CSE6242 Project Final Report

Team 129: Xuanxuan Xue, Xiwen Chen, Jessie Hu, Yujun Kong, Zhongmin Tao, Liuyi Ye

Introduction

Our project was inspired by several recent research which focused on the large-scale outbreak of Covid-19 among nursing homes and their policy responding to pandemics. With further research into this topic, we realized how nursing homes responding to the pandemic can greatly impact their effectiveness of preventing spread of pandemics facility-wide. So far, researchers have yet gone a step further to use multidimensional large-scale data to actively visualize and escalate the results for the wider community to acknowledge this issue. Therefore, we decided to build an interactive platform based on in-depth study of large-scale nursing home data. Our platform will help policy makers and nursing home management better understand what measures can help facilities effectively prepare and respond to pandemics. More importantly, Seniors can utilize our platform to identify outperforming nursing homes. The success of this project will contribute to building a more solid fence for vulnerable residents and staff in nursing homes.

Problem Definition

Our project studies the correlation between nursing home features and their effectiveness in responding to pandemics through machine learning models, based on the existing nursing home data and Covid-19 data. In particular, 9 models are trained and compared for feature selection and correlation study. A rating metrics oriented to nursing home Covid-19 performance is developed to reflect facility effectiveness in responding to pandemic. Our findings are then translated into an interactive platform for statewide exploration of outperforming nursing homes.

Survey

Since covid-19 started spreading in the US, the high case rate and death rate in nursing homes are distinct from the broader community. Seniors, who comprise the majority of the long-term care (LTC) population, are more vulnerable to respiratory diseases, owing to advanced age and pre-existing chronic conditions (Walsh et al., 2013; Cohen et al., 2014; Furmenti et al., 2019). Meanwhile, LTC facilities are well recognized to have easier circulation of a wide range of respiratory viruses (Davidson et al., 2020; Li et al., 2020; Ji et al., 2020). These facts demonstrate the urgent need to understand what features of nursing homes are likely to play roles in spreading contagious infection within the community.

Several studies (Chen et al., 2020; He, Li, & Fang, 2020, Li at al., 2020) showed that nursing home settings, characteristics and policies corresponding to pandemic response can greatly impact their effectiveness in preventing disease spread. A study identified that nursing homes having a higher case rate is more significantly related to the facility size than the type of ownership or rating stars of the facility (Abrams et al., 2020). Arguments are made that proactive measures like visitor restriction, monitor and screening mandates, and adherence to preventive policy may prevent spread (McMichael et al., 2020; Ouslander & Grabowski, 2020). Considine (2019) found that residents face higher risks when accurate assessment is absent before interhospital transfers. In terms of staffing, the lack of qualified nurses (Davidson, 2020; Flynn et al., 2010) and staff interinstitutional mobility (McMichael et al., 2020) brings risk to epidemic transmission.

However, most research is limited to a short period of time and lacks longitudinal tests, which is an important factor of such studies (Sugg et al., 2020). Studies also pointed out that many states still have room for improvement in the support they provide to facilities regarding infection control (R. Dorritie et al). Further studies will bring implication to policy initiatives to better regulate nursing

homes under outbreaks (Smith et al., 2008) and prevent disease spreads within the facilities (Martinez, 2020).

Proposed method

Our innovation

- We developed a novel rating metric to rank the nursing homes mainly based on their competence in response to epidemics. The rating metric is developed based on around 80 indicators related to different aspects of nursing home characteristics. Our ranking system emphasizes on the confirmed death rate measured by "Total Residents COVID-19 Deaths as a Percentage of Confirmed COVID-19 Cases", which is different from all existing ranking systems.
- We visualized the results on an interactive choropleth map and showed top ranked nursing homes in each county. This will allow potential nursing home residents to identify outperforming nursing homes closed to them.
- We displayed significant features that have impact on the rank, this will allow policy makers and nursing home managements to track what strategies can help facilities better respond to pandemic.
- The use of cross-sectional comparison and feature filters in the platform allows stakeholders to actively control predictors of their concerns and investigate their interest of views. Users can associate evidence to their hypotheses and generate perspectives of specific issues.
- The dataset we are using is continuously updated and enlarged. This allows for longitudinal study.

