Assignment 2

Min-Wei,Li

Data 1

Data 1

- data_id = 41283

- Description : The task of Data 1 is to predict whether a customer will churn, with the target variable indicating "churned" or "not churned." This dataset includes 20 features: 6 nominal features such as state and international_plan, and 14 numeric features such as account_length, call usage, and charge data. It contains 5000 samples with no missing values.

- https://www.openml.org/search?type=data&sort=runs&status=any&qualities.NumberOfInstances=between_1000_10000&qualities.NumberOfClasses=%3D_2&order=asc&id=41283

Tuned Models:

Decision Tree:

- Parameter: min_samples_leaf
- Range: [10, 30, 50, 100, 150, 200, 250]
- Purpose: Controls the minimum number of samples required in a leaf node. Helps to prevent overfitting by limiting model complexity.

K-Nearest Neighbors (KNN):

- Parameter: n_neighbors
- Range: (10, 30, 50, 100, 200)
- Purpose: Defines the number of nearest neighbors to consider. Balances model sensitivity with smoothness by adjusting the number of neighbors.

Non-Tuned Models:

Naive Bayes:

- Parameter Tuning: None
- Reason: Default parameters are typically sufficient for good performance with MultinomialNB.

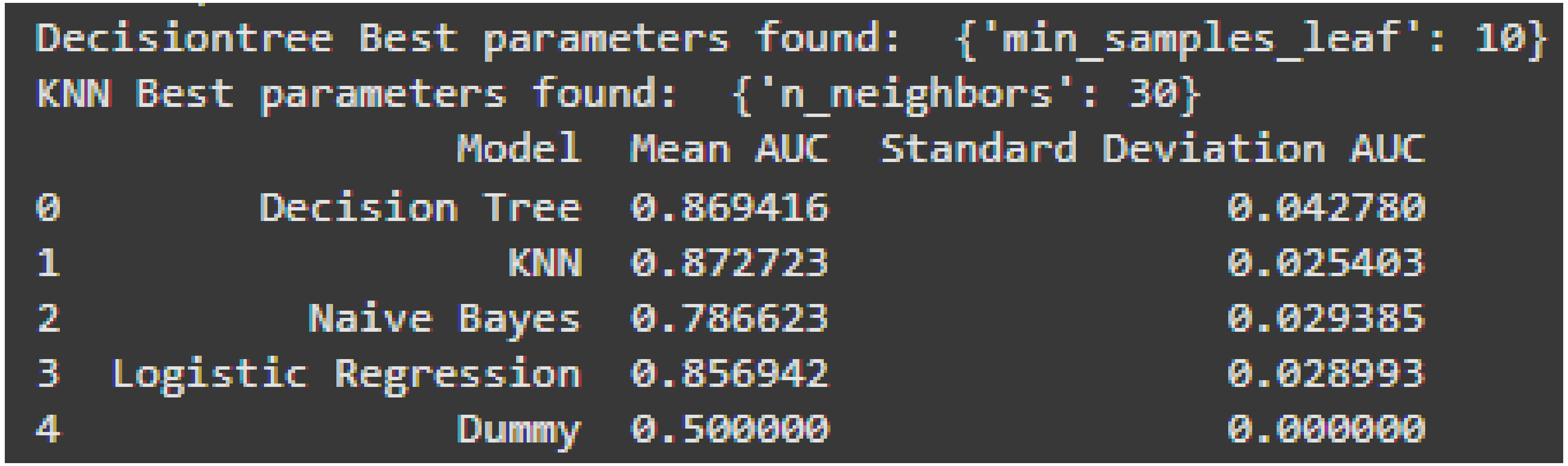
Logistic Regression:

- Parameter Tuning: None
- Reason: Used with default parameters as the model performs reasonably well without additional tuning.

Dummy Classifier:

- Parameter: strategy
- Setting: 'most_frequent’
- Purpose: Provides a baseline by predicting the most frequent class. Used as a simple benchmark to evaluate model performance.

A table showing means and standard deviations



```
Decisiontree Best parameters found:  {'min_samples_leaf': 10}
KNN Best parameters found:  {'n_neighbors': 30}
                 Model  Mean AUC  Standard Deviation AUC
0        Decision Tree  0.869416                0.042780
1                  KNN  0.872723                0.025403
2          Naive Bayes  0.786623                0.029385
3  Logistic Regression  0.856942                0.028993
4                Dummy  0.500000                0.000000
```

Conclusion

- KNN: Best performance with a Mean AUC of 0.872723, optimal parameter n_neighbors=30.
- Decision Tree: Second best, Mean AUC of 0.869416, optimal parameter min_samples_leaf=10.
- Logistic Regression: High stability, Mean AUC of 0.856942.
- Naive Bayes: Moderate performance, Mean AUC of 0.786623.
- Dummy: Baseline model with Mean AUC of 0.5, confirming superior performance of other models.

