가톨릭대학교
THE CATHOLIC UNIVERSITY OF KOREA

데이터 마이닝

팀프로젝트 결과보고서

가톨릭대학교
THE CATHOLIC UNIVERSITY OF KOREA

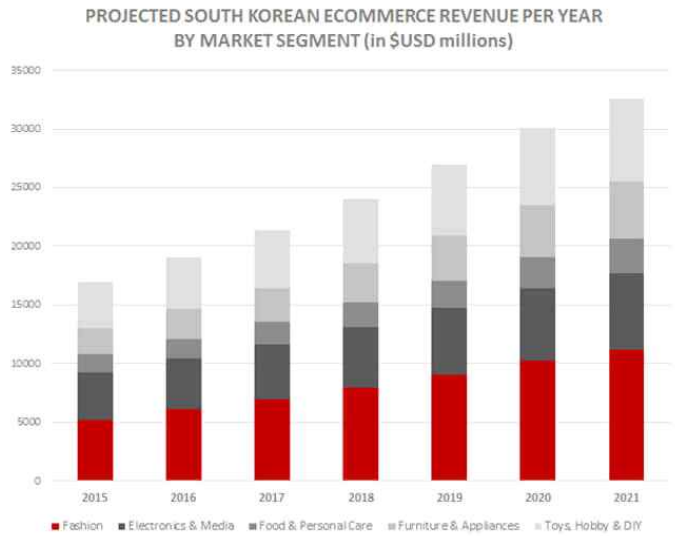
과목명	데이터 마이닝
담당교수	노상욱 교수님
학과	컴퓨터정보공학과
팀 명	E-commerce
팀장 학번 & 이름	201521842 정광진
팀원 학번 & 이름	201420701 유준호
제출일자	2018.12.06.(목)

목 차

Ⅰ. 서론

1. 보고서 개요

소비자의 이탈을 각 종 알고리즘을 이용해 예측함으로써 소비자 별 맞춤대응 전략 및 소비자 이탈 요인 파악을 통해 기업이 경쟁 우위를 얻을 수 있도록 하는 것을 목표로 진행합니다.



Amazon은 미국의 1위 E-commerce 업체이며 18년 1분위 매출은 501억 달러 수준으로 미국의 소비패턴이 점차적으로 E-commerce로 이동하고 있다는 것을 뜻합니다.

이러한 현상에 부응하듯 2017년 Eshopworld에서 발표한 한국의 e-commerce 수익률은 매년 증가하여 2021년도에는 무려 50%이상 증가할 것으로 추측하고 있습니다. 현재 한국의 E-commerce업계는 압도적인 선두주자가 부재하며 경쟁이 매우 심각한 상태입니다. 이 때, 고객의 거래정보를 이용해 고객의 이탈을 예측을 통하여 고객 별로 쿠폰이나 광고를 선별적으로 보여줌으로써 고객의 관리 측면에서 업체는 우위를 취할 수 있습니다. 또한 광고비의 효율적 운용을 통한 비용 감소의 효과를 얻을 것으로 기대됩니다.

2. 작업 환경

(1) 데이터 전처리
- Python (pandas library)
- MySQL
- R

(2) 알고리즘 적용
- Python(scikit-learn library)
- R

(3) 사용 툴
- R studio
- MySQL workbench
- google colaboratory

(4) 시각화 툴
- MS powerBI

(5) 프로그램 데모 제작
- python(pyqt5)

Ⅱ. 본론

1. 데이터 파악

(1) Concept

Kaggle에서 영국의 소매상에서 전자상거래를 통해 거래된 데이터를 통해 데이터 분석을 진행하였습니다.
이를 통해서 clustering을 통해 고객들을 분류하고 이에 따라 이탈한 고객과 비이탈 고객을 나누어서 이탈고객을 1, 비이탈 고객을 0으로 정의합니다.

(2) Instances

총 541910개의 거래내역 데이터로 구성되어 있으며 데이터는 아래와 같이 구성되어 있습니다.

CustomerI	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerI	Country
768535	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	768535	United Kingdom
768535	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	768535	United Kingdom
768535	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	768535	United Kingdom
768535	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	768535	United Kingdom
768535	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	768535	United Kingdom
768535	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/1/2010 8:26	7.65	768535	United Kingdom
768535	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12/1/2010 8:26	4.25	768535	United Kingdom
768535	536366	22633	HAND WARMER UNION JACK	6	12/1/2010 8:28	1.85	768535	United Kingdom
768535	536366	22632	HAND WARMER RED POLKA DOT	6	12/1/2010 8:28	1.85	768535	United Kingdom

(3) Attribute

속성 이름	설명	특징
InvoiceNo	송장 고유 번호	Numeric
StockCode	물품 고유 코드	Nominal
Description	물품에 대한 설명	Nominal
Quantity	물품 구매 갯수	Numeric
InvoiceDate	물품 구매 날짜와 시간	Numeric
UnitPrice	하나의 물품 가격	Numeric
CustomerID	고객 고유 코드	Numeric
Country	구매 지역	Nominal

총 3개의 Nominal, 5개의 Numeric 변수로 구성되어있습니다.

