# Predicting H1N1 Vaccination Rates

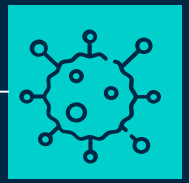
# Meet the Team

Christos Chen

Jasmine Dogu

Brian Wimmer

# TABLE OF CONTENTS



01

## BACKGROUND

General Question & Objectives

02

## HYPOTHESES

Formation

03

## MODELING

Exploration & Tuning

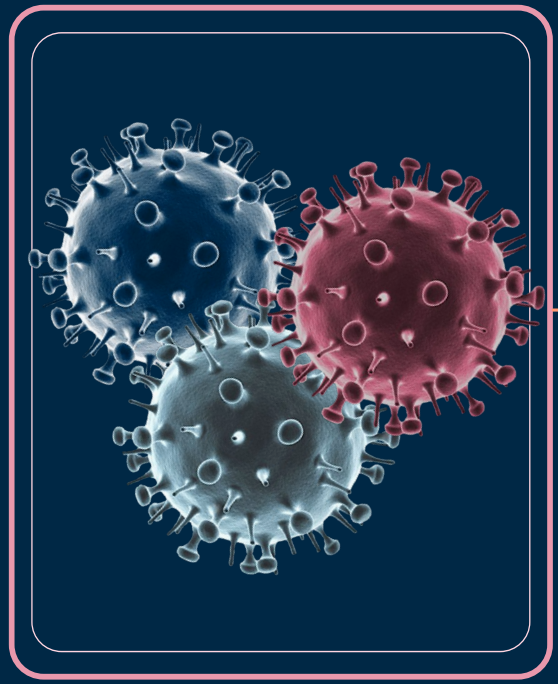
04

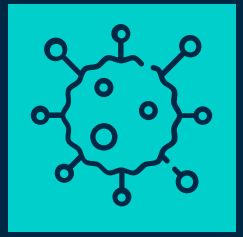
## VALUE

Organizational Benefit & Future Work

# General Question

Will the SVM or the Random Forest Model Predict the Likelihood of a Person Getting the H1N1 Vaccine Better?





# Background

# What is H1N1?

- Subtype of the Influenza A Virus
  - Orthomyxovirus containing the haemagglutinin and neuraminidase glycoproteins
- Symptoms - high fever, sore throat, etc.
- Emerged in 2009 as a novel influenza A virus (Swine Flu)
- Originated in the United States
- More contagious, less existing resistance in the general population

# Hypothesis

# Hypothesis

## Null

The SVM Polynomial Kernel will not perform differently than the Random Forest Classifier Model with regards to its F1 Score

## Alternative

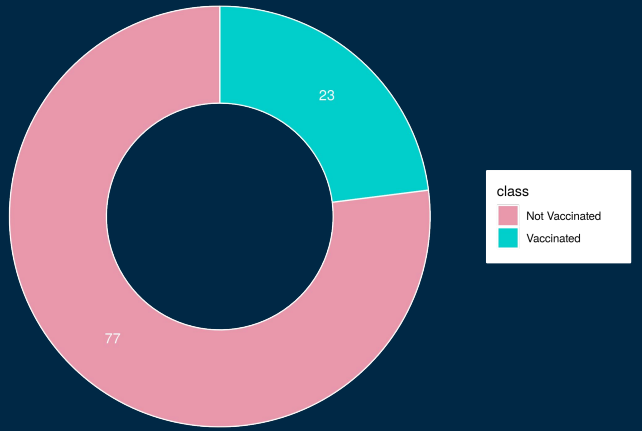
The SVM Polynomial Kernel will perform differently than the Random Forest Classifier Model with regards to its F1 Score

# Model Assessment Metrics

**F1 Score:** A harmonic mean of precision and recall.

Particularly effective:

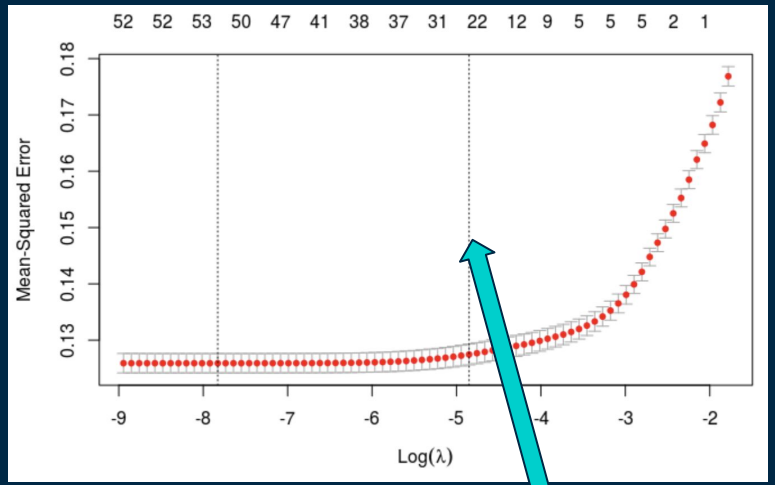
- Intolerance for misclassification
- Imbalanced datasets