Data 2

Data 2

- data_id = 41335

- Description : The goal of this dataset is to train a classification model to predict car ratings, specifically to determine if a car belongs to the "very good" (vgood) class based on various features (such as buying cost, maintenance cost, number of doors, etc.).

- https://www.openml.org/search?type=data&sort=runs&status=any&qualities.NumberOfInstances=between_1000_10000&qualities.NumberOfClasses=%3D_2&order=asc&id=41335

Tuned Models:

Decision Tree:

- Parameter: min_samples_leaf
- Range: [10, 30, 50, 100, 150, 200, 250]
- Purpose: Controls the minimum number of samples required in a leaf node. Helps to prevent overfitting by limiting model complexity.

K-Nearest Neighbors (KNN):

- Parameter: n_neighbors
- Range: (10, 30, 50, 100, 200)
- Purpose: Defines the number of nearest neighbors to consider. Balances model sensitivity with smoothness by adjusting the number of neighbors.

Non-Tuned Models:

Naive Bayes:

- Parameter Tuning: None
- Reason: Default parameters are typically sufficient for good performance with MultinomialNB.

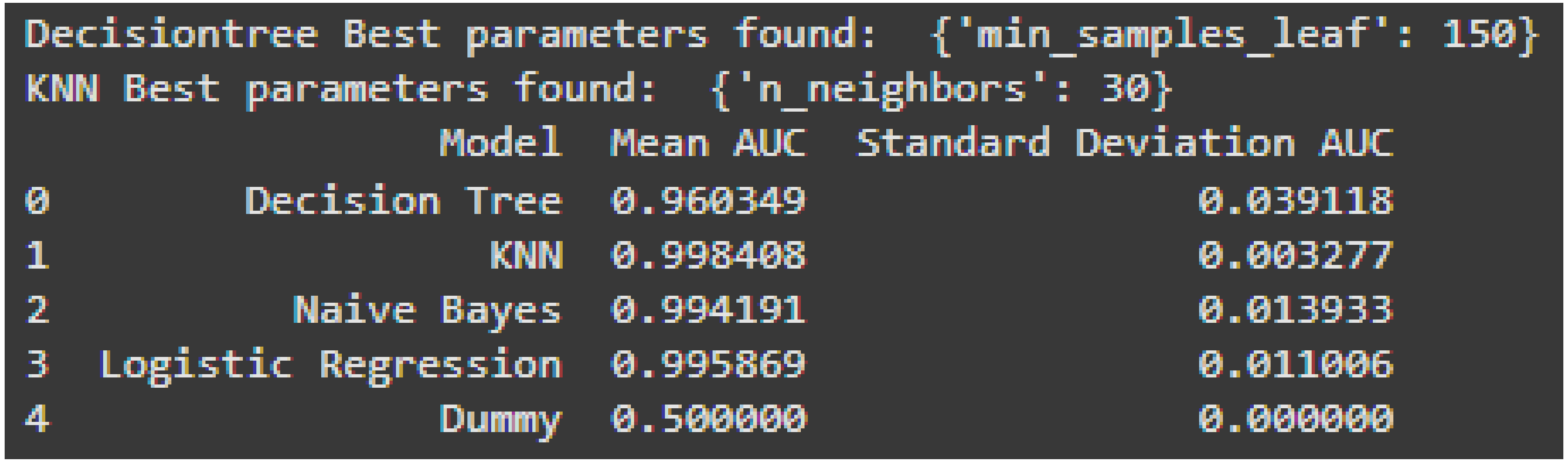
Logistic Regression:

- Parameter Tuning: None
- Reason: Used with default parameters as the model performs reasonably well without additional tuning.

Dummy Classifier:

- Parameter: strategy
- Setting: 'most_frequent’
- Purpose: Provides a baseline by predicting the most frequent class. Used as a simple benchmark to evaluate model performance.

A table showing means and standard deviations



```
Decisiontree Best parameters found:  {'min_samples_leaf': 150}
KNN Best parameters found:  {'n_neighbors': 30}
                 Model  Mean AUC  Standard Deviation AUC
0        Decision Tree  0.960349                0.039118
1                  KNN  0.998408                0.003277
2          Naive Bayes  0.994191                0.013933
3  Logistic Regression  0.995869                0.011006
4                Dummy  0.500000                0.000000
```

Conclusion

- KNN: Best performance with a Mean AUC of 0.998408, optimal parameter n_neighbors=30, and the lowest standard deviation, indicating stable results.
- Logistic Regression: Second best with a Mean AUC of 0.995869 and standard deviation of 0.011006, showing stable performance.
- Naive Bayes: Mean AUC of 0.994191, slightly lower than KNN and Logistic Regression, but still performs well.
- Decision Tree: Mean AUC of 0.960349, optimal parameter min_samples_leaf=150, slightly lower performance but still good.
- Dummy: Baseline model with a Mean AUC of 0.5, indicating that other models perform significantly better in predicting customer churn than random classification.