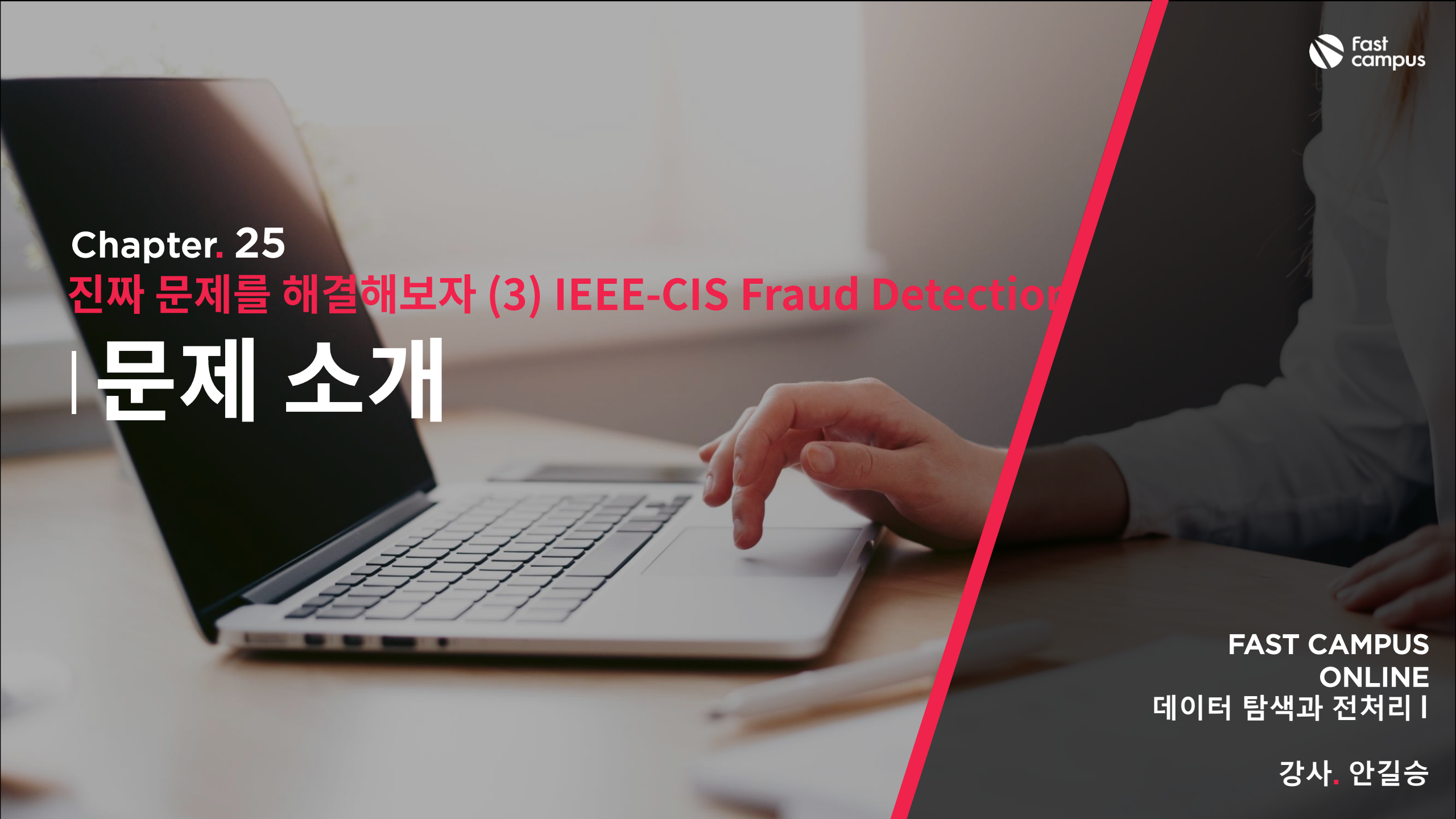


Chapter. 25

진짜 문제를 해결해보자 (3) IEEE-CIS Fraud Detection

문제 소개

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

I 대회 소개

- 출처: 캐글
  - 문제 제공자: IEEE Computational Intelligence Society
  - https://www.kaggle.com/c/ieee-fraud-detection/overview

- 문제 개요: 카드 기록을 활용한 사기 거래 탐지

- 대부분 컬럼이 비식별화되어 있어, 이전 문제와 다르게 **데이터 탐색을 통한 데이터에 대한 이해**가 매우 중요한 상황임

I 사용 데이터

- train.csv: 모델 학습용 데이터 (데이터가 커서 샘플링된 데이터 제공) / test.csv: 모델 평가용 데이터
  - TransactionID: 거래 ID (비식별화)
  - TransactionDT: 거래 시각 (비식별화)
  - TransactionAmt: 거래 금액 (US Dollar)
  - ProductCD: 상품 코드
  - card1 – card6: 카드 관련 정보 (비식별화)
  - P_emaildomain, R_emaildomain: 이메일 정보
  - M1 – M9: 기존 거래와의 매칭 정보
  - isFraud: 사기 거래 여부

