Flu Vaccines Prediction

Jimmy Nguyen

University of San Diego

Master of Science, Applied Data Science

ADS-501-01

SP21

20 February 2021

Business Understanding

Background

Organization:

Our customer, DrivenData Labs, assist "mission-driven organizations" by building solutions through machine learning, unlocking hidden meanings and insights in their data. DrivenData offers their expertise in services of predictive analytics to make a positive impact for their partners. Examples of their services range from optimization in natural language processing, modeling probabilistic pathways of certain events, or developing automation tools in image recognition.

DrivenData states that with projects "involving social impact and where public awareness is an objective" – they tackle problems globally on their crowdsourcing platform, hosting data science competitions for data scientists all over the world. Thus, competitors are able to compete in finding the best solutions through predictive analytics and mathematical models.

Problem Area:

DrivenData is currently hosting an ongoing competition in the public health sector. While vaccines are still under administration logistics with risk factors involved, current data related to vaccinations of COVID-19 patients are not yet available. Therefore, this competition aims to analyze data from "a different recent major respiratory disease pandemic", the H1N1 influenza virus from spring 2009 (DrivenData). DrivenData pointed out that over 150,000 deaths occurred due to this virus also known as the "swine flu". Since then, a vaccine was developed and became publicly available in October 2009. As early of 2010, the current data available in this competition is derived from when "the United States conducted the National 2009 H1N1 Flu Survey" (DrivenData).

In this survey, there were questions asking about the respondents' "social, economic, and demographic background" and their "opinions on risks of illness and vaccine effectiveness" (DrivenData). Also, questions on whether the respondents have received the vaccines as well. DrivenData cited that this data "is provided courtesy of the United States National Center for Health Statistics."

The current status of this project is ongoing until the deadline of March 31, 2022. The prerequisites involve the use of data mining through machine learning algorithms and modeling. This project aims to predict the probabilities whether patients are taking the H1N1 and seasonal flu vaccines. The motivation is to push for vaccinations, as our customer explains that vaccines are "a key public health measure used to fight infectious diseases."

Customer and project (Also see references page for APA-style citations):

- DrivenData - Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines. https://www.drivendata.org/competitions/66/flu-shot-learning/page/210/

Business objectives and success criteria

Business Objectives:

The customer's primary business objective: "Can you predict whether people got H1N1 and seasonal flu vaccines using information they shared about their backgrounds, opinions, and health behaviors?" (DrivenData).

In other words, the customer would like predictive capabilities to improve public health safety based on data from the National 2009 H1N1 Flu Survey (NHFS). We are tasked deliver an analytics solution that can predict probabilities of individuals taking the H1N1 and seasonal vaccines. The customer's satisfaction is dependent on the analytics solution to perform better than a random classifier (better than 50% accuracy).

The business organization also seeks insights from the survey data for data-driven decisions. Listed below is a table of other business goals and questions the customer is interested in.

Goals/Question	How to solve?	Expected Benefits	Attainable?
How do we predict the effect of health-based decisions on vaccinations more accurately?	There are a lot of factors related to this question derived from the dataset. We can hypothesize by looking at the individuals' opinions on the vaccinations from the data. For example, if they think the vaccines are effective or if they think there is a risk to taking the vaccines.	Customer can use these insights to figure out how to better spread public knowledge and announcements on the effectiveness of the next vaccinations. Customer can also efficiently work with hospitals to figure out how to better support individuals who are less likely to take the vaccinations.	Yes
What are the most common characteristics in individuals that do take the vaccinations?	Separate individuals into a group of Yes responses for one vaccination type and analyze the features in the data such as the frequency counts or the average. We can do this through data visualizations.	Customer will be able to better target individuals who are more likely to take vaccinations. This will improve accessibility and efficiency to increase more public health	Yes

		awareness for the less likely group.	
Which factors affects individuals from taking the vaccinations?	Customer will be given insights on why individuals stayed away from vaccinations. More in-depth analysis looking the probable reason related to social-economic background, and knowledge or skepticism on vaccinations.	Customer will find these insights useful to their decision-making on individuals less likely to take vaccines. They will know how to better allocate resources in the right areas to improve overall public health vaccinations.	Yes

Success Criteria:

The primary business objective is to push for more flu vaccinations. The customer's goals are to increase the protection and safety awareness for public health. Thus, an analytics solution that can provide support to public health departments and nudge individuals toward vaccines. This predictive model will help influence the number of vaccinations for flu and future illnesses. Insights we provide will drive efficiency in contacting which individuals that may defer a vaccination. An effective model and a data analysis report will complete the business objectives by the customer's deadline on March 31, 2022, midnight UTC.

Other success criteria will be based on an extensive report and presentation for insights and analysis on the customer's data. Conclusions in statistical testing and storytelling in visualizations will help the customer understand what is going in their data and to drive better decision-making on vaccination inquiries. These fully entails the appropriate business solutions for the customer's business goals and secondary business questions. The customer and their business leaders will make the subjective judgement on our reporting and analysis.

Inventory of resources

Hardware resources:

- Base hardware
    - Personal computer with i5 or i7 CPU and 8 gigabytes of memory RAM and sufficient storage to download datasets.
    - Customer does not provide the hardware for this data mining project

Sources of data and knowledge:

- Our customer has a file called “submission format” for what the expected final outcome should look like: https://www.drivendata.org/competitions/66/flu-shot-learning/page/211/#sub_values
- Data for this project are available to download and is separated into three separate files named: “Training Features”, “Training labels”, “Test Features”. https://www.drivendata.org/competitions/66/flu-shot-learning/data/
- A problem description of the competition
- A data dictionary explaining each feature in the data. https://www.drivendata.org/competitions/66/flu-shot-learning/page/211/#features_list
- An example of how an individual responded to the survey as a single record in the dataset. https://www.drivendata.org/competitions/66/flu-shot-learning/page/211/#features_eg
- A performance metric guideline https://www.drivendata.org/competitions/66/flu-shot-learning/page/211/#metric
- A discussion section on their online crowdsourcing platform for further in-depth discussions with other competitors https://community.drivendata.org/c/flu-shot-learning/33

Output Format:

- Download the datasets here:
- https://www.drivendata.org/competitions/66/flu-shot-learning/data/
- A sample of the dataset will look like the following:

A	B	C	D	E	F	G	H	I	J	K	L	M	N
responde	h1n1_con	h1n1_kno	behaviora	behaviora	behaviora	behaviora	behaviora	behaviora	behaviora	doctor_re	doctor_re	chronic_m	child_und
0	1	0	0	0	0	0	0	1	1	0	0	0	0
1	3	2	0	1	0	1	0	1	1	0	0	0	0
2	1	1	0	1	0	0	0	0	0			1	0
3	1	1	0	1	0	1	1	0	0	0	1	1	0
4	2	1	0	1	0	1	1	0	1	0	0	0	0
5	3	1	0	1	0	1	0	0	1	0	1	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0
7	1	0	0	1	0	1	0	0	1	1	0	1	0
8	0	2	0	1	0	1	1	1	1	0	0	0	0
9	2	1	0	1	0	0	1	0	1	0	0	1	0
10	2	1	0	1	0	1	1	0	0	0	0	1	1
11	1	2	0	1	0	1	0	0	0	0	0	0	0
12	1	1	0	1	0	1	1	0	1	0	1	0	0
13	1	1	0	1	0	1	0	0	1	0	0	1	1
14	3	1	0	1	0	1	0	1	1	0	0	0	0
15	1	1	0	1	0	0	0	0	0	0	1	0	0

Personal Resources:

- Personal computer
  - I5 CPU with 8 gigabytes of memory RAM and 1 terabytes of storage
- Data visualization software applications

    - Tableau
    - Microsoft Excel
- Programming languages solutions
    - Python
    - R-programming

Requirements, assumptions, and constraints

Requirements

Under the customer's rules (DrivenData), this multi-label classification project is hosted by DrivenData until the date of March 31, 2022, midnight UTC. All expected scheduling is by this date only. This is the final deadline for submissions to the customer. However, under future circumstances the deadline to submit business deliverables may be extended. Given the timeline of approximately a year from today (February 12, 2021), a shorter timeline of 7 weeks can be used as an overview of the project in the following schedule of completion.

The schedule of completion can be found here:
https://drive.google.com/file/d/15Fx5mjiyfenLb4HhDUEcQ3hpjiWkvDep/view?usp=sharing

The final business deliverable is a comma separate values (CSV) file under of which contains two columns separated by commas for each row. The values will be the predictions on the probabilities of likelihood for each respondent taking the H1N1 flu and seasonal flu vaccine. This will be the only submission requirement toward the deadline of this project.

The customer will then evaluate the performance of the model based on the prediction outputs. Then the engineering team will apply their private threshold function onto the model's output. This will be used to convert each row of probabilities to output labels of 0s and 1s. Using these output labels, the customer will look at the results of the AUC-ROC curve performance measurement. This is why a requirement is imposed onto the model's final outputs.

All property used and produced in this project/competition including data analysis reports and code will be subject to be licensed under "The MIT License an open-source software license commonly described at http://opensource.org/licenses/MIT)", if the customer deemed our business deliverables to be the winning solutions (DrivenData). Thus, any private sharing of code is not permitted (between teams in the competition).

If chosen as the winner of this competition, there are requirements for documentation on the analytics solutions as well. Solutions are to be documented by using the customer's model documentation template. Under the requirements of the competition, the template is only available when the winner is announced. Therefore, the level of comprehensibility, and maintainability is crucial and should follow universal programming rules and ethics for readability and reproducibility.

There should be clean, well-documented code in the programs. Such as coding conventions for variable names and comments for functions. Directories should be easy to navigate and organized. Outputs for any data visualization should be titled including a short description.

Inclusions for "README" files and a text file for the list of dependencies for the project's virtual environment. We should be able to repeat the same data preparation and modeling process for future data pipelines or deployment. By doing so, if the customer decided to choose our model to be the winner, we can easily transition to the customer's model documentation template.

The project has no submission limits to the customer before the final deadline. However, there is a fixed number for submissions in a day "per-competition basis" (DrivenData).

The use of any external data outside of the project is strictly forbidden. Any use of external data will not be tolerated. The customer's requirements state that we must agree to make no attempt on using any additional data for the completion of this project.

