Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines

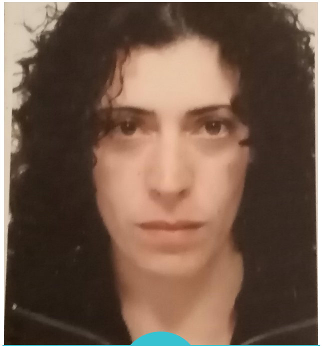
Big Data Analytics
A.A. 2020/21

MaLuCS
Luca Corbucci
Cinzia Lestini
Marco Giuseppe Marino
Simone Rossi

MaLuCS TEAM

LUCA CORBUCCI

Master Degree in Computer Science

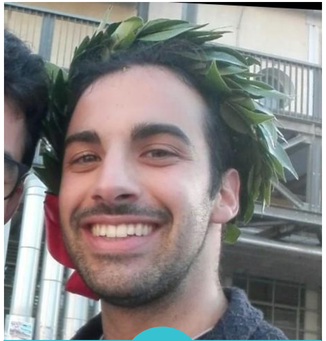


CINZIA LESTINI

Master Degree Data Science and Business Informatics

MARCO GIUSEPPE MARINO

Master Degree in Computer Science



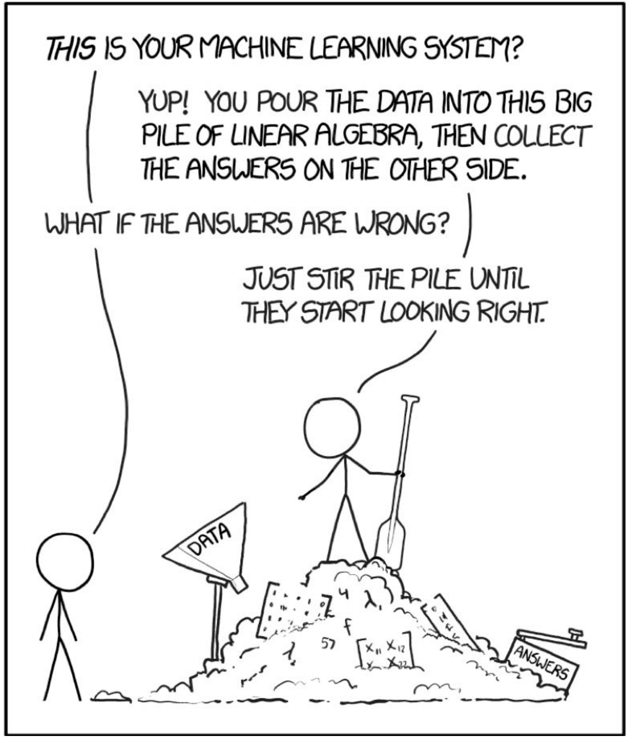
SIMONE ROSSI

Master Degree in Computer Science

INTRODUCTION

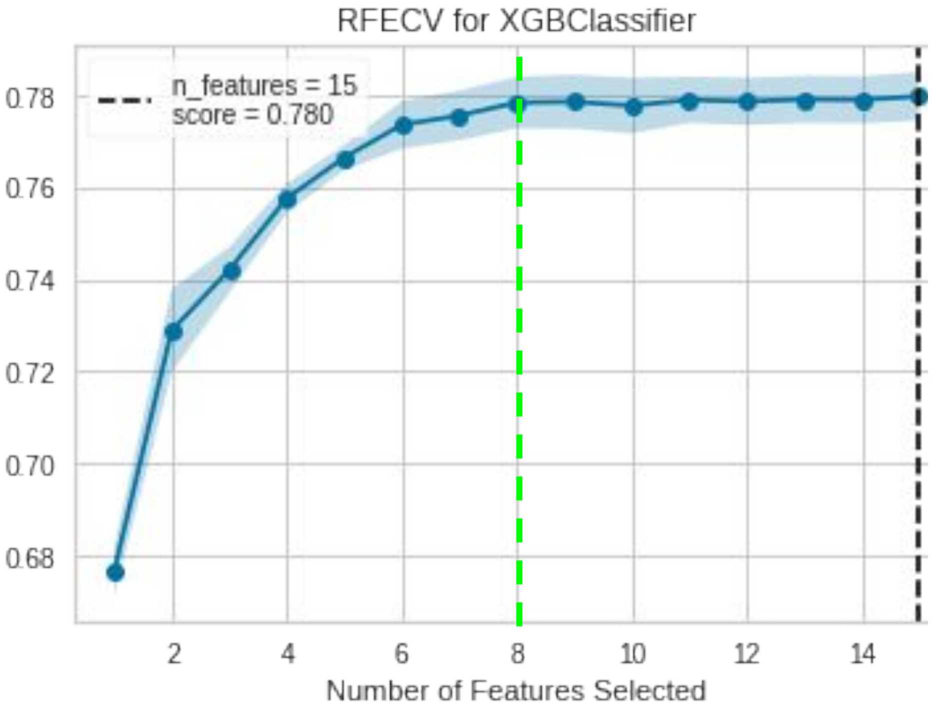
WHY IS IMPORTANT TO UNDERSTAND HOW OUR MODELS WORK?

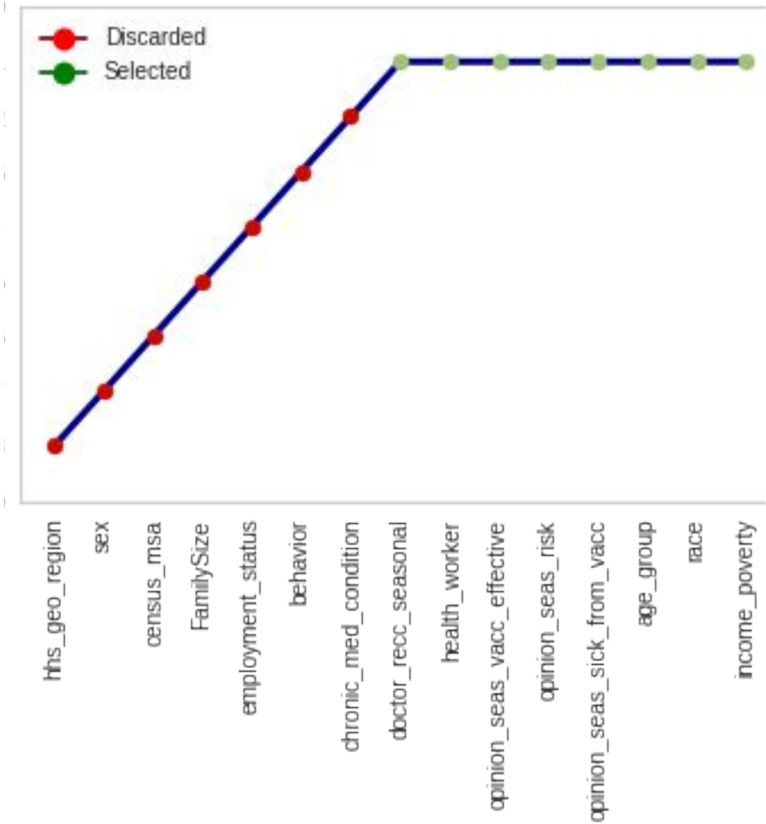
- We can understand whether our models have bias (for example racial bias)
- We can explain to a possible customer why a certain model makes a prediction
- We can understand how each feature affect the final prediction of the model
- In this specific case we can understand why a person is more likely to receive the vaccine



