중고차 가격 예측 데이터 분석 보고서

1. 예측 모델 성능 비교: RandomForest vs XGBoost

- **RandomForest 결과와 XGBoost 결과를 비교 분석한 결과,**
  두 모델 모두 실제 가격과 예측 가격이 비슷하게 나타나는 구간이 많지만,
  XGBoost가 전반적으로 더 일관적이고 안정적인 예측력을 보여줌.
- **공통점:**
  - 두 모델 모두 점(각 샘플)이 완벽 예측선(빨간 대각선) 근처에 빽빽하게 분포해,
    전체적으로 예측값이 실제값과 유사함을 시각적으로 확인할 수 있다.
- **차이점:**
  - **RandomForest**에서는 고가 차량 구간(실제 가격이 4,000만~1억 4,000만 원 이상)에서 일부 예측값이 실제값에 비해 크게 낮게 형성되는 outlier가 더 도드라진다.
  - **XGBoost**는 동일 구간에서 예측력이 개선되어, 극단적으로 벗어나는 outlier가 눈에 띄게 줄었으며,
    전체 분포가 완벽 예측선에 더 가까워졌다.
- **전체적으로 XGBoost 모델이 고가 차량 예측에서 약간의 개선 효과를 보이며, 전체 RMSE, R^2 등의 정량 지표 역시 XGBoost에서 더 좋은 값이 나타날 가능성이 높다.**

---

2. 주요 인사이트 및 분석 결과

- **일반 가격대(0~2,000만 원)에서는 두 모델 모두 예측 정확도가 매우 높으며, 실무에서 충분히 신뢰할 수 있는 결과를 보여줌.**
- **고가 차량(상위 10% 구간)의 경우, 데이터의 불균형과 고유 특성 부족으로 인해 예측값이 실제값에 비해 낮게 형성되는 경향이 여전히 존재.**
- **XGBoost와 같은 고급 모델을 활용하면 outlier에 대한 예측력을 일정 부분 보완할 수 있으며,**
  전체 예측의 안정성과 정확도 향상에 효과적임.

---

3. 실무적 제안 및 적용 방안

1. **일반 가격대(대중차 위주) 가격 예측 시스템에는 현재의 XGBoost 모델을 그대로 적용해도 충분한 실무 성과가 기대됨.**
2. **고가 차량에 대한 예측 성능 개선이 필요한 경우**
    - **고가 차량 데이터의 추가 확보(샘플 수 확대)**
    - **고가 차량 전용 피처 엔지니어링(옵션, 브랜드, 희귀 사양 등 특성화 변수 생성)**
    - **고가 구간에 대한 sample_weight 가중치 적용, 별도 모델링 혹은 앙상블 적용**
      등의 보완이 필요함.
3. **모델 운영 시**
    - 주기적으로 새로운 실거래 데이터를 반영해 학습 데이터의 품질을 높이고, 예측의 최신성과 실효성을 유지할 것을 권장함.
    - RMSE, R^2와 같은 정량적 지표를 계속 모니터링하며,
      고가 차량 등 특수 구간의 성능 개선 여부도 별도 관리할 것.

---

4. 결론

- **XGBoost 기반의 가격 예측 모델은 중고차 실거래 데이터 기반 가격 산정에서 높은 신뢰도와 정확도를 보임.**
- **특정 구간(고가 차량 등)에서의 성능 개선을 위해 추가 데이터 확보, 피처 엔지니어링, 가중치 부여 등 고도화 전략이 필요함.**
- **실무에 즉시 적용 가능하며, 정기적인 데이터 업데이트와 모델 성능 관리 체계가 병행되면 더욱 강력한 예측 시스템으로 발전할 수 있을 것으로 판단된다.**