



SAIT

SUPERVISED LEARNING - REGRESSION TIME SERIES

DATA SCIENCE

Daily Female Birth Dataset

Written by:
Shubh Desai

July 13, 2024

Supervised Learning- Time Series Use Case with Daily Female Birth Prediction

Problem

- **What problem are you solving?**
Predicting the number of female births on a given day.
- **Why is it worth solving?**
Accurate predictions can help in planning and resource allocation in healthcare and related sectors.
- **What is the source of your data and what kind of data are you using?**
The dataset used is the "Daily Female Births" containing the number of daily female births.

Data Collection

This dataset offers details on a range of daily female birth:
Variables:

Date: various date format: (YYYY-MM-DD)

Births: Number of births



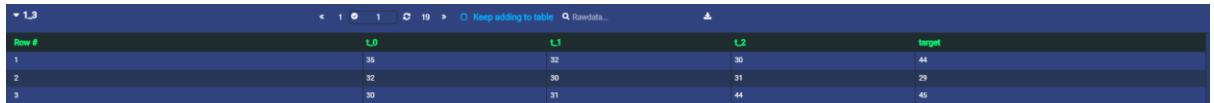
The screenshot displays a data table with three columns: 'Row #', 'Date', and 'Births'. The table contains 20 rows of data, representing the first 20 days of January 1959. The 'Births' column shows the number of female births for each day. The table is presented in a dark blue theme with white text. At the top, there is a navigation bar with a search icon and a 'Keep adding to table' button. At the bottom, a status bar indicates 'Currently loaded pages: 1'.

| Row # | Date | Births |
|-------|------------|--------|
| 1 | 1959-01-01 | 35 |
| 2 | 1959-01-02 | 32 |
| 3 | 1959-01-03 | 30 |
| 4 | 1959-01-04 | 31 |
| 5 | 1959-01-05 | 44 |
| 6 | 1959-01-06 | 29 |
| 7 | 1959-01-07 | 45 |
| 8 | 1959-01-08 | 43 |
| 9 | 1959-01-09 | 38 |
| 10 | 1959-01-10 | 27 |
| 11 | 1959-01-11 | 38 |
| 12 | 1959-01-12 | 33 |
| 13 | 1959-01-13 | 55 |
| 14 | 1959-01-14 | 47 |
| 15 | 1959-01-15 | 45 |
| 16 | 1959-01-16 | 37 |
| 17 | 1959-01-17 | 50 |
| 18 | 1959-01-18 | 43 |
| 19 | 1959-01-19 | 41 |
| 20 | 1959-01-20 | 52 |

Methodology

Data Wrangling (Time Lag & Window):

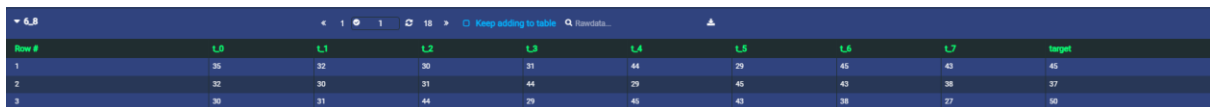
- 1 Time Lag (Forecast Time) & 3 Window Size (Historical Value):



A screenshot of a data table interface. The table has 5 columns: 'Row #', 'L0', 'L1', 'L2', and 'target'. The 'L0' column contains values 35, 32, and 30 for rows 1, 2, and 3 respectively. The 'L1' column contains values 32, 30, and 31. The 'L2' column contains values 30, 31, and 44. The 'target' column contains values 44, 29, and 45. The interface includes a search bar and a 'Keep adding to table' button.

| Row # | L0 | L1 | L2 | target |
|-------|----|----|----|--------|
| 1 | 35 | 32 | 30 | 44 |
| 2 | 32 | 30 | 31 | 29 |
| 3 | 30 | 31 | 44 | 45 |

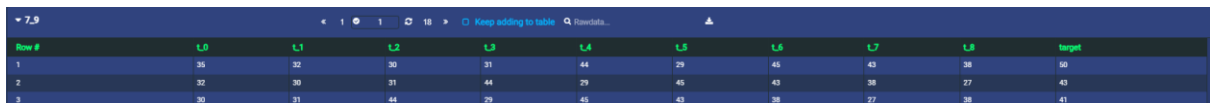
- 6 Time Lag (Forecast Time) & 8 Window Size (Historical Value):



