Regression Problem

Shadi Mahboubpardahi
Politecnico Di Torino
Student Id: 329057
s329057@studenti.polito.it

***Abstract*—This paper presents a solution to the challenge of estimating the coordinates through which a particle of interest has passed through a sensor plain. This challenge is the regression one, and I have used some models to train and predict the optimal position estimation of the location of particles. In this paper, I will discuss the dataset features to see how well they have been used for the models, and what model prediction has the minimum average Eculidian distances as is required in the problem.**

I. Problem Overview

The considered project consists of a regression problem in which the position of particles should be predicted on plain or pads that are a kind of detection sensor. Actually, the process of detecting the position of particles is done by passing them through sensors. As the particles pass, the sensors begin to measure time series signals, and the given dataset consists of some of the characteristics of the signal. These extracted features are described below:

1) **PMAX** : the magnitude of the positive peak of the signal, in mV
2) **Negpmax** : the magnitude of the negative peak of the signal, in mV.
3) **Tmax** : the delay (in ns) from a reference time when the positive peak of the signal occurs.
4) **Area** : the area under the signal
5) **RMS** : the root mean square (RMS) value of the signal.

However, due to the hardware limitations, instead of 18 records, there are 12 records for each particle that are valid. The dataset includes the following information:

Each signal consists of five attributes, each of which has eighteen values. In total, our dataset contains nineteen numerical columns, some of which are used to evaluate particle positions.

II. Proposed approach

To do this project, there are several steps that are explained below:

A. preprocessing

As it is noticed in section I, there is a challenging task to identify valid features, so I need to perform a feature selection phase before applying our regressions. To do that, each attribute value must be evaluated separately. Therefore, I transform the given dataset into 5 separate data frames as I can perform the analysis for each feature separately. I have used

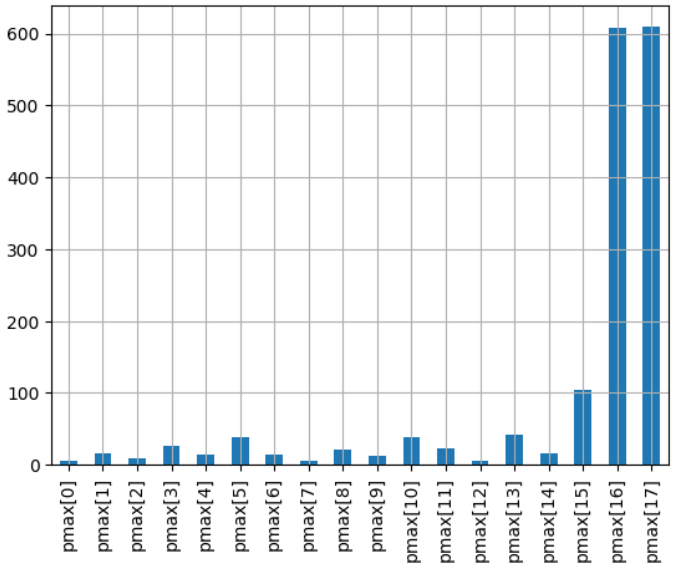

Fig. 1. Pmax Means

statistical methods such as mean and variance to observe and compare each value of features with each other and eliminate those that act differently.
for example, in "Fig. 1" Mean of each values of Pmax are individually plotted.

As can be seen, there are properties with different behaviors. For example, values 15, 16, and 17 have different means values than others. This method is done for the remaining features (Negpmax, Tmax, Area, Rms). The omitted columns per each feature are presented in "Table. I".

TABLE I
Omitted Values By Means

Features Name	Omitted Values
Pmax	15,16,17
Negpmax	16,17
Tmax	0,7,12
Area	15,16,17
Rms	16,17

The other statistical criterion that has been used, is variance. In statistics, variance is a measure of how spread out a set of values is around the mean (average) of the data. As in data analysis, features with low variance, are omitted, that is why,

I use Variance as a feature selection.
in the “Table. II”, considered values have been eliminated.

TABLE II
OMITTED VALUES BY VARIANCE

Features Name	Omitted Values
Pmax	0,7,12
Negpmax	0,5
Tmax	15,16,17
Area	0,7,12
Rms	17

B. Model Selection

two algorithms have been tested:
Random Forest Regression: it is a technique capable of performing regression task using multiple decision trees. It is robust towards outliers and it works well with sparse data. However, as I am dealing with a data set containing a lot of outliers, this algorithm could be highly useful.
Ridge: it is a technique used to prevent overfitting in linear regression models. In linear regression, the goal is to find the coefficients for the features that best fit the training data. However, if there are too many features or if some features are highly correlated, the model can become sensitive to small variations in the data, leading to overfitting.

C. Hyperparameters tuning

By grid search, I obtained different Average Euclidean Distance values for each combination of parameters which are provided in the “Table. III”.

TABLE III
HYPERPARAMETERS GRID

Model	Parameter	values
Random Forest	number of estimators Max Feature	50, 100, 300 ’log2’,’sqrt’
Ridge	alpha	0.1, 0.2, ... , 1

III. RESULT

I will now compare the results obtained with the two model, using the hyperparameter discussed in the previous section.

The best combination of parameters found for the random forest is **number of estimators**=300, **max features**=log2, **min samples leaf**=2, and **min samples split**=1
in my test set the result is 4.662 and Ridge is 17.739. On the evaluation set, I got 5.437 by using the prediction of the best Model (RandomForestRegression).

IV. DISCUSSION

The results obtained show that the Random Forest Regression machine outperforms the L2 (Ridge) LinearRegression, Although it is required to explore further hyperparameters by doing a Grid search on the models. The ensemble model like the random forest regression becomes more powerful as they take advantage of different types of models. Especially

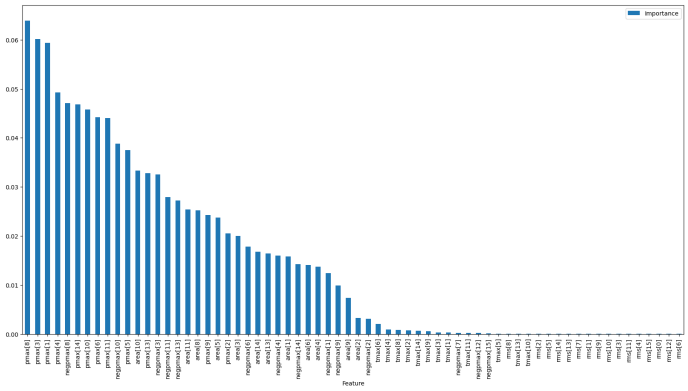

Fig. 2. Feature Importance

in the case of outliers and nonlinear relationships between the features, The usage of RandomForestRegression has been discovered on this dataset more effectively. Furthermore, one of the attributes that make the random forest regression more compelling to use on such a dataset is the interpretability, which shows how much each feature impacts the predictions. As shown in “Fig. 2” some of the features are less important the others which means I can get the same result with less number of features.

On the other hand, the Ridge algorithm is capable of training a linear regression task by avoiding over-fitting and performing much faster than RandomForestRegreesion, However, it is not suitable for this dataset as the features do not have linear relationships with the target, therefore, the linear model can not fully describe the predictions.

REFERENCES

[1] ChatGpt is used for writing some of the definitions in the report.
[2] https://scikit-learn.org/