However, the internal data provided by the customer (DrivenData) is for the public use within the realms of this project/competition. Any other use outside of the competition is not permitted such as for the purpose of applying algorithms to re-identify the respondents who participated in the survey (dataset). This violates all codes of privacy and may be subjected to legal action(s).

Under the governing law and jurisdiction, "DrivenData shall be the sole interpreter of these official rules" (DrivenData). DrivenData states that if there are any disputes or claim in the matter of this competition, it shall "governed by and construed in accordance with the internal laws of the Commonwealth of Massachusetts." Thus, any legal suit arising from the competition will be subject accordingly in the federal courts of the United States. *Governing Law and Jurisdiction: https://www.drivendata.org/competitions/66/flu-shot-learning/rules/*

Assumptions

This dataset came directly from the "National 2009 H1N1 Flu Survey (NHFS)" (DrivenData). The NHFS was a compiled list of "random-digit-dialing telephone survey of households" for the purpose of monitor the H1N1 immunization records (DrivenData). Thus, the timestamp for the snapshot of this dataset is between October 2009 and June 2010. For each phone call, it would be a one-time session of survey questions asking for the respondents' personal information and opinions on the flu vaccinations.

However, at the time when we received the dataset, the customer stated that the targeted population was all people in the U.S. that were 6 months or older. However, during data exploration, we discovered a column/survey question in the dataset specifying the age groups for a respondent to belong in, which ranges from 18-65+ years old only. Therefore, there should be no age groups under 18 years old in the data. This is our first assumption. We assumed that the target population is actually adults 18 years old and above, living in the United States at the time of which the survey was conducted.

The following table below contains information about the columns we made assumptions with during data exploration and mining.

Assumptions	Descriptions
Antiviral Medicine	We assumed that during the data collection phase, respondents understand what antiviral medications are and if they are aware of their own prescriptions
Avoidance with flu-like symptoms	We assumed this information is accurate to put trust in respondent's ability on knowing what flu-like symptoms are
Washing hands	We assumed by the definition of 'frequent' universally should weigh the same. For example, washing their hands once every hour.
Large gatherings	We assumed the number of large gatherings is coming in contact with more than 20 people in a setting.
Doctor recommendations for H1N1 flu or Seasonal vaccine	We assumed that by recommendations they were or were not directly recommended by a doctor in person and not by any other form of recommendations such as indirect advertised emails. We assumed full acknowledgement from the respondent at the time of the recommendation.
Children under 6 months	We assumed what this question is asking is if they come into contact of any child that is under 6 months, however this is most notably easier to know if they are a guardian of a child under 6 months. So, we assumed this is more likely asking if they had a child under 6 months during the time.
Health-care worker	The question asks if the respondent is a healthcare worker. In this case, we made assumptions for the definition of who is considered a health care worker. "A healthcare worker is one delivers care and services to the sick and ailing either directly as doctors and nurses or indirectly as aides, helpers, laboratory technician" (Joseph). By this definition, we assumed a janitor working in a hospital is most definitely not a health-care worker. Joseph, B., & Joseph, M. (2016). The health of the healthcare workers. *Indian journal of occupational and environmental medicine*, *20*(2), 71–72. https://doi.org/10.4103/0019-5278.197518
Opinions about H1N1 or Seasonal flu vaccine's effectiveness	We assumed this originates from what they heard from family and friends which spread news about the vaccines

Age group	Since we are given a range of possible values from 18-65+, we can make assumptions that this dataset only has those age groups. However, if we were to find other possible values outside of these age groups such as those who were under 18 may be deemed dismissible since this is an error in data entry.
Education	The possible range values here can easily confuse the insights about the respondents. Such as what does 12 years of education mean? Does this mean they finish with a high school diploma or did they have to go to adult school for their G.E.D? We assume 12 years of education is the traditional high school completion pathway.
Race	The possible range for values is 'White', 'Black', 'Hispanic', and 'Other or Multiple'. We will make assumptions for Asians or Native Americans to be under 'Other or Multiple.' This is an example of unfairness in choices. As respondents may feel indifferent about being label as 'Other'
Income Poverty	Here we have a value '<= $75,000. Above poverty' and this is confusing to define because depending on the household size, some people may not see this as above poverty. And what is the actual cut off value that will set us to be above poverty? Instead of less than $75,000, we should have value explicitly stating <$40,000 is poverty.
Marital Status	We will make assumptions that anything 'Not Married' is divorced/widowed.
Rent or own	Is it possible to assume values that are neither? Can there be a respondent who was 'homeless' at the time? Rent or Own may be too narrow
Household adults	We will make assumptions that adults in this case are at least the legal age of 18 years old.
Household children	We will make assumptions that children in this case are under 18 years old.

Constraints

- Legal constraints - data contain identifiable information about the respondents in the survey that may be traceable and leaked. (This is further elaborated under the ethical review section)
- No budget constraints
- Timescale constraint – the deadline is March 31, 2022, midnight UTC

- No rights to data sources constraint – data is currently hosted publicly by the customer (DrivenData) with no passwords or limitations on downloads.
- No technical accessibility of data constraints – data is already shared by the customer as a comma separated values (CSV) file. This file can be read through Microsoft's Excel or by a programming language that supports reading and writing CSV files such as Python or R.
- No relevant knowledge constraints – there is a discussion board among coworkers and the customer if any questions arise

The legal issue will now be addressed below using the RESOLVEDD strategy.

Ethical Review (RESOLVEDD Strategy)

1. Review

Based on a list of survey questions, we have a variety of responses that draws the line between being useful to a predictive model to generalizing people to a certain group. There were some questions on the survey that asks specifically about the respondents' opinions on the vaccines or about their level of concerns for the flu types. This will serve as great data to help build a predictive model. Aside from general demographics questions like age groups and sex, we learned about the other survey questions that may be too personal about the respondent.

2. Estimate

Other questions involved very specific details about the respondent. Such as their medical history of any chronic medical conditions or asking about their finances such as current employment status and level of income.

The following questions (features) listed below may be detrimental to modeling:

- Has chronic medical condition?
- Has health insurance?
- Race?
- Education?
- Marital Status?
- Current employment status?
- Employment industry?
- Employment occupation?
- Income level?
- Rent or own property?
- Residence geographic region?

We should be aware of algorithms and their capacity for possible re-identification purposes. Meaning, the survey questions raise risks for privacy protection on original respondents who participated. This will break privacy laws including the HIPAA compliance with the health-related question ('has chronic medical conditions?'). Not to mention, these survey questions

contain limited choices where respondents may feel forced to fit into a category. For example, a survey question that asks about their education level with very few choices to pick from such as: "<12 Years", "12 Years", "Some College", "College Graduate". These choices fail to capture any real insight for modeling. A choice such as less than 12 years of school produces a lot of vagueness. This could represent respondents who may be a high school drop-out, repeated the same grade or was homeschooled.

Another example is the survey question asking about race ('Race'). This survey question only gives respondents the choices of "White", "Black", "Hispanic", and "Other or Multiple". While in fact, Asian and Native Americans should have their own groups listed. Thus, a question like so with absent-minded choices may make the respondents feel indifference. Such an effect may lead to more missing answers (pass/skip) or false answers. This in turn will be the result of poor data quality.

Existing possible questions will in fact do more harm than good. We cannot sacrifice for accuracy in a predictive model by asking sensitive information from respondents in a survey. This addresses the ethical problem on the principle of confidentiality.

3. Solutions
    a. Redesign more responsible survey questions that provide total coverage of all choices for example – respondents may not answer the question about their sex if the only options are male or female when they identify as transgender.

    b. Feature Engineering/Selection – derived new features based on understanding of ethical concerns or replace values to mask (de-identify) sensitive information. For example, create a new education column called highest level of education. The new values can be mapped from the original education column such as "<12 years" to "less than high school" and "12 years" to "high school diploma". Then select a subset of these de-identified features in the dataset for modeling. This will regulate any unethical data that causes a leak in privacy or bias in modeling.
4. Outcomes
    a. The first solution does not fix the current problems in this project. Rather, it addresses a possible approach for future surveys to be conducted on a new set of vaccines. So, moving forward with this solution will only be a waste of time and resources. In the end, the customer may not be able to receive the business deliverable or may be involved in potential legal lawsuits with ethical concerns from the respondents' sensitive information.
    b. We will be able to continue into data preparation as there is no need to throw away any data that may be useful to our models. Rather, this solution will protect respondents' privacy and improve the model's accuracy. This solution will provide richness in data if this solution is done correctly to preserve the respondents' identities.

5. Likely Impact
    a. The first solution will only provide a more consistent selection of choices to each survey question during the data collection phase. However, it is not the immediate solution to be implemented right now for this project. This solution will only be beneficial to future projects with similar goals.
    b. Previous data with ethical concerns will now be transformed into useful features without compromise as long as we do the necessary steps to de-identify any trace-backs to the original respondents' identity. Without this solution, we will be left with only half of the features available to train our models on. However, the impact of this solution may need an extension on the project for any failed attempt during data preparation. Since data cleaning and wrangling to derive new features may take some time to perform.
6. Values
    a. The first solution will not be harm by any violations as it is an efficient approach to data collection on future projects.
    b. The second solution will be delaying the modeling phase to a further date, since this solution will require more time in the data preparation stage. Therefore, no violation will occur as this solution only aims to solve the data privacy issue.
7. Evaluate
    a. This solution should not be considered as it is currently irrelevant and does not offer any positive effects for the current project.
    b. The second solution serves to be the best due to its alignment with the business goals and success with the business deliverable for the customer.
8. Decide
    a. The first solution fails to address any realistic approach that will be useful toward the business deliverable. If we were building a predictive model with this solution, we will waste more valuable time and proceed to violate data privacy laws for negligence to the unethical data usage.
    b. The second solution involves feature engineering and is best selection out of the two proposed. Since some of the features we currently have raises ethical concerns, throwing away these data will cause an inconvenience in modeling. The throwaway data will harm the models further as they do not have enough information to train on. This will most likely be underperforming against competitors. Therefore, the second solution should be chosen because it will boost the model's accuracy and protect the respondents' identities too.
9. Defend
    a. The second solution is most efficient and ideal approach. We can preserve the insights from the original features and reduce the high dimensionality within feature engineering. This solution also serves to protect the respondents' privacy without throwing away meaningful data that can help with modeling. Thus, the customer will be satisfied with a highly accurate model that can make predictions on flu vaccination likelihoods with de-identified data. The second solution is the best choice.