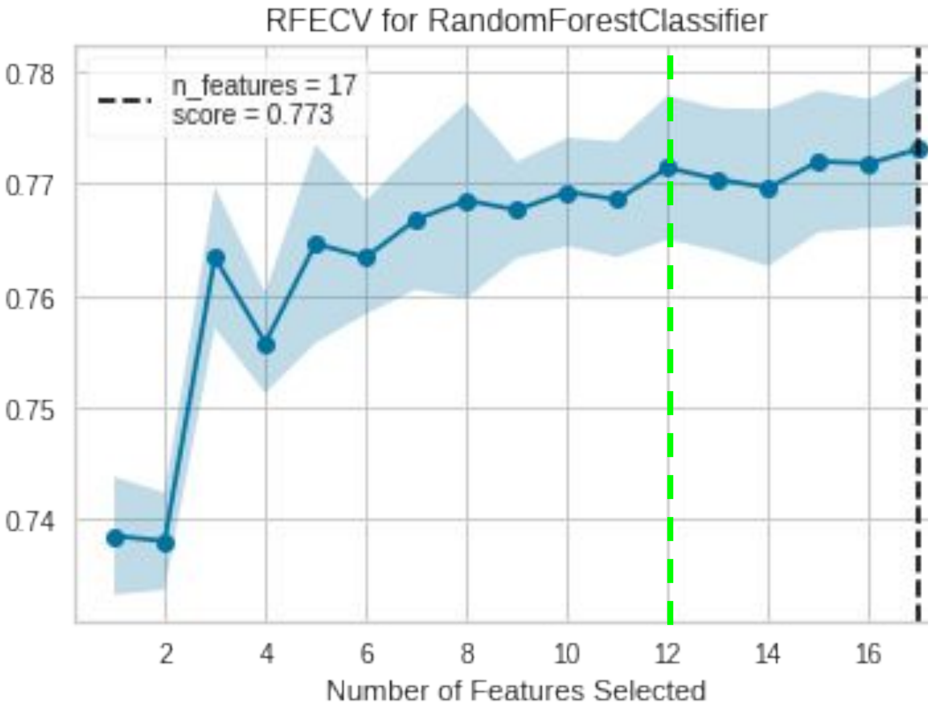
FEATURE SELECTION

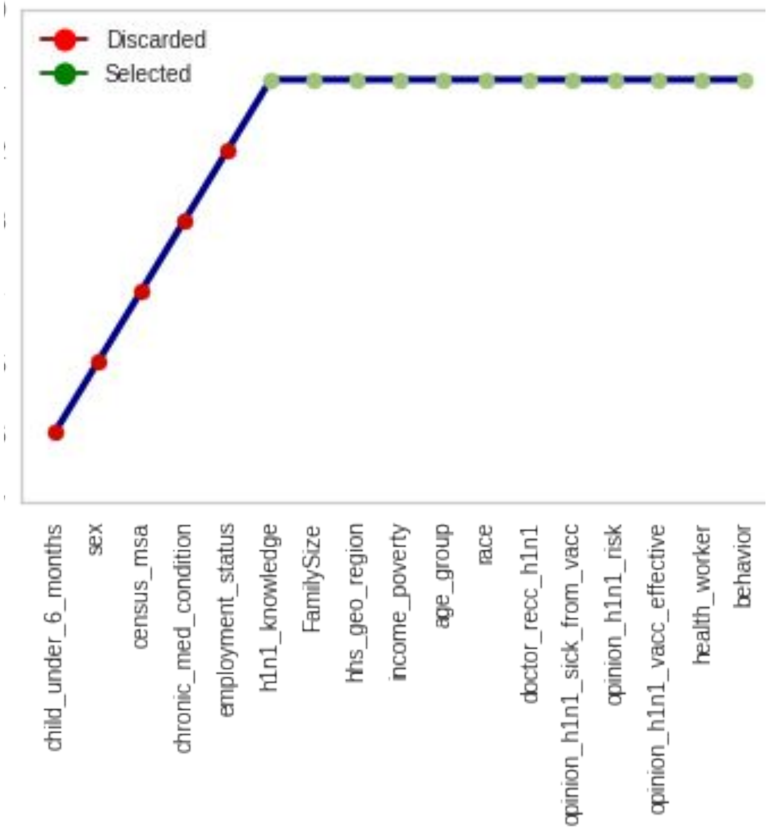
FEATURES SELECTION: Seasonal flu XGBoost



