미국 은행 재무제표 분석 프로젝트

미국 NYSE 상장 은행 98개사 5개년 분기보고서 모델 데이터 전처리 및 머신러닝

도입

- 대다수의 재무분석은 제조판매업 기업 대상
- 금융회사의 회계시스템은 타 산업과 상이
- 미국: 상장되어 정기공시를 진행하는 금융사가 많음

→ 미국의 은행 재무제표 데이터를 분석하면, 새롭고 유의미한 결과를 얻을 수 있을 것

사용 언어	분석 모델	사용 프로그램
➢ Python 3 (100%)	➢ Random Forest ➢ XG Boost ➢ Linear Regression ➢ CNN ➢ DNN ➢ LSTM	➢ VSCode ➢ Jupyter Notebook ➢ Google Colab ➢ MS Excel ➢ Chat GPT ➢ Claude

프로젝트 진행 일정

07.01 브레인스토밍 및 주제 후보선정

07.02 주제 확정 및 데이터 범위선정

07.03 SEC 공시 데이터 자동추출 코드 작성 및 데이터 크롤링

07.04 누락된 데이터 조치, 모델 구상, 미국 재무제표 데이터 특유의 전처리가 난해한 지점들 단순화, 컬럼 선정

07.05 컬럼 선정(계속), 보다 효율적인 재무제표 Raw 데이터 추출 방안 고민

07.06 7월 3일의 자동추출 코드를 사용해 재무제표 엑셀 파일 다운로드, 추출 가능한 항목들을 이용한 효과적 재무분석을 위해 전략 보완

07.07 자동 다운로드 과정에서 누락된 파일 보완, Raw 데이터에서 CSV파일을 추출하는 코드 작성

07.08 크롤링 코드와 CSV추출 코드 보완 및 샘플 데이터셋에 시험적용

07.09 재무제표 항목 수작업 매핑 및 CSV추출 코드 보완

07.10 재무제표 항목 수작업 매핑 및 CSV추출 코드 보완

07.11 자동화를 통한 98개 은행 재무제표 전처리 완료 및 모델링 분석 테스트

07.12 전처리된 데이터셋에 오류 발견, 데이터셋 변경 및 모델링 분석 테스트와 시각화

Web Crawling

미 증권거래위원회 전자공시시스템 (SEC EDGAR)

- 기업별로 회계공시이력 API, CSV 제공
- 공시이력이 아닌 공시 본문이 필요
- Excel 형식의 보고서 파일을 추출하기 위한 동적 크롤링 기법 사용

데이터 전처리

표기가 비슷하고 필수적으로 나타나기에 전처리가 상대적으로 용이한 항목들을 추출

손익계산서 항목	재무상태표 항목
➢ interest income (이자수익) ➢ net interest income (순 이자수익) ➢ non-interest income(비이자수익) ➢ income before tax (세전수익) ➢ tax expense (법인세비용) ➢ net income (순이익)	➢ cash [and other cash equivalents] (현금 [및 현금성자산]) ➢ goodwill (영업권) ➢ loans (대출)* ➢ net loans (순대출)* ➢ total assets (총자산) ➢ total liabilities (총부채) ➢ total [shareholders'/stockholders'] equity (총자본) ➢ retained earnings (이익잉여금. 유보이익이라고도 함)

*: 원 표기상의 전처리 한계로 인한 일부 누락값 존재

데이터 전처리

```python
# 2. 특정 파일과 시트에서 금융 데이터를 추출하는 함수
def extract_financial_data(file_path, sheet_names, data_items, date_columns):
    results = []
    try:
        xls = pd.ExcelFile(file_path)
        available_sheets = xls.sheet_names

        for sheet_name in sheet_names:
            if sheet_name in available_sheets:
                df = pd.read_excel(file_path, sheet_name=sheet_name, header=[0, 1])
                df.columns = [f'{col[0]} {col[1]}' if 'Unnamed' not in col[1] else col[0] for col in df.columns]

                # 시트 이름에 특정 단어가 포함된 경우 확인
                unit_in_thousands = any('in Thousands' in col for col in df.columns)
                unit_in_millions = any('in Millions' in col for col in df.columns)
```

