YUJIN KIM

Data Analysis Portfolio

김 유 진 ✉ yjkimda134@gmail.com ☎ (82+)010-5117-9645

Contents

Self Intro

I am **Yujin Kim**, an aspiring data analyst.

I believe that **growth** comes through **diverse experiences and continuous challenges**.
My goal is to become someone who **seeks opportunities for improvement** while maintaining a **broader perspective**.

☑ Data Analysis Experience

During my internship, I assisted with data organization and preprocessing tasks required for customer data analysis. This experience sparked my interest in **data analysis**. Later, I pursued a **graduate program** to deepen my knowledge in data analysis and **developed analytical skills** through various **data analysis projects** during my studies.

☑ Program planning and operation Experience

As a team member at a public institution, where I gained experience in **planning and operating various programs** on diverse topics. I strived to design programs that reflected societal trends and captured participants' interests. Additionally, while planning and operaing programs, I was able to handle related **administrative tasks** such as budgeting and promotion.

Background

Education

2017.03 - 2021. 08

- Hankuk University of Foreign Studies
- Bachelor of Political Science and Diplomacy

2023.03 - 2025. 02

- Sogang University
- Master of Business Analytics

Work Experience

2021.07 - 2021.12

- Dajeon Social Innovation Center(DSIC) (Space Planning Team Associate)

2022.04 - 2022.07

- Global Santa Fe (NM, USA) (Intern)

2022.08 - 2023.01

- The graduate school of public policy at Sejong University (Administrative Officer)

Skills

- OA(EXCEL, WORD, POWERPOINT etc)
- Computer Language: Python, R, MySQL
- Language: Korean(Native), English
- Program: WordPress, Salesforce, Canvas

Certificate

Word Processor Specialist (2021.10.22)

Advanced Data Analytics Semi-Professional(ADsP) (2023.06.16)

Computer Specialist in Spreadsheet & Database Level-1 (2024.08.16)

Data Mining Project

Travel Insurance Subscription Prediction and Analysis

2023.06

Problem Definition

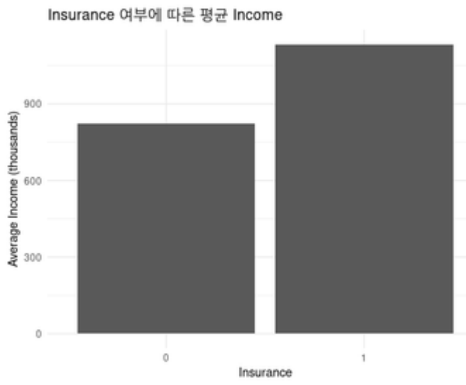
- Post-pandemic boom in the travel industry led to increased competition in travel insurance, with a surge in subscriptions
- **What model can predict travel insurance subscriptions?**
- **What are the characteristics of travelers who subscribe to travel insurance?**

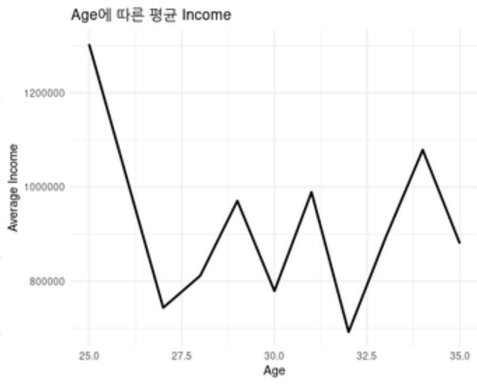
Data Collection and Preprocessing

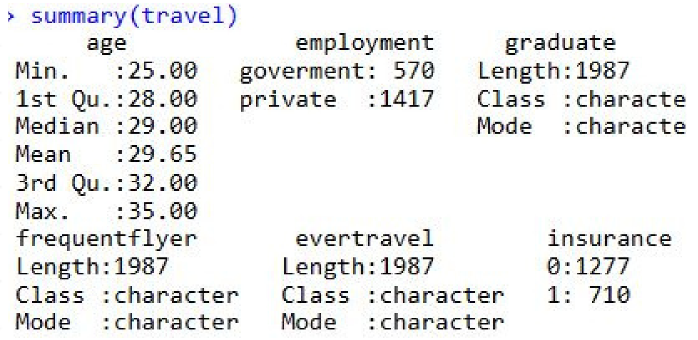
- Data: Travel Insurance Prediction Data
- Source: Kaggle

- Check and handle duplicates and missing values
- Undersampling, split data (train and test) (7:3)
- Balance the y variable ratio in the train set (5:5)
- Cleaning variable types

