Predicting Dengue Spread Using Bayesian Ridge Regression

140068C : P V Bandara
140135F : D M W A Dissanayake
140254T : K L A W P Jayashrimal
140408V : D MG C M Nalinga

Introduction

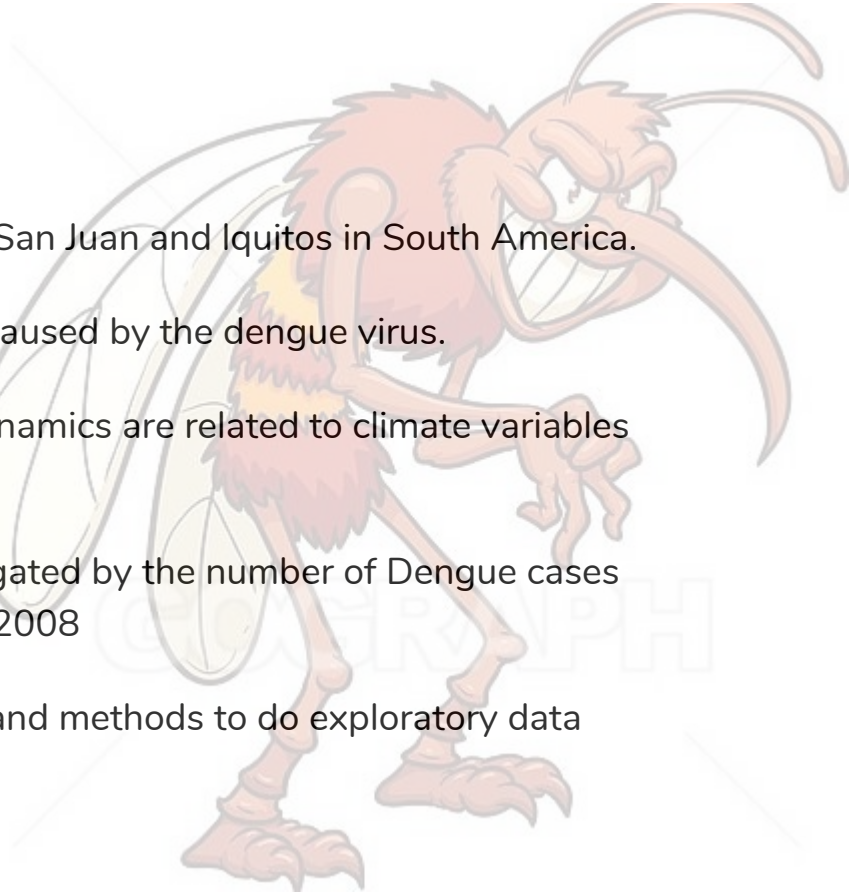
Predict the outbreak of Dengue in the cities of San Juan and Iquitos in South America.

Dengue is a mosquito-borne tropical disease caused by the dengue virus.

carried by mosquitoes and the transmission dynamics are related to climate variables such as temperature and precipitation

given historical data for these two cities aggregated by the number of Dengue cases reported weekly over the period from 1990 to 2008

machine learning model based on time series, and methods to do exploratory data analysis



Description of Data

Data set has collected in the www.drivendata.org website from National Oceanic and Atmospheric Administration (NOAA).

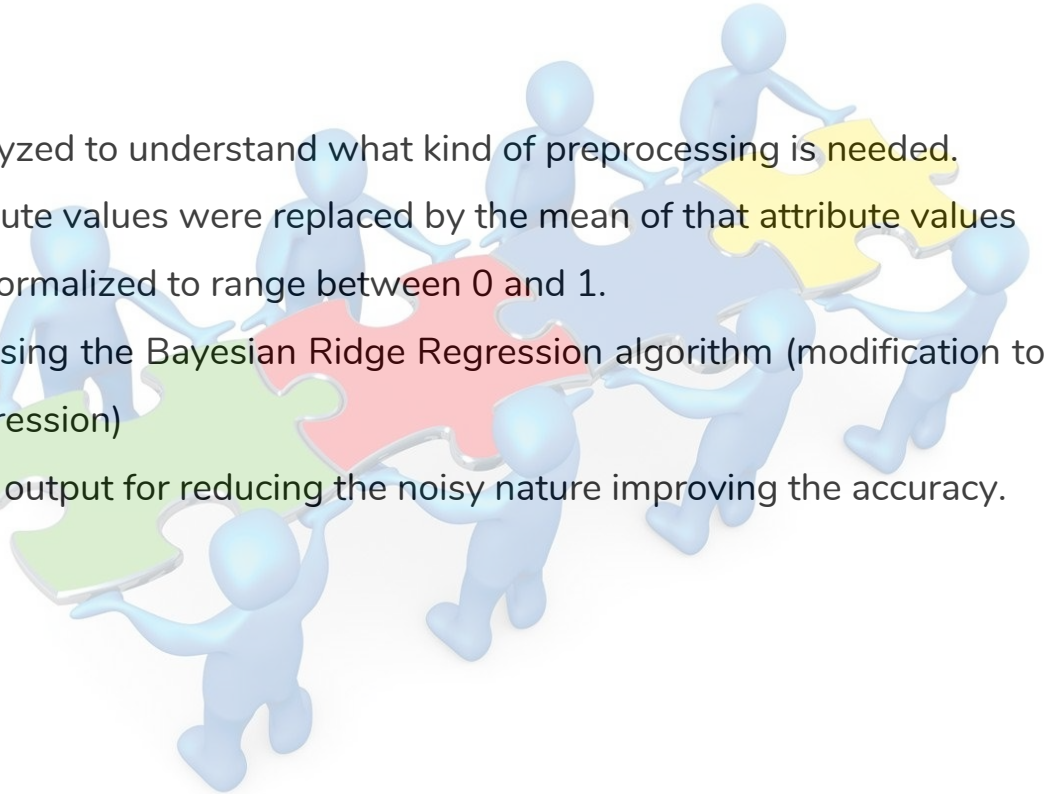
Training data set has contained 25 different attributes.

Temperature and precipitation have different types of data retrieval methods and in NOAA has two types of data recording mechanisms.

Precipitation Estimation from Remotely Sensed Information using ArtificialNeural Networks

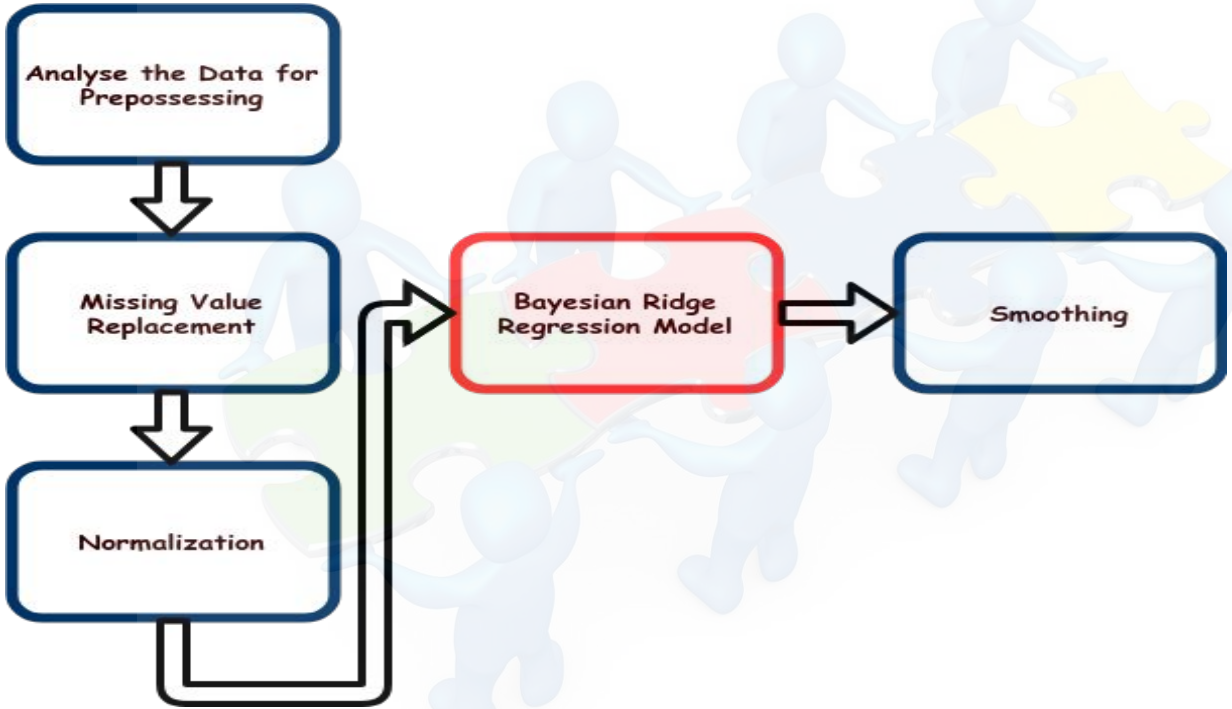
The data training dataset for city ‘sj’ runs back to April 1990 till April 2008 and for city ‘iq’ data record starts from July 2000 and ends at June 2010.

Methodology



1. The data set analyzed to understand what kind of preprocessing is needed.
2. The missing attribute values were replaced by the mean of that attribute values
3. attributes were normalized to range between 0 and 1.
4. model was built using the Bayesian Ridge Regression algorithm (modification to the Bayesian Regression)
5. smoothed out the output for reducing the noisy nature improving the accuracy.

Methodology Ctd..



Results & Analysis

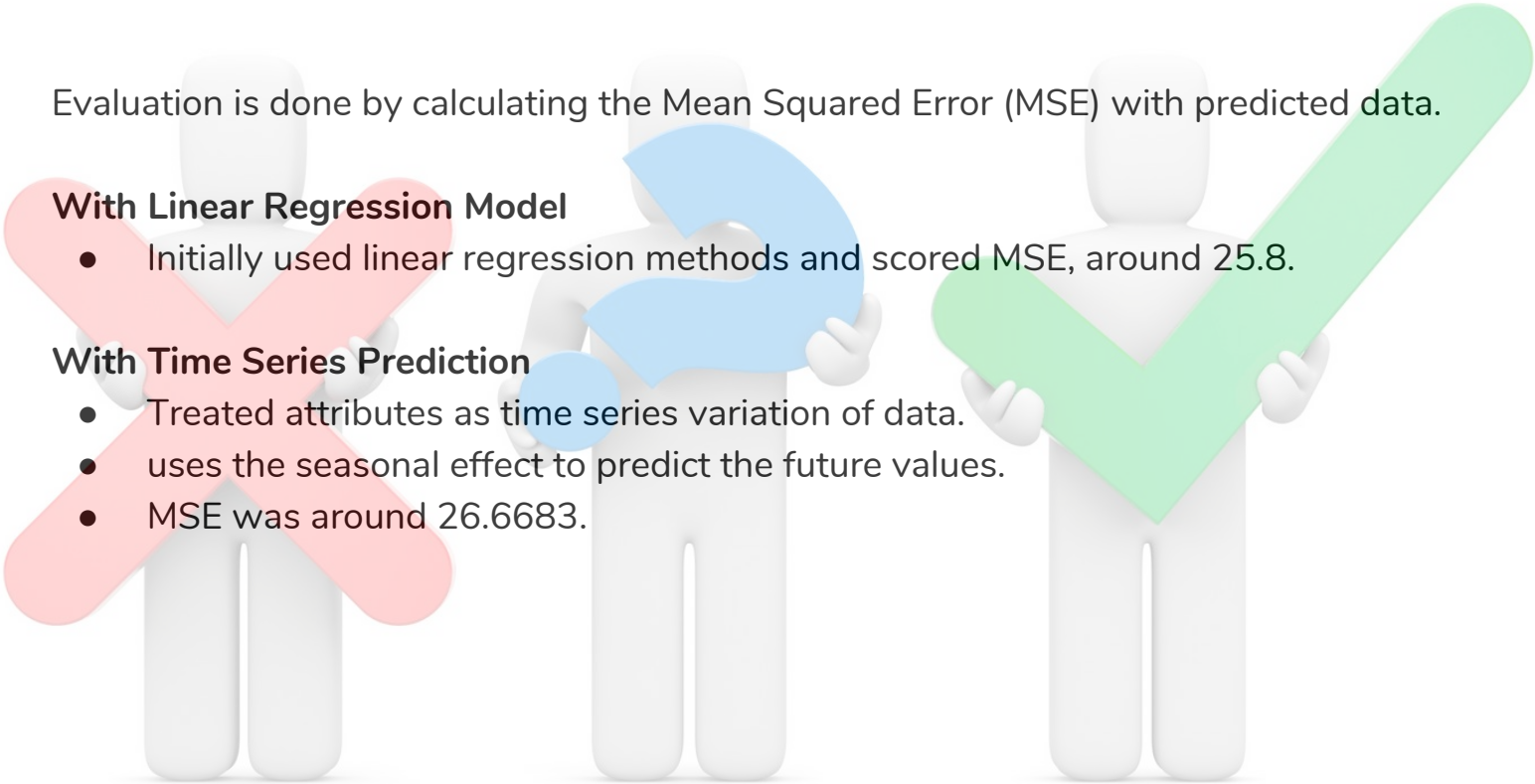
Evaluation is done by calculating the Mean Squared Error (MSE) with predicted data.

With Linear Regression Model

- Initially used linear regression methods and scored MSE, around 25.8.

With Time Series Prediction

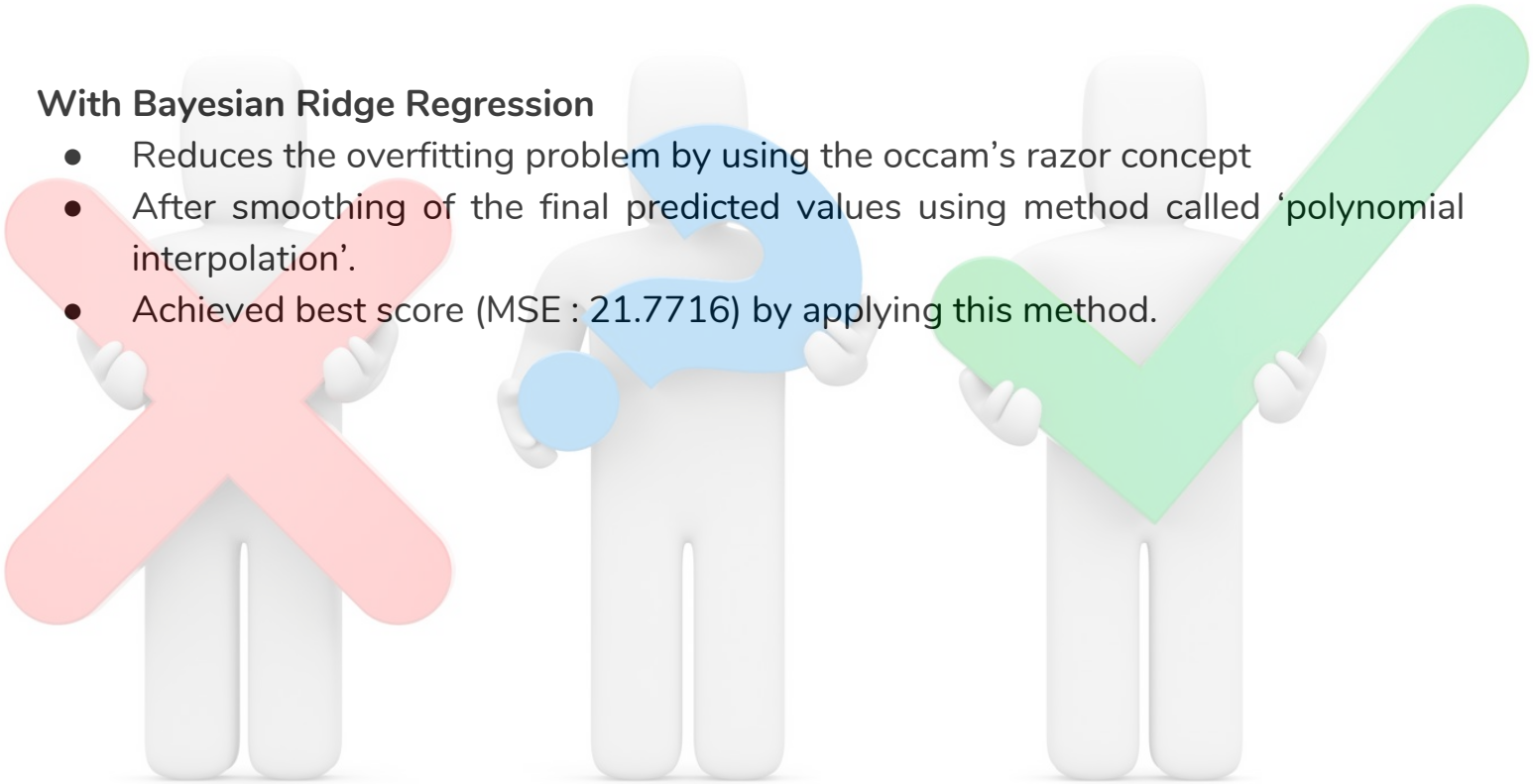
- Treated attributes as time series variation of data.
- uses the seasonal effect to predict the future values.
- MSE was around 26.6683.



