**A Data Mining Approach to Predict Forest Fires**

**using Meteorological Data**

*Madi Yussupov*

**Abstract.**

Forest fires are a major environmental issue, creating economical and ecological damage while endangering human lives. Fast detection is a key element for controlling such phenomenon. To achieve this, one alternative is to use automatic tools based on local sensors, such as provided by meteorological stations. In effect, meteorological conditions (e.g. temperature, wind) are known to influence forest fires and several fire indexes, such as the forest Fire Weather Index (FWI), use such data. In this work, we explore a DataMining (DM) approach to predict the burned area of forest fires.

**Introduction**

One major environmental concern is the occurrence of forest fires (also called wildfires),which affect forest preservation, create economical and ecological damage and causehuman suffering. Such phenomenon is due to multiple causes (e.g. human negligence and lightnings) and despite an increasing of state expenses to control this disaster, each year millions of forest hectares (ha) are destroyed all around the world. In particular, Portugal is highly affected by forest fires . From 1980 to 2005, over 2.7 million ha of forest area (equivalent to the Albania land area) have been destroyed. The 2003 and 2005 fire seasons were especially dramatic, affecting 4.6% and 3.1% of the territory, with 21 and 18 human deaths. Fast detection is a key element for a successful firefighting. Since traditional human surveillance is expensive and affected by subjective

factors, there has been an emphasis to develop automatic solutions.

**Forest Fire Data**

**Number of Instances:** 517

**Attribute information:**

1. **X** - x-axis spatial coordinate within the Montesinho park map: 1 to 9

2. **Y** - y-axis spatial coordinate within the Montesinho park map: 2 to 9

3. **month** - month of the year: "jan" to "dec"

4. **day** - day of the week: "mon" to "sun"

5. **FFMC** - FFMC index from the FWI system: 18.7 to 96.20

#### ***Fine Fuel Moisture Code***

The Fine Fuel Moisture Code (FFMC) is a numeric rating of the moisture content of litter and other cured fine fuels. This code is an indicator of the relative ease of ignition and the flammability of fine fuel.

6. **DMC** - DMC index from the FWI system: 1.1 to 291.3

### ***Duff Moisture Code***

The Duff Moisture Code (DMC) is a numeric rating of the average moisture content of loosely compacted organic layers of moderate depth. This code gives an indication of fuel consumption in moderate duff layers and medium-size woody material.

7. **DC** - DC index from the FWI system: 7.9 to 860.6

### Drought Code

The Drought Code (DC) is a numeric rating of the average moisture content of deep, compact organic layers. This code is a useful indicator of seasonal drought effects on forest fuels and the amount of smoldering in deep duff layers and large logs.

8. **ISI** - ISI index from the FWI system: 0.0 to 56.10

9. **temp** - temperature in Celsius degrees: 2.2 to 33.30

10. **RH** - relative humidity in %: 15.0 to 100

**Relative humidity** (RH) is the ratio of the partial pressure of water vapor to the equilibrium vapor pressure of water at a given temperature. Relative humidity depends on temperature and the pressure of the system of interest.

11. **wind** - wind speed in km/h: 0.40 to 9.40

12. **rain** - outside rain in mm/m2 : 0.0 to 6.4

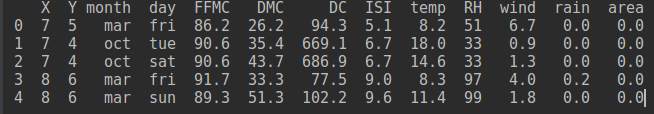
13. **area** - the burned area of the forest (in ha): 0.00 to 1090.84

(this output variable is very skewed towards 0.0, thus it may make

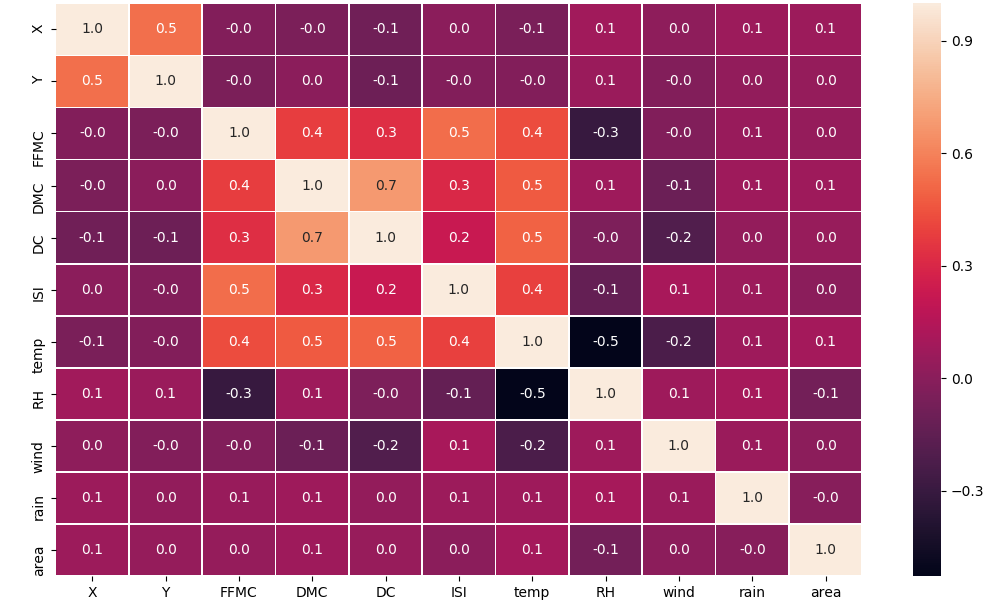
sense to model with the logarithm transform).

