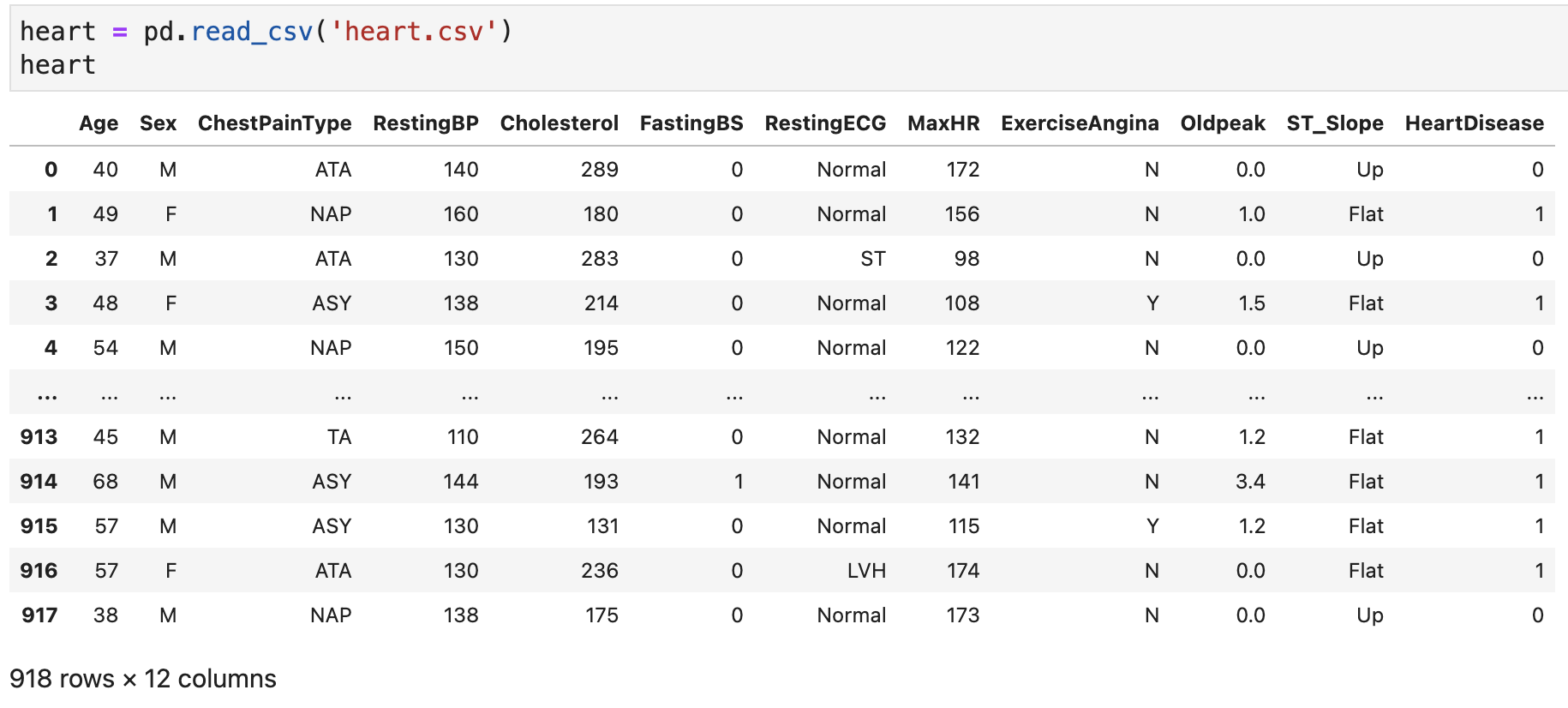
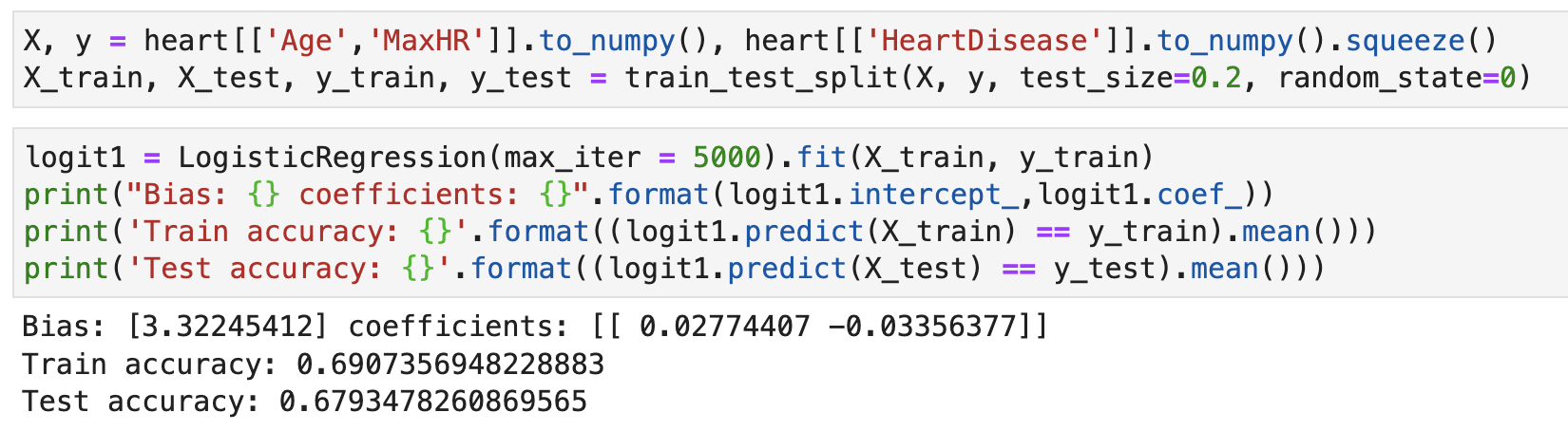
**CS5228 Tutorial 4 – Logistic Regression & Recommender Systems**

