「정보시스템연구」 제30권 제2호
한국정보시스템학회
2021년 6월, pp. 105~126

http://dx.doi.org/10.5859/KAIS.2021.30.2.105

머신러닝을 이용한 국내 수입 자동차 구매 해약 예측 모델 연구: H 수입차 딜러사 대상으로

정동균* · 이종화** · 이현규***

〈목 차〉

Ⅰ. 서 론

1955년 8월 국내 생산 1호 차인 시발(始發) 자동차가 생산되었다. 맨 처음 출발의 의미로 '시발'을 사용하였으며 흙으로 만든 틀에 쇳물을 붓고 주물에 일일이 손으로 구멍을 뚫어가면 엔진을 만들었다고 한다(산업통상자원부, 2020). 1980년대 TV에서 방영된 '전격 Z 작전'은 자동차 '키트'는 주인공과 대화를 하고 또 부르면 혼자 달려오기도 하는 등 지금으로 보면 인공지능, 자율주행 자동차와 흡사한 모습을 보였다. 인공지능 자율주행 기술, 운전자 성향을 맞춘 AI 자동차, 자율주행센서로 탑승자를 보호하는 기술 등 자동차 '키트'는 더 이상 미래의 자동차가 아니다.

자동차 산업은 수만 개의 부품을 만드는 수천 개의 하청 업체와의 수직계열화를 이루며 존재하는 거대한 기간산업이다(김태기, 안창순, 1997). 그러나 현재의 자동차 산업을 들여다보면 이 같은 수직계열의 기본 틀이 완전히 바뀌고 있는 것을 알 수 있다. 3만여 개의 부품으로 구성된 내연 기관 자동차가 전기 자동차로 생산으로 된다면 2만여 개의 부품 모듈로 구성되며 배터리나 전자 제어 장치 등이 독립된 모듈로 되어있다(Hwang, 2019). 이 모듈들을 모아 조립할 수 있는 정도의 기술만 있다면 전기차를 생성할 수 있다. 그 예로 2016년 3월 창업한 중국의 바이톤(Byton)은 2018년 1월 전기자동

* 부경대학교 경영컨설팅 협동과정, jdk1204@gmail.com(주저자)
** 동의대학교 상경대학 정보경영학부, jhlee6050@deu.ac.kr(공동저자)
*** 부경대학교 경영대학 경영학부, hyunqlee@pknu.ac.kr(교신저자)

차를 생산하였으며 업계에서도 완성도 높은 제품으로 평가받았다(KAICA, 2019; 채영석, 2018).

GM이나 다임러, 도요타 등 대표되는 기존 자동차 제조 산업은 완성차 업체 아래에 피라미드 구조의 수많은 부품 업체로 구성되어 있어 신생 업체의 진입이 어려운 즉, 진입장벽이 높은 산업이다. 하지만 2010년 테슬라의 등장으로 기존 자동차 업계에 큰 변화가 생겼다. 내연 기관 자동차 산업을 전기 모터로 대체된 것이다. 테슬라뿐만 아니라 중국의 바이톤을 비롯한 신규 기업이 자동차 산업에 진출하고 있다. 자동차 산업 구조가 바뀌어 가는 것과 동시에 자동차의 모습만 변화하는 것이 아니라 자동차를 이용하는 목적까지도 달라지고 있다. 테슬라는 자동차 안에서 어떻게 시간을 보내는지를 고려하여 차량용 게임을 시장에 내놓고 있으며 구글, 애플 등의 글로벌 IT 기업 또한 변화하는 시대의 패권을 잡기 위해 공룡 소프트웨어 회사들의 경쟁이 치열하다(박남규 외, 2019). 인텔(inter)이나 엔비디아(nvidia) 같은 반도체 업체나 LG화학 같은 배터리 업체, 소프트뱅크, KT 같은 IT 업체에도 자동차로 인한 새로운 기회의 장이 열리게 될 것으로 본다.

자동차는 정보통신기술과 인공지능 기반의 자율주행차, 친환경 전기 및 수소차로 개념을 확대하고 있으며 스마트폰과 O2O 플랫폼 기반의 공유 이동 수단 서비스까지 산업을 확장하고 있다(문화체육관광부, 2020). 또한, 자동차 산업에 ICT 기술로 무장하여 고가의 고급 자동차 브랜드화를 생성시키고 있다. 삼정 KPMG 경제연구원에 따르면 2019년 3월 수입 자동차 대수는 18,078대였지만 2020년 3월수입 자동차 등록 대수는 20,304대로 12%이상 증가한 것으로 나타났다. 글로벌 환경에서의 자동차 수입은 코로나19 여파에도 불구하고 가솔린의 상승세와 전기자동차의 폭발적 인기로 수입차 수요 상승세가 이어지고 있다(삼정KPMG경제원구원, 2020).

고가의 수입차는 재고 부족으로 차량 계약 후 출고 전까지 긴 대기 기간 동안의 해약 발생 가능성이 상대적으로 높다. 해약 가능성이 높은 고객을 집중 관리할 수 있는 기회를 만들고자 본 연구를 기획한다.

머신러닝 기법의 해약 예측은 보험회사의 보험 유지 고객과 이탈 고객 데이터를 활용한 고객 이탈 예측 모델 구현(서광규, 2005), 실제 증권사의 온라인 대 고객 마케팅 CRM에 활용 중인 고객 이탈 예측 및 원인 규명 연구(나광택 외, 2020), 국내 전 업계 대기업 카드사 데이터를 이용하여 6개월 이상 사용하지 않은 휴면 고객 예측 모형 연구(이동규, 신민수, 2018), 통신서비스 이용고객의 행동 패턴을 분석하고 이탈을 예측할 수 있는 이탈 방지 시스템 구현(김상휘 외, 2020)과 같이 보험, 증권, 카드, 통신 서비스 등 다양한 산업 분야에서 고객 이탈을 예측하고 원인 규명을 함으로써 기업의 손실의 줄이고 이익을 극대화하고 있다(이동규 외 2018). 이와 같이 기존 연구는 이미 비용을 지불하고 서비스를 이용 중인 고객이 대상이고 서비스나 상품의 만족도, 고객의 상황에 따라 이탈 또는 해지에 대한 예측 모형으로 구현되었다. 그러나 본 연구는 초기 구매 계약 이후 상품이나 서비스를 이용 전 대기 기간 동안 고객 변심으로 인한 해약 발생 가능성을 머신러닝 기법을 활용하여 예측할 수 있는 방안을 마

련하고 해약 가능성이 높은 고객을 집중 관리할 수 있는 기회를 영업 및 마케팅 담당자에게 제공함으로써 해약을 최소화하고 매출 향상에 기여하고자 한다.

본 연구는 실증 연구로 독일 Mercedes-benz를 판매하고 있는 H 수입차 딜러사에서 범용으로 사용하고 있는 영업지원 시스템(Sales force automation, SFA)의 누적된 계약, 해약, 매출 데이터를 기반으로 머신러닝 기법 중 로지스틱 회귀, 서포트 벡터 머신, 랜덤 포레스트, 가우시안 나이브 베이즈, 인공신경망 모델에 적용하여 각 모델의 성능을 비교함으로써 최적의 해약 예측률을 돌출할 수 있는 모델을 찾고 해약 여부에 영향을 미치는 주요한 변수를 파악하여 계약 고객관리 및 영업, 마케팅 전략 수립에 필요한 자료로 활용하고자 한다. 또한, 성능이 우수한 머신러닝 모델을 기반으로 수입차 딜러뿐만 아니라 국내 자동차 시장에도 적용할 수 있는 해약 예측 모형의 기반이 될 것으로 기대된다.

본 논문은 2장에서 국내 수입자동차 시장의 비즈니스 프로세스와 방법론에 대하여 설명하고 3장에서 본 연구의 프레임 워크와 분석자료, 종속 변수에 영향을 미치는 독립 변수를 파악한 후 4장에서는 각각의 머신러닝 모델에 적용하여 예측 성능을 비교한다. 5장에서는 결론과 시사점 그리고 본 연구의 한계점과 향후 연구 방향을 제시한다.

Ⅱ. 이론적 배경

2.1 수입자동차 계약 분석

1987년 1월 수입차 개방에 따라 한성 자동차(벤츠), 효성 물산(아우디/폭스바겐), 한진(볼보), 코오롱 상사(BMW) 등이 수입자 딜러로 판매를 시작되었다. 독일, 미국, 일본, 스웨덴, 이탈리아, 프랑스, 영국 등 세계 26개 브랜드가 국내 진출하여 매년 수입 및 판매가 증가하여 2012년에 10%의 국내 수입차 점유율을 보이면서, 2017년 15%, 2020년 기준 국내 자동차 시장의 18%를 차지하고 있다(KAIDA, 2019).

국산차의 경우 국내에서 조립, 생산, 판매로 일부 신규 모델 출시에 따른 예약 판매를 제외하면 판매 대상 차량의 재고 수급이 원활한 반면, 고가의 수입차는 재고 보유 및 관리 비용의 문제로 쇼룸 전시차를 포함하여 일부 재고만 보유하고 있어 고객이 원하는 모델, 색상, 옵션을 계약 후 인도까지 상당한 시일이 소요되고 있는 상황이다.

