

**BUSS4933 Advanced Machine Learning for**

**Business Insights**

**(2022S2)**

**Term Project**

Due date: Saturday 12 October 2022

Student ID: 490465932

Predictors of COVID-19 Pneumonia

In December 2019, the Coronavirus Disease 2019 (COVID-19) began its rapid spread worldwide, becoming the fifth documented pandemic throughout history (Moore, 2021). Despite many skepticisms, its asymptomatic and contagious incubation period allowed it to spread effectively in a series of surges (Achenbach, Cha, & Sellers, 2021; Maragakis, 2021). By the end of 2020, there were almost 2 million COVID-19 deaths reported, with an estimate of at least 3 million excess mortalities (World Health Organization, 2021).

Being a significantly pro-inflammatory condition, one of the common mechanisms that COVID-19 kills is through instigating pneumonia as a lung complication (Galiatsatos, 2022). In 2022, about 15% of people with COVID-19 develop serious complications, with 75% being pneumonia (Wiersinga, Rhodes, Cheng, Peacock, & Prescott, 2020; Cleveland Clinic, 2022).

Pneumonia is a respiratory infection that causes the air sacs in one or both lungs to fill with pus and other liquids (Johns Hopkins, n.d.). Although most people respond well to pneumonia treatment, it has the potential to be life-threatening. In particular, for people with impaired immune systems such as elderlies, children, or patients with pre-existing conditions, pneumonia tends to lead to complications (Johns Hopkins, n.d.).

Unlike the standard pneumonia, pneumonia caused by COVID-19 are often severe which are often associated with very high mortality. This is because, it uses the individuals’ immune system to spread, allowing it to last longer and cause widespread damage in multiple areas of the body (Cleveland Clinic, 2022). It also often impairs both lungs, limiting one’s ability to take in oxygen causing agonizing symptoms, such as shortness of breath and coughing, alongside lasting lung injuries that persists months after its recovery (Galiatsatos, 2022).

Unfortunately, a significant proportion of COVID-19 pneumonia patients are presented with a lack of certain symptoms. For example, a significant proportion of COVID-19 patients do not exhibit dyspnoea, despite it being a common predictor of regular pneumonia. This insidious nature of COVID-19 pneumonia worsened prognosis by increasing the difficulty of disease diagnosis, resulting in a sudden increase in need for high-intensity care (Goyal, et al., 2021). This is a problem as a substantial increase in severe cases can overwhelm less prepared hospitals, which doubles the mortality rate of COVID-19 pneumonia from 20% to 40% (Cleveland Clinic, 2022). To exacerbate the issue is the fact that resource-constrained environments are now common due to sudden surges of cases and poor economics. All of these factors associate COVID-19 pneumonia with severe disease development and high mortality (Mahendra, Nuchin, Kumar, Shreedhar, & Mahesh, 2021).

To further complicate the issue, patients without dyspnoea may instead present with other symptoms of hypoxia, such as severe fatigue, exertional fatigue and/or altered mental status (Goyal, et al., 2021). Ultimately, the great difficulty of diagnosis identifies a necessity in discerning symptoms specific to COVID-19 pneumonia to help with diagnosis and reduce its mortality rate.

Logically, to help with diagnosis, we need to identify explainable risk factors associated with COVID-19 pneumonia, especially latent ones. By identifying risk factors, healthcare workers can better predict and prevent the development of pneumonia by employing early interventions. Doing so is more effective than treating severe COVID-19 pneumonia after it arises and is advantageous in conserving resources. Being aware of more risk factors can also further inform the general public on how to keep themselves safe through the understanding of to adopt and avoid. For example, smoking is a known risk factor because it damages the lungs which increases the risk of COVID-19 pneumonia, hence not smoking the risk of COVID-19 pneumonia decreases (Cleveland Clinic, 2022). Accordingly, the differing and deadly nature of COVID-19 pneumonia stresses the importance of exploring risk factors to increase its diagnosis and reduce its mortality rate.

**1 Exploratory Data Analysis (EDA)**

*1.1 Dataset*

The dataset used was from the ‘COVID-19 Case Surveillance Restricted Use Detailed Data’. It is a database of patient-level data recorded in the United States (US) and uploaded on a monthly basis from April of 2020 till the current date November 2022. Despite an incomplete download the entire database, 158 of the files were downloaded and unzipped. The downloaded files contained 1,433,415,248 observations in total. The dataset contained at most 33 features which can be found in Appendix A, with some of the earlier files missing one or two features.

The incompleteness of the dataset is due to the ‘Unknown’, ‘Missing’, and ‘NA’ values which varies by jurisdiction and time period. The value “Unknown” represents when jurisdictions specify that the value is unknown, the value “Missing” represents when jurisdictions do not provide a value, and the value “NA” represents when the value is suppressed as part of privacy protections. For the sake of simplicity, these three values are put under one class ‘Unknown’ for all predictors.

As ‘pna\_ya’ is the target variable, the ‘Unknown’, ‘Missing’ and ‘NA’ values in this feature are removed as they do not provide any useful information. Since there is an ample amount of data, it was assumed that dropping these observations would not lead to any significant impact. Accordingly, out of 1,433,415,248 observations, 1,337,516,466 observations were dropped, which leaves 95,898,782 (6.69% data remaining).

Due to computational power and time, only 3 files with removed observations were used for modelling. The three files were:

1. COVID\_Cases\_Restricted\_Details\_03312021\_Part\_1.parquet
2. COVID\_Cases\_Restricted\_Details\_03312021\_Part\_2.parquet
3. COVID\_Cases\_Restricted\_Details\_03312021\_Part\_3.parquet

These three files represented the data collected during March of 2021. It is important to understand the context around this time in the US to better handle the dataset. For example, earlier in January 2021, vaccines were made available for first responders and individuals 65 and older (macmillan learning, 2022). This may indicate that whether the patient is a healthcare worker would not be a likely predictor of COVID-19 pneumonia as they are unlikely to get COVID-19 pneumonia due to being vaccinated.

These files were combined into one data frame containing 1515261. The dataset also contained all 33 features, however ‘race’ and ‘ethnicity’ was merged into ‘race\_ethnicity\_combined’ which resulted in 32 features. Further feature reduction was conducted by removing all 4 datetime data: ‘cdc\_report\_dt’, ‘cdc\_case\_earliest\_dt’, ‘onset\_dt’, and ‘pos\_spec\_dt’. This is because the features lack value as they only indicate the initial date, but not the duration.

A number of predictors was also removed due to target leakage, which included: hosp\_yn, death\_yn icu\_yn, mechvent\_yn. This is because the pattern within these predictors likely has a causal relationship with the target variable, however in the wrong direction. For this reason, these predictors do not provide healthcare worker with valuable information in predicting COVID-19 pneumonia.

*1.2 Pneumonia*

The target variable ('pna\_yn') is heavily imbalanced, with 'No' making up 93% of the training observations Figure 1. Accordingly, of 1,015,224 observations, only 6.7% of patients presented with pneumonia.

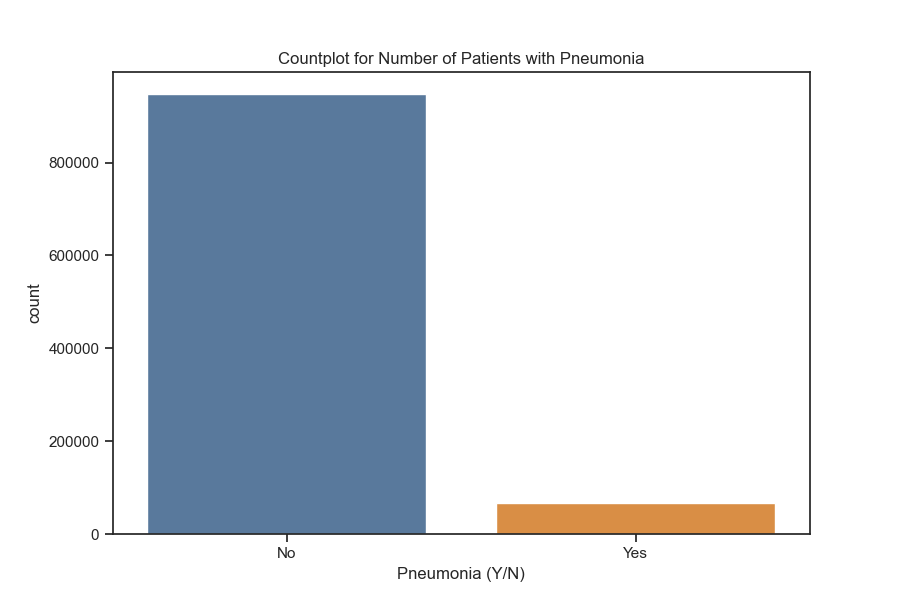


Figure : Pneumonia Countplot

The low proportion of patients with pneumonia was likely because the most prominent variant of COVID-19 in March 2021 was the Omicron. While Omicron was notoriously known to spread quickly by evading the immunity provided by prior infection and vaccination, they do not pose a greater risk of death and severe illness, such as pneumonia. Although it does not provide immunity, vaccinations are also known to reduce the risk of pneumonia (Kelly, 2022; Ito, Kitahara, Miwata, Okimoto, & Takafuta, 2022). Pneumonia was present in 78& unvaccinated patients but only 41% of fully vaccinated patients (Schiebler & Bluemke, 2022). Accordingly, since vaccinations became available to those most vulnerable, which are healthcare workers and elderlies, it makes sense that the rate of pneumonia was lower within the current dataset.

*1.3 Predictors*

Of the remaining 24 predictors, 19 was evaluated to be important based on exploratory data analysis (Appendix B). For each feature, a univariate and bivariate visualization was conducted, before a collective dependence measure was taken. Notable findings have been listed below and a summary table was provided in Appendix D.

