**통합 입찰 공고 내역 데이터를 통한**

**초중고별 상위품목 확인 및 수요 예측하기**

**김태경 김혜인 이은지 이윤진 이지연**

2022년 02월 15일

**목차**

**1. 서론** ·············································································

1.1 분석 추진 배경 및 목적 ····································································

1.2 분석 결과 활용 방안 ········································································

1.3 활용 데이터 및 정제 방안 ································································

**2. 본론** ············································································

2.1 분석 프로세스 ················································································

2.2 분석 내용 및 코드 설명 ··································································

2.3 모델링 평가 및 예측······································································

**3. 결론** ············································································

3.1 결과 해석 ······················································································

3.2 활용방안 및 발전방향 ·····································································

\*목차는 혹시 몰라 첨부했고, 보고서 작성할 때 목차 보면서 흐름 익히고 정리하는 편이 저는 조금 더 편해서 간단하게 만들어놨습니다.

**1. 서론**

**1.1 분석 추진 배경 및 목적**

**1.1.1 추진 배경**

조달청을 통해 물품조달을 할 수 있는 수요기관은 국가기관, 교육기관, 지방자치단체, 정부투자기관 등이 있다. 이 기관들을 기준으로 계약 건수들의 순위를 비교해보고자 최상위기관별 계약변수 보고서를 조사해보았다. 현재 조달청에서 제공하는 최상위기관별 금액 구간에 따른 계약 건수 보고서에 따르면 최근 4개년간 교육부에서 가장 많은 계약 건수를 차지하고 있음을 알 수 있다. 자료를 보면, 2017년 교육부의 계약건수는 약 7300건, 2018년은 약 7900건, 2019년은 약 9600건, 2020년은 약 10800건으로 매년 교육부의 계약 건수는 증가하고 있음을 알 수 있다. 이 자료를 바탕으로 가장 많은 계약건수를 차지하고 있는 교육부가 조달하고 있는 물품은 어떤 것인지 알아보고, 물품 빈도에 대해서 조사해 보았다. 더 나아가 교육부의 하위기관인 학교를 초등학교, 중학교, 고등학교로 세분화하여 상위 조달물품을 조사해보았다.

**1.1.2 분석 목적**

통합 입찰공고 내역 데이터를 살펴보면 업무구분, 공고기관명, 수요기관명에 따라서 조달 물품을 조회할 수 있다. 하지만 학교별 상위 조달 물품을 한 번에 알 수 있는 방법이 없다. 학교 전체의 물품 조달건수는 2017년에는 약 44600개, 2018년에는 약 44400개, 2019년에는 약 44200개로 총 물품 조달건수는 매년 변화가 거의 없음을 알 수 있다. 따라서 매년 수요기관에 따른 상위 수요물품과 수요량을 조사하면, 앞으로의 패턴을 예측할 수 있을 것이다. 수요물품과 수요량 예측을 통해서 공급기관에서 조달물품의 양을 파악할 수 있을 것이고 더 나아가 예산을 효율적으로 짤 수 있을 것이다. 이를 통해서 초등학교, 중학교, 고등학교 별 상위 조달품목을 분석해보고자 한다.

**1.2 분석 결과 활용 방안**

통합 입찰 공고 내역 데이터를 이용하여 초등학교, 중학교, 고등학교별 조달된 물품을 월별로 분석하면 월별 상위 조달 품목이 무엇인지 알 수 있다. 또한 각 월별 상위 조달 품목을 그래프로 시각화하여 나타내면 월별 동향 또한 파악이 가능하다. 2017년부터 2020년까지 총 4개년의 월별 상위 조달 품목이 분석된 데이터를 이용해서 분석 모델을 만들면, 분석 모델을 통해 향후 년도의 조달 품목이 예측이 가능해지고, 공급기관 및 수요기관의 예산 설계를 위한 수요 예측 및 반영이 가능해진다.

