Midterm Q2 - Kaggle Competition

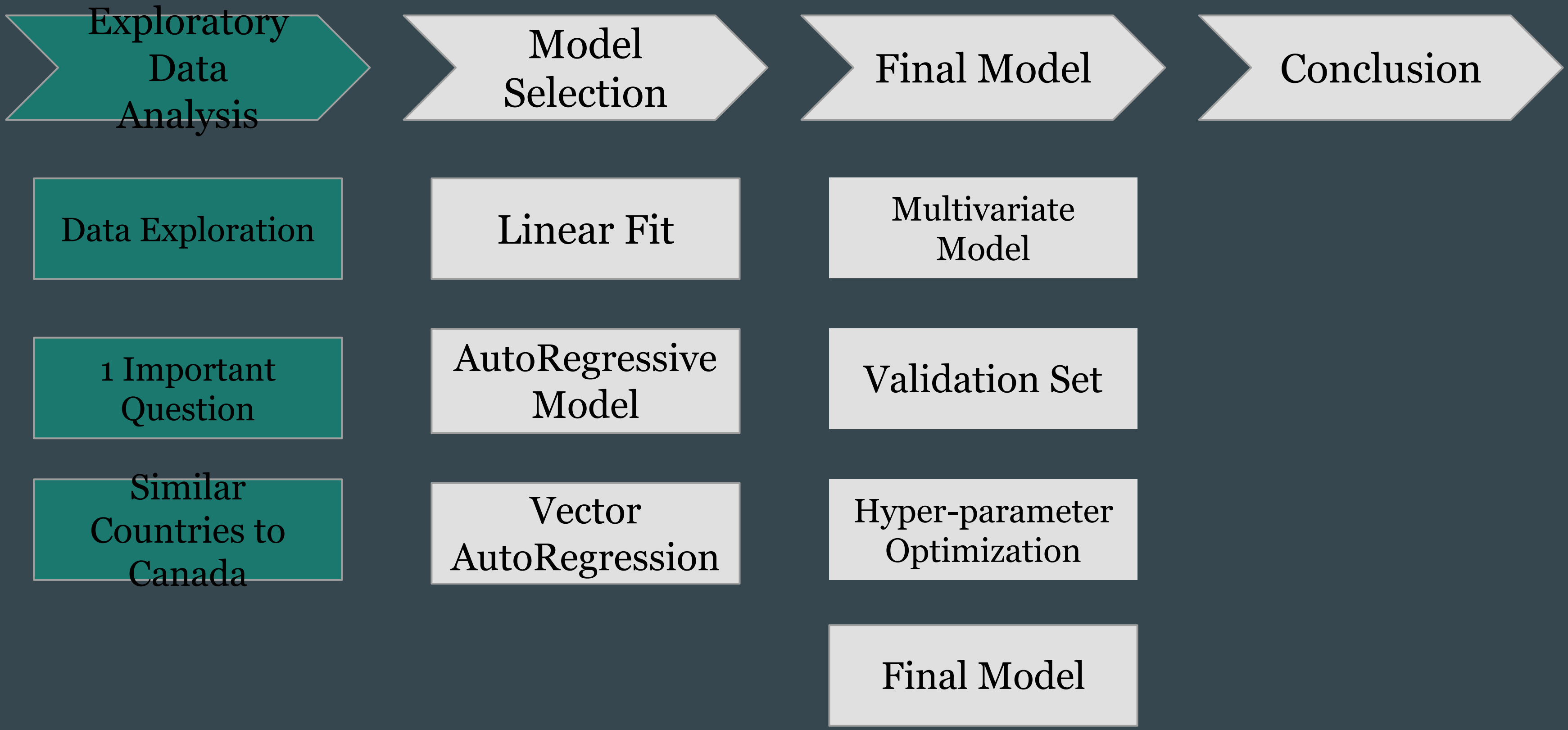
•••

CPSC-340
Machine Learning and Data Mining

Ali Seyfi - Mohamad Amin Mohamadi - Farnoosh Hashemi

Agenda

- Exploratory Data Analysis
  - Data Exploration
  - 1 important question
  - Similar countries to Canada
- Model Selection
  - Linear Fit
  - AutoRegressive Model
  - Vector AutoRegression
- Final Model
  - Multivariate Model
  - Validation Set
  - Hyper-parameter Optimization
  - Final Model
- Conclusion