A screenshot of a data table interface. The table has 10 columns: 'Row #', 'L0', 'L1', 'L2', 'L3', 'L4', 'L5', 'L6', 'L7', and 'target'. The 'L0' column contains values 35, 32, and 30. The 'L1' column contains values 32, 30, and 31. The 'L2' column contains values 30, 31, and 44. The 'L3' column contains values 31, 44, and 29. The 'L4' column contains values 44, 29, and 45. The 'L5' column contains values 29, 45, and 43. The 'L6' column contains values 45, 43, and 38. The 'L7' column contains values 43, 38, and 27. The 'target' column contains values 45, 37, and 50. The interface includes a search bar and a 'Keep adding to table' button.

| Row # | L0 | L1 | L2 | L3 | L4 | L5 | L6 | L7 | target |
|-------|----|----|----|----|----|----|----|----|--------|
| 1 | 35 | 32 | 30 | 31 | 44 | 29 | 45 | 43 | 45 |
| 2 | 32 | 30 | 31 | 44 | 29 | 45 | 43 | 38 | 37 |
| 3 | 30 | 31 | 44 | 29 | 45 | 43 | 38 | 27 | 50 |

- 7 Time Lag (Forecast Time) & 9 Window Size (Historical Value):



A screenshot of a data table interface. The table has 11 columns: 'Row #', 'L0', 'L1', 'L2', 'L3', 'L4', 'L5', 'L6', 'L7', 'L8', and 'target'. The 'L0' column contains values 35, 32, and 30. The 'L1' column contains values 32, 30, and 31. The 'L2' column contains values 30, 31, and 44. The 'L3' column contains values 31, 44, and 29. The 'L4' column contains values 44, 29, and 45. The 'L5' column contains values 29, 45, and 43. The 'L6' column contains values 45, 43, and 38. The 'L7' column contains values 43, 38, and 27. The 'L8' column contains values 38, 27, and 38. The 'target' column contains values 50, 43, and 41. The interface includes a search bar and a 'Keep adding to table' button.

| Row # | L0 | L1 | L2 | L3 | L4 | L5 | L6 | L7 | L8 | target |
|-------|----|----|----|----|----|----|----|----|----|--------|
| 1 | 35 | 32 | 30 | 31 | 44 | 29 | 45 | 43 | 38 | 50 |
| 2 | 32 | 30 | 31 | 44 | 29 | 45 | 43 | 38 | 27 | 43 |
| 3 | 30 | 31 | 44 | 29 | 45 | 43 | 38 | 27 | 38 | 41 |

Univariate Analysis

- Analysed the date column (e.g., months, days) to identify patterns or trends.
- Evaluated data distribution, missing values, and outliers.

Data Preparation

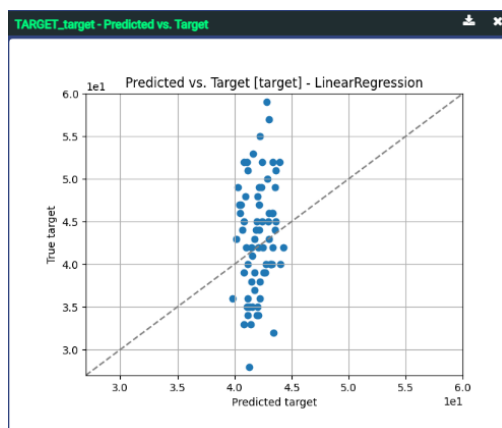
- Extracted relevant features from the date column, such as day of the week, month, and any significant dates.
- Handled missing values and outliers appropriately.
- Normalized/standardized the data if necessary.

