#SASGF

VIRTUAL

SAS® GLOBAL FORUM 2021

Pandemic Pandemonium

Team S  Squatch

Oklahoma State University

SASquatch

Hannah Perz

Maryam Taherirani

Sean Everett

Trinh Phan

Outline

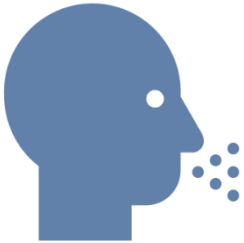
Background

Data Preparation

Analysis

Recommendations

Introduction



January 20, 2020

US COVID-19 Patient Zero
3.5% US Unemployment

April 20, 2020

800,000 COVID-19
Infections in US

May 1, 2020

14.8 % US
Unemployment

Problem Statement

Identify Significant Characteristics of Population Impacted by Job Loss During COVID-19 Pandemic

Explore Correlation Between COVID-19 Infection Rate & Unemployment Rate By US Region

Data Sources

Primary Source

US Census Bureau Household Pulse Survey

Secondary Sources

The New York Times

US COVID Infection Totals

US Pre-COVID Industry Employment Totals

US Population Totals

COVID Policy Indices

Data Collection & Scope

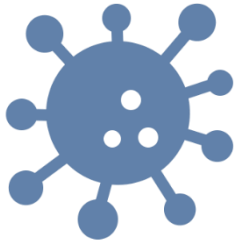
Import Data

Merge 13 Weeks of Pulse Survey

Scope: HPS-Phase1; April 23 – July 21, 2020

Data Transformation

Binary Target Variable: Workloss



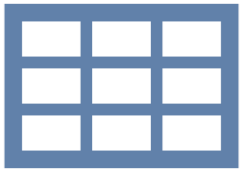
Per Capita COVID-19 Infection Rate Variable

Datasets Joined. 108 Variables, 330,000 Records

Data Cleaning

Filter Records &
Variable Selection

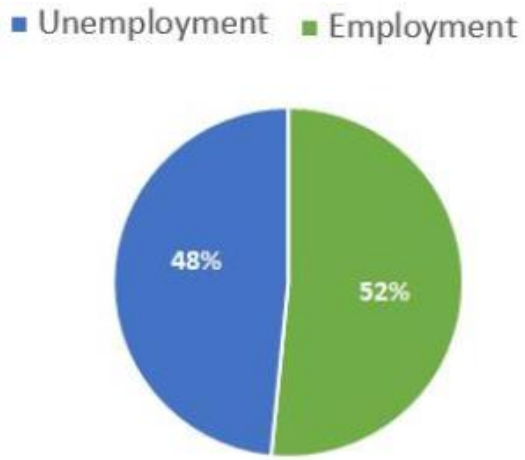
Transform
Missing Values



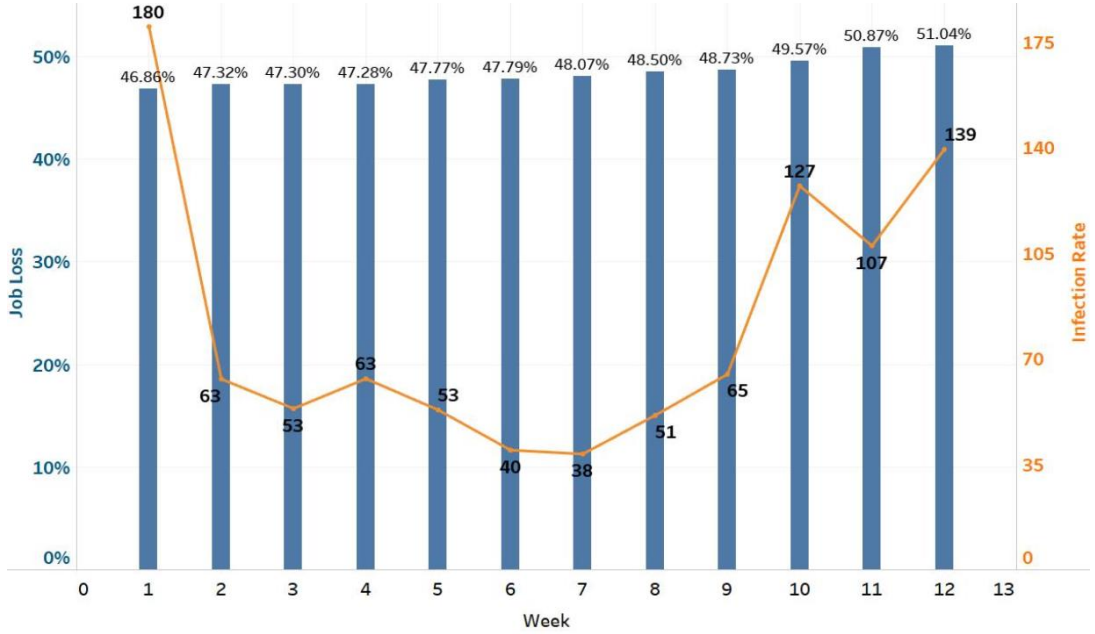
Final Dataset:
273,984 records 49 variables