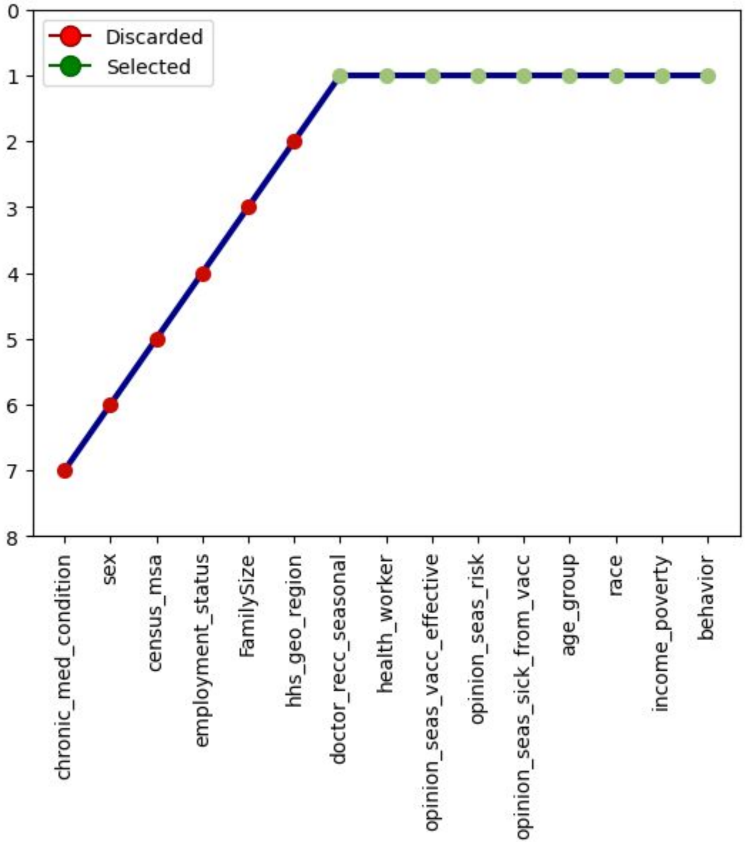


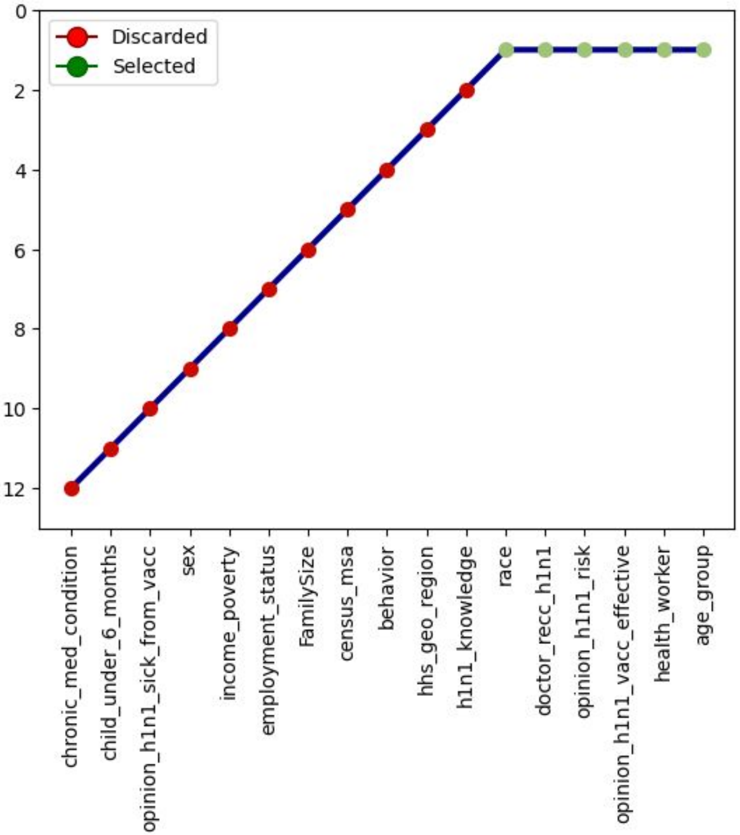
FEATURES SELECTION: H1N1 Random Forest





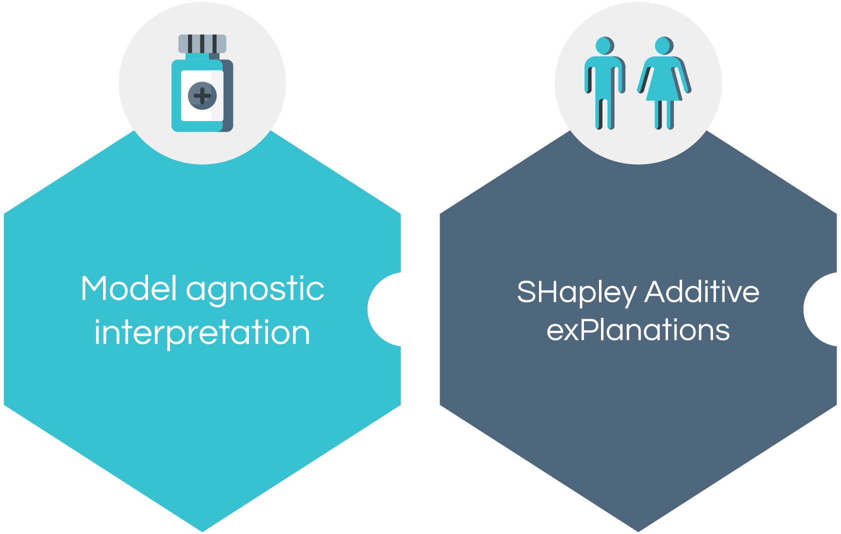
FEATURES SELECTION: Seasonal flu and H1N1 Decision Tree





GLOBAL EXPLANATIONS

GLOBAL EXPLAINER



SKATER

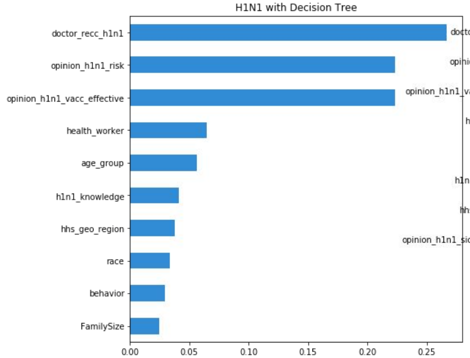
SHAP

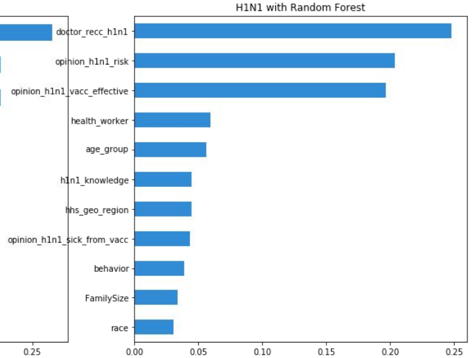
Skater

Interpretation Techniques

- Feature Importances
- Partial Dependence Plots
- Visualizing boundaries
- Surrogate Tree-based Models

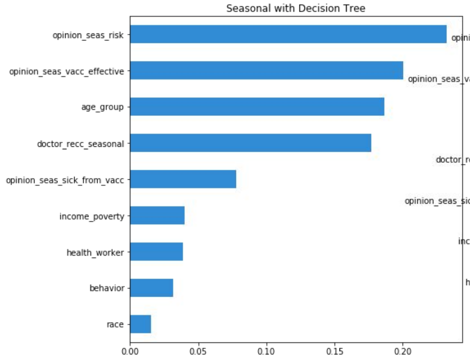
Features Importance

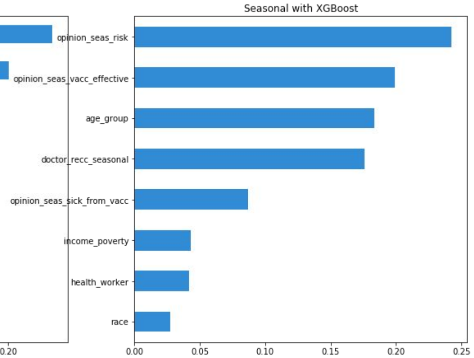




H1N1 Vaccine

" doctor_rec_h1n1"
"Opinion_h1n1_risk"
"opinion_h1n1_effective"



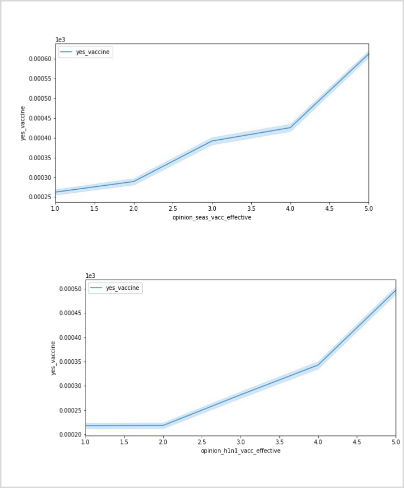


Seasonal Vaccine

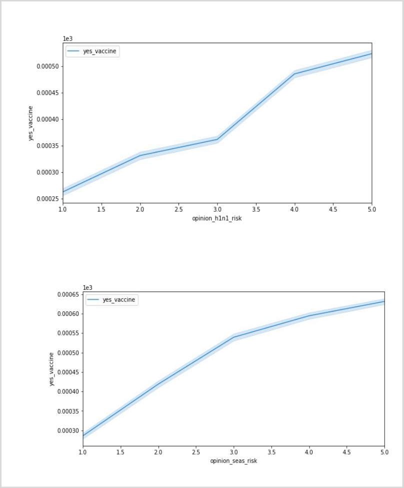
"Opinion_seas_risk"
"opinion_seas_vacc_effective"
"age_grop"