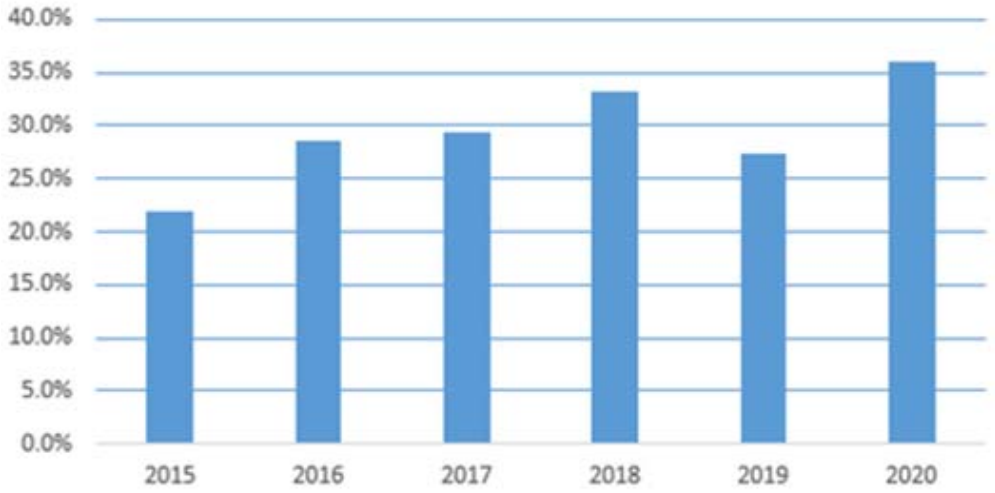
<표 1> 2015년~2020년 국내 H 수입차 딜러사 출고 및 해약 대기 기간 및 대수

대기 기간	출고차량		해약차량		합계
	대수	비율(%)	대수	비율(%)	
1개월 이상 ~ 3개월 미만	3,495	72.3%	3,081	67.8%	6,576
3개월 이상 ~ 6개월 미만	902	18.7%	879	19.3%	1,781
6개월 이상 ~ 9개월 미만	277	5.7%	304	6.7%	581
9개월 이상 ~ 1년 미만	91	1.9%	127	2.8%	218
1년 이상	69	1.4%	152	3.3%	221
합계	4,834	100%	4,543	100%	9,377

<표 1>은 수입차 딜러사가 2015년부터 2020년까지 기간별 출고 및 해약 실적 통계를 나타낸 표이다. 1개월에서 6개월 미만은 해약차량보다 출고차량이 근사하게 높은 것을 나타났지만 6개월 이상의 대기 기간을 갖는 고객이 해약률이 높이지고 있음을 확인 할 수 있다.

<그림 1>의 전체 출고 건 대비 해약률을 보면 2019년 코로나로 인하여 계약, 매출 감소로 해약률이 감소한 것을 제외하면 매년 지속적으로 증가하고 있으며 매출이 증가하면서 계약 및 해약도 증가하겠지만 2020년에는 평균 10대의 계약 중 3.5대의 해약 발생으로 매출에 큰 영향을 미치고 있다.

<표 2>는 계약 차량 출고 후 차량의 고장이나 서비스의 불만 등으로 해약이 발생하는 경우도 있겠지만 계약 후 출고 전 인도일(대기기



<그림 1> 2015년~2020년 국내 H 수입차 딜러사 연간 해약률

<표 2> 2015년~2020년 국내 H 수입차 딜러사 해약 사유 및 대수

해약사유	대수	비율(%)
인도일(대기기간)	1,393	30.7%
개인사유	1,235	27.2%
타 브랜드, 타딜러 선택	854	18.8%
판매조건	398	8.8%
기타	291	6.3%
신용등급	203	4.5%
금융조건(할부, 리스)	113	2.5%
차량품질	27	0.6%
구매보류	20	0.4%
세일즈맨	9	0.2%
합계	4,543	100.0%

간) 지연에 따른 해약이 30.7%를 차지하고 있다. 고가 수입차의 재고 부족으로 인한 차량 계약 후 출고 전까지, 긴 대기 기간 동안의 해약 발생 가능성을 예측하여 해약 가능성이 높은 고객을 집중 관리할 수 있는 기회를 제공함으로써 해약을 최소화하는 방안이 필요해 보인다.

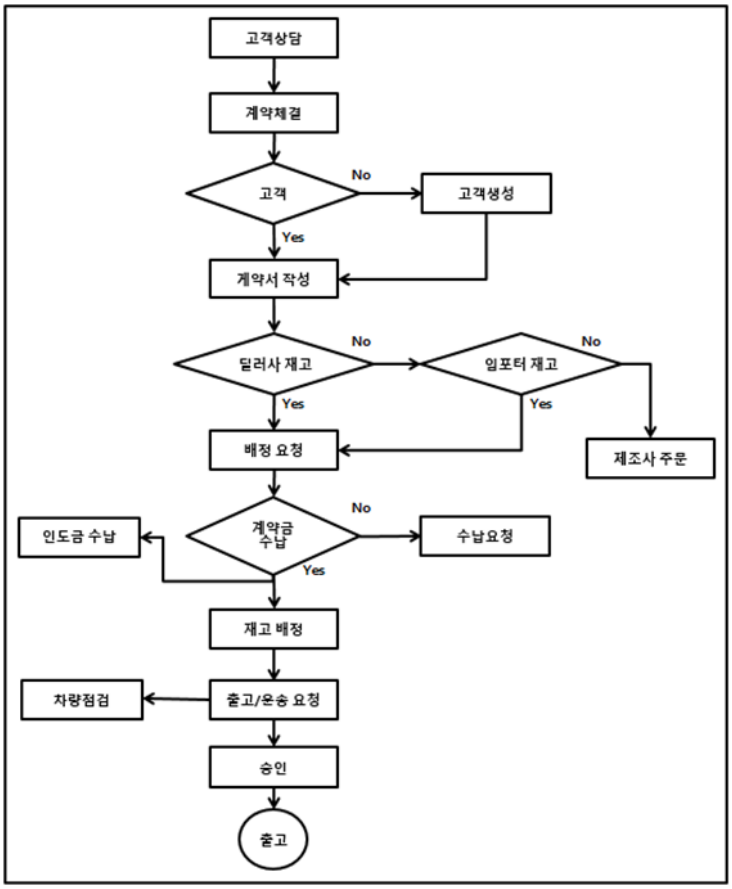
2.2 영업지원시스템(Sales Force Automation, SFA)

영업지원시스템는 회사 내 영업조직의 자동화 시스템으로 개별 영업직원의 동선과 고객 상담의 진행 상황 등을 수집하고, 이를 정보로서 축적·관리해 실질적인 매출로 이어지도록 돕는 것을 주목으로 미국에서 기존의 '사무 자동화(OA, Office Automation)'를 대체하는 모델로 개발되어 인터넷-모바일 등 온라인 통신 인프라의 발전 속에 빠르게 확산되었다(SalesForce.com, 2020).

기업의 전방위 고객 접점 중 영업 및 마케팅과 관련된 각 주체들의 활동 및 업무수행의 효율성과 생산성을 극대화시키며 적극적인 고객관계 유지, 향상과 고객의 가치증대에 그 초점을 두고 있는 기업전략의 일환으로서 이미 많은 구매 기업들에서 실행전략으로 SFA 시스템을 도입, 활용되고 있다(김미정, 2002).

SFA 시스템을 도입하여 실천한 서비스 특정 기업에서는 고객관리, 영업보고, 일정관리, 매출부분, 재무관리 등의 분야에서 50~100% 개선율을 나타냈고 영업 인력의 매출이 12% 정도 향상되었으며 영업조직 간의 정보공유가 용이해지는 등의 효과가 있었다(김태인, 2013). 영업의 성패를 좌우하는 영업 사원의 이동성(Mobility)을 보장해 주기 위해 도입된 모바일 SFA 시스템은 현장에서 실시간으로 업무 처리를 가능하게 함으로써 기업의 영업력 제고에 기여하고 있다(박기호 외, 2003). 국내 모 은행은 고객과의 강력한 상호 작용을 바탕으로 금융 마케팅 및 효과적인 영업 활동을 통하여 고객가치를 높이고 경쟁적 우위 확보 및 기업 이익 극대화를 위한 전략적 방안으로 실시간 고객정보 관리, 대출 영업 인력의 활동 정보 및 실적 분석 관리를 통한 영업력 강화를 위해 금융 SFA을 도입하였다(김미정, 2002). HDC현대 산업 개발은 고객 중심 경영 환경 구축을 위해 세일즈포스 모바일 기반의 실시간 고객 연결 통합 관리 플랫폼을 도입하여 계약, 입주, 하자접수 등의 고객 서비스를 원 스탑으로 처리함으로써 고객 서비스 업무 경험 및 생산성 향상을 도모하였으며 이러한 축적된 데이터를 가공, 분석하여 최적의 의사결정을 위한 환경을 구축하였다(SalesForce.com, 2020).