Descriptive Analysis

Job Loss Distribution

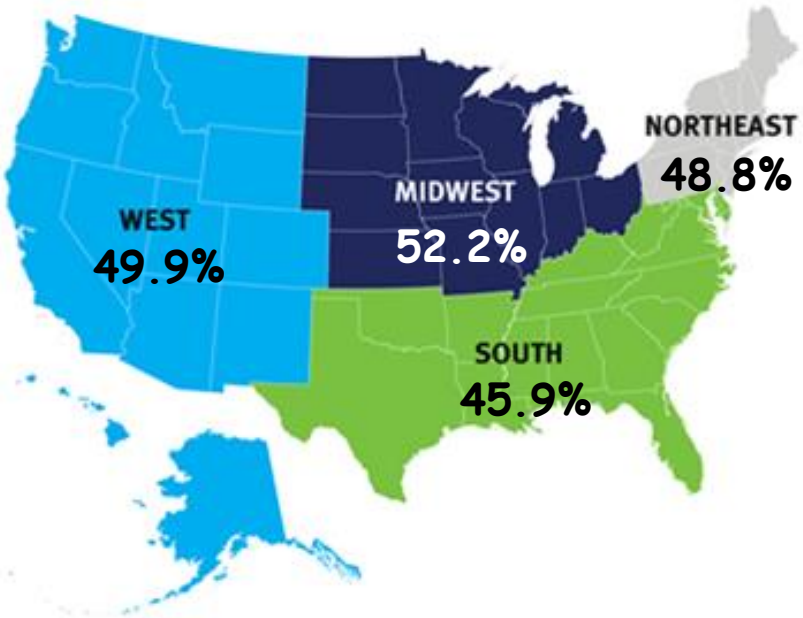


Job Loss vs. Infection Rate

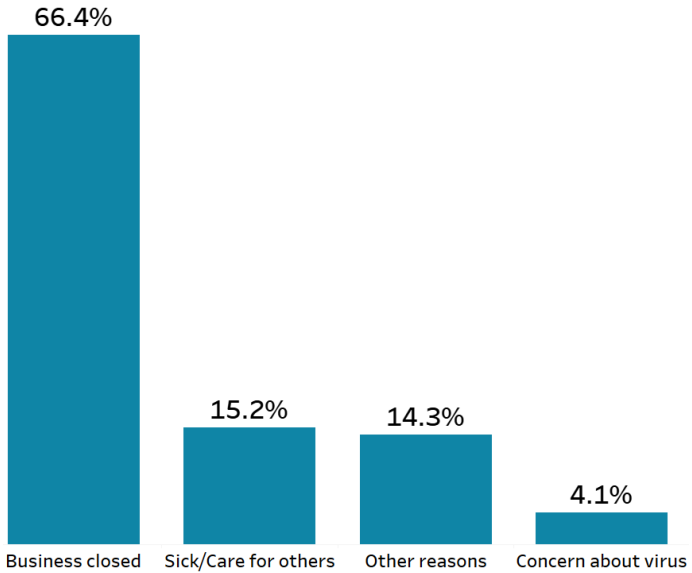


Descriptive Analysis

Job Loss Rate by Region



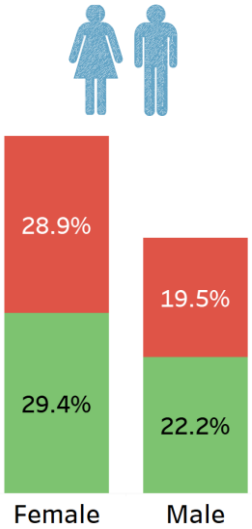
Reason for not working



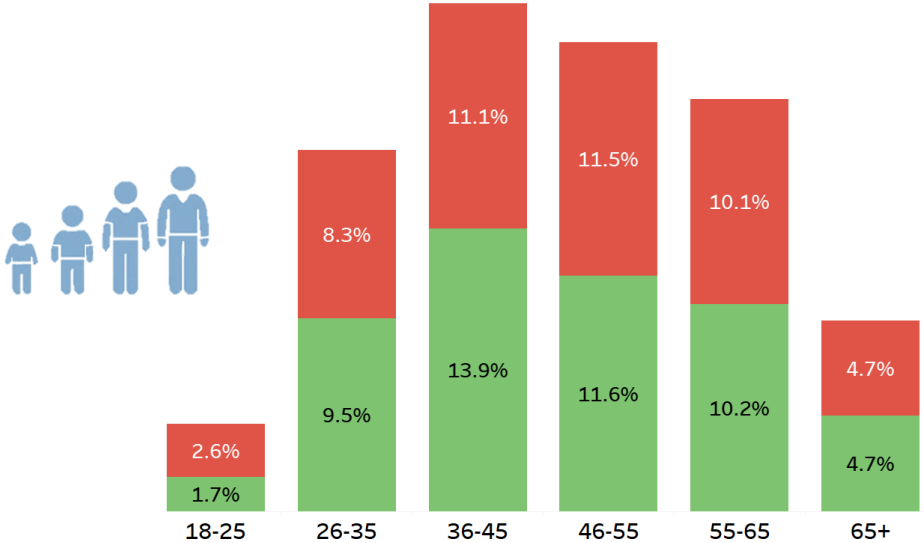
Descriptive Analysis

Percentage of **Job Loss** vs. **Employed** in Different Groups

Gender



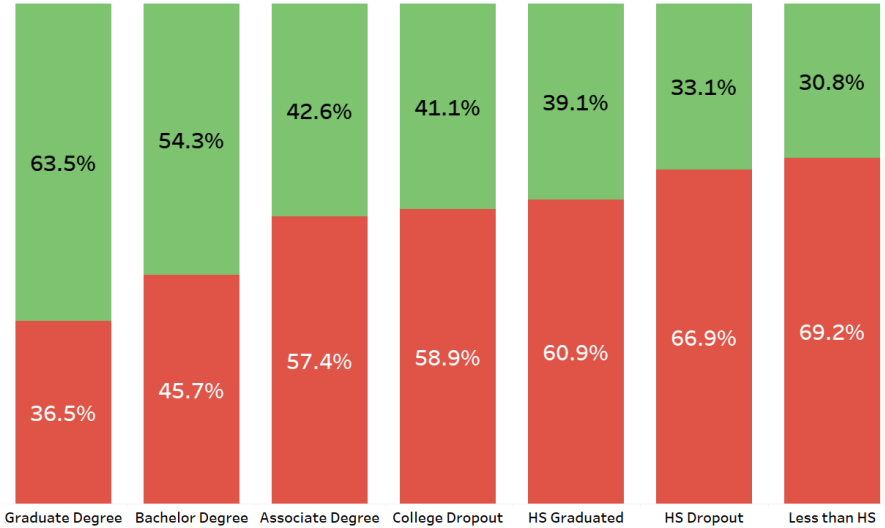