Exploratory Data Analysis

Data Exploration

1 Important Question

Similar Countries to Canada

Model Selection

Linear Fit

AutoRegressive Model

Vector AutoRegression

Final Model

Multivariate Model

Validation Set

Hyper-parameter Optimization

Final Model

Conclusion

Exploratory Data Analysis

Data Exploration

1 Important Question

Similar Countries to Canada

Data Exploration

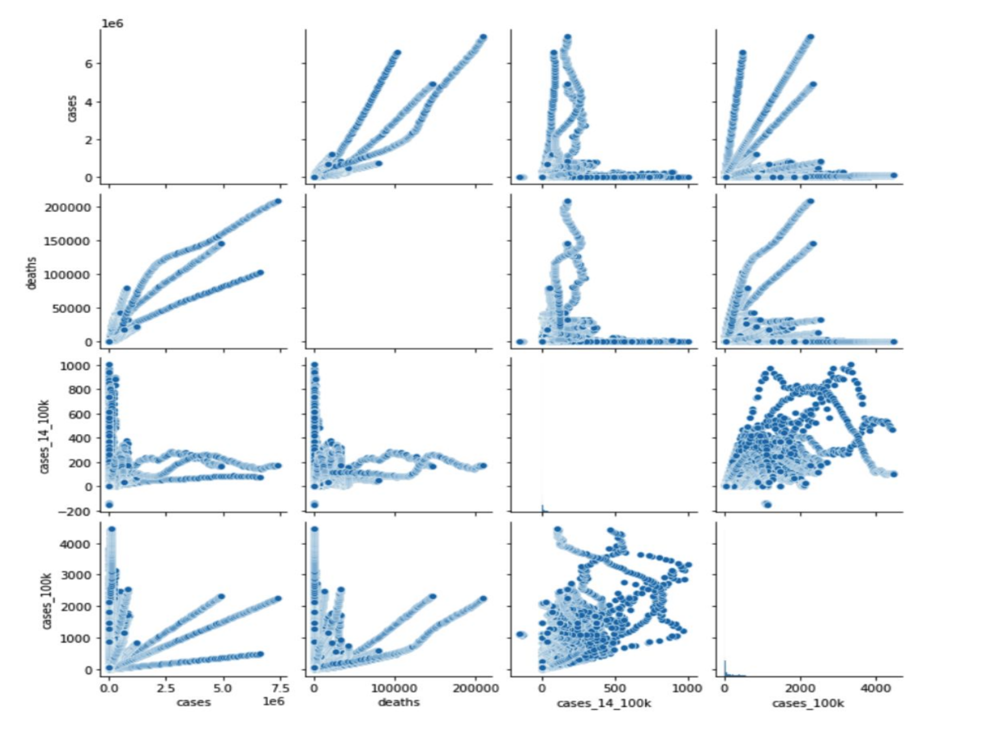
- Negative values and outliers
- A missing country
- Cumulative features
- Slow spread of the pandemic among countries at the beginning
- Feature correlations among countries

Data Exploration

1 Important Question

Similar Countries to Canada

Data Exploration



Data Exploration

1 Important Question

Similar Countries to Canada

1 Important Question

Should we consider all countries?

- The more data, the better model.
- There are many countries with different trends from Canada.

Data Exploration

1 Important Question

Similar Countries to Canada

Similar Countries to Canada

- Using death per 100k correlation >0.9 and p<0.5:

  CM, DE, FI, HU, IE, JE, LT, LV, ML

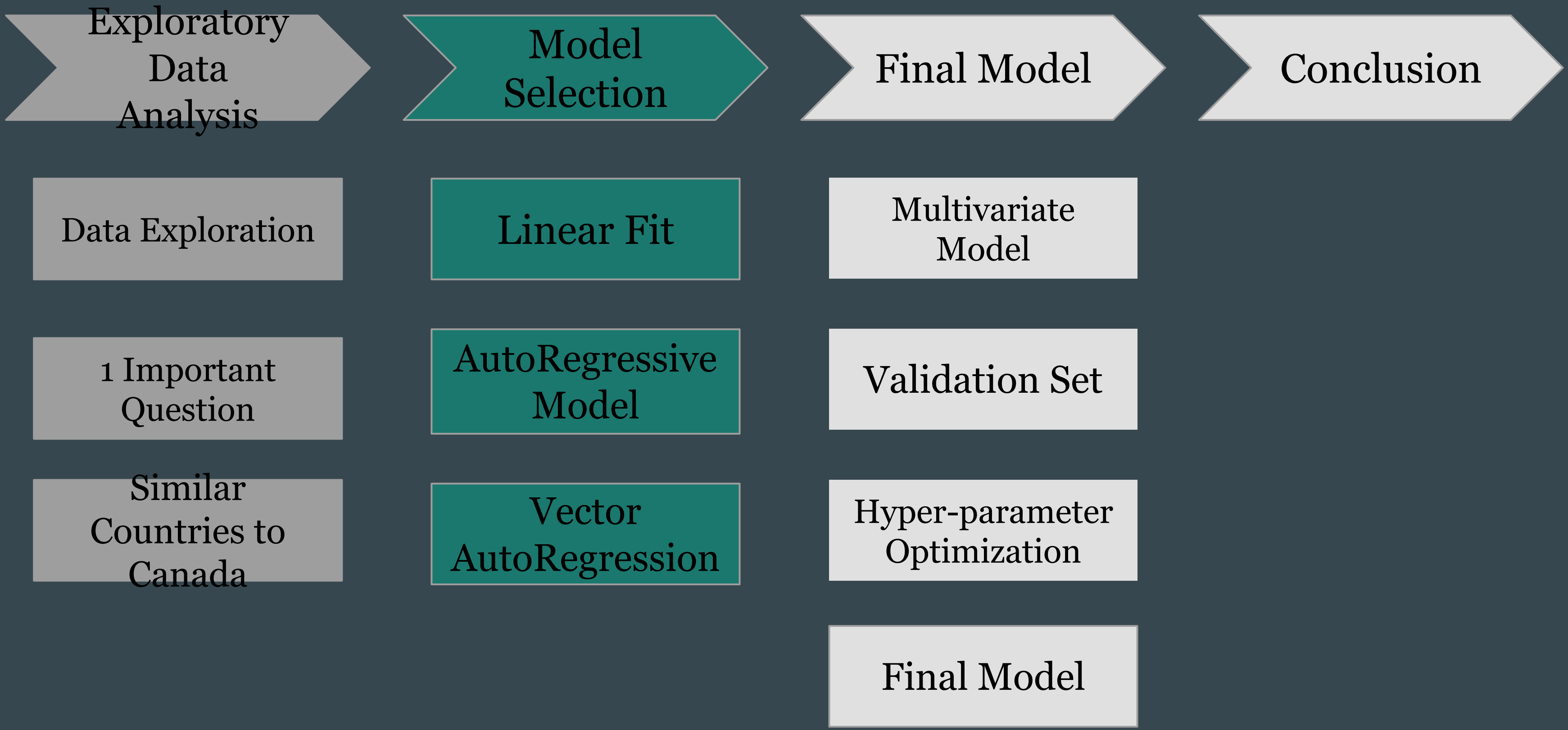
- Using KDTree:

  DE

Data Exploration

1 Important Question

Similar Countries to Canada



Exploratory Data Analysis

- Data Exploration
- 1 Important Question
- Similar Countries to Canada

Model Selection

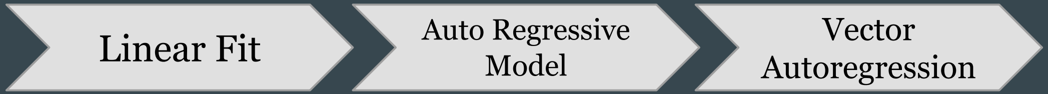
- Linear Fit
- AutoRegressive Model
- Vector AutoRegression

Final Model

- Multivariate Model
- Validation Set
- Hyper-parameter Optimization
- Final Model

Conclusion

Model Selection: Exploration Phase

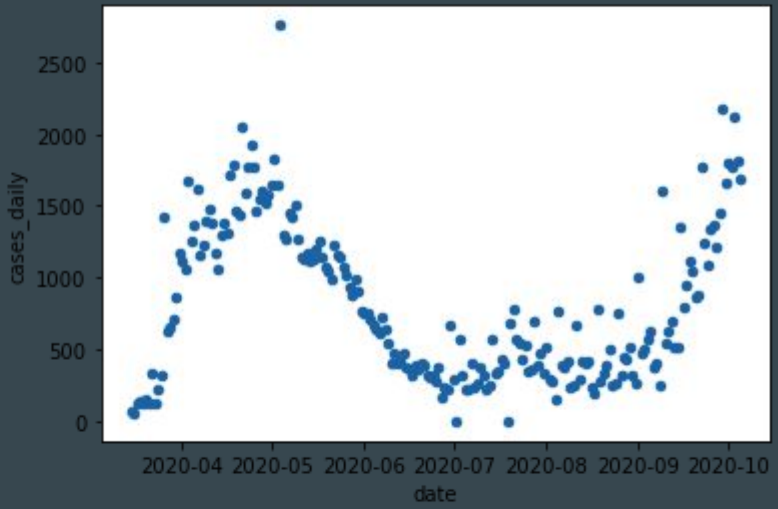


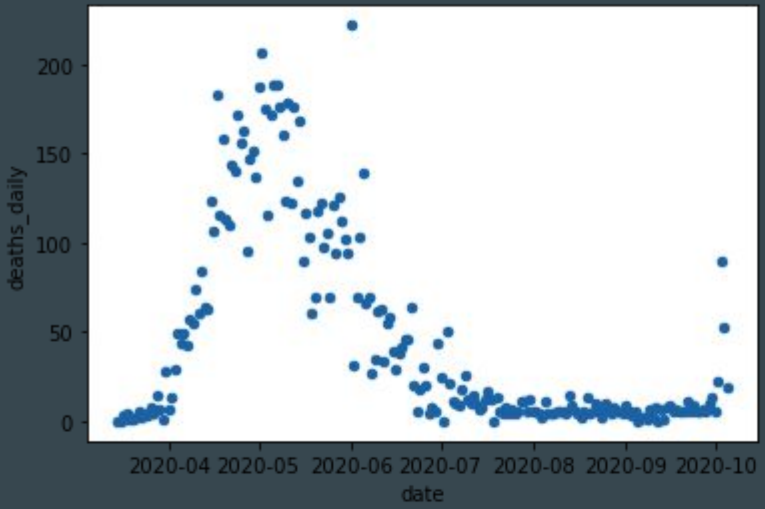
Linear Fit

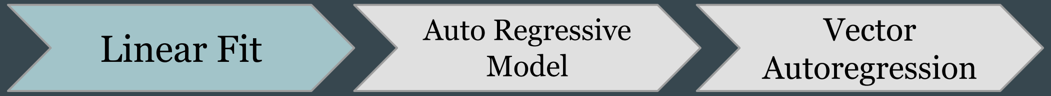
Auto Regressive Model

Vector Autoregression

Linear Fit: Feature Extraction





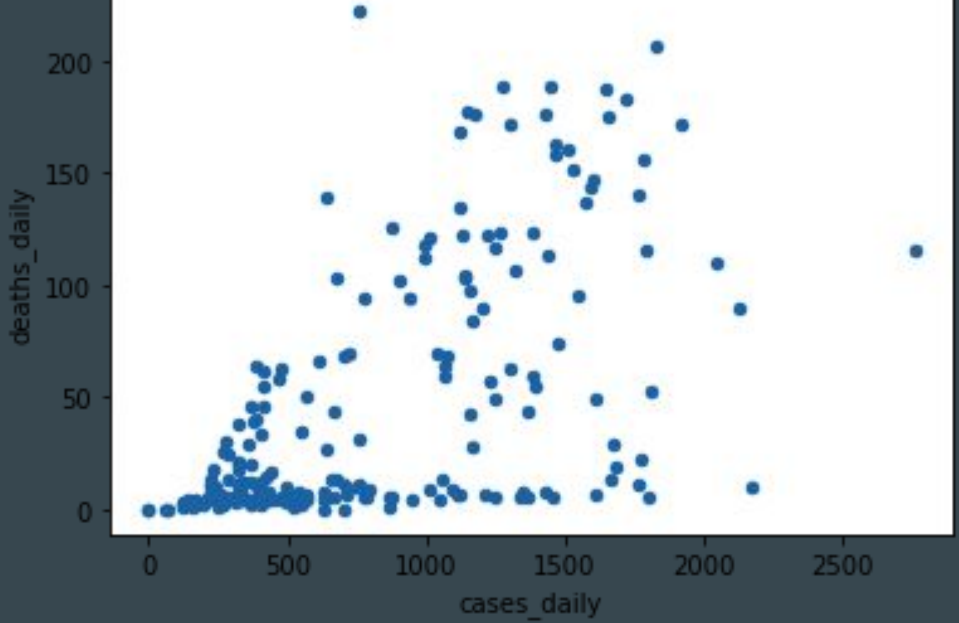


Linear Fit

Auto Regressive Model

Vector Autoregression

Linear Fit: Daily statistics



