**Group 294: Crop Yield Prediction**

|  |  |  |
| --- | --- | --- |
| First Name | Last Name | Email address |
| Vishnu | Pajjuri | [vpajjuri1@hawk.iit.edu](mailto:vpajjuri1@hawk.iit.edu) |
| Aditya | Dubey | [adubey6@hawk.iit.edu](mailto:adubey6@hawk.iit.edu) |
| Amrutha | Gowda | [alakshmanegowda@hawk.iit.edu](mailto:alakshmanegowda@hawk.iit.edu) |

Table of Contents

[**1. Introduction** 2](#_Toc498878359)

[**2. Data** 2](#_Toc498878360)

[**3. Problems to be Solved** 2](#_Toc498878361)

[**4. KDD** 2](#_Toc498878362)

[4.1. Data Processing 2](#_Toc498878363)

[4.2. Data Mining Methods and Processes 2](#_Toc498878364)

[**5. Evaluations and Results** 2](#_Toc498878365)

[5.1. Evaluation Methods 2](#_Toc498878366)

[5.2. Results and Findings 2](#_Toc498878367)

[**6. Conclusions and Future Work** 2](#_Toc498878368)

[6.1. Conclusions 2](#_Toc498878369)

[6.2. Limitations 2](#_Toc498878370)

[6.3. Potential Improvements or Future Work 2](#_Toc498878371)

# **1. Introduction**

A new fast food chain is seeing rapid expansion over the past couple of years. They are now trying to optimize their supply chain to ensure that there are no shortages of ingredients. For this, they’ve tasked their data science team to come up with a model that could predict the output of each food processing farm over the next few years. These predictions could further increase the efficiency of their current supply chain management systems**.**

# **2. Data**

Size of our data is around 2080580 with 17 attributes. There are three datasets which contain farm data, train data and weather data which is mix of categorical and numeric values. Attribute Yield is an independent variable and rest of attributes are dependent variables.

|  |  |
| --- | --- |
| Farm id | Unique farm ids |
| Date | Dates per hour from 2016 in train and from 2017 in test |
| Ingredient\_type | Type of ingredient in the farm : There are 4 types - w,x,y,z |
| Yield | Yield for each farm per hour |

|  |  |
| --- | --- |
| Farm id | Unique farm ids |
| operations\_commencing\_year | Year the farm has started |
| num\_processing\_plants | processing plants present in the location/ farm |
| farm\_area | Area of the given farm |
| farming\_company | The company that owns the farm |
| deidentified\_location | Location of the farm |

|  |  |
| --- | --- |
| timestamp | Dates at which the weather was calculated at each hour |
| deidentified\_location | Location of the farm |
| temp\_obs | Temperature at that hour |
| cloudiness | Cloulds present in the sky at that hour |
| wind\_direction | The direction of the wind at the hour |
| dew\_temp | Dew temperature at the hour |
| pressure\_sea\_level | Pressure sea level at the hour |
| precipitation | Rainfall at the hour |
| wind\_speed | Wind speed at that hour |

