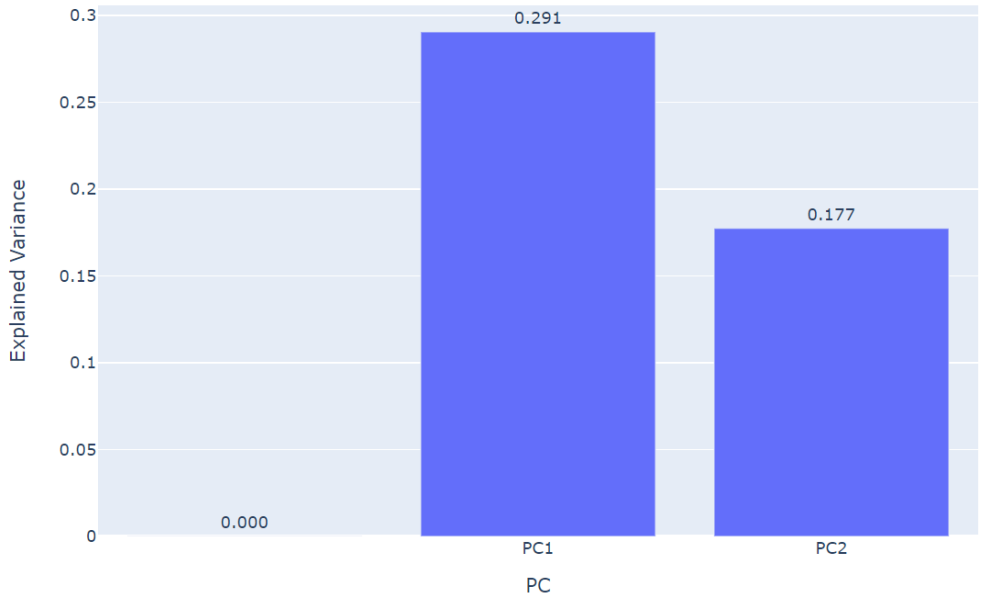
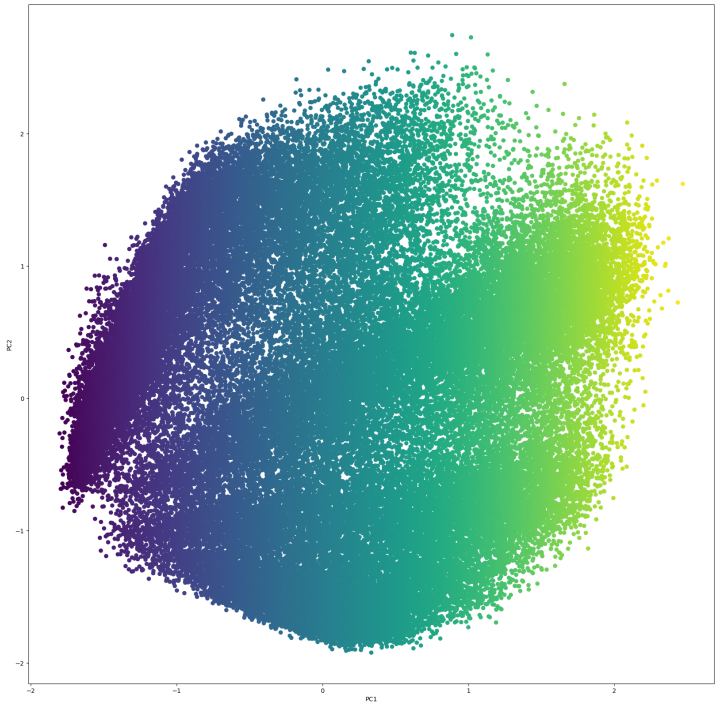
PRINCIPAL COMPONENT ANALYSIS (PCA)

The task at hand is to analyze the Fashion_mnist dataset with PCA.

Upon performing PCA on our dataset, the explained_variances are as shown. PC1 has an explained variance of 0.291, and PC2 has an explained variance of 0.177. We can see that the sum of the first 2 principal components makes up less than 47% of the variance.

