

교육생용 해상운임 ML 10단계 순차 실습 안내

Google Colab에서 데이터 품질부터 머신러닝, 딥러닝 시각화, 업무 시나리오까지

10개 실습 데이터

총 5시간 35분

Test 평가 1회

이 자료의 사용법

- 교육생용 노트북을 열고 각 단계의 설명을 읽은 뒤 TODO 코드를 직접 완성합니다.
- 셀은 위에서 아래로 실행합니다. 중간 셀을 건너뛰면 다음 변수가 생성되지 않습니다.
- 모델 선택과 조절은 Validation에서만 합니다. Test는 최종 모델 확정 후 단 한 번 실행합니다.
- 정답 포함 노트북은 실습을 끝낸 뒤 비교와 복습에 사용합니다.

학습의 최종 결과

- 운임 데이터의 품질 문제와 누출 위험을 설명할 수 있습니다.
- 시간순 분할로 기준선, Ridge, Random Forest를 공정하게 비교할 수 있습니다.
- 신경망의 입력층, 은닉층, 출력층과 학습곡선을 시각적으로 해석할 수 있습니다.
- 시나리오, 지연, 통관보류 결과를 업무 의사결정 언어로 전달할 수 있습니다.

시작 전 10분 | Colab 업로드와 실행

최신 교육생용 노트북을 사용해 수정된 오류 처리와 딥러닝 시각화가 반영됩니다.

순서	교육생이 할 일	완료 기준
1	배포 ZIP과 교육생용 노트북을 PC에 준비합니다.	두 파일의 위치 확인
2	Google Colab에서 파일 > 노트 업로드 를 선택합니다.	새 Colab 탭 표시
3	해상운임_ML_10단계_교육생용.ipynb 를 업로드합니다.	제목에 교육생용 표시
4	첫 Setup 셀을 실행합니다. 데이터를 찾지 못하면 배포 ZIP을 압축하지 않은 상태로 업로드합니다.	10개 CSV 인식
5	1단계부터 셀을 하나씩 실행하고 TODO를 해결합니다.	오류 없이 다음 단계 이동

파일 구성

파일/폴더	용도
해상운임_ML_10단계_교육생용.ipynb	TODO를 직접 완성하는 기본 실습 파일
해상운임_ML_10단계_정답포함.ipynb	완료 뒤 비교와 복습에 쓰는 실행 검증본
data	1~10단계에서 사용하는 CSV 10개
데이터사전.md / 과정개요.md	컬럼 의미, 누출 여부, 단계별 목표와 시간
outputs/day2-freight-automl	모델 비교표, 그림, 보고서 예시

진행 원칙

- 각 단계에서 표 또는 그래프를 확인하고 마지막에 업무 해석 3문장을 작성합니다.
- TODO가 남아 있으면 마지막 Checks 셀이 누락 단계를 안내하고 보고서를 생성을 건너뛰니다. 이것은 오류가 아닙니다.
- Colab 탭은 PC 파일 수정 내용을 자동 반영하지 않습니다. reg_train NameError가 보이면 최신 교육생용 노트북을 다시 업로드합니다.

전체 학습 로드맵

앞 단계의 산출물이 다음 단계의 입력이 됩니다. 순서를 바꾸지 마세요.

단계	주제	시간	핵심 산출물
1	문제와 스키마 이해	20분	프로파일 표
2	데이터 품질 진단	35분	품질 점검표와 정제 규칙
3	EDA와 파생변수	30분	차트 2개와 해석 3문장
4	데이터 누출 실험	25분	누출 블랙리스트
5	시간순 분할	35분	분할/성능 비교표
6	업무 기준선	25분	기준선 지표와 fallback
7	회귀 모델과 Test 1회	70분	모델 비교표와 최종 Test
7-B	딥러닝 구조와 학습 시각화	7단계 포함	구조 그림과 학습곡선
8	견적 시나리오	30분	5개 변화량과 변화율
9	선적 지연 분류	35분	혼동행렬과 재현율
10	B/L OCR와 통관보류	45분	원문대조표와 검토 큐

두 구간으로 운영하는 방법

- 1교시 묶음: 1~4단계 - 데이터 이해, 품질, EDA, 누출 방지
- 2교시 묶음: 5~10단계 - 시간순 검증, 모델링, 딥러닝, 시나리오, 분류
- 시간이 부족하면 7-B는 강사 시연으로 진행하되 Test 1회 원칙은 유지합니다.

1~4단계 | 좋은 모델보다 먼저 좋은 데이터

문제 정의, 품질 점검, 탐색, 누출 차단을 순서대로 수행합니다.

STEP 1

문제 정의와 데이터 프로파일링 | 20분

실습: 예측 목표를 해상운임_USD로 정하고 단위를 USD/컨테이너로 명시합니다. 행/열, 기간, 결측, 중복, 운임 요약을 확인합니다.

확인: 선적ID 유일성, 건적일/선적일 역할, 데이터 기간을 설명합니다.

제출: 프로파일 표

주의: 코드보다 먼저 예측 대상, 예측 시점, 단위, 기준일을 적습니다.

STEP 2

결측, 중복, 논리 오류, 이상치 | 35분

실습: 결측, 중복, 음수 중량, 100 초과 혼잡도, 음수 리드타임, IQR 이상치를 따로 탐지합니다.

확인: 오류 유형별 건수와 처리 규칙이 연결되는지 확인합니다.

제출: 품질 점검표와 정제 규칙

주의: 높은 운임은 성수기나 우회 신호일 수 있으므로 자동 삭제하지 않습니다.

STEP 3

EDA와 업무 파생변수 | 30분

실습: 항로, 월, 컨테이너규격, 계약유형별 운임 분포와 추세를 시각화합니다.

확인: 차트 제목, 축, USD 단위와 근거 기반 해석을 확인합니다.

제출: 차트 2개와 해석 3문장

주의: 상관관계를 가격의 인과효과라고 단정하지 않습니다.

STEP 4

데이터 누출 함정 | 25분

실습: 컬럼을 건적 시점 사용 가능/불가로 분류하고 사후 컬럼 포함/제외 성능을 비교합니다.

확인: 목표값, 실제운송일수, 지연일수, 최종할증료, 청구운임, ID가 최종 입력에 없는지 봅니다.

제출: 누출 블랙리스트

주의: 성능이 지나치게 좋으면 누출부터 의심합니다. 시장값도 건적 시점 스냅샷만 씁니다.

4단계 종료 체크

- 어떤 시점의 정보로 무엇을 예측하는지 설명할 수 있다.
- 오류, 이상치, 실제 이벤트 신호를 서로 구분할 수 있다.
- 최종 모델에서 제외할 누출 컬럼 목록을 만들었다.

5~7단계 | 시간순 검증과 최종 Test 1회

Validation은 반복 사용 가능하지만 Test는 최종 모델 확정 뒤 한 번만 사용합니다.

STEP 5

무작위 분할과 시간순 70/15/15 | 35분

실습: 같은 날짜를 한 구간에 유지하며 Train, Validation, Test를 시간순으로 나누고 무작위 분할과 비교합니다.

확인: 날짜 교집합이 0이고 $\max(\text{Train}) < \min(\text{Validation}) < \min(\text{Test})$ 인지 봅니다.

제출: 분할 기간표와 성능 비교표

주의: 전처리하는 Train에서만 학습하고 아직 Test 목표값을 보지 않습니다.

STEP 6

업무 중앙값 기준선 | 25분

실습: Train의 항로 x 컨테이너규격 x 계약유형 중앙값으로 Validation을 예측합니다. 미등장 조합은 Train 전체 중앙값으로 대체합니다.

확인: MAPE, RMSE, R2와 fallback 건수를 계산합니다.

제출: 기준선 지표와 fallback 규칙

주의: Validation 목표값으로 중앙값을 만들면 누출입니다.

STEP 7

기준선, Ridge, Random Forest 비교 | 70분

실습: 같은 Validation에서 세 모델을 비교해 최종 모델을 고정한 뒤 Train+Validation 재학습과 Test 1회 평가를 수행합니다.

확인: 기준선보다 나쁜 ML은 채택하지 않고 성능, 설명가능성, 운영 난이도를 기록합니다.

제출: 모델 비교표와 최종 Test 지표

주의: Test 결과를 본 뒤 모델이나 트리 수를 바꾸지 않습니다.