**1.3 활용 데이터 및 정제 방안**

| 수집 데이터 | 데이터 설명 | 수집유형 | 데이터  수집 기관 | 데이터  기준일 |
| --- | --- | --- | --- | --- |
| 통합 입찰공고  내역 | 업무구분, 공고게시일, 공고명, 공고기관, 수요기관, 추정가격 등 | CSV | 조달청 | 2017.1~2021.12 |

**2. 본론**

**2.1 분석 프로세스**



방안 (간단한 설명이 있으면 좋을 것 같아 추가하였습니다. 분량이 넘어간다면 삭제할게요)

1. 공공데이터 포탈의 조달청\_통합\_입찰공고 내역을 데이터로 선정
2. 총 N개년 자료를 기준으로 초중고별 월별로 가장 많이 파는 물품을 상위 자료를 선정.
3. Ridge모델과 lightgbm 모델을 이용하여 모델링을 진행.
4. 상위 품목 건수를 2021년 데이터와 비교하여 결과 확인

**2.2 분석 내용 및 코드 설명**

**2.2.1 데이터 수집**

공공데이터포털에 있는 조달청\_통합\_입찰공고 내역을 데이터로 선택하였다. 2017년 1월부터 2021년 12월까지의 자료를 데이터로 사용했다. 2017년부터 2020년까지의 데이터 중 무작위로 추출하여 학습 및 검증데이터로 활용하였으며 2021년 데이터는 예측데이터로 활용하였다. 또한 초중고별 월을 기준으로 EDA를 하기 위해 month칼럼을 추가하였으며 ‘가격’ 칼럼에서 빈 데이터를 0으로 바꾸었다.

**2.2.2 EDA**

1)초중고별 월별 물품 가격의 합(전체년도)

|  | 초등 | 중등 | 고등 |
| --- | --- | --- | --- |
| 2017년 |  |  |  |
| 2018년 |  |  |  |
| 2019년 |  |  |  |
| 2020년 |  |  |  |

월별 물품을 가장 많이 사는 시기를 파악하기 위해 2017년부터 총 4개년 월별 물품 가격의 합을 나타내는 그래프를 작성하였다. 월별 물품을 확인한 결과 초등학교에서는 방학이 끝나고 새학기가 시작되는 전달 2 압도적으로 가장 많이 사는 것으로 추측할 수 있었다. 다른 달은 중고등학교 금액 합과 비슷했지만 2월달엔 평소의 금액보다 2배이상 높았다. 이와 달리 중학교에서는 방학(1월, 7월, 12월)에는 물품 금액의 합이 학기중보다 낮았고 매년 수요가 높은 달이 달랐다. 고등학교에서는 방학과 학기중이 큰 차이가 없었으면 매년 11월 금액 합이 높은 것으로 파악되었다.