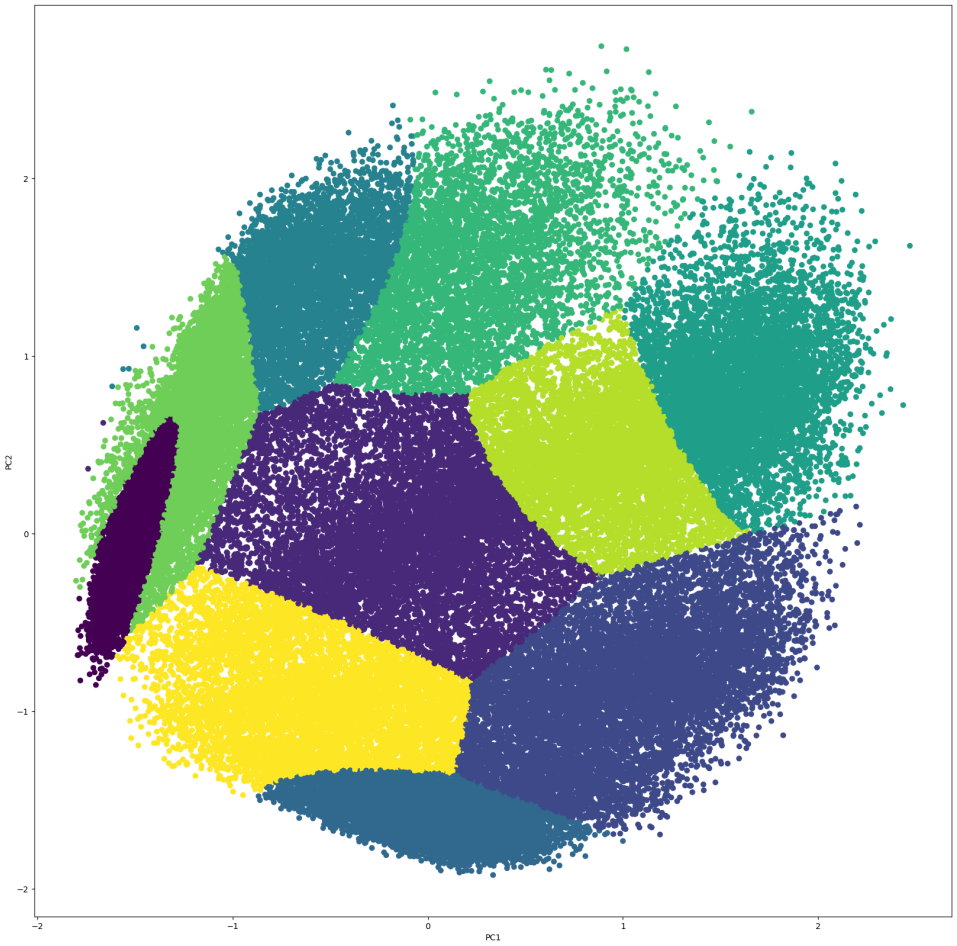


After computing the explained variance, a scatter graph was plotted to visualize the relationship between our principal components. With clusters being differentiated by colour, it can be seen that clusters are “formed” in our scatterplot by the intensity of the colour. With the cluster on the far right being more pale in comparison to the cluster on the far left.



GAUSSIAN MIXTURE MODELS (GMM)

Each of the principal components were then fit to a Gaussian Mixture Model to perform soft clustering as per the plot shown below.



Analyzing the clusters.

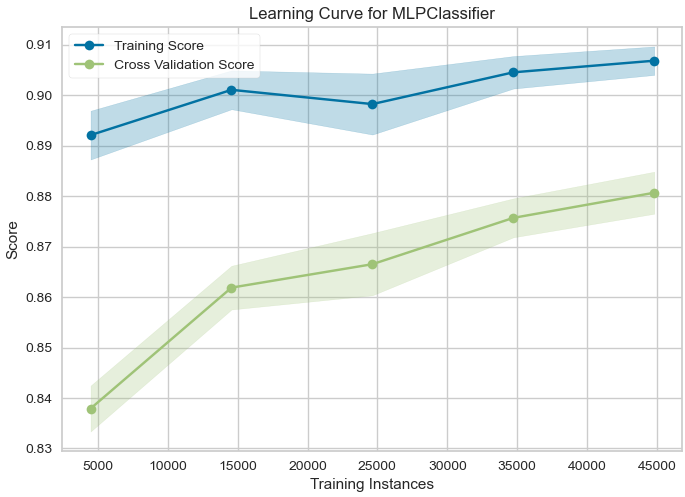
The table below shows the summary of our gaussian mixture model and each of our clusters.

