2024.12.09

스마트시티 문제해결 프로젝트

<<가늠>>

린튼글로벌비즈니스 20211944 임하은
빅데이터 응용학과 2022 박성빈
수학과 2020 원현아

Contents

01.

공모전 소개

공모전 소개

DACON 데이터, AI를 활용한 물가 예측 경진대회 : 농산물 가격을 중심으로

[주제]

국민생활과 밀접한 10개 농산물 품목의 가격 예측

(배추, 무, 양파, 사과, 배, 건고추, 깐마늘, 감자, 대파, 상추)

[문제 상세 설명]

Train Data - 2018년 ~ 2021년의 순 단위(10일)의 데이터

Test Data - 추론 시점 T가 비식별화된 2022년의 순 단위의 데이터 /
평가 데이터 추론은 추론 시점 T 기준으로 최대 3개월의 순 단위의
입력 데이터를 바탕으로 T+1순, T+2순, T+3순의 평균가격을 예측

02.

팀 프로젝트 개요

팀 프로젝트 개요

DACON 데이터, AI를 활용한 물가 예측 경진대회 : 농산물 가격을 중심으로

01 데이터 결측치 확인 및 그룹화

- 시점, 품목명, 품종명, 거래단위, 등급을 기준으로 데이터 그룹화
- 각 그룹의 평균가격과 평년 평균가격을 계산하고, 결측치 처리 진행

02 데이터 병합 및 품목 선정

- 전국도매 데이터와 산지공판장 데이터를 병합하여 활용 가능한 품목 선정
- 사용하기 어려운 품목은 시각화를 통해 변화를 분석하고, 모델에 적합한 데이터만을 사용

03 시점 변환 및 데이터 정규화

- 시점 변환 및 정규화를 통해 LSTM 모델에 적합한 형태로 데이터 준비
- train_data와 test_data로 나누어 모델 학습과 예측에 사용될 데이터 준비

04 LSTM 모델 설정 및 학습

- 시계열 예측을 위한 LSTM 모델을 설정하고, epochs 200, batch_size 32로 학습 진행

팀 프로젝트 개요

DACON 데이터, AI를 활용한 물가 예측 경진대회 : 농산물 가격을 중심으로

05 학습된 모델로 예측

- 학습된 LSTM 모델을 test_data에 적용하여 예측 수행
- 예측 결과는 시간 순서대로 정렬하여 분석에 활용

06 시계열 데이터셋 생성

- 슬라이딩 윈도우 기법을 사용하여 시계열 데이터셋을 생성하고, 과거 데이터를 기반으로 미래를 예측할 수 있도록 데이터셋 준비

07 예측값 복원 및 평균 계산

- 예측된 값을 정규화된 값에서 원래 스케일로 복원하고, 예측값을 세 개씩 묶어서 그룹 평균을 계산

08 예측 결과 분석 및 모델 평가

- 예측 결과를 분석하여 모델 성능을 평가하고, 예측의 정확성을 높이기 위한 모델 튜닝과 하이퍼파라미터 조정 진행

03.

공모전 수상작 분석

DACON 데이터, AI를 활용한 물가 예측 경진대회 : 농산물 가격을 중심으로

- 팀 <나서스>
- 팀 <푸릇푸릇>

팀 <나서스 >

시장 안정성 및 물가 안정을 위한 시장 동향 반영 품목별 가격 전략 모델링

01 데이터 통합

- Train 데이터와 Meta 데이터를 통합하여 분석 가능하도록 구성.
- 품종명 불일치 문제(예: 배추 ↔ 월동배추)를 해결하여 품종 데이터의 신뢰성 확보.
- 데이터 증강 기법을 활용하여 실제 시장 특성을 반영한 추가 데이터 생성.

02 데이터 구조

- 시계열 데이터 학습을 위해 Sliding Window 기법을 적용, 9개 시점 데이터를 기반으로 3개 시점 예측 구조 설계.
- Label Encoding을 통해 범주형 데이터를 구조화하고, Standard Scaling 등 적절한 스케일링 방법으로 데이터 정규화.

