## **The Soldier Fly Hunger Games: Predicting the hunger of insects using historical data**

Kasia Glowacka & Hayley Hubbard

**Problem Introduction**

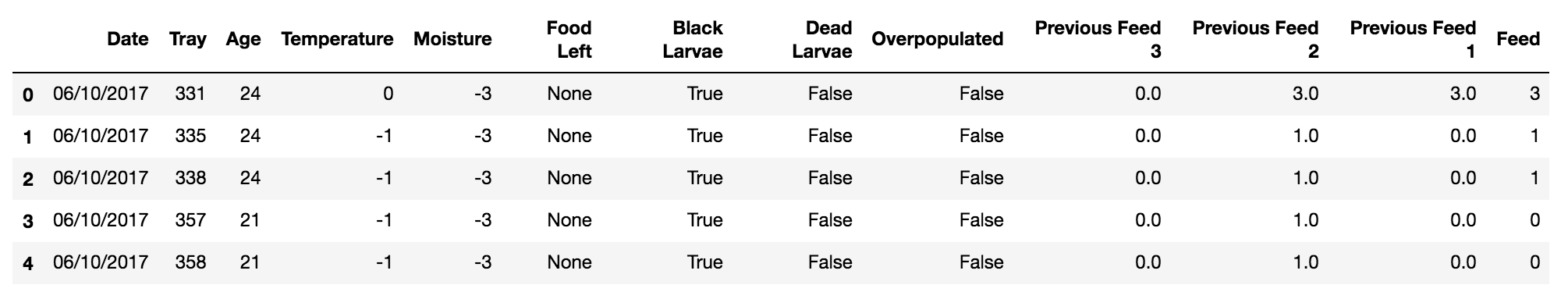
Entomics is a start-up company developing technology solutions for optimising waste conversion. Their main project uses Black Soldier Flies (BSF) to convert organic waste into fertiliser, animal feed and biofuels. Entomics is developing an engineering system to automate the feeding, upkeep and harvesting of the BSF. Currently, Entomics ensure the BSF are fed the appropriate amount of waste by recording factors such as temperature and humidity, and using a basic rules-based system and human intuition to calculate the feeding recommendation. This project will predict how much waste to feed each tray of BSF based on several input factors and be able to predict as well as or better than the rules-based system.

**Dataset**

This project uses Entomics historical daily feeding data, which was collected between 06/10/2017 and 31/05/2018. The dataset consists of 13 features and 11,491 data points (Table 1). Each tray of larvae was checked daily, from tray set-up through to harvest. According to Entomics, trays were usually harvested once the larvae started turning black, because this was an indication that they were starting to pupate.

The dataset includes temperature and humidity of trays, which ranged from -6 to 6 (dry/cold to wet/hot) with 0 being the ideal conditions. Dead larvae present was an indicator of disease and overcrowded trays were also recorded. The previous three feeds, along with the other variables, were used to calculate how much waste to feed the BSF, ranging from 0 to 7L of food.

**Table 1.** First five entries in the Entomics feeding dataset



**Preprocessing**

Through EDA, we identified a number of issues with the quality of the dataset. These are summarised in the table below, along with our solution, where applicable. To prevent data leakage in case of hidden variables, we used GroupShuffleSplit to ensure that each tray is only present in either the training or the test subset. To prepare our data for the models, we used one hot encoding for our categorical variable (‘Food Left’) and removed ‘Date’. Where no previous feed data was available at the start of a tray’s cycle, we replaced the ‘NaN’ values with 0.