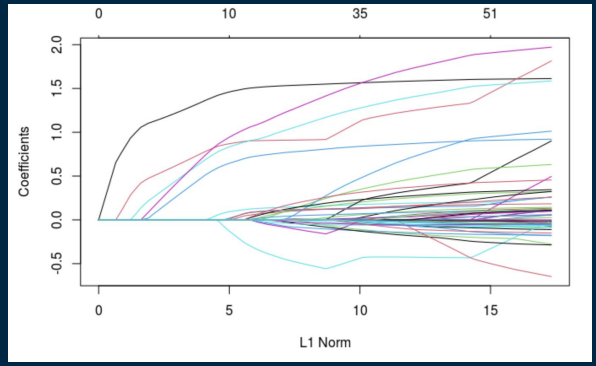

**Kappa:** Performance relative to a randomly-guessing classifier

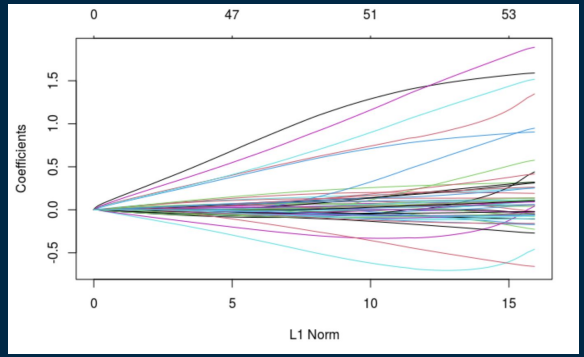
# Modeling

# Lasso Regression



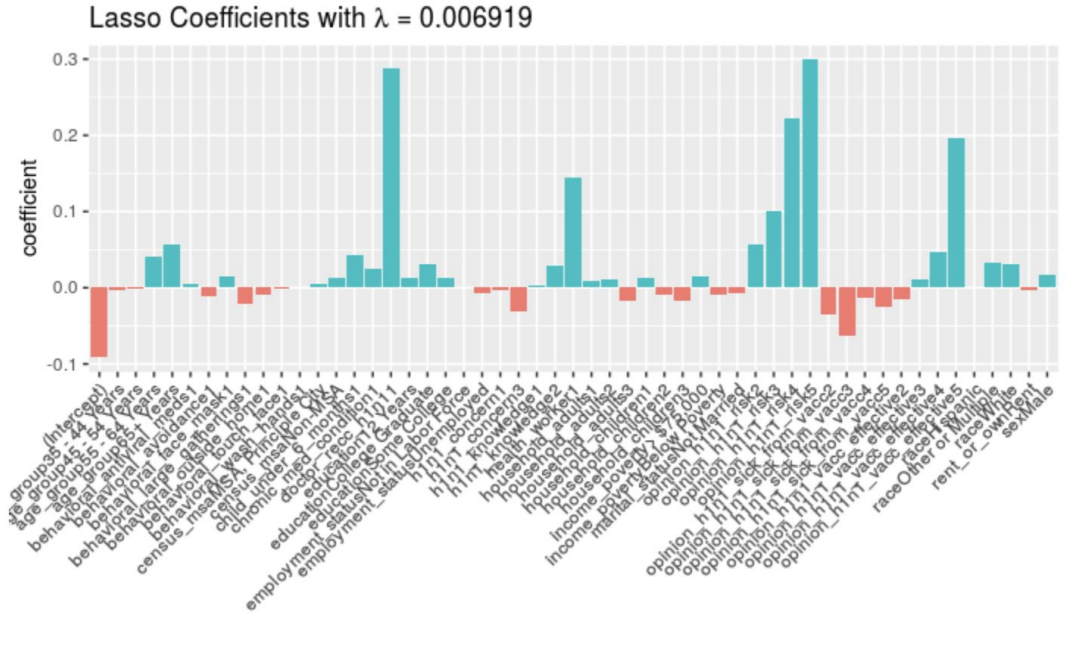

Mean-Squared Error significantly minimized at lambda = 0.006919




*@ lambda = 0.006919*

# Feature Reduction

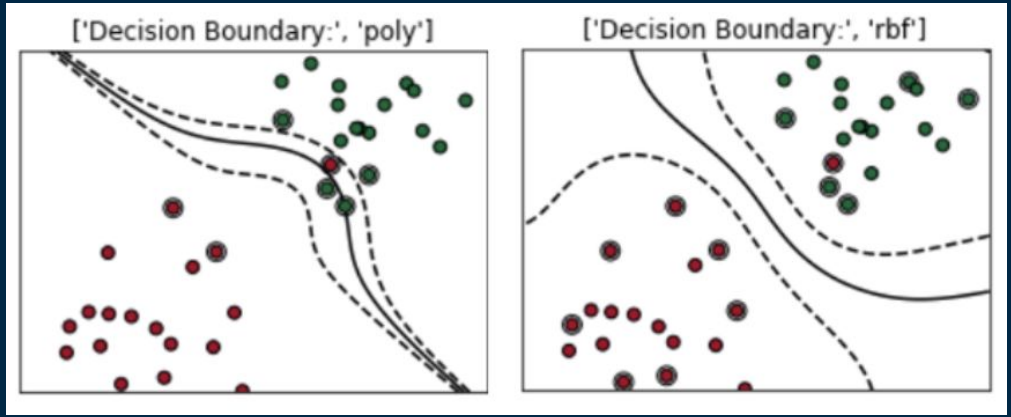



- Removed non-contributing variables
  - **0.3 > Coefficients > -0.3**
- Reduced feature space to **10 distinct variables**

# Support Vector Machine

Explore Polynomial & Radial Kernels

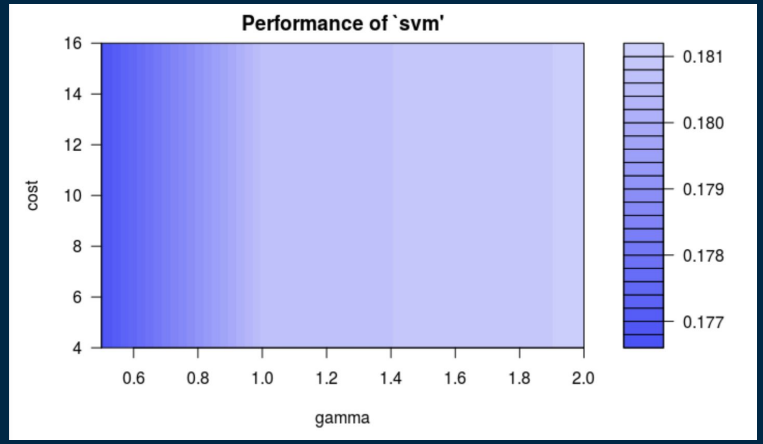
| Kernel | RMSE |
| --- | --- |
| Polynomial | 0.401 |
| Radial | 0.417 |



