생명보험 해지 확률 결과

a) 위 Part 1,2) 에서 구한 최적 예측모형을 이용하여 추정용 자료(estimation data) ;1-5000)에서 예측변수들을 이용하여 각 고객별로 i) Logistic/ GLM 해지확률 , ii) GAM 해지확률 , iii) Cox PHM 해지확률, iv) Logistic GLM / GAM 해지확률 및 Cox PHM score가 상위 10% 에 해당하는 500명 고객들의 지시변수(indicator variables)를 구하시오.

customer	Logistic / GLM	GAM	Cox PHM
1	1	1	0
2	1	1	0
3	1	1	0
…			
499	1	0	0
500	0	0	0

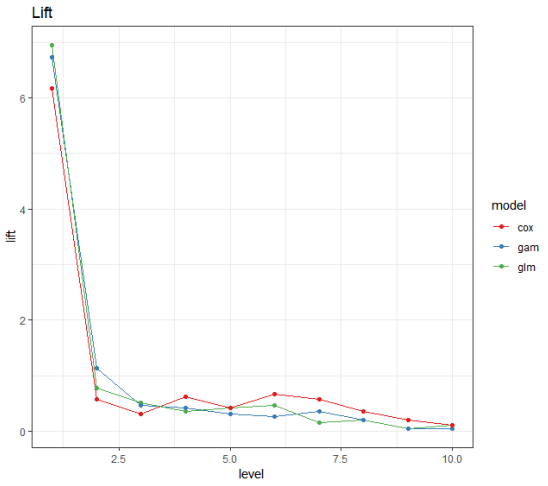
b) 위에서 구한 GLM/ GAM/ CoxPHM 확률 및 점수값의 크기순으로 전체 5000명의 고객을 10개의 구간으로 나눈 후 분석용 자료를 이용하여 각 구간별 실제 해지고객의 백분율 및 Lift 값을 구하여 세 방법을 비교하는 표를 만드시오.

group	Logistic / GLM	GAM	Cox PHM
1	6.9430	6.7358	6.1658
2	0.7772	1.1399	0.5700
3	0.5181	0.4663	0.3109
4	0.3627	0.4145	0.6218
5	0.4145	0.3109	0.4145
6	0.4663	0.2591	0.6736
7	0.1554	0.3627	0.5699
8	0.2073	0.2073	0.3627
9	0.0518	0.0518	0.2073
10	0.1036	0.0518	0.1036

세 모형 별 Lift Chart를 보면 group (등급) 이 낮아질수록 Lift값이 감소하는 것을 볼 수 있어 예측력이 적절한 것을 알 수 있다. 또한, Lift 값이 1등급 (상위10%)에서 세 모형 모두 6.5 근처로 매우 높은 것을 알 수 있다. 두 모형을 비교해 보면, 1등급(상위10%)에서 GLM 모형 (전체 이탈 고객 중 약 70%정도 사전예측)이 GAM모형

(전체 이탈 고객 중 약 67%정도 사전예측), Cox PHM 모형 (전체 이탈 고객 중 약 60%정도 사전예측)보다 예측력이 뛰어나며, 등급이 낮아질수록 Lift값이 안정적으로 감소하는 것을 볼 수 있다. 그러므로 세 모형 중 GLM 모형의 예측력이 가장 뛰어남을 알 수 있다. 하지만, 세 모형 모두 Lift값이 6.5 근처이므로 모두 예측력이 상당히 높은 것으로 판단된다.

b) 위에서 구한 표를 그래프로 그린 Lift Chart를 서로 겹쳐서 그려보고 각 방법의 장단점을 서로 비교 검토하시오.



보험 해약을 우리의 event로 보고 이 event가 발생하기까지의 시간을 하나의 변수로 생각한다면, 이를 censored된 데이터로 생각해 볼 수 있다. 이것만 따진다면, GLM이나 GAM보다는 생존 분석을 이용한 COX 모형이 더 예측 모형에 적합하다고 예상을 했다. 그러나 Lift 값이 GLM이 가장 높았으며, 그 다음 GAM, COX 순으로 측정되었다.

GLM은 종속변수가 이진 0 , 1로 이루어진 로지스틱 회귀분석으로, 독립변수들에 따라 사건이 발생하는 것과 발생하지 않는 것에 대한 확률을 구한다. 종속변수가 단순히 0과 1 로 표현된다는 점에서 보험 계약의 기간이 반영되는 것에 부정확한 결과를 낳을 수 있다. 이것은 새로운 적절한 변수를 넣거나 변환하여 보완될 수 있다. GLM은 다른 두 모형에 비해, 변수에 대한 변화를 주지 않기 때문에 계산 속도가 빠르다. 따라서, 예측력에 도움이 되는 변수만 추가된다면, 가장 효율적인 모형이 될 것이라 생각된다.

GAM은 독립변수 중 비선형인 독립변수가 존재할 경우, 해당 변수들에 대해 적절한 평활함수를 이용하여 종속변수를 예측함으로써 비선형 독립변수의 특성까지 고려하는 통계적 분석방법이다. 따라서, 비선형인 독립변수가 존재하지 않을 경우에는 평활함수를 사용하지 않아, 로지스틱 분석의 결과와 같게 된다. GAM은 적절한 평활함수를 찾지 못하거나 비선형인 독립변수들이 많이 존재할 경우, 계산속도가 느리다는 단점이 있다. 하지만, 비선형인 각 독립변수들에 대한 함수를 정해, 예측함으로써 각각의 독립변수를 고려할 수 있고, 그에 따라 정확도가 향상될 수 있다.

Cox-PHM은 사건의 발생 여부에 따라 중도 절단된 자료를 이용하여 사건발생과 사건 발생 전을 나누어 사건에 대한 확률을 계산한다. Cox-PHM의 경우, 생존시간의 분포에 대한 가정이 필요 없고, 생존회귀계수 베타가 일반 회귀분석에서의 회귀계수와 유사하다. 또한, 위험함수, h(t)에 관해서는 비모수적인 분석이지만, 베타라는 모수를 추정하기 때문에 semiparametric 모형이라고 부른다. 따라서, 생명보험 해약에 따른 고객이탈을 예측하는데 있어 시간을 고려하기 때문에 가장 적절한 모형이라 생각된다. 하지만, Cox-PHM을 이용할 경우, 시작점에 대한 명확한 정의와 시간 측정이 가능해야 하고, 사건의 발생 여부를 정확히 구별할 수 있어야 한다. 또한, 시간에 의존하는 자료를 분석하기 위해 우선 위험함수의 지수파트를 구한 뒤 baseline hazard를 다시 구해야 하기 때문에 복잡하다는 단점이 있다.

c) Part 1,2) 에서 최종 선택한 GLM/ GAM/ Cox PHM 예측모형의 AIC값을 각각 제출하시오.

	Logistic / GLM	GAM	Cox PHM
AIC	1157.243	1187.134	1454.91