Data Science Group Project 2

Introduction

This Project is a continuation of “Data Science group project 1”, where the case study was about the contamination of a Fairfield district’s reservoir fish by mercury. The dataset of that project contains the reservoirs, fish samples, and other possible factors that could contribute to elevated mercury levels in fish and we were said to do some analytical study from that dataset. In this project, another dataset is given that contains more information than the previous dataset. In this project, we built a model that can predict the mercury level in the reservoirs by the linear regression model.

Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent and independent variable. A Linear Regression model’s main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized.

To build this regression model accurately, Data Wrangling forms an essential component of data preparation. It is the process of ‘cleaning’ unstructured data sets so that they can be explored and analyzed more effectively. It can involve –

- selecting the relevant data from a large set
- merging data sets
- fixing/removing any corrupt data
- identifying anomalies or outliers
- standardizing data formats
- checking for inconsistencies, etc.

Ultimately, the goal is to give analysts data in a user-friendly format, while addressing anything that could undermine the data modeling that is to come.

Method

Preparing Dataset & Basic Analysis

Preparing dataset for machine learning projects is a crucial first step.

At first, important libraries required for the project are imported in the project file, i.e- numpy, pandas, matplotlib, seaborn, etc. and also loaded the two datasets.

Analyzed the dataset by observing the shape (rows, columns), object type.

We merged the two datasets into one dataset on Reservoir names.

Finally analyzed the dataset description (count, mean, standard deviation, etc.), information (Null values, total columns, rows).

In this section, the dataset was analyzed by the help of python in built different libraries.

Data Cleaning (Null Values handling) & Train Test Split

Data cleaning is a lot of muscle work. The dataset has no duplicate rows. It had some null values which were fixed by replacing some null values with median, and by removing some rows.

After handling with the null values, the dataset was split into train and test dataset. This was done by a library function from “sklearn.model_selection”. The column “Mercury” was the class label and the remaining columns are the features that were split.

Feature Engineering

In this section, some unimportant columns were removed and the regression model is built.

A training dataset that's machine learning ready typically contains several types of columns (features), while you don't need them all, having as many as possible can help make better predictions. In removing some columns, the term “p-value” can help a lot. A p-value measures the probability of obtaining the observed results, assuming that the null hypothesis is true. The lower the p-value, the greater the statistical significance of the observed difference. From the dataset, those columns are taken where p-value<0.5

Again, the dataset is split into train test dataset.

Finally, the dataset is ready for building the regression model. The “train dataset” was fit by the linear regression function which is in built function of a python library and analyzed the r2 score, mean_squared_error of the model.

Visualization, Cross Validation

In this part, dataset was visualized and cross validation was checked.

Data visualization is an important part of this project. In this part, relation among the features with the label feature was analyzed.

Then k-fold cross validation was applied to the built model to determine the accuracy of the model.

In this method, we split the data-set into k number of subsets(known as folds) then we performed training on the all the subsets but leave one(k-1) subset for the evaluation of the trained model. In this method, we iterated k times with a different subset reserved for testing purpose each time.

Thus r2 score and neg_root_mean_squared_error was found for the built model.

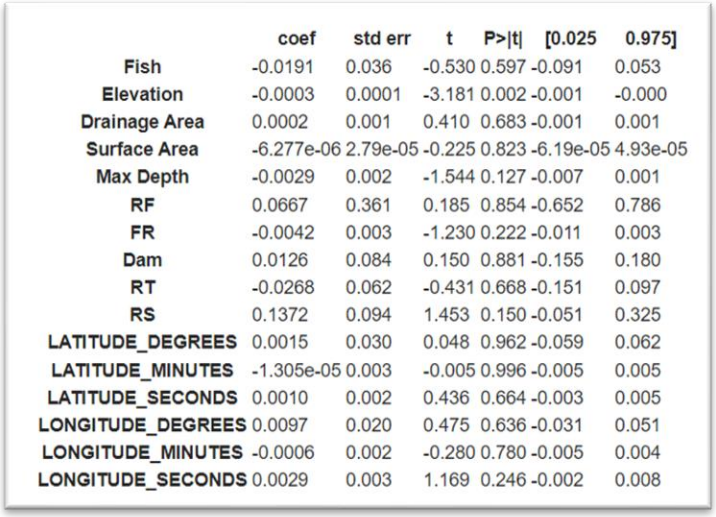
Result and Discussions

For this model, the r2 score was -0.427 and the rmse score is 0.292.

After cross validation, the r2 score was -0.219 and rmse score was 0.323.

The accuracy was low because the dataset was too small and there were some noisy features.

