Employee likelihood to receive H1N1 and Seasonal Flu Vaccines

Salvador R Sanchez Castro

University of San Diego

Master of Science, Applied Data Science

MADS 501: Foundations of Data Science

June 21nd, 2022

Business Understanding

Background

Parker Hannifin Corporation commonly referred to as Parker is a Fortune 250 company and a component of the S&P 500 stock market index, with over 100 years of history and engineering breakthroughs it is a leader in the industry. Parker has expanded to 336 manufacturing facilities with presence in 50 countries and currently has a workforce of 55,000 employees (Parker, n.d.). Parker is extremely committed to maintaining customer satisfaction and OTD. The aerospace division has secured a 10-year LTA with key customers while this ensures business it also burdens heavy contractual responsibilities on monthly customer demand. During the COVID-19 pandemic several production lines in the aerospace division were shut down due to personnel being out sick. Jack Baur in charge of operations for the division is looking to reduce downtime due to infection diseases, with enough immunizations in the shopfloor to produce “herd immunity”, he has commissioned us to provide insight on the probability his employees have taken H1N1 and seasonal flu vaccines.

Business objectives and success criteria

During the COVID 19 outbreak companies got a rude awaking to a new risk factor they should manage in order to maintain customer demand, infectious diseases. Parkers Aerospace main objective is to reduce production down time due to infectious diseases outbreaks, product lines require specialized personnel and although there is an effort to cross train personnel not all positions are applicable and human resources are limited. Parker will not be mandating its employees to be vaccinated, from the study it will gain insight on what factors influence an employee’s decision in order to try to encourage vaccination. With knowledge of the estimated vaccinated population parker will have vaccination risk factor, allowing them to control or mitigate the risk via risk management, at the production line level the

mix of vaccinated vs unvaccinated can you also provide perceptive for production leads to hedge the combination of employees and prevent production lines to have a proportion of unvaccinated employees higher than they are willing to risk.

The success of the model will be measured in hours of downtime due to a lack of personnel. Downtime hours can come from several factors, material shortage, broken machine, personnel shortage, therefor to measure the impact of the produced PA model shall be downtime. Downtime hours should be reduced by 10%, for lack production related to employees taking sick leave. Secondary measures of success can be measured in sick days, if the model can help reduce the number of sick days proportional to the current employee population compared year over year, we can reduce the risk of downtime. The objective for the model shall be to reduce sick days by 20% YOY.

Inventory of resources

The following recourses will be utilized in the study:

Personnel

- **Project Manager** - Manage timeline and milestones, responsible for managing Jira project board and holing daily meetings to review advancements. They will share progress with business stockholders and request additional information if necessary.
- **Data Engineer** – Responsible for obtaining the data, cleaning it and prepare for analysis. Construct ETL pipelines, deploy databases, and design tables.
- **Data Scientist** – Responsible for analyzing the data and developing the statistical model. Shall follow CRISP-DM process for data mining.

Data

- National 2009 H1N1 Flu Survey https://www.drivendata.org/competitions/66/flu-shot-learning/data/

Hardware

- Apple MacBook Air with:
    - Apple M1 Chip
    - 16 GB Memory
    - macOS Monterey version12.3.1

Software

- Python 3.9.7.
- jupyter_core 4.8.1
- pandas 1.3.4
- scipy 1.7.1
- scikit-learn 0.24.2
- seaborn 0.11.2
- spyder 5.1.5
- GitHub at https://github.com/zalvatore/H1N1-and-Seasonal-Flu
- Jira https://olilu.atlassian.net/jira/core/projects/FLU/board
- Dropbox
- Survey Monkey https://www.surveymonkey.com

Requirements, assumptions, constraints, and RESOLVEDD Strategy

Requirements

- The operations/production managers will be the main users for the tool developed.

- The initial study will be for manufacturing employees in the engine composites division further expansion of the study will include the entire division followed by the entire manufacturing population.
- Human resources will have four weeks to deliver survey questionnaire to all employees together their behavioral responses.
- Parker has requested the study be finished before the next flu season, so we have set a completion date of September 30th.
- The study will be evaluated during the next two flu seasons if successful migrating to other divisions in the organization may start.
- To maintain the model when moving to new locations the model must be validated to avoid any shifts due bias in region.
- Parker has two California location and with annual revenues over 25 million they must comply with CCPA. The California Consumer Privacy Act stipulates that the company shall “have mapped data and be able to provide disclosures on all information the company has retained on an employee as far back as Jan. 1, 2022” (Kostal, 2021), data compiled by the study will fall in this category and must comply.
- The repercussions of in the event of a false positive will not cause harm or have a severe impact, accuracy levels above 60% shall be acceptable.
- Due to the sensitivity of the data the employee’s names shale by encrypted and only visible unencrypted to the direct line manager.
- The server hosting the data shall require MFA and SQL access will be managed by roll bases access (RBAC).
- Programing code for the model shall be implemented in Python scripting language with proper commenting and follow PEP 8 code styling (PEP 8, n.d.) .

Assumptions

- IT shall grant all necessary access and privileges to data and SQL servers to the personnel performing the study, technical support should be provided in timely manner if necessary.
- It is implicit the data provided for the training set is accurate and obtained by legal concept from the individuals taking the survey.
- Human resources and legal departments are aware and have bought off the study and its implications. The number of completed and valid surveys shall be a meet a minimum sampling quantity to assure a confidence level of 95%.
- It is assumed that employees with flu vaccine will take fewer sick days, the CDC states that in effectiveness of flu vaccines towards the season strain they are designed for reduces the risk up to 60% (CDC, n.d.).
- As part of adequate regression model, we assume that the independent variables have little or no Multicollinearity.

Constraints

- The budget of the study has been capped at $75,000, this shall include all labor and travel expenses. Parker has agreed to a biweekly meeting of an hour to review progress and any face time needed to resolve limitations.
- Network access is limited and may not be accessed through a non-Parker asset.
- Data for the study and production of model maybe copied and used by personnel performing the study but access to the SQL Server will be limited to being at a Parker facility.

