

**School of InfoComm Technology**

**Machine Learning**

Diploma in Data Science (DS)

Diploma in Information Technology (IT)

October 2022 Semester

**INDIVIDUAL ASSIGNMENT 2**

(40% of Machine Learning Module)

# Deadline for Submission:

**Presentation: 29th Jan 2023 (Sunday), 2359 Hours**

**Report: 11th Feb 2023 (Saturday), 2359 Hours**

|  |  |  |
| --- | --- | --- |
| Student Name | : | Ang Jing Yi Clairer |
| Student Number | : | S10223223K |

**Penalty for late submission:**

10% of the marks will be deducted every day after the deadline.

**NO** submission will be accepted after 18th Feb 2023, 23:59.

Contents

[Deadline for Submission: 1](#_Toc127018292)

[1. Introduction 3](#_Toc127018293)

[2. HR Analytics 3](#_Toc127018294)

[2.1. Load and Sample data 3](#_Toc127018295)

[2.2. Building of Models 4](#_Toc127018296)

[2.2.1. Logistic Regression (Base Model) 4](#_Toc127018297)

[2.2.1.1. Improving Logistic Regression Model (Stats Model - Pvalue) 5](#_Toc127018298)

[2.2.2. Decision Tree (Base Model) 6](#_Toc127018299)

[2.2.2.1. Improving Decision Tree (Grid Search) 7](#_Toc127018300)

[2.2.2.2. Improving Decision Tree (K-Fold) 8](#_Toc127018301)

[2.2.2.3. Improving Decision Tree (Ada Boost) 8](#_Toc127018302)

[2.2.2.4. Improving Decision Tree (Feature Importance) 9](#_Toc127018303)

[2.2.3. Support Vector Machine (SVC Base) 10](#_Toc127018304)

[2.2.3.1. Improving Support Vector Machine (K-Fold) 10](#_Toc127018305)

[2.2.4. Artificial Neural Network (Base) 11](#_Toc127018306)

[2.2.4.1. Improving Artificial Neural Network (Grid Search) 11](#_Toc127018307)

[2.2.4.2. Improving Artificial Neural Network (K-Fold) 12](#_Toc127018308)

[3. Building of Models 12](#_Toc127018309)

[3.1. Importing and Adjusting Dataset 12](#_Toc127018310)

[3.2. Build the Models 14](#_Toc127018311)

[3.2.1. Linear Regression (Base Model) 14](#_Toc127018312)

[3.2.1.1. Improving Linear Regression Model (Stats Model) 15](#_Toc127018313)

[3.2.2. Decision Tree (Base Model) 16](#_Toc127018314)

[3.2.2.1. Improving Decision Tree (Grid Search) 17](#_Toc127018315)

[3.2.2.2. Improving Decision Tree (K-Fold) 17](#_Toc127018316)

[3.2.2.3. Improving Decision Tree (Feature Importance) 18](#_Toc127018317)

[3.2.3. Support Vector Machine (SVR Base) 19](#_Toc127018318)

[3.2.3.1. Improving Support Vector Machine (K-Fold) 19](#_Toc127018319)

[3.2.4. Artificial Neural Network (Base) 20](#_Toc127018320)

[3.2.4.1. Improving Artificial Neural Network (Grid Search) 20](#_Toc127018321)

[3.2.4.2. Improving Artificial Neural Network (K-Fold) 21](#_Toc127018322)

[4. Further Improvements 21](#_Toc127018323)

[4.1. HR Analytics 22](#_Toc127018324)

[4.2. Airbnb 23](#_Toc127018325)

[5. Conclusion 24](#_Toc127018326)

[6. Reflection 24](#_Toc127018327)

[7. Discussion Board 25](#_Toc127018328)

## ****Introduction****

**In this assignment, the objective was to test the user's understanding of various Machine Learning Models through two different datasets, the HR Analytics dataset for classification and the Airbnb dataset for regression. The user was required to apply various models such as Logistic Regression, Decision Tree, Support Vector Machine (SVC), and Artificial Neural Network to both datasets.**

**These are the models the user used:**

|  |  |
| --- | --- |
| **HR Analytics (Classification)** | **Airbnb (Regression)** |
| **Logistic Regression** | **Linear Regression** |
| **Decision Tree** | **Decision Tree** |
| **Support Vector Machine (SVC)** | **Support Vector Machine (SVR)** |
| **Artificial Neural Network** | **Artificial Neural Network** |

