프로세스 마이닝 알고리즘을 활용한 은행 로그 데이터 분석

데이터 및 프로세스 설명

본 분석에서 활용되는 이벤트 로그는 네덜란드 은행으로부터 추출된 실제 추출된 로그이며, 이는 은행에서의 대출 신청 프로세스에 관한 것입니다. 고객이 요청한 금액은 AMOUNT_REQ에 나타나 있으며, Case 속성으로 구성되어 있습니다. 본 이벤트 로그는 세 개의 서브 이벤트 로그를 합쳐서 구성되었으며, 이는 이벤트 로그 내 Activity의 첫 대문자로 구분할 수 있습니다. 분석을 수행할 때, 제공되는 전체 이벤트 로그 단위에서 혹은 각 서브 이벤트 로그 단위에서 분석해도 무방하니 다양하게 적용하기 바랍니다.

데이터 상세 설명

(1) 대출 승인의 전체 과정

대출 신청은 웹페이지를 통해서 제출되며, 먼저 자동화된 제출 내역 확인 과정과 이에 따른 추가 자료 보완이 진행됩니다. 만약 대출 자격을 만족하면, 은행 측 대출 제안이 메일을 통해 고객에게 전달됩니다. 고객으로부터 본 제안 내용을 다시 은행이 받게 되면 평가가 진행되고, 해당 내용이 불완전한 경우 고객에게 다시 연락하여 누락된 정보를 추가합니다. 그 후, 최종 평가가 수행된 후 대출 신청이 승인되고 활성화됩니다.

(2) 이벤트 타입 설명

이벤트 타입	의미
States starting with 'A_'	대출 신청의 상태
States starting with 'O_'	신청된 대출 관련 제안의 상태
States starting with 'W_'	신청된 대출 관련 작업의 상태
COMPLETE	'A_', 'O_' 종류의 작업이 종료됨
SCHEDULE	'W_' 종류의 작업이 Queue에 생성됨
START	'W_' 종류의 작업이 작업자에 의해 처리 시작됨
COMPLETE	'W_' 종류의 작업이 작업자에 의해 처리 종료되며, 그 후 다시 Queue에 돌아감

(3) 이벤트 번역

네덜란드 표기	영어 번역
W_Afhandelen leads	W_Fixing incoming lead
W_Completeren aanvraag	W_Filling in information for the application
W_Valideren aanvraag	W_Assessing the application
W_Nabellen offertes	W_Calling after sent offers
W_Nabellen incomplete dossiers	W_Calling to add missing information to the application

서론 (Introduction)

데이터 분석은 더 많은 정보를 제공하고 더 나은 결정을 내리도록 돕기 위해 사용된다. 목적에 따라 데이터 분석의 용도는 달라지겠지만 결국 새로운 정보를 찾아내 기존보다 더 나은 결과를 도출하는 것을 목적으로 하는 것이 데이터 분석의 필요성이다. 본 분석에서 활용되는 이벤트 로그는 네덜란드 은행으로부터 추출된 로그이며, 이는 은행에서의 대출 신청 프로세스에 관한 것이다. 주어진 이벤트 로그로부터 다양한 프로세스 마이닝 알고리즘을 활용한 데이터 분석을 통해 네덜란드 은행 대출 프로세스에 대한 새로운 Insight 를 도출하는 것을 목적으로 한다. 본 데이터 분석에서는 pm4py 를 사용하였으며 이벤트 로그 분석, 케이스 분석을 통한 다양한 Approach 를 통해 네덜란드 은행 대출 프로세스에 대한 여러가지 특성 및 주요 결과를 보여준다.

데이터 처리 (Data preprocessing)