Partial Dependence plot – Similarity

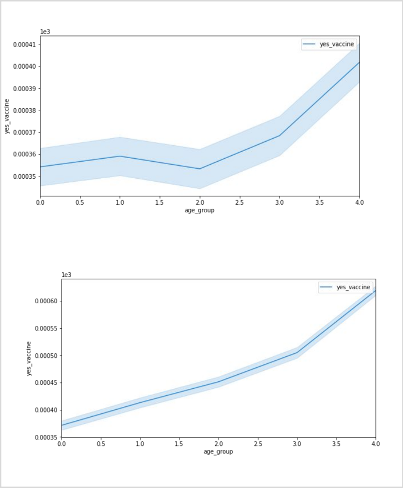
Effectiveness



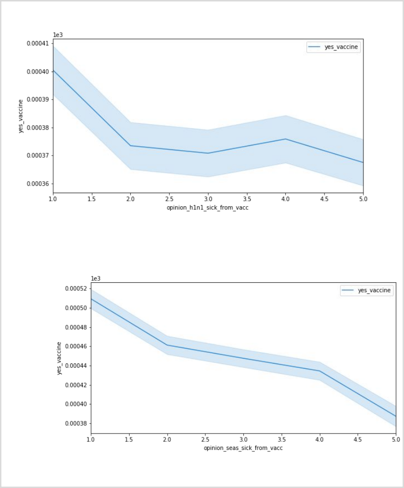
Flu Risk



Age

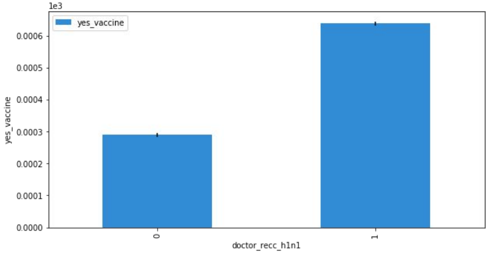


Sick with vaccine

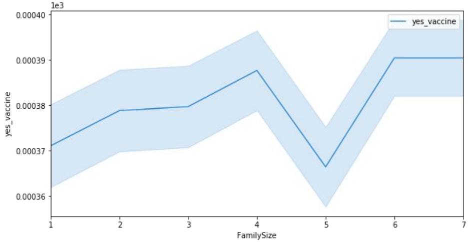


Partial Dependence plot - Others

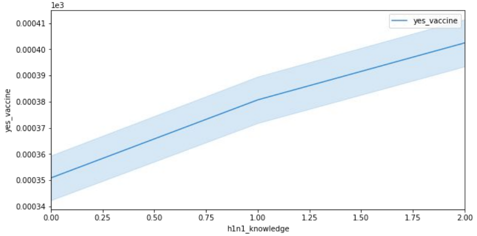
Doctor recommended h1n1



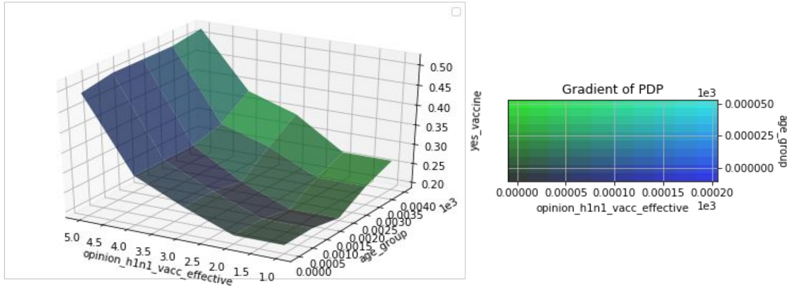
Family Size



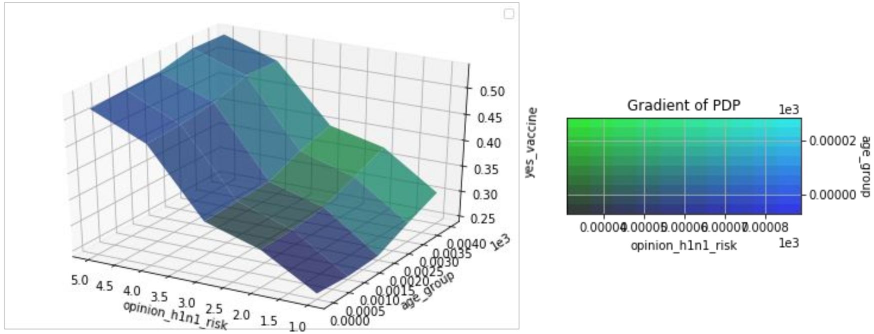
H1n1 knowledge

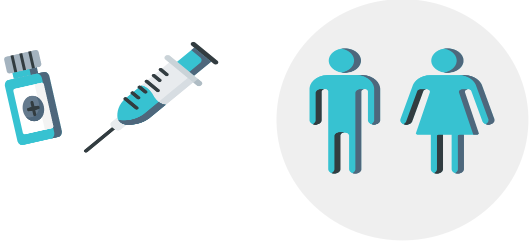


2D Partial Dependence Plot



TEST

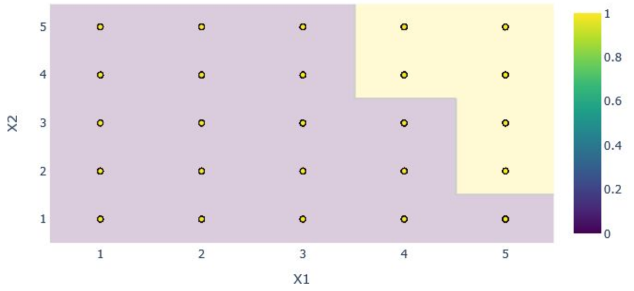




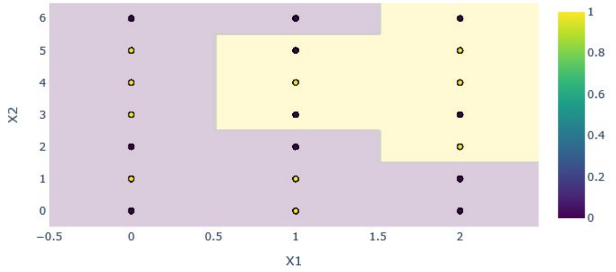
Visualizing Boundaries

Visualizing boundaries we can see the border of the decision between class 0 and 1