**Missing Attribute Values: None**

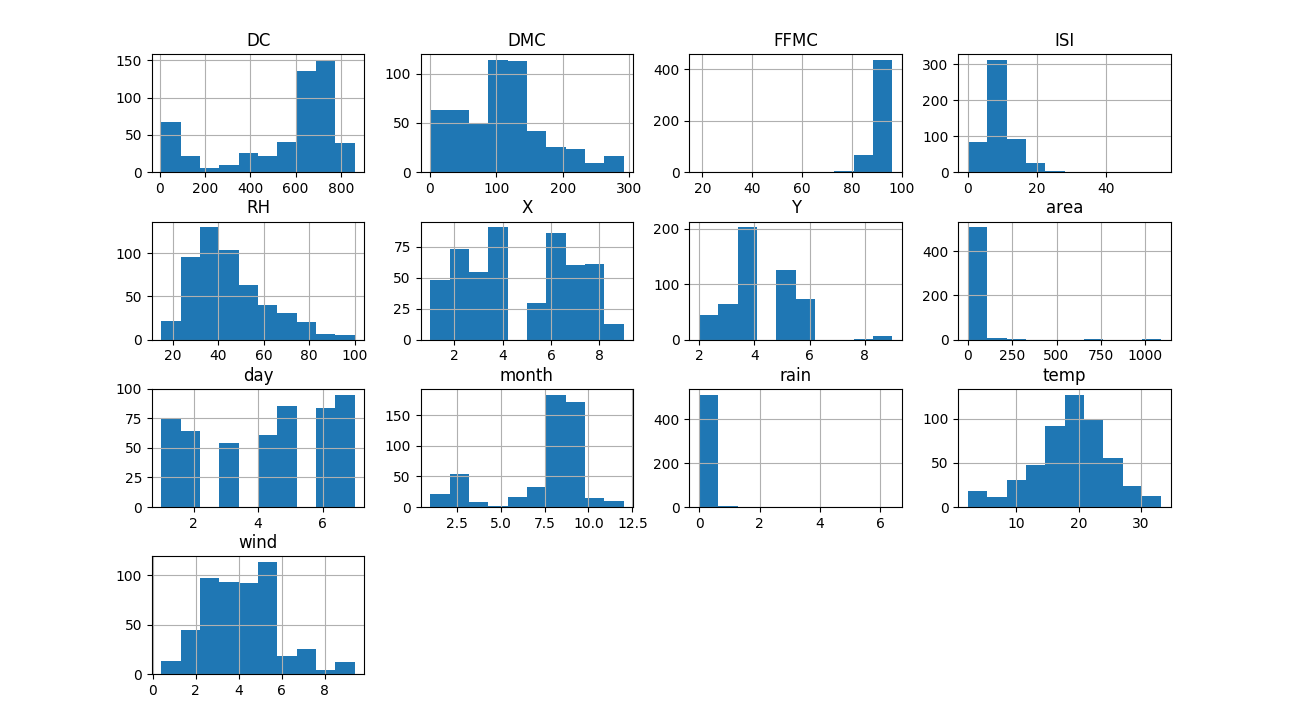
**Data example**

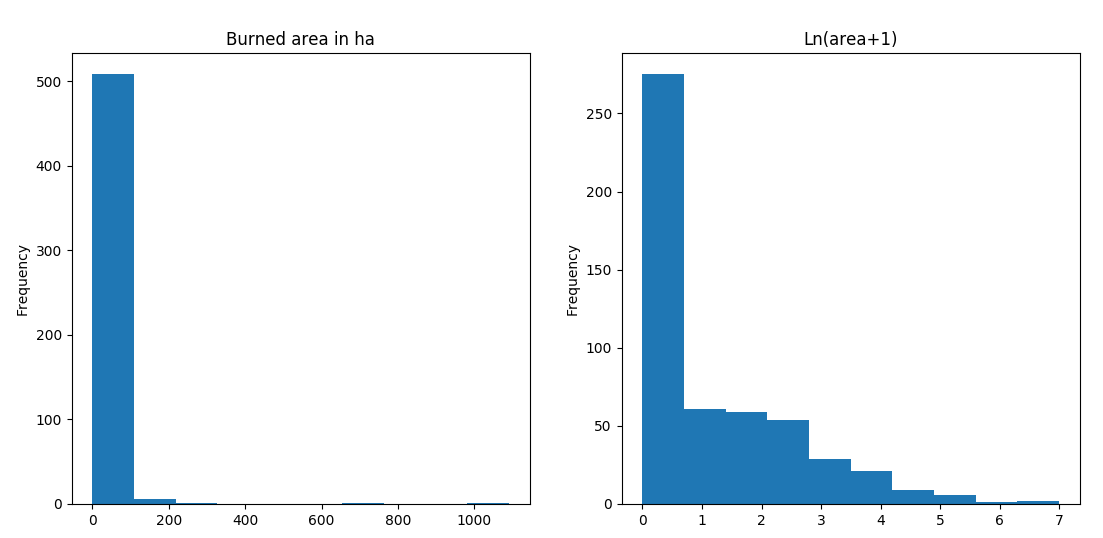


**Correlation** **of attributes**

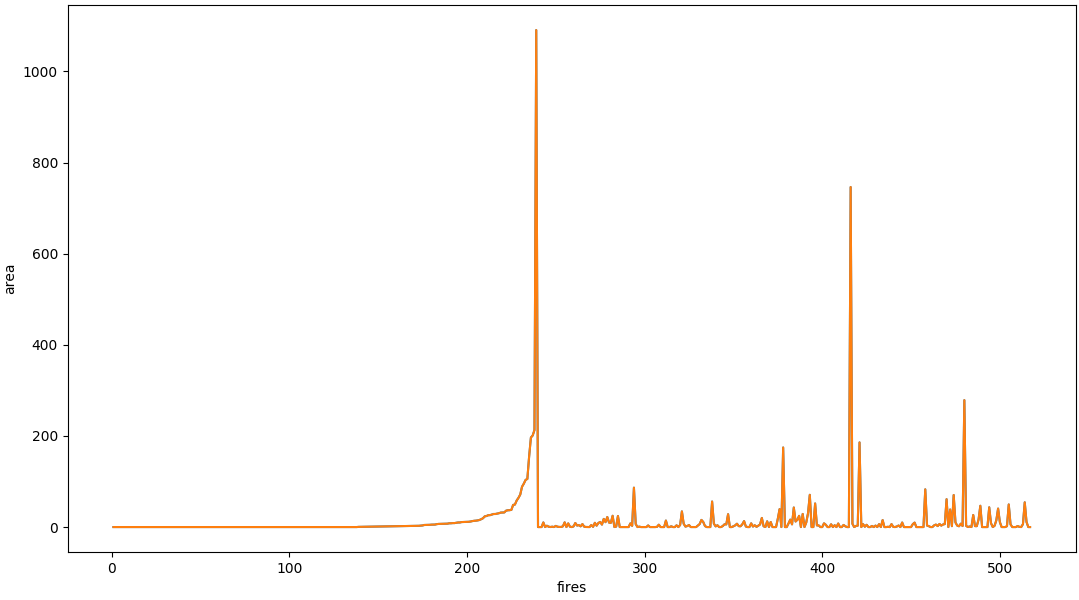


**Attributes Distribution**

****

****

The histogram for the burned area (left) and respective logarithm transform (right)

**Fires Area**

As you can see there are many outliers which makes difficult to predict the burned area. Later I will describe outliers deduction methods.

**Algorithms perfomance with default options**



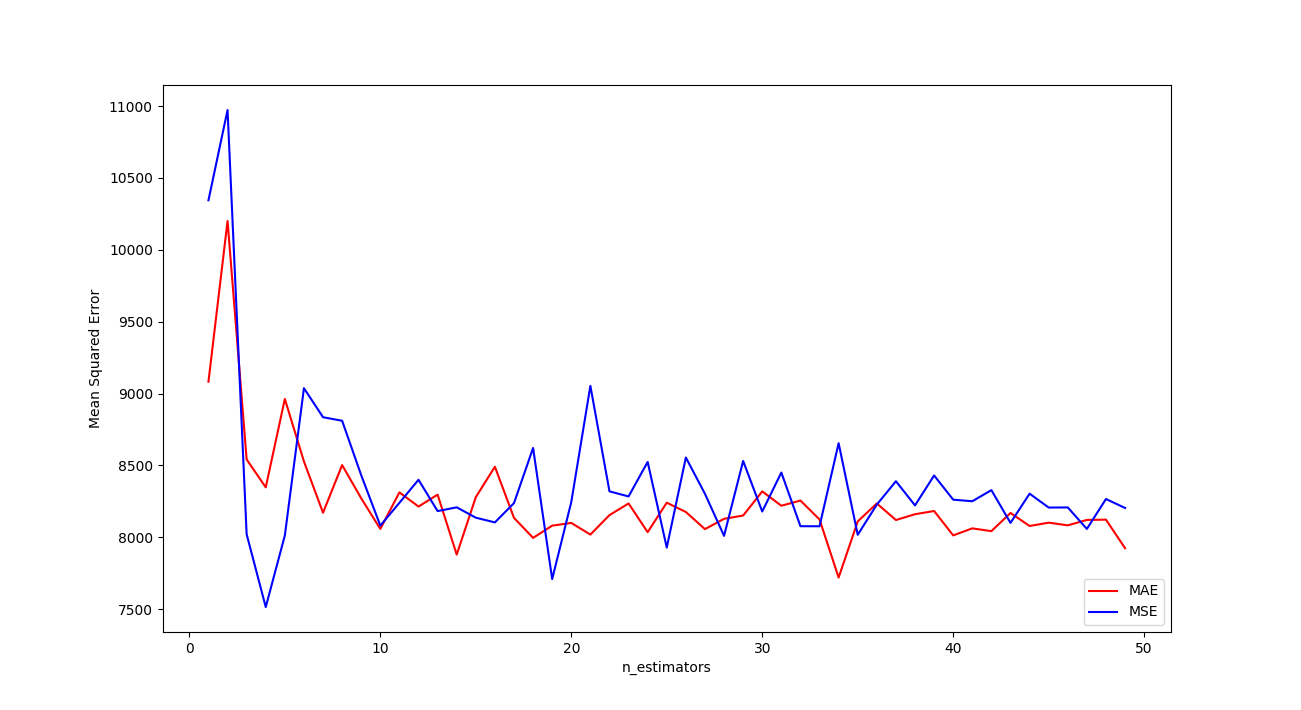
# **RandomForestRegressor**

The main parameters were the n\_estimator and crtierion.

n\_estimators - The number of trees in the forest.

