기업 성공 확률 예측 모델 개발 리포트

"실제 기업 투자 판단에 활용
가능한 확률 예측 모델 개발"

프로젝트 개요

항목	내용
프로젝트 명	기업 성공 확률 예측 모델
플랫폼	**Dacon 해커톤: 기업 성공 확률 예측 AI 경진대회**
모델 목적	기업 특성과 투자 데이터를 바탕으로 0~1의 성공확률 예측
사용 모델	LightGBM + XGBoost + CatBoost 앙상블
최종 점수	MAE: 0.21657 / 상위 약 8% 성적

핵심 전략 요약

전략	설명	기대 효과
1. 정교한 전처리	결측/이상치 처리, 기업가치 문자열 변환, 범주형 인코딩 등 적용	학습 안정성 확보
2. 도메인 기반 파생 변수 생성	업력, ROI, 투자/직원, 매출/고객 등 실제 경영 지표로부터 파생 설계	설명력 강화
3. 모델 앙상블	LGBM 0.4 + XGB 0.3 + CatBoost 0.3 비중 앙상블	일반화 성능 개선
4. StratifiedKFold 적용	qcut 기반 구간 분리로 레이블 분포 보존	데이터 편향 감소
5. Feature Importance 선택	LGBM 기준 Top 25개 피처만 사용	과적합 방지 및 경량화

기술 스택

- Python 3.10
- Pandas / NumPy / scikit-learn
- LightGBM / XGBoost / CatBoost
- Stratified K-Fold Validation
- Matplotlib / Seaborn (분석 시각화)

단계별 개발 내역

◇ **1단계: 데이터 정제 및 전처리**

- 기업가치(백억원): "2500-3500" → 평균치 변환
- 결측치: 평균/중앙값 대체
- 이상치: 상위 1% 클리핑 적용
- 범주형 변수: LabelEncoder 처리

◇ **2단계: 변수 파생 및 피처 엔지니어링**

- 생존연수: 2025 - 설립연도
- ROI: 매출 / 투자
- 투자_직원비: 투자 / 직원 수
- 가치_투자비: 기업가치 / 투자
- 매출_고객비: 매출 / 고객
  도메인 지식을 기반으로 한 실제 지표 활용

◇ **3단계: 모델링 및 앙상블**

- 단일 모델 실험 (LGBM, XGB 등) → 성능 한계 존재
- 예측 분산을 줄이기 위한 Soft 앙상블 적용
- 5-Fold Cross Validation 사용

```
Python                                                    복사  편집
val_pred = (
        lgb_model.predict(X_val) * 0.4 +
        xgb_model.predict(X_val) * 0.3 +
        cat_model.predict(X_val) * 0.3
    )
```

성능 결과

구분	MAE (예측 정확도)
최종 검증 점수	0.21657
공식 제출 점수	0.21667
참여 등수	상위 약 8%

기술적으로 배운 점

항목	내용
모델별 장단점 이해	LGBM → 빠르고 안정적, XGB → 튜닝 민감, CatBoost → 범주형 강점
파생 변수 중요성 체감	단순 수치보다 비율, 상호작용 변수들이 성능에 큰 영향
앙상블 효과	분산 구조가 다른 모델 조합이 성능 안정화에 기여
검증 방법론 중요성	단순 train_test_split보다 StratifiedKFold가 훨씬 견고

개선 여지 및 향후 계획

항목	설명
Meta Stacking 실험	예측값을 메타 모델 입력으로 활용하는 고급 앙상블 시도 예정
Optuna 튜닝	자동 파라미터 탐색 도입 필요성 확인
시계열 요소 반영	설립연도나 업력에 따른 흐름 분석 및 적용
외부 데이터 연계	실제 재무제표/산업분류 등 확장 적용 가능성 고려

마무리

이번 **Dacon 해커톤 프로젝트**를 통해 실제 기업 투자 의사결정에 사용할 수 있는 정량적 판단 기반을 마련하는 모델을 설계하였으며,
전체 ML 파이프라인을 구성하는 과정에서 **모델링 역량, 피처 엔지니어링, 검증 전략 설계** 등의 실무 능력을 크게 향상시킬 수 있었습니다.