Human Centered Data Science 2024 - Project Report

by Ferdinand Daniel König, Henry Khalil El-Jawhari, Johannes Englmeier

Project Goals and Target Audience

Our main goal was to create a understandable interface for our chosen target audience. Initially we planed to have more than one type of target user. The idea was to have a dedicated view respecting the profile of the user, showing only the data in which the user is interested, to increase the focus on relevant parts of the application. The users would have been able to switch the presented perspective on their needs.
In our discussions we identified different possible user types and their different needs. We thought about many different groups that could be interested in an application like ours and their special needs: Executors would need to understand the decisions and should be supported with the explanations, while data scientist maybe would focus on debugging the models regarding the data sheets, labels, shapley values, partial dependence or variable importance. Data providers would require the safety of data provision and need to understand how the product is used, while examiners need transparency regarding the shapley values, partial dependence or variable importance. If the user would be from an operational or management context he/she would like to adapt and improve the decisions by our models.
While creating the prototype and getting the first feedback, we realized that there was not enough time to implement the different perspectives for the different target audiences. Therefore we decided to focus on one group and postponed the development for the other perspectives.
Our resulting target audience is employed at a big pharmaceutical company and wants to implement and roll out a new model to understand the target population, that are more likely to use at least one of the vaccines. We assume that our users have a technical background in computer science and are interested in data science, but not experts.

Dataset

Because we had to choose a medical data set, we decided for a set which consist of data about vaccination of US. American citizens.

Context

In 2009 the H1N1 influenza virus, called the swine flu, spread across the world.
For this reason the United states started a phone survey to ask if respondents received a vaccination for seasonal flu and/or for H1N1. Additionally the interviewers asked about the social, economic and

demographic background of the interviewed, as well as their opinions on risk of illness and vaccine effectiveness. Additionally they were asked about their behaviors towards mitigating transmission. The data set is available to the public at the following website:
https://www.drivendata.org/competitions/66/flu-shot-learning/page/210/

Target variables

The set includes two binary target variables: The first one states, if a respondent received a H1N1 vaccination. The second variable contains the information whether a interviewee received a vaccination for the seasonal flu. The interviewed citizens could have no vaccination, one, or both of them.

Features

The data set contains 35 feature columns in total, containing 13 binary features, all available as two CSV-files, one for training, the other one for testing.

Feature Name	**Feature description**
h1n1_concern	Describes the level of concern about the H1N1 flu: 0 = Not at all concerned 1 = Not very concerned 2 = Somewhat concerned 3 = Very concerned
h1n1_knowledge	Describes the knowledge level about the H1N1 flu: 0 = No knowledge 1 = A little knowledge 2 = A lot of knowledge.
behavioral_antiviral_meds	Describes if the interviewee used antiviral medications
behavioral_avoidance	Describes if the respondent had avoided close contacts to others with flu-like symptoms
behavioral_face_mask	Describes if the respondent has bought a face mask
behavioral_wash_hands	Describes if the respondent has frequently washed hands or used hand sanitizer
behavioral_large_gatherings	Describes if the respondent has reduced the time staying at large gathering
behavioral_outside_home	Describes if the respondent has reduced contact with people from outside their own household
behavioral_touch_face	Describes if the respondent has avoided touching the own nose, mouth or eyes
doctor_recc_h1n1	Describes if the H1N1 vaccine has been recommended to the interviewee by a doctor
doctor_recc_seasonal	Describes if the seasonal flu vaccine has been recommended to the interviewee by a doctor

chronic_med_condition	Describes if the respondent 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 • a weakened immune system caused by a chronic illness or by medicines taken for a chronic illness
child_under_6_months	Describes if the respondent has regular contact with a child under the age of six months
health_worker	Describes if the respondent is a healthcare worker
health_insurance	Describes if the respondent has a health insurance
opinion_h1n1_vacc_effective	Describes the opinion of the respondent about the effectiveness of the H1N1 vaccine: 1 = Not at all effective 2 = Not very effective 3 = Don't know 4 = Somewhat effective 5 = Very effective
opinion_h1n1_risk	Describes the opinion of the respondent, how likely it would be to get sick with H1N1 flu when not having been the vaccinated: 1 = Very Low 2 = Somewhat low 3 = Don't know 4 = Somewhat high 5 = Very high
opinion_h1n1_sick_from_vacc	Describes the respondents worry of getting sick from taking H1N1 vaccine
opinion_seas_vacc_effective	Describes the respondents 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	Describes the respondents opinion on getting the seasonal flu without having been vaccinated: 1 = Very Low 2 = Somewhat low 3 = Don't know 4 = Somewhat high

	5 = Very high
opinion_seas_sick_from_vacc	Describes the respondents opinion on getting sick because of the seasonal flu vaccination 1 = Not at all worried 2 = Not very worried 3 = Don't know 4 = Somewhat worried 5 = Very worried
age_group	Describes the age group of the interviewee
education	Describes the self-reported educational level of the interviewee
race	Describes the race of the respondent
sex	Describes the sex of the respondent
income_poverty	Describes the annual income of the respondent and if the income if below or above the poverty threshold given by the Census 2008
marital_status	Describes if the Interviewee is married or not
rent_or_own	Describes if the respondent own his/her place or rents it
employment_status	Describes if the respondent is employed
hhs_geo_region	Describes in which of the 10 different classes of regions, defined by the US. department of Health , the respondent lives in
census_msa	Describes the respondent’s residence within metropolitan statistical areas (MSA) as defined by the U.S. Census.
household_adults	Describes the number of *other* adults in the household of the respondent(top-coded to 3)
household_children	Describes the number of other children in the household (top-coded to 3)
employment_industry	Describes the type of occupation of the respondent. The values are represented as short random character strings.

The Development Process

For our UI prototype we developed twelve different designs using the crazy 4 method. It has been quite surprising to see how the different perspectives of our group members has been visible in the resulting designs. In the end we decided for the most intuitive design which is now visible in our application.

Even if we had many different ideas, the chosen design was one of our first ideas. In our opinion the quality of the designs created in the crazy 4 process, has been decreasing while progressing.

In the end the crazy 4 approach was helpful to understand the different perspectives and possibilities and therefore beneficial for our prototype.

In the next phase we created a mockup for the UI to get an idea of the appearance of the UI and the feeling when interacting with it. Using Mockup-Tools like Figma, Balsamiq or Sketch seemed a bit too much for a small short project like ours. It is quite an effort to get known to the tools. To really profit in the design phase, it would be necessary to be already familiar with the tools, or have more time to get used to them properly. Otherwise the focus on the design itself can easily get lost, while trying to understand how the tools work. Maybe simpler sketching tools like DrawIO, Miro or Excalidraw are already sufficient to get an idea about the aimed design, event if they are not that interactive.

A quite challenging issue has been the performance of the models' explanations and the models themselves. Because of the long computing times it has been necessary to save already created explanations and models into dedicated pickle-files. Additionally the application faces long loading periods, when the user switches to other tabs. This maybe correlates with the size of the models.

If we would have more time for the project, the performance could be optimized definitely.

For the team, we set up two regular meetings a week, where we did our discussions, compared results and planed the upcoming tasks. We did one meeting in person, while the other one has been done remotely. To improve the communication, we used a Signal messaging group to share thoughts and ask questions. Additionally we set up a git repository for the results.

We tried to separate the work into into several parallel task, that could be split in between the team members, to share the effort, with quite a success.

All in all the process worked quite well and would have been getting even better, if we had more time to get into the project and the process. Also it would have been beneficial if we had known the upcoming tasks more in advance, to achieve better planing and scheduling.