RESOLVEDD

- **Review**: Following the COVID 19 outbreak, vaccination has come to the forefront controversy with parties in favor and against. The study does not intend to isolate or discriminate individuals based on their vaccination status but to provide model that will help production understand their ratio mix of vaccination status employees to reorganize or established risk assessment.
- **Estimate:** If the model is proven to be successful and we can accurate model that predicts which employees will vaccinate and which will not, it is possible for users that have access to the data generate bias against those employees who have different thoughts on vaccination then their own
- **Solutions:** The study shall limit access to the results as much as possible only those employees who required the data shall have access to it. Employee names shall be encrypted and only available to direct line managers, once the manager is no longer responsible for that line their privileges will be revoked. Logs shall be kept of users accessing the data in any suspicious or irregular behavior shall be flagged for review.
- **Outcomes:** It is expected when reviewing the product users can appreciate and review the vaccination rate ratio mix with anonymous employees.
- **Likely:** With limited access to the actual individuals and monitoring who accesses the data we should reduce any in non-intended consequence
- **Values:** Since we are talking about behavior of human beings, we want to avoid any conscious or unconscious bias manages may have on their employees with their own personal decisions. We are not here to promote or reject vaccinations just provide a tool to estimate which individuals are most likely to get vaccinated.
- **Evaluate:** Users of the model should not be discouraged or find lack of value if they do not know the actual individual's name.

- **Decide:** Encrypting the employee will result in the safest way to keep them anonymous, having rolled based access to the information will assure access is granted only to those authorized users
- **Defend:** In many cases managers of big plant may not have personal contact with each employee therefore the solution should not hinder any use of the tool.
-

Risks and contingencies

Risk	**Contingencies**
Data loss, unexpected damage, or loss of hardware	All documents will be working out of a shared directory in Dropbox to assure resilience. Scripts (code) will require the team to push to the data repository at GutHub.
Human Resource unavailability	Our backup solution and coding repository will also help with communication in transfer of knowledge between members of the team. Both data scientists and data engineer will be up to date with their tasks and responsibilities and able to support if ever one of them is not available.
Origination risk to approve study variables	The study in the training data in the test contains several behavioral questions it will be important to validate with human resources that the questions are acceptable and available for parker employees. and available.
Sample Size	After the COVID-19 epidemic vaccines have become a controversial topic, the study is not meant to discriminate any employee who does not wish to be vaccinated but to provide information on the vaccine status as an aggregate. It will be important for Parker to confirm the access to the behavioral information of the employees required for the study.

Terminology

OTD: On Time Delivery

LTA: Long Term Agreements

PA: Predictive Analytics

YOY: Year over year

Dichotomous: Variable with two alternative True or False also known as Boolean

GitHub: tool for software development and version control

h1n1: commonly known as swine flu, strain of influenza

NHFS: National 2009 H1N1 Flu Survey

NCIRS: National Center for Immunization and Respiratory Diseases

CDC: Centers for Disease Control and Prevention

Downtime: Period of time when the factory production line has to stop production

Precision: Proportion of true positives over total positive predictions

Recall: Proportion of true positives over total positive actuals

F_1 measure: Harmonic mean between precision and recall

Harmonic mean: Number of observations divided by the reciprocal of the observations

RESOLVEDD: Review, Estimate, Solutions, Outcomes, Likely, Values, Evaluate, Decide, Defend

CCPA: California Consumer Privacy Act

MFA: Multi Factor Authentication

RABC: Role Based Access Control

ETL: Extract, Transform, and Load

CRISP-DM: Cross International Process for Data Mining.

Data mining goals and success criteria

The study will be designing a classification model**.** The aim for the model is to correctly predict likeliness to receive a flu vaccine, the model will produce a dichotomous result, the likelihood will be True or False. To produce bias in favor of positive result the study shall designate True when the probability of the model produced is above or equal 60%.

To measure the performance of our model we can focus on when predicting an employee is vaccinated (True Prediction) the model is correct, this is called Precision. An alternative is focus on obtaining all the actual employees who are vaccinated (True Actual), thinks called Recall.

When dealing with medical factor such as vaccination it will be important to avoid predications that are incorrect, either False Negatives or False Positives. The study shall use F_1 measure to evaluate the performance of the model, the measure is harmonic mean between precision and recall; F_1 measure is robust and used when incorrect classification is critical. Equitation 1 shows the calculation of the F_1 measure comprised of Precision and Recall.

$$F_1 \text{ measure } = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \quad (1)$$

With a more elevated F_1 measure the model will predict which employees are more probable to be vaccinated thus provided more insight to the company on how to arrange the working teams and reduce downtime.

Project plan/ Order of tasks

Task	Duration (In Days)
Phase 1: Determine business objectives background in business criteria	27
Team formation: Incorporate a of three members, project manager, data engineer, and data scientist.	3
Perform due diligence and initial investigation on parkers background	5
Agreement on business objectives and success criteria by AF and Parker stockholder	3
Deliver inventory of resources. Hardware and software utilized shall be documented, all software in libraries shell detail versions used. Communication and repository tools shall be initiated	2
Deliver list of requirements, and legal constraints	5
Produce RESOLVEDD steps and actions to mitigate ethical concerns.	5
Risk registry shall we developed, and it will include at a minimums Data loss, unexpected damage, or loss of hardware as well as human resource unavailability, and sample size.	3
AF and Parker to buy off on phase one completion.	1
Phase 2: Determine data mining objectives and criteria	10
Data mining objectives and success criteria for study shall be documented, objectives shall be measurable and aligned with business success.	10
Phase 3: Data Preparation: Dive into vaccination likelihood data	16
Initial Data exploration	5

Delivery data quality report. Missing values shall also be explored, and data wrangling will be used if required	4
Develop dashboards of visualizations for factors of the study and interact with the likelihood of vaccination	4
Deploy dashboards to Tableau server to be shared with stakeholders	2
AF shall sign off on finalization of phase three to confirm data is properly prepared for next phase	1
Phase 4: Modeling, Evaluation and Deployment	106
Model design that will evaluate the likelihood of Parker employees to take flu vaccinations.	10
Model shall be evaluated against the criteria established prior and deemed effective or not	5
Evaluate if model is mature enough to be deployed	1
During test period, evaluate if the business achievements have been met.	90
Final evaluation of the project shall be determined as successful or not. If successful Deployment shall be Permanente.	1

For full Gantt chart see Appendix 1.

