**Part1: Regression (15 points)**

**(1) Sales of Toyota Corolla Cars**

The file ToyotaCorolla.xls contains data on used cars (Toyota Corollas) on sale during late summer of 2004 in The Netherlands. It has 1436 records containing details on 38 attributes, including Price, Age, Kilometers, HP, and other specifications. The goal will be to estimate the price of a used Toyota Corolla based on its specifications. Let Y represent the price, X1 the age (“age\_08\_04), X2 Guarantee period, and X3 the automatic (yes=1, no=0).

(1) If he wants to use a simple linear regression model to estimate the selling price of a car, which X variable do you recommend he use?

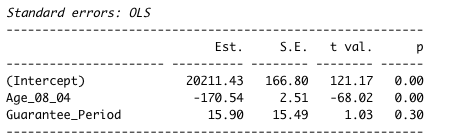
독립변수가 age, guarantee period, automatic중 하나여야 한다면 age를 고를 것이다. Age는 중고차 가격에 가장 큰 영향을 끼치는 요소이다. 따라서 선형회귀분석으로 연식에 따른 감가상각률을 알 수 있다면, 더 좋은 가격을 받을 수 있는 차를 더 낮은 가격에 팔지 않을 것이다.

Guarantee period와 automatic 역시 중요한 영향을 끼칠거라 생각되는 변수이지만, guarantee period는 보증기간이 끝나는 시점 주위로만 가격이 달라질 것이라 예상되기 때문에 simple regression model만 사용해야한다면 전체적인 중고차 생애 주기에 대한 특성을 알기엔 무리가 있다. Automatic은 전체적인 생애 주기에 걸쳐 영향을 줄 것이라 생각되지만, 그 영향력이 age보다 적다고 판단되기 때문에 배제하였다.

만약 위 독립변수를 떠나 전체 변수 중에 하나를 고를 수 있다면, 의미있는 선형회귀 분석이 되기 위해 color또는 fuel type을 선택하겠다. 차량의 연식(age or mfg\_year), 주행거리(km), guarantee period, automatic이 중고차 가격에 영향을 끼친다는 점은 지극히 당연한 사실이므로 제외한다. Statistics Netherlands(CBS)에 따르면 네덜란드에서 신차 구입시,특정 컬러(검정,흰색,회색,파란색)이 특히 선호도가 높다는 분석 결과가 있다. 따라서 이러한 색 선호현상이 중고차에도 적용되는지 알아보기 위해 선택하겠다. 또한 Mark J. Koetse의 네덜란드 대체연료 차량의 선호 요인 논문에 따르면, 대체연료 차량의 주행가능거리와 왕복가능거리가 대체연료차종 선택에 큰 영향을 끼친다고 한다. 따라서 상대적으로 연비가 높은 CNG, Diesel차량이 Gasoline 차량보다 상대적인 선호도가 높은지 분석하기 위해 fuel type을 선택하겠다.

(2) Develop a multiple linear regression model by using X1 and X2. Report coefficients of X1 and X2. Does X2 help to explain the selling price of the cars?

다음과 같은 결과가 나왔다.



Coefficients

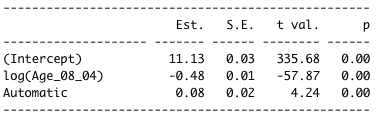
X1(Age) : -170,

X2(Guarantee\_Period) : 15

X2(Guarantee\_period)는 p-value > 0.05이므로 유의미하다고 볼 수 없다. 따라서 X2는 차량 가격을 설명하지 못한다.

(3) Develop a multiple linear regression model by using X1 and X3, and report their coefficients. Does X3 help to explain the selling price of the cars? How much does an automatic option add to the value of a car?

다음과 같은 결과가 나왔다.



\*\*변화량을 알고자 Y와 X1에 로그를 취하였다. Automatic의 경우 binary attribute이기 때문에 Dummy변수와 같이 취급하여 로그를 취하지 않았다.

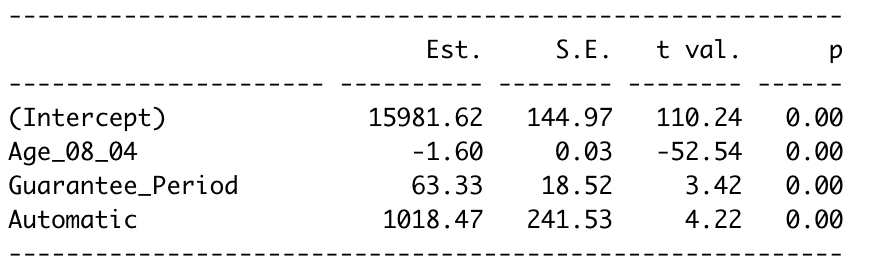
X3(automatic)의 경우 p-value가 2.35e^-5로 p < 0.05이므로 설명력을 갖는다고 볼 수있다. 즉 중고차가 Automatic일 확률이 1%증가할때마다 가격은 8%상승함을 알 수 있다. 즉 사람들이 일반적으로 수동보다 자동 기능을 더욱 선호한다는 것을 알 수 있다.

(4) Develop a multiple linear regression model with X1, X2, and X3 and construct a polynomial term if necessary. Does your polynomial term help to explain the selling price of the cars? Also, provide the scatter plot between a dependent variable and a polynomial term you created.

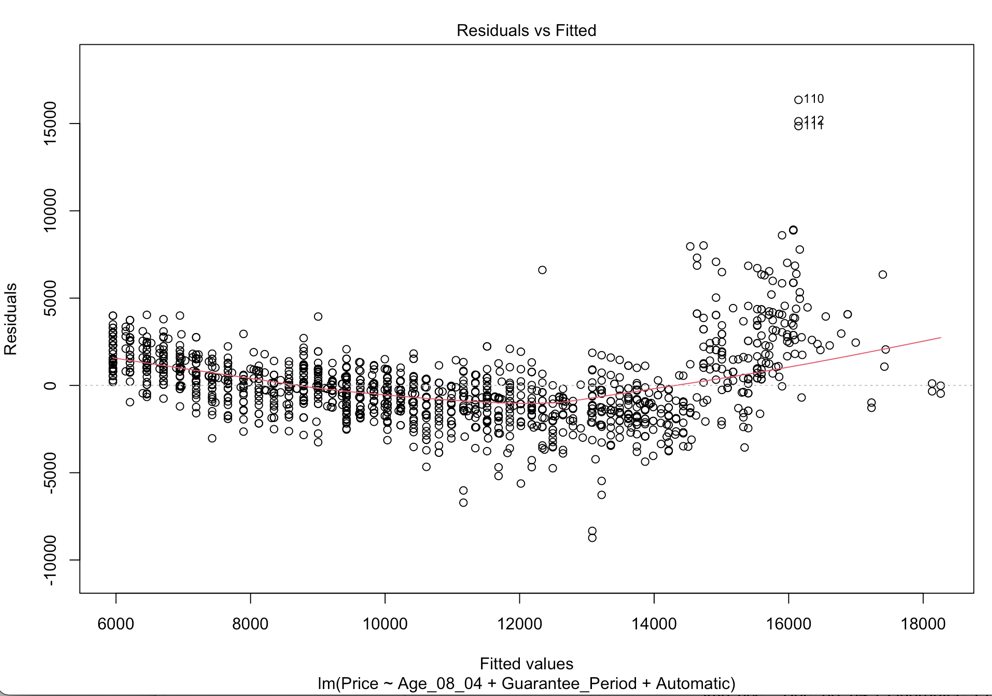
세 변수 중 age variable은 다른 변수보다 훨씬 큰 영향을 미칠 것으로 생각했고, 연식이 얼마 안된 초반보다 후반으로 갈수록 감소폭이 두드러질 것이라 판단하여 제곱하였다.