**Table 2.** Data Quality Issues

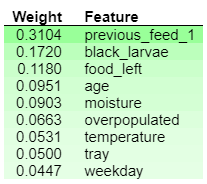
|  |  |
| --- | --- |
| **Issue** | **Attempted Solution** |
| 'Dead Larvae', 'Overpopulated', and 'Black Larvae' values changed from True to False on consecutive days for some trays. We were unable to ascertain whether this was due to some larvae being removed, or due to human error in record-taking. | We removed the ‘Dead Larvae’ variable from our dataset, as we found it had little effect on the accuracy of our predictions. |
| 'Black Larvae' metric was not always ‘True’ on the final day of a tray in production. | In a few trays, ‘Black Larvae’ was ‘True’ within the first seven days. This could have been caused by:  - The entries being accidentally recorded as ‘True’ but the larvae were not black  - A mix up of larvae between trays (i.e. a few older larvae fell in)  - Staff recording the larvae as black when they saw the larvae about to moult (they turn darker when they are about to moult)  Potentially the staff used this metric as a reason to reduce the feed, so we kept the metric in. |
| Missing observations - gaps of one or more days present for most of the trays. | We initially considered removing the affected trays, but this proved difficult due to their large number, and we had to accept this as a dataset limitation. |
| High correlation between Previous Feed 1, 2, and 3 variables. | Removed ‘Previous Feed 2 & 3’ variables. |
| Previous feed data did not always match the ‘Feed’ column from the previous days | The ‘Feed’ column was used to correct for errors in Previous Feed 1. |
| Hidden variation that could not be explained by the available variables - e.g. temperature and moisture were on average significantly higher on Mondays. | Added a ‘weekday’ variable to capture intra-week variation. |
| There were a total of 24 duplicate rows in the dataset (same date & same tray number). There were a total of 12 identical rows (exactly the same data) and 12 conflicting rows (one variable was different between duplicate rows e.g. feed was ‘3’ in the first entry, but ‘4’ in the second) | One row from each identical duplicate was removed. In conflicting duplicates, the first row was kept (based on the given order of data). The conflicting entries were removed this way because the two duplicate entries were very similar (a difference of +1 or -1 in feed or temperature) and it was not possible to know which one was correct. |

**Results**

We tested a range of classification and regression supervised learning models (Table 3). While the achieved accuracy scores were fairly low at 65-70% for the best-performing models, the low Mean Absolute Error (under 1 for all, and under 0.5 for best-performing models) means that our predicted feed values were reasonably close to those calculated by humans. The decision tree classifier had the highest accuracy, but was overfitting the data and made interpreting the results difficult. The model with the best trade-off between interpretability and MAE was XGBoost, with an accuracy of 65.56% and MAE of 0.48 (we were able to obtain similar results using an optimised Random Forest; however, it required a maximum depth of 6 versus 4 for XGBoost, which made individual trees more difficult to interpret and visualise).

**Table 3.** Metrics for selected models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model description** | **Accuracy (%)** | **MAE** | **MSE** | **Mislabelled (out of a total of 3322)** |
| **Logistic regression** | 48.71 | 0.7 | 1.3 | 1704 |
| **Linear regression** | N/A | 0.67 | 0.91 | N/A |
| **Support vector machine** | 56.95 | 0.58 | 1.1 | 1430 |
| **K-nearest neighbours** | 58.19 | 0.62 | 1.2 | 1389 |
| **Decision tree classifier** | 73.99 | 0.39 | 0.78 | 864 |
| **Random forest** | 66.68 | 0.46 | 0.89 | 1107 |
| **XGBoost** | 65.56 | 0.48 | 0.92 | 1144 |

**Conclusions**

We used eli5 to obtain the feature importance of the XGBoost model (Figure 1; note that due to model restrictions the figure only shows the magnitude, not the direction of each variable’s effect on the predicted feed). Previous feed was the most significant feature (even when the ‘food left’ variable was taken into account), which means that if larvae were fed more in the past, it is likely that they will continue to feed on a large amount of food in the future. The presence of black larvae and tray’s age were both important factors, because larvae tend to consume less food when they are very young and just before they are about to pupate. Humidity of a tray was also reasonably important, with higher levels of humidity contributing to lower feed values. On the other hand, we found that temperature had a smaller effect than we had initially anticipated, which might be due to the relatively small variation in the recorded temperature values (86% of the records had temperature value of 0 or 1). The fact that the tray and weekday variables are almost as important means that some of the variance in the ‘feed’ values cannot be explained by the recorded tray conditions, and points to a presence of some hidden variables. To improve future predictions, one could attempt to further optimise the model and predict the missing observations.

**Future**

Without extra features or data points, future work with predicting how much food to feed the BSF is difficult. With the dataset, there is still unexplained variation. For example, it is not currently recorded who is feeding the BSF but human intuition plays a part in how much waste they are fed, which may affect the outcome. Moisture and temperature would be better as recorded values, or if possible, temperature/humidity readings of every five minutes within each tray. The trays may have been hotter when recorded but it’s unknown if they were hotter for one hour or ten hours previously.

The model is only as good as the data put into it, and human intuition may have not resulted in the highest quality of insects every time. The biggest improvement would be working out which trays produced the highest quality of insects. This could be calculated by insect weight and/or nutritional values and working out which features produce the best insects.