EDA

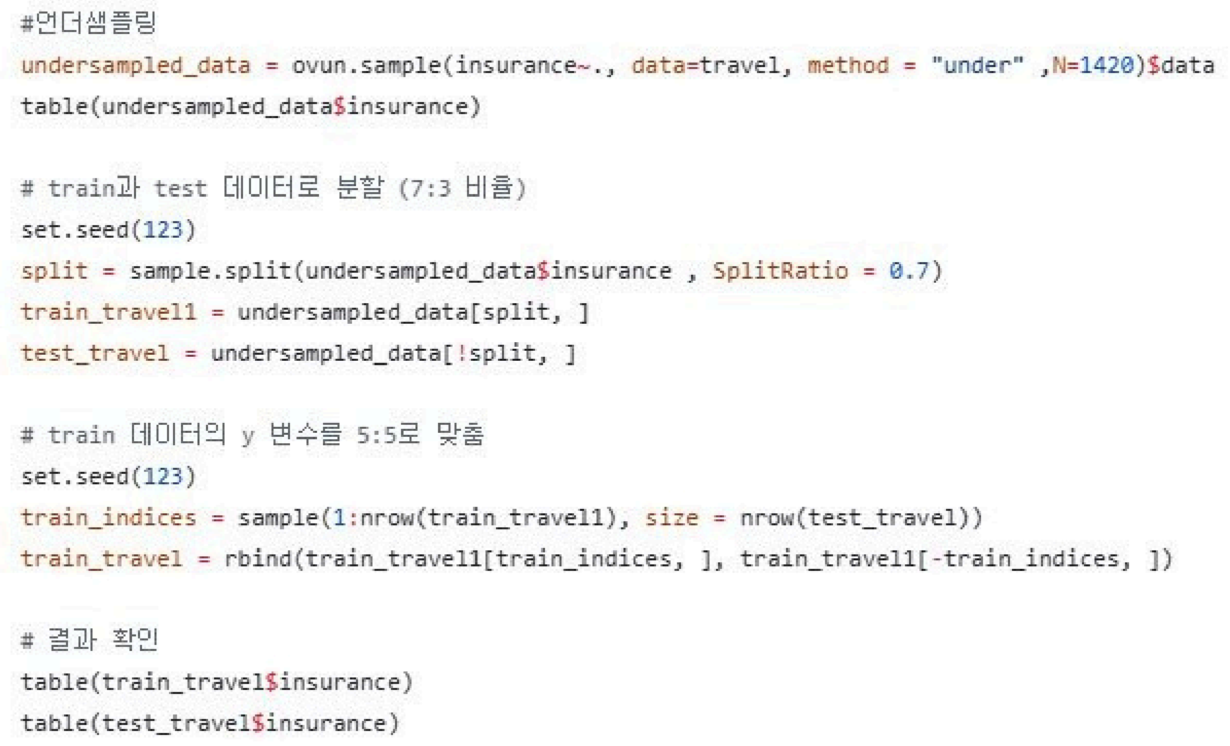






```
> summary(travel)
      age            employment     graduate
 Min.   :25.00   goverment: 570   Length:1987
 1st Qu.:28.00   private  :1417   Class :characte
 Median :29.00                    Mode  :characte
 Mean   :29.65
 3rd Qu.:32.00
 Max.   :35.00
 frequentflyer        evertravel         insurance
 Length:1987         Length:1987         0:1277
 Class :character   Class :character   1: 710
 Mode  :character   Mode  :character
```

Code



```
#언더샘플링
undersampled_data = ovun.sample(insurance~., data=travel, method = "under" ,N=1420)$data
table(undersampled_data$insurance)

# train과 test 데이터로 분할 (7:3 비율)
set.seed(123)
split = sample.split(undersampled_data$insurance , SplitRatio = 0.7)
train_travel1 = undersampled_data[split, ]
test_travel = undersampled_data[!split, ]

# train 데이터의 y 변수를 5:5로 맞춤
set.seed(123)
train_indices = sample(1:nrow(train_travel1), size = nrow(test_travel))
train_travel = rbind(train_travel1[train_indices, ], train_travel1[-train_indices, ])

# 결과 확인
table(train_travel$insurance)
table(test_travel$insurance)
```

Data Mining Project

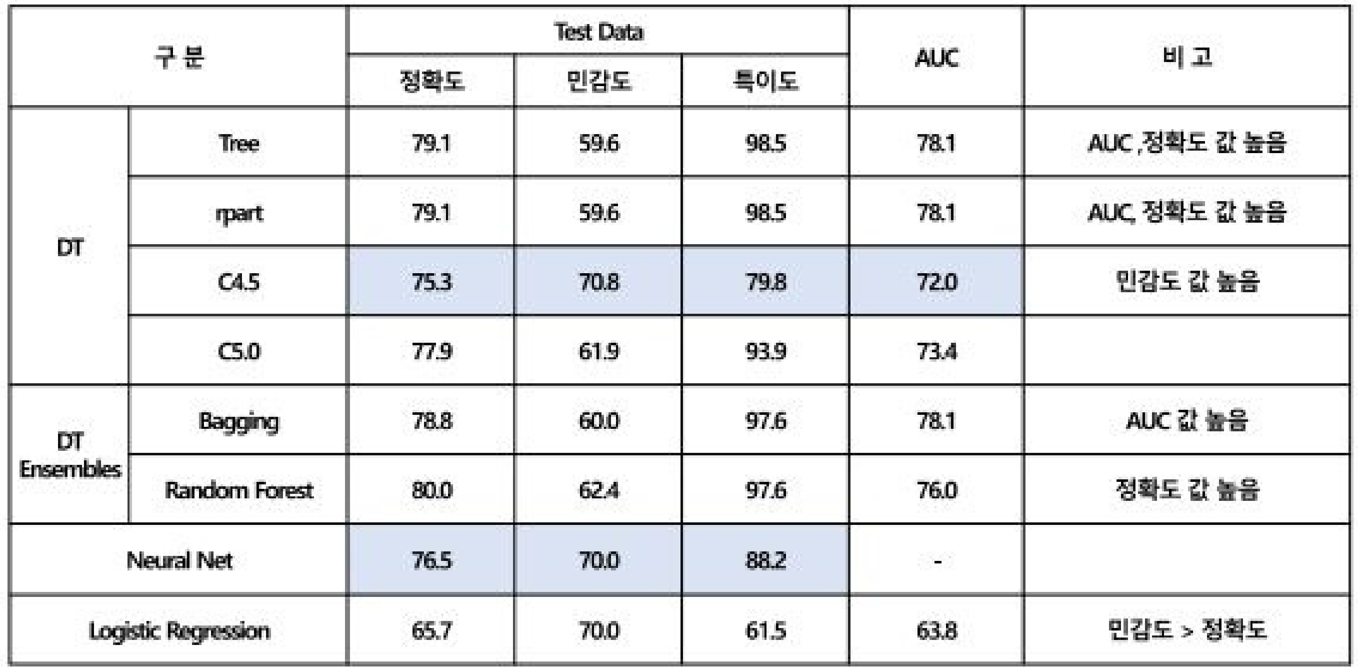
Travel Insurance Subscription Prediction and Analysis

