

TO Small Business

소상공인신용평가모델

To Small Business

CONTENTS

01. 문제인식

02. 프로젝트 목표

03. 대상 선정

04. 벤치마킹

05. 상권분석

06. 변수선정

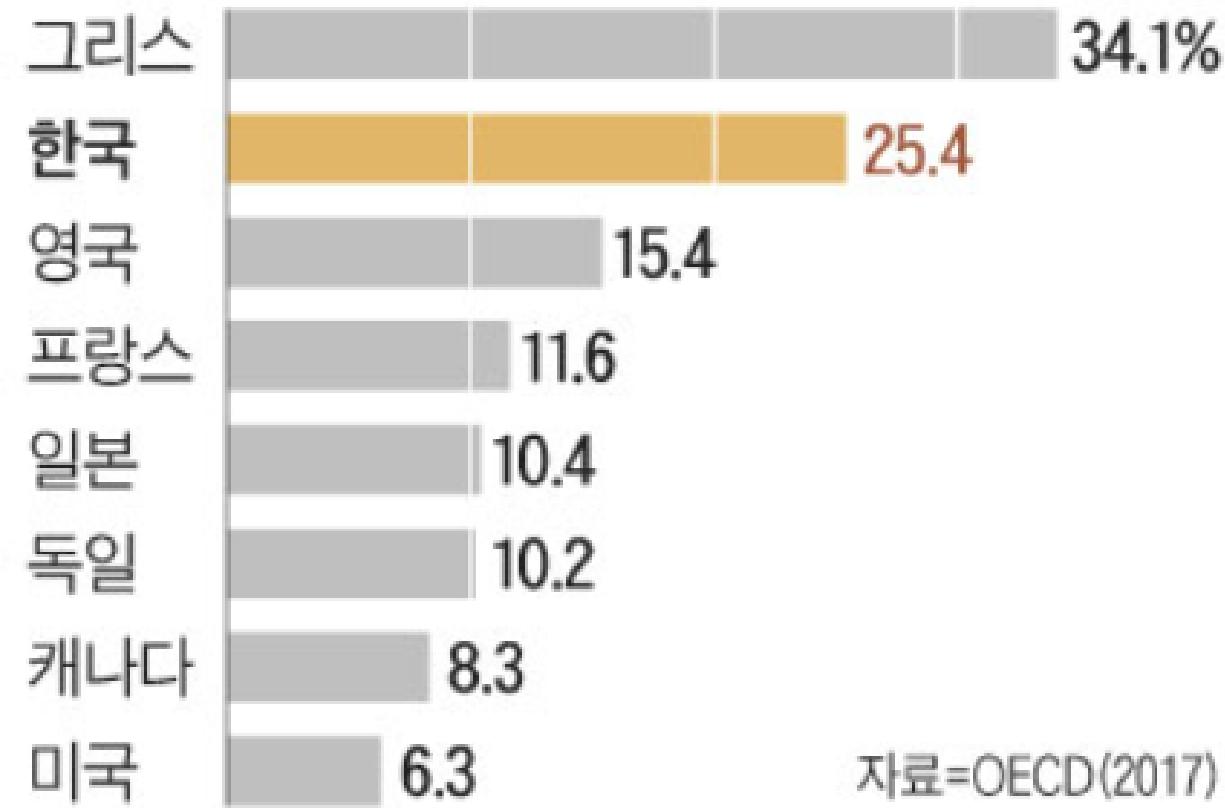
08. 모델링

09. 서비스제공 (Web)

Problem Recognition

문제인식

OECD 주요국의 취업자 중 자영업자 비율



자료=OECD(2017)

자영업자 수



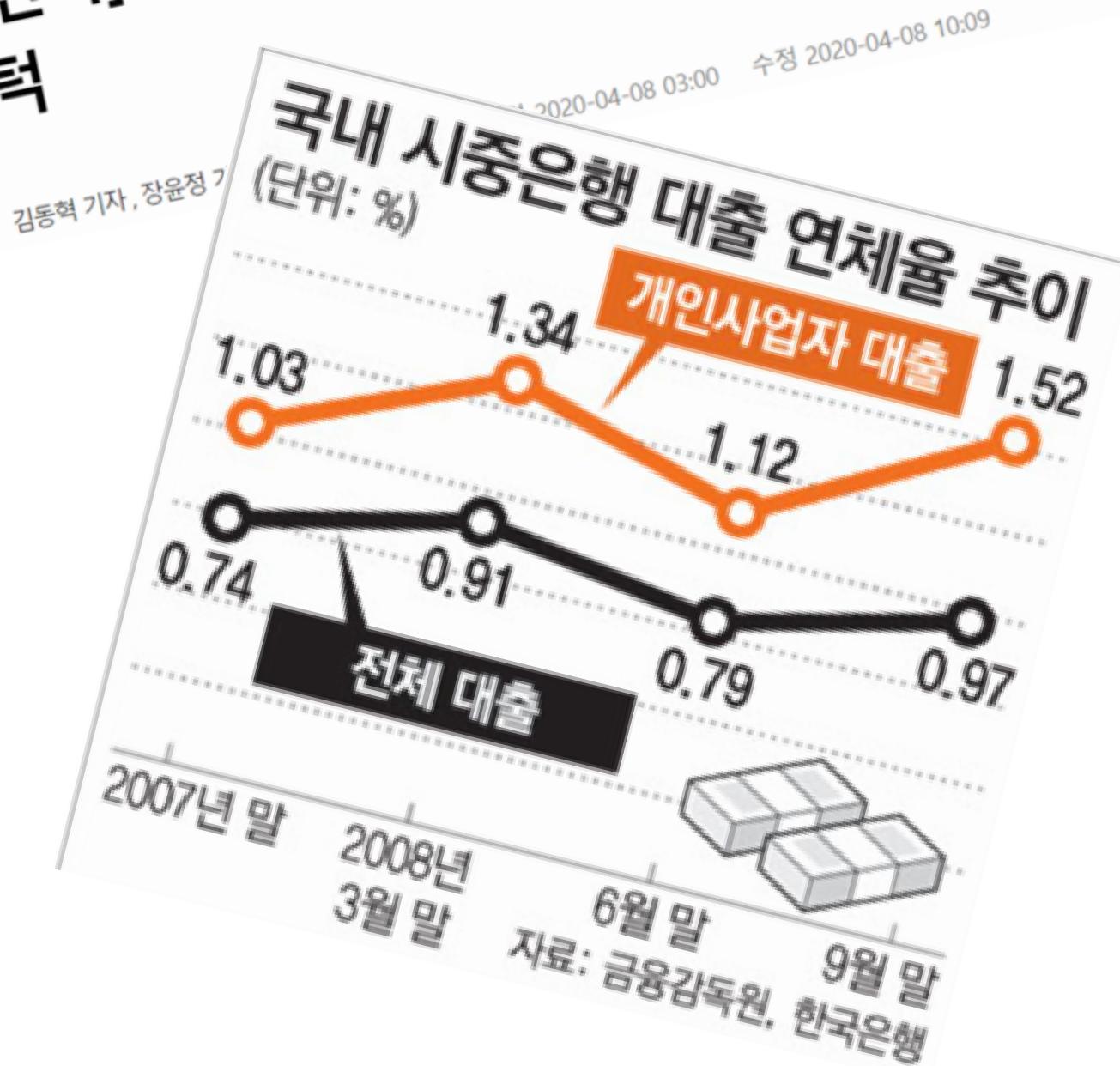
소상공인이 실패하는 이유?



자금부족이 2위 (29%)