Age



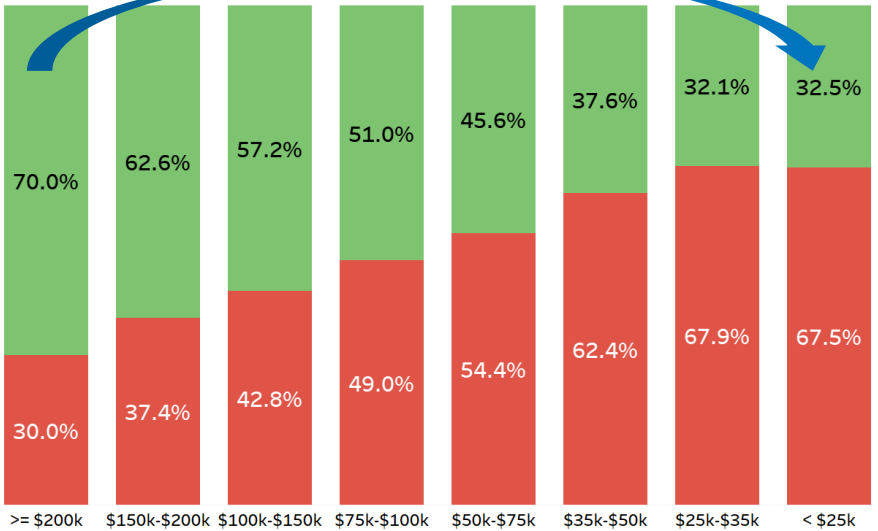
Descriptive Analysis

Percentage of **Job Loss** vs. **Employed** in Different Groups

Education

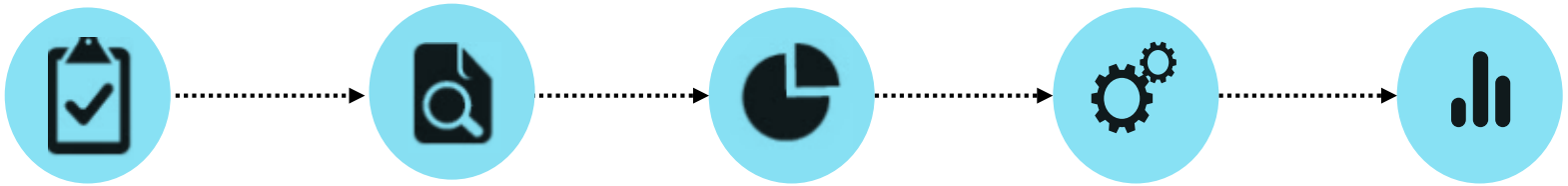


Income



Analysis

Modelling workflow



Analysis

Methods

Dimensionality Reduction

PLS

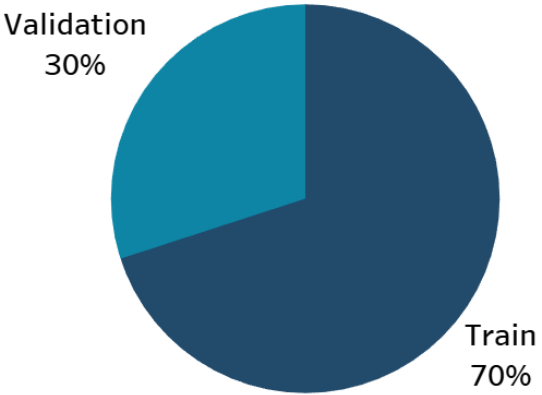
LARS

LASSO

Variable Clustering → Auto-created

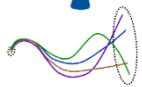