# Hyperparameter Tuning

**Cost:** The cost/penalty of misclassification

**Gamma:** How quickly the class boundaries dissipate as they get further from support vectors




## Bias-Variance Tradeoff

| | Large Gamma | Small Gamma | Large C | Small C |
|---|---|---|---|---|
| **Variance** | Low | High | High | Low |
| **Bias** | High | Low | Low | High |

C = 8 → *Small C*

Gamma = 0.5 → *Small Gamma*

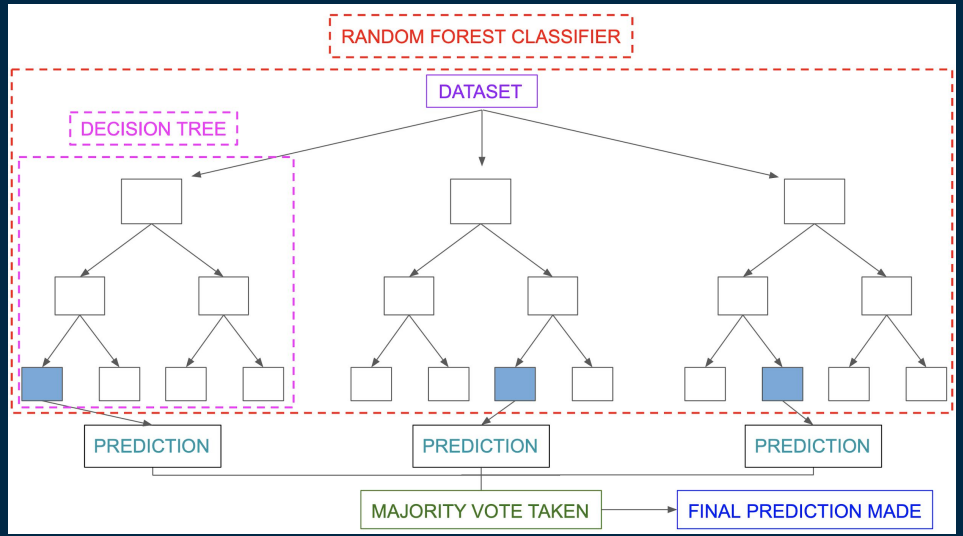
# Support Vector Machine

| | Polynomial Kernel | Radial Kernel | Radial - Feature Reduced | Tuned Radial - Feature Reduced |
|---|---|---|---|---|
| **F1 Score** | 0.1699 | 0.5483 | 0.5734 | 0.6768 |
| **Kappa** | 0.1221 | 0.4529 | 0.4788 | 0.6064 |
| **Specificity** | 0.7832 | 0.8492 | 0.8564 | 0.8779 |
| **Sensitivity** | 0.8376 | 0.7302 | 0.7302 | 0.8778 |
| **Accuracy** | 0.784 | 0.8326 | 0.8375 | 0.8779 |

# Random Forest

| Original Dataset | ntree | mtry |
|---|---|---|
| Initial | 200 | 5 |
| Tuned | 188 | 5 |

| Feature Reduced Dataset | ntree | mtry |
|---|---|---|
| Initial | 200 | 6 |
| Tuned | 104 | 6 |



# Random Forest

| | Random Forest | Tuned Random Forest | Random Forest - Feature Reduced | Tuned Random Forest - Feature Reduced |
|---|---|---|---|---|
| **F1 Score** | 0.8670 | 0.8655 | 0.6646 | 0.6649 |
| **Kappa** | 0.8303 | 0.8285 | 0.5860 | 0.5862 |
| **Specificity** | 0.9453 | 0.9444 | 0.8790 | 0.8791 |
| **Sensitivity** | 0.9298 | 0.9310 | 0.8122 | 0.8116 |
| **Accuracy** | 0.9422 | 0.9417 | 0.8682 | 0.8682 |

# Model Comparison

| | Tuned Radial SVM-Feature Reduced | Tuned Random Forest |
|---|---|---|
| **F1 Score** | 0.6768 | 0.8655 |
| **Kappa** | 0.6064 | 0.8285 |
| **Specificity** | 0.8779 | 0.9444 |
| **Sensitivity** | 0.8778 | 0.9310 |
| **Accuracy** | 0.8779 | 0.9417 |

