Module 7 exercises: Model Selection and Evaluation

Part 1 – Evaluating Regression

A Random Forest regression model has been built to predict the price of a train based upon its destination, duration, and the type of ticket bought.

The following evaluation metrics have been generated for the model. What does each metric tell you? How do you feel about its performance?

The value of the target, price, varies between €16.75 and €342.80.

```
Root Mean Squared Error: €12.68
Mean Absolute Error    : € 8.86
R^2                    :   0.73
```

The model appears to be fairly useful with an R^2 of 0.73. Without further understanding the distribution of prices, the average error stats suggest the model is quite good at predicting price, though €12.68 is a large average error for a train ticket.

Part 2 – Evaluating classification

A) Confusion matrices

Interpret the following confusion matrix for a Logistic Regression model fit to the mortgage applicants dataset:
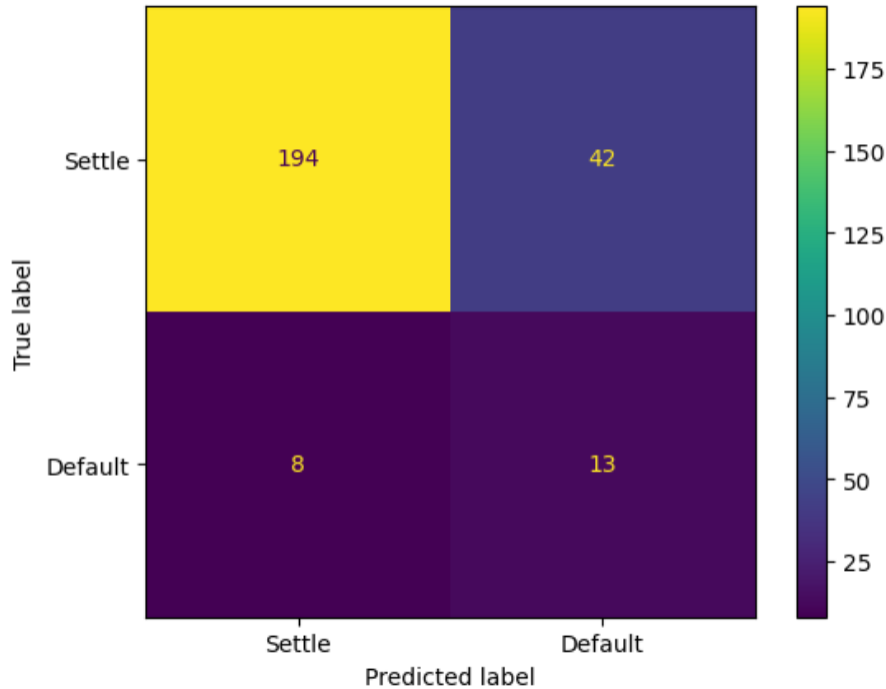
There are very few Default examples in the test dataset. The model predicts well overall, though makes a relatively high number of mistakes in identifying Default (False Negatives) and similarly in identifying Settle (False Positives). A lot of room for improvement!

B) Accuracy, precision, and recall

Evaluate the following classification models, both fit to the mortgage applicants dataset. Which is most performant?

```
LogisticRegression(class_weight='balanced')
                precision    recall  f1-score   support

       Settle       0.96      0.82      0.89       236
      Default       0.24      0.62      0.34        21

     accuracy                           0.81       257
    macro avg       0.60      0.72      0.61       257
 weighted avg       0.90      0.81      0.84       257
```

```
DecisionTreeClassifier(min_samples_leaf=20)
              precision    recall  f1-score   support

      Settle       0.95      0.97      0.96       236
     Default       0.53      0.43      0.47        21

    accuracy                           0.92       257
   macro avg       0.74      0.70      0.72       257
weighted avg       0.92      0.92      0.92       257
```

```
RandomForestClassifier(min_samples_leaf=10, n_estimators=30)
              precision    recall  f1-score   support

      Settle       0.95      0.99      0.97       236
     Default       0.73      0.38      0.50        21

    accuracy                           0.94       257
   macro avg       0.84      0.68      0.73       257
weighted avg       0.93      0.94      0.93       257
```
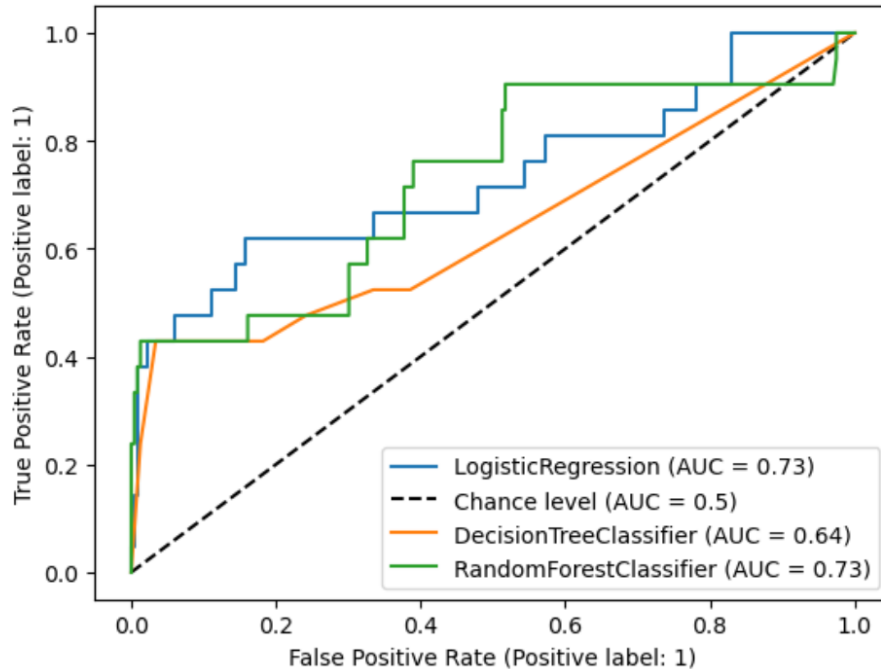
Random Forest appears to perform best overall, with the highest F1 score and Accuracy. Organisational perspective may lead to a different conclusion being drawn.

C) ROC and AUC

Interpret the following ROC plot.

All models predict better than randomly. Logistic Regression and Random Forest have similar performance overall, with Logistic regression more performant at high decision threshold values and Random forest more performant at lower ones.

Part 3 – Models in production

1. Consider the mortgage applicants dataset. What might you want to test about a model you intend to deploy before doing so?

	ID	Income	Term	Balance	Debt	Score	Default
0	567	17626	10 Years	1381	293	228.0	False
1	523	18959	20 Years	883	1012	187.0	False
2	544	20560	10 Years	684	898	86.0	False
3	370	21894	10 Years	748	85	NaN	False
4	756	24430	10 Years	1224	59	504.0	False
...	...	...	...	...	...	...	...
851	71	30191	20 Years	1319	3880	55.0	True
852	932	41669	20 Years	1385	32	780.0	False
853	39	36816	20 Years	1868	3123	366.0	True
854	283	42145	20 Years	1447	2498	422.0	False
855	847	30594	20 Years	1216	2473	179.0	True

856 rows × 7 columns

- *Pertubation testing – Add noise to data and see how model predictions change (lee important here as input quite constrained)*

- *Invariance testing – Change the value of e.g. user ID and ensure prediction of model doesn't change*

- *Directional Expectation testing – Increase income, check likelihood of mortgage approval increases etc.*

Part 4 – Extension

1. Revisit the models you generated in the Regression and Classification exercises and evaluate their performance.