Project Paper
Data Mining
EN.625.740.81.FA23

Eric Farish

November 21, 2023

1 Introduction

The purpose of this analysis is to determine the contributing factors of COVID-19 mortality during a time-period in the United States spanning January 2020 to June 2021. This analysis will also try to determine an order of importance for those factors. The purpose of this analysis is not produce an optimize predictive model. It is focused on understanding the contributed factors to COVID deaths. Performance metrics are only used to evaluate the quality of the models. As these will be regression models, 'RMSE' will be used to evaluate the quality of each model.

The analysis conclusions can be reviewed here.

2 Literature Review

Below is a review of research on COVID-19 comorbidity factors. This research provides justification for inclusion of dataset variables into the models.

1. **Diabetes:** A survey[5] conducted by the Mayo Clinic of 1169 adults aged 18 years of age or older who tested positive found that diabetes was a common comorbidity with 11.2% having this condition. In addition, per the American Diabetes Association[1], persons with diabetes are more likely than others to become severely ill if infected with COVID-19.

2. **Smoking:** As COVID-19 is a disease attacks the lung, it is reasonable to assume smokers are at higher risk of death. This statement[7] from the World Health Organization claims research suggests that smokers are at higher risk of developing severe disease and death.

3. **Income:** An article[2] published by the Lancet found that persons in the lowest income decile had a probability of dying from COVID-19 five times greater than those at the top decile.

4. **Education:** A cross-sectional study[3] published by the American Medical Association found that if all racial and ethnic populations had experienced the same mortality rates as college-educated non-Hispanic White populations, 71% fewer deaths among racial and ethnic minority populations would have occurred.

5. **Social Distancing:** A study[6] published by the NIH showed that higher social distancing was associated with a 29% reduction in COVID-19 incidence and a 35% reduction in COVID-19 mortality.

6. **Vaccinations:** An article[4] published by Lancet estimated that vaccinations prevented 14·4 million deaths from COVID-19 in 185 countries and territories between Dec 8, 2020, and Dec 8, 2021.

3 The Dataset

After deciding I was interested in understanding the factors that contributed to COVID-19 mortality during the pandemic, I started search the internet for data sets to analyze. During that search, I came across this paper. Haratian, Arezoo, et al. (2021) compiled a dataset containing factors they considered relevant for COVID-19 mortality.

The author's dataset contains daily observations for each of the 3142 US counties and covers the time-period of January 2020 through June 2021. Per the abstract of the paper, the data was collected from public online databases. The dataset contains temporal and fixed variables. The the fixed variables are the same for all observations. The data dictionary is here and the data itself is located here. The target variable is **covid_19_deaths**.

There are 992,266 records with no missing values as the authors imputed missing data. The observations for 1181 counties were removed due to missing data that could not be imputed. This leaves data for 1961 counties. There are over 46 variables in the data file, I chose to use 31 of them based on research above.

3.1 ETL Performed

Using the Python library SQLite, SQL was used to aggregates the daily data found in the file "imputed-data.rar" to the county level. Therefore, the 992,266 daily county detail records are aggregated to 1961 aggregate county records. In addition, the following transformation were done.

1. The categorical variables for social GPAs are converted to numerical values and averaged over the period. For example, a grade of "A" was converted to 4.0, "A-" to 3.7, etc.
2. Temporal variables were averaged during aggregation.
3. Fixed variables had their maximum value taken during aggregation.
4. The target variable **covid_19_deaths** is transformed into a new target variable **covid_19_deaths_per_100k**. I thought this variable would be more comparable between counties.

4 Feature Importance Models

Five models will be used to determine the factors for COVID-19 deaths during this time period.

1. A Null Model whose performance all other models must surpass to be part of the analysis.
2. A Ordinary-Least-Squares (OLS) model.
3. A Decision Tree Model.
4. A Random Forest Model.
5. A Multi-layer Perceptron Model

4.1 Null Model

This model takes the average of the aggregated COVID-19 deaths per 100 thousand county residence and uses that as the prediction. The standard error of this model will be compared to the RMSE statistics of the following models. All models must have a RMSE lower than the standard deviation of the Null Model. The average COVID-19 deaths per 100K is 192.16. The standard deviation is 96.77.

4.2 Model 1 - Linear Regression Model

The linear model will be repeatedly fitted to remove statistically insignificant predictors and address model issue like multi-collinearity. In addition, diagnostic tests will be done to make sure the models don't break the OLS regression assumptions below:

1. A linear relationship between response and predictor variables.

2. The homoscedasticity of residuals for different levels of the predictor variables.
3. The Independence (no Autocorrelation) of residuals.
4. The residuals follow a normal distributed.
5. No Multicollinearity between predictor variables.

4.2.1 Build

The **statsmodels** library will be used to build this model. Repeated fits where done to:

1. Remove insignificant predictors.
2. Access Multi-collinearity (VIF).
3. Diagnostic tests to verify OLS assumptions.

The final model consisted of the following statistically significant coefficients.

Model 1: Linear Regression			
Parameter	p-value	CI LB	CI UB
Intercept	0.000	489.349	691.438
percent_of_smokers	0.000	3.168	6.595
percent_of_diabetes	0.000	1.310	3.943
median_household_income	0.000	-0.002	-0.001
population_density	0.020	0.000	0.005
social_distance_gpa_visitation	0.000	42.323	132.268
percent_of_vaccinated_residents	0.000	-1.721	-1.091
age_20_24	0.000	-12.569	-8.400
age_30_34	0.000	-31.250	-18.548
age_55_59	0.000	-31.138	-19.374
age_65_69	0.000	-28.768	-17.517
age_80_84	0.000	19.640	40.914
age_85_or_higher	0.001	6.376	23.699
workplaces_mobility_percent_change	0.010	-2.441	-0.329

All the variables are significant. The F-statistic of 70.48 shows that, overall, the model is statistically significant. Lets now check for multicollinearity.

