ORAL TEMPERATURE PREDICTION WITH MACHINE LEARNING

Sara Sarafimova

Abstract

This project aims to enhance regression task performance using machine learning models combined with feature selection and hyperparameter optimization. Initially, data preprocessing and standardization are conducted. Key features are identified through methods like Univariate Selection and minimal Redundancy Maximal Relevance (mRMR) reducing dimensionality and improving model interpretability. We employ multiple regression models, including K-Nearest Neighbors, Linear Regression, Support Vector Regression, Decision Trees, and ensemble methods like Random Forest and Gradient Boosting. Hyperparameter optimization is achieved through Bayesian Optimisation, aiming to ensure optimal model performance. Cross-validation is used to validate and prevent overfitting. Experiments demonstrate that integrating feature selection and hyperparameter optimization, in this particular problem, sadly did not yield expected improvement. This may be due to insufficient experience of the student.

Introduction

Predicting the oral temperature of the subject is framed as a regression task in this project. This is done with the purpose of measuring the given target in a contactless manner. Habitually, the oral temperature is measured with the probe placed under the subject's tongue and the lips closed around the instrument. However, global pandemics have necessitated distanced interactions whenever possible. Since the oral temperature is a key factor in determining a person's health state, the question examined in this paper is of crucial importance.

The dataset used in this paper contains thermographs of 1020 subjects. 26 features were then extracted from each thermal image.

Materials and Methods

The Process of Collecting Data

Over the course of 18 months, from November 2016 to May 2018, a clinical study was conducted at the Health Centre of the University of Maryland at College Park according to the guidelines of the Declaration of Helsinki. The primary devices used included an oral thermometer (SureTemp Plus 690, Welch Allyn, San Diego, CA) with established clinical accuracy, a webcam (C920, Logitech), two IRTs (the thermographs analysed in this paper are from the IRT-1: 320 x 240 pixels, A325sc, FLIR Systems Inc., Nashua, NH, USA; IRT-2: 640 x 512 pixels, 8640 P-series,

Infrared Cameras Inc., Beaumont, TX, USA), a blackbody (SR-33, CI Systems Inc., Carrollton, TX, USA) as the external temperature reference source (ETRS) for temperature drift compensation, and six models of NCITs. Facial key-points in IRT images were identified by matching landmarks on visible light images to thermal images with an image registration approach, as well as manual labelling. Based on the identified facial key-points, different regions/points on thermal images were defined and the temperatures at these regions were obtained from thermal images (Figure 2). Since IRTs exhibit varying degrees of instability and drift, all IRT-measured temperatures were compensated with a blackbody (ETRS) in the system.
The thermometer was placed under the subject's tongue in a sublingual pocket (heat pocket). The distance (of the subject) from the camera was also documented, as well as the temperature and humidity of the room. Then the temperature was read in two different modes, a "fast" mode in several seconds and a "monitor" mode after 3 minutes. The reference temperature ("aveOralF" – fast mode or "aveOralM" – monitor mode) was calculated as the mean of the two oral temperatures in fast mode or monitor mode (during rounds 1 and 3). All subject data were discarded if the difference between two readings was larger than 0.5°C, due to the likelihood of a measurement error.

The Dataset

The raw dataset used in this paper contains 4080 instances, that is 1020 subjects' thermographs taken in 4 rounds. The 26 facial features extracted were those from the IRT-1, FLIR systems camera. The target variable in this paper is the average oral temperature measured in monitor mode (aveOralM).