본 분석에서 활용되는 네덜란드 은행 대출 프로세스 이벤트 로그를 Dataframe 으로 바꾸어 진행하며 num of events, num of cases, start and end activities, average of case durations 를 보여주는 basic_data_analysis 라는 함수를 사용해 본 데이터 분석을 진행하도록 한다. 본 데이터를 처리하기에 앞서 몇 가지 분석 질의들을 제시하며 데이터를 분석하도록 한다. 첫 번째 질의는 대출 신청 프로세스 관련 총 걸리는 시간이다. 전체 이벤트 로그 데이터를 basic_data_analysis 라는 함수 안에 적용해 본 결과 262200 event 에 대해 13087 개의 case 가 나왔고 이 case 들의 **대출 신청 프로세스 관련 총 걸리는 시간의 평균은 8.62 일**이다.

```
Number of events: 262200
Number of cases: 13087
Start activities: {'A_SUBMITTED-COMPLETE': 13087}
End activities: {'A_DECLINED-COMPLETE': 3429, 'W_Valideren aanvraag-COMPLETE': 2745, 'W_Afhandelen leads-
COMPLETE': 2234, 'W_Completeren aanvraag-COMPLETE': 1939, 'W_Nabellen offertes-COMPLETE': 1289, 'A_CANCELLED-
COMPLETE': 655, 'W_Nabellen incomplete dossiers-COMPLETE': 452, 'O_CANCELLED-COMPLETE': 279, 'W_Beoordelen
fraude-COMPLETE': 57, 'W_Wijzigen contractgegevens-SCHEDULE': 4, 'W_Valideren aanvraag-START': 2, 'A_REGISTERED-
COMPLETE': 1, 'W_Nabellen offertes-START': 1}
Mean of case durations: 8.62
```

<그림 1. 대출 신청 프로세스 관련 총 걸리는 시간>

그리고 대출 신청이 승인된 경우와 거절된 경우의 차이를 살펴보았다. **대출 신청이 승인된 경우** 'A_APPROVED-COMPLETE'에 대한 activity 를 살펴보았으며 pm4py.filter_eventually_follows_relation 을 사용하였다. 99925 event 에 대해 2246 개의 case 가 나왔고 이 case 들의 **대출 신청 프로세스 관련 총 걸리는 시간의 평균은 16.74 일**이 걸렸고, 대출 신청이 거절된 경우 'A_DECLINED-COMPLETE' 대한 activity 를 살펴보았으며 pm4py.filter_eventually_follows_relation 을 사용하였다. 70432 event 에 대해 7635 개의 case 가 나왔으며 **대출 신청이 거절된 경우 대출 신청 프로세스 관련 총 걸리는 시간은 2.05 일**이 걸렸다. 이와 별개로 대출 신청이 취소가 된 경우의 case 인 'A_CANCELLED-COMPLETE' 가 2807 번 나타났는데 이와 관련된 대출 신청 프로세스 관련 총 걸리는 시간의 평균은 18.6 일이 걸렸다. 따라서 대출 신청이 승인된 경우와 취소된 경우 거절된 경우의 걸리는

시간의 차이가 있다는 것을 알 수가 있다.

```
Number of events: 99925
Number of cases: 2246
Start activities: {'A_SUBMITTED-COMPLETE': 2246}
End activities: {'W_Valideren aanvraag-COMPLETE': 2046, 'W_Nabellen incomplete dossiers-COMPLETE': 194,
'W_Wijzigen contractgegevens-SCHEDULE': 4, 'A_REGISTERED-COMPLETE': 1, 'W_Nabellen offertes-COMPLETE': 1}
Mean of case durations: 16.74

Number of events: 70432
Number of cases: 7635
Start activities: {'A_SUBMITTED-COMPLETE': 7635}
End activities: {'A_DECLINED-COMPLETE': 3429, 'W_Afhandelen leads-COMPLETE': 2234, 'W_Completeren aanvraag-
COMPLETE': 1113, 'W_Valideren aanvraag-COMPLETE': 668, 'W_Nabellen incomplete dossiers-COMPLETE': 86,
'W_Beoordelen fraude-COMPLETE': 57, 'W_Nabellen offertes-COMPLETE': 48}
Mean of case durations: 2.05
```

