Team number: Team 2 Project R-7

Student name(s): Krithikashree Lakshminarayanan, Jahnavi Dave, Prananditha Manchikatla

Project title: Covid-19 data analysis and performance comparison of forecasting algorithms like LSTM,XGBoost and SVM (Support Vector Machine).

--

1. Project background and related work

In the past decade machine learning has proven to be extremely useful in real-world problem solving. It has been proven to be useful and advantageous in a variety of fields, including healthcare, where it performed well as a decision support system to aid in the identification of illnesses and the making of medical diagnosis. It is unnecessary to emphasize the profound and irreversible impacts that COVID-19 has had on people's lives all around the world.

Based on this background, we are trying to analyze the ability of ML models to predict the amount of forthcoming COVID-19 patients. To be more specific we are evaluating the performance of 3 standard forecasting methods like LSTM, XGboost and SVC. The project's findings will aim to reinsure that applying these techniques to the current COVID-19 pandemic scenario is a promising strategy.

Before getting into the actual workflow, we read several research papers and we have gone through various methodologies to do these predictions. A few of these techniques have served as the benchmark for our projects. The details are as follows, Firstly, according to authors of [1] to improve decision-making regarding the future course of action, machine learning (ML) based forecasting methods have been demonstrated to be the best way to foresee outcomes in these unexpected scenarios. The authors in [1] have successfully predicted factors such as the number of newly infected COVID 19 people, mortality rates and the recovered COVID-19 estimates in the next 10 days using models like xponential smoothing (ES). And another author in [2], builds a deep learning model for semi-supervised few-shot segmentation (FSS) of Covid-19 with radiographic images. Designing an efficient and precise segmentation of 2019-nCov infection from small-sized annotated lung computed tomography (CT) images is the challenge this work attempts to solve.

Based on these papers and suggestions from our professors we decided on a couple of algorithms. Since the suggested dataset was a time series, we identified the most suitable methods to deal with this kind of data. We finally decided on doing a performance comparison of 3 classification algorithms like LSTM, SVM,Xboost, and did the model evaluation using mean square error (MSE). The model's performance for forecasting future case counts and model evaluation results will be shown in the final report.

2. Project Objective

This project is about contributing towards the pandemic control situation by doing efficient forecasting on the future count of confirmed cases. Before doing that, we must understand the problem (domain knowledge about the current spread rate), the given dataset and which models to use in this type of scenarios. Therefore, the proposed methodology to implement this project successfully is as follows,

1. Understanding the dataset
    - The dataset is derived from a reliable source for COVID-19 Data maintained by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University
    - It consists of the daily confirmed cases, the death and recovery count for each country
2. Data cleaning and preparation
    - Since the data is timeseries there is a high possibility of inconsistency and missing information.
3. Identifying the appropriate algorithms for these scenarios
4. Model building, visualization, and forecasting
5. Model evaluation using prediction loss
6. Evaluating the performance against other models

3. Challenges and/or Motivation

One of the biggest challenges while working on this project is to accurately identify the correct model to represent the data and dealing with time series data will require more attention. It is a little difficult to work with time series data because we will only be working with a single feature i.e., time and there is only one historical sample available for a specific time. For example, to build a model for deaths by covid-19, we will have only a single historical observed time series available for a specific date.to overcome this challenge, a few more features are extracted from the time series data available.

Models build on time series are most likely to overfit on training data. The overfitting problem happens when the data is idiosyncratic - only occurs once, and the past data distribution is not reproduced or repeated in the future. Different types of model evaluation techniques are used to overcome overfitting.

4. Timeline and milestones (e.g., task for each week)

Task	Approximate time for completion
Understanding the problem statement	Oct 1-7
Research on similar problem scenarios	Oct 7-10
Discussion on the end goals of the project	Oct 10-14
Performing the analysis on the given dataset	Oct 14 - 17
Finding the appropriate algorithms for the given scenarios	Oct 17 - 20
Proof of concept to make sure the selected algorithm is appropriate for these analysis	Individual phase through the project
Research on the assigned algorithms	Individual phase through the project
Modeling of these algorithms	Oct 20 – Oct 25
Documentation for the project proposal	Oct 25-Oct 30
Creating the GitHub repo to add all the model code. Analyzing the performance of these models	Oct 31 – Nov 5
Visualizing the results of these modeling and performance evaluation.	Nov 5 – Nov 10
PowerPoint for the presentation	Nov 10 – Nov 15
Report writing	Nov 15 – Dec 5

5. Tools

The entire project is built on google Collab with python as the programming language. The notebooks for each algorithm can be found in the GitHub repository (Link will be attached in the final report). The tools and packages used for each process are listed below,

DATASET ANALYSIS:

1. Pandas: For creating the dataframe from the given dataset
2. Pycountry: For labeling all the countries properly before visualization
3. Matplotlib: For visually analyzing several distributions
4. Ploty: For visually analyzing several distributions
5. Seaborn: For making the graphs more presentable

LSTM:

1. MinMaxScaler from Sklenar: scale or normalize data as the data is too skewed
2. TimeSeriestraingenerator from keras : to generate data in sequences
3. Sequential from keras.models:
4. Dense from keras.layers
5. LSTM from keras.layers
6. Dropout from keras.layers

7. Activation from keras.layers

XGBoost:

1. feature_importances from xgboost: to calculate the importance of each feature considered.
2. XGBRegressor from xgboost: the actual model.
3. mean_squared_error from sklearn.metrics : to evaluate the performance of the model.

Support Vector Regression (SVR):

1. SVR from sklearn.svm library for model
2. mean_squared_error from sklearn.metrics : to evaluate the performance of the model
3. GridSearchCV to find the optimal parameters for Support Vector.

6. Responsibilities for each team member (only required for team project)

Krithikashree Lakshminarayanan	Jahnavi Dave	Prananditha Manchikatla
1. Documentation 2. Dataset analysis • Plot the global spread • Identify the most impacting country 3. LSTM for confirmed cases in USA 4. LSMT model evaluation 5. PowerPoint	1. Documentation 2. Implemented Support Vector Regression Model for predicting COVID cases in USA 3. Fine-Tuned the SVR Model 4. Evaluated the model 5. PowerPoint	1. Documentation 2. Dataset analysis • Plot the current confirmed case count. • Extract more features form the date.(i.e., month, year, week of year) • Perform feature importance. 3. Build and evaluate the XGBoost model. 4. PowerPoint.

References

[1] Machine learning models for covid-19 future forecasting by Ramesh Kumar Mojjada,[a,*] Arvind Yadav,[a] A.V. Prabhu,[b] and Yuvaraj Natarajanc

[2] FSS-2019-nCov: A deep learning architecture for semi-supervised few-shot segmentation of COVID-19 infection by Mohamed Abdel-Basset[a],Victor Chang[b],Hossam Hawash[a,] Ripon K.Chakrabortty[c] Michael Ryan[c]