<그림 2>은 연구 대상 기업인 국내 H 수입차 딜러사는 실시간 재고 조회 및 관리, 고객 관리, 영업 관리, 매출 관리를 위해 2010년 영업지원 시스템(SFA)을 도입하여 영업 지원부서 및 영업점 간에 실시간(Real time) 재고 확인, 계약, 수납, 출고, 차량등록사업소 신차 등록까지 원 스탑으로 업무를 처리하고 있다. 최근에는 반응형 웹기반의 SFA 시스템으로 리뷰얼함으로써 스마트폰이나 테블릿 PC와 같은 다양한 모바일 기기를 이용하여 장소에 구애받지 않고 영업 업무를 진행하고 있다.



<그림 2> H 수입차 딜러사 영업지원 시스템 업무 프로세스

2.3 머신러닝 기반 예측 모델 및 성능 평가

2.3.1 머신러닝 기반 예측 모델

머신러닝 모델 및 데이터를 이용한 예측은 경제, 의료, 교육, 농업 등 각종 산업의 다양한 분야에서 과거 축적된 데이터를 머신러닝 모델을 적용하여 예측된 결과를 주요 의사결정에 활용하고 있다. 예를 들어, 미래 주가의 등락률을 예측하고자 머신러닝 기법 중 합성곱 신경망(Convolutional neural network, CNN) 딥러닝 모델에 차트 이미지 데이터 적용 및 주가패턴 학습을 통하여 매매 빈도 패턴과 예측 시점을 다른 주가 등락 예측 연구들보다 정확도를 향상 시켰다.(송현정, 이석준, 2018). 우리나라의 주요 사망 원인 2위에 해당하는 심혈관 질환 중 허혈성 심장질환에 대하여 실제 병원에서 의료 전문가 및 의사가 작성한 초기 기록의 데이터 셋을 머신러닝의 다층 퍼셉트론, 나이브 베이즈, 서포트 벡터 머신 모델에 학습시켜 확진 환자를 80% 예측함으로써 병원에서 불필요한 검사를 시행할 필요성과 그 경우를 줄이고,

좀 더 효율적으로 진단하는데 기여할 수 있다(박평우 외 2018). 김영식과 김훈호(2019)은 현재 지속적으로 논의 중인 학생들의 사교육 참여 및 지출에 대하여 주요 사교육 참여 변수들을 예측하고 교육학 연구에 머신러닝 기법을 활용하고자 랜덤 포레스트, 나이브 베이즈, 서포트 벡터 머신, 인공신경망 모형에 적용하여 모형간의 예측성과를 비교 분석함으로써 그 활용 가능성을 확인하였다. 이외에 농축산업 분야에서도 머신러닝을 활용하여 농산물 생육에 중

<표 3> 머신러닝 모델 응용 예시

머신러닝 모델	연구자	연구내용
서포트 벡터 머신 로지스틱 회귀 인공신경망	서광규 (2005)	A 보험회사의 2002~2003년의 12개월 보유 고객 데이터베이스 활용하여 고객 이탈 분석 및 예측 모형 구현
결정 트리	임세현과 허연 (2006)	2003년~2004년 온라인 자동차 보험 계약 고객 데이터 기반 고객 이탈 예측
로짓 모형, 인공신경망, 결정 트리,앙상블 모델, k-최근접 이웃	이민수과 최영찬 (2009)	양돈관리프로그램인 Pigplan을 통해 축적된 데이터 셋으로 양돈농장의 모돈의 산차별 생산성 예측.
합성곱 신경망	송현정과 이석준 (2018)	주식 차트 이미지 데이터를 이용하여 주가패턴을 학습하여 미래 주가 등락률 예측.
로지스틱 회귀 서포트 벡터 머신 랜덤 포레스트	이동규 외(2018)	국내 전 업계 대기업 카드사의 6개월간 무실적 휴먼 고객 데이터 기반 휴먼 예측 고객 분석
다층 퍼셉트론, 나이브 베이즈, 서포트 벡터 머신	박평우 외(2018)	심혈관 질환 중 허혈성 심장질환에 대하여 실제 병원에서 의료 전문가 및 의사가 작성한 초기 기록의 데이터 셋으로 확진 환자를 80% 예측
랜덤 포레스트, 나이브 베이즈, 인공신경망, 서포트 벡터 머신	김영식과 김훈호 (2019)	학생들의 사교육 참여 및 지출에 대하여 주요 사교육 참여 요인 예측.
합성곱 신경망 순한 신경망	최준영 외(2019)	딥러닝 기반 미세먼지 측정소 간의 관계와 미세먼지와 기상 데이터의 관계를 학습하여 미세먼지 농도 예측 모델 제시
랜덤 포레스트	김준석 외(2020)	기상청의 GloSea5(Global seasnoal forecast system version 5)의 기후 전망 데이터 셋을 이용하여 장기 농업기상정보 1년의 일평균기온을 예측.
랜덤 포레스트 서포트 벡터 머신	나광택 외(2020)	실제 증권사의 온라인 대 고객 마케팅 CRM에 활용 중인 고객에 대한 이탈 예측 모델링 및 원인 규명
선형 회귀 랜덤 포레스트 에이다부스트	이인지과 윤현식 외(2020)	지역축제 관련 관측 변수 데이터를 이용하여 방문객 수 예측 모형 개발 및 특성 변수들의 영향력 비교
로지스틱 회귀 랜덤 포레스트 XGboost	김상휘 외(2020)	통신 서비스 이용고객의 행동 패턴을 분석하여 고객 이탈을 예측할 수 있는 이탈 방지 시스템 구현

요한 날씨를 예측하고 축산 농가에서는 가축의 생산력을 예측하는 연구가 이루어지고 있다. 농업과 관련된 기상 정보는 기상청의 GloSea5 (Global seasnoal forecast system version 5)에서 기후 전망을 제공하고 있지만 거리에 따른 상관성이 고려된 기계학습법인 공간랜덤포레스트(Spatial random forest)를 적용하여 장기 농업기상정보 1년의 일평균기온을 예측하였다 (김준석외, 2020). 이민수과 최영찬(2009)은 머신러닝 방법론을 사용하여 축산농가의 의사결정지원에 활용할 수 있는 방안을 제시하고자 양돈관리프로그램인 Pigplan을 통해 축척된 자료를 로짓(Logit)모형, 인공신경망, 의사결정나무트리(Decision Tree, DT), 앙상블(Ensemble), k-최근접 이웃(k-Nearest Neighbor, KNN) 모델에 적용하여 예측 모델을 비교 분석함으로써 효율적으로 양돈농장의 모돈의 산차별 생산성 예측 성능을 증명하였다.

본 연구는 머신러닝 모델 중 로지스틱 회귀, 서포트 벡터 머신, 랜덤 포레스트, 가우시안 나이브 베이즈, 인공신경망 모델에 대하여 국내 H 수입자동차 딜러사의 영업지원시스템에 축척된 데이터를 각 모델에 적용 및 학습함으로써 해약 예측 모형을 개발하고 각 모델에 대한 성능을 비교, 평가하여 최적의 자동차 계약에 대한 해약 예측 모형을 구현하고자 한다.

(1) 로지스틱 회귀(Logistic Regression, LR)

로지스틱 회귀는 독립 변수의 선형결합을 이용하여 사건의 발생 가능성을 예측하는데 사용하는 통계기법이다(Cox, 1958). 종속 변수 1개와 여러 독립변수와의 상간 관계를 분석하여 종속변수를 예측하는 모형으로 종속 변수가 범주형 데이터이고 0 또는 1인 경우 사용되는 선형 회귀분석이다. 어떤 값의 독립변수가 주어지더라도 종속 변수가 1의 범주에 포함하는 확률을 예측한다.

하나의 종속 변수와 한 개 이상의 독립변수 사이의 인과관계를 표현하기 위해 가장 잘 적합되고 모수의 수를 절약한 모형을 찾아내는데 궁극적인 목표를 둔다. 수식으로 표현하면 아래와 같다(이장택, 조현식, 2009).

$$P(y=1|x_1,\cdots x_p)=\frac{\exp(\beta_0+\sum_{i=1}^{p}\beta_i x_i)}{1+\exp(\beta_0+\sum_{i=1}^{p}\beta_i x_i)}$$

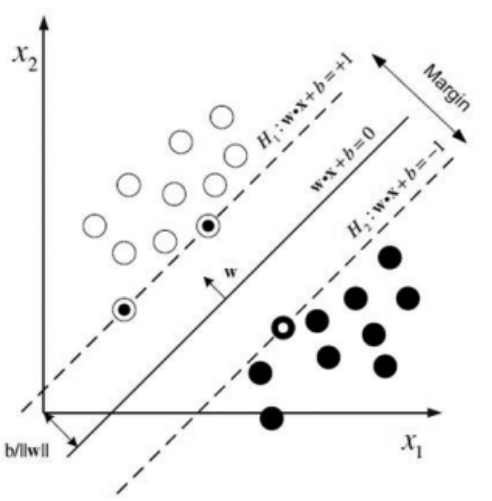
종속변수 : y
독립변수 : $x_1,\cdots,x_p$
추정될 회계 계수 : $\beta_0,\beta_1,\cdots,\beta_p$
로지스틱 회귀모형 로짓 :

$$z=\beta_0+\beta_1 x_1+\cdots+\beta_p x_p$$

(2) 서포트 벡터 머신(Support Vector Machine, SVM)

서포트 벡터 머신은 두 그룹의 분류 문제를 해결하기 위한 새로운 머신러닝 모델로 처음에는 서포트 벡터 네트워크(Support-vector network)로 3가지 아이디어로 제안되었다. 첫째, 두 그룹의 데이터 점들을 분류를 위한 해결책으로 최적의 초평면(Optimal hyperplanes) 찾는 것이고 둘째, 선형분류가 불가능한 데이터 셋에 대하여 비선형분류를 위한 내적연산(Convolution of dot-product), 마지막으로는 소프트 마진(Soft margins)으로 분류 오류의 허용률을 이용한 머신러닝 기법이다(Cortes et al, 1995). <그림 3>과 같이 두 그룹을 분류하기 위한 최적의 분리 초평면(Optimal separating hyperplanes)은 각 그룹의 데이터 점과 평형 분

류 평면과 거리(Margin)가 최대가 되는 평면으로 분리된다(Vapnik, 1996).



