**Goal:** Predict influenza case counts 4 weeks in advance.

**Input:** Weekly number of laboratory-confirmed influenza detections in Canada

**Output:** Prediction of weekly number of laboratory-confirmed influenza detections for 4 weeks into the future

**Evaluation Metrics:** MAE, MSE, RSME

<https://towardsdatascience.com/time-series-forecast-error-metrics-you-should-know-cc88b8c67f27>

Scale dependent means error metrics are expressed in the units of the underlying data

Diagram, timeline

Description automatically generated

**Presentation:** Tableau showing current model predictions over the next 4 weeks, with 95% confidence intervals, RMSE

**Dataset:**

* Weekly positive tests for influenza in Canada (broken down by subtype) from 2015 to present (375 records). Data were provided by 32 sentinel laboratories (Newfoundland (1), Prince Edward Island (1), Nova Scotia (1), New Brunswick (1), Quebec (6), Ontario (17), Manitoba (1), Saskatchewan (2), Alberta (1), and British Columbia (1)
  + Changes in laboratory testing practices may affect the comparability of data to previous weeks or previous seasons.
  + Reach out to someone working in public health to see if these 32 laboratories have remained the same over the last 7 years
  + Double check that flu testing continued throughout COVID

**Data for Features:**

* Monthly score from 2015 to present on google searches in Canada for various search terms:
  + Numbers represent search interest relative to the highest point on the chart for the given region and time. A value of 100 is the peak popularity for the term. A value of 50 means that the term is half as popular. A score of 0 means that there was not enough data for this term.
* Monthly data on mobility – urban transit and air traffic in Canada
* Annual health indicators from 2015-2021 (counts of Canadians meeting criteria for 27 different health indicators)
* News article titles from the health section of CTV news from 4 weeks previous (try out <https://apify.com/lukaskrivka/article-extractor-smart#what-does-smart-article-extractor-do>)
* Add flu hospitalizations?

**Stage 1:** Baseline forecast using weighted moving average

**Stage 2:** Forecast using FB prophet

**Stage 3:** Forecast as a supervised learning problem:

Features to try:

* Moving average of previous 2 weeks?
* Same epi week last year – figure out how to deal with leap year
* health indicator from previous year
* google trends for various symptom searches (fever, chills, headache, sore throat etc.) from 4 weeks previous
* use NLP to include news article titles from the health section of CTV news from 4 weeks previous (try out <https://apify.com/lukaskrivka/article-extractor-smart#what-does-smart-article-extractor-do>)
* mobility data from 4 weeks previous
* look into getting 'FluWatcher' data (weekly self-reports of symptoms from volunteers) from 4 weeks previous

Try various regression models / LSTM

**Stage 5:** Implement online learning so that model can be retrained when new week's numbers come out

* SGDRegressor
* PassiveAggressiveRegressor
* MLPRegressor

**Stage 6:** Deploy the model using Streamlit web app to show predicted cases over the next four weeks

* could also plot predictor features over time? E.g. different google trends, health indicators, mobility?

# Milestone 1 – Research Prior Work

* Finish Friday, November 25th
  + Read 3-4 research papers on influenza forecasting
  + Identify common models, features, evaluation metrics
  + Find a benchmark to evaluate my final model against
  + Review lectures / exercises / walkthroughs on time series forecasting
  + Set up Readme structure

# Milestone 2 – Acquire/Preprocess/Explore Dataset

* Finished

# Milestone 3/4 - Design model, Train and evaluate model

* Finish Saturday, November 26th
  + ~~Baseline Moving Average Model (make 1 step = 4 weeks):~~ 
    - ~~Model A: Use data from prior to March 2020 only~~
    - ~~Model B: Use whole time series~~
      * <https://towardsdatascience.com/multi-step-time-series-forecasting-with-arima-lightgbm-and-prophet-cc9e3f95dfb0>
      * <https://machinelearningmastery.com/multi-step-time-series-forecasting/>
      * <https://machinelearningmastery.com/moving-average-smoothing-for-time-series-forecasting-python/#:~:text=The%20moving%20average%20can%20be,predict%20the%20next%20time%20step>.
      * <https://towardsdatascience.com/defining-the-moving-average-model-for-time-series-forecasting-in-python-626781db2502>
      * <https://www.geeksforgeeks.org/how-to-calculate-moving-average-in-a-pandas-dataframe/#:~:text=In%20Python%2C%20we%20can%20calculate,a%20parameter%20in%20the%20function%20>.
      * <https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc43.htm#:~:text=Whereas%20in%20Single%20Moving%20Averages,forecasting%20than%20the%20older%20observations>.
  + Prophet Model:
    - Model A: Use data from prior to March 2020 only
    - Model B: Use whole time series
      * <https://facebook.github.io/prophet/docs/handling_shocks.html#treating-covid-19-lockdowns-as-a-one-off-holidays>
      * https://facebook.github.io/prophet/docs/diagnostics.html
* Finish by Monday, November 28th
  + Supervised Learning Model – Many Experts Approach
    - Model A: Basic Date Time, Lag, Window Features (e.g., month, day, previous week, two previous weeks, last year)
      * ADD IN PREVIOUS WEEKS, PREVIOUS YEAR AS FEATURES
    - Model B: Google trends features
    - Model C: Mobility features
    - Model D: Health indicators features
    - Model E: Prophet model?
    - Model F: Voting Regressor Model that accepts Models A, B, C, D, E and outputs prediction
  + Finalize project description in spreadsheet