2) 빈도수 상위 20개 물품

| 연도 | 2017년 | | | 2018년 | | | 2019년 | | | 2020년 | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 순위 | 초등 | 중등 | 고등 | 초등 | 중등 | 고등 | 초등 | 중등 | 고등 | 초등 | 중등 | 고등 |
| 1 | 공고 | 구매 | 구매 | 공고 | 구매 | 공고 | 공고 | 구매 | 구매 | 공고 | 중학교 | 구매 |
| 2 | 구매 | 중학교 | 공고 | 구매 | 중학교 | 구매 | 구매 | 중학교 | 공고 | 학년 | 공고 | 공고 |
| 3 | 견적 | 공고 | 자재 | 견적 | 공고 | 견적 | 학년 | 공고 | 견적 | 초등학교 | 구매 | 고등학교 |
| 4 | 제출 | 학년 | 견적 | 제출 | 학년 | 수의 | 초등학교 | 학년 | 고등학교 | 구매 | 학년 | 학년 |
| 5 | 소액 | 입찰 | 제출 | 수의 | 학교 | 고등학교 | 견적 | 학교 | 제출 | 견적 | 입찰 | 견적 |
| 6 | 금액 | 학교 | 고등학교 | 초등학교 | 입찰 | 제출 | 급식 | 입찰 | 수의 | 급식 | 학교 | 학교 |
| 7 | 학교 | 제출 | 수의 | 소액 | 제출 | 소액 | 우유 | 견적 | 소액 | 수의 | 교복 | 자재 |
| 8 | 초등학교 | 견적 | 소액 | 급식 | 견적 | 구입 | 제출 | 제출 | 자재 | 제출 | 하복 | 입찰 |
| 9 | 식용 | 소액 | 실습 | 학년 | 교복 | 자재 | 수의 | 수의 | 학교 | 입찰 | 주관 | 제출 |
| 10 | 학년 | 수의 | 입찰 | 학교 | 수의 | 안내 | 입찰 | 소액 | 안내 | 소액 | 동복 | 수의 |
| 11 | 공통 | 교복 | 구입 | 입찰 | 소액 | 학교 | 소액 | 교복 | 구입 | 간식 | 견적 | 구입 |
| 12 | 전자 | 하복 | 학년 | 우유 | 하복 | 학년 | 학교 | 주관 | 입찰 | 교실 | 제출 | 소액 |
| 13 | 우유 | 단계 | 학교 | 전자 | 동복 | 입찰 | 교실 | 하복 | 실습 | 우유 | 수의 | 실습 |
| 14 | 입찰 | 동복 | 안내 | 식용 | 주관 | 실습 | 업체 | 동복 | 급식 | 업체 | 급식 | 급식 |
| 15 | 안내 | 주관 | 급식 | 안내 | 단계 | 급식 | 간식 | 급식 | 재료 | 학교 | 단계 | 안내 |
| 16 | 업체 | 급식 | 전자 | 공동 | 급식 | 과학 | 서울 | 안내 | 전자 | 선정 | 안내 | 주관 |
| 17 | 선정 | 전자 | 재료 | 업체 | 규격 | 전자 | 안내 | 단계 | 주관 | 납품 | 소액 | 전자 |
| 18 | 공산품 | 규격 | 긴급 | 선정 | 인네 | 재료 | 선정 | 규격 | 하복 | 안내 | 규격 | 교복 |
| 19 | 부식 | 가격 | 식용 | 교실 | 전자 | 교구 | 전자 | 가격 | 교복 | 식용 | 가격 | 하복 |
| 20 | 육류 | 식용 | 기반 | 서울 | 가격 | 도제 | 남품 | 동시 | 물품 | 서울 | 동시 | 긴급 |

-특정 중요 단어 추출

공고명에서 단어만 추출한 다음 단어의 수를 세는 코드를 통해 그래프를 작성하였다.

2018년을 기준으로 초등학교에서 빈도수 상위 20개 물품 중 ‘구매’나 ‘견적’과 같은 단어를 제외한다면 실질적으로 구매하는 물품 종류는 급식, 우유, 전자가 있었다. 따라서 이러한 종류의 물품 수요가 많을 것으로 예측할 수 있다.

같은 방식으로 중학교와 고등학교도 분석한 결과 중학교에서는 급식, 우유, 전자가 있었고 고등학교에서는 급식, 과학, 전자, 교구, 도제가 있었다. 초등학교와 중학교, 고등학교에서 구매하는 물품의 양상이 다르다는 것을 알 수 있다.

**2.3.3 모델링**

-데이터 전처리

| count 값 왜곡 | 로그 변환 | 로그 변환 2번 |
| --- | --- | --- |
|  |  |  |

데이터 전처리 과정에서 count의 값이 비교적 적은 데이터 값에 왜곡되어 분포해 있다. 데이터 분석전, 정확한 결과를 얻기 위해서 count값을 로그 값으로 변환하여 정규성을 높이고자 하였다. 두 번의 log변환을 실행하여 비교적 정규 분포에 가까운 데이터로 변환하였다.