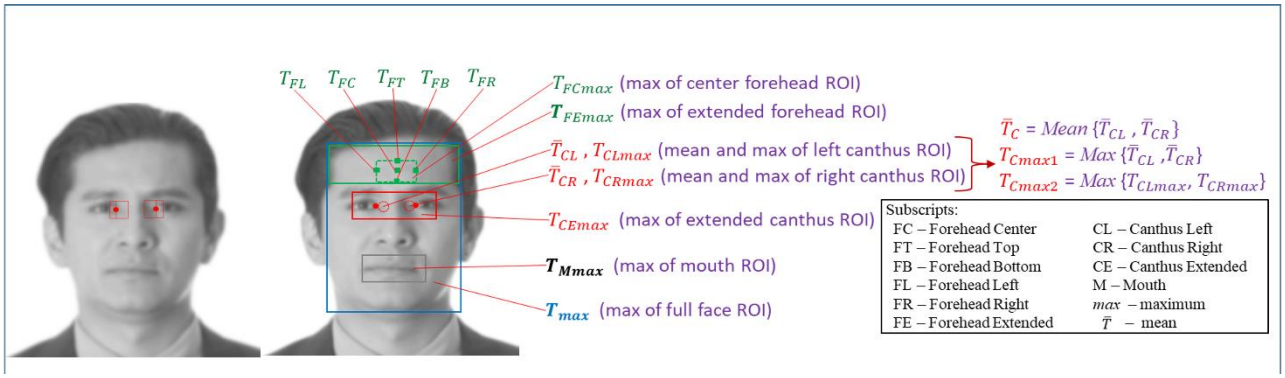

Figure 1 Critical points on thermal images

Feature	Name in papers	Region of interest	Value calculation
Max1R13	TCRmax	A circle with diameter of 13 pixels from the right canthus point to the face centerline	Maximum temperature within the circle
Max1L13	TCLmax	A circle with diameter of 13 pixels from the left canthus point to the face centerline	Maximum temperature within the circle
aveAllR13	TCR	A circle with diameter of 13 pixels from the right canthus point to the face centerline	Average temperature within the whole circle
aveAllL13	TCL	A circle with diameter of 13 pixels from the left canthus point to the face centerline	Average temperature within the whole circle
T_RC		A square of 24x24 pixels around the right canthus, with 2/3toward the face center (dry area, 16x24 pixels) and 1/3 away from the face center (wet area, 8x24 pixels).	Average temperature of the highest four pixels.
T_RC_Dry		The right canthus dry area, a rectangle of 16x24 pixels.	Average temperature of the highest four pixels.
T_RC_Wet		The right canthus wet area, a rectangle of 8x24 pixels.	Average temperature of the highest four pixels.
T_RC_Max		A square of 24x24 pixels around the right canthus, with 2/3toward the face center (dry area, 16x24 pixels) and 1/3 away from the face center (wet area, 8x24 pixels).	Maximum temperature within the square.
T_LC		A square of 24x24 pixels around the left canthus, with 2/3toward the face center (dry area, 16x24 pixels) and 1/3 away from the face center (wet area, 8x24 pixels).	Average temperature of the highest four pixels.
T_LC_Dry		The left canthus dry area, a rectangle of 16x24 pixels.	Average temperature of the highest four pixels.
T_LC_Wet		The left canthus wet area, a rectangle of 16x24 pixels.	Average temperature of the highest four pixels.
T_LC_Max		A square of 24x24 pixels around the left canthus, with 2/3toward the face center (dry area, 16x24 pixels) and 1/3 away from the face center (wet area, 8x24 pixels).	Maximum temperature within the square.
RCC		A square of 3x3 pixels centered at the right canthus point.	Average temperature within the square.
LCC		A square of 3x3 pixels centered at the left	Average temperature within the square.

