**통계적 분석 방법론**

**기말 레포트**

2019021059 조윤영

목차

[**1.** **데이터 설명 및 EDA** 4](#_Toc27515061)

[- 각 변수 별 churn 막대그림 4](#_Toc27515062)

[**2.** **분할표 분석** 7](#_Toc27515063)

[1) Odds ratio 확인 7](#_Toc27515064)

[2) 연관성의 검정 – 카이제곱 검정 7](#_Toc27515065)

[3) 연관성의 검정 – 코크란 맨텔 헨젤 검정 7](#_Toc27515066)

[- Internet Service 8](#_Toc27515067)

[- Contract 8](#_Toc27515068)

[**3.** **로그 선형 모형 및 로짓 모형** 9](#_Toc27515069)

[1) 로그 선형 모형 9](#_Toc27515070)

[- Internet Service (X) & Churn (Y) & Contract (Z) 9](#_Toc27515071)

[- Senior Citizen(X) & Churn (Y) & Gender (Z) 9](#_Toc27515072)

[- Internet Service (X) & Churn (Y) & Dependents (Z) 10](#_Toc27515073)

[2) 로짓 모형 적합 11](#_Toc27515074)

[**4.** **생존분석** 12](#_Toc27515075)

[1) 생명표 방법 12](#_Toc27515076)

[2) 누적한계추정법 – 카플란 마이어 13](#_Toc27515077)

[3) Log-rank test 13](#_Toc27515078)

[4) Cox 비례 위험 모델 14](#_Toc27515079)

[분석에 앞서 주어진 자료가 비례 위험 모형의 가정을 따르는 지 확인해야 한다. 14](#_Toc27515080)

[**5.** **Support Vector Machine** 15](#_Toc27515081)

[**6.** **Tree model** 16](#_Toc27515082)

[**7.** **로버스트 회귀** 18](#_Toc27515083)

1. **데이터 설명 및 EDA**

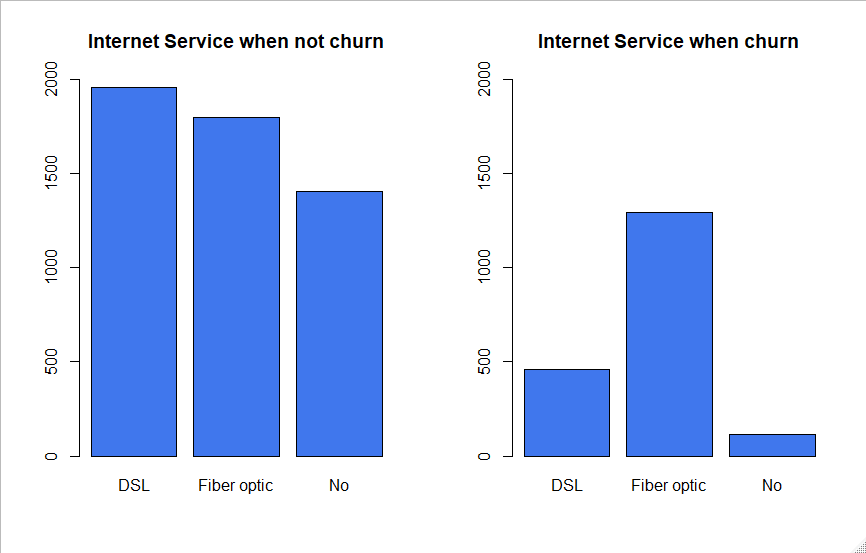
|  |  |  |
| --- | --- | --- |
| Variables |  | Description |
| gender | **C** | 고객의 성별 (female, male) |
| SeniorCitizen | **C** | 고객이 나이-노인 여부 노인일 경우 1 (1, 0) |
| Partner | **C** | 고객의 파트너 여부(Yes, No) |
| Dependents | **C** | 고객의 자녀 여부(Yes, No) |
| tenure | **N** | 고객이 회사와 함께 한 시간 (월 단위) |
| PhoneService | **C** | 고객이 핸드폰 서비스를 사용하는지 (Yes, No) |
| InternetService | **C** | 고객의 internet service (DSL, Fiber optic, No) |
| Contract | **C** | 고객의 계약 기간(Month-to-month, One year, Two year) |
| PaperlessBilling | **C** | 고객이 종이가 아닌 고지서를 받는지 (Yes, No) |
| PaymentMethod | **C** | 고객의 결제 방법 (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)) |
| MonthlyCharges | **N** | 한달동안 고객이 납부한 요금 |
| TotalCharges | **N** | 전체 기간동안 고객이 납부한 요금 |
| Churn | **C** | 고객이탈여부 (Yes or No) |

\* C: Categorical , N: Continuous variable

Internet Service에서 DSL은 전화연결선을 이용한 인터넷을 의미하고, Fiber optic은 작고 유연한 유리 와이어를 통해 이용하는 인터넷이다. Fiber optic은 DSL보다 빠른 속도를 자랑하지만, DSl보다는 덜 안정적이라는 단점이 있다. 따라서 음악이나 영상을 볼 때에는 Fiber optic이 낫지만, 인터넷을 하거나 메일을 보내는 등의 기본적인 활동을 할 때에는 DSL이 더 안정적이라고 할 수 있다.

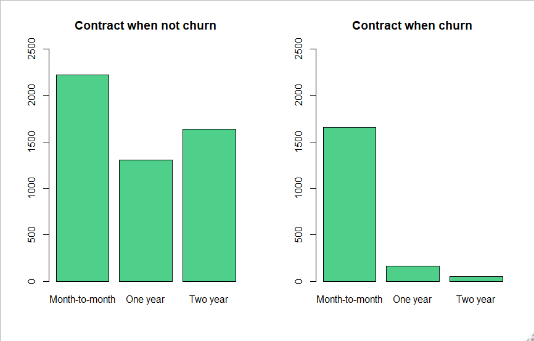
* 변수 별 churn에 따른 막대그래프

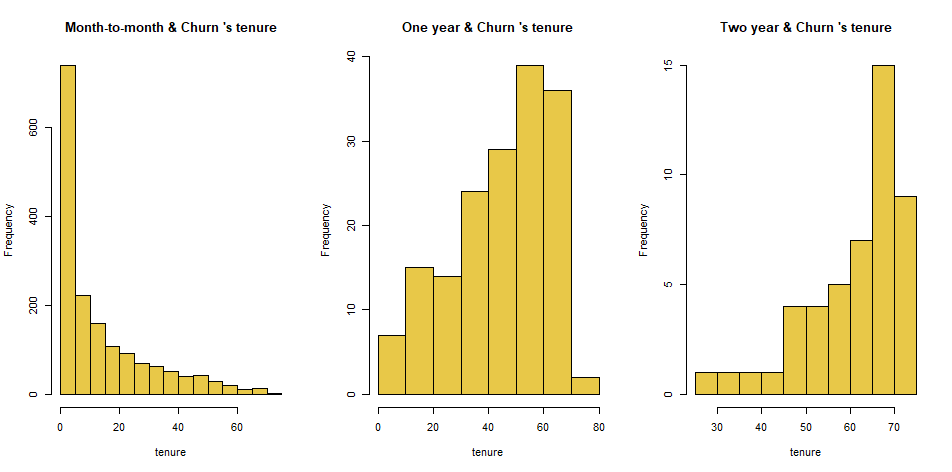
모든 변수가 아닌, 고객이탈에 따라 분포에 차이를 보였던 변수들을 골라 막대그래프를 그려보았다.

