Data Mining Term Project 최종보고서

< 날씨와 배달데이터를 이용한 의사결정지원 모델>



**수업** : 데이터마이닝응용

**담당교수님** : 이석룡 교수님

**팀명** : 배달의 민조

**팀원** : 201703131 전종미

201701573 박지연

201501703 송윤범

201502860 장동혁

**<목차>**

1. 연구개요 및 목표
2. 데이터
   1. 데이터 수집방법
   2. 데이터 전처리
3. 배달음식 데이터 전처리
4. 날씨 데이터 전처리
5. 시간별 배달음식 데이터 전처리
6. 배달음식 데이터와 날씨 데이터 통합
7. 사용할 DM 기능
8. 기초분석
9. 성별, 지점, 연령대, 요일과 총 배달건수의 관계
10. 날씨와 총 배달건수의 관계
11. 시계열 변수와
12. 배달건수의 관계
13. 분석 결과
14. 의사결정나무
15. 신경망 분석
16. 시계열 분석
17. 기대효과
18. **연구개요 및 목표**

하루 평균 2500개의 자영업이 치열한 생존경쟁으로 인해 폐업이 되고 있는 현재, 이를 해결하기 위해 마케팅 컨설팅을 찾는 개인사업자들이 2011년(84만5235명)이후 5년만에 최고치를 기록하였다. 이에 따라 우리 팀은 우후죽순으로 생기는 배달음식점, 주문량의 부정확한 예측으로 인한 재료의 낭비, 날씨에 따른 주문량 차이를 고려하기 어려운 점 등과 같은 문제점이 경쟁시장에서 생존하기 위해서는 해결이 되어야 할 주요 문제점으로 인식을 하였다. 따라서 본 프로젝트는 여러가지 변수들과 배달 주문건수와의 관계를 파악해보고 날씨에 따른 배달음식 주문량의 변화를 예측하여 실제로 자영업자들이 활용 가능한 마케팅 전략을 제언하기 위해 본 모델을 구축해 나갈 예정이다.

1. **데이터**
   1. 데이터 수집방법

**기상자료개방포털**에서 서울지역으로 한정하여 2018년 10월부터 2019년 9월에 해당하는 방재기상데이터를 통해 기온, 풍속, 기압 등과 같은 데이터를 수집하였다. 또한 SK telecom에서 제공하는 **BIG DATA HUB**에서 같은 기간 동안에 해당하는 배달음식 데이터를 수집하였는데, 배달 업종별 데이터에서 성별과 나이대가 나타나 있지 않은 업종도 존재하여, 이 attribute까지 나타내는 치킨, 중국집, 피자에 관련한 데이터만 따로 수집하였다. 또한 시간에 따른 총 배달건수를 각 배달음식에 관하여 나타내기 위해 구별로 나누어진 것이 아닌 시간이 명시되어 있는 서울시 전체 통합 배달 데이터를 따로 수집하였고 시계열 분석을 위하여 4년치 치킨, 피자, 중국음식 데이터를 수집하였다.

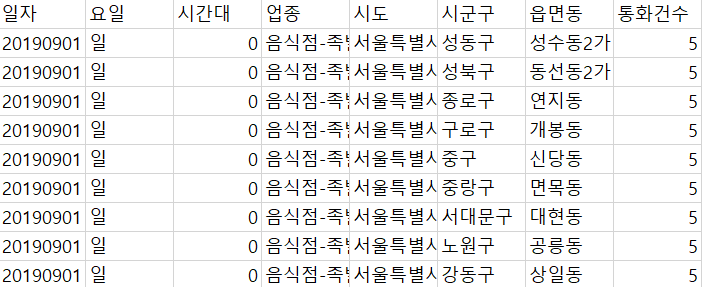
< **기상자료개방포털 –** 날씨데이터**>**



<**SK Telecom BIG DATA HUB –** 구별 배달건수 데이터>



<**SK Telecom BIG DATA HUB –** 시간별 배달건수 데이터>

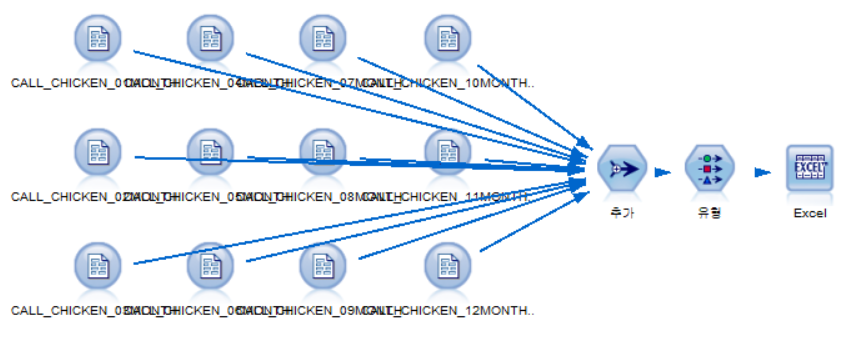


* 1. 데이터 전처리

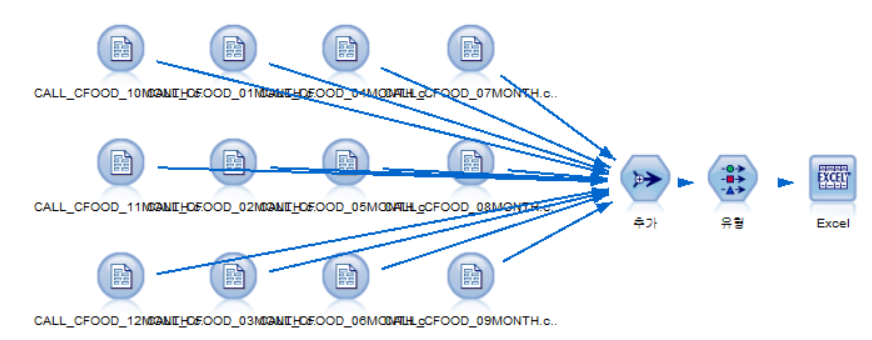
1. 배달음식 데이터 전처리

치킨, 중국음식, 피자 데이터가 각각 월별로 다른 파일로 수집되었기 때문에 추가 노드를 사용하여 하나의 파일로 통합하였다.

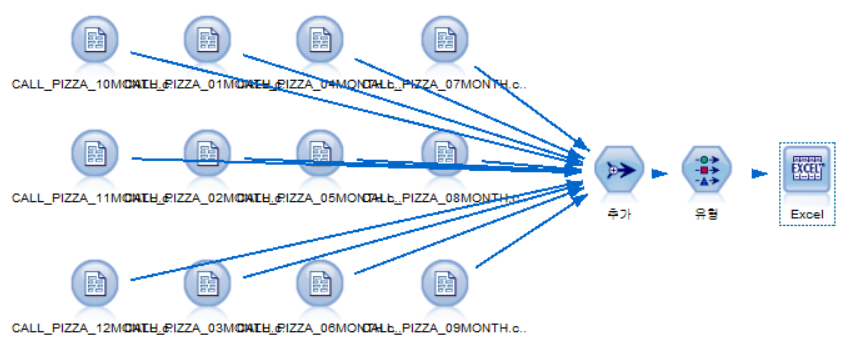
<월별 치킨 데이터 통합>



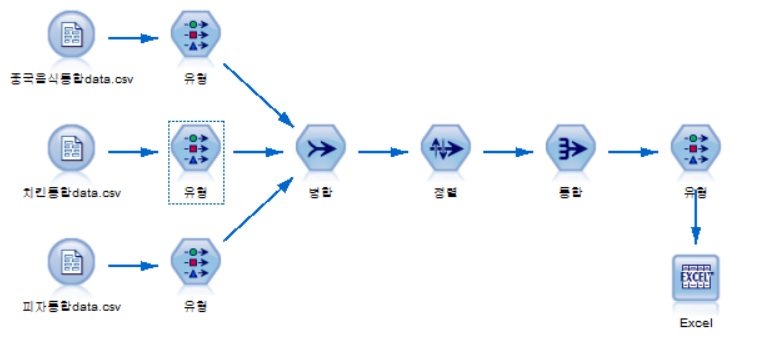
<월별 중국음식 데이터 통합>



<월별 피자 데이터 통합>

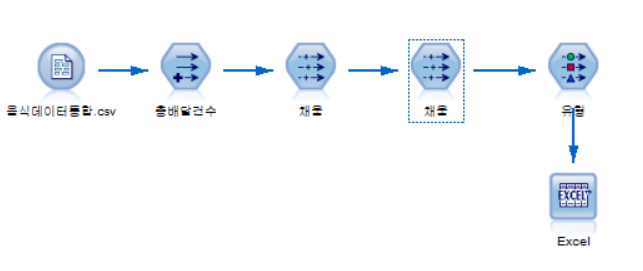


각 음식 데이터들의 통화건수 변수를 중국음식, 치킨, 피자통화건수로 변수명을 바꾸었고 이 데이터들을 통합하기 위해 병합노드를 사용하였다. 또한 읍면동으로 나누어져 있는 데이터를 구별로 나타내기 위하여 통합노드를 사용하여 전처리하였다.



중국음식, 치킨, 피자의 통화건수를 통합한 전체 배달음식 데이터와 날씨 데이터와의 관계를 파악하기 위해 파생노드를 사용하여 총 배달건수 attribute를 만들었고 데이터를 병합했을 때 생기는 null 값을 ‘0’ 으로 처리하기 위해 채움노드를 사용하였다.

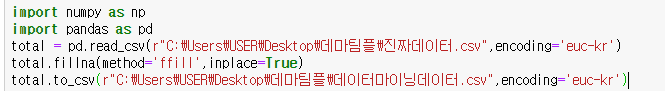
* 배달음식 raw 데이터를 봤을 때 배달건수가 0인 것은 존재하지 않고 데이터가 없는 경우 행 자체가 존재하지 않기 때문에 배달음식을 시키지 않아 데이터가 없는 것으로 판단하여 null 값을 ‘0’ 으로 설정하였다.



1. 날씨 데이터 전처리

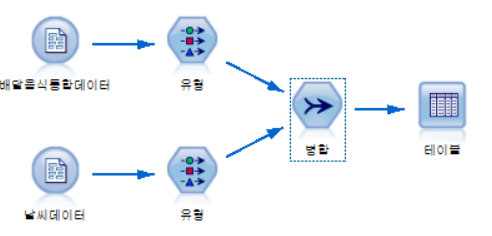
날씨 데이터와 배달음식 데이터의 통합을 위해 숫자로 나와있는 지점 code를 배달음식 데이터와 같이 구명으로 전환하였다. 날씨 데이터는 관측소 별로 나누어져 있어 기상청, 한강, 북악산 등 구에 관한 데이터가 아닌 것들은 분석 목표에 맞지 않기 때문에 사용할 데이터에서 제외하였다.

날씨 데이터의 null 값 처리를 위해 파이썬 코드를 사용하였다.



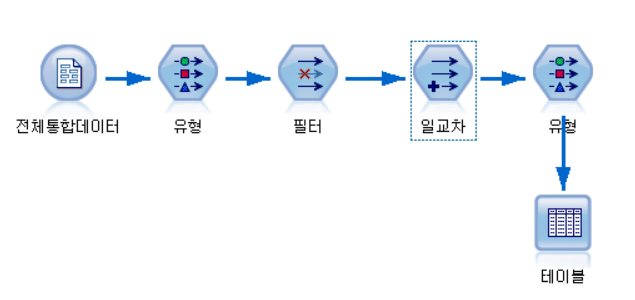
1. 배달음식 데이터와 날씨 데이터 통합

전체 데이터를 한번에 나타내기 위해 배달음식 데이터와 날씨 데이터를 통합하였다.

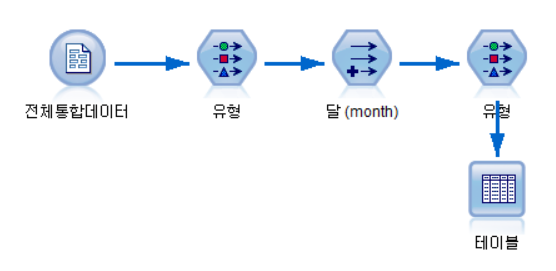


또한 분석에 사용하기 위해 추가적으로 일교차 변수, 시계열 변수인 달(month)를 생성하였고 사용하지 않을 변수를 제거하여 데이터를 생성하였다.

