Ovulation & Extrapolation:

Prediction of oestrus intervals for guide dogs

By Callum Ilkiw, Yiming Ma and Satoshi Komuro. Advised by Prof. Colm Connaughton
In Association with The University of Warwick, EPSRC, MathSys and Guide Dogs UK.

Engineering and Physical Sciences Research Council

WARWICK THE UNIVERSITY OF WARWICK

Guide Dogs

1. The Problem

For a long time now, Guide Dogs UK has been using a simple estimator of **7 months** for the interval period between breeding seasons. Despite this, the last **10 years of breeding data** shows significant variation for individual dogs. This project is using data analysis techniques to produce a mathematical model uses each bitch's details to give a **personalised interval estimation**.

Example Data Point:
Data ID: 23
Dog ID: 54321
Name: Doggo
DOB: 23/03/2003
Breed Name: Labrador
Colour: Black
Pregnant Last Season: 0
Pedigree (Sire): MIKE(12345)
Pedigree (Dam): KIM(12435)
Season Start: 11/11/2007
Age at season: 4.641
Time from previous season: 233
HR Notes: Mating season ? On medication X.
Diet: Brand3 Light
BCS: 4
Weight: 27

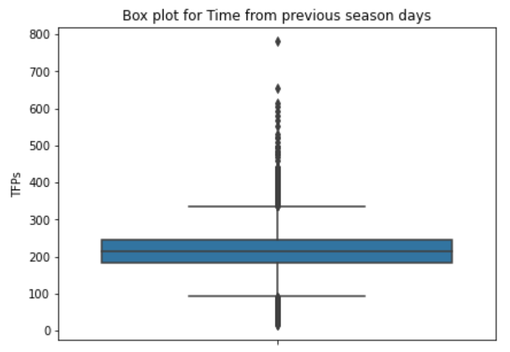
2. Data Overview

We had ***4693 data points*** with ***22 features*** for each point. The feature this project aims to predict is **Time from previous season(TFPS)**. An exploratory analysis of the data revealed some important information and some guidelines for our model. Some key points:

- **Mean TFPS** ≈ 216.7 ≈ 7 months. **Figure 1** shows the full distribution.
- **TFPS Range** 14 – 781
- **Average TFPS** by breed:
  - **Min** German Shepard at 174
  - **Max** Labrador x Golden Retriever* at 250
- **Highest correlation coefficient** to TFPS: Weight at -0.12, **Figure 2**.

The initial data was messy, with missing data, complicated data, and important information hiding in blocks of text. For example, whether a dog had mated or not during a season, was consigned to a small part of the notes section. For missing data, we have used **random assignment** based on estimated distribution and will be moving on to more complex methods. For categorical data, we used the **one-hot encoding** method to ensure there was no accidental "distance" between categories. This method had its limits however, since some categorical data was simply too large for it to be useful e.g. Father.

Understanding the data also involved learning terms exclusive to breeding, e.g. Split Season**.



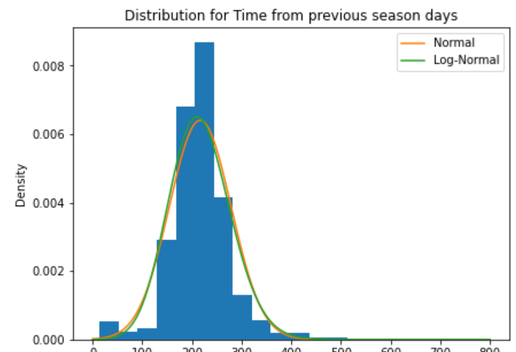


Figure 1 (Top Left): *The distribution of Times from Previous Season (Days) in boxplot.*
Figure 2 (Top Right): *The distribution of Times from Previous Season (Days) in Histogram.*
Figure 3 (Bottom): *Dog weigh(Kg) against Times from Previous Season(Days)*

3.Pipeline construction

The pipeline for this project is based off the Hitchhiker's Guide to DSSG[1]. The general structure is shown in **Figure 4**. Using a formal pipeline structure has many benefits: ensures the raw data is never edited, allows for module-based testing and easy switching between different models.

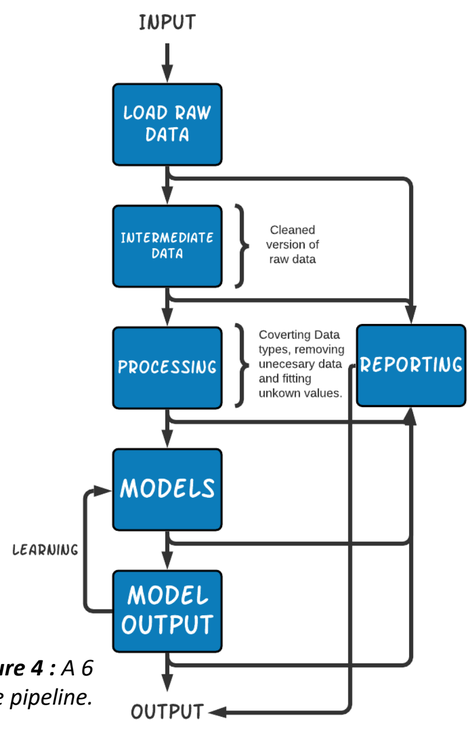


Figure 4 : *A 6 stage pipeline.*

4. Model Overview

During this experimental phase, we tested **many different models**: Null model(always predicts 7 months), Linear Regression, Decision Tree, K-nearest Neighbour, Random Forest, Support Vector Regression, Shallow Neural Network and a Deep Neural Network. Model fitting began with testing the current model and justifying it with a constant model. Minimising the constant model produced a very similar answer ~216 days. After this, several models were tested and the **results are shown below**.

Model	Mean Squared Error	Mean Absolute Error	Median Absolute Error
Null	3774.24	41.64	30.00
Linear Regression	2784.43	34.04	22.88
Decision Tree	5288.31	48.62	31.00
K-NN	3581.81	40.62	29.00
Random Forest	3063.79	36.64	24.57
SVR	2822.91	33.38	21.68
Shallow NN	2647.55	34.16	23.91
Deep NN	2573.97	32.93	21.66

Table 1: *Error measures for several tested models.*

5. Evaluation

Experimental trials show several models giving a significant increase in performance over the current model. Out of these, the 3 best were:

Linear Regression – Fits a line to the data that minimizes the mean squared errors for all data points. $f(\boldsymbol{x}) = a\boldsymbol{x} + b$, $a = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$ and $b = \bar{y} - a\bar{x}$.

Support Vector Regression – Using kernel functions, SVR transforms the data to be almost linearly separable. It then uses decision boundaries to make predictions. $f(\boldsymbol{x}^{(j)}) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(\boldsymbol{x}^{(j)}, \boldsymbol{x}^{(i)}) + c$, where $K(\boldsymbol{x}^{(j)}, \boldsymbol{x}^{(i)}) = \exp(-\gamma\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\|^2)$.

Deep Neural Network – This supervised model has 10 layers of weights that are repeatedly adjusted to produce the most accurate results it can.
Input: $\{\boldsymbol{a}^{[0]} = \boldsymbol{x}\}$, Process: {For $i = 0, 1, \ldots, 8$, $\boldsymbol{z}^{[i+1]} = \boldsymbol{W}^{[i]}\boldsymbol{a}^{[i]} + b^{[i]} \rightarrow \boldsymbol{a}^{[i+1]} = g^{[l+1]}(\boldsymbol{z}^{[i+1]})$ }, Output: $\{y = W^{[9]}\boldsymbol{a}^{[9]} + b^{[9]}\}$.

On Metrics: This analysis uses 3 different metrics, each with advantages and disadvantages. **MSE** is typically the best for analytical problems, or use by a computer, since it is an easily differentiable function. Compared to **Mean AE** and **Median AE**, **MSE** is less intuitive for humans to understand and can make it hard to convey anything about the data. **Mean AE** reduces the effect of outliers from **MSE**, and **Median AE** reduces this effect even further. This can be a positive or negative.

6. Future Work

Directly following this will be moving on to **feature engineering**, and testing more complex models. Following previous research, we will look at the impact of pregnancies in the previous season[2] and the time of year[3]. The final product of the project will be system that Guide Dogs UK can use to predict the interval of future dogs with an **accurate confidence interval**. Learning from **future data points** may be beneficial, but mathematical evidence[4] has shown the contrary to often be true.

References:
(1) Centre for Data Science and Public Policy, University of Chicago, Hitchhiker's Guide to Data Science for Social Good, https://github.com/dssg/hitchhikers-guide .
(2) C. Linde-Forsberg A. Wallén, *Journal of Small Animal Practice*, **Vol. 33**, Issue 2, p.67-70, (1992).
(3) Gavrilovic BB, Andersson K, Linde Forsberg C. Reproductive patterns in the domestic dog--a retrospective study of the Drever breed. Theriogenology. 2008 Sep 15;70(5):783-94. doi: 10.1016/j.theriogenology.2008.04.051. Epub 2008 Jun 25. PMID: 18582927.
(4) J. Liley, S.R. Emerson, B.A. Mateen, C.A. Vallejos, Louis J.M. Aslett and S.J. Vollmer, *"Model updating after interventions paradoxically introduces bias." International Conference on Artificial Intelligence and Statistics. PMLR, 2021*

*Non-pure breed parent
**The bitch goes into season but drops out of season before reaching oestrus, The bitch normally returns to a proper season in a few weeks.