<그림 2. 위 대출 신청이 승인된 경우, 아래 대출 신청이 거절된 경우>

두 번째 질의는 어떤 작업자들이 대출 승인 측면, 생산성 측면 등 업무 효율성이 좋은가 이다. 이 데이터 분석을 진행하기 위해서 Activity 의 종류들을 살펴보았고 이 중 이벤트 타입 설명 중 START, COMPLETE 가 'W_' 종류의 작업이 작업자에 의해 처리가 시작되고 종료됨을 나타내는 것이기에 단순 Resource 가 처리하는 작업 말고 'W_' 종류를 처리하는 작업자들의 업무 효율성을 분석해 보았다. 'W_' 종류의 작업에 대해 START, COMPLETE 가 있는 Activity 를 unique 하게 뽑아 찾아본 결과 'W_Completeren aanvraag', 'W_Nabellen offertes', 'W_Valideren aanvraag', 'W_Afhandelen leads', 'W_Nabellen incomplete dossiers', 'W_Beoordelen fraude' 등 6 가지의 Activity 들에 대해서 START, COMPLETE 가 나타났다. 먼저 생산성 측면에서 업무 효율성에 대해 분석한다. 본 분석에서는 위와 같은 Activity 가 포함된 pm4py.filter_trace_attribute_values 를 사용하여 trace 의 시간을 분석하는 것이 아니라 단순 하나의 Activity 가 START 되고 COMPLETE 되는 것에 대한 시간을 보기 위해 'time:timestamp'의 시간의 차이를 분석하였다. 먼저 pm4py.filter_trace_attribute_values 를 이용해 위와 같은 Activity 가 포함되어있는 trace 를 찾고 datetime 을 사용해 시간의 차이 분석을 진행하였다. 본 분석에서 업무 효율성에 대해 나타난 결과는 'W_Completeren aanvraag'의 경우 Activity Cycle 의 상위 10 개의 평균 시간을 살펴본 결과 8 분 48 초의 시간이 걸렸다. 'W_Nabellen offertes'의 경우 1 분 38 초, 'W_Valideren aanvraag'의 경우 17 분 26 초, 'W_Afhandelen leads'의 경우 2 분 16 초, 'W_Nabellen incomplete dossiers'의 경우 3 분 40 초, 'W_Beoordelen fraude'의 경우 18 초가 걸렸다. 본 분석에서 나타난 결과는 'W_Valideren aanvraag'의 작업이 작업자에 의해 처리가 시작되고 종료될 때까지의 시간이 제일 오래 소요되는 것을 알 수 있고, **'W_Beoordelen fraude'의 작업이 작업자에 의해 처리가 시작되고 종료될 때까지의 시간이 제일 적게 소요되는 것으로 보아 이 작업을 수행했던 작업자들의 생산성 측면에서 업무 효율성이 제일 좋다고 볼 수 있다**. 그리고 위와 관련된 Activity 를 수행할 경우 작업에 Case 마다 작업을 수행했던 Resource 가 다르기 때문에 단순히 어떤 Resource 측면에서 작업 효율성을 본 것이 아니라 Activity 측면에서 분석을 진행하였다.

```
W_Completeren aanvraag :  0 days 00:08:48.612444444
W_Nabellen offertes :  0 days 00:01:38.287222222
W_Valideren aanvraag :  0 days 00:17:26.026222222
W_Afhandelen leads :  0 days 00:02:16.108555555
W_Nabellen incomplete dossiers :  0 days 00:03:40.984444444
W_Beoordelen fraude :  0 days 00:00:18.762777777
```

<그림 3. 생산성 측면 업무 효율성>

다음은 대출 승인 측면에서의 업무 효율성을 분석한다. 대출 승인 측면에서는 'A_APPROVED-COMPLETE' 즉, 대출 승인된 경우 기여한 Resource 에 대해 분석을 진행하였다. 'A_APPROVED-COMPLETE'에 기여한 Resource 라고 해도 'A_DECLINED-COMPLETE'에 기여한 Resource 가 많이 존재하기 때문에 단순 'A_APPROVED-COMPLETE'에만 기여한 Resource 가 대출 승인 측면에서 좋은 업무 효율성이 있다고 판단하였다. 대출이 승인된 경우와 대출이 거절된 경우의 Resource 를 비교해본 결과 **Resource 10124, 10125, 10821 의 작업자들이 기여했던 경우 대출 승인**이 이루어 진 것을 알 수 있다. 이와 반대로 'A_DECLINED-COMPLETE'에만 기여한 Resource 가 존재했는데 **Resource 11304 작업자가 기여했던 경우 대출 승인이 거절**된 것을 알 수 있다.

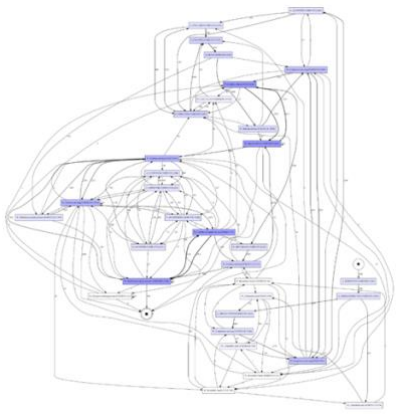
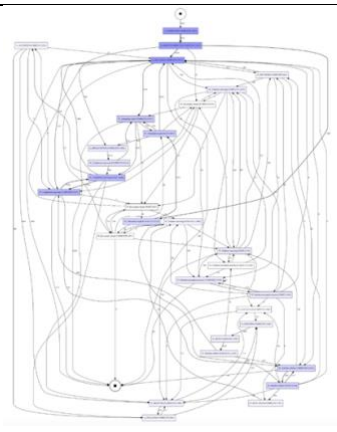
```
array([10124, 10125, 10821])
```