<일교차 변수 생성>

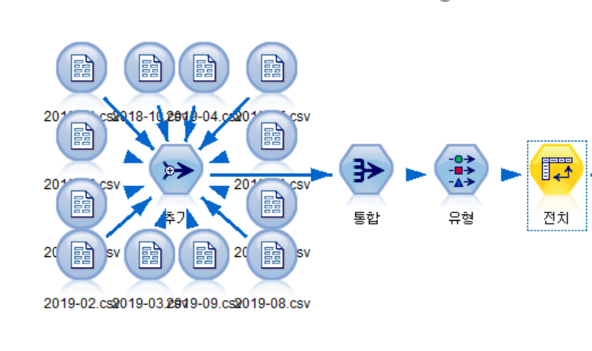


<달(month) 변수 생성>



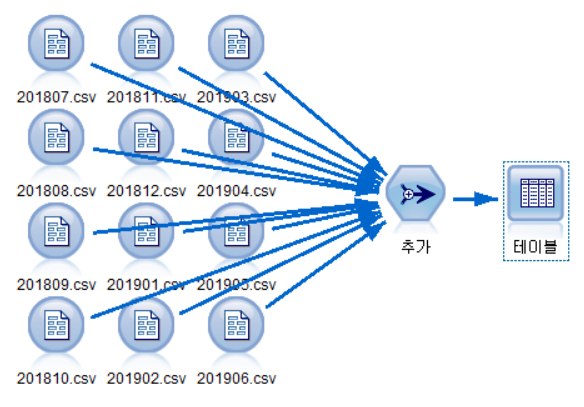
1. 시간별 배달음식 데이터 전처리

시간에 따른 배달음식 건수를 배달 음식별로 나타내기 위해서 월별로 데이터를 통합하였고 전치노드를 사용하여 업종 attribute로 되어있는 배달음식 데이터를 각 배달음식 데이터별로 attribute를 생성하였다.

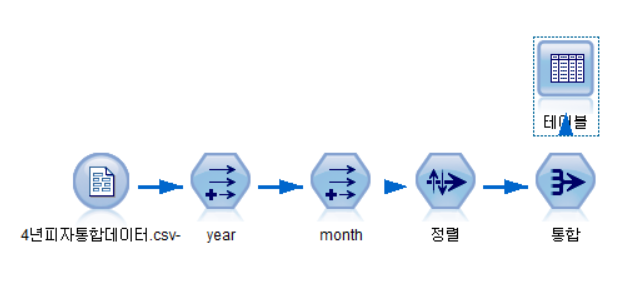


1. 시계열 데이터 전처리

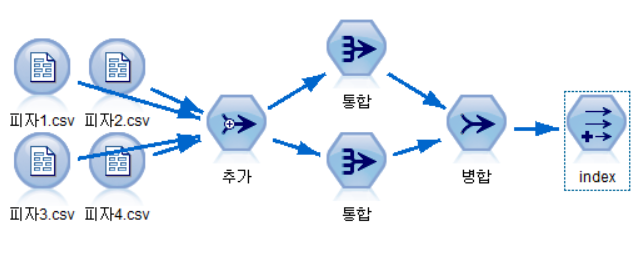
시계열 분석을 위한 치킨, 피자, 중국음식의 4년치 데이터를 통합하기 위해 추가노드를 사용하였다.



시계열 분석에 필요한 year, month 변수를 생성하였고 필요하지 않은 데이터들은 삭제하여 최종 시계열 데이터를 생성하였다.



시계열 분석에 필요한 index 값을 계산하기 위하여 파생노드를 사용하여 index 변수를 생성하였다.

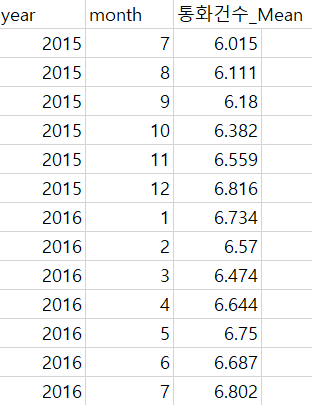


* 치킨과 중국음식도 같은 방식으로 시계열 데이터를 생성하였다.

1. 사용할 최종 데이터



* 최종 데이터를 만든 후 분석에 사용하지 않을 최저 기온 시각(hhmi), 최고 기온 시각(hhmi), 최대 순간 풍속(m/s), 최대 순간 풍속 시간(hhmi), c최대 순간 풍속 풍향(deg) 변수를 삭제하였다.
* 최종적으로 각 행의 치킨, 피자, 중국음식 배달건수를 평균낸 총 배달건수의 평균, 일교차, 달(month) 변수가 추가 되었고 시간 순서대로 정렬하였다.



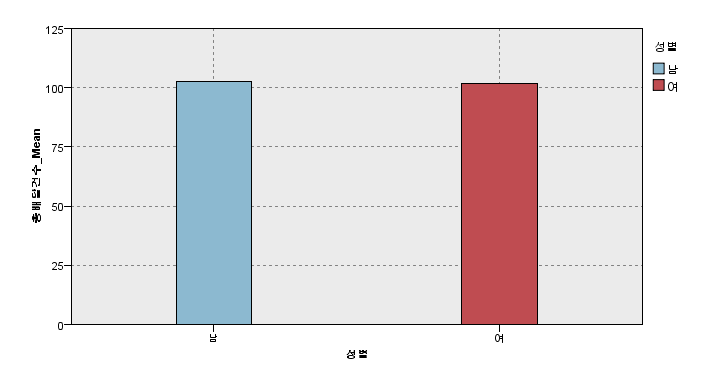
* 사용할 최종 데이터 :

1. 의사결정 나무와 신경망 분석에 사용할 날씨 + 배달데이터
2. 시계열분석 데이터
3. 시간대별 배달 건수 데이터
4. 사용할 DM 기능

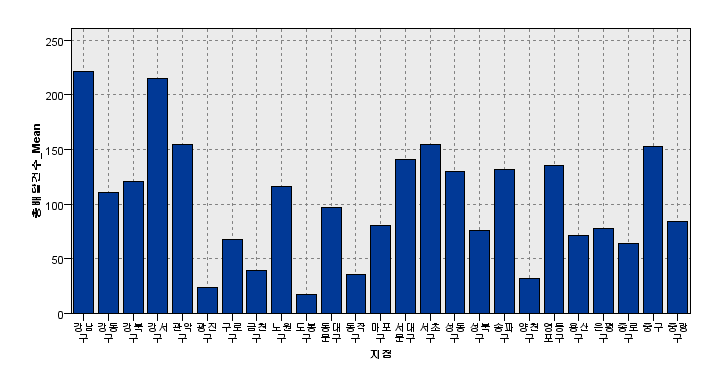
|  |  |
| --- | --- |
| **사용할 DM 기능** | |
| **의사결정나무** | 배달 주문건수에 영향을 미치는 주요 attribute 선정 |
| **시계열 분석** | 시계열 데이터로 미래 추이 예측 |
| **신경망 분석** | 날씨 데이터를 이용한 배달 건수 예측 |
| **군집 분석** | 음식에 대한 주요 고객에 대한 특성 파악 |

<표 : 사용 DM 기능>

1. 기초분석
2. 성별, 지점, 연령대, 요일과 총 배달건수의 관계



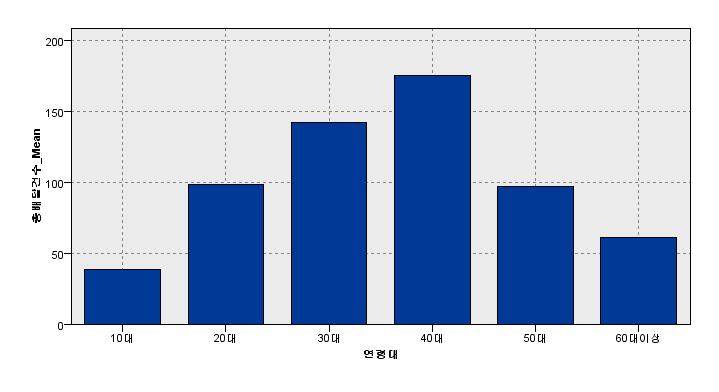
* 성별에 따른 총 배달건수의 평균은 차이가 미비함을 보여준다.



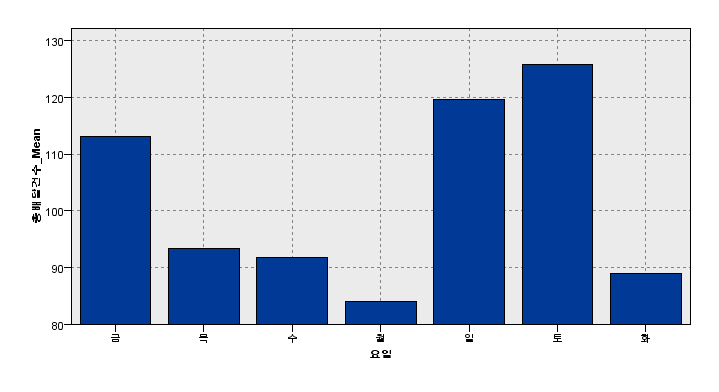
* 구에 따른 총 배달건수의 평균은 위와 같은 분포를 보여주고 있다.

강남구, 강서구에서는 높은 배달건수를 보이며

광진구, 도봉구에서는 낮은 배달건수를 보여주고 있다.

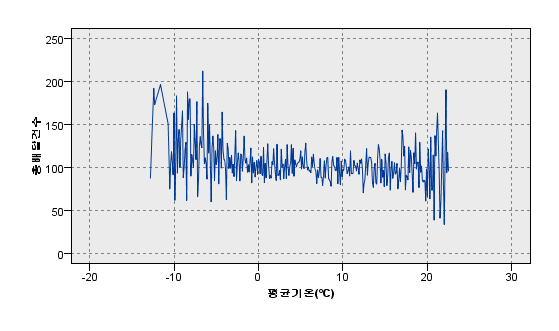


* 연령대에 따른 총 배달건수의 평균은 40대가 가장 높고 10대와 60대는 상대적으로 낮은 평균을 보여준다.

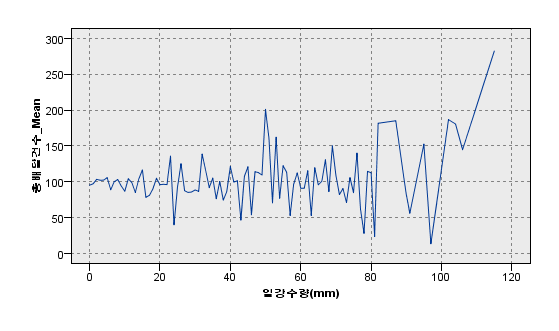


* 요일에 따른 총 배달건수의 평균은 금요일과 주말이 주중에 비해 특히 높은 것을 알 수 있다.

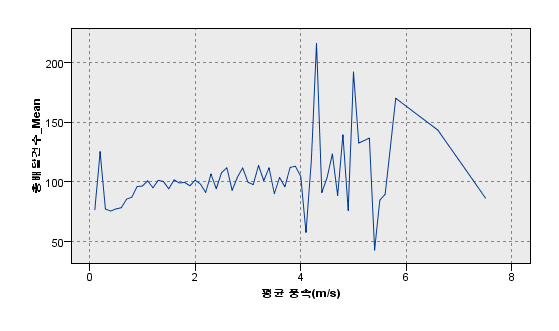
1. 날씨와 총 배달건수의 관계



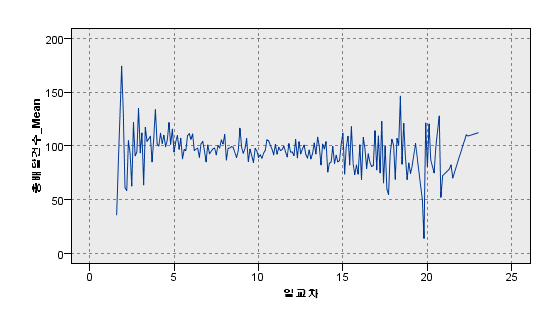
* 평균기온이 높거나 낮을 때 총 배달건수가 많아지는 추세를 보이고 그 외 기온에서는 특정 범위내의 수치를 유지하고 있다.