Data Source

We combined COVID-19 Nursing Home Dataset and Nursing Home Provider Info Dataset from Centers for Medicare & Medicaid Services. These datasets consist of the most comprehensive countrywide nursing home covid-19 data and general info data that we have access to, which makes it an ideal source for answering our research questions. The datasets have records of over 15,000 currently active nursing homes and over 5 months of reported covid-19 data.

Data preprocessing

In order to perform further analysis, we merged the nursing home provider dataset with the Covid-19 dataset. The number of nursing homes reduced to around 14,000. The merged dataset has over 160 nursing home features. In this report, we use the percentile rank of indicator "Total Residents COVID-19 Deaths as a Percentage of Confirmed COVID-19 Cases" as our response variable, considering the population size varies in different nursing homes, to study the effectiveness responding to pandemic. The independent variables of interest include characteristics of nursing homes related to administration, infrastructure, pandemic response policy, and the resources available to nursing homes (supply of N95 masks, Ventilators...etc.). We removed abnormal observations such as those nursing homes with a 'Total Residents COVID-19 Deaths as a Percentage of Confirmed COVID-19 Cases' that exceeds 100%. Since we plan to focus on the overall performance of nursing homes, we decided to keep the last reported observation per nursing home. This date varies from the end of May to recent dates. We compute the average of the resources over the entire period of our dataset and use the average measures as regressors into the algorithms.

Feature Selection

There are more than 160 features in the merged dataset. We performed feature selection in order to avoid overfitting and multicollinearity. The first step is to drop features that have more than 95% missing values, and columns that have irrelevant data such as phone number, name, location,

etc. The second step is to identify highly correlated features by generating correlation matrices. We found eight pairs of features that have correlation scores greater than 0.8. We kept the most relevant one from each pair. After excluding other redundant columns, we kept 80 features.

Modeling

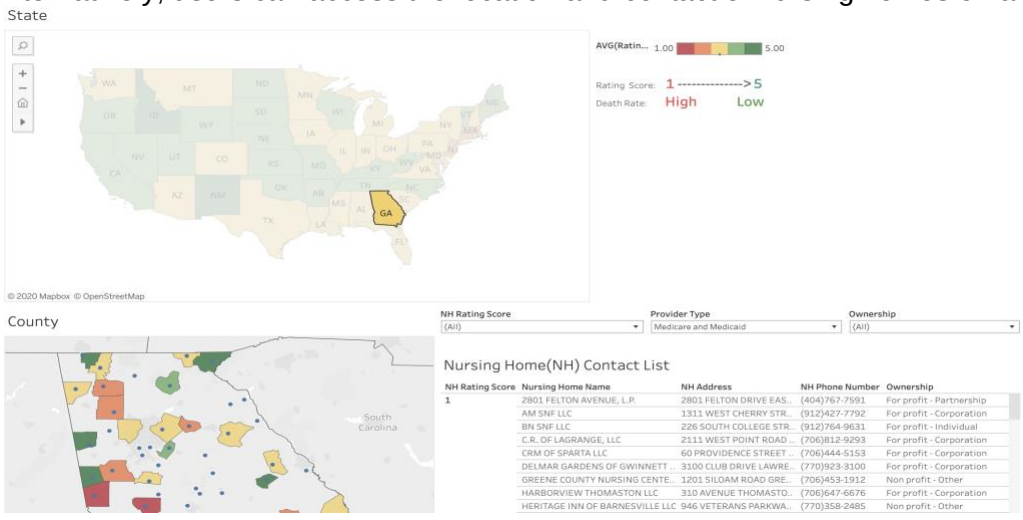
We built nine machine learning models that fit our case in order to compare and cross validate the effectiveness of selected features. The performance of most models was very poor after the first round of feature selection. With further experiment on different features and model tuning, the outcome significantly improved. Following is the summary of how each model was set up.