1. **Churn에 따른 Internet Service의 Bar plot**

internet service를 어떤 것을 이용하느냐에 따른 차이를 볼 수 있다. 이탈 고객 중에는 Fiber optic을 이용했던 고객이 많다.

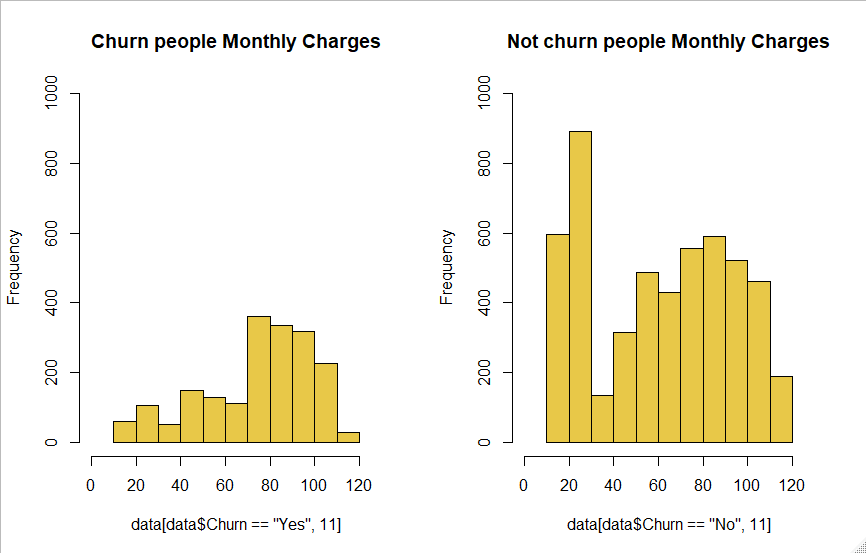
1. **Churn에 따른 Contract의 Bar plot**

어떤 계약을 하느냐에 따른 고객이탈에도 차이가 난다. 이탈하지 않는 고객들의 계약 종류는 Monthly-to month가 가장 많지만 다른 종류의 계약과 큰 차이가 나지 않는 반면, 이탈한 고객들 중에는 월별 계약이 가장 많았다.

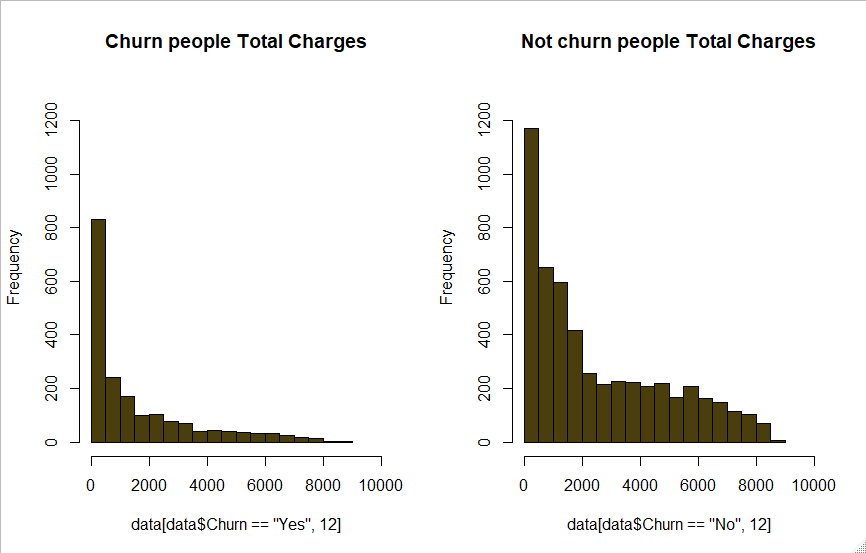


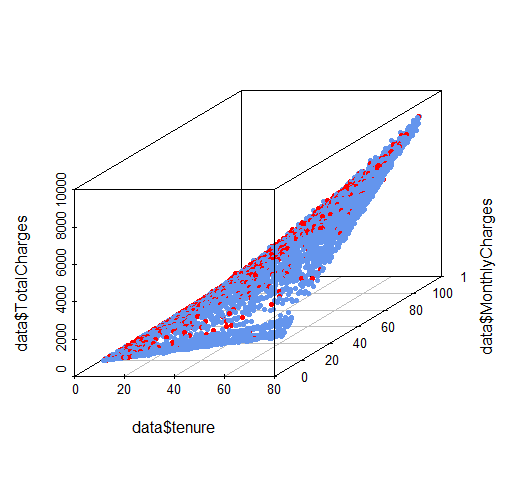
이탈한 사람들의 계약 별 통신사 유지기간을 확인해 보니, 역시 월단위의 계약을 할 경우, 한달만 이용한 뒤 이탈하는 고객이 가장 많았다.

1. **Churn에 따른 Monthly Charges bar plot**

이탈한 사람들의 한달 요금은 많이 나오는 축에 속했다. 반면 이탈하지 않는 사람들의 경우 요금이 아주 적게 나오는 사람들과 그렇지 않은 사람들 두 부류로 나누어 볼 수 있었다.

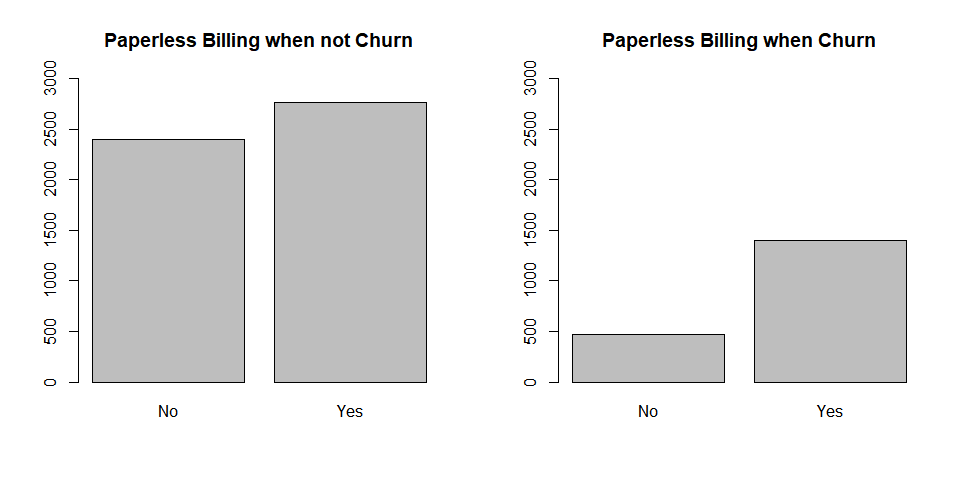
1. **Churn에 따른 Total Charges**

이탈한 사람들과 그렇지 않은 사람들은 비슷한 분포를 보였으며, 다만 이탈하지 않은 사람들의 경우 총 요금이 커질 수 밖에 없기에 오른쪽 꼬리부분이 더 두껍다.

