

|  |
| --- |
| 데이터마이닝 프로젝트 보고서 |
| Predict Future sales |



|  |  |  |  |
| --- | --- | --- | --- |
| 제출일 | 2019. 12. 05 | 전공 | 컴퓨터정보공학부 |
| 과목 | 데이터마이닝 | **학번** | 201620613 |
| 담당교수 | 노상국 | **이름** | 박지은 |

목 차

[I. 서 론 1](#_Toc26412863)

[1. 1. Concept 1](#_Toc26412864)

[1. 2. 시스템 환경 1](#_Toc26412865)

[II. 본 론 1](#_Toc26412866)

[2. 1. Input 분석 1](#_Toc26412867)

[2. 2. Algorithm 선택 2](#_Toc26412868)

[2. 3. Learning Curve 2](#_Toc26412869)

[2. 4. ANOVA Test 4](#_Toc26412870)

[2. 5. Demo Program 4](#_Toc26412871)

[III. 결 론 5](#_Toc26412872)

[3. 1. 결 론 5](#_Toc26412873)

[3. 2. 한계점 5](#_Toc26412874)

1. 서 론

1. 1. Concept

Kaggle의 Playground Prediction Competition중 하나인 Predict Future sales를 도메인으로 설정하였다. 이 Competition은 2013년부터 2015년 10월까지의 특정 가게의 판매 내역을 제공한다. 판매 내역에는 상점명, 상품명, 상품의 분류명, 일 판매량, 판매 가격이 포함되어있다. 제공된 파일을 가지고 2015년 11월의 월 판매량을 예측하는 것이 최종 목표이다.

1. 2. 시스템 환경

Anaconda의 Jupyter Notebook환경에서 Python 언어로 진행하였다. 분석 알고리즘은 Python Library를 이용하였다. 데모 파일은Visual Studio환경에서 C++언어로 제작하였다.

1. 본 론

2. 1. Input 분석

해당 데이터는 2898의 Instance를 가지고 있으며, Attribute는 17개이다. Class는 Item\_cnt\_month, 즉 월 총 판매량으로 지정하였다. Attribute의 상세항목은 다음과 같다.

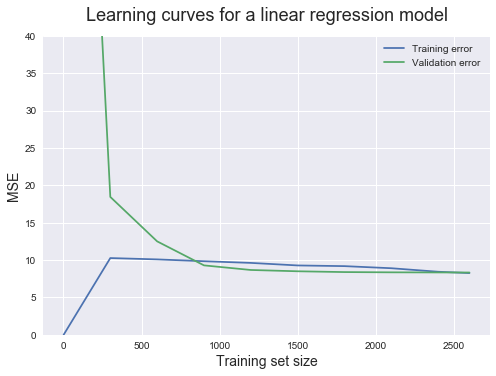
1. Date\_block\_num : 년월 (2013년 1월은 1, 2013년 2월은 2…)
2. Shop\_id : 상점 고유번호
3. Item\_cnt\_month : 이달의 월 총 판매량
4. Item\_category\_id : 상품 유형 고유번호
5. Item\_cnt\_month\_lag\_1 : 1달 전 상품의 월 총 판매량
6. Item\_cnt\_month\_lag\_2 : 2달 전 상품의 월 총 판매량
7. Date\_shop\_avg\_item\_cnt\_lag\_1 : 1달 전 상점별 평균 판매량
8. Date\_shop\_avg\_item\_cnt\_lag\_2 : 2달 전 상점별 평균 판매량
9. Date\_cat\_avg\_item\_cnt\_lag\_1 : 1달 전 상품 유형별 평균 판매량
10. Date\_cat\_avg\_item\_cnt\_lag\_2 : 2달 전 상품 유형별 평균 판매량
11. Item\_avg\_item\_price : 상품 유형별 평균 가격
12. Date\_item\_avg\_item\_price : 상품 유형별 월 평균 가격
13. Month : 월
14. Item\_shop\_first\_sale : 상점 상품 유형 별 첫 판매 이후 지난 개월 수
15. Item\_first\_sale : 상품 유형 별 첫 판매 이후 지난 개월 수
16. Item\_shop\_last\_sale : 상점 상품 유형 별 마지막 판매 이후 지난 개월 수
17. Item\_last\_sale : 상점 상품 유형 별 마지막 판매 이후 지난 개월 수

2. 2. Algorithm 선택

모든 데이터가 Numeric하므로 Linear Regression, XGBoost Regression, LightGBM Regression 총 세 가지 알고리즘을 사용하여 분석한다.

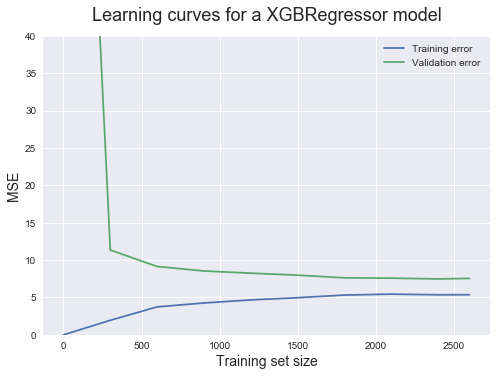
2. 3. Learning Curve

1. Linear Regression



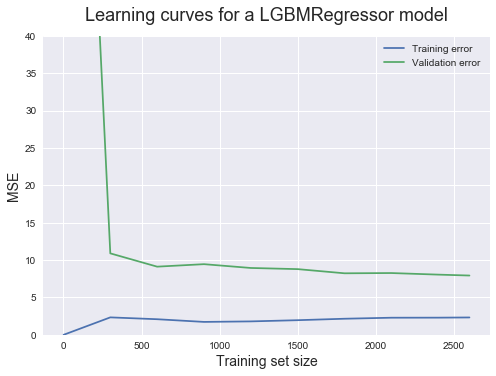
Training set size가 커질 수록 오차율이 점점 줄어드는 것을 확인할 수 있다. 따라서 Linear Regression에 2400개의 데이터를 사용한다.

1. XGBoost Regression



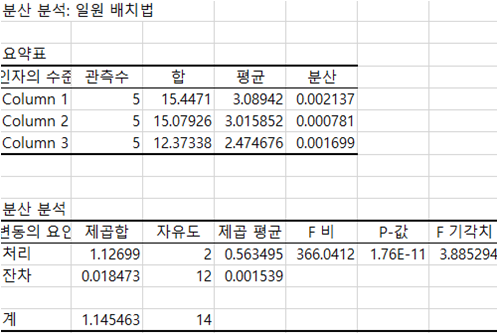
Training set size가 커질 수록 오차율이 점점 줄어드는 것을 확인할 수 있다. 따라서 XGBoost Regression에 2400개의 데이터를 사용한다. 또한 XGBoost Regression의 오차율은 이전의 Linear Regression의 오차율보다 낮은 수치를 보인다.

1. LightGBM Regression



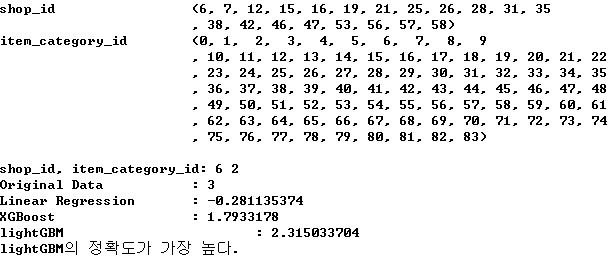
Training set size가 커질 수록 오차율이 점점 줄어드는 것을 확인할 수 있다. 따라서 LightGBM Regression에 2400개의 데이터를 사용한다. LightGBM의 알고리즘은 이전의 Linear Regression, XGBoost Regression과 비교했을 때 가장 낮은 오차율을 보인다.

2. 4. ANOVA Test



Column 1은 Linear Regression의 RMSE이다. Column 2는 XGBoost Regression의 RMSE이며 Column3은 LightGBM의 RMSE이다. 세 알고리즘의 RMSE를 비교하였을 때 LightGBM의 RMSE가 가장 낮고, Linear Regression의 RMSE가 가장 높다. 이 때 F 비는 F 기각치보다 매우 높은 수치를 보인다. 따라서 세 알고리즘은 유의미하다.

2. 5. Demo Program



위의 분석에 따라 간단한 Demo Program을 제작하였다. 상점의 고유번호와 상품 유형의 고유번호를 입력했을 때 세 알고리즘의 예측 결과와 원본 데이터를 비교했다. 위의 그림은 6번 상점의 2번 상품 유형을 입력하였을 때 세 알고리즘 중 LightGBM Regression의 예측 결과가 원본 데이터 값과 가장 유사했다.

1. 결 론

3. 1. 결 론

Linear Regression, XGBoost Regression, LightGBM Regression 세 가지 알고리즘을 적용하였을 때 LightGBM Regression의 정확도가 가장 높았다. 그 다음으로 XGBoost Regression의 정확도가 높았고 Linear Regression의 정확도가 가장 낮았다. 따라서 Predict Futures Sales 데이터에서 LightGBM Regression을 적용하는 것이 가장 적절하다.

3. 2. 한계점

Kaggle에서 제공된 Train데이터의 Instance 수가 과제에서 제한하는 3000개를 초과했다. Instance의 개수를 조절하기 위해 Concept을 상품별 월 판매량 예측에서 상품 유형 별 월 판매수량 예측으로 변경하였다. 또한 최근 2개월간의 데이터만 Instance에 포함하였고 그 중에서도 전체 판매량의 중간 값 이상인 상위 20개 상점의 데이터만 추출하였다. 제한된 Instance로 분석한 결과는 원본 데이터를 가지고 분석했을 때 보다 더 높은 오차율을 보였다. 원본데이터를 가지고 XGBoost Regression, CATBoost Regression, LightGBM Regression으로 분석하고 세 알고리즘으로 예측한 데이터를 XGBoost Regression 알고리즘을 통해 Stacking 기법으로 예측했을때 RMSE는 0.99정도로 낮은 수치를 보였다.