Risks and contingencies

Risks	Details	Contingency Plan
Business	• Competitors will most likely generate a model giving more accurate predictions and higher performance. • Deadline: March 31, 2022	• Communication and collaboration with customer for more in-depth domain knowledge on data – giving rise to better feature engineering and selection. Discussion Board: https://community.drivendata.org/c/flu-shot-learning/33 • Research and experiment hyper-parameters selection and fine-tuning. • Have a selection of different models as back-up. • Although the competition deadline is not aligned with the class's final project deadline, time outside the class may be needed for improving performance if winning the competition is a goal.
Organizational	• N/A	• N/A
Financial	• N/A	• N/A
Technical	• The model does not output the required format of predictions	• Debug code to output CSV file formatted with the required columns. • Debug results to output multilabel probabilities. https://scikit-learn.org/stable/modules/multiclass.html • Research the selected model for its specified outputs. http://scikit.ml/ • More research on how the algorithm is implemented and mathematically how it works. • Make sure model is training on train data and targeting the correct target variables.

Data	• Poor data quality – missing values or incorrect data entries, data may not be relevant	• Discussion board with customer on what some data value means or what it was supposed to be. https://community.drivendata.org/c/flu-shot-learning/33 • Manually encode errors to the right values • Impute missing values by calculating the mean of the observed values for that variable – other methods include clustering or regression algorithms. • Simply exclude missing values. • Feature engineering methods for e.g., one hot dummy encoding.

Terminology

Business terminology

(Terms here are found under Merriam-Webster dictionary)

- H1N1
  - "a virus that is a subtype (H1N1) of the orthomyxovirus (species *Influenza A virus* of the genus *Influenza virus A*) causing influenza A, that infects birds, pigs, and humans, and that includes strains which may occur in seasonal epidemics or sometimes pandemics" (Merriam-Webster).
- Immunization
  - **"**the production of immunity in a living organism against a disease or pathogenic agent" (Merriam-Webster).
- Infectious Disease
  - "a disease (such as influenza, malaria, meningitis, rabies, or tetanus) caused by the entrance into the body of pathogenic agents or microorganisms (such as bacteria, viruses, protozoans, or fungi) which grow and multiply there" (Merriam-Webster).
- Public Health
  - "the art and science dealing with the protection and improvement of community health by organized community effort and including preventive medicine and sanitary and social science" (Merriam-Webster).
- Vaccine
  - "a preparation of killed microorganisms, living attenuated organisms, or living fully virulent organisms that is administered to produce or artificially increase immunity to a particular disease" (Merriam-Webster).

Data Mining Terminology

- All variables in data are defined and provided by the customer as a data dictionary.
    - This can be found in the following link: https://www.drivendata.org/competitions/66/flu-shot-learning/page/211/#features_list
- Area under curve (AUC) – performance evaluation metric for model
    - Full definition can be found here: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc#AUC
- Receiver operating characteristic curve (ROC) - Performance evaluation metric for mode
    - Full definition can be found here: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc#roc-curve

Costs and benefits

Data mining goals and success criteria

Data Mining Goals:

We are tasked with building a multi-label classifier that makes predictions on data from the National 2009 H1N1 Flu Survey (NHFS). This problem currently acts as a classification problem due to its multilabel output nature. This model would essentially predict the likelihood probabilities on individuals that will take their H1N1 and seasonal flu vaccines. Thus, if the customer is satisfied with its performance, this model will serve as a template for making predictions on future vaccines, such as COVID-19. However, the model we deliver will need to perform better than a random classifier (> 50% accuracy) and outperform the customer's competitors.

The data mining goal is a comma-separated values (CSV) file formatted to have three columns, two of which the model will be responsible for making predictions for. This CSV file will have the following three columns: *respondent_id, h1n1_vaccine,* and *seasonal_vaccine*.

Here is a sample of what it should look like:

```
respondent_id,h1n1_vaccine,seasonal_vaccine
26707,0.5,0.7
26708,0.5,0.7
26709,0.5,0.7
26710,0.5,0.7
26711,0.5,0.7
...
```

For everyone by *respondent_id* (each row) will have exactly two float probabilities predicted from the model. The prediction each probability will be between the range of 0.0 and 1.0.

In the end, the model will be given a sample test data to make new predictions and be evaluated on its performance. The completion of this project depends on a successful model that can output accurate multilabel classifications. Thus, if successful the customer will use the model on future vaccinations.

In Summary:

- Use only the data provided by the customer from the National 2009 H1N1 Flu Survey
- Build a multilabel classifier that performs better than 50% accuracy from a random classifier.
    - The model will need to output probabilities of individuals taking the H1N1 and seasonal flu vaccines
- Outputs are required to be in a comma-separated values (CSV) file.

Success Criteria:

The engineering team of the customer's business organization will be the judge for the performance and effectiveness of the model. An evaluation metric based on a universal performance metric called the area under the receiver operating characteristic curve (AUC ROC). The higher the result, the better it performs. This metric represents the performance of a classification model at various thresholds, plotting two parameters: True positive rate and false positive rate. The higher the AUC, the better it performs as a classifier that distinguishes separability between classes. This means an AUC near 1 is a strong indicator of performance, while the opposite near 0. However, when AUC is 0.5, then the model does not have the capability to distinguish the classes and is predicting at random for every data point.

This evaluation metric will score the predictions on each of the two target variables. Thus, the mean of the two scores will be the final score. Objectively, the model should aim to perform better than the customers' competitors. The customer stated that there are other competitors actively building models to attain the highest level of performance. Thus, the model will need to have the highest AUC-ROC score which indicates the best performance. Otherwise, the customer will lose to its competitors. Lastly, the customer will be the one to assess the success criteria with this evaluation metric on our model.

Project plan

- Total project duration: 7 weeks or 35 business days

Order of Tasks

- Please refer to project schedule of completion below: https://drive.google.com/file/d/15Fx5mjiyfenLb4HhDUEcQ3hpjiWkvDep/view?usp=sharing

Initial assessment of tools and techniques

Data Understanding

Initial data collection report

Files:

Our customer has provided us with 3 sets of data.

1. Training Features
2. Training Labels
3. Test Features

We are able to download and use all 3 datasets for analytical purposes in this project.

The location of these datasets can be found here in this link:
https://www.drivendata.org/competitions/66/flu-shot-learning/data/

"Training Features" dataset consists of 36 columns with 26,707records

"Training Labels" dataset consists of 3 columns with 26,707 records

"Test Features" dataset consists 36 columns with 26,708 records

In the beginning of the project, as part of the data preparation phase of the CRISP-DM framework, we should focus on using the following two datasets: "Training Feature" and "Training Labels". Therefore, we can have a look at "Test Features" which acts as a test dataset later after data exploration with the training dataset.

The reason why is because we would need to set up our training dataset in the format of an analytics base table (ABT) for our models to "learn" before applying into a test dataset.

Therefore, we should perform an operation where we are to join the "Training Features" dataset and "Training Labels" dataset to merge as one.

We can do so if working in Python, through an inner join function based on the "respondent_id" column which acts as a unique and random identifier.

After a successful joining, we get a grand total of 38 columns, with 36 descriptive features and 2 target variables.

This merger allowed us to see which individual receive their vaccines in order to train our model(s) under supervised machine learning.

We have encountered no problems and the inner join function works as intended.

Data description report

Volumetric analysis of Data

- Identify Data and Method of Capture:
    - Data was collected from the National 2009 H1N1 Flu Survey (NHFS).

    - NHFS stated that this data "was sponsored by the National Center for Immunization and Respiratory Diseases (NCIRD) and conducted jointly by NCIRD and the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC). The NHFS was a list-assisted random-digit-dialing telephone survey of households, designed to monitor influenza immunization coverage in the 2009-10 season" (Driven Data).

    - The targeted respondents for the data were "all persons 6 months or older living in the United States at the time of the interview. Data from the NHFS were used to produce timely estimates of vaccination coverage rates for both the monovalent pH1N1 and trivalent seasonal influenza vaccines" (NHFS Readme File). ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/nis/nhfs/nhfspuf_readme.txt
- Access Data Sources
    - Data use restrictions are listed below by the customer, DrivenData.
        - "Use the data in these data files for statistical reporting and analysis only."
        - "Make no use of the identity of any person or establishment discovered inadvertently and advise the Director, NCHS, of any such discovery (1 (800) 232-4636)."
        - "Not link these data files with individually identifiable data from other NCHS or non-NCHS data files." https://www.drivendata.org/competitions/66/flu-shot-learning/page/213/
- Report Tables and Their Relations
    - Three tables total: *Training Features, Training Labels,* and *Test Features*
    - *Training Features* data has a relationship with *Training Labels* data based on the column *Respondent_id.*
    - *Test Features* may be use as test data for model to make predictions on but has no relations with the previous two datasets.

Attribute Types and Values

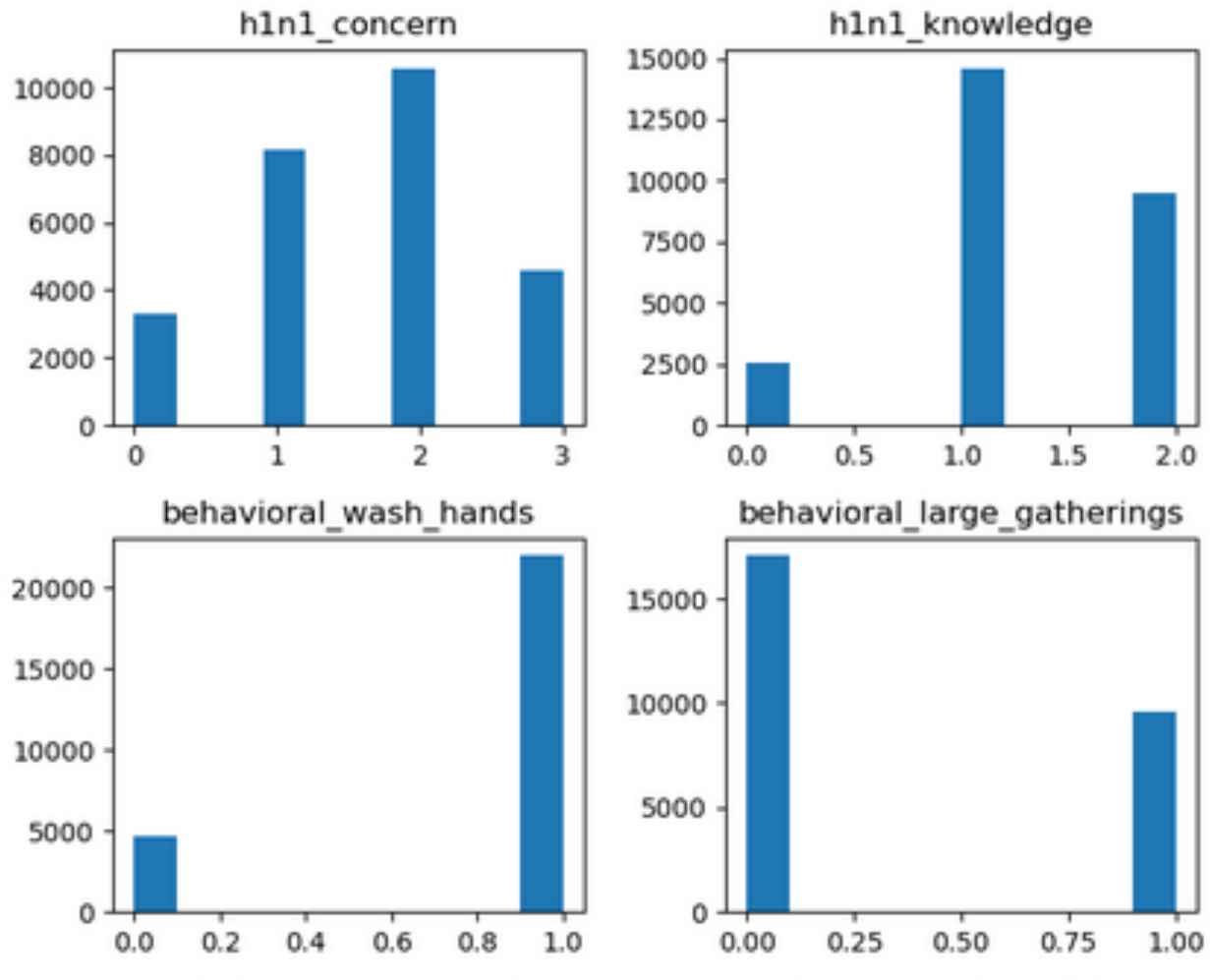
- Check accessibility and availability of attributes
    - All attributes are publicly accessible, and the availability is ready to be analyzed

- However, we did a quality check of missing values in the dataset. Here is what we found:

	Percentage of NA
employment_occupation	50.436
employment_industry	49.912
health_insurance	45.958
income_poverty	16.561
doctor_recc_h1n1	8.088
doctor_recc_seasonal	8.088
rent_or_own	7.646
employment_status	5.478
marital_status	5.272
education	5.268
chronic_med_condition	3.636

- Columns: “employment_occupuation, employment_industry, and health_insurance” contains nearly half (50%) its values missing or marked as ‘NA’ Thus we will need to devise a project plan later on how to deal with missing values, such as imputing them with means or using other machine learning algorithms to fill them in.

- Here is a sample of bar graphs for some of the nominal/ordinal categorical features in our dataset. For each feature, we plotted the response values on the x-axis and the number of responses on the y-axis. For an example, looking at the column ‘h1n1_concern’, this column represents the level of concerns respondents felt about the H1N1 flu. The value ranges from 0 as “not at all concerned”, 1 as “not very concerned”, 2 as “somewhat concerned”, and 3 as “very concerned”.



After looking at the bar graph, we can calculate the percentage of each value in that column to understand how to intrepret it into business terms.

For example, the column “h1n1_concern”, 39.7% or 40% of the respondents are ‘Somewhat concerned’ about the H1N1 flu. This insight will help us build a model weighing the importance of features like this.

	h1n1_concern
0.0	0.124
1.0	0.306
2.0	0.397
3.0	0.172

Here is an attributes table containing variable types, value ranges, business meaning, basic statistic computation and assess relevancy according to data mining goals.

Name of Attribute	Attribute Type	Attribute Value Range	Meaning in Business Terms	Basic Statistics Computation	Relevant?
h1n1_concern	Ordinal Categorial Numeric	0 = Not at all concerned; 1 = Not very concerned; 2=Somewhat concerned; 3 = Very concerned.	Level of concern about the H1N1 flu	Percentages by Value 2.0 0.397 1.0 0.306 3.0 0.172 0.0 0.124 Respondents are often 'Somewhat concerned' about the H1N1 flu	Yes
h1n1_knowledge	Ordinal Categorical Numeric	0 = No knowledge; 1 = A little knowledge; 2 = A lot of knowledge.	Level of knowledge about H1N1 flu.	1.0 0.549 2.0 0.357 0.0 0.094 Half of the respondents has 'a little knowledge' of the H1N1 Flu	Yes
behavioral_antiviral_meds	Nominal Categorical Numeric	0 = No; 1 = Yes.	Has taken antiviral medications.	0.0 0.951 1.0 0.049 95% of the respondents have "not" taken antiviral medications	Yes
behavioral_avoidance	Nominal Categorical Numeric	0 = No; 1 = Yes.	Has avoided close contact with others with flu-like symptoms.	1.0 0.726 0.0 0.274 73% of the respondents has 'avoided close contact with others with flu-like symptoms.'	Yes

behavioral_face_mask	Nominal Categorical Numeric	0 = No; 1 = Yes.	Has bought a face mask	0.0 0.931 1.0 0.069 93% of the respondents have 'not' bought a face mask.	Yes
behavioral_wash_hands	Nominal Categorical Numeric	0 = No; 1 = Yes.	Has frequently washed hands or used hand sanitizer	1.0 0.826 0.0 0.174 83% of the respondents have 'frequently washed hands or used hand sanitizer.'	Yes
behavioral_large_gatherings	Nominal Categorical Numeric	0 = No; 1 = Yes.	Has reduced time at large gatherings	0.0 0.641 1.0 0.359 64% of the respondents have 'not' reduced time at large gatherings	Yes
behavioral_outside_home	Nominal Categorical Numeric	0 = No; 1 = Yes.	Has reduced contact with people outside of own household	0.0 0.663 1.0 0.337 66% of the respondents have 'not' reduced contact with people outside of own household.	Yes
behavioral_touch_face	Nominal Categorical Numeric	0 = No; 1 = Yes.	Has avoided touching eyes, nose, or mouth	1.0 0.677 0.0 0.323 68% of the respondents have 'avoided touching eyes, nose, or mouth.'	Yes
doctor_recc_h1n1	Nominal Categorical Numeric	0 = No; 1 = Yes.	H1N1 flu vaccine was recommended by doctor	0.0 0.78 1.0 0.22	Yes

				78% of the respondents were ‘not’ recommended by doctor to take the H1N1 flu vaccine.	
doctor_recc_seasonal	Nominal Categorical Numeric	0 = No; 1 = Yes.	Seasonal flu vaccine was recommended by doctor	0.0 0.67 1.0 0.33 67% of the respondents were ‘not’ recommended by doctor to take the seasonal flu vaccine.	Yes
chronic_med_condition	Nominal Categorical Numeric	0 = No; 1 = Yes.	Has any of the following chronic medical conditions: asthma or other lung condition, diabetes, a heart condition, a kidney condition, sickle cell anemia or other anemia, a neurological or neuromuscular condition, a liver condition, or a weakened immune system caused by a chronic illness or by medicines taken for a chronic illness.	0.0 0.717 1.0 0.283 72 % of the respondents reported that they did ‘not’ have any of the chronic medical condition.	Yes
child_under_6_months	Nominal Categorical Numeric	0 = No; 1 = Yes.	Reg. Contact with child under the age of 6 months.	0.0 0.917 1.0 0.083 92 % of the respondents does ‘not’ have regular contact with child under the age of 6 months.	Yes

health_worker	Nominal Categorical Numeric	0 = No; 1 = Yes.	Is a healthcare worker	0.0 0.888 1.0 0.112 89% of the respondents are ‘not’ a healthcare worker	Yes
health_insurance	Nominal Categorical Numeric	0 = No; 1 = Yes.	Has health insurance	1.0 0.88 0.0 0.12 88% of the respondents have ‘health insurance’	Yes
opinion_h1n1_vacc_effective	Ordinal Categorial Numeric	1 = Not at all effective; 2 = Not very effective; 3 = Don't know; 4 = Somewhat effective; 5 = Very effective.	Respondent's opinion about H1N1 vaccine effectiveness.	4.0 0.444 5.0 0.272 3.0 0.179 2.0 0.071 1.0 0.034 Only 44% of the respondents believes H1N1 vaccines are ‘somewhat effective’	Yes
opinion_h1n1_risk	Ordinal Categorial Numeric	1 = Very Low; 2 = Somewhat low; 3 = Don't know; 4 = Somewhat high; 5 = Very high	Respondent's opinion about risk of getting sick with H1N1 flu without vaccine	2.0 0.377 1.0 0.309 4.0 0.205 5.0 0.066 3.0 0.042 Majority of the respondents believe they are at a ‘somewhat low’ or ‘very low’ risk of getting sick with	Yes

				H1N1 without H1N1 vaccine.	
opinion_h1n1_sick_from_vacc	Ordinal Categorial Numeric	1 = Not at all worried; 2 = Not very worried; 3 = Don't know; 4 = Somewhat worried; 5 = Very worried	Respondent's worry of getting sick from taking H1N1 vaccine	2.0 0.347 1.0 0.342 4.0 0.222 5.0 0.083 3.0 0.006 Majority of the respondents are 'not very worried' or 'not at all worried' of getting sick from taking the H1N1 vaccine.	Yes
opinion_seas_vacc_effective	Ordinal Categorial Numeric	1 = Not at all effective; 2 = Not very effective; 3 = Don't know; 4 = Somewhat effective; 5 = Very effective.	Respondent's opinion about seasonal flu vaccine effectiveness	4.0 0.443 5.0 0.380 2.0 0.084 1.0 0.047 3.0 0.046 Majority of the respondents believes the seasonal vaccines are 'somewhat effective' or 'very effective'	Yes
Opinion_seas_risk	Ordinal Categorial Numeric	1 = Very Low; 2 = Somewhat low; 3 = Don't know;	Respondent's opinion about risk of getting sick with seasonal flu without vaccine	2.0 0.342 4.0 0.291 1.0 0.228 5.0 0.113 3.0 0.026	Yes

