빅콘테스트 2016 설명회

보험사기자 예측 알고리즘 개발

2016.8.2 | 한화생명 빅데이터센터

한화생명

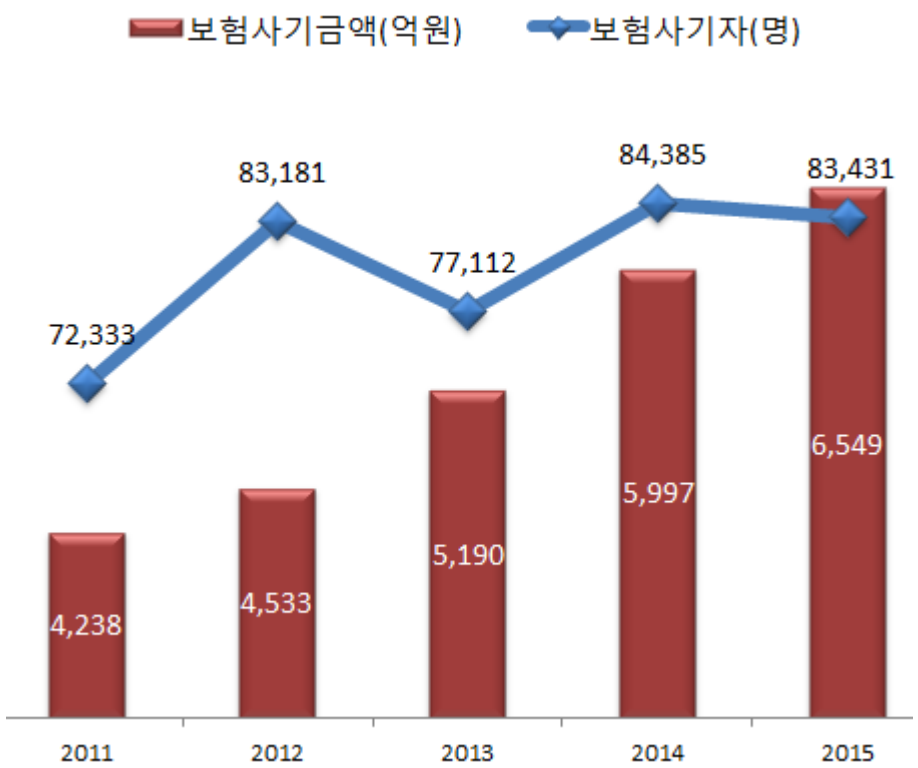
Contents

한화생명

보험사기?

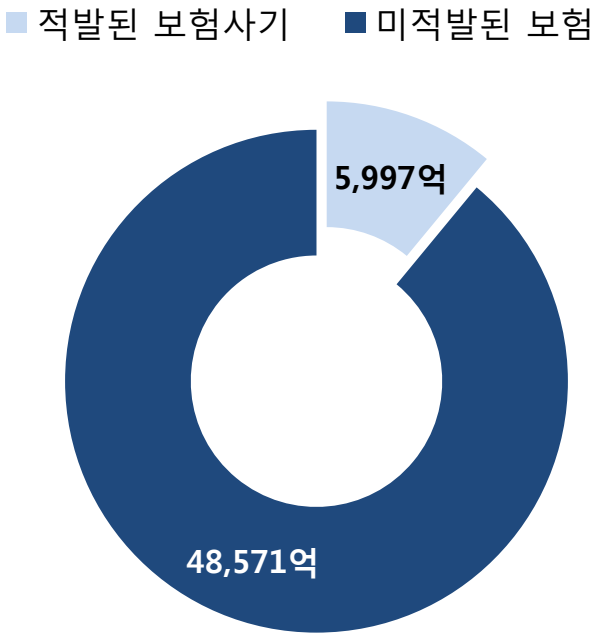
보험사고의 발생, 원인 또는 내용에 관하여 보험자(보험회사)를 기망하여 보험금을 청구하는 행위.「보험사기방지 특별법 제2조」

연도별 보험사기 적발 추이



보험사기 적발률 (2014년 기준)

[출처 : 금융감독원 국회 제출자료]



한화생명

Sustained Painpoint of Insurer

Loss

- **직접 손실**
  - 보험사기에 대한 보험금 지급으로 인한 직접적 손실
- **영업 손실**
  - 보험료 상승으로 인한 보험소비자의 보험수요 감소로 발생하는 영업손실
- **부가적 손실**
  - 보험상품개발 자율성 제약
  - 일반고객 박탈감 증대
  - 보험산업 전체에 대한 신뢰 감소 등

Limitation

- **인력 부족**
  - 보험사기를 전담하여 조사 및 대응할 절대 인력의 부족
- **비용의 한계**
  - 건당 조사비용 발생으로 인한 비용적 한계 존재
- **불확실성**
  - 대부분의 고객은 선량한 고객 (보험금 지급 시 보험사기 Frame이 형성되는 것에 대한 고객의 불쾌감)

한화생명

Data 특성 개요

1. 실제 Data를 기반으로 한 현실세계의 기업 내부 Data

- 콘테스트 용이 아닌 실제 기업내부 Data 기반의 데이터 Sample
- 현실세계 특유의 불친절한 Data 특성 존재
  (Null값이 왜 이렇게 많아? / 값이 이게 맞나? / 다른 형태의 Data가 더 필요한거 아닌가?)

2. 누구나 쉽게 분석을 시도할 수 있는 Data-Set 구성

- Sampling 과정에서 보험사기자의 비율 일정 조정
- Theme 별 Table을 구분하여 구성하여 제공 (데이터 파악에 대한 혼란 최소화)
- 간단한 JOIN 작업만으로 충분히 자신만의 분석 framework을 그려나갈 수 있음.

3. 누구나 한번쯤은 좌절할 수 있는 Data의 한계 존재

- 주어진 Data로 보험사기의 '원인' 에 대한 파악은 실질적으로 불가능
- 추적이 불가능하도록 비식별화 하는 과정 및 군집화 과정에서 데이터정보의 손실 존재함
- 일정 수준의 Data 핸들링 필요

2 데이터 및 변수설명

한화생명

BGCON_CUST_DATA

- 고객의 특성을 나타내는 Data
- SIU_CUST_YN 이라는 최종 보험사기 구분 Factor 포함(학습용 Set에만 결과가 있음)
- 고객의 성/연령/거주지/직업/배우자/소득 및 신용등급 정보 등 포함

BGCON_CNTT_DATA

- 고객들의 계약 속성을 나타내는 Data
- 고객과 연관된 계약들의 상품종류 및 상태변화 및 보험료 수준 등
- 고객테이블과 CUST_ID 값을 Key값으로 하여 Join 가능

BGCON_CLAIM_DATA

- 고객들을 대상으로 한 지급 속성을 나타내는 Data
- 언제 / 어떠한 사유로 / 얼마의 보험금이 지급 되었는지에 대한 정보를 포함함
- CNTT_DATA와 POLY_NO값을 Key로 하여 JOIN 가능

2 데이터 및 변수설명

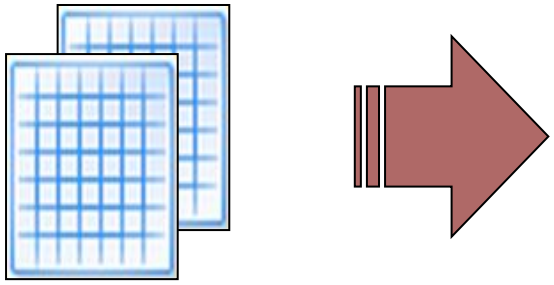
한화생명

BGCON_FMLY_DATA

- 고객간 가족여부를 알 수 있는 Data
- 보험사기의 경우 다수가 연계하여 발생하는 경우가 많으므로 Network 분석 등으로 접근하는 참가자를 위하여 해당 정보를 제공

BGCON_FPINFO_DATA

- 보험설계사 정보. 보험설계사의 재직기간 등을 알 수 있는 Data
- 고객대비 보험에 대한 이해도가 높음. Network 분석을 위한 Data



Data 구성

- 총 5개 Table
- 총 90개 이상의 Columns

※ 개인PC에서의 작업이 가능한 영역내에서의 최선의 Big-Data

한화생명

제출된 Test용 Set의 결과 구분 및 평가 방식

구분		예측 (Predicted)	
		일반고객 (Predicted)	보험사기자 (Predicted)
실제 (Actual)	일반고객 (Actual)	**True/Negative**	**False/Positive**
	보험사기자 (Actual)	**False/Negative**	**True/Positve**

- **Precision : 예측한 실제 보험사기자 수 / 예측한 보험사기자 전체 수**
  ⇒TP/(FP+TP)
- **Recall : 예측한 실제 보험사기자 수 / 실제 보험사기자 전체 수**
  ⇒ TP/(FN+TP)
- **F-measure : Precision 과 Recall의 조화 평균**

$$\Rightarrow \frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}$$

한화생명

집단 지성의 힘

⇒ 제공된 변수들의 조합으로 생각지도 못했던 중요한 변수들이 새롭게 발견되어 향후 활용 가능성 기대

외부 데이터의 연계 활용 가능성

⇒ 제공된 데이터 이외의 공공데이터를 포함한 외부 데이터 연계를 통한 새로운 인사이트 발굴 기대

다양한 분석 알고리즘의 활용

⇒ 새롭고 다양한 분석 알고리즘의 활용으로 한화생명 내부의 실무 분석 담당자들의 신선한 긴장을 기대

Tip!

Q & A