2023.06

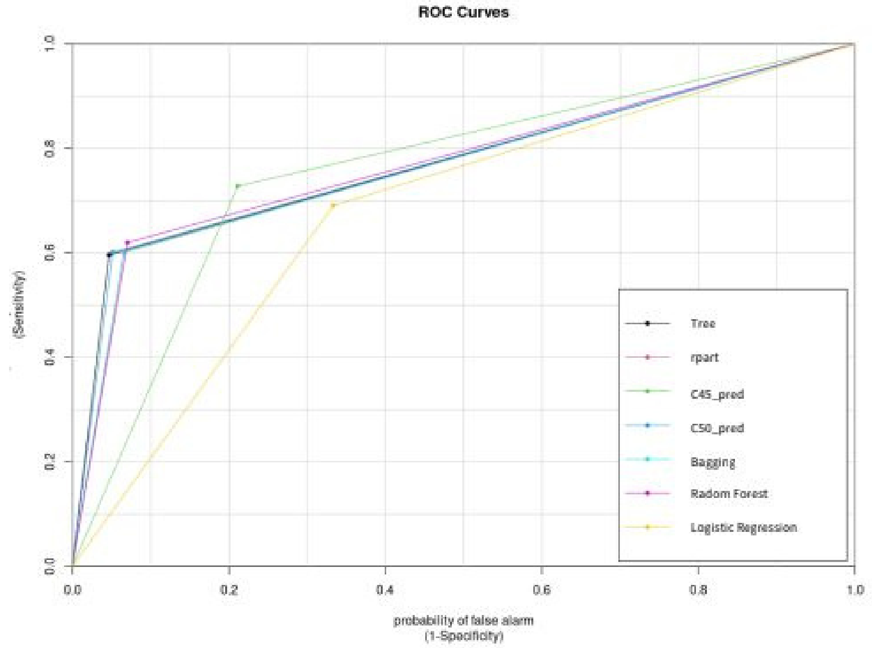
Data Modeling

- Check analysis results for each model
- Models used: decision tree(tree, bagging), rpart, C4.5, C5.0, random forest, neural net, logistic regression
- Model performance comparison method: Compare accuracy, sensitivity, specificity, and AUC values, followed by ROC curve visualization

Modeling result



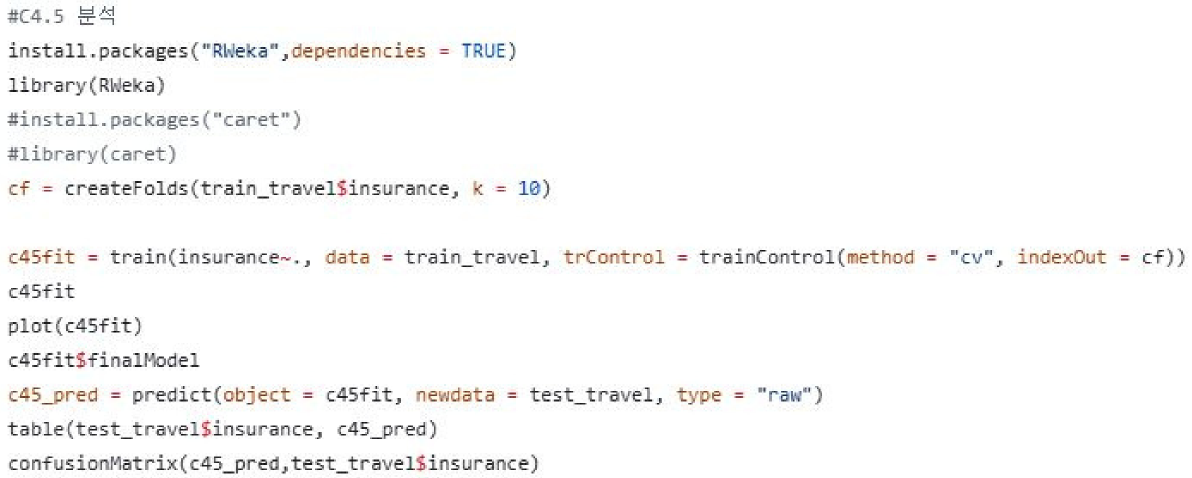
구 분		Test Data			AUC	비 고
		정확도	민감도	특이도		
DT	Tree	79.1	59.6	98.5	78.1	AUC ,정확도 값 높음
	rpart	79.1	59.6	98.5	78.1	AUC, 정확도 값 높음
	C4.5	75.3	70.8	79.8	72.0	민감도 값 높음
	C5.0	77.9	61.9	93.9	73.4	
DT Ensembles	Bagging	78.8	60.0	97.6	78.1	AUC 값 높음
	Random Forest	80.0	62.4	97.6	76.0	정확도 값 높음
Neural Net		76.5	70.0	88.2	-	
Logistic Regression		65.7	70.0	61.5	63.8	민감도 > 정확도



Insight

- C4.5 and Neural Network showed overall superior results
- rpart analysis identified Income as the most important variable
- Age and Family size were found to significantly influence insurance subscription
- **Implications**: Need to promote insurance products to high-income customers and target efficient customer segments
- **Limitations**: limited age group(25-35), insufficient distinction between domestic and international travel insurance.

Code



```
#C4.5 분석
install.packages("RWeka",dependencies = TRUE)
library(RWeka)
#install.packages("caret")
#library(caret)
cf = createFolds(train_travel$insurance, k = 10)

c45fit = train(insurance~., data = train_travel, trControl = trainControl(method = "cv", indexOut = cf))
c45fit
plot(c45fit)
c45fit$finalModel
c45_pred = predict(object = c45fit, newdata = test_travel, type = "raw")
table(test_travel$insurance, c45_pred)
confusionMatrix(c45_pred,test_travel$insurance)
```

TimeSeries Project

Online Grocery Consumption

2023.05 - 2023.06

Problem Definition

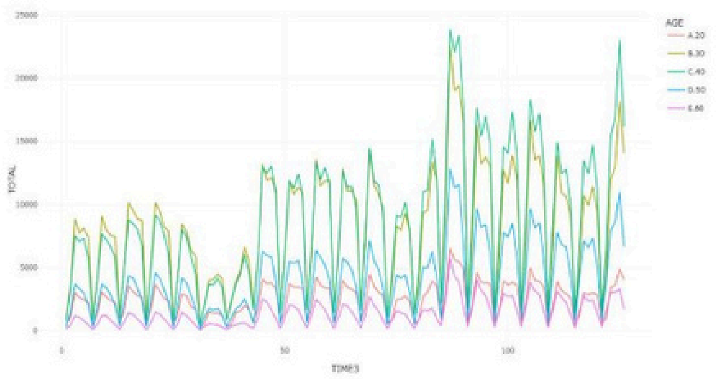
- Consumer purchasing behavior has changed after the COVID-19 pandemic (domestic online shopping transactions increased)
- **How has consumer purchasing behavior changed due to the COVID-19 pandemic?**
- **What are the characteristics of consumers' online purchasing behavior?**