# Wilcoxon Rank-Sum Test

**Main Assumption**

- Two small independent samples compared

Result Details

*W*-value: 0
Mean Difference: -0.2
Sum of pos. ranks: 0
Sum of neg. ranks: 55

*Z*-value: -2.8031
Mean (*W*): 27.5
Standard Deviation (*W*): 9.81

Sample Size (*N*): 10

| Tuned Radial SVM | Tuned Random Forest |
|---|---|
| 0.6768 | 0.8655 |
| 0.6746 | 0.8725 |
| 0.6672 | 0.8677 |
| 0.6771 | 0.8686 |
| 0.6780 | 0.8671 |
| 0.6893 | 0.8658 |
| 0.6683 | 0.8659 |
| 0.6724 | 0.8680 |
| 0.6712 | 0.8686 |
| 0.6764 | 0.8697 |

*Non-parametric version of a Two-Sample T-Test

# We Reject Our Null Hypothesis

The results are significant at p <0.05

- **The value of *z* is -2.8031.**
- **The *p*-value is .00512.**

# SVM vs. Random Forest

**Support Vector Machine**

- Scales well to high dimensional
- Classes are separable
- Generalization in practice to minimize overfitting
- Kernel trick
- Good for imbalanced data
- Highly efficient and accurate

**Random Forest Classifier**

- Works well with non-linear data
- Good for imbalanced data
- Reduced error
- High amounts of data
- Reduced risk of overfitting
- Non-parametric
- Fast/scalable
- Less parameter tuning

*Support Vector Machines work better with high dimensional data, which did not apply to our situation. Because of Random Forest Classifiers ability to work well with non-linear, imbalanced data (like our dataset), it may have performed better.

# Value

# Potential Use Case

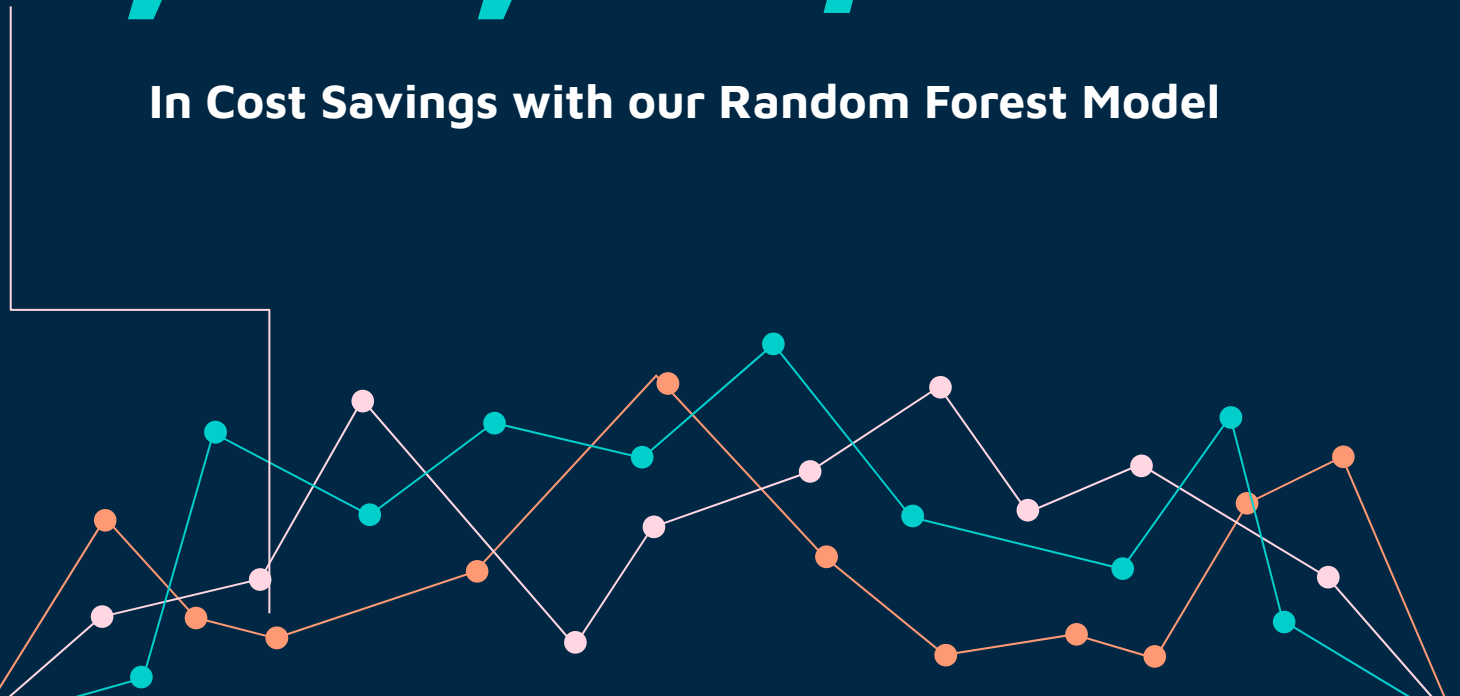
- **Government-based Targeted Marketing Campaign to Encourage Individuals to get Vaccinated**

- **Assumptions:**
  - 500,000 People Outreach/Week
  - Avg Revenue of H1N1 Vaccination is $20
  - Spend $50 marketing to convince those who are believed to not want vaccine
  - Spend $5 marketing vaccine-related resources to those who are believed to want vaccine
  - Population sampled in survey was reflective of U.S. population

| | | Actual | |
|---|---|---|---|
| | | ***Yes Vaccine*** | ***No Vacc*** |
| **Predictions** | ***Yes Vacc*** | $ (15.00) | $ 5.00 |
| | ***No Vacc*** | $ 30.00 | $ 50.00 |

# 2,570,000/week

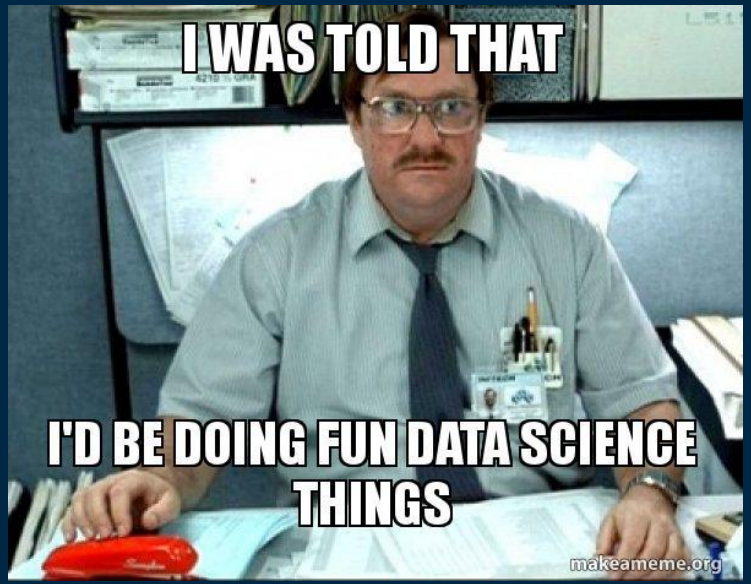
**In Cost Savings with our Random Forest Model**

