Fat prediction with regression

Gilberto Juárez Rangel
(Dated: May 5, 2024)

In this report we utilize the 'Extended Body Fat Dataset' to predict the body fat using different regression models. Before performing any regression model we explore the data in order to clean it and apply feature engineering techniques that allows for a better performance. Finally we compare the different regression models with unseen data.

I. DATASET

The 'Extended Body Fat Dataset' includes 14 features (13 numerical and one categorical) which can be resumed in sex, age and 12 body measurements. The target feature as expected is the percentage of body fat for each person. This dataset includes 252 male measurements and 184 femalee measurements being a total of 436 samples. The names of the features and target are pretty much straight forward so there is no need of explanation for each. The units for each are the following:

- BodyFat (%)
- Sex (M/F)
- Age (years)
- Weight (kg)
- Height (mt)
- Neck (circumference in cm)
- Chest (circumference in cm)
- Abdomen (circumference in cm)
- Hip (circumference in cm)
- Thigh (middle part, circumference in cm)
- Knee (circumference in cm)
- Ankle (circumference in cm)
- Biceps (circumference in cm)
- Forearm (circumference in cm)
- Wrist (circumference in cm)

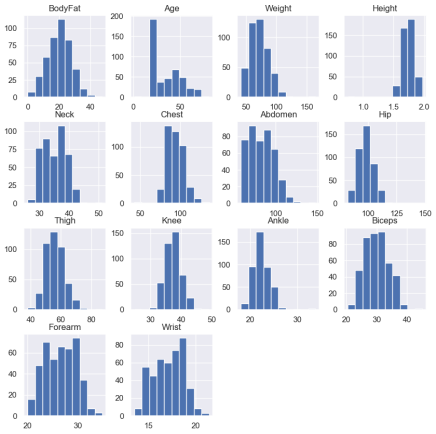

FIG. 1: Histograms of each numerical feature and target

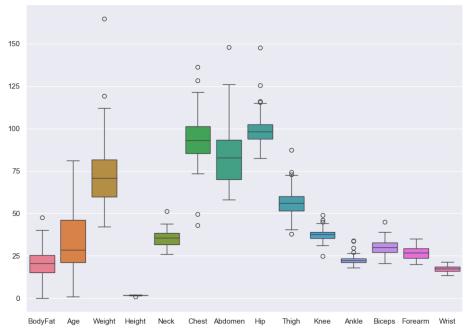

FIG. 2: Boxplots of each numerical feature and target

II. DATA EXPLORATION, CLEANING AND FEATURING

To get some insight of how the data is distributed we used both histograms and boxplots, this is ilustrated in figures 1 and 2. We can observe some skewness in figure 1 for some features and some outliers in figure 2.

In order to clean the data we first checked for any missing values (i.e. null or 'nan' data), for this dataset we found none. Then we proceeded to search for nonsense data (impossible measurements) we finded 3 cases of zero percent body fat which is known to be impossible. After removing those nonsense data we created new boxplots without the nonsense data and removed the outliers. Leaving us with a total of 415 samples.

Once we cleaned our data from outliers and nonsense data, we transformed some variables according to their normaltest pvalue after different transformations, this can be seen in table I:

	Normal test			
Feature	No transf.	Log tansf.	Sqrt transf.	Boxcox transf.
BodyFat	0.139	3.35e-19	2.88e-5	0.1359
Age	1.93e-10	0	3.16e-34	0
Weight	1.76e-4	1.47e-4	9.88e-5	1.04e-4
Height	8.17e-4	4.509e-4	5.84e-4	5.03e-4
Neck	5.62e-33	6.43e-29	1.50e-32	9.58e-33
Chest	3.76e-4	7.78e-5	2.49e-4	1.61e-5
Abdomen	4.42e-18	2.03e-36	2.38e-27	5.42e-37
Hip	0.225	0.554	0.424	0.5609
Thigh	0.045	0.075	0.067	0.071
Knee	0.026	0.011	0.018	0.014
Ankle	0.048	0.131	0.088	0.113
Biceps	1.69e-5	7.82e-5	2.26e-5	2.22e-5
Forearm	2.60e-17	2.78e-17	2.49e-18	8.49e-18
Wrist	8.03e-14	1.419e-11	5.83e-13	1.65e-13

TABLE I: Normal test of each feature after 3 different transformations

For each feature we selected the transformation that perform the best in the pvalue of the normaltest. Finally we used the standard scaler to avoid disproportional coefficients in the lasso and ridge regressions. We also created a side dataset that included polynomial effects.

