ENGR 424 Project

Predicting Daily Mean Temperature Using Linear Regression Model

<u>Prediction Wizards</u>

1- Yousif Faris Robeil 2- Aivn Waleed

3- Raed Ahmed Khalaf

Contents

1 Project Description

1.1 Problem Definition

Temperature is a useful measure for daily life, as people usually check the weather to plan their daily activities, due to the variable nature of temperature with respect to time, the average temperature would represent an intuitive estimate of the temperature for people without need for domain knowledge.

1.2 Objective

The aim of the project is to build a machine learning model pipeline that achieves both simplicity and accurate estimate of the average temperature, also known as the mean temperature. An accurate model is a model that produces predictions compatible with the true values, while a simple machine learning model pipeline is achieved when the method doesn't require high computation power, doesn't require optimization, subtle sampling methods, and all steps are understandable and comprehended by observers with average technical knowledge.

1.3 Responsibilities Distribution

Yousif Faris Robeil:	Data Engineering
Raed Ahmed Khalaf:	Social Media Coverage
Aivn Waleed:	Version Control

2 Method

2.1 Data Collection and Description

A weather dataset collected in Argentina was imported from Kaggle. It is a time series data which, which suggests a stochastic behavior (process). A stochastic process is a phenomenon where observations of the phenomenon have an element of randomness that evolves overtime and this evolution can happen in different ways[1]. Accordingly Temporal order of the data is of significant importance to make proper predictions. Any disturbance to this temporal order will lead to flawed predictions and the predictors must not contain any future value in order to preserve the predictive nature of the machine learning model. The Data spanned 24 years of daily measurements and had the following features; Mean Temperature, Maximum Temperature, Minimum Temperature, Precipitation Sum, and Sunshine Duration, Figure 1 shows the plot of a snippet of mean temperature time series data.

2.2 Data Transformation

most of the data had normal distribution, some of these had skewed normal distribution. Hence their sqrt was taken to give an exact normal distribution,Figure 2 shows data distribution, note that the given data was not cleaned as it already showed the distributions patterns clearly, indicating good quality.

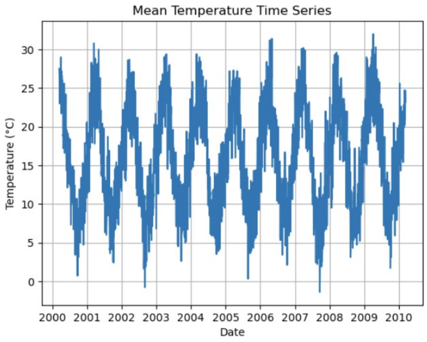


Figure 1: Time Series Plot of Mean Temperature

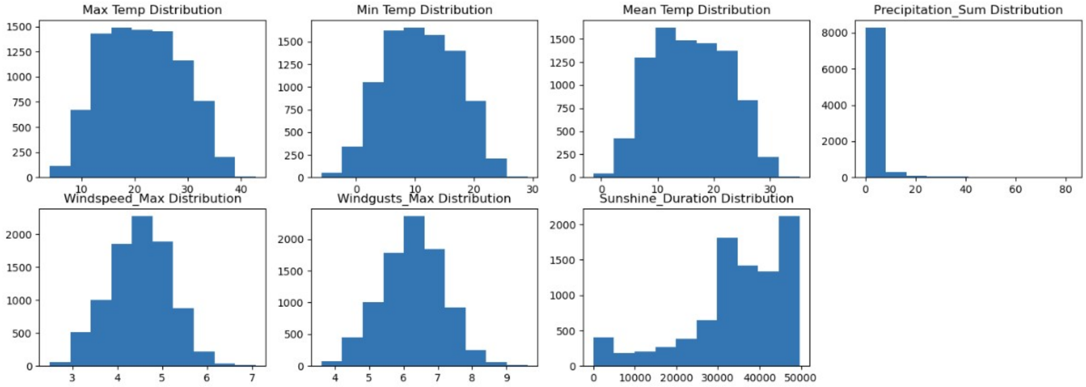


Figure 2: Data Distribution

2.3 Data Split

The first 70% of the data was taken (about 17 years) for training while 30% (about 7) for testing. No validation was used and parameters of the features were chosen by inference rather than tuning

2.4 Data Normalization

To avoid the problem of feature dominance, data were put to the same scale through min max scaler. Note that the scaler was fit to the training data and was used to transform both training and testing data to avoid data leakage, as fitting the scaler to testing data can make the future values distort the scale of the present ones.

