**데이터 사이언스 SCHOOL 사전 과제**

이름 : 김지훈

|  |
| --- |
| 관심 분야의 비즈니스 문제를 데이터 분석 문제로 변환하고, 그 문제를 해결하기 위한 방법을 자세히 서술해주세요.  (데이터 분석 문제로 변환하는 것부터 데이터 수집, 분석까지의 과정을 아는 만큼 정리해주세요. 분석 방법을 자세히 서술하기 어렵다면, 분석 문제를 정의하는 것 까지만 자세히 서술해주셔도 좋습니다.) |

**0. 분석의 이유**

손해보험사에서는 수 많은 업종이 가지는 위험에 대해 담보를 하며 손해율을 관리하기 위해 다양한 언더라이팅 기법을 사용하고 있다. 하지만 모든 계약에 대해서 언더라이터가 보유율을 결정하는 것은 현실적으로 불가능하며 이에 따라 대부분의 손해보험사는 자동 언더라이팅 시스템을 도입함으로서 상당한 수의 계약을 처리하고 있다. 그러나 손해보험사에서 사용하는 자동 언더라이팅 시스템은 각각의 계약에 대한 위험요소를 반영하기보다는 극히 일부의 요소에 전적으로 의존하여 보유율을 결정하고 있다. 자동 언더라이팅 시스템이 처리하는 계약들의 손해율이 높음을 감안할 때 보다 정교한 자동 언더라이팅 시스템을 도입할 경우 손해율을 낮추고 업무의 효율을 극대화 할 수 있을 것이다. 따라서 보다 효율적이고 정교한 자동 언더라이팅 시스템을 만들어보고자 한다.

**1. 데이터 분석 문제로 변환하기**

언더라이터가 보유율을 결정하는 요소는 여러가지가 있다. 그 중에서도 업종, 건물의 구조급수, 가입금액 등은 보유율에 많은 영향을 끼치는 요소이다. Multiple Regression 또는 Deep Learning 모델을 사용할 수 있으며 Multiple Regression을 이용할경우 각각의 Feature의 Weight을 최적화 시킴으로서 주어진 데이터를 통해 최적의 보유율을 산출해볼 수 있다. Deep Learning을 통해서도 최적의 보유율을 산출하는 시스템을 구축해 볼 수 있으며 두 모델 중 더 적합한 모델을 비교해 보는 것 또한 의미가 있을 것으로 생각한다. 최적의 모델을 통해 신규계약의 위험에 대한 정보가 주어졌을 경우 인수여부 및 보유율을 산출하는 시스템을 만들 것이다.

**2. 데이터 수집**

금융정보는 개인정보 규제로 다른 자료에 비해 오픈데이터를 구하기가 어려운 편이다. 따라서 언더라이팅실무에서 결정했던 업종별 보유율 데이터 및 보험통계포털서비스(http://www.insis.or.kr), 보험개발원(http://www.kidi.or.kr)의 자료를 사용하고자 한다. 또한 통계자료에 근거한 손해발생 시나리오를 시뮬레이션하여 현실적인 요소를 반영할 것이다.

**3. 데이터 탐색**

계약의 보유율에 영향을 주는 요소를 시각적으로 표현하여 데이터 분포를 확인하고 각 요소가 최적화 함수에 주는 영향을 개별적으로 표현함으로써 상관관계의 적합성을 확인할 수 있도록 할 것이다.

**4. 데이터 분석**

주어진 데이터를 통해 보유율에 영향을 줄 수 있는 요소를 가공하기 편한 형태로 Data Cleansing한 뒤 Multiple Regression이나 Deep Learning 기법을 이용하여 최적화 함수를 만들 것이다. 각각의 Feature가 가지는 가중치를 시각화하여 보유결정에 중요한 요소를 표현하고 산출된 보유율과 실제 사용되고 있는 기준보유율의 결과를 비교함으로써 효과를 검증할 것이다.

**5. 인사이트**

기존의 보유결정 데이터를 통해 신규계약에 대한 인수여부 및 최적의 보유율을 산출함으로써 손해보험사의 포트폴리오 개선에 기여할 수 있다고 생각한다.