1. **연속형 변수끼리의 Scatter plot**

빨간색이 이탈한 사람들임을 고려하면, 연속형 변수만을 이용한 그래프에서는 이탈자와 이탈하지 않은 고객들에게 일정한 패턴이 보이지 않는다.

1. **분할표 분석**
2. Odds ratio 확인

Churn과 범주가 2개인 변수 간의 odds ratio를 통해 목표인 Churn 변수와의 연관성을 확인했다.

|  |  |  |
| --- | --- | --- |
| Gender | Senior Citizen | Partner |
| 0.96 | 0.43 | 0.50 |
| Dependents | **Phone Service** | **Paperless Billing** |
| 0.40 | 1.10 | 2.58 |

Gender와 phone service는 churn과는 큰 연관성이 없는 것으로 보여 진다. Paperless Billing의 경우 위의 Histogram으로 보이며, 이때 churn한 집단에서 paper billing을 한 사람들이 churn안한 사람들보다 2.6배 많다고 해석할 수 있다. 이는 상식적으로 이해가 가는 것이, paper billing이 아닌 이메일이나 휴대폰, 어플 등으로 고지서를 받는 경우 따로 설정을 해야한다고 추측할 수 있으며, 이는 이 이 통신사에서 제공하는 서비스에 관심이 많은 것으로 해석할 수도 있다.

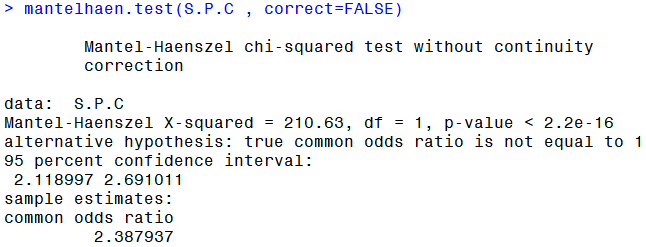
1. 연관성의 검정 – 카이제곱 검정

더불어, 앞서 oddsratio 분석에서 범주가 2개 이상이라 분석하지 못했던 변수들을 분석한다. Internet Service와 Contract 그리고 Payment Method를 카이제곱 검정으로 검정한다.

H01 : Internet Service와 Churn은 독립이다. vs H11 : Not H0   
H02 : Contract와 Churn은 독립이다. vs H12 : Not H0   
H03 : Payment Method와 Churn은 독립이다. vs H13 : Not H0   
세 검정 모두 p-value<0.0001 으로 신뢰수준 0.05하에서 귀무가설을 기각하였다. 세 변수 모두 Churn과 연관이 있다고 볼 수 있다.

H04 : gender와 Churn은 독립이다. vs H14 : Not H04의 경우 p-value=0.49로 신뢰수준 0.05하에서 귀무가설을 기각하지 못하였다. 성별과 고객이탈에는 연관성이 없다.

1. 연관성의 검정 – 코크란 맨텔 헨젤 검정

먼저 oddsratio에서 가장 큰 연관성을 보였던 Paper billing과 Churn의 관계를 검정해보고자 한다. 먼저, Paper Billing이 아닌 형태는 노인들의 경우 설정이 어려울 수도 있으므로, 노인을 층으로 두고 검정하고자 한다. H0 : Paper Billing과 Churn은 독립이다. vs H1 : Not H0 이다.

이때 p-value가 아주 작으므로 귀무가설을 기각한다. 따라서 Paper Billing과 Churn은 Senior여부를 층으로 주어도 독립이 아니다.

* Internet Service

Internet Service의 경우 Fiber optic과 DSL, 그리고 인터넷을 이용하지 않는 범주로 3가지가 있는데, 이는 Churn 여부와 관련이 있는지 분석해보고자 한다. 이때 인터넷의 용도가 다른 노인과의 차이가 있을 수 있으므로 Senior Citizen을 층으로 두었다. H0 : Internet Service와 Churn은 독립이다. vs H1 : Not H0 일때 코크란 맨텔헨젤 검정의 결과로 p-value가 <0.0001로 귀무가설을 기각하였다. 따라서 노인을 층으로 두어도 인터넷 서비스와 이탈은 연관성이 있다고 할 수 있다. 이는 앞서 했던 EDA와도 일치한다. 그렇다면 자녀의 여부도 인터넷 서비스에 영향을 줄 것이라는 생각에 똑같이 층으로 두어 분석해보았으나, 이 또한 아주 작은 p-value로 인터넷서비스와 이탈이 독립이라는 귀무가설을 기각하였다.

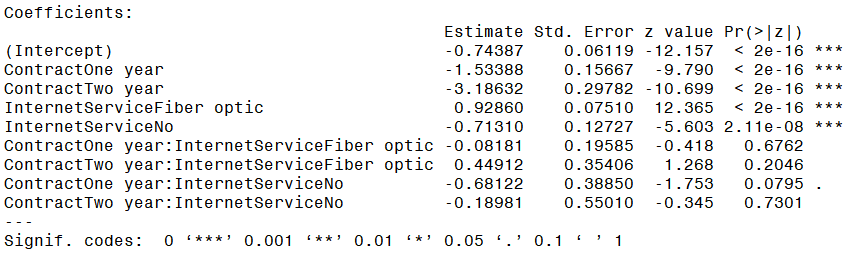
* Contract

함께 사는 이가 있는 것과, 아이가 있다는 것은 어떤 인터넷을 사용하느냐, 또는 언제 통신사를 바꾸냐에 영향을 줄 수도 있으므로, 층을 dependent로 둔 것과, partner로 둔 것으로 두가지 검정을 하고자 한다. H0 : Contract와 Churn은 독립이다. vs H1 : Not H0 이고, Partner를 층으로 두었을 때는 p-value가 <0.0001로 귀무가설을 기각하였으며, Dependent로 두었을 때에도 p-value가 <0.0001로 귀무가설을 기각하였다. 따라서 계약과 이탈에는 연관성이 있다고 할 수 있다.

1. **로그 선형 모형 및 로짓 모형**
2. 로그 선형 모형

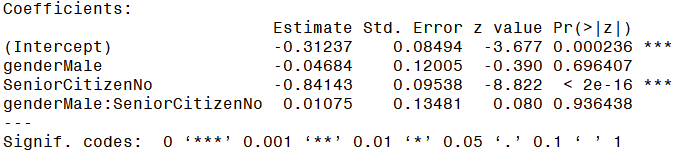
* Internet Service (X) & Churn (Y) & Contract (Z)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 로그-선형 모형 | 독립유형 | (M)(df) | df | (M2|M1)(df) | df |  |
| (XYZ) |  | 0 |  |  |  |  |
| (XY, YZ, XZ) | 부분연관 | 6.0939 | 4 |  |  | 9.488 |
| (XZ, XY) | 조건부 독립 | 1021.8 | 6 | 1015.7061 | 2 | 5.991 |
| (XY, YZ) | 236.99 | 8 | 230.8961 | 4 | 9.488 |
| (XZ, YZ) | 420.07 | 6 | 413.9761 | 2 | 5.991 |
| (X,YZ) | 결합독립 | 1016 | 10 | 1009.9061 | 6 | 12.59 |
| (Y,XZ) | 1800.9 | 8 | 1794.8061 | 4 | 9.488 |
| (Z,YZ) | 1617.8 | 10 | 1611.7061 | 6 | 12.59 |
| (X,Y,Z) | 상호독립 | 2396.9 | 12 | 2390.8061 | 8 | 15.507 |