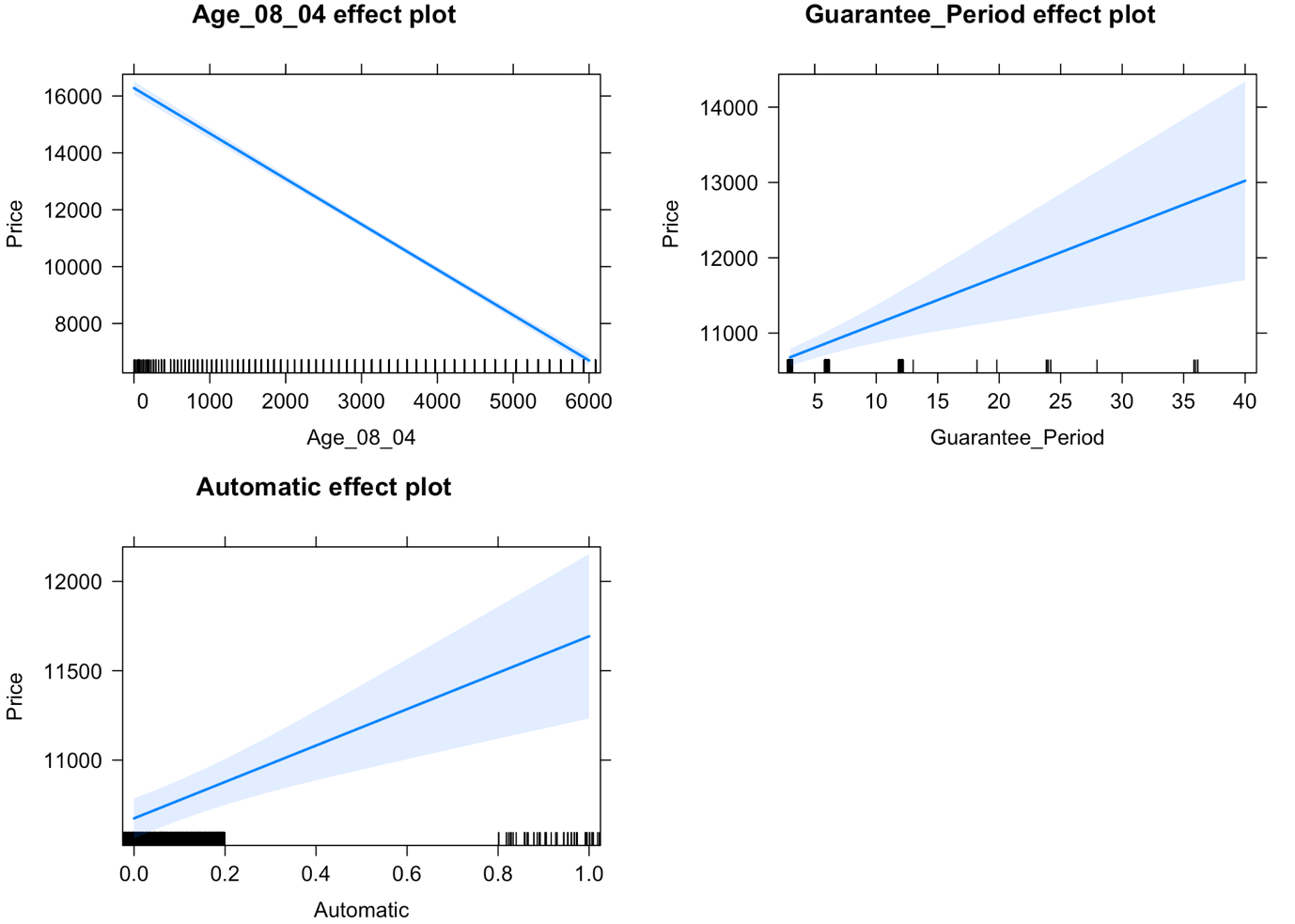
즉 y = B0 + B1X12 + B2X2 + B3X3로 되어있다.

결과는 다음과 같다.



기존에 유의하지 않다고 나왔던 Guarantee\_period가 X1에 의해 영향을 받아 유의하게 나왔다. (독립변수 X1을 제곱하지 않고 multiple linear regression을 돌렸을땐 Guarantee\_period의 P-value는 0.31이 나왔다.) 시각화를 하면 다음과 같다.





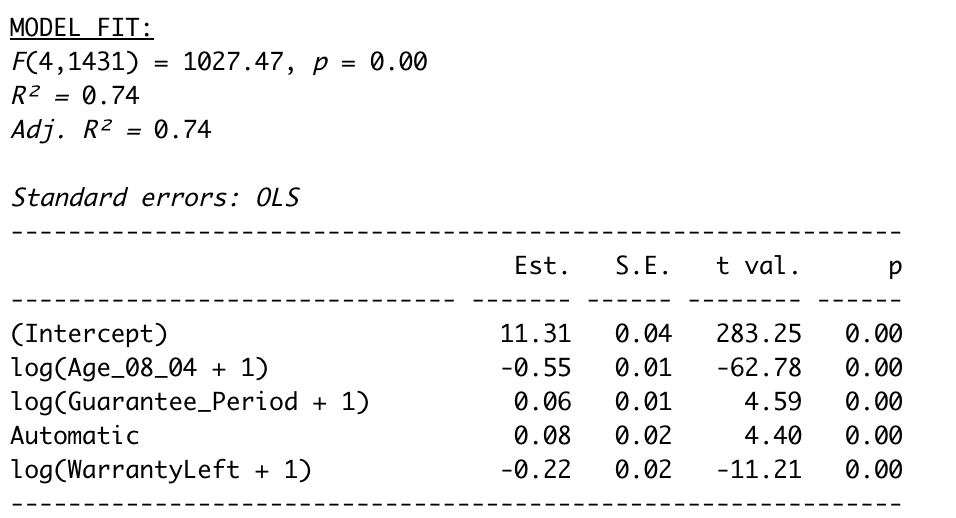
하지만 결과상으로 봤을때 price를 잘 대변하는 모델은 아닌 것으로 보인다. 각 항의 계수를 살펴보면, 일반적으로 가장 큰 영향을 끼칠 것이라 여겨지는 Age가 -1.6으로 영향도가 매우 낮아졌다. 그에 반해 세 독립 변수 중 minor한 요소라 할 수 있는 Automatic의 영향도가 1018로 매우 커진 것을 볼 수 있다. 따라서 polynomial 모델이 Price와의 상관을 잘 드러내지 못한다.

(5) Develop a multiple linear regression model with X1, X2, and X3 and construct an interaction term if necessary. Does your interaction term help to explain the selling price of the cars?

새로운 독립변수로, ‘잔여 Gaurantee\_period’를 추가하였다.