* 100mm 이상의 강수량에서 총 배달건수가 증가하는 추세를 보인다.

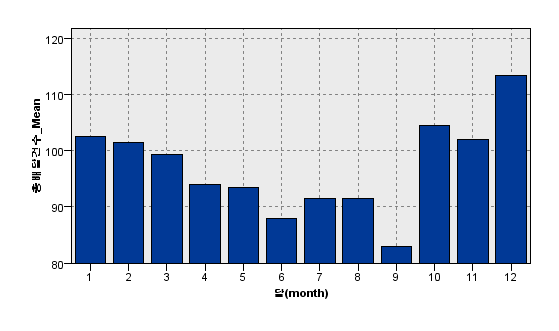


* 4 ~ 6 m/s 에서 총배달건수가 많은 풍속이 있지만 6 m/s 이상의 풍속에서는 감소하는 추세를 보인다.

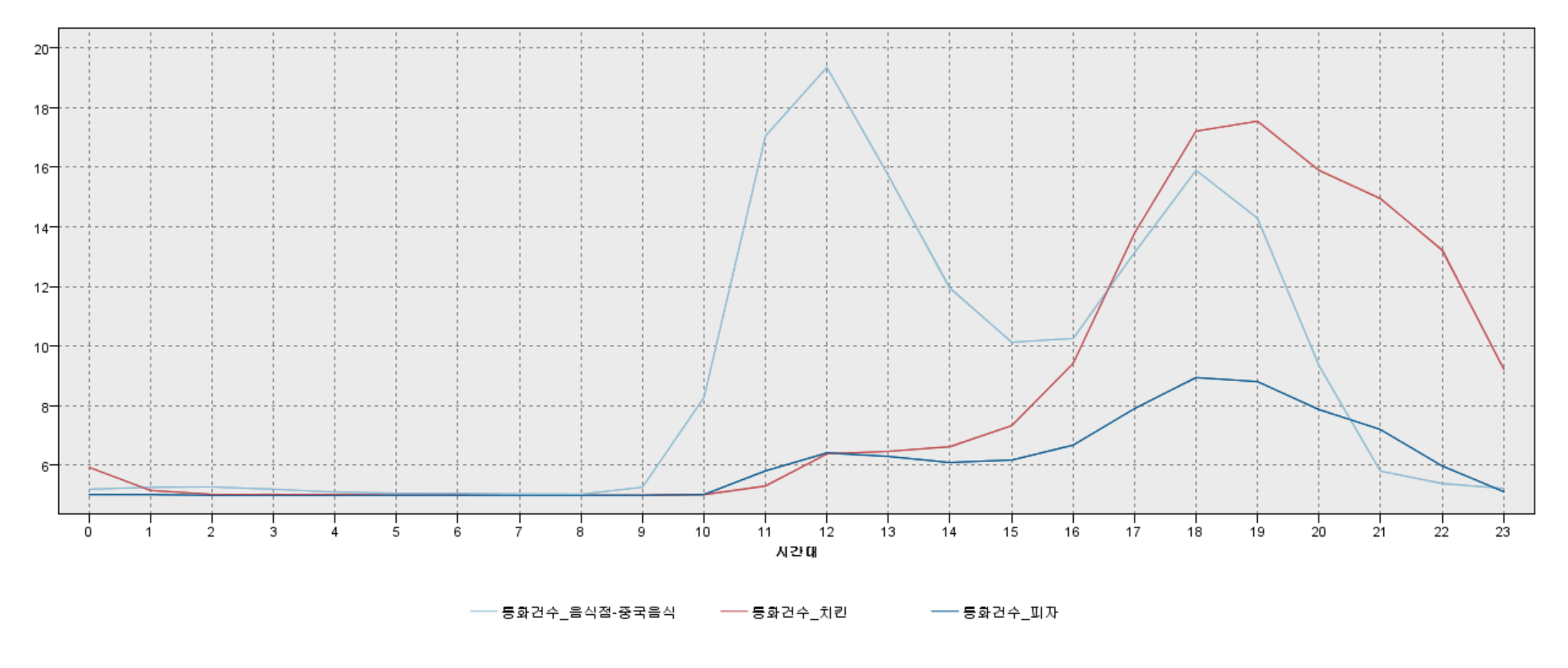


* 일교차가 커 저녁시간에 추워지게 되면 배달건수가 많아질 것이라는 예상과 달리 별 다른 상관관계를 보여주지 못하였다.

1. 시계열 변수와 배달건수의 관계



* 연말에 상대적으로 총 배달건수가 높은 것을 알 수 있다.

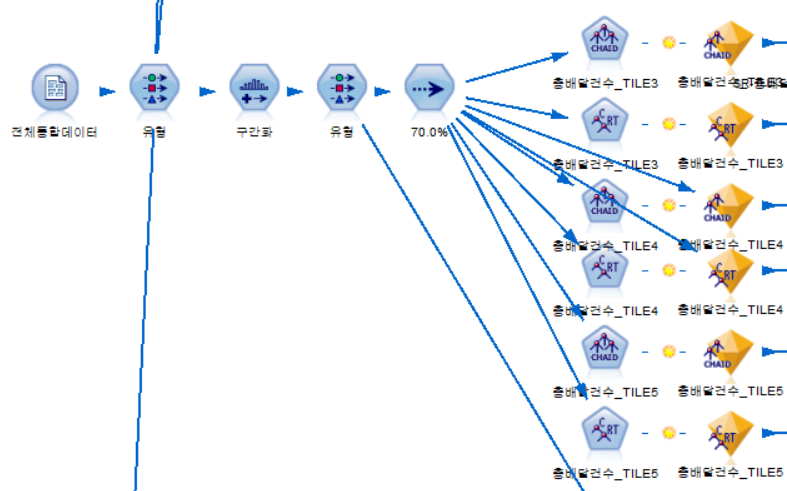


* 중국음식은 점심과 저녁시간 모두 배달건수가 많은 반면 치킨과 피자는 저녁시간에만 배달건수가 많은 것을 볼 수 있다. 피자는 두 음식과 달리 상대적으로 적은 판매를 보이고 있다.

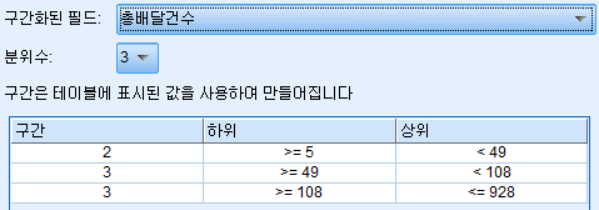
결과적으로 성별에 따라서는 뚜렷한 차이를 보이지 않고 강남구, 강서구의 경우 평균적으로 200건이 넘는 배달건수를 나타내고 있으나 광진구, 금천구, 도봉구, 동작구, 양천구와 같은 경우 현저히 낮은 수치를 보이고 있다. 따라서 광진구, 금천구, 도봉구, 동작구, 양천구의 배달 이용률을 증진시키는 방안을 모색해야 한다. 또한 10대는 경제적 능력이 부족하고 60대는 배달 음식 서비스 이용이 어려울 것으로 예상되기 때문에 10대와 60대의 배달 이용률이 낮은 것을 볼 수 있다. 따라서 배달 이용률이 높은 40대, 30대의 고객 유지에 힘쓰고 10대와 60대는 각 연령대에 맞는 마케팅 방안을 제시해야한다. 요일에 따라서는 금요일과 주말에 많은 분포를 보이고 평일은 상대적으로 적으므로 평일에 배달 서비스 이용률을 높여야 한다. 배달업체는 시간을 고려하여 각 시간대에 적합한 배달음식 서비스를 제공할 수 있어야 하며 날씨와 연관 지어 위와 같은 마케팅 전략을 모색해야 할 것이다.

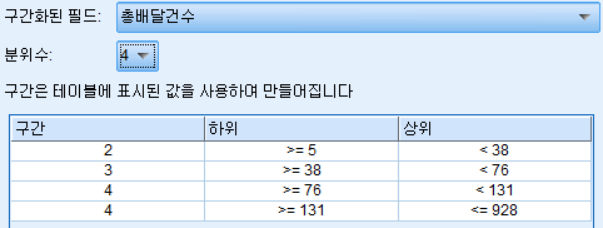
1. **분석 결과**
2. **의사결정나무**

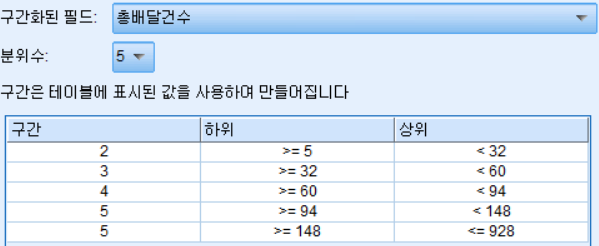
단순 그래프가 아닌 수치적 통계를 나타내 보고자 의사결정나무를 사용하여 어떤 변수가 총 배달건수에 가장 영향을 많이 미치는지를 분석하고자 한다.



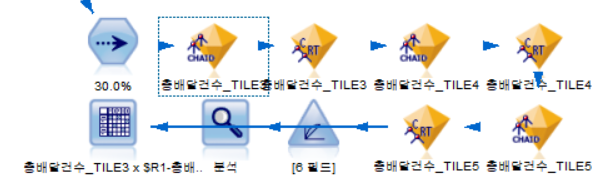
C&RT, CHAID 알고리즘을 사용하기 위해서 목표필드를 변환해야 한다. 구간화 노드를 통해 목표필드인 총 배달건수를 3분위, 4분위, 5분위로 구간화 하였다. 그리고 각각 지점, 성별, 요일, 연령대를 예측필드로 하여 Training Set (70%)에 대한 의사결정나무 분석을 실시하였다.

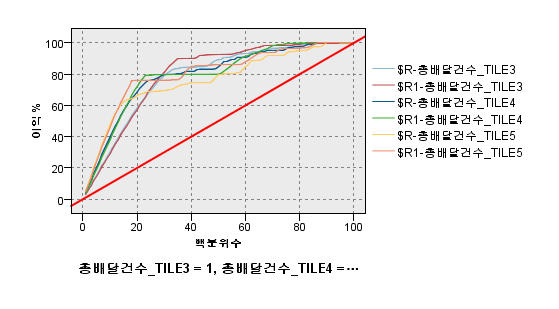




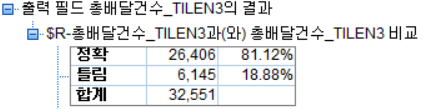


이후 Test Set(30%)에 모델을 적용하고 평가노드를 사용하여 예측 모델의 정확성을 비교하였다.

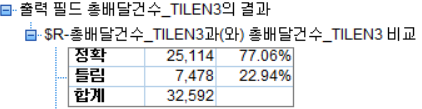




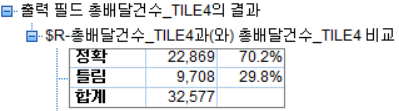
수치적으로 더 자세히 보여주기 위하여 분석노드를 사용하여 각 모델의 정확도를 비교해 보았다.