[단독]高신용 2등급도 대출 퇴짜... 벼랑끝 소상공인에 여전한 은행 문 턱

뉴스 > 사회 > 복지

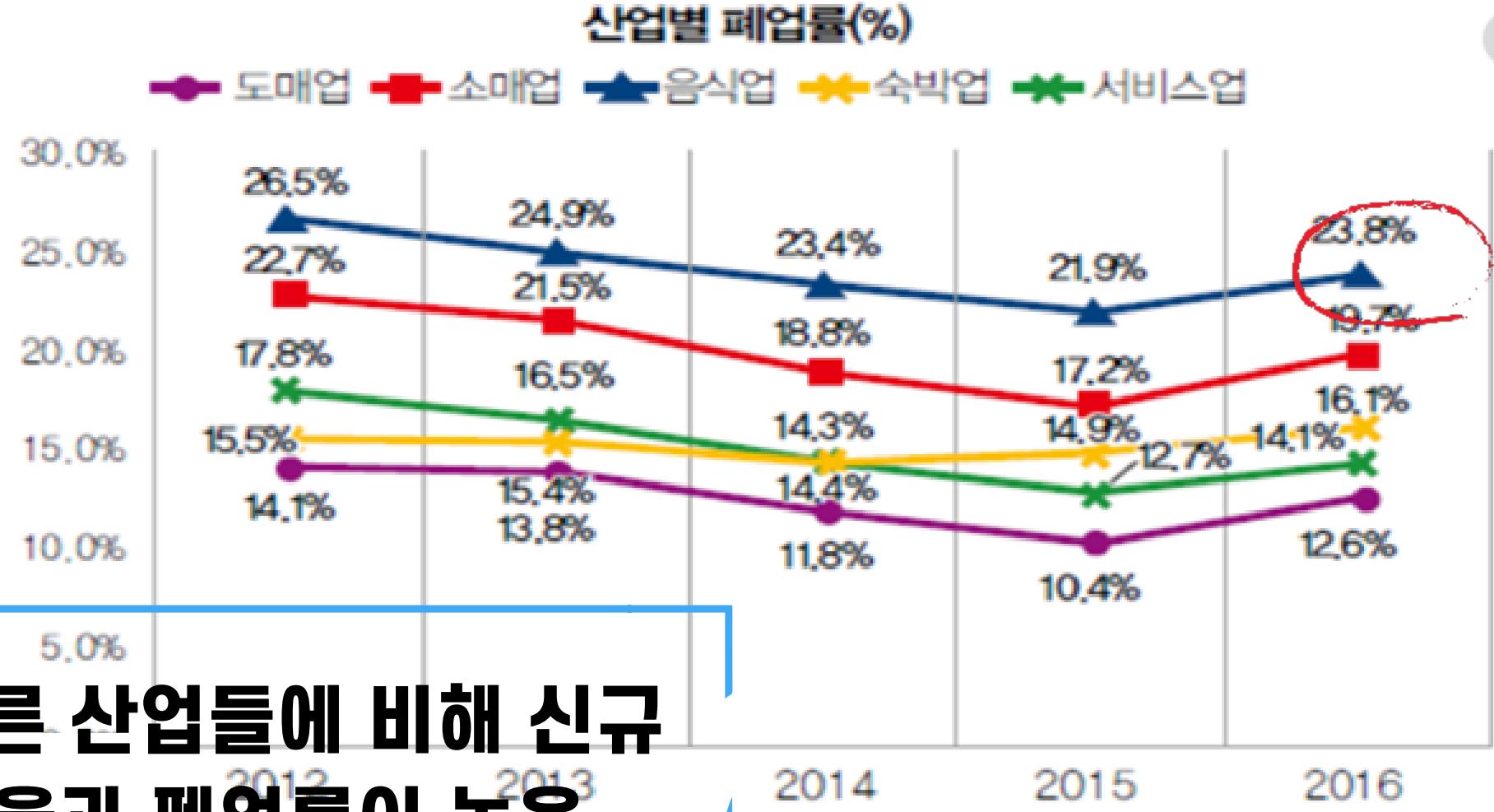
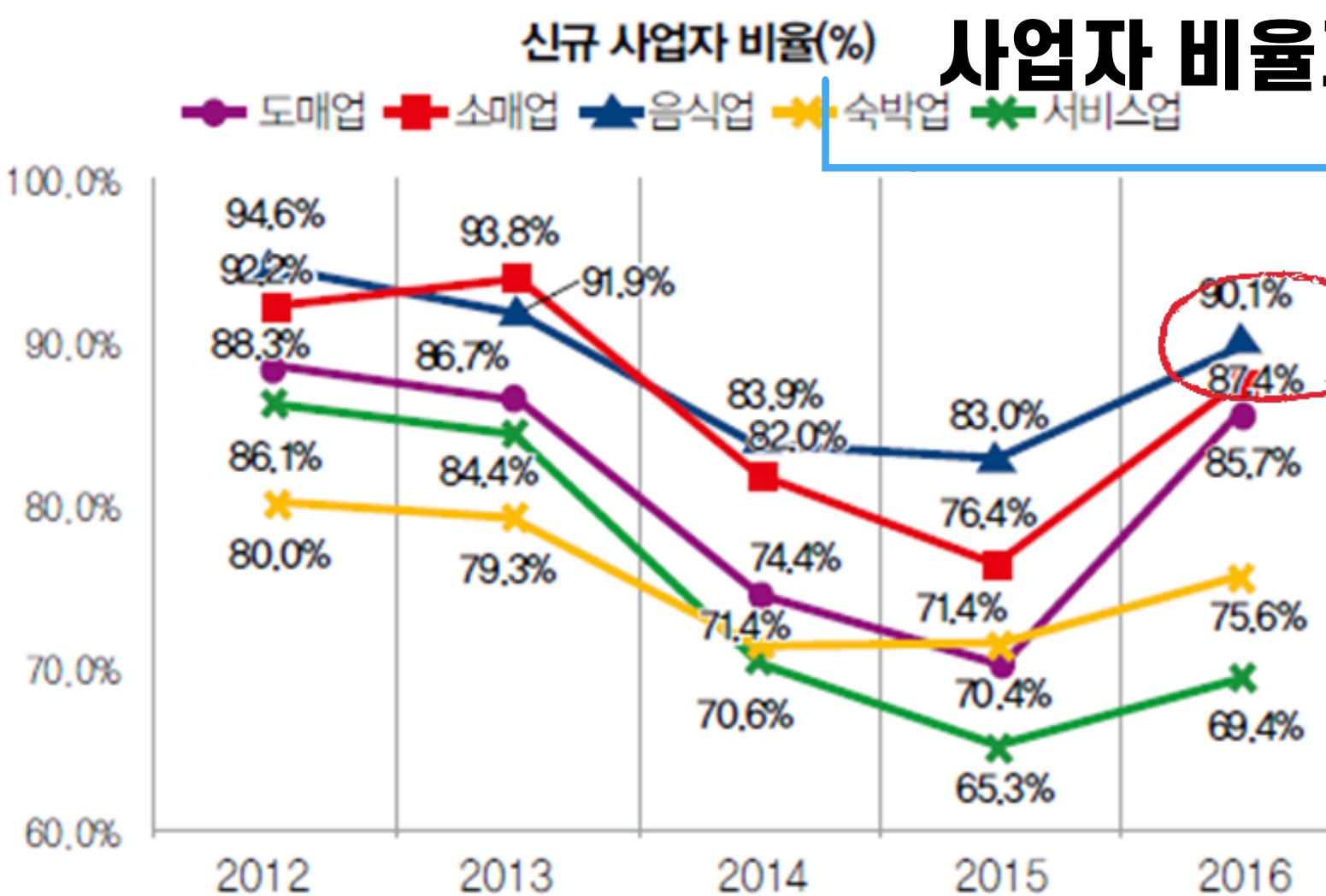


〈표 19〉 자영업자가구와 비자영업자 가구의 재무상태 비교

	단위	전체가구	(단위: 천 개, 만 원, %)		
			자영업자		비자영업자
			임금근로	기타	
전체가구 수	천개	19,463	5,340	14,124	10,454
		(100)	(27.4)	(72.6)	(53.7)
부채보유 가구수	천개	12,360	3,912	8,448	7,052
		[63.5]	[73.3]	[59.8]	[67.5]
부채/자산	(%)	23.6	24.7	22.9	24.1
평균 부채	만원	11,179	14,492	9,645	9,797
		(100)	(129.6)	(86.3)	(87.6)
① 금융부채	(%)	71.0	76.6	67.1	70.5
- 담보대출	(%)	57.3	61.7	54.3	56.6
- 신용대출	(%)	9.6	10.2	9.1	10.1
② 임대보증금	(%)	29.0	23.4	32.9	29.4
...

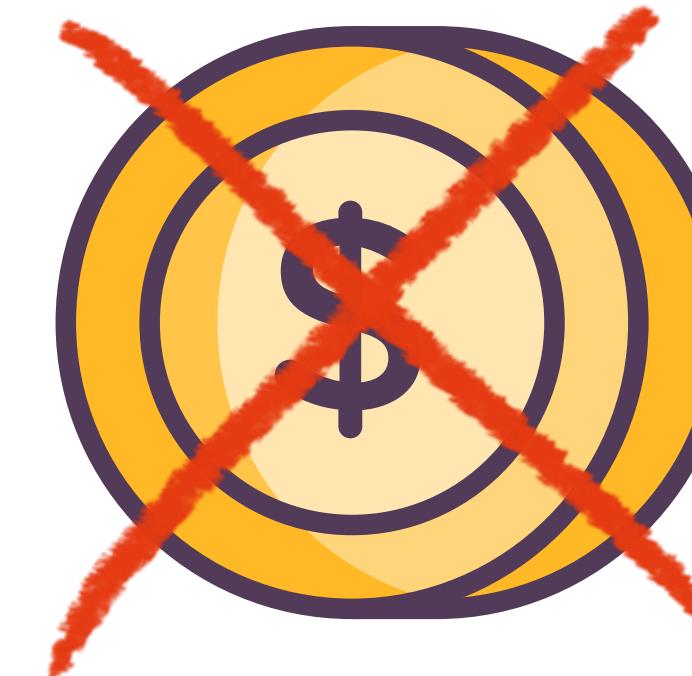
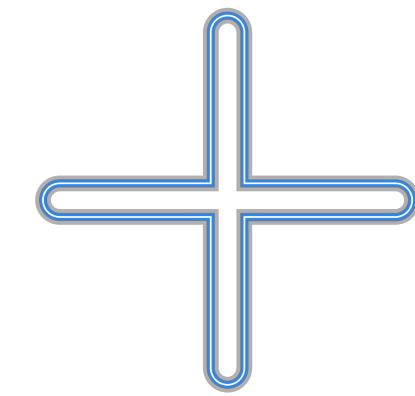


요식업이 다른 산업들에 비해 신규 사업자 비율과 폐업률이 높음



OUR GOAL

재무정보, 비재무정보를 이용한 머신러닝을 통한 신용평가



Machine Learning

고객경험(UX)을 활용한 상권분석



닥터조



5.0

식전 감자조림 소고기 튀김 부터 좋더니 문어와 고추장소스 맛있었다. 성계 선호하시면 성게알을 비빔밥 추천. 전복 좋아하지도 않는데 전복요리가 먹은 날 베스트. 구절편도 깔끔. 항정살 쌈의 두부된장 좋았고 투뿔등심은 양은 적지만 고기는 상당히 좋았다. 옥돔요리 비늘같이 만들어 놓은것 바삭바삭. 정식당을 시그니처 디저트인 돌하르방은 비주얼부터 맛까지 완벽. 음식을 맛과 비주얼 뿐만아니라 서버들의 태도등 괜히 뉴욕점이 빠른시간에 미슐랭 2스타를 받은게 아닌듯. 전체적으로 훌륭했다.



FirstLove1025



3.5

우리나라 최초 아시아 베스트 50 레스토랑에 선정되었고 2015년 아시아 10위에 선정된 레스토랑 뉴욕의 정식당은 미슐랭 가이드에서 매년 별을 받고 있습니다. 자랑스러운 한국의 레스토랑입니다.