		4 = Somewhat high; 5 = Very high.		Respondents' opinions vary between the risk of getting seasonal flu sick 'with' or 'without' seasonal flu vaccine	
opinion_seas_sick_from_vacc	Ordinal Categorial Numeric	1 = Not at all worried; 2 = Not very worried; 3 = Don't know; 4 = Somewhat worried; 5 = Very worried.	Respondent's worry of getting sick from taking seasonal flu vaccine.	1.0 0.454 2.0 0.292 4.0 0.185 5.0 0.066 3.0 0.004 Majority of the respondents are 'not at all worried' or 'not very worried' of getting sick from taking seasonal flu vaccine.	Yes
age_group	Ordinal Categorical Strings	18-34 Years 35-44 Years 45-54 Years 55-64 Years 65+ Years	Age group of respondents.	65+ Years 0.256 55 - 64 Years 0.208 45 - 54 Years 0.196 18 - 34 Years 0.195 35 - 44 Years 0.144 Our respondents are somewhat skewed towards the elderly age group of '65+ years old.'	Yes
education	Ordinal Categorical Strings	< 12 Years 12 Years College Graduate	Self-reported education level	College Graduate 0.399 Some College 0.278 12 Years 0.229	Maybe

		Some Graduate		< 12 Years 0.093 40% of the respondents are 'college graduates'	
race	Nominal Categorical Strings	White Black Hispanic Other or Multiple	Race of respondent	White 0.795 Black 0.079 Hispanic 0.066 Other or Multiple 0.060 Majority of the respondents in this survey are 'White'	Maybe
sex	Nominal Categorical Strings	Male Female	Sex of respondent	Female 0.594 Male 0.406 59% of 'females' responded in this survey	Maybe
Income_poverty	Ordinal Categorical Strings	> $75,000 <= $75,000, Above Poverty Below Poverty 0.121	Household annual income of respondent with respect to 2008 Census poverty thresholds	<= $75,000, Above Poverty 0.573 > $75,000 0.306 Below Poverty 0.121 57 % of the respondents are considered 'Above poverty'	Maybe
Martial_status	Nominal Categorical Strings	Not Married Married	Marital status of respondent	Married 0.536 Not Married 0.464 54% of the respondents are 'married'	Maybe

Rent_or_own	Nominal Categorical Strings	Own Rent	Housing situation of respondent	Own 0.76 Rent 0.24 76% of the respondents ‘own a house’	Maybe
Employment_status	Nominal Categorical Strings	Unemployed Employed Not in Labor Force	Employment status of respondent	Employed 0.537 Not in Labor Force 0.405 Unemployed 0.058 53% of the respondents are ‘employed’	Yes
Hhs_geo_region	Nominal Categorical Strings	Data values are abbrev. strings	Respondent's residence using a 10-region geographic classification defined by the U.S. Dept. of Health and Human Services. Values are represented as short random character strings.	lzgpxyit 0.161 fpwskwrf 0.122 qufhixun 0.116 oxchjgsf 0.107 kbazzjca 0.107 bhuqouqj 0.107 mlyzmhmf 0.084 lrircsnp 0.078 atmpeygn 0.076 dqpwygqj 0.042 Current data values as abbreviated strings are not helpful in determining any insights	No

Census_msa	Nominal Categorical Strings	Non-MSA MSA Not Principal City MSA Principal City	Respondent's residence within metropolitan statistical areas (MSA) as defined by the U.S. Census.	MSA, Not Principal City 0.436 MSA, Principal City 0.294 Non-MSA 0.270 44% of the respondents 'do not live' in a principal city	Maybe
Household_adults	Ordinal Categorical Numeric	0-3	Number of *other* adults in household, top-coded to 3	1.0 0.547 0.0 0.304 2.0 0.106 3.0 0.043 55% of respondents stated there are '1' other adults in household	Yes
Household_children	Ordinal Categorical Numeric	0-3	Number of children in household, top-coded to 3	0.0 0.706 1.0 0.120 2.0 0.108 3.0 0.066 70% of respondents stated they 'don't have any children in household'	Yes
Employment_industry	Nominal Categorical Strings	Data values are abbrev. strings	Type of industry respondent is employed in. Values are represented as short random character strings	Current data values do not contain any meaningful insight.	No
Employment_occupation	Nominal Categorical Strings	Data values are abbrev. strings	Type of occupation of respondent.	Current data values do not contain any meaningful insight.	No

			Values are represented as short random character strings.		
H1n1_vaccine	Nominal Categorical Numeric	0 = No 1 = Yes	Whether respondent received H1N1 flu vaccine	0 0.788 1 0.212 79 % of respondents 'did not received' H1N1 flu vaccine	Yes
Seasonal_vaccine	Nominal Categorical Numeric	0 = No 1 = Yes	Whether respondent received seasonal flu vaccine.	0 0.534 1 0.466 53% of respondents 'did not received' seasonal flu vaccine.	Yes

Data exploration report

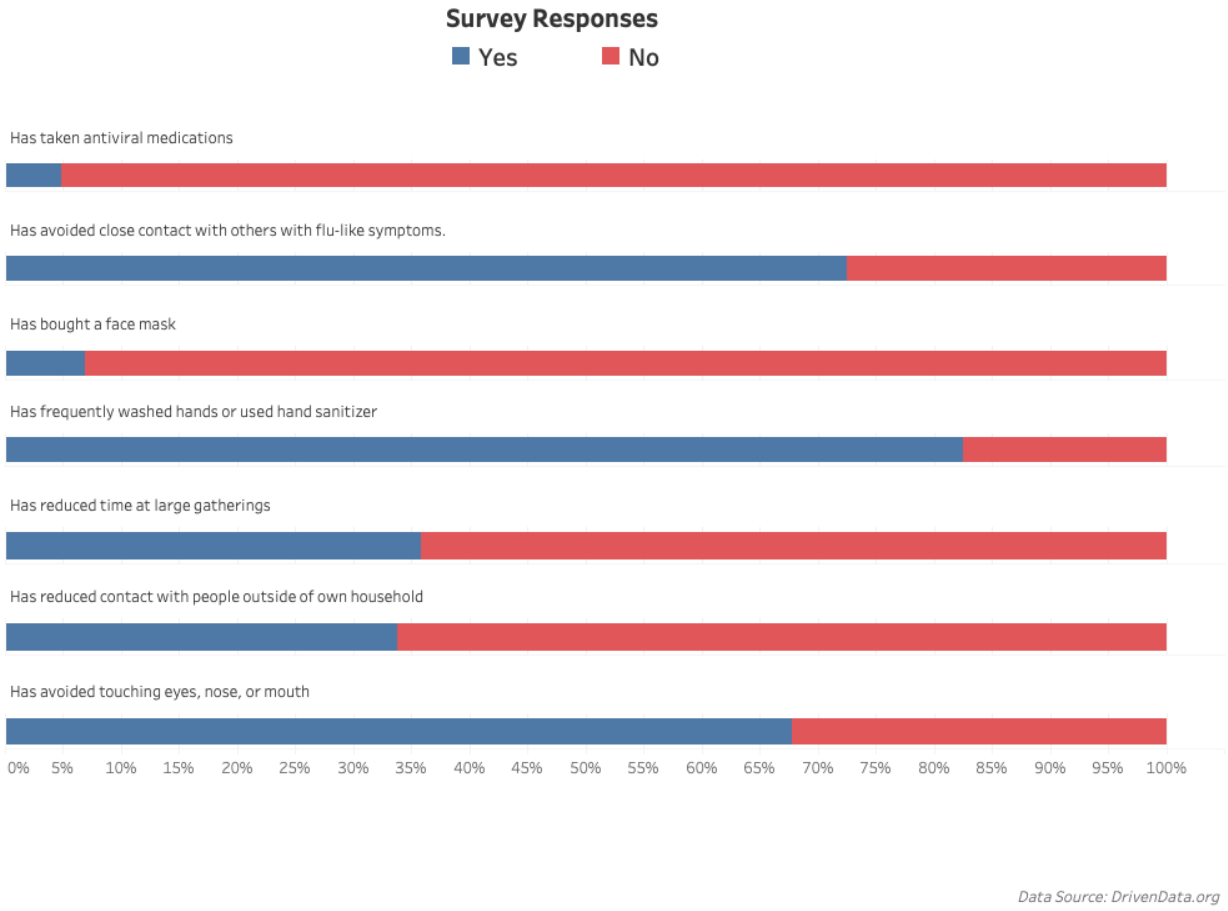
This data exploration report aims to answer the customer's secondary business questions:

1) How do we predict the effect of health-based decisions on vaccinations more accurately?
2) What are the most common characteristics in individuals that do take the vaccinations?
3) Which factors affects individuals from taking the vaccinations?

The previous data description report informs us about each variable/feature/column or in this case survey questions. During the data collection stage, surveys were conducted over phone calls. Each answer serves as a categorical response, for example 'yes' or 'no', or as ratings such as 'not at all concerned', 'not very concerned', 'somewhat concerned', and 'very concerned'. These responses are replaced with numerical entries ranging from '0' to '1', or '1' to '5' and such. Due to the nature of the data, we can create different types of bar charts to compare H1N1 seasonal flu vaccines since we are only dealing with categorical data. The following data visualizations in the next few sections were made in Tableau Desktop with the data provided by the customer (DrivenData).

Behavioral Survey Questions

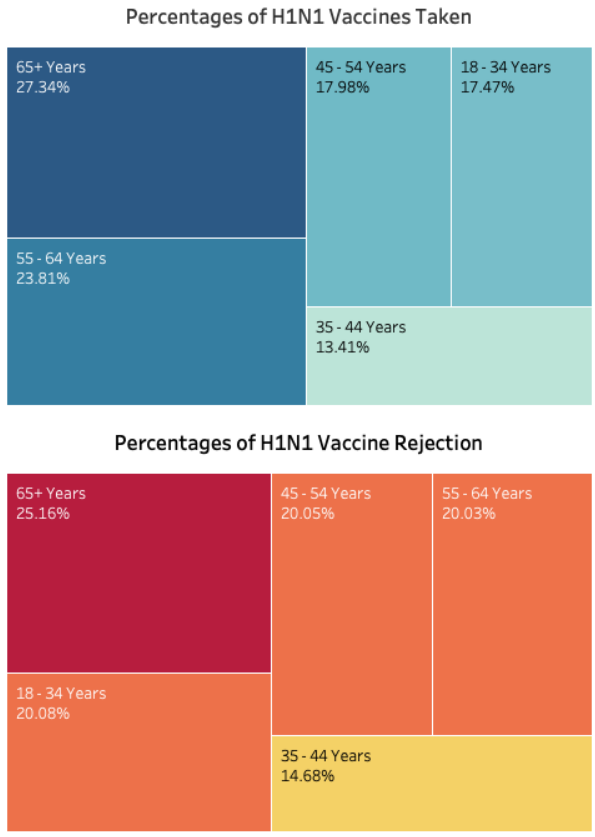
For this first section of the data exploration, we will be looking at questions that were closely related to each other. This following graph displays information about behavioral questions that were asked in the survey. We were able to tell the difference between the questions due to the data dictionary provided by the customer. For each stacked-bar chart, it is associated by the behavioral question that was asked and color coded by the responses: 'yes' and 'no'.