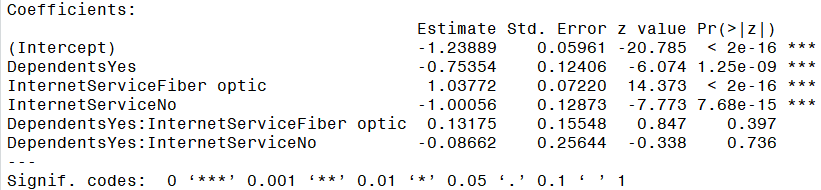
3-way Interaction term이 있는 Saturated model과 부분연관 모형을 비교했을 때 (H0: Reduced model is better.) 귀무가설을 기각하지 않는 것을 생각했을 때, 3-way interaction은 유의수준 0.05 하에서 유의하지 않다고 생각할 수 있다. 위의 두 변수로 Churn을 예측하는 로짓 모형을 적합 시켰을 때 두 변수의 interaction은 유의하지 않다고 할 수 있다.

* Senior Citizen(X) & Churn (Y) & Gender (Z)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 로그-선형 모형 | 독립유형 | (M)(df) | df | (M2|M1)(df) | df |  |
| (XYZ) |  | 0 |  |  |  |  |
| (XY, YZ, XZ) | 부분연관 | 0.006 | 1 |  |  | - |
| (XZ, XY) | 조건부 독립 | 0.499 | 2 | 0.493 | 1 | 3.84 |
| (XY, YZ) | 0.0084 | 2 | 0.0024 | 1 | 3.84 |
| (XZ, YZ) | 148.12 | 2 | 148.114 | 1 | 3.84 |
| (X,YZ) | 결합독립 | 148.14 | 3 | 148.134 | 2 | 5.991 |
| (Y,XZ) | 148.63 | 3 | 148.624 | 2 | 5.991 |
| (Z,XY) | 0.5218 | 3 | 0.5158 | 2 | 5.991 |
| (X,Y,Z) | 상호독립 | 148.65 | 4 | 148.644 | 3 | 7.815 |

가장 간단한 모형을 선택하는 게 가장 좋으므로, 노란색 음영으로 표시된 모형이 가장 좋다고 할 수 있다. 이를 기반으로 두 변수로 Churn을 예측하는 로짓 모형을 세워보았다.

로그 선형모형에서 본 것과 같이, 다른 두 가지 변수들로 Churn을 예측한 결과 Z인 gender 변수가 Churn과 관련이 없고, Interaction term도 없는 것을 확인할 수 있었다.

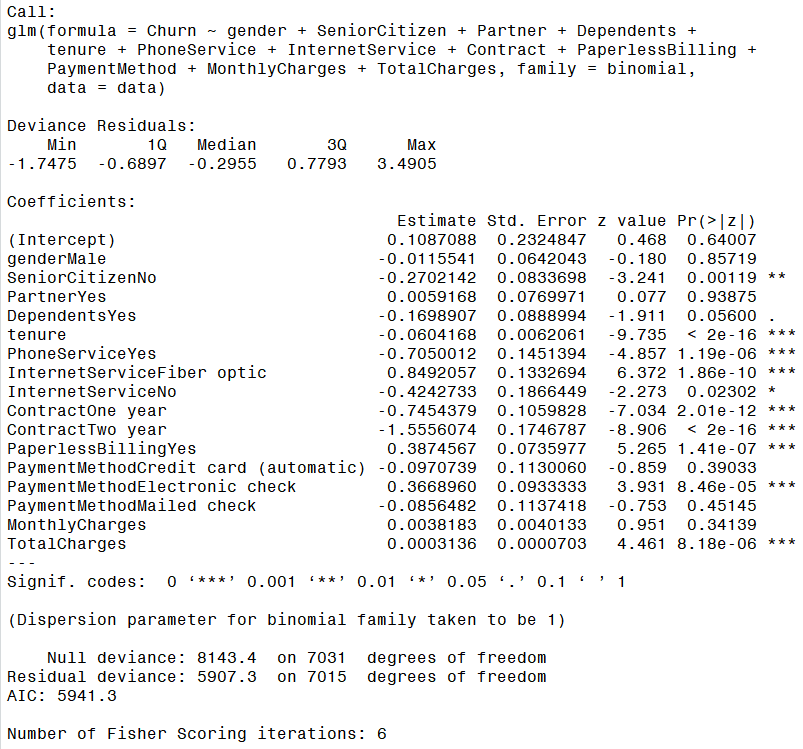
* Internet Service (X) & Churn (Y) & Dependents (Z)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 로그-선형 모형 | 독립유형 | (M)(df) | df | (M2|M1) | df |  |
| (XYZ) |  | 0 |  |  |  |  |
| (XY, YZ, XZ) | 부분연관 | 1.2477 | 2 | - |  | - |
| (XZ, XY) | 조건부 독립 | 102.13 | 3 | 100.8823 | 1 | 3.84 |
| (XY, YZ) | 125.7 | 4 | 124.4523 | 2 | 5.991 |
| (XZ, YZ) | 680.49 | 4 | 679.2423 | 2 | 5.991 |
| (X,YZ) | 결합독립 | 904.76 | 6 | 903.5123 | 4 | 9.488 |
| (Y,XZ) | 881.19 | 5 | 879.9423 | 3 | 7.815 |
| (Z,YZ) | 326.4 | 5 | 325.1523 | 3 | 7.815 |
| (X,Y,Z) | 상호독립 | 1105.5 | 7 | 1104.2523 | 5 | 11.071 |

이 역시도, 자녀 여부와 인터넷 서비스와 이탈에는 연관성이 없다는 것을 확인할 수 있다.

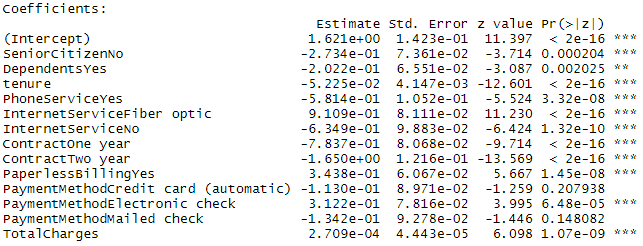
1. 로짓 모형 적합

위의 분석을 기반으로, Churn을 예측하는 로짓모형을 적합시켰다. 먼저 모든 변수를 포함한 모형을 적합시켰다.



Gender와 partner에서 유의미하지 않은 결과를 얻었으며, 결제방식에서 일부 유의미하지 않은 결과를 얻었다. 이를 해결하기 위해 Stepwise 방법을 이용하여 변수를 선택하였다. 그 결과 gender, partner, Monthly Charges가 제거되었다.