(4) Class

시계열(time-series)성을 입히기 위해서 이탈을 다음과 같이 정의하였습니다.
최근성을 나타내기 위해서 최근 21일 동안 구매내역이 없는 고객을 이탈고객으로 정의했습니다. 현재 일자를 기준으로 최근 21일 간 구매내역이 있는 고객과, 최근 21일 이전에 구매내역이 있는 고객과 비교를 통해 이탈자, 비이탈자를 나타내는 Label(Class)값을 만들었습니다. 그에 따라, 전체 데이터는 21일 이전의 구매를 한 고객들의 정보와 이탈여부를 표기한 값으로 구성하였습니다. (최근 21일 간의 구매내역은 이탈여부에 직접적으로 영향이 있기 때문에 전체 데이터에서 제외)

2. 데이터 전처리

(1) Feature Engineering

속성	속성 설명	속성 특징	속성 범위
CustomerID	고객 고유 코드	Nominal	9158명
Total_Quantity	총 구매량	Numeric	1~1,876
Total_Value	총 구매금액	Numeric	9~169,143
StDev_Quantity	구매 량의 표준편차(거래 별 기준)	Numeric	0~488
StDev_Value	구매금액 표준편차(거래 별 기준)	Numeric	0~30,255
AvgTimeDelta	방문 간격의 평균	Numeric	0~43
Recency	현재와 최근 구매의 날짜 차이	Numeric	0~97
TQUT	영수증 당 평균 구매 개수	Numeric	4~23,605
Label	이탈 여부	Numeric	9158명

기존 Raw Data에서 CustomerID를 기준으로 하여 속성들을 재구성 하였습니다.

① Total_Quantitiy : Sum(Quantity)
② Total_Value : Quantitiy * UnitPrice
③ StDev_Value : Std(Quantitiy * UnitPrice)
④ StDev_Quantity : Std(Quantitiy)
⑤ AvgTimeDelta : InvoiceData의 차의 평균
⑥ Recency : 가장 최근 InvoiceData(구매날자)와 현재와의 차이
⑦ TQUT : 고객 데이터를 invoiceNo로 구분하여 개수를 세고 구매 개수의 평균을 산출
⑧ Label : 최근 21일 동안 구매내역이 없는 고객을 이탈고객으로 정의

(2) Feature Engineering Result

CustomerI	Total_Quai	Total_Value	StDev_Qua	StDev_Value	AvgTimeDelta	Recency	Total_Value_per_Un	Label
2105345	9	773	0.353553	77.22127298	0	38	85.88888889	1
2085920	5	400	0	0	0	84	200	0
1976717	72	4939	1.967345	114.5187843	0.333333333	16	123.475	1
768535	52	161	1.474055	4.695451963	5	24	146.1666667	0
1895117	41	2222	0.452816	48.81861472	1.457142857	2	61.72222222	1
1587807	54	2123	2.515217	75.05876645	1.8	15	101.0952381	0
1853100	5	440	0	78.15689349	0	93	73.33333333	1
2158358	5	606	0	146.1427384	0	77	101	1
1950892	86	8945	0.589165	147.1889403	1.369230769	7	135.530303	0
204040	10	324	0.516398	30.4893424	3	71	46.28571429	0
1643824	36	4128	0.912421	219.8993724	3.137931034	4	137.6	0
303767	11	621	1.095445	125.3842095	10.8	19	103.5	0
478502	10	3686	0.46291	800.1060198	2.125	15	409.5555556	1
348508	64	6699	0.494527	101.6954994	1.566037736	10	124.0555556	1
1376944	2	18	0	0	0	66	9	0
1817515	19	549	1.407886	33.84391695	0	73	61	1
1994650	56	4446	0.499471	84.69039498	1.886363636	5	98.8	0
1892383	26	15018	0.732695	987.9299943	2.1	42	715.1428571	1

총 9190개의 unique한 CustomerID로 구분하여 데이터를 재정립하였습니다.