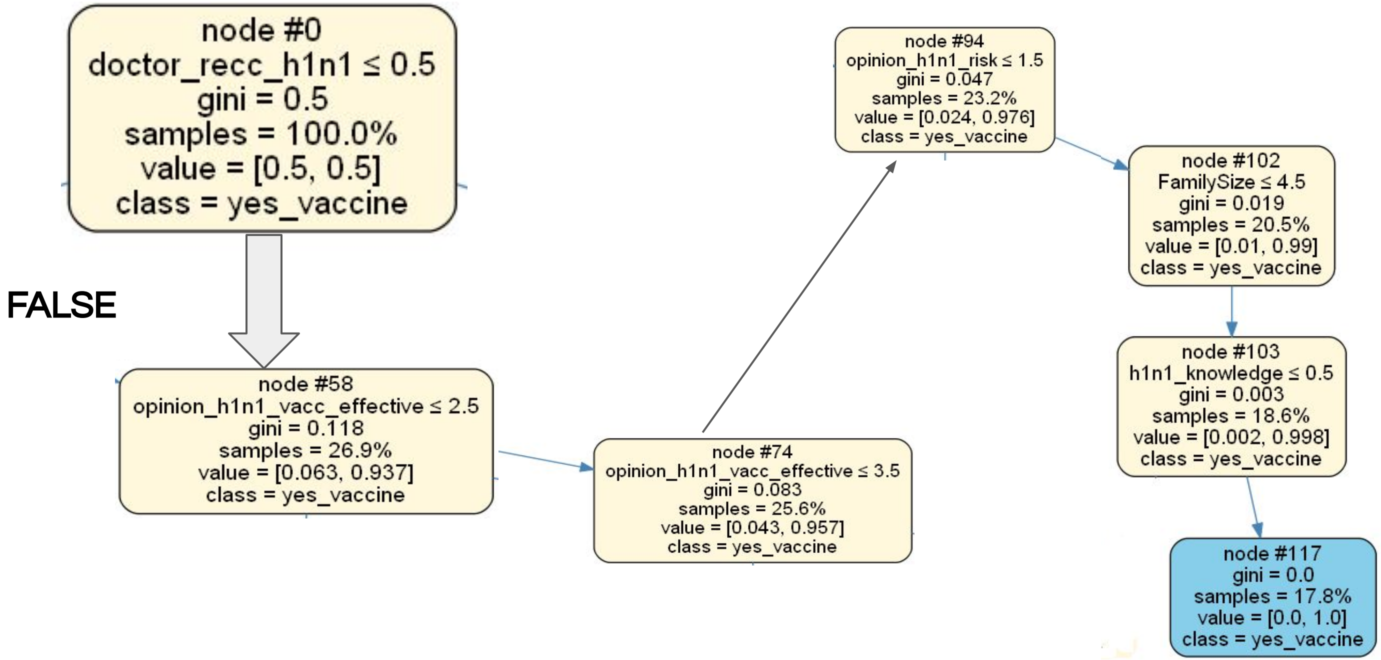
X1= opinion_h1n1_vacc_effective and X2= opinion_h1n1_risk



