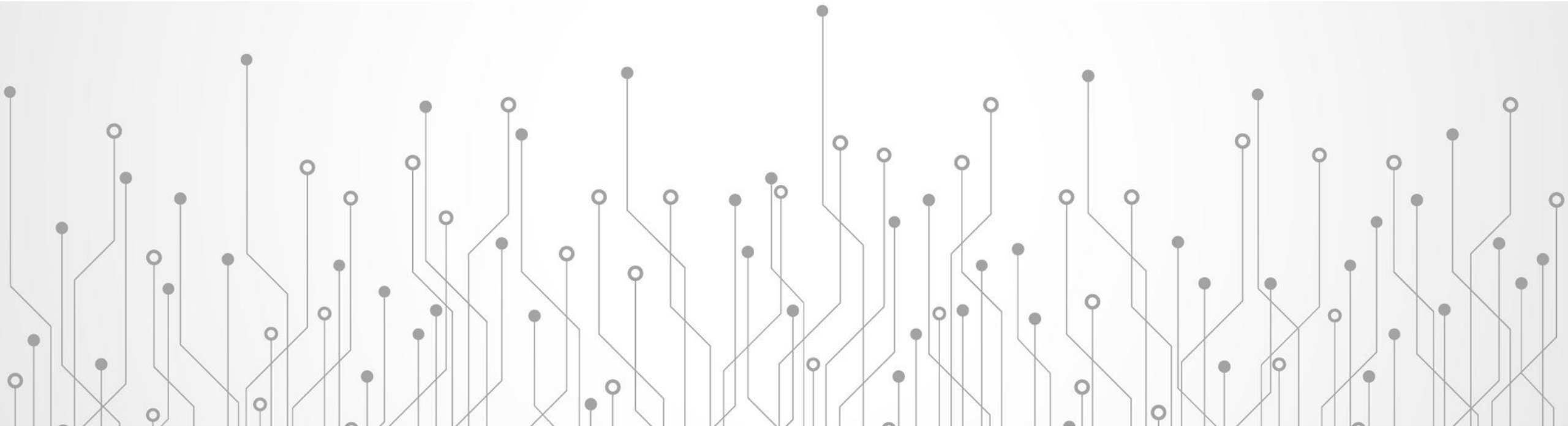
Data Science Project

SHOWCASES



By Kyung Myung Lee

Master of Science in Informatics and Analytics

Master of Science in Computer Science and Engineering

Graduate Certificate in Business Analytics

Contents

Data Analysis

1. Statistical Analysis

- Comparison between cancer centers and genetic counselor utilization (Fisher's exact test and Logistic Regression in R)

  ❖ **Article of Journal of Genetic Counseling (Aug. 2021, co-author) :** https://onlinelibrary.wiley.com/doi/10.1002/jgc4.1495

2. Cluster Analysis

- Segmentation of customers' profile using Clustering technique (Segmentation in SAS Enterprise Miner)

3. Predictive model

- Pneumonia detection using chest x-ray images (Convolutional Neural Networks using Keras (Tensorflow) in Cloud platform (IBM Watson Studio and Amazon AWS)
- Part shortage prediction (Python, Random Forest)

Data Visualization and others

1. Visualization

- Trend and prediction visualization of COVID-19 (Bar/Regression in Python)
- Suffering of deep poverty due to the Great Recession and alleviation by SNAP visualization
- Air pollution and relevant factor trend visualization (R markdown and Shiny, Tableau, and Python)

2. Chatbot

- COVID_19 Screening Bot (Amazon using Node.js, S3 , DynamoDB, Serverless service)

Statistical Analysis

Following NCCN guidelines within one hospital system in the United States: Comparison between cancer centers and genetic counselor utilization

Following NCCN guidelines within one hospital system in the United States: Comparison between cancer centers and genetic counselor utilization

Journal of Genetic Counseling, co-author,
https://onlinelibrary.wiley.com/doi/10.1002/jgc4.1495

Received: 12 February 2021 | Revised: 26 July 2021 | Accepted: 28 July 2021
DOI: 10.1002/jgc4.1495

ORIGINAL ARTICLE

National Society of Genetic Counselors WILEY

Following NCCN guidelines within one hospital system in the United States: Comparison between cancer centers and genetic counselor utilization

Karen Powell[1] | Jonathan Rakestraw[2] | Sat Gupta[3] | Wenhao Shou[3] | Kyung Lee[4] | Ofri Leitner[5]

[1]Genetic Counseling Program, Cone Health Cancer Center, Greensboro, NC, USA
[2]Oncology Informatics System, Cone Health Cancer Center, Greensboro, NC, USA
[3]Department of Mathematics and Statistics, The University of North Carolina, Greensboro, NC, USA
[4]Informatics and Analytics Program, The University of North Carolina, Greensboro, NC, USA
[5]Skypax, Chapel Hill, NC, USA

