 트리의 평균은 충분히 낮은 Variance를 가진다고 판단.

n_estimator가 100일 때 가장 안정적인 CV 결과 발생

n_estimator 100인 모델 채택.

2. 모델 성능

TP 6620 TN 2645
FP 1417 FN 1319



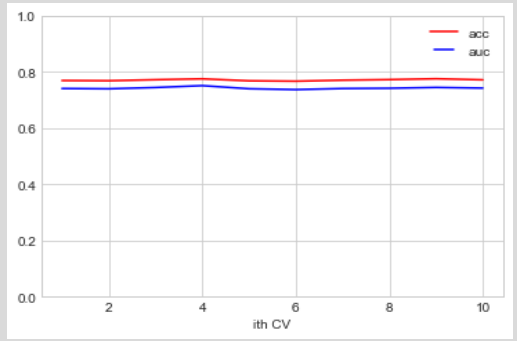

```
Accuracy:  0.7720189984167986

---------Classification Report----------------------------
              precision    recall  f1-score   support

           0       0.67      0.65      0.66      4062
           1       0.82      0.83      0.83      7939

   micro avg       0.77      0.77      0.77     12001
   macro avg       0.75      0.74      0.74     12001
weighted avg       0.77      0.77      0.77     12001

n_estimators:  100 acc: 0.7712664733640691 auc 0.7424600486302686
```



10 Fold CV 성능 모두 안정적

예측 정확도: 0.77

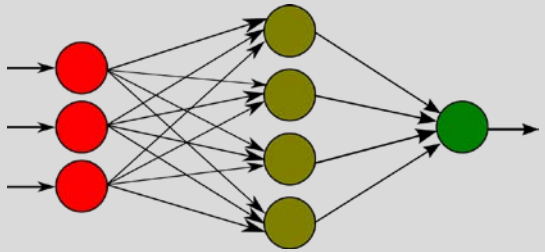
Empirical Analysis

Model 3 & 4

Modeling — Assess — Predict

Multilayer Perceptron

Hidden-layer 크기 결정:
(i) number of hidden layers equals one
(ii) the number of neurons in that layer is the mean of the neurons in the input and output layers.
위 규칙을 지키면 대부분 상황에서 좋은 성능을 보인다는 가이드라인.



10Fold CV Accuracy
매우 불안정
0.35 ~ 0.66

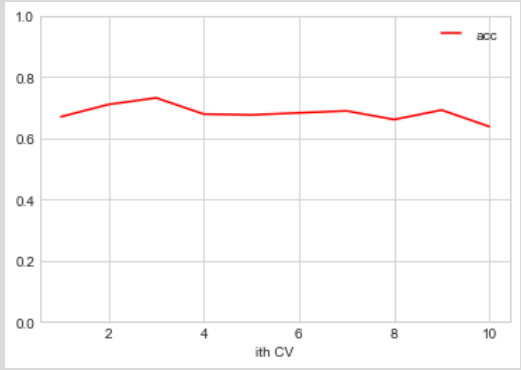
Input 148, hidden Layer: 1 with 148 Neuron. Optimizer: Adam

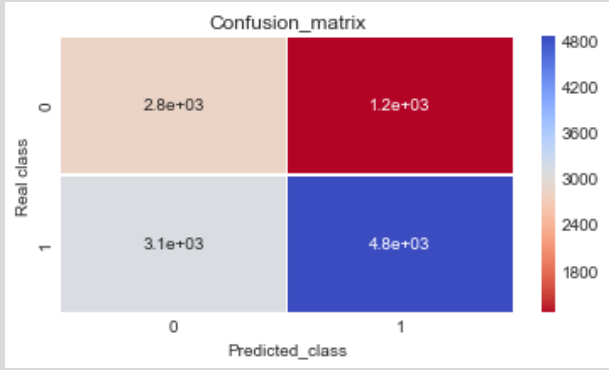
10 Fold Cross Validation, Undersampling / Oversampling 진행 후 모델 트레이닝 실시.

Bagging

Bagging 샘플을 여러 번 뽑아 각 모델을 학습시켜 결과를 집계 (Aggregating) 하는 방법이다.

