수요예측에 사용되는 ML 모델

September 4, 20XX

주요 접근 방법

- Historical Average
  - 방법: 과거(일정 기간)의 데이터에 대한 평균화(smooth)를 한 결과
  - 효과: 일간, 주간 변화는 평균에 의해 수렴되어 큰 변화(variation)가 사라짐
  - 장점: 가장 간단한 방법
  - 단점: 변화가 많은 데이터에 대한 예측이 떨어지고, 특이값(outlier)으로 인한 영향이 큼
  - 주요 모델: Simple Moving Averages, Holt-Winters Exponential Smoothing

- Time Series with Added Regressors
  - 방법: 과거(일정 기간)의 데이터에 Regression 기법을 적용하여 학습
  - 효과: 계절적인 영향이나 다른 독립변수 변수(예: 날씨, 경제지표)들을 모델 학습에 반영할 수 있음
  - 장점: 조금 더 복잡한 요인들을 고려한 예측 가능
  - 단점: 학습 데이터가 부족하면 오버피팅. 한 변수로 인한 과도한 영향을 막기 위해 전처리(예: z_score)가 충분히 되어야 함
  - 주요 모델: Seasonal ARIMAX, Generalized Additive Models

주요 접근 방법

- Machine Learning/Artificial Intelligence
  - 방법: 많은 변수들을 Neural Network에 넣어 귀납적 방법(뉴런 노드의 가중치 최적화)에 대한 학습
  - 효과: 매우 복잡한 변수(조건)들이 있는 환경에서 특정 변수에 대한 특별한 고려 없이 결과 도출 가능
  - 장점: 선형이 아닌 추세나 복합적인 요소(변수)들이 있는 경우의 예측 가능. 사전에 어떤 모델 유형을 사용할지 고민 불필요
  - 단점: 충분한 학습 데이터가 필요함. 결과에 대한 해석이 어려워 모델을 개선하는 방향을 잡기 어렵고 오버피팅 위험성
  - 주요 모델: Random Forest, Gradient Boosted Machines, Neural Networks (LSTM-RNN, CNN) Support Vector Machine

학습 결과에 대한 해석

- 통계적 방법
  - 머신러닝(문제를 풀어가는데)의 연역적 방법이라 할 수 있음
  - a, b, c, d라는 독립변수(결과에 영향을 미치는 요소)들을 가지고 M이라는 모델을 사용해서 결과를 도출했을 때, 어떤 변수가 주성분(가장 크게 영향을 미치는 요소)인지 등을 찾아내는 것
  - 해석력(explainability)이 좋으면 해석 결과를 보고 개선 방향을 찾아 모델의 효율을 높일 수 있음. 또한 새로운 데이터가 들어와 모델에 변화가 생겼을 때 원인(예: 3월에 서울지역 판매 급증)을 찾기 용이 함
  - 단, 변수가 많고 요소들 간의 상관관계가 복잡하면 해석을 하기가 어려워(통계 전문가가 직접 해야 하는 작업이므로) 해석 결과의 신뢰도가 떨어지게 됨 ⇒ 귀납적 방법의 필요성
- Neural Network 기반 머신러닝
  - 귀납적 방법에 해당 – 많은 양의 데이터(정답)를 넣어 가중치를 역으로 추정하는 방법
    - m = ax + by + cz + d 에서 많은 양의 독립변수 x, y, z 값과 종속변수 m 값을 대입하면서 가장 정답에 근접한 a, b, c 값을 찾아 감
  - 독립변수가 많고 서로 간의 관계가 복잡한 경우에도 결국 해답을 넣어 가중치를 추정(귀납적)하는 방법이므로, 데이터만 충분하다면 원하는 결과를 쉽게(전문가가 많은 계산을 해야 하는 통계적 방법에 비해) 찾아낼 수 있음
  - Neural Network 기반의 머신러닝 시스템 구축이 점점 용이(예: 클라우드 환경)해지고 있어 발전 잠재력이 큼
  - 답을 대입해서 가중치를 추정해 가는 귀납적 방법이라, 학습 결과에 대한 해석이 쉽지는 않아서 모델의 품질을 높이는 작업이 용이하지 않음 ⇒ 대신, 주로 입력 데이터에 대한 Feature Engineering에 대한 개선에 주력