In-depth Analysis- Machine Learning

Problem

Astronomers need a metric to predict eclipses so that they can plan accordingly to gather data. The goal is to identify supervised learning techniques to predict the date of future total solar and lunar eclipses.

Hypothesis

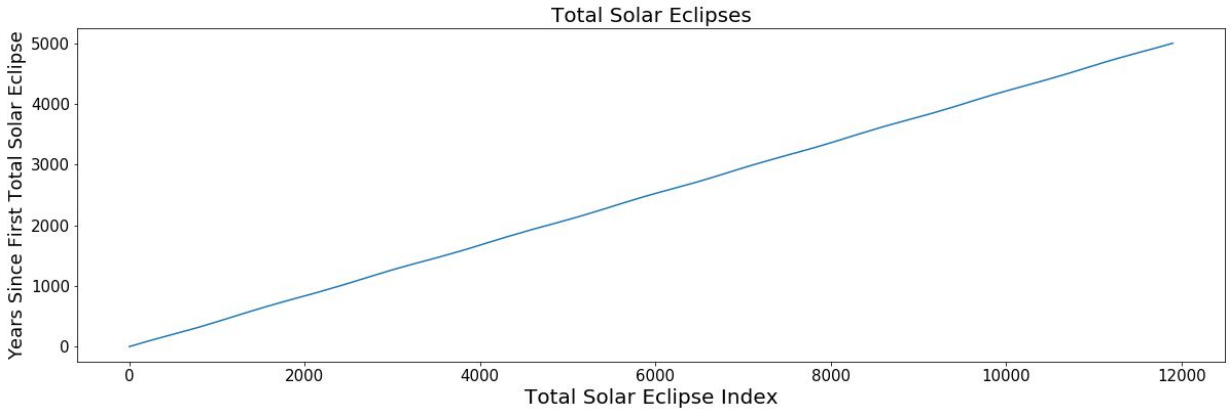
It is possible to predict the time and date of the solar and lunar eclipses from the given data leveraging machine learning regression algorithms.

Preparing Data for Analysis

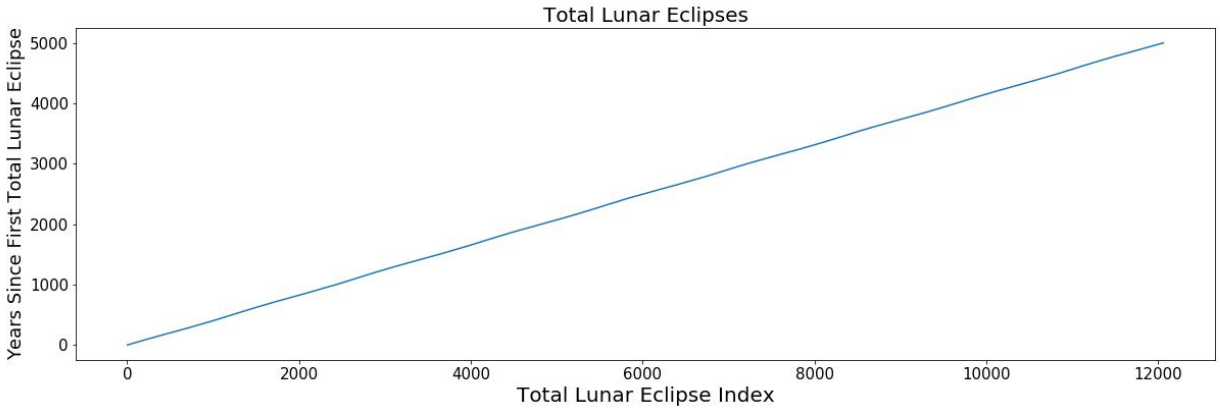
Since the statistical analysis focused on total eclipses, it was decided to do the same for the machine learning aspect of this project. The first step was to convert the total eclipse data into data that can be fit by the regression models. The feature X was defined as the eclipse index and the prediction variable y was the years from the start date of the data, which was 2000 BCE or -1999 in the csv file.

Actual Plots Before Fitting - For reference

Total Solar Eclipses



Total Lunar Eclipses



Models For Analysis

Since the data seemed to be very linear, I used a linear regression model to fit the data.