(3) Train Set, Test Set의 구분

이탈자 수	비이탈자 수
5,508 명	3,651 명

트레인 데이터			테스트 데이터		
크기	이탈자	비이탈자	크기	이탈자	비이탈자
6,300	3,780	2,520	2,700	1,620	1,080

약 6:4의 비율 6:4 & 7:3

이탈자와 비이탈자의 비중이 전체 데이터에서 6:4의 비중 (이탈자 5,508명 : 비이탈자 3,651명)을 가지고 있기 때문에 트레인 데이터와 테스트 데이터를 나눌 때에도 6:4의 비중으로 나눴으며, 7:3의 비율로 트레인 데이터와 테스트 데이터를 전체 데이터에서 나눴습니다.

(4) 알고리즘 설명

알고리즘 적용의 목표는 크게 두 가지로 나눌 수 있습니다. 하나는 속성들을 설명하기 위한 알고리즘의 사용이고 다른 하난 속성들을 사용해 높은 예측의 정확도를 도출해내는데 있습니다. 각 목표에 따라 알고리즘을 두 개의 목표 설정에 맞춰 나눴습니다.

1) 설명을 위한 알고리즘

속성과 이탈의 관계를 파악하기 위해서 OneR과 DecisionTree를 사용하였습니다.

① OneR

OneR은 가장 이탈에 가장 큰 영향을 주는 속성 하나를 찾는데 사용될 수 있습니다. 오직 분류의 정확도가 가장 높은 하나의 속성만 사용해 클래스값을 도출하는 OneR의 특성은 각 속성 중에 어떤 속성을 사용하면 가장 높은 분류 정확도를 얻을 수 있는지 도출해줍니다.

② DecisionTree

결정 트리 학습법(decision tree learning)은 어떤 항목에 대한 관측값과 목표값을 연결시켜주는 예측 모델입니다. 이는 통계학과 데이터 마이닝, 기계 학습에서 사용하는 예측 모델링 방법 중 하나이며, 트리 모델 중 목표 변수가 유한한 수의 값을 가지는 것을 분류 트리라고 합니다. 이 트리 구조에서 잎(리프 노드)은 클래스 라벨을 나타내고 가지는 클래스 라벨과 관련있는 특징들의 논리곱을 나타냅니다.

2) 예측에 대한 알고리즘

이탈 예측의 정확도를 측정하기 위해서 Naive Bayes와 Random Forest를 사용하였습니다.

① Random Forest

기계 학습에서의 Random Forest는 분류, 회귀 분석 등에 사용되는 앙상블 학습 방법의 일종으로, 훈련 과정에서 구성한 다수의 결정 트리로부터 부류(분류) 또는 평균 예측치(회귀 분석)를 출력함으로써 동작합니다.

② Naive Bayes

기계 학습 분야에서의 Naïve Bayes Classification는 특성들 사이의 독립을 가정하는 베이즈 정리를 적용한 확률 분류기의 일종입니다. 통계 및 컴퓨터 과학 문헌에서 , 나이브 베이즈는 단순 베이즈, 독립 베이즈를 포함한 다양한 이름으로 알려져 있으며, 저희가 프로젝트 진행한 것은 가우시안 나이브 베이즈를 이용했으며 계산식은 아래와 같습니다.

$$p(x=v|c)=\frac{1}{\sqrt{2\pi\sigma_c^2}}e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}}$$

3. 알고리즘의 적용 및 해석

(1) OneR

인스턴스의 수	방문 간격의 평균	총 구매금액	총 구매량	구매금액 표준편차	구매량의 표준편차
0~3,000	2	4	1	1	1
3,000~9,000	9	1	0	0	0

이탈에 가장 큰 영향을 주는 속성은 0 ~ 3,000대에서는 총 구매금액(Total_Quantity)가 가장 높았으며 3,000 ~ 9,000대 에서는 방문 간격의 평균(AvgTimeDelta)이 가장 높았음을 확인 할 수 있었습니다. 실제로도 방문 간격의 평균은 얼마나 자주 매장에 방문하는가에 대해 큰 상관관계를 갖고 있기 때문에, 실질적으로도 유의미한 속성임을 확인할 수 있습니다.

인스턴스 수	50	100	300	600	1200	1800	2400	3000	3600
정확도	0.600	0.500	**0.722**	0.600	0.675	0.604	0.585	0.597	0.642
민감도	0.667	0.667	0.815	**1.000**	0.796	0.704	0.708	0.724	0.748
특이도	0.500	0.250	**0.583**	0.000	0.493	0.454	0.399	0.406	0.481

인스턴스 수	4200	4800	5400	6000	6600	7200	7800	8400	9000
정확도	0.625	0.609	0.633	0.647	0.628	0.625	0.641	0.642	0.632
민감도	0.800	0.741	0.774	0.762	0.733	0.757	0.766	0.762	0.749
특이도	0.361	0.411	0.421	0.475	0.471	0.427	0.452	0.463	0.457