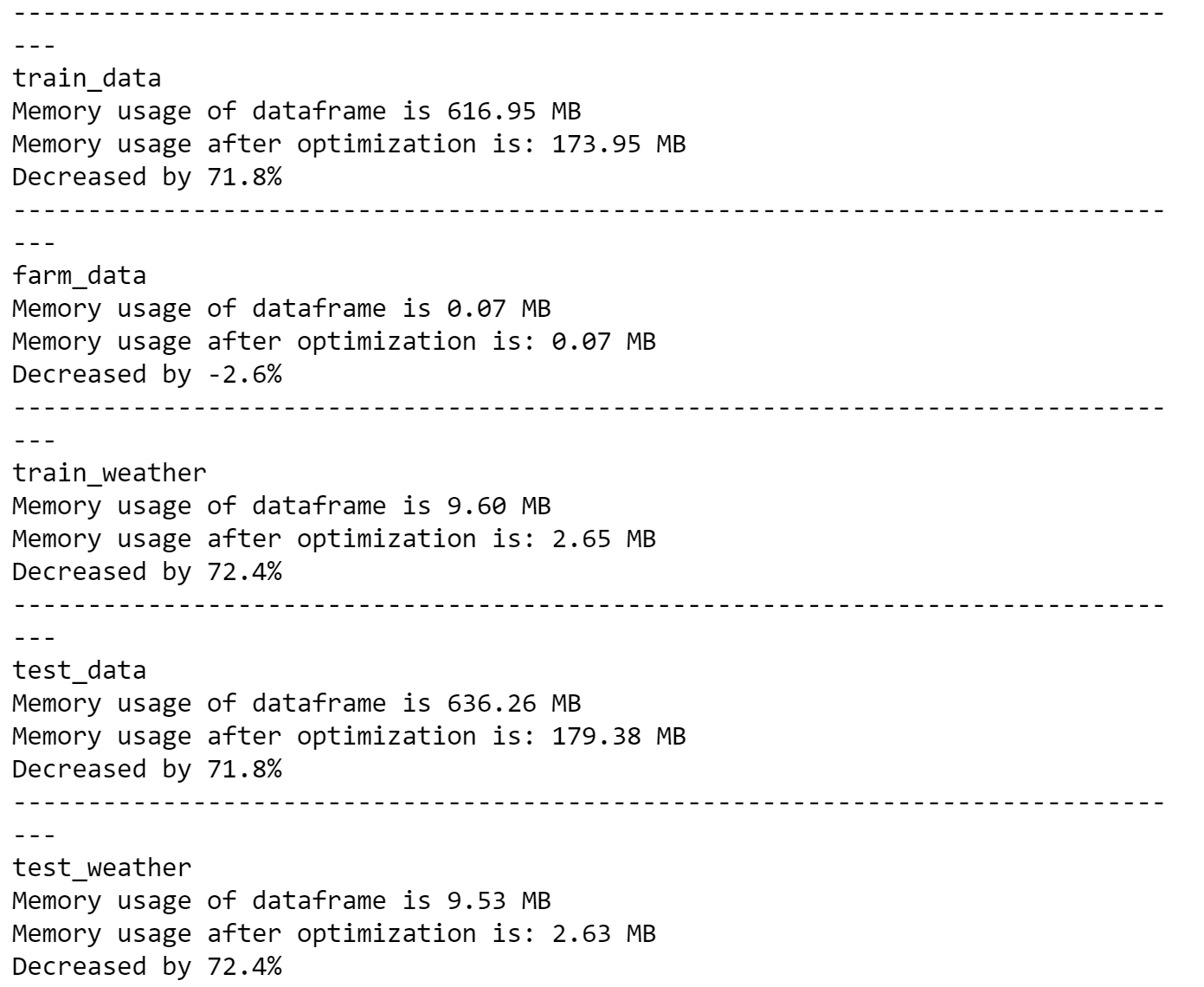
# **3. Problems to be Solved**

* Explore the data and engineer new features
* Predict the yield for each farm crop.
* Given the forecasted demand for the next few months for an ingredient, device a strategy to source it

# **4. KDD**

## 4.1. Data Processing

* Decreased the size of memory by 72% using memory optimization



* Converted date, timestamp to datetime format
* Checking and clearing duplicates
  + **135568** duplicate records have different yield values
    - Dropped the duplicates by copying to another data frame
    - Taken average and replaced the mean for duplicate values
    - Merge the average replaced records back with original data
  + **148920** records are dropped as whole rows are duplicated in the table
* Finding percentage of Null values in each column
  + Dropped columns with missing value percentage > 40%
  + Imputation is performed for records with missing values < 40% using mean.
  + Imputing with mode i.e. as 0 is occurring majority of the times in precipitation
* Merge operation on Data
  + Farm\_data + test -> initial\_merged\_test (+ weather\_test) -> final test
  + Farm\_data + train -> initial\_merged\_train (+ weather\_train) -> final train
* Removed outliers with z-score > 1.96
* Dropping unnecessary columns
  + Index, Date, Timestamp, and weekday
* Label Encoding done for categorical columns

## 4.2. Data Mining Methods and Processes

Extracted Features from Time stamp such as:

* Time Stamp: Extracting features from datetime as weekday, day\_name, dayofyear, day, month, is\_month\_end, is\_month\_start
* Weekend or Weekday: Appending weekend with 0 and weekday with 1
* Morning Evening Night: Segregated time into categorical variables