cluster	0	1	2	3	4	5	6	7	8	9
Number of points	6036	7632	4953	11270	9266	5937	5618	6324	7565	5399

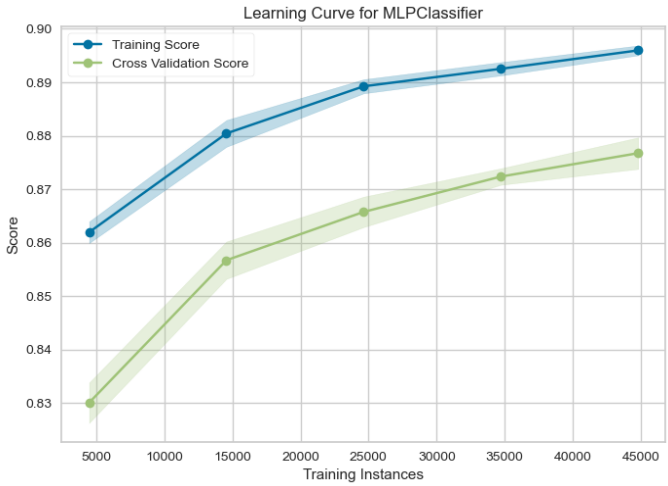
Mean PC1/PC2	0.115/ 1.513	0.0370/ 1.601	1.111/0.4 38	-1.413/ 0.203	1.017/- 0.870	1.643/0.9 01	0.312/0 .322	-0.847/ 1.172	-0.306/ -0.336	-0.710/-0. 990

ARTIFICIAL NEURAL NETWORKS (ANN)

The plots for the training and cross-validation scores are as shown below. As the curve for the training score is above the cv score, it can be concluded that the model is overfitting.



Model 1



Model 2

What are the results for our training dataset?

The MLPClassifier from sklearn used to classify this dataset and the accuracy score of this ANN was computed to be 0.877 (3 d.p) when tested on the testing dataset.

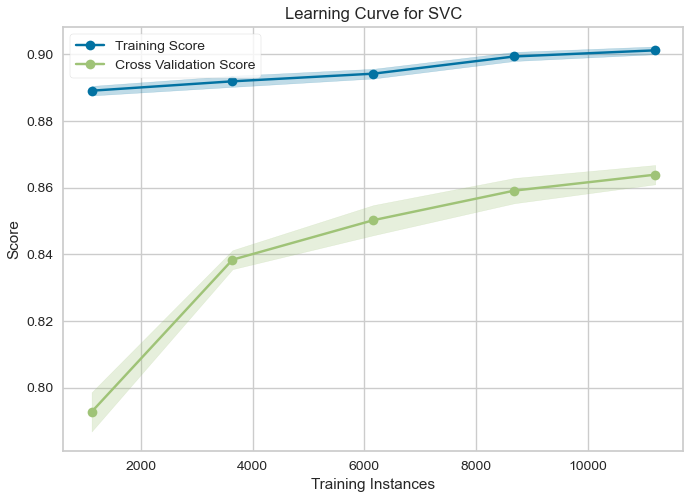
How sensitive is this model to hyperparameters?

Afterwards, the hyperparameters of the model were tuned with RandomizedSearchCV and the training and cross-validation scores were once again plotted. Evidently, the accuracy of the second model has dropped compared to the first model, implying that the model is quite sensitive to the tuning of hyperparameters.