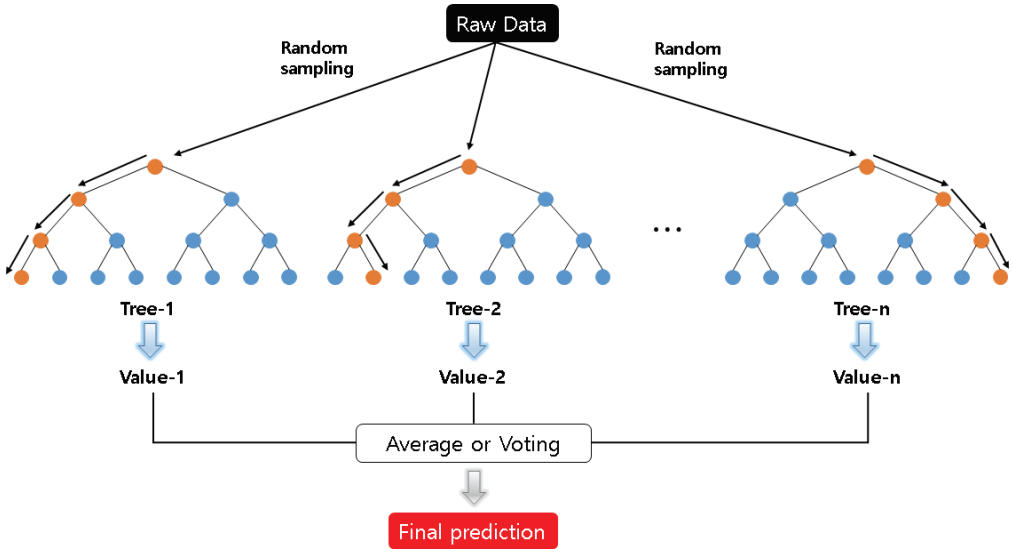
<그림 3> 최적 선형 분리 초평면 및 마진

$H_1 : wx + b = +1$ (1)
$wx + b = 0$ (2)
$H_2 : wx + b = -1$ (3)
w : 분류 경계면과 수직인 법선벡터
b : w와 원점간의 거리
$wx + b$: 두 그룹간 분류 경계면

식(2)와 같이 두 그룹간의 경계평면을 0으로 분류하고 식(2)를 평형으로 플러스 이동 시킨 식(1)과 마이너스 평형으로 이동시킨 식(3)과의 거리를 마진으로, 서포트 벡터 머신은 마진을 최대화로 하는 분류경계 평면을 찾는 기법이다(정성훈, 진창하, 2020).

(3) 랜덤 포레스트(Random Forest, RF)

랜덤 포레스트 기법은 독립적인 결정 트리(Decision Tree)의 결합하여 결정 트리의 과적합(Overfitting)을 개선한 앙상블 분류 예측 기법으로 데이터 셋을 임으로 k개로 분리된 결정 트리 정하고, 독립변수들을 임으로 샘플링하여 각각의 결정 트리에서 예측력이 높은 독립 변수를 기준으로 하위 노드로 분할하고 하위의 각 노드 또한 이와 같은 방법으로 최하위 노드까지 분할함으로서 여러 결정 트리의 앙상블 머신러닝 기법이다(Breiman L, 2001). 일반적으로 변수의 개수가 m개이면 각 분할에서 랜덤으로 m/3개의 변수를 선택하여 결정트리가 만



<그림 4> 랜덤 포레스트(Random Forest 기본 컨셉(김종성 외, 2019)

들어지고 데이터와 변수를 샘플링하여 서로 조금씩 다른 트리들로 구성되어 각 트리들의 예측값은 과적합(Overfitting) 문제가 발생하지 않고 간편하고 빠른 학습과정에도 높은 정확도와 일반화 할 수 있는 성능을 보유한 기법으로 램덤 포레스트의 기본적인 개념은 <그림 4>와 같다(김종성 외, 2019).

(4) 가우시안 나이브 베이즈(Gaussian Naïve Bayes, GNB)

나이브 베이즈는 베이즈 정리을 적용한 쉽고 간단한 확률 분류기로 각 속성의 변수를 독립변수로 간주하고 이미 일어난 사전 확률에 대하여 사후 확률을 추론하여 예측하는 조건부 확률적 기법이다. 레이블링된 지도학습(Supervised learning)에서 효과적으로 학습되어 복잡한 실제 상황에서도 활용될 수 있다(Kamel, et al, 1995).

$$p(C_k|x_1,\cdots,x_n) \quad (1)$$
C_k : k개의 가능한 확률적 결과
$x=(x_1,\cdots,x_n)$: n개의 독립 변수에 대한 벡터값

식(1)을 조건부 확률을 반복적으로 적용하고 독립성을 가정하면 Ck의 조건부분포는 아래 식(2)와 같다.

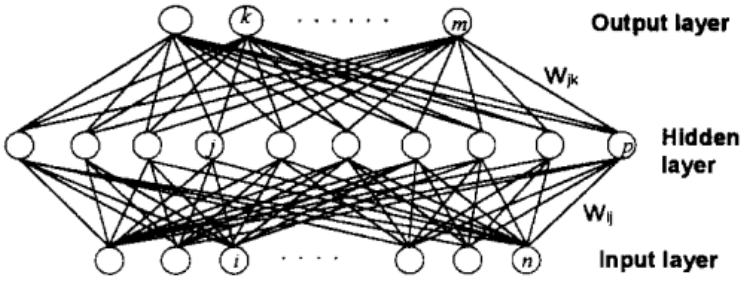
$$p(C_k|x_1,\cdots,x_n)=\frac{1}{Z}p(C_k)\prod_{i=1}^{n}p(x_i|C_k) \quad (2)$$
$Z=p(x)$: 특성값들의 상수

연속적인 데이터를 처리할 경우 모든 클래스와 관련된 연속 값이 가우스 분포(Gaussian distribution)에 따라 분산된다는 가정 하에 훈련 데이터 셋을 클래스 별로 분할하고 모든 클래스의 평균과 표준편자를 계산하여 연속 데이터 셋의 확률을 추정하고자 가우시안 나이브 베이즈 기법을 활용하고 있으며 관련 식(3)은 다음과 같다(Kamel, et al, 1995).

$$P(X=x|C=c)=\frac{1}{\sqrt{\pi\sigma}}e\frac{-(x-\mu)^2}{2\sigma^2} \quad (3)$$
x : 변수, c : 클래스, μ : 평균, σ : 표준편자

(5) 인공신경망(Artificial Neural Network, ANN)

인공신경망은 서로 연결되어 있는 뉴런(Neuron)과 뉴런이 정보를 주고받는 사람이나 동물의 신경계 구조를 기반한 수학적 모형으로 1958년 Rosenblatt의 퍼셉트론(Perceptron) 학습 규칙 개념의 등장에 따라 인공지능 개발 분야에서 큰 기대와 주목을 받았고 이 후 1969년 Minsly and Papert는 퍼셉트론이 간단한 XOR 분류를 수행할 수 없음을 수학적으로 증명되어 그 한계가 밝혀졌다. 하지만, 1986년 Rumelhart et al.이 은닉층을 가진 다층 퍼셉트론을 학습할 수 있는 방법으로 역전파 알고리즘(Backpropagation algorithm) 소개함으로써 인공 신경망에 대한 다양한 연구와 적용이 발전되어 왔다(정성훈, 진창하, 2020). 다층 퍼셉트론 신경망은 입력층(Input layer), 한 개 이상의 은닉층(Hidden layer), 출력층(Output layer) 구성되고 입력층과 출력층의 입력 유닛(Input unit)과 출력 유닛(Output unit)은 각각 분류 파라메터 값과 결과 기댓값을 처리하고 입력층과 출력층 중간의 은닉층은 각 유닛의 입력과 출력 특성을 비선형으로 구현함으로서 신경망의 성능을 향상 시키고 있다.



<그림 5> 역전파 퍼셉트론 일반적인 구조(이석호 외, 1996)

<그림 5>는 은닉층 한 개로 구성된 3층(Three layer) 퍼셉트론 신경망으로 입력층의 입력 유닛에 특성 변수의 입력 신호를 받으면 입력 신호에 따라 가중치(w_i)함께 은닉층의 각 뉴런으로 전달되고 이 신호는 다시 은닉층의 가중치(w_{ij})와 함께 출력측의 뉴런으로 전달(FeedForward)되면 출력층 뉴런의 결과값과 목표값(Target value)을 비교한 후 그 차이(Error)를 감소시키는 연결강도(w_{ik}, w_{ij})를 조정하여 역으로 은닉층 과 입력층으로 전달하여 가중치를 미세조정(편미분)하여 최적의 가중치로 결과값을 학습하는 기법이다(이석호 외, 1996).

2.3.2 머신러닝 모델 예측 성능 평가

머신러닝 모델의 예측 성능 평가는 오차행렬(Confusion matrix) 기반의 이진 분류 결과표를 활용한 성능 지표들로 확인할 수 있다(Han et al., 2011). <표 4>는 분류기의 성능을 결정하는 행렬의 4가지를 설명하고 있다.

- TP(True Positive): 실제 True를 True 예측 카운트
- TN(True Negative): 실제 False를 False 예측 카운트
- FP(False Positive): 실제 False를 True 예측 카운트
- FN(False Negative): 실제 True를 False 예측 카운트

<표 4> 오차행렬(Han et al., 2011)

		Predicted class: Yes	Predicted class: No	Total
Actual class	Yes	TP	FN	P
	No	FP	TN	N
	Total	P'	N'	P+N

<표 5>는 예측 성능 평가 지표로 정확도, 오류율, 재현율, 특이도, 정밀도, F1-score, AUC (Area Under Curve) 등이 있다. 정확도는 전체 데이터 중에 실제 True를 True로 예측한 비율이고 오류율은 전체 데이터 중에 실제 False를 False로 예측한 비율을 나타낸다. 민감도(Sensitivity)라고도 하는 재현율은 실제 True인 것 중에서 True로 예측한 비율을 의미하고 특이도는 실제 False인 것 중에서 False로 예측한 비율이고 정밀도는 True라고 예측한 것들

<표 5> Evaluation measures

Measure	Formula
정확도(Accuracy)	$\frac{TP+TN}{P+N}$
오류율(Error rate)	$\frac{FP+FN}{P+N}$
재현율(Recall)	$\frac{TP}{P}$
특이도(Specificity)	$\frac{TN}{N}$
정밀도(Precision)	$\frac{TP}{TP+FP}$
F-1 Score	$\frac{2\times \mathrm{Precision}\times recall}{\mathrm{Precision}+recall}$

중에 실제 True인 비율을 의미한다. F-1 Score는 정밀도와 재현율을 이용하여 계산할 수 있으며 불균형적인 데이터 셋 평가에 적합하다. 재현율과 1-특이도를 각각 x, y축 그래프를 ROC(Receiver Operating Characteristics) 커브이라 하고 아래 면적의 곡선 AUC라고 한다. 각각의 성능 평가 지표의 최댓값은 1이며 좋은 모델은 1에 가까운 값이다.

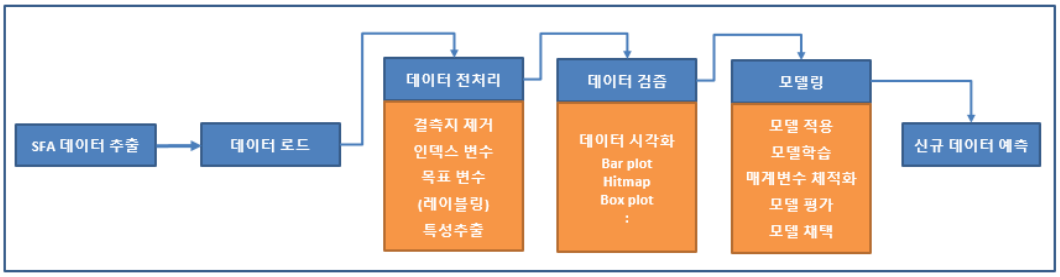
Ⅲ. 연구방법

국내 H 수입차 딜러사의 영업지원시스템의 계약, 해약, 매출 데이터를 기반으로 머신러닝 기법 중 로지스틱 회귀, 서포트 벡터 머신, 랜덤 포레스트, 가우시안 나이브 베이즈, 인공신경망 모델을 파이썬 주피터 노트북을 이용하여 해약 예측 결과를 돌출하였으며 각 모델의 성능을 정확도, 정밀도, 재현율, F1-score, AUC 결과값으로 비교, 평가하였다.

분석 데이터는 2015년부터 2020년의 계약, 해약, 매출 자료 29,073건을 영업지원시스템에서 추출한 후 분석 프로그램 파이썬 주피터 노트북으로 로드하여 데이터 전처리, 검증, 모델링을 진행한 후 신규 데이터로 최종 결과를 예측하였다. <그림 6>은 본 연구의 프레임워크로 영업지원시스템에서 데이터를 추출한 후 데이터 전처리, 검증, 모델링의 절차로 해약 예측 모형을 구현하였다.

<표 6>의 데이터 셋은 15개의 변수 중 1개의 인덱스 변수(계약번호)을 제외하고 13개의 특성 변수와 목표 변수인 해약 여부로 구성하였다.

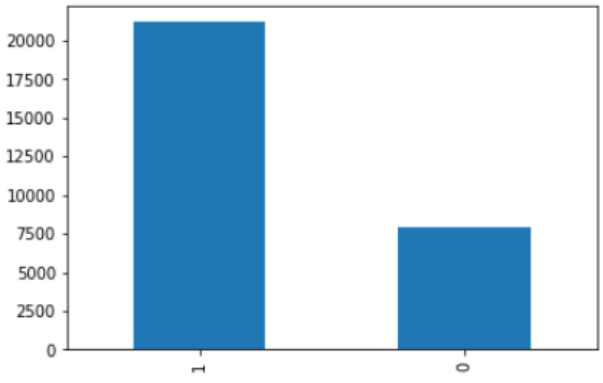
<그림 7>은 전체 데이터 셋에 대한 목표변수인 출고(1) 및 해약(0) 건수를 Bar-plot으로 시각화 한 것으로 각각 출고 73%, 해약 27%를 차지하고 있다.



<그림 6> 본 연구의 프레임워크

<표 6> 데이터 셋 정의

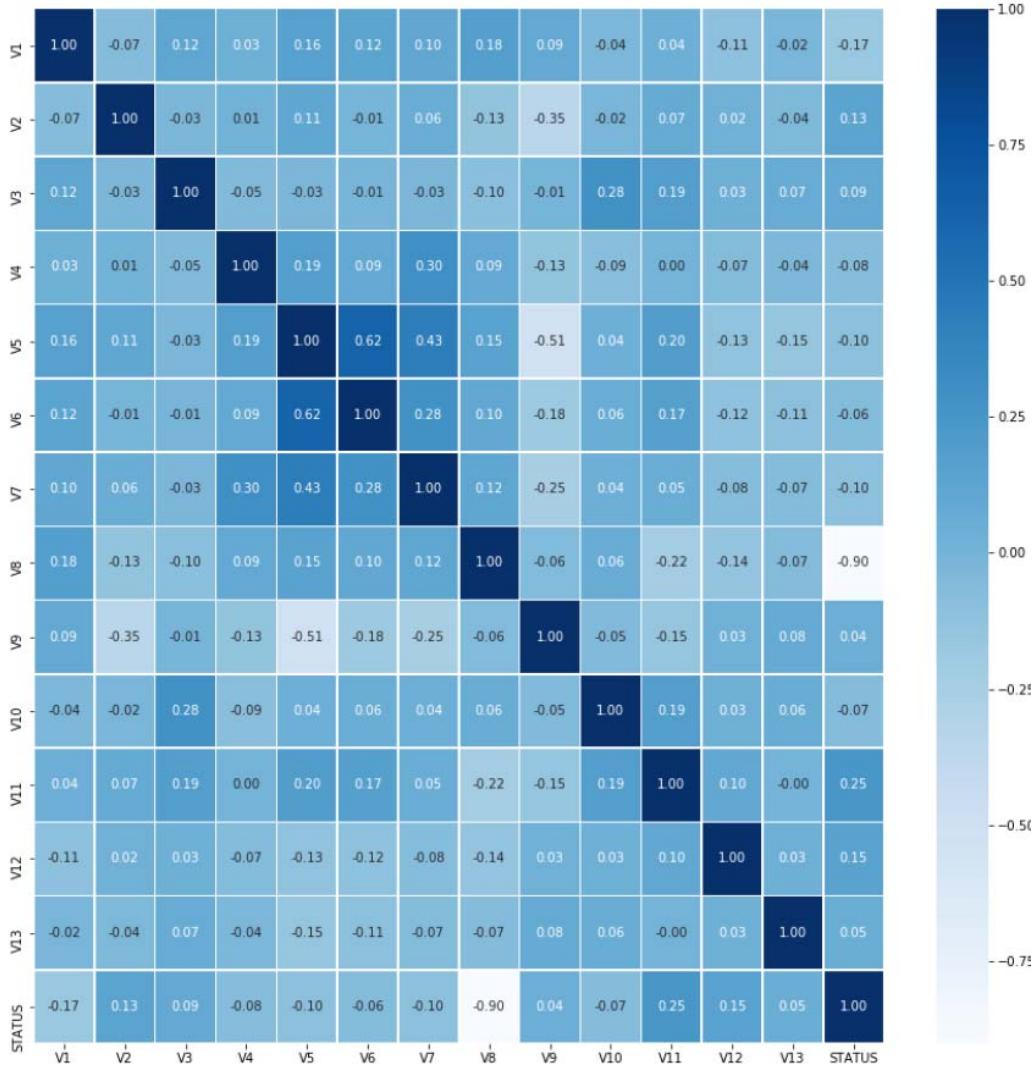
No	변수	설명	데이터	특성 변수	목표 변수
1	ContractNo	계약서 번호			
2	V1	성별	남성, 여성, 미확인	●	
3	V2	인지경로	신규내방, 지인소개, 개인영업활동, 미케팅 활동, 전시장 전화문의, 인터넷/모바일 앱고객	●	
4	V3	계약구분	개인, 법인	●	
5	V4	대표차종	A-Class, B-Class, CLA-Class, C-Class, E-Class, S-Class	●	
6	V5	연식		●	
7	V6	외상색상		●	
8	V7	내상색상		●	
9	V8	시승여부	시승함, 시승안함, 미확인	●	
10	V9	재구매 여부	재구매, 최초 구매	●	
11	V10	판매가		●	
12	V11	할인가격		●	
13	V12	판매조건	현금, 할부, 리스	●	
14	V13	영업사원 직급	주임, 대리, 과장, 차장, 부장	●	
15	STATUS	해약여부	1: 출고, 0: 해약		●



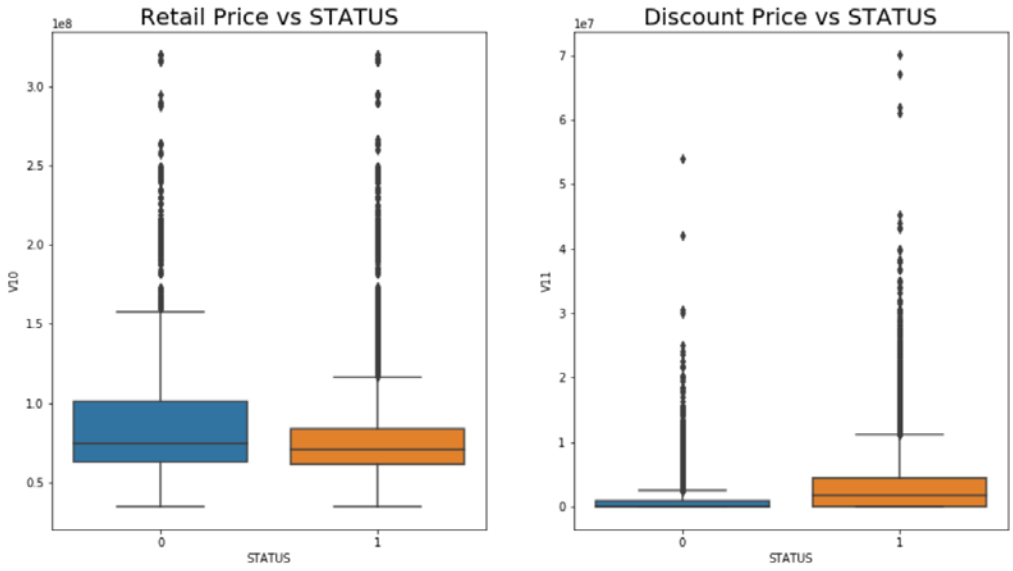
<그림 7> 출고(1), 해약(0) 건수 Bar-plot

데이터 셋의 특성 변수의 상관관계를 파악하고자 Heat map, Box-plot을 이용하여 데이터 시각화를 진행하였다.

<그림 8>은 특성 변수의 상관관계를 한 눈에 파악할 수 있는 Heat map은 성별, 인지경로, 연식, 내부색상, 시승여부, 할인가격, 판매조건이 다른 변수에 비하여 해약여부에 상관관계를 보이고 있으며 특히, 시승여부와 차량의 할인 가격이 높은 상관관계의 경향성을 보이고 있다. 해약여부와 양의 상관관계인 인지경로, 할인 가격, 판매 조건은 출고에 영향을 미치고 있으며 음의 상관관계인 성별, 연식, 시승 여부는 해약에 더 큰 영향이 있다. 특히, 시승여부는 시승을 하지 않은 경우 해약률이 시승한 계약 건보다 높게 나타나고 있고 할인 가격이 있거나 높을수록 출고에 높은 영향을 미치고 있다.



<그림 8> 특성 변수 Heat map



<그림 9> 판매가격, 할인가격 vs 출고(1)/해약(0) Box-plot

<그림 9>은 판매가격과 할인가격에 대한 출고(1) 또는 해약(0)에 대한 Box-plot으로 차량 판매가격이 낮고 할인이 높을수록 해약보다 출고로 이어지는 경향이 높은 것으로 나타나고 있다. 판매 가격의 출고와 해약 1사분위수와 3사분위수의 범위를 보면 각각 6,100만원~8,300만원, 6,300만원~10,100만원으로, 주력으로 판매되고 있는 차량모델의 가격법위이고 판매 가격 8,000만원이 넘는 경우 해약률이 높아지고 있다. 할인가격은 출고 3사분위수 450만원, 해약 3사분위수 100만원으로 할인 프로모션이 적용되어 할인율이 높은 차량이 출고로 이어지고 있다.

Ⅳ. 예측 모델 평가 및 검증

각각의 머신러닝 모델을 학습(Training)시키고 검증(Testing)하기 위하여 분석 데이터 셋 29,073건에 대하여 학습과 검증 데이터셋을 각각 7(20,351건) : 3(8,722건)로 분리하여 모델에 적용 및 학습시켰다. 그리고 분석 데이터 셋에 적합한 각 모델의 성능을 최적화하기 위하여 각 모델의 매개변수를 조정하여 실험하였다. <표 7>은 모델별 주요 매개변수 설정 값, 학습과 테스트 데이터 셋 예측률 그리고 테스트 데이터 셋 에러 건수로 모델 평가결과와 유사하게 나타나고 있다.

실험 후 각 모델의 평가 결과를 보면, <표 8> 및 <그림 10>과 같이 각 모델의 평가 척도가 전체적으로 90.00% 이상을 상회하였으며 실험 모델 중 서포트 벡터 머신과 인공신경망이 전체적으로 비교적 높은 성능을 발휘하고 있다. 특성 변수가 연속적인 데이터 처리에 적합한 가우시안 나이브 베이즈를 제외한 나머지 모델의 경우 매개변수 조정에 따라 유사한 성능을 보였으며, 서포트 벡터 머신은 다소 높은 성능

<표 7> 모델별 매개변수 및 예측률

Model	매개변수	학습 예측률	테스트 예측률	테스트 에러 건수
Logistic regression	C=1.0 penalty='l2' max_inter=100	96.29%	96.00%	349
Support vector machine	C=1.0 gamma='auto' kernel='rbf'	97.55%	97.40%	227
Random forest	n_estimators=20 max_depth=500	99.66%	97.01%	261
Gausian NB	none	92.98%	92.59%	646
Artificial neural network	hidden layer=(10,10,10) activation='relu' max_inter=500	97.63%	97.33%	233

<표 8> 모델 평가

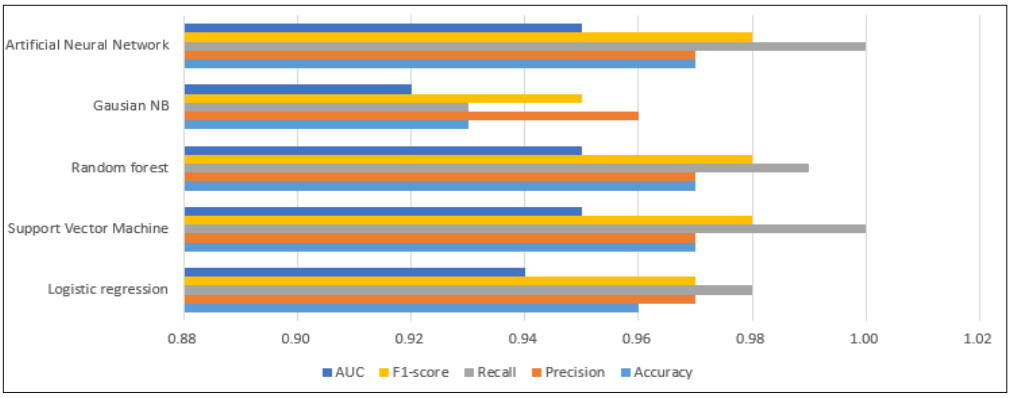
Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic regression	96.00%	96.61%	97.98%	97.29%	94.27%
Support vector machine	97.40%	96.57%	99.95%	98.26%	95.13%
Random forest	97.02%	96.77%	99.25%	97.99%	95.05%
Gausian NB	92.59%	96.43%	93.35%	94.87%	91.93%
Artificial neural network	97.33%	96.66%	99.87%	98.21%	95.11%