- K-Nearest Neighbor (KNN): To tune the best fitted model, the number of nearest neighbors has iterated from 1 to 20.
- Elastic Net: Regularized binary logistic regression that linearly combines L1-norm and L2-norm.
- Lasso regression: Performs L2 regularization, and the minimization objective = LS Obj + α * (SS of coefficients).
- Ridge regression: Performs L1 regularization and the minimization objective = LS Obj + α * (sum of absolute value of coefficients).
- Neural network: We tuned the model and finally set up the model structure to be 5 layers. We also compared the accuracy among three different learning rates.
- Random Forest: We tuned models with different tree-size and tree numbers to get a converged outcome.
- Support Vector Machines: Set a nonlinear kernel model as a classifier.
- Naive Bayes: A simple Gaussian Naive Bayes model.
- Decision Tree: We tuned classification tree model with optimal parameters.

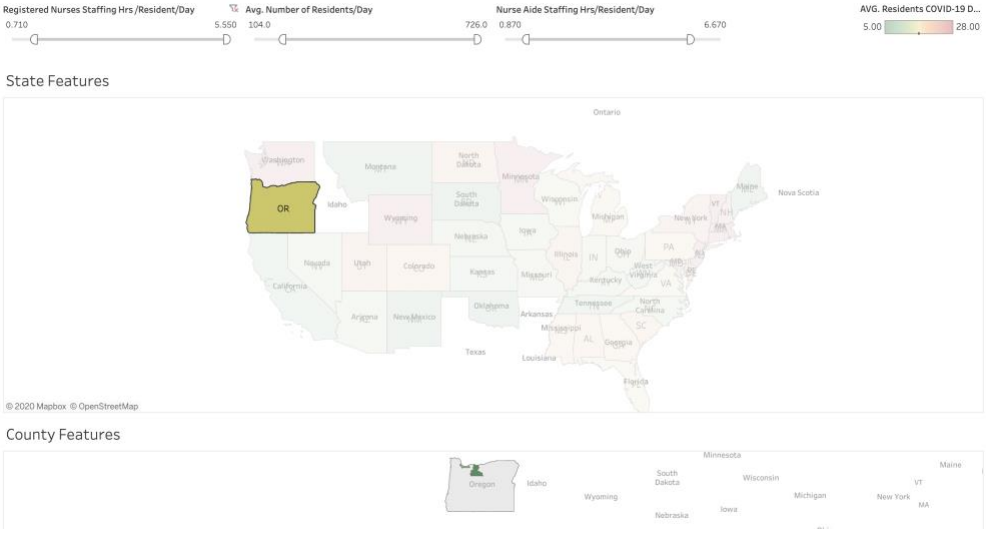
Visualization

We visualized the ranking system and translated our findings into an interactive tool named “Nursing Home Explorer” using Tableau. This tool provides dynamic and user-oriented information of nursing homes by allowing users to generate customized views. It contains three parts.

