Mapping the Seasons of Sickness: Geographic and Seasonal Trends in U.S. Disease Incidence

Health Cats : Bryan Jacobs, Yash Sharma,
Antonio Escalante Jr., & Partha Vemuri

Topic and Motivation

Analyzing	**Analyzing** seasonal patterns in disease incidence across U.S. states to identify periods of heightened risk.
Examining	**Examining** geographic variations to determine how location and seasonal factors influence disease risk.
Informing	Informing public health interventions based on geographic and seasonal trends

Introduction to Data

- Dataset: HealthData.Gov
  - Project Tycho Level 1 Data – Jan 1916 to Jan 2012
- Size: 759,000 rows and 7 columns
- Key columns:
  - epi_week: Epidemiological week
  - state: U.S. state abbreviation
  - disease: Disease name
  - cases: Reported cases
  - incidence_per_100000: Incidence rate per 100,000 population

EDA Highlights

Seasonal incidence rates analyzed using custom season variable

Geographic patterns visualized through heat maps

Bubble maps created to compare total cases and incidence rates by state

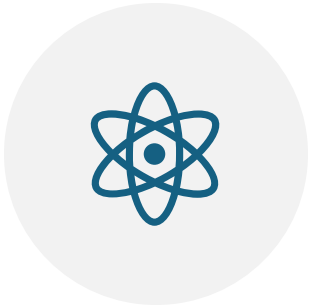
Logistic Regression Model

```
# measles, all years, arizona baseline state

X = measles_df[["state", "season"]]
y = measles_df["risk"]

# Encode Arizona as the baseline state
X_encoded = pd.get_dummies(X, drop_first=True)
baseline_state = "state_Arizona"
if baseline_state in X_encoded.columns:
    X_encoded = X_encoded[[baseline_state] + [col for col in X_encoded.columns if col != baseline_state]]
X_encoded = X_encoded.replace({True: 1, False: 0})

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(X_encoded, y, test_size=0.2, random_state=42, stratify=y)

# Convert target to binary
y_train_binary = y_train.map({'low': 0, 'high': 1})
y_test_binary = y_test.map({'low': 0, 'high': 1})

# Add constant to the predictors
X_train_with_const = sm.add_constant(X_train)
X_test_with_const = sm.add_constant(X_test)

# Fit logistic regression model
logit_model = sm.Logit(y_train_binary, X_train_with_const).fit()

# Model summary
logit_model.summary()
```