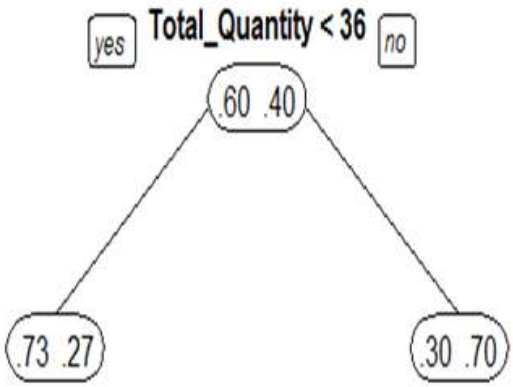
또한 가장 높은 정확도는 인스턴스의 수가 300개일 때 0.722의 정확도를 보임을 확인 할 수 있었습니다.

(2) Decision Tree

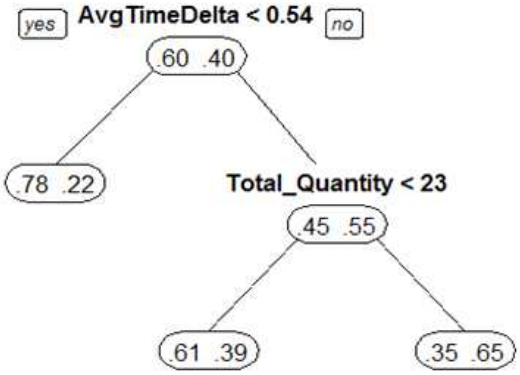
인스턴스 수	50	100	300	600	1200	1800	2400	3000	3600
정확도	0.667	0.400	**0.678**	0.628	0.600	0.569	0.607	0.588	0.632
민감도	0.667	0.389	0.759	**0.935**	0.630	0.627	0.678	0.644	0.687
특이도	0.571	0.313	0.606	**0.632**	0.500	0.462	0.509	0.485	0.540

인스턴스 수	4200	4800	5400	6000	6600	7200	7800	8400	9000
정확도	0.604	0.600	0.609	0.622	0.587	0.620	0.607	0.594	0.603
민감도	0.604	0.668	0.669	0.678	0.646	0.688	0.676	0.626	0.678
특이도	0.505	0.500	0.511	0.527	0.484	0.526	0.509	0.494	0.504

가장 높은 정확도는 인스턴스의 수가 300개일 때 0.678의 정확도를 보임을 확인 할 수 있었습니다.
어떠한 속성이 Tree를 구분하는데 중요한 역할을 하는지 펼쳐보게 된다면,



N = 300일 때



N >= 3000 일때

위와 같이 확인 할 수 있었고, 정확도가 가장 높은 N=300일 때 이탈에 가장 큰 영향을 주는 속성은 Total_Quantitiy(총 구매량) 임을 확인 할 수 있었으며, N이 3000보다 높을 경우 AvgTimeDelta(방문 간격의 평균)이 가장 큰 영향을 주는 것을 파악할 수 있었습니다.

(3) Naive Bayes

인스턴스 수	50	100	300	600	1200	1800	2400	3000	3600
정확도	0.667	**0.700**	0.678	0.667	0.650	0.633	0.613	0.657	0.573
민감도	0.778	0.833	0.889	**0.917**	0.917	0.904	0.484	0.917	0.418
특이도	0.600	0.667	0.684	0.700	0.667	0.613	0.510	0.681	0.418

인스턴스 수	4200	4800	5400	6000	6600	7200	7800	8400	9000
정확도	0.604	0.660	0.465	0.688	0.657	0.585	0.666	0.464	0.665
민감도	0.929	0.947	0.179	0.755	0.930	0.488	0.941	0.175	0.933
특이도	0.695	**0.743**	0.420	0.615	0.701	0.487	0.741	0.420	0.723

가장 높은 정확도는 인스턴스의 수가 100개 일 때 0.700의 정확도를 보임을 확인 할 수 있었습니다.