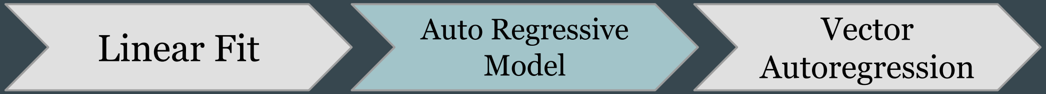
Linear Fit

Auto Regressive Model

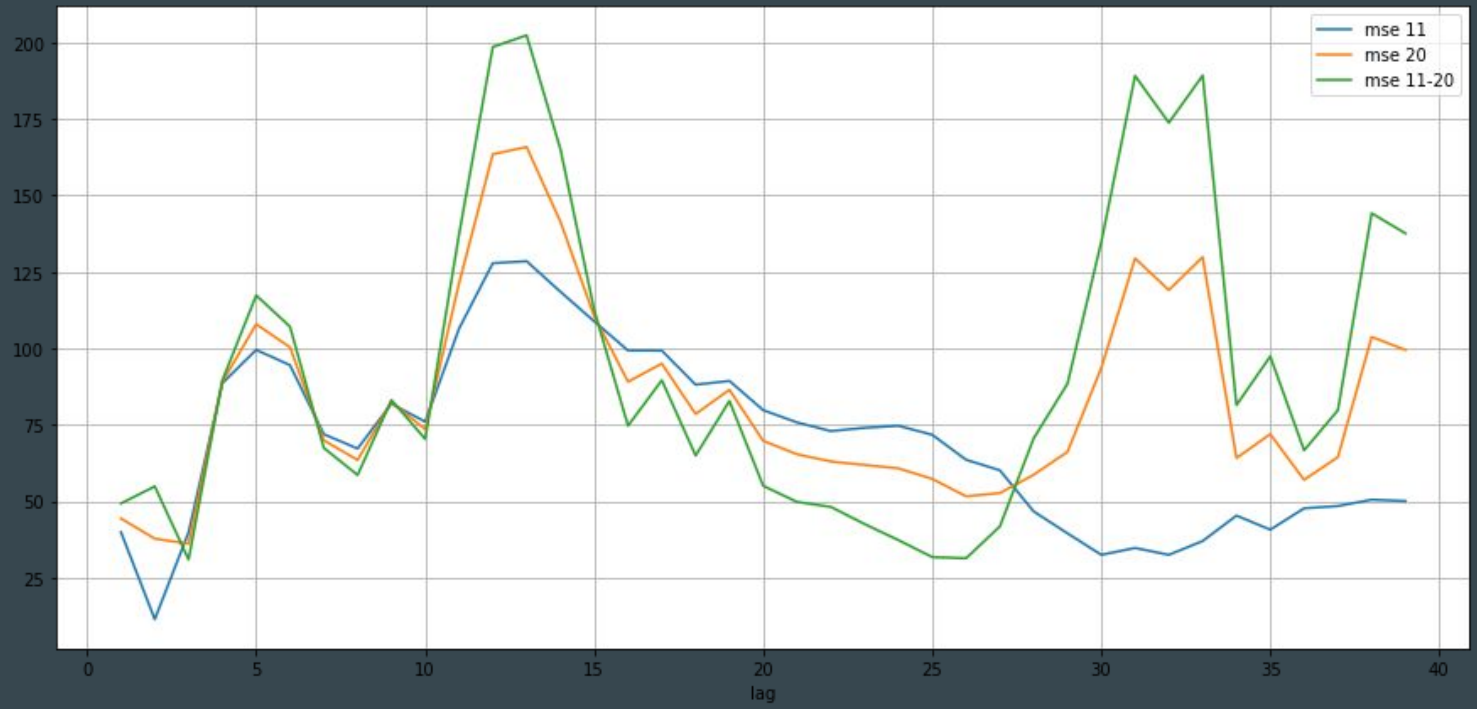
Vector Autoregression

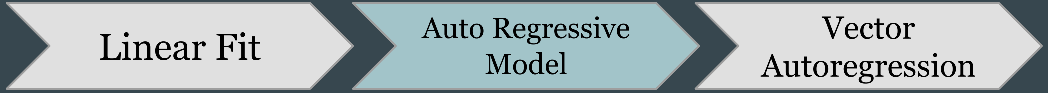
AutoRegressive Model: Introduction

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t,$$

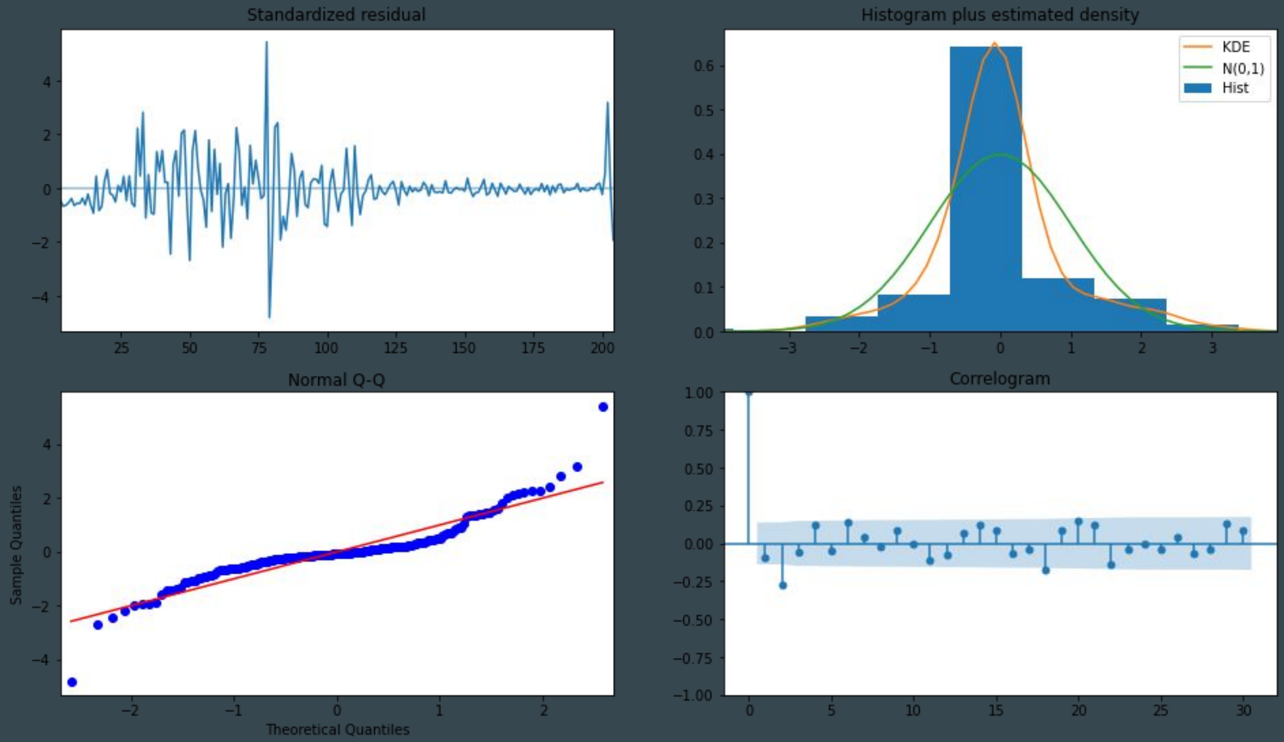


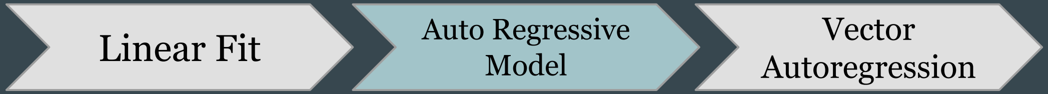
AutoRegressive Model: Lags





AutoRegressive Model: Results



