Octagon Data Science in Health Competition

Alexey Strokach, Dar’ya Redka

Abstract

As part of the 2019 Octagon Data Science in Health Competition, we are provided a dataset containing the type of data that may be obtained in a small asthma study. Our goal was to analyse this dataset and present our most interesting findings and insights that may be gleaned into the disease. First, we performed exploratory data analysis, in which we uncover some idiosyncrasies and limitations of the provided dataset. Next, we combine the provided data into a set of features describing each patient and an outcome—average change in ACQ score after the initial questionnaire—which we expect captures the main effect that this study is trying to capture. We train two machine learning models, a simple decision tree classifier and a gradient boosted decision tree classifier, in order to evaluate how well the available information can predict a decrease in ACQ score for an individual over the course of the study. We find that age is the most predictive feature, with individuals under 35 years of age being less likely to show a decrease in the ACQ score over the course of the study, while individuals who are retired or who have a cardiovascular disease (CVD) score less than 0.4 are more likely to show a decrease in the ACQ score over the course of the study.

Methods

Data preparation

In the first part of our analysis, we combined data from the 6 sheets of the provided excel file into a single dataframe, using the *id* column to join each of the tables. We found that there are two pairs of disjoint tables: *Demographics* and *Demographics(1)* and *medhistory* and *medhistory(1)*. In cases where the tables contained similar information, we attempted to unify this information in a single format, as described below. We also corrected and / or discarded some of the provided data when found that it is likely to be invalid.

Demographics table

The *gender* column in the *Demographics* table takes on values of 0, 1, 2, or 3. Without additional information, it is difficult to interpret this column in reference to the traditional meaning of the word "gender". Furthermore, the values of the gender column show a surprisingly high correlation with the decrease in asthma symptoms as evinced by the decrease in the ACQ score of the participants (~0.4 Spearman correlation, see Table 1). For this reason, *gender* was not used as a feature when training machine learning algorithms.

Demographics(1) table

We converted the *birthyear* column into age in order to be consistent with the values in the *Demographics* column. The age was calculated by subtracting the the birthyear from the year of assessment (the *assess* column) and thereby may be incorrect by one year.

Medhistory(1) table

We added a BMI column using the equation below because we expected that it would correlate better with health-related outcomes than either height or weight.

$$\mathrm{BMI} = \frac{\mathrm{weight}}{\mathrm{height}^2}$$

ACQ table

The dates provided in the ACQ table raise some suspicion. First, there are 12 rows where the date of at least one procedure is in year 1899 or 1990. It appears unlikely that this study would include data collected almost 30 years ago. A more plausible explanation is that those dates were not entered correctly and correspond to the default value of the database used for data entry or storage. There are also 18 rows which contain ACQ scores for dates at some point in the future. While we did not exclude rows containing these exceptional dates from our analysis, we think that these findings are worrisome and would warrant further discussion with the data provider.

Furthermore, in 29 rows in this table, at least one of the provided ACQ scores is greater than 6. Since the ACQ is a 7-point score with ranges from 0 to 6 [1], we considered rows with scores outside this range as corrupt and excluded them from our analysis. Indeed, we can observe distinct differences in the ACQ score trends displayed by individuals with scores within the 0 to 6 range when compared to individuals with scores outside this range (Figure 2A).

Outcome metrics

Throughout our analysis, we utilize two metrics, *ΔACQ* and *ΔACQ-lt1*, in order to capture the changes in ACQ scores throughout the study. We define *ΔACQ* as the difference between the average ACQ score for the four follow-up measurements and the initial ACQ score. The *ΔACQ* outcome metric was chosen to account for patients whose ACQ score goes down initially but then comes back up as well as for patients who missed one or more of the follow-up visits.

$$\Delta ACQ \ = \ \sum_{i=2}^{5} ACQ_i \ - \ ACQ_1$$

We define *ΔACQ-lt1* as value indicating if the ACQ score decreased by more than one between the first measurement and the average of the four follow-up measurements. We chose the -1 threshold because it should correspond to a small but substantial improvement (a difference in ACQ score of 0.5 is thought to be the minimally important difference [4]).