I 사용 데이터의 특징

- 비식별화된 특징이 매우 많음
- 다수 결측이 포함되어 있음
- 클래스 불균형 문제 존재

I 데이터 분리

- df: sampled_train_transaction.csv
- 'TransactionID', 'TransactionDT' 변수는 삭제
- X: df에서 isFraud가 제거된 데이터 프레임
- Y: df에서 isFraud값만 가져온 Series
- Train_X, Test_X, Train_Y, Test_Y

Chapter. 25

진짜 문제를 해결해보자 (3) IEEE-CIS Fraud Detection

변수 탐색

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

I 라벨 확인

- Train_Y에 대해, value_counts를 적용하여 라벨 분포 확인
- 클래스 불균형 문제가 발생할 확률이 매우 높아 보임

변수별 상태 공간 확인

- Train_X에 포함된 모든 컬럼에 대해, 타입, 결측 개수, 상태공간 크기, 상태 공간 일부를 확인
- 이를 바탕으로 결측이 포함된 변수와 그렇지 않은 변수, 연속형 변수와 범주형 변수를 구분

I 탐색을 위한 데이터 준비

- Train_X와 Train_Y를 열 기준으로 병합하여 탐색을 위한 데이터를 준비: Train_df
- 이는 특징에 따른 라벨의 분포를 확인하기 위한 선행 작업임

범주형 변수 탐색

- 탐색 방법
  - 결측을 문자로 변환 (탐색을 위해 임시 변환)
  - 변수별 분포 확인 (bar plot)
  - 변수와 특징 간 관계 확인 (groupby)

- 탐색 결과 활용
  - 주요 값 기준 이진화
  - 더미화

연속형 변수 탐색

- 탐색 방법
  - 히스토그램을 통한 변수별 분포 확인
  - 박스 플롯을 이용한 변수와 라벨 간 관계 파악

- 탐색 결과 활용
  - 연속형 변수 이진화
  - 변수 치우침 제거

- 선형 관계가 보이지 않는 변수가 대부분이어서 트리 기반의 앙상블 모델을 활용하기로 결정

Chapter. 25

진짜 문제를 해결해보자 (3) IEEE-CIS Fraud Detection

데이터 전처리

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

I 이진화

- 특정 값을 갖는 변수로 변환하면, 결측 값은 자동으로 0을 갖게 되므로 결측 처리를 생략할 수 있음
- card4 이진화: american express라는 값을 갖는지 여부를 나타내는 특징으로 변환
- card6 이진화: creidt이라는 값을 갖는지 여부를 나타내는 특징으로 변환
- NA_R_emaildomain 변수 생성: R_emaildomain 변수가 결측인지를 나타내는 특징 생성
- same_email 변수 생성: R_emaildomain과 P_emaildomain이 같은지를 나타내는 특징 생성
- C3_over_1 변수 생성: C3 값이 1보다 크거나 같은지를 나타내는 변수 생성

I 결측 대체 및 더미화 수행

- 더미화를 수행하기 위해, M1 – M9 변수에 발생한 결측 값을 문자로 치환
- ProductCD 변수와 M1– M9 변수에 대한 더미화 수행

I 치우침 해소

- 로그 변환을 통해, TransactionAmt에 있는 변수 치우침 해결

I 연속형 변수를 이진화

- card3이라는 변수가 150인지, 185인지를 나타내는 변수 생성
- card5라는 변수가 226인지를 나타내는 변수 생성

I 결측 대체

- 이진화가 되지 않은 특징에 대한 결측치를 처리하기 위해, Imputer를 사용

I 클래스 불균형 문제 해소

- 사기 거래 수가 정상 거래에 비해 매우 적으므로, 클래스 불균형 문제가 존재할 가능성이 매우 큼
- 이를 방지하기 위해, 언더샘플링과 비용민감모델을 고려 (오버샘플링을 하기에는 샘플이 너무 많음)

