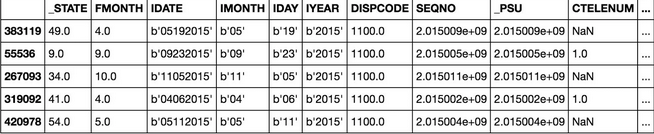
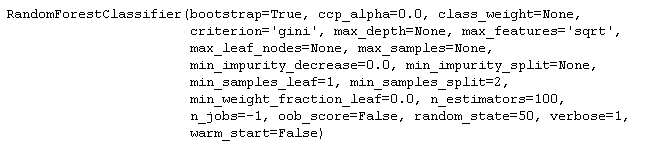
Данные взятые с <https://s3.amazonaws.com/projects-rf/clean_data.csv>



Мы попробовали рассчитать состояние здоровья пациентов в бинарной системе координат. В качестве параметров мы используем социально-экономические и персональные характеристики субъектов. В качестве меток мы используем 0 для плохого здоровья и 1 для хорошего.

В нашем случае стоит проблема несбалансированной классификации (количество объектов одного из классов существенно превосходит количество объектов другого класса.) поэтому мы используем ROC и AUC (<https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc>). 0 – плохо, 1 – хорошо, 0.5 - случайность.

Для решения задачи использовали случайны лес из 100 деревьев с параметрами После финального обучения модели и тестирования модели на тестовой выборки получаем следующий график.

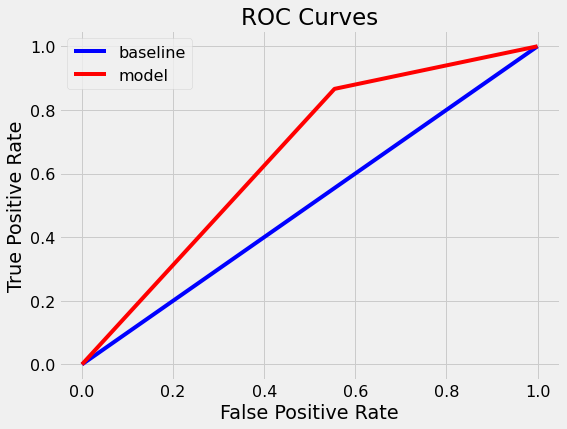


Рисунок 1 ROC для дерева

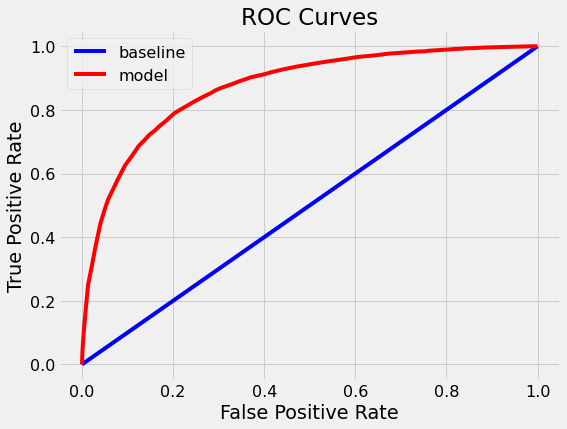


Рисунок 2 ROC для леса

На данном графике можно увидеть, что лес лучше на 0,20. Эффективность у обоих алгоритмов равна 100% так как мы не ограничили глубину дерева и дали дереву готовые ответы, следовательно, переобучили лес.

У случайного леса более низкая погрешность решений и лес лучше обобщает информацию. Так же лес более точный это можно увидеть, посмотрев на матрицу погрешности.

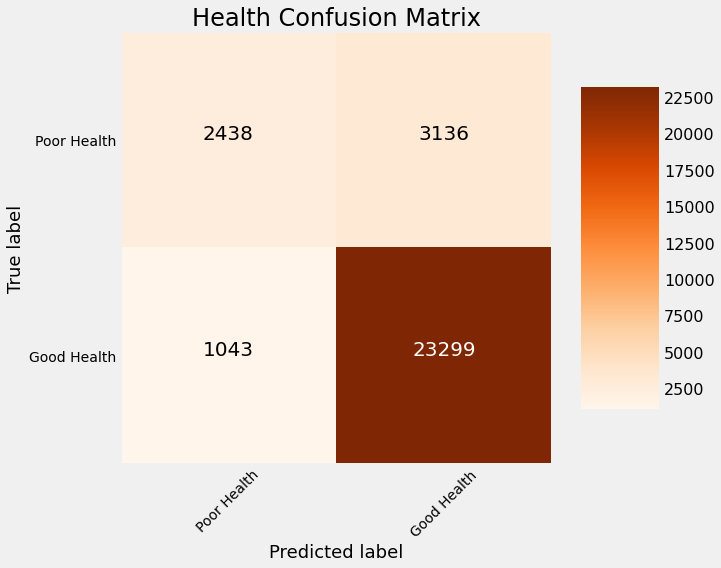


Рисунок 3 Матрица погрешности для леса

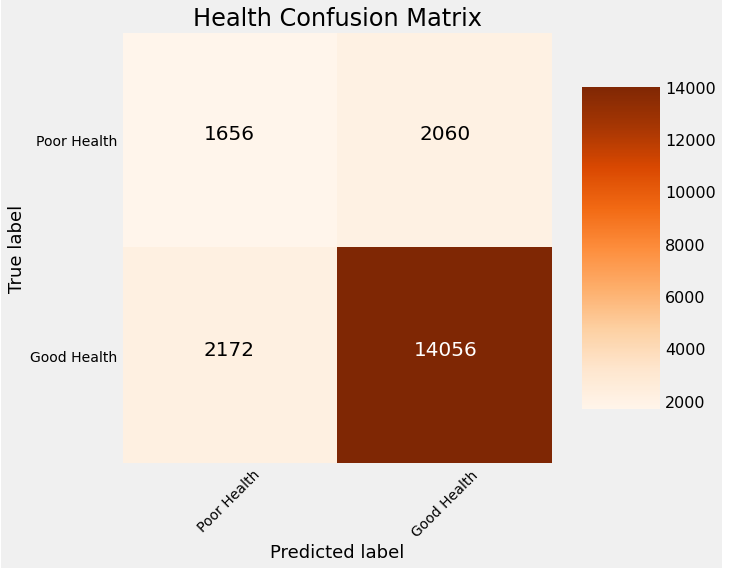


Рисунок 4 Матрица погрешности для дерева

На диаграмме верные прогнозы, сделанные моделью, отображаются в верхнем левом углу и в нижнем правом, а неверные в нижнем левом и верхнем правом.