좋아요 (0)



댓글 (0)

3

사용자에게 신용평가요소에 영향을 미친 변수를 설명해주는 서비스 개발



Home Scoring About Menu Team Gallery Contact

대출할 수 있을까?

STEAK HOUSE HTML CSS TEMPLATE



개인사업자 신용평가

요식업에 종사하는 개인
사업자에 대한 신용평가
모형

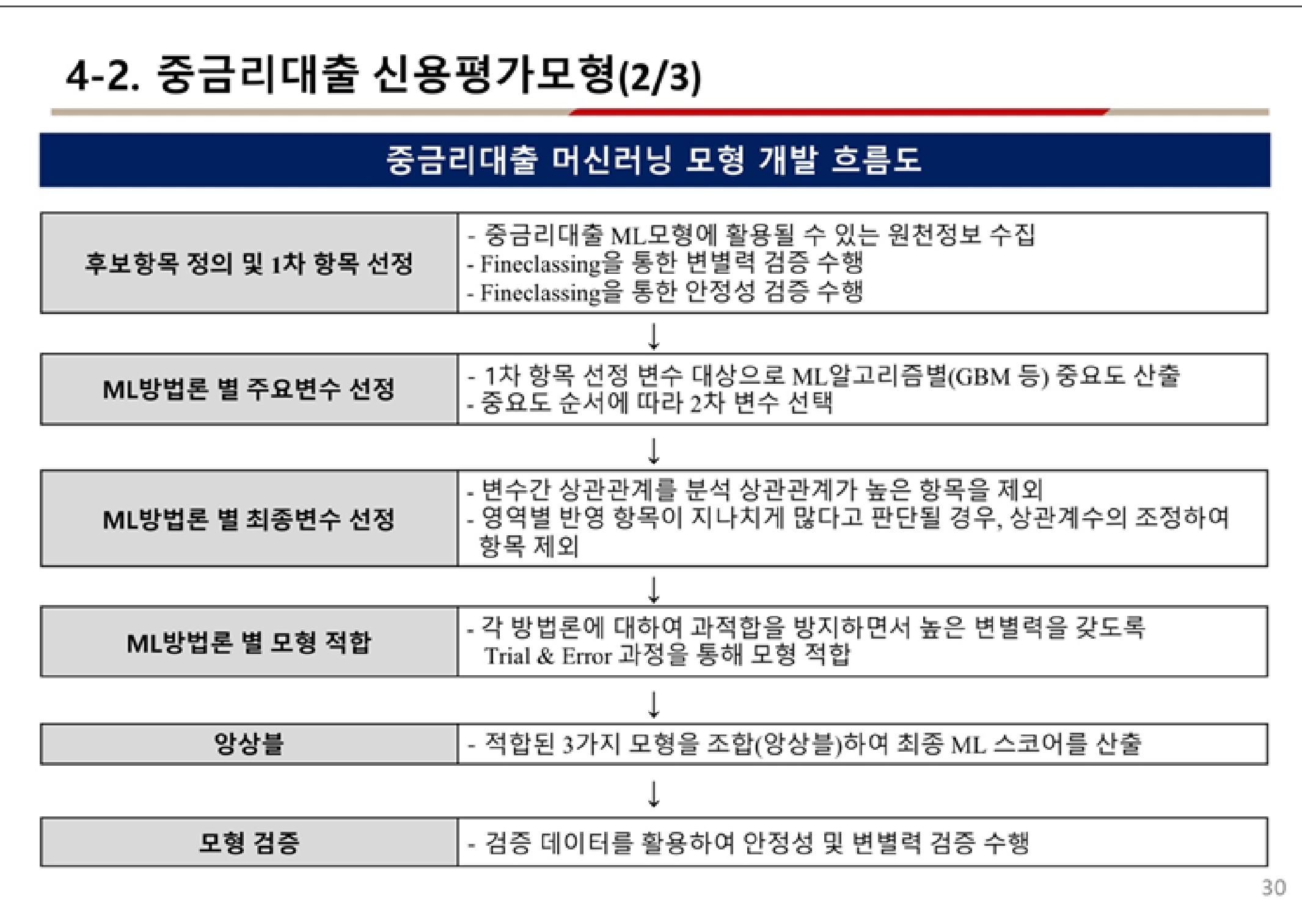


상권 분석

비재무적인 요소들을 넣
어 만든 지역별 상권분석

4-2. 중금리대출 신용평가모형(2/3)

중금리대출 머신러닝 모형 개발 흐름도



Ensemble

Logistic Regression

XGBoost

DNN

5. 상권분석

어떻게 상권분석을 할까?

요식업 && 행정동 코드별로 데이터를 모아 월 평균 매출, 골목 상권 수, 3년 이상 생존율, 리뷰개수를 통한 거래 증감률, 매출액 증감률, 폐업률, GRDP(지역내총생산) 등 재무적 지표와 비재무적 지표(리뷰개수) 상권분석을 수행한다.
식신 데이터의 감성분석을 통해 긍/부정 비율을 구해 행정동 코드별로 시각화 한다.

1 변수 선택하기

상권분석에 필요한 X값 찾기



2 머신러닝 / 선형?

머신러닝을 통한 상권분석?
단순 선형식을 통한 상권분석?



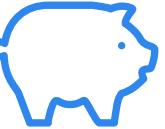
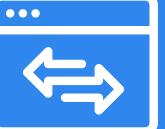
3 상권분석 결과 점수화

행정동 코드별로
상권분석결과 점수화



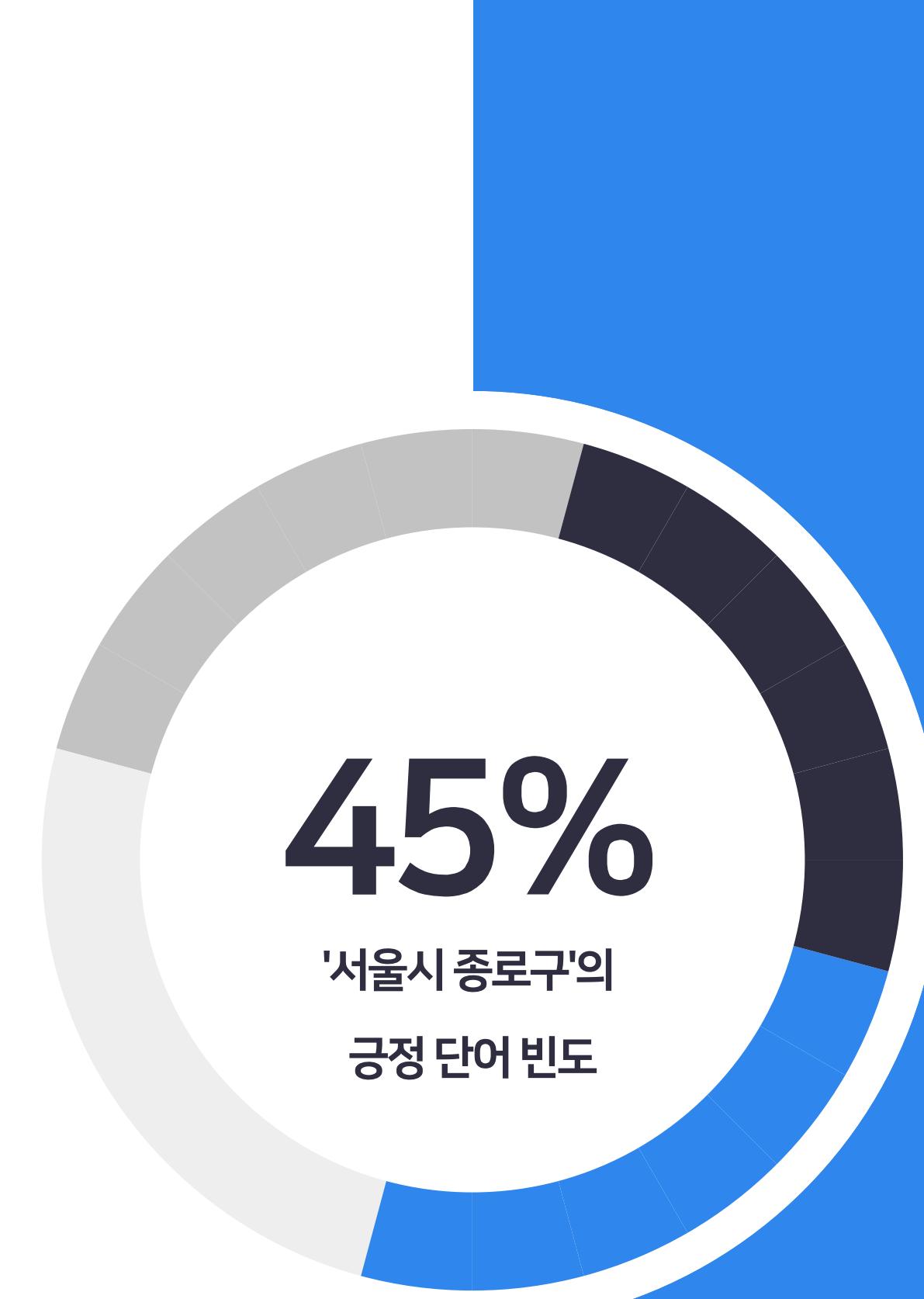
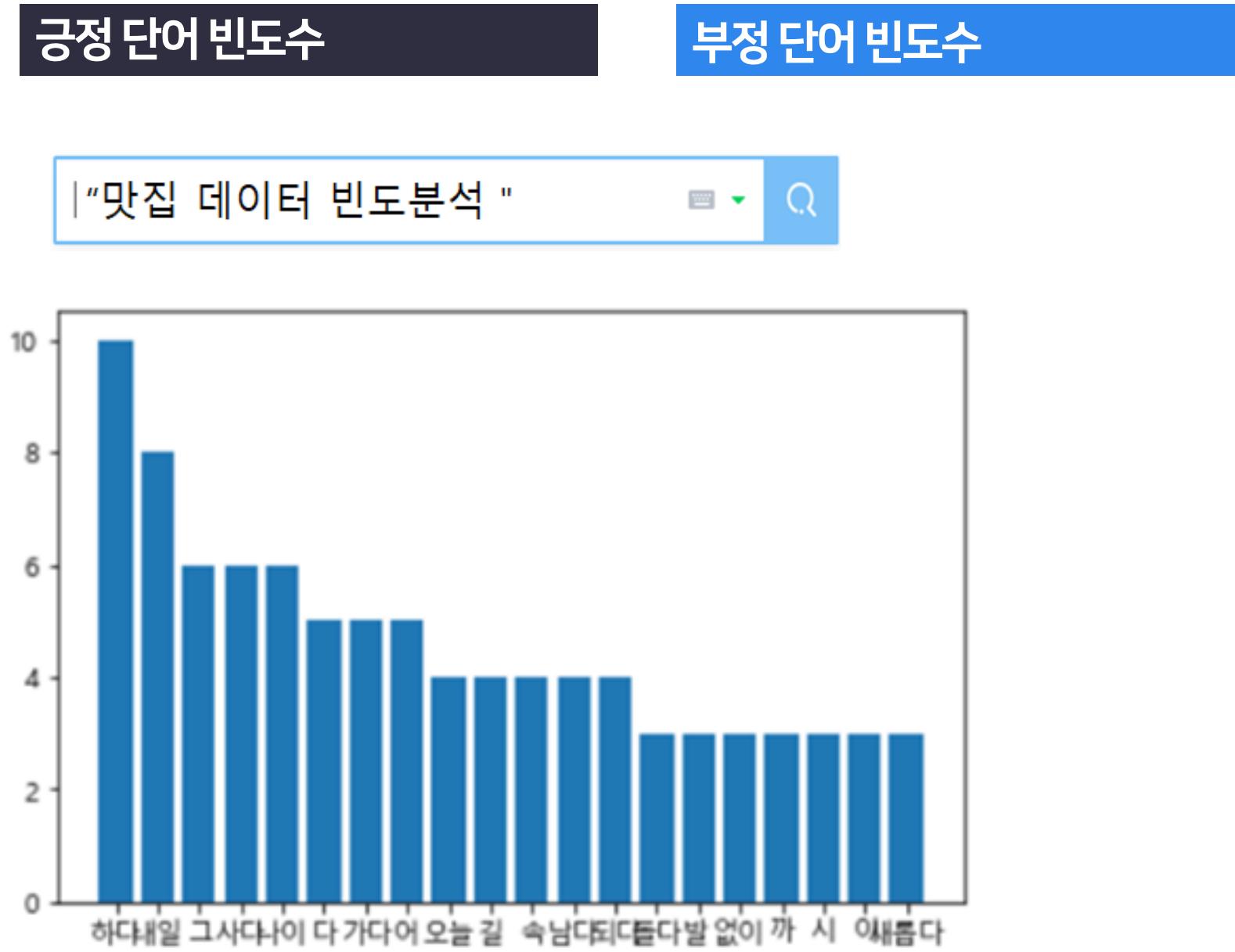
5. 상권분석