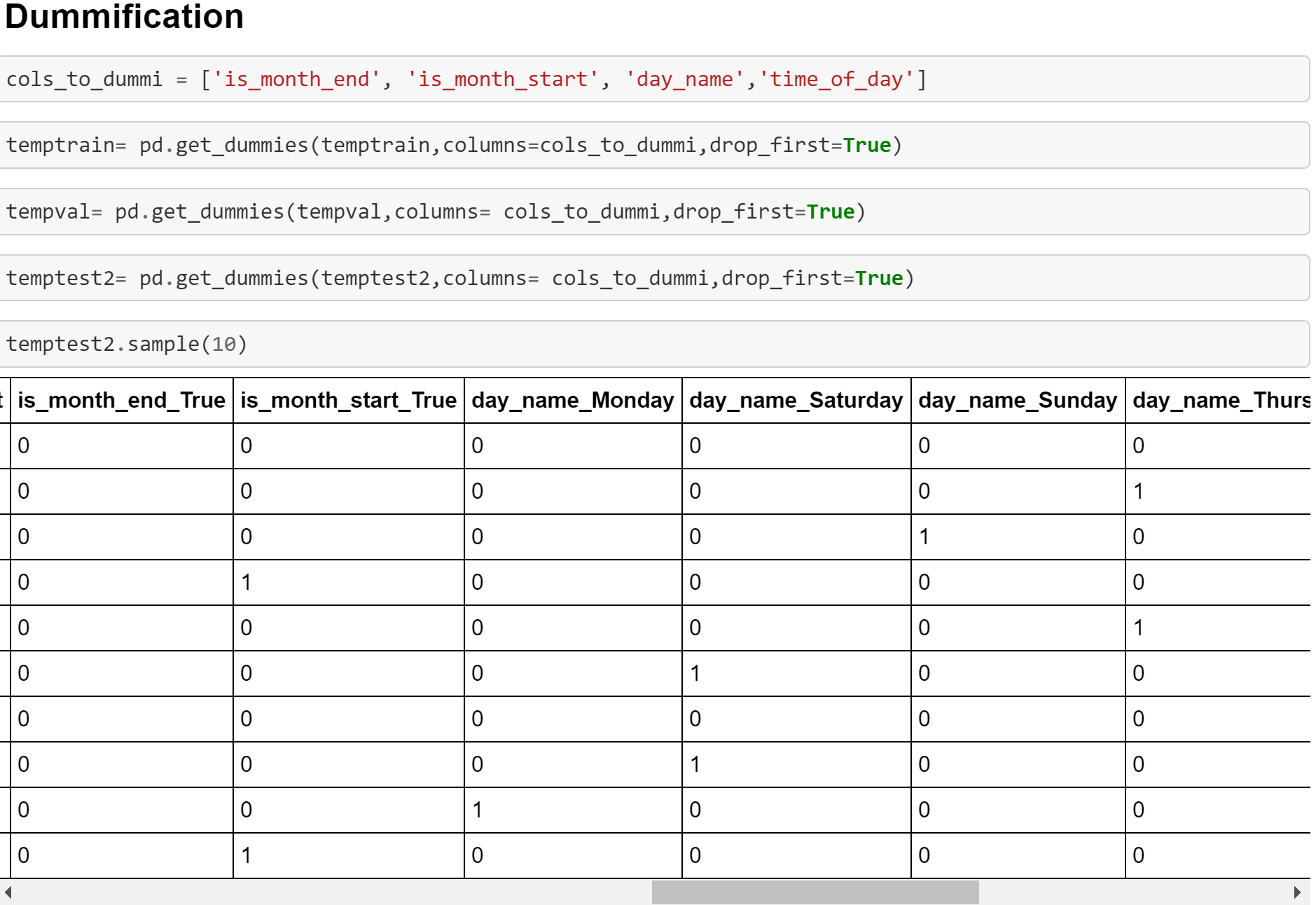
Data validation split is performed

* Hold out evaluation is performed considering the size of data
* Data split is done based on time stamp for train we used 2016 data and test 2017 data for all months and compared yields based on months

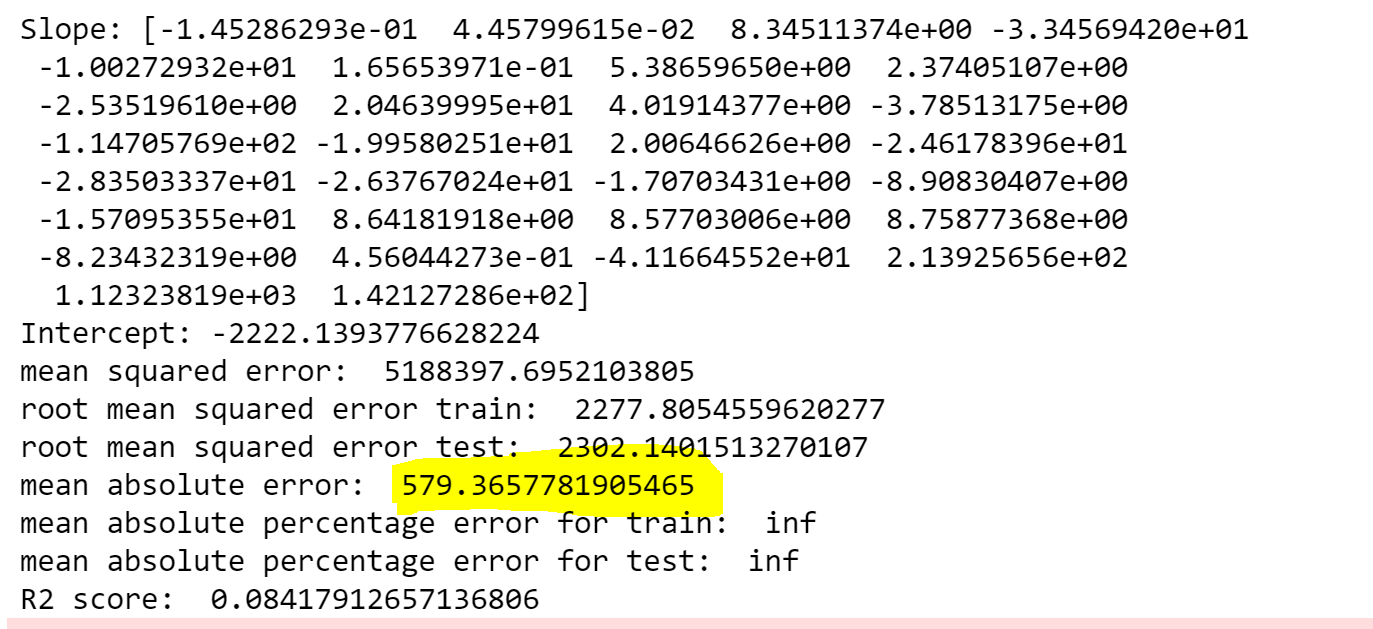
Methods used to perform Data Mining are as follows:

1. LINEAR REGRESSION MODEL

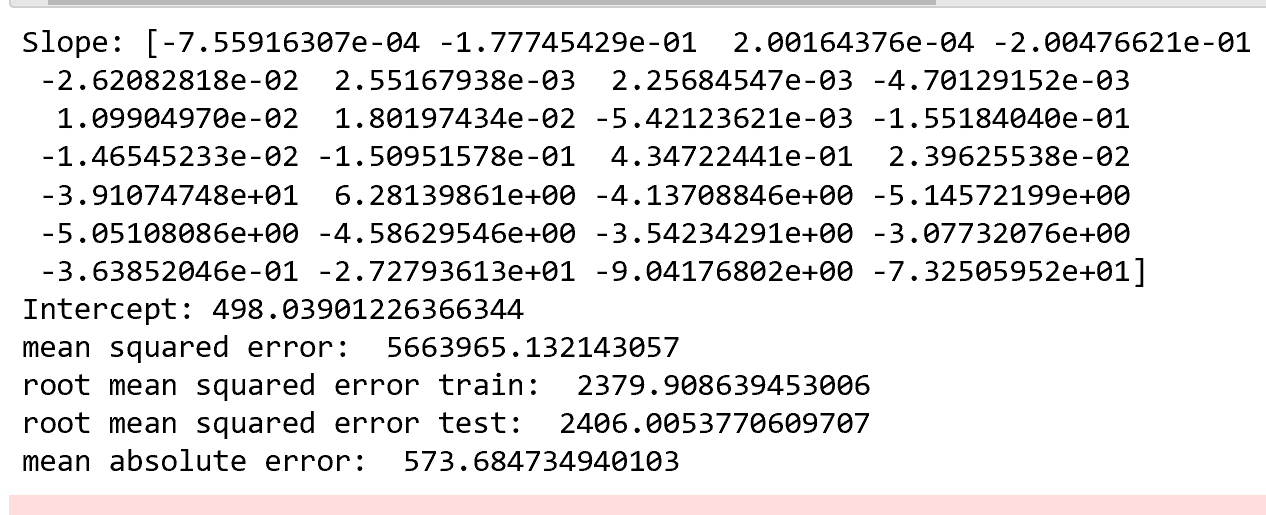
Dummification is performed on columns such as is\_month\_end, is\_month\_start, day\_name, time\_of\_day



After dummification, Linear Regression model is performed which resulted in mean absolute error of 579.36 with an R2 score of 0.084 and mean squared error is 5188397.69.



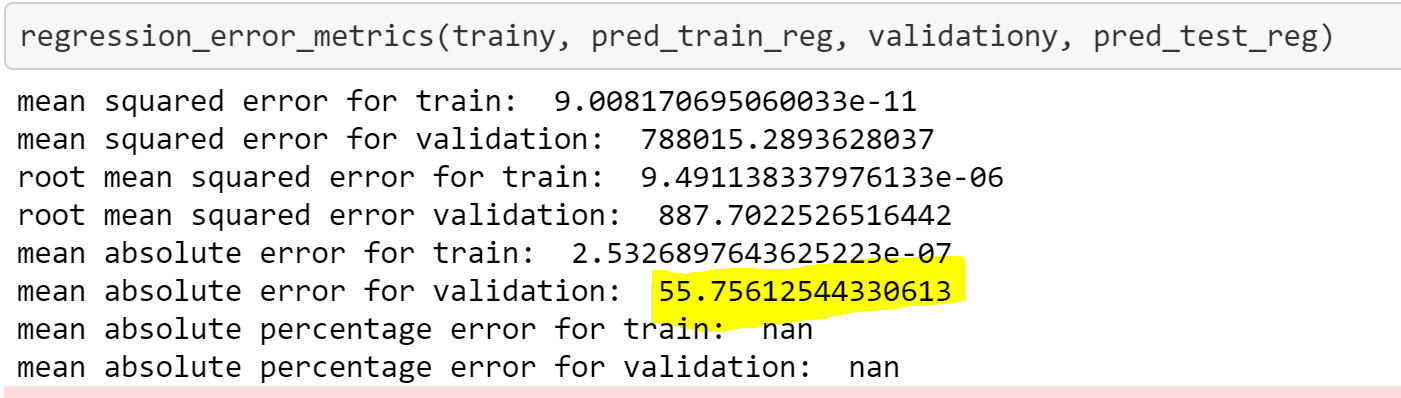
After Changes are performed on Linear Regression, error is 573.68



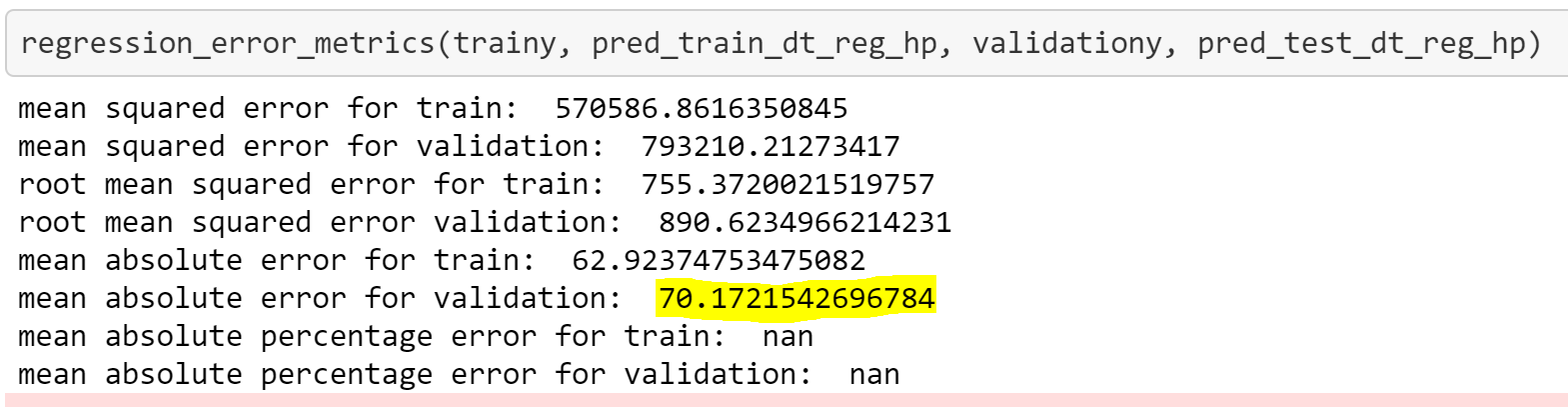
2. Decision Tree

Decision Tree is performed without outliers and with hyperparameter and we get to compare the results.