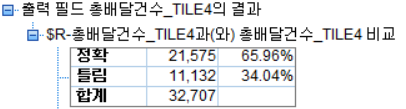
* 3분위 구간화 했을 때 C&RT\_ 총 배달건수의 정확도 : 81.12%



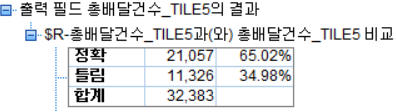
* 3분위 구간화 했을 때 CHAID\_ 총 배달건수의 정확도 : 77.06%



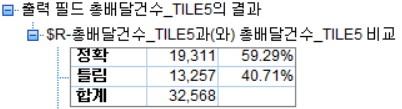
* 4분위 구간화 했을 때 C&RT\_ 총 배달건수의 정확도 : 70.02%



* 4분위 구간화 했을 때 CHAID\_ 총 배달건수의 정확도 : 65.96%

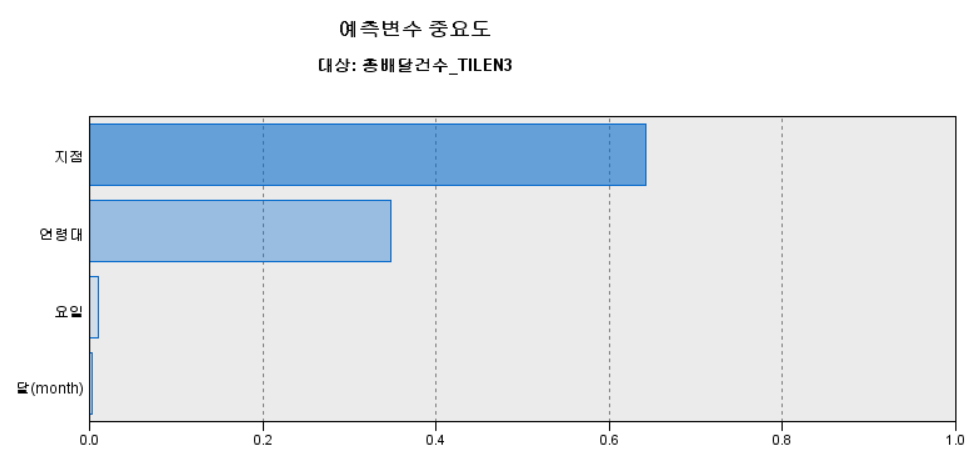


* 5분위 구간화 했을 때 C&RT\_ 총 배달건수의 정확도 : 65.02%



* 5분위 구간화 했을 때 CHAID\_ 총 배달건수의 정확도 : 59.29%

각각의 정확도를 비교했을 때 CHAID 알고리즘을 사용하는 것 보다 C&RT 알고리즘을 사용하는 것이 전체적으로 좋은 것을 볼 수 있다. 또한 총 배달건수를 3분위로 구간화한 것의 정확도가 81.12%로 가장 높으므로 3분위로 구간화한 C&RT 알고리즘을 사용하는 것이 적합하다. 이를 토대로 예측변수 중요도를 알아보고자 한다.

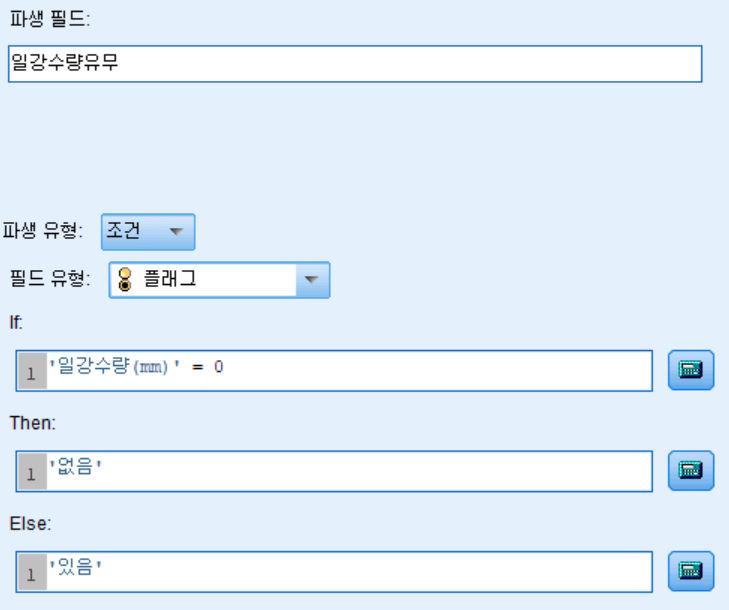


3분위로 구간화한 C&RT 알고리즘을 사용했을 때 예측변수 중요도는 지점과 연령대가 상대적으로 총 배달건수와 연관성이 높고 요일과 달(month)은 연관성이 낮음을 알 수 있다. 즉, 지점과 연령대를 고려하고 위에 제시한 도표그래프를 토대로 마케팅 방안을 제시해야 할 것이다.

1. **신경망 분석**

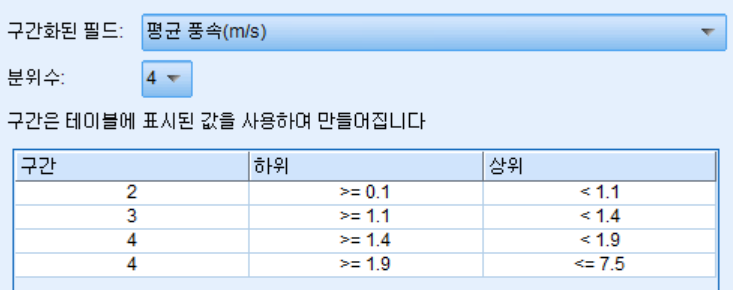
날씨데이터의 강수량, 풍속, 일교차, 평균기온을 예측필드로 하여 연속형 목표필드인 총 배달 건수를 각각 예측해보고자 한다. 먼저 예측필드 값들 중 강수량, 풍속, 일교차, 평균기온은 연속형이기 때문에 구간화를 사용하여 범주화를 시켰다.

<일강수량 구간화>

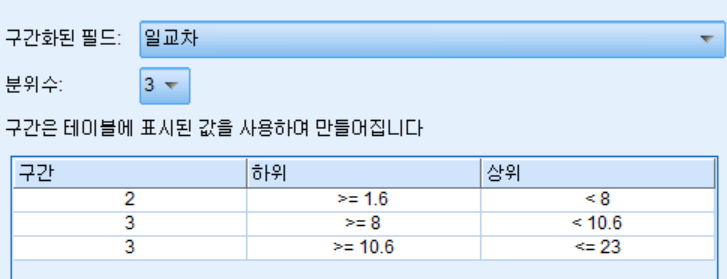


비가 내리지 않아 강수량 데이터가 ‘0’인 값이 많기 때문에 강수량은 구간화 하지 않고 이분형 데이터로 바꿔주었다.

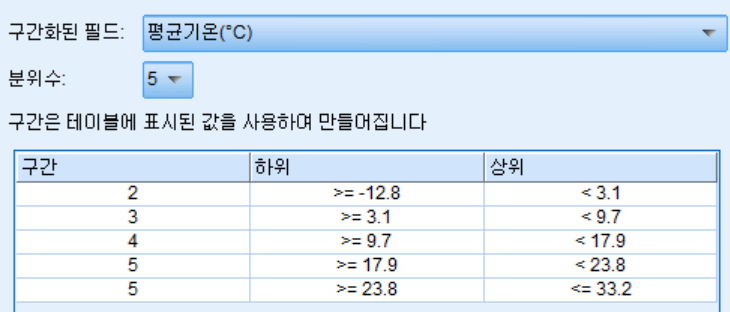
<풍속 구간화>



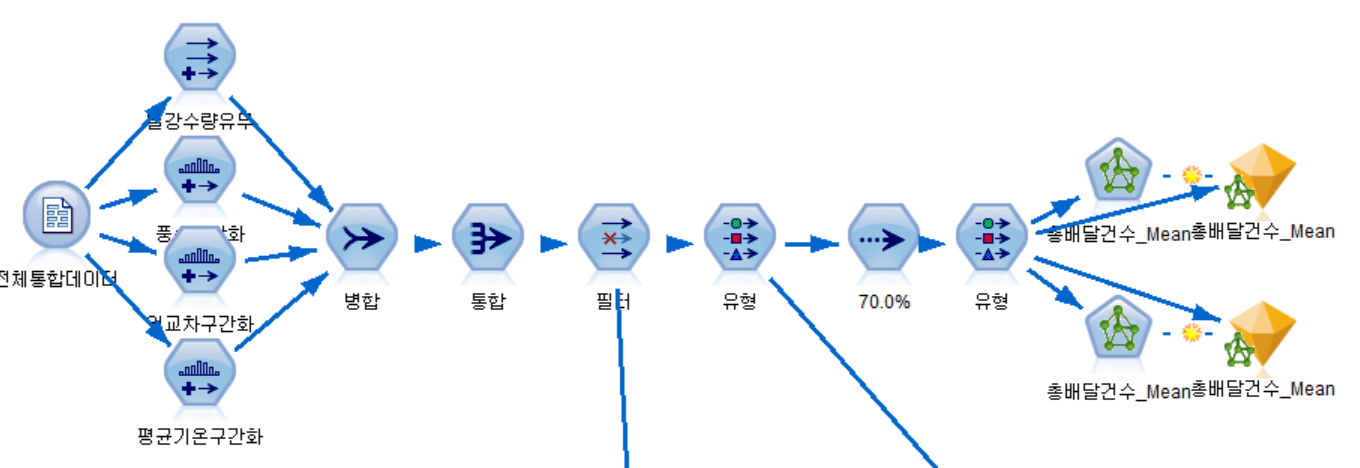
<일교차 구간화>



<평균기온 구간화>



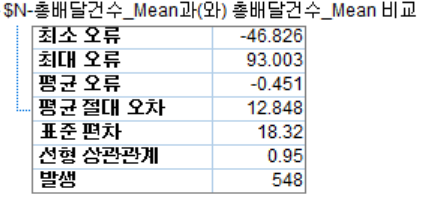
그 다음, Training Set (70%)에 대하여 신경망 분석 기법 MLP와 RBF를 이용해 모델을 생성하였다.



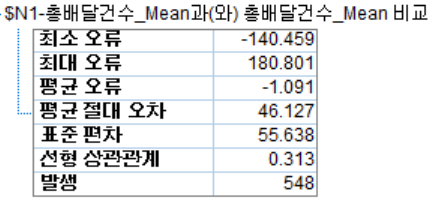
이후 Testing Set (30%)을 이용하여 모델의 적합성을 검증하였다.



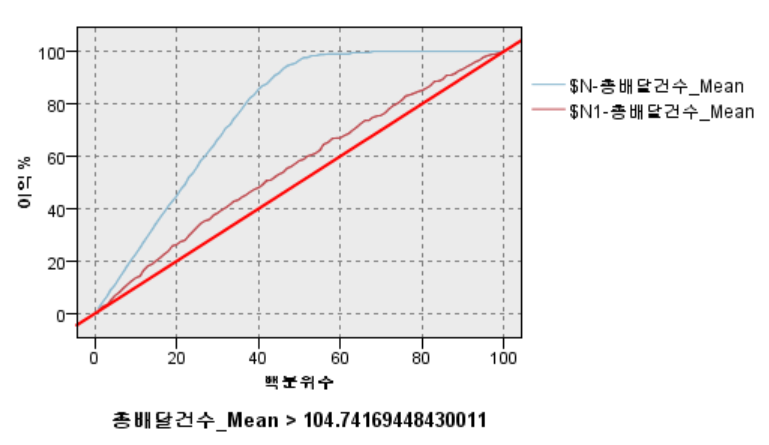
<MLP 기법>



<RBF 기법>



분석노드를 이용해 비교했을 때, 최대 오류 측면에서 MLP 기법의 모델이 수치가 더 적고, 선형 상관관계 측면에서는 MLP 기법의 수치가 0.95으로 RBF 기법의 0.3913에 비해 훨씬 1에 가깝다. 따라서 MLP 기법을 이용한 신경망 모델의 적합성이 더욱 뛰어나다.



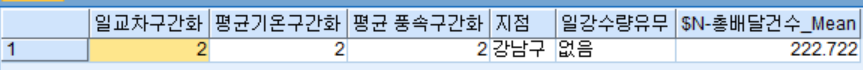
평가그래프를 이용해 비교해보아도 MLP 모델(파란색)이 기존 선형과 더 멀리 떨어져 있는 것을 볼 수 있다. 즉, MLP 기법이 RBF 기법보다 좋다고 말할 수 있다.

생성된 MLP 신경망 모델과 사용자입력 노드를 이용하여 새로운 날씨데이터의 강수량, 풍속, 일교차, 평균기온과 지점을 임의로 지정하여 총 배달건수를 예측해보도록 한다.

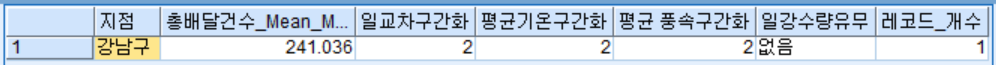




실제 값과 예측 값이 차이가 없는지 확인해보기 위해 위와 같이 임의로 변수를 넣었다.



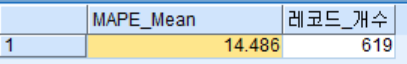
위의 사용자 입력 데이터에 따른 치킨 배달건수 평균의 예측값은 222.722이다.



