**COVID Projections and Responsible Weather Factors for Infections in US**

Amol R, Devendra B, Kushal R, Sanku L and Uday

**Abstract**

The United States have recently become the country with the most reported cases of 2019 Novel Coronavirus (COVID-19). Several research organizations, regulatory agencies and universities have been tracking and forecasting COVID infections and fatalities at country (US) level especially since it peaked in Spring 2020 in US. With this work we aim to provide understanding of the COVID-19 infections in US. We have mainly focused to provide insights into the following areas, a.) Forecast of daily infections for US for the period Nov 2020 – Jan 2021, b.) Identify top 3 States which are going to lead the infections for the said period, c.) Evaluate if demographic distribution and weather parameters are directly responsible for infections.

We have primarily used time series forecasting methods and related models like Prophet and ARIMA to determine the prediction for infections for Nov-Jan 2021. Also, various linear regression and ensemble models were analyzed to determine features responsible for infections.

We also selected 8 States based on their population density to evaluate our selected model (Prophet) for infections projection of top 3 States. Overall, we have based our work on the already defined models and methods and have attempted to determine results by fine tuning data, feature engineering and selection techniques.

**Introduction**

The United States have recently become the country with the most reported cases of 2019 Novel Coronavirus (COVID-19). Several research organizations, regulatory agencies and universities have been tracking and forecasting COVID infections and fatalities at country (US) level especially since it peaked in Spring 2020 in US.

Most of the research and papers published for finding infections in US have been done around finding the role of temperature in these infections. This is mainly due to the reason that with the increasing temperature the entire world also observed dip in infections. However, as time progressed it was evident that temperature alone is not the only parameters playing role in increasing infections. Several other parameters, primarily, Demographics, Medical Conditions of an individual, Income Distribution, Demographic Distribution, Ethnicity, Local weather conditions and access to medical facilities are also playing role in the spread of infection. We therefore decided to consider all these factors while applying regression techniques in feature selection. Secondly, we also attempted to run two models, Prophet and ARIMA to derive the predictions based on the historical data. We obtained this data from Kaggle and applied data cleaning and mangling techniques to prepare as inputs for the above models. In addition, since specific regions are showing increased infections, we formulated a criterion to determine top 3 States which will lead infections.

Related Works:

Zohair Malki [1] proposed the weather parameter model for COVID infections in Italy. The data is however collected for a very short (22 Dec 2019 – Apr 2020). Moreover, the model is based mostly on temperature and humidity variables which do not consider major weather factors like wind speed, dew point, etc. Manas Jyoti [9] et al attempt to find the weather parameter correlation with active cases, recovered cases and deceased, however only air temperature is considered for the analysis. Their paper analyses the cases for Indian subcontinent using intuitionistic fuzzy sets (IFS). Ismail Kirbas [6] et al tried to predict the infection cases using confirmed COVID-19 cases of Denmark, Belgium, Germany, France, United Kingdom, Finland, Switzerland and Turkey. Their paper used models with Auto-Regressive Integrated Moving Average (ARIMA), Nonlinear Autoregression Neural Network (NARNN) and Long-Short Term Memory (LSTM) approaches. Six model performance metrics were used to select the most accurate model (MSE, PSNR, RMSE, NRMSE, MAPE and SMAPE). According to the results of the first step of the study, LSTM was found the most accurate model. However, since the volume/duration of data was only for 3 months the predictions might not be highly accuracy due to the recent non-linearity of the infections.

**Material and Methods**

1. Data Collection

For our study we collected used a Kaggle dataset which had consolidated the data from multiple sources, namely John Hopkins for Case(Infection) count, Death Count, CDC & NOAA for Weather Features like Mean Temperature, Snow, Fog, Precipitation etc and other demographic, per capita income and lifestyle related parameters. The data was collected for all US counties for the period Jan 2020 to Oct 2020 on a daily basis.

1. Prediction for Infection Count for US
2. Prediction for top 3 States in US

……………………END OF ABSTRACT………………………………………>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

INTRODUCTION

…………………………END OF INTRODUCTION>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

<<START OF PROBLEM SOLUTION 1A – PROJECTIONS

………………………..END OF PROJECTIONS SOLUTIONS…>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

<<<<<START OF PROBLEM SOLUTION 1B

In continuation with determining the projections for US and identifying the States which will contribute to infections significantly we also attempted to identify key weather parameters responsible for the increase in infections. The entire data had cumulative infections count and was converted to a daily count format. We ran several linear regression models namely, Lasso & Ridge regressor, Recursive fit regression, SFS (Sequential Forward Selection), SFFS (Sequential Forward Feature Selection), SBS (Sequential Backward Selection). We also applied ensemble regression methods like MLP, Elastic Net, Least Angle Regressor, Extra Tree Regressor, Gradient Booster, Ada Booster, XG Booster, Light GBM Regressor, Random Forest and Extra Tree Regressor to identify the weather features responsible. These regression techniques were used with key features like, Population density, Ethnicity features(number of Black/Hispanic/White/Native American/Asian/Hawaiian), Age features(<17 yrs and 65 yrs <), Per capita income, Lifestyle features (Overcrowding, Life expectancy, Physically active or not), Pre-medical conditions (Diabetes/HIV) and Weather Features (Dew, Precipitation, Mean Temperature, Snow, Fog, Wind Speed).

The scoring method we used to derive the performance of these models are based on Mean Average Error. Based on the performance of these models we derived the importance of the features (From 1-7 out of total 33 features) which yielded results as tabled below.

The first table show the top features using linear regression technique only. It clearly shows that ethnicity features form the top features (1-5) while weather parameter form the second best.

The second table depicts the findings from ensemble regression models. Almost all ensemble techniques have prioritized (1-5) weather features while ethnicity features are second best.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **F1** | **F2** | **F3** | **F4** | **F5** | **F6** | **F7** | **MAE** |
| Lasso | NHS | NHW | NHV | NBL | NAI | N17 | NAS | 25.58 |
| Ridge | NHS | NHW | NAS | NBL | NAI | N17 | NAS | 28.46 |
| SFS | NBL | NAS | N17 | NUS | PVC | WSP | DPT | 11.99 |
| SFFS | NBL | NAS | N17 | NUS | WSP | PTT | SNW | 12.38 |
| SBS | NBL | NAI | NPI | NHS | NHW | PIA | WSP | 11.99 |
| Recursive Reg | NBL | NAS | NHS | NHW | N65 | NDS | NBP | 30.15 |

**NBL**: Number of Black, **NAS**: Number of Asian

**NHS**: Number of Hispanic, **NHW**: Number of Non-Hispanic White

**NAI**: Number of American Indian Alaska Native

**NPI**: Number of Native Hawaiian Pacific Islander

**N17**: Number of Age Less Than 17

**N65**: Number of Age Greater Than 65

**F1 – F7:** Feature Priority

Table 1: Priority of Features - Linear Regression

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **F1** | **F2** | **F3** | **F4** | **F5** | **F6** | **F7** | **MAE** |
| Extra Tree Regressor | MTP | DPT | WSP | NUS | N17 | NBP | PTT | 10.55 |
| Random Forest Regressor | MTP | NBP | DPT | WSP | NUS | NHS | PTT | 10.43 |
| AdaBooster | NUS | WSP | MTP | NBP | DPT | NBL | N17 | 20.88 |
| Gradient Boost Regressor | NUS | NBP | MTP | NHV | N17 | NHS | DPT | 11.11 |
| XGB | NBP | NHV | NHS | NEP | NUS | N17 | NAS | 11.11 |
| LGBM Regressor | MTP | DPT | WSP | POD | PTT | NBL | NHS | 10.82 |
| CatBoost Regressor | MTP | NHV | DPT | NUS | WSP | NBP | PSK | 10.95 |