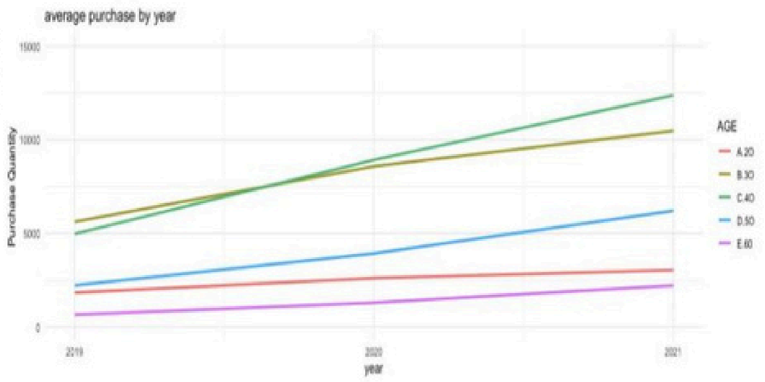
Data Collection and Preprocessing

- Data: 온라인쇼핑 요일/시간대별 이용 특징
- Source: KDX 한국데이터거래소

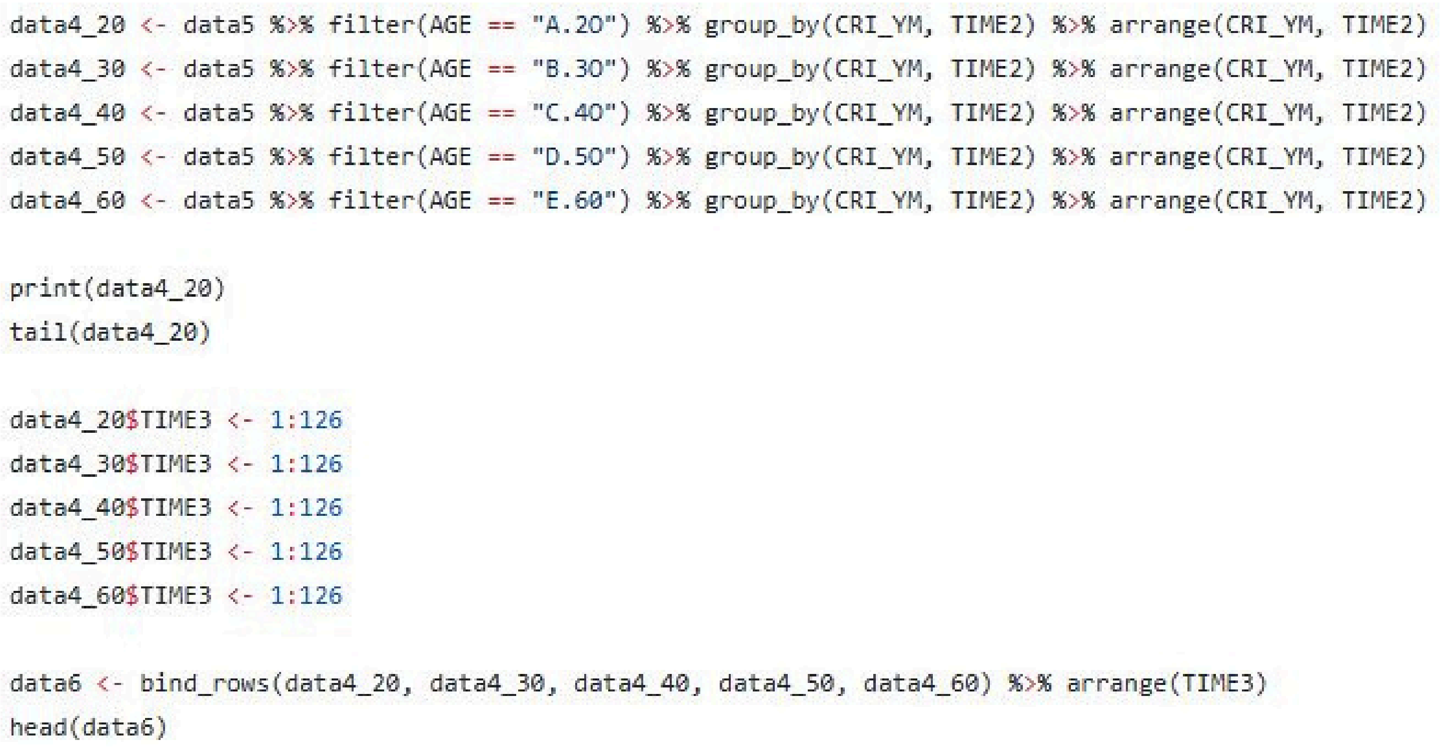
- Extract online grocery shopping data categories
- Check and handle duplicates and missing values
- Verify and adjust the data format (e.g., the time format 'A. 02-06, B. 06-10' was restructured into a new time variable in chronological order)

EDA





Code



```
data4_20 <- data5 %>% filter(AGE == "A.20") %>% group_by(CRI_YM, TIME2) %>% arrange(CRI_YM, TIME2)
data4_30 <- data5 %>% filter(AGE == "B.30") %>% group_by(CRI_YM, TIME2) %>% arrange(CRI_YM, TIME2)
data4_40 <- data5 %>% filter(AGE == "C.40") %>% group_by(CRI_YM, TIME2) %>% arrange(CRI_YM, TIME2)
data4_50 <- data5 %>% filter(AGE == "D.50") %>% group_by(CRI_YM, TIME2) %>% arrange(CRI_YM, TIME2)
data4_60 <- data5 %>% filter(AGE == "E.60") %>% group_by(CRI_YM, TIME2) %>% arrange(CRI_YM, TIME2)

print(data4_20)
tail(data4_20)

data4_20$TIME3 <- 1:126
data4_30$TIME3 <- 1:126
data4_40$TIME3 <- 1:126
data4_50$TIME3 <- 1:126
data4_60$TIME3 <- 1:126

data6 <- bind_rows(data4_20, data4_30, data4_40, data4_50, data4_60) %>% arrange(TIME3)
head(data6)
```