III. REGRESSION MODELS

Once we finished applying cleaning and feature engineering we used different regression models. For each regression model we used the plain dataset, only the men dataset and only the female dataset. We did this since we might suspect that gender might affect considerably the fat metabolism. In total we used 8 regression models for each dataset, we evaluated their effectiveness by using cross validation with 5 splits, the results using the best hyperparameters for each model are shown in table II:

	Mean Cross Validation		
Regression model	Full data	female data	Men data
Linear	0.497	0.532	0.607
Lasso	0.503	0.548	0.613
Ridge	0.504	0.553	0.608
Elastic Net	0.506	0.556	0.614
Linear (P.F.)	0.337	-1.502	-4.078
Lasso (P.F.)	0.524	0.548	0.614
Ridge (P.F.)	0.531	0.555	0.607
Elastic Net (P.F.)	0.525	0.550	0.610

TABLE II: Mean cross validation of each regression model with 3 different dataset splits using 5 splits (P.F. stands for Polynomial Features).

IV. RESULTS AND INSIGHTS

We can observe a great improvement between in most regressions when we consider the gender factor. In this case we must use two regression models, one for the female and another for the men. For the female case, we should use the Elastic Net without polynomial features since this has a better performance in the prediction. For the men case, we can either use the Elastic Net without polynomial features or the Lasso regression with polynomial features. In this case we decided to stay with the Lasso regression with polynomial features since this can gives us an insight of which interactions between features are the most relevant. We can assume this might be similar for the female data.

When doing cross validation, the lasso regression is trained for each subset separately. This can lead to a change of coefficients in each subset. To get the information of the most important coefficients we ran the lasso regression in 100 random train-test splits with test_size = 0.2 and check which features appear the most in the top 5 coefficients, the results can be seen in table III.

	Abdomen	Forearm × Wrist	Chest	Age × Neck	Abdomen × Hip
M Occ.	100	75	72	69	44
	Abdomen	Chest × Hip	Height × Chest	Abdomen × Hip	Age × Knee
F Occ.	100	99	97	96	86

TABLE III: Number of occurrences of top 5 polynomial features in 100 random train-test splits per gender.

In examining the top 5 features of the male dataset, it becomes evident that the 'Abdomen' feature consistently appears across all train-test splits. The subsequent three features are present in no more than 75% of the cases, indicating their significance, nevertheless with a slight dependence on the specific train-test split. Notably, the last feature is observed in only 44 cases, a relatively infrequent occurrence, suggesting it may not be as influential as the others and its inclusion in the top 5 features could be attributed to chance.

Similarly, upon analyzing the top 5 features of the female dataset, we find that the 'Abdomen' feature consistently appears across all train-test splits, underscoring its pivotal role independent of gender. Additionally, the next three features are prevalent in nearly all cases, implying their importance, nevertheless with potential exclusion in rare train-test split scenarios. Lastly, the final feature, although appearing in 86 instances, slightly trails in significance compared to the previous ones.

V. NEXT STEPS

The dataset's size is relatively modest, comprising no more than 450 rows. Given the importance of gender stratification in our analysis, ideally, this would result in approximately 225 rows per gender subgroup. With a dataset of this scale, achieving robust results can be challenging.

To enhance predictive performance, we could explore several paths. Firstly, experimenting with higher-order polynomial features may provide insights into potential nonlinear relationships within the data, potentially improving regression accuracy. Additionally, adopting more sophisticated modeling approaches, such as neural networks, presents an opportunity for enhanced predictive power. However, it's essential to acknowledge that neural networks are often regarded as 'BlackBox' models, meaning they lack interpretability, and determining the relative importance of individual features within the model becomes challenging