**For the HR Analytics dataset, the user used Logistic Regression and Decision Tree as the two models to start with. After evaluating the performance of these two models, the user then improved them to their satisfaction before proceeding to the next model. In the same manner, the user used Support Vector Machine (SVC) and Artificial Neural Network for the HR Analytics dataset.**

**For the Airbnb dataset, the user used Linear Regression and Decision Tree as the two models to start with. Again, after evaluating the performance of these two models, the user improved them to their satisfaction before proceeding to the next model. In the same manner, the user used Support Vector Machine (SVR) and Artificial Neural Network for the Airbnb dataset.**

**Finally, the user compiled all the evaluations into a table, which would be shown later in the report. This table would provide a comprehensive summary of the performance of each model on both datasets and enable the user to compare and evaluate each model's performance. Overall, the goal of this assignment was to provide hands-on experience in applying various Machine Learning models to datasets and to understand their strengths and limitations.**

## ****HR Analytics****

## ****Load and Sample data****

**Graphical user interface, text

Description automatically generated with medium confidence**

**In Assignment 1, the user had already performed stratified sampling on the dataset, which helped to ensure that the data was evenly distributed among different classes. In this next step, the user will further divide the already stratified data into testing and training sets. This is a crucial step in any Machine Learning project as it allows the user to evaluate the performance of their model on unseen data and avoid overfitting, where the model performs well on the training data but poorly on the test data. By dividing the data into testing and training sets, the user can ensure that their model is generalizing well and can make accurate predictions on new data.**

## ****Building of Models****

## ****Logistic Regression (Base Model)****

Graphical user interface, text, application

Description automatically generated

**Logistic regression is to model the relationship between a binary dependent variable and one or more independent variables, in order to make predictions about the likelihood of an event occurring. Thus, the user decided to build logistic regression first as it was the most basic classification model.**

## ****Improving Logistic Regression Model (Stats Model - Pvalue)****

Text

Description automatically generated

Table

Description automatically generated with medium confidence

Graphical user interface, text, application, email

Description automatically generated

**After applying logistic regression on the dataset, the user improved the model using the stats model library by looking at the p-value of each independent variable. The p-value represents the probability of observing the results if the null hypothesis (that the independent variable is not related to the dependent variable) is true. In general, variables with a p-value less than 0.05 are considered to be significant, meaning that there is a strong relationship between the independent variable and the dependent variable.**

**In this case, the user observed that the p-value for "length of service" was 0.024, which was below the threshold of 0.05. As a result, the user dropped this column from the dataset as it was not significantly contributing to the model. This process of improving the model by analyzing the p-values and removing insignificant variables is known as feature selection and is a crucial step in building a high-performing logistic regression model.**

**By following this process, the user improved the logistic regression model and optimized its performance. This concludes the user's work on logistic regression.**

## ****Decision Tree (Base Model)****

Graphical user interface, text, application, email

Description automatically generated Diagram

Description automatically generated with low confidence

**For the next model in the assignment, the user moved on to using a decision tree algorithm. Similar to the logistic regression, the user created new variables to be used specifically in the decision tree model. The decision tree algorithm creates a tree-like structure to model the relationship between the independent variables and the dependent variable. Each internal node in the tree represents a test on a single feature, and each leaf node represents a prediction for the target variable.**

**After running the decision tree code, the user plotted the tree to visualize the structure of the model. On the left side of the tree, the predictions were for employees who were not promoted, while on the right side, the predictions were for employees who were promoted. This visual representation of the decision tree allowed the user to understand the key factors that were driving the predictions made by the model. To further improve the decision tree model, the user implemented four different strategies. These improvements aimed to enhance the model's accuracy and performance which will be shown next.**

## ****Improving**** Decision Tree ****(Grid Search)****

Text, application

Description automatically generated

A picture containing text, map, table

Description automatically generated

**In the first improvement, the user utilized a grid search approach to identify the optimal hyperparameters for the decision tree model. Grid search allows for an efficient and systematic way to search for the best set of hyperparameters by testing different combinations of values. By finding the best hyperparameters, the user could fine-tune the model and optimize its performance. After identifying the best hyperparameters, the user used them to create a new decision tree model and plotted the results. However, due to the limitations of the plotting software and the maximum depth set at 6, the text in the plot was too small to be clearly visible. This could have affected the interpretability of the results, but the user could have taken additional steps to address this issue, such as adjusting the plotting parameters or exploring alternative visualization methods.**