TimeSeries Project

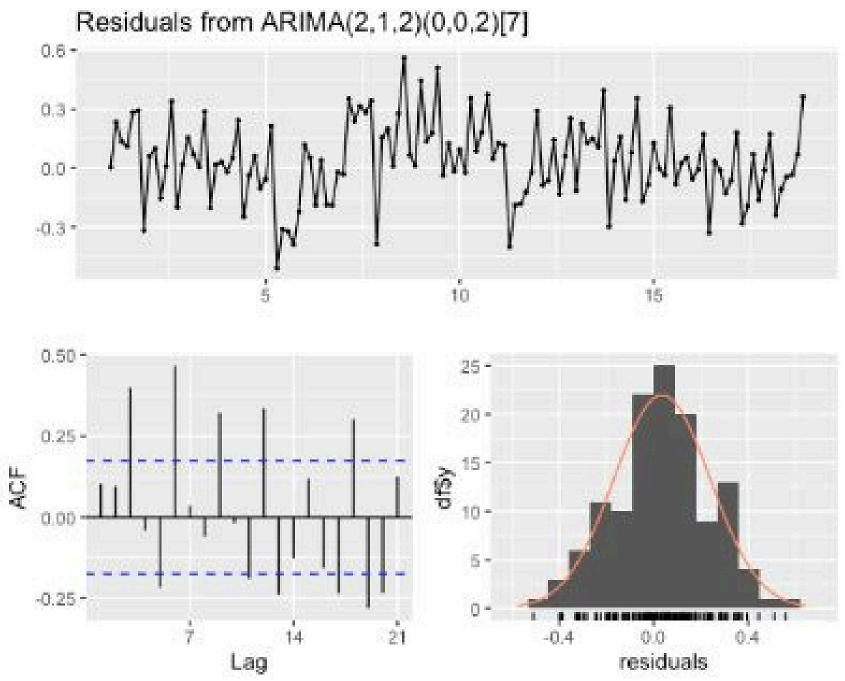
Online Grocery Consumption

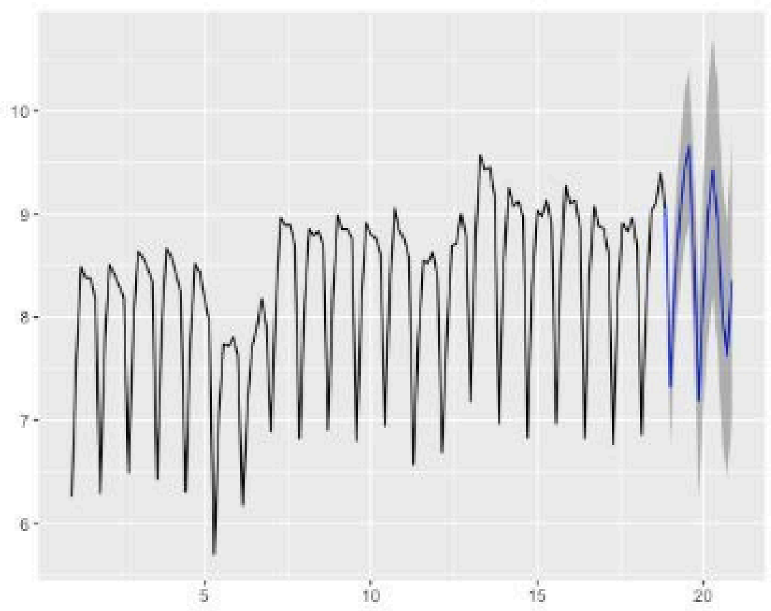
2023.05 - 2023.06

Data Modeling

- SARIMA and SARIMAX time series models were applied, and the best model was selected based on AIC
- To quantitatively assess prediction accuracy, a Paired T-Test was used to check the yearly distribution
- Multiple Regression was applied to examine if the age variable during the COVID-19 situation affected purchase volume

Modeling Result

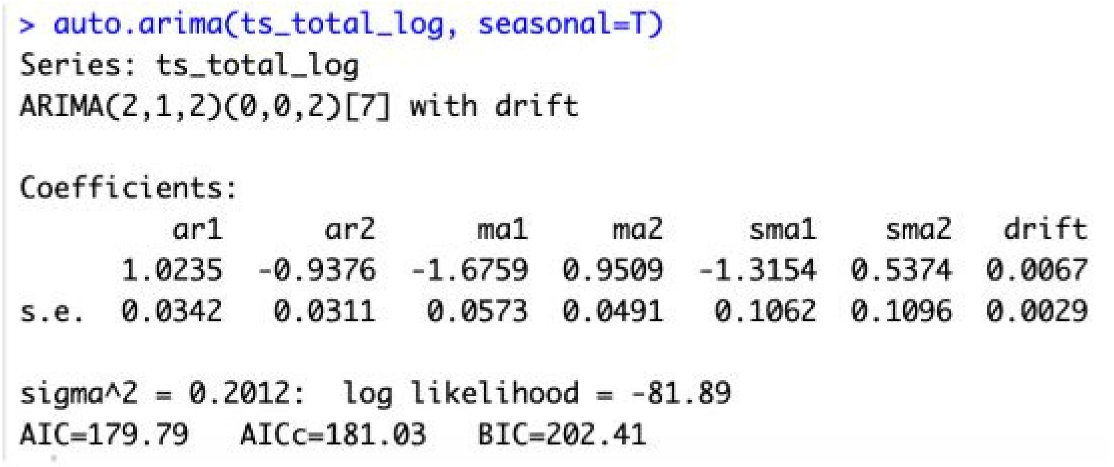




Insight

- SARIMA model showed superior AIC results compared to SARIMAX model
- Paired T-Test revealed a significant difference in the online grocery purchase distribution before and after COVID-19
- Multiple Regression analysis found that year, day of the week, and time had a significant impact on online grocery purchase volume
- **Implications**: Based on the time series analysis results, inventory strategy and marketing targeting strategies need to be established
- **Limitations**: The data is limited to May of 2019, 2020, and 2021