<표 9> 영업지원 시스템 출고/해약 예측 모델 연계 데이터 셋 샘플

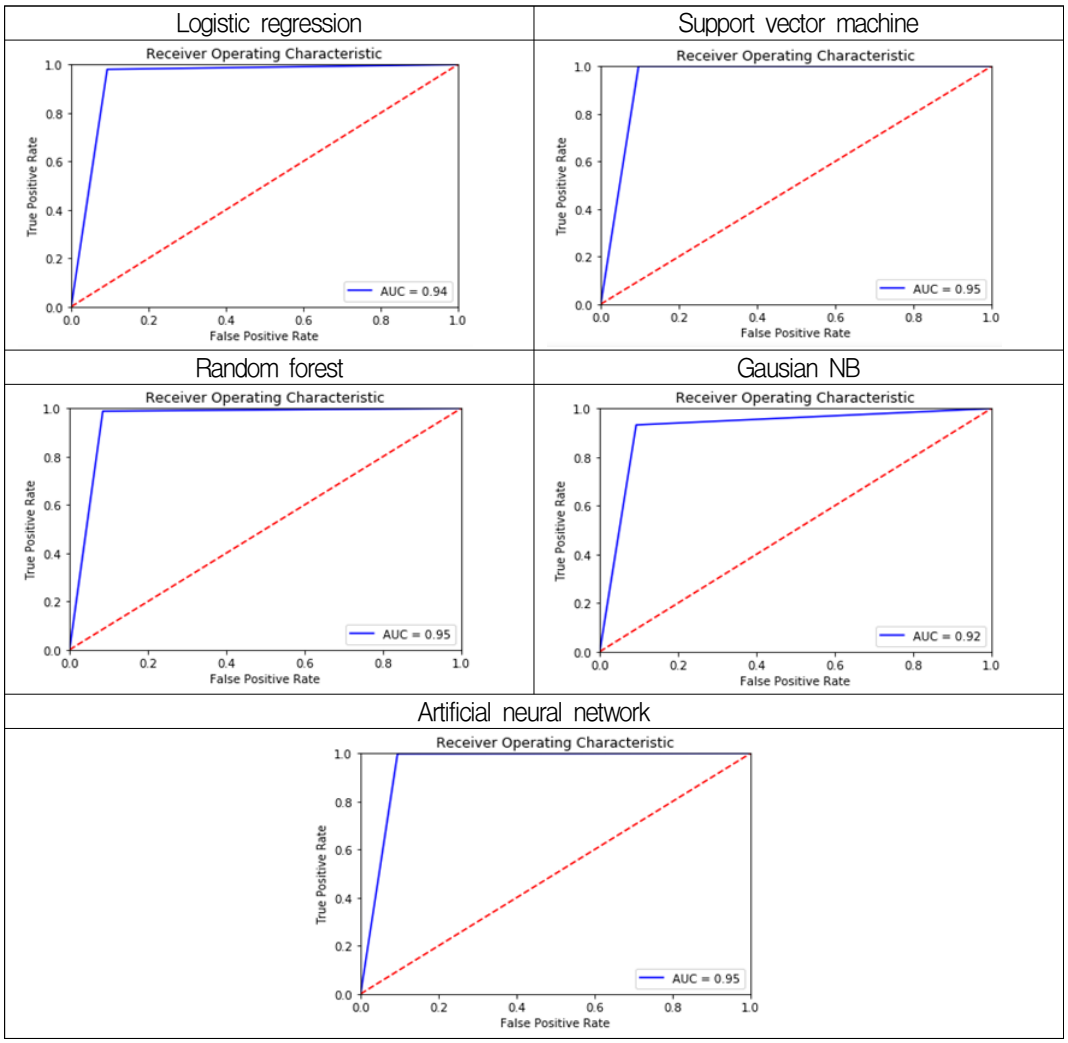
계약서 번호	고객명	출고 예측률	해약 예측률	출고/해약 예측
AR010000	김XX	66.23%	33.77%	출고
AR020000	홍XX	20.09%	79.91%	해약
AR030000	박XX	36.28%	63.72%	해약
AR040000	정XX	52.33%	47.67%	출고

대신에 다른 모델보다 속도가 느린 단점을 확인하게 되었다.

각 모델별로 90%이상의 높은 해약 예측 성능을 발휘하고 있지만, 적은 데이터 수와 널(Null) 값에 대한 다수의 결측치 처리로 과적합 현상을 보이고 있다. 그러나 영업지원 시스템의 실 데이터를 이용하여 <표 9>와 같이 각 계약건들에 대한 출고 및 해약 예측률을 영업지원 시스템과 실시간으로 연계함으로써 고객 영업 및 마케팅 담당자들에게 타켓 고객 관리 효과 기대할 수 있다.



<그림 10> 모델 평가



<그림 11> ROC 커브 & AUC

<그림 11>은 각 모형별 예측성능 지표 중 ROC 커브와 AUC 값을 시각화 한 것으로 으로 ROC 커브(Blue line)의 아래 면적이 넓을수록 AUC 값이 1에 가깝고 예측 성능이 우수하다고 할 수 있다.

Ⅴ. 결과 및 시사점

본 연구는 국내 수입자 H 딜러사 영업지원시스템의 2015년부터 2020년까지의 계약, 해약, 매출 데이터 셋을 머신러닝 모델에 적용하였다. 차량 계약 건들 중 출고 대기 기간 동안 발생할 수 있는 해약을 미리 예측함으로써 해약 가능성이 높은 계약 고객들을 집중 관리할 수 있도록 영업지점과 공유하여 해약을 미연에 방지하고 매출로 이어질 수 있도록 차량 구매 계약에 대한 해약 예측 모형을 구현 하였다.

총 29,073 건의 계약, 매출, 해약 데이터에 대하여 인덱스로 정의한 계약서 번호를 제외하고 13개의 특성 변수와 1개의 목표 변수로 정의하여 5가지 머신러닝 모델에 적용하여 실험하였다. 서포트 벡터 머신과 인공신경망 모형이 정확도 97%, 정밀도 97%, 재현율 99%, F1-Score 98%, AUC 95%로 높은 평가 결과를 돌출되었다. 따라서, 자동차 구매 계약 정보에 대한 해약 예측에 대하여 전통적인 머신러닝 모델뿐만 아니라 인공 신경망을 이용한 딥러닝 모델도 우수한 예측 성능을 발휘한다는 것을 확인하게 되었다. 이는 자동차 구매 계약 정보를 머신러닝 모델에 적용함으로써 해약 예측에 대한 가능성을 증명하게 되었다.

머신러닝 기법을 이용한 고객 해약 예측 또는 고객 이탈 모형에 관한 연구는 주로 보험사, 증권회사, 카드사, 통신 서비스 사업자, 모바일 인터넷 게임회사의 상품이나 서비스 이용 고객을 대상으로 중도 해약이나 중도 이탈 모델이 주로 개발되었다. 그러나, 본 연구는 수입자동차 유통 과정에서 상품이나 서비스 이용 전 초기 계약 이후 수요에 따른 공급 문제로 인한 고객 해약 예측 모델을 구현하였다.

학문적 의의로는 수입자동차 영업 지원 시스템에 축적되어 있는 계약, 매출, 해약 고객 데이터 기반으로, 다수의 머신러닝 모델을 적용한 국내 수입자동차 시장 최초의 계약 고객에 대한 해약 예측 모형 개발연구로, 자동차 계약 데이터를 머신러닝 모델과 융합함으로써 해약 예측이 가능한 일반화된 모형 구현 및 활용 가능성을 제시하였다.

최근 마케팅 및 영업 담당자는 과거 대중 마케팅 전략에서 인스타그램, 페이스북과 같은 SNS(Social network service)의 개인글이나 댓글을 이용하여 개별 고객의 마음을 예측하고 선 제안할 수 있는 고객 핀셋 마케팅 전략이 대세를 이루고 있다. 따라서, 본 연구는 개별 고객의 계약 정보로 해약 여부를 예측하여 집중적으로 해당 고객을 관리함으로써 해약률을 줄이고 매출 증대를 제고할 수 있는 실무적 시사점을 제시하였다.

국내 H 수입차 딜러사의 영업지원시스템의 데이터 셋은 사용자 입력에 의존하여 생성된 데이터로 필수 입력 또는 선택 사항이 아닌 특성 변수의 경우 다수의 결측치 처리와 충분하지 않은 데이터 셋으로 예측모델의 평가 결과가 과적합되는 한계점이 있었다. 또한 특정 딜러사 및 특정 수입 브랜드를 대상의 연구 결과

로, 타 수입 브랜드 및 국산차 구매 계약 데이터 셋을 수용하지 못하여 당장 일반화된 예측 모형을 이용하기에는 연구의 한계점이 있다.