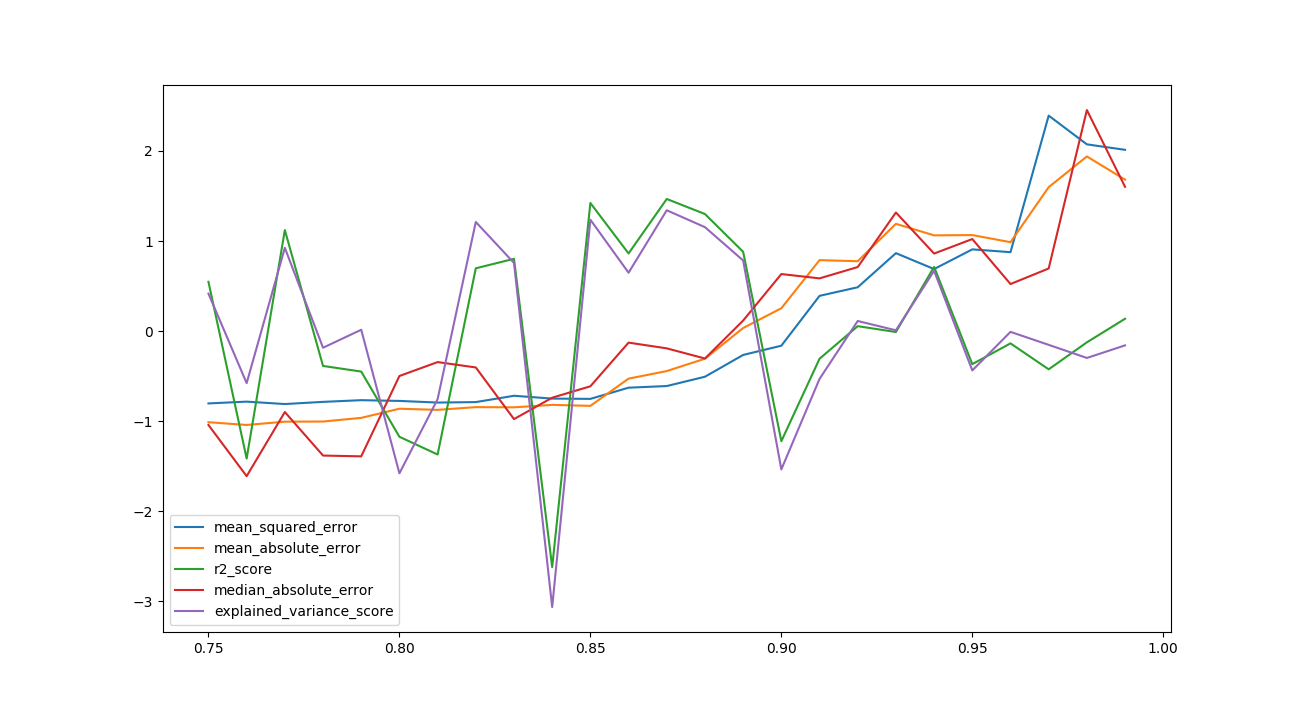
Criterion - The function to measure the quality of a split. Supported criteria are “mse” for the mean squared error, which is equal to variance reduction as feature selection criterion, and “mae”

for the mean absolute error.

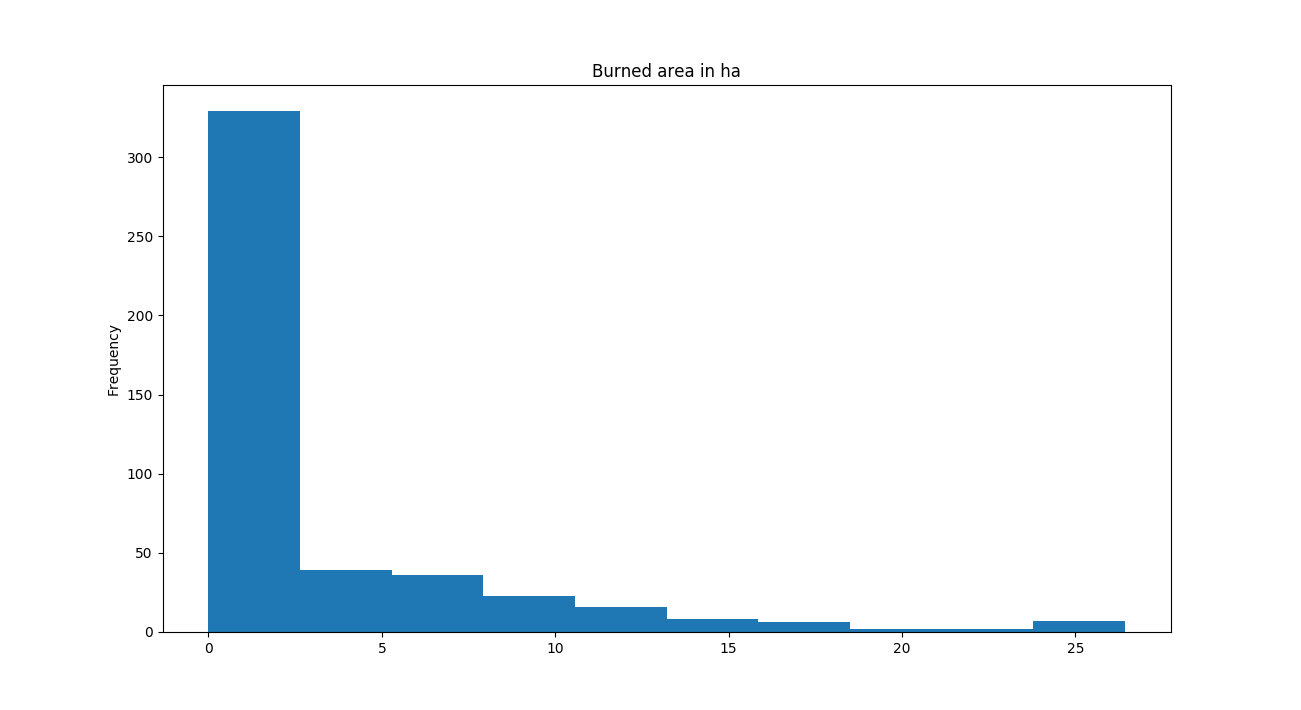


As you can see from the graphics(red line with mae criterion., blu with criterion). MAE criterion with 30 esitimators gives stable results than MSE. But error rate is still big.

**Removing outliers**

Normalized Error rate

As we told earlier there are many outliers there are outliers in dataset. Removing outliers gives better results on algorithm. And there are left records which burned area not higher than 30 ha.



Such methods as feature selection didn’t work random forest regressor and give approximately same results

**mean\_squared\_error 17.2629188092**

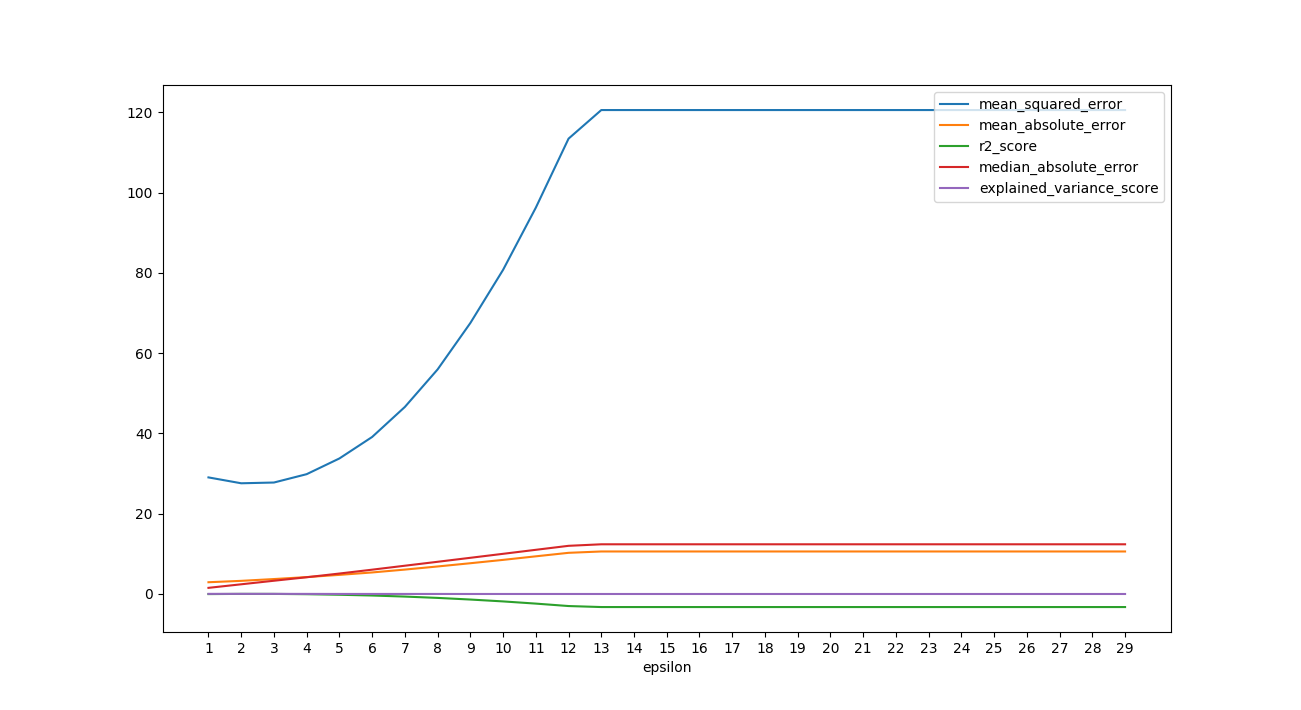
**mean\_absolute\_error 2.38900851582**

**r2\_score -0.208721408017**

**median\_absolute\_error 0.670333333333**

**explained\_variance\_score 0.021621114945**

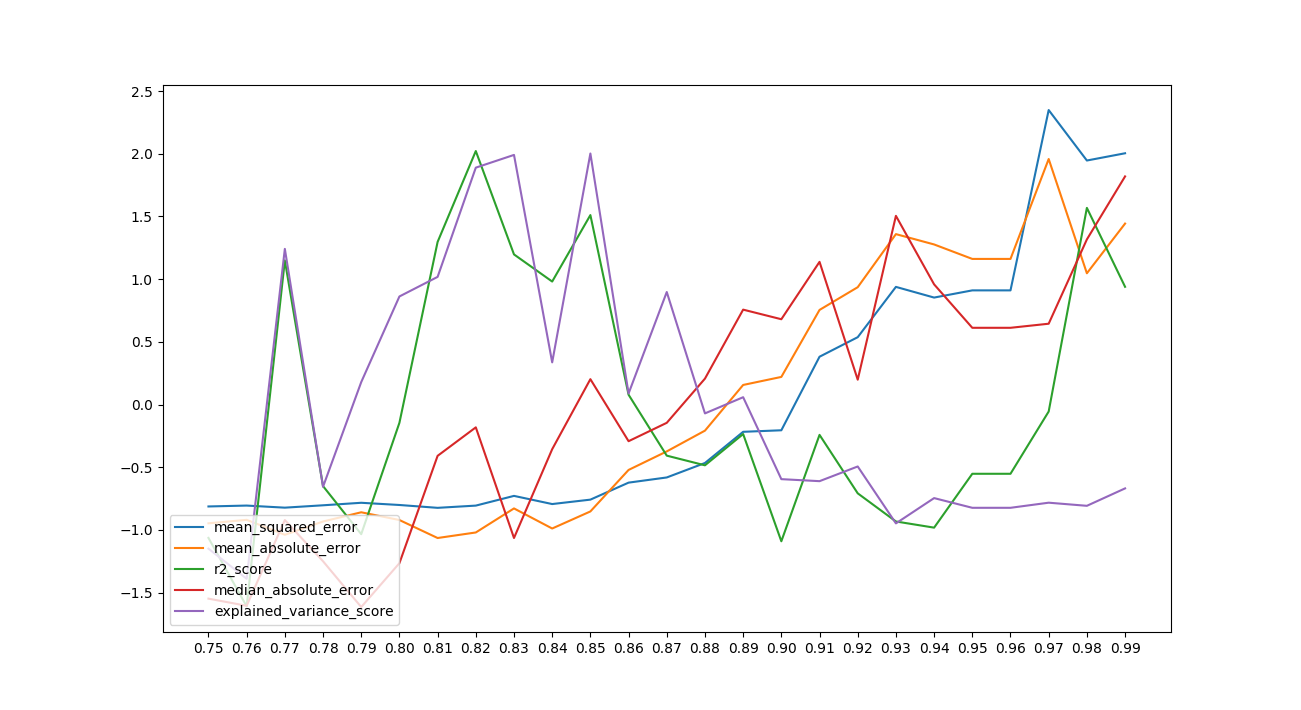
**Support Vector Machine**

****

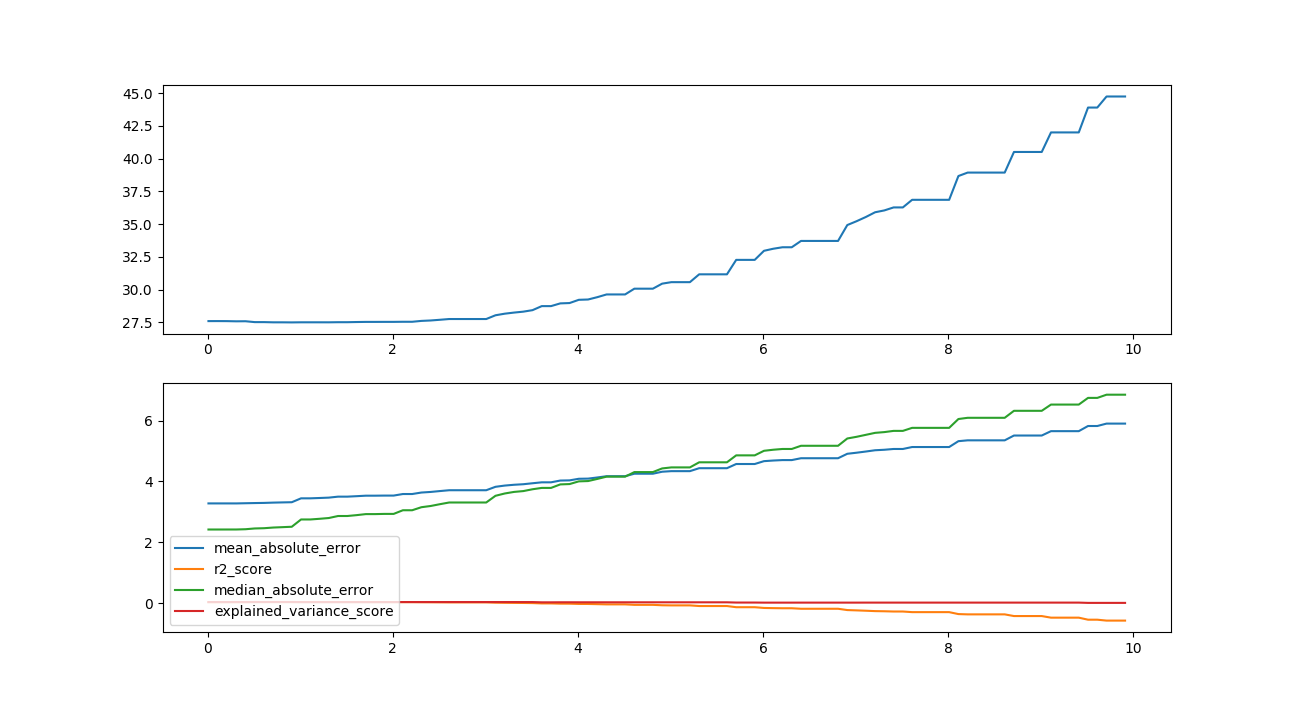
Epsilon in the epsilon-SVR model. It specifies the epsilon-tube within which no penalty is associated in the training loss function with points predicted within a distance epsilon from the actual value.

Best epsilon was 3.

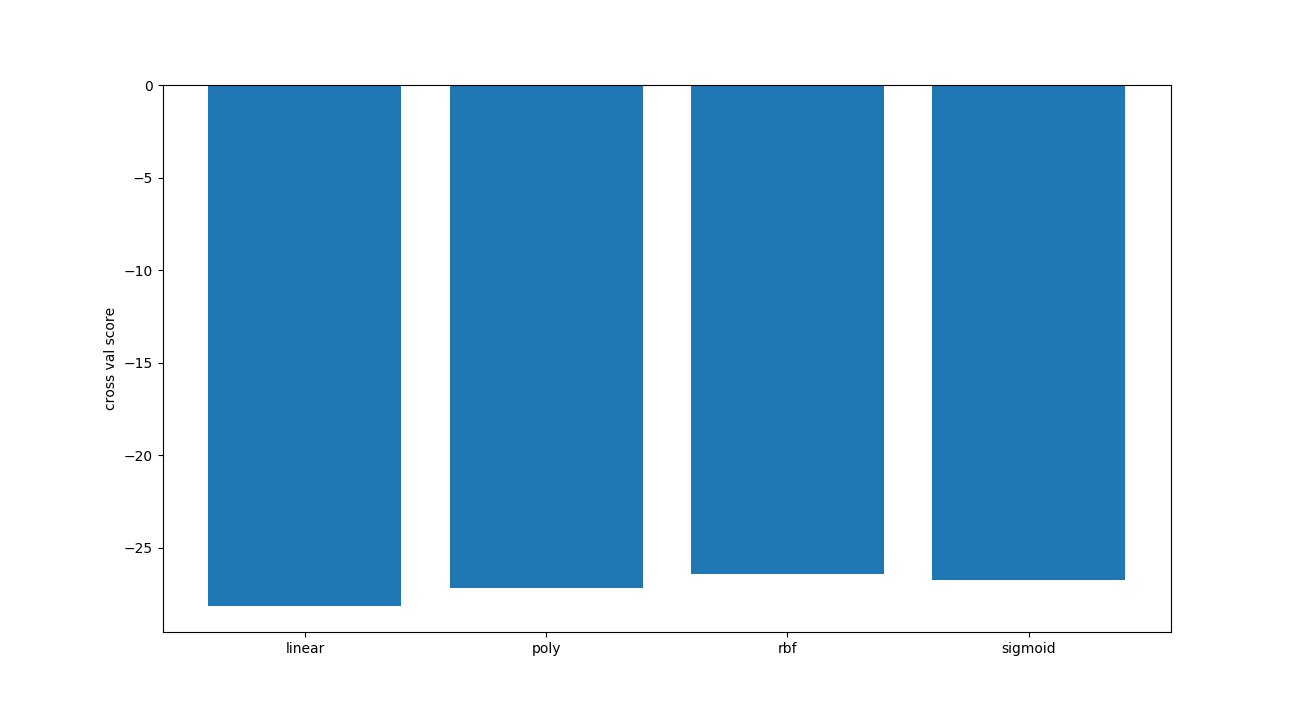
**Removing outliers**

****

As in the previous algorithm best outlier index was 0.83. It means that better to use only that records with area less than 30 ha.



Tolerance for stopping criterion. As you can see from the graph the best result is 2.

SVR has different kernel type the best for he dataset is rbf.

**mean\_squared\_error 27.6978553254**

**mean\_absolute\_error 3.25160600702**

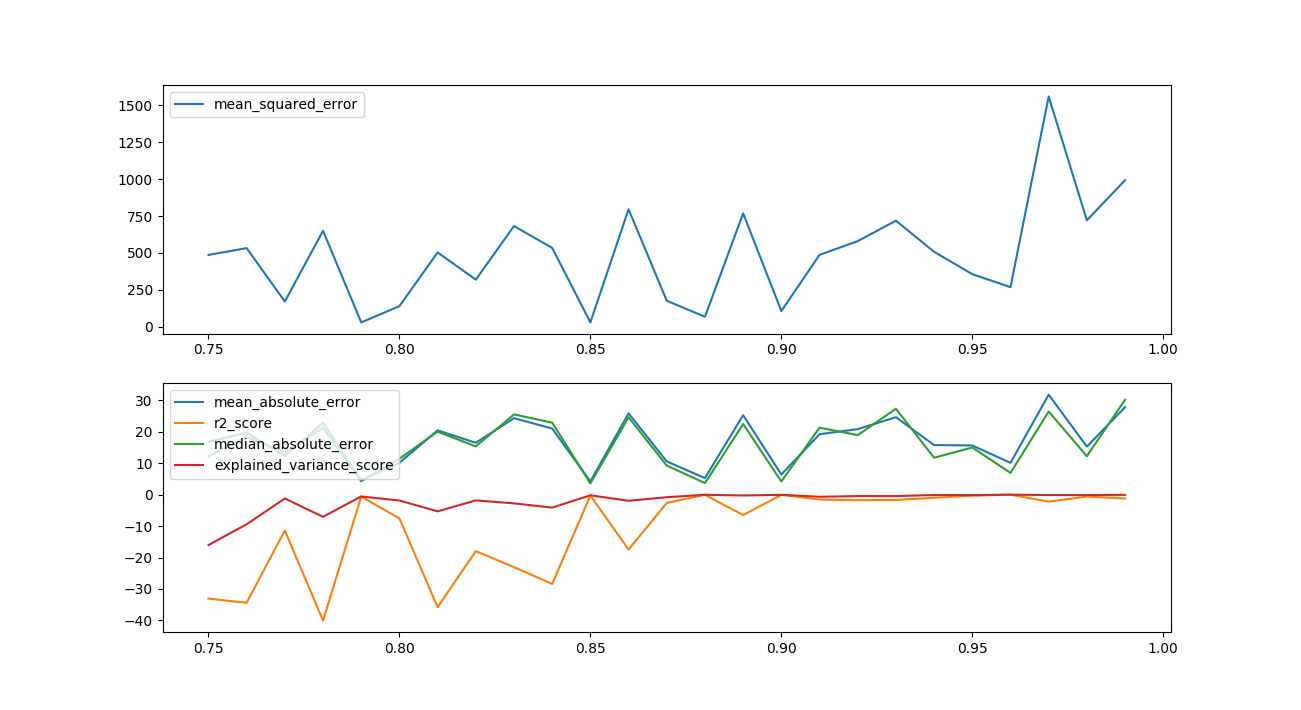
**r2\_score 0.0228243482143**

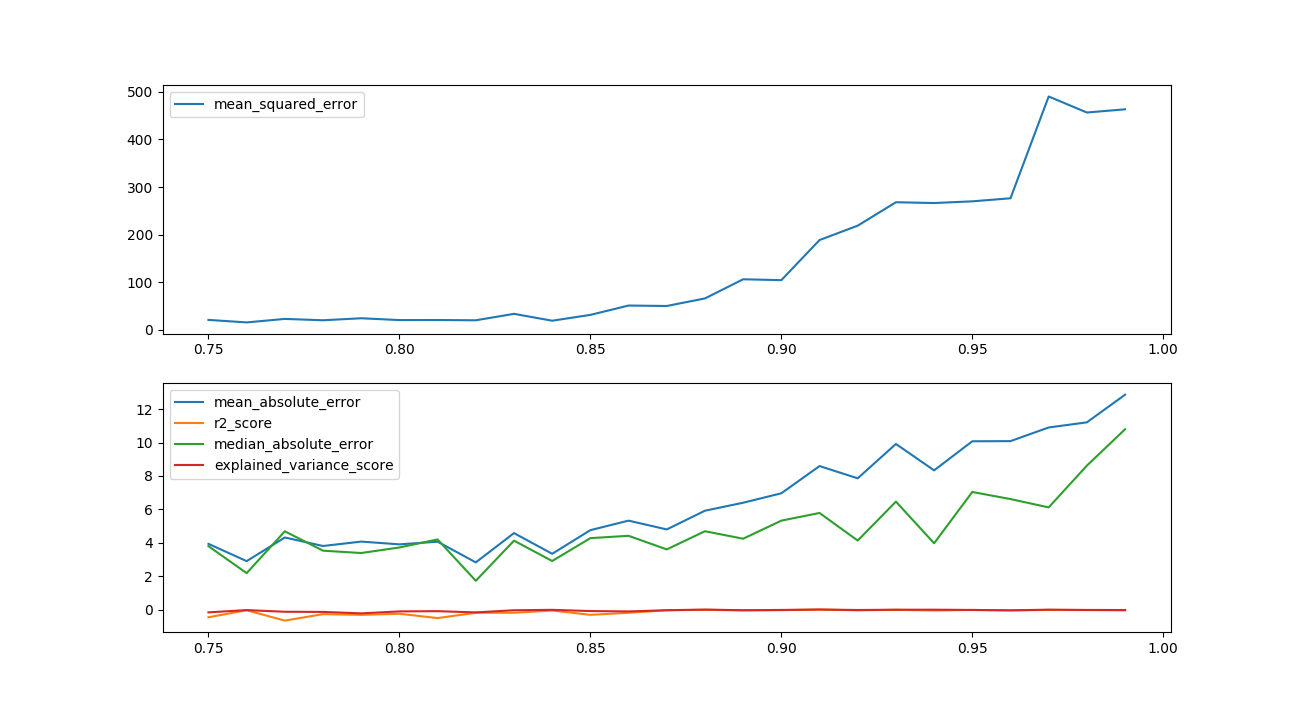
**median\_absolute\_error 2.32201573459**

**explained\_variance\_score 0.0287413250389**

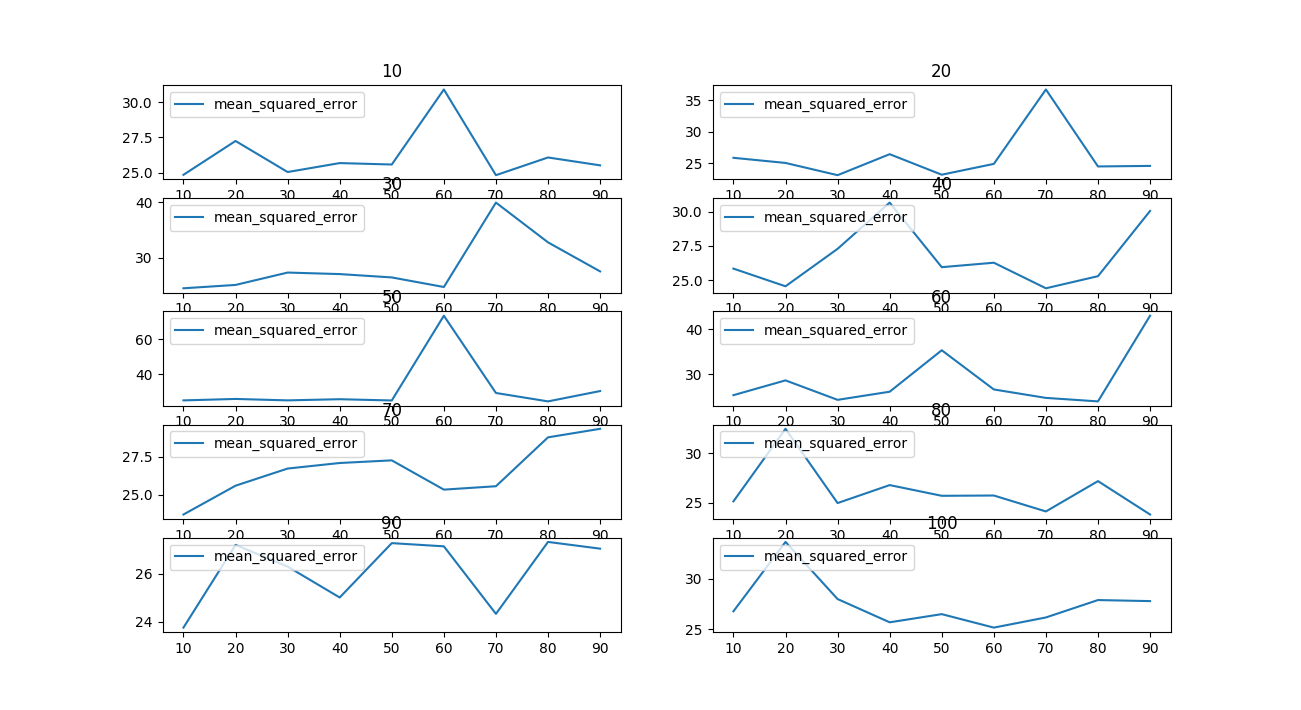
Feature selection and method get dummies didn’t work.

**MLP Regression**

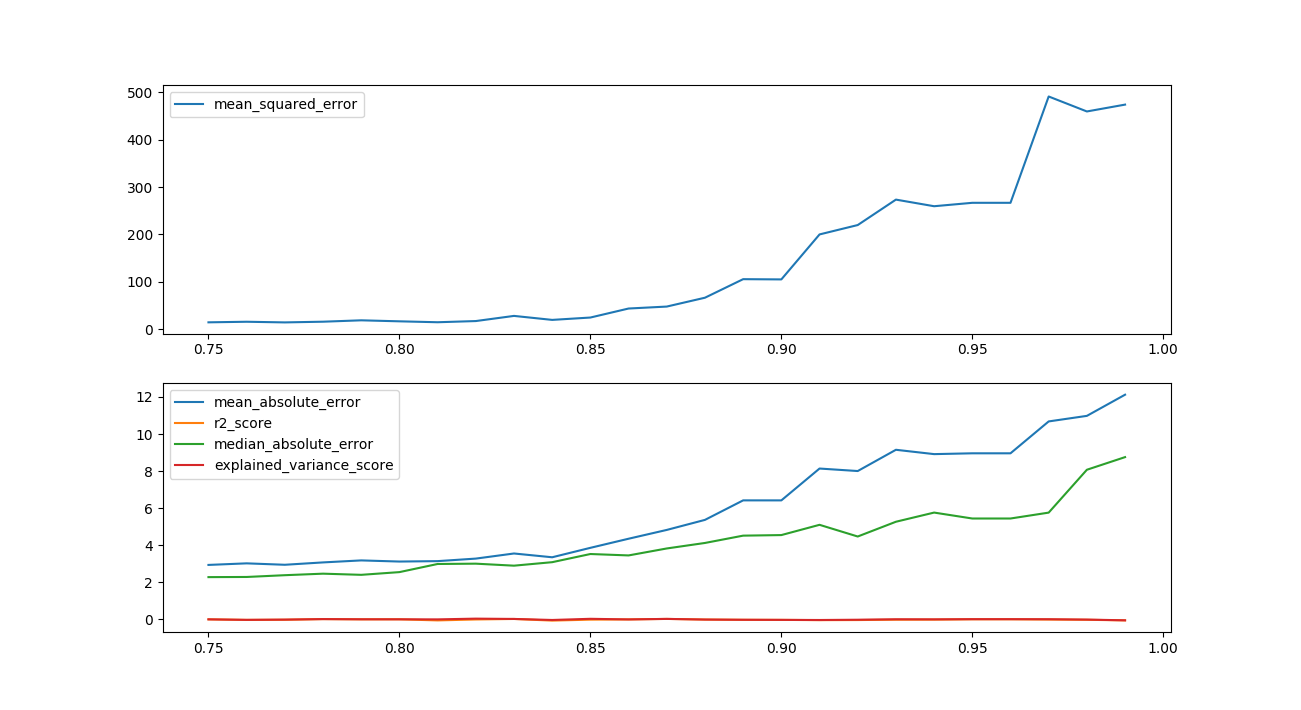
****

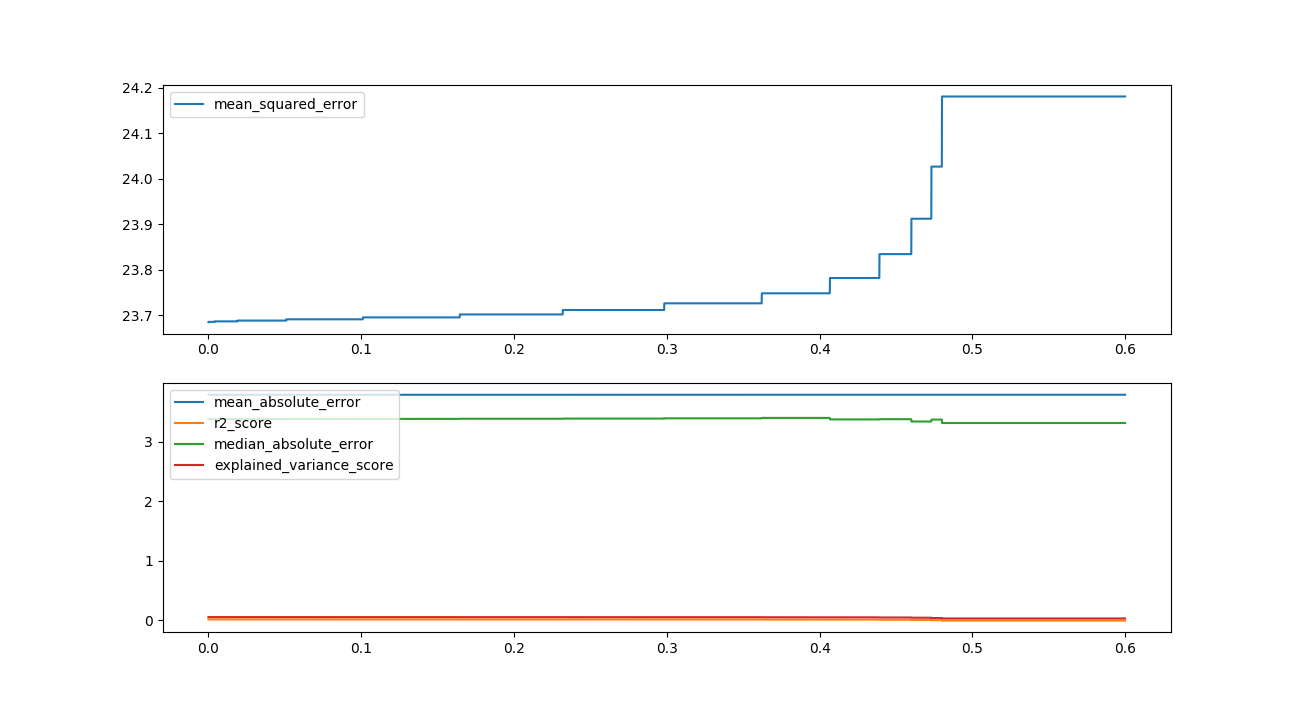
As you can see from the above mlp with outliers doesn’t give stable results.

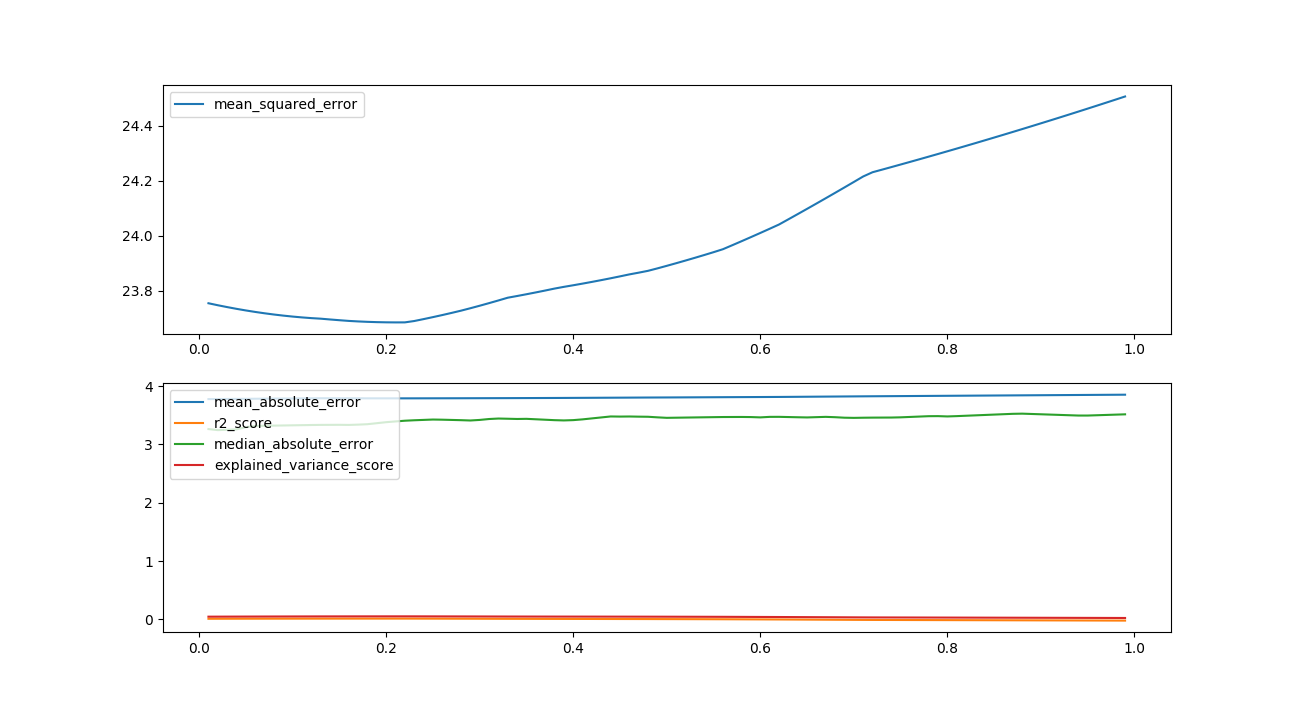
But feature selection gives us better results. Which didn’t give for other algorithms.