1.2.1 Imbalanced classes

Based on univariate analysis, it was found that a large number of features is heavily disproportionate. For example, the pie chart of current\_status in Figure 2 that there is a ratio of 2:23 between its two classes. Without considering ‘unknown’ and ‘other’ classes, ‘current\_status’, ‘race\_ethnicity\_combined’, ‘hc\_work\_yn’, ‘acuterespdistress\_yn’, ‘fever’, ‘sfever\_yn’, ‘runnose\_yn’ ‘sthroat\_yn’, ‘sob\_yn’, ‘nauseavomit\_yn’, ‘abdom\_yn’, and ‘diarrhea\_yn’. Accordingly, 12 out of the 24 predictors were found to have imbalanced class.

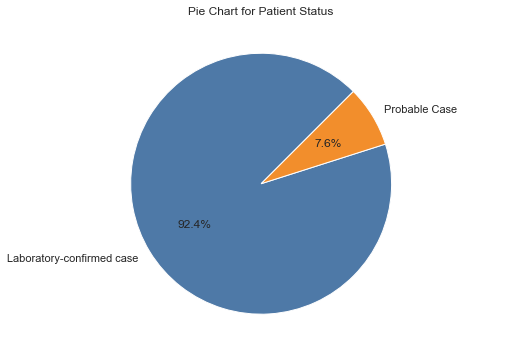


Figure : Univariate pie chart of current\_status

1.3.2 No or week relationship with the target variable

A univariate analysis reveals that the classes in race\_ethnicity\_combined are heavily disproportionate. As seen in the pie chart in Figure 3, starting at ‘White, Non-Hispanic’ and going counter-clockwise, the size of each classes becomes smaller and smaller.

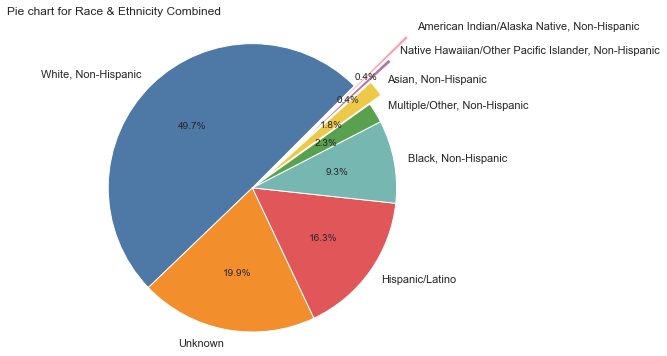
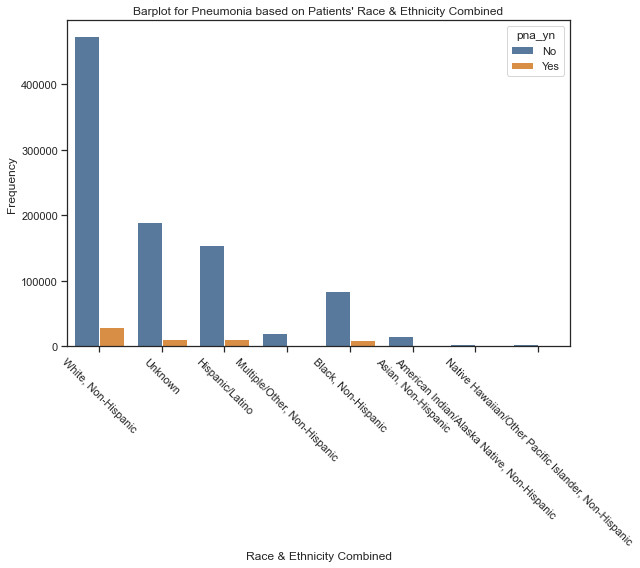


Figure : Race & ethnicity pie chart

Interestingly, a bivariate analysis reveals that the frequency of pneumonia increases with the frequency of the race & ethnicity. This indicates that despite ethnicity, patients have the same rate of obtaining COVID-19 pneumonia. Thus, race\_ethincity\_combined is likely a bad predictor of pna\_yn, due to their lack of relationship.



Similar patterns that indicate a lack or weak relationship with the target variable was found in a number of other variables. These include: ‘hc\_work\_yn’, ‘headache\_yn’, ‘runnose\_yn’, ‘current\_status’, ‘sthroat\_yn’, ‘abdom\_yn’, ‘diarrhea\_yn’, ‘cough\_yn’, and ‘nauseavomit\_yn’. As follows, these variables likely have a weak or no relationship with the target variable. That said, although these variables seemingly have no relationship with the target variable, they were kept as a part of the training data in case they may have a latent relationship with other variables which aids them in predicting the target variable.

1.3.3 Age Group

There seems to be a relationship between age\_group and pneumonia. Despite being normally distributed, as found in the age\_group barplot in Figure 4. A bivariate bar plot of age\_group with consideration of pneumonia in Figure 5 demonstrates that rate of patients obtaining pneumonia increases with age, despite the distribution of the patient age. Hence, age\_group is likely a discriminating feature of pna\_yn.

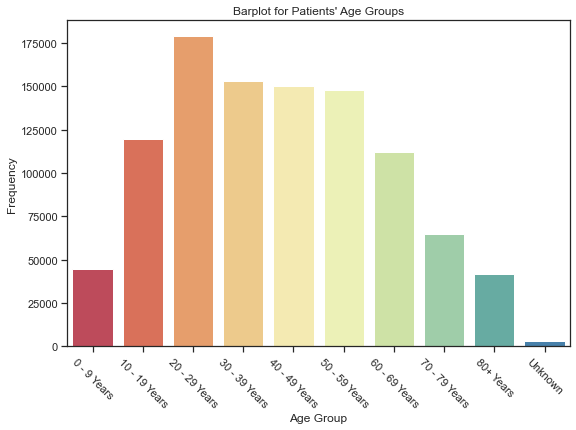


Figure : Univariate barplot of age\_group

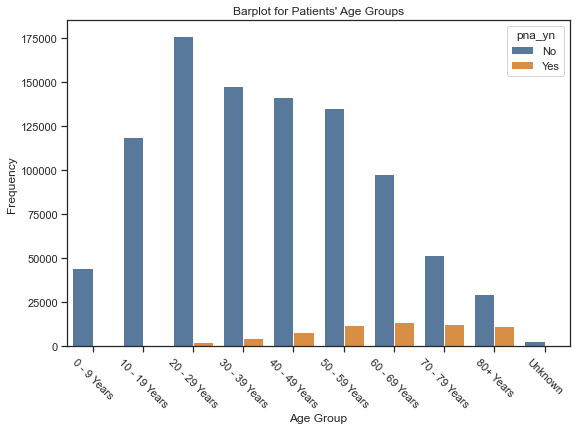


Figure : Bivariate barplot of age\_group and pna\_yn

*1.3.4 Strong relationship with the target variable*

Based on bivariate visualization using barplots, there was a number of variables which demonstrated relationship with pna\_yn, such as abnormal chest x-ray in Figure 6. Obtained from the contingency table in Figure 7, despite the small percentage of patients with an abnormal chest x-ray (3.99%), a large proportion of them had pneumonia (84.69%). A proportion larger than those who did not have abnormal chest x-ray but had pneumonia (1.38%). For this reason, it was determined that

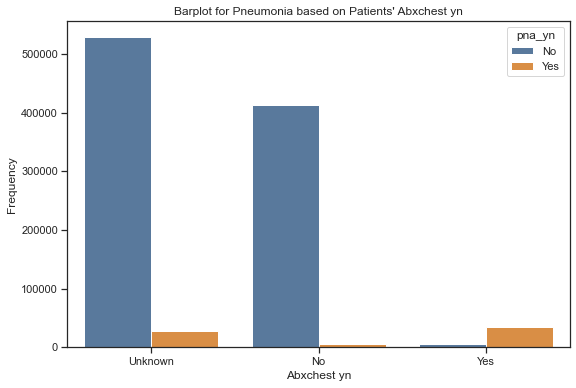
**

Figure : Bivariate bar graph for pneumonia and abnormal chest x-ray

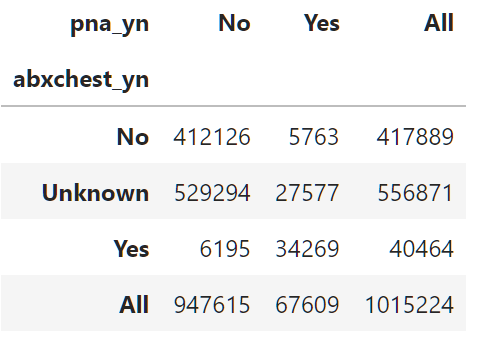
**

Figure : Contingency table between pneumonia and abnormal chest x-ray

A similar pattern was found in other variables including: *‘abxchest\_yn’, ‘acuterespdistress\_yn’, ‘fever\_yn’, ‘sfever\_yn’, ‘myalgia\_yn’, ‘medcond\_yn’, and ‘sob\_yn’. Accordingly, it was also evaluated that these features likely have a strong relationship with the target variable.*

*1.3.5 Requires further analysis*

There were a number of variables which requires further EDA. Univariate and bivariate visualization was difficult to analyze for ‘myalgia\_yn’, ‘diarrhea\_yn’, and ‘sfever\_yn’. Thus, their relationship with the target variable is unclear. As for ‘chills\_yn’ and ‘res\_county’ they lack any visualization. Therefore, there relationship and class distribution are unknown. Accordingly, the dependence measure in the following section would provide a clearer picture of their relationship with the target variable.

*1.4 Dependence measures (PhiK)*

To obtain an understanding of the degree of correlations between the variables, the correlation coefficient Phik was calculated. Similar to Cramer’s , the metric meant for two categorical variables and also based on Pearson’s chi-squared contingency test. It is able to capture non-linear dependencies and less sensitive to outliers compared to Cramer’s . The values of Phik are ranges between 0 and 1, where 0 denotes no correlation and 1 denotes a perfect correlation (Lewinson, 2021). Additionally, to validate the significance of the correlation, a measure of their statistical significance was taken using G-test with a p-value threshold of 0.05 (Baaka, Koopman, Snoek, & Klous, 2019).