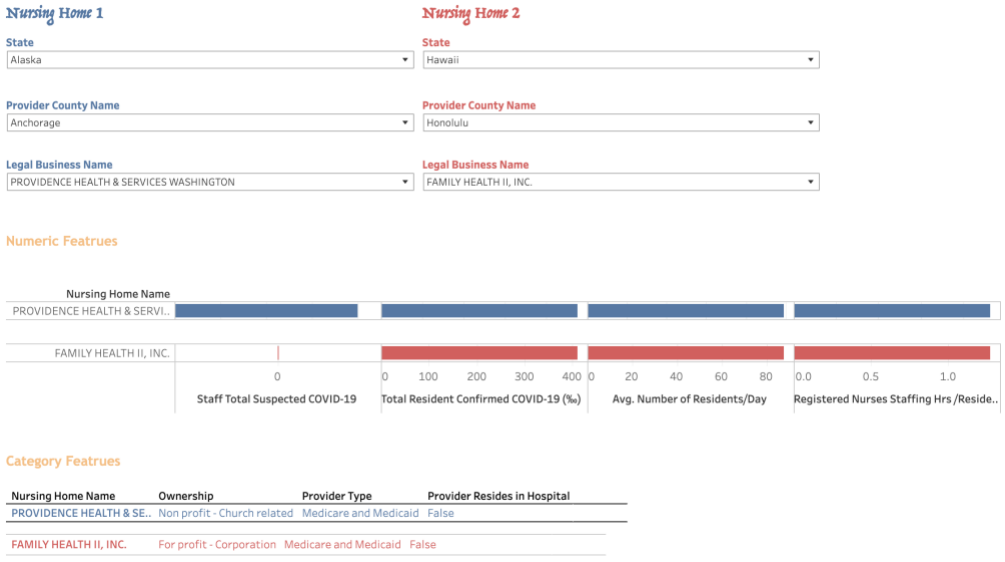
1. A choropleth map with multiple layers to visualize the detailed information of nursing homes and their ranking. The choropleth map can zoom in to state and county level and the colors reflect the ranking within regions. When landing on a specific state, a separate map will automatically display a detailed county level view. By clicking on dots of each county, users will be redirected to the Google map page of the highest ranked nursing home for more details. Alternatively, users can access the location and contact of nursing homes on a side table.



2. Feature range filter sliders are created for the key features to allow users focus on feature values of their interest. Users can visualize the death rate change of each region and nursing home distribution change on the map.



3. Side-by-side bar charts allow users to compare the features of two specific nursing homes. For instance, for two Nursing Homes with the same rating score, if users give more weight on the availability of medical resources to residents, they can use this function to look into the specific difference.



Experiments and evaluation

Feature selection improvement

Question to answer:

- How to improve our method to select features that are significant in predicting nursing home effectiveness?

In our first round of feature selection, all models fail to achieve high accuracy or low mean squared error (MSE) (see Table 1), which means the features we first selected did not reflect the true competence of nursing homes in responding to pandemic.

	KNN	Elastic Net	Lasso Reg	Ridge Reg	Naive Bayes	Decision Tree	Random Forest	SVM	Neural Network
MSE	3.72	1.17	2.80	4.56	3.46	3.92	3.744	5.79	4.29
Accuracy (%)	38.22	N/A	32.19	28.90	14.20	36.72	42.23	37.56	33.75

Table 1: Benchmark models MSE and accuracy comparison

We investigated the selection process. In the initial procedure, we transformed 9 categorical variables to dummy variables and this raised bias in further analysis. Some categorical variables are highly skewed to certain values. For instance, a feature has an overwhelming majority of "False" and barely has "True". These variables have weak explanatory power. At the other end, some categorical variables have over 500 unique values, so that transfer into too many dummy variables and significantly increase model complexity and reduce model performance. Therefore, we selectively drop the categorical variables with weak explanatory power and re-trained all models. All models achieved significantly better results than the benchmark models. Especially, random forest, decision tree, and neural network model outperformed other models (see Table2). The results validate that it is rational to drop those features.

	KNN	Elastic Net	Lasso Reg	Ridge Reg	Naive Bayes	Decision Tree	Random Forest	SVM	Neural Network
MSE	1.70	1.17	0.42	0.42	1.97	0.01	0.08	0.36	0.08
Accuracy (%)	58.50	22.19	62.75	62.30	58.40	98.83	99.16	87.87	95.07

Table 2: Improved models MSE and accuracy comparison

Computational time also significantly reduced. For example, the benchmark KNN model costs over 30 minutes to finish 20 iterations due to enormous size of features, and the improved KNN model only takes a few minutes and achieves higher accuracy.

Evaluation of Selected Features & Dimension Reduce

Question to answer:

- Are the features selected really significant in answering our research questions?
- What are the most significant features?

