**기계학습기초**

**프로젝트 보고서**

**E-Commerce 소비자의 충성도 예측**

**Iot 인공지능 융합전공**

**2조 165987 구태희**

**목차**

[**1. 주제** 3](#_Toc90826896)

[**2. 데이터** 3](#_Toc90826897)

[**2-1. 데이터 수집** 3](#_Toc90826898)

[**2-2. 데이터 분석** 3](#_Toc90826899)

[**2-3. 데이터 전처리** 4](#_Toc90826900)

[**1. Category** 4](#_Toc90826901)

[**2. Brand** 5](#_Toc90826902)

[**3. Event time** 6](#_Toc90826903)

[**4. User session** 7](#_Toc90826904)

[**5. Price** 9](#_Toc90826905)

[**3. 모델** 9](#_Toc90826906)

[**3-1. 모델 선정** 9](#_Toc90826907)

[**3-2. 모델 학습** 10](#_Toc90826908)

[**1. Random forest** 10](#_Toc90826909)

[**2. K-Nearest Neighbor** 11](#_Toc90826910)

[**3. Deep Neural Network** 11](#_Toc90826911)

[**4. 보안점** 13](#_Toc90826912)

[**5. 결론** 13](#_Toc90826913)

# **1. 주제**

우리는 인터넷으로 사고 싶은 물건이 있으면 다양한 e-커머스 플랫폼을 통해서 어떤 플랫폼을 통해 구매하는 것이 합리적인지 고민하고 선택한다. 여러 e-커머스 플랫폼을 비교 분석하여 선택하기도 하고 또는 주로 이용하는 플랫폼만을 사용하기도 한다. 기업은 이익을 증대하기 위해서 소비자의 선택을 받을 확률을 높여야 하며 소비자가 자신의 기업을 주로 사용하게끔 만들어야 한다. 즉, 소비자의 구매 성향을 파악하고 구매율과 재방문율을 높여 기업의 충성기업을 확보하는 것이 주요한 문제가 된다. 기업들은 소비자의 구매 행동 데이터를 통해 어떤 소비자들이 기업의 e-커머스를 사용하지 파악하고 방문 빈도를 통해 소비자의 충성도를 예측할 수 있다면 이를 통해 맞춤 마케팅과 여러 유인 기업을 통해 충성 고객을 확보하기 위한 경영 전략을 수립할 수 있다.

# **2. 데이터**

## **2-1. 데이터 수집**

데이터 샘플은 kaggle의 eCommerce behavior data from multi category store를 이용한다. 해당 데이터 샘플은 한달 간격으로 e-커머스 사용자 데이터가 취합되어 있으며 해당 데이터 샘플의 아카이브를 통해 2019년 10월부터 2020년 4월까지 총 7개월간의 데이터가 존재한다. 이를 병합하여 사용하였으며 전처리 수행 이전 원본 데이터 샘플은 약 57.9GB 크기로 구성되어 있다.



그림 1, <https://www.kaggle.com/mkechinov/ecommerce-behavior-data-from-multi-category-store>