2.5 Feature Extraction

2.5.1 Lag Features

The PACF (Partial Autocorrelation Function), shown in Figure 3, was used to choose the lag features, the chosen lag features were first, second, 4th, and 5th lags of mean temeperature. Moreover, the first lags of the maximum temperature and sunshine duration were considered instead of the present value, as the max temp is acquired at 3:00 PM of the day, while sunshine duration is acquired at the time of sunset, waiting to measure these values will be late in the day, while minimum temperature was not shifted as it is acquired early in the day, at about 6:00 AM, as shown in Figure 4. Accordingly the prediction is made early in the day once the minimum temperature is acquired, would be acquired at 7:00 AM. 11 hours earlier than the case of not shifting the sunshine duration, however the considered strategy would yield less accuracy, as the previous values have weaker correlation with the mean temperature compared to the present values.

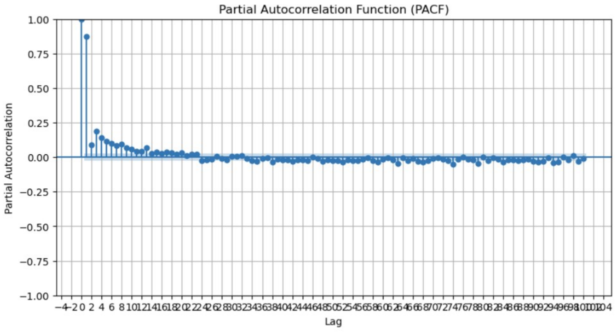


Figure 3: PACF of Training Data

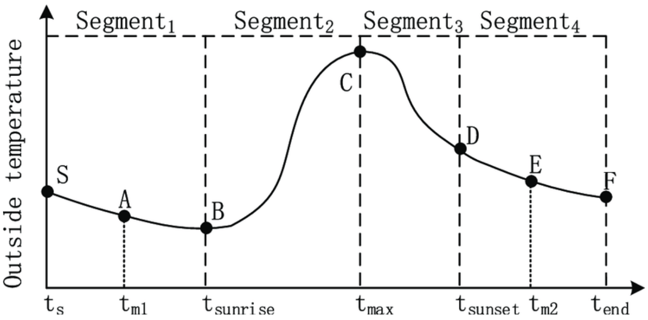


Figure 4: Minimum during sunrise at 6:00 AM

2.5.2 Moving Averages

The ACF (Autocorrelation Function), shown in Figure 5 was used to determine the moving window for the simple moving average (SMA) And half of it is the half life for the exponential moving average (EMA), Figure 6 shown how SMA and EMA follow the general pattern of the mean temperature.

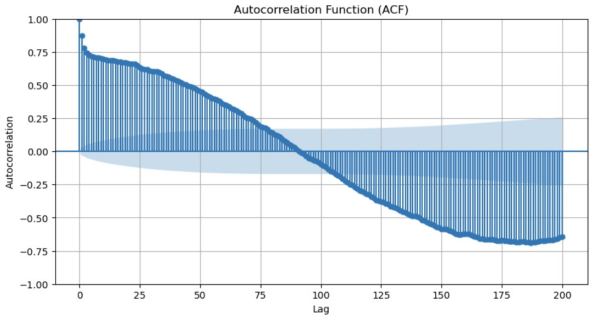


Figure 5: ACF of Training Data

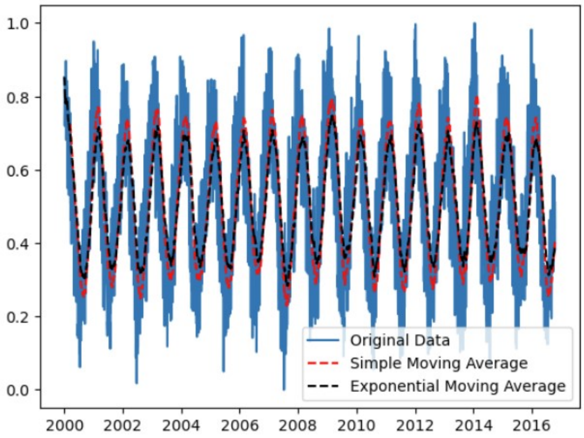


Figure 6: Moving Averages and Daily Mean Temperature

2.6 Feature Selection

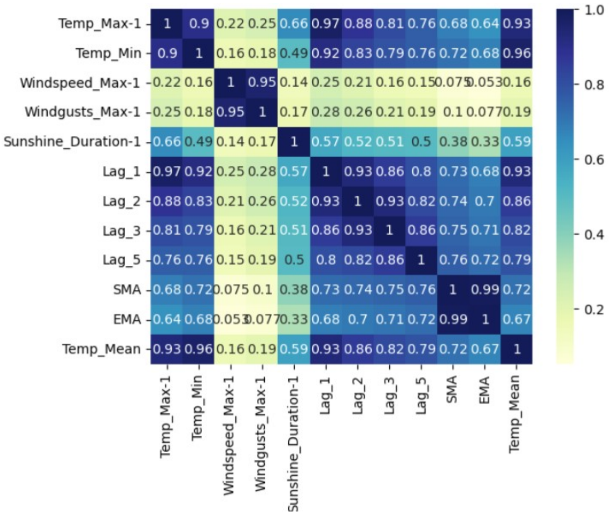
Features were collected based on the correlation matrix in Figure 7, where any feature that had less than 0.5 correlation with respect to the mean temperature was dropped

2.7 Model Selection

As one of the aims of the project is simplicity, a linear regression model is a good candidate for this application. Linear regression requires some assumptions to give good predictions[2], accordingly these assumptions were examined:

2.7.1 Linearity

Linear correlation between features and target variable is assured as the features were already selected based on linear correlation given by the correlation matrix in Figure 7.



	Temp_Max-1	Temp_Min	Windspeed_Max-1	Windgusts_Max-1	Sunshine_Duration-1	Lag_1	Lag_2	Lag_3	Lag_5	SMA	EMA	Temp_Mean
Temp_Max-1	1	0.9	0.22	0.25	0.66	0.97	0.88	0.81	0.76	0.68	0.64	0.93
Temp_Min	0.9	1	0.16	0.18	0.49	0.92	0.83	0.79	0.76	0.72	0.68	0.96
Windspeed_Max-1	0.22	0.16	1	0.95	0.14	0.25	0.21	0.16	0.15	0.075	0.053	0.16
Windgusts_Max-1	0.25	0.18	0.95	1	0.17	0.28	0.26	0.21	0.19	0.1	0.077	0.19
Sunshine_Duration-1	0.66	0.49	0.14	0.17	1	0.57	0.52	0.51	0.5	0.38	0.33	0.59
Lag_1	0.97	0.92	0.25	0.28	0.57	1	0.93	0.86	0.8	0.73	0.68	0.93
Lag_2	0.88	0.83	0.21	0.26	0.52	0.93	1	0.93	0.82	0.74	0.7	0.86
Lag_3	0.81	0.79	0.16	0.21	0.51	0.86	0.93	1	0.86	0.75	0.71	0.82
Lag_5	0.76	0.76	0.15	0.19	0.5	0.8	0.82	0.86	1	0.76	0.72	0.79
SMA	0.68	0.72	0.075	0.1	0.38	0.73	0.74	0.75	0.76	1	0.99	0.72
EMA	0.64	0.68	0.053	0.077	0.33	0.68	0.7	0.71	0.72	0.99	1	0.67
Temp_Mean	0.93	0.96	0.16	0.19	0.59	0.93	0.86	0.82	0.79	0.72	0.67	1

Figure 7: Training Data Correlation Matrix

2.7.2 No Multicollinearity Among Features

This assumption is violated as can be seen in Figure 7.

2.7.3 Homoscedacity

residuals have constant variance, since no increasing or decreasing patterns were observed. Additionally, they fluctuate around the mean 0, which can be seen through residuals plot in Figure 8.

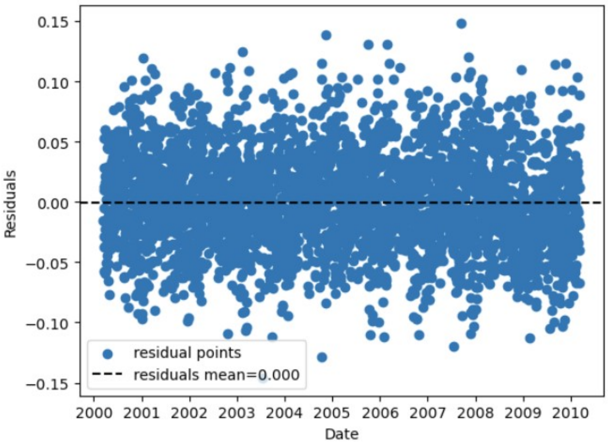


Figure 8: Residuals Plot

2.7.4 Normality

Residuals follow a non-peaked non-skewed normal distribution, as seen through the histogram and qq-plot in Figures 9 and 10.