Code



```
> auto.arima(ts_total_log, seasonal=T)
Series: ts_total_log
ARIMA(2,1,2)(0,0,2)[7] with drift

Coefficients:
         ar1      ar2      ma1     ma2     sma1    sma2   drift
      1.0235  -0.9376  -1.6759  0.9509  -1.3154  0.5374  0.0067
s.e.  0.0342   0.0311   0.0573  0.0491   0.1062  0.1096  0.0029

sigma^2 = 0.2012:  log likelihood = -81.89
AIC=179.79   AICc=181.03   BIC=202.41
```

Master's Thesis

Performance Improvement of Stochastic Self-Attention Recommendation Systems

2024.07-2024.12

Purpose

- Recommendation systems are a key technology in modern marketing
- Recent LLMs like GPT and BERT are based on the Transformer architecture, where Attention is the core
- The STOSA algorithm, introduced in 2022, uses self-attention for recommendation systems with probabilistic embedding
- This paper aims to enhance the STOSA algorithm's performance by comparing different methods for measuring the distance between probability distributions in self-attention

Method

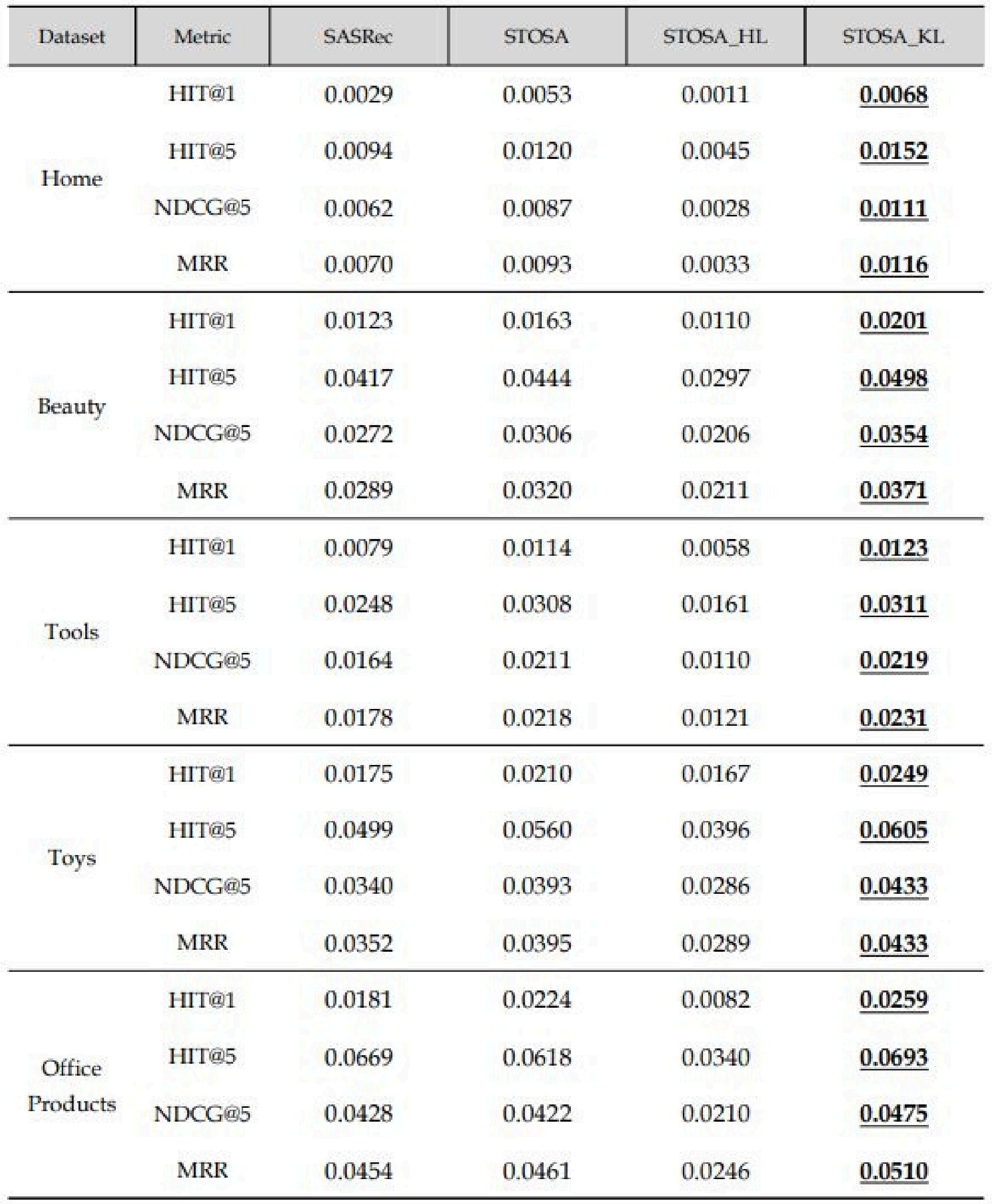
- Data: Amazon Reviews 데이터 (Home, Beauty, Tools, Toys, Office Products)
- Parameters: Reference the parameters used in the original STOSA paper
- Performance comparison method: Hit Ratio(**HIT@1, HIT@5**), Mean Reciprocal Rank(**MRR**), Normalized Discounted Cumulative Gain(**NDCG@5**)
- Probability distribution distance metrics: **Wasserstein distance**(STOSA), **Kullback-Leibler divergence**, **Hellinger distance**