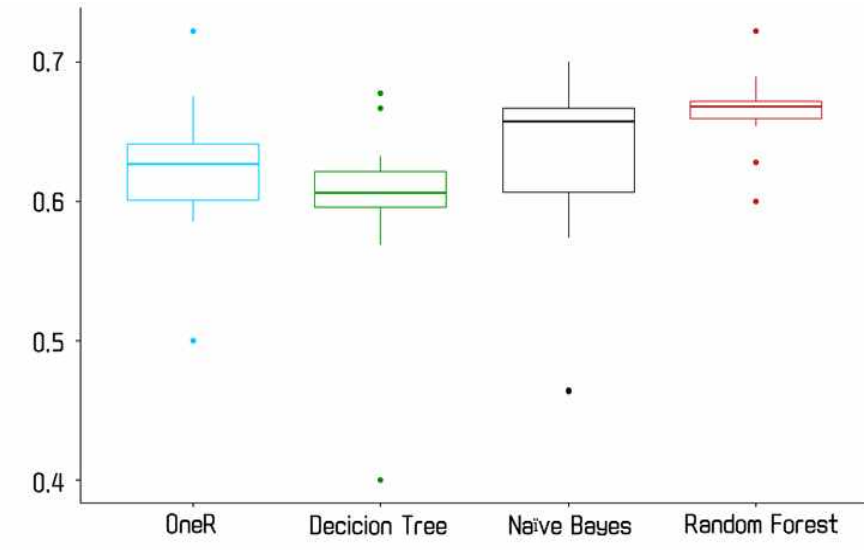
(4) Random Foreset

인스턴스 수	50	100	300	600	1200	1800	2400	3000	3600
정확도	0.600	0.667	**0.722**	0.628	0.689	0.656	0.669	0.681	0.660
민감도	0.778	0.833	0.907	**0.972**	0.801	0.784	0.817	0.794	0.819
특이도	0.500	0.625	**0.762**	0.727	0.636	0.588	0.620	0.624	0.609

인스턴스 수	4200	4800	5400	6000	6600	7200	7800	8400	9000
정확도	0.662	0.659	0.672	0.662	0.670	0.669	0.669	0.654	0.673
민감도	0.804	0.806	0.812	0.806	0.795	0.821	0.808	0.789	0.819
특이도	0.604	0.601	0.621	0.605	0.611	0.622	0.616	0.587	0.626

가장 높은 정확도는 인스턴스의 수가 300개일 때 0.722의 정확도를 보임을 확인 할 수 있었습니다.

(5) 알고리즘의 정확도 분포 시각화



Boxplot으로 나타낸 알고리즘의 정확도 분포

Boxplot을 통해 알고리즘 Accuracy를 종류별로 중앙값(mean), 상위 사 분위수(25% Quartile), 하위 사 분위수(75% Quartile), 최소 데이터 값(Minimum data value) 및 최대 데이터 값(Maximum data value)가 포함되며, 이상치를 눈으로 쉽게 확인 할 수 있습니다.

이를 통하여 Random Forest가 가장 정확도의 분포가 오밀조밀하게 뭉쳐있는 것을 볼 수 있으며 일정한 성능을 자랑한다고 볼 수 있습니다 이에 반해 Naive Bayes의 경우 분포가 광범위하게 분포되어있음을 볼 수 있습니다. 이것은 알고리즘의 성능의 신뢰도를 평가하는데 기여한다고 볼 수 있습니다.

즉, 위 시각화된 자료에 따라 Random Forest가 가장 신뢰도 높은 알고리즘임을 확인 할 수 있었습니다.

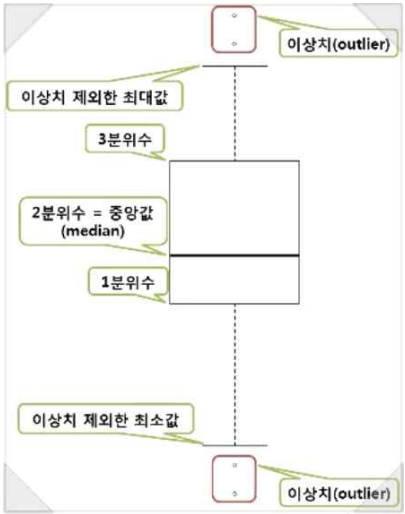
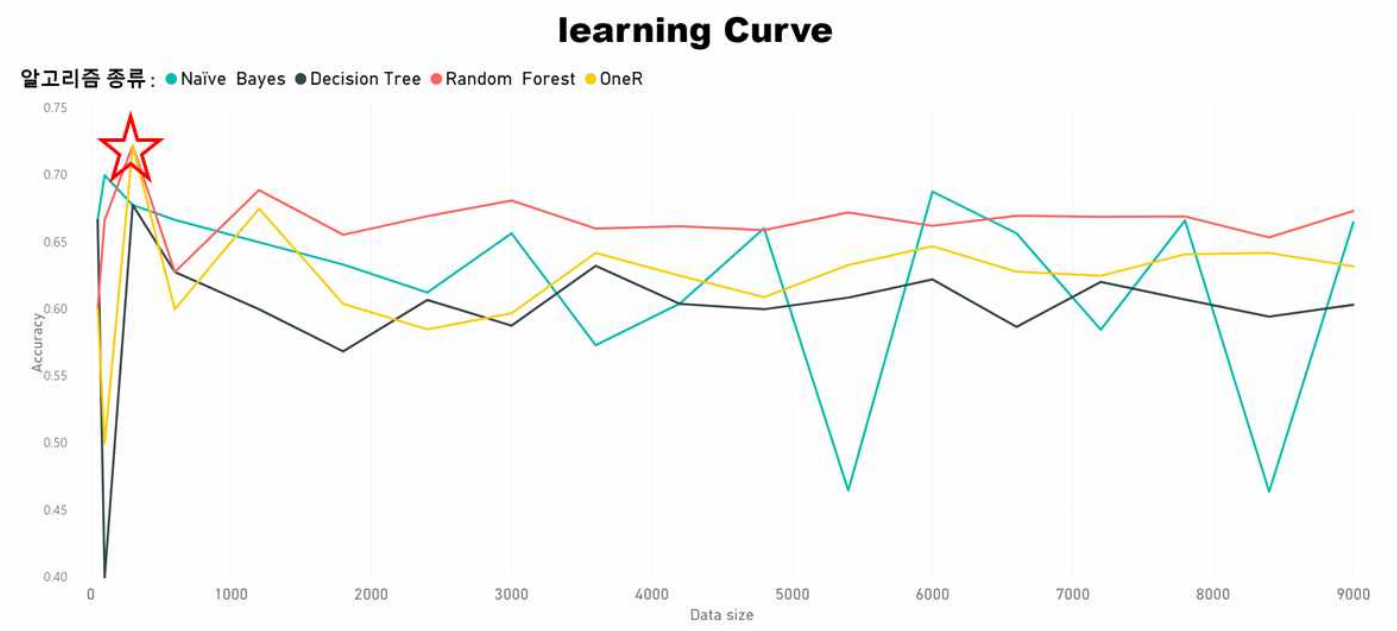