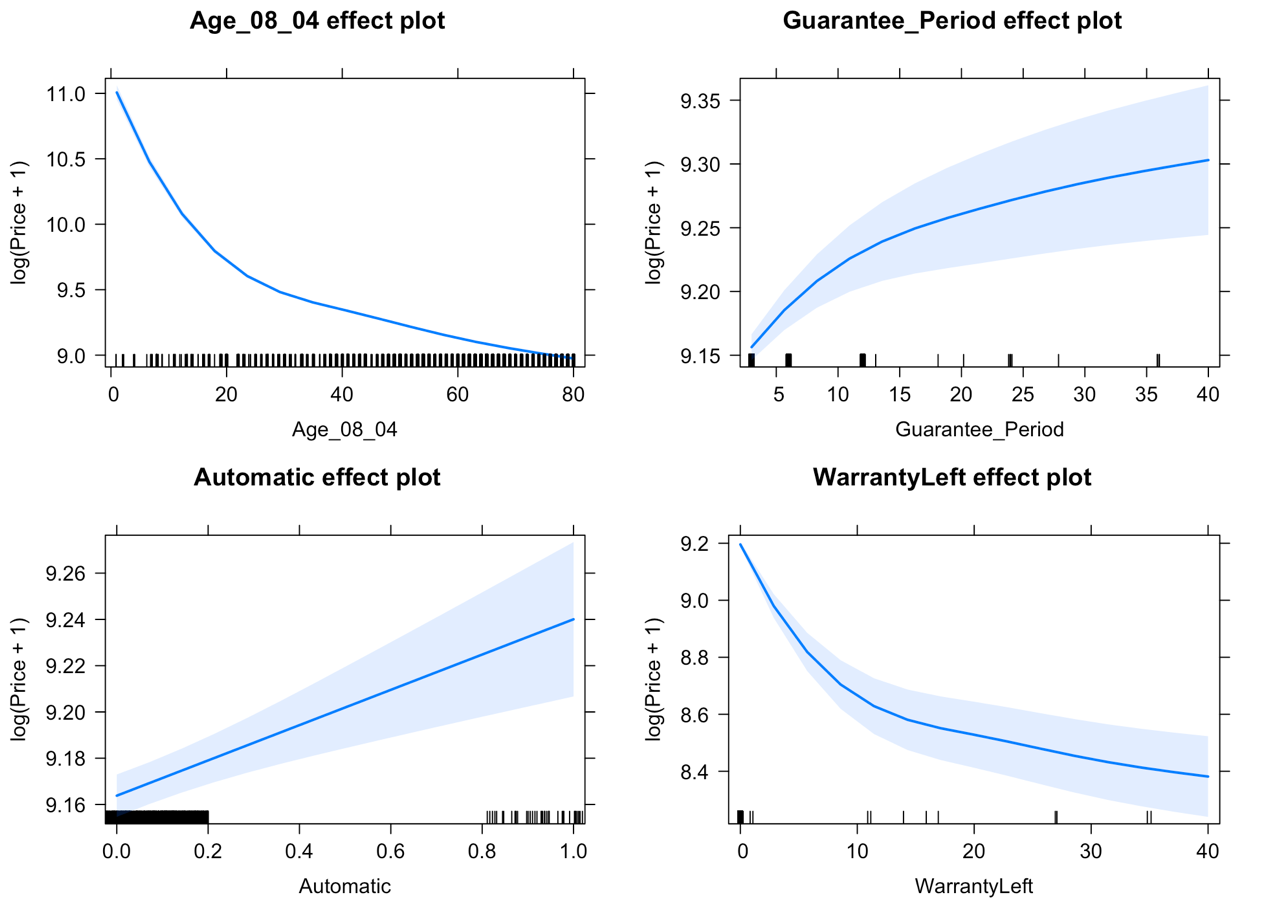
과정 :

다중공선성을 예방하기 위해 (Age-Guarantee\_period)한 값에 +값은 0으로(보증기간 초과), -값은 절대값을 부여하였다. 퍼센티지를 알아보기 위해 회귀분석에 로그를 취하였다. (단, binary value인 automatic은 제외)



모든 항목의 p-value < 0.05이므로 유의하다고 할 수 있다.

Interaction term은 selling price를 설명하는데 도움이 된다고 할 수 있다. 단순한 보증 기간보다, 잔여보증기간이 16% 더 높은 영향을 끼친다는 사실을 알 수 있기 때문이다.

이를 그래프로 나타내면 다음과 같다.

(6) Consider the p-values for each Bi in this model (question (5)). Do these values indicate that any of the independent variables should be dropped form the model? Provide your FINAL model as an equation (e.g., 0.34+0.32Color- 0.12Price ~~~).

각 독립변수에 대한 p-value는 다음과 같다.

Intercept : 2e-16

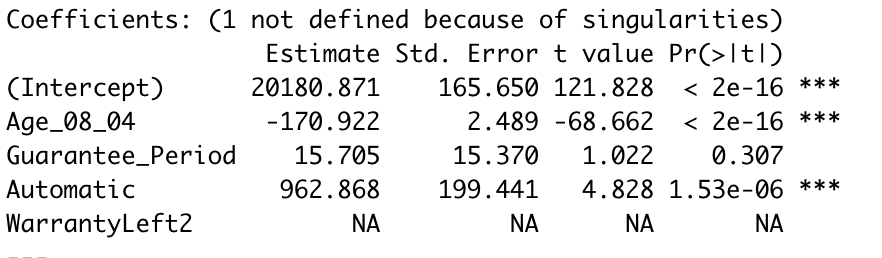
Age\_08\_04 : 2e-16

Guarantee\_period : 4.89e-6

Automatic : 1.19e-5

WarrantyLeft : 2e-16

모든 독립변수의 p-value가 유의미하게 나왔던 이유는 모델링할때 다중공선성을 피하고자 전처리를 진행했기 때문에 가능했다. 만약WarrantyLeft를 전처리하지 않고 단순히(Age – Gurantee Period)로만 진행했을 경우엔 다음과 같은 결과가 나온다.



R에서 다중공선성이 너무 높아 자동으로 배제를 하는 결과가 나타난다. 또한 guarantee period의 p-value가 0.307로 유효하지 않게 나오게 된다. 이 경우 1) Interaction term과 Guarantee\_period를 드랍하거나, 2) 다른 전처리를 통해 네 개의 독립변수 모두 유효하게 만들어야 한다.

**Final Model Equation :**

**Y = -0.55\*Age + -0.22\*WarrantyLeft + 0.08 \* Automatic + 0.06 \* GuaranteePeriod + 11.31**

(7) Interpret the values of adjusted R2 for each model. Which model do you recommend Roger use?

모델에 따른 Adjusted R2 는 다음과 같다.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 독립변수 | Age  Guarantee | Age  Auto | Age  Guarantee  Auto | Age^2  Guarantee  Auto | Age  Guarantee  Auto  WarantyLeft |
| Adj. R2 | 0.7683 | 0.7004 | 0.70 | 0.67 | 0.7718 |

Adjusted R2는 통상 모델에 대한 설명력을 나타내는 것이라고 한다. Adjusted R2에만 의존하여 모델을 선정할 경우, Age, Guarantee period 혹은 Age, Guarantee period, Auto, WarrantyLeft를 선택해야 할 것이다.