		canthus point.	
canthiMax	TCEmax	Extended canthi area	Maximum temperature within the extended canthus area.
canthi4Max		Extended canthi area	Average temperature of the highest four pixels within the extended canthus area.
T_FHCC	TFC	Center point of forehead, a square of 3x3 pixels.	Average temperature within the square.
T_FHRC	TFR	Right point of the forehead, a square of 3x3 pixels.	Average temperature within the square.
T_FHLC	TFL	Left point of the forehead, , a square of 3x3 pixels.	Average temperature within the square.
T_FHBC	TFB	Bottom point of the forehead, a square of 3x3 pixels.	Average temperature within the square.
T_FHTC	TFT	Top point of the forehead, , a square of 3x3 pixels.	Average temperature within the square.
T_FH_Max	TFEmax	Extended forehead area	Maximum temperature within the extended forehead area.
T_FHC_Max	TFCmax	Center point of forehead, a square of 3x3 pixels.	Maximum temperature within the square.
T_Max	Tmax	Whole face region	Maximum temperature within the whole face region.
T_OR		Oral/mouth Region	Average temperature of the highest four pixels within the mouth region.
T_OR_Max		Oral/mouth Region	Maximum temperature within the mouth region.

Table 1 Descriptions of 26 facial variables (unit: °C)

Feature	Definition
T_offset1 (°C)	Temperature difference between the set and measured blackbody temperature.
aveOralF (°C)	Average oral temperature measured twice with the oral thermometer under fast mode.
aveOralM (°C)	Average oral temperature measured twice with the oral thermometer under monitor mode.
Gender	Male or female
Age	Age range of the subject
Ethnicity	Ethnicity of the subject
T_atm (°C)	Ambient temperature
Humidity (%)	Relative humidity
Distance (m)	Distance between the subjects and the IRTs.
Cosmetics	"1" means cosmetics applied. "0" means no cosmetics applied. Self-reported.
Round	Round of pictures (1, 2, 3 or 4)

Date	Date when the data were collected.

Table 2 Definitions of other variables

The Methods

Purely informational features such as “SubjectID”, “Date” and “Round” were dropped.

As most real-life problems with data analytics, this dataset too had missing values, 1917 in total. Twenty-nine features had missing values: all the facial features, “T_offset”, “Cosmetics” and “Distance”. The method for filling missing values was chosen based on analysis. The missing values in “T_offset” and the 26 facial features were filled with the mean value of the measurements for each subject. The missing values in the “Distance” column were determined to be from only two subjects, so the aforementioned technique could not be used. Instead, the missing values were filled with the mean value of the entire column. Since the “Cosmetics” feature can be either “0” or “1”, the missing values were filled with “0”.

To ensure optimal performance for our machine learning models, the columns “Age”, “Ethnicity” and “Gender” have been encoded. This step was necessary because the columns contain string values, while our machine learning models require numerical values. The encoding process involves assigning a number to a string or category that these features can take. As a final step of the data preprocessing procedure, the data was standardized (excluding the three aforementioned features, for obvious reasons). It has been proven that models perform better on standardized data.

When referring to the ‘original’ dataset later on, we refer to the dataset with dropped features, filled missing values, encoded features and standardized values.

Exploratory Data Analysis helps us better understand the problem we are working on and later choose the models more wisely. The distribution of the target variable is useful information since it allows us to identify outliers. It is shown in Picture 1. The average of the target variable is 37,02.
The correlation matrix shown in Picture 2 reveals the correlation coefficients between all the features and the target variable. All coefficients are positive, which is intuitive, since with increase in the temperature of any of the facial points, the oral temperature increases, too.