그림 20 boxplot 설명

(6) Learning Curve



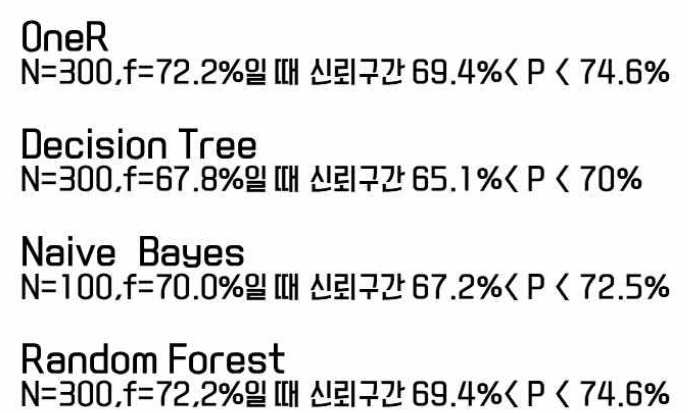
4가지 알고리즘을 Test Data에 따라 Accuracy를 그래프로 그려보았을 때의 결과 화면입니다. 300에서 가장 높은 정확도를 자랑하며 확대해서 보게 된다면,

위와 같이 노란색인 OneR과 Random Forest가 0.722로 가장 높은 정확도를 보임을 알 수 있었습니다.

4. 베르누이 분포

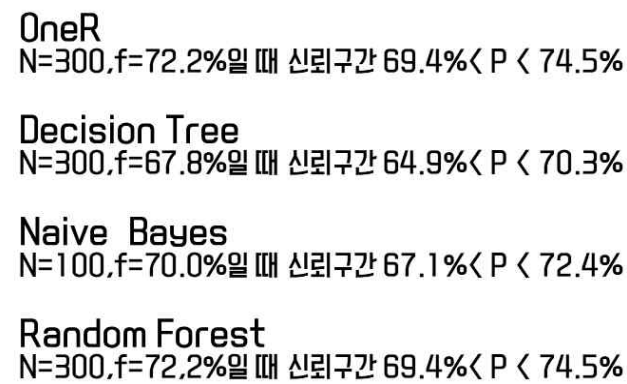
각 알고리즘에서 도출된 정확도가 통계적으로 신뢰도에 따라 어느 정도의 신뢰구간을 가지고 있는 지를 확인할 수 있습니다. 신뢰도가 높아짐에 따라 신뢰구간이 좁아짐을 확인할 수 있으며 각 알고리즘이 최고성능을 나타내는 N지점에서 신뢰도 80%, 신뢰도 90%, 신뢰도 99% 수준의 신뢰구간은 아래와 같습니다.

(1) 신뢰도 80%, z = 1.28



신뢰도가 80%일 때의 신뢰구간입니다.

(2) 신뢰도 90%, z = 1.64



신뢰도가 90%일 때의 신뢰구간입니다.