## **2-2. 데이터 분석**

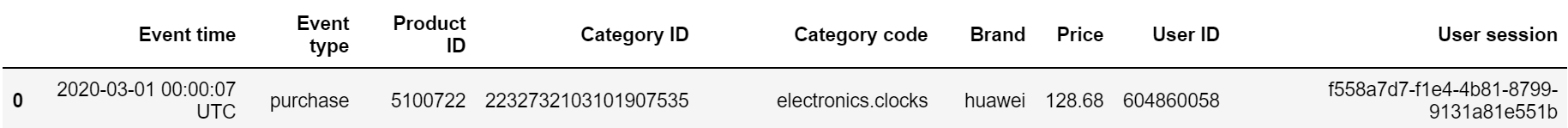


그림 2, eCommerce behavior data 예시

eCommerce behavior data는 Event time, Event type, Product ID, Category ID, Category Code, Brand, Price, User ID, User session 총 9개의 열로 구성되어 있으며 사용자 각각의 행동에 따른 정보가 담겨있다. Event time은 사용자 행동, 이벤트가 발생한 시간으로 협정 세계시에 해당하는 UTC로 표기되어 있다. Event type은 이벤트의 종류를 의미하며 사용자가 상품을 구매한 행동은 purchase, 상품을 클릭한 행동은 view, 상품을 장바구니에 담은 행동은 cart, 상품을 장바구니에서 제거한 행동은 remove로 총 4개의 범주로 구성된다. Product ID는 이벤트가 발생한 상품의 고유 ID로 각 ID는 숫자로 표기된다. 데이터 샘플에서 ID에 대응하는 상품을 찾을 수 없기 때문에 해당 열을 삭제한다. Category ID는 이벤트가 발생한 상품의 카테고리의 ID, Category Code는 상품의 세부 카테고리를 의미한다. Category ID의 경우 정수로 구성되며 ID의 의미가 Category Code와 중복되기 때문에 삭제한다. Category Code의 경우 문자열로 메인.세부1.세부2.세부3 총 4단계로 카테고리가 . 을 통해 구분되어 있다. Brand는 이벤트가 발생한 상품의 브랜드로 총 4116개의 브랜드가 데이터 샘플에 존재한다. Price는 이벤트가 발생한 상품의 가격, User ID는 사용자의 ID로 한 명의 사용자는 고유한 ID를 가진다. User session은 이벤트가 발생할 당시의 사용자의 세션으로 사용자 세션은 사용자의 최초 행동부터 일정기간 유지되며 일정기간이 지나고 나서 해당 사용자의 행동이 등장한다면 세션은 변경된다. 따라서 동일한 세션에서 여러 이벤트가 발생할 수 있으며 사용자의 이용 빈도에 따라 동일한 User session의 등장 빈도 결정된다.

우리는 e-커머스 사용자가 최초 구매 이후 다시 해당 e-커머스를 방문하여 구매할 확률, 재 방문을 통한 재 구매를 통해 사용자의 충성도를 예측하기 위해 Event type 중 상품을 구매한 행동에 해당하는 purchase에 해당하는 데이터 샘플을 사용한다. 그리고 User session을 통해서 개별 사용자의 구매를 위한 방문 빈도와 각 방문 별 구매 횟수를 계산할 수 있다. 동일한 User session을 통합하여 각 방문에 대해 구매한 품목을 확인하고 총 방문 횟수를 통해 충성도를 예측한다.

e-커머스 사용자 행동 데이터 샘플에서 구매 행동에 관한 이벤트에 해당하는 purchase 데이터 샘플만을 추출하여 약 570만개의 데이터를 수집하였으며 추출한 데이터 샘플에서 앞서 이야기한 목적에 맞게 전처리를 수행하여 해당 데이터로 모델 학습을 수행한다.

## **2-3. 데이터 전처리**

### **1. Category**

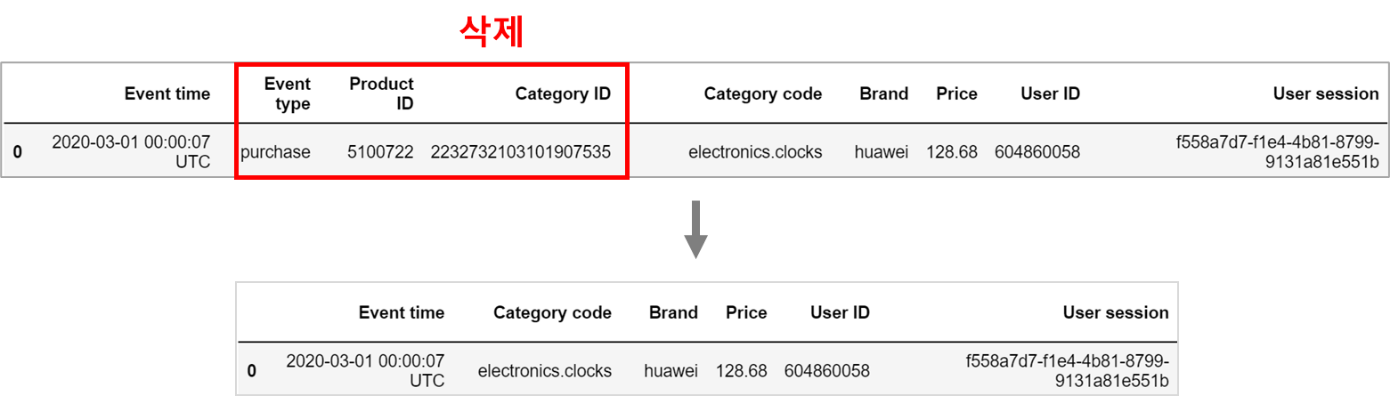


그림 3, Event type, Product ID, Category ID 삭제

먼저 데이터 분석 과정에서 불필요하다고 판단된 열인 Product ID와 Category ID를 삭제하고 Event type의 경우 purchase에 해당하는 데이터 샘플만을 추출하였기 때문에 더 이상 의미가 없어 삭제한다.