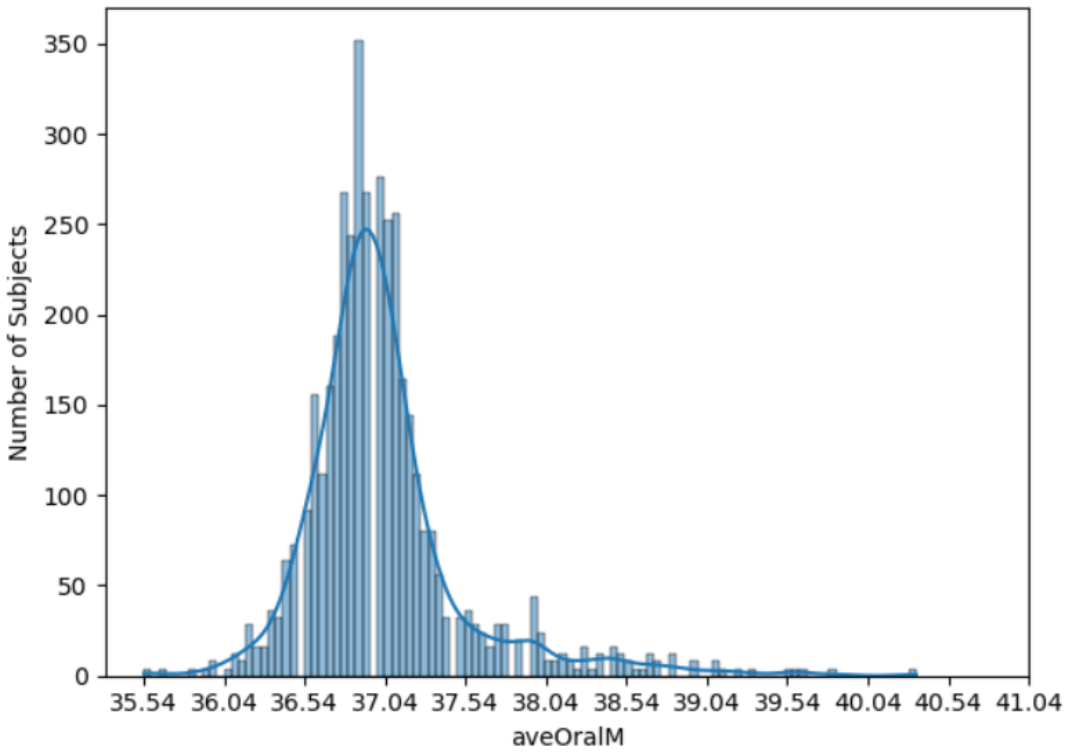

Figure 2 Distribution of the Target Variable

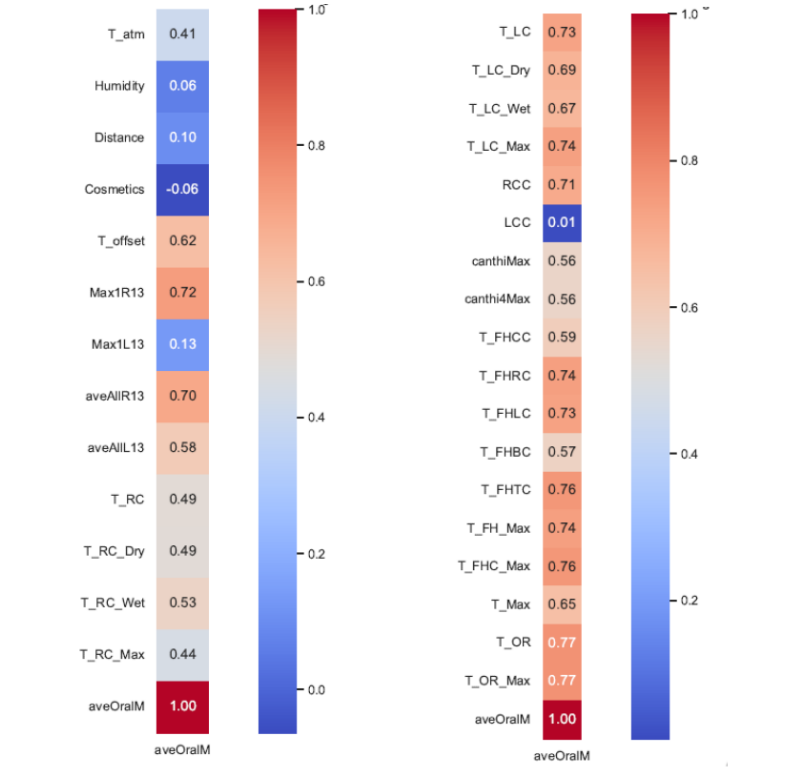

Figure 3 Correlation Matrix between the Target Variable and the Features

The Models

Initially, eight machine learning models with default parameters were trained on the original dataset. The models trained are: Dummy Regressor, Linear Regression, Regression Tree, K-Nearest Neighbors Regressor (KNN), Polynomial with Degree 3, Extreme Gradient Boost (XGBoost) and Support Vector Regression (SVR). In the first evaluation, eigthy percent of the data was used for training, and twenty percent was used for testing. In the second evaluation, 15-fold cross-validation (CV) was used. In a 15-fold CV, the instances are split into 15 equal parts. In the first cycle, one-fifteenth of the data is used for testing, and the rest of it is used for training. In the second cycle, the second-fifteenth of the data is used for testing, the rest for training, and so on. The metrics used in evaluating the errors are: mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE) and R2 score.

To reduce overfitting, improve accuracy and reduce training time, feature selection was performed. Feature selection chooses the *n* best features from the dataset based on the method used. After trying several inputs, it was concluded that working with 20 features is most reasonable. The first method, Univariate Selection, performs statistical tests that can be used to select those features that have the strongest relationship with the output variable. The second method, minimum Redundancy Maximum Relevance (mRMR) is a heuristic algorithm that finds a close to optimal subset of features by considering both the features' importances and the correlations between them. Even if two features are highly relevant, it may not be a good idea to add both to the feature set if they are highly correlated. Adding both would increase the model complexity (increasing the possibility of overfitting) but would not add significant information, due to the correlation between the features. The relevance and redundancy are computed using the mutual information (Information gain). This method also chose 20 best features. Nineteen out of the twenty chosen features were the same in both models. We continued using the 20 features chosen by the mRMR method.

