Study for the amount of imports in KOTRA forecasing based on Ensemble Model with Machine Learning and Data Analysis

CUAI 4 기 데이터 분석팀
강영훈(경영학), 김민주(경영학), 정욱준(경영학),

[요약] 프로젝트 목표, 과정, 결과를 요약하여 200 자 이내로 기술

Machine Learning 을 통한 데이터 분석에 다양한 모델은 크게 단일 모델과 앙상블(Ensemble) 모델로 나눌 수 있다. 단일 모델은 하나의 모델만 사용해 최종값을 도출하는 모델을 의미하고 앙상블(Ensemble)모델은 여러 개의 모델을 적절하게 결합해 최종값을 도출하는 모델을 의미한다. 성능향상 및 과적합(Overfitting) 문제를 해결할 수 있는 앙상블(Ensemble) 모델을 통해 KOTRA 한국 수입액 예측을 수행하고자 한다.

1. 서 론

기계학습에서 모델링을 향상시키기 위해서는 교차검증(cross validation), 피처 엔지니어링(feature engineering) 하이퍼파라미터 튜닝(hpyer parameter tuning), 알맞은 알고리즘(algorism) 선택이 있다. 다양한 방법을 종합적으로 사용해서 최적의 모델링 결과를 찾기위해 노력한다. 주어진 KOTRA 데이터는 시계열(time series) 데이터가 아니었으며, 여러 데이터를 종합한 단적인 데이터의 형태를 보였다.
따라서 데이터에 특성에 맞는 최적의 알고리즘을 랜덤포레스트(Random forest), XGBoost, LightGBM 과 같은 앙상블 모델을 생각했다.
앙상블이란 어떤 데이터의 값을 예측한다고 할 때, 여러 개의 모델을 조화롭게 학습시켜 그 모델들의 예측 결과들을 이용하여 더 정확한 예측값을 구하는 것이다. 여러 개의 결정 트리(Decision Tree)를 결합하여 하나의 결정트리보다 더 좋은 성능을 내는 머신러닝(Machine Learning) 기법이다.
앙상블 학습법에는 두 가지가 있다.
배깅(Bagging)과 부스팅(Boosting)이다.

하이퍼 파라미터 튜닝(hpyer parameter tuning)의 경우 그리드 서치(grid search)와 임의 탐색 등을 통해 진행하였다. 또한 실제 변수들의 제거 방법으로 A/B 테스트를 통해 RMSE 에 부정적인 영향을 주지 않는 변수들을 전진선택법(Forward Selection)을 통해 선택하였다.
피처 엔지니어링(feature Engineering)의 경우 주어진 KOTRA 데이터에 무역지표에 근거를 바탕으로 파생변수(Derived Variable)를 만들고 추가적으로 무역데이터를 이용하였다.
실제 주어진 KOTRA 데이터의 2017 년 수입액을 예측하는 모델링을 진행하고, RMSE 가 가장 낮은 최적의 값을 찾고 2018 년 데이터에 동일하게 진행할 것이다.

2. 본 론

*내용 왼쪽 정렬
*자료 삽입 시 **그림은 아래 캡션, 표는 위 캡션** 이용, 캡션은 8p, 왼쪽 정렬로 설정
*필요에 따라 아래와 같이 소제목 사용 (구분없이 작성하여도 무관)

1) Ensemble Model Inforamtion

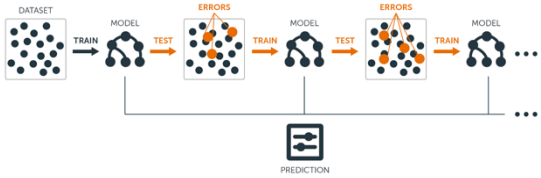

Figure 1. Ensemble models using Decision Trees as weak learners

위의 구조에서 보이듯 앙상블(Ensemble)모델은 가중치를 활용하여 약 분류기를 강 분류기로 만드는 방법인 부스팅(Boosting)과 샘플을 여러 번 뽑아(Boostrap) 각 모델을 학습시켜 결과물을 집계(Aggregration)하는 배깅(Bagging)이 있다.
부스팅은 배깅에 비해 error 가 적은 편이다. 즉, 성능이 좋다. 하지만 속도가 느리고 오버피팅(Over fitting)이 될 가능성이 있다. KOTRA 수입액 예측의 경우 2017 년의 데이터로 모델링을 학습하고 2018 년 데이터를 통해 수입액을 예측하므로 일반화 성능이 중요하다 판단했다.

따라서 일반화 성능을 고려한 최적의 모델링 방법을 고려하였다.

모델명	Cross_val_score.mean
Random Forest	2.009
Xgboost	2.062
Lightgbm	1.975

(각 모델별 cross_val_score.mean값)

2) Feature engineering

Machine Learning 에서 과적합을 피하거나 모델링의 일반화 성능을 높이기 위해서는 무엇보다 데이터가 중요하며, 데이터내에 피처(Feature)를 엔지니어링(engineering) 하는 방법은 유용한 방법 중 하나이다.

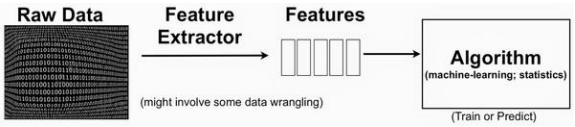

Figure 2. Feature Engineering, Techniques, examples, and case studies.

KOTRA 데이터내에 피처들을 통한 파생변수 생성과 새로운 무역 데이터를 이용하여 진행하였다.