4.2.2 Multicollinearity

The variance inflation factor (VIF) will be used to check for multi-collinearity in the remaining predictors. This statistic quantifies the extent of correlation between one predictor and the other predictors in a model. Predictors with high VIF values indicate correlation between the variables. Significant multicollinearity makes it difficult to accurately assess the contribution of predictors to a model. This is due to coefficient estimates and p-values in the regression output being unreliable. Per **An Introduction to Statistical Learning** (James et al, 2nd edition), VIF values exceeding 5 indicates a problematic amount of collinearity. Below are the top 5 largest parameters VIF statistics.

Model 1: VIF Values	
Parameter	VIF Value
Intercept	699.94
age_80_84	3.38
median_household_income	3.14
age_65_69	3.04
age_85_or_higher	2.75

Most of the multicollinearity is coming from the intercept term and, thus, can be ignored. Based on the results or the predictor variables, there is a moderate but not problematic amount of collinearity in the design matrix. Moving on, lets now check for the normality of the model residuals.

4.2.3 Residuals vs fitted plot

Residual plots are useful for identifying non-linearity and heteroscedasticity. Below is such a plot for the current linear model. Also included in this plot is a red line which indicates the fit of a locally weighted scatterplot smoothing (lowess), a local regression method. It can be interpreted as follows: The fit is almost equal to the dotted horizontal line where the residuals are zero. This is an indication for a linear relationship. For non-linear relationships, the red line will deviate strongly from the horizontal line.

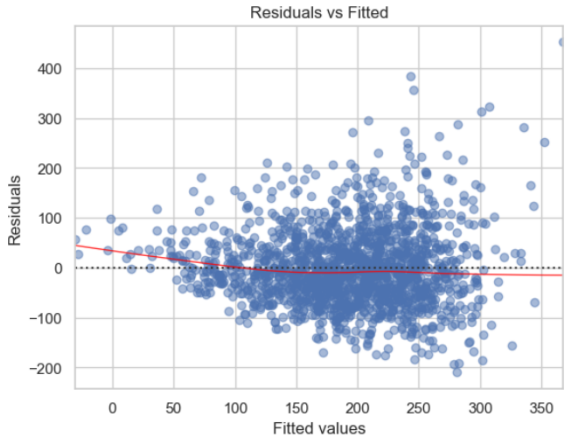

Figure 1: Model 1 Residuals Plot

There is some indication of non-normal distribution of residuals. But the deviation does not appear to be strong. Lets take another view of the same issues.

4.2.4 Normal Q-Q Plot

Below is a Normal Q-Q Plot. This plot shows if the residuals are normally distributed. A good normal Q-Q plot has all of the residuals lying on or close to the red line.

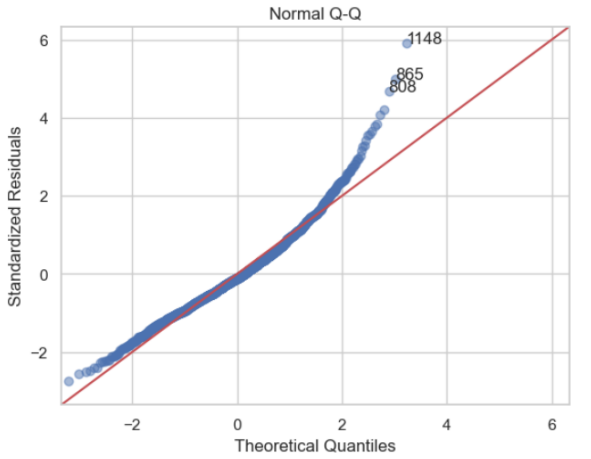

Figure 2: Model 1 Residuals Plot

There appears to be significant deviation from normality in the upper quantiles of COVID-19 deaths. To try to address this, I'll first do pair-wise plots of the individual variable's relationship with the dependent variable

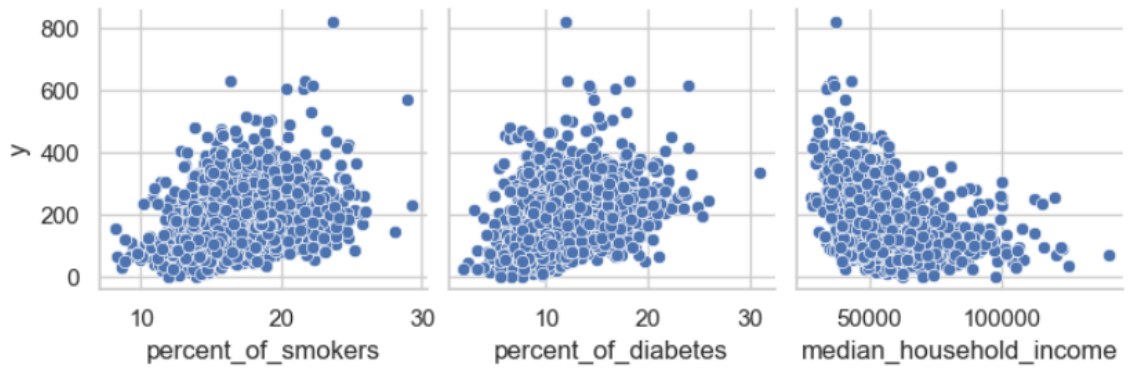

Figure 3: Residuals Plot

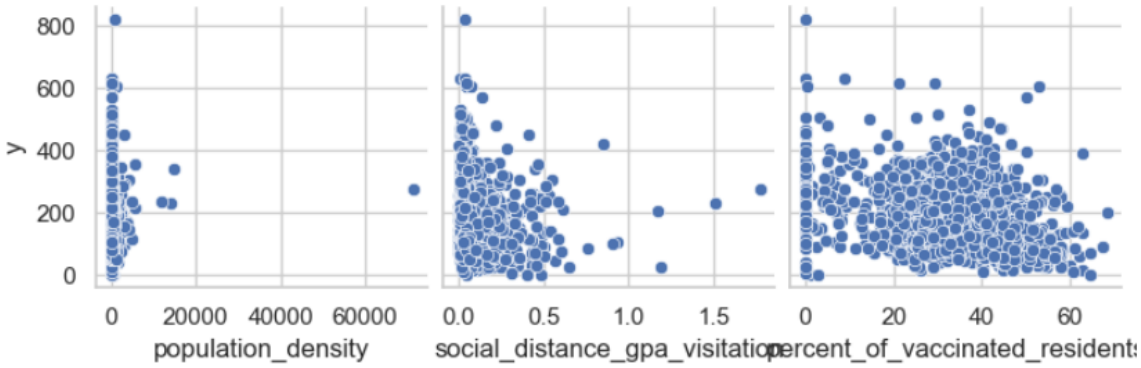

Figure 4: Residuals Plot

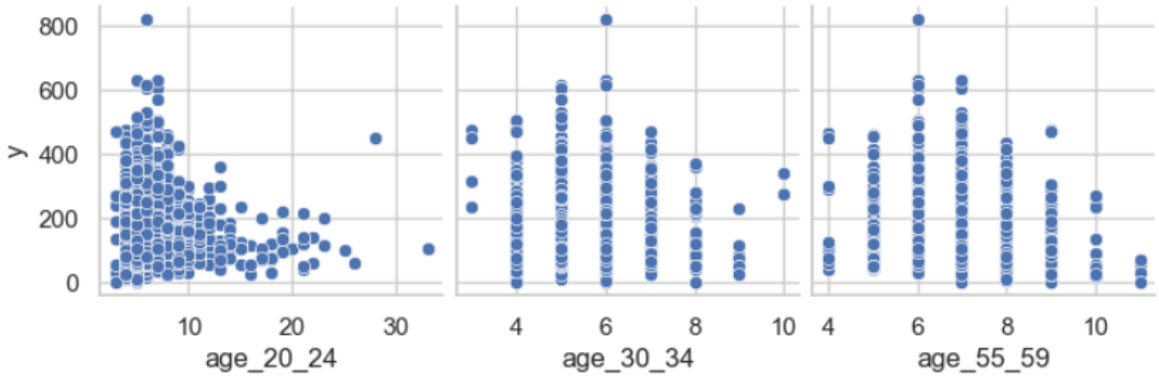

Figure 5: Residuals Plot

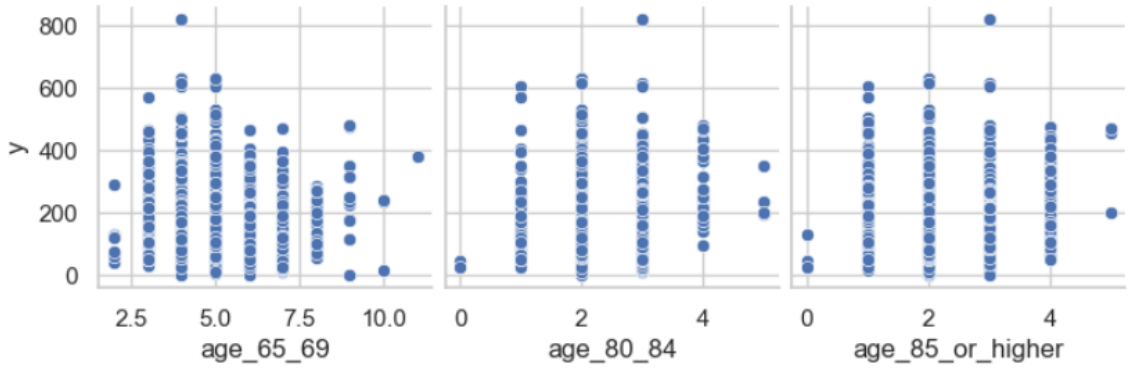

Figure 6: Residuals Plot

Figure 7: Residuals Plot

Based on the pair-wise plots above, **median_household_income** and **age_20_24** appear to have logarithmic relationship with the response variable. Therefore, I tried applying a log transformation to these two variables and refitted the model.

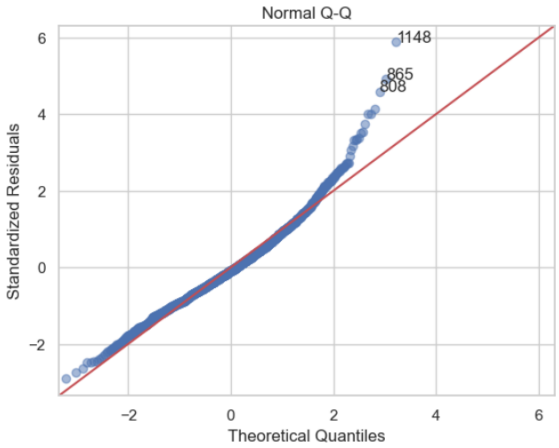

Figure 8: Normal QQ Plot

The impact of the two transformations is marginal, at best. There seems to be an improvement in the lower quartiles. The transformations are not helping.

