ccdefault Data Frame

1. Compare the models' performances (e.g., AUC).

Along our study we have compared three different models: Binomial Logistic regression, Decision Tree and Random Forest, obtained its Root Mean Squared Error (RMSE) and the prediction probability of each client being default. Now we are going to describe briefly how these different models work and the results we obtained:

- Binomial Logistic Regression: we chose the Binomial Logistic Regression instead of the Multinomial Logistic Regression because by means of using the sigmoid function it is able to distinguish between two classes as we just have default equals to 0 or 1, we will not need further classification or computing effort.

- Decision Tree: it is a flow-chart tree like structure which is constructed in a top-down recursive divide-and-conquer manner. To assure that the result is not overfitted it is needed to have a test set and training set of data.

- Random Forest: it builds multiple decision trees that are most of the time trained with the bagging method. Each tree is just making a model with some of the features of the database, not all of them, and for every tree the selected features are changed randomly. After computing all the multiple trees, it merges them together to get a more accurate and stable prediction.

Logistic regression (λ=0.01; α = 0.1; iter=10)		**Decision tree regression**	**Random forest (numTrees = 40)**	
RMSE	**ROC**	**RMSE**	**RMSE**	**RMSE crossval.**
0.000	0.999	0.000	0.067	0.020

Table 1. Results after the implementation of three models.

2. Defend your choice of best model (e.g., what are the strength and weaknesses of each of these models?).

The weaknesses of logistic regression are that the independent variables must be properly selected, otherwise the model will have little to no predictive value. Also, that these variables must be independent from one to the other because if not, the model will tend to overweight those interdependent variables.

For decision trees is very important to choose the right input parameters because a small variation in the input can cause a big difference in the output. Yet, decision trees are very useful in classification problems, but it is always necessary to have a test set to overcome overfitting of the model.

On the other side, random forest is used to reduce overfitting as it creates a set of trees with low variance and low biases. However, it is not an easy to interpret model.

Regarding our results for ccdefault we can say that the best model for the provided dataset would be or binomial logistic regression or decision tree model as the provided RMSE is 0.000 by both. Having such a low result would make us think that the model is overfitted, but as our dataset was split in a train and test set, we can be confident enough that it is not.

3. What more would you do with this data? Anything to help you devise a better solution?

A way to improve our algorithm would be by adding cross validation to test for the selected model/s and from the output we obtain re-train our model to improve its performance. In the model, we used our test set to train our hyperparameters so that to avoid overfitting and improve regularization. However, these steps should be executed in a loop until the best result is obtained.