하지만 Adjusted R2는 모델 선정에 기준이 되지 못한다. 값이 낮아도 상관관계 자체가 유의미할 수 있기 때문이다. 따라서 Adj. R2에 관계 없이 가장 경영학적인 implication이 높은 모델을 선택해야 할 것이다. 따라서 Roger은 네 개의 독립변수 모두에 대하여 유의했고, 잔여보증기간에 대한 상관관계를 알 수 있는 마지막 모델(Age, Guarantee period, Auto, WarrantyLeft)을 선택하는 것이 경영학적으로 의미가 있다고 볼 수 있겠다.

From the question (6), create dummy variables for the categorical variables (e.g., Fuel\_Type) and split the data into training (60%) and validation (40%).

(8) Develop a multiple regression model with all variables you have and report your model as an equation. Also, please compare adjusted R square values between two models (one from (6) and the other from (8)) models. Then, Which model do you recommend Roger use? Justify your answer.

Dummy variable을 다음과 같이 만들었다.

1) Color( default : Red) 2) Fuel\_Type(Default : CNG) 3) CC(Default : CC > 1600)

모든 변수를 넣었을때 Adj. R2 : 0.91

R2 값 자체만 놓고 본다면, 모든 변수를 넣은 회귀식이 더 성능이 좋다고 할 수 있겠다. 그러나 회귀식 결과를 자세히 보면, 유의하지 않은 P값을 가진 변수는 무려 17가지나 된다. 즉, 회귀식 자체가 가격과 다른 변수들의 상관을 정확히 나타내고 있지 않다는 것을 의미한다. 또한 모든 변수의 특성은 당연히 가격에 영향을 미친다. 따라서 이러한 회귀식에서 어떤 비즈니스 인사이트를 도출해내는 것은 매우 어렵다. 따라서 기존 final model을 선택하는 것이 바람직해보인다.

**(2) Blog Post Response (20 points)**

### Background

This data originates from blog posts. The raw HTML-documents of the blog posts were crawled and processed. The prediction task associated with the data is the prediction of the number of comments in the upcoming 24 hours. In order to estimate this situation, we choose a basetime (in the past) and select the blog posts that were published at most 72 hours before the selected base date/time. Then, we calculate all the features of the selected blog posts from the information that was available at the basetime, therefore each instance corresponds to a blog post. The target (or depend variable) is the number of comments that the blog post received in the next 24 hours relative to the basetime.

### Variable Descriptions

1...50: Average, standard deviation, min, max and median of the

Attributes 51...60 for the source of the current blog post. With source we mean the blog on which the post appeared. For example, myblog.blog.org would be the source of

the post myblog.blog.org/post\_2010\_09\_10

51: Total number of comments before basetime

52: Number of comments in the last 24 hours before the basetime

53: Let T1 denote the datetime 48 hours before basetime, Let T2 denote the datetime 24 hours before basetime. This attribute is the number of comments in the time period between T1 and T2

54: Number of comments in the first 24 hours after the publication of the blog post, but before basetime

55: The difference of Attribute 52 and Attribute 53

56...60: The same features as the attributes 51...55, but features 56...60 refer to the number of links (trackbacks), while features 51...55 refer to the number of comments.

61: The length of time between the publication of the blog post and basetime

62: The length of the blog post

63...262: The 200 bag of words features for 200 frequent words of the text of the blog post

263...269: binary indicator features (0 or 1) for the weekday (Monday...Sunday) of the basetime

270...276: binary indicator features (0 or 1) for the weekday (Monday...Sunday) of the date of publication of the blog post

277: Number of parent pages: we consider a blog post P as a parent of blog post B, if B is a reply (trackback) to blog post P.

278...280: Minimum, maximum, average number of comments that the parents received

281: The target: the number of comments in the next 24 hours (relative to basetime)

### Questions

Using the dataset above, perform a regression analysis, and comment on your results.

1. Do your best to create your final model and show all your steps (e.g., how to select your variables)

predictive regression analysis를 진행하였다.

Target : basetime 이후 24시간동안 받은 comment

과정 :

V1-V50, 200 bags of words를 제외한 모든 attribute를 가지고 Backward Elimination을 통해 RMSE가 최소화되는 변수를 추렸다.(단, 다음 핵심 변수들은 comment Prediction에 중요하다고 판단되어 제거하지 않았다.)

1. 24시간 전 코멘트 개수(52)

2. 포스팅 후 24시간 코멘트 개수(54)

3. Basetime기준 0~-24시간, -24~-48시간 사이에 달린 코멘트 개수(55)

4. 1,2,3번 기준에 대한 trackback 개수(57,59,60)

5. basetime – posting upload hour : 이 기간이 길어질수록 블로그의 유저 유입이

적어진 다고 볼 수 있다. (61)

6. Basetime의 요일 : basetime이 주말/ 주일인지가 comment에 영향을 끼칠

가능성이 높다 판단하였다. (264~269)

7. average of parents’ comment received : parent posting의 comment 개수 대신

선정하였는데, 그 이유는 그 parent posting안에 trackback한 블로그가 많으면 유입이 분산되기 때문에 활동성이 높은 유저가 나의 블로그로 유입이 되지 않기 때문이다.(280)

Interaction term은 다음과 같이 선정하였다.

1. 24시간전 trackback \* comment 개수 (57 \* 52)

이유 : 코멘트와 trackback의 개수는 활동성이 높은 유저의 유입을 의미한다. 트랙백을 달면 더 많은 코멘트가 달릴 가능성이 높고, 코멘트가 많을 수록 트랙백으로 들어와서 참여율 높은 유저들이 지나치지 않고 코멘트 남길 가능성이 높다. (답장을 잘 받을거라 생각해서)

2. Total number of trackbacks(56) / average Number of comments that parents

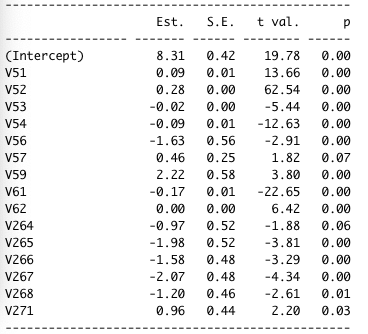
received(280)

이유 : trackback의 업로드 수에 비해 parent의 comment 개수의 비율이 적다면, trackback을 통해 유저가 내 블로그에 유입되는 비율이 적다는 것을 의미한다. 이는 코멘트 개수에 영향을 미칠 것이다. 반대의 경우도 마찬가지다.

