

Executive Summary

We find that from section 3.3.1, a Generalised Additive Model is most appropriate in fitting the dataset due to the presence of non-linear relationships between variables. This model produces the greatest predictive measures while still maintaining interpretability. 3.3.2 explores models focused on predictive performance, which yield a similar model that incorporates smoothing splines. Exploring this dataset, we find that patients older and exhibiting more health conditions were more likely to die from COVID as expected, however vaccinated patients did not indicate a decreasing likelihood to succumb to COVID.

Modelling relationship between patient characteristics and COVID-19 deaths

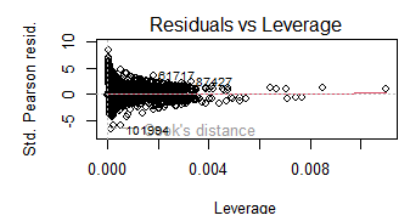
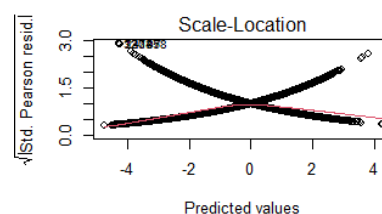
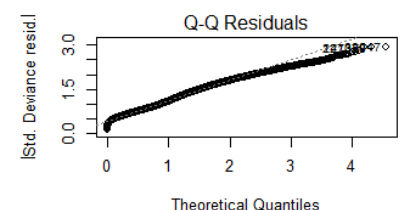
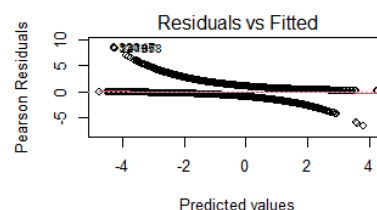
The data set was first cleaned by converting all 'True' 'False' variables into binary outcomes of 1 and 0 respectively. New variables for the age group of patients and duration of a patient's stay in hospital and ICU are created while also totalling the amount of admission conditions and death conditions for each patient (*Data Cleaning.1*). For this section, the data was split into 60% for training the models to find optimal hyperparameters before conducting 10-fold cross validation on the remaining 40% to compare models created. When making predictions, a threshold of 0.5 was chosen to balance the predictions for whether a patient dies from COVID while also maintaining interpretability as in the context of an actuary we wish to make accurate yet interpretable models.

Initially, a naive implementation of logistic regression with the binomial family for binary outcomes is fitted with all created and standard variables was formed before conducting both forwards and backwards stepwise selection to omit predictors deemed inappropriate on the basis of the AIC measure. Checking this model, we find that there exists high collinearity between age and age.group variables as expected, with a VIF factor over 8. As a result, this variable was removed despite a lower AIC as it severely reduced interpretability and stability of other coefficients. The remaining variables after stepwise selection are placed into a new model. After performing 10-fold CV we attain the following measures for the logistic stepwise model (*Measures.1*). From this model, we have an AUC of 0.7922 which measures the ability of the model to accurately distinguish between classes (*Logistic.1*).

Accuracy	73.337%
Precision	68.721%
Sensitivity	61.227%
F1 Score	0.6475

In analysing the coefficients of this model, there are non-significant variables (*Logistic.2*).

Analysing the diagnostic plots of this model there are clear trends in their residuals which is indicative of heteroskedasticity, where the assumption

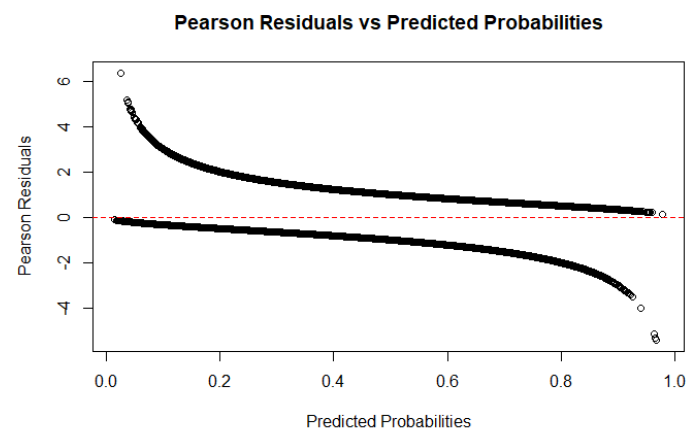


of a constant variance may not be valid. Additionally, the Scale-Location plots indicate this absence of homoscedasticity. While the general trend is close to the trend line on the QQ plot, the data points fall below the trend in the top right, indicating the residuals have flatter tails compared to the normal distribution. However, the leverage plot indicates the presence of only a few high leverage points.

Applying regularisation to the logistic model with all variables excluding ‘age.group’ due to multicollinearity, an elastic net method which balances both lasso and ridge regression is used. After tuning all hyperparameters (*Construction.1*), the final regularised model returned the following measures after CV. The regularised model offers very slight improvements over the stepwise logistic model in accuracy, sensitivity and F1 score while returning almost identical precision scores. The model also has a lower AUC of 0.7916 indicating the model is slightly worse in differentiating classes (*Regularisation.1*).

Accuracy	73.379%
Precision	68.572%
Sensitivity	61.779%
F1 Score	0.6499

Examining the residuals vs fitted plot (*Regularisation.2*), we find that the non-linear trend is still prevalent, indicating that the residuals are not properly accounted for in the model. Thus, despite offering very minor improvements in predictability, the model includes more variables (*Regularisation.3*) than the stepwise model as coefficients are only shrunk not removed, in addition to requiring more tuning of hyperparameters, overall increasing complexity.



In attempting to address the non-linearity in the residual plots, a Generalised Additive Model is fitted, applying a smoothing spline on variables for age and preAdmCond. All variables excluding age.group are used (*Construction.2*). The AUC of 0.793 indicates a slight improvement in model fit (*Spline.1*). From the QQ plot (*Spline.4*), it appears that this model better captures the assumption of normal errors, as the majority of data points fall along the trend line, and there is an absence of falling-off when approaching the top right in the previous two models. However, the residuals plot (*Spline.4*) still indicates a similar trend previously exhibited, despite fitting non-linear predictors, indicating that this model is only a slight improvement in fitting for the assumptions. This could potentially be attributed to being able to only smooth two terms, with most being binary variables which are harder to capture. Conducting CV to compare this model, we find the following measures. We find while the measures are very similar, there are minor

Accuracy	73.395%
Precision	68.568%
Sensitivity	61.859%
F1 Score	0.6504

improvements in accuracy, sensitivity and the F1 score, indicating that this model, while more complex, is appropriate.

Fitting a random forest using the same variables excluding age.group, we choose the number of trees to be 3 due to high computational time with the given amount of predictors. Conducting CV on this yields the following measures. We find that across all measures, the random forests are much worse with approximately 5% decrease in measures when compared to the GAM model. While increasing the amount of trees can potentially improve accuracy and other related measures, due to computation limits it is not possible to proceed further. However, examining the variable importance plot determined by the random forest, we find that clearly age is the most important variable, followed by the patient's ICU status and duration of hospital stay. This indicates that appropriately fitting these variables are important in making accurate predictions.

Accuracy	70.563%
Precision	64.182%
Sensitivity	59.796%
F1 Score	0.6190

The model that is ultimately chosen as the best is the spline model (GAM). This model possesses the best measures when using 10-fold CV to compare against other models, despite this improvement only being minor compared to the stepwise and regularised models. The random forest model is deemed inappropriate due to its significantly worse predictive power and worse interpretability due to the complexity of deep trees as it is difficult to determine which variables are most important in determining a patient's COVID death. The computational limit of random forests also favours the use of other models. Comparing the remaining three models, despite the spline being unable to capture similar non-linear trends among residuals, the spline model's QQ plot indicates the model better captures the distribution assumption of the errors in comparison to the stepwise and logistic model whose residuals indicate a flatter tail than that of the assumed normal distribution. The variable importance plot presented from the random forest which indicated the importance of age further signified the need to fit this variable appropriately, which the spline model best captures, as the variable plot indicates significant non-linearities at extreme data points which the other linear models would fail to capture. Examining the coefficients of the spline model (*Spline.2*), we find variables such as dyspnoea, hematologic, diarrhea, immuno are removed from the model indicating that these predictors are not appropriate for determining covidDeaths. Both terms where a smoothing spline is applied are significant at a level of 1% indicating they are appropriate predictors. Excluding pneumopathy, obesity, hepatic, renal and durationICU, all variables are significant at a level of 5%. Differentiating between characteristics of a patient that are expected to contribute to surviving or succumbing to COVID, we find that the vaccine has a negative coefficient indicating that vaccinated patients are less likely to die to COVID. For all other variables, as they are health conditions, a positive coefficient is more expected as worse health can be attributed to increasing likelihood to succumb to COVID. However, negative coefficients are also present for variables fever, cough, sorethroat, asthma,

preAdmCondSymp, preAdmnCondDis, durationHosp, respdistress, hepatic and renal. This could be potentially attributed to these health conditions having only a minor effect on a patient's overall health status and thus do not strongly contribute to covidDeaths hence the negative coefficient. Thus, the spline model was chosen as it had superior predictability while maintaining interpretability strengths of the stepwise and regularised model, in addition to increasing capability in modelling the non-linear trends of the residuals.

Predicting at high-risk individuals COVID-19 admission

Since the focus is now on prediction power of a model, all variables are initially included in the construction of each model including those newly created, excluding those known only upon death or discharge. A lower threshold of 0.3 is chosen as prediction in the healthcare context, we may wish to minimise the false negative rate. We are more inclined to predict a patient will die to COVID, as predicting a non-death in the case of a death is extremely detrimental in the medical context. Hospitals are more likely to take precautionary measures when dealing with patient health. Thus, it is more appropriate to increase sensitivity by capturing a greater proportion of positives, as predicting too many deaths is more appropriate than predicting too few. For this section, the dataset is split into 60% training and 20% for validation and test as we wish to obtain final measures for the selected predictive model.

A logistic regression with stepwise selection is conducted with all chosen variables. After stepwise selection, it is found that all variables are kept (*Logistic.3*). Despite present multicollinearity as present with previous analysis, the age.group variable is kept due to increased predictive power. Applying regularisation to a model with all variables (*Regularisation.4*), using the elastic net model and conducting CV to find optimal hyperparameters, we find that the optimal lambda is also 0.01994 and the alpha value is 0. The 0 alpha value indicates that ridge regression is preferred, as all variables have some strength in predicting the outcome variable. A Generalised Additive Model is fitted with all variables, applying a smoothing spline on variables age and preAdmCond. A random forest was fit with all variables similarly, while specifying the number of trees to be 3 due to computational limits. Comparing all models on a validation set we find the following measures.

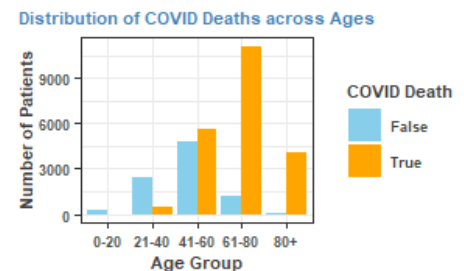
	Stepwise	Regularisation	GAM	Random Forest
Accuracy	55.159%	54.440%	55.159%	59.564%
Precision	46.844%	46.414%	46.842%	49.819%
Sensitivity	84.360%	85.743%	84.305%	58.720%
F1 Score	0.6024	0.6023	0.6022	0.5390

We can see that a lowered threshold has resulted in a much higher sensitivity across all models but the random forest. Across all measures, the random forest model is either comparable or significantly worse other than the accuracy measure. However this measure is rather one dimensional as it does not consider precision or sensitivity which the F1 Score does, hence

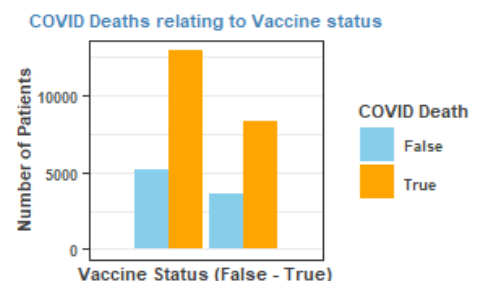
greater weight is placed on this measure when considering predictability. While all other three models of stepwise, regularisation and GAM exhibit similar performance, we will choose the GAM model as our best predictive model for 3.3.2 (*Spline.3*). This is due to the model's non-linear capabilities, in examining the diagnostic plots for residuals in 3.3.1, we find that all three models exhibit similar trends in the residuals vs fitted plots indicating that all models are unable to capture the residuals heteroskedasticity. However, the GAM is an improvement when examining the QQ plot as this indicated a slightly better fit of the error assumption. In addition to the variable plots produced (*Construction.2*) we find that both age and preAdmCond exhibit non-linearity particularly at extreme values. Thus for unseen datasets with potential non-linear relationships, the GAM model is much better suited for predictions. The importance of modelling the non-linear trend of age is amplified in the random forest's variable importance plot where age is significantly of greater importance in comparison to other variables. Thus, the GAM model offers greater flexibility and potentially greater predictive power despite its performance measures being on par with other models for this data set. Choosing this model, we get measures on an unseen test dataset and we get the following results.

Accuracy	55.790%
Precision	46.929%
Sensitivity	85.789%
F1 Score	0.6067

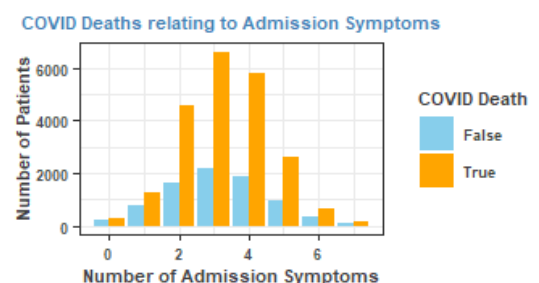
Making predictions on the Kaggle dataset and exploring the characteristics of patients that were more likely to die to COVID, we find that age contributes heavily to death. There exists a clear trend where among patients that are older, they are significantly more likely to succumb to COVID. Fitting a logistic regression model, we find that the age coefficient of ~ 0.119 (*EDA.1*) is statistically significant, further reinforcing this relationship.



However, there is a lack of reduction in covidDeaths in vaccinated patients, in fact patients that are vaccinated are exhibiting a greater proportion of deaths. This could be potentially due to patients that are vaccinated are already high-risk patients as many countries prioritise vaccinating those higher in age. The negative coefficient (*EDA.2*) from the model further indicates the lacking effectiveness of the vaccine in preventing COVID deaths.



Expectedly, there exists a relationship between the amount of COVID related health symptoms a patient exhibits and whether they die to COVID. The significant positive coefficient (*EDA.3*) further indicates the existence of this relationship. Hence, as expected, patients that enter hospital with more health conditions are more likely to succumb to COVID.



Technical Appendix

Data Cleaning.1

Variables are explained as follows: age.group - patients are divided into 5 groups ranging from 0 to 80+, preAdmCond - all health conditions a patient exhibits upon admission, preAdmCondSymp - these are a subset of preAdmCond and are symptoms linked to COVID, preAdmCondSympComm - these are conditions that are most common with COVID not just any symptom potentially caused by COVID, preAdmCondDis - these are conditions that are diseases and not linked with COVID, deathCond - these are condition known upon a patient's death.

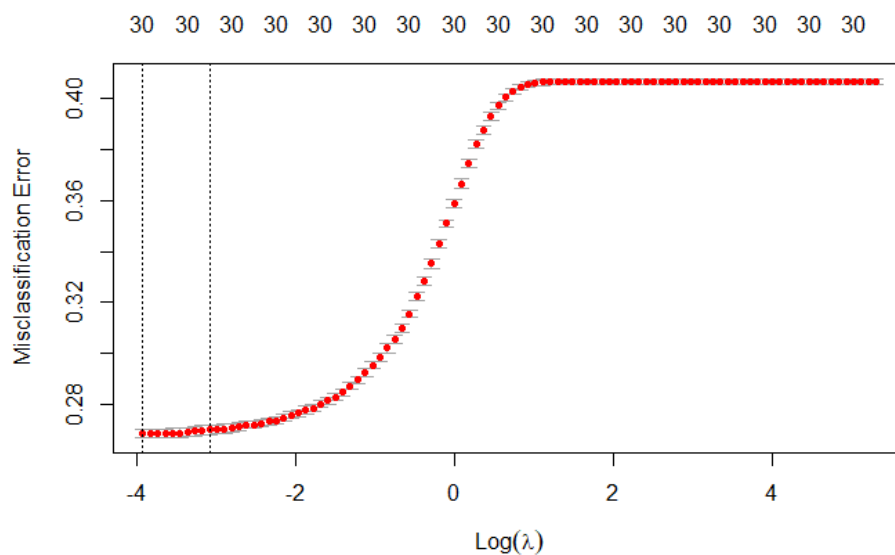
Measures.1

Precision is defined as the TP over the TP + FP while Sensitivity is measured as the TP over the TP + FN. Precision measures the accuracy of positive predictions while Sensitivity measures the ability of the model to identify all relevant instances. The F1 score looks to balance both these effects.

Construction.1

This is related to the construction of the regularisation method, finding the hyperparameters and tuning them to optimal values.

An optimal trade-off point between lasso regression which shrinks coefficients to 0 which may improve interpretability and ridge regression which minimises coefficients however does not often shrink to 0 is found using 10-fold cross validation. Using the accuracy measure for comparison, it is actually found that the most appropriate alpha value is 0, indicating that ridge regression is preferred as all variables offer some strength in prediction.

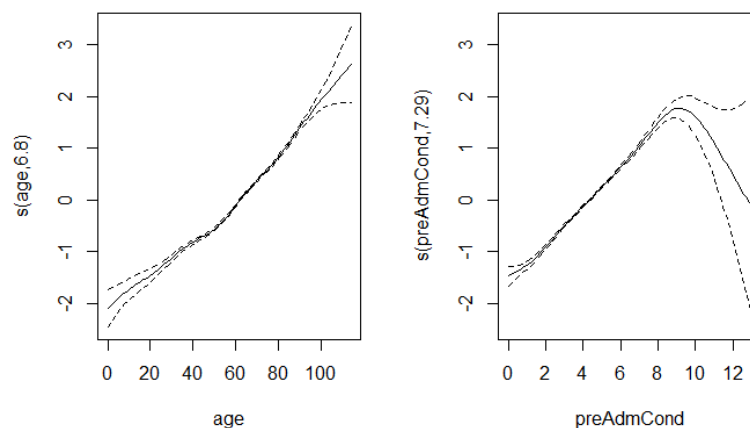


The optimal lambda value found using R's built in CV while specifying comparison using the misclassification error, was found to be 0.01994. This value determines the degree to which

coefficients are penalised, the graph indicates that to minimise misclassification errors, coefficients should be penalised heavily.

Construction.2

The variable plots indicate the age has a close to linear relationship in the middle ages of 40-80, as patients increase in age, they are more likely to die to COVID. However, this trend is harder to capture at both younger and older ages, as nonlinear trends are present, requiring a higher effective degree of freedom of 6.8 in the smoothed spline. Examining preAdmCond, there is a similar linear trend from 0-8 before presenting volatile trends after which can be potentially attributed to smaller sample size of patients exhibiting more than 8 pre-admission health conditions, decreasing the ability of the model to capture these trends. This is further evident in the high effective degrees of freedom as more flexibility is needed to capture these trends.



Logistic.1

ROC curve for stepwise model in 3.3.1



Logistic.2

Logistic stepwise model used for 3.3.1

Coefficients:

Estimate	Std. Error	z value	Pr(> z)
----------	------------	---------	----------

(Intercept)	-4.2312936	0.0468127	-90.388	< 2e-16	***
sexM	0.1356032	0.0154575	8.773	< 2e-16	***
vaccine	-0.5314910	0.0170230	-31.222	< 2e-16	***
fever	-0.0393078	0.0159261	-2.468	0.0136	*
cough	-0.1578261	0.0173855	-9.078	< 2e-16	***
dyspnoea	0.1801704	0.0209319	8.607	< 2e-16	***
oxygensat	0.2519243	0.0209134	12.046	< 2e-16	***
diarrhea	-0.0439006	0.0219458	-2.000	0.0455	*
downsyn	0.4582325	0.1072612	4.272	1.94e-05	***
asthma	-0.1048616	0.0407568	-2.573	0.0101	*
diabetes	0.1819572	0.0160897	11.309	< 2e-16	***
neurological	0.4747343	0.0341638	13.896	< 2e-16	***
pneumopathy	0.2409668	0.0364907	6.604	4.02e-11	***
obesity	0.2058347	0.0204340	10.073	< 2e-16	***
age	0.0426151	0.0005853	72.806	< 2e-16	***
deathCond	0.5995511	0.0246762	24.297	< 2e-16	***
icu	1.8489753	0.0199340	92.755	< 2e-16	***
durationHosp	-0.0063620	0.0008632	-7.370	1.70e-13	***
respdistress	-0.2939405	0.0302918	-9.704	< 2e-16	***
cardio	-0.5889619	0.0290326	-20.286	< 2e-16	***
durationICU	0.0018629	0.0012146	1.534	0.1251	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 127384 on 94325 degrees of freedom
 Residual deviance: 101964 on 94305 degrees of freedom
 AIC: 102006

Number of Fisher Scoring iterations: 4

Logistic.3

Variables used after stepwise selection, used for all models in 3.3.2

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-2.710979	0.106198	-25.528	< 2e-16	***
sexM	0.181104	0.014107	12.838	< 2e-16	***
vaccine	-0.495666	0.015589	-31.796	< 2e-16	***
fever	-0.039112	0.014601	-2.679	0.007392	**
cough	-0.227271	0.015972	-14.230	< 2e-16	***
sorethroat	-0.053943	0.019095	-2.825	0.004728	**
dyspnoea	0.380859	0.018472	20.619	< 2e-16	***
oxygensat	0.422803	0.018433	22.938	< 2e-16	***
diarrhea	-0.111778	0.020700	-5.400	6.67e-08	***
vomit	-0.056819	0.025449	-2.233	0.025572	*
hematologic	0.292146	0.072475	4.031	5.55e-05	***


```

downsyn      0.490193    0.099150    4.944 7.66e-07 ***
asthma       -0.165541    0.037116   -4.460 8.19e-06 ***
diabetes      0.222989    0.014751   15.117 < 2e-16 ***
neurological  0.394408    0.031541   12.505 < 2e-16 ***
pneumopathy  0.280426    0.033179    8.452 < 2e-16 ***
obesity       0.371985    0.018591   20.009 < 2e-16 ***
age           0.034539    0.001311   26.338 < 2e-16 ***
age.group21-40 -0.292485    0.110037   -2.658 0.007859 **
age.group41-60 -0.411561    0.116298   -3.539 0.000402 ***
age.group61-80 -0.270397    0.127917   -2.114 0.034529 *
age.group80+  -0.203143    0.142343   -1.427 0.153540
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

(Dispersion parameter for binomial family taken to be 1)

```

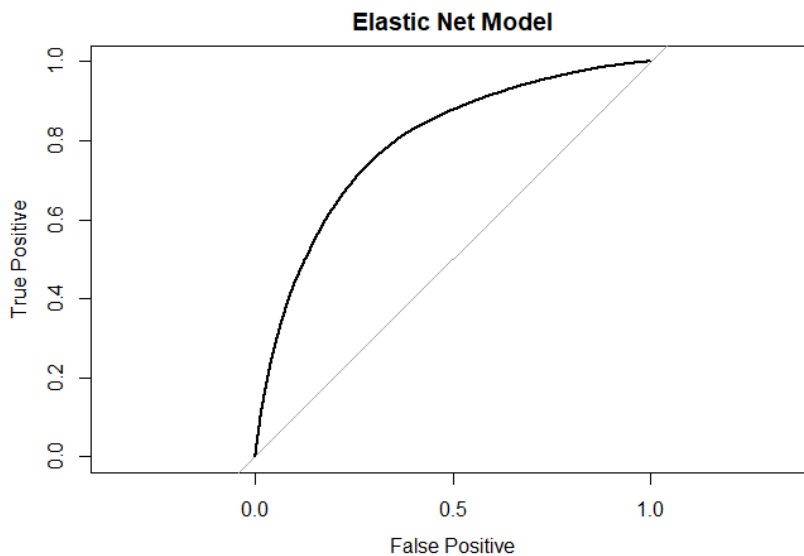
Null deviance: 127384 on 94325 degrees of freedom
Residual deviance: 117996 on 94304 degrees of freedom
AIC: 118040

```

Number of Fisher Scoring iterations: 4

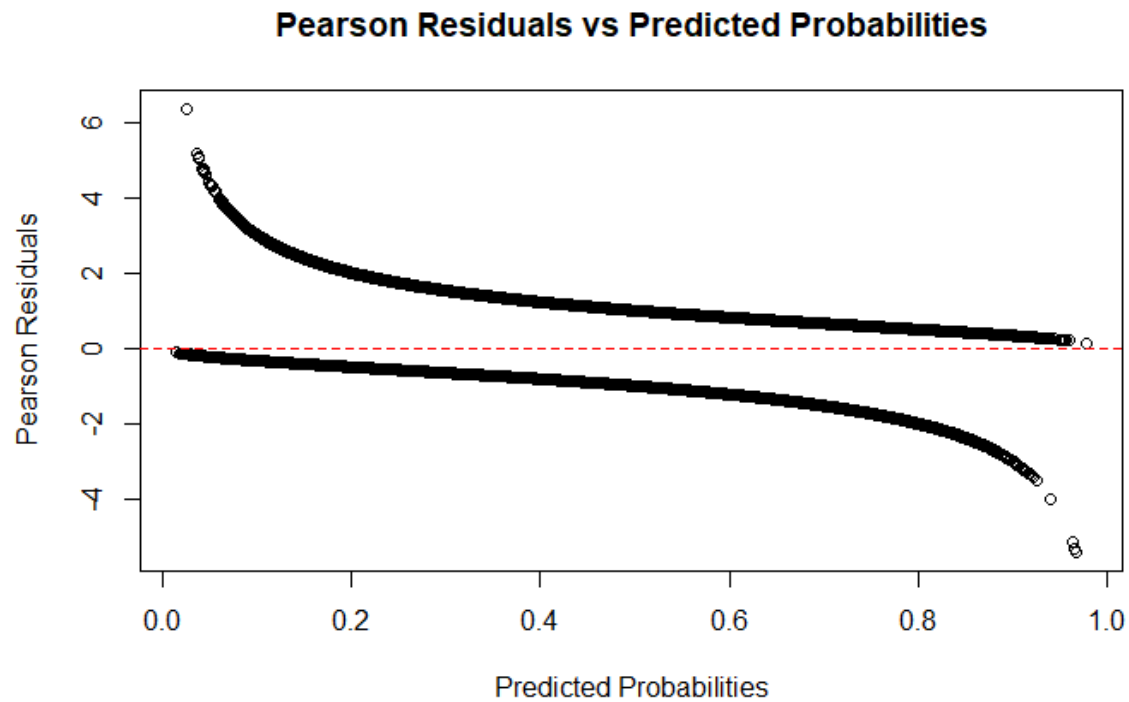
Regularisation.1

ROC Curve for the regularised model using an elastic net approach in 3.3.1



Regularisation.2

Residuals plot for the regularised model in 3.3.1 which exhibits similar trends to the logistic model selected using stepwise selection.



Regularisation.3

Regularisation Model for 3.3.1

	s1
(Intercept)	-3.236258882
sexM	0.087930937
vaccine	-0.304792816
fever	-0.060804125
cough	-0.149161720
sorethroat	-0.052106228
dyspnoea	0.159828078
oxygensat	0.219764453
diarrhea	-0.058956841
vomit	-0.004104070
hematologic	0.040885475
downsyn	0.204540199
asthma	-0.185490663
diabetes	0.101090741
neurological	0.335791601
pneumopathy	0.171608008
obesity	0.040381110
age	0.029285569
preAdmCond	-0.008066060
preAdmCondSymp	0.003041548
preAdmCondDis	0.098531568
preAdmCondSympComm	0.009577306

deathCond	0.108862051
icu	1.359895251
durationHosp	-0.005239881
respdistress	0.170257342
cardio	-0.036992150
hepatic	0.347148192
immuno	0.366936860
renal	0.378771980
durationICU	0.009266991

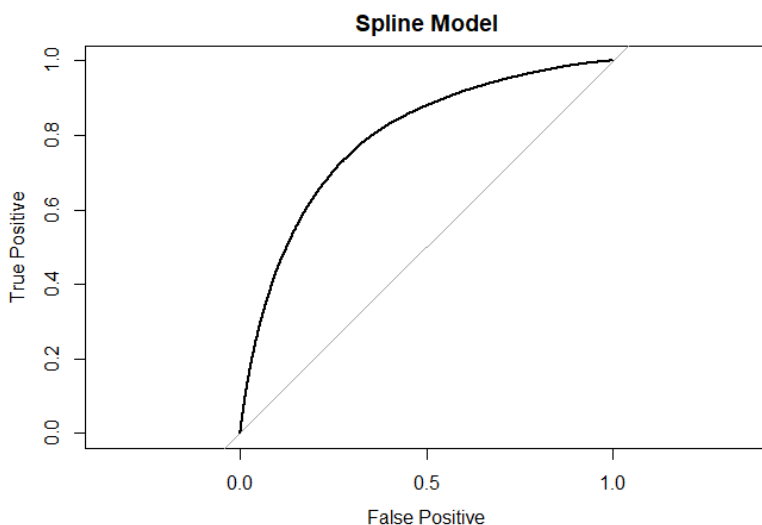
Regularisation.4

Regularisation model used for 3.3.2

	s1
(Intercept)	-1.835799510
sexM	0.129563644
vaccine	-0.292477720
fever	-0.072241268
cough	-0.193698413
sorethroat	-0.083316617
dyspnoea	0.271136644
oxygensat	0.314331009
diarrhea	-0.084719584
vomit	-0.071892235
hematologic	0.127618287
downsyn	0.172539924
asthma	-0.253669508
diabetes	0.081881844
neurological	0.237227657
pneumopathy	0.178270608
obesity	0.141795811
age	0.016096289
age.group21-40	-0.294993786
age.group41-60	-0.187224625
age.group61-80	0.142180333
age.group80+	0.369598011
preAdmCond	-0.001291268
preAdmCondSymp	0.008332775
preAdmCondDis	0.107594121
preAdmCondSympComm	0.022976742

Spline.1

ROC Curve for the GAM model with splines.



Spline.2

Model selected for 3.3.1

Parametric coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	0.0000000	0.0000000	NaN	NaN	
sexM	0.1429412	0.0155220	9.209	< 2e-16	***
vaccine	-0.9340681	0.0173840	-53.732	< 2e-16	***
fever	-0.2240827	0.0264953	-8.457	< 2e-16	***
cough	-0.3350387	0.0282412	-11.863	< 2e-16	***
sorethroat	-0.2151630	0.0306810	-7.013	2.33e-12	***
dyspnoea	0.0000000	0.0000000	NaN	NaN	
oxygenstat	0.0705468	0.0326997	2.157	0.030974	*
diarrhea	0.0000000	0.0000000	NaN	NaN	
vomit	0.0484563	0.0401240	1.208	0.227177	
hematologic	0.0000000	0.0000000	NaN	NaN	
downsyn	0.3432923	0.1361675	2.521	0.011699	*
asthma	-0.2232643	0.0909271	-2.455	0.014072	*
diabetes	0.0931337	0.0826611	1.127	0.259872	
neurological	0.3368782	0.0884013	3.811	0.000139	***
pneumopathy	0.1222969	0.0894448	1.367	0.171535	
obesity	0.0897339	0.0835226	1.074	0.282658	
preAdmCondSymp	-0.4276536	0.0238253	-17.950	< 2e-16	***
preAdmCondDis	-0.2726827	0.0813269	-3.353	0.000800	***
preAdmCondSympComm	0.2323293	0.0312337	7.438	1.02e-13	***
deathCond	0.6347123	0.0437161	14.519	< 2e-16	***
icu	1.8507930	0.0199559	92.744	< 2e-16	***
durationHosp	-0.0063202	0.0008643	-7.313	2.62e-13	***
respdistress	-0.3281053	0.0471277	-6.962	3.35e-12	***
cardio	-0.6098937	0.0456271	-13.367	< 2e-16	***
hepatic	-0.0186156	0.0844390	-0.220	0.825511	
immuno	0.0000000	0.0000000	NaN	NaN	
renal	-0.0487562	0.0577667	-0.844	0.398659	
durationICU	0.0020719	0.0012138	1.707	0.087838	.

```

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
              edf Ref.df Chi.sq p-value
s(age)        6.799  7.645   5479 <2e-16 ***
s(preAdmCond) 7.289  7.827   3081 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Rank: 42/47
R-sq.(adj) =  0.252   Deviance explained = 20.1%
UBRE = 0.080054   Scale est. = 1           n = 94326

```

Spline.3

Model selected for 3.3.2

Parametric coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	0.0000000	0.0000000	NaN	NaN	
sexM	0.1820576	0.0141188	12.895	< 2e-16	***
vaccine	-0.9666225	0.2251487	-4.293	1.76e-05	***
fever	-0.4382476	0.0373213	-11.743	< 2e-16	***
cough	-0.5381484	0.0405865	-13.259	< 2e-16	***
sorethroat	-0.4549902	0.0277176	-16.415	< 2e-16	***
dyspnoea	0.0000000	0.0000000	NaN	NaN	
oxygensat	0.0757197	0.0857596	0.883	0.37727	
diarrhea	-0.0528540	0.0366069	-1.444	0.14879	
vomit	0.0000000	0.0000000	NaN	NaN	
hematologic	0.5267770	0.1153188	4.568	4.92e-06	***
downsyn	0.7288962	0.1334798	5.461	4.74e-08	***
asthma	0.0000000	0.0000000	NaN	NaN	
diabetes	0.4454552	0.0896917	4.967	6.82e-07	***
neurological	0.6118293	0.0938805	6.517	7.17e-11	***
pneumopathy	0.4973155	0.0952064	5.224	1.76e-07	***
obesity	0.5874418	0.0905608	6.487	8.77e-11	***
age.group21-40	-0.1605494	0.1409830	-1.139	0.25479	
age.group41-60	-0.1986026	0.1597270	-1.243	0.21373	
age.group61-80	-0.0534208	0.1659868	-0.322	0.74758	
age.group80+	-0.0431050	0.1718383	-0.251	0.80193	
preAdmCondSymp	-0.5380703	0.2260831	-2.380	0.01731	*
preAdmCondDis	-0.6887988	0.2407346	-2.861	0.00422	**
preAdmCondSympComm	0.4587664	0.0318716	14.394	< 2e-16	***
age	0.0192101	0.0161549	1.189	0.23439	
oxygensat:asthma	0.0633708	0.0976286	0.649	0.51627	
cough:oxygensat	-0.0883215	0.0404424	-2.184	0.02897	*
fever:oxygensat	0.0159808	0.0372309	0.429	0.66775	
oxygensat:age	0.0002805	0.0011096	0.253	0.80045	

```

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Approximate significance of smooth terms:

	edf	Ref.df	Chi.sq	p-value
s(age)	2.811	3.672	8.415	0.06006 .
s(preAdmCond)	6.057	6.660	23.773	0.00069 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

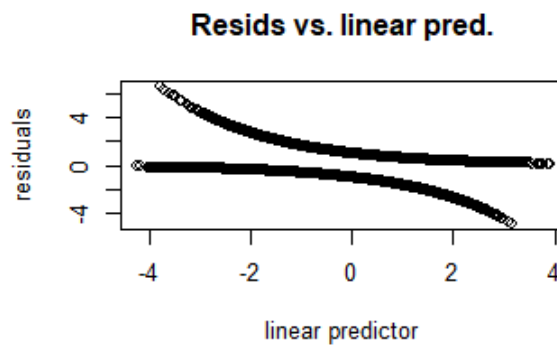
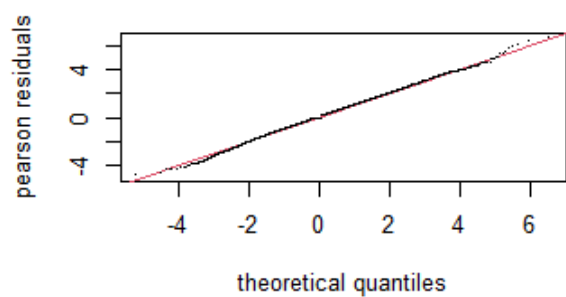
Rank: 42/47

R-sq.(adj) = 0.0952 Deviance explained = 7.41%

UBRE = 0.25115 Scale est. = 1 n = 94326

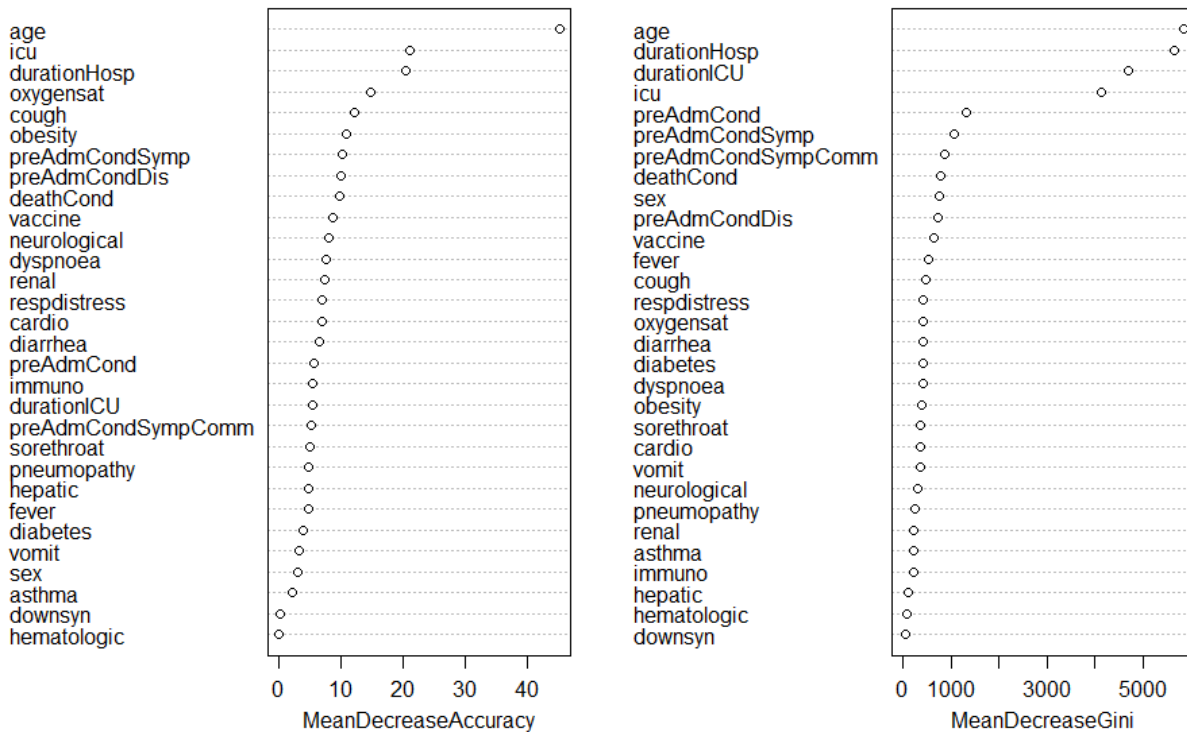
Spline.4

This plot is for residuals of the GAM model.



Random Forest.1

rfor.model



EDA.1

Model for age against covidDeath

Call:

```
glm(formula = covidDeath ~ age, family = binomial(), data = COVID.kaggle)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-5.848194	0.080952	-72.24	<2e-16 ***
age	0.119174	0.001464	81.38	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 36177 on 29999 degrees of freedom
Residual deviance: 23720 on 29998 degrees of freedom
AIC: 23724

Number of Fisher Scoring iterations: 5

EDA.2

Model for vaccine status against covidDeath.

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.92114	0.01648	55.880	< 2e-16 ***
vaccine	-0.07504	0.02590	-2.897	0.00376 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 36177 on 29999 degrees of freedom
Residual deviance: 36169 on 29998 degrees of freedom
AIC: 36173

Number of Fisher Scoring iterations: 4

EDA.3

Model for the amount of conditions a patient exhibits upon admission.

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.779121	0.033490	23.264	< 2e-16 ***
preAdmCondSymp	0.034780	0.009668	3.597	0.000321 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 36177 on 29999 degrees of freedom
Residual deviance: 36164 on 29998 degrees of freedom
AIC: 36168

Number of Fisher Scoring iterations: 4

Generative AI Usage Appendix

Prompt: Using a confusion matrix, how to calculate accuracy, precision, sensitivity and f1 score?

Response:

Definitions

- True Positive (TP): Number of positive samples correctly predicted as positive.
- True Negative (TN): Number of negative samples correctly predicted as negative.
- False Positive (FP): Number of negative samples incorrectly predicted as positive.
- False Negative (FN): Number of positive samples incorrectly predicted as negative.

Formulas

1. Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

2. Precision (also known as Positive Predictive Value):

$$\text{Precision} = \frac{TP}{TP + FP}$$

3. Sensitivity (also known as Recall or True Positive Rate):

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

4. F1 Score:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$$

Example Calculation

Suppose you have the following confusion matrix:

$$\begin{array}{ll} \text{TN} = 50 & \text{FP} = 10 \\ \text{FN} = 5 & \text{TP} = 35 \end{array}$$

Using the formulas:

1. Accuracy:

$$\text{Accuracy} = \frac{35 + 50}{35 + 50 + 10 + 5} = \frac{85}{100} = 0.85 \text{ or } 85\%$$

2. Precision:

$$\text{Precision} = \frac{35}{35 + 10} = \frac{35}{45} \approx 0.777 \text{ or } 77.7\%$$

3. Sensitivity:

$$\text{Sensitivity} = \frac{35}{35 + 5} = \frac{35}{40} = 0.875 \text{ or } 87.5\%$$


4. F1 Score:

$$\text{F1 Score} = 2 \times \frac{0.777 \times 0.875}{0.777 + 0.875} \approx 2 \times \frac{0.679}{1.652} \approx 0.821 \text{ or } 82.1\%$$

In R

Here's how you could compute these metrics in R given a confusion matrix:

r

 Copy code

```
confusion <- table(predicted, actual)

TP <- confusion[2, 2]
TN <- confusion[1, 1]
FP <- confusion[2, 1]
FN <- confusion[1, 2]

accuracy <- (TP + TN) / (TP + TN + FP + FN)
precision <- TP / (TP + FP)
sensitivity <- TP / (TP + FN)
f1_score <- 2 * (precision * sensitivity) / (precision + sensitivity)

list(
  Accuracy = accuracy,
  Precision = precision,
  Sensitivity = sensitivity,
  F1_Score = f1_score
)
```

Replace `'predicted'` and `'actual'` with your actual vectors of predictions and true values. This code calculates and returns accuracy, precision, sensitivity, and F1 score.