이때 가장 큰 coefficient를 가진 변수들은 위의 음영으로 칠해진 변수들과 같다. 이탈하는 것에 비해 이탈 안 할 오즈가 계약을 월 단위로 하는 것에 비해 2년 계약을 할 때 exp(-1.65)=0.192배 낮다. 이탈하는 것에 비해 이탈 안 할 오즈가 인터넷을 DSL로 쓰는 것에 비해 Fiber optic을 이용할 때 exp(-0.9109)=0.4배 낮다. 다만 이 모형에서 아래의 표를 얻을 수 있었는데, 아래의 표를 보아서는 좋은 모델이라고 할 수 없다.

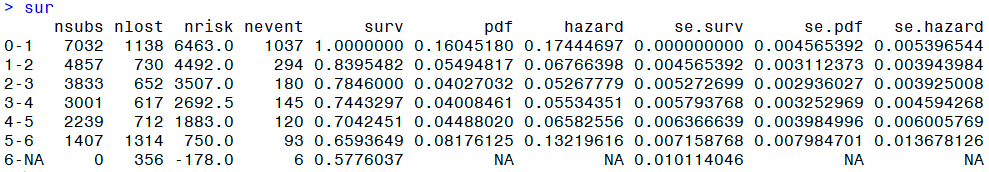


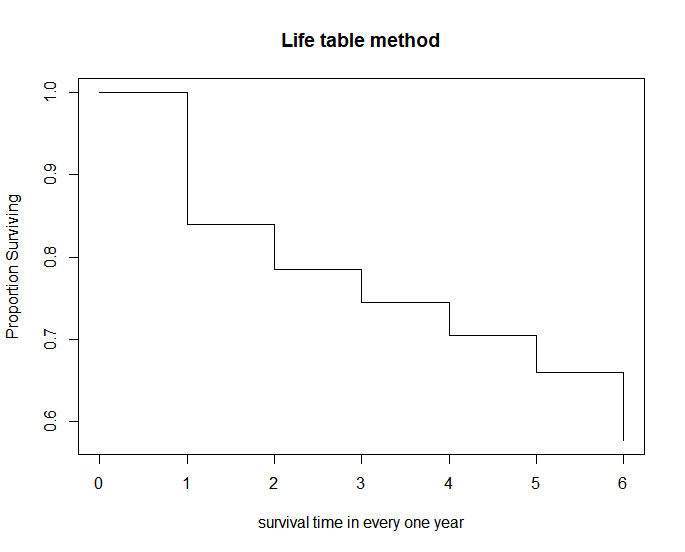
|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Precision | F1\_score | Recall |
| 0.77 | 0.6 | 0.69 | 0.76 |

위의 표를 보아서는 좋은 모델이라고 할 수 없다.

1. **생존분석**
2. 생명표 방법

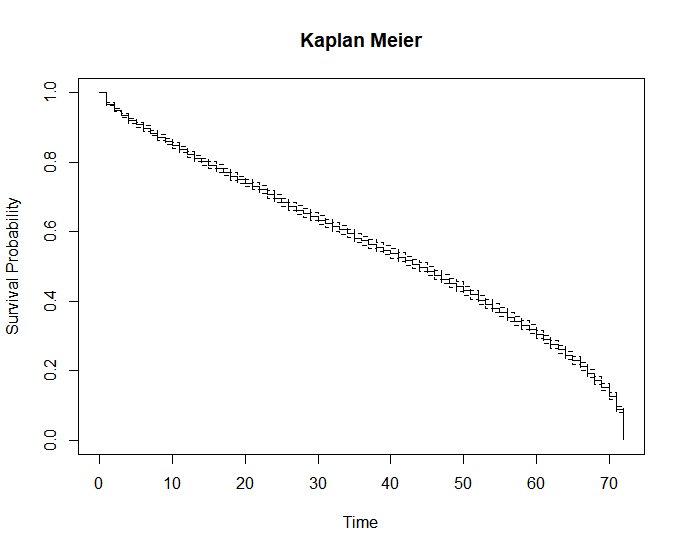
총 관측 기간이 72개월이기에, 이를 12개월씩 해서 총 6개의 구간으로 나누었다.



각 구간에서 위와 같은 생존율을 확인할 수 있었다.

다만 각 구간에서 censor된 데이터 많기 때문에 이를 보완하기 위해 카플란 마이어 분석또한 진행하였다.

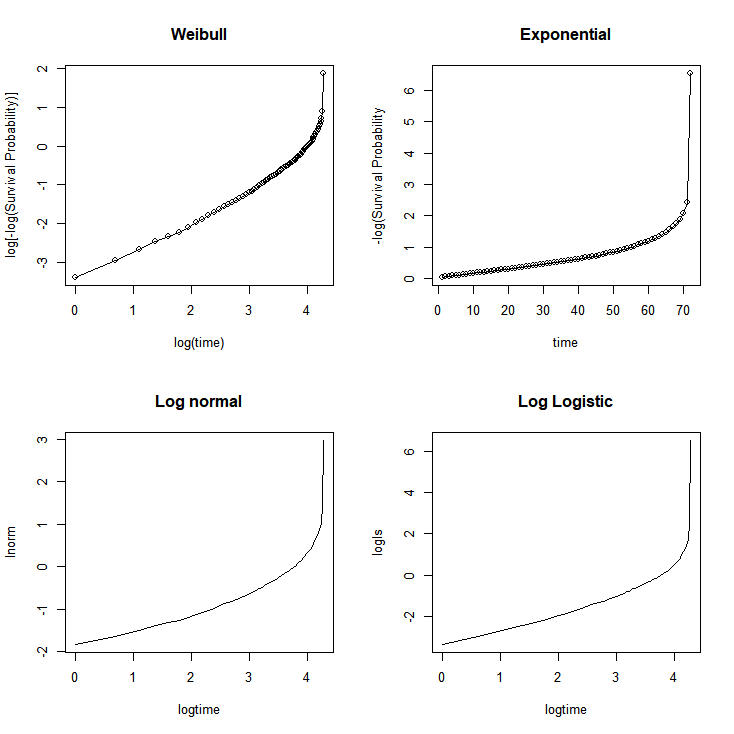
|  |
| --- |
| Time n.risk n.event survival std.err  1 7032 233 0.96687 0.002134  2 6419 115 0.94954 0.002637  3 6181 106 0.93326 0.003030  4 5981 93 0.91875 0.003335  5 5805 69 0.90783 0.003545  6 5672 70 0.89662 0.003746  7 5562 80 0.88373 0.003960  ...  67 944 88 0.19223 0.005667  68 846 91 0.17155 0.005456  69 746 87 0.15155 0.005225  70 651 108 0.12641 0.004886  71 532 164 0.08744 0.004222  72 362 356 0.00145 0.000591 |

1. 누적한계추정법 – 카플란 마이어

위와 같은 결과를 얻었다.