## ****Improving**** Decision Tree (K-Fold)

Text

Description automatically generated

**In the second improvement, the user employed a k-fold cross-validation technique to evaluate the accuracy of the decision tree model. K-fold cross-validation involves dividing the original dataset into k equally sized folds, where k-1 folds are used as training data and the remaining fold is used as test data. This process is repeated k times, with each fold serving as the test set once, to produce k different accuracy scores. The final accuracy is then calculated as the average of the k accuracy scores. By using k-fold cross-validation, the user could obtain a more robust estimate of the decision tree's performance and assess its generalization ability on unseen data. This helped to prevent overfitting and produce more reliable results.**

## ****Improving**** Decision Tree (Ada Boost)

Text, Word

Description automatically generated

**To further improve the accuracy of the model. Ada boost works by combining multiple weak learners to create a single, strong learner. This approach helps in overcoming the overfitting or underfitting problems that can arise in decision tree models. By combining the results of multiple weak learners, the model is able to make better predictions and provide a more accurate representation of the data.**

## ****Improving**** Decision Tree (Feature Importance)

Text

Description automatically generated

Text

Description automatically generated

**Lastly, the user used feature importance to sort the features based on their importance and dropped the least important features, which in this case were gender and number of trainings. This concluded their work on decision tree and the various improvements made to enhance its accuracy.**

## Support Vector Machine (SVC Base)

Text

Description automatically generated

**SVC stands for Support Vector Classification. It is a type of supervised machine learning algorithm used for solving classification problems. It is a part of the Support Vector Machine (SVM) algorithm family. SVC works by creating a hyperplane or line that separates the classes in the feature space. The goal is to find the hyperplane that maximizes the margin between the classes, which is the distance between the hyperplane and the closest data points. These closest data points are known as support vectors and the hyperplane is built based on their locations. SVC is commonly used in solving complex classification problems where data points are not linearly separable. This is the base code for it.**

## ****Improving**** Support Vector Machine (K-Fold)

Text

Description automatically generated

**After running the base model, the user then aimed to improve the model's performance. One of the methods used was K-Fold Cross Validation. This involves dividing the data into k-subsets and using k-1 subsets for training while using the remaining subset for testing. This process was repeated k times, and the average performance was recorded. This helped to reduce overfitting and provide a better estimation of the model's performance on unseen data. By using K-Fold Cross Validation, the user was able to evaluate the model's performance more accurately and adjust if needed to improve the model's performance.**

## Artificial Neural Network (Base)

Graphical user interface, text, application

Description automatically generated

**The Artificial Neural Network (ANN) is a powerful machine learning algorithm modeled after the structure and function of the human brain. The base model for ANN in this assignment trains the dataset by creating a network of artificial neurons that are connected through layers. Each neuron receives input from the previous layer, processes that input through its activation function, and then passes the processed information to the next layer. The goal of the ANN is to learn the complex relationships between the input and output variables in the dataset and make accurate predictions based on those relationships. By using backpropagation, gradient descent, and other optimization techniques, the ANN continually updates its weights and biases to improve its predictions over time. Overall, ANNs are highly versatile and can be used for a variety of tasks, including classification, regression, and image recognition.**

## ****Improving**** Artificial Neural Network (Grid Search)

Text, application

Description automatically generated

**Grid search is a method for hyperparameter tuning that is widely used in machine learning. It essentially involves specifying a range of values for each hyperparameter, and the algorithm will search for the combination of hyperparameters that provides the best performance. This is done by running the model multiple times, each time with a different combination of hyperparameters, and evaluating the model's performance using a metric such as accuracy or mean squared error. The process is repeated until the best hyperparameters are found, at which point the model can be re-run using these hyperparameters to get the final results. This approach helps to avoid overfitting and to improve the performance of the model by finding the hyperparameters that work best with the given data.**

## ****Improving**** Artificial Neural Network (K-Fold)

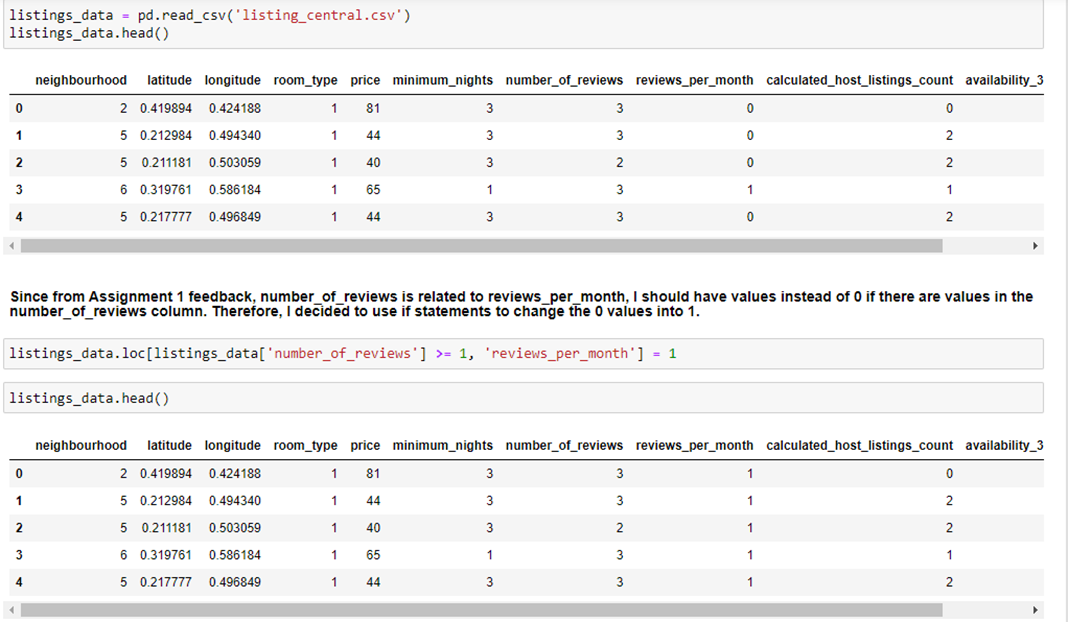
Text

Description automatically generated

**Additionally, the user utilized K-fold cross validation to improve the performance of the ANN model. This technique involves dividing the dataset into k subsets and training the model k times, each time using a different subset as the test set and the remaining subsets as the training set. The results from each iteration are then averaged to obtain a more robust evaluation of the model's performance. By using K-fold cross validation, the user was able to assess the generalization performance of the ANN model and ensure that it was able to accurately predict outcomes on new data.**

## ****Building of Models****

## Importing and Adjusting Dataset

****

**From the feedback the user gotten back from Assignment 1, the user took note of the relationship between number\_of\_reviews and reviews\_per\_month and discovered that they were not aligned. In order to rectify this, the user decided to make a necessary adjustment to ensure that the 0 values in reviews\_per\_month were changed to 1 if there were values in the number\_of\_reviews column. This step was taken to ensure that the results of the analysis would be more accurate and meaningful, considering the relationship between these two important variables. The user's goal was to obtain results that would provide a clearer picture of the relationships between the variables in the dataset and the target variable, which in this case was the price of the Airbnb listings.**

**In additional, from the general feedback received, the outliers for price were not removed and therefore, the user went ahead and eliminated the outliers in the price column. By removing the outliers, the user was able to improve the accuracy of the predictions and get a better understanding of the real trends and patterns in the data.**

**Before:**

Chart

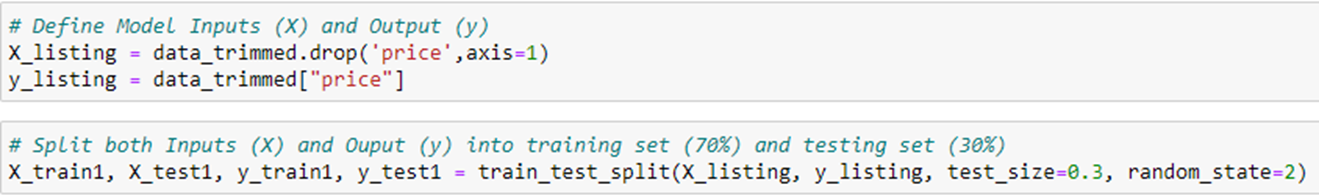
Description automatically generated

**After:**

Chart, histogram

Description automatically generated

**After settling the outliers and the columns relations, the user then split the dataset.**

****

## ****Build the Models****

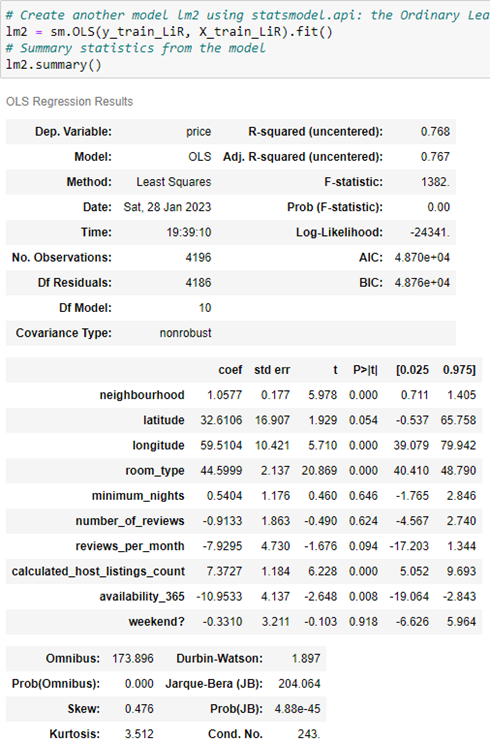
## ****Linear Regression (Base Model)****

Graphical user interface, text, application

Description automatically generated

**The linear regression model was used as the first step in the regression process. This is because linear regression is a simple and straightforward method for modeling the relationship between dependent and independent variables. The user performed the linear regression by fitting the linear equation to the data and observing the relationship between the predictor and target variables. This model was a good starting point for understanding the basic relationship between the variables and could be used to make predictions about the target variable based on the predictor variables.**

## Improving Linear Regression Model (Stats Model)

****

Graphical user interface, text, application

Description automatically generated

**In the improved linear regression model, I utilized the stats Model to assess the significance of each variable through the p-value method. After examining the results, I selected the top three variables with the largest p-value, which were "number of reviews", "weekend", and "minimum nights". I then proceeded to remove these three variables from the dataset in order to obtain a more optimized and efficient model.**

## Decision Tree ****(Base Model)****

Text

Description automatically generated

Table

Description automatically generated with medium confidence

Letter

Description automatically generated with low confidence

**Next model, the user used decision tree. Above are the codes and plots.**

## Improving Decision Tree (Grid Search)

Graphical user interface, text, application, email

Description automatically generated

**For the improvement in decision tree, the user used grid search which is a technique to find the best hyperparameters from the set of values provided. The aim of grid search is to find the best parameters for a machine learning model that result in the highest accuracy. This was done by evaluating a model for each combination of hyperparameters and choosing the combination that yields the best results. Once the best hyperparameters were determined, they were then used to fit the model and improve its performance.**

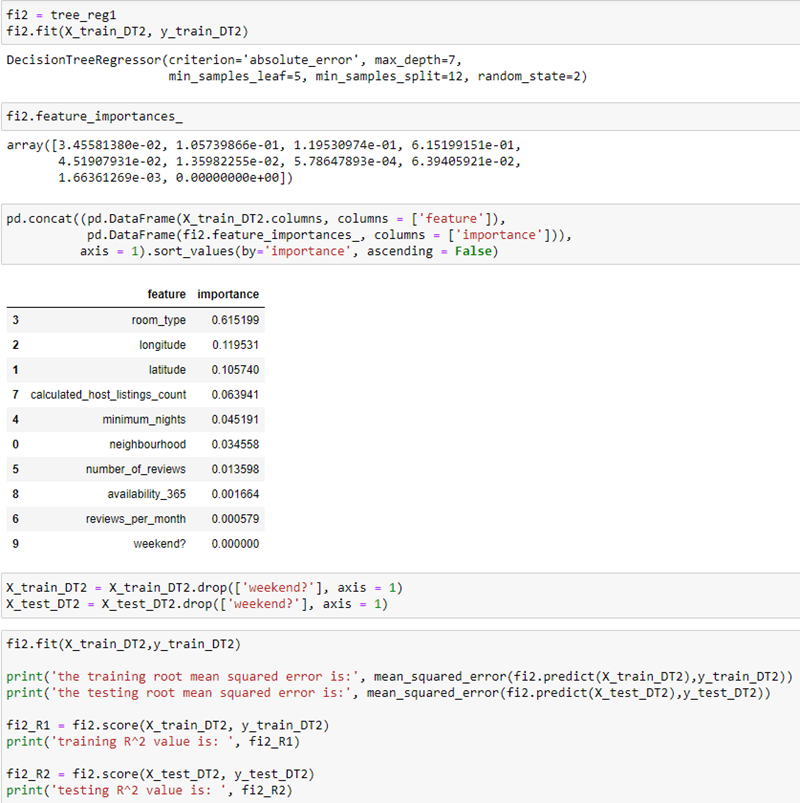
## Improving Decision Tree (K-Fold)

A picture containing text

Description automatically generated

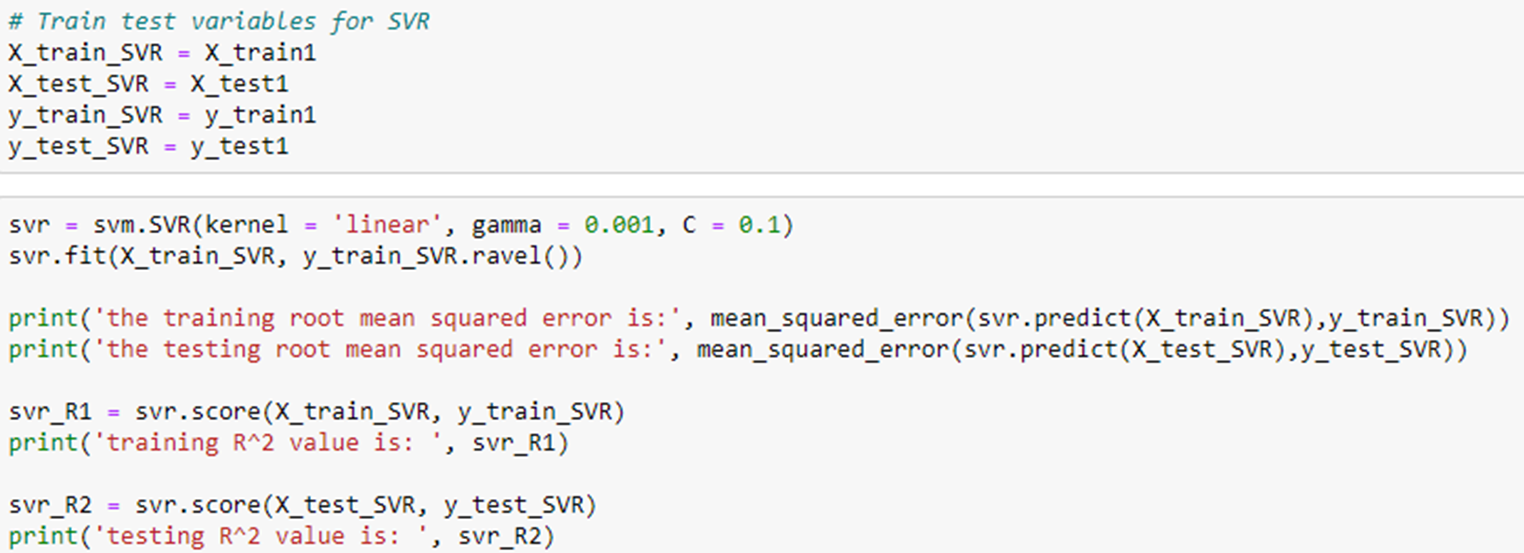
**In addition to using grid search to find the best hyperparameters, the user also used k-fold cross validation to validate the decision tree model. The k-fold cross validation involves dividing the dataset into k subsets, training the model on k-1 subsets, and testing the model on the remaining subset. This process is repeated k times, with each subset used as the test set once. The average accuracy of the model across all k iterations is then calculated, providing a more robust estimate of the model's performance on unseen data. This helps to minimize overfitting and generalize the model's performance to new, unseen data. By using k-fold cross validation, the user was able to better understand the performance of the decision tree model and ensure it was a good fit for the dataset.**

## Improving Decision Tree (Feature Importance)

****

**The last improvement made to the decision tree in regression was using feature importance. This technique involves sorting the features based on their importance and then removing the least important feature. In this case, the least important feature was identified as "weekend" and was removed from the model. This helps to reduce the complexity of the model and prevent overfitting by eliminating features that have limited impact on the target variable. By using feature importance, the user was able to optimize the decision tree and improve its predictive performance.**

## Support Vector Machine (SVR Base)

****

**The user used SVR (Support Vector Regression) from the Support Vector Machine (SVM) algorithm for the next model in regression. The code provided by the user is the base code for implementing SVR. SVR is a type of regression algorithm that is commonly used for solving regression problems, especially for non-linear regression problems. SVR is used to fit a regression model to the training data to make predictions for the target variable. The SVR algorithm uses the concept of Support Vector Machines to fit the regression model to the training data and make predictions. The base code provided by the user is the starting point for implementing SVR to solve the regression problem.**