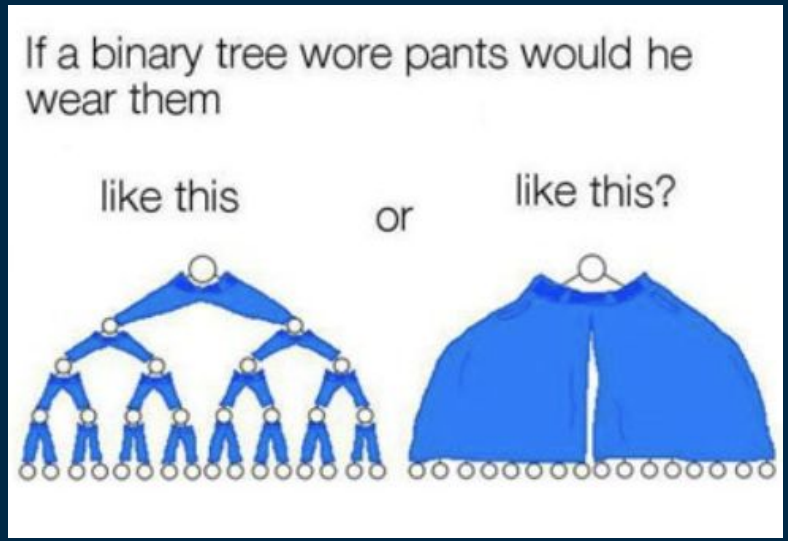


# Future Analysis

| | |
|---|---|
| Other Viruses | More **insight** into factors that make a person more/less likely to get a vaccine (ex. COVID-19) |
| Include Seasonal Data | Develop **trends** between seasonal flu and H1N1 data and assist public health officials with running targeted campaigns |
| Decode Variables | Ability to **identify** encoded geographic, occupation, and employment industry data to access resources and infection risks |
| XGBoost | **Expand** & expound analysis to utilize XGBoost to cross-examine potential overfitting with random forest |

# Thank you!


I WAS TOLD THAT
I'D BE DOING FUN DATA SCIENCE THINGS
makeameme.org


If a binary tree wore pants would he wear them
like this
or
like this?

# Citations


https://towardsdatascience.com/support-vector-machine-simply-explained-fee28eba5496
https://towardsdatascience.com/a-guide-to-svm-parameter-tuning-8bfe6b8a452c
https://medium.com/analytics-vidhya/machine-learning-decision-trees-and-random-forest-classifiers-81422887a544
https://towardsdatascience.com/what-is-out-of-bag-oob-score-in-random-forest-a7fa23d710
ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NIS/nhfs/nhfspuf_DUG.PDF
https://www.cdc.gov/h1n1flu/surveillanceqa.htm
https://towardsdatascience.com/simplifying-precision-recall-and-other-evaluation-metrics-d066b527c6bb
https://www.jihongzhang.org/post/2019-02-19-lasso-regression-with-glmnet/

# Appendix A: Cost Calculations

**Cost Matrix**

| | | Actual: *Yes Vaccine* | Actual: *No Vacc* |
|---|---|---|---|
| | ***Yes Vacc*** | $ (15.00) | $ 5.00 |
| **Predictions** | ***No Vacc*** | $ 30.00 | $ 50.00 |

**Assumptions**

**$5 expended if we think they want vaccine**
**$50 expended if we think they don't want vaccine**
**$20 Revenue if someone gets vaccine**

**Population Matrix**

Assuming population reach of **500,000** /week

| | | Actual: *Yes Vaccine* | Actual: *No Vacc* |
|---|---|---|---|
| | ***Yes Vacc*** | 107069 | 21425 |
| **Predictions** | ***No Vacc*** | 7931 | 363575 |

**COST SAVINGS**

$ 2,570,170 / week

| | | |
|---|---|---|
| ***People*** | **115,000** | **385,000** |
| ***Base Rates*** | 0.23 | 0.77 |
| SUM | 115000 | 385000 |

**Positive Pred Rates**

| | | Actual: *Yes Vaccine* | Actual: *No Vacc* |
|---|---|---|---|
| | ***Yes Vacc*** | 0.931034483 | 0.05564996 |
| **Predictions** | ***No Vacc*** | 0.068965517 | 0.94435004 |
| | | 1 | 1 |

Cost Matrix relative to Status Quo = "predictions that all people don't want the vaccine"

| | | Actual: *Yes Vaccine* | Actual: *No Vacc* |
|---|---|---|---|
| | ***Yes Vacc*** | $ (15.00) | $ (45.00) |
| **Predictions** | ***No Vacc*** | $ - | $ - |

**Confusion Matrix**

| | | *Yes Vaccine* | *No Vacc* |
|---|---|---|---|
| | ***Yes Vacc*** | 1107 | 262 |
| | ***No Vacc*** | 82 | 4446 |