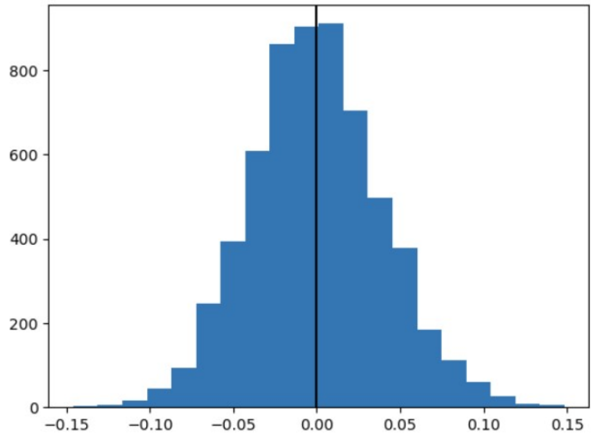


Figure 9: Residuals Distribution

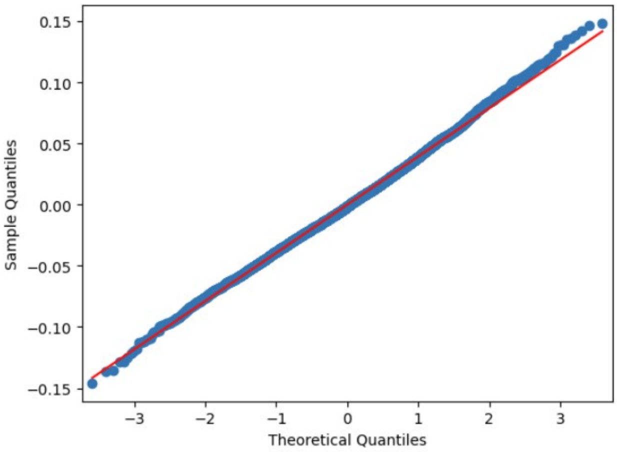


Figure 10: Residuals Q-Q Plot

2.7.5 Residuals are independent

Residuals have no autocorrelation, according to ACF plot found in Figure 11.

2.7.6 Residuals have no correlation with the features

No linear Correlation between residuals and the features were detected according in the correlation matrix in Figure 12.
In summary, all of the assumptions of linear regression were satisfied except multicolliniarity. On the other hand, the linear regression model will still not have an issue when it comes prediction accuracy as suggested by the literature[2].

2.8 Model Outcome Evaluation and analysis

The used Evaluation metrics for model output were the following, R^2 to assess the goodness of fit, MAPE to assess model accuracy, and the root mean squared

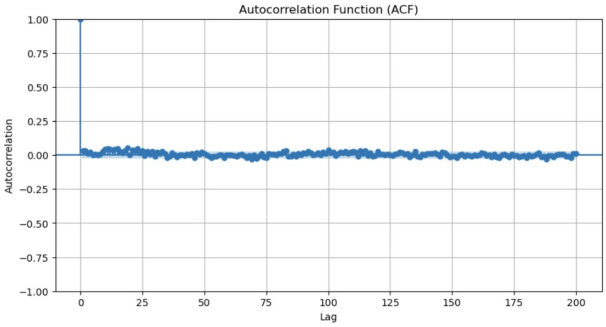
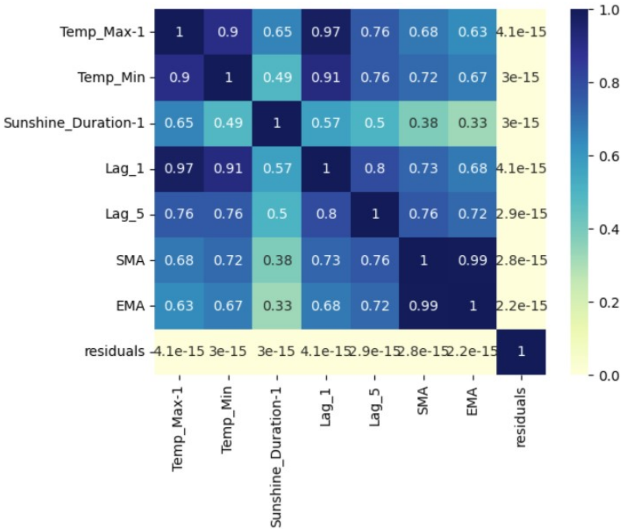