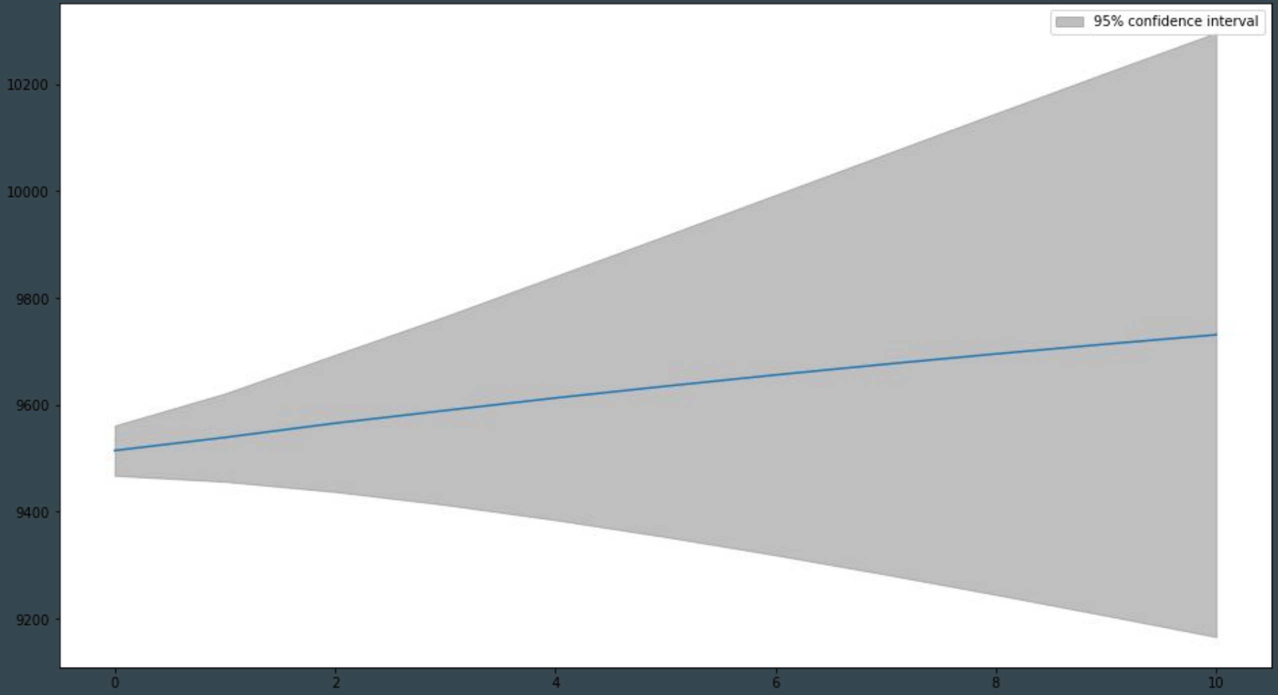


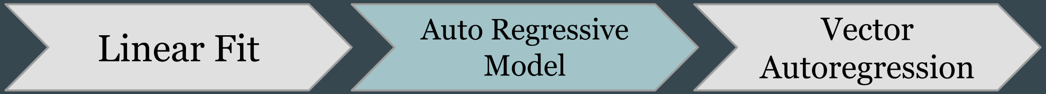
Linear Fit

Auto Regressive Model

Vector Autoregression

AutoRegressive Model: Results



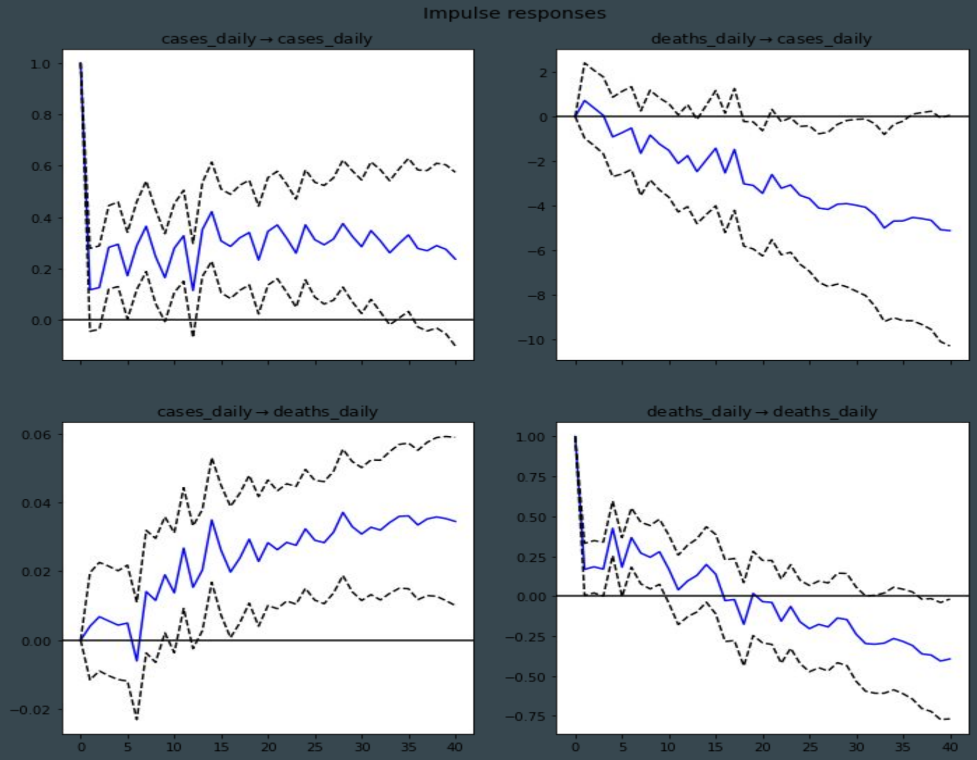


Vector AutoRegression: Introduction

$$
\begin{aligned}
Y_t &= \nu + A_1 Y_{t-1} + \ldots + A_p Y_{t-p} + u_t \\
u_t &\sim \text{Normal}(0, \Sigma_u)
\end{aligned}
$$