To determine if the residuals deviation from normality is significant, I did a Shapiro-Wilk test. This tests the null hypothesis that the data was drawn from a normal distribution. I retrained using the original dataset and calculate the test test statistic 0.9657 and p-value 7.8027e-19. The very small p-value provides evidence that the assumption of normality of the residuals can be rejected.

My last attempt was creating a log-linear model by applying a natural-log transformation to the response variable. Below is the Normal Q-Q plots.

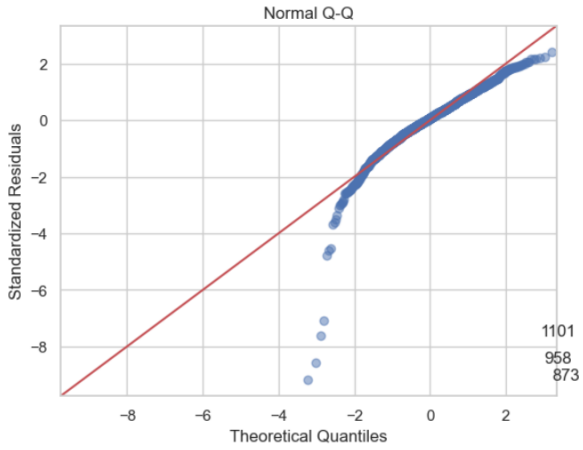

Figure 9: Normal QQ Plot

This appears to create even more problems. The non-normality of the residuals (and my inability to fix it) is a violation of the OLS assumption of normality of residuals. The impact of this violation results in the confidence intervals and p-values for parameters estimations not being reliable and, therefore, not a good model to use for feature importance for this dataset. Despite this, I will still evaluate model's quality by examining its **RMSE**.

4.2.5 Model Evaluation

Thirty averages of 10-fold cross validations were performed. I did this 30 times to leverage the Central Limit Theorem and produce performance metrics that are normally distributed.

The average RMSE is 78.48 with SD 0.4616. The average Adjusted R^2 is 0.32 with SD 0.0098. The test dataset RMSE is 78.9931 and the Adjusted R^2 was 0.3185. The model performance is better than the Null Model.

4.2.6 Summary

I could not create an OLS model that satisfied the assumptions of OLS. In particular, the normality of residuals assumption appears to be violated. This may also indicate the relationship between the predictors and the response may not be linear. These violations make the linear model parameters estimates unreliable. Therefore, for this dataset, OLS is not a good model to use for feature importance despite having performance results better than the Null Model.

4.3 Model 2 - Random Forest Regression

Decision Trees and Random Forest regressors do not make the assumptions that OLS regression does. In particular, the linear relationship between the predictors and the target variable. Their only assumptions are that there is some predictive power for the features in the model and that the decision trees are not correlated. Based on my analysis, I think here is predictive value in the features. As for the correlation between trees, the Random Forest algorithm address this by choosing 'm' variables randomly to build each estimator tree.

A drawback of Decision Trees and Random Forest models is that they can easily over-fit the data. To avoid this, 30 averages of 10-fold cross validation will be performed.

Both Decision Trees and Random Forest track the importance of features in reducing the entropy present in the data. After an optimized tree is built, the importance of each feature will be evaluated.

The Random Forest model is fitted first.

4.3.1 Build

A random forest model was optimized using randomized search grid cross-validation. $RMSE$ is be the performance metric.

4.3.2 Model Evaluation

Using the optimized model, thirty averages of 10-fold cross validations was performed. The average RMSE was 71.88 with a standard deviation 0.3319. The average of Adjusted R^2 was 0.43 with standard deviation of 0.0062. On the test dataset had a RMSE of 69.81 and a Adjusted R^2 of 0.47.

4.3.3 Model Interpretation

Each estimator in the random forest calculates the importance of a feature according to its ability to increase the pureness of the decision tree leaves. The random forest algorithm takes this calculation, averages them across the trees and then normalizes the values. Below is the plot for the random forest regression estimator features importance.