Some features were removed as those had a higher p-value such as Drainage Area, Surface Area, RF, Dam, RT, and some of the latitude, longitude. By observing p-values, it is noticed that some features were very important to built the model, such as – Elevation can be influence more in predicting the mercury level, max dept, longitude degrees, latitude degrees etc.



	coef	std err	t	P>|t|	[0.025	0.975]
Fish	-0.0191	0.036	-0.530	0.597	-0.091	0.053
Elevation	-0.0003	0.0001	-3.181	0.002	-0.001	-0.000
Drainage Area	0.0002	0.001	0.410	0.683	-0.001	0.001
Surface Area	-6.277e-06	2.79e-05	-0.225	0.823	-6.19e-05	4.93e-05
Max Depth	-0.0029	0.002	-1.544	0.127	-0.007	0.001
RF	0.0667	0.361	0.185	0.854	-0.652	0.786
FR	-0.0042	0.003	-1.230	0.222	-0.011	0.003
Dam	0.0126	0.084	0.150	0.881	-0.155	0.180
RT	-0.0268	0.062	-0.431	0.668	-0.151	0.097
RS	0.1372	0.094	1.453	0.150	-0.051	0.325
LATITUDE_DEGREES	0.0015	0.030	0.048	0.962	-0.059	0.062
LATITUDE_MINUTES	-1.305e-05	0.003	-0.005	0.996	-0.005	0.005
LATITUDE_SECONDS	0.0010	0.002	0.436	0.664	-0.003	0.005
LONGITUDE_DEGREES	0.0097	0.020	0.475	0.636	-0.031	0.051
LONGITUDE_MINUTES	-0.0006	0.002	-0.280	0.780	-0.005	0.004
LONGITUDE_SECONDS	0.0029	0.003	1.169	0.246	-0.002	0.008

Figure 1: Columns and Their p-values

It is important to see the scatter plot of each feature with the label feature ("Mercury").

For this model, the elevation feature can play a vital role predicting the mercury level in the reservoirs. This can be seen from the scatter plot between the elevation and mercury. It creates an equal distribution both side of a straight line.

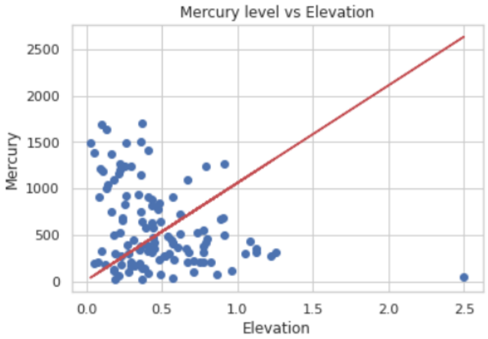


Figure 2: Elevation vs Mercury

Another important feature in this dataset is the max depth.

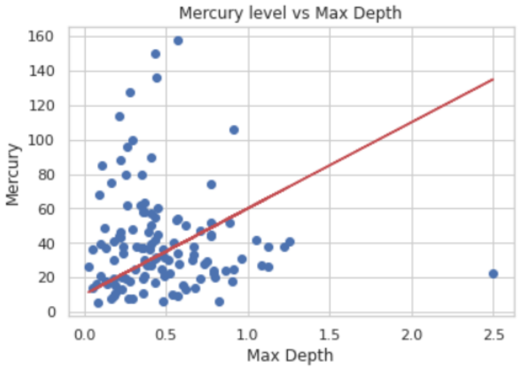


Figure 3:Mercury vs Max Depth

FR plays less significant influence as there are some outliers in this. But it can be a vital feature if there were more observations in the dataset.

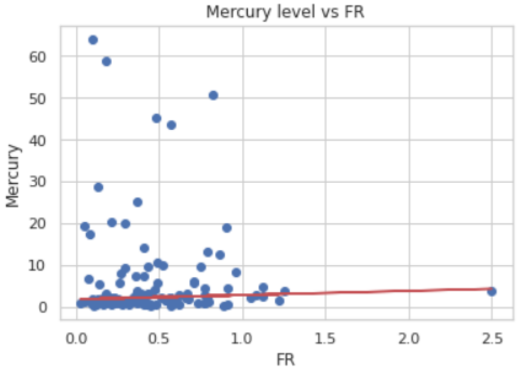


Figure 4: Mercury vs FR

Longitude_seconds also influence to predict the mercury level.

