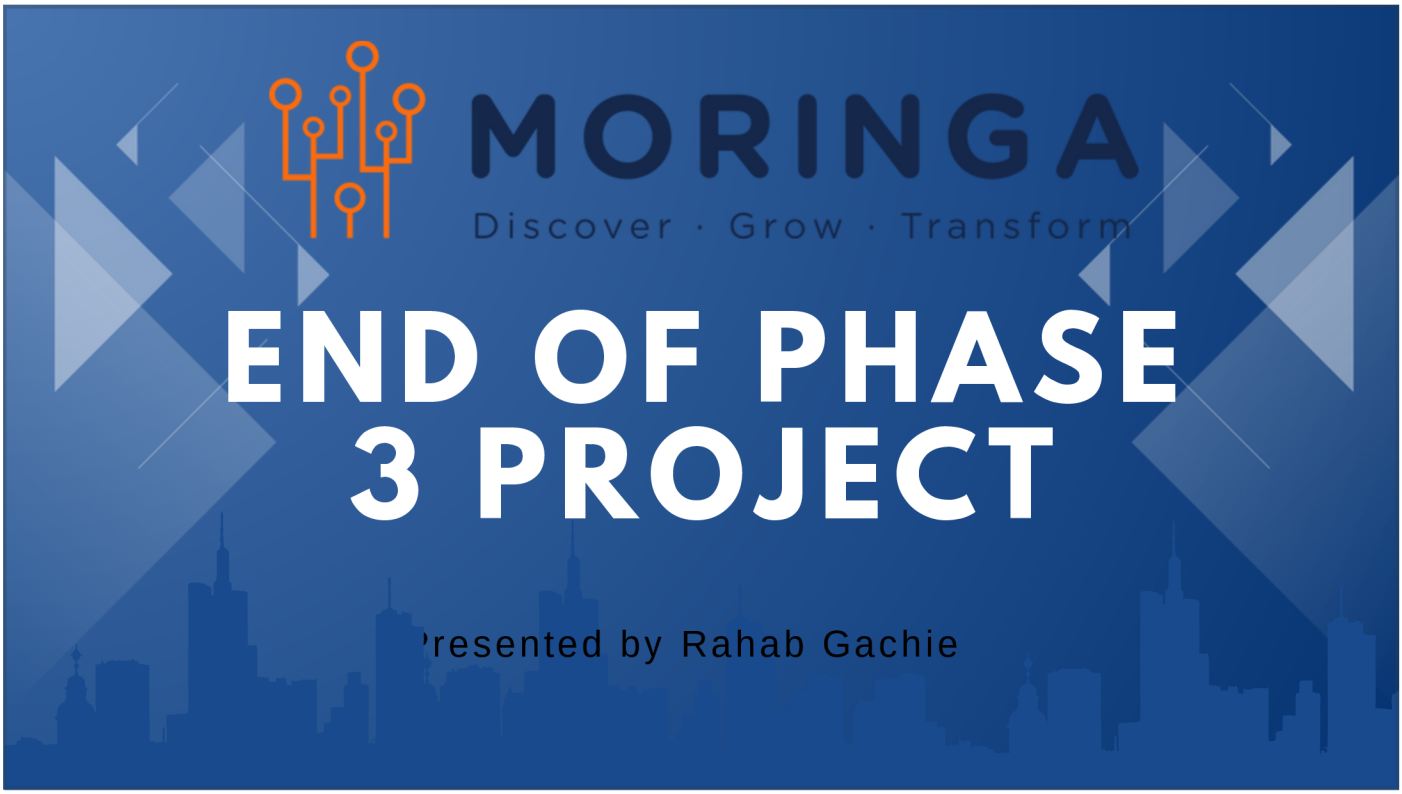

MORINGA
Discover · Grow · Transform
END OF PHASE 3 PROJECT
Presented by Rahab Gachie

PROJECT OBJECTIVE

To build a predictive model that supports a pharmaceutical company in launching and scaling a seasonal flu vaccine product line by increasing uptake and optimizing distribution.

OVERVIEW

A pharmaceutical company wants to start manufacturing the seasonal flu vaccine as a new product line. They would like to know how they can increase vaccine uptake and optimize distribution strategies to ensure that they get to majority of the population. To get better insights the company would want to make data driven decisions.