Time Lag & Window:

1 Time Lag (Forecast Time) & 3 Window Size (Historical Value):

| ▼ Model Container @ 1_3 (13) | | | | | | | | |
|------------------------------------|---------|------------------|------|-----------------------------|------|-------------------------------------|--------|--|
| Select Dataset | | | | Select Regression Algorithm | | | | |
| 1_3@1_3 | | | | LinearRegression | | | | |
| Q 13 Model Versions... | | | | | | Auto Pilot Create New Model Version | | |
| Version-Tag | Dataset | Algorithm | Rank | Error | Doc. | Publish | Delete | |
| <input type="checkbox"/> v.1-v.b3c | 1_3-1_3 | LinearRegression | 1 | 5.32 | | | | |
| <input type="checkbox"/> v.8-v.a0b | 1_3-1_3 | LassoRegressor | 1 | 5.32 | | | | |
| <input type="checkbox"/> v.3-v.a7d | 1_3-1_3 | ElasticNet | 1 | 5.32 | | | | |

Target vs Predicted



Evaluation Metrics

Used metrics such as Mean Squared Log Error, Median Absolute Error, Mean Absolute Error (MAE), and Mean Squared Error (MSE).

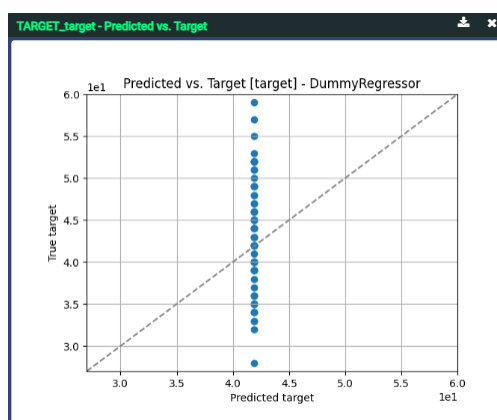
| Regression | Mean Squared Log Error | MAE | MSE | Median Absolute Error |
|------------------|------------------------|------|------|-----------------------|
| LinearRegression | 0.02 | 5.32 | 6.45 | 4.75 |
| LassoRegressor | 0.02 | 5.31 | 6.46 | 4.79 |
| ElasticNet | 0.02 | 5.31 | 6.46 | 4.77 |

This window size 1 and 3 shows the top three algorithms with the lowest error rate, and the target / predicted graph refers to the dots that are close to the median line. Each algorithm establishes evaluation metrics mean and median absolute, squared log, and error.

6 Time Lag (Forecast Time) & 8 Window Size (Historical Value):

| Model Container @ 6_8 (13) | | | | | | | |
|-------------------------------------|---------|---------------------|-----------------------------|-------|------|------------|--------------------------|
| Select Dataset | | | Select Regression Algorithm | | | | |
| 6_8@6_8 | | | LinearRegression | | | | |
| Q 13 Model Versions... | | | | | | | |
| | | | | | | Auto Pilot | Create New Model Version |
| Version-Tag | Dataset | Algorithm | Rank | Error | Doc. | Publish | Delete |
| <input type="checkbox"/> v.3-v.917 | 6_8-6_8 | DummyRegressor | 1 | 5.38 | | | |
| <input type="checkbox"/> v.11-v.9e1 | 6_8-6_8 | SVMRegressor | 2 | 5.45 | | | |
| <input type="checkbox"/> v.7-v.966 | 6_8-6_8 | KNeighborsRegressor | 3 | 5.92 | | | |

Target vs Predicted



Evaluation Metrics

Used metrics such as Mean Squared Log Error, Median Absolute Error, Mean Absolute Error (MAE), and Mean Squared Error (MSE).

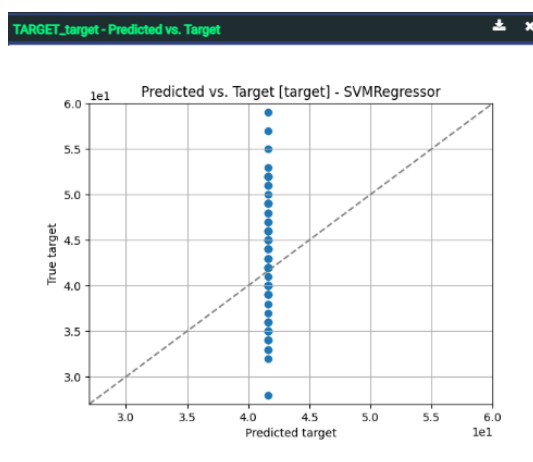
| Regression | Mean Squared Log Error | MAE | MSE | Median Absolute Error |
|----------------|------------------------|------|------|-----------------------|
| DummyRegressor | 0.02 | 5.38 | 6.64 | 4.91 |
| SVMRegressor | 0.02 | 5.44 | 6.70 | 4.59 |
| KNeighbors | 0.02 | 5.92 | 7.11 | 5 |