<그림 4. 대충 승인된 경우>

```
array([11304])
```

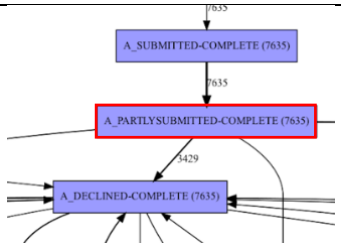
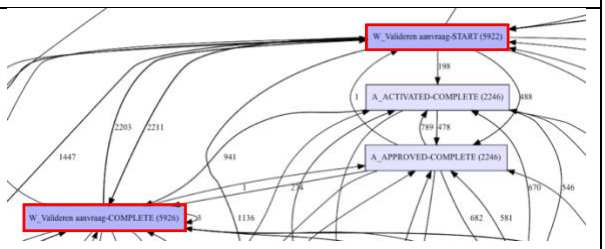
<그림 5. 대출 거절된 경우>

세 번째 질의는 대출 신청 프로세스 모델의 생김새이다. 대출이 승인된 경우와 거절된 경우의 차이를 보여주며, 이를 분석한다. 대출 신청 프로세스를 가장 직관적으로 볼 수 있는 모델은 바로 Directed Follows Graph(DFG) 이다. 이 DFG 를 이용하여 대출이 승인된 경우와 거절된 경우를 보았을 때, 직관적으로 차이를 바로 알 수 있다. 일단 Case 에서부터 크기의 차이가 있다. 대출이 승인된 경우에는 99925 개의 이벤트와 2246 의 케이스가 존재하고, 대출이 거절된 경우에는 70432 개의 이벤트와 7635 개의 케이스가 존재하는 것을 알 수 있다. 이를 통하여 단순 프로세스 모델을 구현해보았을 때 두 모델의 차이점이 있음을 알 수 있다.

대출 승인된 경우	대출 거절된 경우
	

<테이블 1. 왼쪽 : 대출 신청 승인된 경우, 오른쪽 : 대출 신청 거절된 경우>

네 번째 질의는 대출 신청 승인의 결과를 이끄는 주요한 요인은 무엇인가이다. 위 테이블의 그림을 세부적으로 살펴보자. 위 그림과 같이 대출이 승인된 경우의 분석을 할 경우 다음과 같이 주요한 요인을 찾을 수 있다. 대출 승인이 거절된 프로세스의 DFG 를 살펴보면 'A_PARTLYSUBMITTED-COMPLETE' 에서 'A_DECLINED-COMPLETE'로 가는 Arc 가 7635 개의 Case 중 3429 개의 Case 가 존재한다. 즉, 대출 승인이 거절된 경우 중 약 45%의 Case 가 'A_PARTLYSUBMITTED-COMPLETE'에서 나타나는 것을 보아 해당 Activity 가 대출 승인을 이끄는 1 차 주요한 요인이라 볼 수 있다. 대출 신청이 승인된 경우를 살펴보자. 'A_PARTLYSUBMITTED-COMPLETE' Activity 가 실행된 후 다른 주요한 요인을 찾게 되면 'W_Valideren aanvraag'에 관련된 Activity 즉, 응용 프로그램 평가에 대한 Activity 를 수행할 경우 대출 신청 승인의 결과를 이끄는 2 차 주요한 요인이라고 볼 수 있다.

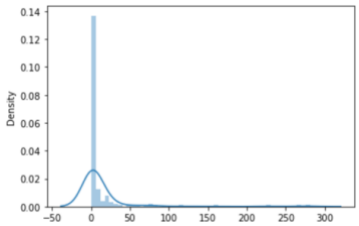
1 차 주요 요인	2 차 주요 요인
	

<테이블 2. 왼쪽 : 1 차 주요요인, 오른쪽 : 2 차 주요 요인>

마지막 질의는 대출이 승인되었을 경우 고객이 요청한 금액('AMOUNT_REQ')이 어느 정도였을 경우가 많은 지 존재 여부이다. 본 분석을 진행하기 위해서는 대출이 승인되었을 경우의 Case 를 찾아 'case:concept:name'의 unique 한 값들로 Groupby 를 해주어 Case ID 마다 고객이 요청한 금액 즉, 'AMOUNT_REQ'를 구하였다. 그리고 'AMOUNT_REQ'들의 value_counts 를 통하여 'AMOUNT_REQ' 별 대출이 승인된 경우의 수를 알 수 있다. 총 대출 승인 횟수인 2246 번 중 아래와 같이 'AMOUNT_REQ'이 15000$일 경우에 281 건으로 가장 많은 대출 승인되었고, 그다음으로 5000$가 269 건, 10000$가 226 건, 25000$가 160 건, 20000$가 117 건 순으로 대출 승인이 이루어진 것을 알 수 있다. 아래 Density plot 과 같이 고객이 요청한 금액별로 많이 존재하는 Case 도 있지만 'AMOUNT_REQ' 금액이 1 개씩 존재하는 경우가 많이 존재한다는 것을 확인할 수 있다.