3) Predictor variable

피처 엔지니어링을 통한 변수들을 선택하기 전 모든 변수들 중 기본 변수 2 개만 선택하고 나머지를 하나씩 선택하며 RMSE 에 부정적 영향을 주는 변수를 제거하는 전진 선택법(Forward Selection)을 사용하였다. Machine Learning 모델링을 통해 feature importance 도 확인하며 중요한 변수들을 다시 한번 확인하였다.

변수(variable)				RMSE	변수 선택 유무
A변수	B변수			RMSE = 2.11	시작
A변수	B변수	C변수		RMSE = 2.05	C변수 선택
A변수	B변수	C변수	D변수	RMSE = 2.18	D변수 제거
A변수	B변수	C변수	E변수	RMSE = 2.01	E변수 선택

(A/B테스트 예시 RandomForest, XGBoost, LightGBM)

3) Hyper parameter tuning

중요한 변수와 알고리즘을 선택 후 우리는 일반화 성능이 가장 높다고 판단되는 랜덤 포레스트(Random Forest)를 주 모델로 선정하였다. 일반적으로 부스팅 계열의 모델이 랜덤포레스트 모델보다 성능이 높지만, 부스팅 계열의 모델은 오답에 가중을 두는 방식으로 훈련하므로 학습 데이터 기간이 짧다면 과적합의 가능성이 높다. 따라서 부스팅(Boosting) 계열 모델의 사용을 지양하였다.

임의 탐색(random search)과 그리드 서치(grid search)를 통해 하이퍼파라미터 튜닝 결과 값을 확인하였다.

튜닝 방법	max_features	min_samples_leaf	n_estimators	RMSE
그리드 탐색	8	5	300	0.6657
임의 탐색	4	4	436	0.6657

(하이퍼 파라미터 튜닝 결과 값)

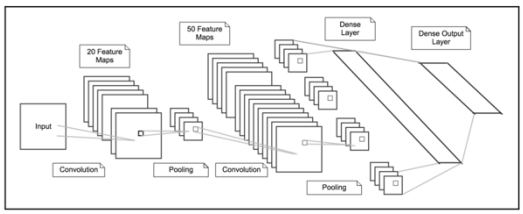

Figure 3. Machine Learning Model Evaluation and Hyper-Parameter Tuning

3. 결 론

다음 표는 최종 데이터셋에 적합한 모델 두 가지(랜덤 포레스트(random forest), LightGBM))을 선택한 후, 하이퍼 파라미터 튜닝을 통해 최종적으로 5 개의 모델을 구축했다.

모델명	사용 된 데이터 변수	하이퍼 파라미터 튜닝	RMSE
LightGBM	Input_var1	1000, 0.05,dart	1.5409
Random Forest	**Input_var1**	**8,1,500**	**0.6921**
Random Forest	Input_var1	3,2,500	1.1016
Random Forest	Input_var2	4,2,500	1.2275
Random Forest	Input_var3	7,4,500	1.2964

(최종 모델 구축 및 결과 값 예시)

분석 결과 Random Forest 모델을 사용하여 input_var1 의 변수를 이용한 하이퍼 파라미터 튜닝 값 8, 1, 500 시 가장 낮은 RMSE 0.6921 을 확인하였다. 특히 각 모델링에서 주요한 feature 들을 A/B 테스트를 통해 하나의 리스트에 담아 사용함으로써 모델링 마다 변수 선택이 비교 가능하였다.이를 바탕으로 2018 년 KOTRA 데이터 한국 수입액 예측도 동일하게 진행하였다.
아쉬운 점은 처음에 주어진 KOTRA 데이터 자체가 시계열 (Time series) 데이터가 아닌 단적인 데이터로 존재하였으며, 시간과 특정한 주기성에 영향을 받을 수 있는 수입액을 예측하는 것에 논리적인 오류를 범할 확률이 높았다.

참고 문헌

J. M. Yoon, "Effectiveness Analysis of Credit Card Default Risk with Deep Learning Neural Network," Journal of Money & Finance, vol. 33, no. 1, pp. 151-183, Mar. 2019.
Kaggle. UCI Credit Card Dataset [Internet]. Available: https://www.kaggle.com/uciml/default-of-credit-card-client s-dataset.
A. Shen, R. Tong, and Y. Deng, "Application of Classification Models on Credit Card Fraud Detection," in 2007 International Conference on Service Systems and Service Management, pp. 1-4, Jul. 2007.
B. M. Pavlyshenko, "Machine-Learning Models for Sales Time Series Forecasting," Data, vol. 4, no. 1, Apr. 2019.
B. Scholkopf, C. J. C. Burges, A. J. Smola, 'Ad-¨ vances in kernel methods', The MIT Press 1999
C.L. Wilson and M.D.Garris, Handprinted character database 3, February 1992, http://www.nist.gov/srd/nistsd19.htm, National Institute for Standards and Technology, Advanced Systems Division pg.43
D.M.J. Tax, 'One-class classification' , PhD Thesis, Delft University of Technology, http://www.ph.tn.tudelft.nl/˜davidt/thesis.pdf ISBN: 90-75691-05-x, 2001
D.M.J. Tax, R.P.W. Duin, 'Support Vector Data Description' , Pattern Recognition Letters, December 1999, vol. 20(11-13), pg. 1191-1199
Dacon. Korea data competition platform [Internet]. Available: https://dacon.io/. [13] R. J. A. Little and D. B. Rubin, Statistical Analysis with Missing Data, 2nd ed. Hambrug, NJ: John Wiley & Sons Inc., 2014. Dacon. Korea data competition platform. Card Sales Prediction contest [Internet]. Available: https://dacon.io/competitions/official/140472/overview/.
Documents for Catboost [Internet]. Available: https://catboost.ai/.