**1)초중고별 lightGBM 모델**

GBM(Gradient Boosting Machine), 부스팅 알고리즘은 여러 개의 학습기를 순차적으로 학습-예측하면서 잘못 예측한 데이터에 가중치 부열을 통해 오류를 개선하면서 학습된다. 손실함수를 최소화하는 방향성을 가지고 가중치 값을 업데이트에 경사하강법을 이용하는 모델이다. 일반적인 GBM모델은 노드 분할 방식이 ‘균형 트리 분할’이지만, 최대 손실값을 가지는 노드를 중심으로 계쏙해서 분할하는 ‘리프 중심 분할’방식을 사용한다. 따라서 다른 모델들의 비해 트리가 깊어지기 위해 소요되는 시간과 메모리를 절약할 수 있다는 장점이 있다. 본 모델링 상위 품목 예측 및 정확도를 높이기 위해서 사용되었으므로 비대칭적인 규칙 트리가 생성되더라도 균형 트리 분할 방식보다 예측 오류 손실을 최소화하는데 가장 큰 장점이 있었다.LightGBM은 과적합 문제가 있지만, 공식 문서에서 ‘적다’라는 기준은 약 만건 이하로 본 자료의 데이터의 개수가 커서 과적합에 민감하지 않을 것으로 파악했다.

-lightGBM 모델 분석과정

① lightgbm 데이터셋 변환

데이터 전처리 결과 파일인 df\_XX\_final파일에서 tag, rank, month, year을 x로 가져왔고 y로 count\_log값을 가져와 train 데이터와 test데이터를 7:3으로 나누었다. lgb.Dataset을 이용하여 lightgbgm 데이터 셋으로 변환하였다.

② 주요 파라미터를 세팅

learing \_rate : 부스팅 스텝을 반복적으로 수행할 때 업데이되는 학습률 값으로, learning\_rate를 작게하여 예측 성능을 향상시키고자 하였다.

boosting : 일반적인 그래디언트 부스팅 결정 트리를 사용하였다.

objective: 최솟값을 가져야 할 손실함수로 회귀로 정의하였다.

metric: 모델 구현시 손실을 정하기 위해 mse를 사용하였다.

③ 파라미터 세팅이후 모델을 만든 후 test 데이터로 예측한다.

주요 파라미터 세팅 이후 predict를 이용하여 test데이터로 예측한다.

④ MSE와 R2값으로 에러 측정

⑤ 산점도를 그려 fit 값을 산점도로 확인

**2)초중고별 릿지 모델**

-Ridge 모델 분석과정

①피처 인코딩과 피처 벡터화

선형 회귀의 경우 원-핫 인코딩 적용이 훨씬 선호되므로 원-핫 인코딩을 사용하였고 비교적 단어가 짧게 구성되어 있어 Count 기반의 벡터화를 적용하였다. 그리고 모델링을 할 때 희소 인코딩 변환된 데이터 세트를 hstack()을 이용해 모두 결합하였다.

② 적당한 알파값 찾기

alpha\_list = [0.001,0.01,0.1,1,10,100]을 만들어 어떤 알파값이 올 때 릿지회귀 모델의 socre값이 가장 높은지 확인하여 최적의 알파값을 설정하였다. 초중고 모두 alpha 값이 0.001일때 가장 score가 높았다. 따라서 alpha=0.001인 모델을 통해 데이터를 학습시켰다.

③ 모델링

train\_test\_split함수를 이용하여 전체 데이터의 20%를 평가 데이터 세트로 만들었다. train데이터를 이용해 Ridge모델을 train하여 ridge\_el, ridge\_mi, ridge\_hi에 모델을 저장하였다.

**2.3 모델링 평가 및 예측**

**2.3.1 평가**

| lightGBM 모델 / fit 모델 | 릿지 회귀 모델 / fit 모델 |
| --- | --- |
| Mean squared error: 0.0005735451707996335  R2 score: 0.9915099327767475 | Mean squared error: {:,2f} 0.0033814860315223797  R2 score : 0.9466511319347999 |
|  |  |

lightGBM 모델의 R2 score는 0.99로 데이터의 99%를 설명해주고 릿지 회귀 모델의 R2 score는 0.94로 데이터의 94%를 설명해 준다. 따라서 릿지 회귀모델보다 lightGBM이 단어의 언급 수를 예측하는 데에 더 적절한 모델이다.