Without Outliers:

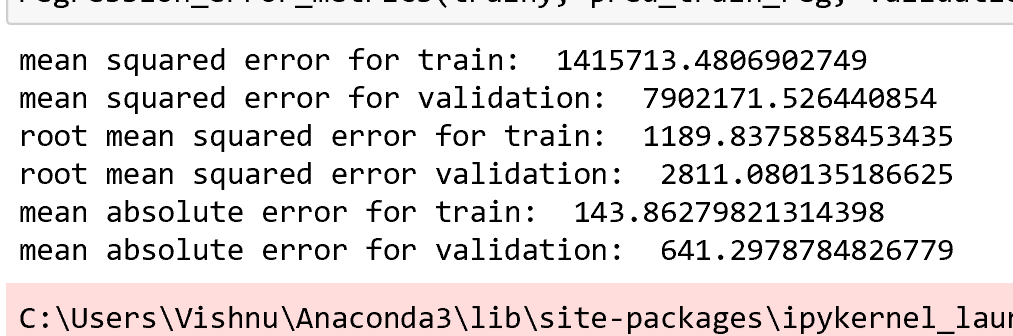


Hyperparameter:

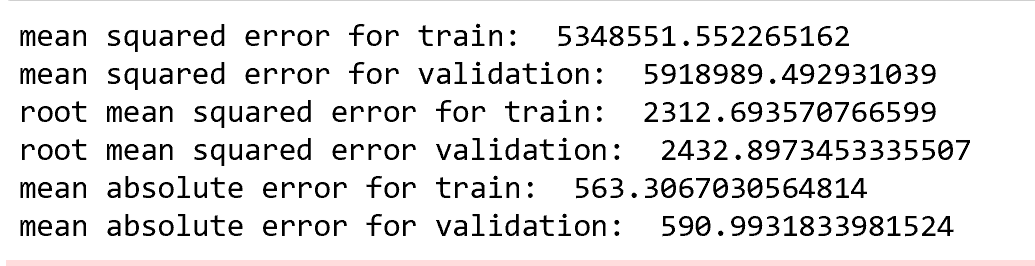


After Changes are performed on Decision Tree, we got below error:

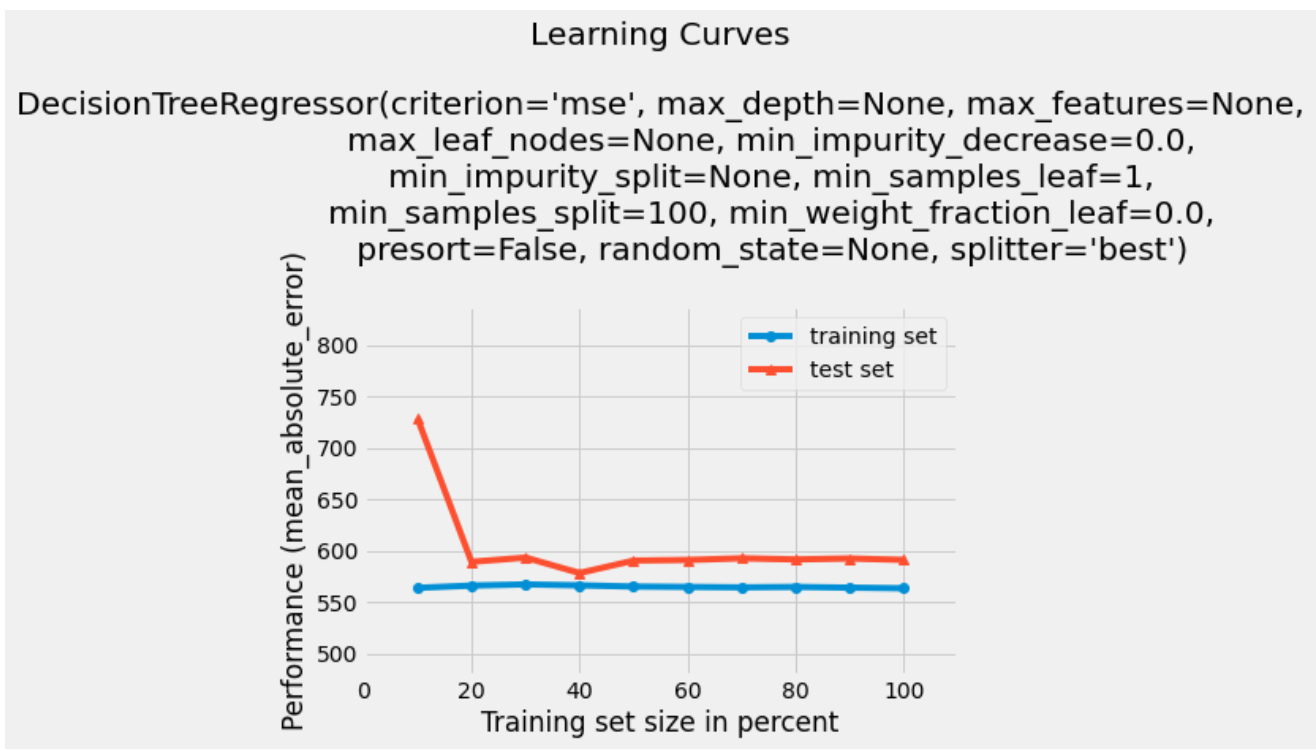
Without Outliers:



Hyperparameter:

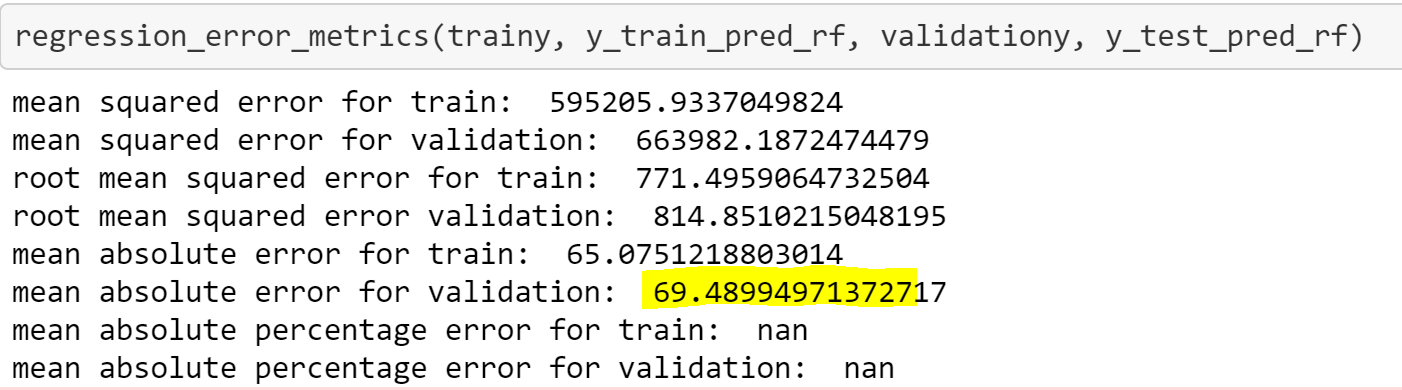


Learning Curve:

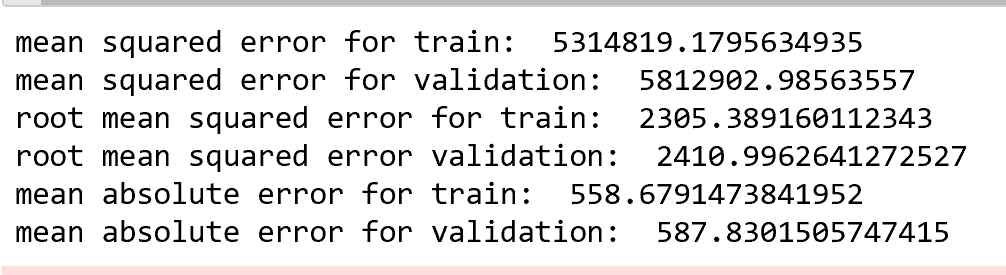


3. Random Forest

Before changes we got Mean absolute error of 69.489 and mean squared error was 663982.18



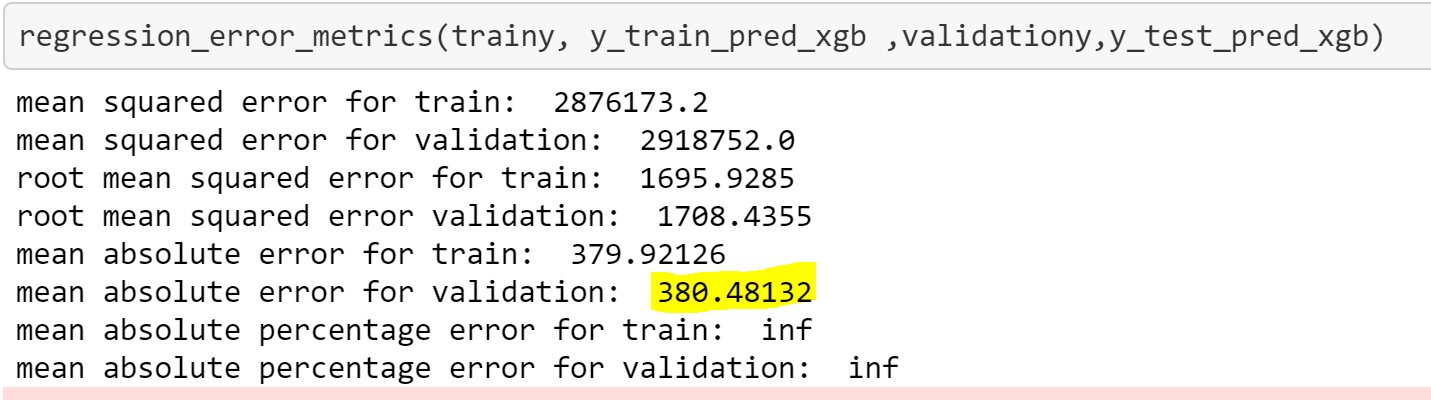
After changes performed on Random Forest, we got 587.83 mean absolute error



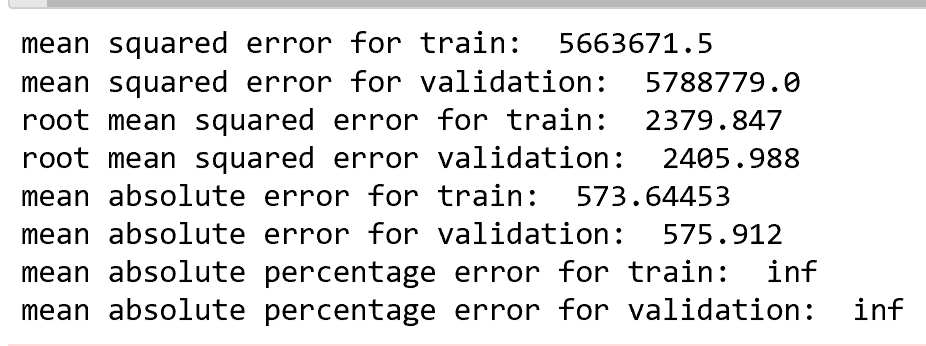
4. XG Boost

Preprocessing: Label Encoding done on columns dayofyear, day and hour

MAE value for XG Boost is 380.48132

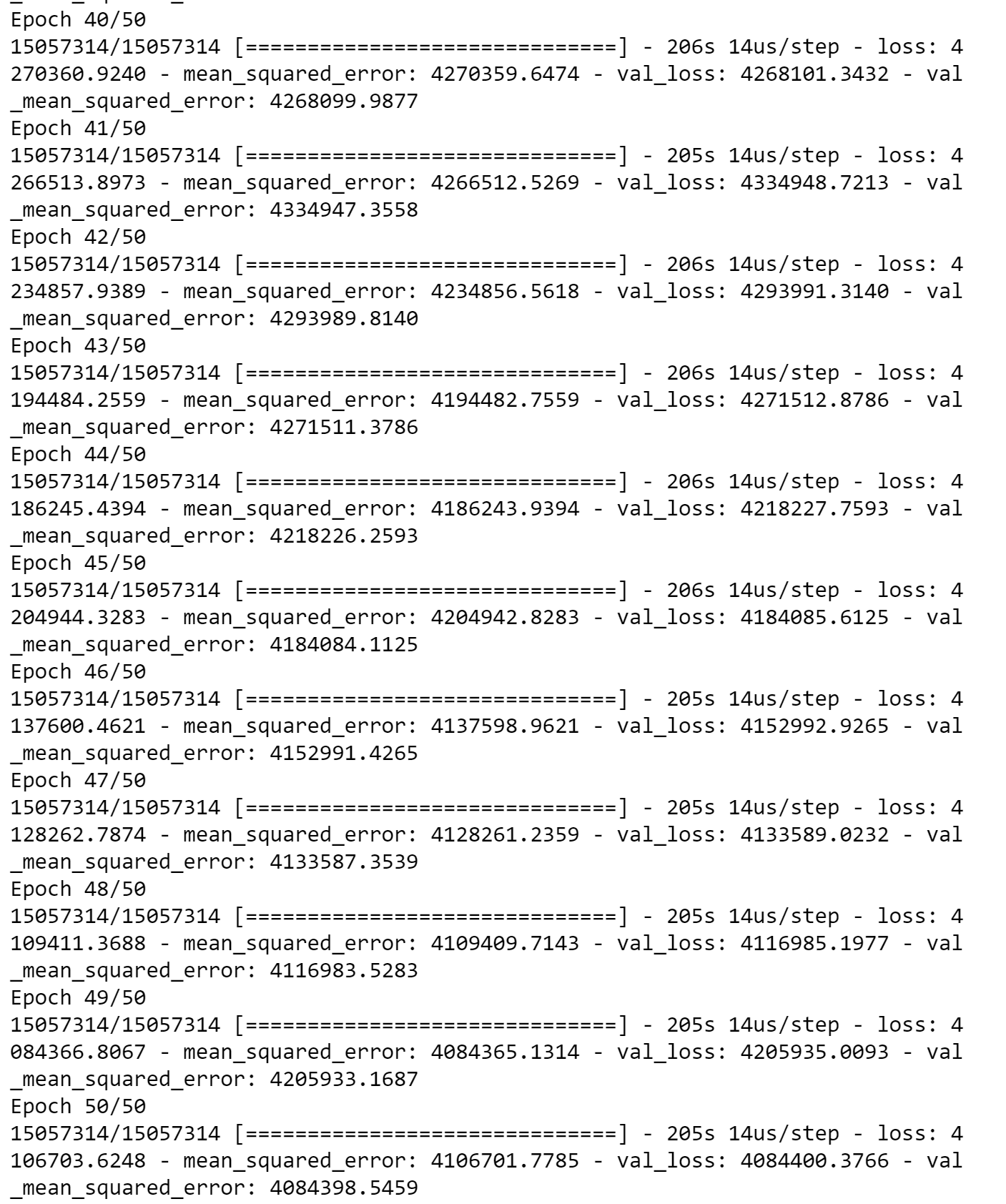


After changes performed on XG Boost, we got below results with mean absolute error is 575.912