Data Understanding

Initial data collection report

The data for the study was obtained via a random-digital-dialing telephone survey during the 2009-10 season, the survey was denoted NHFS (The National 2009 H1N1 Flu Survey) and was sponsored by the NCIRD and conducted together with the CDC and NCHS (Data Driven, n.d.). The data is sample of

the national population limited to individuals six months or older. The data was compiled into the following data sets:

File	Description	Observations	Columns
Training Features	Training set features.	26,707	36
Training Labels	Labels for the training set.	26,707	3
Test Features	Test set features.	26,708	36

The “Training Features” will be used to develop the model, and the “Test Features” to validate the model. “Training Labels” include Boolean field indicating if respondent to the survey received h1n1 vaccine or season flu vaccine. The “Training Features” dataset has 35 variables and one identifier column. Data files are stored in CSV format and will be accessed vias pandas library.

Data description report

The data set in ‘Training_Features.csv’ contains Float and Object data types that will data type updates in order to optimize the model. Table 1 shows features that shall be updated to Boolean data type.

Figure 1

H1n1 Boolean Features in "Training Features" dataset

	Unique Obs
marital_status	2
health_insurance	2
health_worker	2
child_under_6_months	2
doctor_recc_seasonal	2
doctor_recc_h1n1	2
behavioral_outside_home	2
chronic_med_condition	2
behavioral_wash_hands	2
behavioral_face_mask	2
behavioral_avoidance	2
behavioral_antiviral_meds	2
sex	2
behavioral_large_gatherings	2
behavioral_touch_face	2

The data sets contains categorical feature, special consideration must also be taken into account to be processed for a regression model, the feature must be coded into n minus one (n equal to number of categories) dichotomous variables. Table 2 shows variables that will require coding with dummy variables.

Table 2

Categorical *features in "Training Features" dataset*

	Unique Obs
employment_status	3
income_poverty	3
h1n1_knowledge	3
census_msa	3
household_children	4
education	4
h1n1_concern	4
household_adults	4
race	4
opinion_h1n1_risk	5
opinion_seas_sick_from_vacc	5
opinion_seas_risk	5
opinion_seas_vacc_effective	5
opinion_h1n1_sick_from_vacc	5
age_group	5
opinion_h1n1_vacc_effective	5
hhs_geo_region	10
employment_industry	21
employment_occupation	23

Table 3 is an example of the first three rows of the training set.

Table 3

"Training Features" dataset example

	Row 1	Row 2	Row 3
respondent_id	0	1	2
h1n1_concern	1.0	3.0	1.0
h1n1_knowledge	0.0	2.0	1.0
behavioral_antiviral_meds	0.0	0.0	0.0
behavioral_avoidance	0.0	1.0	1.0
behavioral_face_mask	0.0	0.0	0.0
behavioral_wash_hands	0.0	1.0	0.0
behavioral_large_gatherings	0.0	0.0	0.0
behavioral_outside_home	1.0	1.0	0.0
behavioral_touch_face	1.0	1.0	0.0
doctor_recc_h1n1	0.0	0.0	NaN
doctor_recc_seasonal	0.0	0.0	NaN
chronic_med_condition	0.0	0.0	1.0
child_under_6_months	0.0	0.0	0.0
health_worker	0.0	0.0	0.0
health_insurance	1.0	1.0	NaN
opinion_h1n1_vacc_effective	3.0	5.0	3.0
opinion_h1n1_risk	1.0	4.0	1.0
opinion_h1n1_sick_from_vacc	2.0	4.0	1.0
opinion_seas_vacc_effective	2.0	4.0	4.0
opinion_seas_risk	1.0	2.0	1.0
opinion_seas_sick_from_vacc	2.0	4.0	2.0
age_group	55 - 64 Years	35 - 44 Years	18 - 34 Years
education	< 12 Years	12 Years	College Graduate
race	White	White	White
sex	Female	Male	Male
income_poverty	Below Poverty	Below Poverty	<= $75,000, Above Poverty
marital_status	Not Married	Not Married	Not Married
rent_or_own	True	True	True
employment_status	Not in Labor Force	Employed	Employed
hhs_geo_region	oxchjgsf	bhuqouqj	qufhixun

The data contains gaps that will require management, Table 4 shows the percentage of missing records for features with more than five percent missing values.

Table 4

Percentage of Null values for features in "Training Features" dataset

employment_occupation	50.27
employment_industry	49.70
health_insurance	45.78
income_poverty	16.84
doctor_recc_h1n1	8.09
doctor_recc_seasonal	8.09
rent_or_own	7.62
employment_status	5.51
marital_status	5.40
education	5.27

The following are the description for each independent variable:

- h1n1_concern - Level of concern about the H1N1 flu.
    - 0 = Not at all concerned; 1 = Not very concerned; 2 = Somewhat concerned; 3 = Very concerned.
- h1n1_knowledge - Level of knowledge about H1N1 flu.
    - 0 = No knowledge; 1 = A little knowledge; 2 = A lot of knowledge.
- behavioral_antiviral_meds - Has taken antiviral medications. (binary)
- behavioral_avoidance - Has avoided close contact with others with flu-like symptoms. (binary)
- behavioral_face_mask - Has bought a face mask. (binary)
- behavioral_wash_hands - Has frequently washed hands or used hand sanitizer. (binary)
- behavioral_large_gatherings - Has reduced time at large gatherings. (binary)
- behavioral_outside_home - Has reduced contact with people outside of own household. (binary)
- behavioral_touch_face - Has avoided touching eyes, nose, or mouth. (binary)
- doctor_recc_h1n1 - H1N1 flu vaccine was recommended by doctor. (binary)
- doctor_recc_seasonal - Seasonal flu vaccine was recommended by doctor. (binary)
- chronic_med_condition - Has any of the following chronic medical conditions: asthma or an 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. (binary)