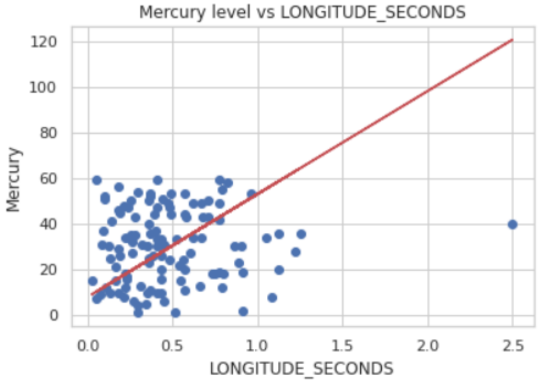


Figure 5:Lat vs Mercury

Some other scatter plots from the dataset-

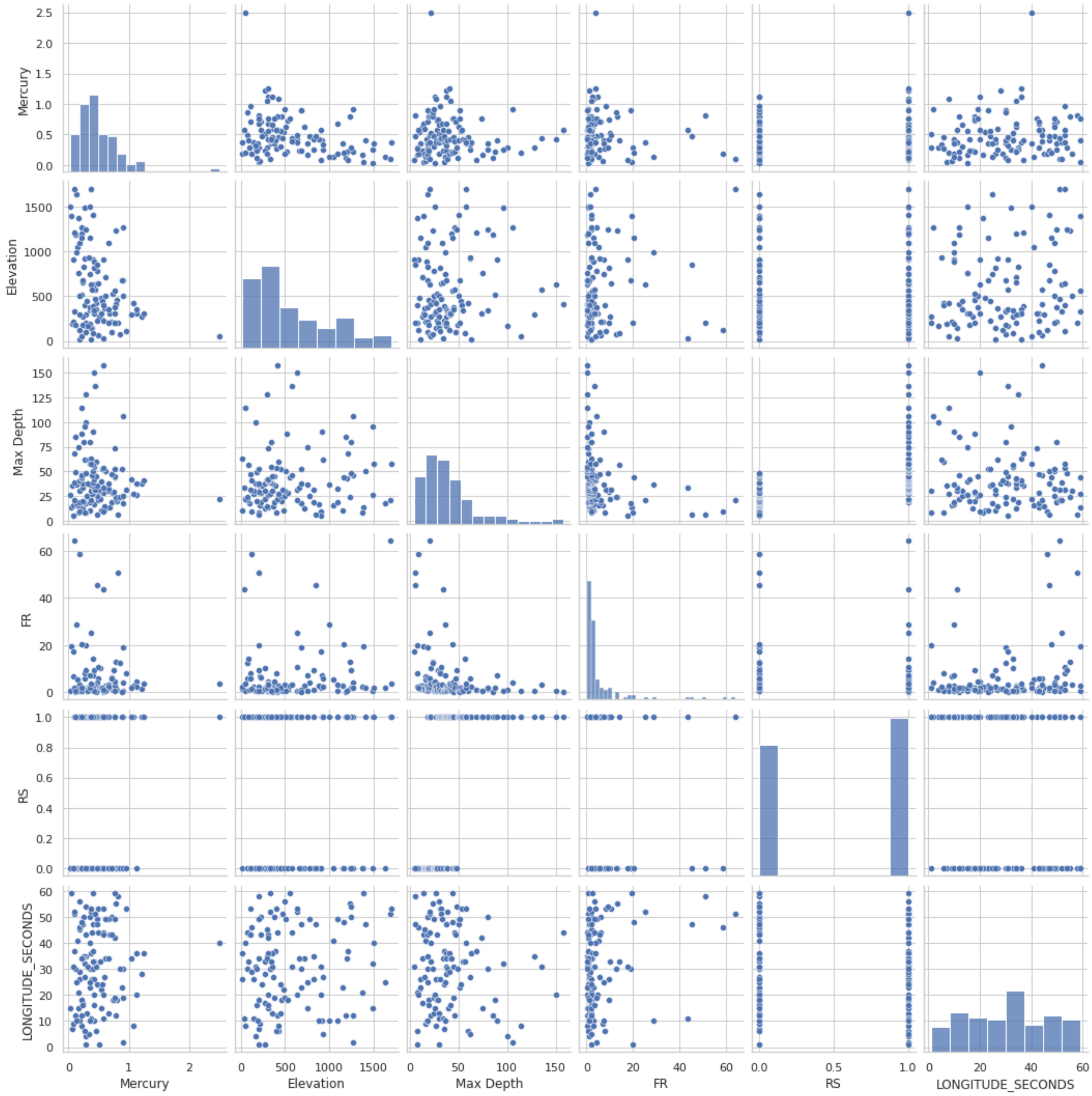


Figure 6:Subplot grid for plotting pairwise relationships in the dataset

It can be seen that some of the features are important predicting the mercury level which consequently gave us the built model. Therefore, feature selection, cleaning data and analyzing data played a vital role for building the linear regression model.

Conclusion

The model gives us a negative accuracy in r2 score. It happens whenever model's predictions are worse than a constant function that always predicts the mean of the data. The dataset records were noisy and there were a little amount of records. This may be an important cause for this negative r2 score. But for this dataset, cross validation was near around the actual score, so the model was rightly built. Moreover, the rmse score was good and it indicates the accuracy of the model.

The right features were selected in feature engineering part by analyzing the p-values and null values were handled efficiently.

References

1. https://www.geeksforgeeks.org/ml-linear-regression/#:~:text=Linear%20Regression%20is%20a%20machine,relationship%20between%20variables%20and%20forecasting. (Linear regression Model)
2. https://www.jobsity.com/blog/a-guide-to-data-wrangling-in-python?fbclid=IwAR3xm3WtnIIWRNsp3E-x1-JvM4NsOR4Euy11l7dVsUFdJqnW7UvmeVL5Qc0 (Data Wrangling Process)
3. https://www.youtube.com/watch?v=V8CNBXxyjF0&ab_channel=DhanashriKolekar (Structure of the project)
4. https://www.geeksforgeeks.org/data-cleansing-introduction/ (Data Cleaning Process)
5. https://www.investopedia.com/terms/p/p-value.asp#:~:text=A%20p%2Dvalue%20measures%20the,is%20generally%20considered%20statistically%20significant. (P-value Significance)