This window size 6 and 8 shows the top three algorithms with the lowest error rate, and the target / predicted graph refers to the dots that are close to the median line. Each algorithm establishes evaluation metrics mean and median absolute, squared log, and error.

7 Time Lag (Forecast Time) & 9 Window Size (Historical Value):

| ▼ Model Container @ 7_9 (12) | | | | | | | |
|-------------------------------------|-----------|----------------|--------|-----------------------------|------|----------------------------|--------|
| Select Dataset | | | | Select Regression Algorithm | | | |
| 7_9@7_9 | | | | LinearRegression | | | |
| Q 12 Model Versions... | | | | 🚀 Auto Pilot | | 🔧 Create New Model Version | |
| Version-Tag ⬆ | Dataset ⬆ | Algorithm ⬆ | Rank ⬆ | Error ⬆ | Doc. | Publish | Delete |
| <input type="checkbox"/> v.4-v.bc9 | 7_9-7_9 | DummyRegressor | 1 | 5.38 | | | |
| <input type="checkbox"/> v.12-v.a63 | 7_9-7_9 | SVMRegressor | 2 | 5.45 | | | |
| <input type="checkbox"/> v.5-v.96b | 7_9-7_9 | LassoRegressor | 3 | 5.94 | | | |

Target vs Predicted



Evaluation Metrics

Used metrics such as Mean Squared Log Error, Median Absolute Error, Mean Absolute Error (MAE), and Mean Squared Error (MSE).

| Regression | Mean Squared Log Error | MAE | MSE | Median Absolute Error |
|----------------|------------------------|------|------|-----------------------|
| DummyRegressor | 0.02 | 5.37 | 6.64 | 4.92 |
| SVMRegressor | 0.02 | 5.44 | 6.70 | 4.59 |
| LassoRegressor | 0.02 | 5.93 | 7.23 | 4.87 |

This window size 7 and 9 shows the top three algorithms with the lowest error rate, and the target / predicted graph refers to the dots that are close to the median line. Each algorithm establishes evaluation metrics mean and median absolute, squared log, and error.

Modelling Process

Algorithm Selection

Used regression algorithms like Linear Regression, Elastic Net, SVM Regressor, Dummy Regressor, Lasso Regressor, K Neighbors Regressor

Linear Regression:

Linear regression uses the relationship between the data-points to draw a straight line through all them. This line can be used to predict future values.

SVM Regressor:

Creates a Linear SVR object using the Vertica SVM (Support Vector Machine) algorithm. This algorithm finds the hyperplane used to approximate distribution of the data.

Elastic Net:

Elastic net linear regression uses the penalties from both the lasso and ridge techniques to regularize regression models

Dummy Regressor:

This regressor is useful as a simple baseline to compare with other (real) regressors. Do not use it for real problems

Lasso Regressor:

Lasso regression—also known as L1 regularization—is a form of regularization for linear regression models

K Neighbors Regressor:

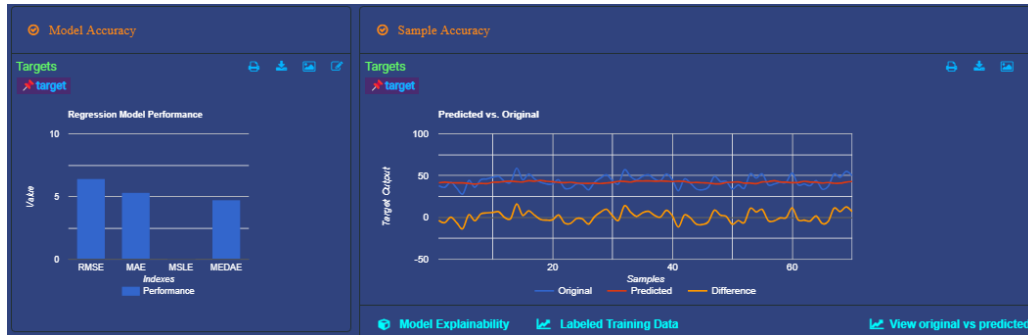
The target is predicted by local interpolation of the targets associated of the nearest neighbors in the training set