SUPPORT VECTOR MACHINE (SVM)

As the SVM takes too long to run on a large training set, only 20% of the entire dataset has been used as training data. The SVM is then fitted and the score of our model has a value of 0.868 (3dp), implying that the accuracy of our model when classifying testing data is 86.8%.

Afterwards, we tuned our hyperparameters to optimize our SVM. We then created a new model with the best hyperparameters and the learning curve for our model as well as the cross-validation curve is then shown as below. The plot below shows the learning curve for our model on the training score and cross validation score. Similar with our ANN, it can be seen that the model is indeed overfitting as the curve for the training score goes above the curve for the cross validation score.

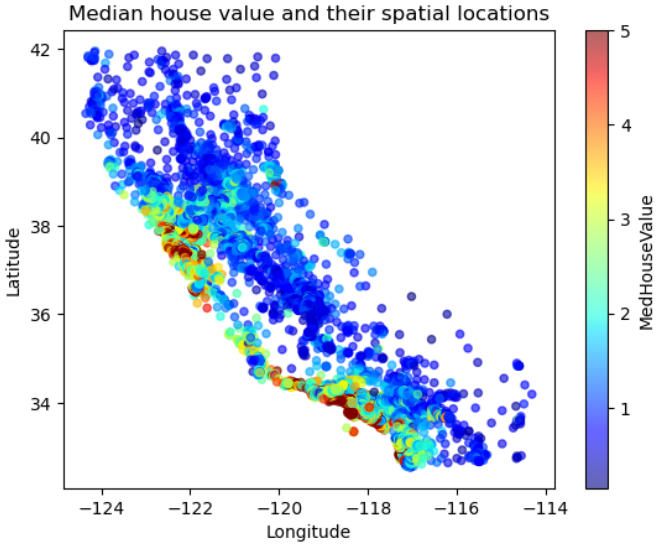


Comparing ANN and SVM

After analysing both ANN and SVM, it is clear that not only does our ANN model have a higher accuracy, the time taken to fit our training set to our ANN model is far less than SVM. Hence, ANN is a better model for classification.

BAYESIAN LINEAR REGRESSION

The diagram below shows a 2D scatter plot to show the relationship between the Median House Value and its spatial locations.



The diagram shows that the median house values increases at a certain spatial location, this could suggest that houses are more expensive at those certain locations.

Data Preprocessing
Before running our data through Cart algorithms and Ensemble methods, we first have to pre-process our data. When pre-processing our data, 4 criterias have been taken into account.

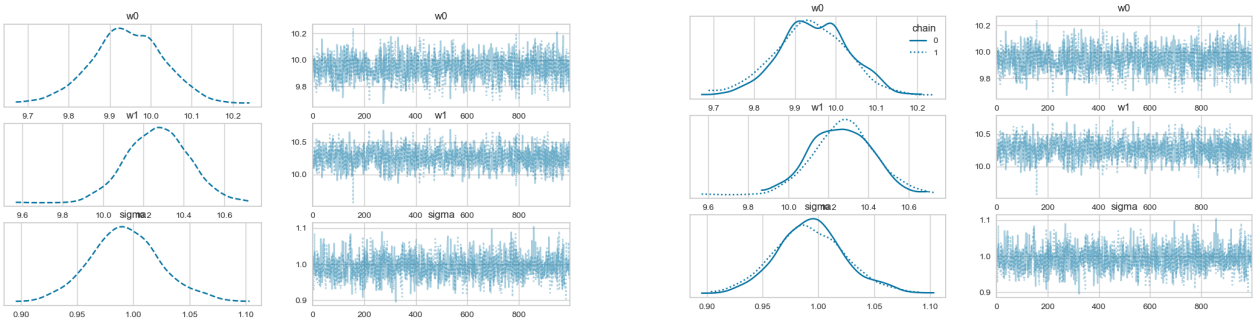
Order of data
By checking our dataset in a pandas dataframe, it is clear that the data has been set in an ascending order, therefore we need to reshuffle our dataset.

Missing values
When checking our dataframe for null values, no missing values can be seen, so no change is needed.

Duplicated rows
when checking our dataframe for duplicate rows, no duplicate rows can be found, hence no issues.

Structural error
No issues as we can see that each row has its correct data type. Therefore the only change needed to clean our data would be to reshuffle our dataframe.

We then used pymc to obtain our posterior distributions for our dataset. The results are as shown below.

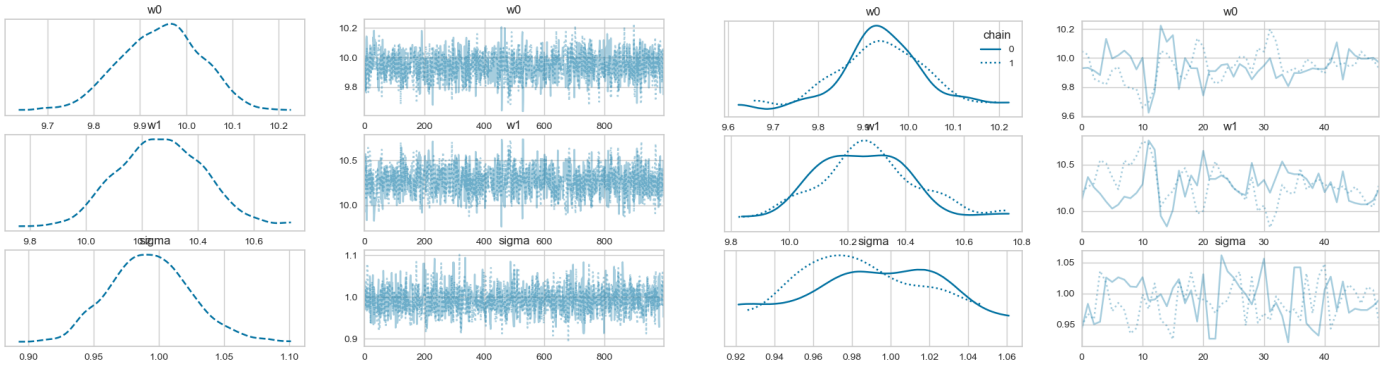


The table below describes the summary of our plot.

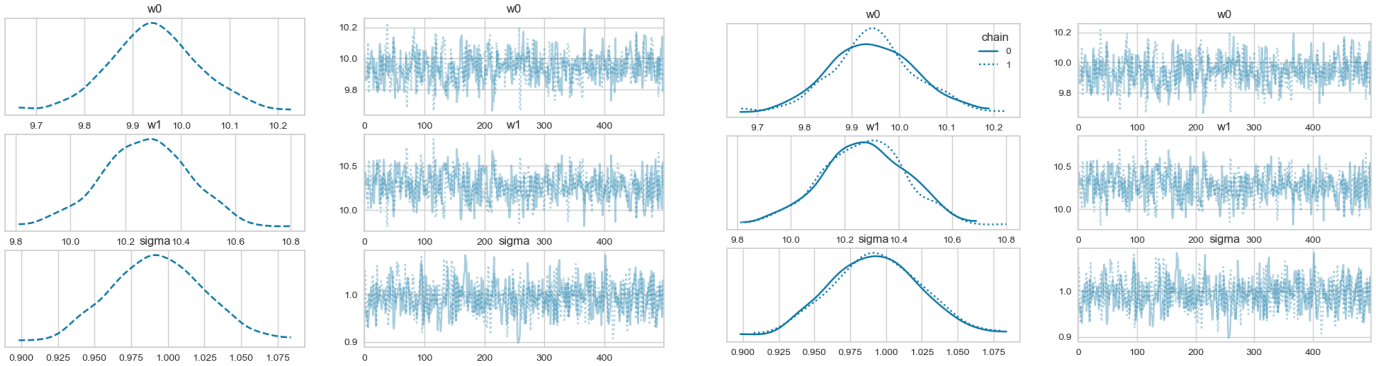
	mean	sd	hdi_3 %	hdi_97 %	msce_ mean	mcse_sd	ess_bul k	ess_tail	r_hat
w0	9.945	0.088	9.781	10.111	0.001	0.000	18723.0	19658.0	1.0
w1	10.269	0.153	9.965	10.538	0.001	0.001	18373.0	20467.0	1.0

sigma	0.994	0.032	0.931	1.050	0.000	0.000	25989.0	24861.0	1.0

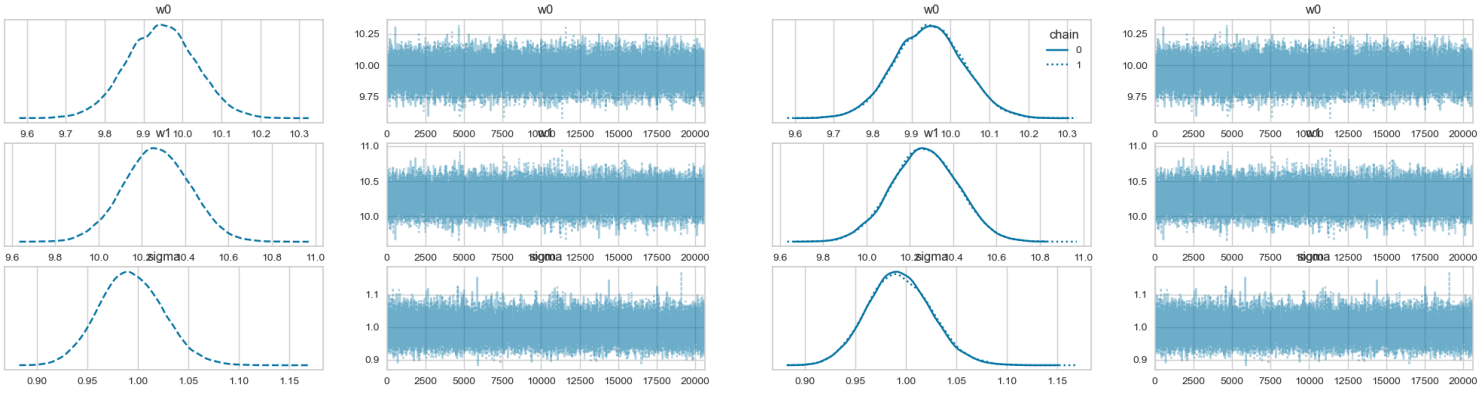
pymc is then run a few more times with 50, 500 and 20640 random samples and the results are as shown.



50 samples



500 samples



20640 samples

The most conclusive difference, when comparing these 3 diagrams, would be that as the number of sample size increases, the more bell shaped the curve becomes. This is because as the data size increases, the more normally distributed the posterior distributions become.