Linear Fit
Auto Regressive Model
Vector Autoregression

Vector AutoRegression: Impulse Responses

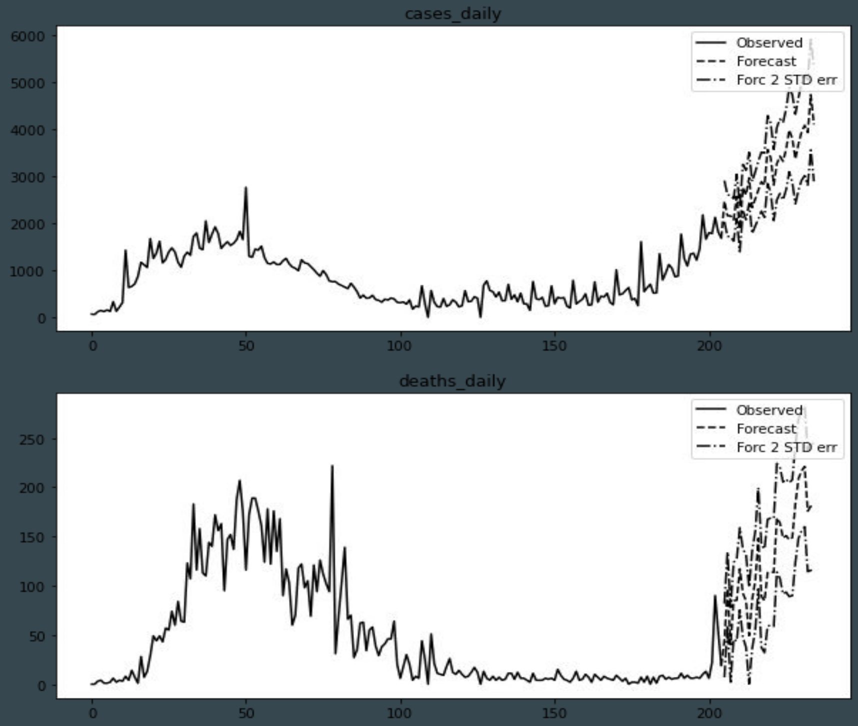


Linear Fit

Auto Regressive Model

Vector Autoregression

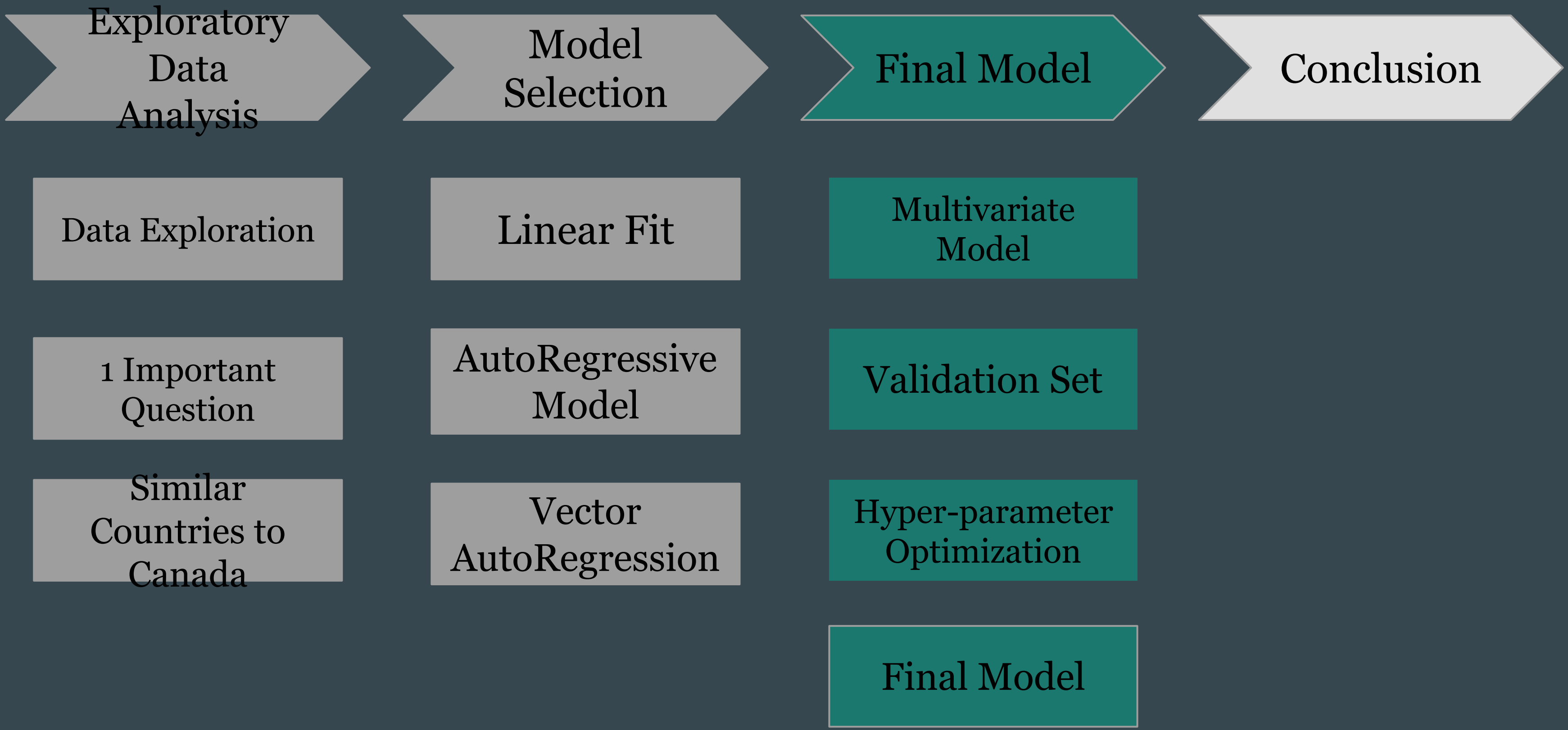
Vector AutoRegression: Predictions



Linear Fit

Auto Regressive Model

Vector Autoregression



Exploratory Data Analysis

Data Exploration

1 Important Question

Similar Countries to Canada

Model Selection

Linear Fit

AutoRegressive Model

Vector AutoRegression

Final Model

Multivariate Model

Validation Set

Hyper-parameter Optimization

Final Model

Conclusion

Final Model

Multivariate Model

Validation Set

Hyper-paramete r Optimization

Final Model

Moving towards our Multivariate AutoRegression Model

- From the nature of the problem, given the data of cases, we can predict the deaths better
- Important Notice! We only change the matrix X. [The vector y is the same as naive AutoRegression Model]
- $$y = \begin{bmatrix} d_K \\ d_{K+1} \\ \vdots \\ d_T \end{bmatrix} \qquad X = \begin{bmatrix} x_K^T \\ x_{K+1}^T \\ \vdots \\ x_T^T \end{bmatrix} = \begin{bmatrix} 1 & d_1 & d_2 & \cdots & d_{K-1} & c_i & c_j & \cdots & c_k \\ 1 & d_2 & d_3 & \cdots & d_K & c_j & c_m & \cdots & c_l \\ \vdots & \cdots & \cdots & \cdots & \vdots & \vdots & \cdots & \cdots & \vdots \\ 1 & d_{T-K} & d_{T-K+1} & \cdots & d_{T-1} & c_n & c_o & \cdots & c_p \end{bmatrix}.$$
- ds are the 'death' time series, and cs are 'cases' with selected lags

Choosing the Validation Set

- In order to decrease the error, we do not predict the cases.
- For the validation set, last 10 or 5 days? [5, because of the following reasons]
    - Average case confirmation to death time for people is 14 days, for +60 y.o. People this average is 11.5 starting from 6, and for young people is 19 starting from 14.
    - Canada does not face any health care shortage, we guess that most death cases are the old people.
    - So 6 might be a good number for case lags.
    - Validation set must be smaller than 6 in size.
    - So 5 is the biggest number smaller than 6.