For the current dataset, the calculated pair-wise Phik coefficients are shown in Figure 8, which were all determined to be statistically significant (p-value>0.05) as found in Appendix .

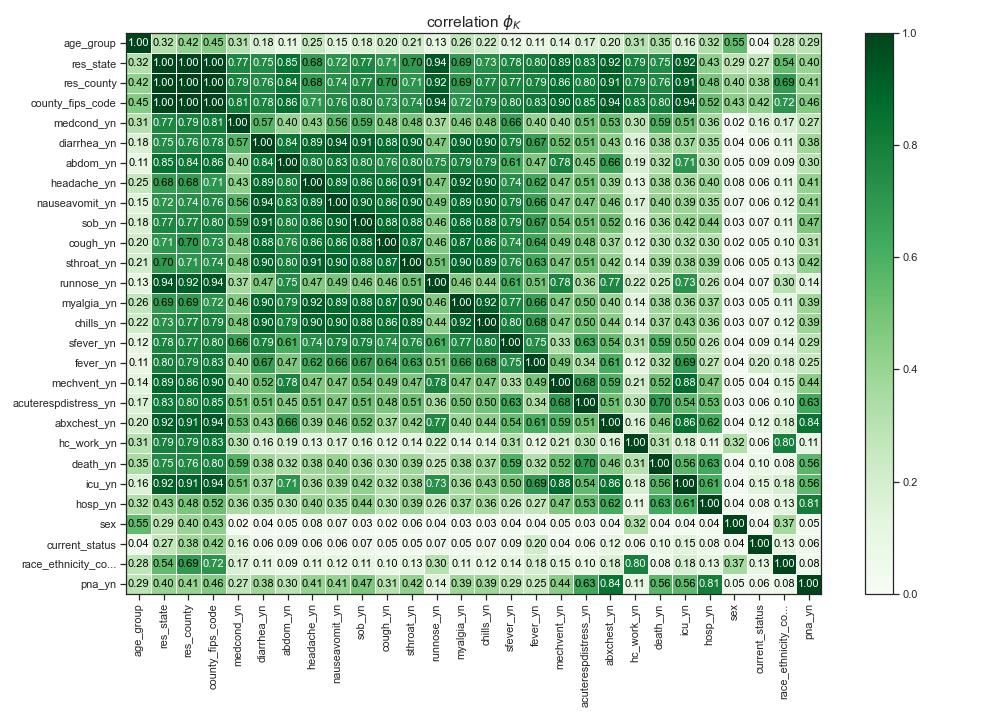


Figure : Heatmap of Phik pair-wise correlation matrix

There is a perfect correlation between ‘county\_fips\_code’, ‘res\_county’, and ‘res\_state’. Not only do they have a correlation of 1 with each other, but the exhibit very similar relationship with other features. Since county\_fips\_code exhibit very high cardinality (46309 classes) compared to res\_county (901 classes) and res\_state (56 classes), ‘county\_fips\_code’ was dropped to prevent multicollinearity and conserve computational time. On the other hand, ‘res\_state’ was dropped as it exhibit perfect collinearity with ‘res\_county’, however ‘res\_county’ was able to provide more refined geographical information. A more refined information provides greater predictive value and information of the ‘res\_state’ may be derived from ‘res\_county’, but not the other way around.

Although PhiK matrix demonstrates that there are a number of other features which exhibits low correlation with pna\_ya, they will be kept. These variables include ‘current\_status’, ‘sex’, ‘race\_ethnicity\_combined’, ‘runnose\_yn’, and ‘hc\_work\_yn’. This is in case they provide latent information and so will be removed later on depending on feature importance analysis.

**2 Feature engineering**

*2.1 ‘Unknown’ class*

To deal with all three variants of missing values, an ‘unknown’ class was created to denote missingness. This is because the ‘missingness’ can provide valuable information, if the values are not missing by random. Due to time constraint, it was assumed that all ‘Unknown’, ‘Missing’, and ‘NA’ values were missing not at random. Then they were merged into one big ‘Unknown’ class by replacing all 'Missing' and ‘NA’ values with ‘Unknown’.

*2.2 Categorical encoding*

The majority of features which exhibits the classes 'yes', 'no' and 'unknown’ were label encoded. The 'res\_county' feature was target encoded because it had a high number of classes. The 'age\_group' feature was manually encoded by taking the mean of each bins, with the exception of the '80+' bin which was converted into '85' for the sake of uniformity. Encoding methods such as one-hot encoding was avoided due to the high number of features, which will lead to a very large dataset. Target encoding was avoided so as to avoid overfitting. A summary of the encoding method used can be found in Appendix C.

*2.3 Feature scaling*

Feature scaling is important for the two models implemented logistic regression (LR) and Linear Support Vector Classifier (SVC).

Since LR uses Gradient Descent during optimization, feature scaling is important because it speed up the convergence process. Additionally, the logistic regression algorithm tends to assume that features with greater magnitude are more relevant. Thus, scaling these features into the same range allows the model to truly evaluate the best features based on their predictive value, regardless of their relevance (Bhor, 2022).

SVC uses the distance between data to create a decision boundary. Accordingly, it is important to scale the features so that the mapping of the values is within the same range. This allows for the SVC to improve the classifier performance significantly (Tokuç, 2022).

Two main methods of feature scaling are standardization and normalization; however, standardization was chosen for the current task. Standardization shifts the distribution of each feature so that they have a mean of 0 and standard deviation of 1. Standardization was chosen because compared to normalization, it centers the data and tends to train more generalizable models. To scale the features, the transformation in Equation 1 was applied, where denotes an observation, is the mean, and is the standard deviation of the feature values (Tokuç, 2022).

*2.4 Train Test Split*

The dataset was split into a training set with the test set making up 33% of the entire dataset. There was no split for validation set as cross-validation (cv) was used for hyperparameter tuning.

**3 Methodology**

Two main algorithms were selected for the current dataset: Logistic Regression (LR) and Linear Support Vector Classifier (LinearSVC). These models were hyperparameter tuned and variants of these models were created based on further feature selection. The models were evaluated based on the f1-score.

*3.1 Evaluation metric*

The f1-score (f1) was chosen as the evaluation metric for the current project. It is singular a measure of accuracy in the statistical analysis of binary classification (Wikipedia, 2022). It evaluates the harmonic mean between precision and recall measured by Equation 2 (Wood, n.d.).

Based on equation 2, the f1-score was derived from a confusion matrix. The score is limited to the range of 0 to 1, where 0 indicates no precision or recall and 1 indicates perfect precision and recall (Wikipedia, 2022).

The f1-score was chosen in particular by considering the dataset and the problem at hand. As found in the EDA, a large proportion of the features are imbalanced. Moreover, since the current project aims to help health-care workers, there is an imbalance in the cost based on the type of false predictions. Compared to false positive which at most will stress out parties involved; whereas a false negative may result in death. Hence, the f1-score was chosen in particular because they are able to handle a large class imbalance and because they are robust to differing cost of false positive or false negatives.

*3.2 Logistic regression*

3.2.1 Definition

A logistic regression is a supervised learning algorithm used in binary classification {0,1}. It models the relationship between features to predict the probability the input belongs to the default output class (Brownlee, A Tour of Machine Learning Algorithms, 2020). It does so by modelling the relationship between variables using a linear and then transforming its predictions into probability outputs using a logistic function (Brownlee, Logistic Regression for Machine Learning, 2016).

Given features , the relationships between features are modelled in a linear equation denoted by Equation 3.1.

where represents the intercept and is the coefficients of features .

The output of the linear equation is transformed into a probability using the logistic (aka sigmoid) function in Equation 3.2, where denotes the probability that the input belongs to the default class 0.

To classify the output value of the sigmoid function, a probability threshold is set. Given a probability threshold of 0.5, if is greater than 0.5, the input will be classified as 0, otherwise 1.

A logistic regression algorithm was chosen to model the current dataset due to a number of reasons. Most importantly, it is a classification algorithm that predicts a binary outcome which is suitable for the current project. It is also highly interpretable as it helps determine the probabilities and its direction between any two classes by predicting the likelihood of an event (Joby, 2021). Due to its simplicity, it is effective to train and easy to implement. Its model coefficient may be used as indicators of feature importance. Regularization techniques may also be implemented to avoid over-fitting (Rout, 2022)

3.2.2 Hyperparameter tuning

To enhance the performance of the logistic model, two hyperparameters model were tuned. The ‘penalty’ which determines its regularization method and its ‘solver’ which determines how the algorithm optimizes the problem. The two parameters were tuned using a randomized search method and stratified k-fold cross-validation where . Notably, a stratified k-fold cross validation was implemented due to the imbalanced nature of the dataset. This is so that the training and validation set exhibits the same imbalanced as the test data so that the validation f1-score returned will be similar to the test f1 score.

Given the two penalty options {‘l1’ and ‘elasticnet’} and two solver options {‘sag’ and ‘saga’}, ‘l1’ and ‘saga’ was evaluated to be the best hyperparameters. An ‘l1’ penalty means that a penalty term equal to the absolute value of the magnitude will be added, allowing it to limit the size of its coefficient. With the ‘l1’ penalty, the algorithm is also known as the lasso regression (Glen, n.d.). Whereas a ‘saga’ solver means that the model uses utilizes a variant of the stochastic average gradient method which is able to support the ‘l1’ penalty (Yahya, 2018). Accordingly, a hyperparameter tuning found that the best logistic regression for the current dataset implements a lasso penalty term and utilizes a variant of the stochastic gradient method for the problem optimization.

3.2.3 Models

Firstly, a benchmark LR model was created by training the model using the entire dataset with the best hyperparameters - this model is denoted as LR (all). Since this model exhibits target leakage a following LR model was trained using the reduced 19 features and denoted as LR (eda). To analyze the model its feature importance was plotted and its regularization path was graphed.

The importance of each feature was derived from its coefficient which can be found in Figure 9. This is because the magnitude of the coefficient demonstrates its contribution towards predicting the probabilistic output. By this logic, features with low coefficients may be removed as they do not contribute in predicting the final output.

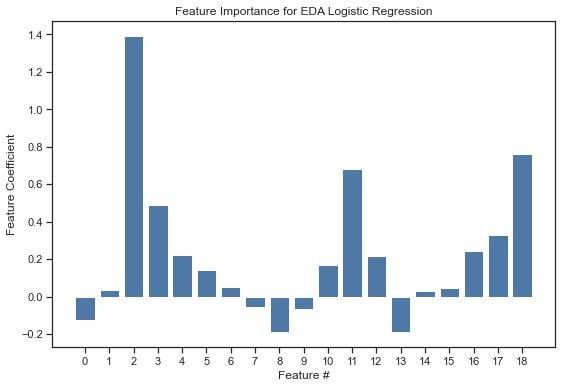
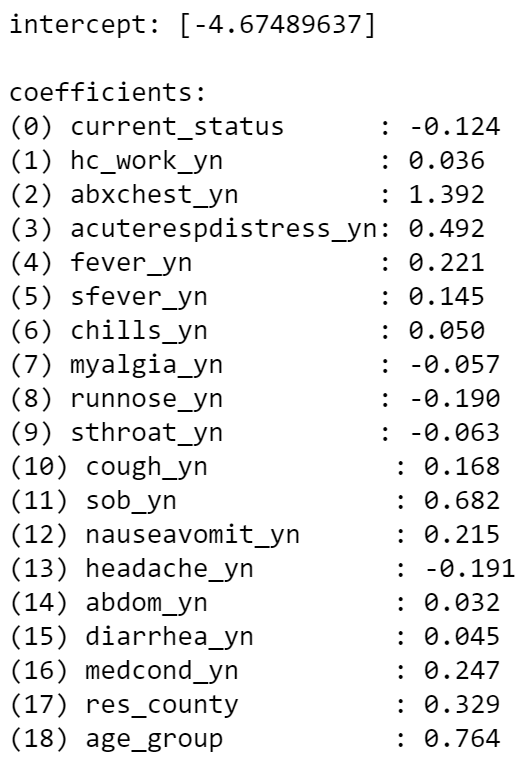
 

Figure : Feature importance of LR (eda)

Based on Figure 9, the feature importance of the model may be distinguished. Since the current report aims to identify crucial predictors to aid health-worker identify COVID-19 pneumonia, it is important to have a limited but robust number of features for health-care workers to remember. For this reason, a threshold of 0.100 was selected as the cutoff point for feature reduction based on feature importance. In the order of importance derived from the table in Sorted coefficients in LR (eda) in comparison to relationships derived in EDA, the 5 most influential features include: ‘abxchest\_yn’, ‘age\_group’, ‘sob\_yn’, ‘acuterespdistress\_yn’, and ‘res\_county’. All 5 features have a coefficient greater than 0.300. The next following features are ‘medcond\_yn’, ‘fever\_yn’, ‘nauseavomit\_yn’, ‘cough\_yn’, ‘sfever\_yn’, and ‘chills\_yn’. In an attempt to reduce features, further exploration will be conducted by training a model using these 10 identified predictors.

As a side note, the feature importance seems to reflect the relationships evaluated based on the EDA (Appendix F). Accordingly, future work may train the model based on only the features determined to have a strong relationship with the target variables. However, this was not completed within the current project.

Following the feature importance, a regularization path was plotted to visualization of how the lasso regularization term of the logistic regression behaves. Found in Figure 10, the 19 coefficients are plotted against the weight vector of the regularization term . is a sklearn hyperparameter that denotes the strength of the penalty term. For example, will lead to a sparser solution than , therefore a will result in less coefficients. Accordingly, from the left-hand side to right-hand side of the x-axis, the models are ordered from strongest regularized to least regularized (Gramfort, n.d.).

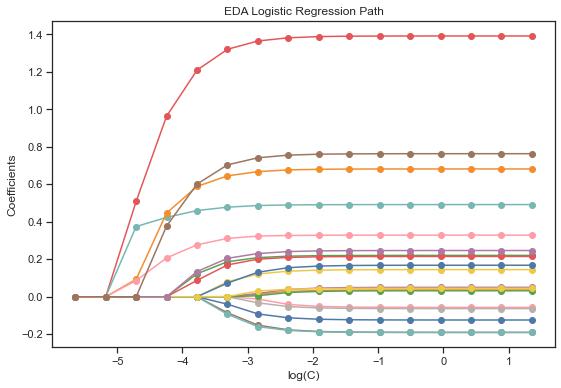


Figure : Regression path for LR (eda)

Based on Figure 10, when is too small, the LR (eda) could not fit the coefficients to any features. However, when C is too high, too many unnecessary coefficients were fit into the model. Interestingly, at approximately before and after , the

Based on the regularization path of the LR (eda) model, there was a notable change in the number of predictors at approximately before and after . When , the highest number of features was 5, but the next growth in feature when resulted in the addition of 3 features simultaneously but at a relatively small magnitude. With relatively low coefficient it was determined that these features do not pose great importance. The following features added had even lower coefficient, emphasizing their lack of predictive value. To achieve the balance in the trade-off between complexity and interpretability, a was manually chosen. As further exploration, a model will also be created with 5 features chosen based on feature importance.

3.2.4 Further feature reduction

A further feature reduction was explored to obtain the most defining identifier for COVID-19 Pneumonia. Two models were explored based on conservative and non-conservative manual and naïve feature selection. For the non-conservative feature selection, 5 features were retained as per the regression path analysis. This model was denoted as LR (non-conservative). Its feature importance graph and coefficient may be found in Figure 11.

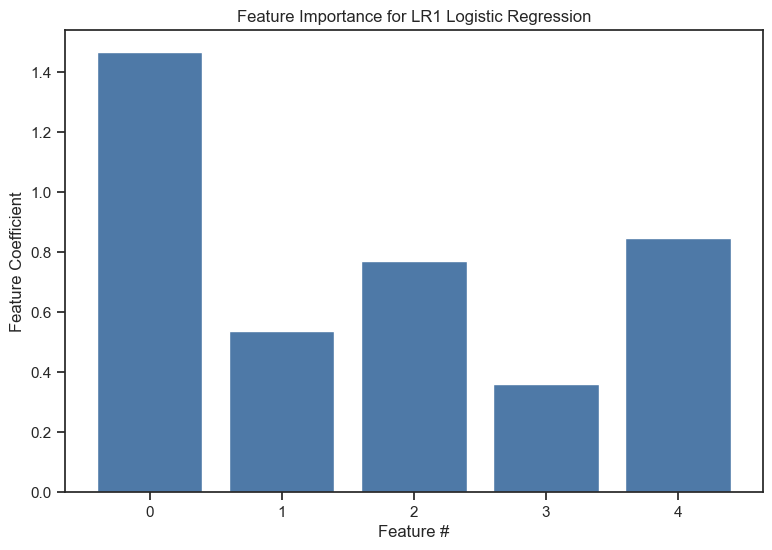
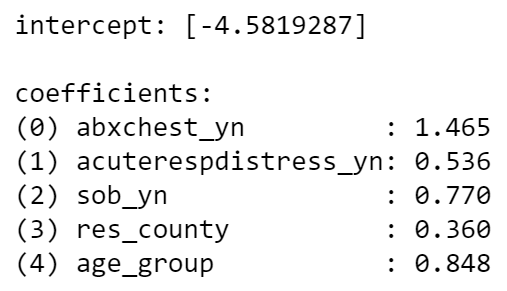
 

Figure : Feature importance and coefficients of LR (non-conservative) model

For the conservative feature selection, 10 features were retained. This is by setting a minimum threshold of 0.1 for coefficients. Accordingly, a feature was kept if its associated coefficient was greater than 0.1 ), otherwise it was removed. This model was denoted as LR (conservative). Its feature importance and coefficient may be found in Figure 12.

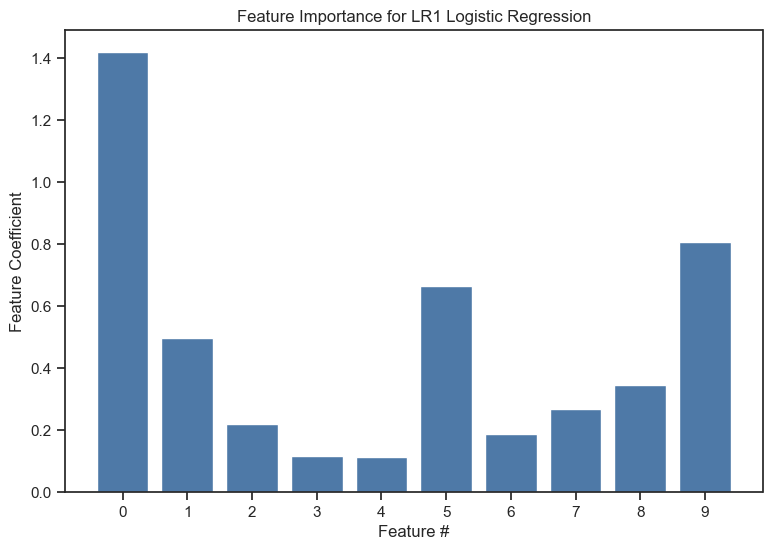
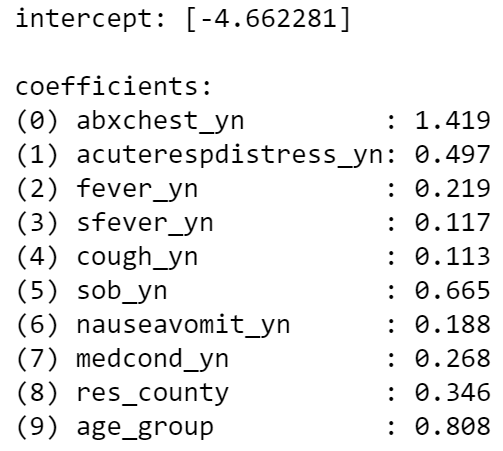
 

Figure : Feature importance and coefficients of LR (conservative) model

Model

*3.3 Linear Support Vector Classifier*

3.3.1 Definition

Linear support vector classifier (Linear SVC, but will be denoted as SVC) is a class of support vector machines capable of performing binary classification (sckit-learn developers, n.d.). Support vector machines are a set of supervised learning method which attempts to separate the different classes of the target variable using hyperplanes. The hyperplanes are calculated based on the predictors. To optimize the model, the distance between the hyperplanes to the nearest data point are maximized. An example of its results based on 2 features may be found in Figure 13.