요식업 && 행정동 코드별로 데이터를 모아 월 평균 매출, 골목 상권 수, 3년 이상 생존율, 거래 증감률, 폐업률, GRDP(지역내총생산) 등 재무적 지표로 상권분석을 수행한다.

	월 평균 매출 행정구 코드별 음식점업을 중심으로 추정 매출 통계를 구함		거래 증감률 네이버 리뷰 개수의 변화를 크롤링 하여 행정구 코드별로 거래 증감률을 구함 (단위 : 년)		매출액 증감률 사업장 데이터를 행정동 코드별로 구분하여 년간 매출액의 증감률을 구함		폐업률 사업장 데이터의 2017년 ~ 2020년 기간 동안의 폐업률을 계산해 상권 안정도를 측정		GRDP 각 시도내에서 얼마 만큼의 부가가치가 발생되었는가를 나타냄
---	---	---	--	---	--	---	---	---	---

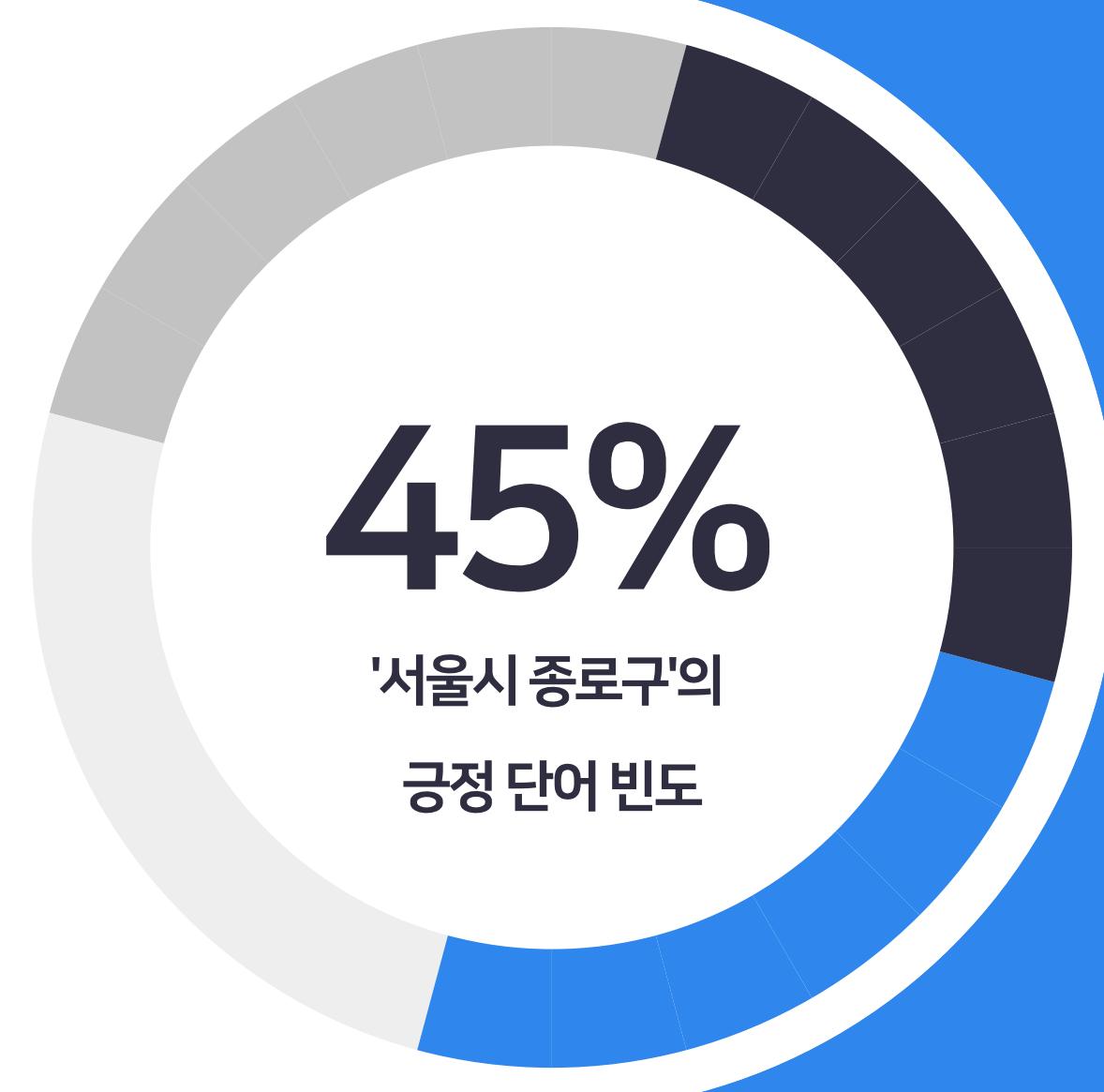
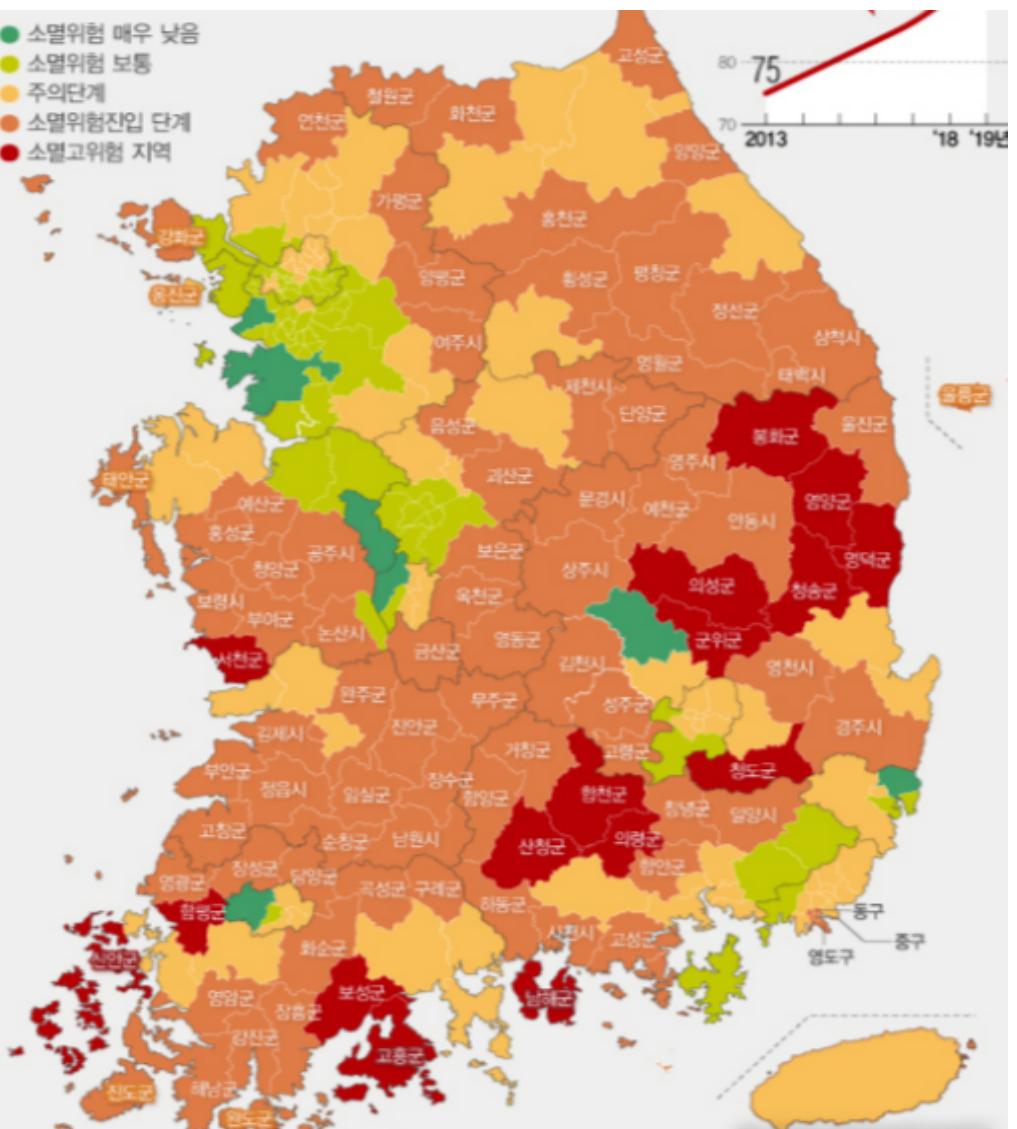
5. 상권분석