최종 데이터셋

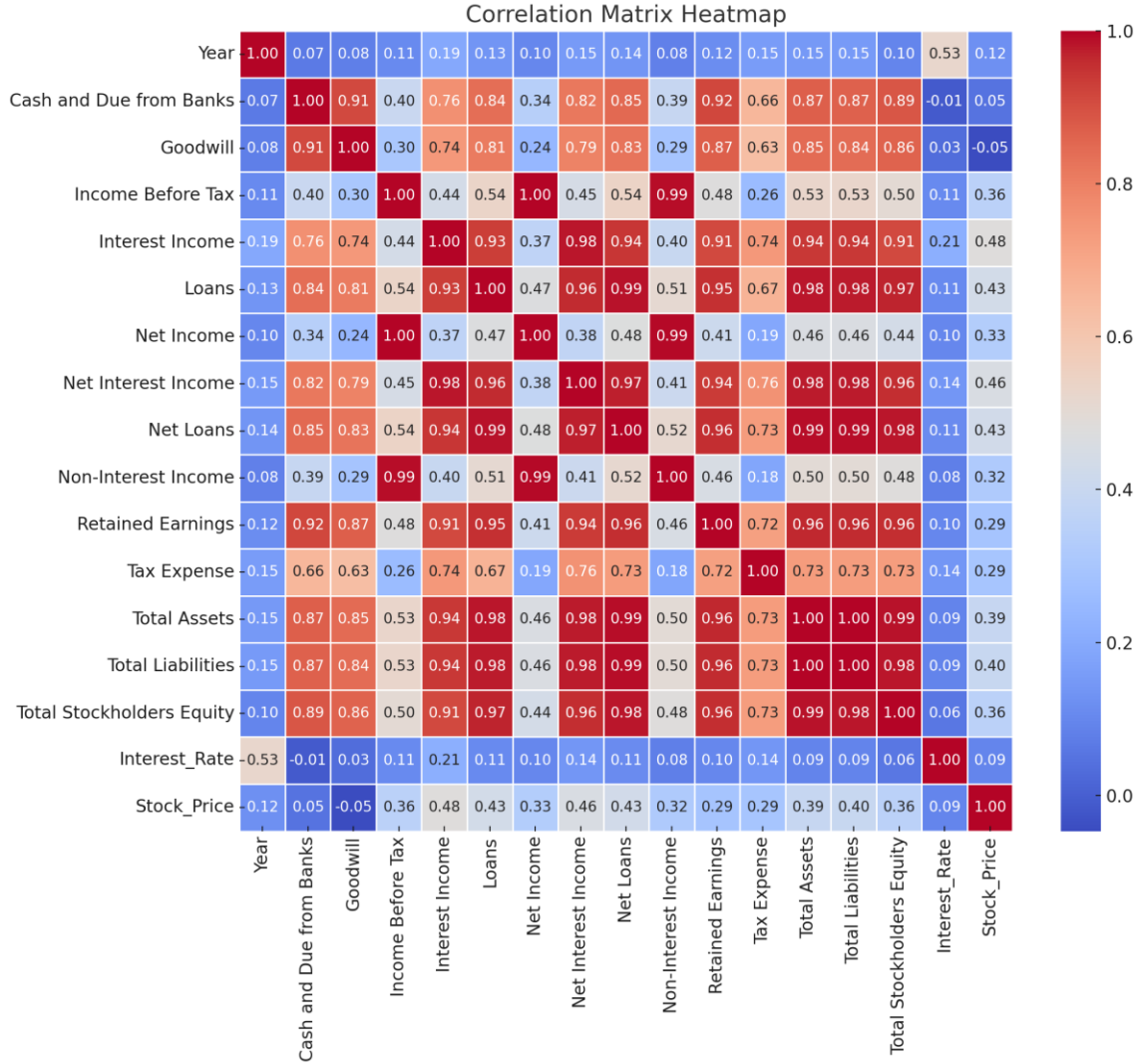
Name	Year	Quarter	Cash and	Goodwill	Income Be	Interest In	Loans	Net Incom	Net Intere	Net Loans	Non-Intere	Retained E	Tax Expen	Total Asse	Total Liab	Total Stoc	Interest_Ra	Stock_Price
BANF	2018	Mar. 31	181.9	79.8	36.9	70.9	4984.5	29.6	63	4932.9	30.1	661.3	7.3	7615.6	6777.5	838.1	1.625	46.171
BANF	2018	Jun. 30	188.5	79.7	39.8	75.1	5007.5	30.6	64.9	4955.3	30.4	684.4	9.2	7623	6761	862	1.875	51.681
BANF	2018	Sep. 30	185	79.7	41.9	77.4	4947.5	32.9	65.7	4895.7	32.8	707.5	9	7602.4	6717.6	884.8	2.125	52.52
BANF	2018	Dec. 31	228.4	79.7	41.2	79.8	4976	32.7	66.9	4924.6	31.9	722.6	8.4	7574.3	6671.5	902.8	2.375	43.934
BANF	2019	Mar. 31	186	79.7	41	80.9	5042.5	31.8	66.9	4989.6	32	744.7	9.2	7709	6781.1	927.9	2.375	46.193
BANF	2019	Jun. 30	185.4	79.7	43.8	83.1	5094.4	34.2	68.8	5039.3	34.1	769.1	9.7	7642	6685.6	956.4	2.375	49.589
BANF	2019	Sep. 30	225.5	147	43	86.4	5606.8	33.4	72.3	5550.9	35.6	792	9.6	8388.8	7409.1	979.8	1.875	49.645
BANF	2019	Dec. 31	222	148.6	41.8	86.3	5662.1	69.6	73.9	5607.9	35.5	815.5	6.2	8565.8	7560.8	1005	1.625	56.257
