Vacation Preferences Prediction with Classification Model (Mountains/Beaches)

Introduction

This project aims to build a classification model to predict an individual's preference for vacation venues (mountains/beaches) based on demographics. The dataset is collected from Kaggle, containing over 50K records of individual demographics (e.g. age, preferred activities, favourite season, etc.) and their preference (1 means prefers mountains, 0 for beaches)

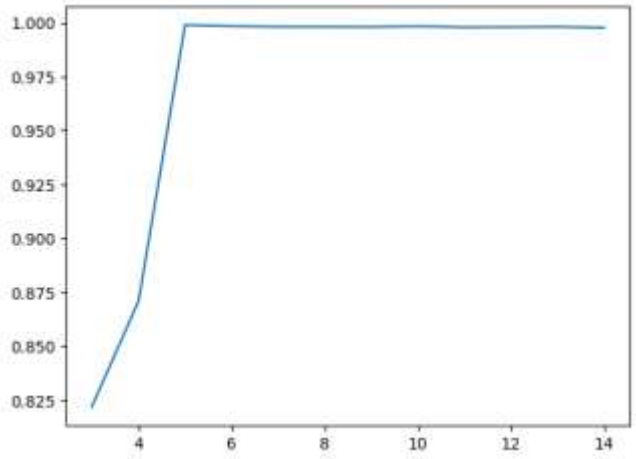
Building a Classification Model to Predict an Individual's Preference (Mountains/Beaches)

Firstly, The dataset is inspected to identify duplicated and blank rows and deal with them if they exist. Then, the target class (Preference) is inspected and found to be **unbalanced**, with about 75% preferring beaches, indicating that the **accuracy score is not a suitable metric** to evaluate the data model as it may misleadingly indicate high performance resulting from correct prediction of the majority class. Consequently, the **confusion matrix** of the predictions and the Area Under the Curve (**AUC**) of the Receiver Operating Characteristic curve (**ROC curve**) will be the **key metrics** to evaluate the data models in this case.

Next, feature engineering is conducted to prepare the dataset for training. The categorical features having more than two unique categories (Preferred_Activities, Location, etc.) are one-hot encoded and the last variables produced *(last category)* are removed to reduce the redundancy of the features. Highly correlated and low-variance features are also detected and removed from the dataset. Lastly, different **numbers of features with the strongest relationship to the target** feature are tested and as a result, **N = 5** is the optimum number of features for the model performance *(accuracy score is used for this testing of features selection for its simplicity, the performance of the models built will not be evaluated by using this approach)*. The 5 features selected and the line graph of different numbers of features used in the model and their respective accuracy score are shown below:

```
['Proximity_to_Mountains', 'Proximity_to_Beaches', 'hiking', 'skiing', 'sunbathing']
```

Selected Features



▲ Numbers of Features Used in the Model vs Accuracy Score

After determining the features of the data model, different algorithms are used to build the classification model, including Logistic Regression, Decision Tree Classifier and K-Nearest Neighbors Classifier, with $\frac{2}{3}$ of the dataset as the training set and the remaining as the testing set. The algorithms are then evaluated for their prediction on the testing set with the confusion matrix, precision and recall scores, ROC curve and AUC. The result shows that the **Logistic Regression** classification model performs the **best**, with its **least false classification in the confusion matrix** and **highest AUC** in the ROC curve (0.**999990**). This model is then further evaluated with **cross-validation** and the result shows the model's **consistent performance** with the **average AUC = 0.99990** with a **standard deviation of 0.0000072.**

Logistic Regression

```
[[13030     5]
 [   13  4259]]
```

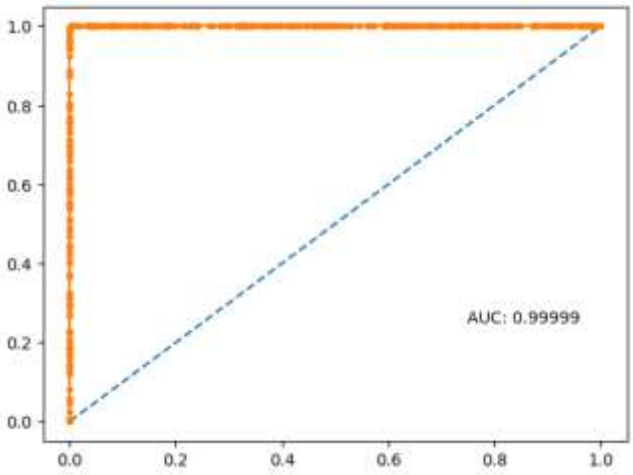
Decision Tree Classifier

```
[[12993    42]
 [   34  4238]]
```

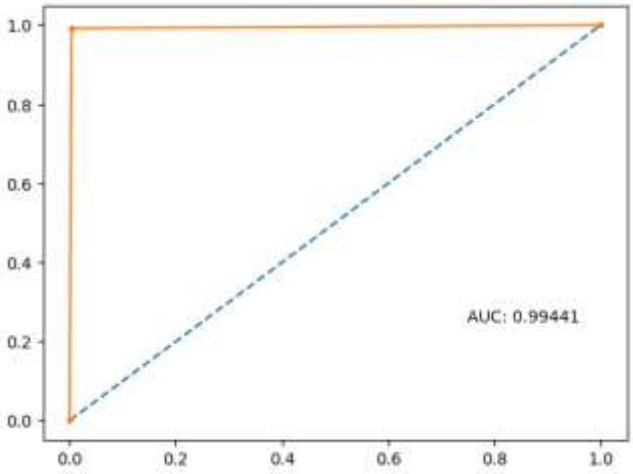
K-Nearest Neighbors

```
[[13015    20]
 [   24  4248]]
```

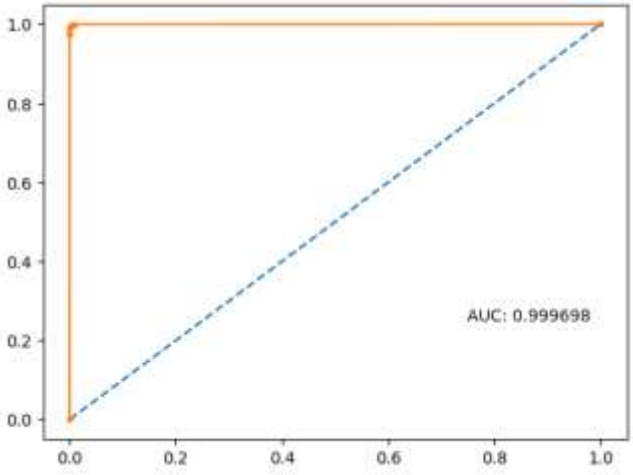
▲ Confusion Matrix



▲ Logistic Regression ROC Curve



▲ Decision Tree Classifier ROC Curve



▲ K-Nearest Neighbors Classifier ROC Curve