

# 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.

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

## 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,  $\Delta ACQ$  and  $\Delta ACQ\text{-}lt1$ , in order to capture the changes in ACQ scores throughout the study. We define  $\Delta ACQ$  as the difference between the average ACQ score for the four follow-up measurements and the initial ACQ score. The  $\Delta ACQ$  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  $\Delta ACQ\text{-}lt1$  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]).

$$\Delta ACQ\text{-}lt1 = \Delta ACQ < -1$$

## 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  $\Delta ACQ$ : 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  $\rho$  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  $\Delta ACQ-lt1$ : 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  $\Delta ACQ-lt1$  (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  $\Delta ACQ-lt1$  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  $\rho$ : 0.3177, 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/questionnaires/acq.php>