중고차 시세 예측 관련 논문 요약

2015019898 이동환

1. 서론

전체 자동차 시장에서 중고차 시장이 차지하는 비율은 신차보다 크다. 중고차 시장에 대해 연구하는 것은 의미가 있고 시장 안에서 가장 중요한 요소는 가격이다. 온라인 검색은 정보 불균형으로 일어날 수 있는 영향을 줄인다. 결과적으로, 소비자들은 더 나은 결정을 할 수 있다. 우리는 소비자들의 의사결정을 돕는 방법론과 웹사이트를 소개한다. 방법론은 모든 중고품에 적용 가능 하다.

생존 분석에서는 대상들이 죽을때까지 혹은 실험이 끝날때까지 관찰된다. 고려되는 데이터는 일단 대상의 어떤 특징이 변하지 않는(그대로 있는) 시간이다. 예를 들어서 대상이 끝까지 살아남으면 해당 데이터는 중도절단된다. 중도절단 데이터로는 언제 이벤트가 일어나는지 알 수 없지만 중도 절단된 해당시간 이후라는 것은 알 수 있다. 생존 분석은 의료과학에서 주로 쓰이는데 특히 새로운 치료법이나 약물 실험에 쓰인다.

이번에는, 생존 분석을 통해 이벤트 타임(판매가 일어나는 시간)을 분석한다. 호가는 판매가 일어나는 시간을 결정하는 요인이다. 본 논문에선 호가랑 세부사항이 주어진 차량에 대해 미리 정해놓은 기간(여기선 30일)안에 차량이 팔릴 가능성을 알려주는 방법론을 소개한다.

2. 방법론

터키에서 가장 유명한 중고품 사이트에서 3개월치 총 **83만개**의 데이터를 수집했다. 최대 30일까지만 관찰했다(즉 30일이 넘으면 중도절단). 이전 데이터 수집 단계에서, 딜러들이 임의로 판매목록을 삭제했다가 다시 올리는 것을 발견했다. 우리는 private seller의 판매목록을 더 선호했다. 그 후에 연식이 매우 오래된 차인데 주행거리가 매우 낮은 것들과 같은 차량을 데이터에서 지웠다. 10년이상 된 차도 지웠다. 다 준비를 하니 **45만3천개**의 데이터가 남았다.

이 분석에서 주로 나타나는 단점은 예정된 이벤트가 일어날 때까지는 정보를 수집할 수 없다는 점이다. 관찰 기간이 끝난 이후에도 이벤트가 안일어날 수도 있다. 이렇게 중도절단된 데이터도 더 나은 결과를 위해 쓰이긴 한다. **생존분석의 접근법 중 비모수적 방법론의 장점은 단순한 계산과 결과의 명료함이다. Kaplan-Meier은 비모수적 방법론중의 제일 유명한 방법이다. 반면 모수적 방법은 분포에 대한 가정이 타당하지 않더라도 비편향적이다.** Cox 회귀 모델은 견고함 때문에 생존분석 모델중 가장 널리 쓰인다. 그 모델은 공변량 집합과 생존 사이의 관계를 검사한다. cox 회귀에서 위험함수는 죽음의 상대적 위험을 측정한다. 비례위험모형은 상대적 위험에 대한 공변량의 영향이 항상 상수라는 것을 내포한다. 시간에 영향을 받는 공변량이 있으면, 비례 가정이 깨진다. 확장된 cox 회귀 모형은 시간의존적 공변량이 위험함수 h(t)에 예측변수(독립변수)로 포함되어있으면 쓸 수 있다.

3. 생존분석을 통한 이벤트 타임 예측

우리가 분석을 해봤을 때, 제조사랑 모델 별로 데이터를 묶는게 좋았다. 504개의 페어가 있었고 그 중 몇몇 페어의 데이터는 굉장히 적었다. 데이터가 적은 페어들은 쓰기 힘들어서 100개 이하의 데이터가 있는 것들은 제거했다. 결국 147개의 페어가 남았고 약 **44만8천개**의 차량이

남았다.

cox 회귀는 SAS의 PHREG(parametric proportional hazards regression) 통해서 할 수 있고 변수를 5개 이하로 제한했다. 카이스퀘어에 기초해서 PHREG는 5개의 베스트 변수를 뽑아냈다. 기본적으로 cox 회귀 147개에 가장 많이 포함된 공변량 7개를 표1에 나타냈다. '가격'같은 경우는 147 cox 회귀 모델중 141번 쓰였다. '연간 키로미터'는 유도된변수(인위적으로 만든) '킬로미터'를 '연식'으로 나누어서 만들었다. 이런 특정 변수도 중요한 인자로 작용할 수 있다는 것을 보여주는 논문이 있다.

4. 예측 모델에 기초한 로지스틱 회귀
이번 섹션에서는 로지스틱 회귀에 기초한 예측 모델을 소개한다. 로지스틱 회귀의 특징 때문에 우리 모델은 차 팔릴 확률을 알 수 있다. 일단, 독립변수들의 선형 함수이고 buyer/seller에게 차가 팔릴 확률 중요한 정보를 준다. 모수 결정되면, 차가 팔릴 p(확률) 구할 수있다. 로지스틱 회귀의 모수는 최대우도추정법으로 구할 수 있다. 로지스틱 회귀 모델은 각각의 제조사/모델 페어 이용해서 만들어진다. 측정된 모수는 SQL 표에 저장된다. 우리는 이 모듈이 우리 의사결정 툴에 중요한 요소라고 생각한다. 예측 모델은 처음보는 데이터 포인트에도 테스트를 해봐야 하지만 우리의 처음 목적은 모아진 데이터에 대해 이용 가능성을 보여주는 것이었다.

5. 결론 및 토의
우리는 e-commerce 사이트에서 모은 데이터로부터 성공적으로 생존 분석 모델을 적용시켰다. 다른데도 적용할 수있지만 우리는 일단 중고차 시장 기저에 깔려있는 프로세스를 이해하기 위한 통계 모델을 만들어 보았다. 최근 연구에 비추어 보면, 추후에 가능한 주제 중 하나는 온라인과 오프라인 사이의 가격 차이를 이해하기 위한 가격 정보를 포함하는 것이 될 수 있겠다. **가격결정모델을 넘어, 소비자들에게 결정을 돕는 툴을 제공하는 것은 아마 중고차 시장가 결정에 도움이 될 것이다.** 미래의 데이터에도 잘 맞는 예측 모델을 도와주는데 계속적인 데이터 수집은 도움이 될 것이다. 로지스틱 말고도 이용할 수 있다. 그러나 많은 주의가 요구될 것이다.