(3) 신뢰도 98.93%, z = 2.3

OneR
N=300,f=72.2%일 때 신뢰구간 69.2%< P < 74.3%

Decision Tree
N=300,f=67.8%일 때 신뢰구간 64.8%< P < 70.1%

Naive Bayes
N=100,f=70.0%일 때 신뢰구간 64.5%< P < 73.4%

Random Forest
N=300,f=72.2%일 때 신뢰구간 69.2%< P < 74.3%

신뢰도가 99%일 때의 신뢰구간입니다.

5. ANOVA(Analysis of variance)

(1) 분산분석 : 일원배치법

각 알고리즘이 최고성능을 가진 N값의 정확도의 통계적 유의성을 확인하기 위해서 분산분석을 실행합니다. OneR, Decision Tree, Random Forest의 경운 N=300일 때 가장 높은 정확도를 가졌고, Naive Bayes의 경우에는 N=100일 때 가장 높은 정확도를 가졌음을 Learning Curve에서 확인 할 수 있었습니다. 이를 바탕으로 각 알고리즘이 최고성능을 가진 N의 값을 가진 집합을 전체데이터에서 5번 더 출력했고, 이에 대한 정확도를 출력, 정리해 분산분석을 실행했으며 아래와 같은 결과를 얻었습니다.

요약표

인자의 수준	관측수	합	평균	분산
OneR	5	3.244444	0.648889	0.001642
Decision Tree	5	3.381111	0.676222	0.001438
Naive Bayes	5	3.332222	0.666444	0.001889
Random Forest	5	3.386667	0.677333	0.000286

분산 분석

변동의 요인	제곱합	자유도	제곱 평균	F 비	P-값	F 기각치
처리	0.0026	3	0.000867	0.659665	0.588728	3.238872
잔차	0.021019	16	0.001314			
계	0.023619	19				

(2) 결과

결과적으로 0.659665의 F값을 얻을 수 있었으며, 이 값을 통해 귀무가설의 기각 가능여부를 확인하면 아래와 같습니다.

f(0.05, 3, 16) = 3.24

6.59 > 3.24 (신뢰도 95%)

-> Null Hypothesis를 기각 가능

신뢰도가 95%일 때에 Null Hypothesis기각이 가능함을 확인했습니다.

f(0.01, 3, 16) =5.29

6.59 > 5.29 (신뢰도 99%)

-> Null Hypothesis를 기각 가능

신뢰도가 99%일 때에 Null Hypothesis기각이 가능함을 확인했습니다.

이를 통해 신뢰도가 95%일 때뿐만 아니라 99%일 때도 Null Hypothesis기각이 가능함을 확인했고 통계적으로 알고리즘의 정확도 값이 유의미함을 확인 가능했습니다.

5. 프로그램 데모

(1) 프로그램 설명

Input Data	CustomerID
Output Data	이탈여부에 따른 이미지

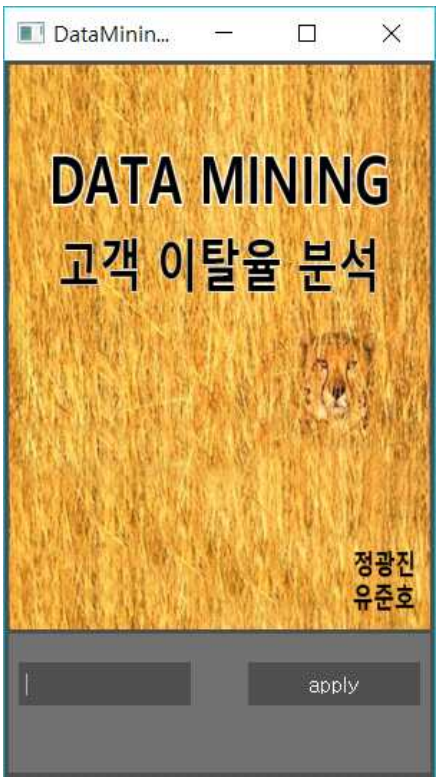
일일이 속성값을 입력하기에는 불편함을 유발한다고 생각하였습니다. 그러므로 예측 값에 따른 Label을 미리 지정하여 엑셀로 저장하고 이에 따른 이벤트 발생만 가능하도록 설계 하였습니다.

프로그램 플로우

이탈한 고객 입력	쿠폰 전송
비이탈 고객 입력	광고 전송

이탈한 고객에게는 쿠폰을 전송함으로써 재방문을 유도하여 소비를 촉발시키는 것이 목표이고,
비이탈 고객에게는 좀 더 소비를 촉발시키기 위해 광고를 보냄으로써 소비를 유도하도록 하였습니다.

(2) 시연된 프로그램 화면



초기화면

CustomerID가 이탈한 고객일 때

CustomerID가 비이탈 고객일 때

Ⅲ. 결론

소비자의 이탈을 예측하는 것은 기업이 고객을 대상으로 하는 쿠폰 발송, 광고 발송과 같은 전략적인 의사 결정에 사용될 수 있음을 확인할 수 있습니다. 이 예측을 고객 관리에 대한 비용 감소를 위해 사용한다면 알고리즘의 '예측'의 측면에서 예측의 정확도를 중요시하며, 그에 따라 비이탈자와 이탈자의 구분에 영향을 끼치는 속성을 알 수는 없지만 가장 높은 정확도에 안정적인 성능을 보이는 'Random Forest' 알고리즘을 사용하는 것이 통계적으로 유의하다는 것을 Learnig Curve, 베르누이, ANOVA를 통해서 확인이 가능했습니다.

높은 예측의 정확도를 통해 발행된 쿠폰과 광고는 단기적으로 기업에게 고객관리의 경쟁우위를 가질 수 있게 합니다. 하지만 장기적으로 보았을 때, 어떤 속성이 이탈에 영향을 크게 미치는 지를 설명하지 못한다면 기업은 장기적인 고객관리는 불가능합니다. 이 때 주목될 수 있는 알고리즘이 'OneR'과 'Decision Tree'입니다. 'OneR'의 경우 손쉽게 어느 속성이 이탈에 큰 영향을 주는지 파악 할 수 있게 했는데 선정된 규칙을 통해 확인이 가능합니다. 그리고 'Decision Tree'는 직관적으로 어떤 방식으로 나무가 구성이 되었는지를 확인함으로써 어떤 속성을 통해 이탈과 비이탈의 구분 규칙이 생성이 되는지 시각적으로 볼 수 있습니다. 그 규칙은 구매금액(Total_Quantity)과 방문 간격의 평균(AvgTimeDelta)임을 파악할 수 있었으며 이에 따라 기업은 어떻게 각 속성들의 수치를 어떻게 증가시킬지를 고민할 수 있습니다. 구매금액(Total_Quantity)을 높이기 위해서 다량의 구매 시의 할인 정책 제공할 수 있으며 단순한 전략적 의사결정을 넘어서 전술적 의사결정이 가능함을 확인할 수 있는 부분입니다. 또한 방문 간격의 평균(AvgTimeDelta)을 줄이기 위해 신제품에 대한 메일발송 등의 장기적인 고객관리 방안을 도출할 수 있는 근거를 'OneR'과 'Decision Tree'을 통해 얻을 수 있었습니다.

프로젝트의 진행은 컴퓨터 공학부의 전공수업이지만 컴퓨터공학부의 학생과 경영학과의 학생이 함께 진행했습니다. 각 학생들은 프로젝트를 진행하면서 의사소통 능력의 중요성을 실감하고 협업을 통한 최고의 결과를 도출하기 위해 토론하고 함께 고민했으며 그에 따라 유의미한 결론을 도출할 수 있는 프로젝트를 진행했습니다. 현재 진행 상황은 다양한 알고리즘을 적용하지 못했으며, 이탈의 정의 부분에서도 그 근거가 부족하기 때문에 한계를 가지고 있는 것은 사실입니다. 하지만 알고리즘들의 사용방법에 대한 고민과 최고성능의 알고리즘을 파악하는 프로세스의 진행, 알고리즘의 사용이 통계적으로 유의미한지 아닌지를 검증하는 과정에 대한 경험은 추후 공모전, 포트폴리오 작성 및 대외활동 나아가 취업 후의 업무에서도 유용하게 쓰일 것으로 기대됩니다.

Ⅳ. 참고문헌

서적

Ian H. Witten, Eibe Frank, Mark A. Hall(2011). Data Mining: Practical Machine Learning Tools and Techniques

Decision Tree
https://ko.wikipedia.org/wiki/%EA%B2%B0%EC%A0%95_%ED%8A%B8%EB%A6%AC_%ED%95%99%EC%8A%B5%EB%B2%95

Naive Bayes
https://ko.wikipedia.org/wiki/%EB%82%98%EC%9D%B4%EB%B8%8C_%EB%B2%A0%EC%9D%B4%EC%A6%88_%EB%B6%84%EB%A5%98#%EA%B0%80%EC%9A%B0%EC%8B%9C%EC%95%88_%EB%82%98%EC%9D%B4%EB%B8%8C_%EB%B2%A0%EC%9D%B4%EC%A6%88

Random Fore
https://ko.wikipedia.org/wiki/%EB%9E%9C%EB%8D%A4_%ED%8F%AC%EB%A0%88%EC%8A%A4%ED%8A%B8