03 가격 변동 분석 및 맞춤형 전처리

- 품목별 데이터 특성을 고려한 스케일링 기법 선택:
  - 큰 변동 폭: Standard Scaling.
  - 작은 변동 폭: No Scaling.
  - 이상치 처리 필요: Robust Scaling.
  - 특정 범위 제한: MinMax Scaling.

04 시계열 데이터의 계절성과 추세 반영

- 계절적 패턴 학습을 위해 Cycling Transform(월별/일자별)과 Binary Transform(peak/bottom)을 적용.
- 계절성 데이터를 정량화하여 모델이 계절적 변동을 효과적으로 학습하도록 설계.

팀 <나서스 >

시장 안정성 및 물가 안정을 위한 시장 동향 반영 품목별 가격 전략 모델링

05 외부 데이터 활용

- 기상 데이터와 시장 데이터를 통합해 시장 가격에 영향을 미치는 요인을 추가.
- 통계적 분석을 통해 파생 변수 생성 및 데이터의 시계열 특성 강화.

06 검증 전략 설계

- 월별 3단계 검증(1순, 2순, 3순)을 통해 검증 데이터와 훈련 데이터 간의 시점 간격 최소화.
- 장기 예측의 어려움을 보완하기 위해 3순(21~31일)에 대해 특별 처리 전략 적용.

07 품목별 최적화 모델링

- XGBoost: 양파, 깐마늘, 무, 대파.
- Extra Trees + XGBoost: 배추, 사과, 감자, 상추.
- RANSAC: 배, 건고추.

08 계절성과 품목 특성 기반 최적 모델 학습

- 계절성을 반영한 데이터 구조와 품목별 특성에 맞춘 최적 모델을 구성하여 정확도를 극대화.

공모전 수상작 분석

팀 <푸릇푸릇 >

농산물 특성 및 통계적 분석 기반 시계열 변화를 반영한 품목별 농산물 가격 예측 AI 모델링

01 데이터 통합

- Train 데이터와 Meta 데이터를 통합하고 품종명 불일치 문제를 해결.
- 신선식품의 순수입량을 계산하여 주요 변수로 추가.

02 데이터 전처리

- Null 값 처리 (Imputation) 및 Data Leakage 방지를 위해 과거 데이터만 활용.
- 날씨 데이터를 순 단위로 변환하여 일관성 확보.

03 파생 변수 생성

- 3순~0순(40일) 평균 가격의 평균값 및 표준편차 계산.
- Mann-Kendall Test와 Theil-Sen 기울기 분석을 통해 미래 가격 추정치를 생성.

04 주요 변수 선정

- 문헌조사 및 변수 중요도 평가(SHAP, Permutation, Feature Importance)를 통해 품목별 중요한 변수를 선택.
- Backward Selection 방식으로 불필요한 변수를 제거하여 최적 변수 집합 도출.

팀 <푸릇푸릇 >

농산물 특성 및 통계적 분석 기반 시계열 변화를 반영한 품목별 농산물 가격 예측 AI 모델링

05 시계열 특성 반영

- 계절성과 트렌드 데이터를 활용해 모델에 시계열적 특성을 학습시킴.
- Cycling Transform(월별/일자별)과 Binary Transform(peak/bottom)을 적용.

06 모델링 및 하이퍼파라미터 튜닝

- XGBoost, Random Forest, AutoGluon 모델 적용.
- Grid Search를 통해 colsample_bytree, max_depth, n_estimators 등 최적의 하이퍼파라미터 설정.

07 검증 및 모델 평가

- 품목별 최적화된 변수와 모델을 사용해 예측 정확도(NMAE) 0.2484 달성.
- 날씨 및 수입량 변수 활용으로 성능 개선.

08 모델 최적화 및 적용

- XGBoost 모델을 활용하여 높은 해석력과 빠른 추론 시간(12.2초)을 구현.

04.