데이터에서 해당되는 레코드만을 선택했을 때 실제 총 배달건수의 평균은 241.036으로 나왔다.

위의 예측값은 임의적으로 넣은 값이기 때문에 정확도가 떨어질 수 있다. 따라서 수치적으로 더 자세한 정확도를 알아보기 위해 MAPE를 사용하여 정확도를 검증하였다.

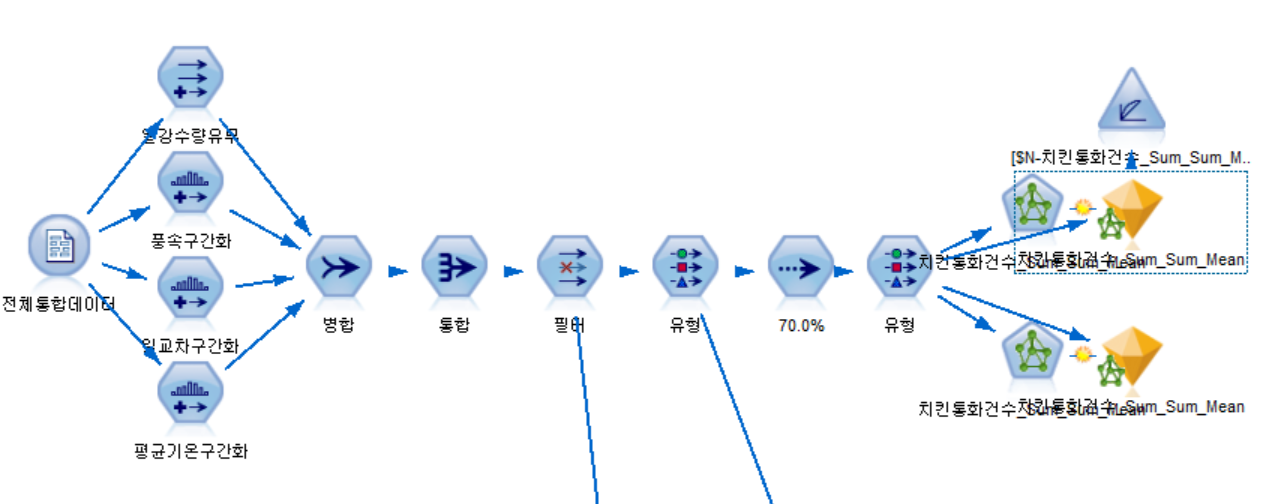
MAPE = (abs(실제 총 배달건수의 평균 – 예측 총 배달건수의 평균) / 실제 총 배달건수의 평균) \* 100



MAPE를 계산한 결과 14.486로 비교적 정확한 예측을 했다고 할 수 있다.

날씨와 총 배달건수와의 관계만이 아닌 자영업자한테 영업에 도움이 되는 더 세부적인 예측값을 제언해주기 위해 연령대와 성별을 추가적으로 예측변수에 넣고 예측 목표값을 치킨, 피자, 중국집으로 나눠 위와 같은 방법으로 더욱 세부적인 분석을 해 보았다.

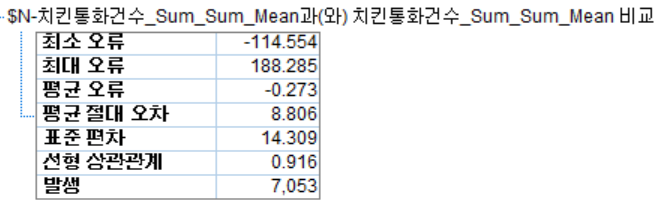
그 다음, Training Set (70%)에 대하여 신경망 분석 기법 MLP와 RBF를 이용해 모델을 생성하였다.



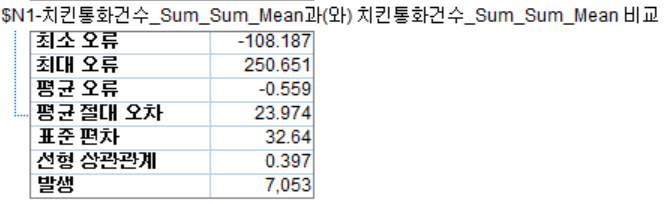
이후 Testing Set (30%)을 이용하여 모델의 적합성을 검증하였다.



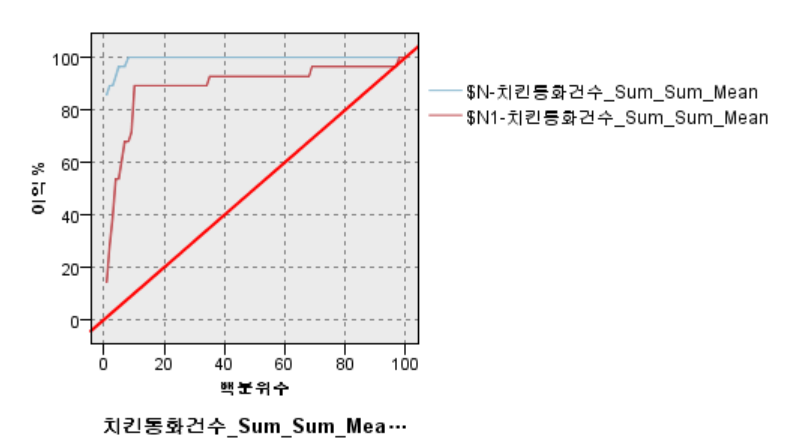
<MLP 기법>



<RBF 기법>



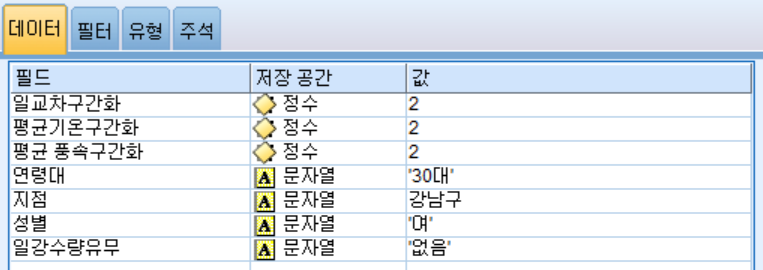
분석노드를 이용해 비교했을 때, 최대 오류 측면에서 MLP 기법의 모델이 수치가 더 적고, 선형 상관관계 측면에서는 MLP 기법의 수치가 0.916으로 RBF 기법의 0.397에 비해 훨씬 1에 가깝다. 따라서 MLP 기법을 이용한 신경망 모델의 적합성이 더욱 뛰어나다.



평가그래프를 이용해 비교해보아도 MLP 모델(파란색)이 기존 선형과 더 멀리 떨어져 있는 것을 볼 수 있다. 즉, MLP 기법이 RBF 기법보다 좋다고 말할 수 있다.

생성된 MLP 신경망 모델과 사용자입력 노드를 이용하여 새로운 날씨데이터의 강수량, 풍속, 일교차, 평균기온과 배달음식 데이터의 지점, 연령대, 성별을 임의로 지정하여 치킨 배달건수를 예측해보도록 한다.

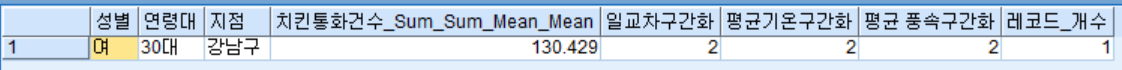




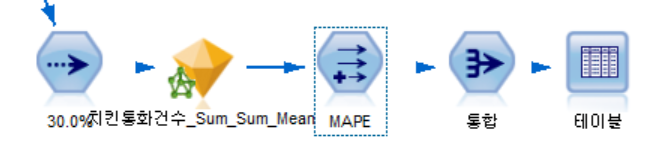
실제 값과 예측 값이 차이가 없는지 확인해보기 위해 위와 같이 임의로 변수를 넣었다.



위의 사용자 입력 데이터에 따른 치킨 배달건수 평균의 예측값은 114.204이다.



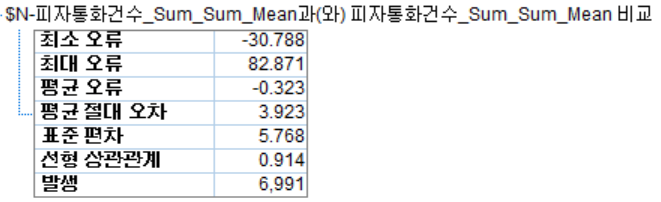
데이터에서 해당되는 레코드만을 선택했을 때 실제 치킨 배달건수의 평균은 130.429으로 나왔다.

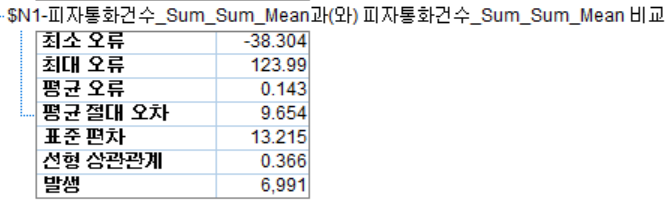


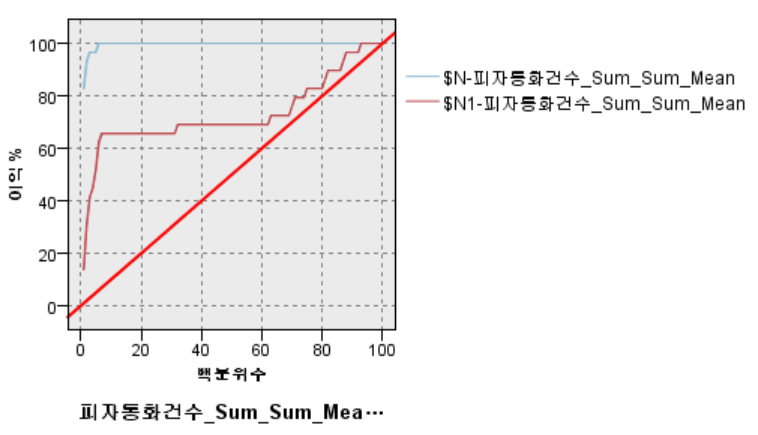


MAPE를 계산한 결과 30.541로 비교적 합리적 예측을 했다고 할 수 있다.

위와 같은 방식으로 피자와 중국음식 배달건수에 대한 신경망 분석을 하였을 때 MLP 기법이 적합하여 치킨과 똑같은 조건으로 피자와 중국음식의 배달 건수를 예측해 보았다.





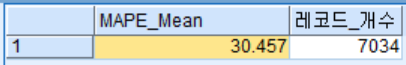




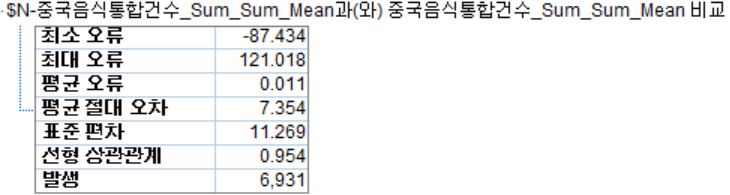
위의 사용자 입력 데이터에 따른 피자 배달건수 평균의 예측값은 46.857이다.



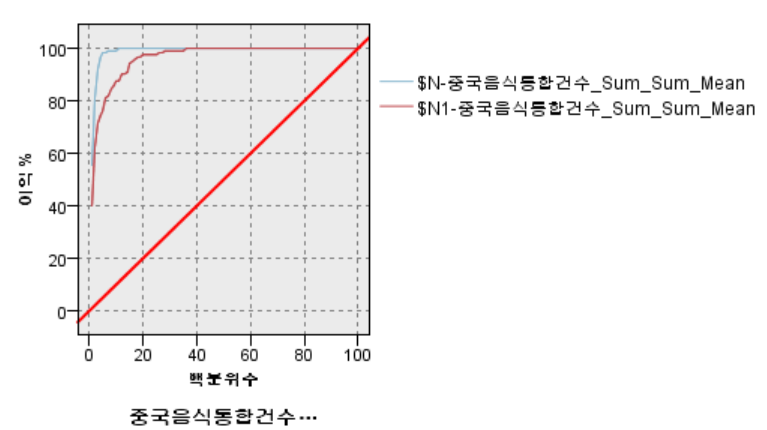
데이터에서 해당되는 레코드만을 선택했을 때 실제 피자 배달건수의 평균은 37.787으로 나왔다.



MAPE를 계산한 결과 30.457로 비교적 합리적 예측을 했다고 할 수 있다.

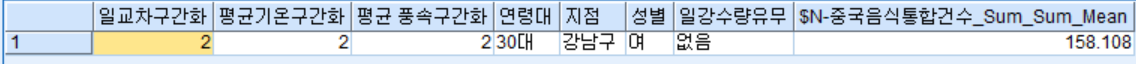








위의 사용자 입력 데이터에 따른 중국음식 배달건수 평균의 예측값은 154.571이다.



데이터에서 해당되는 레코드만을 선택했을 때 실제 중국음식 배달건수의 평균은 158.108으로 나왔다.



MAPE를 계산한 결과 30.301로 비교적 합리적 예측을 했다고 할 수 있다.

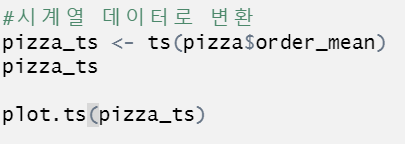
따라서 치킨, 피자, 중국음식의 배달건수를 예측하였을 때 모두 MAPE 값이 30 내외로 비교적 합리적 예측을 했다고 할 수 있다. 즉, 서울시 내의 여러 구의 자영업자들은 이 예측 방식을 이용하였을 때 가상의 날씨에 따른 고객의 치킨, 피자, 중국음식 배달건수를 알 수 있고 연령대, 성별에 따라 알맞은 마케팅을 할 수 있을 것이다. 신경망 분석은 과과거의 모든 데이터를 분석할 필요가 없이 배달건수를 간편하게 예측해 볼 수 있는 장점이 있다.

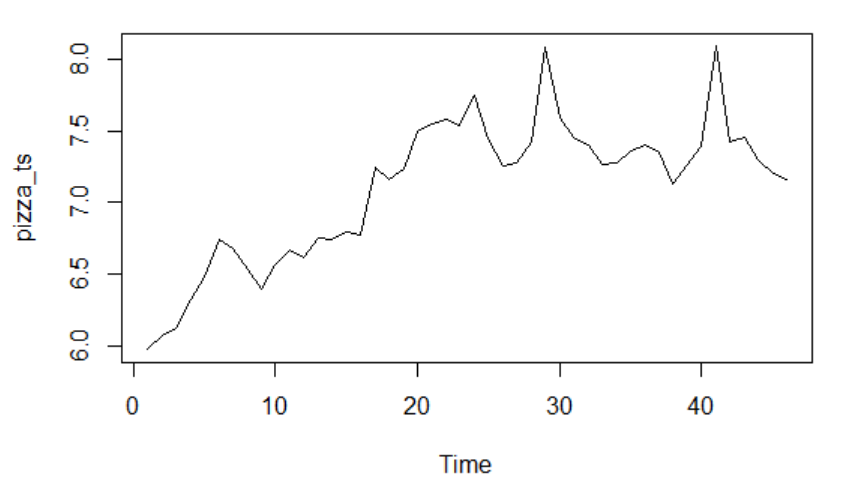
1. **시계열 분석**

4년동안의 치킨, 피자, 중국음식의 배달건수 추세를 알아보고 향후 5달의 배달건수를 예측하기 위하여 R프로그램을 사용하여 시계열 분석을 하였다.

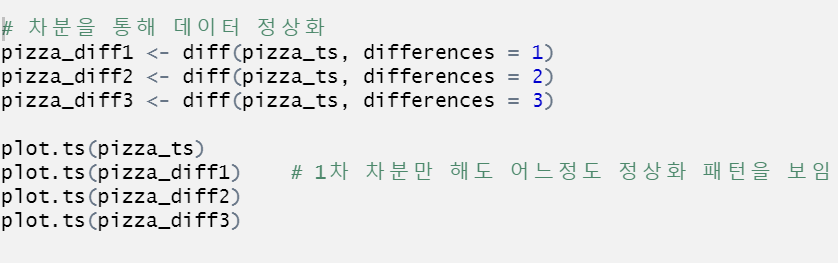


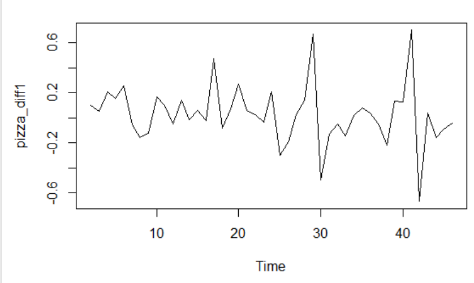
* 시계열 데이터로 변환



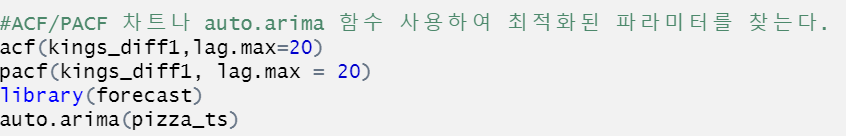


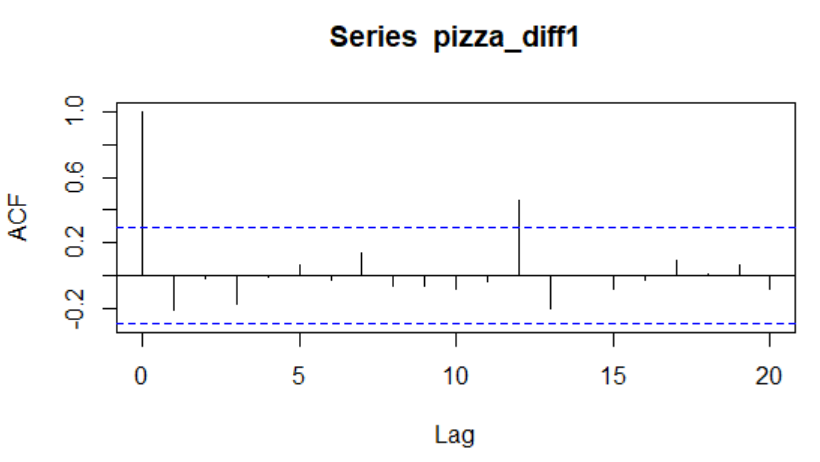
* 차분을 통해 데이터 정상화

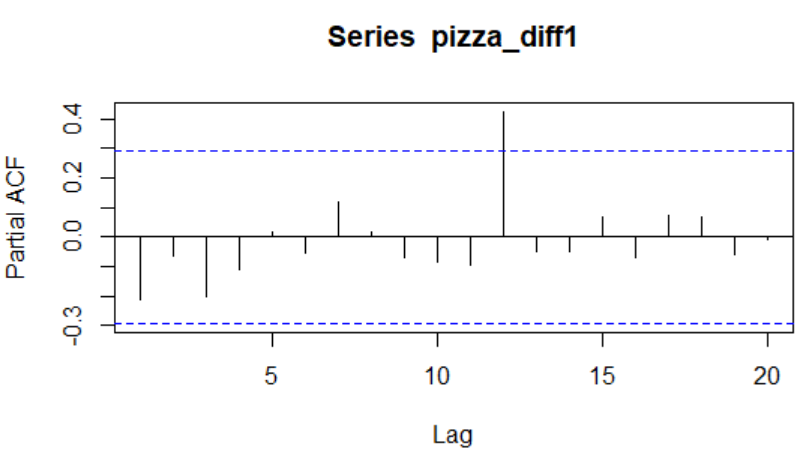


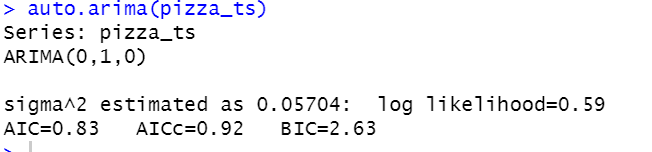


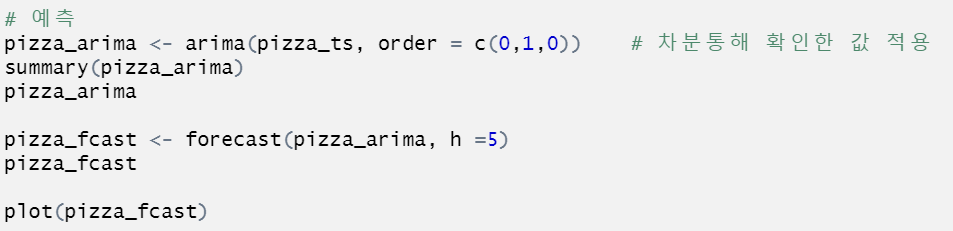
* ACF/PACF 차트와 auto.arima 함수 사용하여 최적화된 파라미터를 찾음.