This bar chart displays tells us that respondents often did the following: 'avoid close contact with others with flu-like symptoms', 'has frequently washed hands or used hand sanitizer' and 'has avoided touching eyes, nose, or mouth'. Meanwhile there are overwhelmingly high percentages of 'No' response such as buying a face mask, reducing time at large gatherings, and reducing contact with people from outside of their own household. These behavioral questions serve as the first layer into understanding people's reactions and actions toward H1N1 and seasonal vaccines. We can see that there is a pattern to respondents that they tend to be outside in public as long as they frequently wash their hands or avoid people with flu-like symptoms. However, we will need to go explore other survey questions in preparation of our model building. This bar chart would help us answer the first question in the list of secondary objectives the customer is interested in.

Age groups and H1N1 Flu Vaccines

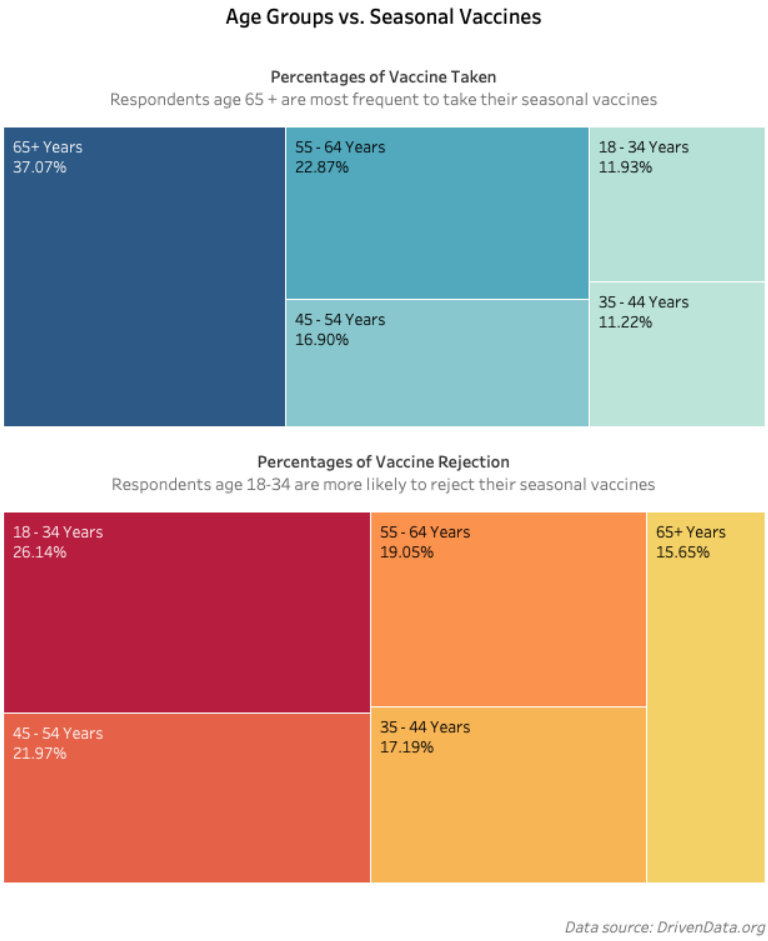
The analysis of these variables will help us determine what we should include ('feature selections') in our models. Since one of the secondary business questions is about common characteristics of the respondents, we chose to explore age groups and their respective decisions on receiving the H1N1 and seasonal vaccines. The following graphs are two separate tree maps displaying the percentages of 'Yes' and 'No' responses based on their age groups out of all groups.



The first tree map in shades of blue describes that the age groups of 55-64 and 65+ years old leads the percentages for people who took the H1N1 Vaccines. While the bottom tree map in shades of red represents the age groups that rejected the H1N1 vaccines. However, we notice there are no discernable patterns between the age groups as they are similarly proportionate for their rejection decision of H1N1 vaccines. Therefore, these graphs tell us that there is a good mix of people between age groups that took or rejected vaccines, but no real relationship can be determine since there is no significance between proportions. Next, we will examine a similar tree map based on seasonal vaccines instead.

Age Groups and Seasonal Flu Vaccines

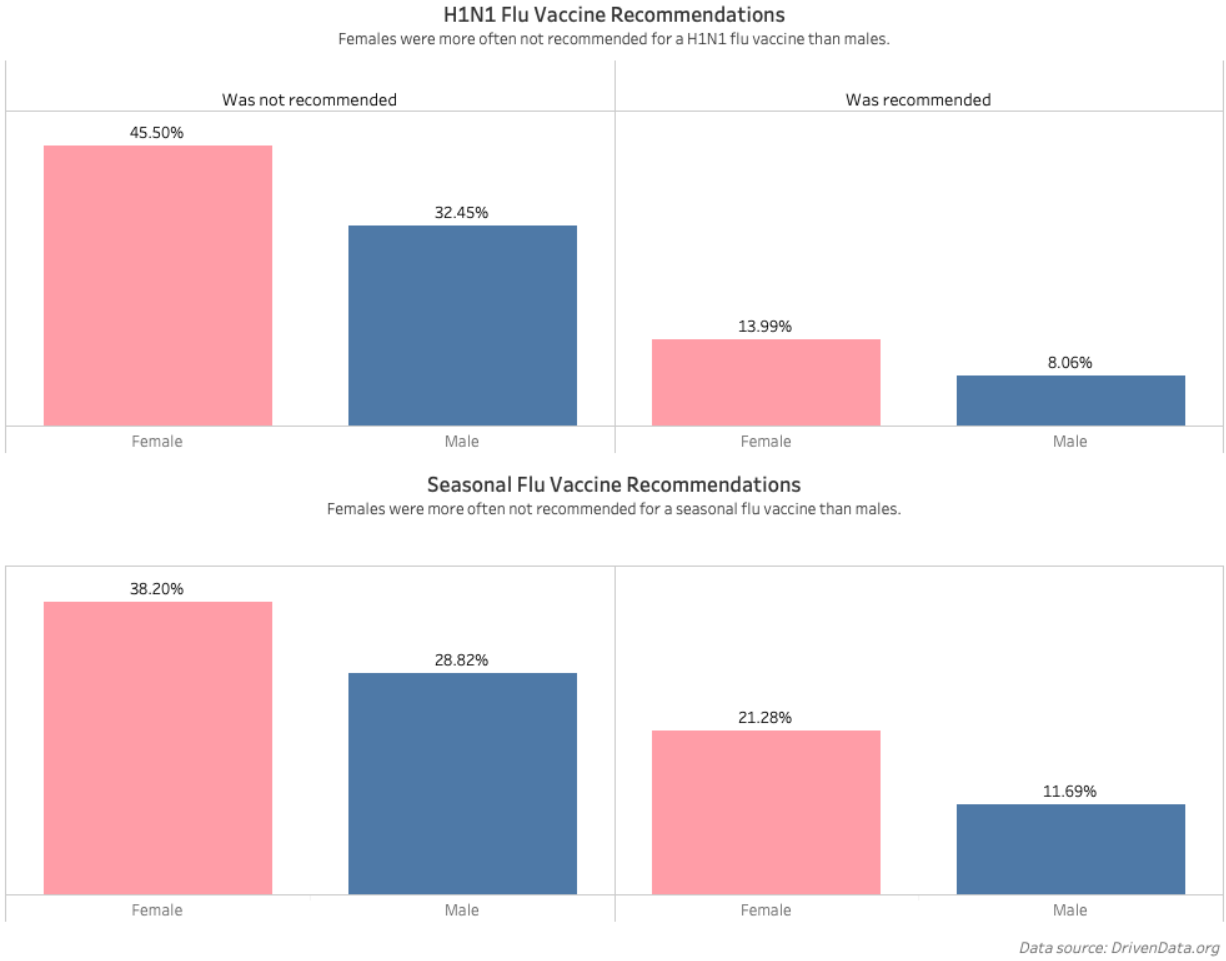
The following tree maps represent the percentages of age groups responding to ‘Yes’ or ‘No’ if they had taken the seasonal flu vaccine. The first tree maps represent age groups responding to ‘Yes’. Here we are able to see right away people 65 and older are the most frequent age group to take their seasonal vaccines.



However, on the bottom tree map, younger people in the age group of 18-34 years old leads the percentages of seasonal flu vaccine rejections. From these graphs, the insights we have are that 65 years old and older are the most frequent to take their seasonal vaccines, while 18-34 years old are more likely to reject them. As we dive deeper into the characteristics of the respondents, we should keep track of what we know so far. After looking at the demographics of age, we should consider exploring gender and their decisions on the vaccines.

Gender and Doctor's Vaccine Recommendations

Other variables may or may not have direct or indirect influences toward a person deciding to take or reject their vaccinations. In the data, we have available information about the respondents' gender and two other variables called 'doctor_recc_h1n1' and 'doctor_recc_seasonal'. The first variable was a survey question asking if the H1N1 flu vaccine was recommended by a doctor and the latter was similar but with the seasonal flu vaccine. Here, we can speculate that if a patient was recommended a vaccine for either H1N1 or seasonal flu, we hypothesize that this factor may increase the probability of taking either vaccines.