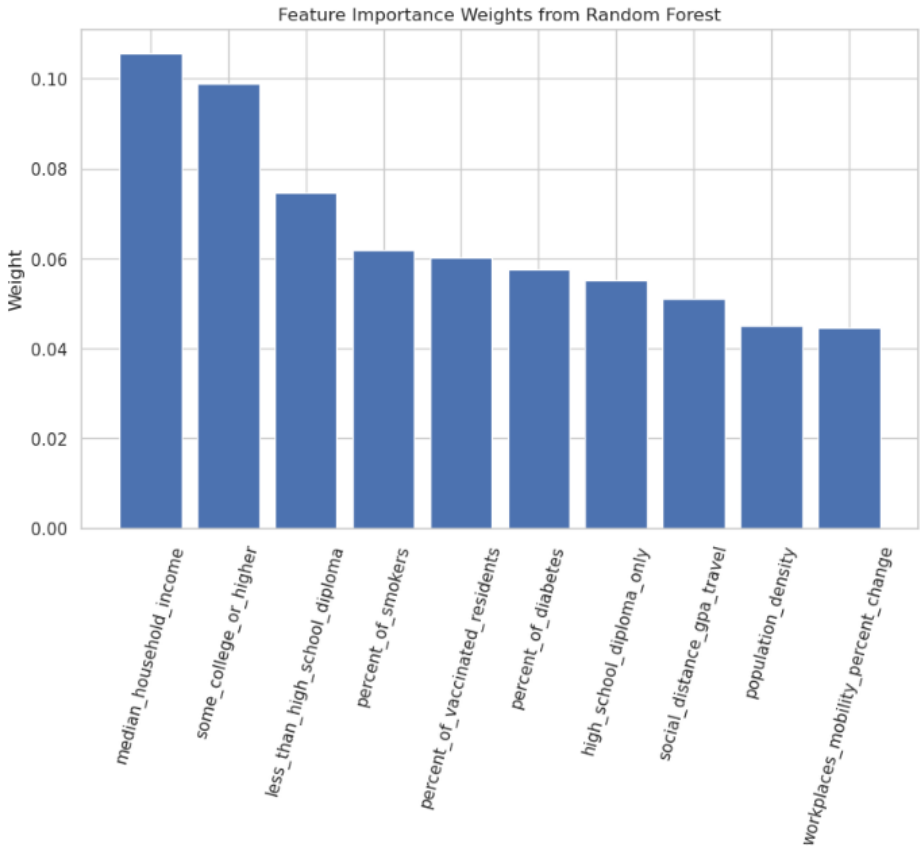

Figure 10: Feature Importance

The 5 most import features are college education, income, vaccination status, population density, and smoking. The top 10 important features account for 66% of the impurity reduction of the model.

Impurity-based feature importance ranks the numerical features to be the most important features. The impurity-based importance is a training dataset statistic and therefore does not reflect the model's ability to predict feature importance on unseen data.

To address this, the Permutation Importance of the Random Forest features were computed. This procedure removes each feature from the model and then re-calculates the model performance and records any decrease in performance. These calculations are done using the held-out test dataset.

Per the **scikit-Learn** documentation: The permutation feature importance is defined to be the decrease in a model score when a single feature value is randomly shuffled. This procedure breaks the relationship between the feature and the target, thus the drop in the model score is indicative of how much the model depends on the feature. This technique benefits from being model agnostic and can be calculated many times with different permutations of the feature.

Below, the Permutation Importance is calculated.

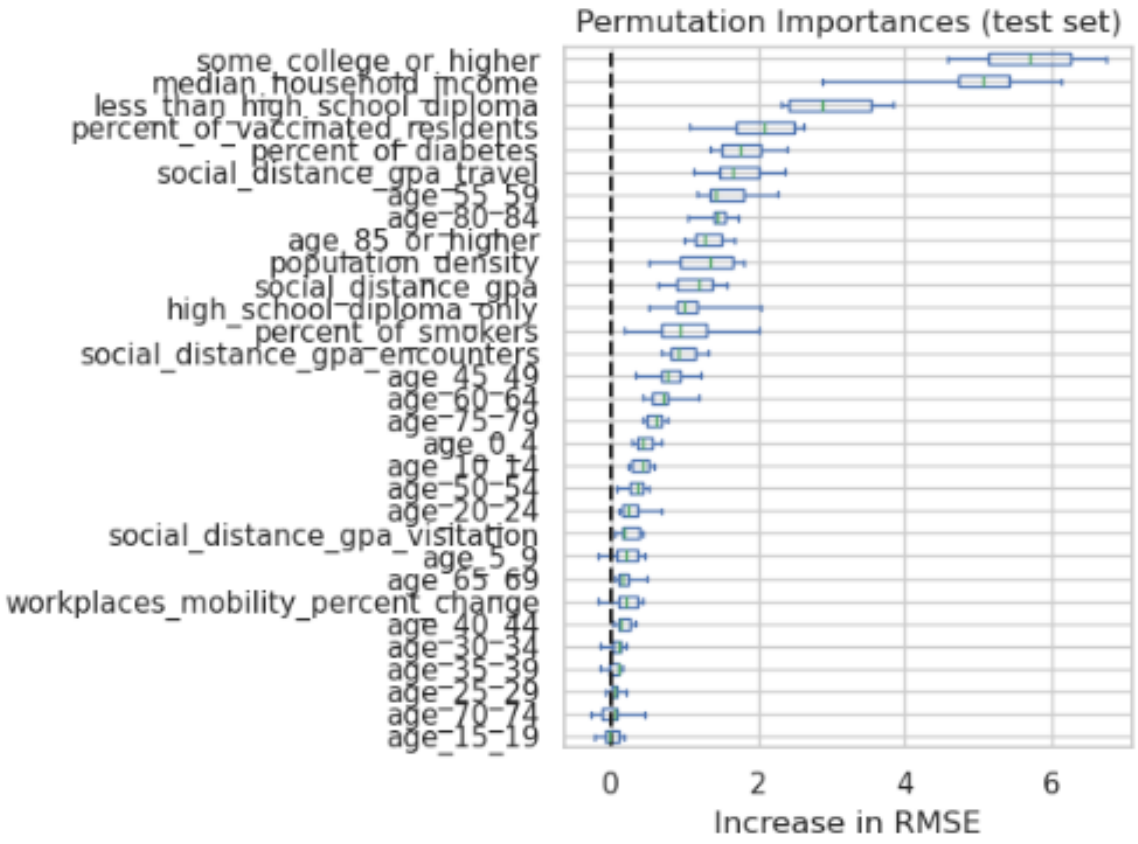

Figure 11: Permutation Importance

These results mostly agree with the top 10 features obtained from the training data. However, the feature importance from the test data has the feature **age_80_84** in its top 10 where the training data did not.

4.4 Model 3 - Decision Tree Regression

Above, many different decision tree estimators are used in the Random Forest model. Here a single decision tree will be evaluated.

4.4.1 Build

A single decision tree was optimized using a Randomized Grid Search.

4.4.2 Model Evaluation

The average of 30 10-fold average RMSE was 84.06 with standard deviation of 0.4263. The average Adjusted R^2 was 0.22 with standard deviation of 0.0096. The test dataset had an R^2 of 0.19 and a

RMSE of 86.08.