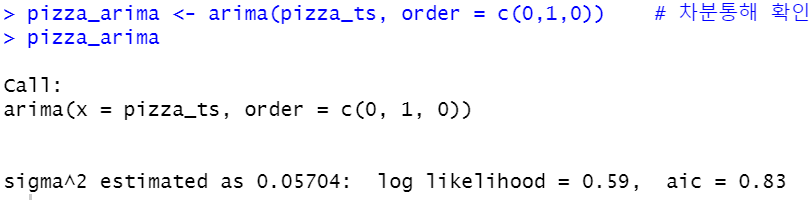


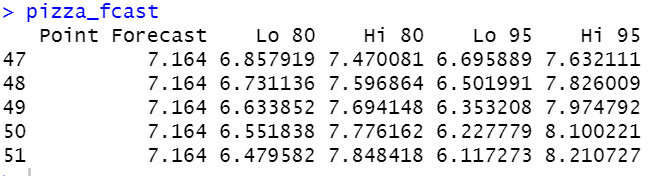


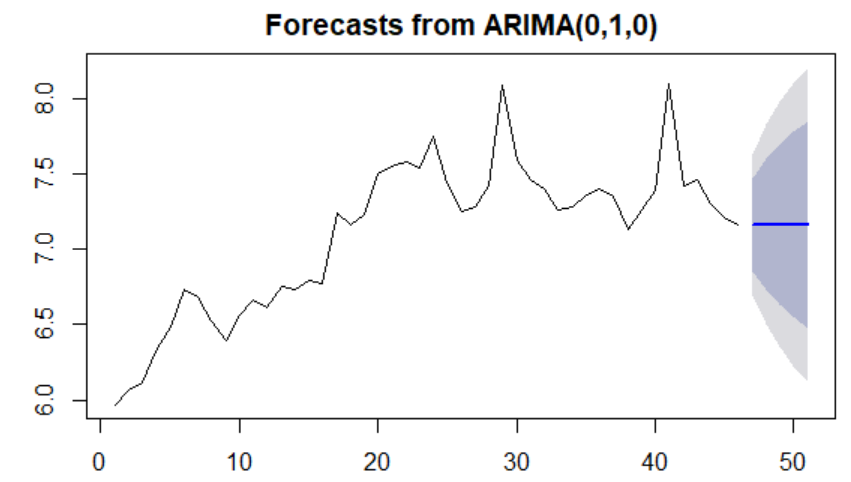




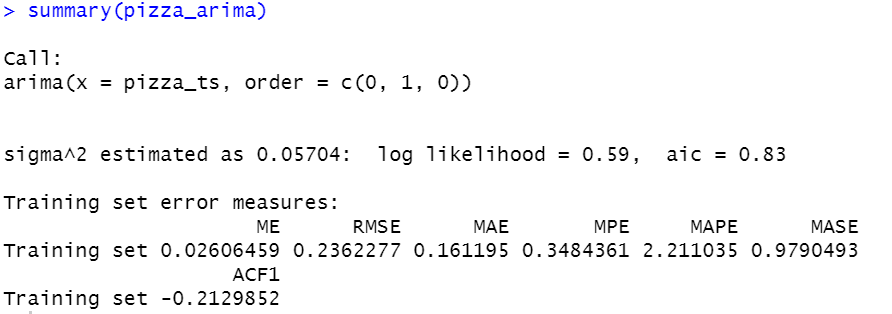








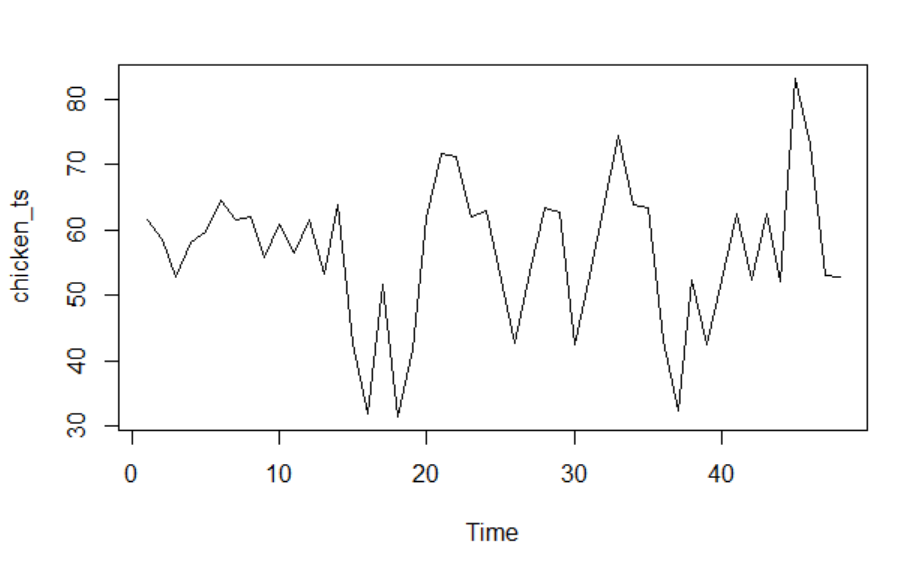
* 모델 평가



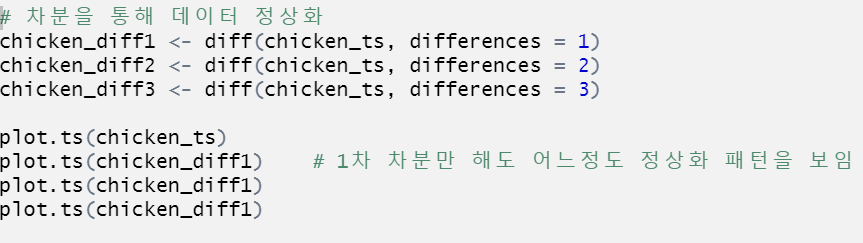
1. MAPE를 봤을 때 2.21%로써 모델의 예측은 매우 정확하다고 볼 수 있다.

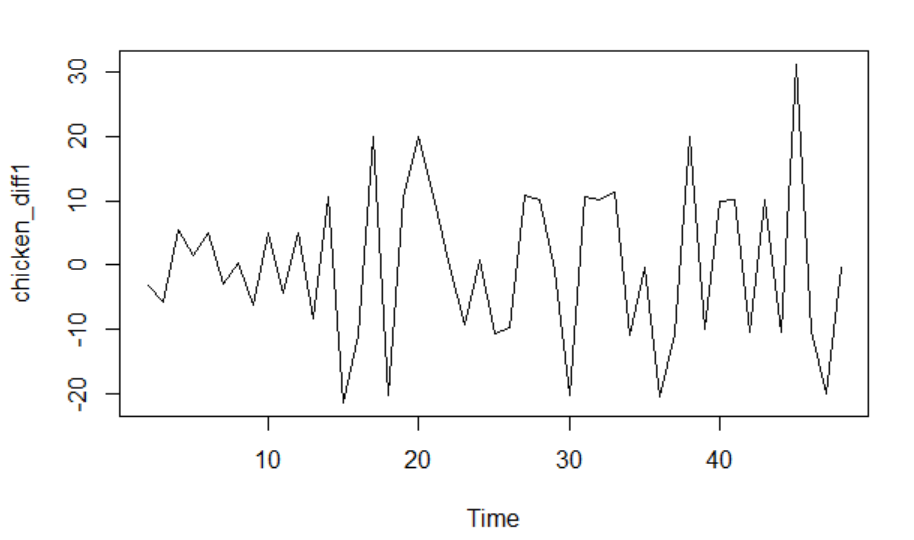
# 치킨

치킨도 위와 같이 데이터 불러오고 시계열 변수로 변환했다.

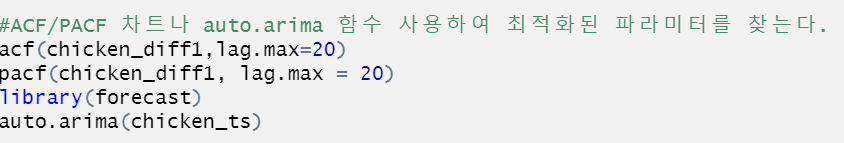


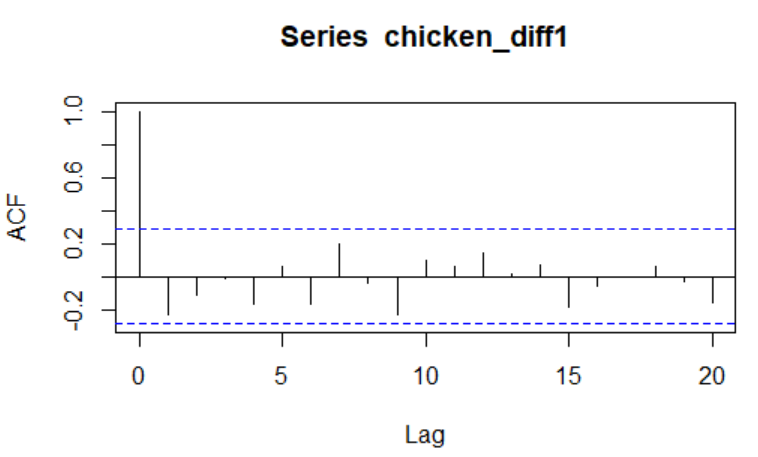
* 차분을 통해 데이터 정상화

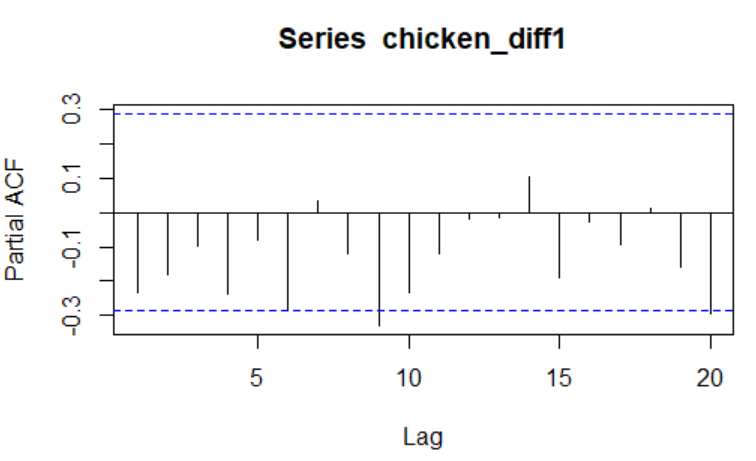


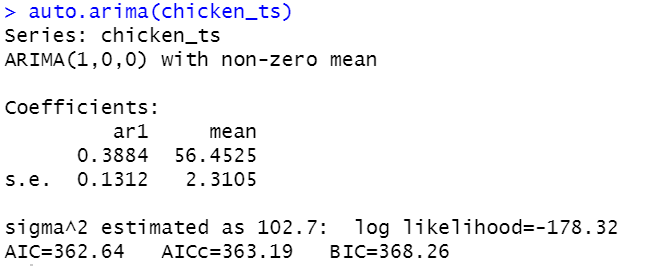


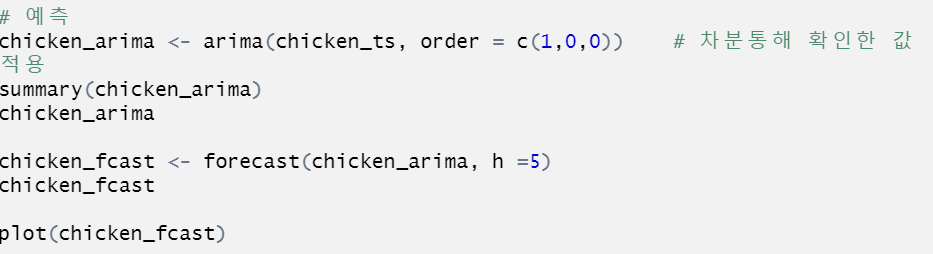
* ACF/PACF 차트와 auto.arima 함수 사용하여 최적화된 파라미터를 찾음.