$$\Delta ACQ\ lt1 \ = \Delta ACQ \ < \ -1$$

Machine learning models

We used the decision tree classifier implemented in scikit-learn [2] and a stochastic gradient boosted decision tree classifier implemented in LightGBM [3]. For each classifier, the parameters were selected using grid-search and the accuracy of the models were evaluated using stratified 10-fold cross-validation. The accuracies displayed in Figure 4 and 5 correspond to predictions made by the models on the test subsets of each train-test split that was evaluated during cross-validation.

Software availability

The code required to reproduce the analysis presented in this report can be accessed at: https://github.com/dashapyly/octagon-data-science-competition.

Results / Discussion

First, we calculated the Spearman’s correlation coefficient between the demographic and medical history features and *ΔACQ*: the difference in ACQ score between the first administration of the questionnaire and the average of the subsequent four administrations of the questionnaire (Table 1). However, we do not observe any statistically-significant correlations. While *gender* shows the highest Spearman's ρ of -0.39 (p value: 0.068), as discussed above, the *gender* column appears suspicious and was not used in our analysis.

Next, we calculated the Spearman's correlation coefficient between the demographic and medical history features and *ΔACQ-lt1*: a binary value indicating whether the average ACQ score for follow-up treatments 2 to 5 is more than one point lower than the ACQ score obtained during the first measurement (Table 2). In this case, we observe that *age* and *retire* show significant correlations (before adjusting for multiple comparisons). In order to examine in more depth the ability of the various features in our dataset to predict the change in ACQ scores, we trained a decision tree classifier (Figures 3 and 4) and a stochastic gradient-boosted decision tree (GBT) model (Figure 5) to predict ACQ-lt1.

The model was trained to predict the difference between the average ACQ score for visits 2-5 and the ACQ score for visit 1. We trained a simple decision tree classifier to predict whether the average ACQ score reported during follow-up visits was at least 1 point lower than the ACQ score during the first visit.If "split", result contains numbers of times the feature is used in a model.

Due to the small size of the dataset, it is difficult to make decisive conclusions about the effect that any of the provided demographic and medical history features have on the change in ACQ scores. The strongest evidence that we have is for age and retirement status, which have Spearman's correlation of 0.202 and 0.266 with *ΔACQ-lt1* (p-values of 0.0157 and 0.0238, respectively; Table 1). Both age and retirement status were two of the strongest features in the GBDT classifier (Figure 5), while age was also the strongest feature in the decision tree classifier (Figure 3). While cardiovascular disease (CVD) showed a relatively modest Spearman's correlation of -0.124 (p-value: 0.205), this feature was an important component of both machine learning models and, in the case of GBDT, we observe an interesting trend where CVD is especially predictive of *ΔACQ-lt1* when it occurs in older patients (Figure 5D, E).

We observe that pneumonia comorbidity has a strong correlation with the first ACQ score (Spearman's : 31.77, p-value: 0.16, data not shown). This correlation is not significant, largely due to the small number of participants with pneumonia (N=21). However, it would be reasonable for pneumonia to have an effect on asthma symptoms as both disorders negatively affect the respiratory system. Therefore, this correlation warrants further investigation, and people with pneumonia should potentially be excluded from future asthma studies in order to prevent the introduction of confounding factors.

References

1. Juniper EF, O'Byrne PM, Guyatt GH, Ferrie PJ, King DR. Development and validation of a questionnaire to measure asthma control. Eur Respir J. 1999;14: 902–907.
2. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12: 2825–2830.
3. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems 30. Curran Associates, Inc.; 2017. pp. 3146–3154. Available: http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf
4. American Thoracic Society - Asthma Control Questionnaire (ACQ). [cited 4 Nov 2019]. Available: https://www.thoracic.org/members/assemblies/assemblies/srn/questionaires/acq.php

Appendix

Tables

Table 1. Spearman's correlation coefficient between the demographic and medical history features provided in the dataset and ΔACQ, the change in ACQ scores between the first administration of the questionnaire and the average of the subsequent four administrations of the questionnaire. N corresponds to the number of rows where the given feature is not null.