# Milestone 5 - Version 1 of Capstone

* Finish by Tuesday, November 29th
  + Learn how to use Tableau
  + Draft presentation
  + Clean up code
  + Draft readme
  + Clean up github

# Milestone 6 – Model Development

* Finish by Friday, December 2nd
  + LSTM model
  + Online Learning
  + Hyperparameter tuning with different models
  + Dimensionality reduction
  + Stretch: Try an end-to-end approach and train a single model with all features, try adding article titles from CTV using NLP, Transformers, Particle Filtering

# Milestone 7 - Make figures

* Finish Saturday, December 3rd
  + Work on Tableau figures
  + Work on presentation
  + Work on Readme

# Milestone 8 – Finish Presentation

* Finish December 5th
  + Finishing touches on presentation, practice presenting

|  |  |  |  |
| --- | --- | --- | --- |
| Model – Post-COVID DATES | Test MSE | Test RMSE | Test MAE |
| 2-week moving average – 1 week ahead | 156456.326 | 395.546 | 201.757 |
| 2-week moving average – 2 weeks ahead | 452901.212 | 672.979 | 363.14 |
| 2-week moving average – 3 weeks ahead | 769671.565 | 877.309 | 493.923 |
| 2-week moving average – 4 weeks ahead | 1262253.255 | 1123.5 | 697.043 |
| ARIMA(2,0,7) – 1 week ahead | 435890.265 | 660.22 | 660.22 |
| ARIMA(2,0,7) – 2 weeks ahead | 2555784.276 | 1598.682 | 1476.698 |
| ARIMA(2,0,7) – 3 weeks ahead | 3603337.539 | 1898.246 | 1604.237 |
| ARIMA(2,0,7) – 4 weeks ahead | 3256009.613 | 1804.442 | 1383.402 |

|  |  |  |  |
| --- | --- | --- | --- |
| Model – Pre-COVID DATES | Test MSE | Test RMSE | Test MAE |
| 2-week moving average – 1 week ahead | 177356.256 | 421.137 | 251.062 |
| 2-week moving average – 2 weeks ahead | 519380.628 | 720.681 | 445.739 |
| 2-week moving average – 3 weeks ahead | 933826.007 | 966.347 | 619.11 |
| 2-week moving average – 4 weeks ahead | 1434981.361 | 1197.907 | 770.789 |
| ARIMA(2,0,6) – 1 week ahead | 19515.161 | 139.697 | 139.697 |
| ARIMA(2,0,6) – 2 weeks ahead | 244579.525 | 494.55 | 467.69 |
| ARIMA(2,0,6) – 3 weeks ahead | 50539.367 | 224.81 | 185.685 |
| ARIMA(2,0,6) – 4 weeks ahead | 785688.536 | 886.391 | 852.602 |
| Linear Regression – 1 week ahead | 1782145.805 | 1334.97 | 1334.97 |
| Random Forest (season) – 1 week ahead | 11828.738 | 108.76 | 108.76 |
| Random Forest (season) – 2 weeks ahead | 97081.911 | 311.58 | 311.25 |
| Random Forest (season) – 3 weeks ahead | 206036.315 | 453.912 | 443.407 |
| Random Forest (season) – 4 weeks ahead | 119252.581 | 345.33 | 228.905 |
| Random Forest (flu symptoms) – 1 week ahead | 435719.268 | 660.09 | 660.09 |
| Random Forest (flu symptoms) – 2 weeks ahead | 424747.631 | 651.727 | 650.498 |
| Random Forest (flu symptoms) – 3 weeks ahead | 295923.824 | 543.989 | 513.751 |
| Random Forest (flu symptoms) – 4 weeks ahead | 243593.986 | 493.552 | 468.207 |

|  |  |  |  |
| --- | --- | --- | --- |
| Benchmark Models | Test MSE | Test RMSE | Test MAE |
| Comparative\_evaluation\_time\_series\_models\_predicting\_influenza\_Iran (random forest) |  | 22.78 | 14.99 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |