Introduction:

This problem revolves around predicting the number of rented bikes from a bike rental shop using data which includes climate conditions and time information. Since the target variable, the count of rented bikes, is a continuous variable, and we are interested in making a prediction in a continuous state space, we will be looking at regression machine learning algorithms to solve our problem.

Methods:

To start, we use the Linear Regressor algorithm as a baseline model to compare our results against. This is because Linear Regression is a robust algorithm, providing fewer worries about overfitting, and so we would expect it to perform at a good level without any configurations. This is preferable to sklearn’s Dummy Regressor since the Linear Regression algorithm is likely to produce a viable model to predict the count of rental bikes. Thus, should provide suitable competition to any proposed model.

The two candidate algorithms to solve our regression problem are a Random Forest Regressor and an XGBoost Regressor. This is because ensemble algorithms have a good record for predictive accuracy, we can control the fit of the model easily and they also have a reputation for stability. We also have some intricacies that differentiate the Random Forest and XGBoost Regressors. XGBoost is a boosting algorithm that provides greater regularization capabilities as well as an improved speed.

We will be using several techniques to effectively test the performance and robustness of each model. This will allow us to make a fair comparison between our two models to choose the most appropriate. Firstly, we will be splitting our data into three samples: data to train the model with, data to validate the model and tune our hyperparameters and data that will act as our unseen data to test the performance results from the tuning. The training set will account for 60% of the data and the remaining 40% being split evenly between the test and validation datasets.

To evaluate our models, we will be using the as our performance metric. The of a model is defined as the proportion of variance in the dependent variable that we can predict from the independent variables. Mathematically, for observation , prediction and mean observed value, we have that

The benefit of using this metric is that, unlike Root Mean Squared Error (RMSE), it is not attached to any units, thus we can get a clear view of the model performance without needing to give context to a figure. However, a downside of the that the figure does not provide us with any more information about the residuals since it is detached from the predictions. Further, it does not highlight bias in the model predictions, so it is important to be careful of an overfitted model.

To choose our hyperparameters, we will be using the Optuna framework to assist us with optimisation. Optuna is a Bayesian approach to finding the optimal parameters which deals with the search problem probabilistically. It uses a Tree-structured Parzen Estimator which sequentially builds models to estimate hyperparameter performance based on past trials, and thus selects new hyperparameters to evaluate. What this does is provide a search algorithm that performs better than the more commonly used optimisers such as a Randomised Search or Grid Search in a clean structure.

Results:

Hyperparameters are parameters within the algorithm that can be used to adjust the learning process in a variety of ways. They can be split into two types of hyperparameters; ones to control the model complexity and ones to control the regularisation of the model – one might argue that an algorithm’s learning rate and optimiser fall slightly outside this grouping since they sit ‘on top’ of the algorithm.

Chart, bar chart

Description automatically generatedStarting with the results of our baseline Linear Regression Model, as displayed in Figure 1, we found that this baseline produced the following for the training and validation datasets respectively: 0.6847 and 0.6595. These scores are close, suggesting we have avoided mass-overfitting, and show a suitably large level of performance.

As previously mentioned, we will be utilising the XGBoost and Random Forest Regressors to model our problem. To start, we fitted each model to the training set without any hyperparameter adjustments. This is because the evaluation of these models will give us an indication to what hyperparameters we should be adjusting to get the best out the model. An overfit model would mean we would want to reduce the model complexity and increase regularisation with the opposite required for an underfit model.

As you can see in Figure 1 both the XGB and RF models perform at a higher level than our baseline with in the range of 0.90 - 1.00. An exceptional start. However, this plot does show that we have overfitted our models since there is a great different in between the train and validation datasets in both cases. Telling us that we need to increase our regularisation and reduce our model complexity.

We created an OptunaSearchCV instance for each model to help tune our hyperparameters. OptunaSearchCV includes a cross-validation component in which the training data was randomly split into ten folds to improve the robustness of our model. For each search, we performed 200 trials whereby each trial would select a value for each hyperparameter from the given options, train the model using these hyperparameters – with our cross-validation element – and produce the mean value of our cross-validation scores. OptunaSearchCV uses these scores, and the Tree-structured Parzen Estimator sampler to final the maximal score.

With regards to hyperparameters, we shall start with our XGB model. Two intuitive ways to control model complexity were min\_child\_weight, to control how finely the data is split for each leaf, and n\_estimators, the number of decision trees generated in the boosting process. We decided to tune n\_estimators as opposed to using the early\_stopping\_rounds hyperparameter in XGBoost’s *fit* method. This is so we could maintain consistency across our cross-validation folds as n\_estimators can be reproduced, unlike early stopping.

We also chose to include subsample and colsample\_bytree into our tuning pool to introduce a stronger element of randomness into our training. Subsample denotes the proportion of observations to be randomly sampled for each tree and colsample\_bytree is similar, but it indicates the fraction of columns to be randomly sampled for each tree. This was done so that our training data was more representative of the entire population such that we did not draw conclusions specific to our training data. We also used three regularisation hyperparameters: gamma, reg\_lambda and reg\_alpha. Each hyperparameter regularises the model at a different stage. Gamma controls the complexity of the gradient boosting; reg\_alpha smooth the prediction line by penalising the features that cause the loss function to increase (L1 regularisation); and reg\_lambda, L2 regularisation, regularises the weights of the features. Finally, we also looked at tuning the learning\_rate of the model since a high rate will cause the model the take larger steps in the training’s optimisation phases and thus avoid overfitting.

With regards to the chosen hyperparameters for our RF model, we also selected max\_depth and n\_estimators – both denoting the same aspect of the model as with XGB. However, we also introduced min\_samples\_split and min\_samples\_leaf. Min\_samples\_split controls the minimum number of samples needed for a node to be split and min\_samples\_leaf works in a similar way to XGB’s min\_child\_weight whereby it states the minimum number of samples on a leaf node. RF-specific hyperparameters we also chose to tune were max\_features, the number of features to use when looking for an optimal split, and bootstrap hyperparameter. This is used to determine if we want to use bootstrap samples to build decision trees which is useful for the model only being representative of the training data.

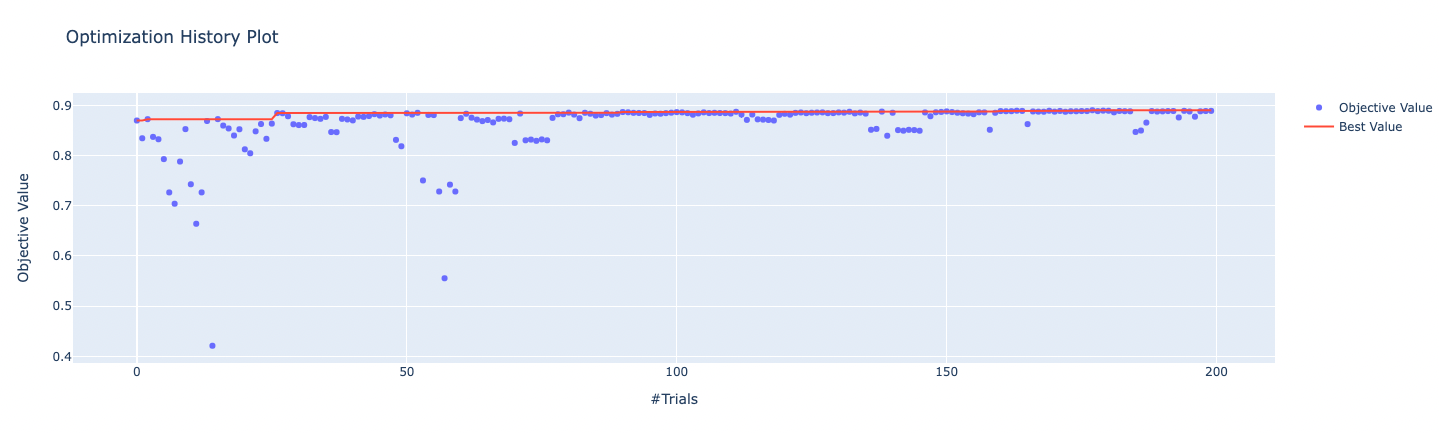


Figure 2

After 200 trials, the optimised hyperparameters produced an XGB model which was evaluated at the following for train and validation respectively: 0.9852 and 0.8782. As you can see from Figure 2, our search algorithm found the area in which the optimal values lay very quickly and then worked on perfecting it – note that the objective value is the . Figure 3 shows a the hyperparameters trialled by Optuna with darker coloured points showing us the later trials. As we can see, a lot of hyperparameters were tested, but the algorithm hones into the optimal areas very quickly. The max\_depth is a notable graph since, later trials do seem to be spread a little more evenly across the values. We can use Figures 2 and 3 in conjunction to see how the performance changed with different suggestions of hyperparameters – we can see the change in performance of trials 70-76 is due to the search algorithm exploring a max\_depth of two instead of five.

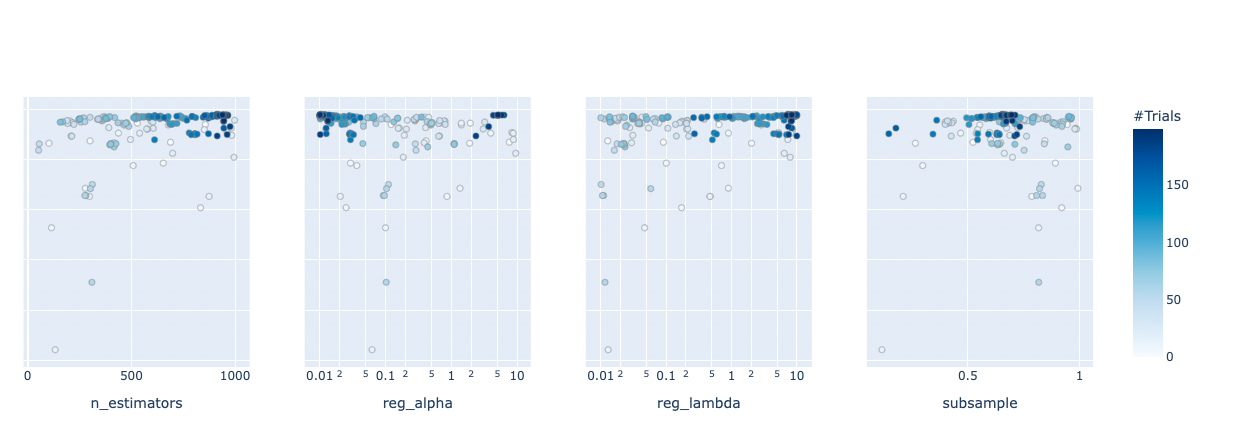
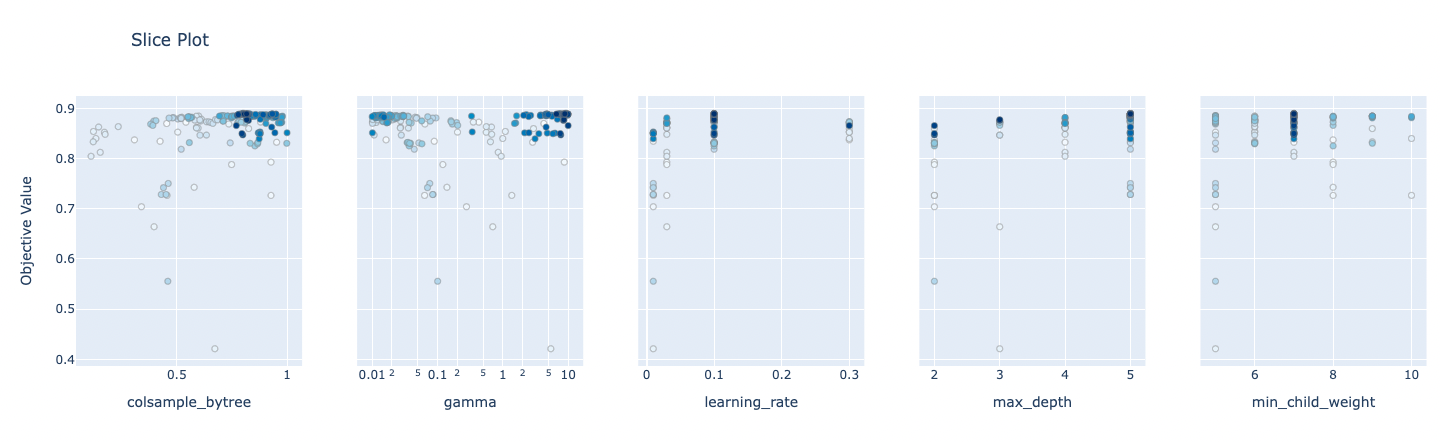
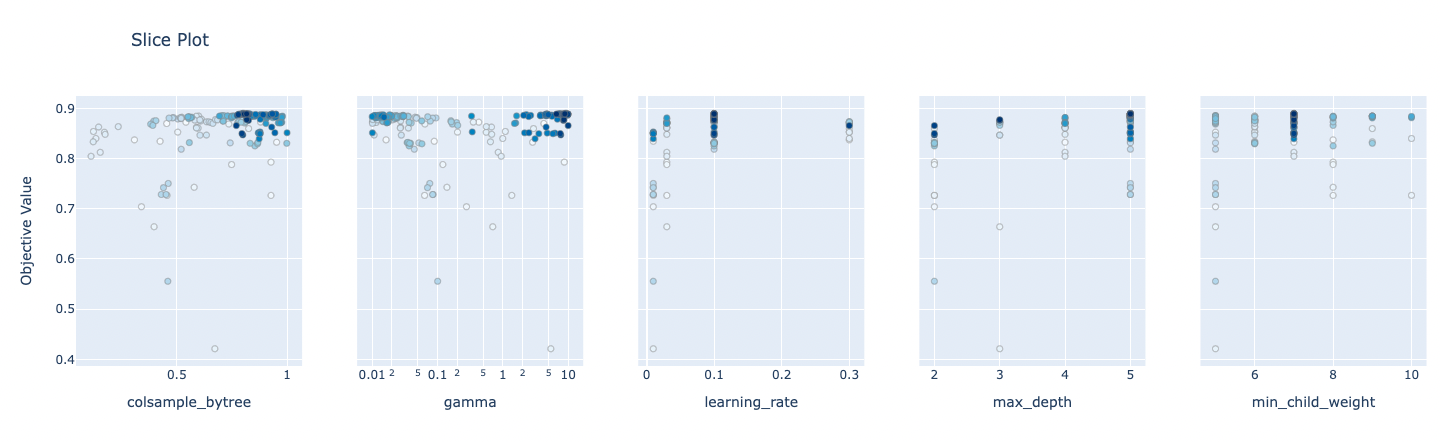


Figure 3



Moreover, Figure 4 shows us which hyperparameters had the greatest impact on the fit of our model. We can see subsample and learning\_rate to be frontrunners, but max\_depth and n\_estimators also show to have a large impact on results. The values suggest a better model than the baseline, but the model is still overfit. Thus, we undertake some manual adjustments to max\_depth and learning\_rate to reduce the model complexity – these hyperparameters chosen due to the impact shown in Figure 4 and the fact that they are intuitive complexity-reduction hyperparameters as they reduce the size of the model. We did consider changing subsample, but it did not yield any strong results in terms of fit. Figure 5 shows the values that the manual adjustments yielded. Taking into account both performance and fit, we decided max\_depth = 3 and learning\_rate = 0.03 were the most appropriate values since they are close together, but not at a large detriment to performance.

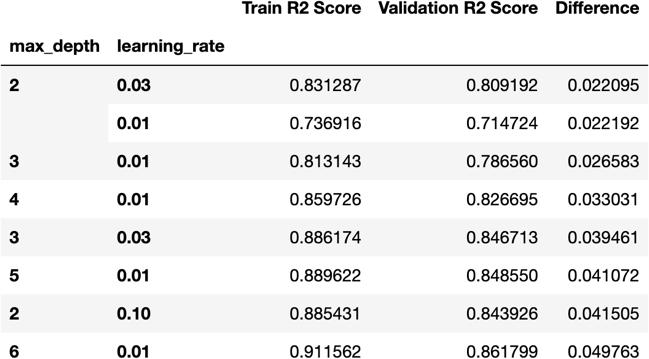


Figure 5

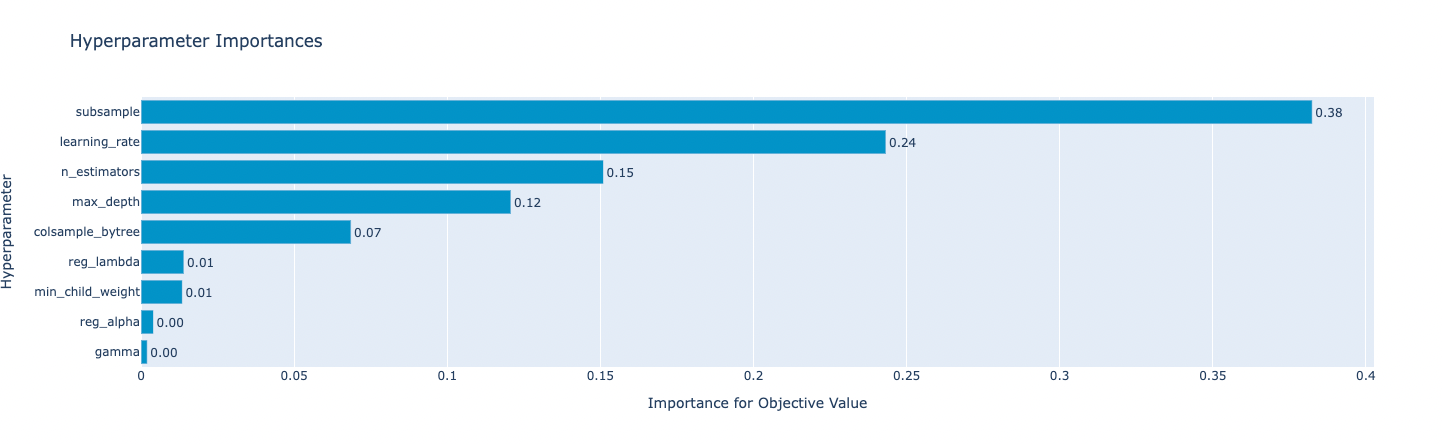


Figure 4

With the same cross-validation setup, 200 trials of the RF model gave tuned for train and validation respectively: 0.8620 and 0.7909. This is again a massive improvement; however, the RF model has not yielded the same performance as the XGB model. Figure 6 shows us the optimisation history of the hyperparameter search for this algorithm. We see that the search algorithm took more trials to find the optimal space of hyperparameters with great progress made in trial 101 where the search algorithm found max\_depth = 15 really boosted the performance of the algorithm. Figure 7, a snapshot of max\_depth trials, supports this since later trials are either 14 or 15. We also found that min\_samples\_leaf was the more influencial hyperparameter on the performance of the model. However, when it came to manual adjustments to correct the minor overfitting, we did not see much change in the results.

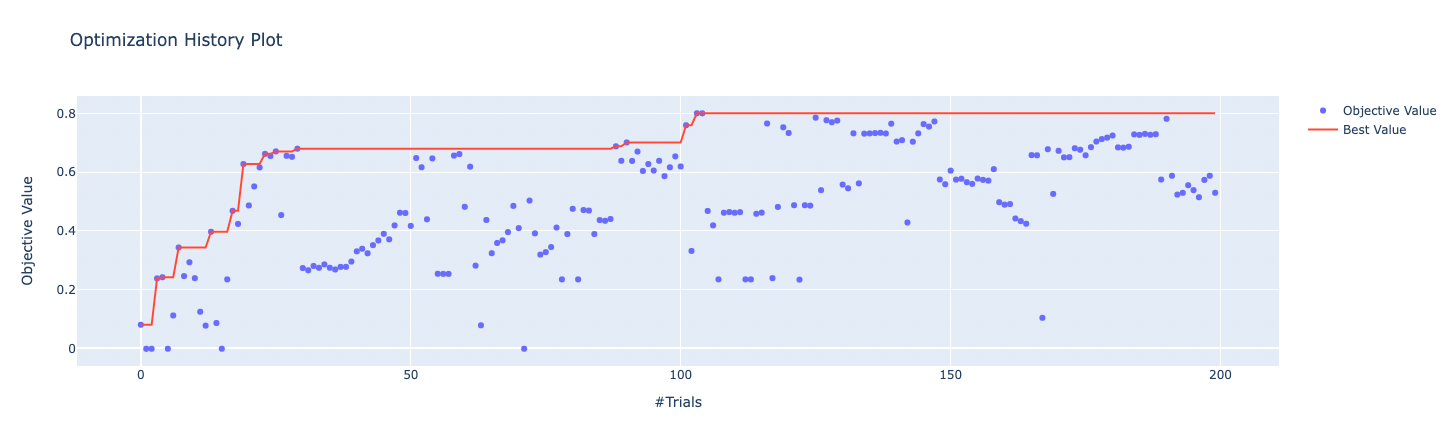


Figure 6

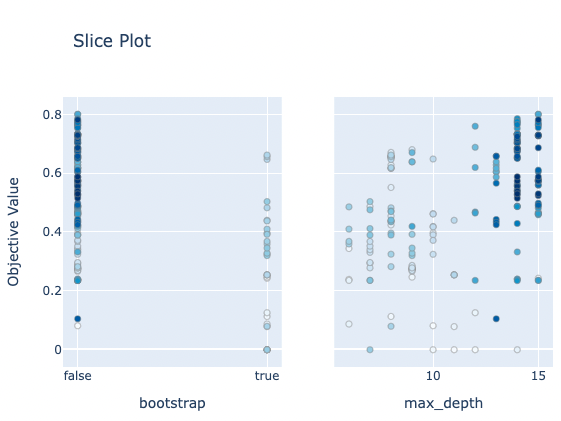
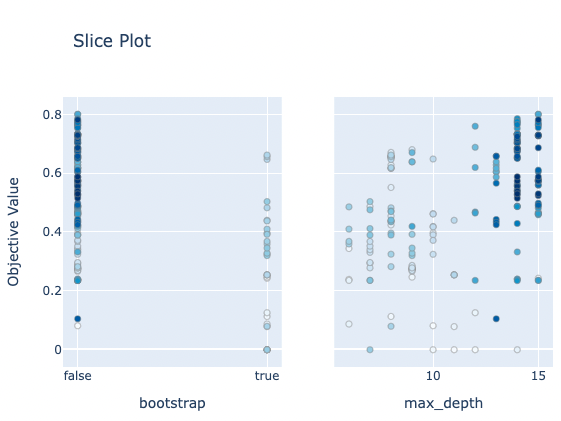


Figure 7

Snapshot

Chart, bar chart

Description automatically generatedTo compare our two models fairly, we use the test dataset. This is so that we can be confident that the model we choose is robust enough to generalise to unseen data with a high-performance level. The test dataset qualifies for our unseen data as it was split from the training data before we started modelling and so will display the qualities of each model on unseen data. Figure 8 shows us the performance metrics for each model on the testing dataset. We can see the XGBoost model performs with the highest scores on our unseen dataset with a score of 0.8545 in comparison to the Random Forest and baseline models evaluating to 0.7927 and 0.6664 respectively. We choose the XGBoost model since it performs at a high level on unseen data and this value is close the performance score when evaluated on the test data. This shows us that our chosen model generalises well to unseen data and predicts the number of rental bikes out each hour with reliable performance.