Train

Validation

최종 모델 고정

Test 1회

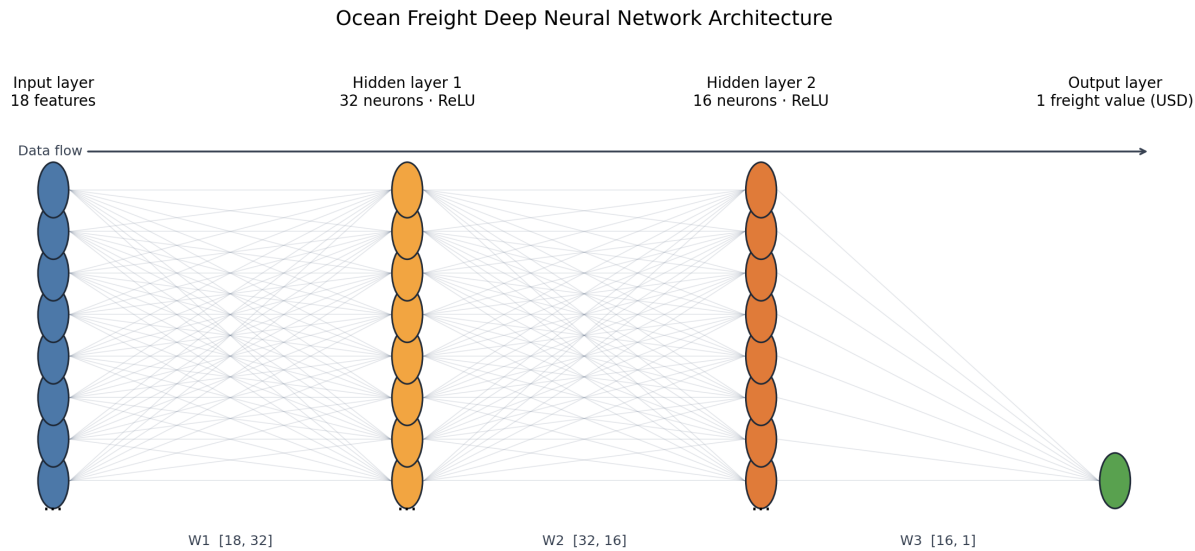
조절 가능한 횟수와 고정 규칙

- RF_TREE_COUNT: 50~1000, 기본값 220. 예: 100, 220, 500을 Validation에서 비교합니다.
- 트리 수는 학습 반복량입니다. Test 평가 횟수와는 다른 개념입니다.
- Test 실행 횟수는 1로 고정합니다. 모델 선택과 튜닝이 끝난 뒤에만 실행합니다.

7-B | 딥러닝 구조와 학습 과정 시각화

딥러닝은 Validation에서 구조와 학습을 이해하는 보조 실습입니다. 최종 Test는 추가 사용하지 않습니다.

신경망 구조



입력층은 운임 특성을 받고, 두 은닉층이 비선형 패턴을 학습하며, 출력층 1개가 예상 해상운임을 반환합니다.

조절값

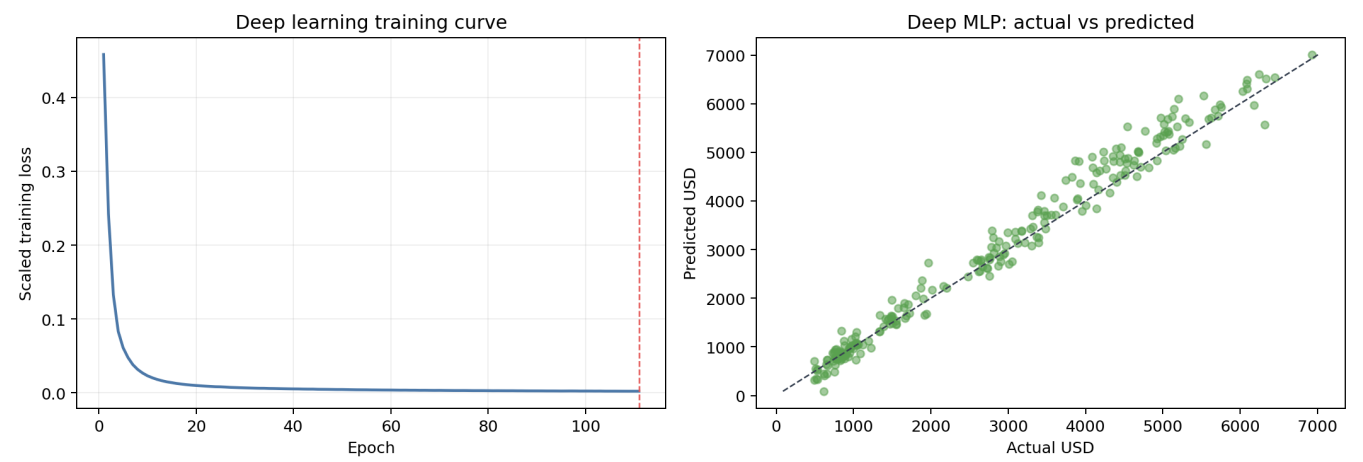
항목	기본값	범위	해석
DEEP_EPOCHS	250	50~500	최대 학습 반복 횟수
DEEP_HIDDEN_1	32	8~128	첫 번째 은닉층 노드 수
DEEP_HIDDEN_2	16	4~64	두 번째 은닉층 노드 수

해석 포인트

- 조기 종료가 작동하면 실제 epoch는 설정한 최대값보다 작을 수 있습니다.
- 학습 손실은 감소하지만 Validation 성능이 나빠지면 과적합을 의심합니다.
- 작은 표형 데이터에서는 Random Forest가 신경망보다 우수할 수 있습니다.
- 목적은 구조와 학습 과정의 이해이며 신경망 사용 자체가 목표는 아닙니다.