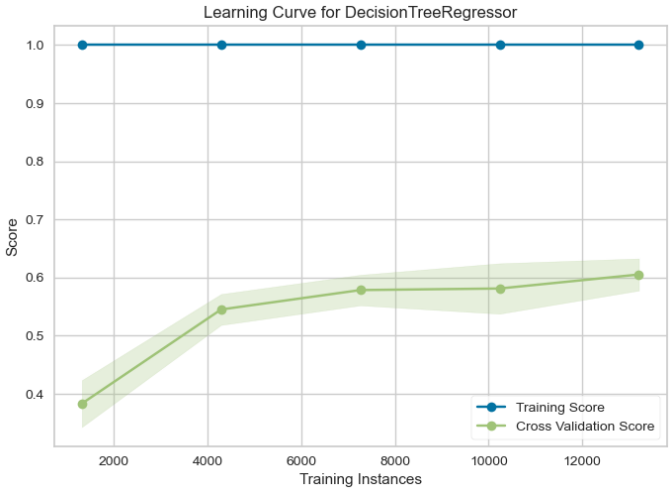
CART DECISION TREES

CART Decision Trees algorithm works as follow

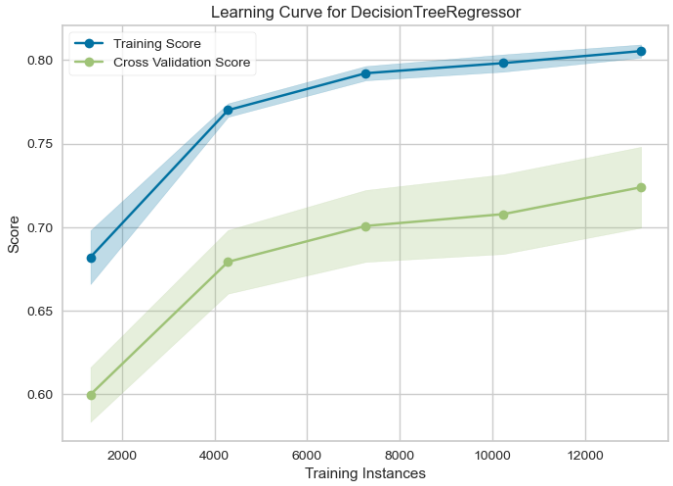
1. Given a dataset, the root node first finds the best split of the dataset.
2. The splitted data is then passed down the tree into the leaf nodes, the leaf nodes then once again find the best split of the data.
3. The splitting continues until no further best split can be found.

Which hyperparameters have the strongest effect on our model?

The algorithm used for our decision tree was the decision tree regressor algorithm. To find the best fitted hyperparameters for our model, the GridSearchCV function was once again used. 2 learning curves were then plotted.



Model 1



Model 2

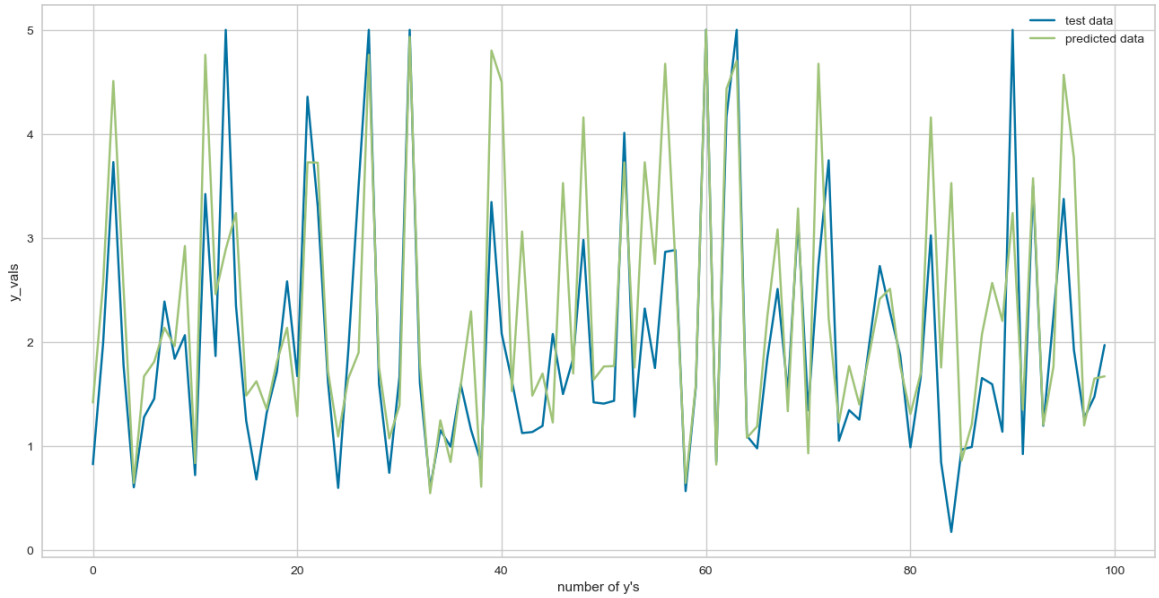
After tuning our hyperparameters, the best hyperparameters which was computed had a max depth of 90, minimum samples leaf of 20 and minimum sample split of 40.

How does the tuning of hyperparameters affect training time?

Unfortunately, the time taken to train model would increment as the number of max depth increases, because the model would take a longer time to give a better cross validation score.

What are the results of the chosen setup on the test set?

The diagram below shows a plot of the test values of y and the predicted test values of y. It can be seen that the predicted values of y have a relatively high accuracy to the test set of y.



In what situations is a decision tree better than linear regression?