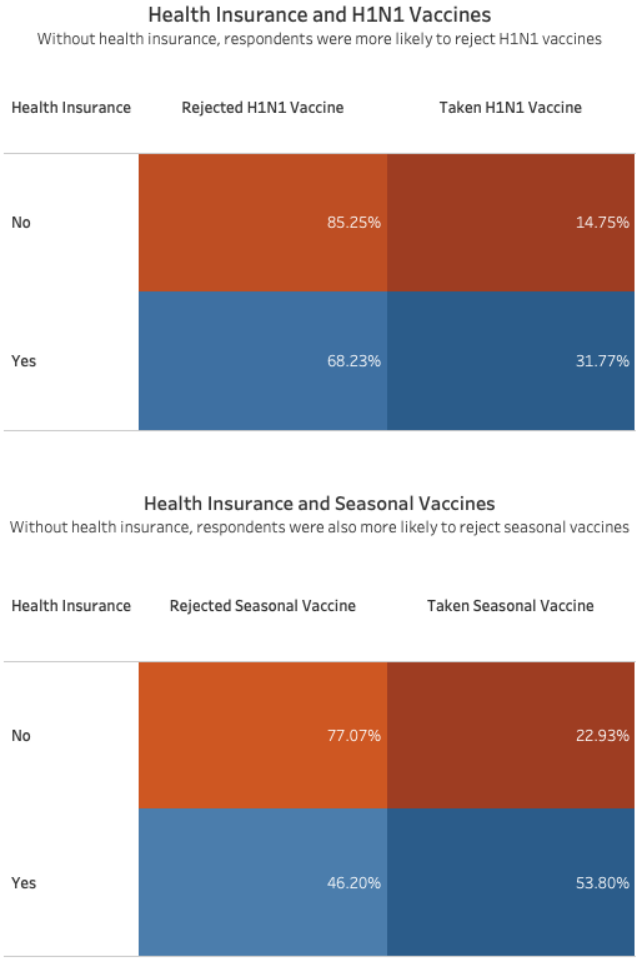


Let us explore the data between gender and vaccine recommendations. For H1N1 flu vaccines, we can see that both males and females were very often left without a recommendation. Then, looking closely again, females were more likely than males for no recommendations for this vaccine. For the bottom graph, we can see a similar pattern occurs with a smaller difference. Both males and females were often not recommended with the seasonal flu vaccine and females were more likely than males with no recommendations. These variables should be taken into consideration for further hypothesis testing to validate our claims. However, these variables are in full consideration for model building as we enter the data preparation phase later on.

So far, we know that the respondents do take precautions such as being sanitary, younger people from 18-34 years old were more likely to neglect their seasonal flu vaccine and that females were more often not recommended a H1N1 or seasonal flu vaccine. These insights will help us narrow down the factors that affect individuals from taking the vaccines. Moving on, we should look at respondents with or without health insurance.

Health Insurance and Flu Vaccines

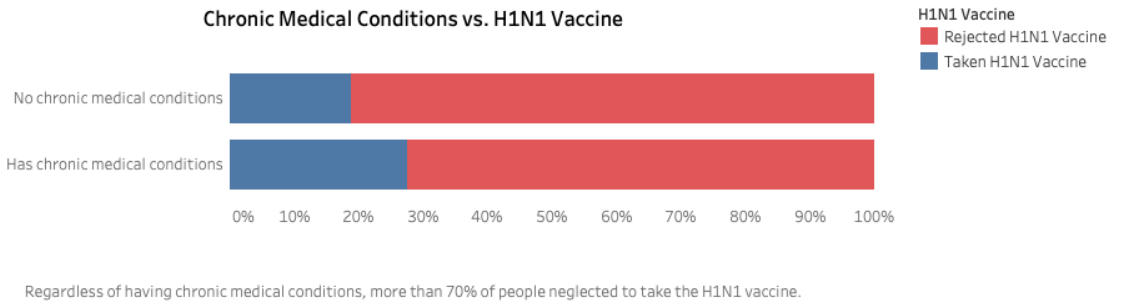
Here we have a variable called ‘health_insurance’. This was a survey question asking the respondents if they have health insurance or not. This is an extremely important key attribute that may play a role in affecting their vaccine decisions. We will be exploring these relationships in a cross-table diagram on how having health insurance influence H1N1 and seasonal flu vaccines.



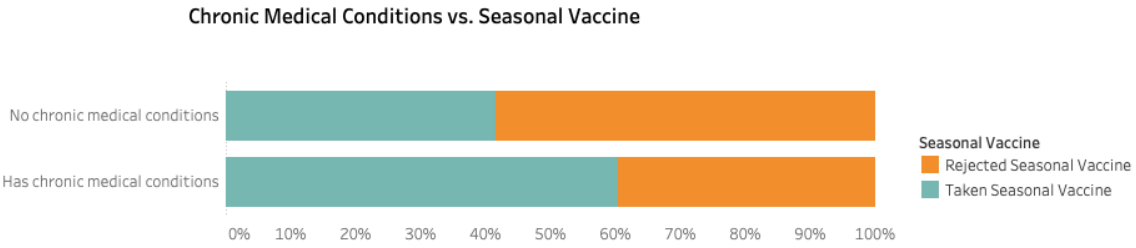
In the first table, we can see that without health insurance, respondents were undoubtedly opting out of taking the H1N1 vaccine. On the other hand, even if the respondents were to have health insurance at the time, they were twice as likely to reject the H1N1 vaccine then to take it. This is atonishing information because the respondents may have their mind already set on opting out of the H1N1 flu vaccine regardless of having health insurance or not. While on the second table, having health insurance, increases the chances of a seasonal vaccine by nearly 7%. Although, without health insurance, respondents were most certainly opting out of the seasonal vaccine. This provide very valauble insight to understanding how health insurance would not increase the likelihood of vaccinations for the H1N1 flu. As we diverged from demographics data of the respondents (age and sex), we should continue on this path to understand the relationship between vaccines and health-based factors.

Chronic Medical Conditions and Vaccines

In the data, we have a variable called ‘chronic_med_condition’. This survey question asks respondents if they have any of the following chronic medical conditions: asthma, diabetes, heart, lungs, or kidney conditions, or in general a weakend immune system due to a chronic illness. We will explore these health conditions and the respondents’ decisions on vaccinations. As we start noticing a pattern that people tend to ignore the H1N1 flu vaccine, we speculate that having a chronic medical condition will also not play a role or has any influence.



Regardless of having chronic medical conditions, more than 70% of people neglected to take the H1N1 vaccine.



People are more likely to take the seasonal vaccine if they have chronic medical conditions

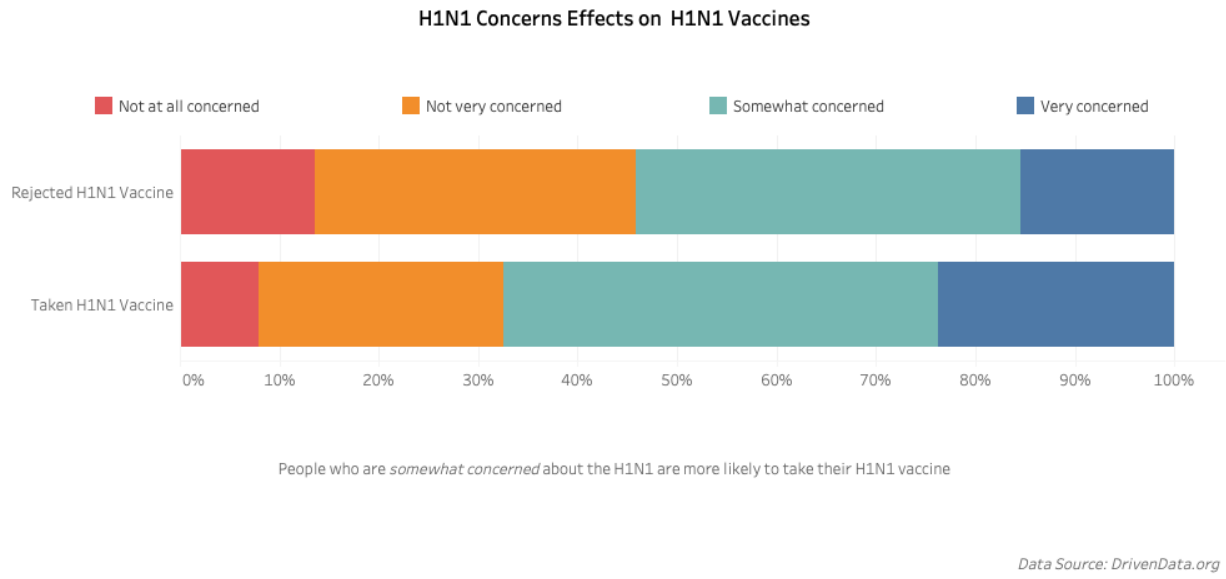
Data source: DrivenData.org

Here we can note the same overwhelming difference between chronic medical conditions and decisions on the H1N1 vaccine. Regardless of having said conditions, more than 70% would rather opt of the H1N1 vaccine. For those that does have these medical conditions, we saw a negible difference in choosing to take this vaccine. Onward to the seasonal flu vaccine, in the bottom graph we can spot the difference right away as more people are encouraged to take the seasonal vaccine if they do have chronic medical conditions. This bring things to a lighter note as we can see that people do take the seasonal flu vaccines more willingly to be safe. Since this pattern continues to develop as we explore health based factors with H1N1 flu vaccinations, we

are collecting common characteristics on respeondents. Lastly, we should analyze the respondents' opinions on the H1N1 flu for a better look at why most opted out.

H1N1 Concerns and H1N1 Vaccines

This variable in the data is called 'h1n1_concern' and in business terms it means the level of concern about the H1N1 flu. The possible answers are ranked from 0-3, for level 0 at not at all concerned, not very concerned, somewhat concerned to very concerned at level 3.



This stacked bar chart displays information about the population's concerns on the H1N1 flu. At majority level, respondents were between 'not very concerned' and 'somewhat concerned' than their respective ends of the spectrum. There was no significant difference to note between the decision of taking or opting out of the H1N1 vaccine. We can see there was a very minor difference between percentages that took or rejected the H1N1 vaccine. Thus, this variable can tell us a lot about the H1N1 flu discoveries and break outs. Were people not well-informed? Or do they simply not care enough? We are able to observe that the majority of the population did not prioritize the H1N1 vaccine. However, for those who were concerned enough, there was a higher likelihood they took the H1N1 flu vaccine. As we wrap up the data exploration stage, we

will conclude with a summary of our findings to answer the customer's secondary business questions.