Feature name	Spearman's ρ	p-value	N
gender	-0.3964	0.0678	22.0000
retire	-0.1995	0.0930	72.0000
arthritis	0.1514	0.1214	106.0000
vaccine	-0.1507	0.1230	106.0000
CVD	0.1419	0.1469	106.0000
age	-0.1186	0.1582	143.0000
sex	-0.1190	0.1935	121.0000
smoking	-0.1234	0.2077	106.0000
allergy	0.1093	0.2648	106.0000
polyps	-0.1040	0.2885	106.0000
eos300	0.0901	0.2969	136.0000
Prev-biologic	0.0944	0.3359	106.0000
race	-0.0390	0.6436	143.0000
co-yellowfever	-0.0366	0.6644	1.0000
co-heartfailure	0.0977	0.6818	20.0000
co-asthma	-0.0827	0.7289	20.0000
co-pneumonia	0.0766	0.7413	21.0000
co-diabetes	0.0727	0.7541	21.0000
co-copd	0.0966	0.7653	12.0000
disab	0.0251	0.7664	143.0000
work	-0.0141	0.8676	143.0000
weight	0.0162	0.8691	106.0000
bmi	-0.0147	0.8813	106.0000
height	0.0129	0.8959	106.0000
co-arthritis	-0.0101	0.9043	5.0000
demographics_table	0.0009	0.9911	143.0000
medhistory_table	0.0009	0.9911	143.0000

Table 1. Spearman's correlation coefficient between the demographic and medical history features provided in the dataset and *ΔACQ-lt1*, a boolean value which is true if the ACQ scores decreased by more than one point between the first administration of the questionnaire and the average of the subsequent four administrations of the questionnaire. N corresponds to the number of rows where the given feature is not null.

Feature name	Spearman's ρ	p-value	N
age	0.2017	0.0157	143.0000
retire	0.2663	0.0238	72.0000
sex	0.1554	0.0887	121.0000
co-diabetes	-0.3464	0.1240	21.0000
race	0.1172	0.1632	143.0000
gender	0.2844	0.1996	22.0000
CVD	-0.1242	0.2047	106.0000
co-pneumonia	-0.2066	0.3689	21.0000
co-yellowfever	0.0734	0.3835	1.0000
arthritis	-0.0702	0.4746	106.0000
allergy	-0.0691	0.4818	106.0000
disab	-0.0576	0.4945	143.0000
Prev-biologic	-0.0655	0.5049	106.0000
vaccine	0.0649	0.5086	106.0000
work	0.0503	0.5506	143.0000
co-asthma	0.1307	0.5828	20.0000
co-heartfailure	-0.1307	0.5828	20.0000
eos300	-0.0468	0.5885	136.0000
co-copd	0.1690	0.5995	12.0000
smoking	0.0377	0.7013	106.0000
medhistory_table	0.0181	0.8306	143.0000
demographics_table	0.0181	0.8306	143.0000
weight	-0.0203	0.8366	106.0000
height	-0.0169	0.8638	106.0000
co-arthritis	0.0129	0.8785	5.0000
bmi	0.0065	0.9469	106.0000
polyps	0.0032	0.9744	106.0000

Table 2. LightGBM parameters which were evaluated using grid-search. The accuracy of each trained classifier was evaluated using 10-fold cross-validation, and the parameters that produced a classifier with the highest cross-validation accuracy are provided in the "Final value" column.