When it comes to deciding whether to use a decision tree or linear regression, you tend to use decision tree algorithms when there are non linear relationships in your dataset.

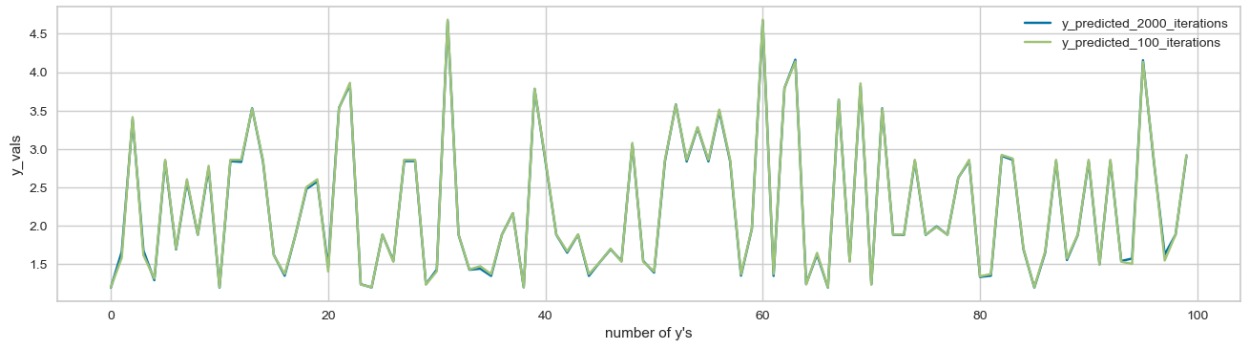
ENSEMBLE METHODS

The ensemble method chosen was a Random Forest Regressor algorithm. In terms of how this algorithm compares to other ensemble classification algorithms, the few key benefits are as listed below:

1. Random Forest algorithms tend to avoid overfitting issues when it comes to classification problems.
2. You can use this classifier on both classification and regression problems.
3. The Random Forest algorithm can extract features from raw data.

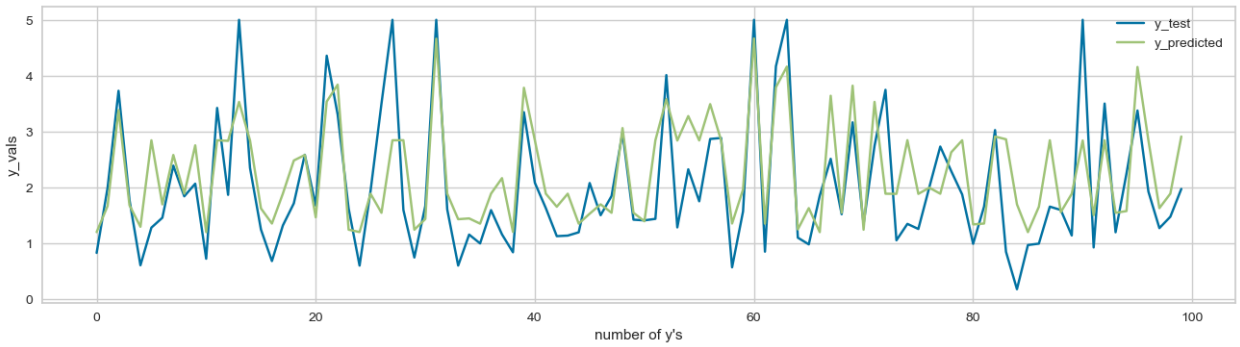
How is your method affected by the number of base models in the ensemble? Use a plot to support your discussion.

When plotting 2 models with different number of base models, it can be seen that there is barely any difference in the predicted values.



What are the results of your chosen setup on the test set?

The results of our chosen setup on the test set are as shown below, the accuracy of our model to predict y values have come quite close with little variance to the test data.



How is the Random Forest Algorithm compared to that of decision tree and bayesian linear regression?

When comparing our Random Forest Algorithm to that of decision tree or linear regression, it is clear that the predicted values of the random forest algorithm has much lesser variance to that of the decision tree as well as perform “better” than bayesian linear regression as this dataset is a non linear dataset.