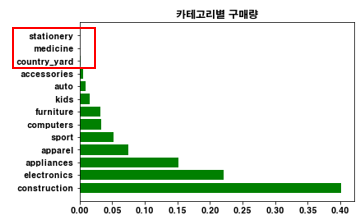


그림 4, 메인 카테고리로 분류, 2%미만 카테고리 확인

구매한 상품의 카테고리를 구분하여 소비자가 어떤 카테고리에 해당하는 상품을 구매하였는지 파악하기 위해 이에 해당하는 Category code를 수정한다. Category code는 총 4단계의 카테고리가 .으로 구분되어 있고 세부 카테고리까지 모두 고려할 경우 범주가 928개로 학습 시 데이터에 대한 분별력이 매우 감소한다. 세부1 카테고리까지 만을 사용하는 경우에도 128개의 범주가 존재하여 Category code에서 첫 번째에 해당하는 메인 카테고리를 기준으로 상품의 카테고리를 확인한다. . 으로 구분되어 있는 세부 카테고리를 모두 제거하여 메인 카테고리 만을 사용하였을 때 총 13개의 범주가 등장하며 이 중 데이터 샘플에서 차지하는 비율이 2% 미만이 되는 stationary, medicine, country\_yard에 해당하는 데이터는 전체 데이터 샘플에 대한 설명력이 매우 약하기 때문에 제거한다. 최종적으로 Category code는 10개의 카테고리로 구분된다.

### **2. Brand**

구매한 상품의 브랜드를 파악할 수 있는 Brand의 경우 데이터 샘플에서 총 4081개가 등장한다. 모든 브랜드를 고려하여 사용할 경우 데이터 샘플이 매우 희소해질 수 있는 문제가 있다고 판단된다. 브랜드는 상위 브랜드에 분포가 몰려 있어 등장 빈도가 적은 브랜드는 전체 데이터 분포에서 각각이 차지하는 비율이 매우 작다. 그러므로 앞서 처리한 Category code를 이용하여 파악한 10의 카테고리에서의 브랜드 출현 빈도에 따라 각 카테고리별 상위 5위까지의 브랜드를 사용, 나머지 브랜드는 데이터 샘플에서 제외한다. 브랜드의 경우 여러 카테고리에서 중복적으로 등장할 수 있다. 따라서 특정 카테고리에서 상위 5위 안에 들지 못하지만 다른 카테고리에서 등장할 경우 5위안에 들지 못해도 해당 카테고리에서 브랜드가 등장한다. 이러한 과정을 거친 후 브랜드의 개수는 총 39개로 줄어든다.

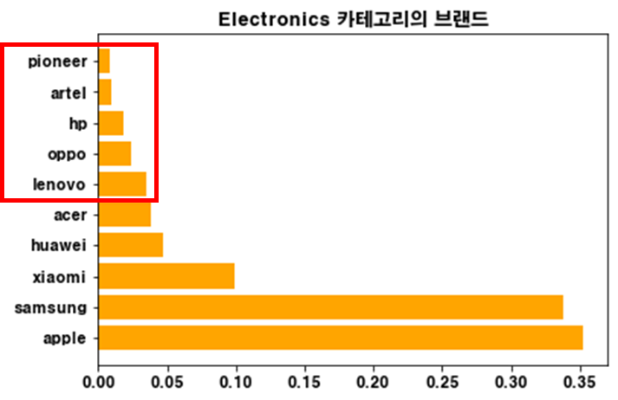


그림 5, Electronics 카테고리의 상위 5위 브랜드와 다른 카테고리 상위 5위 브랜드 중 Electronics에 포함된 브랜드

Category code 와 Brand에서 데이터처리를 통해 약 438만개의 데이터 샘플이 남게 된다.

### **3. Event time**

이벤트가 발생한 UCT 시간은 Event time은 문자열로 구성되어 있다. 이를 Datetime 타입으로 변환하여 시간과 날짜를 계산한다. 시:분:초에 해당하는 시간은 분, 초를 제외하여 24시간 형태로 구분한다. 24시간으로 나뉜 시간에 따른 구매 빈도를 확인할 수 있다. 그리고 날짜를 통해 요일을 계산하여 각 요일 별 사용자의 구매 빈도를 확인할 수 있다.