Analysis Result

1번)

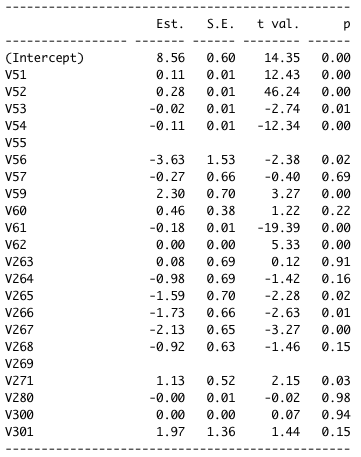
Backward Elimination만 진행한 분석 결과



RMSE : 31.24

2번)

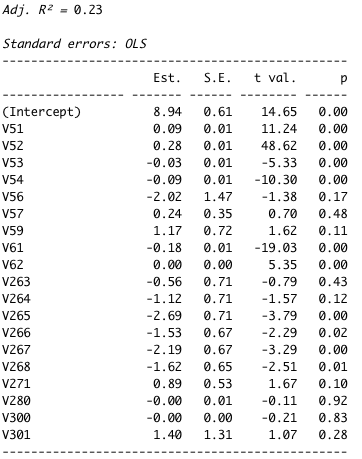
Backward Elimination + 핵심 변수 + Interaction term(V300,V301) 추가



RMSE : 32.94

3번)

Backward Elimination + Interaction term(V300,V301)만 추가



RMSE : 31.3

1. Suggest your final model as an equation. (based on Adjusted R-Square)

해당 모델은 Predictive regression analysis이다. 따라서 Adj. R-Square값보다 RMSE값이 모델을 평가하는데 더 중요하다. Minimun RMSE를 가진 모델은 1번 모델이다(RMSE = 31.24)

하지만 핵심변수와 interaction term이 추가되어있는 2번 모델은 1번 모델보다 RMSE값은 높지만, 더 많은 비즈니스 인사이트를 함의하고 있다. 따라서 2번 모델을 final 모델로 지정한다.

Final Model :

Y = 8.56 + 0.11 V51 + 0.28 V52 + -0.02 V53 + -0.11 V54 + -3.63 V56 + 0.27 V57 + 2.30 V59 + -0.18 V61 + -0.01 V62 + -0.5 V263 + -1.12 V264 + -2.69 V265+ -1.53 V266 + -2.19 V267+ -1.62 V268 + 0.89 V271 + -0.01 V280 + -0.01 V300 + 1.40 V301

1. Include at least one interaction term into your final model

Interaction term을 위에서 이미 넣고 진행하였다. Interaction term은 다음과 같다.

1. 24시간전 trackback \* comment 개수 (57 \* 52) <- V300

이유 : 코멘트와 trackback의 개수는 활동성이 높은 유저의 유입을 의미한다. 트랙백을 달면 더 많은 코멘트가 달릴 가능성이 높고, 코멘트가 많을 수록 트랙백으로 들어와서 참여율 높은 유저들이 지나치지 않고 코멘트 남길 가능성이 높다. (답장을 잘 받을거라 생각해서)

2. Total number of trackbacks(56) / average Number of comments that parents received(280) <- V301

이유 : trackback의 업로드 수에 비해 parent의 comment 개수의 비율이 적다면, trackback을 통해 유저가 내 블로그에 유입되는 비율이 적다는 것을 의미한다. 이는 코멘트 개수에 영향을 미칠 것이다. 반대의 경우도 마찬가지다.

1. Discuss the interpretations in plain English and various output statistics (i.e., interpret the coefficient of each independent variable, significance of each independent variable, which independent variables are important? etc).

Comments within 24 hours (V52, Coefficient: 0.28) are more affected than total comment (V51, Coefficient 0.11). Comments after 24 hours do not affect.

However, in the trackback, the total trackback has the greatest impact (V56, coefficient:-3.63), rather, the more total trackback, the less comment is available, indicating that users who come in on the trackback only look at the posts or have no activity on the blog. However, within 24 hours of the trackback posting, there is a lively comment (V59, Coefficient: 2.3), so **updating the new trackback just before basetime can receive more comments.**  
  
 Total number of trackbacks (V56) / average number of accounts that parameters received (V280), the second interaction term, shows a high amount of correlation (Coefficient: 1.97). This can be seen that the more trackback is done on the lower number of comment posts, the more Comments run. Therefore, it is possible to **track back the post that seems to have less interaction than to track back the post that is popular and active among the posts**.

Negative correlation with the number of comments on weekdays. However, Tuesday's posting received more comments.

1. Provide all your processes in a word file including results, interpretation, and your comments, R-codes, etc.

# Part 3: German Credit Case (30 points)

### Background

The German credit dataset has 21 variables and 1,000 records, each record being a prior applicant for credit. Assume that you are the expert in data mining at this company. You need to form two groups – those with good credit rating or those with bad credit rating.

For these 1,000 applicants for credit, you need to evaluate them based on these 21 predictors. It is possible that not all predictors (or variables) are useful.

### Your Task

You need to use K-mean Clustering classification tree, kNN and neural network (or logistic regression) to split the customers into good credit or bad credit based on the above variables. First, please evaluate which variables can impact a customer’s credit rating. **Select 10~21 out of 21 variables that you think are useful for identifying a customer’s credit rating**. Explain briefly how and why you picked up the 10 or 21 variables. Now, use R with 2 customer groups.

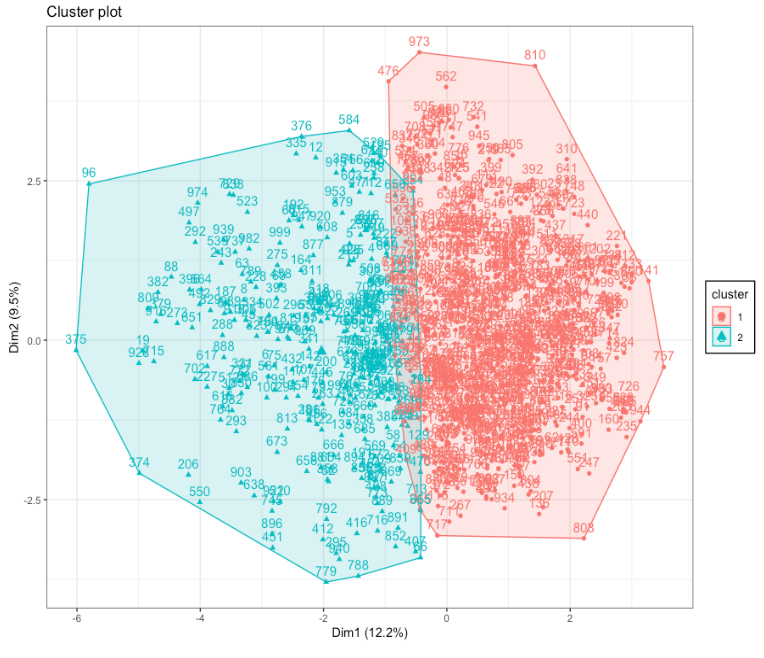
선정항목은 다음과 같다.