**Q1: Logistic Regression**

Cardiovascular disease (CVD) is the leading cause of death globally, representing about 32% of global deaths[[1]](#footnote-1). We will use a “Heart Failure Prediction Dataset”[[2]](#footnote-2), a dataset with 1190 observations, and 11 common clinical features. (Optional: the Jupyter notebook for the analysis in this question can be found in the class Canvas under tutorials).



For simplicity and visualization, we will focus on only 2 features, age (in years) and maximum heart rate, denoted Age and MaxHR respectively. We first split the data, then fit a simple logistic regression model:



**1a)** Interpret the coefficients and comment on the accuracy achieved.

Higher age leads to higher heart disease risk (specifically, since 1/0.277 ≈ 36, an increase in 36 years corresponds to an increase in the log-odds of heart disease by 1).

(Note: the logistic regression model predicts the positive class with probability , and the logit or “log-odds ratio” refers to log(/(1- )). Logistic regression assumes this to be a linear function of the parameters, i.e. )

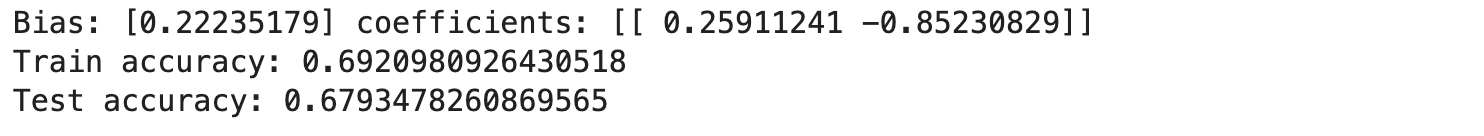
Lower max heart rate leads to higher heart disease risk (since 1/0.0335 ≈ 30, a decrease in 30 BPM corresponds to an increase in the log-odds of heart disease by 1)[[3]](#footnote-3).

The accuracy achieved is generally low, which is not surprising since we are using a very limited set of only 2 features. The training and test accuracy are relatively close, suggesting that not much overfitting is happening (also not surprising since this is a simple model with only 2 features).

**1b)** Let us now standardize the features, as follows:



We then re-run logistic regression as before.



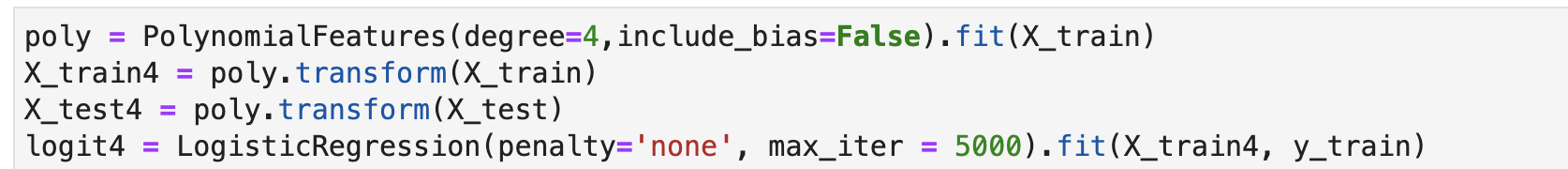
First notice that the coefficients have greatly changed. Explain your interpretation of these new coefficients.

* An increase of one standard deviation of Age corresponds to 0.259 **increase** in the log-odds of heart disease
* An increase of one standard deviation of MaxHR corresponds to 0.852 **decrease** in the log-odds of heart disease.

