Covid Data Analysis

Group 13: Yunhe Jia, You Wu, Yixuan Zeng, Meilin Li, Muhuan Lyu

Contents

Dataset

Both of the datasets are from the CDC.

- Dataset1: COVID-19 Case Surveillance Public Use Data with Geography
  - This patient-level dataset includes demographics and geography features such as sex, ethnicity, exposure history, county and state of residence, death or not, etc.

- Dataset2: COVID-19 Vaccinations in the United States by County
  - Dataset 2 is a aggregated data that includes covid-19 vaccine administration and vaccine equity data at county level.

URLs:

- Dataset1: https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4
- Dataset2: : https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh

Data Pre-processing

- Build an ETL pipeline for data pre-processing:
  - Covid Data
    - Drop records with unknown status of death
  - Vaccination Data
    - Group data to month level and count number of vaccinated
  - Merged two datasets
    - Merge covid and vaccination data by year-month, state and county.

- Execution Time: 0.46s
- Cluster : 8 Node i3.xlarge cluster with 9.1 LTS (includes Apache Spark 3.1.2, Scala 2.12)

Analytical Goals

- Predict the probability of death of patients
- Predict the amount of Covid death with time series model
- Predict the cumulative vaccinated population with time series model

Implementation

Analytical Goal #1: Predict the probability of death of patients

- We used four different models to predict the probability of death of patients:

Model	Accuracy	Area under ROC	Area under PR	F1	Execution Time
Logistic Regression	0.945	0.908	0.563	0.932	12.77 sec
Decision Tree	0.935	0.526	0.306	0.925	4.63 sec
Random Forest	0.946	0.910	0.573	0.930	15.69 sec
K-means	0.812	-	-	-	2.59 sec

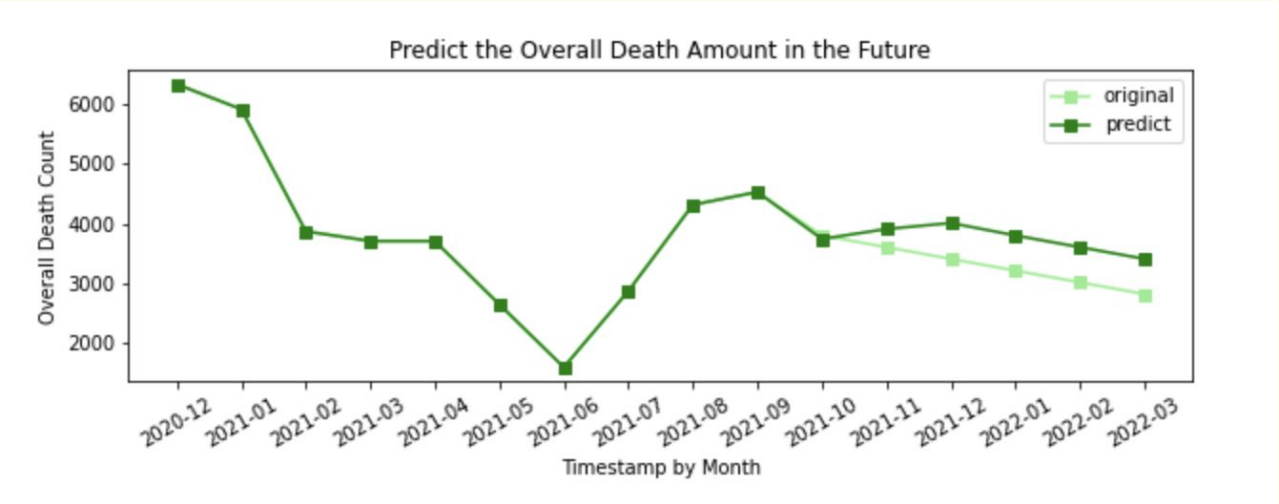
- Number of instances: 4
- Machine type: i3.xlarge
- Disk: 1 x 950 NVMe SSD
- Memory size: 30.5 GB

Implementation

Analytical Goal #2: Predict Amount of death cause by Covid-19 with time series model

Prediction plot

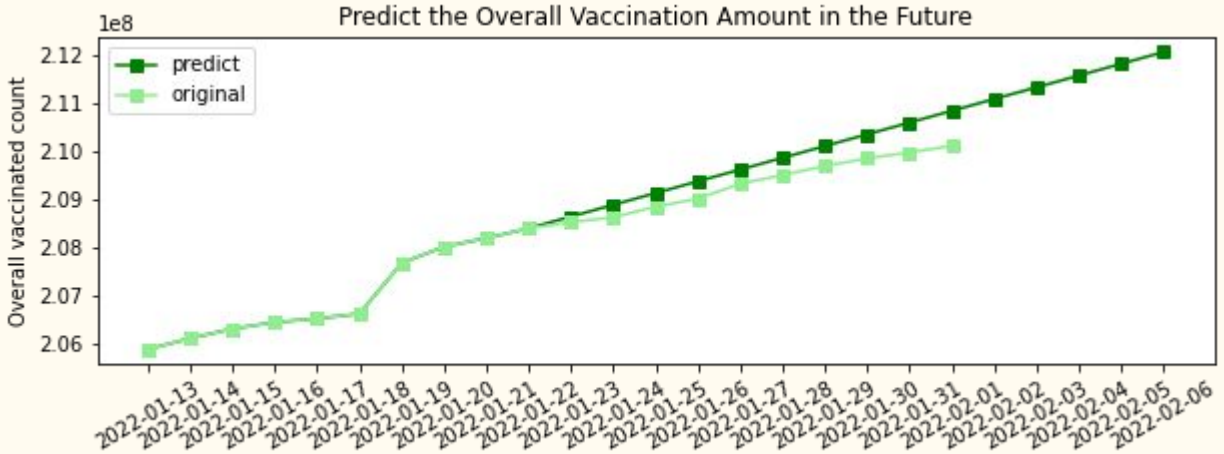
The time series data is non-stationary, which shows a large volume of fluctuations. The main reason is very likely to be the random occurrence of the virus mutations. As a result, the time series model works poorly in long run.



Implementation

Analytical Goal #3: Predicting the cumulative vaccinated population with time series model

Fitted an additive time series model on the two-year vaccinated data and predicted for the next 5 days. We obtained the predicting result that the trend will be linearly increasing over date and will reach 2.12 billion on Feb 6.



Conclusion

- ML Model:
  - ML models can be implemented to predict probability of death of a patient.
  - The Random Forest model shows a best performance and least execution time compared to Decision Tree, Logistic Model and K-means.
- Time Series Model:
  - Time series model can help to forecast infected/vaccinated amount in the future.
  - Time series model works well on fitting overall trend and short-time prediction. However bias may increase in long term run.

Lessons Learned

- It’s meaningful to leverage knowledge learned from class to solve real-world problem.
- Thanks to Pyspark, MongoDB which helped us a lot on processing and storing large dataset.
- Build our baseline model first, and then try other models and compare the results with it to choose the most suitable model.
- Use multiple metrics to evaluate models, they will evaluate performances of models from different perspectives.
- Don’t forget to pause clusters of mongodb/databricks when finished work or they will send you “surprising” bills.

Thank you