|  |  |  |
| --- | --- | --- |
| 선정항목 | 제외항목 | 이유 |
|  | "OBS." | 분석에 의미가 없는 항목 |
| "CHK\_ACCT" |  | 계좌 상태는 rating에 중요한 지표 |
| "DURATION" |  | 신용거래기간 역시 ratingdp 중요한 지표 |
| "HISTORY" |  | Histroy는 신용도를 평가할때 기준이 된다 생각한다 |
| "NEW\_CAR" |  | 이용자의 자산에 대한 지표 |
| "USED\_CAR" |  | 이용자의 자산에 대한 지표 |
|  | "FURNITURE" | 액수가 크지 않은 금액이라 배제하였다 |
|  | "RADIO.TV" | 상동 |
| "EDUCATION" |  | 교육기간은 소득에 영향을 미치므로 포함 |
| "RETRAINING" |  | 보유액 역시 변제능력에 영향을 미친다 |
| "AMOUNT" |  | 대출가능금액은 신용평가에 있어 중요한 척도라 판단 |
| "SAV\_ACCT" |  | 당좌예금은 지급능력에 영향을 끼친다 |
| "EMPLOYMENT" |  | 고용상태는 변제능력에 영향을 끼친다 |
| "INSTALL\_RATE" |  | 할부금의 액수 크기가 클수록 Rating에 부정적인 영향을 끼친다고 판단 |
|  | "MALE\_DIV" | 인간관계가 rating과 상관관계가 낮을거라판단 |
|  | "MALE\_SINGLE" | 상동 |
|  | "MALE\_MAR\_or\_WID" | 상동 |
|  | "CO.APPLICANT" | 상동 |
| "GUARANTOR" |  | 보증인이 있으면 신용평가가 좋아진다 |
|  | "PRESENT\_RESIDENT" | Rating과 크게 연관성이 있다고 생각되지 않는다 |
| "REAL\_ESTATE" |  | 부동산 소유 유무는 영향을 끼친다고 판단하였다 |
| "PROP\_UNKN\_NONE" |  | 상동 |
|  | "AGE" | 나이가 실질적인 영향을 크게 끼치지 않는다 판단하였다 |
| "OTHER\_INSTALL" |  | Install\_Rate과 같은 이유로 포함하였다 |
| "RENT" |  | Real\_Estate과 상동한 이유 |
| "OWN\_RES" |  | 위와 상동 |
| "NUM\_CREDITS" |  | 액수의 크기가 신용도에 영향을 끼친다 |
|  | "JOB" | Job이 실질적인 Rating에 큰 영향을 주지 않는다 판단하였다. |
|  | "NUM\_DEPENDENTS" | 부양가족에 따른 영향은 크게 중요하지 않다고 판단하였다 |
| "TELEPHONE" |  | 당시 휴대전화가격은 높았기 때문에 경제력의 척도가 된다 판단하였다. |
|  | "FOREIGN" | 외국인의 유무는 중요하지 않다 판단 |

### Questions

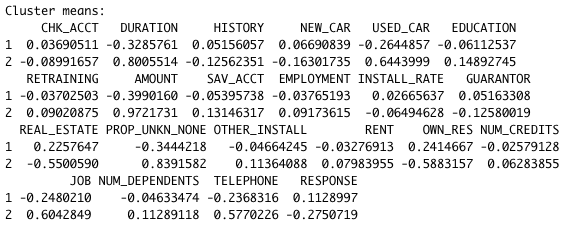
1. Please see the clustering result’s outputs. Each record belongs to cluster 0 or cluster 1. Please use your judgment based on the data whether cluster 0 or cluster 1 is “good rating” or “bad rating” Explain your judgment.

위 요인을 가지고 모델링을 한 결과 다음과 같은 결과가 나왔다.

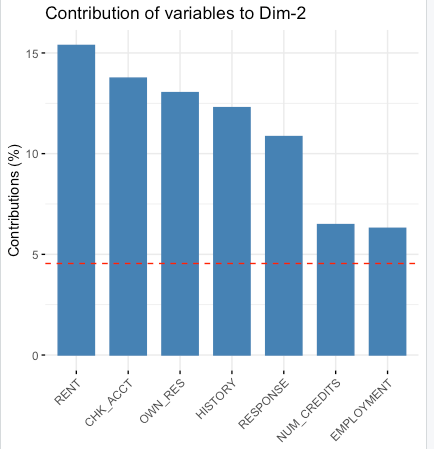
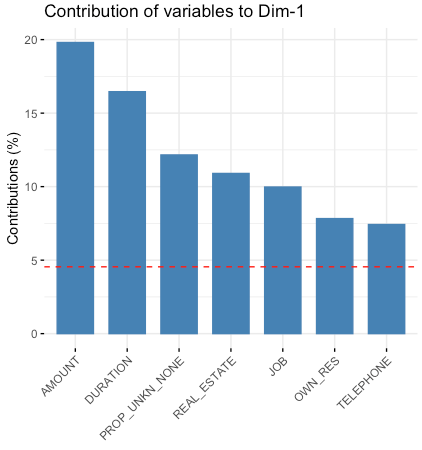


클러스터가 Dim1에 따라 나뉜 것을 파악했다. 클러스터의 특성을 파악하기 위해 다음과 같은 정보를 활용했다.

1. kmeans의 수치적인 결과값



2. 차원을 구성하고 있는 요인, 공헌도



1번 클러스터 : 신용거래이력이 좀 더 많고, 신차신용거래가 많으며, credit amount가 매우 낮다. 통잔 잔고가 적으며, 대부분 부동산을 소유하고 있다.

2번 클러스터 : 신용거래이력이 적으며, 대부분 중고차를 보유하고 있다. credit amount가 높고, 통장 잔고가 많다. 대부분 부동산을 소유하고 있지 않으며, Rent를 많이 이용한다.

Dimension contribution chart에서는 모델링 결과 특성상 특히 Dimension1 chart를 자세히 볼 필요가 있다. Amount와 Duration, Applicant owns no property, Real\_estate가 Contibution의 60%를 차지하고 있다. 이는 신용거래한도와 신용기간, 부동산 소유의 유무가 가장 큰 영향을 끼쳤다는 것을 의미한다.

따라서 둘을 고려한 결과, **1번 클러스터 : Good rating, 2번 클러스터 : Bad rating으로 선정했다.**

1번 클러스터의 특징은 대부분 부동산을 소유하고 있다는 점이다. 즉 유동자산보다 비유동자산이 더 많다. credit amount가 낮은 점, 신용거래이력이 많은 점을 봤을때 부동산 투자를 위한 신용거래를 많이 한 것으로 보인다. 이 영향으로 현금자산이 상대적으로 적고, 그렇기에 신용거래가 활발하다. 하지만 신용거래이력과 credit amount는 부동산 투자로 인한 영향도 있다 볼 수 있으므로, 위 두 요인이 rating에 끼치는 영향을 오히려 적다고 판단할 수 있겠다. 오히려 부동산의 소유로 인해 좋은 rating을 받을 수 있다.

2번 클러스터는 신용거래이력이 적고 credit amount가 높다. 보유하고 있는 현금자산도 1번 클러스터에 비해 많은 편이다. 하지만 대부분 중고차를 보유하고 있고, Rent의 비율이 높은 점을 봤을때, 특별한 투자 없이 월급 등의 단기현금자산으로 자산운용을 하고 있을 가능성이 높다. 즉, good rating을 받을만한 기타 자산이 없는 것이다. 이 군집은 당월 근로 소득, 연금 등으로 바로 rent fee를 내고, 투자 목적보다는 생활 목적의 자산 운용을 할 가능성이 높다.중고차 보유의 비중이 높은 것은 큰 돈을 운용할 여건이 되지 않다는 것을 나타내기도 한다. 따라서 Bad rating 군집으로 선정하였다.

1. **Now compare the actual credit rating against that of R clustering results.**

Since you have many observations, obviously it becomes difficult to evaluate how many were correct or wrong. So, use Pivottable!

You need to create a Pivot Table like this:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| Count of Correct? | Predicted Response | |  |
| RESPONSE | 0 | 1 | Grand Total |
| 0 | 124 | 176 | 300 |
| 1 | 167 | 533 | 700 |
| Grand Total | 291 | 709 | 1000 |

1. Now, interpret the result. Do you think the clustering method was a good one? Why or why not?

정확도 : 65.7%

Prediction을 위한 모델으로써는 적절하지 못하다. Attributes에 따라서 정확도가 달라질텐데, clustering analysis에서는 정확도를 최대화할 수 있도록 모델링을 하기가 번거롭다. 따라서 고객군의 일반적인 특성을 파악하는데에는 더없이 좋은 모델이지만, rating을 예측하는데에는 좋은 모델이 되기 힘들다.

