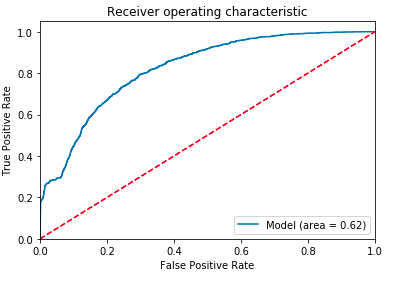
**Варіант 2**

**Байесів наївний класифікатор**

Спершу було побудовано байесів класифікатор на основі всіх колонок, було отримано наступну точність 0.785679 та наступну ROC-криву:

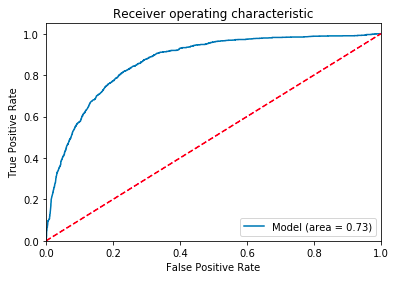


Далі було побудовано (2^14 - 1 ) наївних байесівських класифікатори, кожен з яких на вхід приймав певну підмножину із колонок (всього колонок 14) вхідних даних та оцінював результуючу змінну.

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

Загалом точність варіювалась від 0.645782 (['education', 'race', 'sex']) до 0.822642 (['age', 'education-num', 'marital-status', 'hours-per-week']). При використанні всіх колонок точність моделі становила 0.785679.

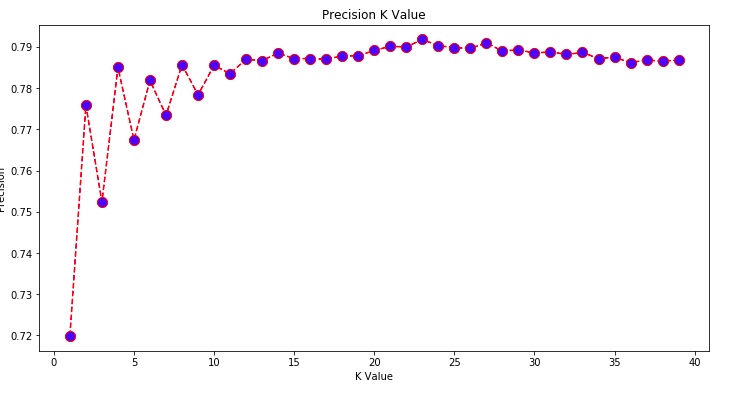
Для найкращої моделі було проведено проведено ROC-аналіз, який має наступний вигляд:



**k-nn класифікатор**

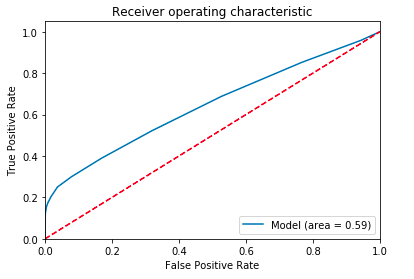
Для k-nn класифікатора було побудовано 40 моделей, змінюючи k від 1 до 40, що використовувли лише числові колонки. Загалом точність варіювалась від 0.719874 (k = 1) до 0.791812 (k = 23) .

Далі побудуємо графік, як змінювалась точність в залежності від значення k:



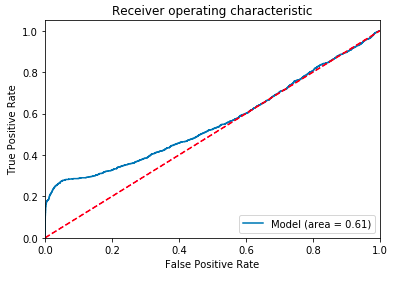
Бачимо, що в загальному точність росте при збільшенні k і починає стабілізуватись при k>12, тому пропонуємо взяти k = 14 як більш оптимальну модель, адже його точність мало відрізняється від найбільш оптимальної (0.788497 для k = 14 та 0.791812 для k = 23) проте буде вимагати менших обчислень під час використання.

Для цієї моделі побудуємо ROC-криву:



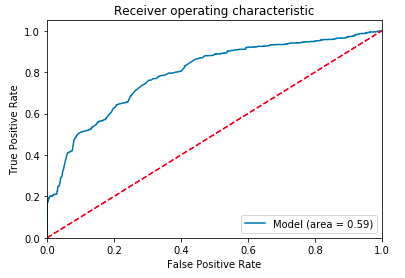
**Логістична регресія**

Із використанням логістичної регресії спершу було побудовано модель, що використовувала лише числові колонки. Для неї було отримано точність 0.787502 і наступну ROC-криву:



Далі ми використовували факторизовані дані колонок рядкового типу. І знову перебравши всі можливі підмножини колонок побудували моделі, найгірша з яких мала точність 0.712912 (['marital-status', 'sex']), а найкраща — 0.827615 (['education-num', 'occupation', 'relationship', 'capital-gain', 'native-country']).

Для найкращої побудували ROC-криву:



Також варто зауважити, що побудова всіх моделей логістичної регресії тривала значно довше, аніж побудова моделей байесових класифікаторів. (Точний час не засікався).

Порівняння алгоритмів класифікації:

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

|  |  |  |
| --- | --- | --- |
| Модель | Точність | Площа під ROC-кривою |
| Наївний байесів класифікатор на основі всіх колонок | 0.785679 | 0.62 |
| Наївний байесів класифікатор на колонках ['age', 'education-num', 'marital-status', 'hours-per-week'] | 0.822642 | 0.73 |
| K-nn класифікатор при k = 14 | 0.788497 | 0.59 |
| Логістична регресія на основі числових колонок | 0.787502 | 0.61 |
| Логістична регресія на колонках ['education-num', 'occupation', 'relationship', 'capital-gain', 'native-country'] | 0.827615 | 0.59 |

Бачимо, що із в цілому всі методи дають високу точність, при цьому байесів класифікатор та логістична регресія можуть бути чутливим до вибору колонок. K-nn моделі можуть залежати від k, проте з ростом k точність стабілізується, тому оптимально вибирати значення k, де відбулась дана стабілізація.

Стосовно точності найркащими були моделі наївний байесів класифікатор та логістична регресія при певних колонках. Проте з точки зору ROC-аналізу найкращою буде наївний байесів класифікатор на колонках ['age', 'education-num', 'marital-status', 'hours-per-week'].

**Додаткове завдання**

В межах додаткового завдання для виділення ознак було здійснено обчислення частоти використання слів у вхідних повідомленнях для тренувальної вибірки.

Отримані частоти використовувались як набір ознак для байесового наївного класифікатора. Побудована модель мала дуже високу точність рівну 0.986547 і наступна ROC-криву:

