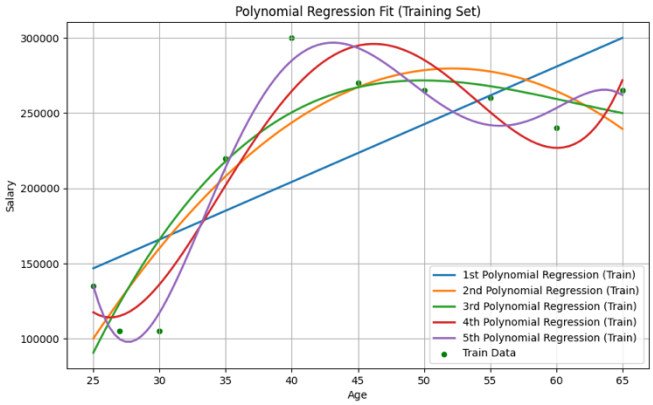
## 금융과인공지능 1차 과제

국제통상학부 2021510010 이상민

1. 이 문제에서는 급여(Salary)와 연령(Age) 간의 관계를 분석하기 위해 다항 회귀 분석을 수행한다. 이를 위해 1차부터 5차까지의 다항 회귀 모형을 적용하여 각 모형의 성능을 평가하고, Valid Set에서의 일반화 성능을 통해 최적의 모형을 선정한다. 기존 분석에서 2차 모형이 가장 적합하다고 평가되었음을 바탕으로, 3차와 4차 모형의 성능을 추가로 검토하고, 최종적으로 과적합의 위험성을 최소화하면서도 일반화 성능이 가장 우수한 모형을 결정하는 것이다.

1-1) 1차부터 5차까지의 다항 회귀 모형을 Training Set에 적용하여 적합화 결과를 시각화 하였다. <그림 1>은 Training Set에 대한 적합화 결과를 나타내며, <표 1>은 차수 별 RMSE를 정리한 표이다.



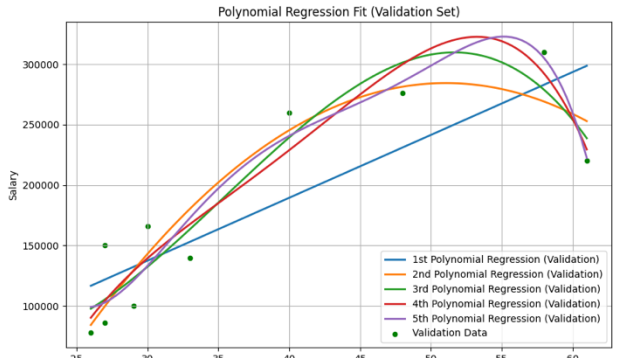

<그림 1>

| Training Set RMSE Values | |
|---|---|
| Degree | Training RMSE |
| 1 | 47179.09307 |
| 2 | 31242.11685 |
| 3 | 30346.97183 |
| 4 | 20704.2035 |
| 5 | 12240.10454 |

<표 1>

Training Set에서 차수가 높아질수록 RMSE 값이 감소하는 것을 확인 할 수 있다. 특히, 5차 모형이 가장 낮은 RMSE 값을 기록하였다. 그러나, 차수가 높아질수록 Training Data에 과도하게 적합(Overfitting)될 가능성이 있으므로, Validation Set에서의 일반화 성능을 추가로 평가할 필요가 있다.

1-2) Valid Set에 대한 각 다항 회귀 모형을 적용하고 적합화 결과를 시각화 하였다. <그림 2>는 Valid Set에 대한 적합화 결과이며, <표 2>는 차수 별 RMSE를 정리한 표이다.



