리조트 호텔 예약취소 고객의 특징 파악 연구

-로지스틱 회귀모형과 의사결정나무 분석을 중심으로-

담당교수 김 규 성

서울시립대학교 통계학과 2018580008

문혜현

2020년 12월

목차

1. 서론

1.1 연구목적

2020년 코로나 사태 직후 항공길과 국내여행길이 막히면서 많은 숙박업소들이 예약취소를 줄줄이 맞게 되었다. 이와 같이 거의 대부분의 예약이 취소되는 사례는 매우 특별한 경우이다. 하지만 코로나와 같이 특별한 시기가 아닐 때에도 숙박업소는 예약취소를 빈번하게 받는다. 예약취소가 예약날짜에 임박해 일어나게 되면 숙박업소는 불가피하게 피해를 받는다. 다른 손님을 받을 기회를 잃고, 예약에 맞춰 준비한 것들이 필요가 없어지기 때문이다. 이에 불가피한 사정이 아닌 이상 숙박업소는 무분별한 예약에 따른 피해를 막기 위해 자사의 규칙 및 위약금을 명시하며 법적으로도 날짜에 따른 위약금을 명시하고 있다. 실제로 이러한 규제가 생긴 뒤 예약이 취소되는 건수가 많이 줄어들었지만 여전히 예약취소는 의도적이든 불가피하든 숙박업에서 계속해서 일어나고 있다.

따라서 본 연구에서는 숙박업소 중 호텔의 예약취소가 숙박고객의 어떤 특성에 따라서 일어날 확률이 높은지 로지스틱 회귀모형과 의사결정나무를 통해 세부적으로 파악하고자 한다. 본 연구의 결과는 호텔에서 예약취소를 최소화하기 위한 정책을 만드는 데 도움을 줄 수 있고 호텔이 초과예약(overbooking)을 고려할 때 참고할 수 있는 유익한 자료가 될 것이다.

1.2 문헌연구

반종삼(2010)은 휴양콘도미니엄인 H리조트 운영사례를 통해 예약취소 및 NO-SHOW가 주는 피해를 추정하며 이를 최소화하기 위해 예약운영방식 상의 허점을 조사하고 벌점제도를 도입한다. 결과적으로 벌점제도에 대한 효과가 통계적으로 유의하지는 않았으나 NO-SHOW의 감소추세가 있음을 입증했다. [1]

천송이(2013)는 예약부도가 나타나는 사례를 로지스틱 회귀분석과 의사결정나무 기법을 적용하여 예약부도확률이 강한 집단의 특성을 파악했다. 예약부도 가해경험, 예약부도 피해경험, 예약취소경험 순으로 예약부도에 크게 영향을 미친다는 결과를 도출했다. 그리고 이 세 인자를 중심으로 예약부도를 낼 사람들을 로지스틱 회귀모델로 예측을 한 결과

---

[1] 반종삼(2008), "휴양콘도미니엄 객실 가동율에 영향을 미치는 예약취소 및 NO-SHOW에 관한 연구-H리조트 운영사례를 중심으로"

약 91.3 %의 정확도를 보였다.[2]

다음 두 문헌을 통해 본 연구에서 예약을 취소하는 고객의 특성파악을 바탕으로 예약취소를 줄이는 정책을 만드는 것이 실제로 효과가 있을 수 있다는 가능성을 암시할 수 있다. 또한, 본 연구에서 도출한 예약취소 확률을 높이는 특성들을 예측인자로 초과예약 모델을 만드는 데 도움이 될 수 있다는 것을 알 수 있다. 선행연구에서 예약취소고객의 특성을 파악하는 데 로지스틱 회귀모형과 의사결정나무 분석이 용이하게 사용되었으므로 본 연구에서도 로지스틱 회귀모형과 의사결정나무를 이용해 분석하겠다.

선행연구는 특정 리조트를 선정하여 연구가 이루어졌거나 700명 이하의 표본을 기반으로 인구통계학적인 특성, 부도의 피해, 가해 경험 등을 조사하였다. 이에 본 연구는 여러 리조트 호텔의 4만 개 이상의 예약을 담은 큰 데이터와 이전 선행연구에서 설명되지 않았던 변수들을 이용해 연구를 진행함으로서 선행연구와 차별점을 두고 새로운 예약취소 요인을 밝혀내겠다.

1.3 데이터 설명

Hotel Booking demand dataset을 사용했고 포르투갈에 위치해 있는 호텔의 예약취소와 수립에 관한 데이터이다. Property Management System(PMS)에서 추출하였으며 Algarve 지역에 위치한 리조트 호텔의 총 40,060개의 예약과 Lisbon에 위치한 도시 호텔의 79,330개의 예약으로 표본이 이루어져 있다. 본 연구에서는 리조트 호텔의 표본만을 사용한다. 2015년 7월 1일부터 2017년 8월 31일까지 도착예정으로 되어 있는 예약에 한해서 데이터가 있으며 29개의 변수들로 구성되어 있다. [표 1]에 각 변수가 무엇을 의미하는지 상세히 설명했다.

[표 1]Hotel Booking demand dataset 변수

ADR	숙박기간 동안의 거래액을 총 숙박기간으로 나눈 값	Babies	아기의 수
Adults	어른의 수	BookingChanges	체크인 될 때까지의 예약변경 수
Agent	예약을 한 여행 에	*ReservedRoomType*	방의 유형에 대한

[2] 천송이(2013), "리조트 예약부도 방지를 위한 연구-의사결정나무 분석과 로지스틱 회귀분석의 적용을 중심으로"

	이전시의 이름		코드
ArrivalDateDayOfMonth	도착날짜의 날짜	DistributionChannel	Travel Agents, Tour Operators 분류
ArrivalDateMonth	도착날짜의 월 이름	IsCanceled	예약이 취소됨(1), 수립됨(0)
ArrivalDateWeekNumber	도착날짜의 몇 번째 주	IsRepeatedGuest	재방문일 경우(1), 첫 방문일 경우(0)
ArrivalDateYear	도착날짜의 연도	LeadTime	PMS에 예약을 입력한 날과 도착한 날 사이의 기간
AssignedRoomType	예약된 방의 유형	MarketSegment	고객 유형
Meal	예약된 식사 종류	C*ompany*	회사예약이름
C*ustomer type*	Contract, Group, Transient, Transient-pary 네 가지로 분류	C*ountry*	고객의 국적
Days in waiting list	예약이 손님에게 확인되기까지 걸린 기간	P*revious Booking Not Canceled*	재예약 고객중 이전 예약에 대해서 취소하지 않은 예약 개수
TotalOfSpecialRequests	고객이 주문한 특별한 요구의 개수	*PreviousCancellations*	재예약 고객중 이전 예약에 대해서 취소한 예약개수
RequiredCardParkingSpaces	예약된 주차공간개수	*ReservationStatus*	마지막예약상태,Canceled, Check-out, NO-Show 세 가지로 분류됨
ReservationStatusDate	마지막 예약상태가 찍힌 날짜	*StaysInWeekNights*	호텔에서 머물 동안 평일 개수
staysInWeekendNights	호텔에서 머물 동안 주말 개수		

1.4 분석방법

(1) 로지스틱 회귀(Logistic Regression)

회귀분석은 목표 변수가 입력 변수들에 의해서 어떻게 설명 또는 예측되는지를 알아보기 위해 자료를 적절한 함수식으로 표현하여 분석하는 통계적 방법을 말한다. 로지스틱 회귀분석은 종속변수가 명목변수이고 값이 이항형이거나 순서형으로 나타내는 경우에 이용된다. 종속변수와 독립 변수 간의 관계를 S자형으로 가정한 로지스틱 회귀모형은 아래와 같다.