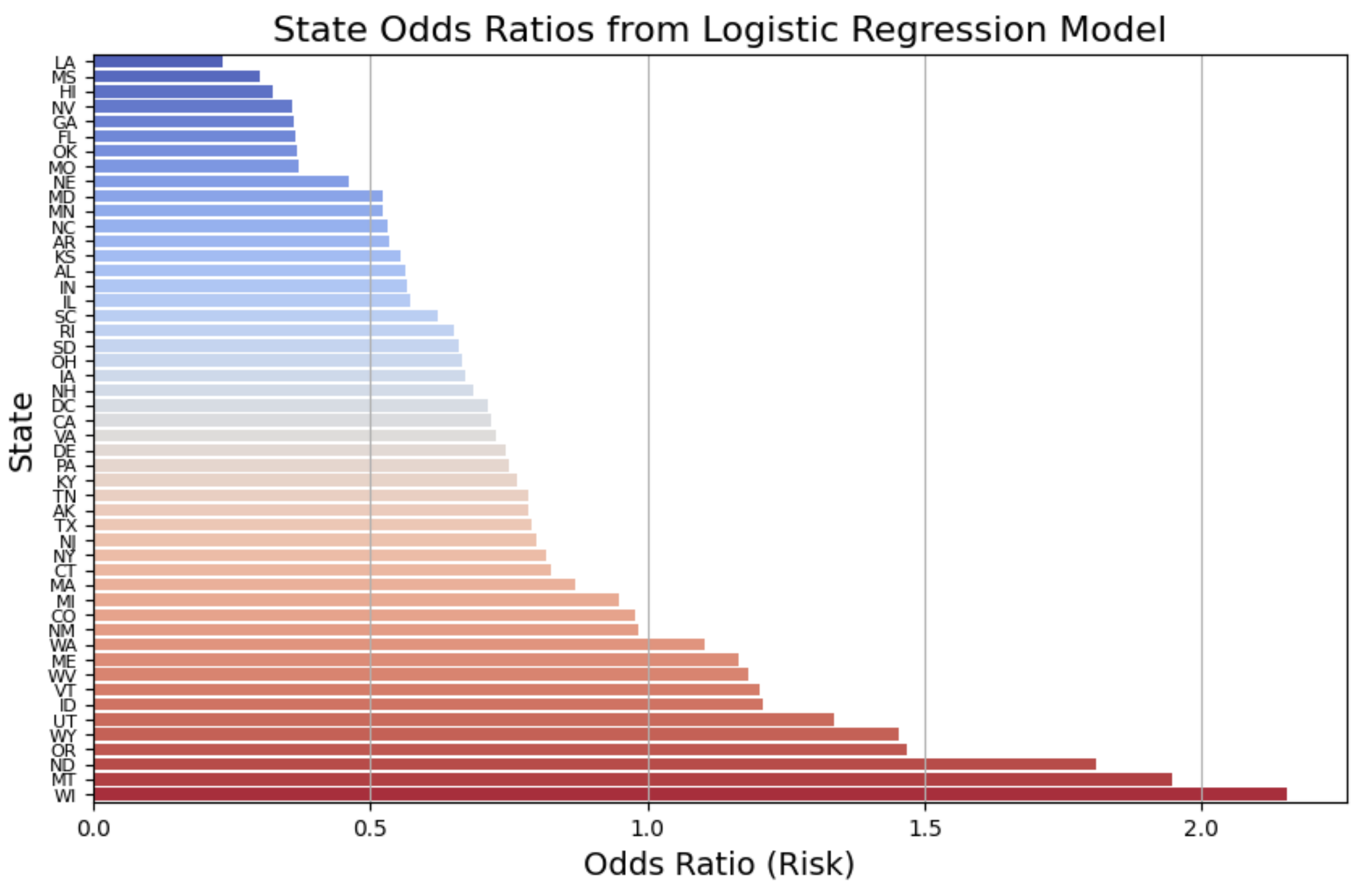
LOGISTIC REGRESSION TO CLASSIFY STATES AS HIGH-RISK OR LOW-RISK

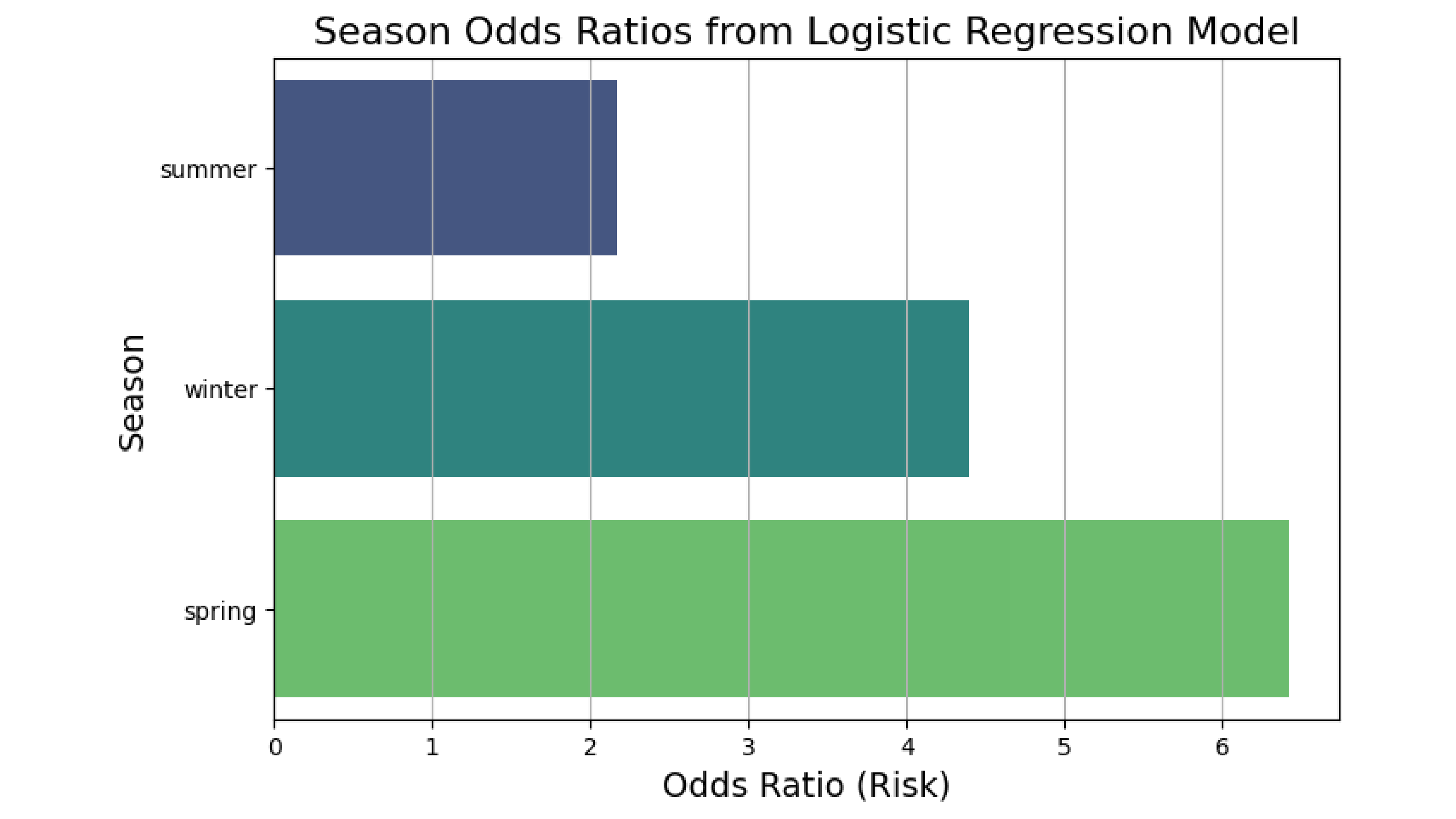
PREDICTORS: STATE AND SEASON

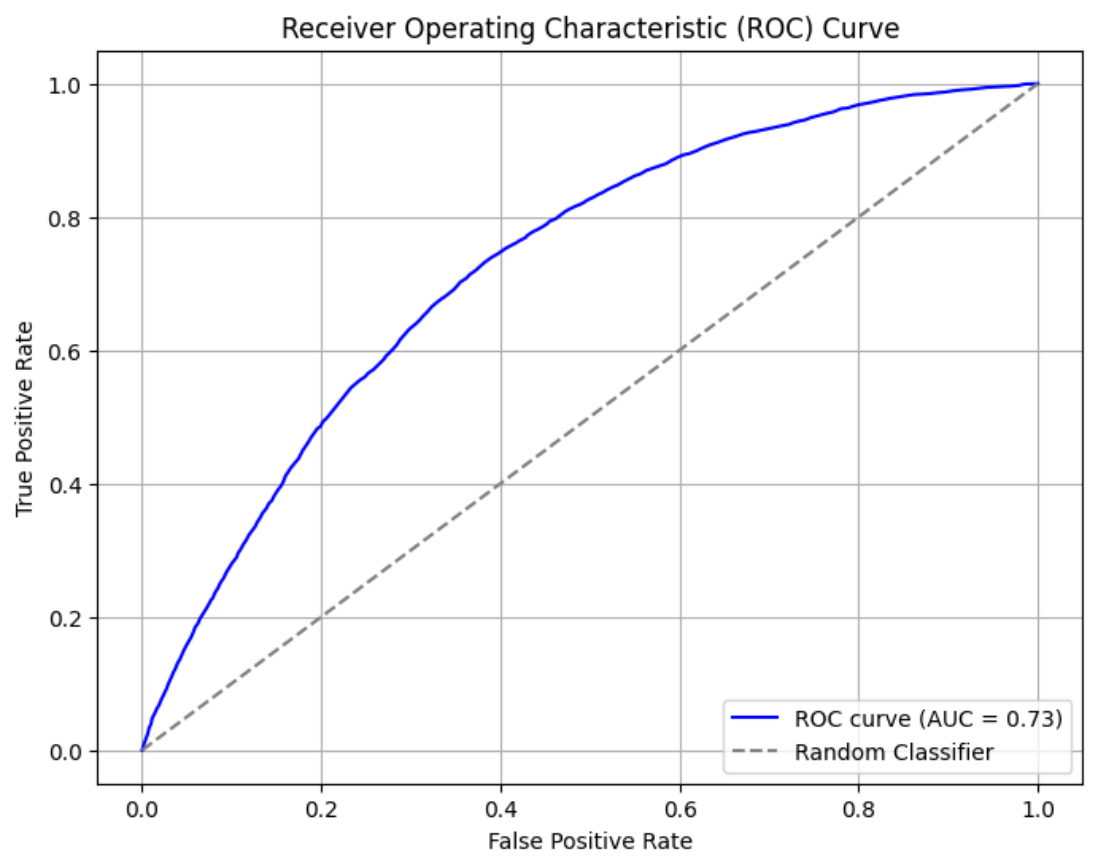
EVALUATION METRICS: ACCURACY, PRECISION, RECALL, AUC, AND F1-SCORE



DATA SPLIT: 80% TRAINING, 20% TESTING


State Odds Ratios from Logistic Regression Model
State
LA
MS
HI
NV
GA
FL
OK
MO
NE
MD
MN
NC
AR
KS
AL
IN
IL
SC
RI
SD
OH
IA
NH
DC
CA
VA
DE
PA
KY
TN
AK
TX
NJ
NY
CT
MA
MI
CO
NM
WA
ME
WV
VT
ID
UT
WY
OR
ND
MT
WI
0.0
0.5
1.0
1.5
2.0
Odds Ratio (Risk)


Season Odds Ratios from Logistic Regression Model
summer
winter
spring
Season
0
1
2
3
4
5
6
Odds Ratio (Risk)



Model Evaluation

- Accuracy – 0.67
  - 67% of predictions were correct
- Precision – 0.64
  - 64% true positives
- Recall – 0.59
  - Identified 59% of actual positive cases
- F1-Score – 0.62
  - Indicates a moderate balance between identifying positives and avoiding false positives
- AUC – 0.73
  - 73% chance of correctly distinguishing between high and low risk

Conclusions

The model can determine one's risk of getting measles relatively well based on state and season

This is based on the AUC value, where 0.5 would be random guessing, 0.73 is notably better than random guessing

The risk of getting measles is greatest in the spring, the odds of being at a high risk for measles in spring are ~6.5 times greater than in the fall

The risk of getting measles is greatest in Wisconsin and least in Louisiana

If you move from AZ to WI your odds of being at high risk for measles are ~2.8 times greater, and if you move from AZ to LA your odds of being at high risk for measles are ~0.3 times as great

Conclusions and Future Work