The RMSE of the Decision Tree model is inferior to the OLS and Random Forest models but better than the Null Model.

4.4.3 Model Interpretation

The feature importance of this model will be obtained using the same approach as was used for the Random Forest model. Below is the feature importance.

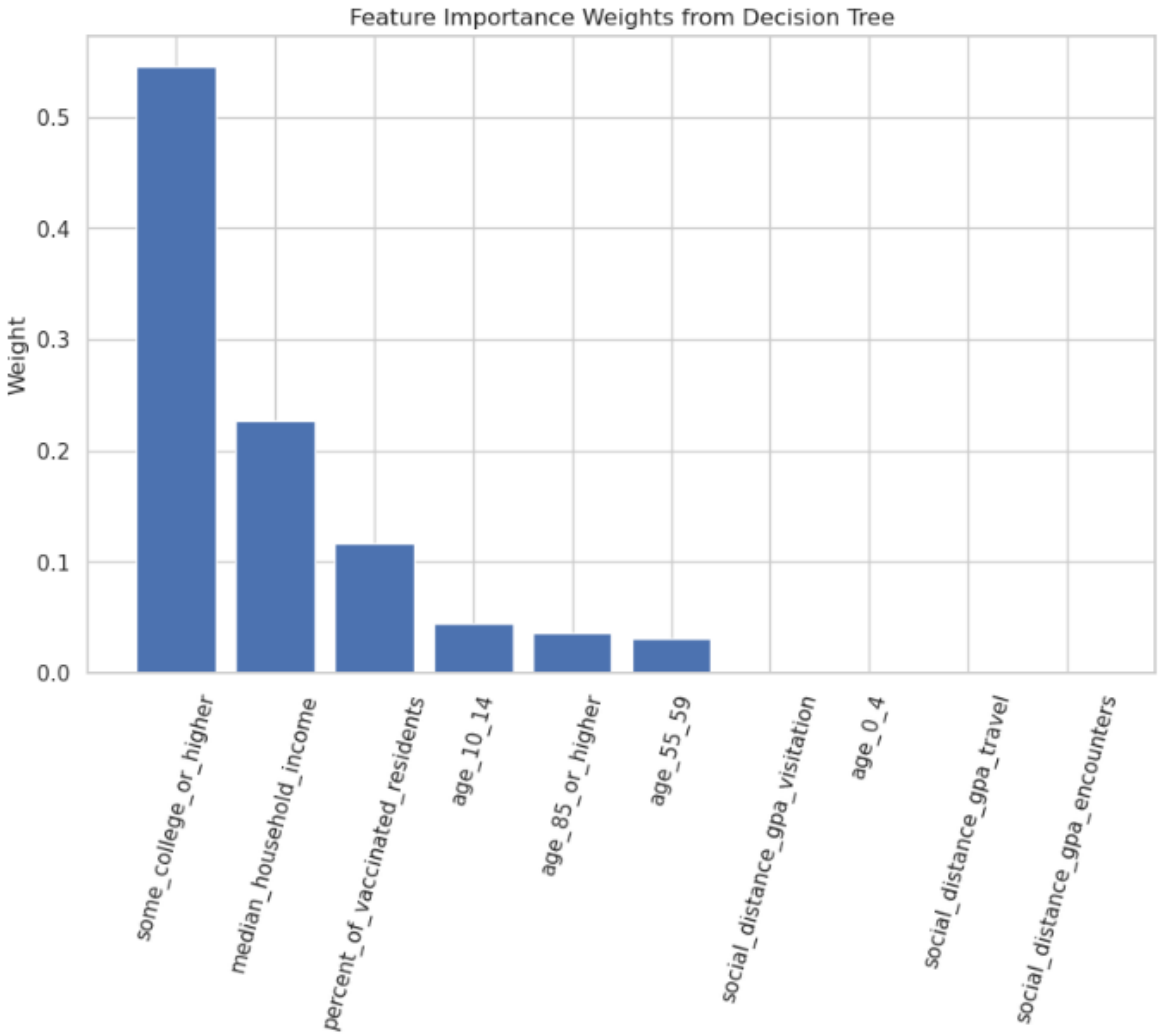

Figure 12: Feature Importance

Below is the Permutation Importance.