- child_under_6_months - Has regular close contact with a child under the age of six months. (binary)
- health_worker - Is a healthcare worker. (binary)
- health_insurance - Has health insurance. (binary)
- opinion_h1n1_vacc_effective - Respondent's opinion about H1N1 vaccine effectiveness.
    - 1 = Not at all effective; 2 = Not very effective; 3 = Don't know; 4 = Somewhat effective; 5 = Very effective.
- opinion_h1n1_risk - Respondent's opinion about risk of getting sick with H1N1 flu without vaccine.
    - 1 = Very Low; 2 = Somewhat low; 3 = Don't know; 4 = Somewhat high; 5 = Very high.
- opinion_h1n1_sick_from_vacc - Respondent's worry of getting sick from taking H1N1 vaccine.
    - 1 = Not at all worried; 2 = Not very worried; 3 = Don't know; 4 = Somewhat worried; 5 = Very worried.
- opinion_seas_vacc_effective - Respondent's opinion about seasonal flu vaccine effectiveness.
    - 1 = Not at all effective; 2 = Not very effective; 3 = Don't know; 4 = Somewhat effective; 5 = Very effective.
- opinion_seas_risk - Respondent's opinion about risk of getting sick with seasonal flu without vaccine.
    - 1 = Very Low; 2 = Somewhat low; 3 = Don't know; 4 = Somewhat high; 5 = Very high.
- opinion_seas_sick_from_vacc - Respondent's worry of getting sick from taking seasonal flu vaccine.
    - 1 = Not at all worried; 2 = Not very worried; 3 = Don't know; 4 = Somewhat worried; 5 = Very worried.
- age_group - Age group of respondent.
- education - Self-reported education level.
- race - Race of respondent.
- sex - Sex of respondent.
- income_poverty - Household annual income of respondent with respect to 2008 Census poverty thresholds.
- marital_status - Marital status of respondent.
- rent_or_own - Housing situation of respondent.
- employment_status - Employment status of respondent.
- hhs_geo_region - 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.
- census_msa - Respondent's residence within metropolitan statistical areas (MSA) as defined by the U.S. Census.
- household_adults - Number of *other* adults in household, top-coded to 3.
- household_children - Number of children in household, top-coded to 3.
- employment_industry - Type of industry respondent is employed in. Values are represented as short random character strings.

- employment_occupation - Type of occupation of respondent. Values are represented as short random character strings.

Data exploration report

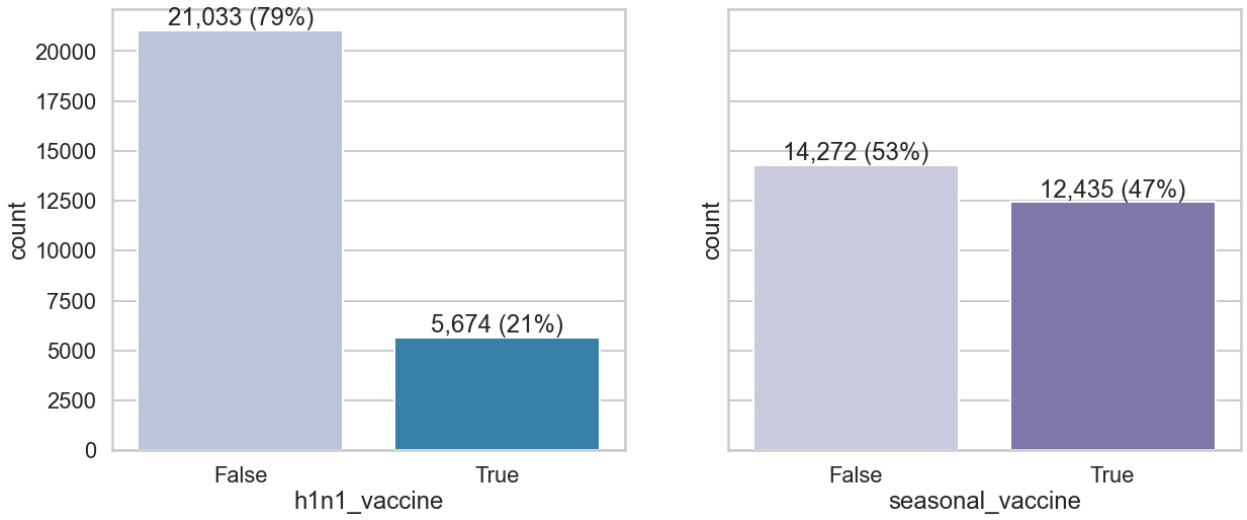
A vaccination outcome has two possible outcomes for each observation (person), vaccinated or not vaccinated, with one trial (one vaccination event). The probability mass function for a Bernoulli distribution (vaccination phenomena) is given by Equation 2.

$$P(x) = \begin{cases} 1-p, x=0 \\ p, x=1 \end{cases} \quad (2)$$

The distribution is represented in Figure 1, utilizing the data in "Training Features", x – axis is shared between charts for observations on h1n1 vaccination status and seasonal vaccination status while the hue represents True or False vaccinated. Chance for seasonal vaccination in the sample is parallel with 47% True and 53% False while h1n1 vaccination is much more leaning towards False with 79%.