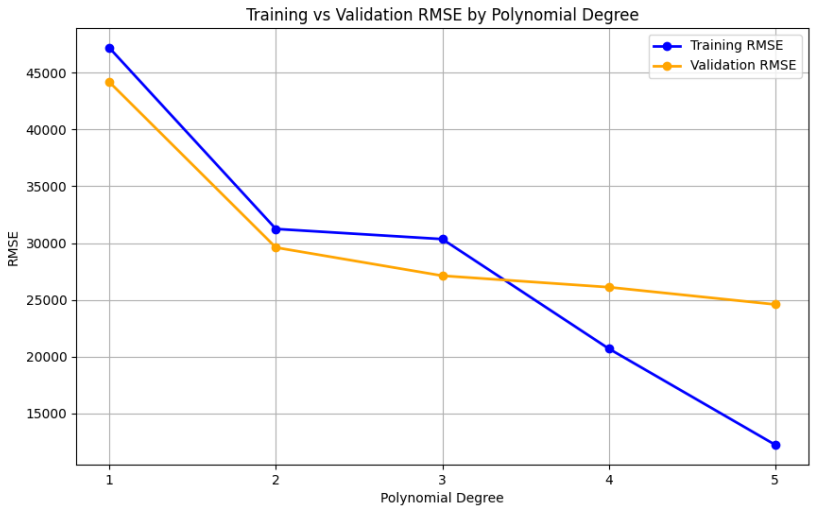

<그림 2>

| Validation Set RMSE Values | |
|---|---|
| Degree | Validation RMSE |
| 1 | 44166.51161 |
| 2 | 29606.84494 |
| 3 | 27121.76099 |
| 4 | 26113.81122 |
| 5 | 24594.34247 |

<표 2>

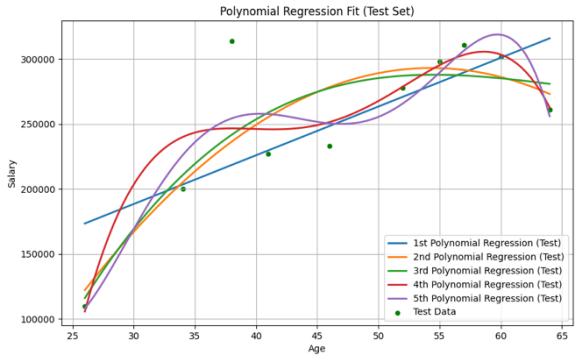
Valid Set에서의 결과를 분석한 결과, 2차 모형이 Training Set과 Valid Set 모두에서 안정적인 성능을 보였다. 5차 모형은 Training Set에서는 매우 좋은 성능을 보였지만, Valid Set에서는 성능이 크게 저하되며 과적합의 위험이 존재한다.

<그림 3>은 Training Set과 Valid Set의 RMSE 값을 그래프로 시각화한 것이다. 이를 확인해 봤을 때 2차 모형에서 RMSE의 차이가 가장 적은 것을 확인 할 수 있다. 따라서 Training Set과. Valid Set에서 RMSE 차이가 가장 적은 2차 모형이 과 적합 문제를 최소화하면서도 일반화 성능이 가장 우수한 모형으로 판단된다.




<그림 3>

1-3) 최종적으로, 1차부터 5차까지의 다항 회귀 모형을 Test Set에 적용하여 각 모형의 성능을 평가하였다. <그림 4>는 Test Set에 대한 적합화 결과이며, <표 3>은 각 모형의. RMSE 값을 정리한 표이다.




<그림 4>

| Test Set RMSE Values | |
|---|---|
| Degree | Test RMSE |
| 1 | 41457.41794 |
| 2 | 31710.3467 |
| 3 | 31330.93712 |
| 4 | 26321.10097 |
| 5 | 24572.04545 |

<표 3>

Test Set에서도 2차 모형이 Valid Set에서 보였던 성능을 유지하면서, 가장 우수한 성과를 기록하였다. 이는 Training Set과 Valid Set를 일관되게 반영한 결과이며, 과 적합 문제를 피하면서도 일반화 성능이 뛰어난 것으로 판단된다. RMSE 차이 또한 2차 모형이 가장 작은 것을 <그림 5>에서 확인할 수 있다.

2-1) 요인로딩은 <표 4>와 같다.

| | PC1 | PC2 | PC3 | PC4 |
|---|---|---|---|---|
| 부패지수 | 0.594089 | -0.155184 | 0.29164 | 0.733431 |
| 평화지수 | -0.530424 | -0.039876 | 0.842391 | 0.086246 |
| 법률위험지수 | 0.0585023 | -0.134562 | 0.430976 | -0.673721 |
| 실질GDP성장률 | 0.153154 | 0.977865 | 0.139939 | 0.027201 |

<표 4>

PC1은 주로 부패지수와 평화지수에 의해 설명되며, 이는 국가의 정치적 불안정성 및 사회적 부패와 관련이 있을 가능성이 크다. 부패가 높고 평화가 낮을수록 PC1 값이 증가하는 경향을 보일 것이다.