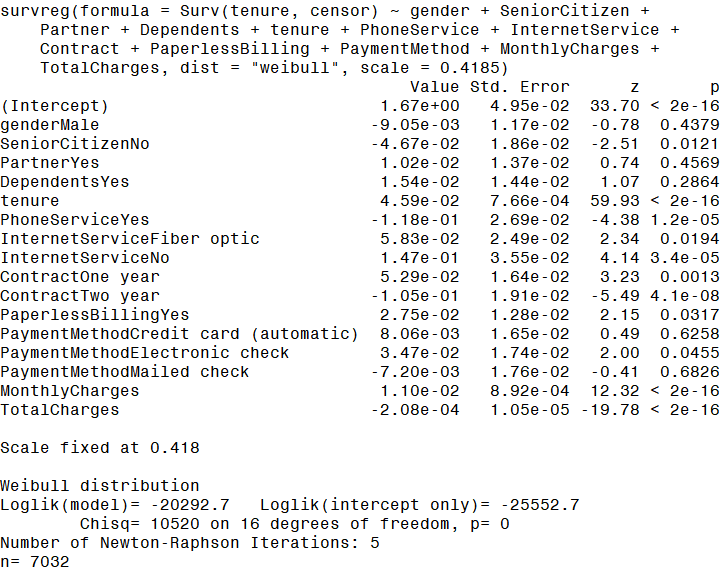
1. Log-rank test

각 변수로 처리로 두고, 처리에 따른 Churn의 생존함수를 비교해보고자 하였다. Log-rank test를 하기 위해서는 각 그래프가 겹치지 않는 형태여야 하기 때문에 먼저 변수별로 그래프를 그려보았다. 이때 Gender와 Phone Service의 경우 겹치는 형태를 가지고 있고, Payment Method도 두개의 범주에서 겹치는 형태를 보이므로 이 세 변수를 빼고 log-rank test를 진행하였다(H0: 생존함수가 같다 H1: not H0). 그 결과, 모든 변수에서 귀무가설을 기각하여 처리 별(변수 별) 생존함수가 모두 다르다고 결론지을 수 있다.

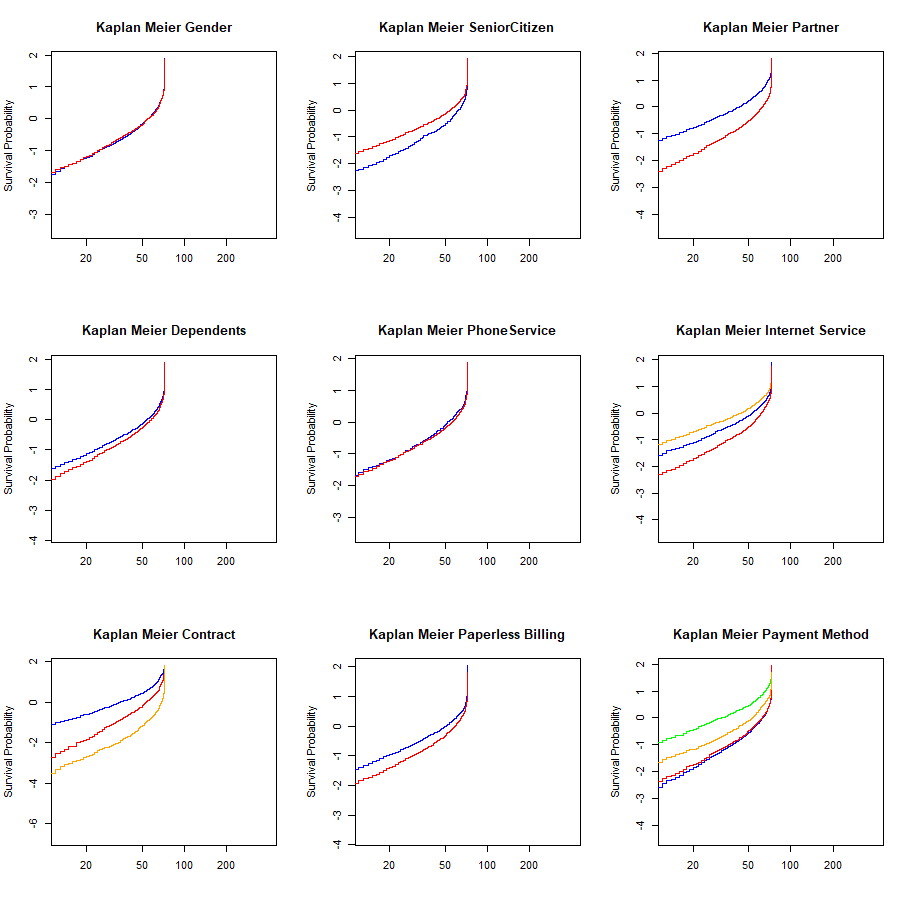
4) 모수적 모델

통신사 데이터가 어떤 분포를 따르는지 확인하기 위해 그래프를 그려보았다. 어떤 그래프도 직선의 형태를 보이지 않기에 분포를 가정하기 어렵다고 판단하였다. 다만 와이블 분포에 가장 가까워 보이므로 와이블 분포라고 가정하고 분석을 진행한다.

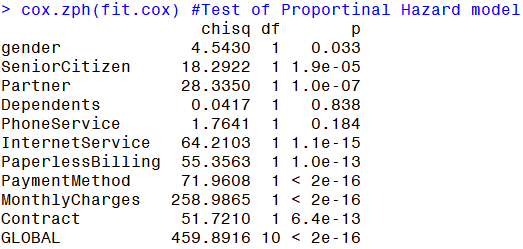
음영 처리된 변수들이 유의하게 나온 변수들이라고 할 수 있다.

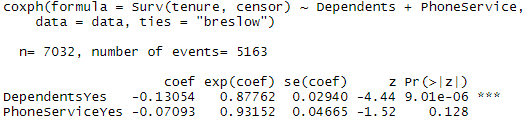


1. Cox 비례 위험 모델

분석에 앞서 주어진 자료가 비례 위험 모형의 가정을 따르는 지 확인해야 한다.

각 그래프가 겹치지 않는 형태여야 하는데, Gender와 Phone Service의 경우 겹치는 형태를 가지고 있고, Payment Method도 두개의 범주에서 겹치는 형태를 보이므로 이 세 변수에서는 비례위험모형가정을 만족하지 않는다고 판단하였다. 따라서 이 세 변수를 빼고 진행하였다. 다만 비례위험 모형가정을 검정해보았을 때에는 비례위험모형가정을 만족한다는 귀무가설을 기각하지 않는 변수는 Dependents와 Phone Service 뿐이었다.



비례위험모형가정을 만족하는 두 변수로만 Cox 비례 위험 모형에 적합해 보았다. 자녀여부는 굉장히 유의하게 나타났으며 Phone Service는 유의하지는 않았지만 stepwise 과정에서 탈락되지 않고 남아있었다.

1. **Support Vector Machine**

고객의 이탈을 예측하는 것이 목적이기에, Support Vector Machine을 통해서 분류 모형을 적합하고자 한다. 먼저, 아무런 튜닝 없이 가장 기본적인 모델을 적합 시켜 보았다.