식신데이터의 감성분석을 통해 긍/부정단어 빈도를 측정하고
대한민국 지도를 그려 시각화한다.



5. 상권분석

식신데이터의 감성분석을 통해 긍/부정단어 빈도를 측정하고
대한민국 지도를 그려 시각화한다.



variable choice

6. 변수 선장

자료_은행
개인 자료 구문코드
사업체 구문코드
자료종류 기준
신문도록
책과 같
한글 및 무한한 보급여부
작가 저작 여부
주제요약
(주제) 불꽃군 페스티벌
(주제) 무한한 자산보유수
(주제) 무한한 자산공식규격
대출보증(기본)수
불꽃제작기록
신문대출작성
주제급보대출작성
주제와급보대출작성
기타대출작성
불꽃군 신문과급보작성
불꽃군 산화기자급보작성
불꽃보급보대출미사용
불꽃군 투표가드미분급액
불꽃군 투신문가드미분급액
불꽃군(신문가드)의시불이분급액
불꽃군(신문가드)할부이분급액
CDA
CSIR
대출면세금액(단기)
대출면세금액(중기)
대출면세금액(장기)
카드면세금액(단기)
카드면세금액(중기)
카드면세금액(장기)
불꽃로켓별매매가
불꽃로켓별전세가
대출총기기수(미해자)(1개월내신규개설)(기밀예신)
전통문화극장선물대출개설일자(선풀인증)(장기)(기타자금부록)(1개월내신규개설)(기밀예신)
대출총기기수(미해자)(기밀예신)
지속가능한환경대출총기기수(미해자)(기밀예신)
지속가능한환경대출총기기수(미해자)(기밀예신)
환경자금(신문등보급체계적금액)(대출총액증급액)(1개월내신규개설)(미해자)
지속가능한환경대출총기기수(미해자)(기밀예신)
대출총액증급액(미해자)(기밀예신)
대출총기기수(미해자)(기밀예신)
불꽃보급보대출총기기수(미해자)(기밀예신)
지속가능한환경대출총기기수(미해자)(기밀예신)
대출총기기수(미해자)(기밀예신)
불꽃로켓별총대출총기기수(미해자)(기밀예신)
면세기기수(미해자)(별인카드)
면세금액(미해자)(별인카드)
면세금액(미해자)(기밀예신)(별인카드)
미해자면세총기기수(기밀예신)(별인카드)
면세기기수(미해자)(별인카드)
최장면세총기수(액제크림)(기밀예신)(사업자카드)
면세제세총기수(1년내신도자)(액제크림)(기밀예신)(사업자카드)
최장신문가드개설일자로부터 최과세기수(별인카드)
1개월전(15일)카드일시를최고총미분급액(별인카드)
1년내(15일)카드일시를최고총미분급액(별인카드)
카드총이분급액(별인카드)
불꽃카드기기수(별인카드)
스폰서업체기수(액제크림)
최장면세기자료로부터 최과세기수
사업자별부정면세증명으로자본보유증명(영업인)

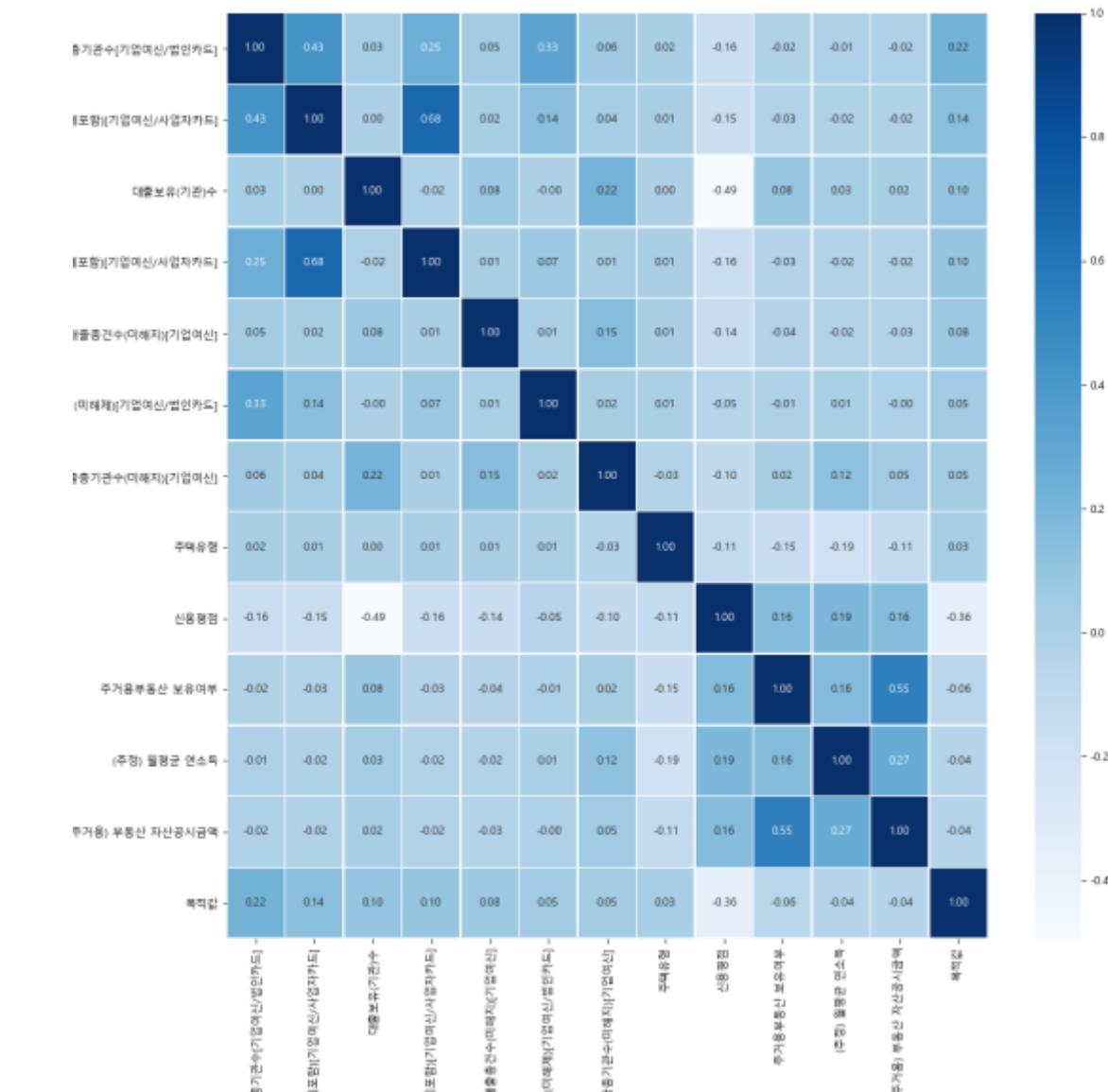
독립 변수 목록 약 40개

종속변수 : 목적값(연체여부)