**1c)** Notice that the accuracy has changed slightly as a result of the standardization, from 0.6907 to 0.6921. Recall that during lecture we discussed that logistic regression is invariant to scaling of the features (since any scaling of the features can be “undone” by the opposite scaling to the coefficients). In this case, why does the accuracy change?

As mentioned in lecture, scikit-learn logistic regression uses L2 regularization by default (with a default “inverse penalty” parameter of C=1). As such, the two versions of logistic regression above (with and without standardization) are not equivalent as they apply different regularization penalties from one another.

**1d)** To allow for a more flexible model, we can use polynomial features, which allow interactions between the two variables up to a certain limit of the degree, as follows:



X\_train4 and X\_test4 contain 14 columns. Explain why that is the case.

X\_train has 2 columns: for convenience, let’s call the variables u and v. If we allow polynomials of degree up to 4, we have the following

u, v,

u2, uv, v2,

u3, u2v, uv2, v3,

u4, u3v, u2v2, uv3, v4

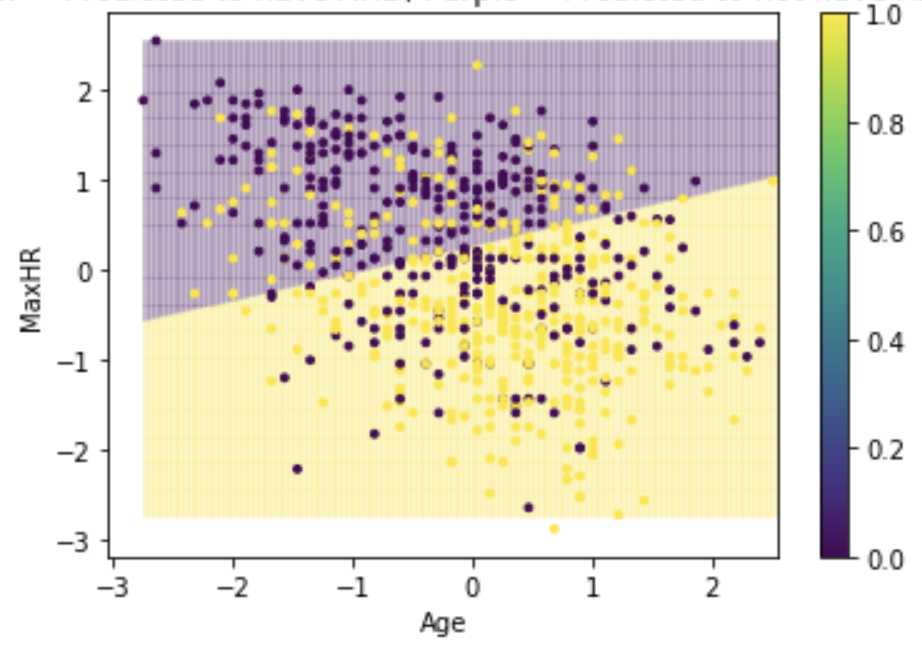
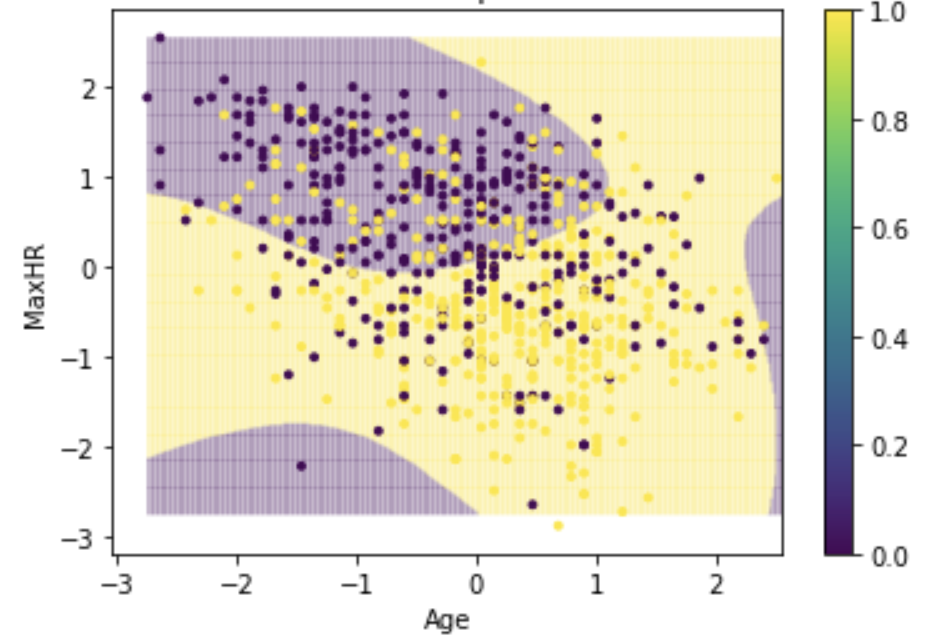
(note that “1” is not present since we set include\_bias=False)

**1e)** The results for logistic regression on this new set of features is as follows.



The following plots show the points and decision boundaries for the two models.