When I build net for mlp mse was in range of 25-30.So I decide to use 20,30,20 which gives more stable result.

**LASSO**

First of all I used outliers deduction. As expected it gives better results at area less than 30ha.

From parameters of the algorithm tolerance gives better result at 0,1

 Also one of the parameters Alpha gives better result 0.2

**Conclusion**

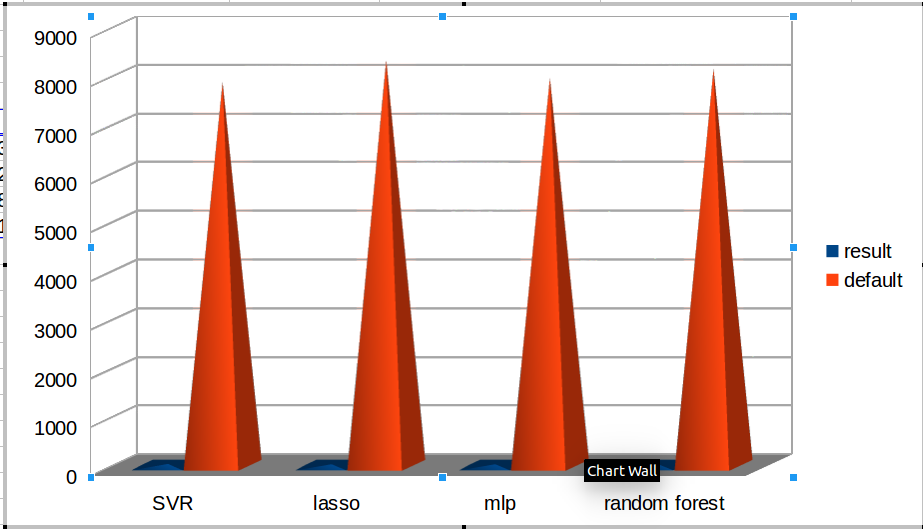
In the conclusion I can say the outliers deduction gives better result on algorithm and the biggest improvement was with this method, other methods like get\_dummies doesn’t give significant difference.

Feature selection is really important for multilayer perceptron regressor .Without feature selection MLP doesn’t give stable result, feature selection has been based on ExtraTreeRegressor, that features were TMP, RH and ISI(which has been described earlier). For other Algorithms like LASSO, Random Forest and SVR feature selection doesn’t give significant improvement.

In the end I can say the better result can be achieved on small forest fires which are the most, approximately less than 10ha.

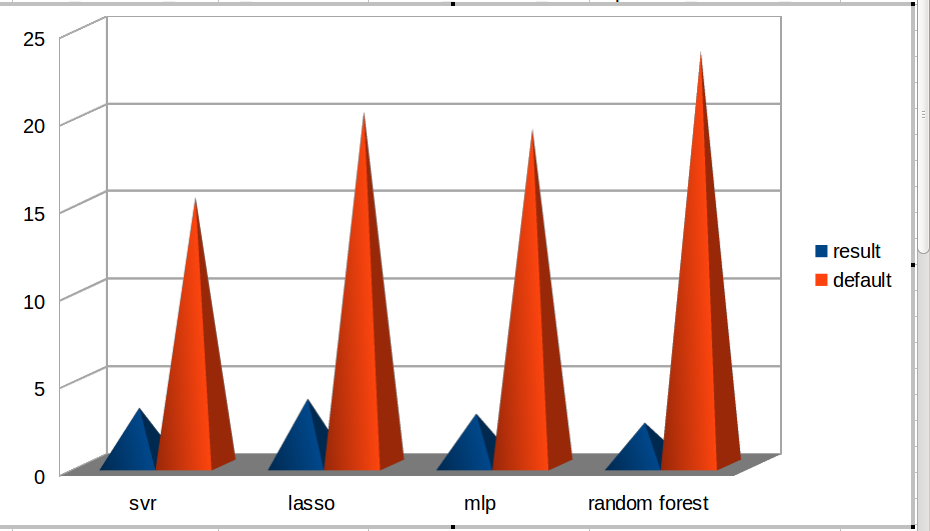
Chart result you can see below will give you understanding of improvement.

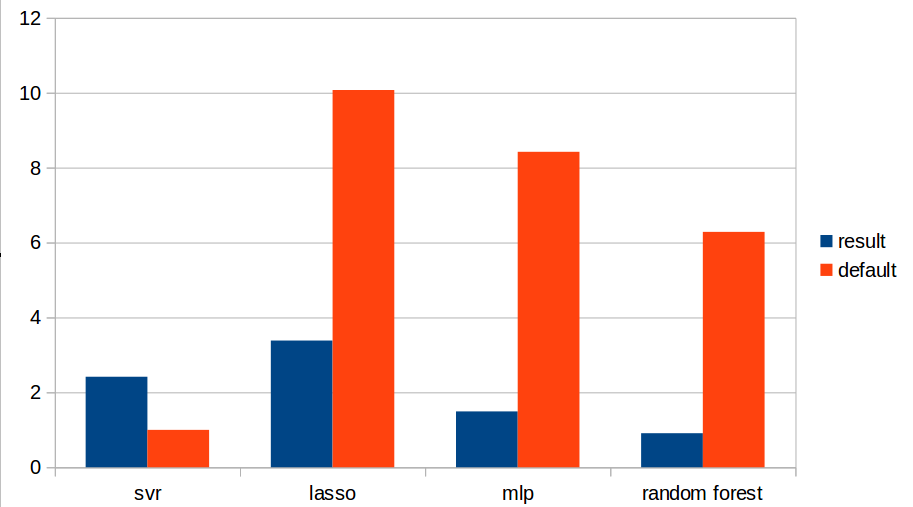
Mean Squared Error

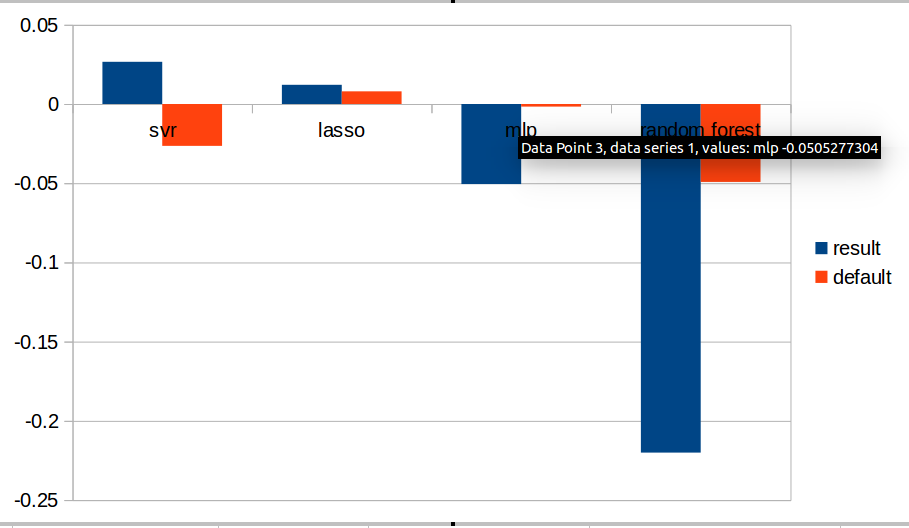


As you can see from the above chart mse has dramatically decreased, even so you can’t see result of improved algorithms.

Mean absolute error



Median Absolute Error

r2 score

explained variance score