프로젝트와 수상작 비교 분석

프로젝트와 수상작 비교 분석

팀 <나서스>/<가늠>프로젝트 비교분석

데이터 활용

<가늠>	<나서스>
• 주어진 데이터만 사용, 데이터 증강을 시도하지 않음. • 데이터 부족 문제를 해결하기 위해 별도의 데이터 보완이 이루어 지지 않음. • 데이터의 다양성과 양적 한계로 인해 모델이 학습할 수 있는 패턴 제한적	• 외부 meta 데이터와 결합하여 데이터 증강 • 품종명이 불일치 하다면 유사 트렌드를 가진 품종명으로 대체

프로젝트와 수상작 비교 분석

팀 <나서스>/<가늠>프로젝트 비교분석

전처리 및 피처

<가늠>	<나서스>
• 간단한 스케일링을 적용해 데이터 정규화 • 추가적인 피처 생성 X, 시계열 특성 반영 작업 부족 • 한계점 : 추세, 계절성을 모델에 반영하지 못함	• Sliding Window 기법으로 데이터를 재구조화 하여 Train 데이터 강화 (9개의 과거 평균 시점 데이터를 사용해 다음 3개 시점을 예측하는 방식) -> 이 방법은 모델이 과거 가격의 변화와 미래 연속성을 학습할 수 있도록 지원 • Sin/Cos 변환 : 계절성과 주기성을 반영한 피처를 생성해 월별/일자별 주기를 수치화 • 적절한 스케일링 사용 : 가격 변동 폭에 따른 RobustScaler(이상치 고려), StandardScaler(변동 폭 조정)

프로젝트와 수상작 비교 분석

팀 <나서스>/<가늠>프로젝트 비교분석

모델

<가늠>	<나서스>
• 단일 LSTM 모델을 사용하여 시계열 데이터의 패턴 학습 • 한계점 : 단일 모델로 모든 품목의 특성을 반영하기 어려웠음.	• 품목별 특성에 따라 XGBoost, Extra Trees, RANSAC 모델을 선택 • 모델별 최적화 : 그리드 서치를 활용해 하이퍼파라미터 조정

프로젝트와 수상작 비교 분석

팀 <나서스>/<가늠>프로젝트 비교분석

계절성 반영

<가늠>	<나서스>
• 계절성을 고려하지 못하고 시계열 데이터를 단순 시점 단위로 학습 • 월별 / 일자별 반복되는 패턴이 모델에 반영되지 못함	• Sin/Cos 변환으로 주기적 특성을 가진 데이터를 모델이 학습할 수 있도록 변환 • 월별 데이터는 $\sin(2\pi * \frac{month}{12})$, $\cos(2\pi * \frac{month}{12})$ 로 변환 • 피크/바닥 분석 : 특정 기간의 가격 피크/ 바닥을 이진값으로 표시해 극단값 처리

프로젝트와 수상작 비교 분석

팀 <나서스>과 <가능>프로젝트 차별점

[데이터 활용]
증강으로 추가된 데이터의 확장 뿐 아니라 실제 시장 특성 반영

[전처리 및 피처]
단순 스케일링을 넘어, 시계열 데이터의 복잡한 패턴을 학습 가능하게 구조화

[모델]
단일 모델에 의존하지 않고 품목 특성에 맞는 모델링 진행

[계절성 반영]
계절성 데이터의 정량화를 통해 모델이 계절적 패턴을 학습

프로젝트와 수상작 비교 분석

팀 <푸릇푸릇>/<가늠>프로젝트 비교분석

데이터 활용

<가늠>	<푸릇푸릇>
• 주어진 데이터만 사용, 데이터 증강을 시도하지 않음. • 데이터 부족 문제를 해결하기 위해 별도의 데이터 보완이 이루어 지지 않음. • 데이터의 다양성과 양적 한계로 인해 모델이 학습할 수 있는 패턴 제한적	• 수출입 데이터를 활용해 신선식품의 순수입량을 계산하여 데이터 강화 • Data Leakage를 방지하기 위해 2개월 이전 데이터를 사용하여 순수입량 값으로 정리 • Null 값은 동일 시점의 평균 값으로 Imputation 처리하여 데이터 완성도 개선

프로젝트와 수상작 비교 분석

팀 <푸릇푸릇>/<가늠>프로젝트 비교분석

전처리 및 피처

<가늠>	<푸릇푸릇>
• 간단한 스케일링을 적용해 데이터 정규화 • 추가적인 피처 생성 X, 시계열 특성 반영 작업 부족 • 한계점 : 추세, 계절성을 모델에 반영하지 못함	• 날씨 데이터를 월 단위에서 순 단위로 변환하고 관측 시점을 통일 • 수출입 데이터를 기반으로 신선식품의 순수입량을 계산하고, 이를 주로 피처로 사용 • 과거 40일간 평균값, 표준편차 등 통계적 변수 생성 • 가격 기울기 및 증감 트렌드 계산 • 통계적 기울기를 기반으로 미래 가격 추정값 추가