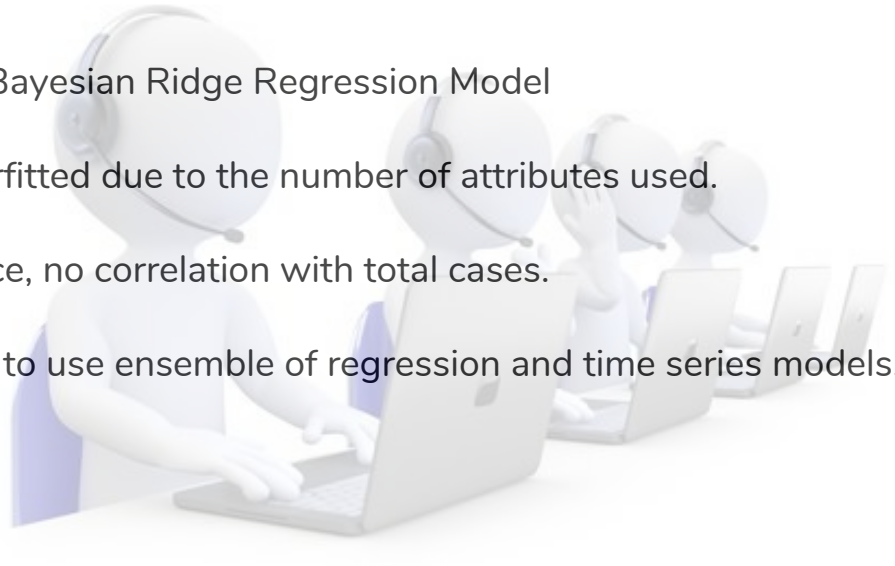
Results & Analysis Ctd..

With Bayesian Ridge Regression

- Reduces the overfitting problem by using the occam's razor concept
- After smoothing of the final predicted values using method called 'polynomial interpolation'.
- Achieved best score (MSE : 21.7716) by applying this method.



Conclusion



Winning performance from Bayesian Ridge Regression Model

Linear regression model overfitted due to the number of attributes used.

Did not use all attributes since, no correlation with total cases.

Future Improvement : Hope to use ensemble of regression and time series models.

Thank you!