**Tell us how you validate your model, which, and why you chose such evaluation technique(s).**

The scoring mechanism that I used was ROC AUC, which is suitable for classification problems with imbalanced classes. I used k-fold cross validation (5 folds) to choose the best algorithm as well as to tune the model parameters of the chosen model. Initially I had only done validation using a single train test split, but the results fluctuated depending on how the dataset was split. K-fold cross validation overcomes this problem by repeating this process multiple times so that the average ROC AUC score obtained is a good representation of the performance of the model.

**What is AUC? Why do you think AUC was used as the evaluation metric for such a problem? What are other metrics that you think would also be suitable for this competition?**

AUC is area under the curve. ROC AUC is the area under the receiver operating characteristic curve (ROC). ROC is a plot of the true positive rate against the false positive rate. It is a measurement of the performance of the classification model at various threshold settings. It is the ideal evaluation metric for imbalanced classification problems as compared to a simple accuracy metric. For example, for a classification problem that is 99% majority class and 1% minority class, just classifying every observation to be 0 (majority class) will give an overall accuracy of 99%, which sounds high, but the number of observations we have gotten right for the minority class would in fact be 0.

The F1 score, which is a harmonic mean of the precision and recall rates would also be suitable for this competition. However, the probability thresholds might need to be adjusted to obtain the best F1 score, as the default threshold of 50% might not be the most optimal.

**What insight(s) do you have from your model? What is your preliminary analysis of the given dataset?**

Tuning a model does not bring about significant improvements to the model performance. Having better features or more data contributes more significantly to model improvement than model tuning. The number of observations is large but the number of features is on the low side. However, the available features are strong predictors of the outcome and hence the high AUC score.

**Can you get into the top 100 of the private leaderboard, or even higher?**

Currently I am not in the top 100 but it is possible with better/more feature engineering. The feature engineering I have done are to drop the income column, fill in missing values for number of dependents and remove entries where ages are below 20.