Master’s Thesis

Performance Improvement of Stochastic Self-Attention Recommendation Systems

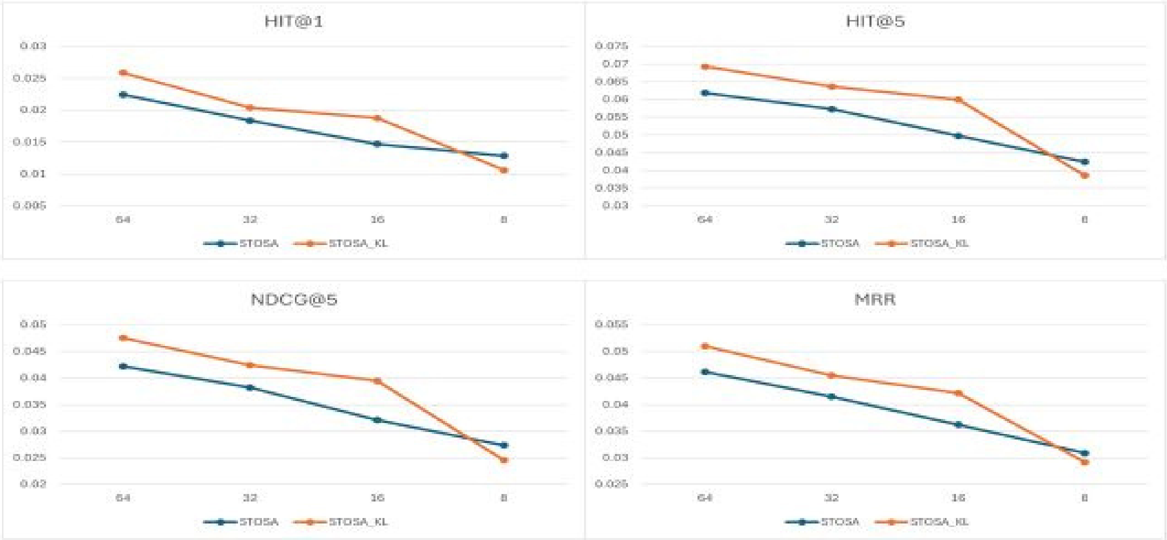
2024.07-2024.12

Results



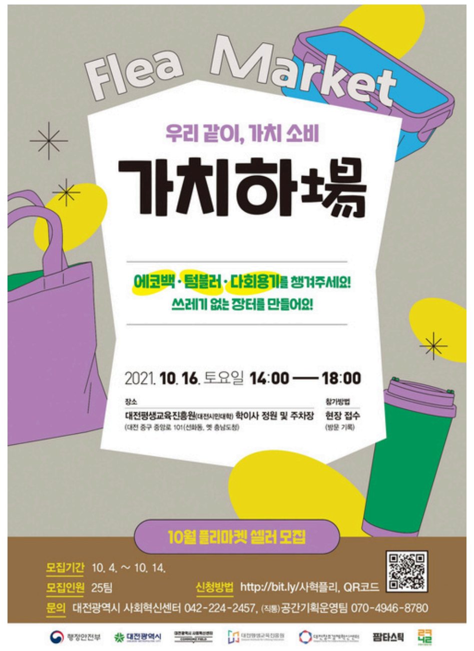
Dataset	Metric	SASRec	STOSA	STOSA_HL	STOSA_KL
Home	HIT@1	0.0029	0.0053	0.0011	**0.0068**
	HIT@5	0.0094	0.0120	0.0045	**0.0152**
	NDCG@5	0.0062	0.0087	0.0028	**0.0111**
	MRR	0.0070	0.0093	0.0033	**0.0116**
Beauty	HIT@1	0.0123	0.0163	0.0110	**0.0201**
	HIT@5	0.0417	0.0444	0.0297	**0.0498**
	NDCG@5	0.0272	0.0306	0.0206	**0.0354**
	MRR	0.0289	0.0320	0.0211	**0.0371**
Tools	HIT@1	0.0079	0.0114	0.0058	**0.0123**
	HIT@5	0.0248	0.0308	0.0161	**0.0311**
	NDCG@5	0.0164	0.0211	0.0110	**0.0219**
	MRR	0.0178	0.0218	0.0121	**0.0231**
Toys	HIT@1	0.0175	0.0210	0.0167	**0.0249**
	HIT@5	0.0499	0.0560	0.0396	**0.0605**
	NDCG@5	0.0340	0.0393	0.0286	**0.0433**
	MRR	0.0352	0.0395	0.0289	**0.0433**
Office Products	HIT@1	0.0181	0.0224	0.0082	**0.0259**
	HIT@5	0.0669	0.0618	0.0340	**0.0693**
	NDCG@5	0.0428	0.0422	0.0210	**0.0475**
	MRR	0.0454	0.0461	0.0246	**0.0510**

*가장 성능이 높게 나온 값에 대해 **굵은 글자체**와 밑줄을 표시하였음.



- The results show that the **STOSA_KL model using Kullback-Leibler outperforms** others
- The original STOSA algorithm also shows high performance (2nd)
- Additionally, performance comparisons based on parameter settings confirm that **STOSA_KL consistently outperforms** other models.
- A comparison of algorithm performance by embedding dimension showed that STOSA performs better in lower dimensions, but as the dimension increases, STOSA_KL shows superior performance.

Extra Projects list

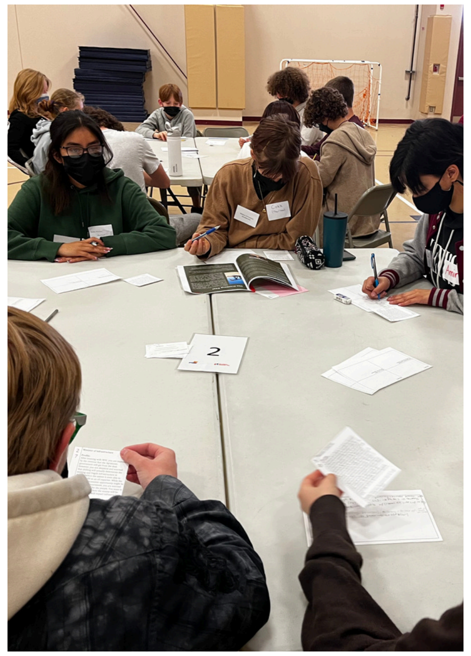

Flea Market
우리 같이, 가치 소비
가치하場
에코백·텀블러·다회용기를 챙겨주세요!
쓰레기 없는 장터를 만들어요!
2021. 10. 16. 토요일 14:00 — 18:00
10월 플리마켓 셀러 모집