**WSP**: Wind Speed, **MTP**: Mean Temperature**,**

**DPT**: Dew Point, **PTT**: Precipitation, **SNW**: Snow,

**POD**: Population Density

**NHV**: Number of HIV Cases,

**NEP**: Num Unemployed

**NPP**: Number of Primary Care Physicians

**NBP**: Num of people below poverty,

**PSK**: Percent Smokers

**PIA**: Percent Physically Inactive,

**NRU**: Num of Rural

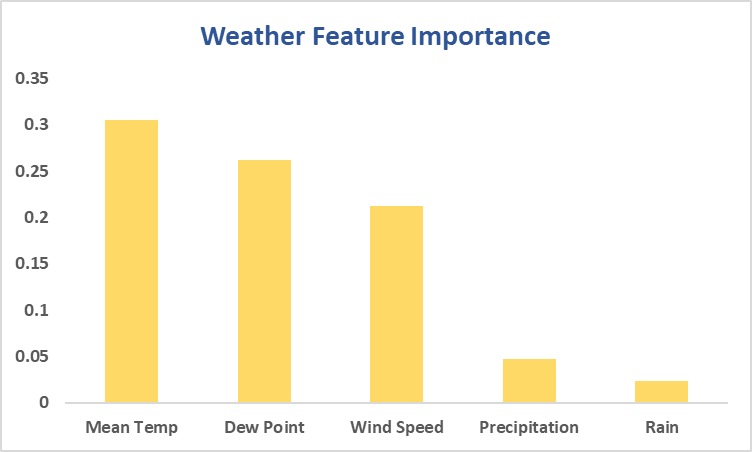
**NUS**: Num of Uninsured,

**PVC**: Percent Vaccinated

**NDS:** Num Disabled

Table 2: Priority of Features - Ensemble Method

After evaluating the results of feature priority for both Linear and Ensemble methods there is no clear indication that weather parameters are the only top features impacting infections though both the models enlist at least 3 weather parameters in top 7 feature list. Moreover, the models are also exhibiting that ethnicity and age factors are also playing major roles. As our study is focused to find the weather parameters, we will circumvent only those weather features which are highlighted by all the models in their top 7 features list. The below graph depicts the top weather features:



Apart from the temperature parameters, key ethnicity and age parameters which seem to impact infections are,

1. Number of Black,
2. Number of Hispanics,
3. Number of Non-Hispanic Whites,
4. Age greater than 65yrs
5. Age less than 17yrs

REFERENCES

[1] Zohair Malki, El-Sayed Atlama, Aboul Ella Hassanienc, Guesh Dagnewd, Mostafa A Elhosseini, Ibrahim Gad, Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches

[2] Sayantani Basu, Roy H. Campbell, Going by the numbers: Learning and modeling COVID-19 disease dynamics, Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, United States

[3] Sarbjit Singh, Kulvinder Singh Parmar, Sidhu Jitendra Singh Makkhan, Jatinder Kaur, Shruti Peshoria, Jatindar Kumar, Study of ARIMA and least square support vector machine (LS-SVM) models for the prediction of SARS-CoV-2 confirmed cases in the most affected countries

[4] Harun Kinac, Mehmet Güray Ünsal b, Resat Kasap, A close look at 2019 novel coronavirus (COVID 19) infections in Turkey using time series analysis & efficiency analysis

[5] Amit Singhal, Pushpendra Singh, Brejesh Lall, Shiv Dutt Joshi, Modeling and prediction of COVID-19 pandemic using Gaussian mixture model

[6] Ismail Kirbas, Adnan Sözenb, Azim Dogus¸ Tuncer, Fikret Sinasi Kazancıoglu, Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches

[7] Md. Sabbir Hossain, Sulaiman Ahmed, Md. Jamal Uddin, Impact of weather on COVID-19 transmission in south Asian countries: An application of the ARIMAX model.

[8] Dimitrios Paraskevis, Evangelia Georgia Kostaki, Nikiforos Alygizakis, Nikolaos S. Thomaidis, Constantinos Cartalis, Sotirios Tsiodras, Meletios Athanasios Dimopoulos, A review of the impact of weather and climate variables to COVID-19: In the absence of public health measures high temperatures cannot probably mitigate outbreaks

[9] Manash Jyoti Borah, Bipan Hazarika, Sumati Kumari Panda, Juan Jose Nieto, Examining the correlation between the weather conditions and COVID-19 pandemic in India: A mathematical evidence

[10] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, “Epidemic processes in complex networks,” Reviews of modern physics, vol. 87, no. 3, p. 925, 2015.

[11] A. Y. Pastore Piontti, M. F. D. C. Gomes, N. Samay, N. Perra, and A. Vespignani, “The infection tree of global epidemics,” Network Science, vol. 2, no. 1, pp. 132–137, 2014.

[12] L. Pellis, F. Ball, S. Bansal, K. Eames, T. House, V. Isham, and P. Trapman, “Eight challenges for network epidemic models,” Epidemics, vol. 10, pp. 58–62, 2015.

[13] J. O. Wertheim, A. J. Leigh Brown, N. L. Hepler, S. R. Mehta, D. D. Richman, D. M. Smith, and S. L. Kosakovsky Pond, “The global transmission network of HIV-1,” Journal of Infectious Diseases, vol. 209, no. 2, pp. 304–313, 2014.

[14] K. Mizumoto, K. Kagaya, A. Zarebski, and G. Chowell, “Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020,” Eurosurveillance, vol. 25, no. 10, p. 2000180, 2020.

[15] W. Gleissner, “The spread of epidemics,” Applied Mathematics & Computation, vol. 27, pp. 167–171, 1988.