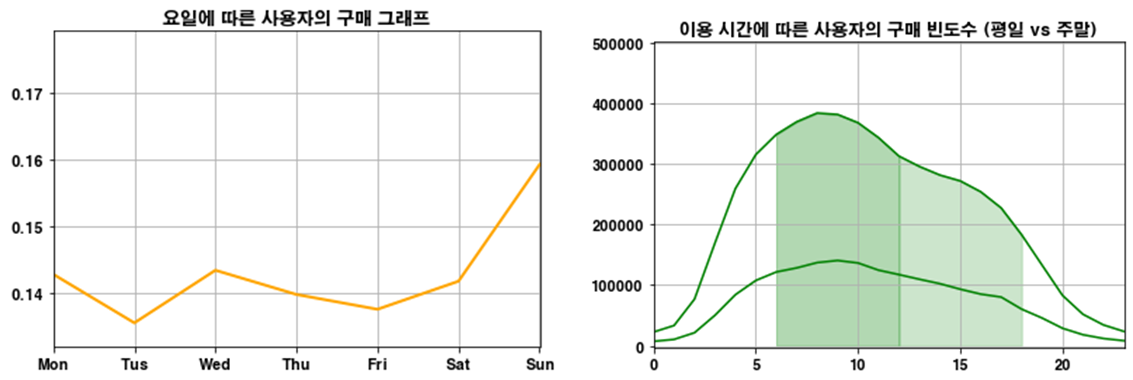


그림 6, 요일과 시간에 따른 구매 분포

나뉜 시간과 요일을 통해 카테고리 별로 특정 시간과 요일에 따른 특징이 있는지 역시 확인이 가능하다.

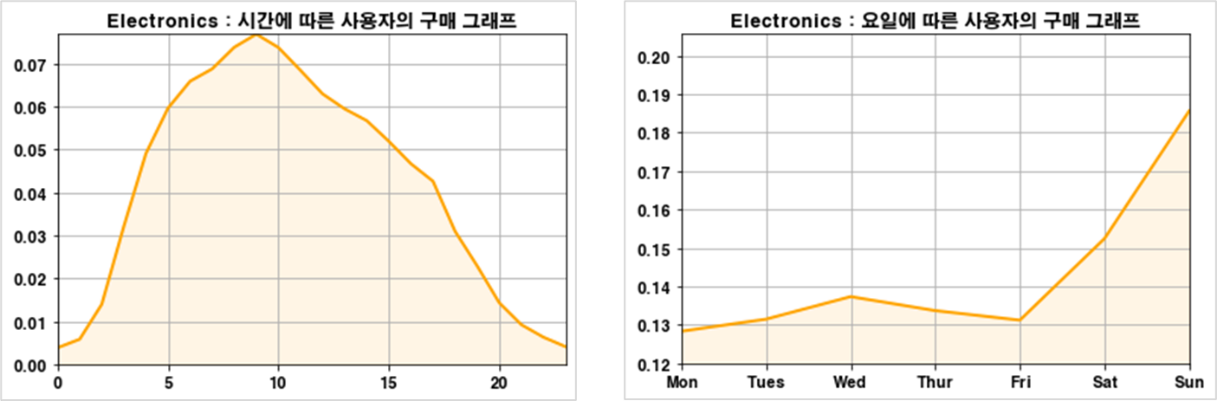


그림 7, Electronics 카테고리의 시간별, 요일 별 구매 분포

평균적으로 시간대의 경우 점심 시간대, 요일의 경우 주말에 해당하는 토요일과 일요일에 구매 빈도가 상대적으로 높음을 파악할 수 있다. 이를 통해 하루를 24시간에서 6시간 단위로 나누어 dawn, morning, afternoon, evening 4개의 범주로 묶어서 처리하고 요일의 경우 월, 화, 수, 목, 금요일은 주중에 해당하는 Weekday, 토요일과 일요일은 주말에 해당하는 Weekend로 2개의 범주로 구분한다. 행동 발생의 시간을 하루의 시간대와 주중, 주말 여부를 통해 구분했기 때문에 기존의 열인 Event time은 삭제한다.

텍스트이(가) 표시된 사진

자동 생성된 설명

그림 8, Event time을 Event timeline과 Weekday로 구분

시간대를 나타내는 Event timeline과 주중, 주말 여부를 나타내는 Weekday의 추가로 데이터 샘플의 특성은 7개로 구성되어 있다. 이후 범주형으로 구성된 Brand, Category, Event timeline, Weekday에 대해서 One-hot 인코딩을 수행한다. One-hot 인코딩 이후 데이터 샘플의 특성은 65개가 된다.

### **4. User session**

각 사용자의 고유한 ID인 User Id와 사용자별 접속 기록인 User session을 통해 사용자의 각 접속에 대한 구매 이력을 파악할 수 있다. User session은 특정 사용자가 접속하고 나서 일정 시간이 지나면 변경되기 때문에 User session은 각 접속, 사용자의 e-커머스 방문에 대한 기록이 된다. User ID를 키로 사용하여 User session을 통합하면 한 명의 사용자에 대해 총 방문 기록과 각 방문에 대한 전체 구매 이력을 파악할 수 있다.