Comment on the difference between the two models and their performance.

Due to the increased flexibility of the 2nd model, it is able to fit the training dataset more closely, thus increasing the training accuracy. However, this has overfit the data, causing the test accuracy to decrease.

From the plots, the data is very noisy, and a linear decision boundary is fairly reasonable. There is little additional gain from the more flexible model, and it seems to overfit in the purple regions at the bottom left / right.

**Q2: Recommender Systems**

Reviews are helpful in our daily lives for deciding which restaurants, hotels, products, services etc. we should buy. However, fake reviews (purchased by the business, or even by competitors) have become a common issue which can mislead customers.

**2a)** What are some indicators that a product has fake reviews, that you have found useful in your experience?

* Look for reviews mentioning receiving free product in exchange for ratings (indicating financial incentives which can lead to bias)
* Uninformative, overly generic, or exaggerated reviews (more likely to be purchased)
* Large number of similar (highly positive / negative) reviews in a short period of time
* Lack of “helpfulness” votes, if available
* Reviewer history (lack of profile information, or new reviewer), if available

**2b)** In practice, this can be challenging. Some of the following are genuine TripAdvisor reviews, while the rest are written by paid crowdworkers in Amazon Mechanical Turk. Which of the following do you think are genuine?

1. The best service!!! The staff here was incredible. You never had to lift a finger. The room was huge as hotel rooms go. The view was phenomenal. Location great. What a great weekend. We did stop for a drink at the Palm Restaurant and unfortunately, it closes at 11 p.m. Too bad!!! The Lobby Lounge is open later but drinks costs twice as much as the bar at the Palm.

2. The Swissotel Chicago is a delight to visit. Located in downtown Chicago this hotel has nine different styles of room accommodations to serve everyones taste. They have children friendly rooms as well as the exquisite presidential suite. Rooms have exceptional views of downtown Chicago. This hotel also has a penthouse fitness center and pool for entertainment. Other places to stop by to eat is the Palms, Geneva and the Lobby Lounge.

3. The Swissotel Chicago is a very mediocre hotel, the service is always poor, and the room service food always comes cold, unless it’s supposed to be cold than it comes warm. I would rather stay at a super 8 than this place again.

4. I received the type of room that I had reserved. Cleanliness seems to be an issue with the maid staff. The carpet had debris and the bathroom needed attention. Room service was adequate but not great. The walls seem to be paper then as I heard one neighbor’s TV for most of the night and the other neighbor’s late enjoyment.

5. Overall I had a VERY Bad experience when hosting a meeting for company, but when my parents came in from out of town and wanted to stay at the Swissotel, I thought we should give them another try. Turns out my first impression was correct, this is the LEAST friendly hotel I’ve ever been in! Even the concierge was rude and disagreeable. I could go on and on, but instead I’ll just say GO SOMEWHERE ELSE!

6. My boyfriend and I were amazed by the breathtaking view of Lake Michigan! We are from Texas so the view of the city was very important to us. We had no problems at this hotel... rooms, service, location were top notch.

2, 3, 4 are fake; the rest are genuine.

**2c)** In the lecture, we came across 2 basic problems when building recommendation systems: **popularity bias** and **cold-start problem**. Briefly describe both problems in your own words.

• Popularity bias: only a small subset of items gets regularly recommended (”rich get richer”)

• Cold-start: when a new *user* joins the platform, no or not much information about that user is available to provide proper personalized information. The same could be true for a new *item* but possibly, item-item similarity can be used to make (halfway decent) recommendations.

**2d)** Content-based recommender systems represent items as feature vectors (item profiles) to calculate distances/similarities between them.

For the following 3 categories of items, what are arguably useful information to create an item profile to allow for meaningful recommendations:

• Electronic devices (e.g., phone, cameras, laptops)

• Property/Housing

• News articles

• Electronic devices: technical features, price

• Property/Housing: area size, floor height, age, location

• News articles: source, text features

**2e)** Considering your answers in (d), what do you think are the most important difference(s) between these categories in terms of how you might expect content-based recommender systems to perform? (This is an open question).

In some categories, we have relatively **informative features** which could relate to product quality in a stronger way: e.g. technical features of electronic devices, and features of property like location and area size which are highly informative. In other categories, features may be less informative, or the quality may be highly subjective, e.g. news articles, making it difficult for content-based recommendation systems to perform well. In the latter case, we would expect to have to rely more on collaborative filtering rather than content.

1. <https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)> [↑](#footnote-ref-1)
2. <https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction> [↑](#footnote-ref-2)
3. If this seems surprising, note that this is maximal heart rate, not resting heart rate. [↑](#footnote-ref-3)