15000	281
5000	269
10000	226
25000	160
20000	117

<그림 6. 'AMOUNT_REQ' Top 5>



<그림 7. Density plot>

데이터 분석 결과 (Results)

첫 번째 질의 같은 경우 대출 신청 프로세스 관련 총 걸리는 시간 평균이 8 일인데 비해 대출이 승인된 경우에는 16 일이 소요되고 대출이 거절된 경우에는 평균 2 일이 소요되기 때문에 대출이 승인되기까지 시간이 많이 소요된다는 것을 알 수 있다. 두 번째 질의에서는 어떤 작업자들이 대출 승인 측면, 생산성 측면 등 업무 효율성이 좋은 지 분석해 보았다. 생산성 측면에서는 작업자에 의해 'Activity'의 START-COMPLETE 되는 시간을 분석하였는데, 이 작업에서 평균 작업 승인 시간이 제일 적게 소요된 'W_Beoordelen fraude' Activity 를 처리하는 작업자들의 업무 효율성이 좋다고 볼 수 있으며 'W_Afhandelen leads'와 관련된 Activity 경우에는 생산성 측면에서 대출 승인 평균 소요 시간이 오래 걸리기 때문에 이와 관련된 작업자들의 업무 효율성이 좋지 않다고 볼 수 있다. 대출 승인 측면에서는 Resource 10124, 10125, 10821 의 작업자들이 대출 신청에 기여했던 경우 대출 신청이 승인되었음을 볼 수 있다. Resource 11304 작업자의 경우에서는 대출 승인이 거절된 경우가 있으므로 대출 승인 측면에서 업무 효율성이 좋지 않다는 결과를 도출할 수 있다. 세 번째 질의에서는 DFG 를 통하여 대출이 승인된 경우와 대출이 거절된 경우의 모델을 비교했을 때 직관적으로 두 모델의 차이점을 확인할 수 있으며 이를 세부적으로 분석하게 되면 네 번째 질의에서 1 차 주요 원인인 'A_PARTLYSUBMITTED-COMPLETE' 와 2 차 주요 원인인 'W_Valideren aanvraag'에 관련된 Activity 의 유무에 따라 대출 신청이 승인되고 거절됨을 알 수 있다는 결론을 도출할 수 있다. 마지막 질의로 대출이 승인되었을 경우 고객이 요청한 금액이 어느 정도의 금액일 때 대출 승인이 많이 되었는지 유무이다. 이 분석을 통하여 15000$일 경우에 281 건으로 대출 승인이 많이 된 것을 알 수 있으며 Density plot 을 통하여 고객 대출 금액별 승인 횟수의 분포를 확인할 수 있었다.

결론 (Conclusion)

본 데이터 분석을 통해 네덜란드 은행 대출 프로세스에 대한 다양한 정보들을 도출해 보았다. 대출 신청 승인과 거절의 차이점과 대출 신청 프로세스의 세부 내용들에 대한 분석을 토대로 대출 신청이 어떻게 진행되는지 알아보았고, 대출 신청이 승인되는 주요한 원인을 도출해낼 수 있었다. 본 분석에서 여러 방면으로 데이터를 분석하려 하였지만 본 분석에서 다루지 못한 질의들도 존재한다. '대출 신청 프로세스 시간이 오래 걸리는 경우 이에 대한 원인이 고객인지 혹은 은행인지'에 대한 질의에 대해서는 조금 더 분석을 진행해야 할 필요가 있는 문제다. 이외에도 Process Enhancement 측면에서 다양한 추가 분석 및 pm4py 외 다른 프로그램들을 사용하여 데이터를 분석해보면 좋을 것 같다는 생각이 든다. 최종적으로 본 분석을 통하여 네덜란드 은행 대출 프로세스 분석에 이전에 발견하지 못했던 새로운 Insight 분석 향상을 기대한다.