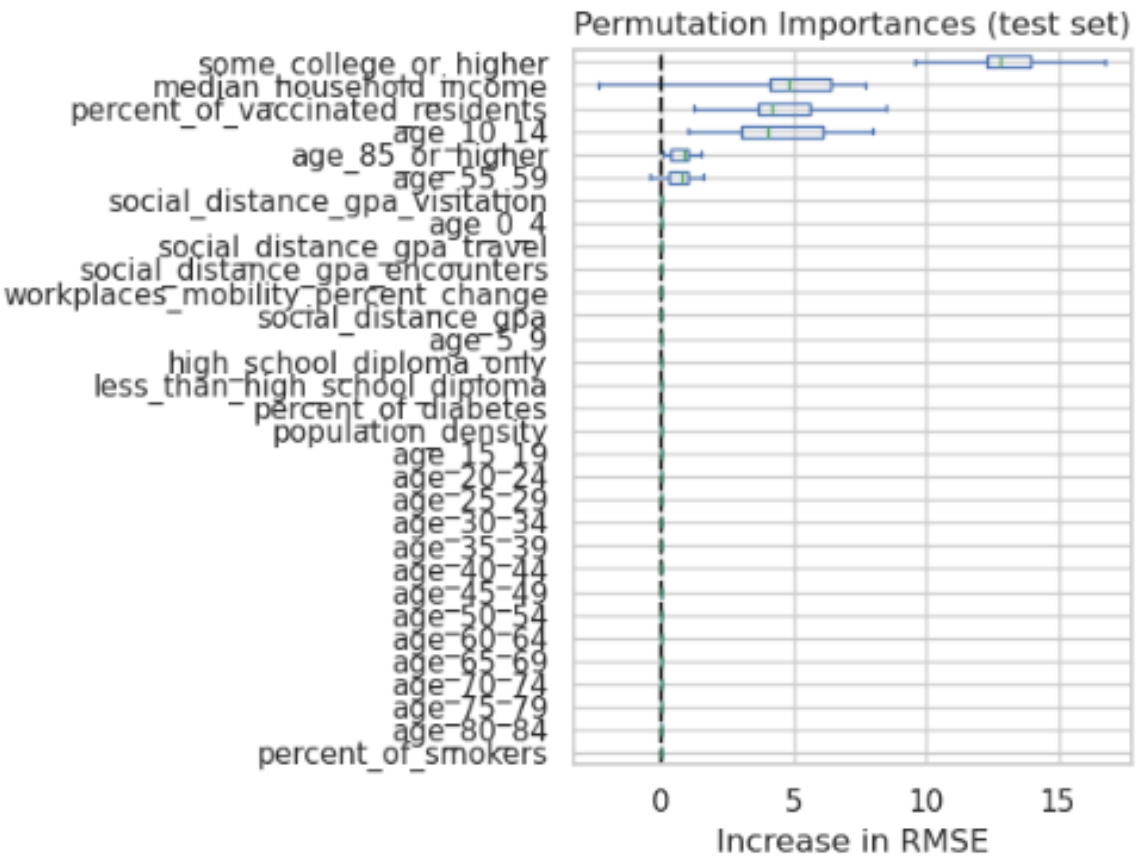

Figure 13: Permutation Importance

4.5 Model 4 - Multi-layer Perceptron (MLP)

4.5.1 Build

A hyper-parameter randomized grid search was performed to find an optimized MLP model. Please the Jupyter Notebook accompanying this distribution for the code.

The best performing model had the following configuration.

1. Solver: Adam
2. Epochs: 2000
3. Hidden Layers: [500, 300, 100, 50, 10]
4. Batch Size: 25
5. Alpha: 0.0001
6. Activation: ReLu

4.5.2 Model Evaluation

Thirty averages of 10-fold cross-validation were calculated and the model's performance was tested on a held-out test dataset. The CV RMSE and R^2 where 81.44 and 0.25. The test dataset CV RMSE and R^2 were 79.90 and 0.30.

This better than the Decision Tree and worse then the Random Forest regressor.

4.5.3 Model Interpretation

As MLP models don't have an intrinsic feature importance metric, only Permutation Importance will be calculated. For more details on this, see the discussion in the Random Forest model section of this document. Below a chart of the results.

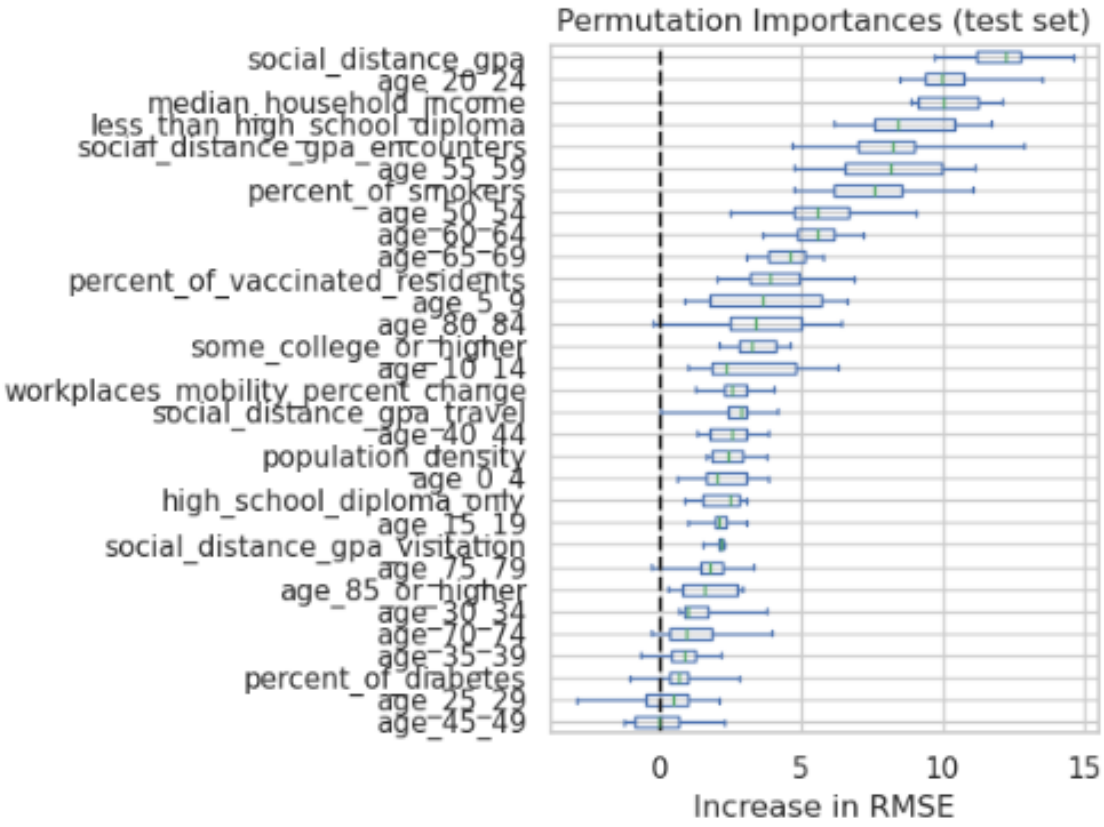

Figure 14: Permutation Importance

5 Summary of Results and Conclusions

Per analysis above, the OLS model rejected for the purposes of feature importance due to its unreliable parameter estimates. As Permutation Importance is a better indication of feature importance on unseen data, this metric will be used to determine feature importance. Model performance is used as a indicator of model quality for this analysis. Below, is a summary the performance of models developed for this analysis.

