# Origin of Music Source

## Dataset:

959 music clips have been collected from various parts of planet. From each music clip 116 sound synthesis quantities have been extracted and are reported along with the origin of the music clip (the coordinate of the city of the origin).

Input variables: Var1 – Var116

sound and wave synthesis features extracted from each music clip. No information is available about the nature of these features.

Output variables:

Latitude: Latitude of the city the music originates from.

Longitude: Longitude of the city the music originates from.

## Approach:

Since the input features of the data had no information related to them, it was not a straight task to identify the important features by visual exploration.

Usually the regression models predict one response variable. In this problem. We are expected to predict two response variables in our problem.

The first impression is that the problem can be explored and solved either by regression or classification.

### Exploratory data analysis:

All the values in the file are numeric. So, they can be used directly to model and predict. There are no Nan or missing values in the file.

Below map plot shows that the locations can be classified based on the country/city. But it is a regression problem. So classification is not a good option.



The regression problem can be addressed either by predicting both latitude and longitude at the same time, called multivariate regression model or separate models can be built to predict latitude and longitude individually, called univariate regression models in general.

On constructing a basic regression model, it was identified that the first 72 features are unique and important. The remaining features are collinear to the previous features.

### Multivariate Models**:**

Train data: 80% sample

Test data: Remaining 20% sample

Regression:

* A normal regression model can be used to predict two response variables at the same time
* We used both latitude and longitude as response variable to build a model and predict the values.
* Parameter tuning and feature selection options like Lasso, Ridge were not available for Multivariate regression model.
* The calculated MSE was 2127.

Package: lm(formula). Coding platform: R

Multivariate Random forest:

* Since we got a basic regression score from multivariate regression model, we decided to try a couple of other multivariate prediction models.
* We fit a random forest with number of trees as 500, max features as 10 and minimum leaves as 5.
* The MSE was above 2500.

Package: MultivariateRandomForest. Coding platform: R

Neural network:

* The next feasible model which supports multi response regression is a neural network. A neural network was trained with different parameters, layers and neurons.
* The lowest MSE we got was 6076.8.
* Below are the list of neurons and layers we tried.
* 116 neurons, 1 hidden layer
* 80,30 neurons, 2 hidden layers
* 30,15,5 neurons in 3 hidden layers – MSE 6076.8

Package: neuralnet. Coding platform: R

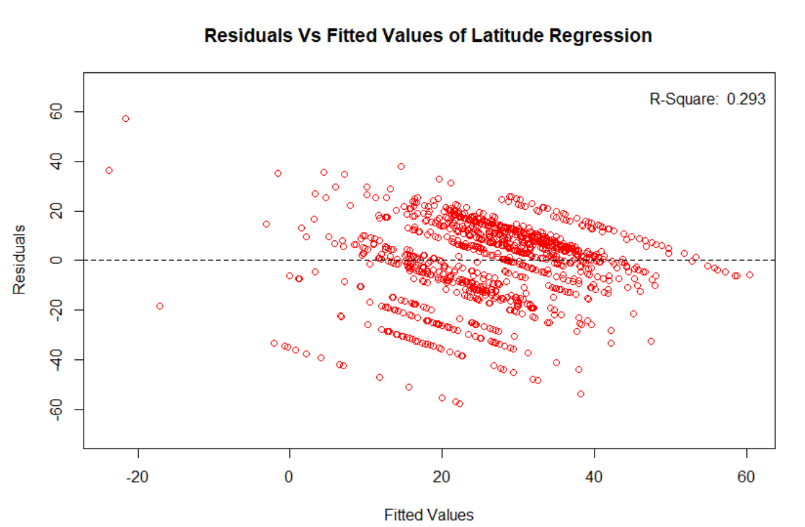
### Univariate Models**:**

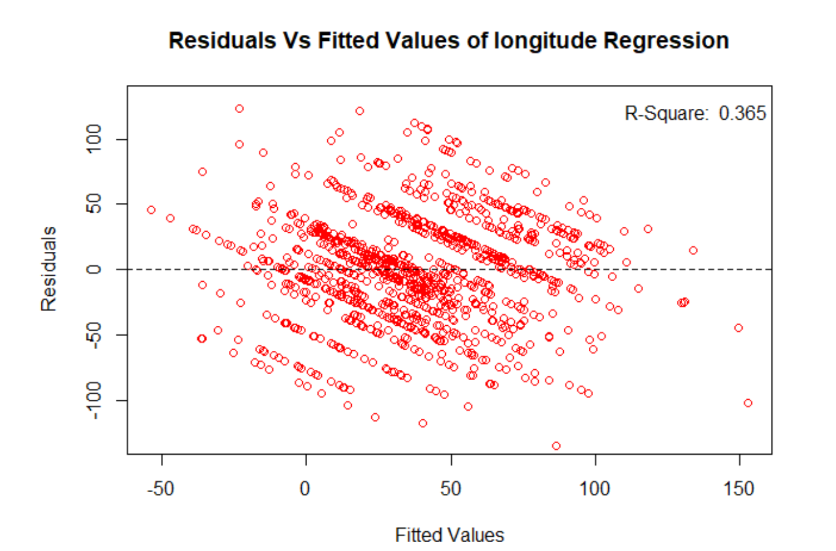
To explore the possibility of getting better results, we attempted to predict latitude and longitude separately by building a separate model for each.