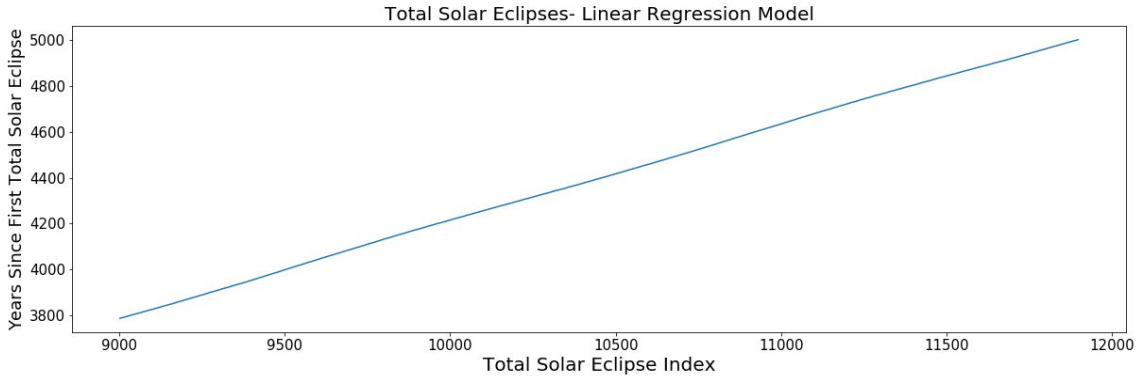
Linear Regression

Steps

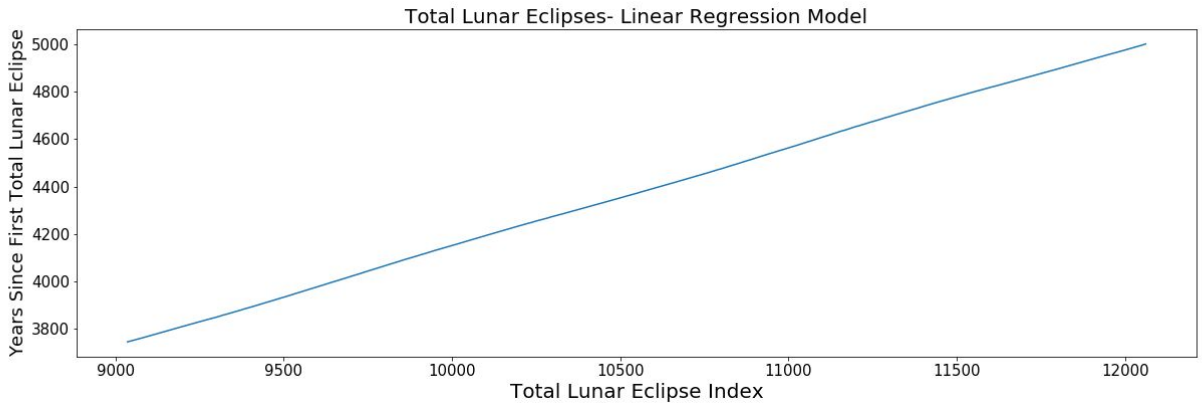
1. Created a function that converts the data into the appropriate linear regression format. X = total eclipse index , y = years from start date of dataset
2. Split the data in order into train and test sets. Train = 0.75 Test = 0.25
    a. It was done in order to keep the data in order of date.
3. Fitted the training data to linear regression model.
4. Predicted the model on testing data and computed root mean square error

Results

Total Solar Eclipses RMSE = 4.8 yrs



Total Lunar Eclipses RMSE = 5.4 yrs



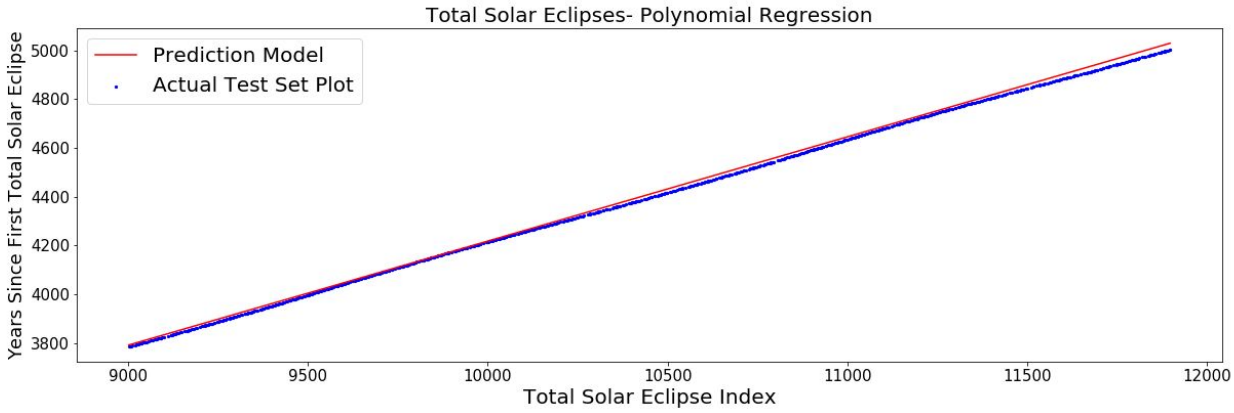
Polynomial Regression

Steps

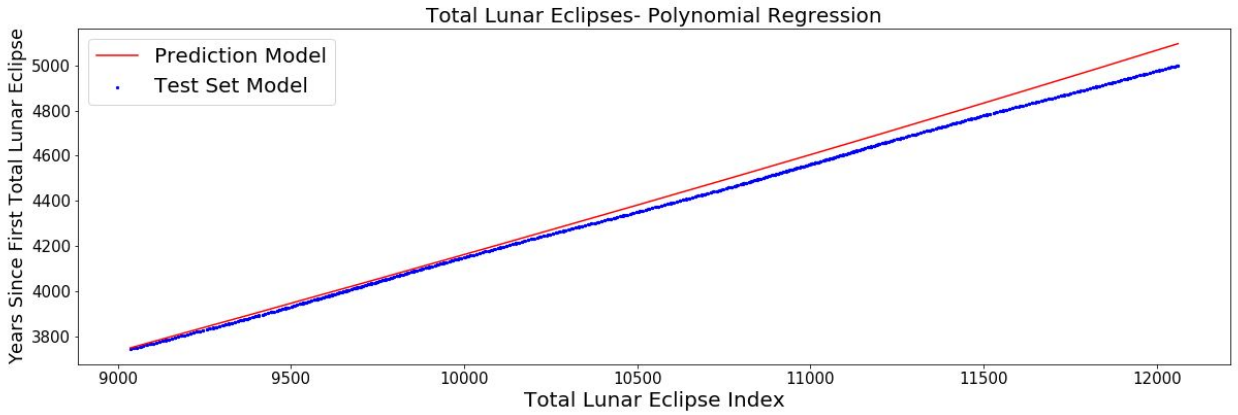
1. The data was already converted into the correct format and split into train and test sets in previous steps.
2. Performed a manual gridsearch to test for the best degree parameter for the Polynomial Features. Best degree = 4
3. Fitted the training data to polynomial regression model.
4. Predicted the model on testing data and computed root mean square error.

Results

Total Solar Eclipses RMSE = 4.6 yrs



Total Lunar Eclipses RMSE = 4.4 yrs



Ridge Regression

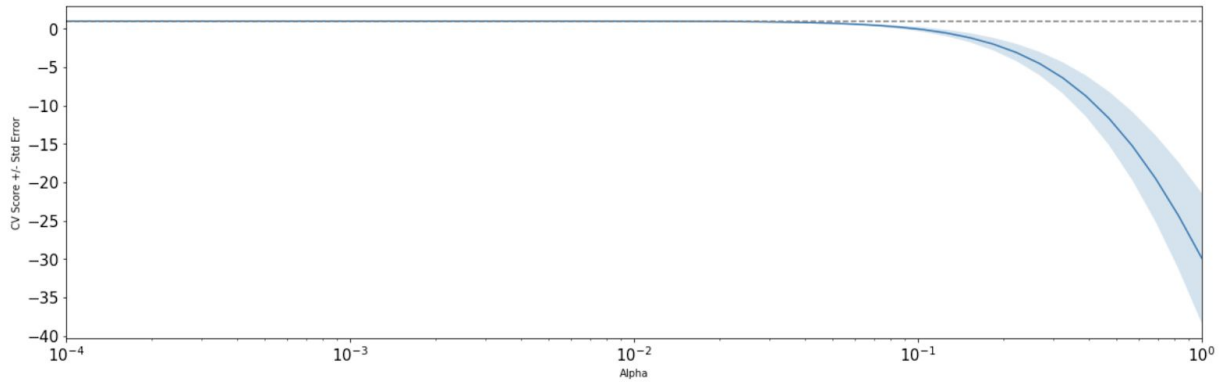
Steps

1. The data was already converted into the correct format and split into train and test sets in previous steps.
2. Performed GridSearchCV to find the best alpha parameter for the model and cross validation of 5.
3. Plotted the alphas vs the cross validation of 10 scores for each alpha for visual analysis.
4. Fitted the training data to ridge regression model.
5. Predicted the model on testing data and computed root mean square error.