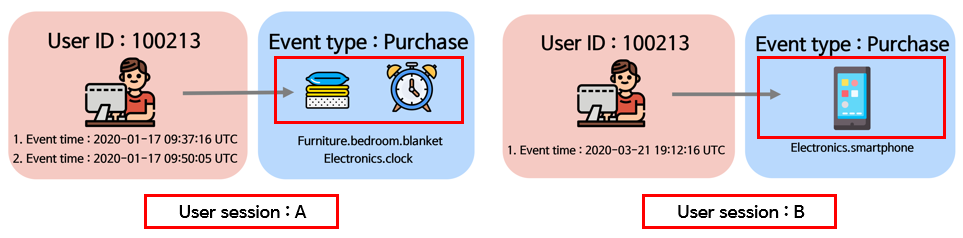


그림 9, User session 별 구매이력

각각의 세션 별 파생된 구매 이력을 통합한다.

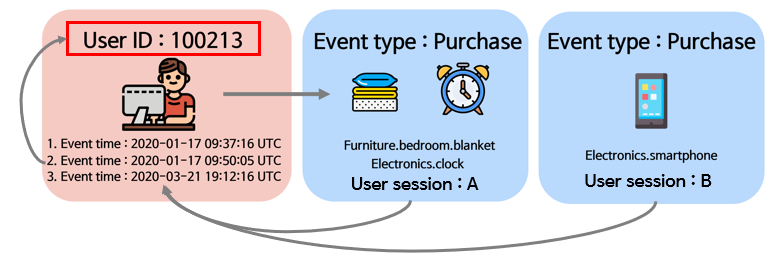


그림 10, User ID 별 구매이력

User ID와 User session을 Event time에 따라 정렬하여 사용자의 세션 별 구매 이력을 시간 순으로 확인할 수 있다. 시간 순으로 정렬된 User session을 통해서 각 세션이 사용자의 몇 번째 방문인지를 파악하여 session으로 표현한다. 또한 사용자의 총 방문 횟수를 통해 재방문 횟수를 나타내는 total\_session을 계산할 수 있다.



그림 11, User ID를 키로 생성한 각 세션에 대한 기록 session, 총 방문횟수 total\_session

사용자의 재방문 횟수에 해당하는 total\_session을 통해 우리는 예측하고자 하는 소비자 충성도에 해당하는 Customer royalty를 획득하여 종속변수로 사용한다. 소비자의 충성도는 재방문 횟수에 따라 분포를 확인하고 충성도가 높은 상위 클래스의 비율을 일반적으로 생각할 수 있는 분포와 유사하게 만들도록 경계 값을 설정한다. 재방문하여 구매하지 않은 사용자는 Bronze(26%), 재방문하여 1~3회 추가 구매한 사용자는 Silver(32%), 재방문하여 4~10회 추가 구매한 사용자는 Gold(20%), 재방문하여 11~39회 추가 구매한 사용자는 Platinum(15%), 재방문하여 40회 이상 추가 구매한 고객은 Dia(5%)로 5개의 범주로 구분한다.

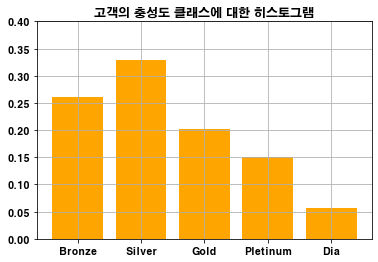


그림 12, Customer royalty 분포

### **5. Price**

User session으로 통합된 구매 금액에 해당하는 Price의 분포를 확인한다. 이후 IQR을 사용하여 제한선을 설정하고 전체 6.7%에 해당하는 이상치를 제거한다. 이상치를 제거한 후 Price의 경우 One-hot 인코딩된 다른 값들과 스케일의 차이가 존재하기 때문에 분포의 평균화 표준편차를 이용하여 정규화 작업을 수행한다.

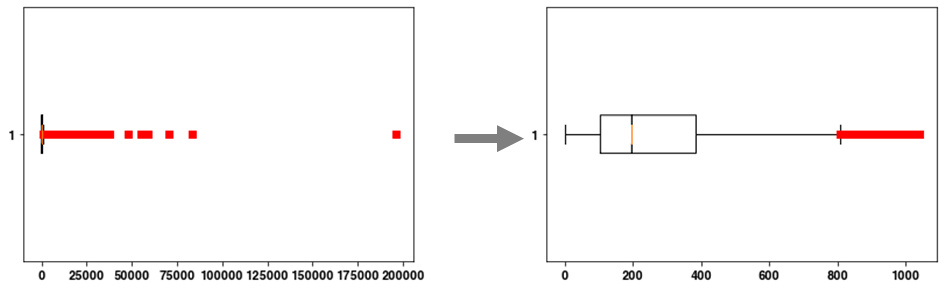


그림 13, 이상치 제거 후 Price

최종 데이터 샘플은 65개의 특성이 있는 약 325만개로 구성되어 있다.