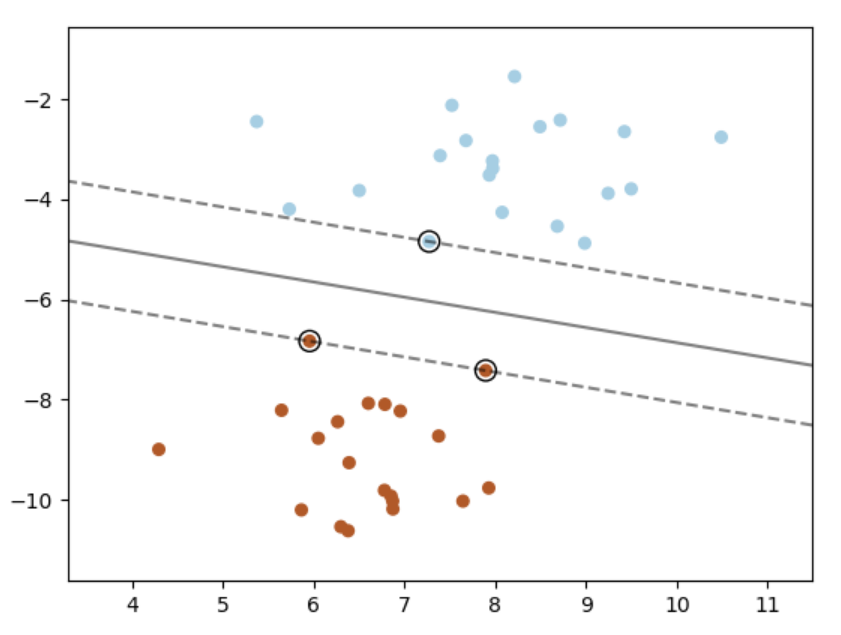


Figure : Sample Support vector machine decision boundary (hyperplane in 2-dimensions)

SVC was the second machine learning algorithm chosen to model the current dataset due to a number of reasons. Primarily, the target variable is a categorical and thus require a classification method. SVC are able to work effectively in high dimensional space. It is memory efficient; however, this is because it only utilizes a subset of training points called support vector, therefore requires a large number of feature reduction. That said, this aligns with the current project as filtering out the important features is beneficial for applications (sckit-learn developers, n.d.). For these main reasons, Linear SVC was chosen as the second model for the current dataset.

3.3.2 Hyperparameter tuning

To obtain the best performing SVC model, the ‘penalty’ hyperparameter was tuned. Between ‘l1’ and ‘l2’ regularization, a grid search stratified 5-fold cross-validation found that an ‘l1’ penalty term provided a better f1-score. The use of ‘l1’ penalty term leads to sparse vector, meaning that it will indirectly conduct feature selection.

3.3.3 Model

First a benchmark model, which will be denoted as SVC (all), was created by training the SVC on all features. Then a model denoted as SVC (eda) was trained by the reduced 19 features based on EDA. Its top features were visualized in Figure 14. It may be seen that there is an excessive number of features used and a large proportion did not have much contribution.

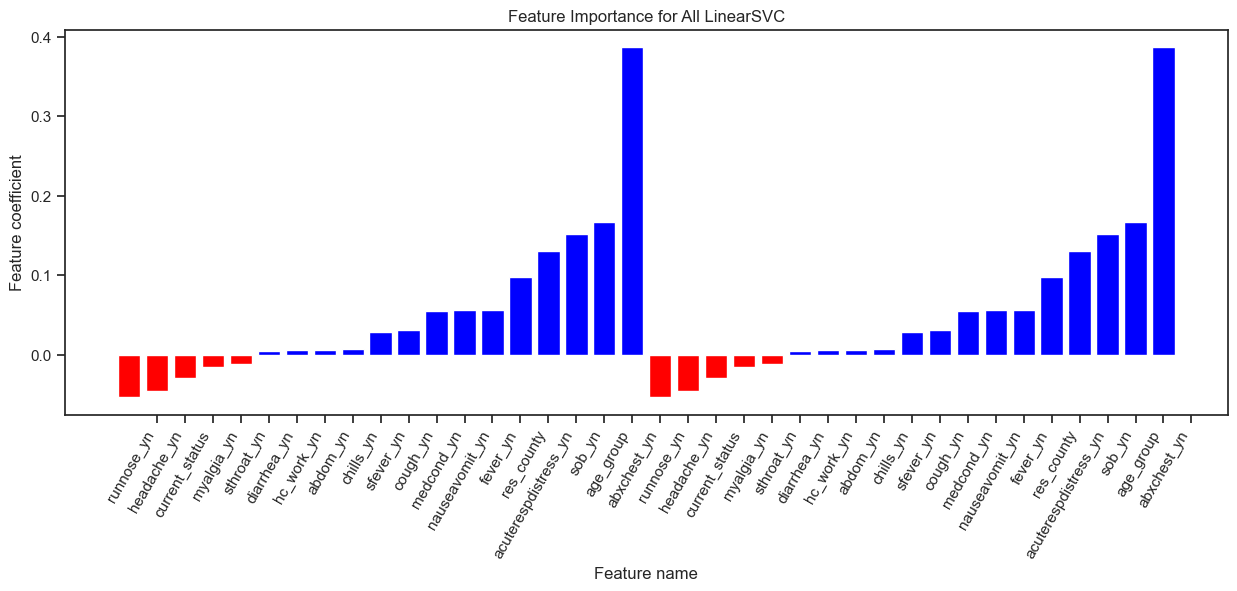


Figure : SVC (eda) top Features

To conduct feature reduction, features with coefficients within the range of [-0.01, 0.01] were dropped. This included ‘hc\_work\_yn’, ‘chills\_yn’, ‘abdom\_yn’, and ‘diarrhea\_yn’. In retrospect, a larger number of features should have been prune; however, at the time the predictive power of the model was of concern. Accordingly, if future work were to take place, it is recommended to drop coefficients with values ranging between [-0.1, 0.1] instead.

3.3.4 Further feature reduction

After the removal of the 4 features, the remaining features was used to train another SVC model, denoted as SVC (model) model. Its feature importance was graphed in Figure 15.

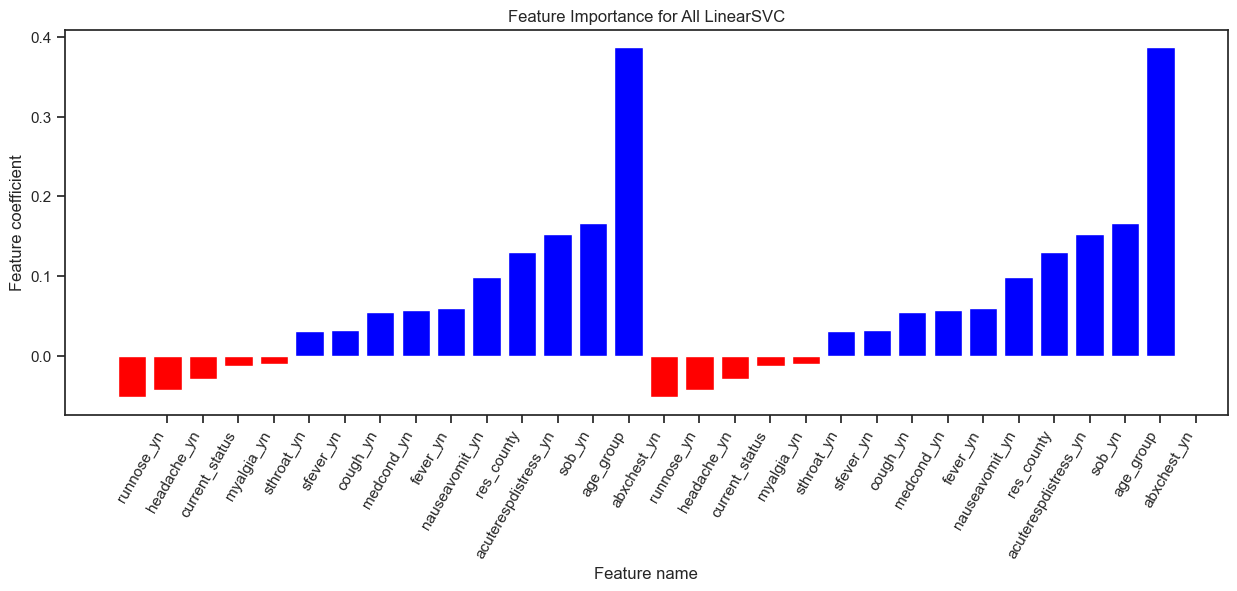


Figure : SVC (model) top features

The feature importance of SVC (model) and SVC (eda) returned very similar results. This is a good indication that removing these variables did not have a significant impact on the model. Hence, they were unnecessary variables with low predictive value. As previously mentioned, an attempt to further reduce the features should have been made. That said it is important to note that naively removing features based on coefficients may not be the best techniques and there are other methods of better selecting features to remove in future work.

**4 Results**

*4.1 Experiments*

|  |  |  |
| --- | --- | --- |
| **Model** | **F1 valid** | **F1 test** |
| LR (all) | 0.747 | 0.751 |
| LR (eda) | 0.711 | 0.716 |
| LR (non-conservative) | 0.700 | 0.704 |
| LR (conservative) | 0.709 | 0.713 |
| SVC (all) | 0.742 | 0.747 |
| SVC (eda) | 0.699 | 0.703 |
| SVC (model) | 0.698 | 0.702 |

Figure : Table of f1-scores of the different models

Evidently, models that were trained based on all features (LR and SVC all) resulted in better predictive performance. However, these models are not valid as they utilize predictors which will cause target leakage. Of the two models based on features selection derived from domain knowledge, visualization, and statistical parameters (LR and SVC eda), SVC (eda) was able to better predict whether patients have COVID-19 pneumonia.

For both LR and SVC, a further feature reduction was conducted based on feature importance. For LR a non-conservative LR (non-conservative) and conservative approach was taken LR (conservative). While the non-conservative approach found a decrease in predictive power, the conservative approach returned relatively similar results as LR (eda). Due to the high cost associated with false negative predictions, it was determined that the LR (conservative) had better predictive value than the LR (non-conservative) model.

Similarly, a further reduction based on feature importance was conducted for the SVC model (SVC model). After further feature reduction, its predictive performance of the SVC model did not change. Accordingly, between SVC (eda) and SVC (model), SVC (model) was determined to be better due to exhibiting a more conservative support vector as well as being more time efficient.

Between the further reduced model SVC (model) and LR (conservative), LR (conservative) was determined to be better. This is because not only did it return a higher f1-score, but it utilized less features. This is important as it allows health-care workers to focus on remembering and identifying a small but robust number of indicators for COVID-19 pneumonia.

Therefore, the best model was determined to be the LR (conservative). This is because both model with all features exhibited target leakage and are not useful for the current aim. Between the remaining SVC models and LR models, the LR models were better due to its smaller pool of features and lower computational time. Ultimately, LR (conservative) was determined to be the best as it has the greatest predictive power which is important in the health setting, alongside a decreased number of features which has practical advantages.

*4.2 Best model*

The best model was denoted as LR (conservative). The model implemented a lasso penalty and utilized a variant of stochastic average gradient specific for lasso penalties. To train it, 10 features were specified which includes abnormal chest x-ray, acute respiratory distress, presence of fever, claims of subjective fever, shortness of breath, nausea and/or vomit, pre-existing medical conditions, residential county, and age groups. The returned trained intercept and coefficients are presented in Equation 4.1 (Choueiry, n.d.).

To interpret the intercept, inverse logit is applied following the formula in Equation 4.2 (Choueiry, n.d.).

Accordingly, the probability that a COVID-19 patient will have COVID-19 pneumonia is 9%.

The most notable feature is the age group (age\_group). Based on the magnitude of the coefficient and bivariate visualization of the feature, it was inferred that people of old age are more likely to obtain COVID-19 pneumonia. Another notable feature is dyspnea (sob\_yn). In particular, if a patient exhibits shortness of breath, then they have a moderate chance of exhibiting COVID-19 pneumonia.

To interpret the logistic regression coefficients, they are converted into the odds ratio. Since all coefficients are positive, Equation 4.3 was used to obtain Figure 14 (Choueiry, n.d.).

|  |  |  |
| --- | --- | --- |
| **Features** | **Coefficient** | **Odds ratio** |
| abxchest\_yn | 1.419 | 4.133 |
| sfever\_yn | 0.808 | 2.243 |
| medcond\_yn | 0.665 | 1.944 |
| age\_group | 0.497 | 1.644 |
| cough\_yn | 0.346 | 1.413 |
| nauseavomit\_yn | 0.268 | 1.307 |
| sob\_yn | 0.219 | 1.245 |
| fever\_yn | 0.188 | 1.207 |
| acuterespdistress\_yn | 0.117 | 1.124 |
| res\_county | 0.113 | 1.120 |

Figure : Coefficients to Odds Ratio

The most significant odds ratio for COVID-19 patients are their chest xray. Based on Figure 14, patients with abnormal chest x-ray have approximately 4.133 times the odds of having pneumonia than patients with a normal chest x-ray (Choueiry, n.d.). Similarly, patients with subjective fever have 2.253 times the odds of obtaining COVID-19 than those who does not report any subjective fever. Interestingly, there is a greater odd of having COVID-19 pneumonia for patients with subjective fever, than a normal fever who only has a ratio odd of 1.207. It may be theorized that the data reflect the behaviour to provide more attention to patients with a fever, and thus they are less likely to develop COVID-19 pneumonia.

Alternatively, it can also be said that COVID-19 patients with pre-existing medical conditions have 94.4% more odds of having COVID-19 pneumonia (Choueiry, n.d.). Similar statements may be made with the rest of the results, such as COVID-19 patients with a cough is 41.3% more likely of having COVID-19 pneumonia.

With a validation f1-score of 0.709 and test f1-score of 0.713, the model does not exhibit any signs of overfitting. However, despite exhibiting the highest f1-score, a score of is not good enough. This is mainly attributed to the lack of transformation to deal with the imbalanced classes. Moreover, dropping missing values from the predictors likely disrupted the model performance as well. On the bright side, the feature selection process was able to effectively select the integral features for the current model which does not exhibit target leakage.

*4.3 Further application (LR non-conservative)*

Although, the non-conservative model did not return the best score, it is able to provide valuable information. With relatively similar predictive power, it was able to model the probability of COVID-19 pneumonia with a scarce number of features. Hence, healthcare workers can utilize these features on a daily basis to notice the onset of COVID-19 Pneumonia. These predictors include ‘res\_county’, ‘acuterespdistress\_yn’, ‘sob\_yn’, ‘age\_group’, and ‘abxchest\_yn’. Accordingly, notable predictors of COVID-19 pneumonia include: the patient’s county, whether they have acute respiratory distress syndrome, whether they have shortness of breath, the patient’s age, and whether they have an abnormal chest x-ray.

**5 Evaluation & Future work**

As mentioned, in the field of medicine, an inaccurate prediction has drastic consequences. For this reason, it is integral to further improve the performance of the model. There are a number of improvements possible for future work which includes feature transformation, treating missing values and outliers, and also using more training data (Ray, 2020).

The majority of predictors (12 out of 24) within the dataset are heavily imbalanced. However, the current project did not apply any techniques to create a more balanced dataset. This likely negatively impacted the performance of the model by creating a bias towards the majority class. This is a problem because the current dataset aimed to predict the minority class, the chances of patients having pneumonia (Cheruku, 2019). Accordingly, possible techniques future work may consider includes under-sampling of the abundant class or over-sampling of the minority class can be used to balance a new dataset. Accordingly, further feature engineering can be implemented to improve the model performance.

Similarly, further feature engineering can take place by handling the missing values. Due to the missing values, only approximately 7% of the dataset was used. This loss of data may have been unavoidable as these observation in particular were missing for the target feature. However, imputing missing values for the predictors may improve the performance of the model. For the current dataset in particular, ‘Unknown’ and ‘Missing’ may be imputed, it may be useful to keep the ‘NA’ values as a class as they represent classified information. This is because the ‘Unknown’ and ‘Missing’ values were missing at random and therefore calls for deletion or imputation, whilst ‘NA’ are not missing at random and so can be kept as an ‘NA’ class. Possible methods of handling categorical missing values include imputing it with the most frequent value – a crude method but timely method; otherwise by developing a model to predict missing values and imputing its prediction – a extensive but rewarding method. Hence, to improve model performance, it is recommended that future work attempts to impute ‘Unknown’ and ‘Missing’ values, whist maintaining the ‘NA’ class.

The current project takes on a very small proportion of this database, therefore it is possible to add more data in an attempt to increase the model’s predictive power. As previously mentioned, the ‘COVID-19 Case Surveillance Restricted Use Detailed Data’ database contains observations from April 2020 and continues to grow. That said, as the current world is recovering from COVID-19 both from vaccines and weaker variants existing, valuable information is more likely to be derived from data that was uploaded early on. This may also explain the moderate level of predictive power in the current model as the majority of patients by March 2021 no longer catches the more severe variants of COVID-19 which causes pneumonia. Hence, it is recommended that future work within the current field explore the data before January 2021, which is before the vaccines were rolled out.

# References

Achenbach, J., Cha, A. E., & Sellers, F. S. (2021, March 9). *A viral tsunami: How the underestimated coronavirus took over the world*. Retrieved from The Washington Post: https://www.washingtonpost.com/health/2021/03/09/coronavirus-spread-world/

Baaka, M., Koopman, R., Snoek, H., & Klous, S. (2019, March 9). A new correlation coefficient between categorical, ordinal. *Computational Statistics and Data analysis*. doi:10.48550/ARXIV.1811.11440

Bhor, Y. (2022, August 26). *Guide for building an End-to-End Logistic Regression Model*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/09/guide-for-building-an-end-to-end-logistic-regression-model/

Brownlee, J. (2020, February 24). *A Gentle Introduction to the Fbeta-Measure for Machine Learning*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/fbeta-measure-for-machine-learning/#:~:text=F2%2DMeasure&text=It%20has%20the%20effect%20of,negatives%20than%20minimizing%20false%20positives.

Brownlee, J. (2020, August 14). *A Tour of Machine Learning Algorithms*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

Cheruku, S. K. (2019, November 3). *What is imbalanced dataset and its impacts on machine learning models?* Retrieved from Linkedin: https://www.linkedin.com/pulse/what-imbalanced-dataset-its-impacts-machine-learning-models-cheruku#:~:text=The%20problem%20with%20training%20the,important%20over%20not%20having%20cancer).

Cleveland Clinic. (2022, October 8). *COVID Pneumonia*. Retrieved from Cleveland Clinic: https://my.clevelandclinic.org/health/diseases/24002-covid-pneumonia#:~:text=How%20many%20people%20with%20COVID,severe%20cases%20of%20COVID%2D19.

Galiatsatos, P. (2022, February 28). *COVID-19 Lung Damage*. Retrieved from Johns Hopkins: https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/what-coronavirus-does-to-the-lungs#:~:text=The%20pneumonia%20that%20COVID%2D19,breath%2C%20cough%20and%20other%20symptoms.

Glen, S. (n.d.). *Regularization: Simple Definition, L1 & L2 Penalties*. Retrieved from Statistics How To: https://www.statisticshowto.com/regularization/#:~:text=Penalty%20Terms&text=L1%20regularization%20adds%20an%20L1,Lasso%20regression%20uses%20this%20method.

Goyal, D., Inada-Kim, M., Mansab, F., Iqbal, A., McKinstry, B., Naasan, A. P., . . . Burke, D. (2021, October 19). Improving the early identification of COVID-19 pneumonia: a narrative review. *BMJ Open Respiratory Research, 8*(1). doi:10.1136/bmjresp-2021-000911

Gramfort, A. (n.d.). *Regularization path of L1- Logistic Regression*. Retrieved from scikit learn: https://scikit-learn.org/stable/auto\_examples/linear\_model/plot\_logistic\_path.html

ILO, FAO, IFAD & WHO. (2020, October 13). *Impact of COVID-19 on people's livelihoods, their health and our food systems*. Retrieved from World Health Organization: https://www.who.int/news/item/13-10-2020-impact-of-covid-19-on-people's-livelihoods-their-health-and-our-food-systems#:~:text=The%20economic%20and%20social%20disruption,the%20end%20of%20the%20year.

Ito, N., Kitahara, Y., Miwata, K., Okimoto, M., & Takafuta, T. (2022, May 4). Can the Omicron variant of COVID‐19 cause pneumonia in young patients without risk factors? *Clin Case Rep, 10*(5). doi:10.1002/ccr3.5684

Johns Hopkins. (n.d.). *Pneumonia*. Retrieved from Johns Hopkins: https://www.hopkinsmedicine.org/health/conditions-and-diseases/pneumonia

Kelly, P. (2022, November 8). New COVID-19 variant leads to increase in cases. *Australia Government: Department of Health and Aged Care*. Retrieved from https://www.health.gov.au/news/new-covid-19-variant-leads-to-increase-in-cases

Lewinson, E. (2021, August 9). *Phik (𝜙k) — get familiar with the latest correlation coefficient*. Retrieved from Towards Data Science: https://towardsdatascience.com/phik-k-get-familiar-with-the-latest-correlation-coefficient-9ba0032b37e7

macmillan learning. (2022, March). *Information about COVID-19*. Retrieved from macmillan learning: https://covid19.macmillanlearning.com/

Mahendra, M., Nuchin, A., Kumar, R., Shreedhar, S., & Mahesh, P. A. (2021). Predictors of mortality in patients with severe COVID-19 pneumonia - a retrospective study. *Adv Respir Med, 89*(2), 135-144. doi:10.5603/ARM.a2021.0036.

Maragakis, L. (2021, October 21). *Coronavirus Second Wave, Third Wave and Beyond: What Causes a COVID Surge*. Retrieved from Johns Hopkins Medicine: https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/first-and-second-waves-of-coronavirus

Moore, S. (2021, September 28). *History of COVID-19*. Retrieved from News Medical: https://www.news-medical.net/health/History-of-COVID-19.aspx

Park, J. Y., Freer, R., Stevens, R., Soneji, N., & Jones, N. (2021, March 4). *The accuracy of chest CT in the diagnosis of COVID-19: An umbrella review*. Retrieved from THe Centre for Evidence-Based Medicine: https://www.cebm.net/covid-19/the-accuracy-of-chest-ct-in-the-diagnosis-of-covid-19-an-umbrella-review/

Pawar, M. (2021, November 2). *More than 144 million jobs lost, COVID caused a global employment shift*. Retrieved from Charles Sturt University: https://news.csu.edu.au/opinion/more-than-144-million-jobs-lost,-covid-caused-a-global-employment-shift

Ray, S. (2020, June). *8 Proven Ways for improving the “Accuracy” of a Machine Learning Model*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2015/12/improve-machine-learning-results/

Schiebler, M., & Bluemke, D. (2022, February 1). Seeing Is Believing: COVID-19 Vaccination Leads to Less Pneumonia at Chest CT. *Radiology, 303*(3). doi:https://doi.org/10.1148/radiol.220129

Tokuç, A. A. (2022, November 10). *Why Feature Scaling in SVM?* Retrieved from Baeldung: https://www.baeldung.com/cs/svm-feature-scaling#:~:text=Feature%20scaling%20is%20crucial%20for,data%20points%20from%20different%20classes.

Wiersinga, W., Rhodes, A., Cheng, A., Peacock, S., & Prescott, H. (2020, July 10). Pathophysiology, Transmission, Diagnosis, and Treatment of Coronavirus Disease 2019 (COVID-19): A Review. *JAMA, 324*(8), 782-293. doi:10.1001/jama.2020.12839

Wikipedia. (2022, November 6). *F-score*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/F-score

Wood, T. (n.d.). *F-Score*. Retrieved from DeepAI: https://deepai.org/machine-learning-glossary-and-terms/f-score#:~:text=The%20F%2Dscore%2C%20also%20called,positive'%20or%20'negative'.

World Health Organization. (2021). *The true death toll of COVID-19: estimating global excess mortality*. Retrieved from World Health Organization: https://www.who.int/data/stories/the-true-death-toll-of-covid-19-estimating-global-excess-mortality

Yahya. (2018, September 18). *Logistic regression python solvers' definitions*. Retrieved from Stackoverflow: https://stackoverflow.com/questions/38640109/logistic-regression-python-solvers-definitions

Zheng, Z., Yao, Z., Wu, K., & Zheng, J. (2020, April 29). The diagnosis of pandemic coronavirus pneumonia: A review of radiology examination and laboratory test. *J Clin Virol, 128*(104396). doi:10.1016/j.jcv.2020.104396

# Appendix A

## Features and their details

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Description** | **Source** | **Values** | **Type** |
| current\_status | Current status of the persons’ COVID-19 tests? | Case Report Form | Laboratory-confirmed case Probable case | String |
| cdc\_report\_dt | Date case was first reported to the CDC | Calculated | YYYY-MM-DD | Date |
| cdc\_case\_earliest\_dt | Date related to illness/ Date received by CDC | Calculated | YYYY-MM-DD | Date |
| sex | Sex | Case Report Form | Male Female Unknown Other Missing | String |
| age\_group | Age group | Calculated | 0 - 9 Years 10 - 19 Years 20 - 39 Years 40 - 49 Years 50 - 59 Years 60 - 69 Years 70 - 79 Years 80 + Years Missing NA | String |
| race | Race | Calculated | American Indian/Alaska Native Asian Black Multiple/Other Native Hawaiian/Other Pacific Islander White Unknown Missing NA | String |
| ethnicity | Ethnicity | Case Report Form | Hispanic Non-Hispanic Unknown Missing NA | String |
| county\_fips\_code | County FIPS Code | Calculated | NA if suppressed null if could not be matched | String |
| res\_county | County of residence | Case Report Form | NA if suppressed Missing if not available | String |
| res\_state | State of residence | Case Report Form | NA if suppressed Missing if not available | String |
| onset\_dt | Date of symptom onset | Case Report Form | YYYY-MM-DD | Date |
| pos\_spec\_dt | Date of first positive specimen collection | Case Report Form | YYYY-MM-DD | Date |
| hosp\_yn | Was the patient hospitalized? | Case Report Form | Yes No Unknown Missing | String |
| icu\_yn | Was the patient admitted to an intensive care unit (ICU)? | Case Report Form | Yes No Unknown Missing | String |
| death\_yn | Did the patient die as a result of this illness? | Case Report Form | Yes No Unknown Missing | String |
| hc\_work\_yn | Is the patient a health care worker in the United States? | Case Report Form | Yes No Unknown Missing | String |
| pna\_yn | Did the patient develop pneumonia? | Case Report Form | Yes No Unknown Missing | String |
| abxchest\_yn | Did the patient have an abnormal chest X-ray? | Case Report Form | Yes No Unknown Missing | String |
| acuterespdistress\_yn | Did the patient have acute respiratory distress syndrome? | Case Report Form | Yes No Unknown Missing | String |
| mechvent\_yn | Did the patient receive mechanical ventilation (MV)/intubation? | Case Report Form | Yes No Unknown Missing | String |
| fever\_yn | Fever >100.4F (38C) | Case Report Form | Yes No Unknown Missing | String |
| sfever\_yn | Subjective fever (felt feverish) | Case Report Form | Yes No Unknown Missing | String |
| chills\_yn | Chills | Case Report Form | Yes No Unknown Missing | String |
| myalgia\_yn | Muscle aches (myalgia) | Case Report Form | Yes No Unknown Missing | String |
| runnose\_yn | Runny nose (rhinorrhea) | Case Report Form | Yes No Unknown Missing | String |
| sthroat\_yn | Sore throat | Case Report Form | Yes No Unknown Missing | String |
| cough\_yn | Cough (new onset or worsening of chronic cough) | Case Report Form | Yes No Unknown Missing | String |
| sob\_yn | Shortness of breath (dyspnea) | Case Report Form | Yes No Unknown Missing | String |
| nauseavomit\_yn | Nausea or Vomiting | Case Report Form | Yes No Unknown Missing | String |
| headache\_yn | Headache | Case Report Form | Yes No Unknown Missing | String |
| abdom\_yn | Abdominal pain | Case Report Form | Yes No Unknown Missing | String |
| diarrhea\_yn | Diarrhea (≥3 loose/looser than normal stools/24hr period) | Case Report Form | Yes No Unknown Missing | String |
| medcond\_yn | Pre-existing medical conditions? | Case Report Form | Yes No Unknown Missing | String |

# Appendix B

## Feature Selection

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Features** | **ALL** | **EDA** | **SVC** | **LR non-con** | **LR con** |
| current\_status | keep | keep | drop | drop | drop |
| cdc\_report\_dt | drop | drop | drop | drop | drop |
| cdc\_case\_earliest\_dt | drop | drop | drop | drop | drop |
| sex | keep | drop | drop | drop | drop |
| age\_group | keep | keep | keep | keep | keep |
| race | keep | drop | drop | drop | drop |
| ethnicity | keep | drop | drop | drop | drop |
| county\_fips\_code | keep | drop | drop | drop | drop |
| res\_county | keep | keep | keep | keep | keep |
| res\_state | keep | drop | drop | drop | drop |
| onset\_dt | drop | drop | drop | drop | drop |
| pos\_spec\_dt | drop | drop | drop | drop | drop |
| hosp\_yn | drop | drop | drop | drop | drop |
| icu\_yn | drop | drop | drop | drop | drop |
| death\_yn | drop | drop | drop | drop | drop |
| hc\_work\_yn | keep | keep | drop | drop | drop |
| pna\_yn | target | target | target | target | target |
| abxchest\_yn | keep | keep | keep | keep | keep |
| acuterespdistress\_yn | keep | keep | keep | keep | keep |
| mechvent\_yn | drop | drop | drop | drop | drop |
| fever\_yn | keep | keep | keep | drop | keep |
| sfever\_yn | keep | keep | keep | drop | keep |
| chills\_yn | keep | keep | drop | drop | drop |
| myalgia\_yn | keep | keep | keep | drop | drop |
| runnose\_yn | keep | keep | keep | drop | drop |
| sthroat\_yn | keep | keep | keep | drop | drop |
| cough\_yn | keep | keep | keep | drop | keep |
| sob\_yn | keep | keep | keep | keep | keep |
| nauseavomit\_yn | keep | keep | keep | drop | keep |
| headache\_yn | keep | keep | keep | drop | drop |
| abdom\_yn | keep | keep | drop | drop | drop |
| diarrhea\_yn | keep | keep | drop | drop | drop |
| medcond\_yn | keep | keep | keep | drop | keep |
| **Number of features** | 24 | 19 | 14 | 5 | 10 |

# Appendix C

## Encoding Method

|  |  |
| --- | --- |
| **Predictors** | **Encoding Method** |
| pna\_ya | Binary encoding |
| current\_status | Label encoding |
| sex | Label encoding |
| age\_group | Manually binned |
| race\_ethnicity\_combined | Target encoding |
| county\_fips\_code | Target encoding |
| res\_county | Target encoding |
| res\_state | Target encoding |
| hosp\_yn | Label encoding |
| icu\_yn | Label encoding |
| death\_yn | Label encoding |
| hc\_work\_yn | Label encoding |
| abxchest\_yn | Label encoding |
| acuterespdistress\_yn | Label encoding |
| fever\_yn | Label encoding |
| sfever\_yn | Label encoding |
| chills\_yn | Label encoding |
| myalgia\_yn | Label encoding |
| runnose\_yn | Label encoding |
| sthroat\_yn | Label encoding |
| cough\_yn | Label encoding |
| sob\_yn | Label encoding |
| nauseavomit\_yn | Label encoding |
| headache\_yn | Label encoding |
| abdom\_yn | Label encoding |
| diarrhea\_yn | Label encoding |
| medcond\_yn | Label encoding |

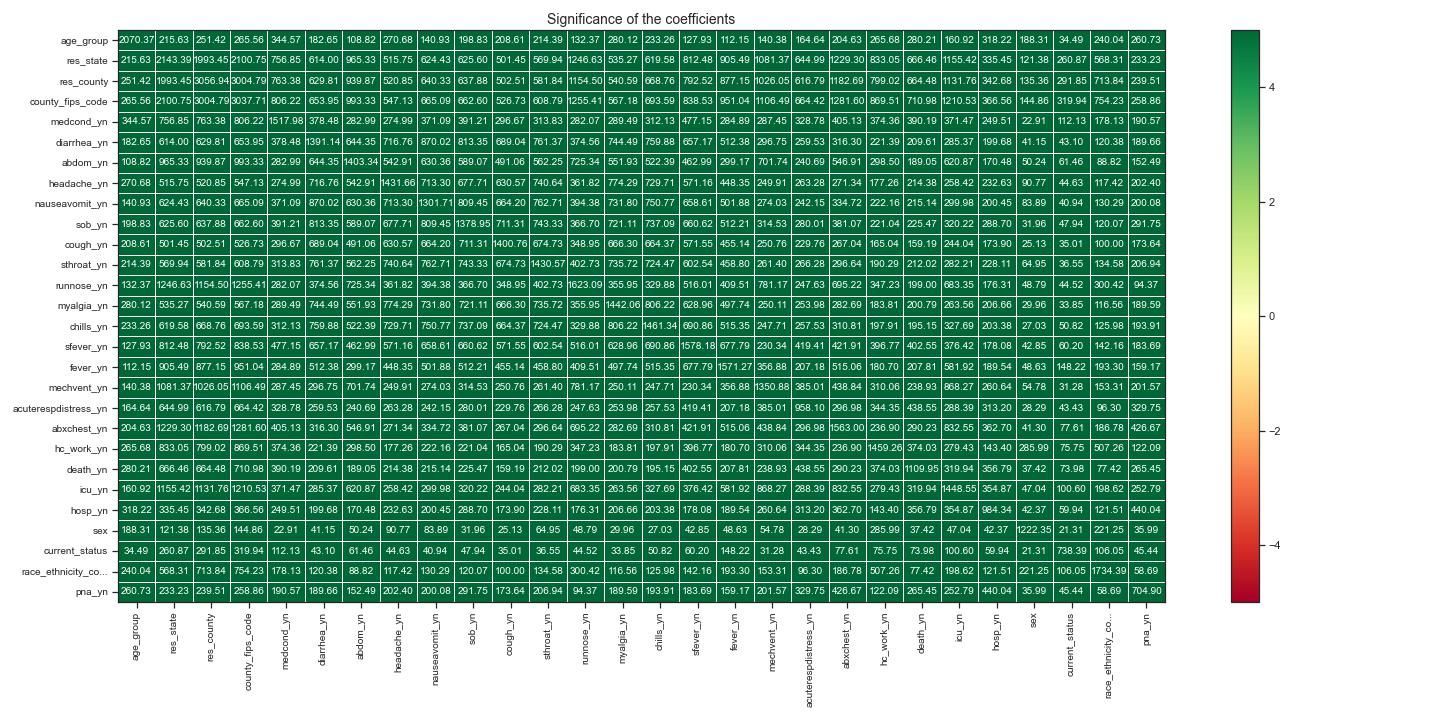
# Appendix D

## Predictors’ relationship with target variable based on EDA

|  |  |
| --- | --- |
| **Feature** | **Relationship?** |
| Intercep | na |
| headache\_yn | no |
| runnose\_yn | no |
| current\_status | no |
| sthroat\_yn | no |
| myalgia\_yn | maybe |
| abdom\_yn | no |
| hc\_work\_yn | no |
| diarrhea\_yn | maybe |
| chills\_yn | na |
| sfever\_yn | maybe |
| cough\_yn | no |
| nauseavomit\_yn | no |
| fever\_yn | yes |
| medcond\_yn | yes |
| res\_county | na |
| acuterespdistress\_yn | yes |
| sob\_yn | yes |
| age\_group | yes |
| abxchest\_yn | yes |

# Appendix E

## Statistical Significance of Phik



# Appendix F

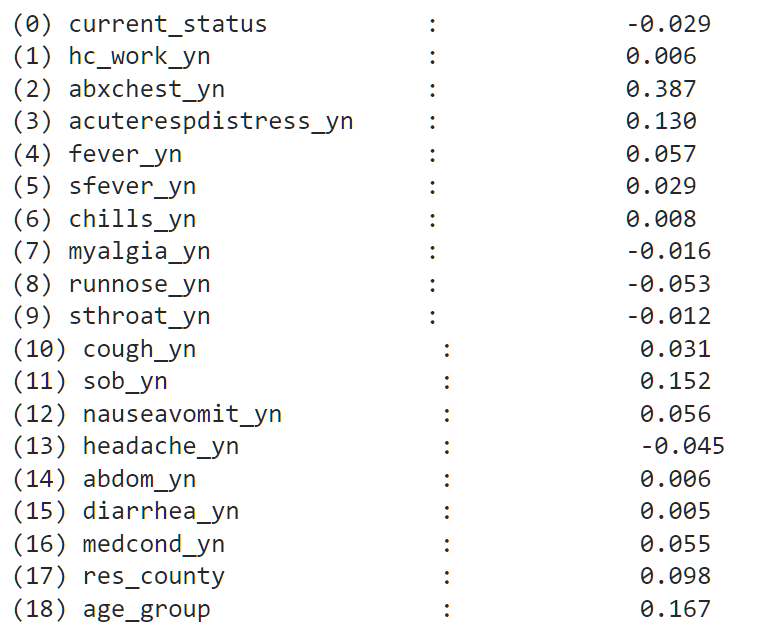
## Sorted coefficients in LR (eda) in comparison to relationships derived in EDA

|  |  |  |
| --- | --- | --- |
| **LR\_eda** | **COEF** | **Relaionship?** |
| Intercep | -4.6749 | na |
| headache\_yn | -0.19067 | no |
| runnose\_yn | -0.19014 | no |
| current\_status | -0.12406 | no |
| sthroat\_yn | -0.06322 | no |
| myalgia\_yn | -0.05672 | maybe |
| abdom\_yn | 0.031906 | no |
| hc\_work\_yn | 0.036282 | no |
| diarrhea\_yn | 0.044977 | maybe |
| chills\_yn | 0.050217 | na |
| sfever\_yn | 0.144841 | maybe |
| cough\_yn | 0.168268 | no |
| nauseavomit\_yn | 0.215071 | no |
| fever\_yn | 0.220733 | yes |
| medcond\_yn | 0.246795 | yes |
| res\_county | 0.329079 | na |
| acuterespdistress\_yn | 0.492012 | yes |
| sob\_yn | 0.68221 | yes |
| age\_group | 0.763616 | yes |
| abxchest\_yn | 1.392121 | yes |

# Appendix G

## SVC coefficients

SVC (eda)



SVC (model):