Results

Ridge regression is just linear regression with a normalization term to help reduce overfitting. For both solar and lunar eclipses models, it was observed that the best alpha parameter seemed to be the lowest alpha. It implies that the best model is one where you use as little regularization as possible.

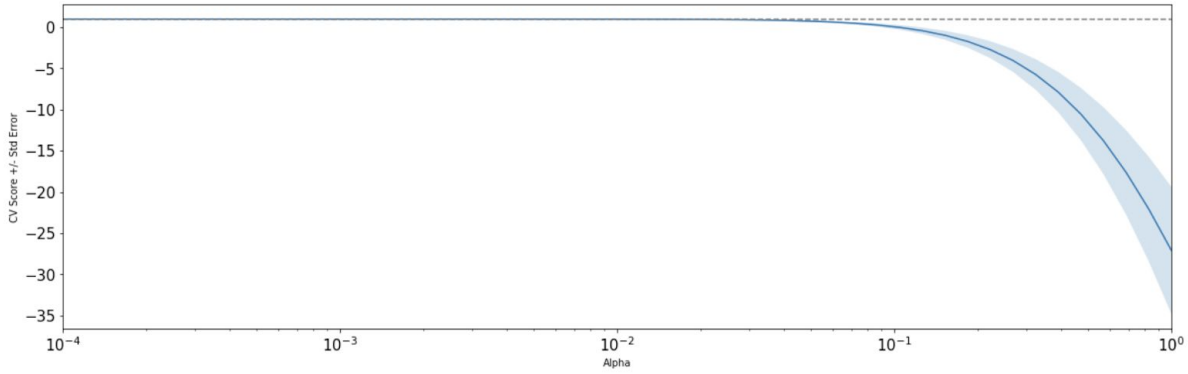
Total Solar Eclipses RMSE = 4.9



Grid search and RMSE results

```
[0.99958553 0.99970982 0.9995001  0.99979139 0.99958706]
Root Mean Squared Error: 4.942123596833923
```

Total Lunar Eclipses RMSE = 5.3



Grid search and RMSE results

```
[0.99938511 0.99968353 0.99971629 0.99971051 0.99981713]
Root Mean Squared Error: 5.304404730656201
```

Lasso Regression

Steps

1. The data was already converted into the correct format and split into train and test sets in previous steps.
2. Performed GridSearchCV to find the best alpha parameter for the model and cross validation of 5.
3. Plotted the alphas vs the cross validation of 10 scores for each alpha for visual analysis.
4. Fitted the training data to lasso regression model.
5. Predicted the model on testing data and computed root mean square error.

Results

Lasso regression had a similar result as ridge regression with the best alpha parameter being the smallest alpha, or no normalization.

Total Solar Eclipses RMSE = 4.8 yrs

Total Lunar Eclipses RMSE = 5.3 yrs

Bayesian Regression

Steps

1. The data was already converted into the correct format and split into train and test sets in previous steps.
2. Performed GridSearchCV to find the best alpha and lambda parameters for the model with cross validation of 5.
3. Fitted the training data to bayesian regression model.
4. Predicted the model on testing data and computed root mean square error.

Results

Total Solar Eclipses RMSE = 4.8

Total Lunar Eclipses RMSE = 5.4

Conclusion

The results of our analysis using different Machine Learning models are tabulated below. This shows that our data is highly linear and hence polynomial regression performs poorly. The other regressors are all linear predictors - Ridge and Lasso regressions are just Linear Regression with L1 and L2 norm penalty terms and hence the hyperparameter for these learned from grid search result in convergence of these two models to linear regression. We see that Bayesian Regression also does not show an increase in performance, despite our 4-hyperparameter grid search. Hence, the best modelling technique given this data is probably linear regression - we conclude that further investigation into the data must be performed as the current data is not suitable for modelling.

Regressor	Solar Eclipse RMSE	Lunar Eclipse RMSE
Linear Regression	4.820 yrs	5.456 yrs
Polynomial Regression	4.621 yrs	4.417 yrs
Ridge Regression	4.942 yrs	5.304 yrs
Lasso Regression	4.826 yrs	5.281 yrs
Bayesian Regression	4.820 yrs	5.456 yrs