Figure 11: Residuals ACF



	Temp_Max-1	Temp_Min	Sunshine_Duration-1	Lag_1	Lag_5	SMA	EMA	residuals
Temp_Max-1	1	0.9	0.65	0.97	0.76	0.68	0.63	4.1e-15
Temp_Min	0.9	1	0.49	0.91	0.76	0.72	0.67	3e-15
Sunshine_Duration-1	0.65	0.49	1	0.57	0.5	0.38	0.33	3e-15
Lag_1	0.97	0.91	0.57	1	0.8	0.73	0.68	4.1e-15
Lag_5	0.76	0.76	0.5	0.8	1	0.76	0.72	2.9e-15
SMA	0.68	0.72	0.38	0.73	0.76	1	0.99	2.8e-15
EMA	0.63	0.67	0.33	0.68	0.72	0.99	1	2.2e-15
residuals	4.1e-15	3e-15	3e-15	4.1e-15	2.9e-15	2.8e-15	2.2e-15	1

Figure 12: Residuals and Features Correlation Matrix

error's percentage ratio to the mean known as coefficient of variation to assess the predictions variability around the true values. The metrics for training and testing data respectively were of values; 0.958 and 0.960 R^2, 7.78% and 7.57% coefficient of variation, and 6.16% and 5.89% MAPE. It can be seen that the model had very close performance in both training and testing data, indicating good generalization, R^2 was close to 1, indicating good fit, and with low coefficient of variation and MAPE, indicating good precision and accuracy. A Prediction interval, with 90% confidence level was generated using gradient boosting regressor (GBR). Figure 13 shows how the prediction curve, the black one, lies inside the interval most of the time, indicating reasonable predictions. Note that the odd outcome of having the model to perform better on testing data than training data doesn't consistently hold when different snippets or split of training data and testing data were considered to evaluate the model or retrain it. Indicating that it is an outcome of luck, changing the considered samples to hide this outcome was avoided as it would be a biased approach. As the method followed the proper practices to avoid leakage and preserving the temporal order of the data. One conclusion can be drawn about this specific data from this outcome is that the mean temperature of the region of interest in Argentina is tending to follow linear relationships with the used predictive features better as future goes, as the problem of interest has stochastic nature. There is no guarantee that the tendency towards order will keep holding as time goes given that stochastic processes can take variable paths to change their current random behavior.

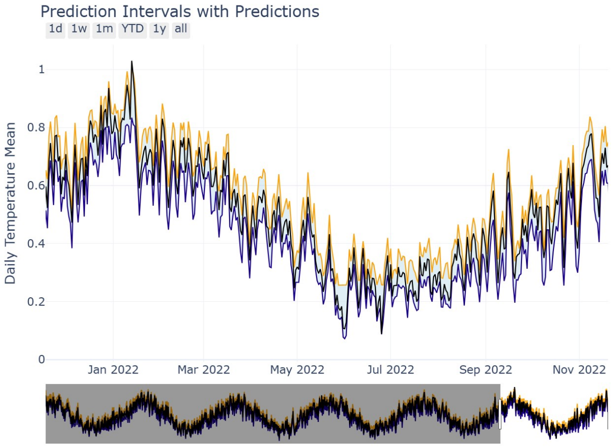


Figure 13: Prediction Interval of 2022 predictions

3 Conclusion

The Linear Regression Model performed accurately and can be used as a reliable source to make reasonable daily prediction of the average temperature, achieved at 7:00 AM everyday. Moreover, the simplicity condition was achieved as the model didn't require high computation power, nor it involved optimization, since all hyperparameters where inferred rather than tuned, the straight forward 70/30% train/test rule of thumb was followed, and all procedures are expected to be understandable and comprehended by observers with average technical knowledge.

4 References

[1] Peter Kempthorne et al, (2013), ”S096 Lecture 5 Stochastic Processes I”, .
Available: https://ocw.mit.edu/courses/18-s096-topics-in-mathematics-with-applications-in-f
f5784e4facf3de690210d17c97358eba_MIT18_S096F13_lecnote5.pdf

[2] D. J. Mundfrom, M. D. Smith, and L. W. Kay, “The effect of multicollinearity on prediction in regression models,” General Linear Model Journal, vol. 44, no. 1, pp. 24–28, 2018. Available: https://www.glmj.org/archives/articles/Mundfrom_v44n1.pdf

Appendix

Our Git Hub Repository: https://github.com/ra22213/ML-Project-Temawork.git