7-B 결과 읽기 | 학습곡선과 실제값-예측값

손실곡선의 방향과 Validation 오차를 함께 봅니다.



그래프를 읽는 순서

순서	확인 항목	판단 기준
1	학습 손실	반복이 진행될수록 안정적으로 감소하는지 확인
2	조기 종료	개선이 멈춘 지점에서 학습이 종료되었는지 확인
3	실제값-예측값	점들이 45도 기준선 주변에 모이는지 확인
4	Validation 지표	MAPE, RMSE, R2, WAPE, bias를 Random Forest와 비교
5	결론	성능이 낮아도 구조와 과적합 신호를 설명하면 목표 달성

실습 기록 예시

- 설정값: epoch, 첫 번째 은닉층, 두 번째 은닉층을 기록합니다.
- 관찰값: 실제 종료 epoch와 Validation 지표를 기록합니다.
- 판단: Random Forest 대비 장점과 단점을 한 문장씩 적습니다.

8~10단계 | 예측을 업무 의사결정으로 연결

회귀를 시나리오로 해석하고 지연과 통관보류 분류까지 확장합니다.

STEP
8

5개 견적 시나리오 | 30분

실습: 기본 조건과 유가 +10%, 혼잡 +20점, 홍해 우회, 계약 전환, 복합 악화를 비교합니다.

확인: 기본 대비 변화 USD와 변화율, 학습범위 초과 여부를 봅니다.

제출: 시나리오 결과표와 업무 해석

주의: 조건부 추정이며 확정 견적이나 인과효과가 아닙니다.

STEP
9

3일 이상 선적 지연 분류 | 35분

실습: Logistic과 Random Forest를 비교하고 임계값별 혼동행렬, Precision, Recall, F1, PR-AUC를 확인합니다.

확인: 지연 누락 FN과 허위 경보 FP 비용을 비교해 임계값을 고릅니다.

제출: 혼동행렬, 양성 Recall, 임계값 근거

주의: 불균형 데이터에서는 정확도만으로 선정하지 않습니다.

STEP
10

B/L OCR 검증과 통관보류 분류 | 45분

실습: B/L 7개 필드를 원문과 대조하고 보류 Recall과 PR-AUC로 사람 검토 우선순위 큐를 만듭니다.

확인: 보류사유, 검사결과, 해제일 같은 사후정보가 입력에 없는지 확인합니다.

제출: 원문대조표와 통관 검토 큐

주의: HS코드는 후보만 제시하고 '전문가 확인 필요'로 표시합니다. 자동 거절에 쓰지 않습니다.

업무 해석 3문장 틀

구분	작성 질문
1. 관찰	어떤 수치와 패턴이 나타났는가?
2. 한계	조건부 추정, 표본 범위, 누출 위험, 오분류 비용은 무엇인가?
3. 행동	예산, 계약, surcharge, 조기 부킹, 사람 검토 중 무엇을 제안할 것인가?

최종 제출 체크리스트

아래 항목을 모두 확인한 뒤 정답 포함 노트북과 비교합니다.

- [] 1단계 프로파일 표

[] 2단계 품질 점검표와 정제 규칙

[] 3단계 EDA 차트 2개와 해석 3문장

[] 4단계 누출 블랙리스트

[] 5단계 시간순/무작위 분할 비교표

[] 6단계 기준선 지표와 fallback 건수
- [] 7단계 모델 비교표와 최종 Test 1회 결과

[] 7-B 신경망 구조, 학습곡선, Validation 지표

[] 8단계 5개 건적 시나리오 결과

[] 9단계 지연 분류 혼동행렬과 임계값 근거

[] 10단계 B/L 원문대조표와 통관 검토 큐

자주 만나는 문제

증상	확인하고 해결
reg_train is not defined	앞의 시간순 분할 TODO를 실행합니다. 계속되면 NEW 실습의 최신 교육생용 노트북을 다시 업로드합니다.
데이터 폴더를 찾지 못함	Setup 셀 안내에 따라 배포 ZIP을 압축하지 않은 상태로 업로드합니다.
마지막 보고서가 생성되지 않음	Checks 셀이 표시한 TODO를 해결합니다. 건너뛰기는 정상 보호 동작입니다.
딥러닝 epoch가 설정값보다 작음	조기 종료가 개선 정체를 감지한 정상 결과입니다. 실제 종료 epoch를 기록합니다.
Random Forest 결과가 달라짐	SEED 20260722를 확인하고 런타임 재시작 뒤 위에서부터 실행합니다.

모델 선택과 튜닝은 Validation에서만 수행합니다.
Test는 최종 모델을 확정된 후 단 한 번만 평가하며,
Test 결과를 보고 모델을 다시 바꾸지 않습니다.