In order to cross validate the significance of features we selected from previous steps. We adopted another method for feature selection. Since the random forest model has the best accuracy, we used the feature importance module in the random forest model. We compared the outcome with our selected features and selectively kept features that exist in both methods with high importance. Additionally, we decided to keep the top 5 most important nursing home features in our demo, in order to improve the user experience.

User experience survey

Question to answer:

- Are the functions and interface of our explorer user-friendly?
- Is the tool self-explanatory enough in using?

We developed a user experience survey to evaluate the user experience of our interactive explorer using Google Survey platform. The purpose of this survey is to understand how useful and comfortable the users find the tool for exploring the different features of nursing homes and their correlation to the facilities' effectiveness of responding to Covid-19 pandemics. We invited more than 25 users to participate in the survey. We received constructive feedback and improved our tool accordingly. Overall, the tool was highly rated in terms of performance and user experience.

Contribution

Name	Contribution
Xuanxuan Xue	Topic research, literature survey, data exploration, data ETL, modeling, improve models/ further analysis, report documentation
Xiwen Chen	Topic research, literature review, data exploration, data ETL, data analysis, modeling, proposal and report compile
Jessie Hu	Topic research, literature review, data exploration, data ETL, data analysis, data visualization, survey, feature selection
Yujun Kong	Topic research, literature review, coding, proposal compile, poster
Zhongmin Tao	Topic research, literature review, modeling, feature selection, advanced analysis, survey
Liuyi Ye	Topic research, literature review, data exploration, data cleaning, data analysis, poster

Conclusions and discussion

In this study, we explored two large nursing home datasets and combined them to investigate the performance of nursing homes during COVID-19. We built a novel metric to assess the performance of the nursing home and developed robust models for feature selection and performance prediction. We built an interactive platform on the features as predictors to nursing home practice during pandemic and received fairly good feedback. There are still rooms in improving our visualization tool to be more user-friendly for senior adults in the future based on the population of our user. Overall, we fulfilled our plan and our results will bring implication to residents in the nursing homes, the policy makers and nursing home designers.

During this study, we realize that more constructive works can be done in studying the performance predictors of nursing homes in combating epidemics. Though the result of this study is based on Covid-19 data, it has meaningful implications on broader respiratory epidemics studies, as research has shown the similarity in epidemic transmission among nursing homes. The dataset we are using is continuously updated and enlarged. This provides us opportunities to monitor how features change can impact performance change in a longitudinal study. With that said, the models we trained can be further tuned in predicting future performance.

Reference

Abrams, H., Loomer, L., Gandhi, A., & Grabowski, D. (2020, July 07). Characteristics of U.S. Nursing Homes with COVID-19 Cases. Retrieved October 06, 2020, from https://onlinelibrary.wiley.com/doi/full/10.1111/jgs.16661

Center for Medicare & Medicaid Services (2020, September 20). Retrieved October 06, 2020, from https://data.cms.gov/stories/s/COVID-19-Nursing-Home-Data/bkwz-xpvg

Chen, M., Chevalier, J., & Long, E. (2020, July 30). Nursing Home Staff Networks and COVID-19. Retrieved October 06, 2020, from https://www.nber.org/papers/w27608

Cohen, C., Herzig, C., Carter, E., Pogorzelska-Maziarz, M., Larson, E., & Stone, P. (2014, April). State focus on health care-associated infection prevention in nursing homes. Retrieved October 08, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4030678/

Considine, J., Street, M., Bucknall, T., Rawson, H., Hutchison, A. F., Dunning, T., . . . Hutchinson, A. M. (2018). Characteristics and outcomes of emergency interhospital transfers from subacute to acute care for clinical deterioration. *International Journal for Quality in Health Care, 31*(2), 117-124. doi:10.1093/intqhc/mzy135

Davidson, P., & Szanton, S. (2020, May 11). Nursing homes and COVID-19: We can and should do better. Retrieved October 06, 2020, from https://onlinelibrary.wiley.com/doi/full/10.1111/jocn.15297