X1= income_poverty and X2= behavior



Surrogate Tree Models

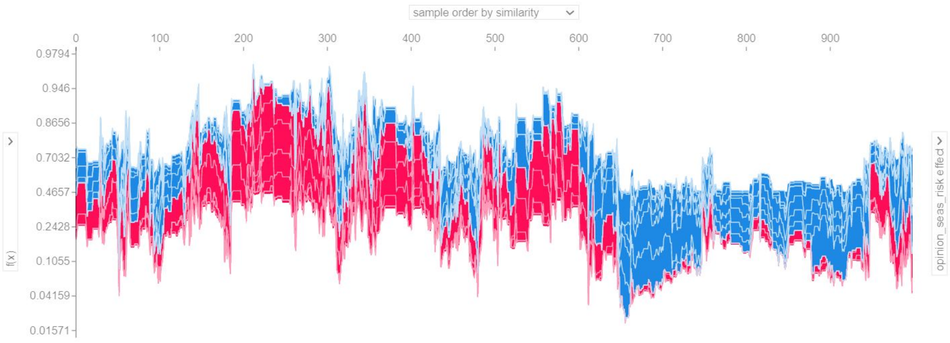


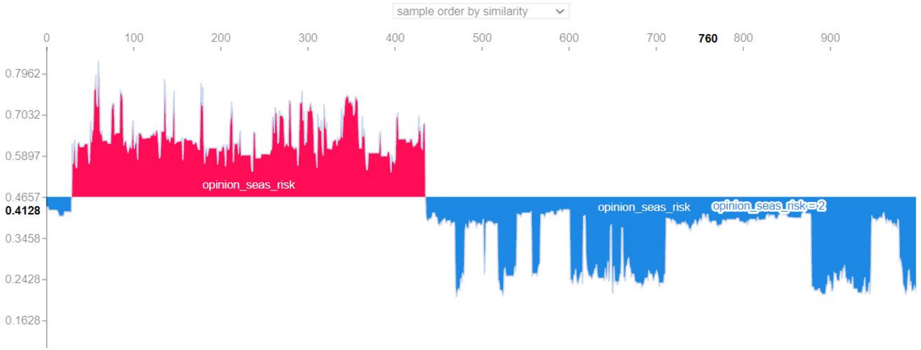
SHapley Additive exPlanations

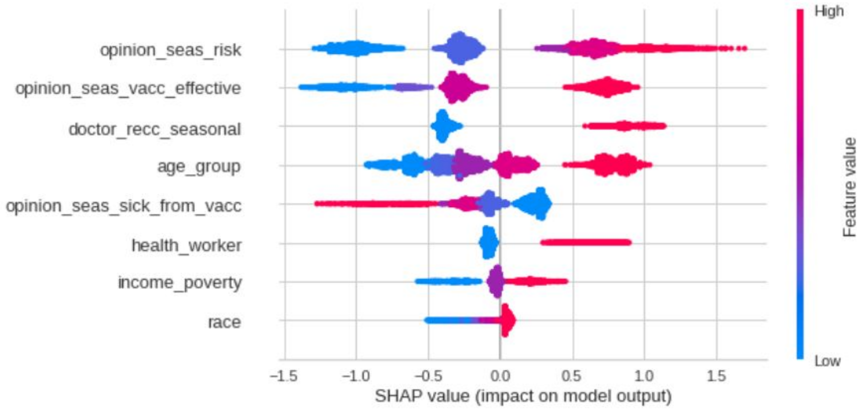
SHAP GLOBAL

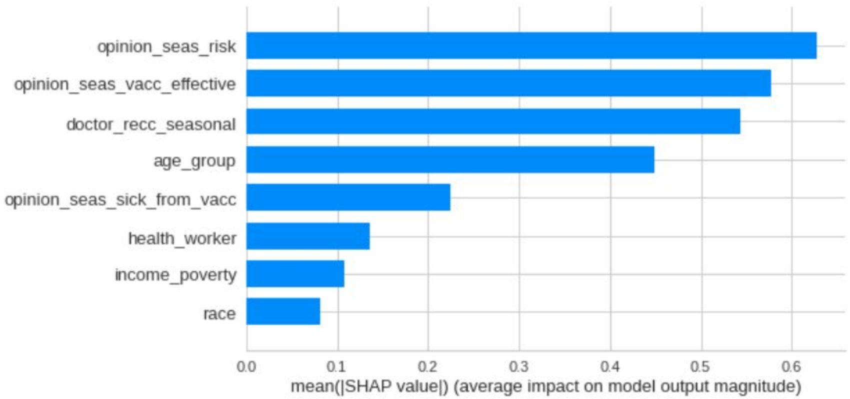
Interpretation Techniques

SHAP GLOBAL: Seasonal flu XGBoost

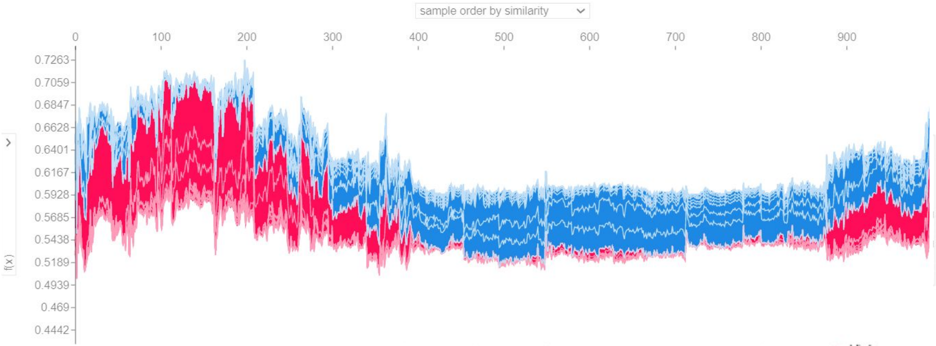


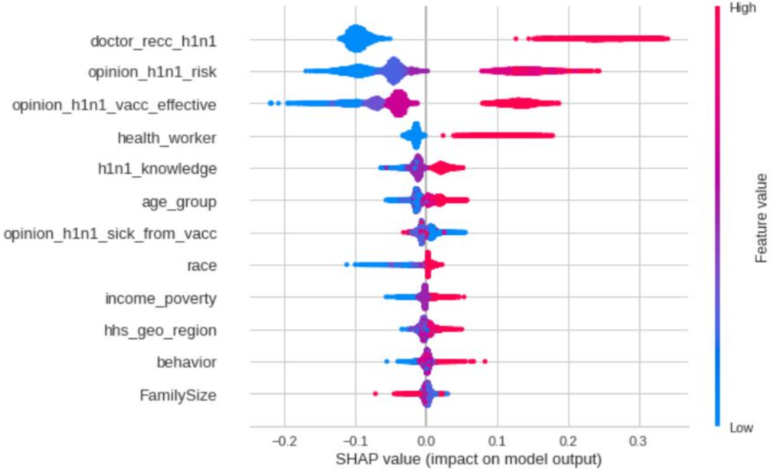


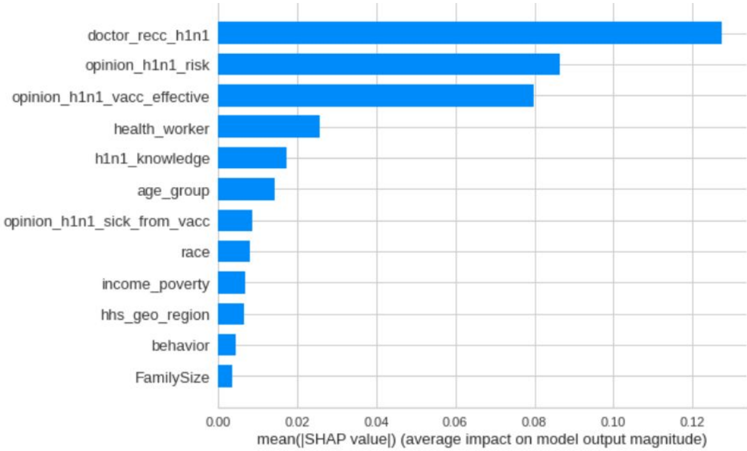




SHAP GLOBAL: H1N1 Random Forest







SHAP GLOBAL: Seasonal flu and H1N1 features influence

- opinion_seas_risk : 3|4|5 -> 1 and 1|2 -> 0
- opinion_seas_vacc_effective: 5 -> 1 and 1|2|3|4 -> 0
- doctor_recc_seasonal: 0 -> 0 and 1 -> 1
- age_group: <= “45 - 54 Years” -> 0 and >= “55 - 64 Years” ->1
- opinion_seas_sick_from_vacc: > 1 -> 0 and 1 -> 1
- health_worker: 0 -> 0 and 1 -> 1
- income_poverty: “ > $75000” -> 1 and <= “$75000 Above Poverty ” -> 0
- race: White -> 1 and Other/Multiple|Black|Hispanic -> 0

- doctor_recc_h1n1: 0 -> 0 and 1 -> 1
- opinion_h1n1_risk : 3|4|5 -> 1 and 1|2 -> 0
- opinion_h1n1_vacc_effective: 5 -> 1 and 1|2|3|4 -> 0
- health_worker: 0 -> 0 and 1 -> 1
- h1n1_knowledge: 0|1 -> 0 and 2 -> 1
- age_group: <=“45 - 54 Years” -> 0 and >=“55 - 64 Years” -> 1
- opinion_h1n1_sick_from_vacc: 2|3|5 -> 0 and 1|4 -> 1
- race: White -> 1 | Other/Multiple , Black|Hispanic -> 0
- income_poverty: “ > $75000” -> 1 and <= “$75000 Above Poverty ” ->0
- hhs_geo_region, behavior and FamilySize not discriminant

opinion_*_vacc_effective=4	Seasonal	H1N1
Not Vaccinated	7538	9937
Vaccinated	4553	2137

LOCAL EXPLANATIONS

LOCAL EXPLAINER

We used Lime to generate a visualization of the most important features for the classification

LIME

We used Shap to generate a visualization of the most important features for the classification

SHAP

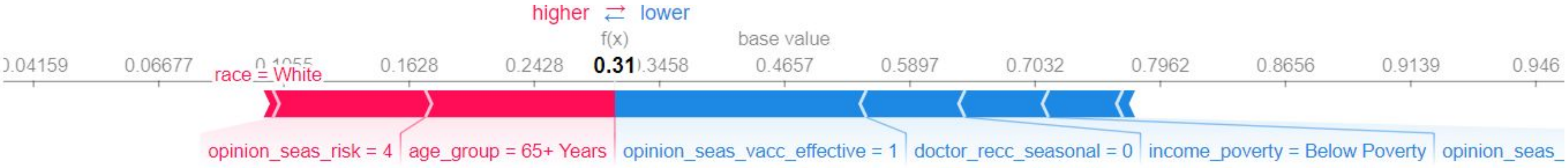
We used Lore to generate rules that can explain an instance.

LORE

SHAP

The local Shap plot shows how the expected probability of classification for a record shifts from its base value for the influence of different features.