10 Fold Cross Validation, Undersampling / Oversampling 진행 후 모델 트레이닝 실시.





10Fold CV Accuracy 안정적

예측 정확도: 0.64

Empirical Analysis

Model Selection

Modeling — Assess — Predict

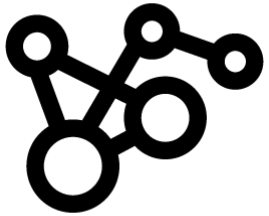
성능 차이

Logistic Regression	0.58
Random Forest	0.77
Multilayer Perceptron	0.35
Bagging	0.64

Random Forest 모델 채택

결과 분석

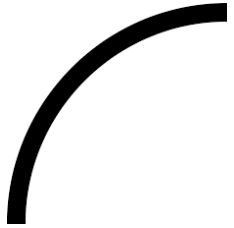
트리기반 모델들이 전반적으로 더 좋은 성능을 보인다.



상호작용 효과

01

분류 문제



비선형적 관계

Random Forest의 성능이 Bagging보다 우수하다.

Empirical Analysis

Textmining

Crawling → Counter Vector 만들기 → EDA

텍스트 마이닝을 통해 지역별 중요 키워드를 추출하여, 사업자들에게 상대적으로 집중해야할 정보 제공

업체 선정 후 리뷰 크롤링

KoNLPy

- 오픈소스 소프트웨어
- 한국어 정보처리 파이썬 패키지
- 자연어처리(NLP)

Wordcloud를 통한 시각화

- 각 지역별 소비자 리뷰 분석
- 분석 후 시각화
- 판단 핵심 요인 분석

Empirical Analysis

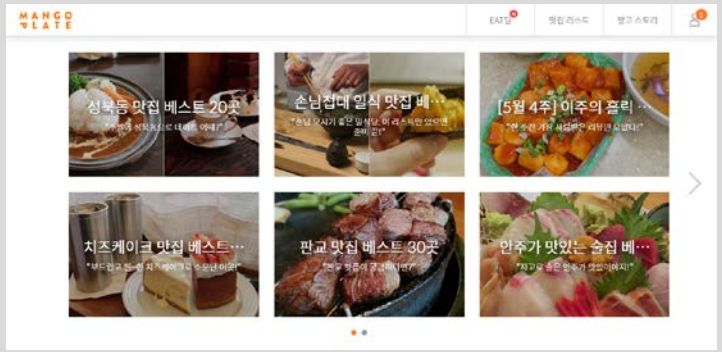
Textmining

Crawling → Counter Vector 만들기 → EDA

크롤링 업체 선정

크롤링 후보

MANGO PLATE



맛집을 추천하는 컨텐츠 위주로 구성

요기요

서비스 사용자만 쓸 수 있는 클린리뷰

▼

망고플레이트 좋은 리뷰만 존재, Sampling 과정에서 Bias 발생
→ **무의미**

▼

업체 '요기요'로 크롤링 진행
다양한 리뷰 크롤링 가능

파이썬 request를 써서 크롤링한 결과

각 동당 30개 음식점에 한해서 리뷰 크롤링 진행

법정동	별점	리뷰
법정동 300개	맛 양 배달	자연어 리뷰

Empirical Analysis

Textmining

Crawling — **Counter Vector 만들기** — EDA

형태소 분석
크롤링한 음식점 리뷰에 대해 형태소 분석 실시.

형태소 분석이란:
어떤 대상 어절을 최소의 의미 단위인 '형태소'로 분석하는 것을 의미한다.

Konlpy에서 제공하는 Okt 한글 자연어 처리기 사용.

```
In [10]: okt.morphs('맛있어해요 맛있었다 맛있었던 맛있었어요 맛있죠 맛있지 맛있음 맛있습니다',norm=True,stem=True)
Out[10]: ['맛있다', '해', '요', '맛있다', '맛있다', '맛있다', '맛있다', '맛있다', '맛있다', '맛있다']
```

예시:
"배달 빠르고 맛있게 잘 먹었어요.이런 행사 자주했음 좋겠어요."

↓

'배달', '빠르다', '맛있다', '자다', '먹다', '이렇다', '행사', '자주', '하다', '좋다'

카운트 벡터
각 법정동의 리뷰 특징을 관찰하기 위해,
각 법정동에 작성된 리뷰에 쓰인 형태소 카운트 벡터 생성.

Sklearn에서 제공하는 CountVectorizer사용.

옥수동		옥천동		청파동3가	
많다	983	보다	175	빠르다	1245
보다	995	오다	218	자다	1287
오다	1183	너무	238	많다	1378
너무	1496	빠르다	255	너무	1710
좋다	1970	시키다	312	좋다	2429
시키다	2305	좋다	406	시키다	2652
배달	3001	하다	485	하다	3303
하다	3047	배달	623	배달	3306
먹다	4989	먹다	736	먹다	6558
맛있다	5620	맛있다	949	맛있다	7453

Empirical Analysis

Textmining

Crawling → Counter Vector 만들기 → EDA

1. 대부분의 리뷰에서 공통적으로 많이 쓰이는 단어가 많음
→ 동별 특징 파악하기 힘듦

"맛있다 / 잘먹었다" 등의 리뷰가 대다수

2. 리뷰 단어들이 제각각
→ 완벽한 categorize가 어려움

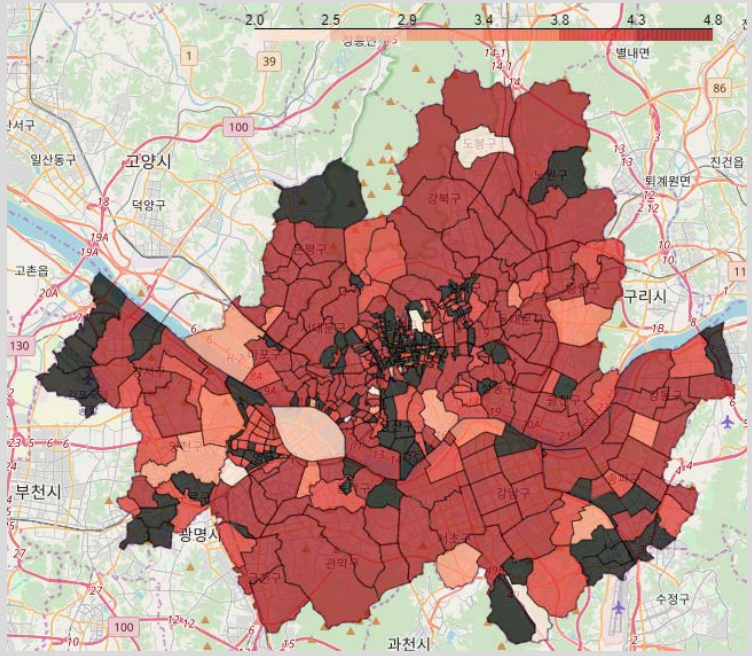
따라서 리뷰의 맛, 양, 배달 별점을 활용

Empirical Analysis

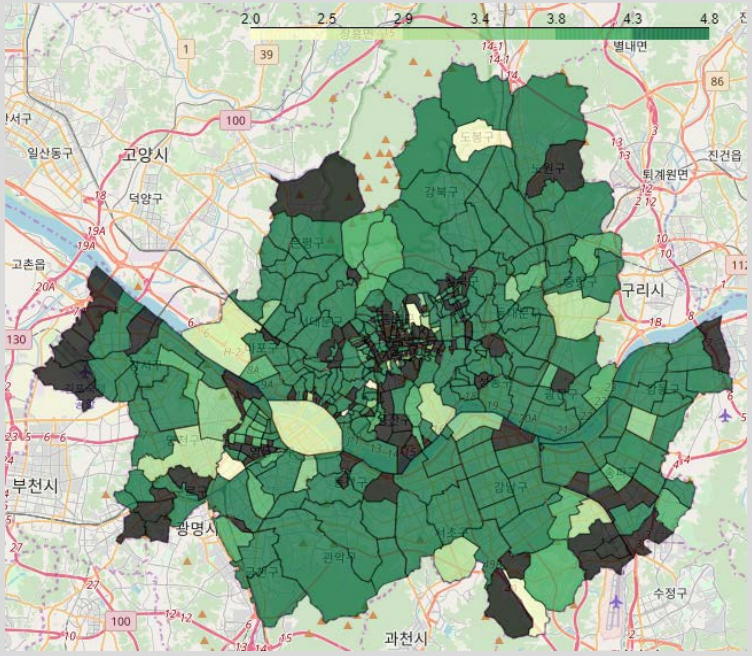
Textmining

Crawling → Counter Vector 만들기 → **EDA**

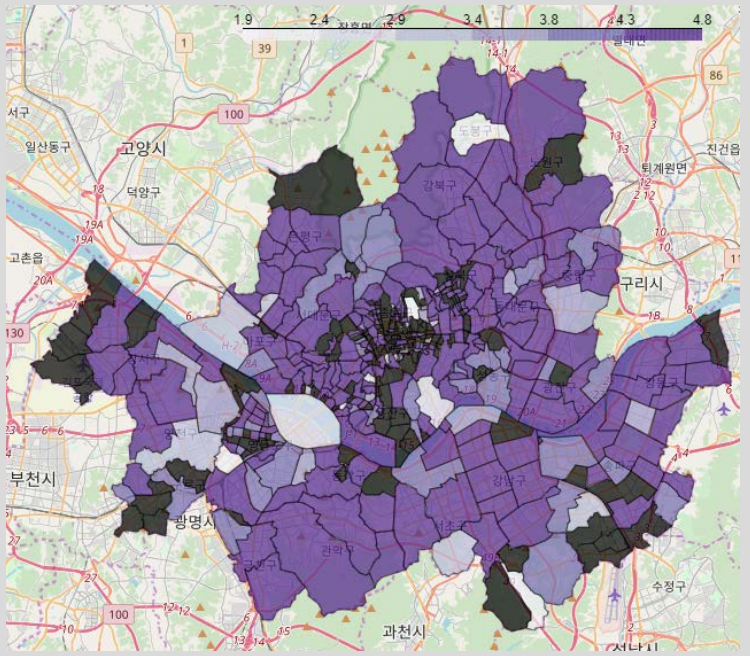
맛



양



배달



- ✓ 동별 배달음식점의 고객 별점을 기반으로 맛, 양, 배달 요소에 대한 민감도를 나타냈다.
- ✓ 전체적으로 비슷한 흐름을 보이지만, 남창동, 동소문동 1가 등 몇몇 지역에서는 요인 별 평점에서 큰 차이를 보여주었다.

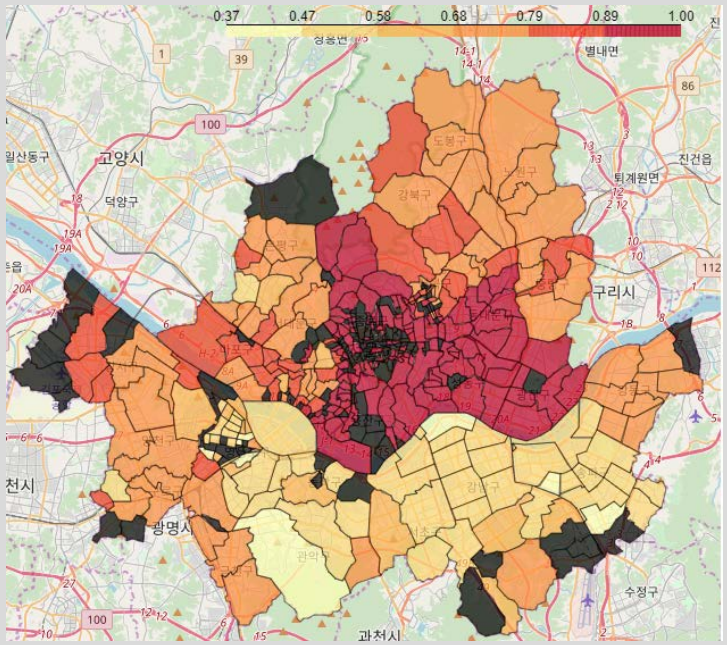
Conclusion

Predict

In [1]: Forecast('가락동','한식음식점')