0 : 연체 무

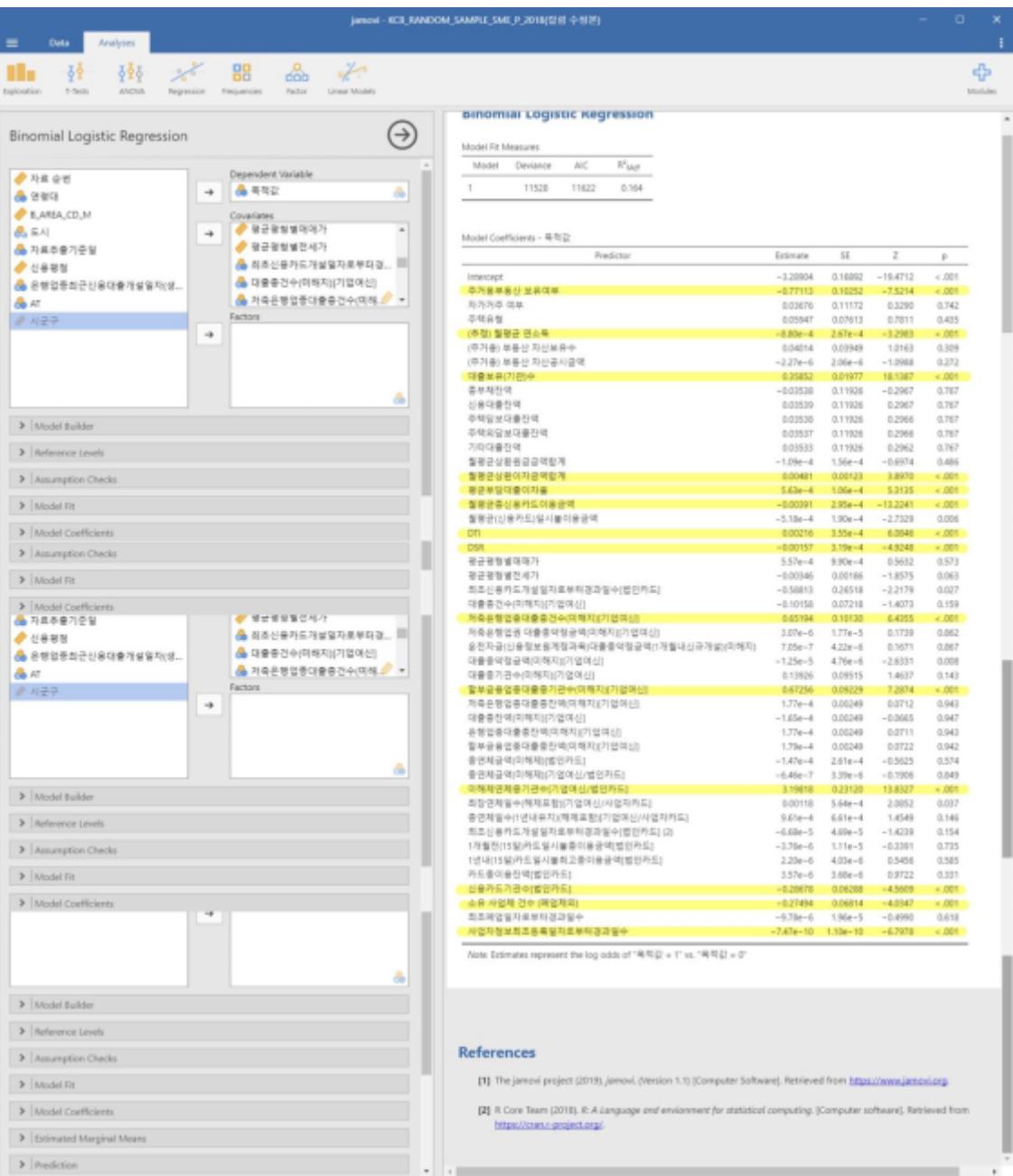
1 : 연체 유(90일 이전)

2 : 연체 유(90일 이후)



상관관계 확인

6. 변수 선정



자모비를 이용해 p-value 구하기

종속변수 Y를 목적값으로 두고 '자모비' 프로그램으로
적절한 X값을 찾는다.

1. P-Value < 0.001 인 X 후보 구하기

2. Stepwise Selection 으로 X 후보 구하기

단계적 선택법으로 추려냄

Step: AIC=8394.36

목적값 ~ 신용평점 + `주거부동산 보유여부` + 월평균상환이자금액합계 + 평균부당대출이자율 + 월평균상환카드이용금액 + DTI + DSR + 평균평형별매가 + `최초신용카드개설일자로부터경과일수 [법인카드] ... 25` + `저축은행연중대출총건수 (미해지) [기업여신]` + `대출총액정점액 (미해지) [기업여신]` + `대출총기관수 (미해지) [기업여신]` + `할부금융연중대출총기관수 (미해지) [기업여신]` + `저축은행연중대출총잔액 (미해지) [기업여신]` + `은행연중대출총잔액 (미해지) [수]` + `할부금융연중대출총잔액 (미해지) [기업여신]` + `총연체금액 (미해제) [기업여신]` + `미래체연체총기관수 [기업여신/법인카드]` + `최장연체일수 (해제포함) [기업여신]` + `소유 사업체 건수 (폐업체외)` + `최초폐업일자로부터경과일수` + `사업자정보최초등록일자로부터경과일수`

6. 변수 선정



3. 다중공선성 판단하기

평균부담대출이자율, 월평균 신용카드 이용금액
DSR, DTI

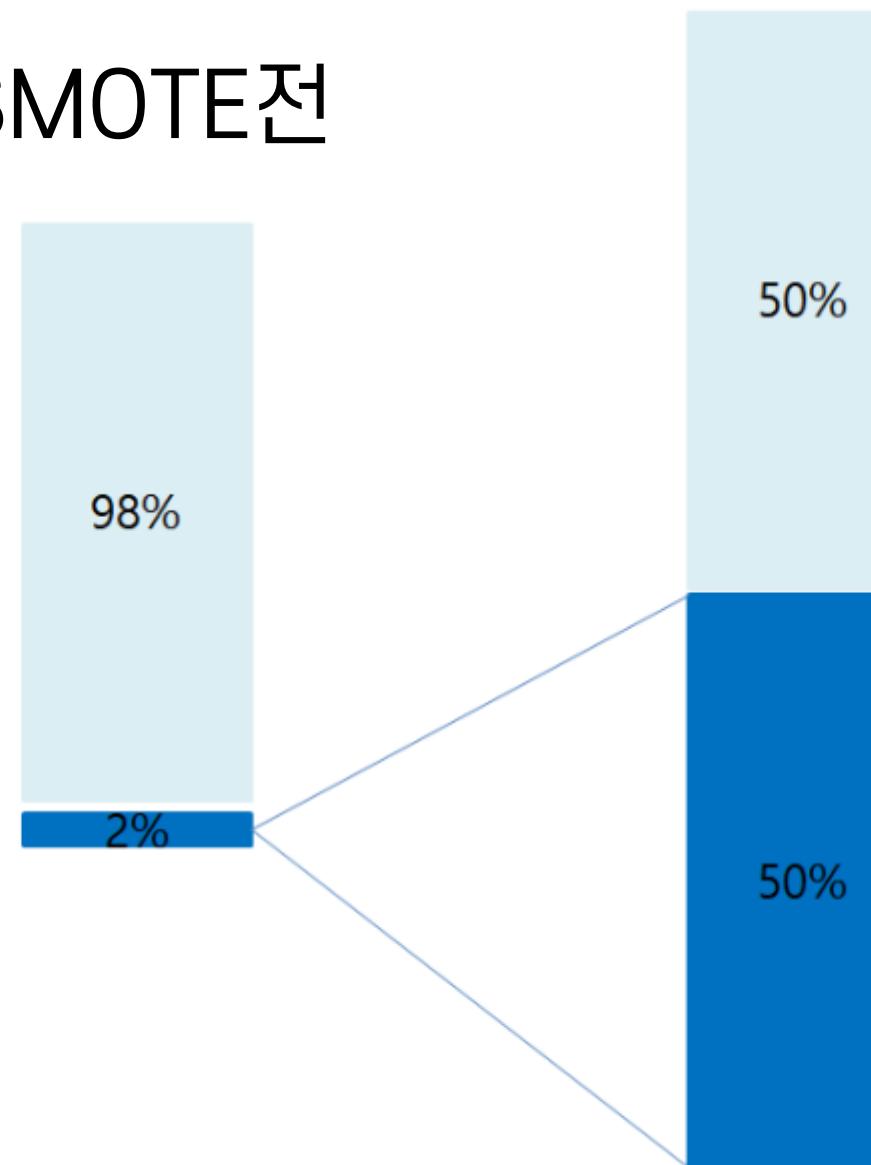
최종 변수

'목적값','주거용부동산 보유여부','(추정) 월평균 연소득','대출보유(기관)수', '월평균상환이자금액 합계','평균부담대출이자율', '월평균총신용카드이용금액','DTI','DSR', '저축은행업종대출총건수 (미해지)[기업여신]', '할부금융업종대출총기관수(미해지)[기업여신]'

7. 모집단 수 조정

SMOTE후

SMOTE전



SMOTE 기법을 활용하여
데이터 불균형 해결

In [58]: 1 data_1

Out [58]:

	주거용부동산 보유여부	월평균상환원금금액합계	평균부담대출이자율	월평균총신용카드이용금액	DTI	DSR	저축은행업종대출총잔액(미해지)[기업여신]	할부금융업종대출총기관수(미해지)[기업여신]
0	0	0	386	82	117.3	117.3	0	0
1	0	0	0	0	0.0	0.0	0	0
2	0	667	150	150	36.4	36.4	0	0
3	1	131	129	106	107.1	221.5	0	0
4	0	0	258	92	72.8	110.6	11700	1
...
95764	1	363	567	154	133.2	147.7	0	0
95765	1	26	168	65	37.6	40.7	0	0
95766	1	0	402	261	0.1	45.4	0	0
95767	0	0	189	185	0.0	0.0	0	0
95768	1	0	16	16	28.2	28.2	0	0