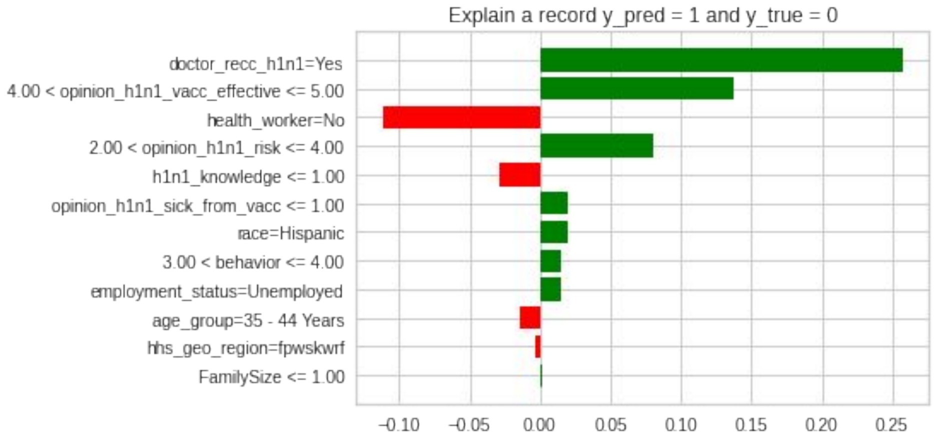
Here we consider an instance with a good classification output (No Vaccine) .



LIME

The LIME's output is a list of explanations, reflecting the contribution of each feature to the prediction of a data sample.

This provides local interpretability, and it also allows us to determine which feature changes will have the most impact on the prediction.



LORE

Lore gave us a set of rules and a set of counterfactuals that could lead the model to a different classification.

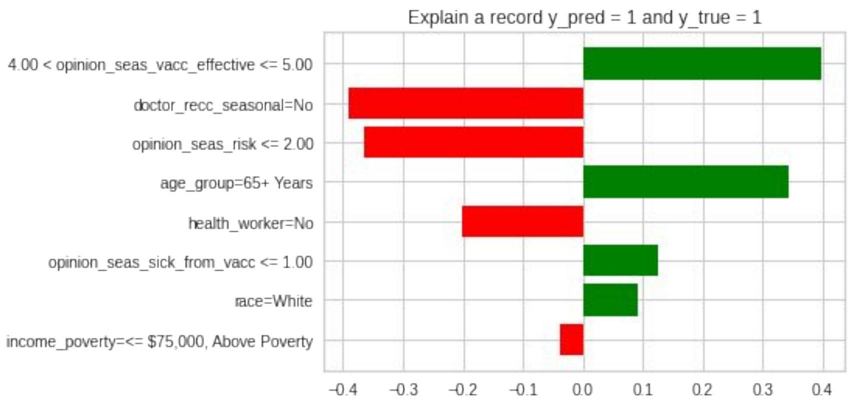
Here we consider a misclassified instance (Vaccine instead of No Vaccine)

- **Rules =** {doctor_recc_seasonal <= 0.50, race != Black
  age_group != 18 - 34 Years
  opinion_seas_risk > 1.50
  opinion_seas_sick_from_vacc <= 1.50
  opinion_seas_vacc_effective > 3.50 }
  --> { seasonal_vaccine: 1 }

- **Counterfactuals =** { { race = Black },
  { age_group = 18 - 34 Years },
  { opinion_seas_risk <= 1.50 },
  { opinion_seas_sick_from_vacc > 1.50 },
  { opinion_seas_vacc_effective <= 3.50 } }

EXPLANATION COMPARISON – SEASONAL DATASET

LIME



LORE

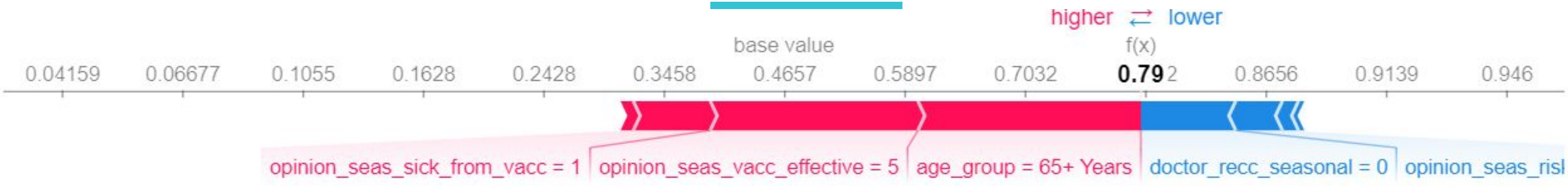
Rules:

- Age_group != 18-34 Years
- Opinion_seas_vacc_effective > 4.50
- Opinion_seas_sick_from_vacc <= 4.50

Counterfactual:

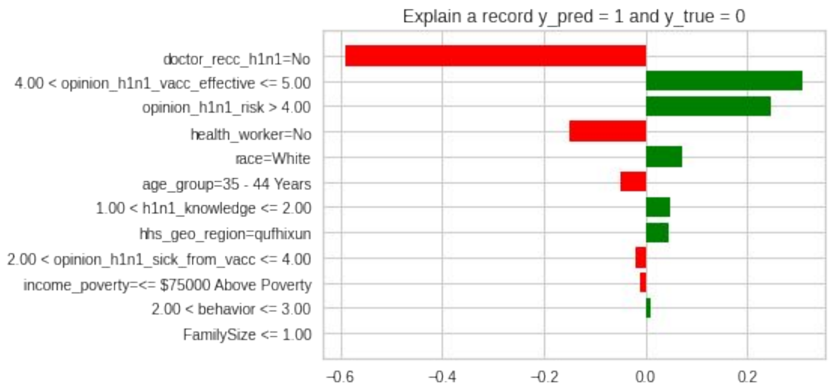
- Opinion_seas_sick_from_vacc > 4.50
- Opinion_seas_vacc_effective <= 4.50