PC2는 주로 경제적 성장을 설명하며, 실질 GDP 성장률이 높을수록 PC2 값이 증가하는 것으로 보인다.

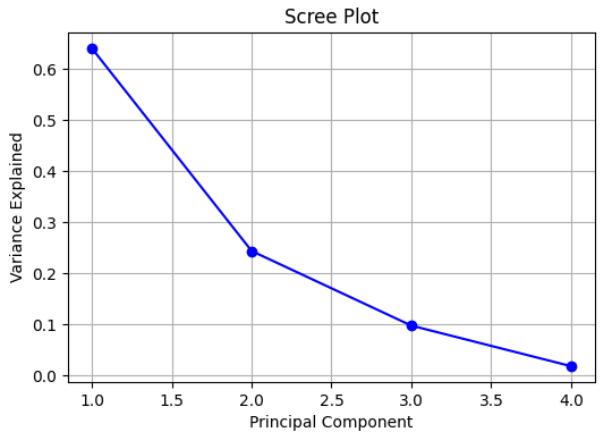
PC3는 주로 평화지수와 법률적 안정성을 설명하며, 평화가 높고 법률 위험이 낮을수록 PC3 값이 높아질 것이다.

PC4는 부패와 법률적 안정성과 관련이 있으며, 부패가 높고 법적 위험이 낮을 때, PC4값이 증가한다.

2-2) 국가 위험 데이터의 값들은 아래 <표 5>와 같고, 이를 통해 작성된 Scree Plot은 <그림 5>와 같다.

| 만기 | PC1 | PC2 | PC3 | PC4 |
|---|---|---|---|---|
| 요인점수 (분산) | 2.58250012 | 0.98295186 | 0.39396962 | 0.07363625 |
| 요인점수 (표준편차) | 1.6070159 | 0.99143929 | 0.62766999 | 0.27136 |
| EVR | 0.64033302 | 0.24372372 | 0.09768509 | 0.01825817 |

<표 5>




<그림 5>

PC1, PC2, PC3, PC4는 각각 64.03%, 24.37%, 9.77%, 1.83%의 분산을 설명한다.

<그림 5>의 Scree Plot을 보면 PC1과 PC2가 데이터의 주요 변동성을 설명하고 있으며, 이후의 주성분은 설명력이 급격히 감소하는 것을 알 수 있다. 특히, PC3 이후로 설명된 분산이 매우 낮은 수준으로 떨어진다.

일반적으로 Scree Plot에서 급격한 기울기 변화를 확인할 수 있는 “엘보우 (Elbow)” 지점에서 주성분을 자르는 것이 좋은 기준이다. 이 경우 PC2에서 급격한 변동성이 끝나고, 그 이후의 주성분은 설명된 분산 비율이 매우 낮기 때문에 PC1과 PC2만 사용하는 것이 적절해 보인다. PC1과 PC2를 사용하면 데이터의 88.4%의 변동성을 설명할 수 있다.

2-3) PC1의 요인 로딩을 살펴보면 부패지수가 0.594089, 평화지수가 -0.530424로 큰 가중치를 가지고 있다. 이에 반해 법률위험지수와 실질GDP성장률은 매우 작거나, 비교적 작은 가중치를 가지고 있다. 이를 바탕으로 PC1은 주로 부패와 평화지수에 의해 설명되는 것을 알 수 있다. 이는 정치적/사회적 불안정성을 반영하는 요인으로 해석할 수 있다. 부패가 심하고 평화 수준이 낮은 국가일수록 PC1 값이 커지며, 이는 국가의 정치적 리스크가 크다는 의미이다.

PC2의 요인 로딩은 실질GDP성장률의 값이 0.977865로 매우 큰 가중치를 가지고 있으며, 이는 PC2가 주로 경제 성장을 설명하는 요인임을 나타낸다. 경제 성장이 빠른 국가일수록 PC2 값이 커진다. 이는

국가의 경제적 안정성 및 성장성을 나타낸다.

종합적으로 고려해보면 PC1과 PC2의 주성분을 통해 국가의 위험을 정치적/사회적 리스크와 경제적 리스크로 나누어 설명할 수 있다. PC1과 PC2에서 모두 법률위험지수의 값이 작았기 때문에 법률위험지수에 리스크는 잘 알기 힘들 것이다.