Multivariate Model

Validation Set

Hyper-parameter Optimization

Final Model

Hyper-parameter Optimization

- Training data : until October 20th
- Validation data : October 21-25th
- Hyper-parameters:
  - The beginning index of the data
  - The lag of deaths used in feature space
  - The lag of cases used in the feature space
  - The number of consecutive cases used starting from the lag of cases values onward
- Lowest validation error achieved : 1.446

Multivariate Model

Validation Set

Hyper-paramete r Optimization

Final Model

Final Model

- In order to get better results, use median value of predictions of the best 200 models.
- Why median? [To avoid outliers in prediction of some models]
- Results [in compare with real results]:

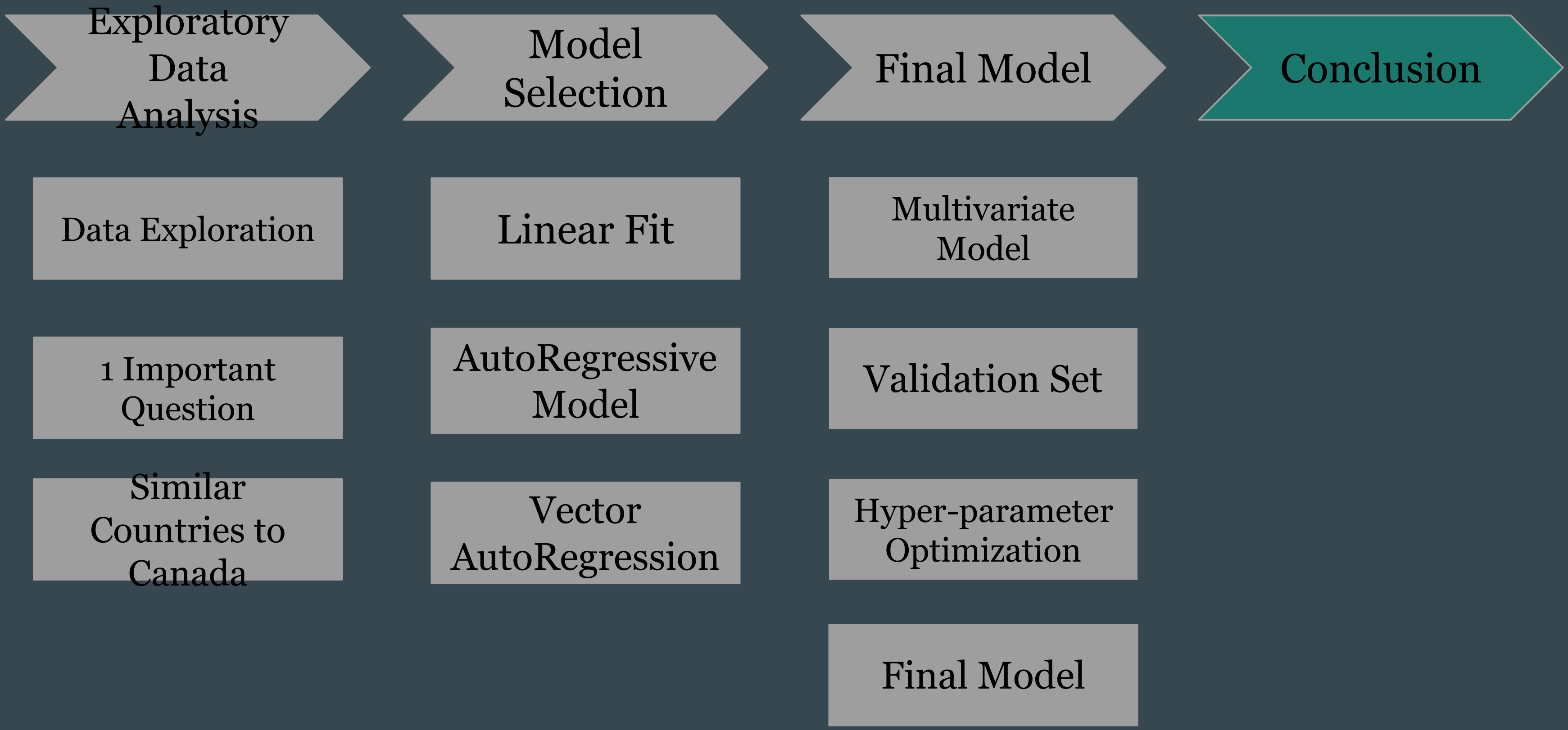
	1	2	3	4	5
Predicted	9947.368	9974.076	10003.066	10035.501	10063.897
Real	9946	9973	10001	10032	10074

Multivariate Model

Validation Set

Hyper-parameter Optimization

Final Model



Exploratory Data Analysis

Data Exploration

1 Important Question

Similar Countries to Canada

Model Selection

Linear Fit

AutoRegressive Model

Vector AutoRegression

Final Model

Multivariate Model

Validation Set

Hyper-parameter Optimization

Final Model

Conclusion

Conclusion

- Linear Regression models can perform a good job but only for near future.
- Feature selection is the most important step.
- Look at the nature of the problem! [Cases with lags are better than same day]
- Never give up!
  - [We found the best features and parameters of our model in the last hour of the competition]

Thank you for your attention!

•••