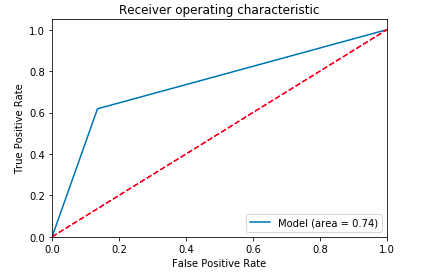
**Варіант 2**

**Дерево рішень**

Спершу було побудовано дерево рішень на основі всіх колонок, було отримано наступну точність 0.803083 та наступну ROC-криву:

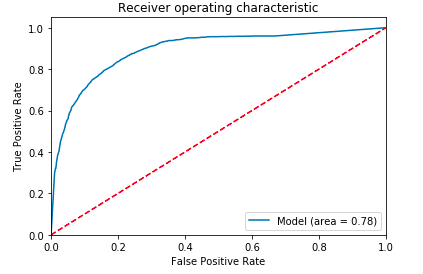


Далі було побудовано (2^14 - 1 ) дерев рішень, кожен з яких на вхід приймав певну підмножину із колонок (всього колонок 14) вхідних даних та оцінював результуючу змінну.

Для побудованих моделей оцінювалась точність їх роботи, як відношення кількості правильно класифікованих записів до їх загальної кількості із тестової вибірки.

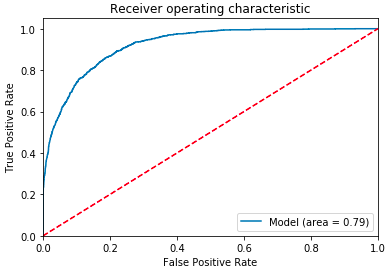
Загалом точність варіювалась від 0.668987 (['age', 'workclass', 'fnlwgt']) до 0.857782 (['workclass', 'education-num', 'marital-status', 'occupation', 'capital-gain', 'capital-loss'])

Для найкращої моделі було проведено проведено ROC-аналіз, який має наступний вигляд:



**AdaBoost**

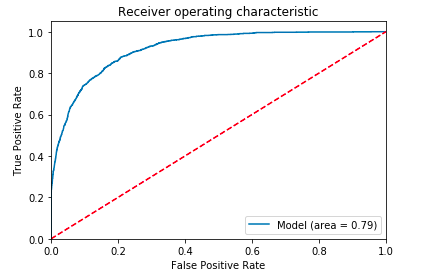
Спершу було побудовано AdaBoost на основі всіх колонок, було отримано наступну точність 0.862258 та наступну ROC-криву:



Далі аналогічно було побудовано моделі на підмножинах колонок.

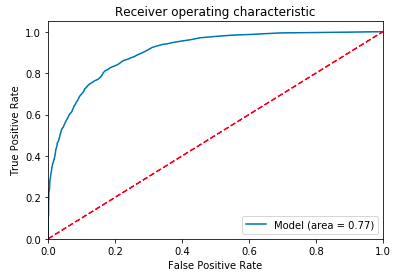
Загалом точність варіювалась від 0.742417 ['fnlwgt', 'marital-status', 'race', 'sex', 'native-country'] до 0.865407 (['age', 'workclass', 'fnlwgt', 'education-num', 'marital-status', 'occupation', 'relationship', 'capital-gain', 'capital-loss'])

Для найкращої моделі було проведено проведено ROC-аналіз, який має наступний вигляд:



**Випаковий ліс**

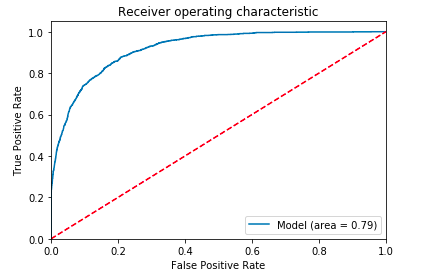
Спершу було побудовано випадковий ліс на основі всіх колонок, було отримано наступну точність 0.850655 та наступну ROC-криву:



Далі аналогічно було побудовано моделі на підмножинах колонок.

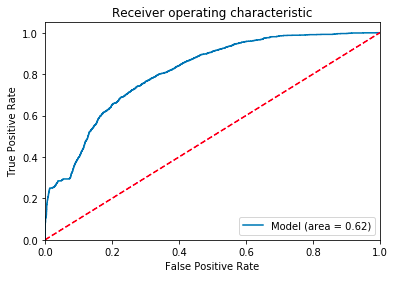
Загалом точність варіювалась від 0.666335 (['workclass', 'fnlwgt']) до 0.857782 (['education-num', 'occupation', 'relationship', 'capital-gain', 'capital-loss']).

Для найкращої моделі було проведено проведено ROC-аналіз, який має наступний вигляд:



**Машина опорних векторів**

Було побудовано машину опорних векторів на основі всіх числових колонок, було отримано наступну точність 0.791149 та наступну ROC-криву:



Порівняння алгоритмів класифікації:

Вибравши найкращі моделі з кожного типу класифікаторів, отримуємо наступні дані:

|  |  |  |
| --- | --- | --- |
| Модель | Точність | Площа під ROC-кривою |
| Дерево рішень на основі всіх колонок | 0.803083 | 0.74 |
| Дерево рішень на колонках ['workclass', 'education-num', 'marital-status', 'occupation', 'capital-gain', 'capital-loss'] | 0.857782 | 0.78 |
| AdaBoost | 0.862258 | 0.79 |
| AdaBoost на колонках ['age', 'workclass', 'fnlwgt', 'education-num', 'marital-status', 'occupation', 'relationship', 'capital-gain', 'capital-loss'] | 0.865407 | 0.79 |
| Випадковий ліс | 0.850655 | 0.77 |
| Випадковий ліс на колонках ['education-num', 'occupation', 'relationship', 'capital-gain', 'capital-loss'] | 0.857782 | 0.77 |
| Машина опорних векторів | 0.791149 | 0.62 |
|  |  |  |

Бачимо, що із в цілому всі методи дають високу точність, при цьому моделі можуть бути чутливим до вибору колонок. При оптимальному виборі колонок більшість моделей дали практично однакову точність, близьку 0.85.

Найкращою моделлю серед побудованих виявилась AdaBoost на колонках ['age', 'workclass', 'fnlwgt', 'education-num', 'marital-status', 'occupation', 'relationship', 'capital-gain', 'capital-loss'].