DATA UNDERSTANDING

Source: 2009 National H1N1 Flu Survey (CDC)
Population: U.S. residents aged 6 months+
Datasets:
Training features (demographics, behavior, opinions)
Training labels (received vaccine or not)
Test set (for prediction)

Modeling & Evaluation

Before embarking on the modeling, the data was prepared in the following ways:

Handling the missing values: The missing data in the numerical columns was imputed with the mean and categorical columns with the most frequent value.
Feature scaling: The numerical features are standardized using StandardScaler in the preprocessing pipeline, 0,1.
Data balancing: This is important to avoid bias in regression and classification. Logistic regression is configured to use balanced class weight so that it assigns more importance to cases from minority class during training.
Two models were build using Logistic Regression and Decision Tree algorithms and thereafter hyperparameters were employed to tune the two models.
Hyperparameters are crucial for optimizing the performance of machine learning models. A combination of Regularization Strength (C), Penalty type and Solver were used.

Model Comparison

Metric	Logistic Regression	Decision Tree
ROC AUC	0.855	0.690
F1-Score (class 1)	0.769	0.667
Interpretability	High	Medium

After modeling with the 2 algorithms, Logistic Regression performed best with a ROC AUC of 0.85 which identified strong predictors of the vaccine uptake. The ROC AUC score shows how well the model separates the two classes,that is the vaccinated vs non-vaccinated.
Logistic regression was consistent in showing better class separation and generalization on the validation set. Also, it provided clear, interpretable coefficients showing the direction and strength of influence for each feature. This gives the much needed insights to stakeholders to help them understand why certain populations are more or less likely to get vaccinations.

KEY PREDICTIVE FACTORS

Top positive predictors:

-Doctor recommendation

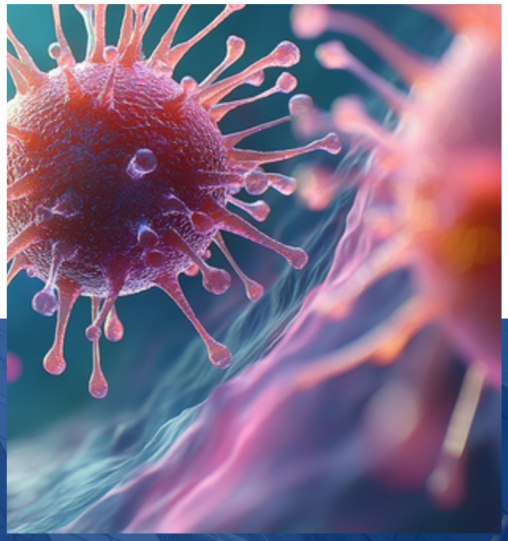
-Perceived vaccine effectiveness

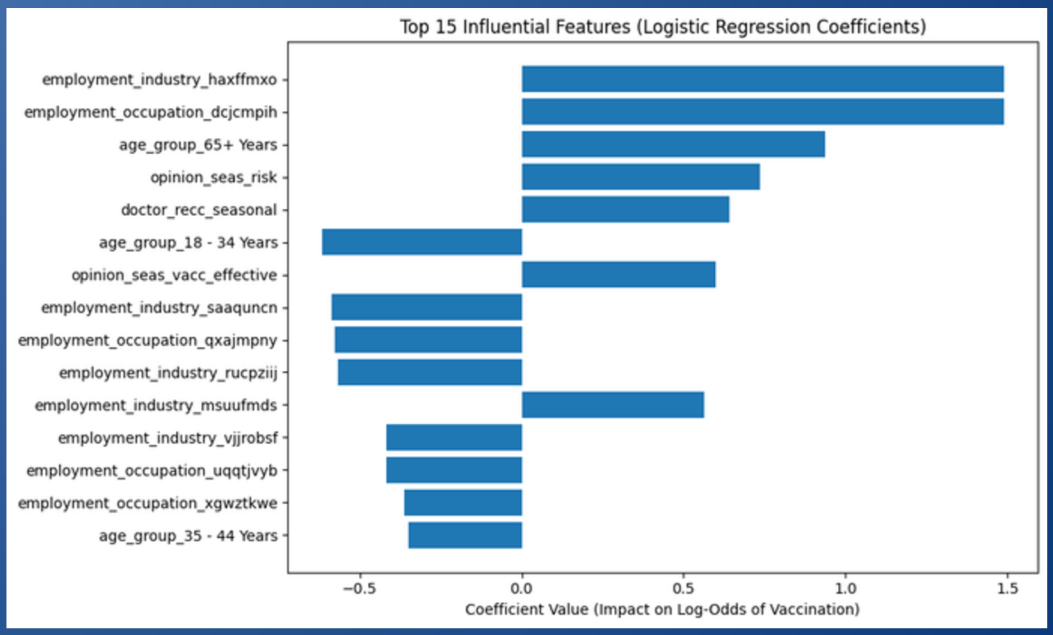
-Perceived flu risk

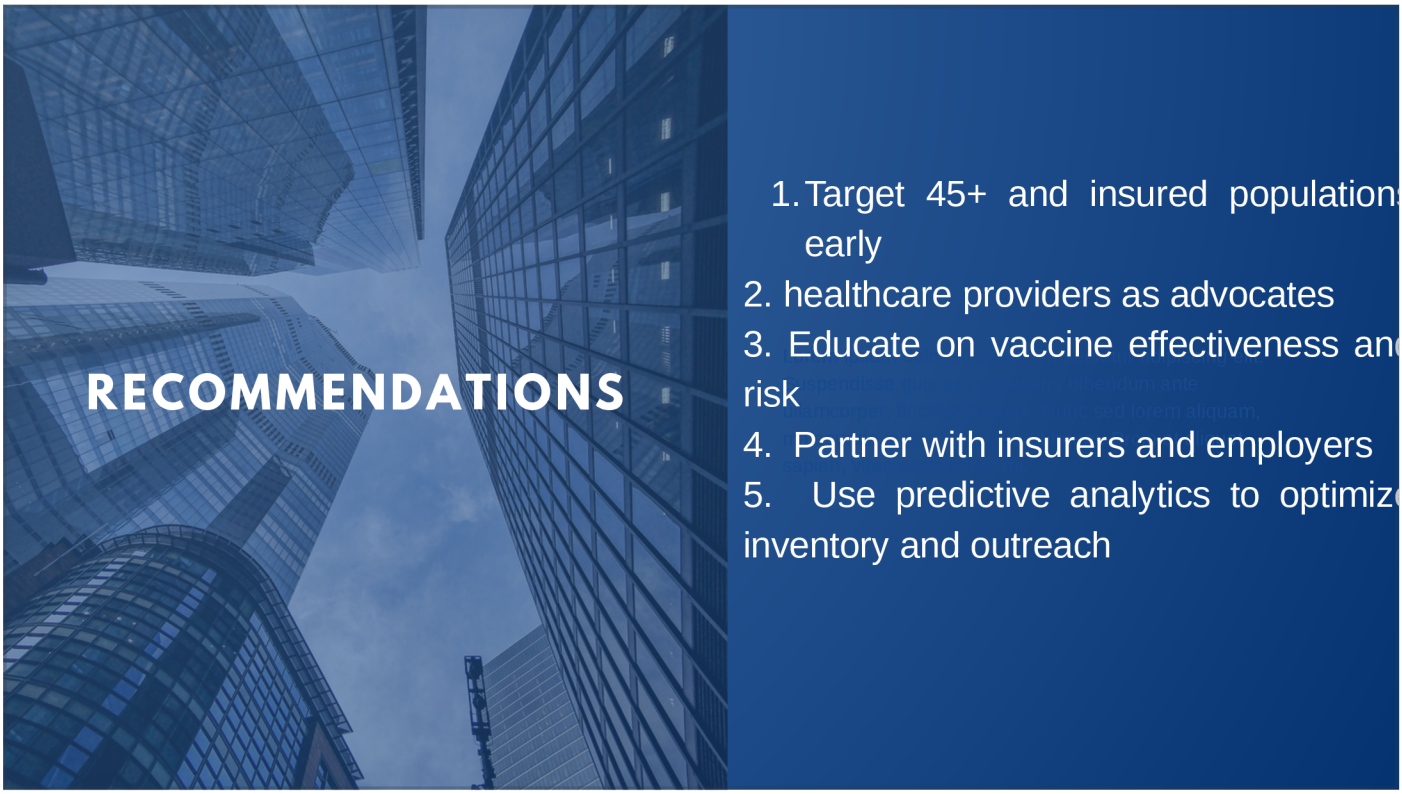
Top negative predictors:

-Young age groups (18–44)

-Certain occupational categories




Top 15 Influential Features (Logistic Regression Coefficients)
employment_industry_haxffmxo
employment_occupation_dcjcmpih
age_group_65+ Years
opinion_seas_risk
doctor_recc_seasonal
age_group_18 - 34 Years
opinion_seas_vacc_effective
employment_industry_saaquncn
employment_occupation_qxajmpny
employment_industry_rucpziij
employment_industry_msuufmds
employment_industry_vjjrobsf
employment_occupation_uqqtjvyb
employment_occupation_xgwztkwe
age_group_35 - 44 Years
−0.5
0.0
0.5
1.0
1.5
Coefficient Value (Impact on Log-Odds of Vaccination)


RECOMMENDATIONS
1. Target 45+ and insured population early
2. healthcare providers as advocates
3. Educate on vaccine effectiveness and risk
4. Partner with insurers and employers
5. Use predictive analytics to optimize inventory and outreach

NEXT STEPS

1.Explore more advanced models (Random Forest, XGBoost)
2.Deploy model into digital tools for field reps
3.Monitor and refine with ongoing data


THANK
YOU