Assignment

**Class Assessment 2**

**COURSE – INT247 – Machine Learning**

**Foundation**

|  |  |  |  |
| --- | --- | --- | --- |
| r. No. | Registration No | Name of Student | Roll No |
| 1 | 11803997 | GOVIND GANGA | RKM002B41 |

**Submitted To Dr. Lovi Raj Gupta,Sanjay Kumar Singh**

Lovely Professional University

Jalandhar, Punjab, India.



**Abstract**

A Machine learning model to predict CPU performance.

Machine Learning is a very powerful technique to analyse data, find some important insights and make a architecture which can predict future decision based on data. “Machine Learning is the science of getting computers to learn and act like humans do, and improve their learning over time in autonomous fashion, by feeding them data and information in the form of observations and real-world interactions.”

In this project, with the Relative CPU Performance Data we are going to train a machine learning model which can predict future Relative performance of the CPU based on required features value. We are going to train three machine learning algorithms Linear Regression, Polynomial Regression and Random Forest and compare each model based on root mean square error and R square. Both Linear Regression and Polynomial Regression are Statistical learning whereas, Random Forest is an ensemble learning based on decision trees.

After that we will interpret our project and predictions that were made by the model. In this process we are going to analyse about the particular prediction, what are the factors that are important most and what are least. We will interpret are results using tree interpreter and waterfall charts.

**About Dataset**

**Title:** Relative CPU Performance Data

**Source Information**

* Creators: Phillip Ein-Dor and Jacob Feldmesser
* Ein-Dor: Faculty of Management; Tel Aviv University; Ramat-Aviv; Tel Aviv, 69978; Israel
* Donor: David W. Aha (aha@ics.uci.edu) (714) 856-8779
* Date: October 1987

**Number of Instances:** 209

**Number of Attributes:** 10 (6 predictive attributes, 2 non-predictive,

1 goal field, and the linear regression's guess)

**Attribute Information:**

1. vendor name: 30 different vendors name

2. Model Name: many unique symbols

3. MYCT: machine cycle time in nanoseconds (integer)

4. MMIN: minimum main memory in kilobytes (integer)

5. MMAX: maximum main memory in kilobytes (integer)

6. CACH: cache memory in kilobytes (integer)

7. CHMIN: minimum channels in units (integer)

8. CHMAX: maximum channels in units (integer)

9. PRP: published relative performance (integer)

10. ERP: estimated relative performance from the original article (integer)

**Missing Attribute Values:** None

**Flowchart of the project**

Improve the model

Performance measure

**RMSE & R2 score**

Taking care of missing values

**Analyse and Interpret Result**

Performance measure

**RMSE & R2 score**

Performance measure

**RMSE & R2 score**

**Random Forest**

**Polynomial Regression**

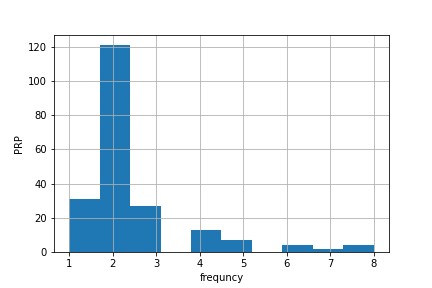
**Linear Regression**

**Training and validation split**

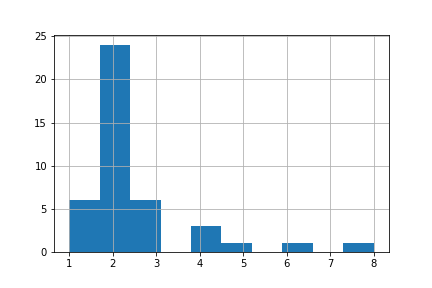
**EDA**

**DATASET**

**Splitting training and validation data**

****

So using stratified sampling from Sklearn we distributed the data in the same proportion. The following plot is of the validation data and having the same proportion of the distribution.

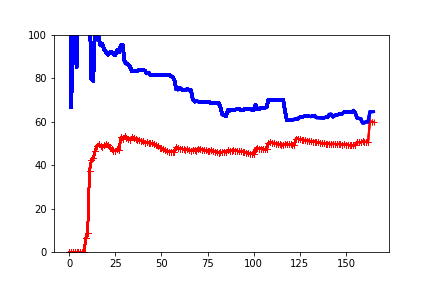


Linear Regression

**RMSE on training data: 59.592999**

**RMSE on validation data: 64.296008**

**R2 score of training data 0.8474442173588106**

**R2 score of validation data 0.8834651638700353**

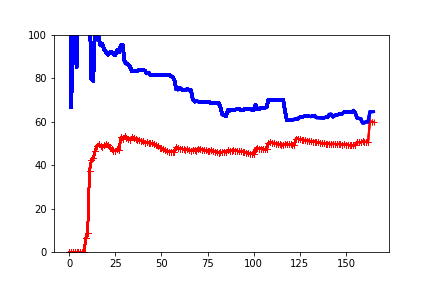
Polynomial Regression

**RMSE on training data: 21.357588**

**RMSE on validation data: 60.999958**

**R2 score of training data 0.9804051198002194**

**R2 score of validation data 0.8951069231748879**

****

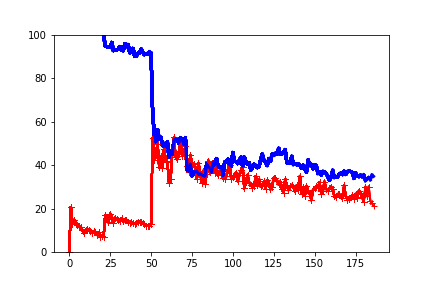
Random forest

**RMSE on training data: 19.242774**

**RMSE on validation data: 36.043999**

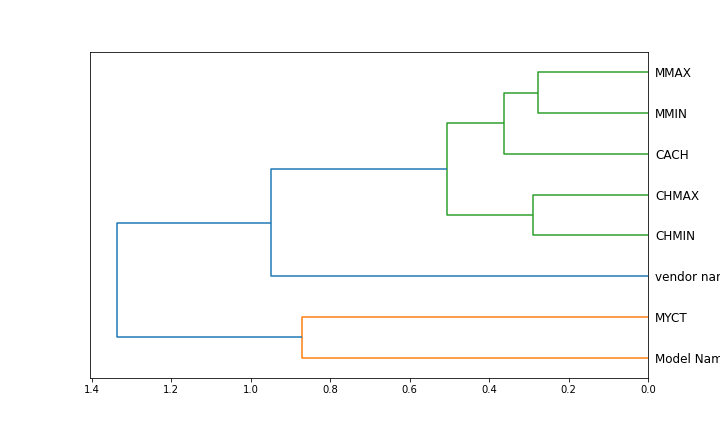
**R2 score of training data 0.9862452369529565**

**R2 score of validation data 0.914446537829289**

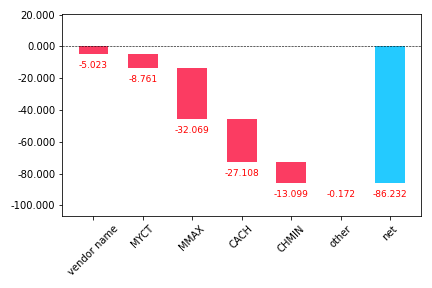
****

**Model Interpretation**

**Dendograph of Colinearity**

****

**Waterfall Chart of the model**

****

Reference: -

**Fastai library** <https://www.fast.ai/>

**Scikit Learn Library** https://scikitlearn.org/stablescikit-learn