Correspondence
Karen Powell, Genetic Counseling Program, Cone Health Cancer Center, Greensboro, NC, USA.
Email: Karen.powell@conehealth.com

Abstract

Genetic testing is an instrumental tool used to determine whether an individual has a predisposition to certain cancers. Knowing of a hereditary cancer predisposition may allow a patient and their family to consider high-risk screening or risk-reducing options. Genetic counselors work with physicians to identify patients at increased risk for genetic testing using available guidelines such as those provided by the National Comprehensive Cancer Network (NCCN). Information within one hospital system's cancer registry was used to identify individuals who qualify for genetic testing. This includes patients with a history of cancer of the breast (diagnosis ≤45, triple negative (TN) ≤60, and male), ovaries, colon (diagnosis ≤50), or uterus (diagnosis ≤50). Within this hospital system's registry, there are six cancer centers. Data were collected from cancer centers that utilized genetic counselors (GCs), and cancer centers that did not (non-GC) to determine whether there was a difference in genetic testing rates between GC and non-GC cancer centers. An analysis of 695 patients demonstrated a significantly higher proportion of eligible patients undergoing genetic testing at the GC cancer centers than at the non-GC cancer centers (91.6% versus 68.7%, $p < .001$). Further analysis of specific cancers showed a significantly higher uptake of genetic testing for eligible patients with colon cancer (90.8% versus 50%, $p < .001$), breast cancer ≤45 (99.5% versus 86%, $p < .001$), and ovarian cancer (91.3% versus 62.8%, $p < .001$) at the GC cancer centers than at the non-GC cancer centers. There was no significant difference in the proportion of testing of TN breast cancer ≤60 or uterine cancer ≤50 between cancer centers. These data suggest that having a GC working within a cancer center increases the ability to identify and offer testing to patients who meet NCCN genetic testing criteria based on their cancer type.

KEYWORDS
cancer, genetic counselor, genetic services, genetic testing, NCCN guidelines

Cluster Analysis:

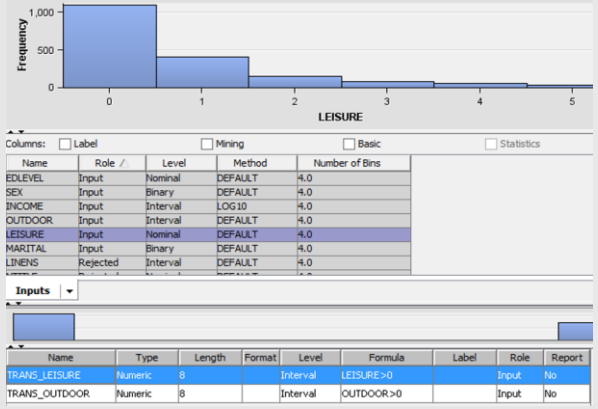
Revealing Customer Profiles of Macy’s Department Store

Cluster analysis

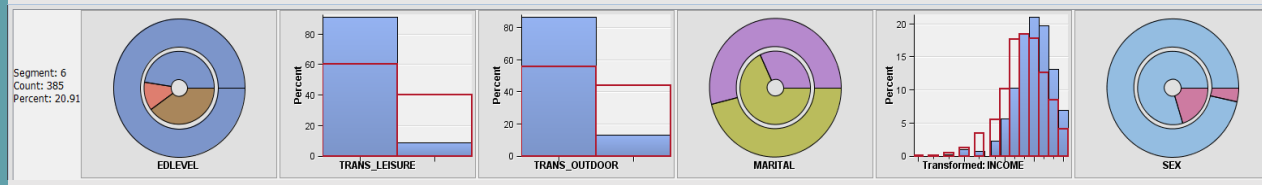
Problem	• Support a marketing campaign of Macy's Department store
Modeling	• Segment customer profiles to target
Dataset	• Customer dataset
Technique	• Clustering technique
Tools and Env.	• SAS Enterprise Miner

Outliers detected

Transform into categorical var.



Segmentation by multiple variables

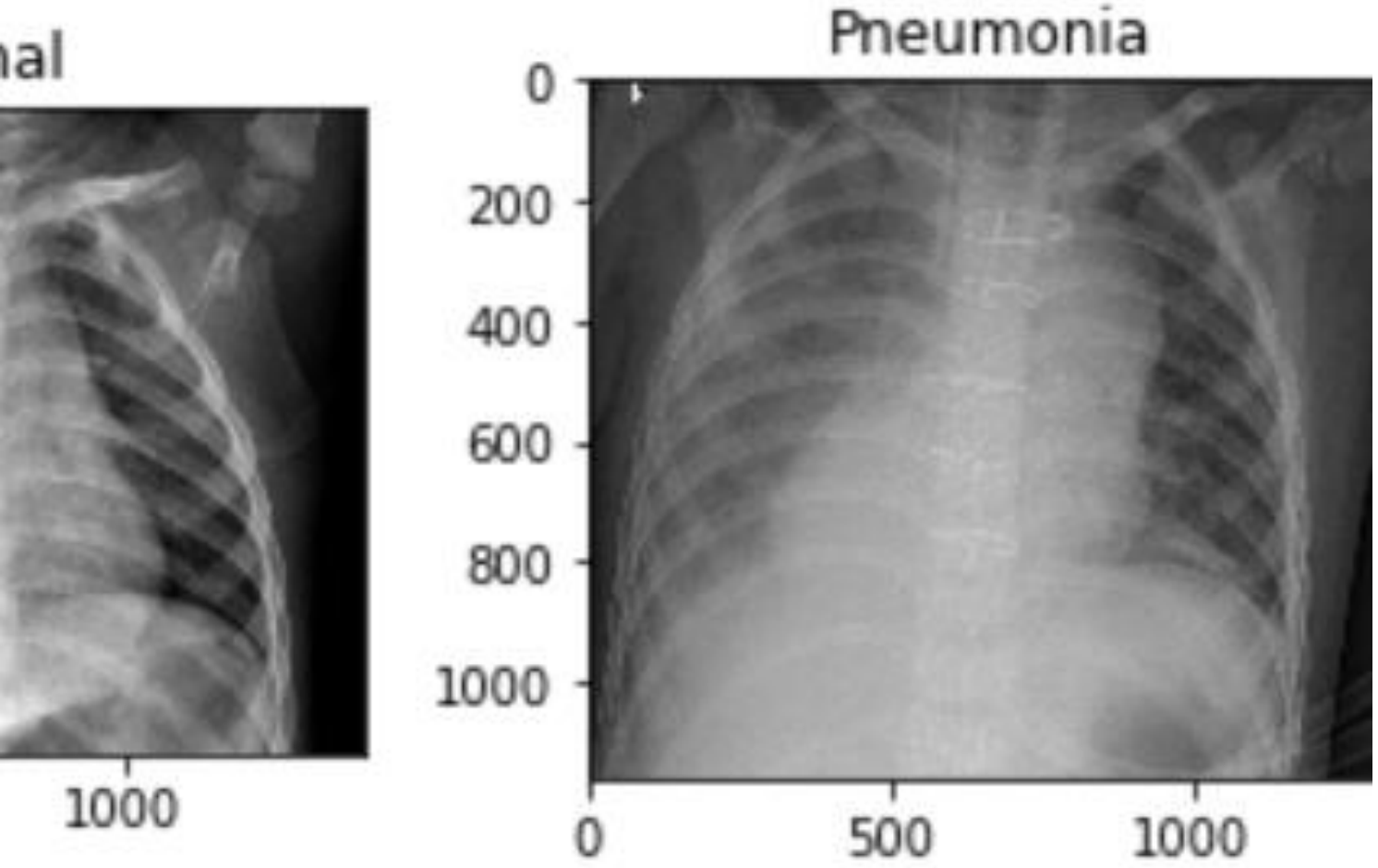


Findings

The example of the segment by multiple variables represents a college-educated customer profile that consists of dominant females and minor males who are almost half and half (approximately 54%:46%) in marital status, bought leisure and outdoor items with a lower percentage, and has a higher average income when compared to the overall distributions.