프로젝트와 수상작 비교 분석

팀 <푸릇푸릇>/<가늠>프로젝트 비교분석

모델

<가늠>	<푸릇푸릇>
• 단일 LSTM 모델을 사용하여 시계열 데이터의 패턴 학습 • 한계점 : 단일 모델로 모든 품목의 특성을 반영하기 어려웠음.	• Tree 기반 모델(Random Forest, XGBoost, AutoGluon) 사용 • 변수 중요도 분석 (Feature Importance, SHAP Importance, Permutation Importance)을 통해 품목별 맞춤형 변수 선택 • 모델별 최적화 : 그리드 서치를 활용해 하이퍼 파라미터 조정

프로젝트와 수상작 비교 분석

팀 <푸릇푸릇>/<가늠>프로젝트 비교분석

계절성 반영

<가늠>	<푸릇푸릇>
• 계절성을 고려하지 못하고 시계열 데이터를 단순 시점 단위로 학습 • 월별 / 일자별 반복되는 패턴이 모델에 반영되지 못함	• 데이터를 순 단위로 변환하여 계절적 패턴 반영 • Mann-Kendall Test : 데이터를 기반으로 계절적 트렌드와 변화를 분석 • 과거 데이터를 기반으로 한 가격 기울기 및 증감 트렌드 추가로 계절성 보강

프로젝트와 수상작 비교 분석

팀 <푸릇푸릇>과 <가늠>프로젝트 차별점

[데이터 활용]
데이터를 통합, 보완하여 신뢰성과 경제적 의미를 모델에 반영

[전처리 및 피처]
통계적 분석과 파생 변수로 데이터의 시계열적 특성을 강화

[모델]
다양한 모델을 활용해 품목별 데이터 특성에 맞춘 최적의 모델 구성

[계절성 반영]
계절성 데이터를 분석하고 파생 피처로 활용하여 모델이 학습할 수 있도록 설계

05.

결론 및 학습점

결론 및 학습점

프로젝트 결론

- **데이터 활용**: 우리는 제공된 데이터를 중심으로 분석을 진행했으나, 수상 팀들은 데이터 증강 및 통합을 통해 실제 시장의 특성과 경제적 의미를 모델에 반영함.
- **전처리 및 피처 엔지니어링**: 단순 스케일링에 그친 우리와 달리, 수상 팀들은 시계열 데이터의 특성을 강화하고 복잡한 패턴을 학습할 수 있는 구조화를 진행함.
- **모델링 전략**: 우리는 단일 LSTM 모델에 의존했으나, 수상 팀들은 품목별 데이터 특성에 맞는 다양한 모델을 활용하며 최적화를 이룸.
- **계절성 반영**: 계절성을 정량화하고 이를 파생 변수로 활용한 수상 팀들과 달리, 우리는 계절성을 고려하지 않아 모델이 해당 패턴을 학습하지 못함.

결론 및 학습점

프로젝트 진행 후 학습점

- **데이터 활용의 중요성**: 단순히 제공된 데이터에 의존하지 않고, 증강 및 통합 과정을 통해 더 깊이 있는 분석 가능
- **전처리 및 피처 엔지니어링의 확장성**: 단순한 스케일링이 아닌 데이터의 본질적 특성을 드러낼 수 있는 통계적 분석 및 파생 변수 생성이 모델 성능 향상에 핵심이라는 점
- **다양한 모델의 필요성**: 단일 모델이 모든 데이터에 최적일 수 없으며, 데이터 특성에 따라 모델을 선택하거나 하이브리드 접근 방식을 사용하는 것이 효과적임
- **계절성 반영**: 계절적 패턴이 강한 데이터를 다룰 때, 이를 정량화하고 모델이 학습할 수 있도록 설계하는 것이 중요하다는 점

Thank You