Variable Clustering → 4 Clusters 

Data Partitioning



Models Created

Decision Tree 

Ensemble 

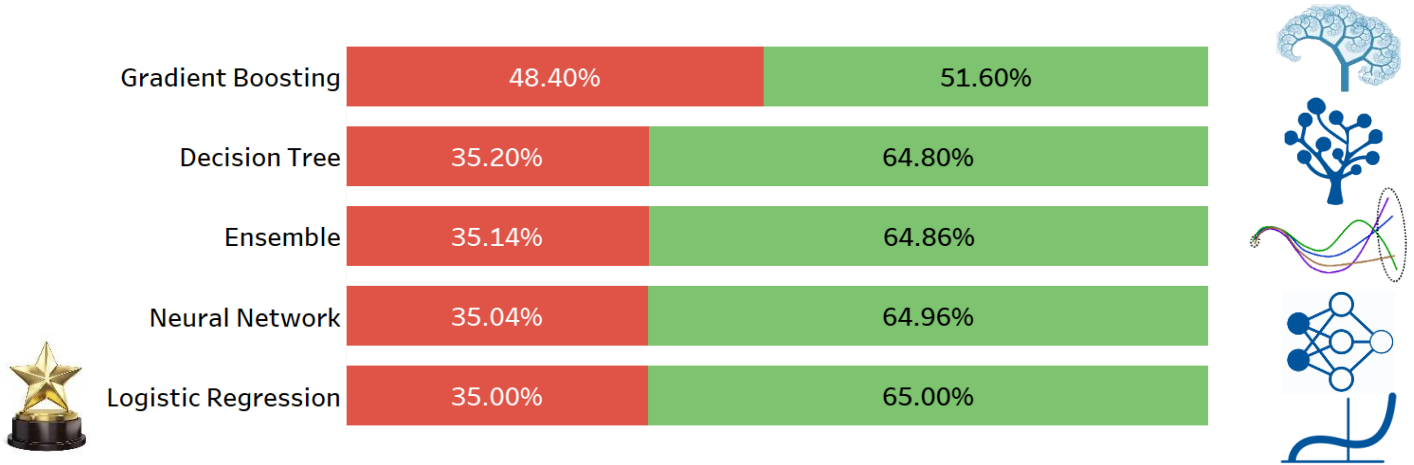
Gradient Boosting 

Logistic Regression 

Neural Network 

Analysis

Modeling Result – Accuracy and Misclassification Rate on Validation Set



Analysis

Model Performance

Predicted Job Loss	Actual Job Loss		
	No (0)	Yes (1)	Total
No (0)	30,410	16,448	46,858
Yes (1)	11,982	23,358	35,340
Total	42,392	39,806	82,198

