**Федеральное государственное автономное образовательное**

**учреждение**

**высшего образования**

**"Национальный исследовательский университет**

**"Высшая школа экономики"**

Московский институт электроники и математики им. А.Н. Тихонова НИУ ВШЭ

Департамент компьютерной инженерии

**Курс: Проектный семинар:**

**“Python в науке о данных“**

ОТЧЕТ

по поиску оптимальной модели прогнозирования

Содержание

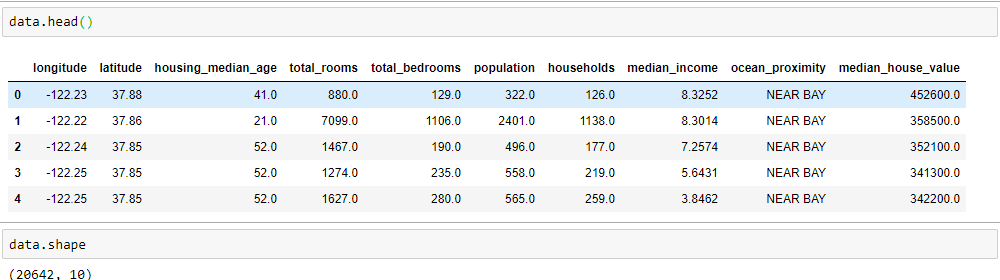
Описание данных и выбор признаков……………………………………………....3

Метрики качества классификации……………….………………….........................5

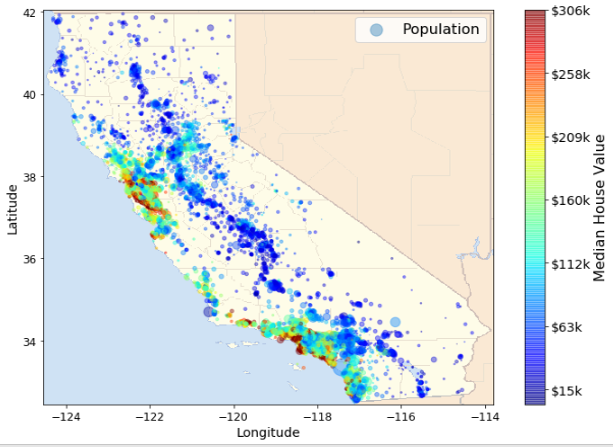
Выбор алгоритма……………….……………………………….............................................6

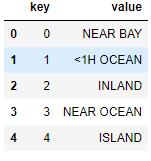
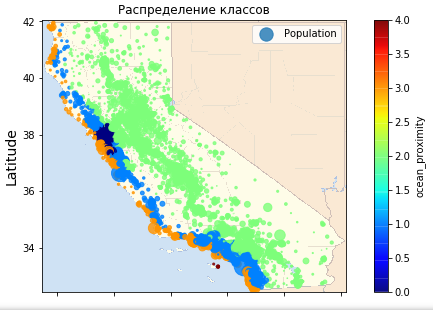
Описание данных и выбор признаков

Исходные данные:

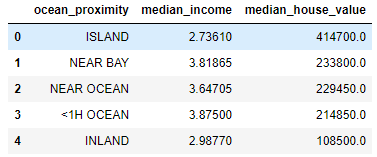


Визуализация:





Как можно заметить, данные выражают вполне логичную закономерность: дома, находящиеся ближе к океану стоят дороже, чем дома, находящиеся в глубине материка 🡪



Но стоит заметить, что распределение среднего дохода не слишком логично, как признак, характеризующий цену дома.

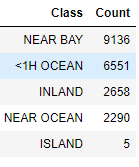
В нашем проекте была поставлена задача классифицировать дома на зоны проживания, пользуясь признаками в данных, которые вводит пользователь

В нашем исследовании были рассмотрены три варианта выборки признаков:

1. Классификация по долготе и широте
2. Классификация по долготе, широте, среднему доходу и средней стоимости дома
3. Классификация по всем признакам

Также была определена проблема в данных:

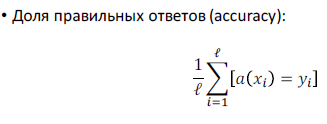
1. Несбалансированность классов:

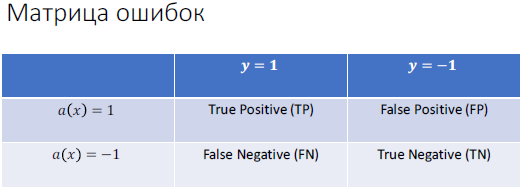


Например, могут возникнуть проблемы с предсказанием класса ISLAND

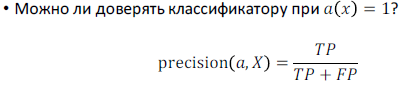
Метрики качества классификации

1. Accuracy score

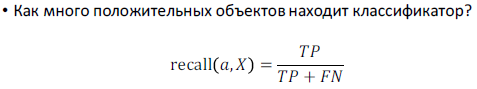




1. Precision score



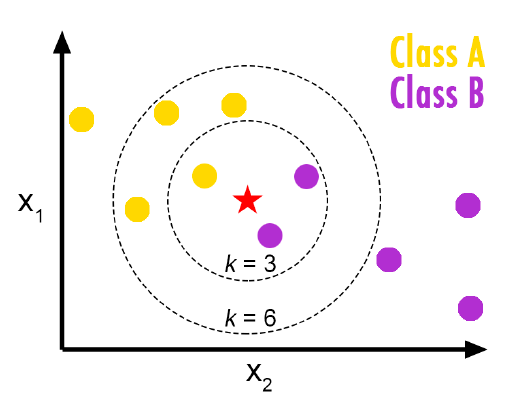
1. Recall score



Качество исследования будет проверятся на всех трех метриках, но стоит отметить, что в данной ситуации можно было бы обойтись только Precision score.