Figure 1

H1n1 and Seasonal Vaccination Status



The data contains 35 features, the study shall determine the importance and relationship of the different variables. For seasonal vaccination the age group appears to have a positive association, the higher the age group the higher probability to have the vaccination, see Figure 2, while with h1n1 vaccination there seems to be no difference between age groups, see Figure 3. We establish alternative hypothesis that Age, Sex and Income have a statistical significance, with a corresponding null hypothesis were Age, Sex and Income have no statistical significance.

Figure 2

Seasonal Vaccination by Age Group

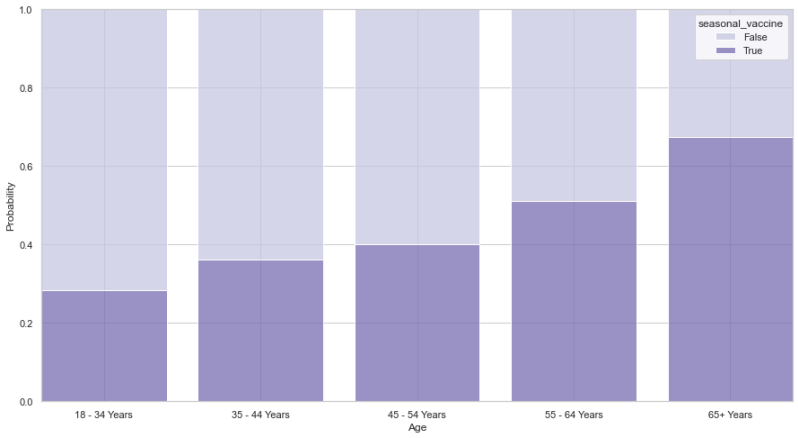
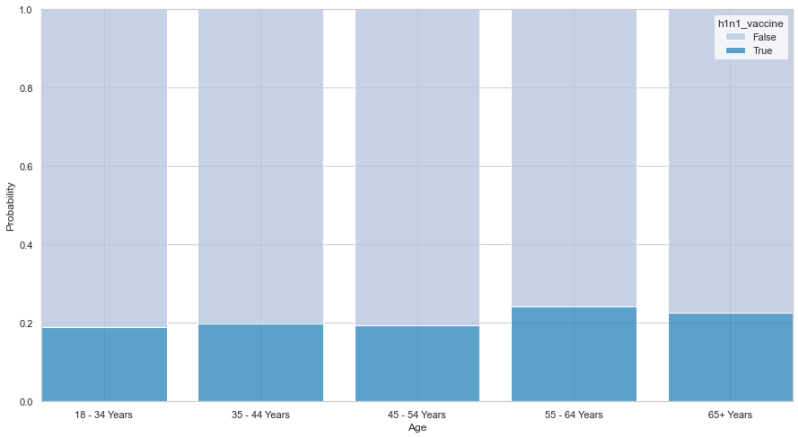


Figure 3

H1n1 Vaccination by Age Group



Employment industry appears to be a contributing factor for both vaccines the industry “haxffmxo” and “fcxhlnwr” have higher probability then other industries for season vaccine, see Figure 4 as well as for h1n1 vaccination, see Figure 5. This may not be a relative factor for Parker as most of its divisions would all in similar industries.

Figure 4

Seasonal Vaccination by Employment Industry

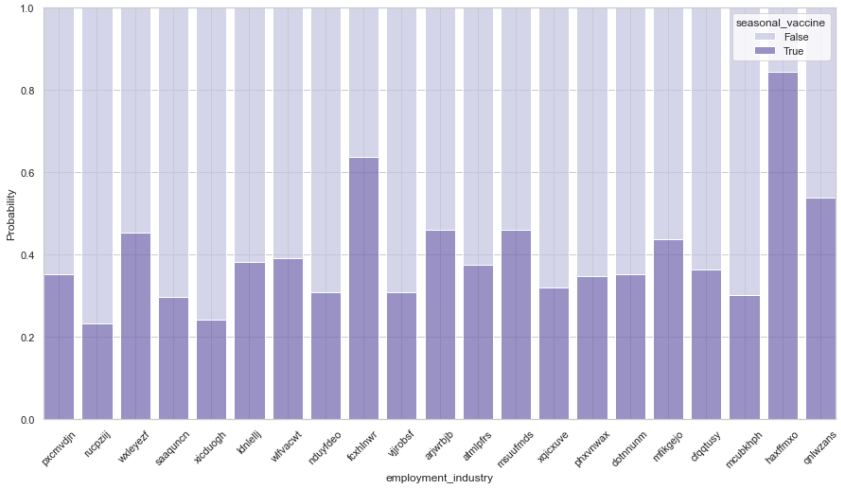
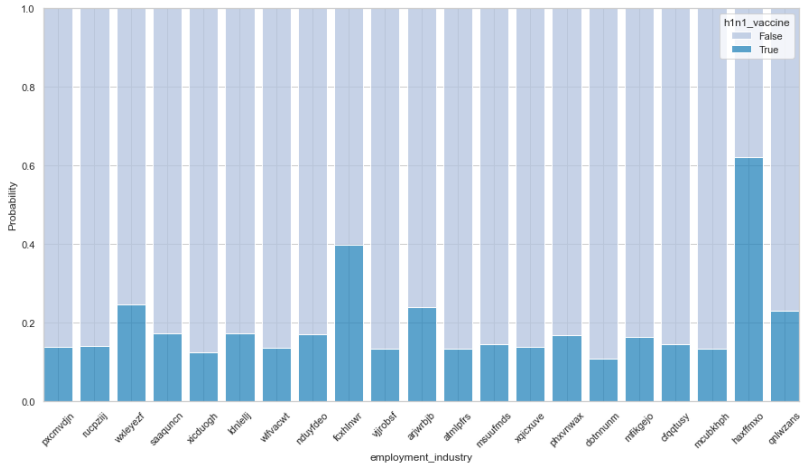