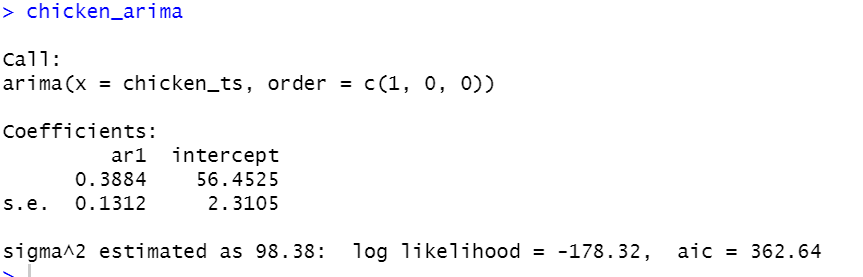


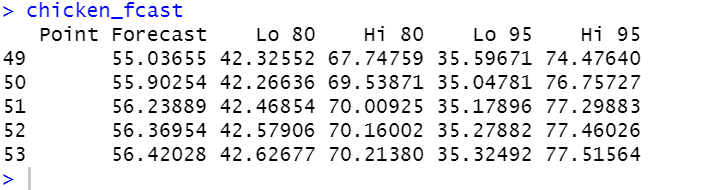


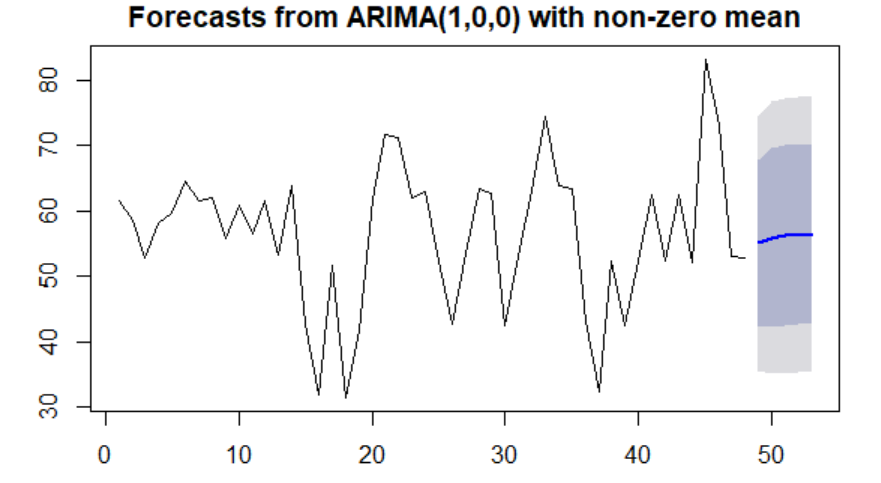




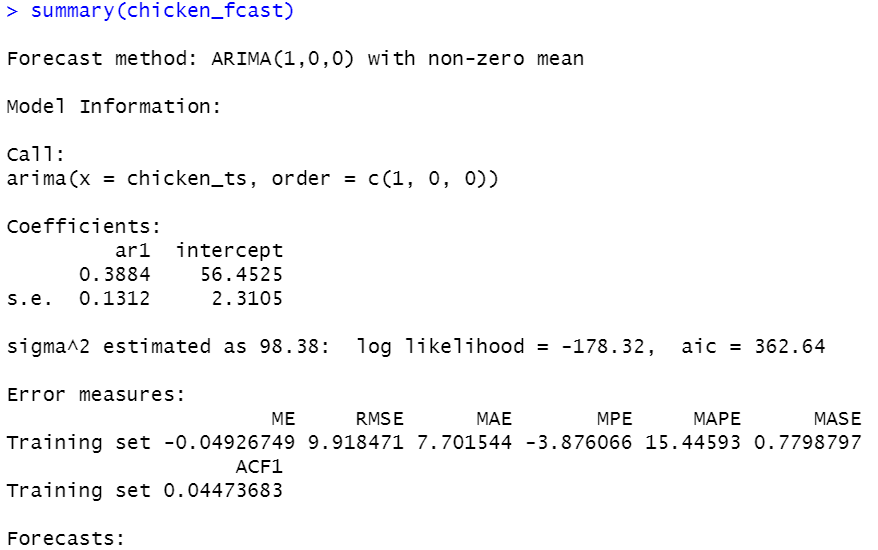








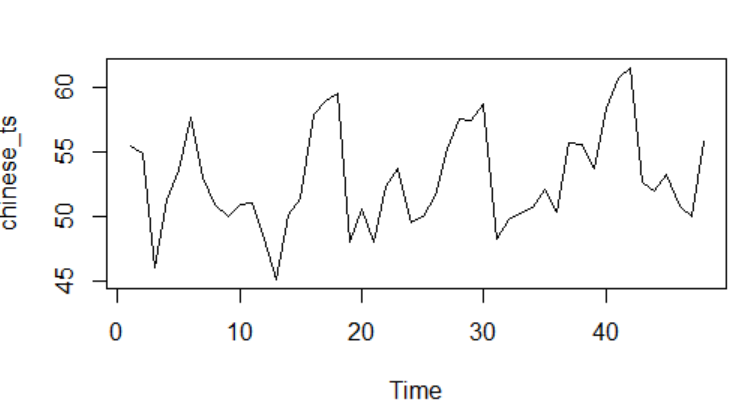
* 모델 평가



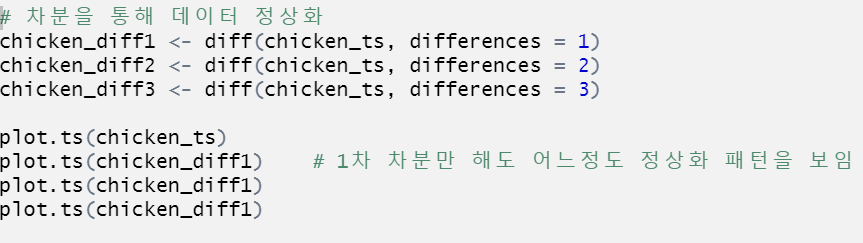
1. MAPE를 봤을 때 15.2%로써 모델의 예측은 피자보다는 아니지만 비교적 정확한 예측이라고고 볼 수 있다.

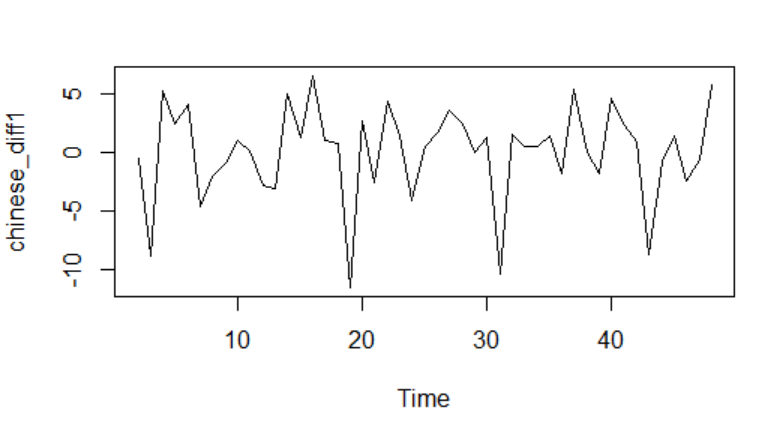
# 중국음식

중국음식도 위와 같이 데이터 불러오고 시계열 변수로 변환했다.

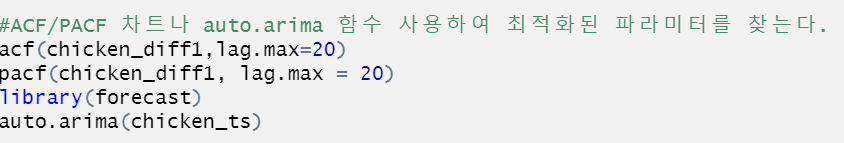


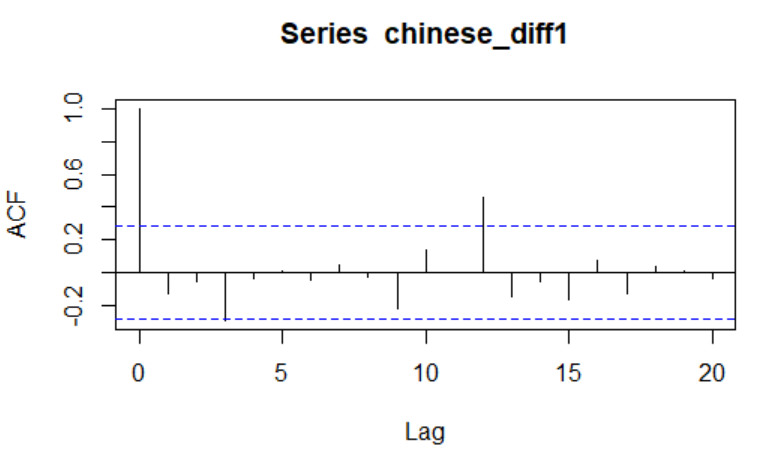
* 차분을 통해 데이터 정상화

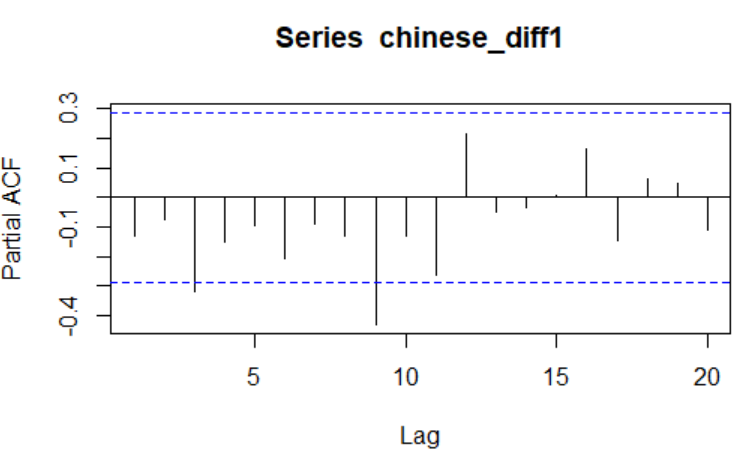


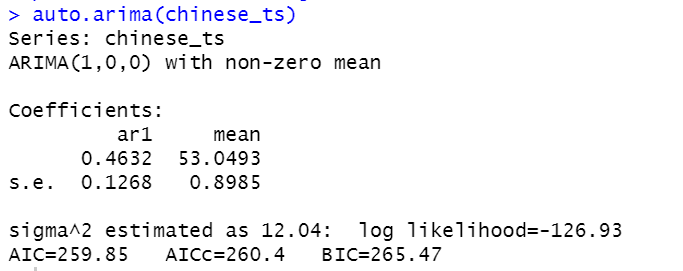


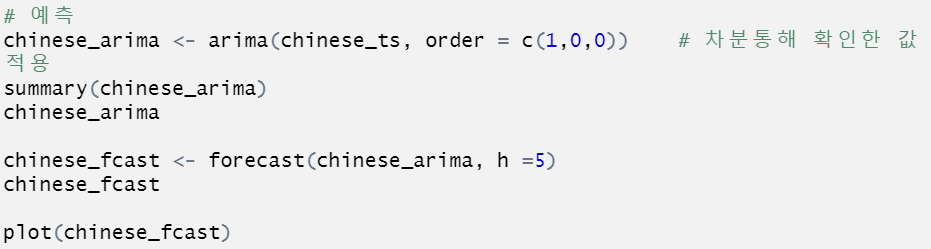
* ACF/PACF 차트와 auto.arima 함수 사용하여 최적화된 파라미터를 찾음.

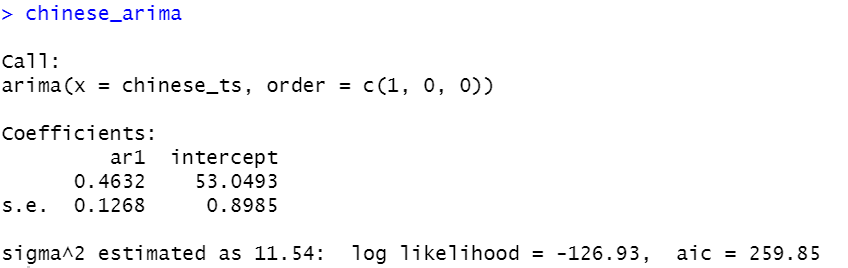


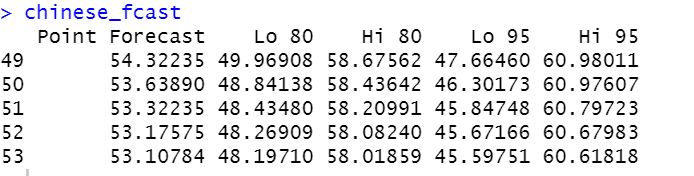


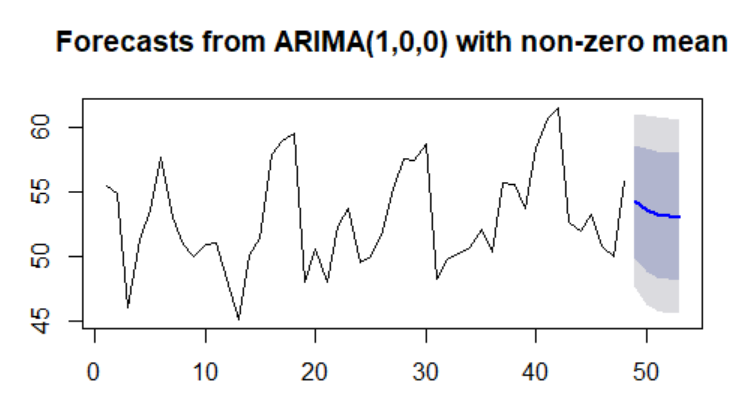




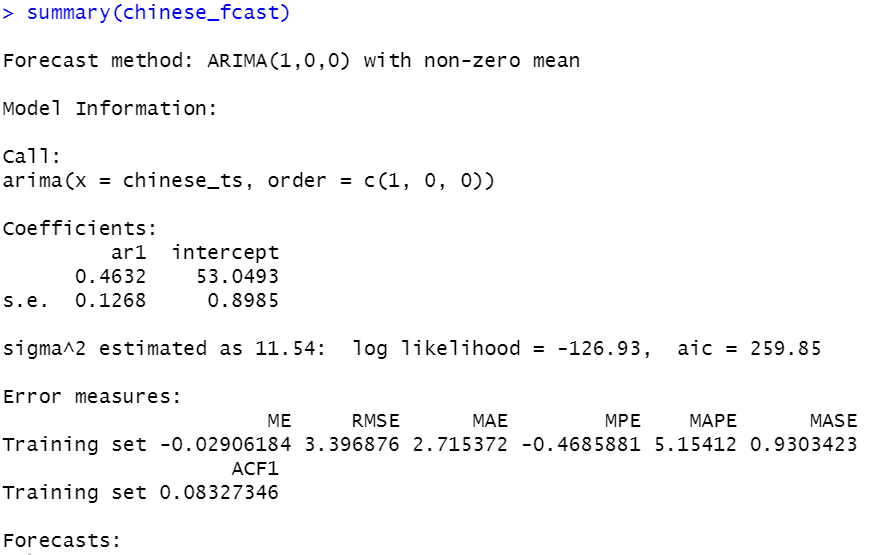








* 모델 평가



1. MAPE를 봤을 때 5.15%로써 모델의 예측은 피자와 마찬가지로 10% 미만이기 때문에 매우 정확한 예측이라고 할 수 있다.
2. **기대효과**

날씨에 따른 배달음식 주문량을 예측하여 영세 자영업자 및 소상공인이 적극적으로 활용하기 어려웠던 날씨 빅데이터를 손쉽게 활용, 이를 기반으로 배달음식의 수요를 예측하여 창업자와 자영업자들에게 날씨에 따른 식자재 관리 활용 및 타게팅 전략 등을 제안함으로써 사업활성화에 기여할 수 있다.