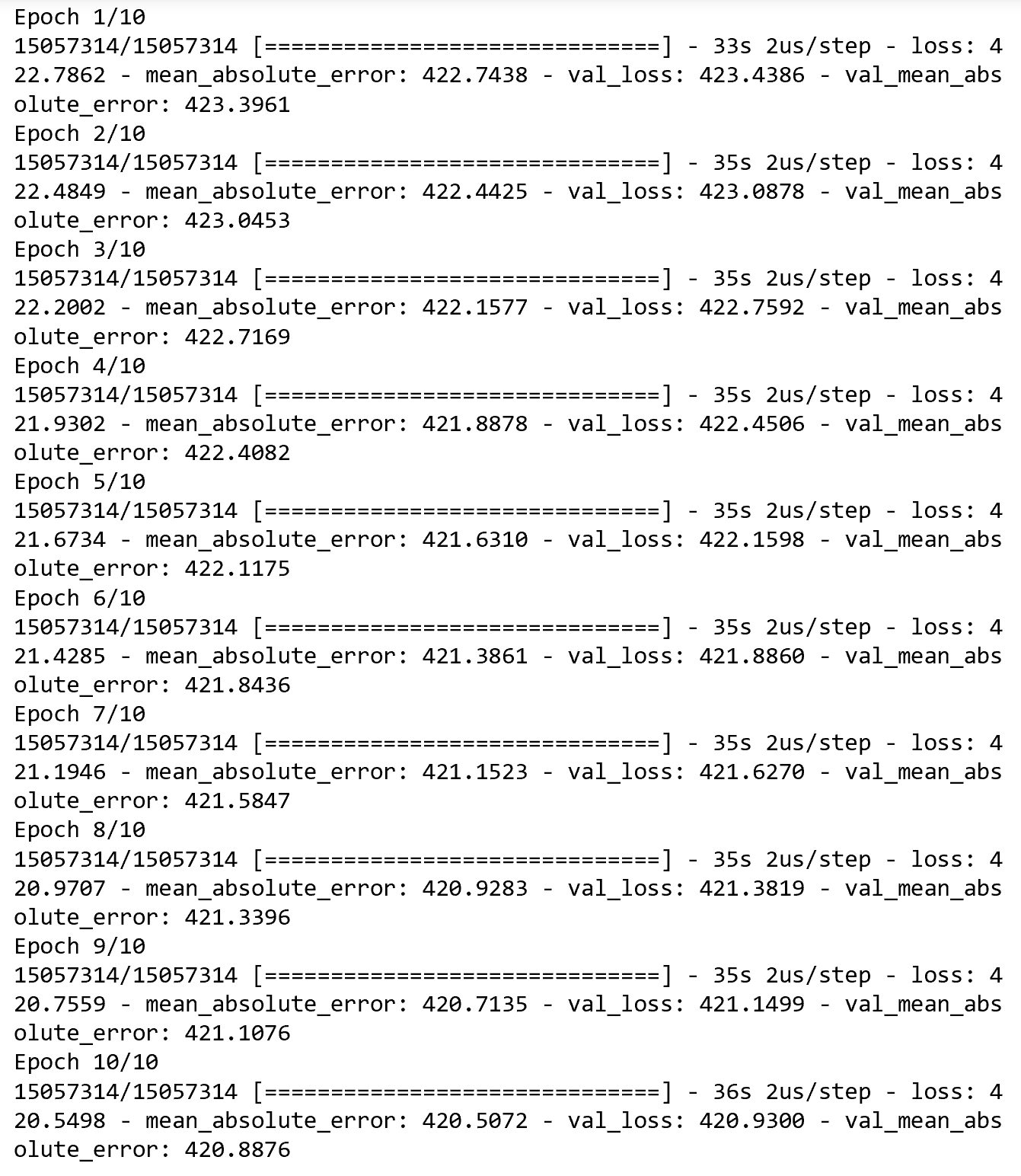


5. Neural Network

Before changes we have Mean squared Error 4084398.5459



After changes performed on XG Boost, we got below results with mean absolute error of 420.93



# **5. Evaluations and Results**

## 5.1. Evaluation Methods

1. Compared the errors of all the models performed and evaluated the results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Linear Regression | Decision Tree without outliers | Decision Tree with Hyperparameter | Random Forest | XG Boost | Neural Networks |
| 2277.8 | 9.49 | 755.37 | 771.49 | 1695.92 | 4106701.77 |

**After changes evaluation:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Linear Regression | Decision Tree without outliers | Decision Tree with Hyperparameter | Random Forest | XG Boost | Neural Networks |
| 2277.8 | 9.49 | 755.37 | 771.49 | 1695.92 | 4106701.77 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Linear Regression | Decision Tree without outliers | Decision Tree with Hyperparameter | Random Forest | XG Boost | Neural Networks |
| 2277.8 | 9.49 | 755.37 | 771.49 | 1695.92 | 4106701.77 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Linear Regression | Decision Tree without outlier | Decision Tree with Hyperparameter | Random Forest | XG Boost | Neural Network |
| 573.8 | 564.6 | 574.37 | 587.83 | 573 | 420.88 |

1. Compared Yield produced - Actual Versus Original Demand:

Predicted yield for every month for 2017 is presented below



**After changes:**



## 5.2. Results and Findings

Decision Tree without outliers gives the least error when compared to all the other models, from which we can say that is the best model to use for crop yield production.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Linear Regression | Decision Tree with outlier | Decision Tree without outlier | Decision Tree with Hyperparameter | XG Boost | Neural Network |
| 573.8 | 143.86 | 564.6 | 574.37 | 573 | 420.88 |

Extra Yield is produced after comparing the actual consumption with the original demand, which makes it easier to know and grow crops according to the future requirements.

# **6. Conclusions and Future Work**

## 6.1. Conclusions

* Decision Tree is the best suited model, among all compared model
* Yield produced monthly is more than sufficient for consumption
* Extra yield produced can be stored future consumption

## 6.2. Limitations

* Used 2 years of data for crop prediction, number of years considered can be more to get more efficient result based on consumption growth every year.
* With historical data future demands can be predicted for next couple of years without limiting to one or two years.

## 6.3. Potential Improvements or Future Work

* Density Based clustering Technique can be used on Crop yield Prediction
* Factors affecting the crop yield production can be identified and work towards efficient way of yield prediction.