향후에는 차량 구매 계약 데이터 수집 시 포괄적으로 일반화를 적용할 수 있은 구매 계약, 해약, 매출의 특성 변수 데이터 셋을 대상으로, 다수의 머신러닝 예측 모델을 결합하여 성능을 향상 시킬 수 있는 앙상블 러닝(Ensemble Learning) 기법의 예측 모델을 구현하고자 한다.

참고문헌

김미정, “SFA(Sales Force Automation)의 도입과 구축 및 실행에 있어서의 핵심 성공 요인에 관한 연구: 전사적 업무 프로세스 통합의 관점에서”, 서울대학교 석사 학위 논문, 2002.

김영식, 김훈호, “머신러닝 기법을 활용한 사교육 참여 예측 모형 탐색”, 교육재정경제연구, 제28권, 제3호, 2019, pp 29-52.

김종성, 이준형, 김동현, 최창현, 이명진, 김형수, “머신러닝 기반의 호우피해 발생확률 예측 모형 개발”, 한국방재학회 논문집, 제19권, 제6호, 2019, pp. 115-127.

김준석, 양미연, 윤상후, “기계학습과 GloSea5를 이용한 장기 농업기상 예측: 고랭지 배추 재배 지역을 중심으로”, 디지털융복합 연구, 제18권, 제4호, 2020, pp. 243-250.

김태기, 안창순. “한국 자동차산업의 국제경쟁력 분석”, 무역학회지, 제22권, 제3호, 1997, pp. 139-157.

김상휘, 김기원, 김유성, 윤태영, 전재완, “머신러닝, 딥러닝을 이용한 통신서비스 이용고객 분석 및 이탈 예측”, 2020 온라인 추계학술발표대회 논문집, 제27권, 제2호, 2020, pp. 568-571.

김태인, “SFA(sales forces automation) 도입이 서비스 기업의 업무성과에 미치는 영향”, 한국엔터테인먼트산업학회 학술대회, 논문집, 2013, pp. 39-43.

나광택, 이진형, 깅은찬, 이효찬, “증권 금융 상품 거래 고객의 이탈 예측 및 원인 추론”, 한국 빅데이터 학회지, 제5권, 제2호, 2020, pp. 215-229.

박기호, 정재곤, 황명화, “P2P LBS를 활용한 모바일 영업자동화(SFA) 시스템에 관한 연구”, 한국GIS학회지, 제11권, 제1호, 2003, pp. 61-72.

박남규, 전영신, 장완진. “자동차 산업 100 년 역사에 도전하는 작은 거인: Tesla.” Korea Business Review, 제23권 제3호, 2019, pp. 49-68.

박평우, 김민구, 임홍석, 윤덕용, 이석원, “허혈성 심장질환 진단을 위한 기계 학습 알고리즘 비교 연구”, 정보과학학회 논문지, 제45권, 제4호, 2018, pp. 376-389.

서광규, “Support Vector Machine을 이용한 고객 이탈 예측 모형에 관한 연구”, 안전경영과학회지, 제7권, 제1호, 2005, pp. 199-210.

송현정, 이석준 “딥러닝을 활용한 실시간 주식 거래에서의 매매 빈도 패턴과 예측 시점에 관한 연구: KOSDAQ 시장을 중심

으로", 정보시스템연구, 제27권, 제3호, 2018, pp. 123-240.

이동규, 신민수, "카드 산업에서의 고객 휴면 예측", 한국IT서비스학회 2018추계 학술대회, pp. 404-407.

이민수, 최영찬, "머신러닝을 활용한 모돈의 생산성 예측 모델", 농촌지도와 개발 제16권, 제4호, 2009, pp. 939-965.

이석호, 강대천, 이찬, 강무진, "실험계획법을 이용한 다층 퍼셉트론 인공 신경망의 구조 설계", 한국정밀공학회 학술발료대회 논문집, 1996, pp. 536-540.

이장택, 조현식, "로지스틱 회귀모형을 이용한 프로야구 홈경기의 이점에 관한 연구", 한국자료분석확회, 제11권, 제1호, 2009, pp. 533-543.

이인지, 윤현식, "머신러닝을 활용한 지역축제 방문객 수 예측모형 개발", 한국정보시스템 연구, 제29권, 제3호, 2020, pp. 35-52.

임세현, 허연, "의사결정나무를 이용한 온라인 자동차 보험 고객 이탈 예측과 전략적 시사점", 한국경영정보학회, 제8권, 제3호, 2006, pp. 125-135.

정성훈, 진창하, "머신 러닝 방법을 이용한 오피스 임대료 산정 - 랜덤포레스트, 인공신경망, 서포트 벡터 머신 활용 중심으로", 부동산학 연구, 제26집, 제2호. 2020, pp. 23-53.

최준영, 이동현, 김준영, 정교민, "딥러닝 기반의 미세먼지 농도 예측", 한국컴퓨터종합학술대회 발표 논문집, 2019, pp. 859-861.

채영석. "중국의 자동차산업, 시장 독재가 시작된다.", 오토저널, 제40권, 제8호, 2018, pp. 54-57.

Hwang, U. "4 차 산업혁명 시대의 국내외 전기자동차 보급 확대정책과 전력망 재편 방향." 전기의세계, 제68권, 제1호, 2019, pp. 8-16.

Breiman, L., and "Random forests", *Machine Learning*, Vol. 45, No.1, 2001, pp. 5-32

Cox, D.R. "The Regression Analysis of Binary Sequences", *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 20, No. 2, 1958, pp. 215-232.

Cortes C., Vapnik V., "Support-Vector Networks", *Machine Leaming*, Vol. 20, 1995, pp. 273-297.

Han, J., Kamber, M., and Pei, J.,, "Data Mining: Conception and Techniques Third Edition", *Mogan Kaupmann Publisers*, 2011, pp. 302-308.

Kamel, H., Abdulah, D., Al-Tuwaijari, and Janal M, "Cancer Classification Using Gaussian Naive Bayes Algorithm", *2019 International Engineering Conference (IEC) Engineering Conference (IEC)*, 2019, pp. 165-170.

Vapnik V., "Support Vector Machines", *The Nature of Statistical Learning Theory (2nd)*, John Wiley & Sons: New York, N.Y, 1996, pp. 181-223

삼정KPMG경제원구원, "코로나19에 따른 자동차산업 동향 및 대응전략", 2020.

산업통상자원부 “시바-르에서 완전자율주행까지” http://www.motie.go.kr/, 2020.
KAIDA, “한국수입자동차협회 Homepage”, https://www.kaida.co.kr/, 2019.
KAICA, “한국자동차산업협동조합 Homepage”, http://www.kaica.or.kr/, 2020.
Salesforce.com “세일즈포스닷컴 코리아 Homepage”, https://salesforce.com/, 2020.

정 동 균 (Jung, Dong Kun)

부경대학교 경영학과 석사 학위를 취득하고 부경대학교 경영컨실팅 협동과정 박사 과정을 수료하였으며. 주요 관심분야는 BigData, Machine Learning, Sales Force Automation 등이다.

이 종 화 (Lee, Jong Hwa)

부경대학교 경영학 석사와 박사학위를 취득하였다. 현재 동의대학교 정보경영학부 e비즈니스학전공 교수로 재직하고 있으며, 주요 관심분야는 BigData, Mining, Content Analysis 등이다.

이 현 규 (Lee, Hyun Kyu)

연세대학교 경영학 박사학위를 취득하였다. 현재 부경대학교 경영학부 교수로 재직하고 있으며, 주요 관심분야는 정보시스템전략, Data-Mining & Analysis 등이다

<Abstract>

A Study on the Prediction Model for Imported Vehicle Purchase Cancellation Using Machine Learning: Case of H Imported Vehicle Dealers

Jung, Dong Kun · Lee, Jong Hwa · Lee, Hyun Kyu

Purpose

The purpose of this study is to implement a optimal machine learning model about the cancellation prediction performance in car sales business. It is to apply the data set of accumulated contract, cancellation, and sales information in sales support system(SFA) which is commonly used for sales, customers and inventory management by imported car dealers, to several machine learning models and predict performance of cancellation.

Design/methodology/approach

This study extracts 29,073 contracts, cancellations, and sales data from 2015 to 2020 accumulated in the sales support system(SFA) for imported car dealers and uses the analysis program Python Jupiter notebook in order to perform data pre-processing, verification, and modeling that is applying and learning to Machine learning model after then the final result was predicted using new data.

Findings

This study confirmed that cancellation prediction is possible by applying car purchase contract information to machine learning models. It proved the possibility of developing and utilizing a generalized predictive model by using data of imported car sales system with machine learning technology. It can reduce and prevent the sales failure as caring the potential lost customer intensively and it lead to increase sales revenue by predicting the cancellation possibility of individual customers.

Keyword: Machine Learning, Sales Force Automation, Imported Car Business, Cancellation Prediction.

* 이 논문은 2021년 5월 8일 접수, 2021년 5월 27일 1차 심사, 2021년 6월 17일 게재 확정되었습니다.