Model	CV RMSE	CV R2	Test RMSE	Test R2
Random Forest	72.05	0.43	70.23	0.46
Linear Regression	78.48	0.32	78.99	0.32
MLP	81.44	0.25	79.90	0.30
Decision Tree	84.59	0.21	84.87	0.21

Figure 15: Performance Summary

The Random Forest model had the best performance. The permutation importance of this model's

features will be given a greater weight in determine feature importance. The other two models will be examined for overlap with the Random Forest model.

Below is a summary of the feature importance of each model.

Permutation Importance					
Random Forest		MLP		Decision Tree	
Feature	**Δ RMSE**	**Feature**	**Δ RMSE**	**Feature**	**Δ RMSE**
College Or Higher	5.70	Social Dist. GPA	12.01	**College Or Higher**	13.07
Median HH Income	4.94	Age 20 to 24	10.19	**Median HH Income**	4.60
LT HS Diploma	2.84	**Median HH Income**	10.18	**% Vaccinated**	4.51
% Vaccinated	2.05	**LT HS Diploma**	8.80	Age 10 to 14	4.39
% Diabetes	1.80	Soc. Dist. Encter.	8.37	Age 55 to 59	0.76
Social Dist. Travel	1.72	Age 55 to 59	8.12	Age 20 to 24	0.68
Age 55 to 59	1.60	% Smokers	7.56		
Age 80 to 84	1.43	Age 50 to 54	5.76		
Age 85 or Higher	1.30	Age 60 to 64	5.55		
Population Density	1.30	Age 65 to 69	4.52		

❑ Top 5 for RF appear at least one other model: **Education, Income, % Vaccinated**.

Figure 16: Permutation Importance

The top 5 import features for the Random Forest model are:

1. Percentage of Residence with College or Higher Education.
2. Median Household Income of County Residence.
3. Percentage of Residence With Less Than A High-school Diploma.
4. Percentage of Residence Vaccinated.
5. Percentage of Residence with Diabetes.

The first four of these factors have overlap with at least one of the other two models. Theses four features can be summarized a pertaining to education, income, and percentage of residence vaccinated. Therefore, per this analysis, these three factors were the most important factors determining COVID-19 mortality in U.S. counties over the period January 2020 to June 2021.

6 Further Research

Based on this analysis, below are areas requiring further research.

1. The Percentage of Residence Vaccinated is self-explanatory as to why it is one of the determinants of COVID-19 mortality. Income and education, however, are not. Further work needs to be done to determine the causative factors underlying these two proxies.
2. The best model, the Random Forest Regressor, only explained 46% of the test dataset target variable variance. A better performing model should be investigated.
3. During this research, is was fortunate to identify Permutation Importance as a way of evaluating feature importance for artificial neural networks. Another approach I did not have time to evaluate is SHAP (SHapley Additive exPlanations). Like Permutation Importance, its a model agnostic approach to determine feature importance for models and can be used for ANNs.

References

[1] American Diabetes Association, "How covid-19 impacts people with diabetes," *American Diabetes Association*. [Online]. Available: https://www.diabetes.org/coronavirus-covid-19/how-coronavirus-impacts-people-with-diabetes

[2] E. O. Arceo-Gomez, R. M. Campos-Vazquez, G. Esquivel, E. Alcaraz, L. A. Martinez, and N. G. Lopez, "The income gradient in covid-19 mortality and hospitalisation: An observational study with social security administrative records in mexico," *The Lancet Regional Health - Americas*, vol. Volume 6, 2021. [Online]. Available: https://www.thelancet.com/journals/lanam/article/PIIS2667-193X(21)00111-3/fulltext

[3] S. Justin M. Feldman and M. Mary T. Bassett, MD, "Variation in covid-19 mortality in the us by race and ethnicity and educational attainment," *American Medical Association*, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8611482/

[4] P. Oliver J Watson, M. Gregory Barnsley, P. Jaspreet Toor, PhD adn Alexandra B Hogan, P. Peter Winskill, and P. Prof Azra C Ghani, "Global impact of the first year of covid-19 vaccination: a mathematical modelling study," *The Lancet Infectious Diseases*, 2022. [Online]. Available: https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(22)00320-6/fulltext

[5] L. T. Sanjeev Nanda, "A midwest covid-19 cohort for the evaluation of multimorbidity and adverse outcomes from covid-19," *NIH: National Library Of Medicine*, 2021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33855875/

[6] M. Trang VoPham, PhD, P. Matthew D. Weaver, S. Jaime E. Hart, M. Mimi Ton, P. Emily White, and P. Polly A. Newcomb, "Effect of social distancing on covid-19 incidence and mortality in the us," *NIH: National Library Of Medicine*, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7310657/

[7] World Health Organization, "Who statement: Tobacco use and covid-19," *World Health Organization*, 2020. [Online]. Available: https://www.who.int/news/item/11-05-2020-who-statement-tobacco-use-and-covid-19