Parameter name	Values attempted	Final value
boosting_type	`["gbdt"]`	`"gbdt"`
learning_rate	`[0.001, 0.005, 0.01, 0.05, 0.1, 0.2]`	`0.1`
num_leaves	`[2, 4, 6, 8, 12, 16, 24, 32]`	`6`
num_iterations	`[5, 10, 20, 50, 100, 200]`	`5`
min_sum_hessian_in_leaf	`[0]`	`0`
min_data_in_leaf	`[1, 5, 10, 20]`	`10`
max_bin	`[15, 31, 63, 123, 255]`	`31`

Figures

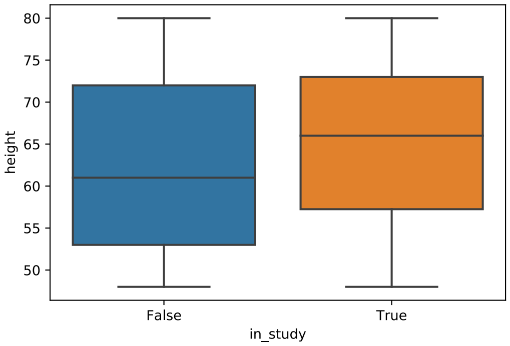

Figure 1. Box-plot showing the difference in height between the individuals who were enrolled in the study (as evinced by their presence in the ACQ table) and individuals who were *not* enrolled in the study. The difference is significant (p-value < 0.05; two-sided t-test).

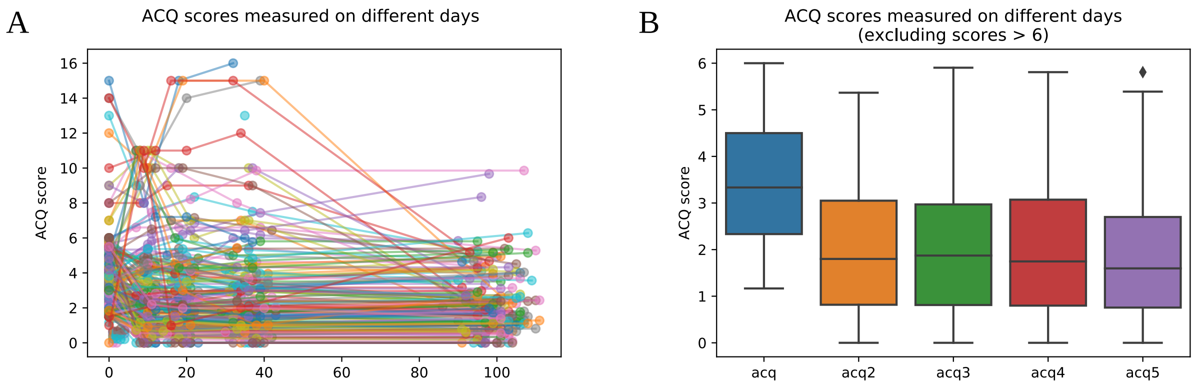

Figure 2. **(A)** Line-plot showing changes in ACQ scores of all individuals enrolled in the study. Most individuals had four follow-up measurements around 10, 20, 40, and 100 days after the initial survey was conducted. Participants whose ACQ score was reported to be higher than 6 were excluded from our analysis since the maximum ACQ score is 6, and the pattern of ACQ scores for those individuals is distinctly different from the pattern shown by the majority of the participants. **(B)** Box-plot showing the distribution in ACQ scores for the initial visit and the four subsequent follow-ups for participants whose reported ACQ score was ≤ 6 throughout the study.

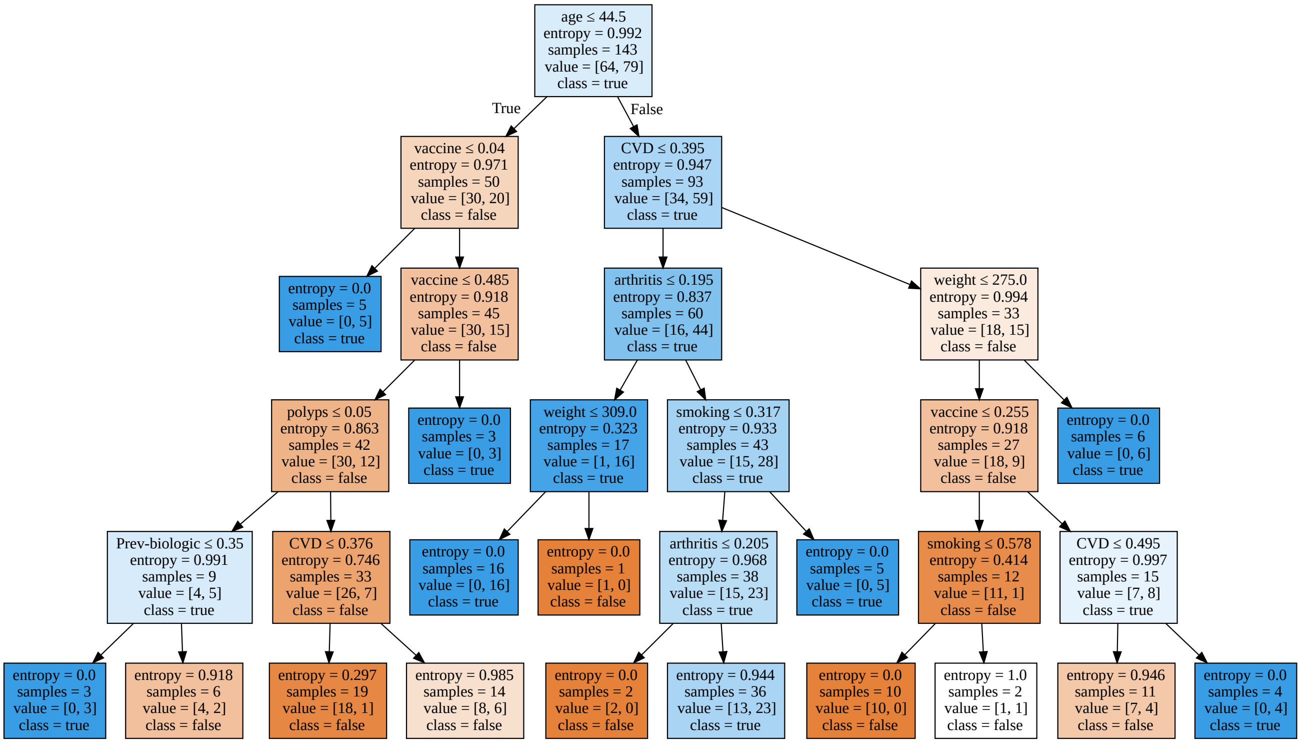

Figure 3. Visualization of a decision tree trained to predict whether the ACQ score of an individual will decrease by at least 1 point on average over the four follow-up surveys. A decision tree allows for intuitive visualization of how the interaction between different features is predictive of a particular outcome.

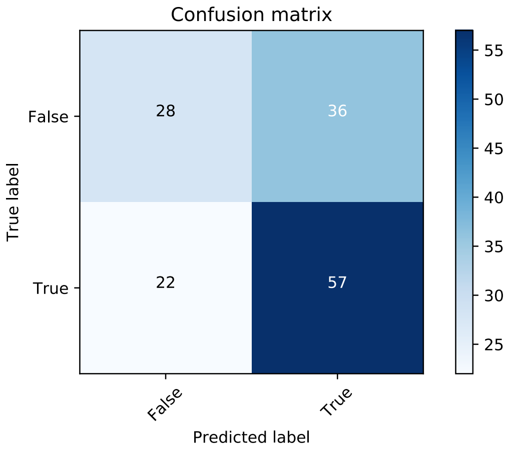

Figure 4. A confusion matrix corresponding to the decision tree displayed in Figure 3. The predicted values shown in the matrix correspond to predictions made by the trained decision trees in 5-fold cross-validation. The overall validation accuracy of the decision trees is 59.44%.

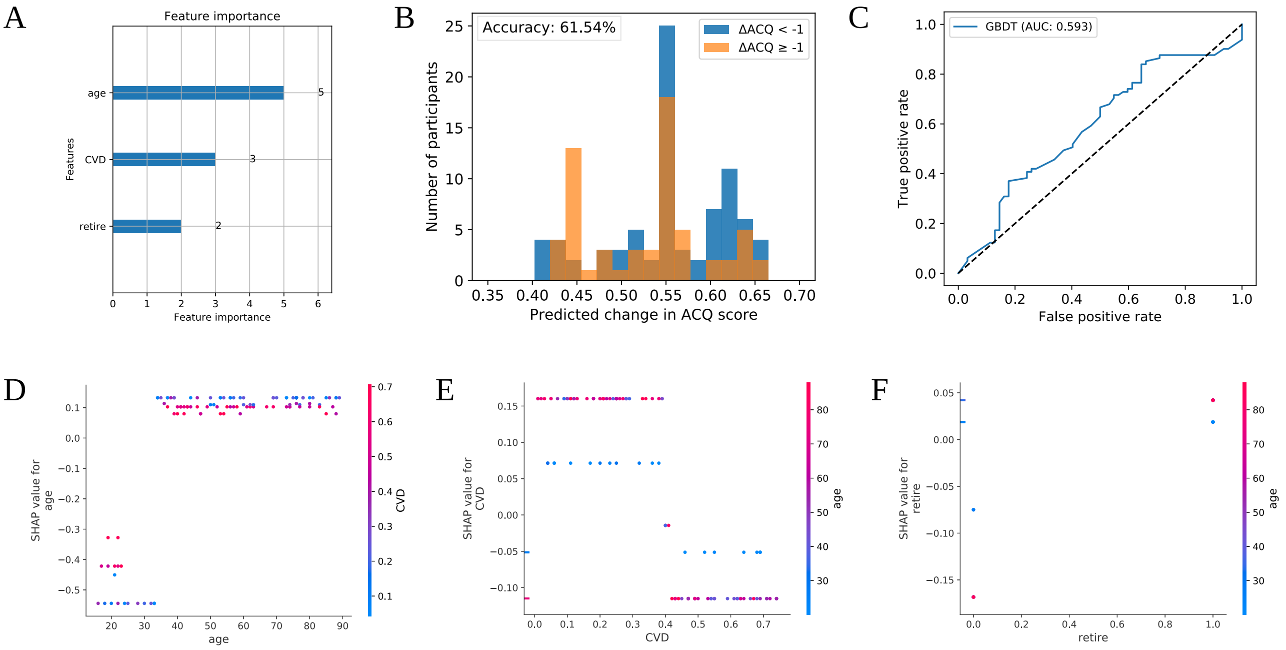

Figure 5. **(A)** Feature importance of a stochastic gradient-boosted decision tree classifier (SGBDT) trained to predict whether an individual will have a decrease in their average ACQ score over the four follow-up surveys of at least 1 point. Age is the most important feature, involved in 5 splits inside the decision trees, followed by CVD, involved in 3 splits, and *retire*, involved in 2 splits. **(B)** Histogram showing predictions made by trained SGBDT classifiers during 10-fold cross-validation for individuals who either showed a decrease in the average ACQ score over the four follow-up surveys of at least 1 point (blue) or did not show such a decrease (orange). With a threshold of 0.5, the SGBDT classifier shows an accuracy of 61.54%, which is marginally higher than the accuracy achieved by the decision tree classifier (see Figure 4). **(C)** The receiver operator characteristic curve showing the trade-off between the false positive rate and the true positive rate for predictions made by trained SGBDT classifiers during 10-fold cross-validation. The area under the receiver operator characteristic curve is 0.593, which is higher that what would be expected from a classifier guessing at random (black dashed line). **(D-F)** Partial dependence plots showing the effect that *age* **(D)**, *CVD* **(E)**, and *retire* **(F)** features have on the predictions made by the classifier. An age greater than 35, a CVD less than 0.4, and being retired, are all associated with a higher probability of an at least 1 point decrease in the ACQ score over the course of the study. In younger individuals, the CVD score has a decreased effect on the prediction of the network **(E)**, while in people with a low CVD score, age has an *increased* effect on the prediction of the network **(D)**.