Best Result (Linear Regression)

Linear Regression is the best model in a window frame of one time lag (forecast time) and three window sizes (historical value)

The error rate is 5.32, and the graph's dots are close to the median

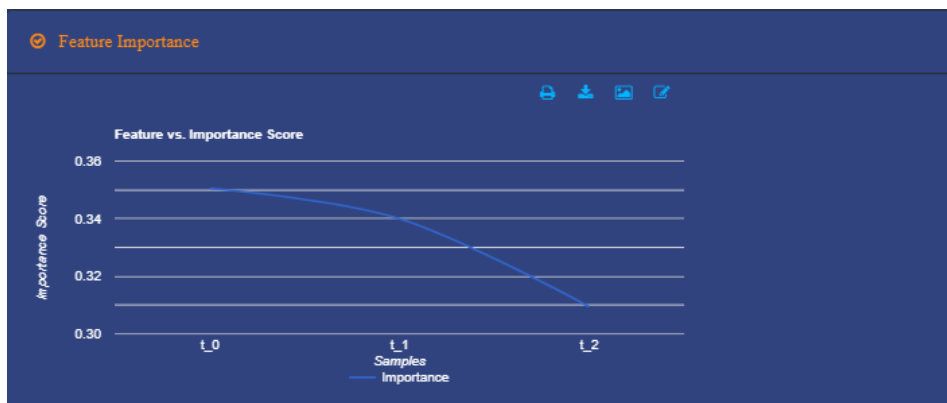
Model Performance



Performance Metrics



Feature Importance



Result Summary

Presented the results of the best-performing model Linear Regression with

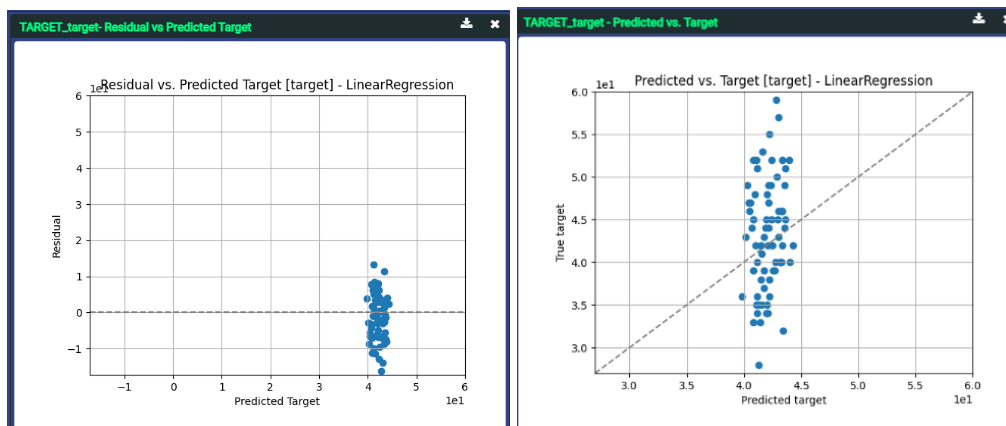
Mean Squared Log Error: 0.02162

MAE: 5.32119

MSE: 6.45583

Median Absolute Error: 4.75796

Target vs Predicted



Conclusions

1. Future Improvements

- Discussed potential improvements like incorporating more features, using more advanced models, or refining the current model.

2. Real-life Application

- Explained how this predictive model could be used in real-life scenarios, such as hospital staffing or resource planning.

3. Value to Client

- Highlighted the value this solution could bring to stakeholders, such as improved planning and resource allocation.

4. Learnings

- Summarized the key learnings and insights gained from the project.