* Linear Regression
* Ridge Regression
* Lasso Regression
* Elastic net Regression
* Random forest

**Linear Regression:** A linear regression is a method in statistics which is used to determine the relationship between two continuous variables. A simple linear regression fits a straight line through the set of n points.

When we applied Linear Regression, we are unable to get the better accuracy values and got MSE values of 460.56 for Latitude and 2130.65 for Longitude, as our data is not distributed linearly. As a next step we have started working with Ridge regression as there is some collinearity between the features as explained above.

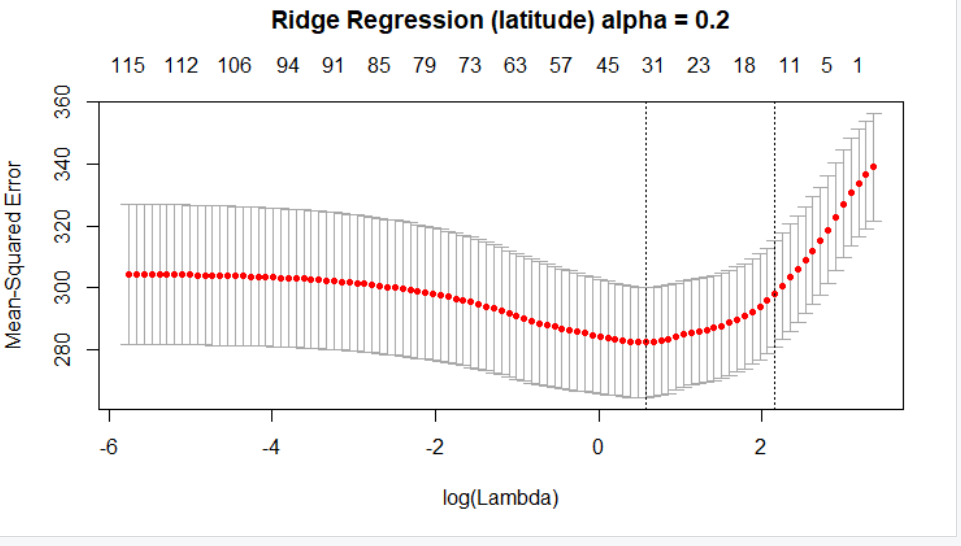


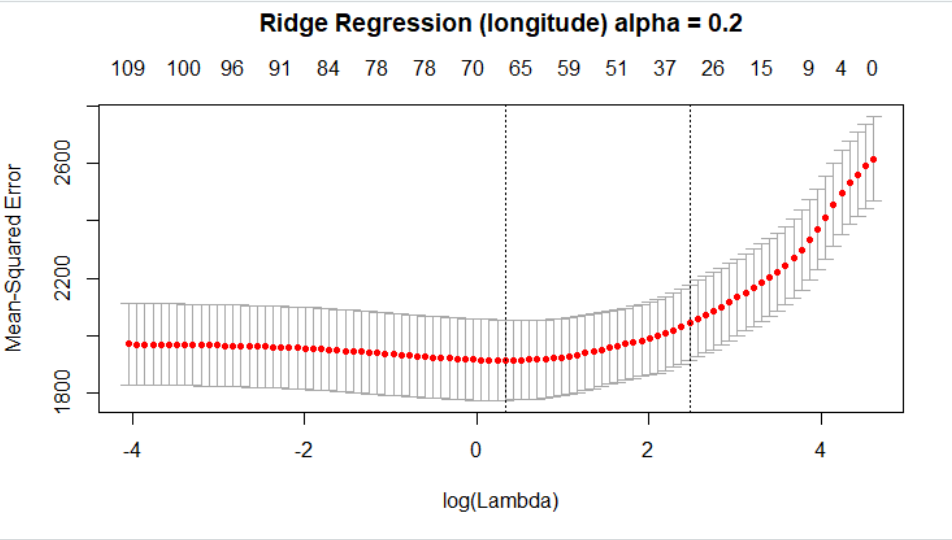


**Ridge Regression:** Ridge Regression is a remedial measure taken to alleviate multicollinearity amongst regression predictor variables in a model. Often predictor variables used in a regression are highly correlated. It performs the following operations.

* Performs L2 regularization, i.e. adds penalty equivalent to square of the magnitude of coefficients
* Minimization objective = LS Obj + α \* (sum of square of coefficients)

By using L2 regularization we were able to reduce the error when compared to Linear regression. To check if the model improves by using L1 regularization, we applied Lasso Regression.

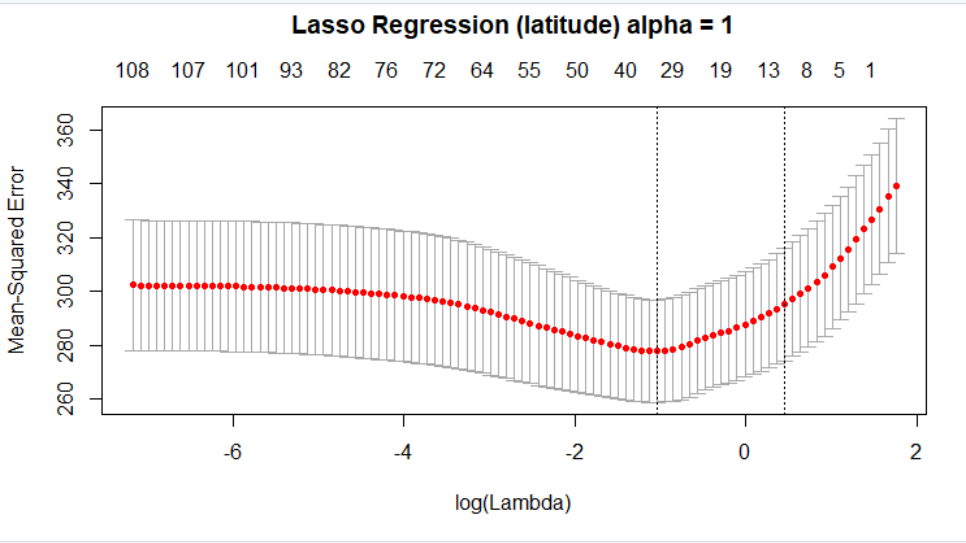


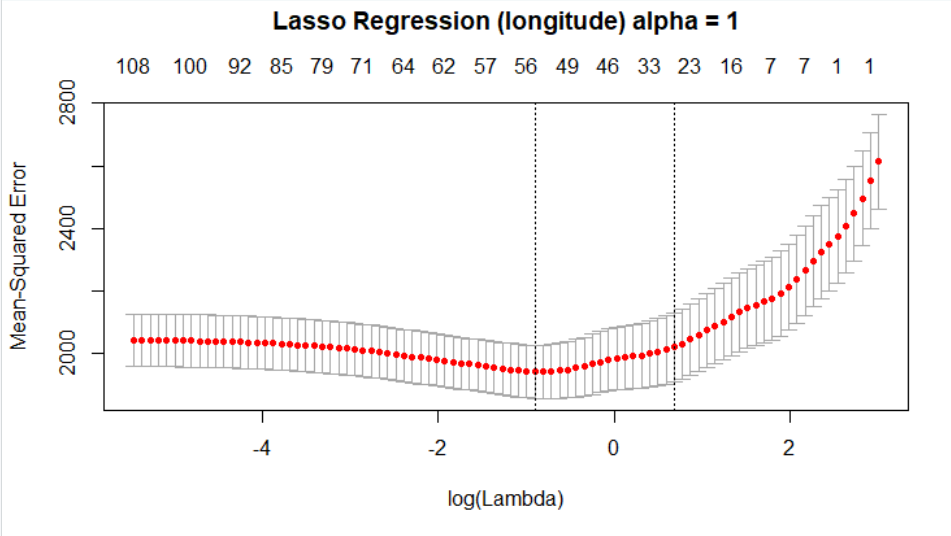


**Lasso Regression**: Least absolute shrinkage and selection operator is a [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis) method that performs both [variable selection](https://en.wikipedia.org/wiki/Variable_selection) and [regularization](https://en.wikipedia.org/wiki/Regularization_(mathematics)) in order to enhance the prediction accuracy and interpretability of the [statistical model](https://en.wikipedia.org/wiki/Statistical_model) it produces. It performs the following operations.