L

L



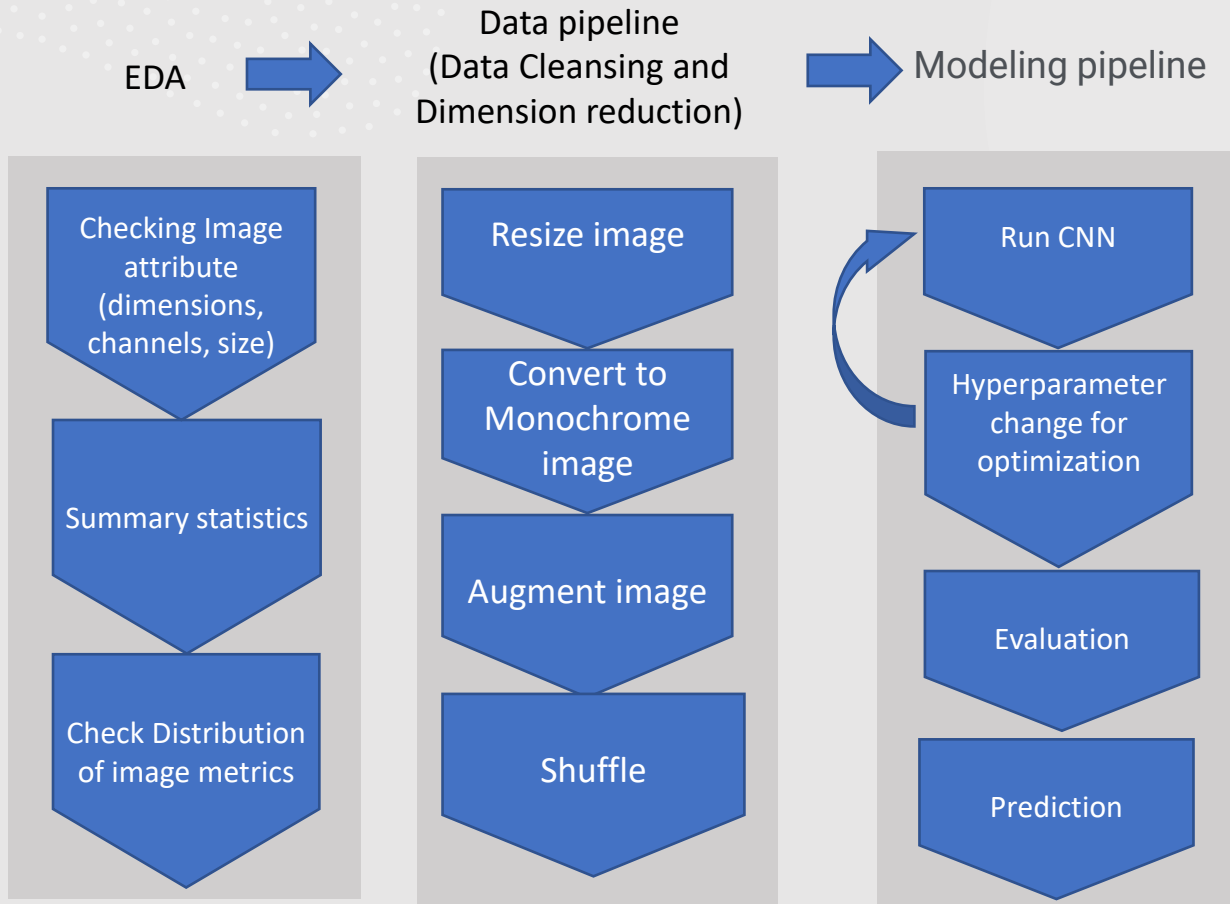
Predictive Model:

Pneumonia Detection using Convolutional Neural Network

*X-ray images from Dataset: Kermany, Daniel; Zhang, Kang; Goldbaum, Michael (2018), “Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification”, Mendeley Data,v2 http://dx.doi.org/10.17632/rscbjbr9sj.2 (License: CC BY 4.0)

Pneumonia Detection

Problem	• Children(age 1 ~ 5) are too young to express symptoms, especially before they get seriously ill.
Model Solution	• Binary image classification ∵ Earlier pneumonia detection matters.
Dataset	• 5,863 X-ray images (JPEG) and 2 classes (Pneumonia/Normal)
Technique	• CNN(Convolutional Neural Network)
Tools and Env.	• Pandas, numpy, keras(tensorflow) libraries on IBM Watson Studio



Pneumonia Detection

Exploratory Analysis /
Data Cleansing/Dimension Reduction

Imbalance labels (1,341 vs 3,875)

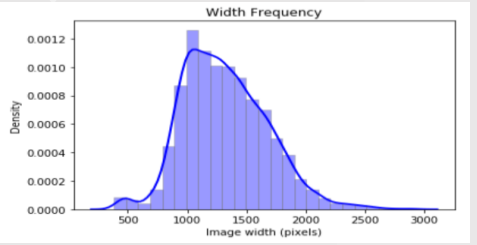
Train Normal dataset:

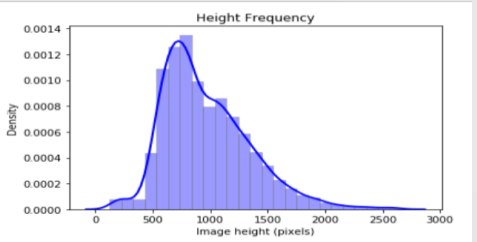
	width	height
count	1341.000000	1341.000000
mean	1667.734526	1381.431022
std	289.210512	326.320734
min	912.000000	672.000000
25%	1466.000000	1152.000000
50%	1640.000000	1328.000000
75%	1824.000000	1542.000000
max	2916.000000	2663.000000

Train Pneumonia dataset:

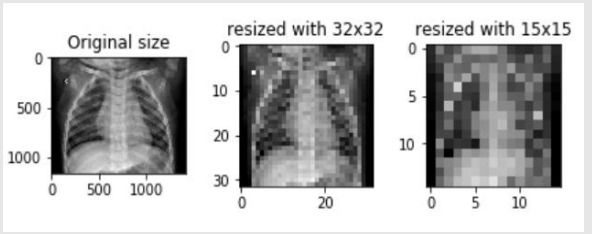
	width	height
count	3875.000000	3875.000000
mean	1200.483613	825.026839
std	291.305676	277.073758
min	384.000000	127.000000
25%	1000.000000	640.000000
50%	1168.000000	776.000000
75%	1368.000000	968.000000
max	2772.000000	2304.000000

Image Width/Height Distribution





Size reduction



Conversion to Gray channel

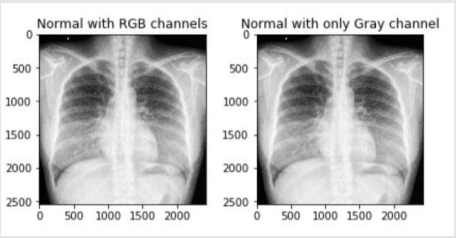
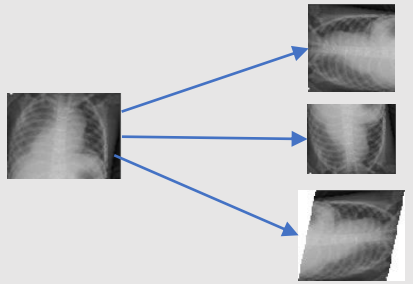


Image augmentation(Oversampling effect for training data)



Rotate

Flip

Skew

Pneumonia Detection

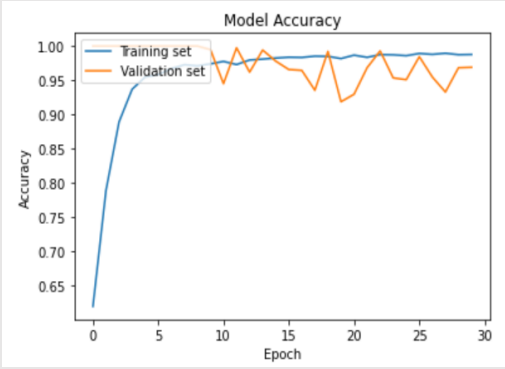
Modeling and Performance Evaluation

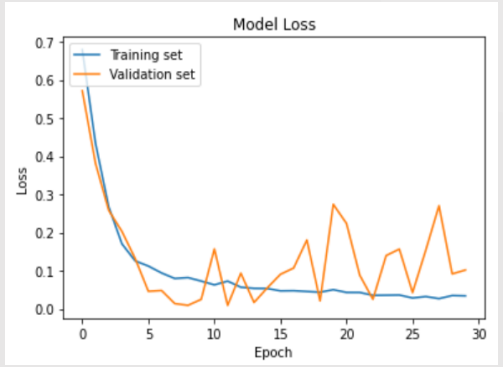
Model	Accuracy (test set)	F1 Score	Training vs Test data ratio (80 : 20)
Model 1 (32 x 32)	83%	88%	7,750 vs 155
Model 2 (60 x 60)	76%	84%	7,750 vs 155

Convolutional Neural Network

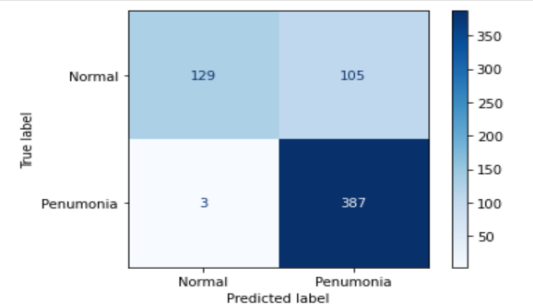
Model implementation in Python:
https://github.com/kmleeDS/portfolio/blob/main/Showcase_attached_to_JobKorea_CNN_32x32_Images.ipynb

Accuracy and Loss



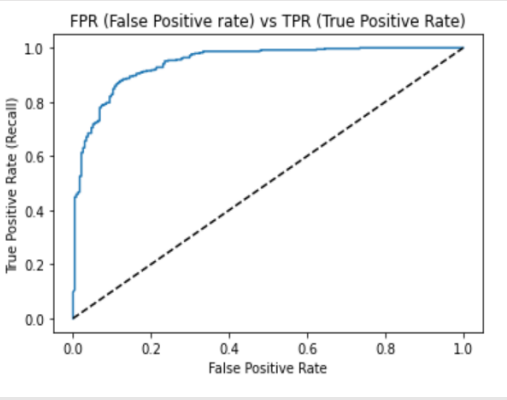


Confusion Matrix (Model 1)




```
accuracy of chest X-ray for pneumonia:0.83
Precison of chest X-ray for pneumonia:0.79
Recall of chest X-ray for pneumonia:0.99
f1 of chest X-ray for pneumonia:0.88
```

ROC with ACU curve (Model 1)



Predictive Model:

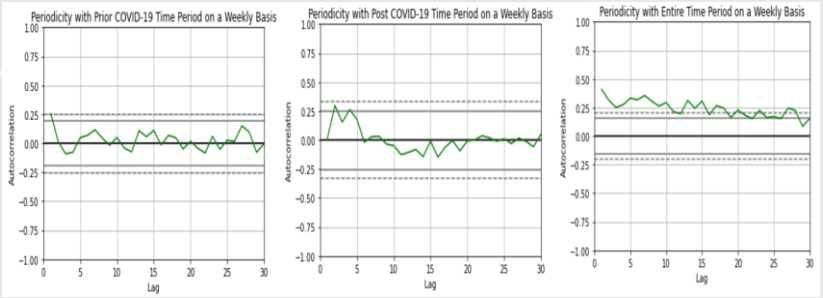
Part Shortage Prediction at truck production

