**WannaVetaran Feaure Engineering 보고서**

1. **결측치 처리**
2. 평수: 결측치 비율이 높아 ‘없음’으로 처리
3. 결재형태: 가장 많은 유형인 ‘신용카드’로 변경
4. 고객가입일: ‘2020-06-23’ 가장 많은 고객가입일로 변경
5. 부재중여부: ‘모름’으로 처리
6. 우선청소: ‘없음’으로 처리
7. 매니저사용휴대폰: 가장 많은 유형인 ‘안드로이드’로 변경
8. 매니저이동방법: 가장 많은 유형인 ‘대중교통’ 으로 변경
9. CS교육이수여부: 가장 많은 유형인 0으로 변경
10. 청소교육이수여부: 가장 많은 유형인 0으로 변경
11. **Feature 생성**
12. **접수월**

접수일에서 접수월을 추출

1. **접수일일**

접수일에서 접수일을 추출

1. **서비스이용기간**

서비스일자에서 최초서비스일을 뺀 날짜 기간

1. **매니저연령대**

매니저생년월일을 기준으로 1960년대생부터 10년 단위로 연령대 구분

1. **반려동물여부**

반려동물이 없으면 1, 있으면 0으로 구분

1. **종합지수**

매니저 매칭성공여부에 영향을 줄 것이라고 가정한 6개의 Feature를 수치화하여 이 점수가 높을수록 매칭성공에 여부에 긍정적인 영향을 줄 것이라고 가정

1. **서비스시간대**

서비스시작시간을 시준으로 12시 이전 그룹을 오전 12시 이후 그룹을 오후로 구분

1. **주거가중치**

집의 평수가 클수록, 주택의 유형이 일반주택이 아닐수록 청소서비스를 이용할 확률이 높을 것이라고 가정함

1. **서비스주소\_광역**

근무가능지역을 광역권 지역으로 분류

1. **서비스 시간**

서비스종료시간에서 서비스시작시간을 뺀 서비스 시간

1. **서비스대기기간**

최초서비스일에서 접수일을 뺀 것으로 접수일로부터 최초서비스일까지 걸린 시간을 계산

1. **매니저나이**

현재년도에서 매니저생년월일을 뺀 매니저 나이

1. **서비스이용기준**

전체회차의 평균값인 4.64를 반올림 한 5를 기준으로 장기고객과 단기고객을 분류

1. **고객가입일**

최초서비스일에서 고객가입일일을 뺀 기간으로 고객이 서비스를 탐색한 기간을 의미

**2-1 Feature 범주화**

접수일, 최초서비스일, 서비스일자, 고객가입입, 매니저성별, 매니저생년월일, 매니저생년월일, 접수월, 접수일일, 접수요일, 매니저나이, 매니저연령대의 11개 feature를 ‘object’형으로 변경

**2-2 Drop features**

**Train 제거**

train에서 매니저최초가입일, 매니저최초서비스일, 매니저주소는 Test와 동일하게 제거

**Train, Test 공통 제거**

접수시각 -> 결측치 비율이 너무 높아 제거

매칭성공여부 -> 제거

서비스시작시간, 서비스종료시간, 고객가입일 -> 시계열 데이터 제거

SEQ -> 제거

**2-3 범주형 데이터 원핫인코딩**

**3 불균형 데이터 처리**

데이터 불균형을 해결하기 위하여 SMOTETomek모델을 사용하여 oversampling을 진행함

최종 31062개의 데이터를 가지고 모델링을 진행함.

**4 모델링**

여러가지 모델링을 진행해본 결과 ExtraTreeClassifier를 사용하여 최종 모델링을 진행함.