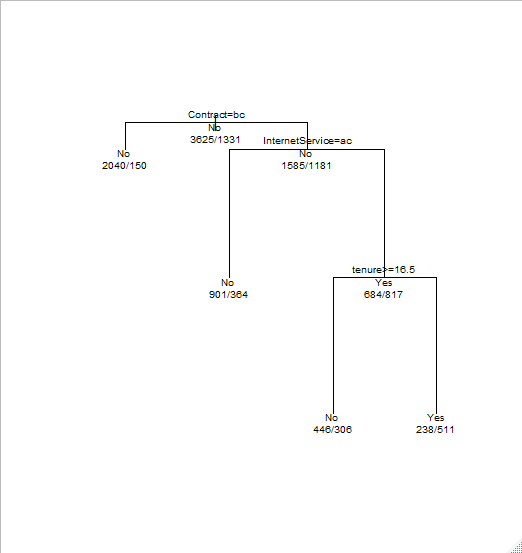
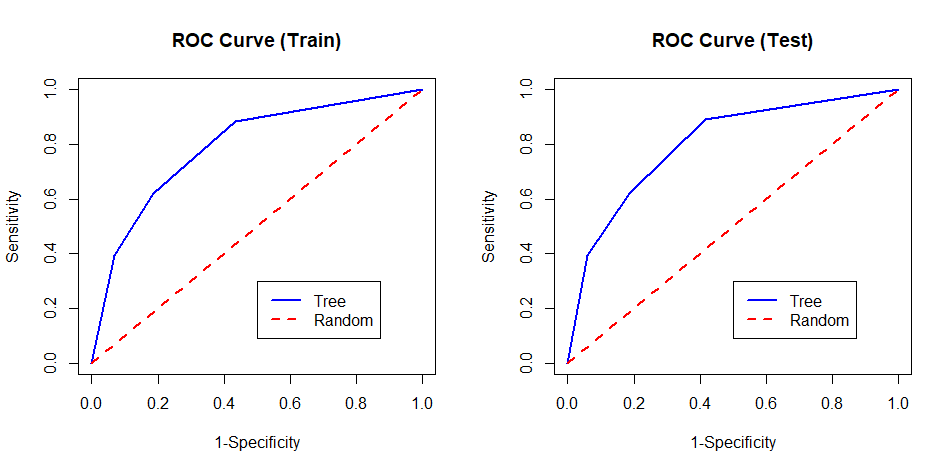
|  |  |  |  |
| --- | --- | --- | --- |
|  | Default | Tuned | Tuned with Class weight |
| C | 1 | 0.1 | 0.1 |
| Gamma | 1 | 10 | 1 |
| Accuracy | 0.79 | 0.81 | 0.78 |
| Precision | 0.87 | 0.90 | 0.82 |
| F1-score | 0.86 | 0.88 | 0.85 |
| Recall | 0.85 | 0.85 | 0.88 |

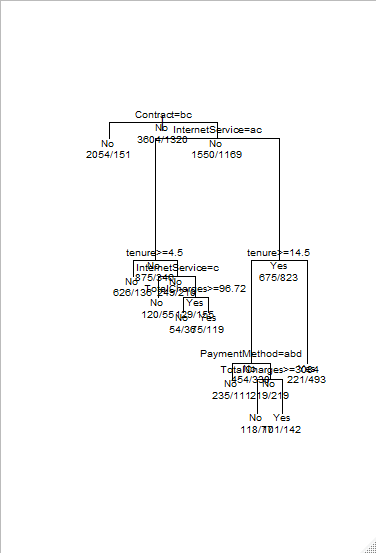
초모수를 튜닝하고, 데이터 불균형을 해결하고도 좋지 않은 결과를 보였다. 서포트 벡터 머신은 어떤 선형적인 선으로 데이터를 분류하는 모델임을 고려했을 때, 앞선 EDA에서 본 것과 같이 선형적으로 분리하기 어렵다고 생각할 수 있다.

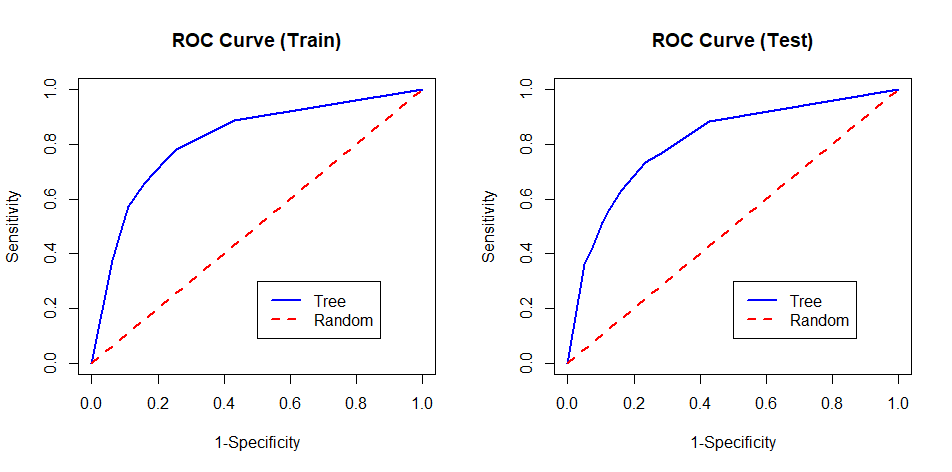
1. **Tree model**

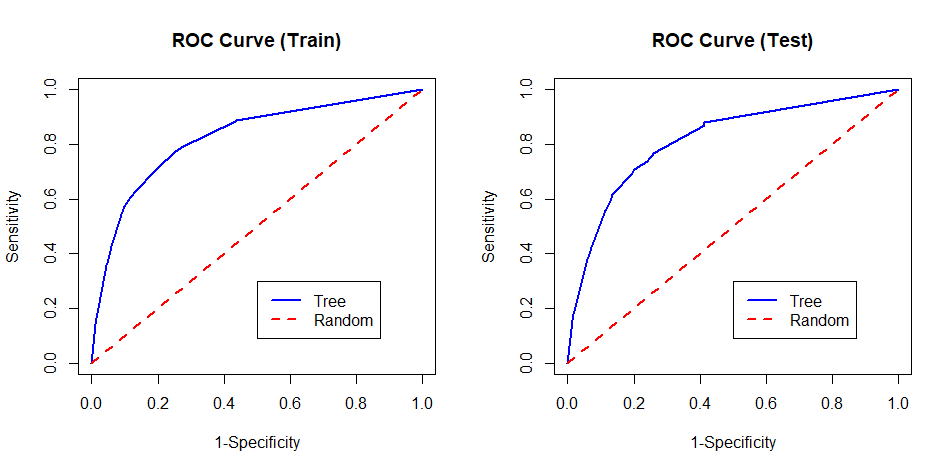
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Default | Pruned | Pruned with Class weight | Bagging | Random Forest |
| Accuracy | 0.80 | 0.80 | 0.80 | 0.90 | 0.88 |
| Precision | 0.93 | 0.90 | 0.89 | 0.79 | 0.73 |
| F1-score | 0.87 | 0.87 | 0.87 | 0.87 | 0.84 |
| Recall | 0.82 | 0.84 | 0.85 | 0.98 | 0.98 |