In Summary

- Assumptions
    - There were assumptions on the behavioral aspect of the respondents during the H1N1 flu season. We assumed that people were not really concerned as long as they stay hygienic by washing their hands and avoid contact with flu-like symptoms people.
    - Then we made assumptions about the relationships of demographics influencing vaccinations such as the higher the age group, the higher the likelihood for them to take a vaccine.
    - We assumed for people without health insurance, they were more likely to opt out of a vaccine. We saw this to be true for both types of vaccinations.
    - Lastly, we looked at more health-related factors such as
- Constraints
    - There are constraints on the number in each class/level for people who took or rejected the H1N1 flu vaccine because we may have to compensate for the proportion of classes
    - A lot of missing values in 'Health_insurance' column – may bias analysis
- Conclusion
    - We observed that respondents were able to be out in public and no changes were added to their lifestyles as long as they take precautions during the flu seasons.
    - Next, H1N1 flu vaccines does not differ between age groups. Meanwhile, we saw that the higher the age group, the more likely they were to take a seasonal vaccine instead.
    - Moving on to sex demographics, we observed an unfair percentage of females not receiving a vaccine recommendation by doctors than males.
    - Regardless of health insurance, more respondents opted out of the H1N1 flu vaccine. While if respondents did have health insurance, seasonal flu vaccines are more in favor.
    - We noted that regardless of chronic medical conditions, it has no effect on respondents taking the H1N1 vaccine, as we observed that more people opted out for both groups. While for seasonal flu vaccines, respondents followed the yearly trend of getting a seasonal flu vaccine if they struggle with health issues.
    - Lastly, we observed that respondents who are 'somewhat concerned' are more likely to take the H1N1 flu vaccine.

This concludes the data exploration report as we dive into the data quality report.

Data quality report

In our initial data description report, we found that there are no continuous features recorded in our data. All of the available variables/features in our dataset are listed as categorical data types. So, there is no need to perform an outlier check or produce histograms to look at its distribution shapes. As we explored the data in the previous report, we produced a collection of bar plots to visualize pairs of categorical features instead of drawing insights from scatter plots and correlation matrices.

However, as we go through the data quality report, we will need to note data quality issues such as missing values, irregular cardinality, since we are dealing with categorical data types directly from the valid data. The valid data is the only data provided by the customer, which we will only use for building the predictive models.

The following report gives us information about each feature such as the number of counts, the percentage of missing values, the cardinality number of each, the top mode value along with its frequency number and percentage, and the second top mode value along with the same associations.

	Feature	Count	% Miss	Card.	Mode	Mode Freq.	Mode %	2nd mode	2nd Mode Freq.	2nd Mode %
0	h1n1_concern	26707	0.003	4	2.0	10575	0.396	1.0	8153	0.305
1	h1n1_knowledge	26707	0.004	3	1.0	14598	0.547	2.0	9487	0.355
2	behavioral_antiviral_meds	26707	0.003	2	0.0	25335	0.949	1.0	1301	0.049
3	behavioral_avoidance	26707	0.008	2	1.0	19228	0.720	0.0	7271	0.272
4	behavioral_face_mask	26707	0.001	2	0.0	24847	0.930	1.0	1841	0.069
5	behavioral_wash_hands	26707	0.002	2	1.0	22015	0.824	0.0	4650	0.174
6	behavioral_large_gatherings	26707	0.003	2	0.0	17073	0.639	1.0	9547	0.357
7	behavioral_outside_home	26707	0.003	2	0.0	17644	0.661	1.0	8981	0.336
8	behavioral_touch_face	26707	0.005	2	1.0	18001	0.674	0.0	8578	0.321
9	doctor_recc_h1n1	26707	0.081	2	0.0	19139	0.717	1.0	5408	0.202
10	doctor_recc_seasonal	26707	0.081	2	0.0	16453	0.616	1.0	8094	0.303
11	chronic_med_condition	26707	0.036	2	0.0	18446	0.691	1.0	7290	0.273
12	child_under_6_months	26707	0.031	2	0.0	23749	0.889	1.0	2138	0.080
13	health_worker	26707	0.030	2	0.0	23004	0.861	1.0	2899	0.109
14	health_insurance	26707	0.460	2	1.0	12697	0.475	0.0	1736	0.065
15	opinion_h1n1_vacc_effective	26707	0.015	5	4.0	11683	0.437	5.0	7166	0.268
16	opinion_h1n1_risk	26707	0.015	5	2.0	9919	0.371	1.0	8139	0.305
17	opinion_h1n1_sick_from_vacc	26707	0.015	5	2.0	9129	0.342	1.0	8998	0.337
18	opinion_seas_vacc_effective	26707	0.017	5	4.0	11629	0.435	5.0	9973	0.373
19	opinion_seas_risk	26707	0.019	5	2.0	8954	0.335	4.0	7630	0.286
20	opinion_seas_sick_from_vacc	26707	0.020	5	1.0	11870	0.444	2.0	7633	0.286

21	age_group	26707	0.000	5	65+ Years	6843	0.256	55 - 64 Years	5563	0.208
22	education	26707	0.053	4	College Graduate	10097	0.378	Some College	7043	0.264
23	race	26707	0.000	4	White	21222	0.795	Black	2118	0.079
24	sex	26707	0.000	2	Female	15858	0.594	Male	10849	0.406
25	income_poverty	26707	0.166	3	<= $75,000, Above Poverty	12777	0.478	> $75,000	6810	0.255
26	marital_status	26707	0.053	2	Married	13555	0.508	Not Married	11744	0.440
27	rent_or_own	26707	0.076	2	Own	18736	0.702	Rent	5929	0.222
28	employment_status	26707	0.055	3	Employed	13560	0.508	Not in Labor Force	10231	0.383
29	census_msa	26707	0.000	3	MSA, Not Principle City	11645	0.436	MSA, Principle City	7864	0.294
30	household_adults	26707	0.009	4	1.0	14474	0.542	0.0	8056	0.302
31	household_children	26707	0.009	4	0.0	18672	0.699	1.0	3175	0.119
32	h1n1_vaccine	26707	0.000	2	0	21033	0.788	1	5674	0.212
33	seasonal_vaccine	26707	0.000	2	0	14272	0.534	1	12435	0.466

The next section of this data quality report will include a data quality plan on how to deal with any messy data we have encountered.

Data Quality Plan

This table will report the data quality issues we encountered along with the necessary actions to handle them. It will serve as a reminder for data quality issues.

Feature	**Data Quality Issue**	**Potential Handling Strategies**
Health_insurance	Contains 46% of missing data	May be considered for complete removal – or use complete case analysis
Income_poverty	Contains 17% of missing data	Imputation based on mode
Behavioral questions • Behaviorial_antiviral_meds • Behaviorial_avoidance • Behaviorial_face_mask • Behaviorial_wash_hands • Behaviorial_large_gatherings • Behaviorial_outside_home • Behaviorial_touch_face	Redundant or repetitive information	May provide similar information and may be redundant – consider Principal component analysis?
Opinion questions • Opinion_h1n1_vacc_effective • Opinion_h1n1_risk • Opinion_h1n1_sick_from_vac • Opinion_seas_vacc_effective • Opinion_seas_risk • Opinion_seas_risk_from_vacc	Redundant or repetitive information	May provide similar information and may be redundant – consider Principal component analysis?

Education	Contains ordinal categorical data: '< 12 Years', '12 Years', and 'College Graduate' – might be harmful to modeling	Consider feature engineering – one hot encoding or replacing string data with numerical data representing ratings from 0 to 2 to match with education level. This may reduce complexity in data for modeling
race	Contains nominal categorical data: 'White', 'Black', 'Hispanic', 'Other or multiple'	One hot encoding or string conversion to numerical data types 0 – 'White' 1 – 'Black' 2 – 'Hispanic' 3 – 'Other or multiple'
sex	Contains nominal categorical data that can be converted into numerical for better modeling accuracy	One hot encoding or string conversion to numerical data types 0 – male 1 - female
Income_poverty	Contains ordinal categorical data: 'Below Poverty', '<=$75,000, Above Poverty', '>$75,000', – might be harmful to modeling	One hot encoding or string conversion to numerical data types representing ratings from 0 to 2 to match with education level. This may reduce complexity in data for modeling
Martial_status	Contains nominal categorical data that can be converted into numerical for better modeling accuracy	One hot encoding or string conversion to numerical data types
Own_rent	Contains nominal categorical data that can be converted into numerical for better modeling accuracy	One hot encoding or string conversion to numerical data types

Employment_status	Ordinal categorical data that can be converted into numerical data	One hot encoding or string conversion to numerical data types representing ratings from 0 to 2 to match with education level. This may reduce complexity in data for modeling
Hhs_geo_region	Data is incomplete and contain errors from the valid data customer has provided – values are represented as short random character strings (abbreviated)	May need to be consider for full removal Ultimately, does not provide any useful insight during data exploration
Census_msa	Contains nominal categorical data that can be converted into numerical for better modeling accuracy	Consider feature engineering – one hot encoding ranging or replacing string data with numerical data representing ratings from 0 to 2 to match with education level. This may reduce complexity in data for modeling
Employment_industry	Data is incomplete and contain errors from the valid data customer has provided – values are represented as short random character strings (abbreviated)	May need to be consider for full removal Ultimately, does not provide any useful insight during data exploration
Employment_occupation	Data is incomplete and contain errors from the valid data customer has provided – values are represented as short random character strings (abbreviated)	May need to be consider for full removal Ultimately, does not provide any useful insight during data exploration

In conclusion, there are a handful of columns that needs to be removed completely due to data entry errors from the original valid data. These columns contain values that cannot be corrected or retrieved from the customer directly. It is more optimal to remove them. Another issue we ran into was redundancy. Some survey questions (variables/features/columns) extract the same type of information from the respondents. Thus, we may consider feature selection such as principal component analysis or other machine learning algorithms such as a random forest classifier to look at entropy in relation to its relevancy with the target features. Other data quality issue is a matter of small fixes such as converting ordinal categorical data to numerical data types which may be performed in the data preparation phase. Lastly, there are only 2 columns contain missing values over 10% but under 50%. Therefore, we may be able to perform imputation or a complete case analysis as well as flagging certain instances to handle them better.

This marks the end of our data quality report.

References

DrivenData. (n.d.). Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines. Retrieved January 17, 2021, from https://www.drivendata.org/competitions/66/flu-shot-learning/page/210/