1. If net profit for correct prediction was $100 and -$300 for bad choice (i.e., predicted by good, but the customer was bad) did the firm benefit from data mining? Note – if a customer was indeed good, but data mining predicted “bad” is lost opportunity which costs the firm $100 as well (Please estimate the cost manually (e.g., use Excel))

1) profit from correct prediction : 100 \*(124 + 533) = 65,700($)

2) deficit from bad choice : -300 \*176 + -100 \* 167 = -69,500($)

total earning from analysis : -3,800($)

따라서 기업은 analysis model을 사용하면 손실을 입는다.

1. Try all these processes (including variable selection) by using classification tree, kNN & neural network (or logistic regression) techniques and compare the results with cluster one. Which one is better? You need to answer this question based on both the error rate and cost!

1)위와 상동한 attribute로 Classification tree를 돌린 결과(Train : Test = 7:3)

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| Count of Correct? | Predicted Response | |  |
| RESPONSE | 0 | 1 | Grand Total |
| 0 | 46 | 35 | 81 |
| 1 | 41 | 178 | 219 |
| Grand Total | 87 | 213 | 300 |

정확도 : 75%

Cost

from correct prediction : 100 \*(46 + 178) \* 10/3 = 74,666($)

2) deficit from bad choice : (-300 \*35 + -100 \* 41) \*10/3 = -48,666($)

total earning from analysis : 26,000 ($)

2) kNN으로 돌린 결과

1. 같은 attribute로 kNN를 돌린 결과(Train : Test = 7:3, k = 22)

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| Count of Correct? | Predicted Response | |  |
| RESPONSE | 0 | 1 | Grand Total |
| 0 | 27 | 66 | 93 |
| 1 | 14 | 193 | 207 |
| Grand Total | 41 | 259 | 300 |

정확도 : 73.4%

2. Best K로 모델링을 한 결과( k = 9)

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| Count of Correct? | Predicted Response | |  |
| RESPONSE | 0 | 1 | Grand Total |
| 0 | 36 | 57 | 93 |
| 1 | 17 | 190 | 207 |
| Grand Total | 53 | 247 | 300 |

정확도 : 75.3%

(k = 9)

Cost from correct prediction : 100 \*(36 + 190) \* 10/3 = 75,334($)

deficit from bad choice : (-300 \*57 + -100 \* 17) \*10/3 = -62,677 ($)

total earning from analysis : 12,657($)

3) Logistics regression analysis를 이용한 결과(Train : Test = 7:3)

Tree와 같은 변수를 사용하였다.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| Count of Correct? | Predicted Response | |  |
| RESPONSE | 0 | 1 | Grand Total |
| 0 | 47 | 34 | 81 |
| 1 | 31 | 188 | 219 |
| Grand Total | 78 | 222 | 300 |

정확도 : 78.3%

Cost from correct prediction : 100 \*(47 + 188) \* 10/3 = 78,333($)

deficit from bad choice : (-300 \*34 + -100 \* 31) \*10/3 = -44,333 ($)

total earning from analysis : 34,000($)

결론

모든 모델 중 error rate이 가장 낮고, 그에 따른 total earning이 제일 높은 모델은 logistics regression model이다. Logistics regression model이 정확도 78.3%, total earning $34,000으로 제일 좋은 아웃풋을 보였다. Clustering은 고객의 군집별 특성을 파악하고, 그걸 통해 마케팅 전략을 수립한다거나 promotion 전략을 짜는데 용이하다.

하지만 기존 데이터를 가지고 prediction을 하기 위한 용도로써는 적절하지 않다. 이는 tree와 다른 모델의 error rate차이에서도 잘 드러난다. Cost 측면에서도, error rate 측면에서도 clustering analysis는 prediction에 있어 가장 안좋은 퍼포먼스를 보였다. 따라서 목적에 따른 신중한 analysis를 선정하는 것이 중요하다.

**Part1: Optimization**

# (1) Electro-poly Corporation Manufacturer’s Make VS. Buy Decisions (5 points)

The Electro-poly Corporation is the world’s leading manufacturer of slip rings. A slip ring is an electrical coupling device that allows current to pass through a spinning or rotating connection – such as a gun turret on a ship, air-craft, or tank. The company recently received a $750,000 order for various quantities of three types of slip rings. Each slip ring requires a certain amount of time to wire and harness. The following table summarizes the requirements for the three models of slip rings.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Model 1 | Model 2 | Model 3 |
| Number ordered | 3,000 | 2,000 | 900 |
| Hours of wiring required per unit | 2 | 1.5 | 3 |
| Hours of harnessing required per unit | 1 | 2 | 1 |

Unfortunately, Electro-Poly does not have enough wiring and harnessing capacity to fill the order by its due date. The company has only 10,000 hours of wiring capacity and 5,000 hours of harnessing capacity available to devote to this order. However, the company can subcontract any portion of this order to one of its competitors. The unit costs of producing each model in-house and buying the finished products from a competitor are summarized below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Model 1 | Model 2 | Model 3 |
| Cost to Make | $50 | $83 | $130 |
| Cost to Buy | $61 | $97 | $145 |

Electro-Poly wants to determine the number of slip rings to make and the number of buy in order to fill the customer order at the least possible cost.

# What to do:

(a) Write the decision variables, objective function, and constraints.

Decision variables :

x1 : number of Model 1 to make

x2 : number of Model 2 to make

x3 : number of Model 3 to make

y1 : number of Model 1 to buy

y2 : number of Model 2 to buy

y3 : number of Model 3 to buy

Constraints :

1. 2x1 + 1.5x2 + 3x3 <= 10000
2. 1x1 + 2x2 + 1x3 <= 5000
3. X1 + y1 =3000
4. X2 + y2 = 2000
5. X3 + y3 = 900
6. X1,x2,x3,y1,y2,y3 >= 0

Objective function :

- min( 50x1 + 83x2 + 130x3 + 61y1 + 97y2 + 145y3)

(b) Use the template and solve the problem using an Excel solver: what is your optimal solution?

|  |  |  |  |
| --- | --- | --- | --- |
| Number to | Model1 | Model2 | Model3 |
| Make | 3000 | 550 | 900 |
| Buy | 0 | 1450 | 0 |
| Total | 3000 | 2000 | 900 |

다음과 같이 만든다. 이럴 경우, **total cost = $453,300**

# (2) Building Apartment (10 points)

A real estate developer is planning to build an apartment building specifically for graduate students on a parcel of land adjacent to a major university. Four types of apartments can be included in the building: efficiencies, and one-, two, or three-bedroom units. Each efficiency requires 500 square feet; each one-bedroom apartment requires 700 square feet; each two-bedroom apartment requires 800 square feet; and each three-bedroom unit requires 1,000 square feet. The developer believes that the building should include no more than 15 one-bedroom units, 22 two-bedroom units, and 10 three-bedroom units. Local zoning ordinances do not allow the developer to build more than 40 units in this particular building location, and restrict the building to a maximum of 40,000 square feet. The developer has already agreed to lease 5 one-bedroom units and 8 two-bedroom units to a local rental agency that is a "silent partner" in this endeavor. Market studies indicate that efficiencies can be rented for $350 per month, one-bedrooms for $450 per month, two-bed rooms for$550 per month, and three-bedrooms for $750 per month. How many rental units of each type should the developer include in the building plans in order to maximize the potential rental income from the building?