Chapter. 25

진짜 문제를 해결해보자 (3) IEEE-CIS Fraud Detection

모델 학습

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

비용 민감 모델: 파라미터 그리드

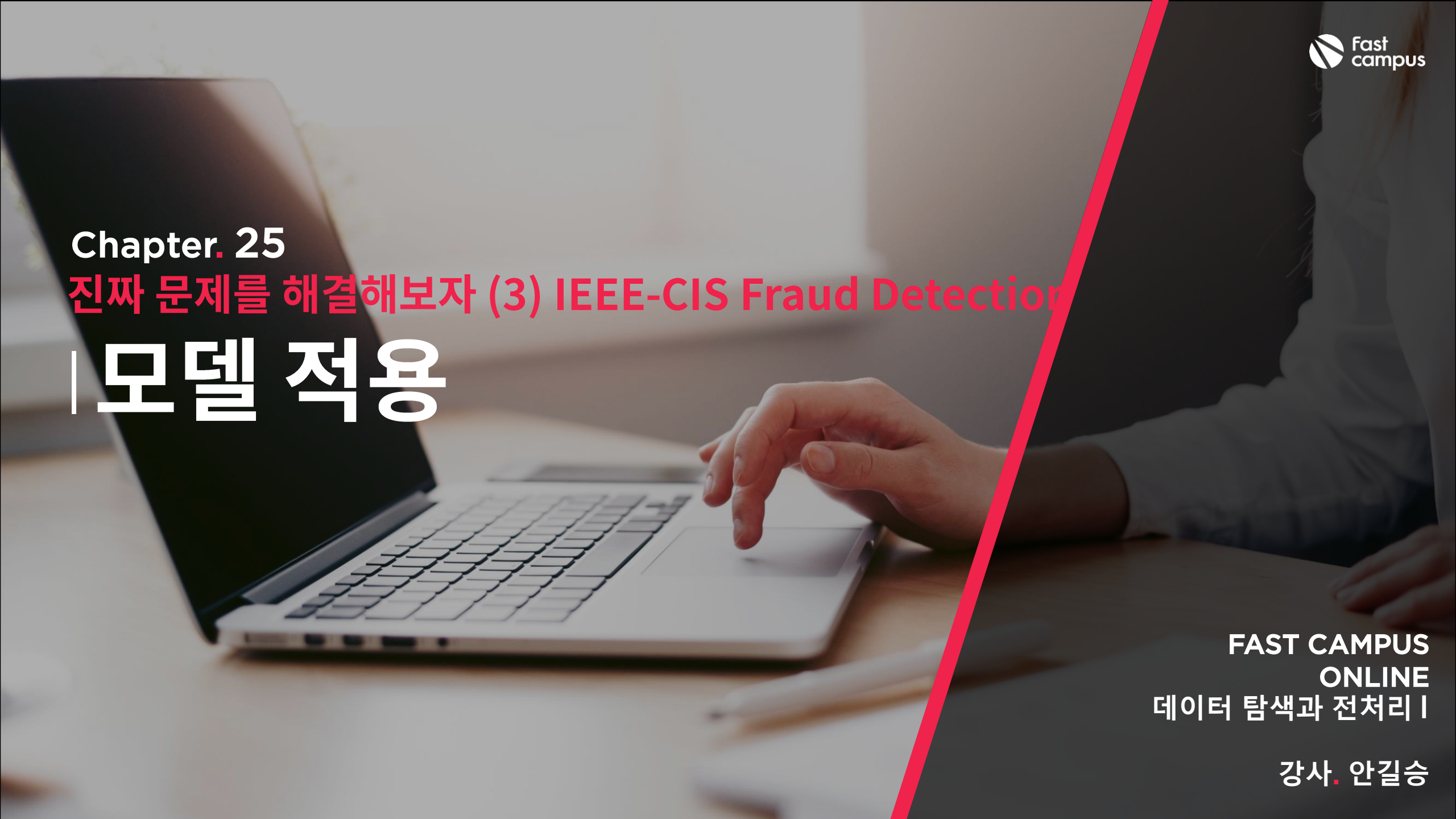
- 고려하는 파라미터 조합은 다음과 같음

특징 선택 기준	선택하는 특징 개수	모델	파라미터
상호 정보량	{5, 10, 15, 20, 25, 30}	Random Forest Classifier	✓ n_estimators: {100, 200} ✓ max_depth: {3, 4, 5, 6} ✓ Class_weight, $w \in${1, 0.9, 0.7, 0.5} (w는 다음과 같이 반영: {1: $w * C$, 0: 1}, 여기서 C는 클래스 불균형 비율
		XGB Classifier	✓ n_estimators: {100, 200} ✓ max_depth: {3, 4, 5, 6} ✓ learning_rate: {0.05, 0.1, 0.2} ✓ Class_weight, $w \in${1, 0.9, 0.7, 0.5} (w는 다음과 같이 반영: {1: $w * C$, 0: 1}, 여기서 C는 클래스 불균형 비율

일반 모델: 파라미터 그리드

- 고려하는 파라미터 조합은 다음과 같음

특징 선택 기준	선택하는 특징 개수	모델	파라미터	Near Miss 파라미터
상호 정보량	{5, 10, 15, 20, 25, 30}	Random Forest Classifier	✓ n_estimators: {100, 200} ✓ max_depth: {3, 4, 5, 6}	✓ version: 2 ✓ Sampling_strategy: {1: w}, 여기서 $w \in \{1, 0.9, 0.7, 0.5\}$
		XGB Classifier	✓ n_estimators: {100, 200} ✓ max_depth: {3, 4, 5, 6} ✓ learning_rate: {0.05, 0.1, 0.2}	



Chapter. 25

진짜 문제를 해결해보자 (3) IEEE-CIS Fraud Detection

모델 적용

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

파이프라인 구축

- 새로운 데이터에 대한 예측을 수행하기 위해, 하나의 함수 형태로 파이프라인을 구축함
- 파이프라인을 사용하여 새로운 데이터를 예측함