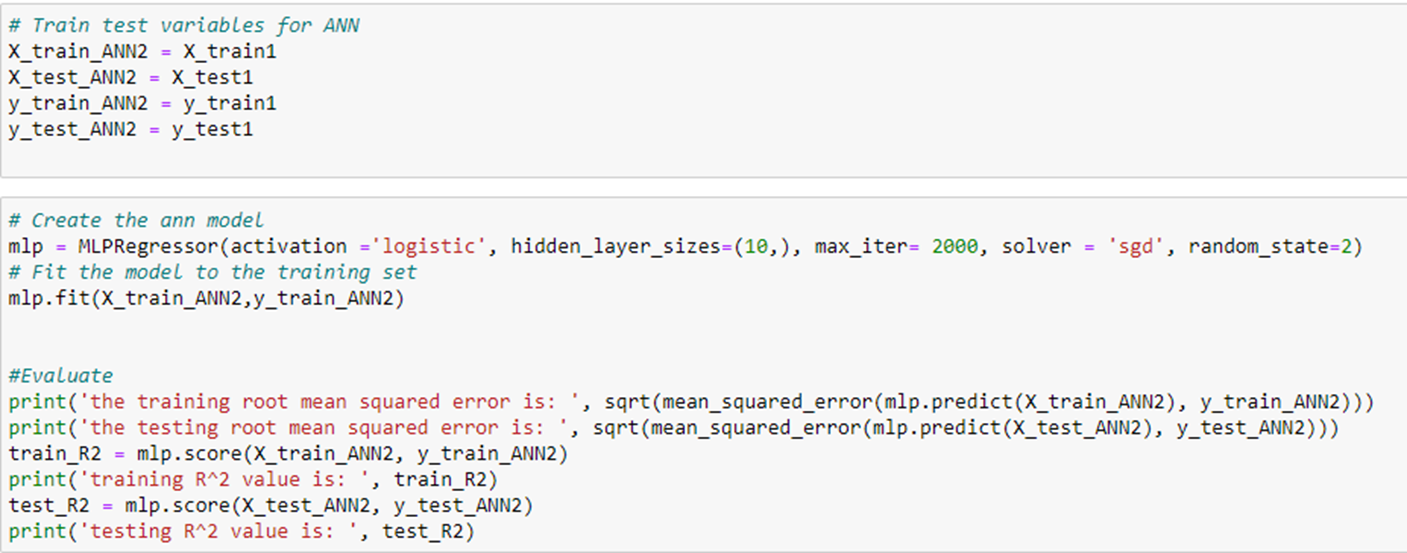
## Improving Support Vector Machine (K-Fold)

Text

Description automatically generated

**For the improvement of the SVR model, the user utilized k-fold cross validation. This technique splits the dataset into a specified number of folds and trains the model on different subsets of the data. The results from each fold are then averaged to obtain a more robust evaluation of the model's performance. By using k-fold cross validation, the user can ensure that the SVR model is being trained and evaluated on a diverse range of data, reducing the risk of overfitting or underfitting. Additionally, it provides a more reliable estimate of the model's accuracy and helps identify any trends or patterns in the performance of the model.**

## Artificial Neural Network (Base)

****

**The last model the user created is an Artificial Neural Network (ANN). The base code represents the foundation for the model. The code sets up the basic structure of the ANN, including the input and output layers, as well as the number of hidden layers and nodes in each layer. The base code is a starting point for the user to build and improve upon. By having a solid base code, the user can experiment with different configurations, algorithms, and parameters to optimize the performance of the model.**

## Improving Artificial Neural Network (Grid Search)

Graphical user interface, text, application

Description automatically generated

**The use of grid search in the ANN model helps to identify the best hyperparameters among the given options, which can significantly impact the accuracy of the model. After finding the best hyperparameters, the model is built using these hyperparameters, providing a more optimized model that is likely to perform better in terms of accuracy and prediction performance. By using grid search, the user is able to fine-tune the hyperparameters in the ANN model to find the best combination that results in the highest accuracy, ensuring that the model is as optimized as possible. This can lead to improved results and better prediction performance, making the model more useful for real-world applications.**

## Improving Artificial Neural Network (K-Fold)

Text

Description automatically generated

**The last improvement the user applied to their model was using K-Fold cross-validation to evaluate the performance of the Artificial Neural Network (ANN) on the dataset. K-Fold cross-validation is a widely used technique in machine learning that helps to overcome the limitations of using a single train-test split. In this technique, the dataset is split into k-folds, where each fold is used as the test set one time, and the remaining k-1 folds are used for training the model. This provides a better estimate of the model's generalization performance, as the model is evaluated on multiple unseen samples. The user used K-Fold cross-validation to evaluate the ANN on their dataset and ensure that their model was able to generalize well to unseen data.**