BANF	2020	Mar. 31	192	149.9	28.2	84	5989.9	22.6	74.1	5919.8	35.1	826.9	5.6	8669.1	7645.7	1023.4	0.125	30.218
BANF	2020	Jun. 30	205.2	149.9	25.3	81.5	6675	20.7	77.2	6585.5	32.1	837.2	4.6	9612.5	8578.3	1034.2	0.125	37.093
BANF	2020	Sep. 30	226.1	149.9	25.6	79.2	6612.1	20.9	75.9	6506	34.6	846.9	4.7	9618.9	8575.1	1043.8	0.125	37.664
BANF	2020	Dec. 31	280.5	149.9	44.4	82.4	6394.5	35.4	79.5	6303.1	35.4	871.2	9	9212.4	8144.5	1067.9	0.125	54.582
BANF	2021	Mar. 31	274.1	149.9	52.2	80	6358.4	42.5	77.2	6267.5	39.9	898	9.7	10549.3	9454.6	1094.7	0.125	66.12
BANF	2021	Jun. 30	268.3	149.9	62.9	84.9	6191.2	48.2	82.4	6107.3	44.6	935.1	14.7	11015.3	9883.7	1131.6	0.125	58.677
BANF	2021	Sep. 30	274.1	149.9	48.3	83.2	6016.9	38.8	80.2	5930.5	39.8	950.6	9.5	11302.8	10155.9	1146.9	0.125	56.812
BANF	2021	Dec. 31	228.8	149.9	45	78.9	6169.4	38.1	75.9	6085.5	45.7	977.1	6.9	9405.6	8233.9	1171.7	0.125	67.078
BANF	2022	Mar. 31	274.9	176.6	43.7	78.5	6494.3	35.9	75.5	6407.1	43.6	1001.2	7.8	12624.4	11456.6	1167.8	0.375	79.507
BANF	2022	Jun. 30	289	183.6	55.2	91.5	6613.3	44.7	86.9	6526.3	42.6	1034.1	10.5	12530.1	11344.4	1185.7	1.625	91.841
BANF	2022	Sep. 30	227	182.1	68.3	114	6827.8	55.4	100.9	6737.9	49.3	1076.3	13	12452.4	11257.2	1195.1	3.125	86.175
BANF	2022	Dec. 31	259	182.1	70.2	135.8	6943.6	57.1	110.4	6850.8	48.2	1120.3	13	12387.9	11137	1250.8	4.375	85.304
BANF	2023	Mar. 31	212.8	182.1	74.3	145.4	7118.6	57.5	109.2	7023.8	47.8	1164.7	16.8	12332.1	11021.2	1310.9	4.875	80.757
BANF	2023	Jun. 30	221.1	182.1	70	150.8	7298.7	55	105.9	7201.8	48	1206.5	15	12020.3	10679.5	1340.8	5.125	89.835
BANF	2023	Sep. 30	202.7	182.3	65.2	160.2	7472.6	51	104.3	7374.8	44.4	1241.5	14.2	12114.6	10744	1370.6	5.375	85.07
BANF	2023	Dec. 31	225.5	182.3	60.4	167.5	7656.6	49	105.1	7559.8	45.2	1276.3	11.5	12372	10938.2	1433.9	5.375	95.943
BANF	2024	Mar. 31	183.5	182.3	64.2	171.6	7781.9	50.3	106.1	7684.6	44.9	1312.5	13.9	12602.4	11133.1	1469.3	5.375	87.15
BPOP	2018	Mar. 31	280.1	627.3	113.5	453.1	24224.8	91.3	393	23962	113.5	1261.8	22.2	45756.8	40691.9	5064.9	1.625	34.494
BPOP	2018	Jun. 30	400.6	627.3	251.2	480.9	24752.7	279.8	414.1	23965.5	234.8	1515.1	-28.6	47535.2	42245.5	5289.7	1.875	37.885
BPOP	2018	Sep. 30	400.9	687.5	182.7	528.4	26662	140.6	451.5	25878.5	151	1629.7	42	47919.4	42675.1	5244.3	2.125	42.946