When referring to the (new) dataset later, we refer to the dataset with 20 best features.

The same eight models were trained on the new dataset. As a last step, we performed Hyperparameter Optimization on the 3 best performing models: KNN Regressor, XGBoost and Random Forest Regressor. Hyperparameter Optimization refers to the tuning of parameters on models with the purpose of achieving the smallest error. It works on the following principle: we provide the algorithm a set of values for each parameter, and it calculates the error for each combination of hyperparameter values. The number of evaluations was 30, the cross-validation – 3 and the scoring – mean absolute error. We only conducted Bayesian Optimization, assuming it is the best performing one. Bayesian Optimization uses the previous

outcomes to focus the search and reduce the number of iterations. The parameters we got are the following:

For KNN: n_neighbors = 4, weights = distance, p = 1 (Manhattan distance)
For XGBoost: alpha = 0.3, max_depth = 9, min_child_weight = 1, learning_rate = 0.097
For Random Forest: criterion = absolute_error, max_Features = None, min_samples_leaf = 9, min_samples_split = 9, n_of_estimators = 550

Results

In the following section, we will compare the performance metrics of the models trained with the 80/20 percentage split and the 15-fold cross-validation technique. The models are trained with default parameters on the original dataset. “Model name” refers to the model trained with 80/20 percentage split, “Model name 2” refers to the model trained with the 15-fold CV technique. The model names are abbreviated as follows: D – Dummy, LR – Linear Regression, RT – Regression Tree, RF – Random Forest, KNN – K-Nearest Neighbors, PR – Polynomial Regression, XGB – Extreme Gradient Boost, SVR – Support Vector Regression.