# What to do:

(a) Write the decision variables, objective function, and constraints.

Decision variables :

|  |
| --- |
| 1. number of efficiencies to build(x1) |
| 2. number of one-bedrooms to build(x2) |
| 3. number of two-bedrooms to build(x3) |
| 4. number of three-bedrooms to build(x4) |
| Objective Function  MAX(350x1 + 450x2 + 550x3 + 750x4)  Constraints   |  | | --- | | 1. 500x1 + 700x2 + 800x3 + 1000x4<= 40000 | | 2. x1 + x2 + x3 + x4 <= 40 | | 3. 5 <= x2 <=15 | | 4. 8 <= x3 <= 22 | | 5. x4 <= 10 | |

(b) Set-up your own Excel Spreadsheet

(c) Solve the problem using an Excel solver: what is your optimal solution?

Efficiency : 0

one-bed : 8

two-bed : 22

three-bed : 10

**Total income : $23,200**

# (3) Power distribution (10 points)

Assume that you need to distribute electricity to each city efficiently. The following table gives the distances between Boston, Chicago, Dallas, Los Angeles, and Miami. Each city requires 40,000 kilowatt hours (kWh) of power, and Chicago, Dallas, and Miami are capable of producing 70,000 kWh. Assume that shipping 1000 kWh over 100 miles costs $4. From where should power be sent to minimize the cost of meeting each city’s demand?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Boston | Chicago | Dallas | Los Angeles | Miami |
| Chicago | 983 | 0 | 1205 | 2112 | 1390 |
| Dallas | 1815 | 1205 | 0 | 801 | 1332 |
| Miami | 1539 | 1390 | 1332 | 2757 | 0 |

# What to do:

(a) Write the decision variables, objective function, and constraints.

|  |
| --- |
| Decision Variables |
| 1. Amount of Boston |
| 2. Amount of Dallas |
| 3. Amount of LA |
| 4. Amount of Miami |
| 5. Amount of Chicago |

|  |  |
| --- | --- |
| objective function : | |
| minimize the cost of distributing from supply site to each city | |
|  | |
| constraints |
| ChiBo + ChiChi + ChiDal + ChiLA + ChiMia <= 70000 |
| DalBo + DalChi + DalDal + DalLA + DalMia <= 70000 |
| MiaBo + MiaChi + MiaDal + MiaLA +MiaMia <= 70000 |
|  |
| ChiBo + DalBo + MiaBo >= 40000 |
| ChiDal + DalDal + MiaDal >= 40000 |
| ChiChi + DalChi + MiaChi >= 40000 |
| ChiLA + DalLA + MiaLA >= 40000 |
| ChiMia + DalMia + MiaMia >= 40000 |

(b) Set-up your own Excel Spreadsheet

(c) Solve the problem using an Excel solver: what is your optimal solution?You need to find the **minimum** cost of distributing electricity to each city.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Boston | Chicago | Dallas | LA | Miami |
| Chicago | 30000 | 40000 | 0 | 0 | 0 |
| Dallas | 0 | 0 | 30000 | 40000 | 0 |
| Miami | 10000 | 0 | 10000 | 0 | 40000 |

**Minimum cost : $3,610**

# Part2: Association Rules

# Catalog Cross Selling Case (10 points)

### Background

Exeter, Inc., is a catalog firm that sells products in a number of different catalogs that it owns. The catalogs fall into eight basic categories:

1. Clothing (2) Housewares (3) Health (4) Automotive (5) Personal electronics (6) Computers (7) Garden (8) Jewelry.

The costs of printing and distributing catalogs are high. By far the biggest cost of operation is the cost of promoting products to people who buy nothing. Having invested so much in the production of artwork and printing of catalogs, Exeter wants to take every opportunity to use them effectively. One such opportunity is in cross selling – once a customer has “taken the bait” and purchases one product, try to sell them another while you have their attention.

Such cross promotion might take the form of enclosing a catalog in the shipment of a purchased product, together with a discount coupon to induce a purchase from that catalog. Or it might take the form of a similar coupon sent by e-mail, with a link to the Web version of that catalog.

But which catalog should be enclosed in the box or included as a link in the e-mail with the discount coupon? Exeter would like it to be an informed choice – a catalog that has a higher probability of inducing a purchase than simply choosing a catalog at random.

### Questions

Using the dataset CatalogCrossSellData.csv, perform an association rules analysis, and comment on the results. Answer the following questions based on the association rules:

Assume it costs $3 to mail each distinct catalog to each user. Exeter has 50,000 customers.

1. How much would it cost the firm to send all catalogs to all customers?

단순 배송비 : 3(cost) \* 50,000(customers) = $150,000

만약 catalog 하나를 만드는데 $X만큼 든다면,

8X(categories) \* 3(cost) \* 50,000(customers) = ($)1,200,000X

2. Now from the association rules that you generated identify the top 5 rules. Provide the rationale and interpret the rules in plain English using confidence and lift ratios.

When selecting rules, it is preferable to consider lift ratio rather than confidence. The reason for this is that the confidence is a good indicator of the number of transactions containing both / the number of transactions containing antecedent, it abbreviates about the all transaction factors.

Therefore, simple confidence alone does not, in substance, know how much influence the rule has on the entire transaction. Therefore, it is desirable to choose rules based on the lift ratio. The following is a summary of the top five rules based on Lift ratio.

(Setup constraints : support >0.2, confidence >= 0.8)

Top5 rules satisfying setup contrints were selected based on the lift ratio.   
1) {Clothing} => {Jewerly}    
2) {Computers,Health} => {Personal.Electronics}  
3){Automotive,Personal.Electronics} => {Computers}    
4){Computers} => {Personal.Electronics}  
5) {Personal.Electronics} => {Computers}    
  
rule 4 and 5 were excluded because they were a natural result, and then successors with high lift ratio following top5 were selected.   
  
6){Personal.Electronics,Health} => {Computers}    
7) {Personal.Electronics,Health} => {Automotive}

**So Final Top 5 rules are as follows :**

|  |  |  |
| --- | --- | --- |
| Clothing | => | Jewerly |
| Computers,Health | => | Personal.Electronics |
| Automotive,Personal.Electronics | => | Computers |
| Personal.Electronics,Health | => | Computers |
| Personal.Electronics,Health | => | Automotive |

3. Assume a customer just bought an automotive product. If you have a limit of $6 to spend on promotion per customer, which catalogs would you give this customer and why?

Rule에 따르면 Automotive는 personal electrics또는 computers랑 묶여있다. automotive product 구매자는 IT기기에 관심이 많을 가능성이 높다. 따라서 computers, 혹은 automotive interest에 관련된 personal electronics, 예를 들어 휴대용 공기청정기 같은 상품을 메인으로 한 catalog를 제공할 것이다.

### What to Turn in?

Submit a Word document with answers for the following questions and Excel spreadsheet (Of course, R output too)

1. The above Pivot table output (you can use Excel spreadsheet for Pivottable)