95769 rows × 9 columns

SMOTE 전

In [59]: 1 X

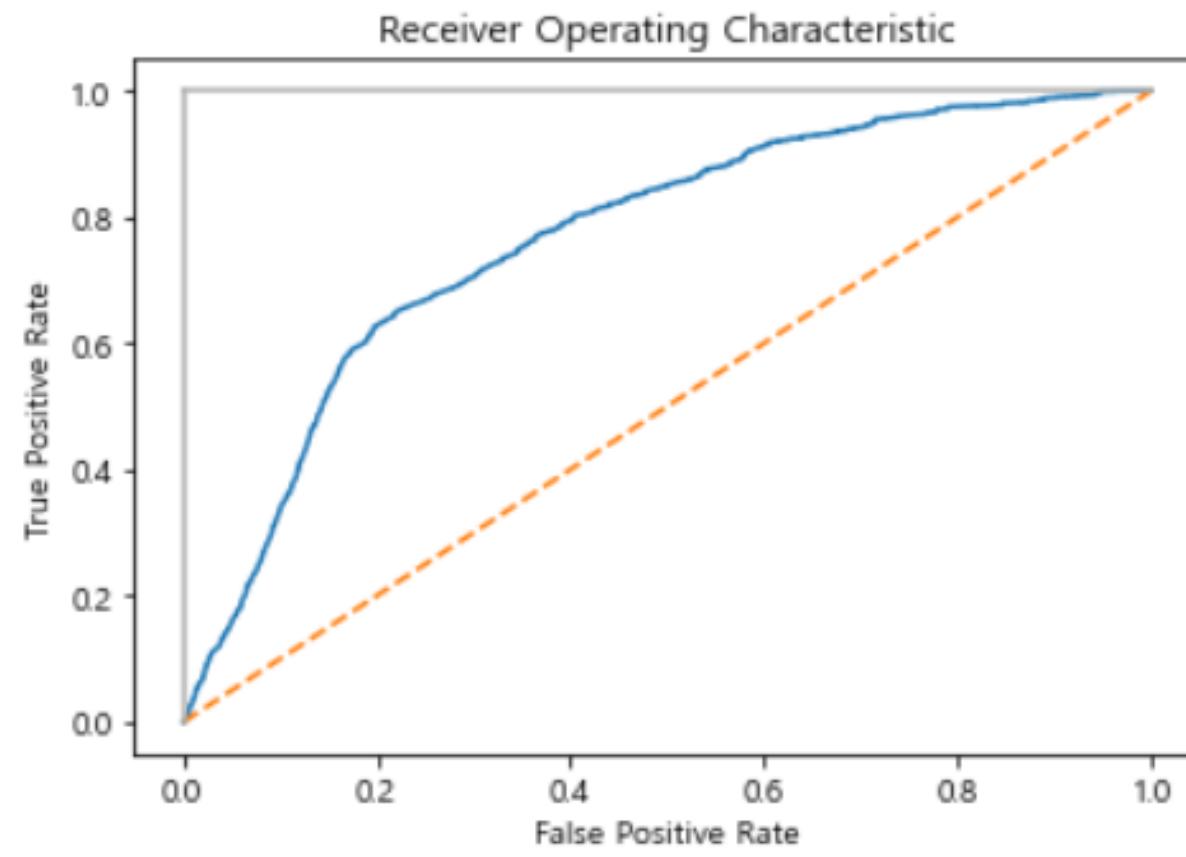
Out [59]:

	주거용부동산 보유여부	월평균상환원금금액합계	평균부담대출이자율	월평균총신용카드이용금액	DTI	DSR	저축은행업종대출총잔액(미해지)[기업여신]	할부금융업종대출총기관수(미해지)[기업여신]
0	0	0	386	82	117.300000	117.300000	0	0
1	0	0	0	0	0.000000	0.000000	0	0
2	0	667	150	150	36.400000	36.400000	0	0
3	1	131	129	106	107.100000	221.500000	0	0
4	0	0	258	92	72.800000	110.600000	11700	1
...
187273	0	0	75	0	109.995877	87.862630	0	0
187274	0	0	0	0	377.376925	131.092884	0	0
187275	0	327	213	138	38.000000	54.600000	0	0
187276	0	79	389	272	54.200000	87.800000	0	SMOTE 후 1
187277	0	55	291	67	25.300000	26.200000	0	0

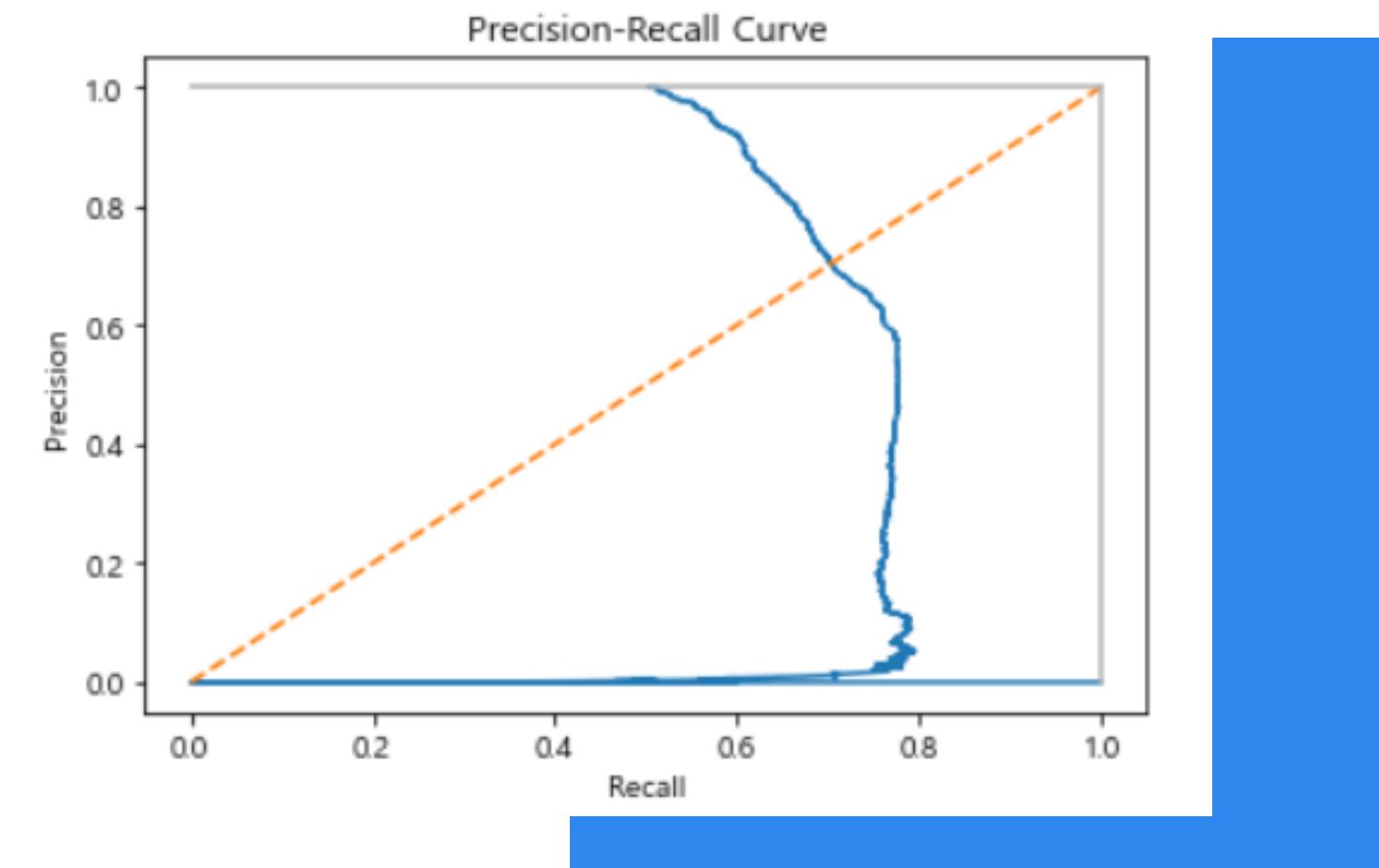
187278 rows × 9 columns

8. 모델링(1)