Figure 5

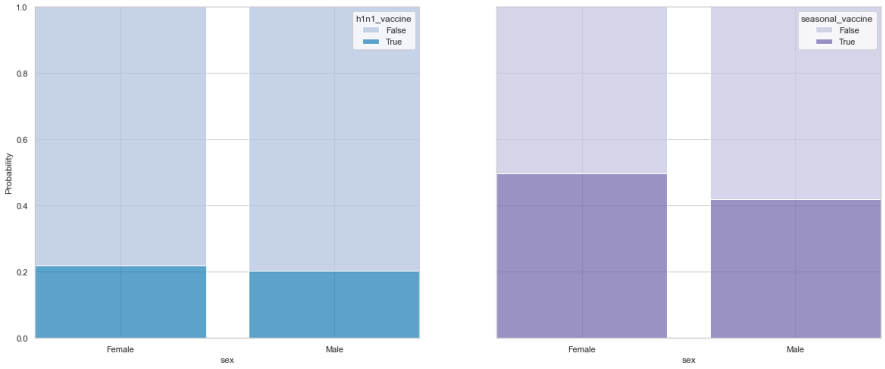
H1n1 Vaccination by Employment Industry



It would be hypotheses that there is no difference between sexes on vaccination rates but once reviewing the data Females have higher probability of taking season vaccine over males, with respect to h1n1 vaccine sex does not appear to be a contributing factor, see Figure 6. This may help towards the reducing downtime due to sick time by making sure we have a common sex ration in our value stream groups.

Figure 6

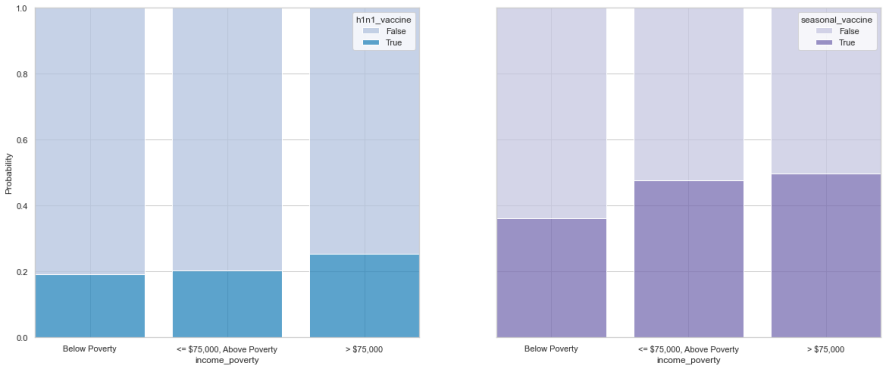
H1n1 and Seasonal Vaccination by Sex



Income appears to have an impact as well, season vaccinations has higher for income above poverty level, while h1n1 is higher until income is above 75K, see Figure 7. This feature may not be as impactful as our working groups compensation gap does not have a high distribution.

Figure 7

H1n1 and Seasonal Vaccination by Income



Data quality report

As part of the data exploration study Table 4 was constructed. The “% Miss” column shows the percentage of missing values per feature; “health_insurance”, “employment_industry”, and “employment_occupation” have more than 45% missing values and shall be dropped from the study. Further study should examen the survey and data collection plan to account for large missing values in these categories.

The cardinality (unique options) for each feature do not highlight any large abnormality, hhs_geo_region has a cardinality of ten which represents geographic classification defined by the U.S. Dept. of Health and Human Services, further study may require more in-depth regions.

The data set contains approximately 80% subjects that identified race as white this large amount may have bias on race, conclusion on the study should be mindful of underestimating other races due to lack of sample sizes.

Figure 4

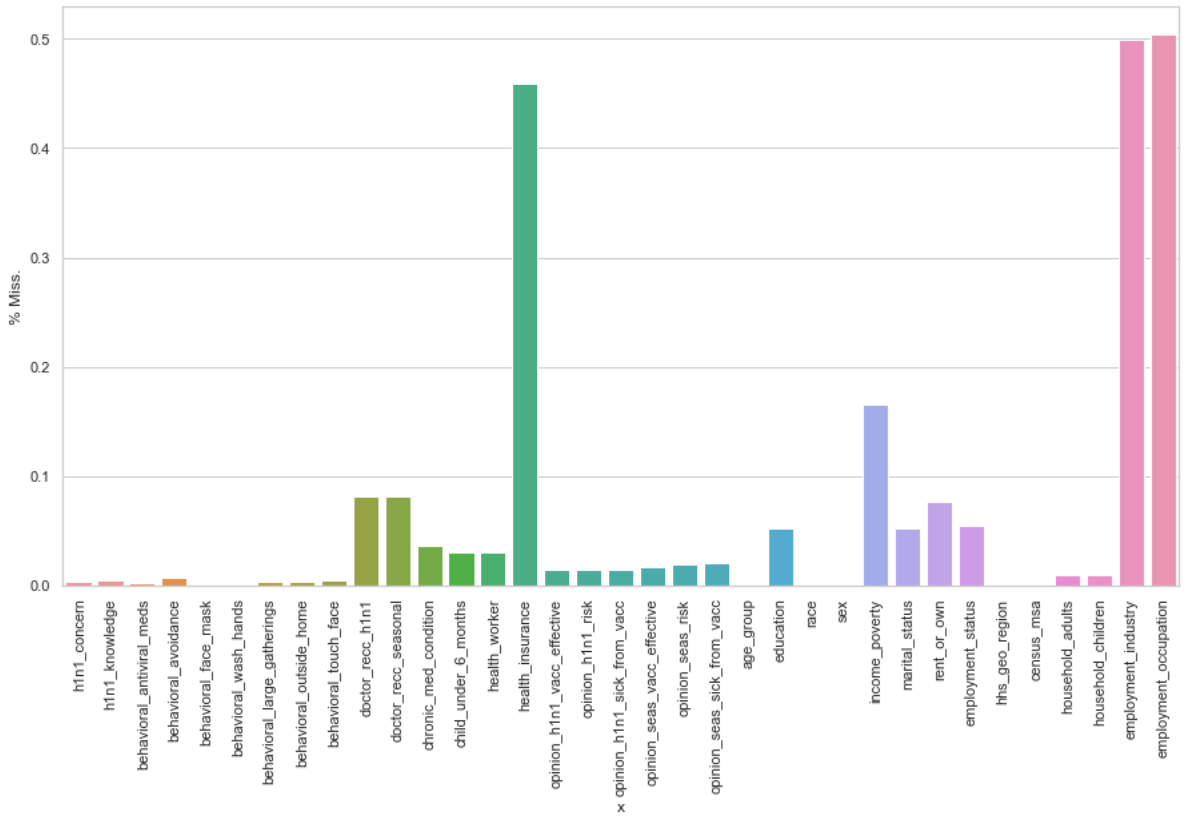
H1n1 Data Quality Report for Training Features