Target Data: 주가

주가	Linear Regression	Random Forest	XG Boost
MSE	0.3861	0.1153	0.1275
RMSE	0.6214	0.3395	0.3571
R2 Score	0.5199	0.8566	0.8414
F1 Score	0.4615	0.4828	0.7619

Target Data: 금리

금리	Linear Regression	Random Forest	XG Boost
MSE	2.1956	0.3185	0.1904
RMSE	1.4818	0.5644	0.4363
R2 Score	0.3659	0.9080	0.9450
F1 Score	0.8333	0.9667	0.9667


Correlation Matrix Heatmap
Year 1.00 0.07 0.08 0.11 0.19 0.13 0.10 0.15 0.14 0.08 0.12 0.15 0.15 0.15 0.10 0.53 0.12
Cash and Due from Banks 0.07 1.00 0.91 0.40 0.76 0.84 0.34 0.82 0.85 0.39 0.92 0.66 0.87 0.87 0.89 -0.01 0.05
Goodwill 0.08 0.91 1.00 0.30 0.74 0.81 0.24 0.79 0.83 0.29 0.87 0.63 0.85 0.84 0.86 0.03 -0.05
Income Before Tax 0.11 0.40 0.30 1.00 0.44 0.54 1.00 0.45 0.54 0.99 0.48 0.26 0.53 0.53 0.50 0.11 0.36
Interest Income 0.19 0.76 0.74 0.44 1.00 0.93 0.37 0.98 0.94 0.40 0.91 0.74 0.94 0.94 0.91 0.21 0.48
Loans 0.13 0.84 0.81 0.54 0.93 1.00 0.47 0.96 0.99 0.51 0.95 0.67 0.98 0.98 0.97 0.11 0.43
Net Income 0.10 0.34 0.24 1.00 0.37 0.47 1.00 0.38 0.48 0.99 0.41 0.19 0.46 0.46 0.44 0.10 0.33
Net Interest Income 0.15 0.82 0.79 0.45 0.98 0.96 0.38 1.00 0.97 0.41 0.94 0.76 0.98 0.98 0.96 0.14 0.46
Net Loans 0.14 0.85 0.83 0.54 0.94 0.99 0.48 0.97 1.00 0.52 0.96 0.73 0.99 0.99 0.98 0.11 0.43
Non-Interest Income 0.08 0.39 0.29 0.99 0.40 0.51 0.99 0.41 0.52 1.00 0.46 0.18 0.50 0.50 0.48 0.08 0.32
Retained Earnings 0.12 0.92 0.87 0.48 0.91 0.95 0.41 0.94 0.96 0.46 1.00 0.72 0.96 0.96 0.96 0.10 0.29
Tax Expense 0.15 0.66 0.63 0.26 0.74 0.67 0.19 0.76 0.73 0.18 0.72 1.00 0.73 0.73 0.73 0.14 0.29
Total Assets 0.15 0.87 0.85 0.53 0.94 0.98 0.46 0.98 0.99 0.50 0.96 0.73 1.00 1.00 0.99 0.09 0.39
Total Liabilities 0.15 0.87 0.84 0.53 0.94 0.98 0.46 0.98 0.99 0.50 0.96 0.73 1.00 1.00 0.98 0.09 0.40
Total Stockholders Equity 0.10 0.89 0.86 0.50 0.91 0.97 0.44 0.96 0.98 0.48 0.96 0.73 0.99 0.98 1.00 0.06 0.36
Interest_Rate -0.53 -0.01 0.03 0.11 0.21 0.11 0.10 0.14 0.11 0.08 0.10 0.14 0.09 0.09 0.06 1.00 0.09
Stock_Price 0.12 0.05 -0.05 0.36 0.48 0.43 0.33 0.46 0.43 0.32 0.29 0.29 0.39 0.40 0.36 0.09 1.00
Year
Cash and Due from Banks
Goodwill
Income Before Tax
Interest Income
Loans
Net Income
Net Interest Income
Net Loans
Non-Interest Income
Retained Earnings
Tax Expense
Total Assets
Total Liabilities
Total Stockholders Equity
Interest_Rate
Stock_Price
1.0
0.8
0.6
0.4
0.2
0.0

문제점 1.

CONSOLIDATED STATEMENTS OF INCOME (unaudited) - USD ($) shares in Thousands, $ in Thousands	3 Month
	Mar. 31, 2018
Interest income:	
Loans, including fees	$ 285,340
Investment securities available for sale	23,928
Trading account assets	54
Mortgage loans held for sale	379
Federal Reserve Bank balances	1,750
Other earning assets	1,683
Total interest income	313,134

Income Statement - USD ($) $ in Thousands	3 Month
	Mar. 31, 2018
INTEREST INCOME	
Loans	$ 45,165
Investment securities and other	1,717
TOTAL INTEREST INCOME	46,882

CONSOLIDATED STATEMENT OF INCOME (UNAUDITED) - USD ($) shares in Millions, $ in Millions	3 Months Ended	
	Mar. 31, 2018	Mar. 31, 2017
Revenues		
Interest revenue	$ 16,332	$ 14,521
Interest expense	5,160	3,566
Net interest revenue	11,172	10,955