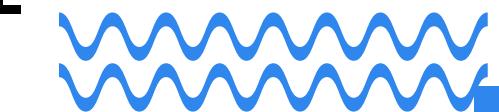
1) Logistic regression



ROC곡선



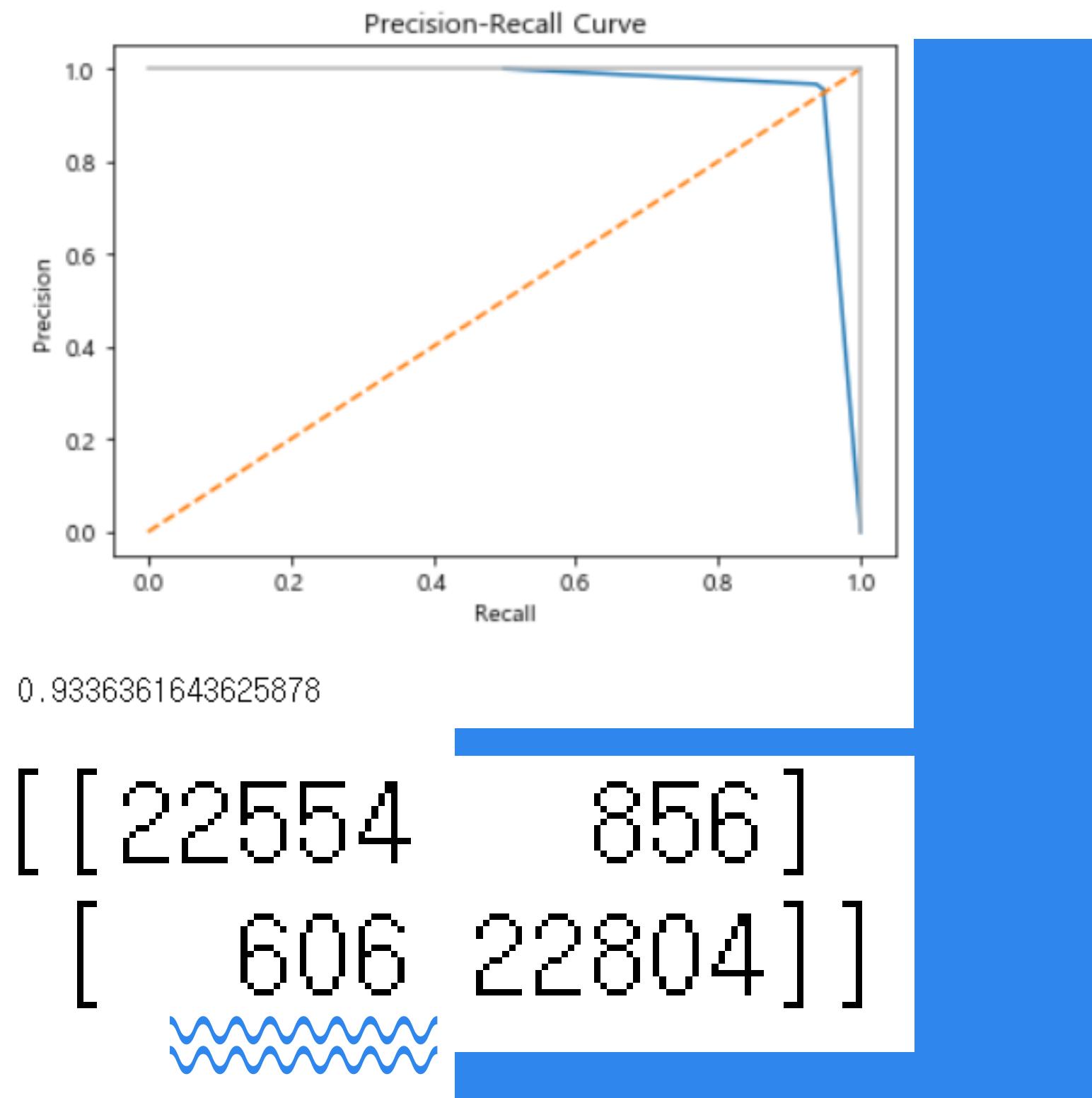
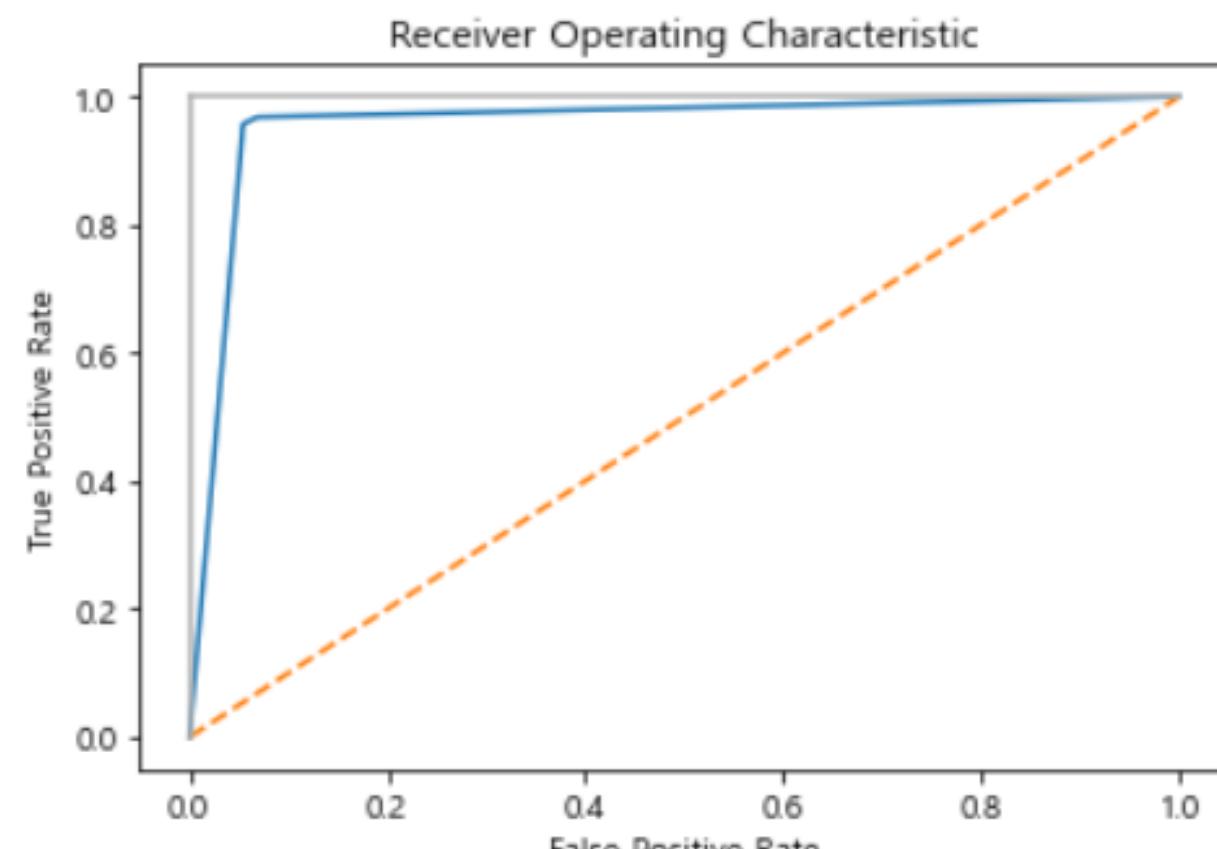
[[13674 9736]
[4712 18698]]



=> 정확도 : 69.1%

8. 모델링(1)

2) Random Forest



=> 정확도 : 96.8%

8. 모델링(1)

3) XGBoost

```
[[ 19939  3471 ]  
 [ 2281 21129 ]]
```

=> 정확도 : 87.7%

8. 모델링(2)

해야할 모델링 !

1. DNN 등 다른 모델들을 놓고 돌려보기
2. Clustering을 통해 신용등급 나누기

```

1 import flask
2 from flask import Flask, request, render_template
3 from sklearn.externals import joblib
4 import numpy as np
5 from scipy import misc
6
7 app = Flask(__name__)
8
9
10 # 메인 페이지 라우팅
11 @app.route("/")
12 @app.route("/index")
13 def index():
14     return flask.render_template('index.html')
15
16
17 # 데이터 예측 처리
18 @app.route('/predict', methods=['POST'])
19 def make_prediction():
20     if request.method == 'POST':
21
22         # 업로드 파일 처리 분기
23         file = request.files['image']
24         if not file: return render_template('index.html', label="No Files")
25
26         # 이미지 평생 정보 읽기
27         # 알파 채널 값 제거 후 1차원 reshape
28         img = misc.imread(file)
29         img = img[:, :, :3]
30         img = img.reshape(1, -1)
31
32         # 입력 받은 이미지 예측
33         prediction = model.predict(img)
34
35         # 예측 값을 1차원 배열로부터 확인 가능한 문자열로 변환
36         label = str(np.squeeze(prediction))
37
38         # 숫자가 10일 경우 0으로 처리
39         if label == '10': label = '0'
40
41         # 결과 리턴
42         return render_template('index.html', label=label)
43
44
45 if __name__ == '__main__':
46     # 모델 로드
47     # ml/model.py 선 실행 후 생성
48     model = joblib.load('./model/model.pkl')
49     # Flask 서비스 스타트
50     app.run(host='0.0.0.0', port=8000, debug=True)

```

Flask를 이용한 머신러닝 코드 연동

main.py

```
1 import scipy.io
2 from sklearn.utils import shuffle
3 from sklearn.ensemble import RandomForestClassifier
4 from sklearn.model_selection import train_test_split
5 from sklearn.externals import joblib
6
7 # Google 주소 숫자 인식 모델 생성
8
9 # 로드 mat 파일
10 train_data = scipy.io.loadmat('extra_32x32.mat')
11
12 # 학습 데이터, 훈련 데이터
13 X = train_data['X']
14 y = train_data['y']
15
16 # 매트릭스 1D 변환
17 X = X.reshape(X.shape[0] * X.shape[1] * X.shape[2], X.shape[3]).T
18 y = y.reshape(y.shape[0], )
19
20 # 셔플(섞기)
21 X, y = shuffle(X, y, random_state=42)
22
23 # 학습 훈련 데이터 분리
24 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.05, random_state=42)
25
26 # 랜덤 포레스트 객체 생성 및 학습
27 clf = RandomForestClassifier()
28 clf.fit(X_train, y_train)
29
30 # 모델 저장
31 joblib.dump(clf, '../model/model.pkl')
```

Colored by Color Scripter 

models.py

8. Web Service 제공 - Index 화면



8. Web Service 제공 - Main 화면

To Small Business

TSB

Home Scoring About Menu Team Gallery Contact

대출할 수 있을까?

SEAK THOUSE HTML CSS TEMPLATE

개인사업자 신용평가
요식업에 종사하는 개인
사업자에 대한 신용평가
모형

상권 분석
비재무적인 요소들을 넣어
만든 지역별 상권분석

Q

<상권 분석>

1. 5가지의 특징들을 설정하고 자료를 모으고 있는데 저 특징들을 각각 변수로 넣어야할지 아니면 특징들을 파악해 한 변수로 모델에 넣어야할지 궁금합니다.
2. 상권 분석 사이트가 있어서 거기서 각 구별 상권의 등급을 파악해 엑셀 파일로 남겨뒀는데 머신러닝식으로 가야할지 궁금합니다.

<데이터>

1. 받은 데이터가 사업하는 개인에 대한 데이터입니다. 그치만 요식업인지 알 수 없어서 데이터를 요청한 상태입니다. 만약 사업장 코드를 받지 못하면 어떤 식으로 분석해야하는지 궁금합니다.
2. DSR, DTI가 9990이면 이상치인가요?
3. '월평균 연소득'의 정확한 뜻이 궁금합니다. 단위가 십만인데 1년인지 월인지 궁금합니다. 인터넷에 안나와요 ,,

<머신러닝>

1. 마지막에 검증 형식으로 ks통계량을 이야기 하셨는데 ks검증에서 나온 통계량 맞나요?
2. 변수의 ks통계량이 0.1이상이어야 한다고 하셨는데 변수를 선택할때 각각의 통계량을 이야기 하는건가요?