	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2nd Mode	2nd Mode Freq.	2nd Mode %
h1n1_concern	26,615	0.3%	4	2	10,575	40%	1	8,153	31%
h1n1_knowledge	26,591	0.4%	3	1	14,598	55%	2	9,487	36%
behavioral_antiviral_meds	26,636	0.3%	2	0	25,335	95%	1	1,301	5%
behavioral_avoidance	26,499	0.8%	2	1	19,228	72%	0	7,271	27%
behavioral_face_mask	26,688	0.1%	2	0	24,847	93%	1	1,841	7%
behavioral_wash_hands	26,665	0.2%	2	1	22,015	82%	0	4,650	17%
behavioral_large_gatherings	26,620	0.3%	2	0	17,073	64%	1	9,547	36%
behavioral_outside_home	26,625	0.3%	2	0	17,644	66%	1	8,981	34%
behavioral_touch_face	26,579	0.5%	2	1	18,001	67%	0	8,578	32%
doctor_recc_h1n1	24,547	8.1%	2	0	19,139	72%	1	5,408	20%
doctor_recc_seasonal	24,547	8.1%	2	0	16,453	62%	1	8,094	30%
chronic_med_condition	25,736	3.6%	2	0	18,446	69%	1	7,290	27%
child_under_6_months	25,887	3.1%	2	0	23,749	89%	1	2,138	8%
health_worker	25,903	3.0%	2	0	23,004	86%	1	2,899	11%
health_insurance	14,433	46.0%	2	1	12,697	48%	0	1,736	7%
opinion_h1n1_vacc_effective	26,316	1.5%	5	4	11,683	44%	5	7,166	27%
opinion_h1n1_risk	26,319	1.5%	5	2	9,919	37%	1	8,139	30%
opinion_h1n1_sick_from_vacc	26,312	1.5%	5	2	9,129	34%	1	8,998	34%
opinion_seas_vacc_effective	26,245	1.7%	5	4	11,629	44%	5	9,973	37%
opinion_seas_risk	26,193	1.9%	5	2	8,954	34%	4	7,630	29%
opinion_seas_sick_from_vacc	26,170	2.0%	5	1	11,870	44%	2	7,633	29%
age_group	26,707	0.0%	5	65+ Years	6,843	26%	55 - 64 Years	5,563	21%
education	25,300	5.3%	4	College Graduate	10,097	38%	Some College	7,043	26%
race	26,707	0.0%	4	White	21,222	79%	Black	2,118	8%
sex	26,707	0.0%	2	Female	15,858	59%	Male	10,849	41%
income_poverty	22,284	16.6%	3	<= $75,000, Above Poverty	12,777	48%	> $75,000	6,810	25%
marital_status	25,299	5.3%	2	Married	13,555	51%	Not Married	11,744	44%
rent_or_own	24,665	7.6%	2	Own	18,736	70%	Rent	5,929	22%
employment_status	25,244	5.5%	3	Employed	13,560	51%	Not in Labor Force	10,231	38%
hhs_geo_region	26,707	0.0%	10	lzgpxyit	4,297	16%	fpwskwrf	3,265	12%
census_msa	26,707	0.0%	3	MSA, Not Principle City	11,645	44%	MSA, Principle City	7,864	29%
household_adults	26,458	0.9%	4	1	14,474	54%	0	8,056	30%
household_children	26,458	0.9%	4	0	18,672	70%	1	3,175	12%
employment_industry	13,377	49.9%	21	fcxhlnwr	2,468	9%	wxleyezf	1,804	7%
employment_occupation	13,237	50.4%	23	xtkaffoo	1,778	7%	mxkfnird	1,509	6%

Figure 5 shows the percentage of missing values for each feature, it clearly shows that employment_ocupation, employement_industry, and health insurance are features that may need to be discard because practically half of the subjects are missing values.

Figure 5

Percentage of Missing Values Per Feature



References

Parker. (n.d.) *About Parker.* Parker. Retrieved May 14, 2022, https://www.parker.com/portal/site/PARKER/menuitem.f830ba32f37af5fe2c5c8810427ad1ca/?vgnextoid=7de94bad565e4310VgnVCM10000014a71dacRCRD&vgnextfmt=default

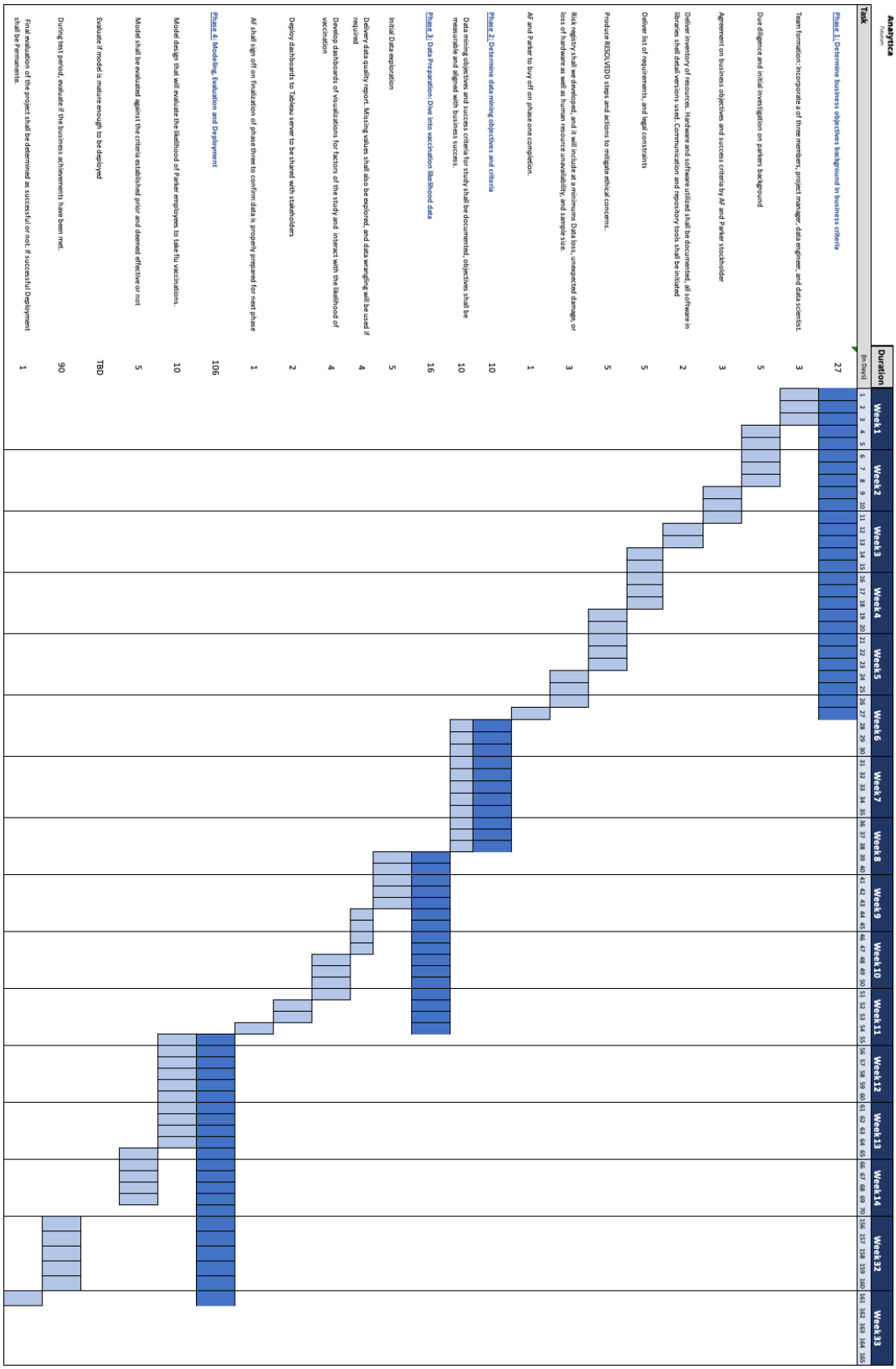
Driven Data (n.d.). *Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines.* Driven Data. Retrieved May 14, 2022, https://www.drivendata.org/competitions/66/flu-shot-learning/page/211/

CDC (n.d.). *Vaccine effectiveness: How well do flu vaccines work?.* Centers for Disease Control and Prevention. Retrieved May 21, 2022, https://www.cdc.gov/flu/vaccines-work/vaccineeffect.htm

Kostal, S. (2021, July 7). *Here's what California employers should know about CCPA compliance*. SHRM. Retrieved June 6, 2022, from https://www.shrm.org/resourcesandtools/legal-and-compliance/state-and-local-updates/pages/what-california-employers-should-know-about-ccpa-compliance.aspx

PEP 8 (n.d.). *Python enhancement proposals*. PEP 8 – Style Guide for Python Code. (n.d.). Retrieved June 6, 2022, from https://peps.python.org/pep-0008/

Appendix 1


Analytica
Task
Duration
(in Days)
Week1
Week2
Week3
Week4
Week5
Week6
Week7
Week8
Week9
Week10
Week11
Week12
Week13
Week14
Week32
Week33
Phase 1: Determine business objectives background in business criteria
27
Team formation: Incorporate 4 of three members, project manager, data engineer, and data scientist.
3
Due diligence and initial investigation on partners background
5
Agreement on business objectives and success criteria by AF and Parker stockholder
3
Deliver inventory of resources, Hardware and software utilized shall be documented, all software in libraries shall detail versions used. Communication and repository tools shall be initiated
2
Deliver list of requirements, and legal constraints
5
Produce RESOLVED steps and actions to mitigate ethical concerns.
5
Risk registry shall we developed, and it will include at a minimums Data loss, unexpected damage, or loss of hardware as well as human resource unavailability, and sample size.
3
AF and Parker to buy off on phase one completion.
1
Phase 2: Determine data mining objectives and criteria
10
Data mining objectives and success criteria for study shall be documented, objectives shall be measurable and aligned with business success.
10
Phase 3: Data Preparation: Dive into vaccination likelihood data
16
Initial Data exploration
5
Delivery data quality report. Missing values shall also be explored, and data wrangling will be used if required
4
Develop dashboards of visualizations for factors of the study and interact with the likelihood of vaccination
4
Deploy dashboards to Tableau server to be shared with stakeholders
2
AF shall sign off on finalization of phase three to confirm data is properly prepared for next phase
1
Phase 4: Modeling, Evaluation and Deployment
106
Model design that will evaluate the likelihood of Parker employees to take flu vaccinations.
10
Model shall be evaluated against the criteria established prior and deemed effective or not
5
Evaluate if model is mature enough to be deployed
TBD
During test period, evaluate if the business achievements have been met.
90
Final evaluation of the project shall be determined as successful or not. If successful Deployment shall be Permanente.
1