Out [1]: 가락동은 한식음식점 사업에 적합하지 않습니다.
RF Classifier 2년 생존 성공확률: 12%

한식음식점 동별 2년 생존율



가락동 배달 리뷰 주요 단어

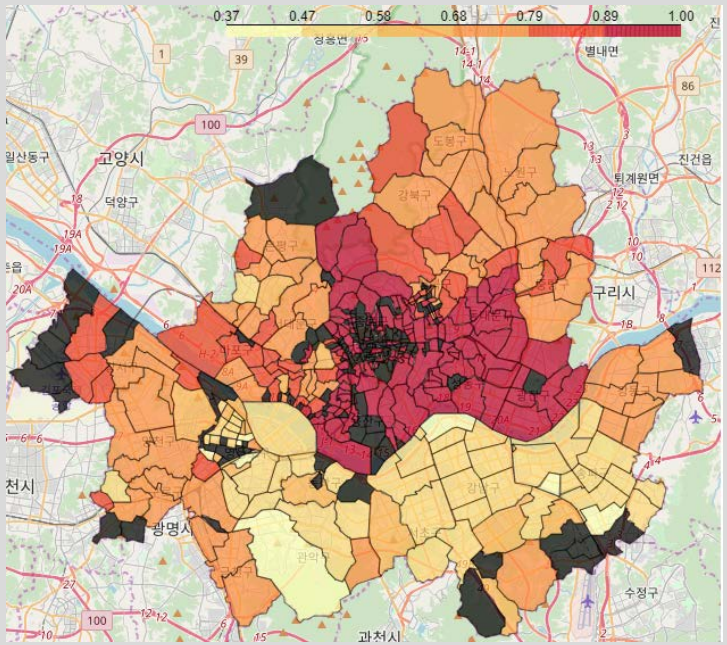
Conclusion

Predict

In [1]: Forecast('옥수동','한식음식점')

Out [1]: 옥수동은 한식음식점 사업에 적합합니다.
RF Classifier 2년 생존 성공확률: 82%

한식음식점 동별 2년 생존율



옥수동 배달 리뷰 주요 단어

Conclusion

Development

한계점 및 개선방향

- ✓ 위치 기본 단위 '법정동'의 크기 불균등 → 법정동이 큰 경우 하나의 수치로 온전하게 표현하기 힘듬 → 좌표별 구분이 보다 적합할지도
- ✓ 상권 데이터에 모든 법정동이 포함되지 않음 → 온전한 데이터를 획득할 필요
- ✓ 상권별 데이터 중 단기간 정보 존재 → 장기간 데이터를 확보할 필요
- ✓ 사용자 리뷰 텍스트를 보다 정밀하게 처리하고, 원데이터에 매칭되는 리뷰를 얻는다면, 보다 정밀한 분석이 기대된다.

유의한 변수를 더 얻을 수 있을 것으로 기대

활용방안

- ✓ 일반음식점 개업을 준비하거나 계획중인 자영업자들에게 데이터 기반의 의사결정 tool 제공
- ✓ 요식업 외에도 다양한 유통업, 소매점 등 다양한 사업 분야에 접목 가능

기대효과

- ✓ 창업자들에게 유의한 정보를 제공해줌으로써, 일반음식점의 실패율을 감소시킬 수 있을 것이라 기대된다.
- ✓ 음식점 창업주들의 대다수가 생업을 위해 큰 투자를 감수한다는 점을 감안한다면, 이런 추천은 사회적 자원낭비를 줄일 수 있을 것이다.

마스크사조

김지환 백상현 이태욱 한재현

Reference

<https://ratsgo.github.io/from%20frequency%20to%20semantics/2017/05/10/postag/>
형태소 분석기 성능 비교

<https://iostream.tistory.com/144>
한국어 형태소 분석기 성능 비교

<https://pythonhealthcare.org/2018/06/02/85-using-free-text-for-classification-bag-of-words/>
Using free text for classification – 'Bag of Words'

<http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE01474739&language=ko_KR>
뉴스 텍스트 마이닝과 시계열 분석을 이용한 주가예측

<http://www.riss.kr/search/detail/DetailView.do?p_mat_type=be54d9b8bc7cdb09&control_no=0f894f453f93b7b8ffe0bdc3ef48d419#redirect>
텍스트마이닝을 이용한 고객 불만사항에 대한 사례연구