**2.3.2 예측 및 확인**

앞에서 진행한 방법과 동일하게 단어의 수를 뽑았다. 더 적절한 모델이라고 판단했던 lightGBM 모델을 통해 2021년도 초중고별 물품 수요를 예측해보고 실제 데이터와 맞는지 확인해보았다. 초중고별 예측 그래프는 다음과 같다.

[lightgbm 모델을 통한 2021년 초중고별 물품 수요 예측]

| 초등 MSE / fit 결과 | 중등 MSE / fit 결과 | 고등 MSE / fit 결과 |
| --- | --- | --- |
| Mean squared error: 0.00951089583604252  R2 score: 0.8734895483031015 | Mean squared error: 0.007516211044374512  R2 score: 0.8936700346856074 | Mean squared error: 0.004610424669397066  R2 score: 0.8910876634086649 |
|  |  |  |

초등, 중등, 고등 모두 R2 score이 0.87, 0.89, 0.89로 데이터의 87%, 89%, 89%를 설명하는 것을 알 수 있으며 이는 LightGBM 모델링이 잘 적합되었다는 결론을 도출 할 수 있다.

**3. 결론**

**3.1 결과 해석**

우선, 데이터 전처리와 eda를 통해 2017년도부터 2020년까지 4년치의 데이터를 분석해 공고명의 단어 빈도 수를 확인하고 월별 가격 동향.및 구매 빈도를 확인하였다. 위의 데이터를 이용해 모델링을 진행하였다. lightGBM 모델과 릿지 회귀 모델을 비교해 보았을 때, lightGBM 모델의 R2 score는 0.99, 릿지 회귀 모델은 0.94로 단어의 언급 수를 예측하는 데에 lightGBM 모델이 더 적합함을 알 수 있다. 따라서 lightgbm 모델을 채택하여 2021년도 물품 수요 예측을 진행하였다. R2 score를 확인해 보았을 때 예측 데이터의 약 90%를 설명해주며 이는 모델링이 잘 적합되었음을 알 수 있다.

**3.2 활용방안 및 발전방향**

위의 분석 결과를 통해 초, 중, 고등학교 별 상위 조달 품목 파악이 가능하다. 2017년부터 2020년까지의 데이터를 기준으로 초등학교의 상위 조달 품목은 급식, 우유 등이고, 중,고등학교의 상위 조달 품목에는 급식, 과학, 전자, 교복 등이다.

또한 분석을 통해 월별 동향 파악이 가능하다. 특이점이 있다면 초, 중, 고등학교는 새학기가 시작하기 직전인 2월에 구매 빈도가 증가하는 추세를 보였다. 월별 물품 가격의 합을 확인해 보았을 때 초등학교는 방학이 끝나고 학기가 시작되기 전인 2월과 8월에 물품 가격의 합이 가장 높았다. 반면, 중, 고등학교는 2학기 중간 지점인 10, 11월에 물품 가격의 합이 가장 높았다.

마지막으로, 분석 모델을 통해 향후 년도의 조달 품목 예측이 가능하다. 2017년부터 2020년의 데이터를 이용해 품목들의 가격 합, 월별 구매 빈도 등을 분석하였다. 다음 그래프는 ‘급식’이라는 품목의 4년간의 동향을 확인해본 그래프이다. 4년간 특정 월에 빈도가 높아짐을 확인할 수 있다. 이런 경향을 확인함으로써 공급기관 및 수요기관은 조달 품목에 대한 예측과 대응이 가능하다. 실제로 구매 빈도의 예측값과 2021년 실제 데이터를 비교해 보았을 때, 초등학교 데이터는 오차 약 0.009, 중학교 데이터는 오차 약 0.007, 고등학교 데이터는 오차 약 0.004로 0에 수렴함을 알 수 있다. 따라서 앞으로도 데이터가 계속 누적된다면, 같은 모델을 통해 향후 예측값을 얻을 수 있을 것이고 위의 결과를 바탕으로 공급기관 및 수요기관의 협업 부서에서 예산 설계 시 수요 예측 및 반영이 가능할 것으로 본다.