Accuracy = 65%

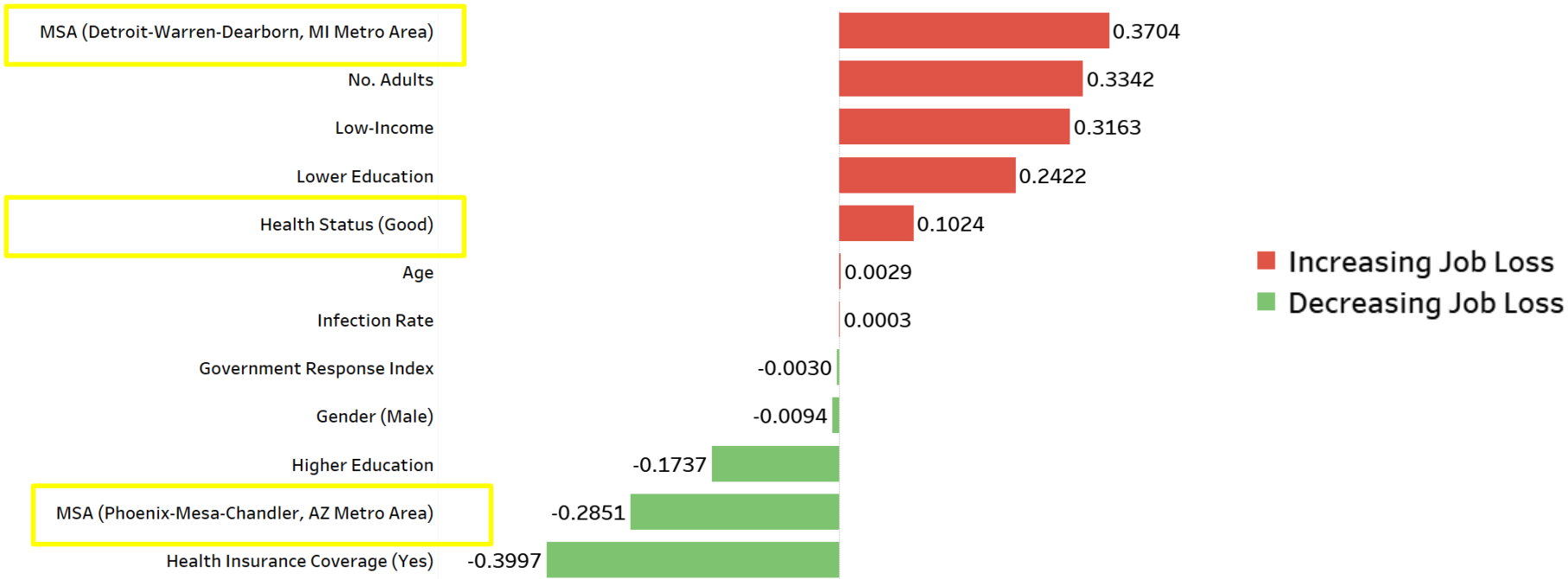
Misclassification Rate = 35%

Sensitivity: 59%

Specificity: 72%

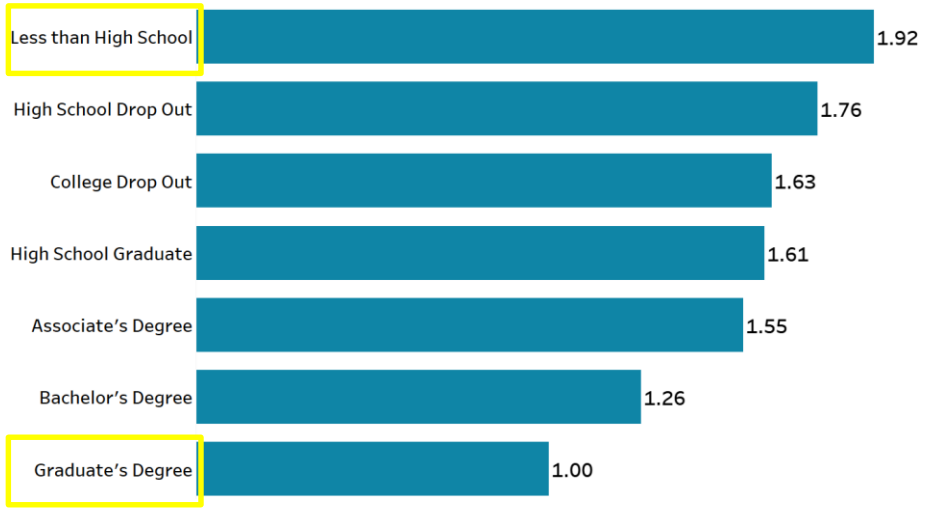
Analysis

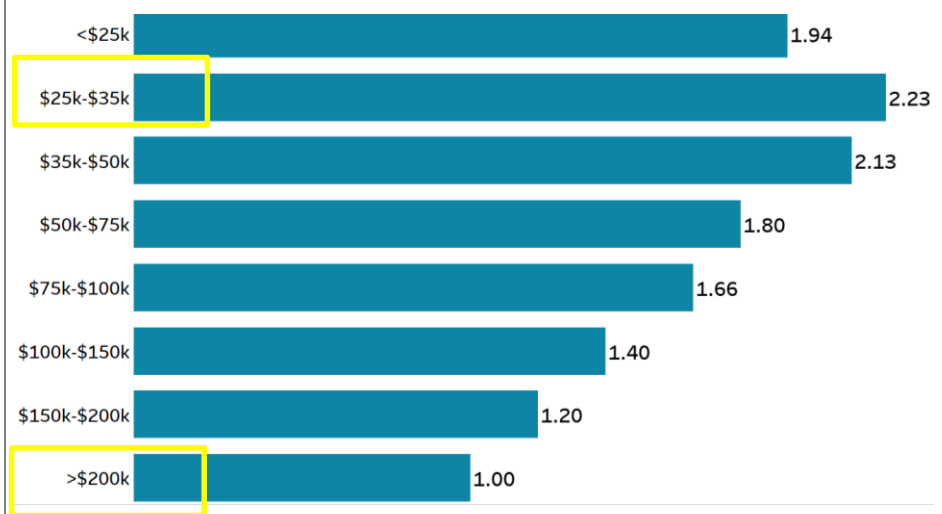
Variable Estimates



Analysis

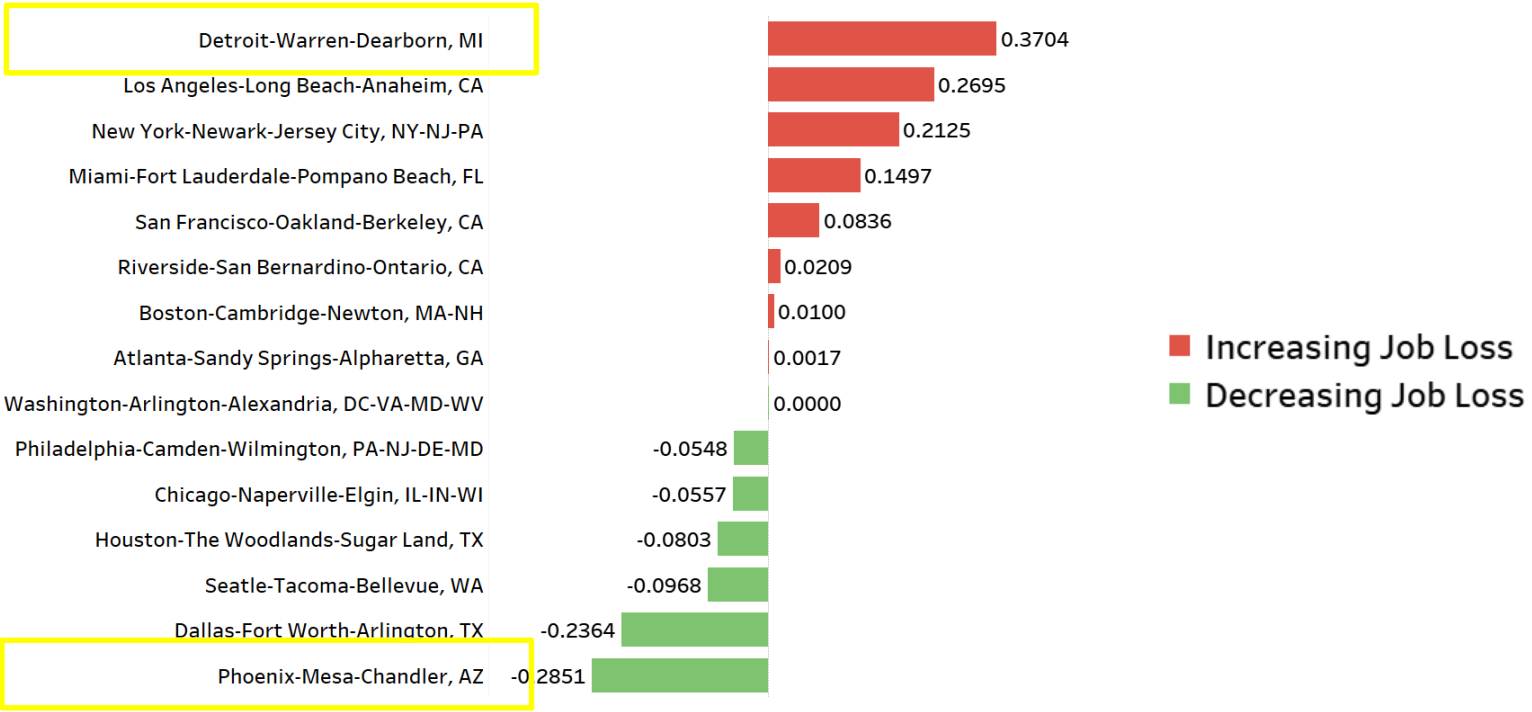
Odds Ratio for Education and Income Levels





Analysis

Metropolitan Statistical Area Estimates

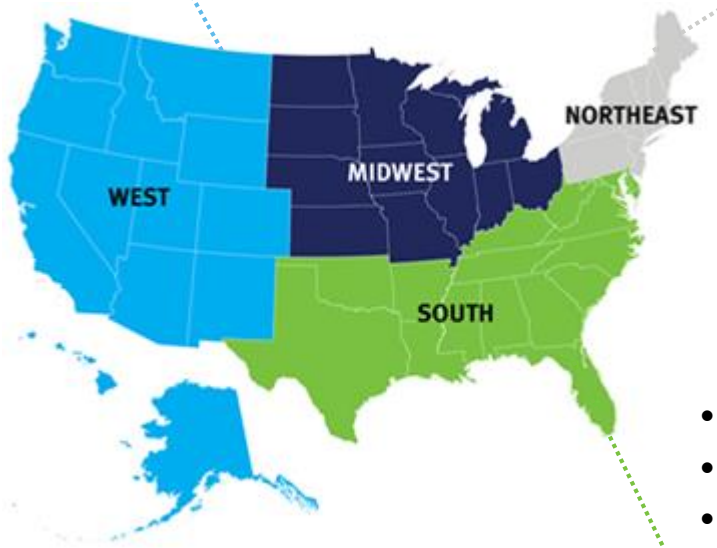


Significant Factors by Region

Shared Significant Factors:

- Education
- Race
- Tenure
- Marital Status
- Income
- Health Insurance
- Health Status
- Number of Adults in Household
- Metropolitan Statistical Area

- Age
- Infection rate



- Infection rate
- Number of kids

- Age
- Gender
- Infection rate
- Number of kids

Recommendations

Education incentive

Women at workplace

Dependent and respite care

Stimulus packages

Future Scope

Expand data to all respondents

Industry Data

Socio-economic Data

Hindsight is 2020

Expectation vs what the data says; let the data be the guide

Personal interpretation may bias variables selected

Experiment with more advanced models, like time series, or different methods of handling missing values

Thank you!

Contact Information

Hannah Flynt	Hannah.flynt@okstate.edu
Maryam Taherirani	Maryam.taherirani@okstate.edu
Sean Everett	Sean.everett@okstate.edu
Trinh Phan	Trinh.phan@okstate.edu