SHAP



EXPLANATION COMPARISON – H1N1 DATASET

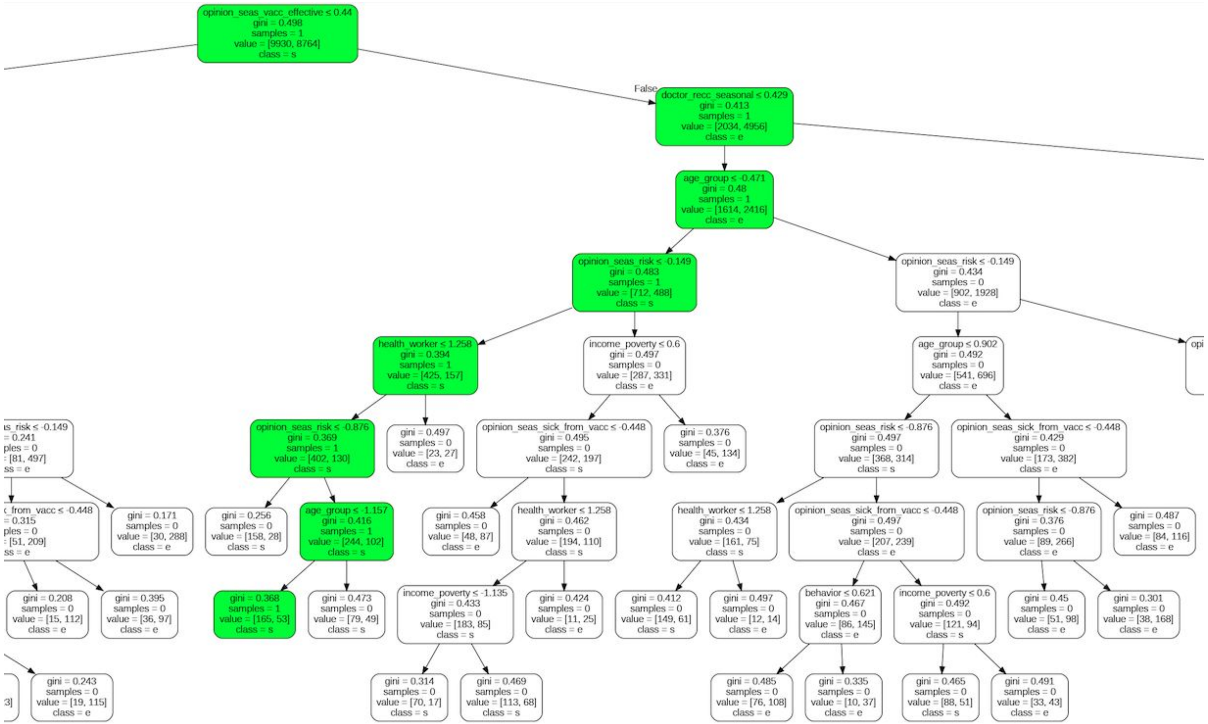
LIME



SHAP

PATH IN THE DECISION TREE

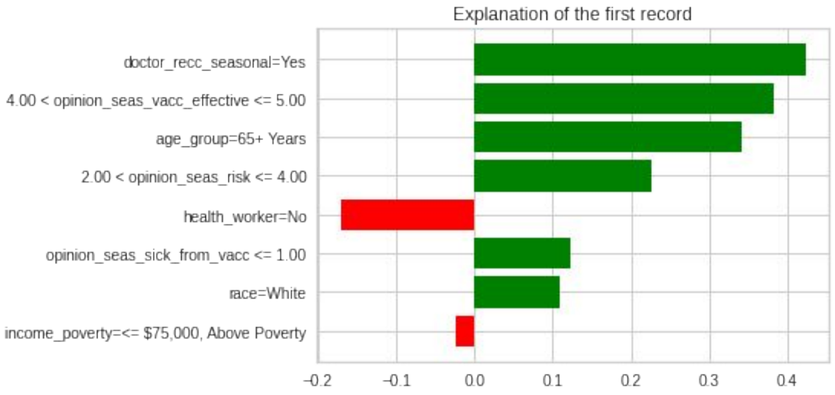
This is an example of the Decision Tree reasoning to classify a “vaccinated” person



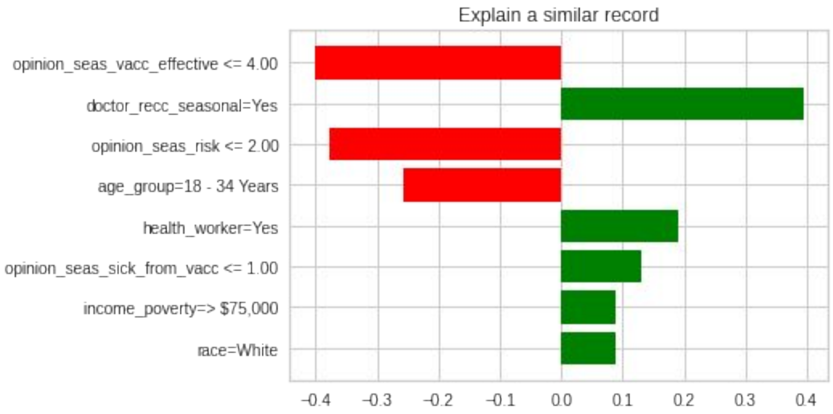
COMPARISONS

TARGET INSTANCE vs MOST SIMILAR INSTANCE

TARGET INSTANCE

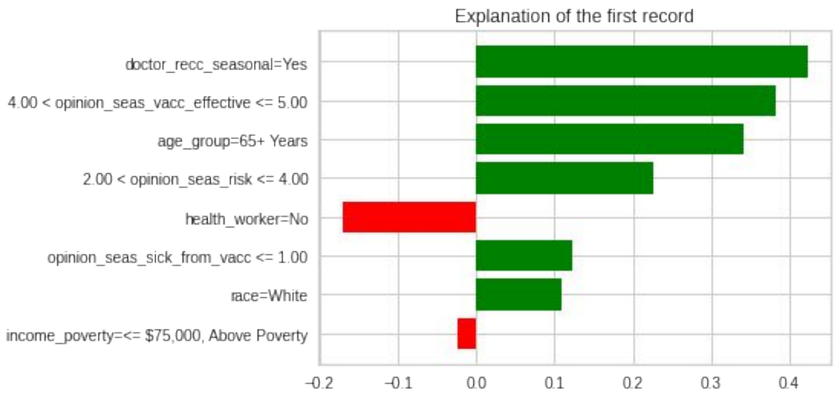


MOST SIMILAR INSTANCE

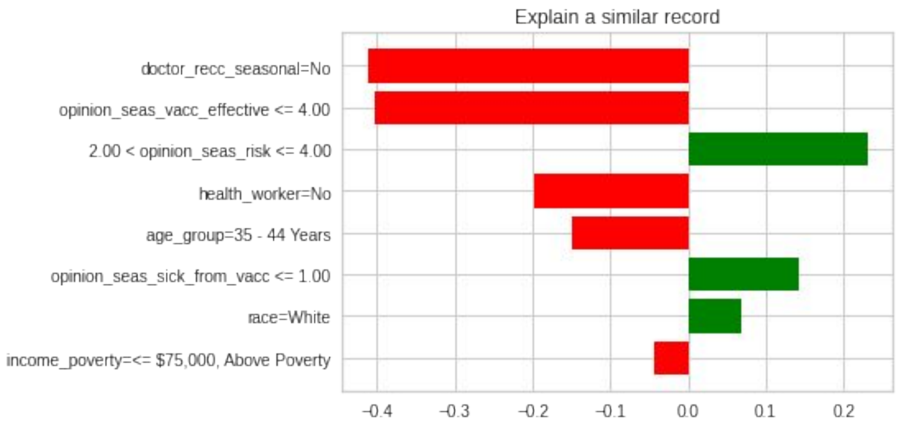


TARGET INSTANCE vs 2° MOST SIMILAR INSTANCE

TARGET INSTANCE



MOST SIMILAR INSTANCE



CONCLUSIONS

WHAT WE CAN LEARN FROM THIS MILESTONE?

From our research, young people are less inclined to vaccinate.

TARGET OF AN AWARENESS CAMPAIGN

We can consider these research to predict who is more likely to receive a vaccine.

PREDICTIONS FOR NEXT VACCINATION CAMPAIGNS

We can extract informations about the subset of populations that are more at risk of not receiving the vaccine

MOST AT RISK POPULATION GROUP

CONCLUSIONS

The third Milestone led us to an in depth knowledge of the models we used during the previous Milestone.

We were able to understand what features are more important for classification purposes.

Considering the importance of these features we were also able to understand whether our model had some bias or not.

THANKS!

Do you have any questions?