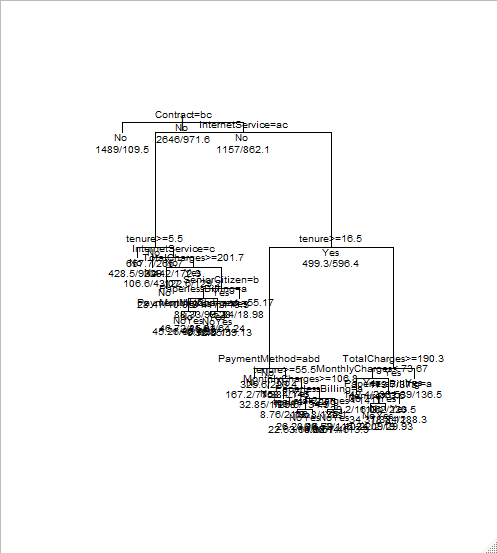
* 1st Model with default setting



* 2nd Model with pruning

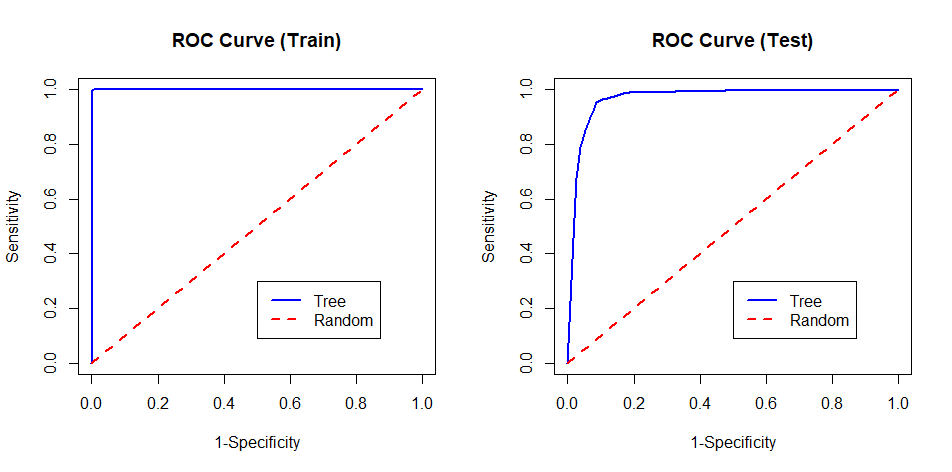


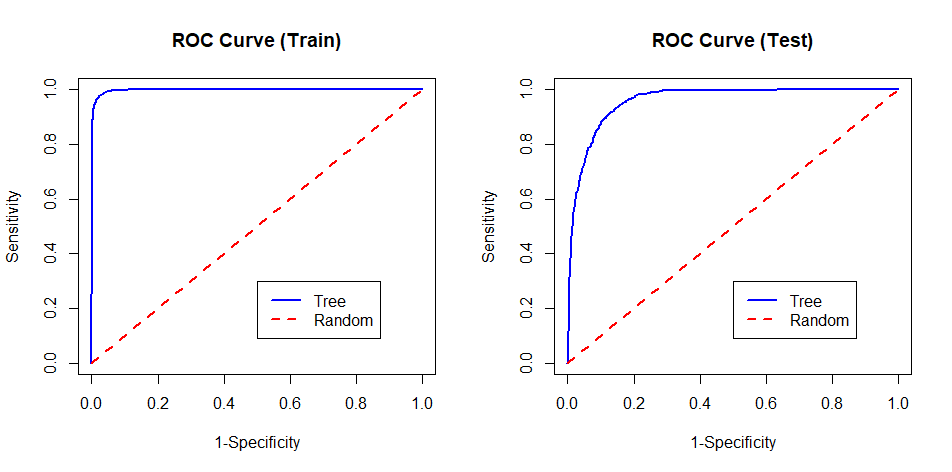
* 3rd Model with pruning and oversampling



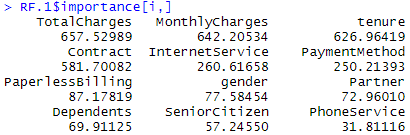
세 번째 모델까지는 rpart를 이용한 의사결정 나무의 모델이다. 세 모델 모두 크게 좋은 결과를 주지는 못했다. 또한, 세 모델 모두 Contract, Internet Service, tenure 순으로 가지치기를 했으며, 이를 통해 모두 다른 의사결정 나무 임에도 불구, 계속해서 같은 변수가 뽑힌다는 것을 확인할 수 있다. 이때 반복해서 뽑힌 dominant variable이 EDA에서 확인했던 변수들임을 알 수 있었다. 세 번째 모델에서는 불균형한 데이터를 oversampling을 통해 해결하고자 하였으나, 적합이 크게 잘되었다고 볼 수는 없다. 오히려 앞선 모델과 차이가 크지 않았다.

* 4th model with Bagging

ROC커브를 보았을 때에도 앞선 의사결정 나무들 보다는 훨씬 좋은 모델을 보여주고 있으며, 의사결정나무 뿐만 아니라 로지스틱과 SVM을 포함해서도 가장 좋은 예측력을 보였다.

* 5th model with Random Forest

랜덤 포레스트 모델의 경우 다른 의사결정 나무 모형보다는 좋은 예측력을 보였지만, 배깅 모델에 비해서는 약간 떨어지는 예측력을 보였다.



랜덤 포레스트에서 불순도를 낮추는데 중요한 역할을 한 변수를 순서대로 보았을 때, 다른 나무들과 달리 Total Charges와 Monthly Charges가 큰 역할을 차지하고, 이용기간과 계약, 인터넷 서비스가 뒤를 이었다.

1. **로버스트 회귀**

* Monthly Charges

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Ls | M | Lms | Lts | Quantile 0.5 | Quantile 0.25, 0.75 |
| MSE | 56.13099 | 56.74966 | 100.5942 | 2698197020 | 58.78896 | 89.10936 |

로버스트 회귀와 분위 회귀를 Monthly Charges에 이용해 보았으나, 일반 회귀 성능에 미치지 못했다.

1. **결론**

먼저 고객의 이탈을 예측하는 데에 있어서는 다른 모델보다 앙상블 모델이 좋은 예측력을 보였다. 그 중에서도 배깅 모델이 가장 좋은 예측력을 보였다. EDA에서 알아본 Churn에 따른 차이가 컸던 변수들이 실제로도 고객이탈과 독립이 아님을 확인할 수 있었으며, 의사결정나무 모델에서도 가장 지배적인 변수로 이용되었다. 다만 랜덤 포레스트에서는 전체 기간동안 납부한 요금과 한달에 납부한 요금이 가장 중요한 변수로 사용되었다. 이는 랜덤포레스트가 각 노드에서 사용할 변수를 무작위로 뽑아 사용하는 특성을 반영하여 더 다양한 모형을 만들 수 있어 찾아볼 수 있었던 장점이다. 또한, EDA에서 연속형 변수로 산점도를 그려보았을 때 어떠한 선형적인 선으로 분류하기 어려워 보이던 것과 더불어, 위에 그린 막대그래프 이외의 다른 변수에서는 고객 이탈에 따른 차이가 없었던 점을 고려하면 선형적인 선으로 분류해 내는 SVM이 좋지 않은 결과를 보여준 것을 납득하기 어렵지 않다.