Condensed Consolidated Statement of Comprehensive Income Condensed Consolidated Statement of Comprehensive Income - USD ($) $ in Millions	3 Months Ended	
	Mar. 31, 2018	Mar. 31, 2017
Financing Revenue and Other Income [Line Items]		
Interest and Fee Income, Loans and Leases	$ 1,543	$ 1,368
Interest and Dividend Income, Securities, Operating, Available-for-sale	176	134
Interest Income, Deposits with Financial Institutions	15	5
Operating Leases, Income Statement, Lease Revenue	382	543
Total financing revenue and other interest income	2,116	2,050

문제점 1. 같은 개념, 표기는 천차만별

Income statement = Income statements = Statements of income
= Condensed Statements of income = Statement of income
=**손익계산서**

Income before tax = income before provisions for income taxes = income before tax expense
= Income before income taxes = **세전수익**

Non-interest income = noninterest income = non interest income = **비이자수익**

Total shareholders' equity = total stockholders' equity= **주주자본**

……

거의 모든 대상 항목들에서 최소 3가지의 표기 유형이 나타남

문제점 2. 데이터의 위치도 분류도 천차만별

Loans, held for sale at fair value	15,937	31,055
Loans, net of allowance for loan losses of $40,810 at March 31, 2018 and $40,599 at December 31, 2017	4,805,758	4,776,318

계정명이 들어 있는 컬럼에 계정값이 들어 있는 경우가 다수 존재

CONSOLIDATED STATE INCOME (UNAUDITED) shares in Millions, $ in	
Revenues	
Interest revenue	
Interest expense	
Net interest revenue	
Commissions and fees	
Principal transactions	
Administration and other fidu	
Realized gains on sales of inv net	

Condensed Consolidated Statement of Comprehensive Income Condensed Consolidated Statement of Comprehensive Income - USD ($) $ in Millions	3 Months Ended	
	Mar. 31, 2018	Mar. 31, 2017
Financing Revenue and Other Income [Line Items]		
Interest and Fee Income, Loans and Leases	$ 1,543	$ 1,368
Interest and Dividend Income, Securities, Operating, Available-for-sale	176	134
Interest Income, Deposits with Financial Institutions	15	5
Operating Leases, Income Statement, Lease Revenue	382	543
Total financing revenue and other interest income	2,116	2,050

어떤 은행에서는 Interest revenue 계정 하나로,

어떤 은행에서는 이자수익 원천마다 구분해서 이자수익을 인식

요약

1. 전처리가 지나치게 난해한 데이터셋을 선정함.
2. 매핑만 완료되면 자동화가 가능했으나, 표기와 형식이 상이한 동시에 방대한 Raw Data
3. 따라서 수작업으로 매핑 - 매우 비효율적
4. 팀원 간 도메인 지식의 교집합이 매우 적고, 사전조사가 충분치 않았음.
5. 그러나 모델의 결과물에서 유의미한 상관관계를 도출할 수 있었다: 조금 더 시간을 가지고 제대로 작업할 수 있다면 유용한 모델을 만들 수 있었을 것으로 기대됨.

CREDITS

- EDGAR 크롤링 코드 (1. url 동적 수집 2. url에 해당하는 페이지에서 정적 크롤링을 통해 파일 다운받는 코드) 작성
- CNN, DNN 모델 작업

- 은행 재무제표 독해 및 추출할 변수 지정
- 용어 매핑 가이드라인 제시 및 영문 자료 독해
- 보고서 및 PPT 슬라이드 제작

- EDGAR 크롤링 코드 (1. url 동적 수집 2. url에 해당하는 페이지에서 정적 크롤링을 통해 파일 다운받는 코드) 작성
- Linear Regression, Random Forest, XG Boost 모델 작업
- 시각화

- 금리 데이터 수집
- NYSE 상장 기업 중 3Q-1K 형식을 가진 은행 리스트업
- 리스트 상의 은행 재무제표(2018 1Q~2024 1Q) 수집 및 전처리
- 엑셀 데이터 추출 함수 제작