Part Shortage Prediction

Problem	• Missing of even one single part -> Truck production delay -> late delivery
Model Solution /Expectation	• Prediction of missing with factors including time factor ∵ Proactive action-taking matters.
Dataset	• By-product data (+500,000) generated on the shop floor and vendor name data
Technique	• Autocorrelation, RF, and LSTM as a baseline
Tools and Env.	• pandas, numpy, libraries in Python

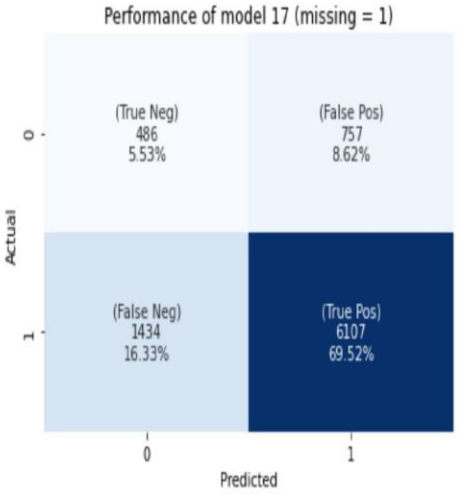
Missing value map

Relationship between time factor and frequency of part-missing



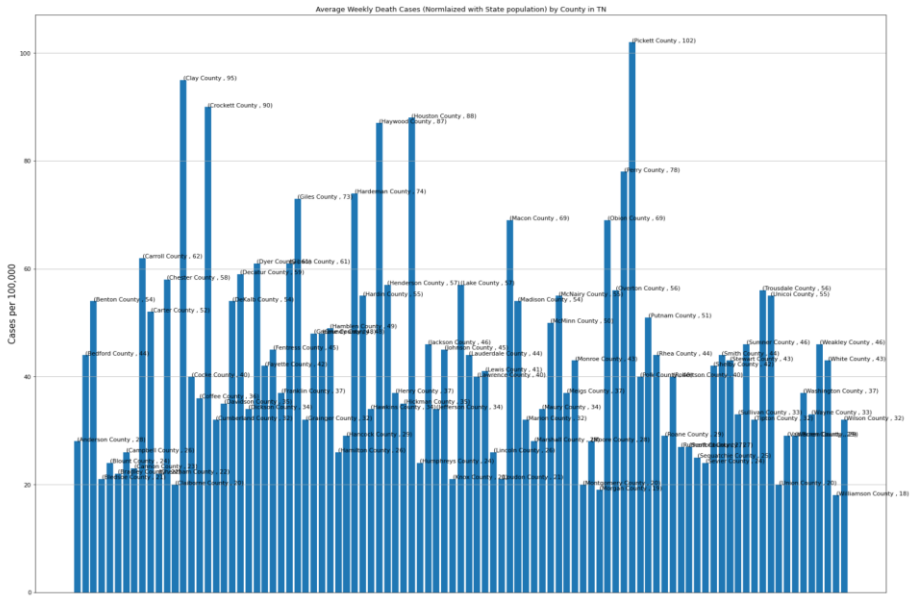
Random Forest Classification

	precision	recall	f1-score	support
0.0	0.25	0.39	0.31	1243
1.0	0.89	0.81	0.85	7541
accuracy			0.75	8784
macro avg	0.57	0.60	0.58	8784
weighted avg	0.80	0.75	0.77	8784

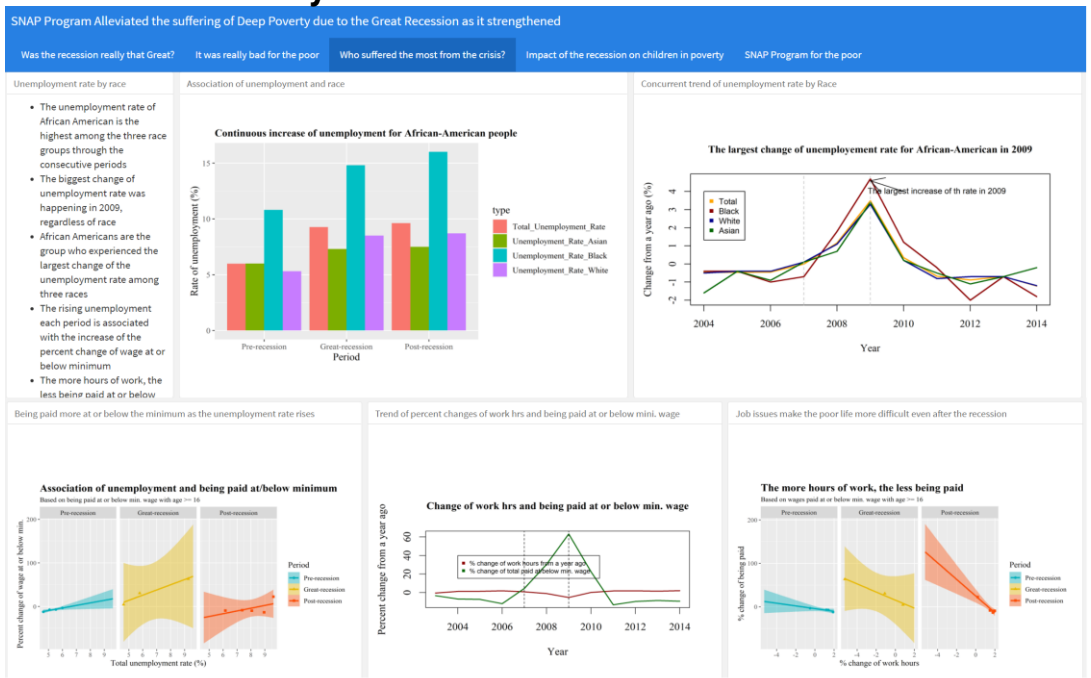


Data visualization showcases

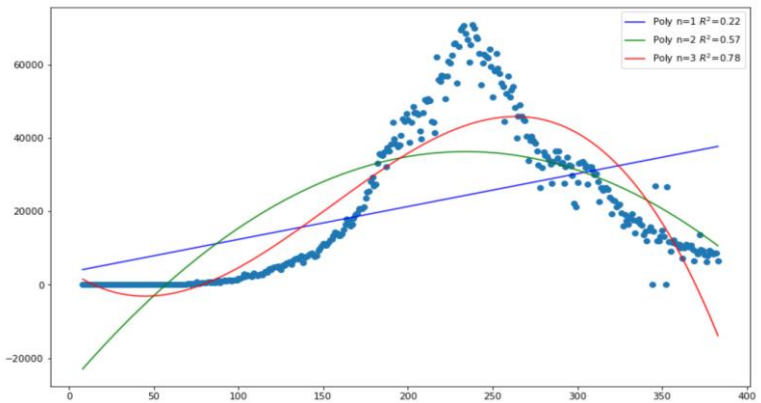
Avg. Death of County Death per 100,000 (TN, US)



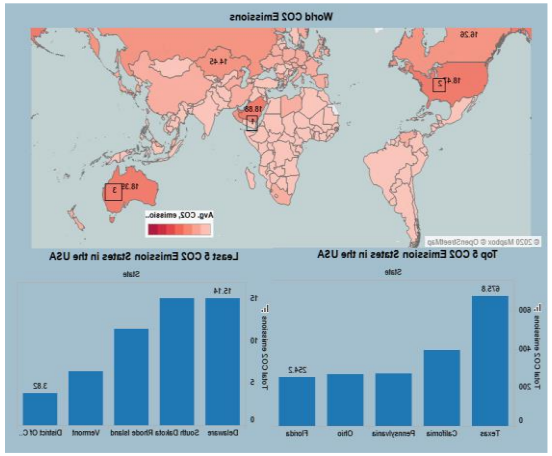
Suffering of deep poverty due to Great Recession and the alleviation by SNAP

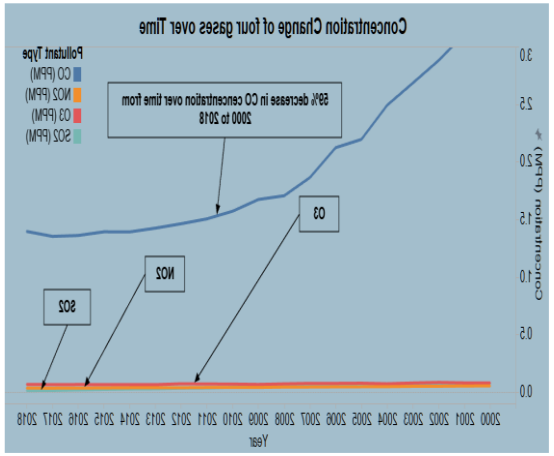


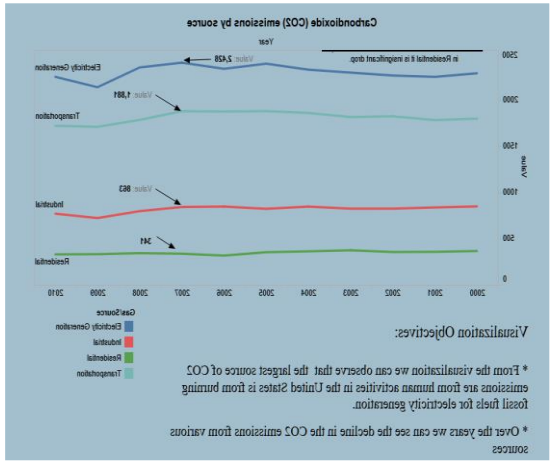
COVID-19 Infection trend per day linear and polynomial regression

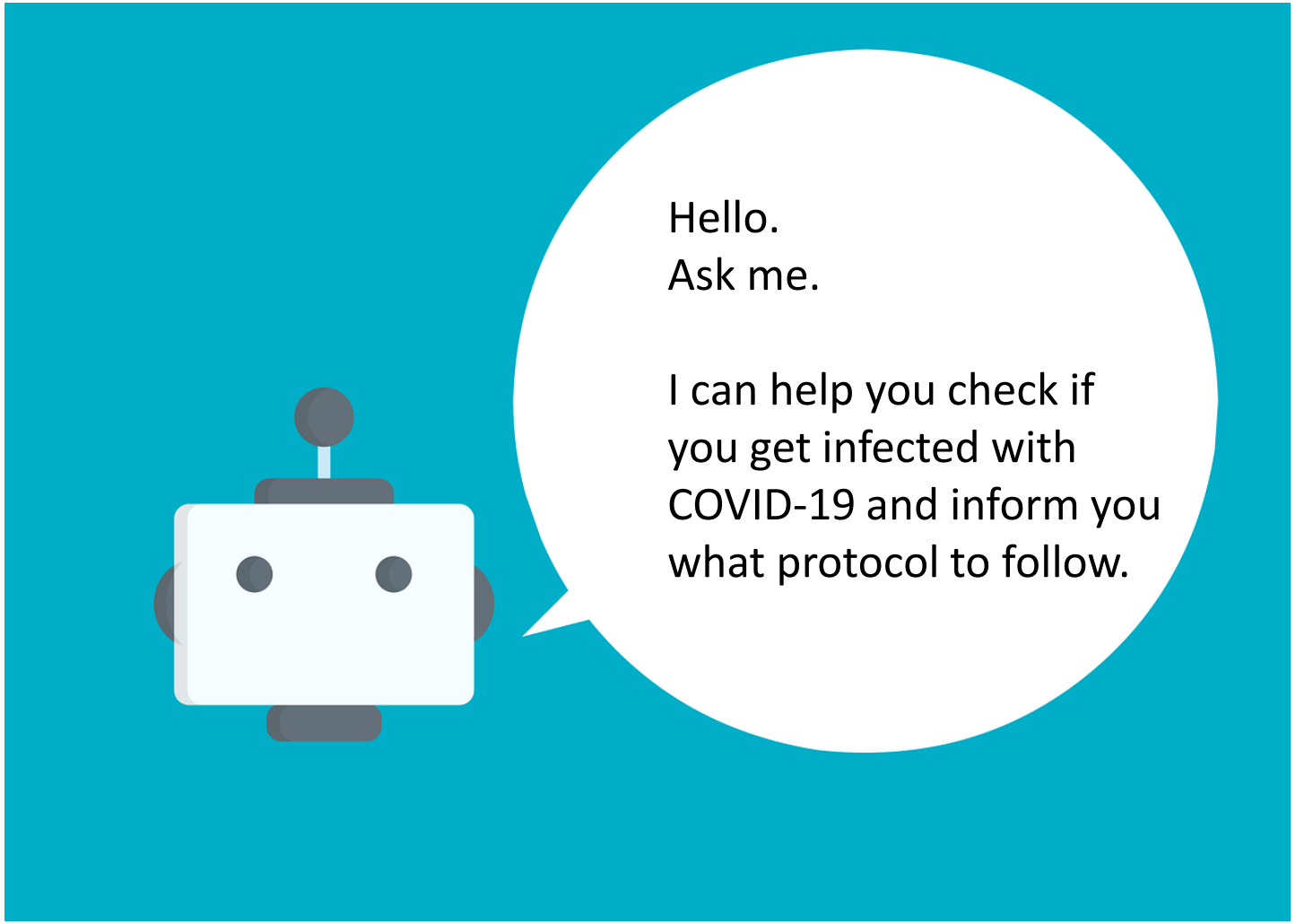


Air Pollution and relevant pollution factors in the U.S.









COVID-19 Symptomatic or Asymptomatic Screening ChatBot using AWS Service)

Chatbot Design

Problem	• Demand of COVID-19 screening service about symptomatic/asymptomatic/other symptoms
Solution	• Chatbot for COVID-19 Screening and providing protocols to follow
Reference	• COVID-19 Screening Protocol • https://github.com/CDCgov/covid19healthbot/blob/master/screening_protocols/covid_19_screening_protocol_cdc_apple.pdf
Development Tools and Env.	• Lex, S3, and DynamoDB on AWS, Node.js

* AWS images are used only for showcasing personal work.

AWS Service components and communication design