# **3. 모델**

## **3-1. 모델 선정**

데이터 전처리 과정으로 획득한 소비자의 충성도를 나타내는 Customer royalty를 종속변수로 사용하여 분류 모델을 통해 구매 이력을 통해 어떤 충성도를 나타내는 클래스가 될지 예측한다.

분류를 위한 전통적인 기계학습 모델인 Random forest와 K-Nearest Neighbor를 학습하고 두 모델의 정확도를 비교해보고 직접 Deep Neural Network 모델을 설계하여 기계학습 모델과의 차이가 존재하는지 확인한다.

## **3-2. 모델 학습**

전처리가 완료된 데이터 샘플을 학습을 위한 Training(80%), 추론을 위한 Validation(20%)로 랜덤 샘플링한다. 랜덤 샘플링 과정에서 각 클래스 별 비율이 유사하게 Training과 Validation으로 나눠지게 되며 각 샘플들은 독립 변수와 레이블 값이 되는 종속 변수를 가지고 있다.

텍스트이(가) 표시된 사진

자동 생성된 설명

그림 14, Training set, Validation set 분리

본격적인 모델 학습에 앞서 주성분 분석을 통해 데이터 샘플의 각 특성들이 종속 변수인 Customer royalty에 대한 설명력을 충분히 가지고 있는지 확인한다.

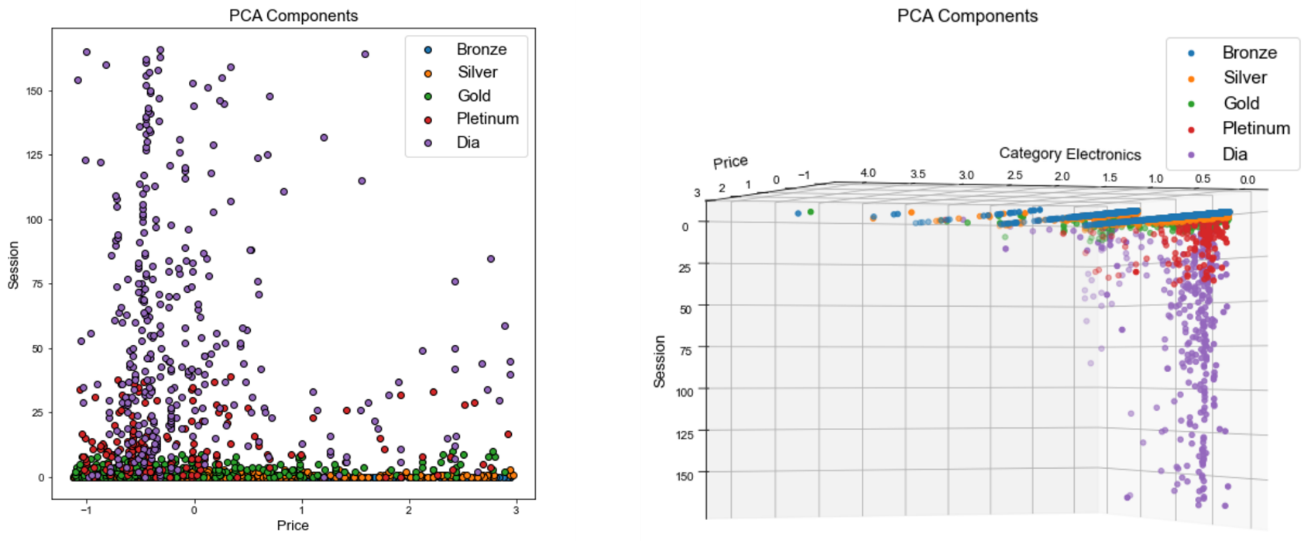


그림 15, PCA를 이용한 주성분 분석

공분산 행렬(PCA)을 활용한 주성분 분석을 진행한 결과 주요 성분이 되는 Session과 Price에서 종속 변수 Customer royalty에 대한 설명이 Session으로 치우쳐 있다. 분류 문제를 해결하기에 현재 데이터 샘플의 한계가 존재하며 이를 염두에 두고 모델 학습을 진행하여 정확도를 예측한다.

### **1. Random forest**

의사 결정 규칙을 나무구조를 통해 데이터를 분류하거나 예측하는 의사결정나무, Decision Tree를 앙상블 방식으로 사용한 알고리즘에 해당하는 Random forest의 경우, 모델 학습 시간은 약30분 정도로 다른 모델과 비교하여 예측 시간이 빠르다. 하지만 학습 과정의 정확도를 나타내는 Train Accuracy가 79.89%, 추론 과정의 정확도를 나타내는 Validation Accuracy가 60.98%로 학습 과정과 추론 과정에서의 정확도 차이가 약 19%으로 크게 나타났다. 이는 과적합(overfitting)의 가능성이 높게 나타나는 Decision Tree의 단점의 반영되어 전체 데이터 샘플을 설명하기에는 모델의 유연성이 떨어지는 일반화 능력의 부재로 예측된다.

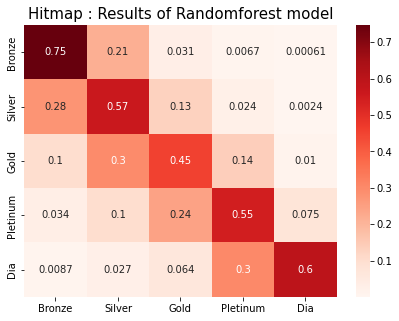


그림 16, Random forest의 Confusion matrix

### **2. K-Nearest Neighbor**

데이터의 입력을 통해 근접한 데이터들의 종류를 확인하고 분류하는 Classification의 대표적인 지도 학습 모델인 K-Nearest Neighbor의 경우, 약 325만개의 샘플을 가진 데이터를 사용하여 학습과 추론 과정에서 약 12시간이 소요되었으며 이는 우리가 사용한 모델 중 제일 느린 예측 시간을 보여줬다. 예측 시간이 매우 느려 최근접 이웃의 수에 해당되는 탐색할 이웃 수 k를 3으로 고정하여 학습하였다. Train Accuracy는 61.21%, Validation Accuracy는 57.01%로 학습과 추론의 정확도의 차이가 Random forest보다 감소하였지만 전체 모델의 정확도가 낮게 나타났다. One-hot 인코딩으로 인해 종속 변수를 제외한 특성이 64개로 많은 특성을 고려하기에는 알고리즘이 상대적으로 단순하고 특성이 0인 데이터 샘플이 자주 등장하여 가지고 있는 데이터로 정확도 개선에 어려움이 존재한다.

### **3. Deep Neural Network**

Deep Neural Network 모델 구현에서 입력으로 Customer royalty를 제외한 64개의 특성을 주고 은닉층을 설계한다. 하나의 은닉층은 완전연결층(FC), 배치 정규화(batch normalize), 활성함수(ReLU)로 구성되어 있다. 총 6개의 Layer로 구성되며 마지막에 완전연결층을 통해 출력으로 5개의 충성도 클래스로 분류될 수 있게 출력 벡터의 크기를 조정한다. Softmax함수를 통해 출력을 확률로 바꾸어 가장 확률이 높은 클래스를 선택한다. 손실함수, Cross Entropy를 통해 loss를 계산하고 역전파 과정을 통해 가중치를 개선한다. 과적합을 방지하기 위한 규제 기법으로 dropout을 사용하여 학습 과정에서 20%의 노드를 mask한다. Learning rate는 0.001로 설정하였으며 Optimizer는 Adam(0.9, 0.999)를 사용하였다.

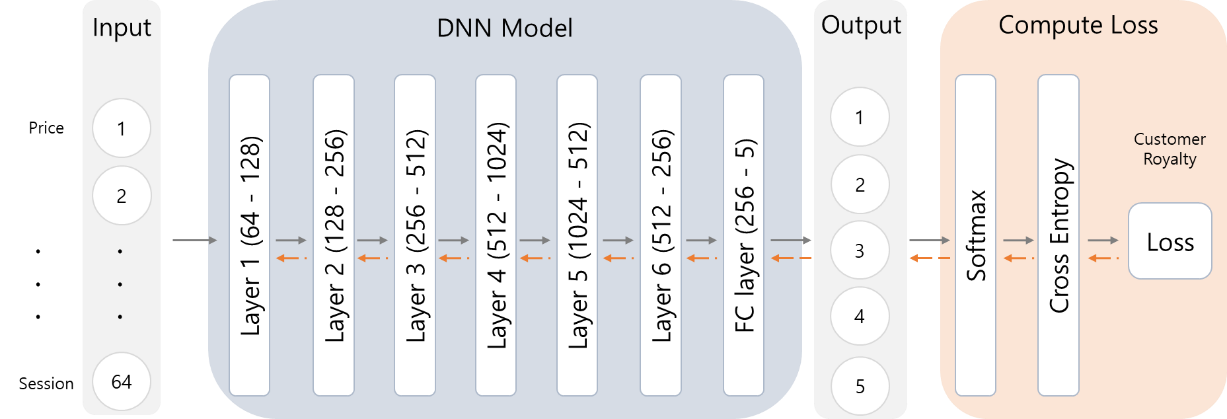


그림 17, 설계한 DNN Model

layer 개수는 고정하고 완전연결층의 채널 수를 바꿔가며 학습에 사용되는 파라미터 수를 조정하여 파라미터 수에 따른 정확도와 학습 과정에서 오답을 고려하여 정확도를 향상하기 위한 방법인 Boosting 기법을 적용한 결과 역시 비교하였다.

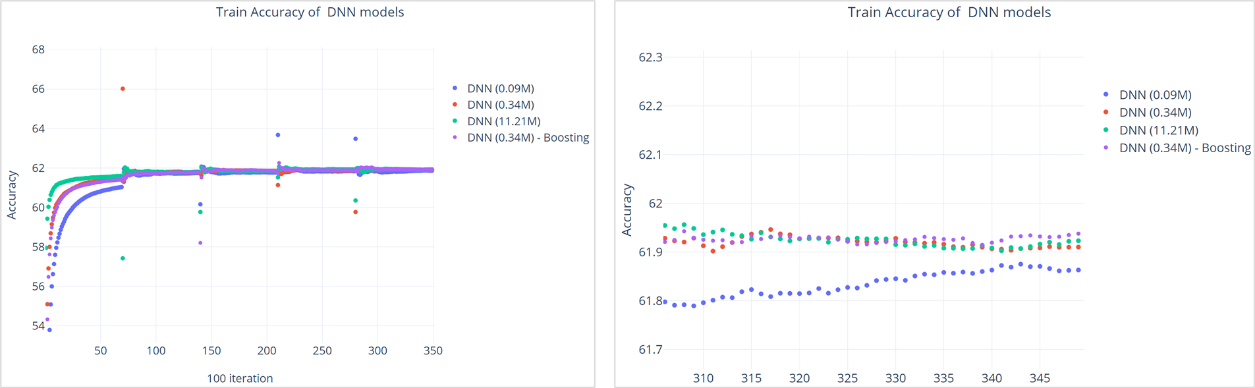


그림 18, 파라미터 수, Boosting에 따른 정확도 비교

학습 초기 과정에서의 차이는 존재하지만 epoch이 진행될수록 큰 차이를 보이지 않았다.

테이블이(가) 표시된 사진

자동 생성된 설명

그림 19, DNN Model에서의 Confusion matrix

|  |  |  |
| --- | --- | --- |
| **Model** | **Train Accuracy** | **Validation Accuracy** |
| **Random forest** | 79.89 | 60.98 |
| **KNN** | 61.21 | 57.01 |
| **DNN(0.09M)** | 66.01 | 66.81 |
| **DNN(0.34M)** | 63.67 | 66.89 |
| **DNN(11.21M)** | 62.06 | 66.90 |

Table 1, 모델 정확도 비교

학습 모델의 정확도를 비교한 결과 Train Accuracy는 Random forest에서 제일 높게 나왔지만 Validation Accuracy와 차이가 큼을 확인할 수 있다. DNN의 경우 파라미터 수에 따라 Train Accuracy의 차이가 존재하지만 Validation Accuracy에서 약 66%의 정확도로 비슷한 결과가 나타났다.

# **4. 보안점**

데이터 수집과 전처리 과정을 통해 획득한 e-커머스를 통해 한 번이라고 구매 경험이 있는 소비자의 구매에 관한 데이터를 통해 단순히 재방문의 여부(ex. 재방문 1, 재방문 하지 않음 0)을 예측하는 것은 일정 수준의 정확도 이상을 기대할 수 있으며 어려운 문제가 아니다. 하지만 재방문 횟수를 예측하고 이를 통해 Customer royalty를 예측하는 것은 전자의 문제보다 정확도가 떨어짐을 확인할 수 있었다.

학습에 사용된 데이터는 소비자에 관한 개인 정보가 없이 상품 브랜드, 카테고리, 가격, 구매 시간과 요일 등 단순 구매 이력 정보로 소비자의 특성을 표현하기에 능력이 부족한 것으로 판단된다.

# **5. 결론**

e-커머스 기업에서 소비자의 충성도를 예측하기 위해 이러한 모델을 사용한다면 그들은 소비자의 구매이력 뿐만 아니라 가입정보를 비롯한 소비자의 특성을 나타낼 수 있는 데이터가 존재할 것이다. 적절한 전처리 과정과 효율적인 모델의 선정이 이루어진다면 기업은 고객의 충성도를 높은 정확도로 예측할 수 있을 것이다. 이를 통해 기업은 예측된 소비자의 충성도를 통해 높은 충성도가 예측되는 고객 관리를 통한 충성 고객 확보 전략을 수립할 수 있다. 또한 낮은 충성도가 예측된 다면 그들을 끌어오기위한 유인 기법을 고민할 필요가 있다. 이를 통해 기업은 소비자 확보를 통한 이익의 증대를 소비자는 기업들의 마케팅을 통해 다른 e-커머스들과 비교하며 합리적인 소비를 할 수 있을 것으로 예상한다.