Metric	*D*	*D2*	*LR*	*LR2*	*RT*	*RT2*	*RF*	*RF2*
MAE	0.34095	0.33276	0.22503	0.22615	0.20780	0.20703	0.15998	0.15164
MSE	0.29816	0.25934	0.08961	0.08666	0.07680	0.07661	0.04748	0.04240
RMSE	0.54604	0.50827	0.29935	0.29420	0.27713	0.27603	0.21789	0.20533
R2	-0.0070	-0.0008	0.69734	0.65737	0.74062	0.69289	0.83966	0.83258

Table 3 Performance measures of the models

Metric	*KNN*	*KNN2*	*PR*	*PR2*	*XGB*	*XGB2*	*SVR*	*SVR2*
MAE	0.13620	0.12344	0.60037	0.69658	0.15025	0.14863	0.21547	0.21040
MSE	0.04071	0.03605	1.60236	62.2298	0.04052	0.04066	0.08831	0.08633
RMSE	0.20178	0.18921	1.26584	4.62784	0.20130	0.20160	0.29717	0.29363
R2	0.86249	0.85848	-4.4119	-287.65	0.86314	0.84196	0.70174	0.66557

Table 4 Performance measures of the models (2)

In this next section, the comparison is done between the models trained with the 15-fold CV technique on the old dataset and the new dataset. “Model name 2” refers to the model trained on the original dataset, “Model name 20” refers to the model trained on the new dataset with 20 features.

Metric	*D2*	*D20*	*LR2*	*LR20*	*RT2*	*RT20*	*RF2*	*RF20*
MAE	0.33276	0.33291	0.22615	0.24014	0.20703	0.21581	0.15164	0.19472
MSE	0.25936	0.25957	0.08666	0.09586	0.07661	0.08389	0.04240	0.06582
RMSE	0.50827	0.50910	0.29420	0.30953	0.27603	0.28948	0.20533	0.25646

R2	-0.0008	-0.0023	0.65737	0.62844	0.69289	0.67458	0.83258	0.74439

Table 5 Performance measures of the models (3)

Metric	*KNN2*	*KNN20*	*PR2*	*PR20*	*XGB2*	*XGB20*	*SVR2*	*SVR20*
MAE	0.12344	0.20884	0.69658	1.7	0.14863	0.20469	0.21040	0.27936
MSE	0.03605	0.07811	62.2298	3001.6	0.04066	0.07305	0.08633	0.14868
RMSE	0.18921	0.27930	4.62784	25.7	0.20160	0.27010	0.29363	0.38558
R2	0.85848	0.69646	-287.65	-14002	0.84196	0.71566	0.66557	0.42272

Table 6 Performance measures of the models (4)

The results obtained from the Hyperparameter Optimization method are as follows:

Metric	*KNN BO*	*XGB BO*	*RF BO*
MAE	0.10383	0.14498	0.20303

Table 7 Performance measures of the models (5)

“Model name BO” refers to the model with optimized parameters.

Tables and Figures

Discussion

After analysing Table 3 and Table 4, we came to some conclusions. Evidently, Polynomial Regression with Degree 3 shows devastating results. It is most likely that this model is not suitable for our problem. When it comes to the validation techniques, 15-fold CV showed better results than the 80/20 percentage split. However, these improvements were in the second decimal, at best. This begs the question whether it is better to use a more complex validation technique, when the results are not satisfyingly better. We decided that it is worth it, so we used it in the next evaluation. After training the same models on the new dataset, with 20 best features, we came to

an intriguing outcome. The errors did not improve! In fact, they increased. This may stem from either overfitting or insufficient information for the model to effectively learn from. The poor performance of the models after hyperparameter optimization was not expected. Improvement was detected only in KNN, the best model throughout the whole process.

Citations

Wang, Q., Zhou, Y., Ghassemi, P., McBride, D., Casamento, J.P., & Pfefer, T.J. (2021). Infrared Thermography for Measuring Elevated Body Temperature: Clinical Accuracy, Calibration, and Evaluation. Sensors (Basel, Switzerland), 22, 215.

Wang, Q., Zhou, Y., Ghassemi, P., Chenna, D., Chen, M., Casamento, J., Pfefer, J., & McBride, D. (2023). Facial and oral temperature data from a large set of human subject volunteers (version 1.0.0)

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220