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

어떤 사건이 발생할 확률을 예측하는 것이 되므로 종속 값은 0과 1사이의 값을 가지게 된다. 분석 결과 종속 변수 값, 즉 확률이 0.5보다 크면 그 사건이 일어난 것이고 0.5보다 작으면 그 사건이 일어나지 않는 것으로 예측하게 된다.

오즈비(odds ratio)는 한 변수의 어떠한 관측값에서의 오즈에 대해 한 단위 증가했을 때의 관측값에서의 오즈의 비를 나타낸 것이다.

$$e^{\beta_i} = \frac{\pi(x_i+1)/(1-\pi(x_i+1))}{\pi(x_i)/(1-\pi(x_i))}$$

β_i가 양수이면 단위가 1씩 증가할 때, 어떤 사건이 일어날 오즈가 증가하고 β_i가 음수이면 단위가 1씩 증가할 때, 어떤 사건이 일어날 오즈가 감소한다.

(2) 의사결정나무(Decision Tree)

의사결정나무는 의사결정규칙을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류하거나 예측을 수행하는 분석방법으로 나무구조로 표현이 된다. 다양한 계산방식이 있으며 본 연구에서는 R에 내장되어있는 rpart패키지의 함수 rpart와 party패키지의 함수 ctree를 사용하였다.

Rpart 나무 모형은 모든 입력변수에 대해 모든 가능한 분할 지점을 조사한다. 구해진 분할 지점 중에서 순수도를 증가시키는 최적의 분할 기준을 선택한다. 분류 나무에서는 지니 지수(Gini index)를 순수도를 증가시키기 위한 기준으로 사용한다. 부모 마디에서 자료가 선택된 분할 기준을 충족하면 왼쪽 자식 마디로, 충족하지 않으면 오른쪽 자식마디로 보낸다. 동일한 방법으로 새로 만들어지는 자식 마디마다 반복하여 분할한다. 이 과정은 가장 하위 마디에 속한 자료의 개수가 지정된 최솟값에 도달하면 분할을 중지한다.

party패키지의 함수 ctree는 조건부 추론방법에 이론적 바탕을 둔 불편의 반복분할 방법으로 다중검정을 이용한 정지 규칙을 이용하고 있다. 변수 선택은 다음과 같은 과정을 거친다. 우선 반응변수와 각각의 설명변수들 간의 독립성 검정에 대하여 유의확률을 구한다. 유의확률을 구할 때는 적절한 통계량을 설정하고 통계량의 조건부 확률분포를 이용하여 유의확률을 구한다. 조건부 분포를 알지 못할 경우에는 조건부 몬테칼로 방법을 이용하여 근사적 분포를 계산하고 이를 이용하여 유의확률을 구한다. 마지막으로 다중검정을 고려하기 위해 구한 각 유의확률들의 보정유의확률을 구하고 이를 이용하여 변수를 선택한다.

1.5 결과 활용 및 기대 효과

로지스틱 회귀분석과 의사결정나무를 통해 예약취소확률을 높이는 예약특징을 분석하여 예약취소 현상의 원인을 규명하고 예약취소를 방지하는 방법을 규명하고자 하였다. 결과적으로 이러한 연구를 통해 호텔산업의 실무자들에게 합리적인 마케팅 시사점을 제공할 수 있을 뿐만 아니라 초과예약(overbooking)을 예측할 때 사용할 수 있는 예측인자에 대한 정보도 제공한다. 이로서 리조트 호텔의 예약취소 방지와 초과예약 도모를 통해 기업과 사회의 이익을 도모할 수 있다.

2. 본론

2.1 분석 과정 소개

우선 Resort hotel 표본이 29개의 변수를 모두 적합하는 것은 모형의 복잡성을 매우 높이므로 알맞은 변수들을 선택하려 한다. 변수들 사이의 피어슨 상관계수 비교, Stepwise Selection 그리고 모든 변수들을 적합한 로지스틱 회귀모형의 왈드검정을 통해서 적합한 변수들을 선택한다.

분석의 목적은 어떤 예약정보가 예약을 취소시킬 확률을 높이는 지를 아는 것이다. 따라서 타겟 변수로 예약취소와 예약수립을 각각 1과 0으로 둔 이항변수로 한다. 이에 맞추어 로지스틱 회귀모형을 사용하며 앞서 선택한 변수들을 이용하여 fischer's socring을 사용한 최대우도추정법으로 모수를 추정한다.

최적의 모형을 도출하기 위해 모형을 진단할 필요가 있다. 모형을 진단하는 기준은 모

형에 대한 검정으로 우도비검정(Likelihood Ratio Test), 계수에 대한 검정으로 Wald 통계량, 적합결여 검정으로 편차(Deviance)와 피어슨(Pearson) 카이제곱 검정, 그리고 피어슨 잔차와 편차 잔차이다. 이 기준들이 최적이 되도록 로지스틱 회귀모형을 적합한다.

2.2 데이터 분석 및 결과 설명

2.2.1 변수 선택

(1) 피어슨 상관계수

명목형 변수는 one-hot encoding을 적용하였고, 순서가 있는 변수는 label-encoding을 적용하여 수치형으로 만들었다. 각 변수 사이의 상관계수를 구하였으며 타겟 변수인 예약취소와의 상관계수가 0.08 이상인 변수들을 추린 결과 다음과 같다.

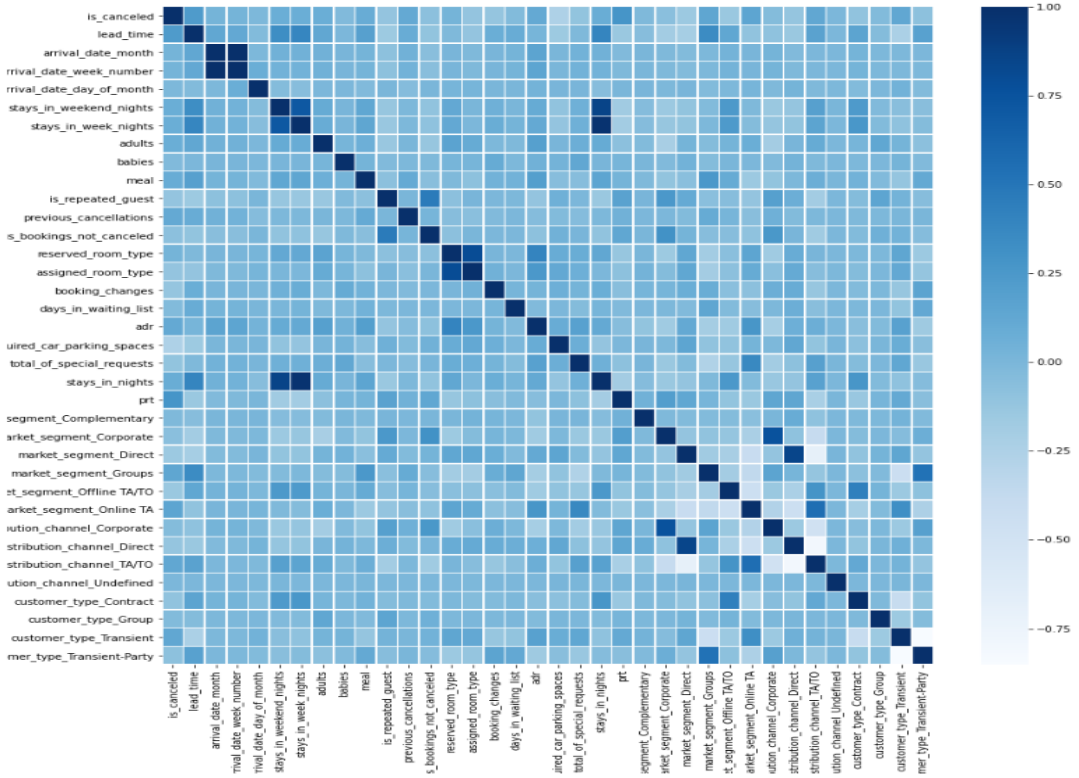
prt	Distribution channel
required_car_parking_spaces	booking_changes
lead_time	previous_cancellations
Market segment	assigned_room_type
customer_type	adr
stays_in_nights	is_repeated_guest
meal	total_of_special_requests
Stays_in_week_nights	

또한, 독립변수가 독립이라는 가정 하에 로지스틱 회귀 모형을 적합하는 것이므로 상관계수가 상대적으로 높게 나온 독립변수 쌍을 찾아본 결과 총 5쌍이 있었고, 의미적으로도 유사하다.

Stay_in_nights	Staay_in_week_nights	stay_in_weekend _nights
Assigned_room_type	reserved_room_type	
Market_segment_channel	Distribution_channel_corporate	Distribution_channel_discrete
Previous_bookings_not_cancelled	is_repeated_guest	
Arrival_date_week_number	arrival_date_month	

위의 과정을 한 눈에 볼 수 있도록 [그림 1]에 나타내 보았다.

[그림 1] 피어슨 상관계수 그래프



(2) 단계적 선택법

SAS의 stepwise 옵션을 사용하여 모든 변수에 대하여 단계적 선택법을 통해 변수를 선택하였다. 그 결과 다음 변수들이 선택이 되었다.

Prt	Market Segment

(3) 로지스틱 회귀에 모든 변수를 넣고 적합

로지스틱 회귀모형에 모든 변수를 넣고 적합했을 때, 유의하지 않은 변수들을 살펴본 결과, 유의하지 않은 변수들은 아래와 같다.

Babies	Required_car_parking
Distribution_channel	Stays_in_weekend_nights
meal	

(4) 의미가 없거나 의미가 유사한 변수

Days_in_waiting_list , Arrival_date_week_number, Arrival_date_year의 경우 수치의 의미를 해석하기 어렵거나 예약취소와 관련이 없다고 판단하여 제거한다. 피어슨 상관계수에서 수치상으로 비슷한 변수뿐만 아니라 의미상으로 비슷한 변수를 판별해 보았다.

Lead_time	reservation_status_date
Is_cancelled	reservation_status

(5) 결론

예약취소와 높은 상관을 가진 변수, 단계적 선택법을 통해 선택된 변수, 모든 변수를 넣은 로지스틱 회귀모형에서 유의한 변수를 비교하여 공통적으로 나오는 변수들은 아래와 같다.

prt	Is_repeated_guest
lead_time	Assigned_room_type
market_segment	adr
Customer_type	Stays_in_nights
Booking_changes	Previous_cancellations

방식들 간의 차이로 공통적으로 선택되지 못한 변수들은 정확한 통계적 해석을 위해 우선적으로 모형에 넣는다. 하지만, 의미적으로 유사한 변수들과 상관계수가 높은 쌍들의 경우 더 유의하게 나왔거나 다양한 방식에서 더 많이 선택된 변수만을 선별하여 넣는다.

required_car_parking_spaces	adults
total_of_special_requests	meal
arrival_date	

본 연구에서는 유의하지는 않으나 선행연구에서 영향이 있다고 언급된 변수가 있는지 검토한 결과 없음을 확인했다. 따라서, 변수선택 수행 결과 위의 변수들로 첫 모형 적합을 시도한다.

2.2.2 로지스틱 회귀분석

(1) 변수선택법으로 선택한 변수들로 적합한 로지스틱 회귀모형

변수선택법으로 선택한 변수들로 로지스틱 회귀모형을 적합한 결과는 [표 2]이고, 유의하지 않은 회귀계수들이 발견되었다. 회귀계수에 대한 왈드 카이제곱 통계량에 대해서 유의수준 5%에서 유의하지 않은 변수들로 arrival_date_day_of_month, customer_type_contract, customer_type_group, market_segment_complementary, meal, required_car_parking이 있다. 따라서 이 모형은 적절하지 않다. 유의하지 않은 변수들 중 arrival_date_day_of_month, meal, required_car_parking을 제거하고 다시 모형 적합을 시도해 본다.

[표 2](1) 로지스틱 회귀 모형

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-3,8648	0,1277	916,0296	<,0001
adr		1	0,00623	0,000283	486,3554	<,0001
adults		1	0,3138	0,0386	66,0018	<,0001
arrival_date_day_of_		1	0,00106	0,00167	0,4034	0,5254
arrival_date_month		1	-0,0274	0,00508	29,1216	<,0001
assigned_room_type		1	-0,1116	0,00763	213,8557	<,0001
booking_changes		1	-0,3949	0,0272	210,1002	<,0001
customer_type	Contract	1	-0,1630	0,1083	2,2631	0,1325
customer_type	Group	1	-0,3294	0,2271	2,1042	0,1469
customer_type	Transient	1	0,9059	0,0815	123,4862	<,0001
is_repeated_guest		1	-2,6124	0,1502	302,3288	<,0001
lead_time		1	0,00603	0,000187	1039,7182	<,0001
market_segment	Complementary	1	0,0977	0,1841	0,2818	0,5955
market_segment	Corporate	1	-0,2801	0,0721	15,0705	0,0001
market_segment	Direct	1	-0,7432	0,0557	177,8911	<,0001
market_segment	Groups	1	0,8636	0,0629	188,3474	<,0001
market_segment	Offline TA/TO	1	-0,9199	0,0557	272,5548	<,0001
meal		1	-0,0588	0,0343	2,9460	0,0861
previous_cancellatio		1	2,7536	0,1348	417,1243	<,0001
prt		1	2,0143	0,0336	3589,1733	<,0001
required_car_parking		1	-17,7337	118,2	0,0225	0,8808
total_of_special_req		1	-0,6572	0,0220	892,4459	<,0001
stays_in_nights		1	0,0574	0,00506	128,6841	<,0001

(2) (1)에서 유의하지 않은 변수들을 제거한 후 적합한 로지스틱 회귀모형

(1)에서 유의하지 않은 계수를 제거하여 로지스틱 회귀모형을 적합했고 결과는 [표 2]이다. 회귀계수에 대한 왈드 카이제곱검정 통계량이 대부분의 변수에서 유의해 졌으며 market_segment_complementary, customer_type_contract, customer_type_group은 유의하지 않는다.

[표 3](2) 로지스틱 회귀 모형

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-3,8422	0,1150	1116,2648	<0001
adr		1	0,00510	0,000255	398,7098	<0001
adults		1	0,2750	0,0367	56,1302	<0001
arrival_date_month		1	-0,0285	0,00486	34,3405	<0001
assigned_room_type		1	-0,1282	0,00719	317,4921	<0001
booking_changes		1	-0,4615	0,0264	305,2119	<0001
is_repeated_guest		1	-2,4210	0,1386	305,0501	<0001
lead_time		1	0,00625	0,000178	1233,9338	<0001
market_segment	Complementary	1	0,1074	0,1775	0,3658	0,5453
market_segment	Corporate	1	-0,2426	0,0701	11,9837	0,0005
market_segment	Direct	1	-0,8221	0,0534	237,3328	<0001
market_segment	Groups	1	0,8880	0,0596	221,6228	<0001
market_segment	Offline TA/TO	1	-0,8640	0,0539	256,8444	<0001
customer_type	Contract	1	-0,1858	0,1058	3,0868	0,0789
customer_type	Group	1	-0,2649	0,2190	1,4627	0,2265
customer_type	Transient	1	0,8507	0,0788	116,5235	<0001
previous_cancellatio		1	2,4974	0,1165	459,4448	<0001
prt		1	1,8766	0,0314	3581,0235	<0001
total_of_special_req		1	-0,6592	0,0209	992,8329	<0001
stays_in_nights		1	0,0662	0,00485	186,0043	<0001

(3) 유의성이 상대적으로 약한 변수들을 제거했을 때의 로지스틱 회귀모형

더 단순한 모형을 만들어 보기 위해 변수의 왈드카이제곱 검정통계량이 비교적 낮은 Adults, arrival_date_month, customer_type 변수를 제거하고 남은 변수들로 로지스틱 회귀모형을 적합했고 결과는 [표 4]이다. 모든 변수들이 왈드카이제곱 검정에서 유의함을 보인다.

[표 4](3) 로지스틱 회귀 모형

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.0628	0.0542	3193.1311	<.0001
adr		1	0.00545	0.000242	508.6993	<.0001
assigned_room_type		1	-0.1146	0.00704	264.7234	<.0001
booking_changes		1	-0.5305	0.0263	408.1527	<.0001
is_repeated_guest		1	-2.4462	0.1381	313.9707	<.0001
lead_time		1	0.00598	0.000171	1225.9321	<.0001
market_segment	Complementary	1	0.3373	0.1762	3.6640	0.0556
market_segment	Corporate	1	-0.4335	0.0674	41.2983	<.0001
market_segment	Direct	1	-0.5999	0.0523	131.6089	<.0001
market_segment	Groups	1	0.3533	0.0510	47.9113	<.0001
market_segment	Offline TA/TO	1	-0.8506	0.0517	270.4003	<.0001
previous_cancellatio		1	2.5492	0.1144	496.8228	<.0001
prt		1	1.9568	0.0309	4006.1110	<.0001
total_of_special_req		1	-0.6617	0.0207	1021.0052	<.0001
stays_in_nights		1	0.0685	0.00473	210.4102	<.0001

(4) 최적의 모형 선택

(2)와 (3)의 모형 비교를 통해 최종 모형을 결정한다. (2)는 독립변수가 19개이고 왈드 카이제곱검정에서 유의하지 않은 회귀계수가 2개 있었으며, (3)은 독립변수가 14개이고 모든 회귀계수가 유의하다.

(2)는 편차 적합결여검정(Deviance Goodness-of-Fit Test)에서 귀무가설을 채택하므로 saturated model과 full model이 비슷하며 모형이 적절함을 나타낸다. 하지만, 피어슨 적합결여검정(Pearson Goodness-of-Fit Test)에서 귀무가설을 기각하므로 관측치와 평균 사이가 크며 모형이 적절치 않다는 것을 알려준다. (3)은 편차와 피어슨 적합결여검정 모두에서 모형이 적절하지 않다는 결과가 나왔다.

(2)와 (3)의 우도비 검정을 살펴보면 두 모형 모두 귀무가설인 $H_0: \beta_1 = \beta_2 = \cdots \beta_3 = 0$을 기각하고 적어도 하나의 회귀계수가 유의하다는 것을 알 수 있다. AIC는 (2)에서 31794.234, (3)에서 32632.561로 (2)의 AIC가 더 작기 때문에 (2)의 모형이 더 적절하다고 판단할 수 있다.

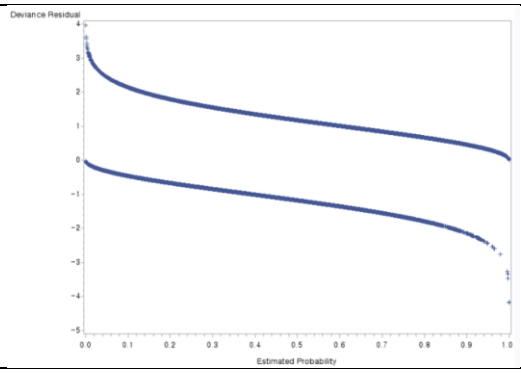
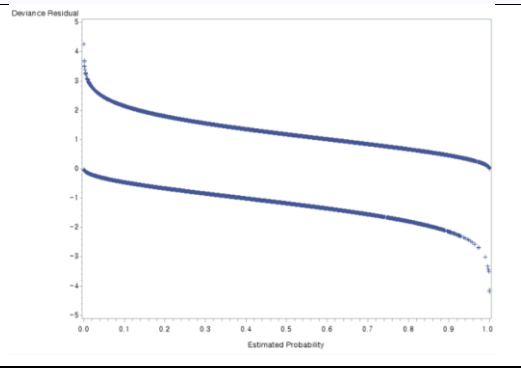
R-square를 살펴보면 (2)는 약 30%, (3)은 28.8%의 설명력을 가지고 있어 크게 차이나지는 않는다.

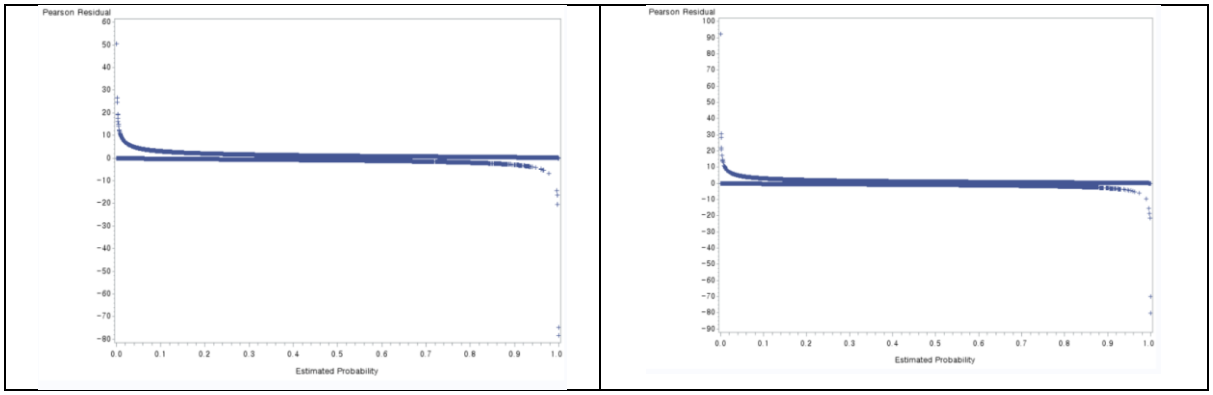
(2)와 (3)의 편차 잔차(Deviance Residual)와 적합값 그래프, 피어슨 잔차(Pearson Residual)와 적합값 그래프를 통해 이상치를 살펴보려 한다. 이상치로 보이는 점들이 보

이므로 추가적인 조치가 필요해 보인다.

위의 내용들을 종합하여 (2) 모형을 본 연구에서의 호텔예약 데이터에 대한 적합모형으로 선택한다. 그 이유는 우도비검정에서 유의하며 대부분의 회귀계수에 대한 왈드 검정에서 유의하다. 또한, 편차적합 결여검정에서 모형이 적합하다고 결론이 나왔음을 통해 모형이 어느 정도 적합하다는 것을 알 수 있다. (3) 적합결여검정 결과 모형이 적절하지 않다고 나왔으며 (2) 모형의 AIC가 더 낮고 R-square가 더 높음을 통해 (2) 모형을 선택한다. (3) 모형이 더 단순하나 충분히 많은 데이터가 있으므로 더 단순한 모형을 꼭 선택할 필요가 없다.

[표 5] (2)와 (3) 로지스틱 회귀 모형 비교

<table>
<tr><th>(2)로지스틱 회귀 모형</th><th>(3)로지스틱 회귀 모형</th></tr>
<tr><td>
<table>
<tr><th colspan="5">Deviance and Pearson Goodness-of-Fit Statistics</th></tr>
<tr><th>Criterion</th><th>Value</th><th>DF</th><th>Value/DF</th><th>Pr > ChiSq</th></tr>
<tr><td>Deviance</td><td>31333.7631</td><td>32E3</td><td>0.9751</td><td>0.9993</td></tr>
<tr><td>Pearson</td><td>56185.3735</td><td>32E3</td><td>1.7484</td><td><.0001</td></tr>
</table>
</td><td>
<table>
<tr><th colspan="5">Deviance and Pearson Goodness-of-Fit Statistics</th></tr>
<tr><th>Criterion</th><th>Value</th><th>DF</th><th>Value/DF</th><th>Pr > ChiSq</th></tr>
<tr><td>Deviance</td><td>31983.3256</td><td>31E3</td><td>1.0217</td><td>0.0035</td></tr>
<tr><td>Pearson</td><td>62237.8754</td><td>31E3</td><td>1.9881</td><td><.0001</td></tr>
</table>
</td></tr>
<tr><td>
<table>
<tr><th colspan="3">Model Fit Statistics</th></tr>
<tr><th>Criterion</th><th>Intercept Only</th><th>Intercept and Covariates</th></tr>
<tr><td>AIC</td><td>45634.731</td><td>31794.234</td></tr>
<tr><td>SC</td><td>45643.288</td><td>31965.366</td></tr>
<tr><td>-2 Log L</td><td>45632.731</td><td>31754.234</td></tr>
</table>
</td><td>
<table>
<tr><th colspan="3">Model Fit Statistics</th></tr>
<tr><th>Criterion</th><th>Intercept Only</th><th>Intercept and Covariates</th></tr>
<tr><td>AIC</td><td>45634.731</td><td>32632.561</td></tr>
<tr><td>SC</td><td>45643.288</td><td>32760.910</td></tr>
<tr><td>-2 Log L</td><td>45632.731</td><td>32602.561</td></tr>
</table>
</td></tr>
<tr><td>
<table>
<tr><th colspan="4">Testing Global Null Hypothesis: BETA=0</th></tr>
<tr><th>Test</th><th>Chi-Square</th><th>DF</th><th>Pr > ChiSq</th></tr>
<tr><td>Likelihood Ratio</td><td>13878.4975</td><td>19</td><td><.0001</td></tr>
<tr><td>Score</td><td>11460.7481</td><td>19</td><td><.0001</td></tr>
<tr><td>Wald</td><td>7766.3967</td><td>19</td><td><.0001</td></tr>
</table>
</td><td>
<table>
<tr><th colspan="4">Testing Global Null Hypothesis: BETA=0</th></tr>
<tr><th>Test</th><th>Chi-Square</th><th>DF</th><th>Pr > ChiSq</th></tr>
<tr><td>Likelihood Ratio</td><td>13030.1705</td><td>14</td><td><.0001</td></tr>
<tr><td>Score</td><td>10785.1124</td><td>14</td><td><.0001</td></tr>
<tr><td>Wald</td><td>7594.4743</td><td>14</td><td><.0001</td></tr>
</table>
</td></tr>
<tr><td>
<table>
<tr><td>R-Square</td><td>0.3031</td><td>Max-rescaled R-Square</td><td>0.4361</td></tr>
</table>
</td><td>
<table>
<tr><td>R-Square</td><td>0.2876</td><td>Max-rescaled R-Square</td><td>0.4138</td></tr>
</table>
</td></tr>
<tr><td> 
</td><td> 
</td></tr>
</table>



(5) 이상치를 제거한 최종 로지스틱 회귀 모형

최적의 모형으로 선택된 (2)모형에서 편차 잔차와 피어슨 잔차가 ±2이상인 것들을 제거한 뒤 (2)모형에서 사용된 변수들과 똑같은 변수들로 로지스틱 회귀모형을 적합하였다. 각 회귀계수에 대한 왈드 검정에서 유의수준 5%에서 Customer_type_contract만 유의하지 않고, 다른 모든 회귀계수들은 유의함을 볼 수 있다.

[표 6](5) 이상치를 제거한 최종 로지스틱 회귀 모형

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-6.0728	0.1733	1227.3519	<.0001
adr		1	0.00940	0.000321	856.2966	<.0001
adults		1	0.5419	0.0480	127.7021	<.0001
arrival_date_month		1	-0.0700	0.00638	120.4222	<.0001
assigned_room_type		1	-0.2110	0.00911	536.7204	<.0001
booking_changes		1	-1.1389	0.0425	717.4418	<.0001
is_repeated_guest		1	-4.2149	0.2610	260.8703	<.0001
lead_time		1	0.0104	0.000234	1982.0197	<.0001
market_segment	Complementary	1	0.5630	0.2075	7.3650	0.0067
market_segment	Corporate	1	-0.3135	0.0881	12.6733	0.0004
market_segment	Direct	1	-1.7568	0.0713	606.8101	<.0001
market_segment	Groups	1	1.6239	0.0804	407.6604	<.0001
market_segment	Offline TA/TO	1	-1.4118	0.0679	432.7379	<.0001
customer_type	Contract	1	0.0493	0.1598	0.0951	0.7577
customer_type	Group	1	-0.7442	0.3673	4.1063	0.0427
customer_type	Transient	1	1.5477	0.1291	143.7819	<.0001
previous_cancellatio		1	4.9134	0.2144	525.0818	<.0001
prt		1	2.8428	0.0431	4346.1327	<.0001
total_of_special_req		1	-1.2441	0.0290	1836.1358	<.0001
stays_in_nights		1	0.1115	0.00609	335.0015	<.0001

[표 2]에서 나온 결과를 오즈비(odds ratio)를 사용해 해석한다. 회귀계수가 양수로 나온 변수들은 1단위씩 증가할수록 이전 오즈에 비해 취소할 확률이 증가하는 것을 의미한다. adr, adults, lead_time, market_segment_complementary, market_segment_groups,

customer_type_transient, previous_cancellation, prt, stays_in_nights가 이에 해당한다.

양수이면서 높은 오즈비를 가지는 계수는 previous_cancellations, prt, customer_type_transient이다. 이 계수들에 집중하여 해석하면 과거의 예약취소개수가 하나 더 늘어날수록 이전 odds에 비해 136배 증가한다. 자국민이 외국인의 odds에 비해 17배 증가한다. 한시적으로 머무르는 개인 고객은 한시적으로 머무르는 무리의 odds에 비해 11배 증가한다. 즉, 과거의 예약취소가 많을수록 외국인에 비해 자국민이 한시적으로 머무르는 무리보다 개인이 더 예약취소를 할 확률이 높다.

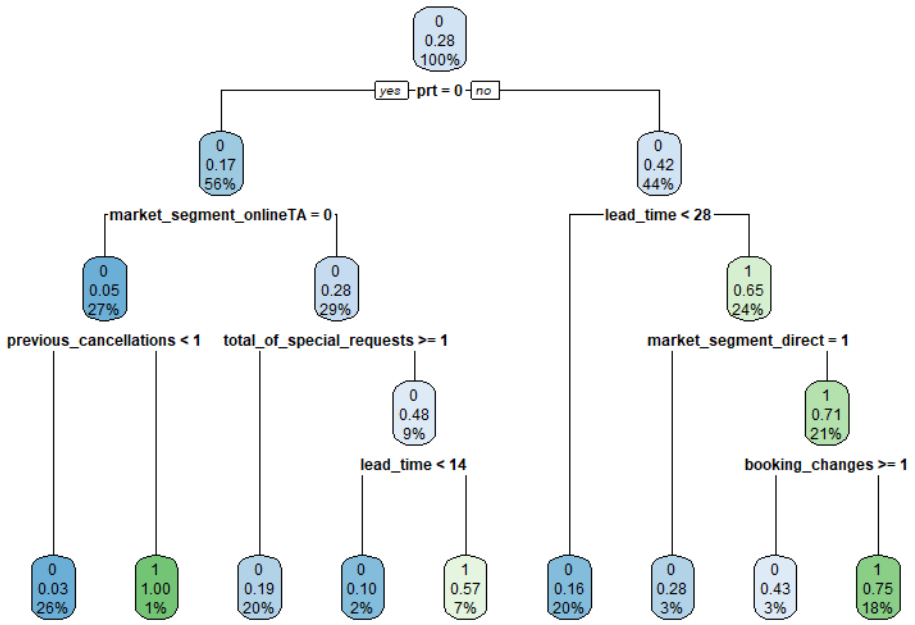
회귀계수가 음수로 나온 변수들은 1단위씩 증가할수록 이전 오즈에 비해 취소할 확률이 감소하는 것을 의미한다. Arrival_date_month, assigned_room_type, booking_changes, is_repeated_guest, market_segment_corporate, market_segment_direct, market_segmet_offlineTA/TO, customer_type_group, total_of_special_requests가 이에 해당한다. 음수이면서 절댓값이 큰 오즈비를 가지는 계수는 is_repeated_guests, booking_changes, total_of_special requests, markget_segment_direct이다. 이 계수들에 집중하여 해석하면 이전에 호텔을 방문했던 고객은 방문하지 않았던 고객의 오즈의 0.015배로 감소한다. 예약변경이 하나 더 늘어날수록 이전 오즈에 비해 0.32배 감소한다. 특별한 요구가 하나 더 늘어날수록 이전 오즈에 비해 0.28배 감소한다. 직접 호텔을 이용하는 고객은 온라인 투어를 이용하는 고객의 오즈에 비해 0.04배 감소한다. 즉, 이전에 방문했던 고객이고, 예약변경이나 특별한 요구가 많을수록, 직접 호텔을 이용하는 경우 예약취소확률이 낮다.

2.2.3 의사결정 나무(Decision Tree)

(1) Rpart 패키지

R 프로그램에 내장되어 있는 Rpart 패키지의 rpart 함수를 이용해 의사결정나무를 만들었다. 최종적으로 결정한 로지스틱 회귀모형에 사용했던 변수들을 입력변수로 하였고, 타겟변수는 예약취소(1), 예약수립(0)으로 하였다. 아래 그래프는 의사결정나무를 도식화한 것이다.

[그림 2] rpart 의사결정나무 그래프



각 노드에 표시되어 있는 라벨들을 해석하면 다음과 같다. 예를 들어, prt에 대한 답으로 Yes를 얻은 자식 노드의 가장 위에 있는 라벨 0은 자신의 노드 안에 있는 데이터 중 1보다 0이 더 많다는 것을 의미한다. 그리고 중간의 라벨 0.17은 자신의 노드 속 1의 비율을 의미하며 가장 아래에 있는 라벨 56%는 전체 데이터 중 자신의 노드 속 데이터의 비율을 의미한다.

라벨이 1이 붙은 소집단 3개에 대해서 왼쪽부터 해석해 본다.

전체 데이터 중 1%를 차지하는 소집단은 외국인이고, 온라인 투어가 아니고, 이전에 취소를 한 경험이 있는 집단이다.

전체 데이터 중 7%를 차지하는 소집단은 외국인이고, 온라인 투어에 속하며 특별한 요청을 하지 않았고, 방문날짜까지 약 2주 이상일 때 예약한 집단이다.

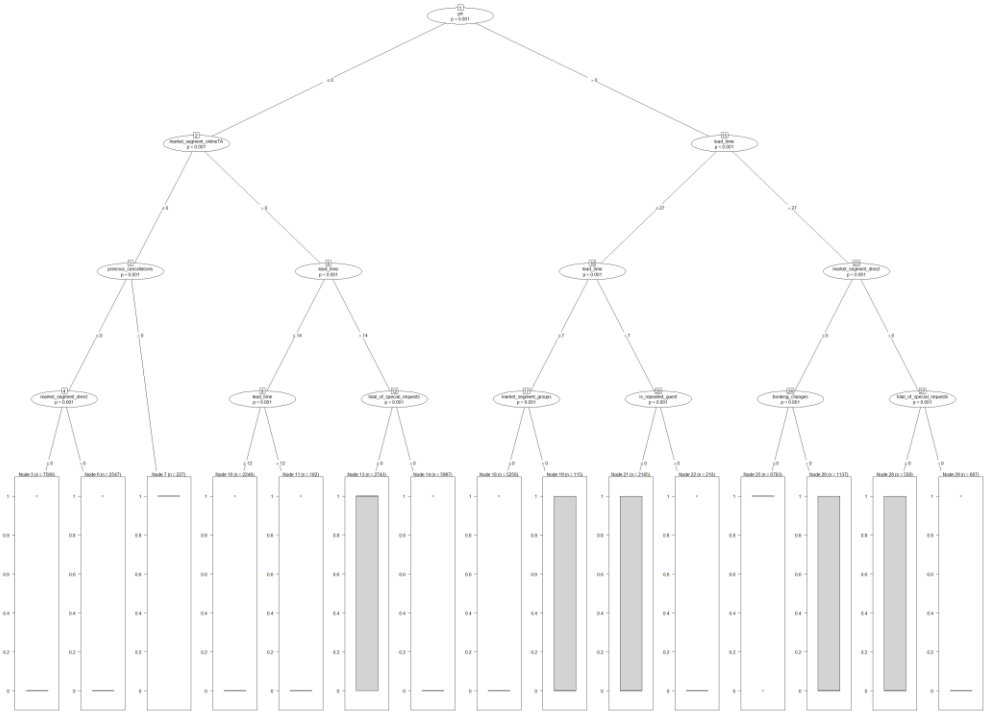
전체 데이터 중 18%를 차지하는 소집단은 자국민이고, 방문날짜까지 약 한 달 이상일 때 예약했고, 직접 예약하지 않았고, 예약변경을 하지 않은 집단이다.

세 집단의 예약취소 패턴의 공통점은 직접 예약하지 않고, 특별한 요청이나 예약변경 등의 호텔의 수고를 필요로 하는 일이 적을수록 예약을 취소하는 경향이 있다.

(2) Party 패키지

R 프로그램에 내장되어 있는 Party 패키지의 ctree 함수를 이용해 의사결정나무를 만들었다. [그림 3]은 유의수준 5%에서 유의성을 중심으로 tree를 만든 결과이다.

[그림 3] ctree 의사결정나무 그래프



rpart에서의 노드 이름과 ctree에서의 노드 이름이 비슷함을 볼 수 있다. 특히, prt, market_segment_online_TA, lead_time, previous_cancellations가 모두 두 tree 내에서 상위 노드에 있다. 하지만 수치기준이 다르다는 점과 분류 결과가 rpart에서와 다르게 형성되었다는 점에서 다른 관점으로 해석할 수 있는 여지가 있다.

각 노드의 라벨은 다음과 같이 해석된다. "P<0.001"이 의미하는 바는 기준값인 0.05보다 유의확률이 작아서 그 노드의 변수를 선택했다는 것이다. Terminal node에 있는 그래프는 노드 내의 데이터 중 예약취소(1)의 비율을 나타낸다.

소분류된 집단 중 예약취소의 특성을 나타낸다고 말할 수 있는 1의 비율이 높은 집단 5개에 대해 가장 왼쪽부터 차례대로 해석하겠다.

첫번째 집단은 외국인이고, 온라인 투어에 참여하고, 방문날짜까지 2주 이상일 때 예약했으며 특별한 요청이 없는 경우로 2743건이 이에 해당한다.

두번째 집단은 자국민이고, 일주일 안에 예약했고, 그룹으로 예약했을 때로 115건이 이에 해당한다.

세 번째 집단은 자국민이고, 일주일~한 달 안에 예약했고, 재방문 고객이 아닌 집단으로 2185건이 이에 해당한다.

네 번째 집단은 자국민이고, 방문날짜까지 한 달 이상일 때 예약했고, 직접 예약하지 않았고, 예약변경을 신청한 집단으로 1137건이 이에 해당한다.

다섯 번째 집단은 자국민이고, 방문날짜까지 한 달 이상일 때 예약했고, 직접 예약했으며 특별요청이 없는 집단으로 558건이 이에 해당한다.

다섯 집단의 예약취소 패턴의 공통점을 살펴보면 자국민이고, 호텔과 직접적으로 연락이 닿지 않는 상태에서 특별한 요청이나 예약변경 등의 호텔의 수고를 필요로 하는 일이 적을수록 예약을 취소하는 경향이 있다. Rpart의 의사결정나무에서와 같은 패턴임이 확인되어 매우 두드러진 특징임을 알 수 있다.

2.3 분석의 타당성

2.3.1 최종 로지스틱 회귀모형을 이용한 분석의 타당성

최종 로지스틱 회귀모형은 2.2.2의 (5)에서 적합한 로지스틱 회귀모형이다. 여러 검정과 진단 방식을 통해 모형의 타당성을 보인다.

편차 적합결여검정(Deviance Goodness-of-Fit Test)에서 귀무가설을 채택하므로 saturated model과 full model이 비슷하며 모형이 적절하다. 피어슨 적합결여검정(Pearson Goodness-of-Fit Test)에서도 귀무가설을 채택하므로 관측치와 평균 사이가 크지 않으며 모형이 적절하다.

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	20805.3981	31E3	0.6814	1.0000
Pearson	21468.1654	31E3	0.7031	1.0000

우도비 검정을 살펴보면 유의수준 5%에서 귀무가설인 $H_0: \beta_1 = \beta_2 = \cdots \beta_3 = 0$을 기각하고 적어도 하나의 회귀계수가 유의하다는 것을 알 수 있다. AIC는 2.2.2의 (2) 로지스틱 회귀모형의 AIC=31794.234, (3) 로지스틱 회귀모형의 AIC=32632.561보다 작은 21085.819이므로 (2)와 (3)의 모형보다 적절하다. R-square는 0.42로 꽤 높은 설명력을 가

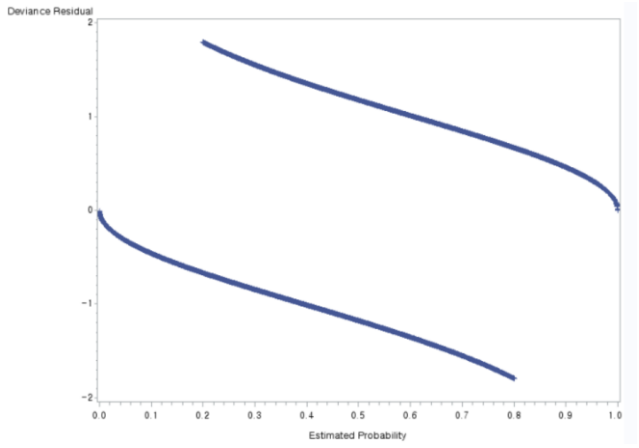
지고 있다.

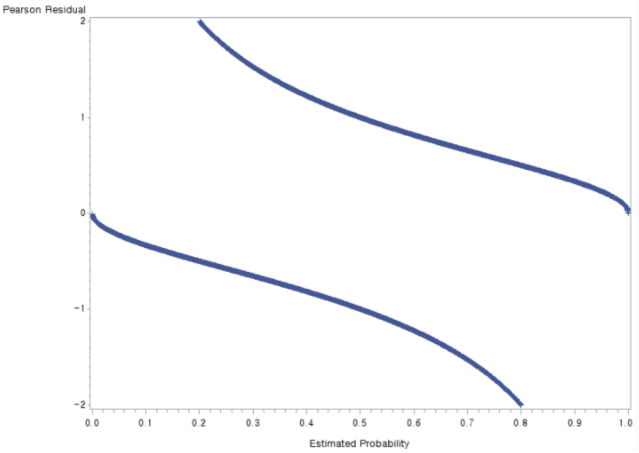
Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	41232.100	21085.819
SC	41240.607	21255.943
-2 Log L	41230.100	21045.819

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	20184.2817	19	<.0001
Score	14813.5526	19	<.0001
Wald	7347.4579	19	<.0001

R-Square	0.4244	Max-rescaled R-Square	0.6274

(2)와 (3)의 편차 잔차(Deviance Residual)와 적합값 그래프와 피어슨 잔차(Pearson Residual) 그래프를 통해 이상치가 잘 제거되었음을 알 수 있다.





위의 모형진단을 통해 19개의 독립변수를 가진 다음 로지스틱 회귀모형은 적절하다는 것을 보일 수 있다.

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = -6.0728 + 0.00940x_{adr} + 0.5419x_{adults} - 0.07x_{3arrival\_date\_monh} - 0.211x_{assigned\_room\_type} - 1.1389x_{booking\_changes} - 4.2149x_{is\_repeated\_guest} + 0.0104x_{lead\_time} + 0.563x_{ms\_complementary} - 0.3135x_{ms\_corporate} - 1.7568x_{ms\_direct} + 1.6239x_{ms\_groups} - 1.4118x_{ms_{offlineTA/TO}} + 0.0493x_{cs\_contract} - 0.7442x_{cs\_group} + 1.5477x_{cs\_transient} + 4.9134x_{previous\_cancellations} + 2.8428x_{prt} - 1.2441x_{total\_of\_special\_requests} + 0.1115x_{stays\_in\_nights}$$

2.3.2 의사결정나무를 이용한 분석의 타당성

rpart 함수를 이용해 만든 의사결정 나무와 ctree 함수를 이용해 만든 의사결정 나무는 가정과 노드의 분리 기준이 다르기 때문에 어느 모형이 더 나은지를 섣불리 판단할 수 없다. 또한, 의사결정나무의 특성상 분석 방식이 매우 다양하고 따라서 결과도 다양하여 회귀분석과 같이 적합성을 평가할 수 있는 공식적인 기준이 없다.

이에 정확성(Accuracy) 지표를 이용해 간접적으로 의사결정나무의 타당성에 대해 논하겠다. 호텔 예약 데이터 전체를 이용해 rpart 의사결정나무 모델과 ctree 의사결정나무 모델을 만든다. 각각의 모델에 대해 호텔 예약 데이터 전체를 예측해 보았을 때 Rpart 의사결정나무 모델의 정확성은 약 82.5%, Ctree 의사결정나무 모델의 정확성은 약 82.5%으로 나왔다. 높은 정확성을 근거로 본 연구에서 사용한 두 의사결정나무 모델을 통한 분석이 신빙성이 있다고 말할 수 있다.

3. 결론

3.1 분석 결과 요약

예약취소(1), 예약수립(0)인 이항변수를 타겟변수로 두어 독립변수 19개에 대해 로지스틱 회귀모형을 적합했다. 이상치 제거, 적합결여 검정, 왈드 검정 등을 통해 타당한 분석임을 증명했다. 로지스틱 회귀 결과 회귀계수 중 양수이며 절대값이 비교적 큰 변수들은 prt, previous_cancellation, customer_type_transient이다. 즉, 과거의 예약취소가 많을수록, 외국인에 비해 자국민이, 한시적으로 머무르는 무리보다 한시적으로 머무르는 개인이 더 예약취소를 할 확률이 높다. 또한, 회귀계수 중 음수이며 절대값이 비교적 큰 변수들은

is_repeated_guests, booking_changes, total_of_special requests, markget_segment_direct이다. 결과적으로 이전에 방문하지 않았던 고객이고, 예약변경이나 특별한 요구가 적을수록, 직접 호텔을 이용하는 것에 비해 온라인 투어가 더 예약취소를 할 확률이 높다.

Rpart함수를 이용한 의사결정나무와 ctree 함수를 이용한 의사결정나무에서도 로지스틱 회귀분석의 결과와 마찬가지로 직접 호텔을 이용하지 않고, 특별한 요청이나 예약변경 등의 호텔의 수고를 필요로 하는 일이 적을수록 예약을 취소하는 패턴을 발견하였다.

따라서 자국민, 이전 취소경험, 고객의 유형, 재방문 여부, 예약변경개수, 특별요청개수, 고객의 구분에 주의하여 호텔의 예약취소 방지 대책을 세우는 것이 좋으며 초과예약 예측 모델을 만들 때 참고할 만한 예측인자가 될 것이다.

3.2 분석의 장점 및 한계점 설명

예약취소(1)와 예약수립(0)이라는 이항변수를 설명하는데 로지스틱 회귀모형은 최적화되어 있다. 독립 변수가 타겟변수에 얼마만큼의 영향을 주는지 수치를 통해 정확히 확인할 수 있어 해석이 용이하고 여러 진단과 검정방법을 통해 모형의 적절성을 판단할 수 있었다. 하지만, 본 연구에서 사용한

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

모형은 선형성만을 다루기 때문에 선형이 아닌 독립변수에 대해서 적절한 분석이 되지 못한다. 또한, 두 독립변수 사이의 교호작용을 고려하지 않았기 때문에 이에 대한 추가적인 연구가 필요하다.

의사결정나무는 선형성, 정규성 등의 가정이 필요하지 않아 모든 변수의 선형성 가정이 만족되지 않는 호텔 예약 데이터를 분석하기에 적합했다. 의사결정나무의 결과 해석이 쉽고, 로지스틱 회귀모형에서 부족했던 교호작용에 대해 살펴볼 수 있다는 장점이 있었다. 하지만 연속형 변수를 비연속적인 값으로 취급하여 분리의 경계점에서 오류가 클 가능성과 각 변수의 타겟변수에 대한 수치적 영향력을 알 수 없다는 점이 한계점으로 인식된다.

4. 참고문헌

데이터 출처 사이트 : https://www.kaggle.com/jessemostipak/hotel-booking-demand

반종삼(2008), "휴양콘도미니엄 객실 가동율에 영향을 미치는 예약취소 및 NO-SHOW에 관한 연구-H리조트 운영사례를 중심으로"

천송이(2013), "리조트 예약부도 방지를 위한 연구-의사결정나무 분석과 로지스틱 회귀분석의 적용을 중심으로"