Dean, A., HR, A., AC, R., Al., E., EJ, C., TM, M., . . . MA, D. (2020, September 10). Mortality Rates From COVID-19 Are Lower In Unionized Nursing Homes. Retrieved October 06, 2020, from https://www.healthaffairs.org/doi/full/10.1377/hlthaff.2020.01011

Dorritie, R., Quigley, D., Agarwal, M., Tark, A., Dick, A., & Stone, P. (2020, February 14). Support of nursing homes in infection management varies by US State Departments of Health. Retrieved October 07, 2020, from https://www.sciencedirect.com/science/article/abs/pii/S0195670120300554

Flynn, L., Liang, Y., Dickson, G., & Aiken, L. (2010, November 04). Effects of Nursing Practice Environments on Quality Outcomes in Nursing Homes. Retrieved October 08, 2020, from https://onlinelibrary.wiley.com/doi/full/10.1111/j.1532-5415.2010.03162.x

Furmenti MF;Rossello P;Bianco S;Olivero E;Thomas R;Emelurumonye IN;Zotti CM; ;. (2019, February 19). Healthcare-associated infections and antimicrobial use in long-term care facilities (HALT3): An overview of the Italian situation. Retrieved October 07, 2020, from https://pubmed.ncbi.nlm.nih.gov/30790605/

He, M., Li, Y., & Fang, F. (2020, June 15). Is There a Link between Nursing Home Reported Quality and COVID-19 Cases? Evidence from California Skilled Nursing Facilities. Retrieved October 06, 2020, from https://www.sciencedirect.com/science/article/abs/pii/S1525861020305211

Ji, Y., Ma, Z., Peppelenbosch, M. P., & Pan, Q. (2020). Potential association between COVID-19 mortality and health-care resource availability. *The Lancet Global Health, 8*(4). doi:10.1016/s2214-109x(20)30068-1

Li, Y., Temkin-Greener, H., Shan, G., & Cai, X. (2020, July 21). COVID-19 Infections and Deaths among Connecticut Nursing Home Residents: Facility Correlates. Retrieved October 06, 2020, from https://onlinelibrary.wiley.com/doi/full/10.1111/jgs.16689

Martinez, M. (2018, November 8). The calendar of epidemics: Seasonal cycles of infectious diseases. Retrieved October 06, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6224126/

McMichael, T., Al., E., For the Public Health–Seattle and King County, Author AffiliationsFrom Public Health–Seattle and King County (T.M.M., E. J. Anderson and Others, J. H. Beigel and Others, & M. J. Mina and Others. (2020, September 29). Epidemiology of Covid-19 in a Long-Term Care Facility in King County, Washington: NEJM. Retrieved October 07, 2020, from https://www.nejm.org/doi/full/10.1056/NEJMoa2005412

Ouslander, J., & Grabowski, D. (2020, September 02). COVID-19 in Nursing Homes: Calming the Perfect Storm. Retrieved October 08, 2020, from https://onlinelibrary.wiley.com/doi/full/10.1111/jgs.16784

Smith, P., Shostrom, V., Smith, A., Kaufmann, M., & Mody, L. (2008, July 23). Preparedness for pandemic influenza in nursing homes: A 2-state survey. Retrieved October 06, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3319409/

Sugg, M., Spaulding, T., Lane, S., Runkle, J., Harden, S., Hege, A., & Iyer, L. (2020, August 25). Mapping community-level determinants of COVID-19 transmission in nursing homes: A multi-scale approach. Retrieved October 06, 2020, from https://www.sciencedirect.com/science/article/pii/S0048969720354759

Walsh, E., Shin, J., & Falsey, A. (2013, August 06). Clinical Impact of Human Coronaviruses 229E and OC43 Infection in Diverse Adult Populations. Retrieved October 07, 2020, from https://academic.oup.com/jid/article/208/10/1634/841065