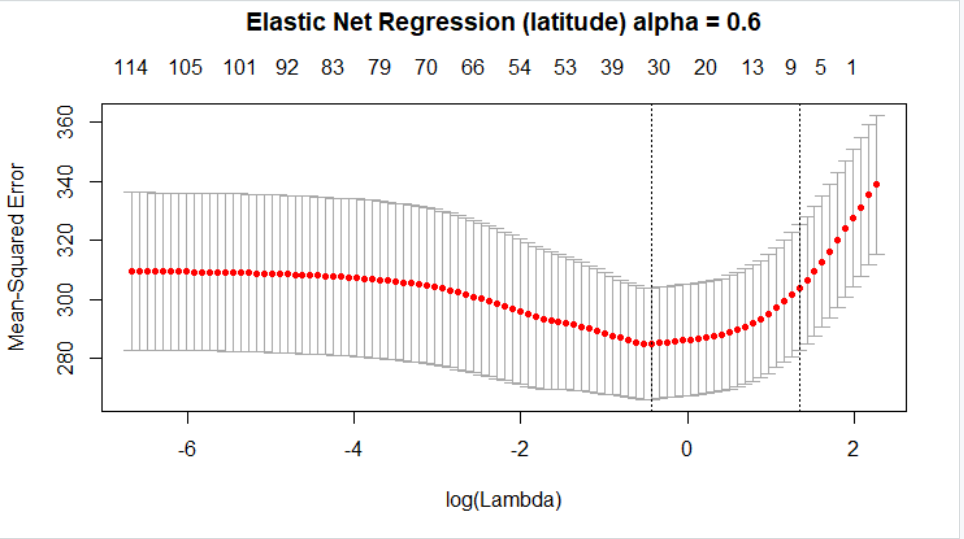
* Performs L1 regularization, i.e. adds penalty equivalent to absolute value of the magnitude of coefficients
* Minimization objective = LS Obj + α \* (sum of absolute value of coefficients)

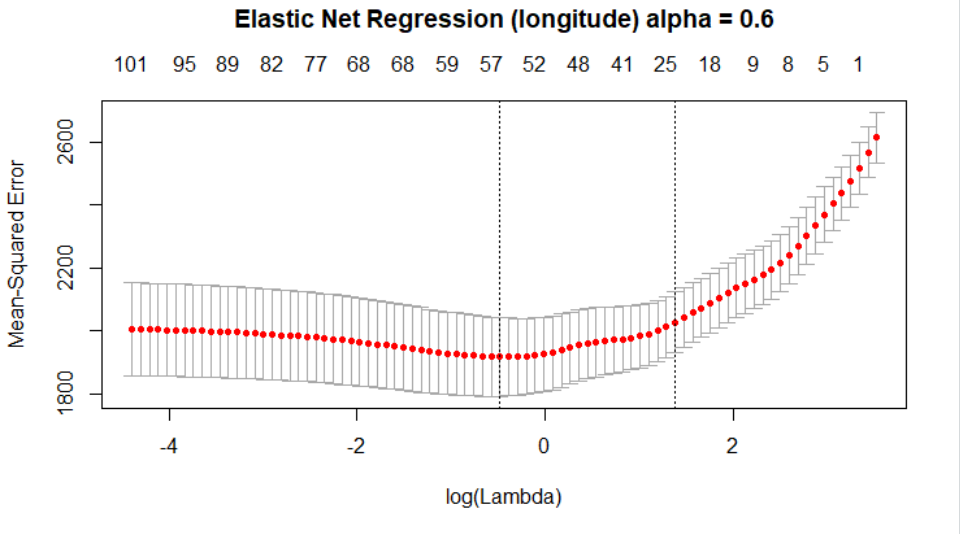
The error was reduced to 306.45 for Latitude and 2003.95 for Longitude, which is very less when compared to Linear Regression, but to get more better results we thought to use Elastic net both L1 and L2 regularization which is explained as below.





**Elastic net Regression**: Elastic net is always preferred over lasso & ridge regression because it solves the limitations of both methods. L1 regularization produces sparse solutions but tends to select the feature most strongly correlated with the outcome and zero out the rest. Moreover, in a data set with n observations, it can select at most n features. L2 regularization is suited to deal with ill-posed problems resulting from highly (or perfectly) correlated features. So by combing these two regularizations using elastic net, we are able to get better results for MSE values when compared to the above values.





The MSE values of Latitude and Latitude using different models are given in the output section below.

**Output and Models**

|  |  |  |
| --- | --- | --- |
|  | MSE Values | |
|  | Latitude | Longitude |
| Linear regression | 450.56 | 2130.65 |
| Ridge regression | 307.30 | 2004.07 |
| Lasso regression | 306.45 | 2003.95 |
| Elastic net regression | 204.6 | 1734.85 |
| Random Forest | 320.68 | 2205.06 |

# Final prediction:

After working with several univariate and multivariate models, we predicted the latitude and longitude values with Elastic net regression model with which we were able to achieve an RMSE of 44 on test data and the final prediction on test data was much lower with an RMSE value of 42.1.