Project	탄조중립 플리마켓
Organization	DSIC
Duration	2021.07 - 2021.12
Overview	Planning and Operating NetZero Flea Market and performance
Role	Program Planning and Operation, Administrative Processing

COMMONZ FIELD 대전
2021
사회혁신
국제 컨퍼런스
International Social Innovation Conference 2021
2021.11.05.(Fri) — 06.(Sat)
YOUTUBE LIVE

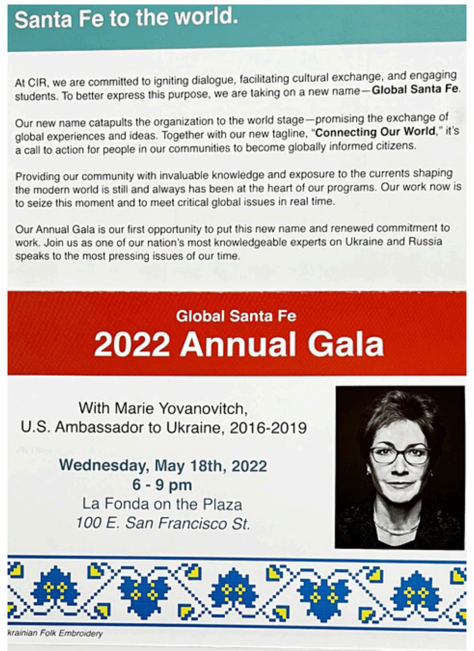
Project	사회혁신 국제 컨퍼런스
Organization	DSIC
Duration	2021.10 - 2021.11
Overview	탄소중립 도시로의 전환을 위한 국제적 사회혁신 아젠다 형성 및 네트워킹
Role	국내외 기관 연락 보조 및 자료 번역, 운영 보조

Project	공공공간 활성화 프로젝트
Organization	DSIC
Duration	2021.10 - 2021.12
Overview	Developing Strategies for Activating Public Spaces online
Role	Program Planning and Operation, Administrative Processing



Project	NextGenSim
Organization	Global Santa Fe (NM, USA)
Duration	2022.04
Overview	Social Education Through Simulated Policy Legislation Based on Key Issues
Role	Assisting in Program Management, Program Documentation

Extra Projects list



Santa Fe to the world.

At CIR, we are committed to igniting dialogue, facilitating cultural exchange, and engaging students. To better express this purpose, we are taking on a new name—**Global Santa Fe**.

Our new name catapults the organization to the world stage—promising the exchange of global experiences and ideas. Together with our new tagline, "**Connecting Our World**," it's a call to action for people in our communities to become globally informed citizens.

Providing our community with invaluable knowledge and exposure to the currents shaping the modern world is still and always has been at the heart of our programs. Our work now is to seize this moment and to meet critical global issues in real time.

Our Annual Gala is our first opportunity to put this new name and renewed commitment to work. Join us as one of our nation's most knowledgeable experts on Ukraine and Russia speaks to the most pressing issues of our time.

Global Santa Fe

2022 Annual Gala

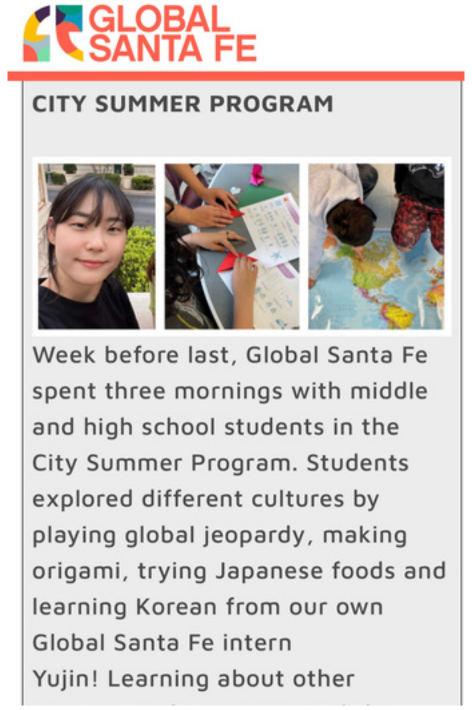
With Marie Yovanovitch,
U.S. Ambassador to Ukraine, 2016-2019

Wednesday, May 18th, 2022
6 - 9 pm
La Fonda on the Plaza
100 E. San Francisco St.

krainian Folk Embroidery

• Project	Annual Gala
• Organization	Global Santa Fe (NM, USA)
• Duration	2022.05
• Overview	Fundraising through Hosting Lectures and Networking Events
• Role	Assisting in Program Planning and Operation, Program Documentation

• Project	Booktalk & Luncheon
• Organization	Global Santa Fe (NM, USA)
• Duration	2022.05
• Overview	Organizing and Hosting Lectures with Various Speakers
• Role	Assisting in Program Promotion and Operation, Program Documentation



GLOBAL SANTA FE

CITY SUMMER PROGRAM

Week before last, Global Santa Fe spent three mornings with middle and high school students in the City Summer Program. Students explored different cultures by playing global jeopardy, making origami, trying Japanese foods and learning Korean from our own Global Santa Fe intern Yujin! Learning about other

• Project	City Summer Program
• Organization	Global Santa Fe (NM, USA)
• Duration	2022.07
• Overview	Planning and Operating Cultural Experience Programs for Local Students
• Role	Introducing Korean Culture and Teaching Korean, Assisting in Program Planning and Management

• Project	Customer Data Analysis
• Organization	Global Santa Fe (NM, USA)
• Duration	2022.07
• Overview	Analyzing Donor Characteristics and Strategic Planning
• Role	Data Extraction from Salesforce, Data Preprocessing

Thank you

김 유 진 ✉ yjkimda134@gmail.com 📱 (82+)010-5117-9645