## ****Further Improvements****

**The XGBoost algorithm is a popular algorithm in the machine learning community, particularly for regression problems. In order to compare the results from the other models and determine if it would be a better fit for the problem at hand, the user decided to add it to their existing models and test its performance after presenting their current models.** **The end goal of adding this last model was to find the best performing model among all the models tested and select the one that would provide the most accurate predictions/performance.**

## HR Analytics

**Graphical user interface, text, application

Description automatically generated**

**Overall Accuracies:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Logistic Regression** | **Decision Tree** | **Support Vector Machine (SVC)** | **Artificial Neural Network** | **XGBoost** |
| **Base:** | **Train: 0.728**  **Test: 0.727** | **Train: 0.670**  **Test: 0.691** | **Train: 0.713**  **Test: 0.714** | **Train: 0.504**  **Test: 0.490** | **(Current Best)**  **Train: 0.823**  **Test: 0.820** |
| **Stats Model**  **(P-value)** | **Train: 0.727**  **Test: 0.726** | **-** | **-** | **-** | **-** |
| **Grid Search** | **-** | **Train: 0.786**  **Test: 0.790** | **-** | **Train: 0.804**  **Test: 0.799** | **-** |
| **K-Fold**  **(Cross-Value)** | **-** | **Train: 0.795**  **Test: 0.790** | **Train: 0.713**  **Test: 0.713** | **(Initial Best)**  **Train: 0.814**  **Test: 0.810** | **-** |
| **Ada Boost** | **-** | **Train: 0.804**  **Test: 0.805** | **-** | **-** | **-** |
| Feature  Importance | **-** | **Train: 0.786**  **Test: 0.790** | **-** | **-** | **-** |

## Airbnb

**Text, application

Description automatically generated**

**Overall Accuracies:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Linear Regression** | **Decision Tree** | **Support Vector Machine (SVR)** | **Artificial Neural Network** | **XGBoost** |
| Base | **Train: 0.111**  **Test: 0.092** | **Train: 0.403**  **Test: 0.380** | **Train: 0.061**  **Test: 0.030** | **Train: 0.031**  **Test: 0.009** | **Train: 0.554**  **Test: 0.458** |
| **Stats Model**  **(P-value)** | **Train: 0.111**  **Test: 0.093** | **-** | **-** | **-** | **-** |
| **Grid Search** | **-** | **Train: 0.536**  **Test: 0.479** | **-** | **Train: 0.111**  **Test: 0.092** | **-** |
| **K-Fold**  **(Cross-Value)** | **-** | **Train: 0.529**  **Test: 0.380** | **Train: 0.056**  **Test: -0.002** | **Train: 0.107**  **Test: 0.036** | **-** |
| **Feature Importance** | **-** | **(Initial/ Current Best)**  **Train: 0.538**  **Test: 0.480** | **-** | **-** |  |

## ****Conclusion****

**To conclude, the user started with a thorough understanding of the data and its characteristics, including identifying any patterns, trends, or relationships between the different features. They then performed data cleaning and preprocessing steps to ensure that the data was ready for analysis. This included filling missing values, removing outliers, and encoding categorical variables. The user then applied various machine learning models, each with its own strengths and weaknesses, to perform the analysis. The Logistic Regression model was used for binary classification problems, the SVC for multi-class classification, and the ANN for regression problems. The user also applied the Decision Tree and SVR models, both of which are commonly used in various machine learning applications. To improve the performance of each model, the user utilized techniques such as Grid Search, K-Fold cross-validation, and feature importance. This allowed them to find the best hyperparameters for each model and evaluate their performance on the dataset. In the end, the user compared the performance of all the models to find the best one. They also added XGBoost as the final model, which helped to validate the results and provide a more comprehensive understanding of the data.**

## ****Reflection****

**Throughout this assignment, the user could have done better and have much more models if it was not for the time constraints. The user also could have done more research for the regression dataset to help improve the models’ performance even more. Overall, the user is still satisfied with what they have achieved for each dataset. Despite these limitations, the user was able to apply various machine learning techniques and make useful predictions on both the classification and regression datasets. The user can continue to build upon their knowledge and improve their skills by attempting similar assignments and exploring new and more advanced techniques.**

## ****Discussion Board****

**Graphical user interface, text, application, email

Description automatically generated**

**Text, letter

Description automatically generated**

**Graphical user interface, text, application, email

Description automatically generated**