Выбор алгоритма

1. Метод k – ближайших соседей:



Обучим knn на признаках набора 2 (см. 3 варианта выборки признаков), качество прогноза будет оцениваться на тестовом наборе данных, равном 1 / 3 от всего датасета, число соседей во всех вариантах равно 3:



Обучим knn на признаках набора 3:



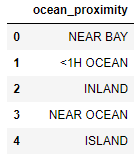
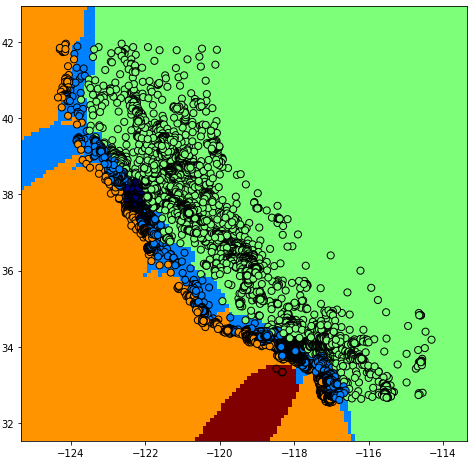
Результаты не слишком впечатляют. Скорее всего дело в том, что выбранные наборы признаков не репрезентативны для предсказания класса. Это значит, что их не стоит использовать для дальнейших проверок гипотез

Попробуем обучить knn на признаках набора 1:



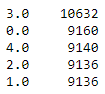
Похоже, что мы нашли оптимальный набор признаков для предсказаний

Так же можно посмотреть на примере разделяющей поверхности, как knn распределил объекты на классы по обучающей и тестовой выборке вместе:



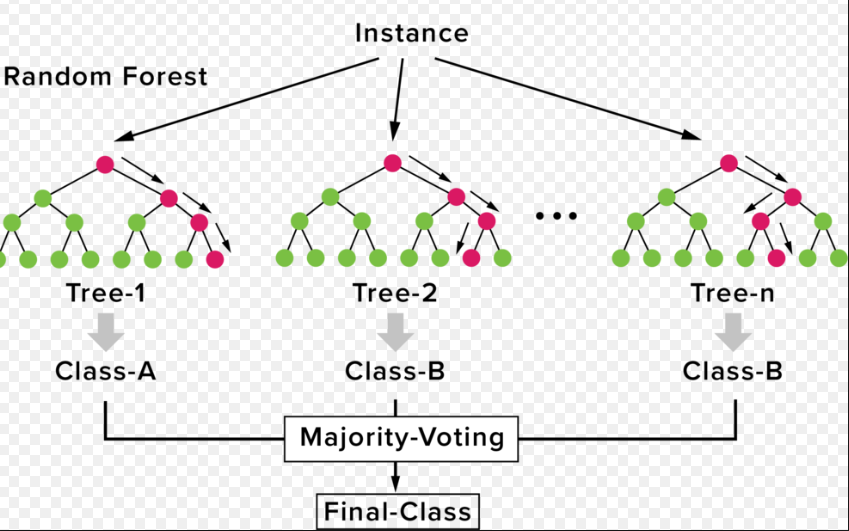
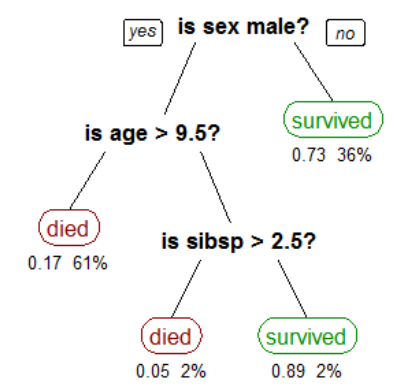
Наш скор уже достаточно высок, но можно сделать его еще выше, сбалансировав классы:



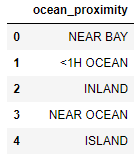
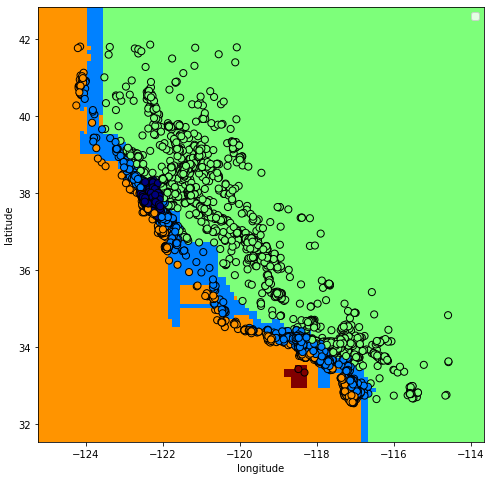
 

Как можно видеть, решение проблемы несбалансированность классов привело к построению самой точной модели

1. Случайный лес



Обучим случайный лес на признаках 1 набора, число деревьев = 500, максимальная глубина дерева = 20:





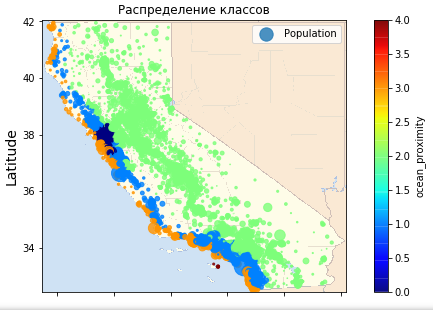
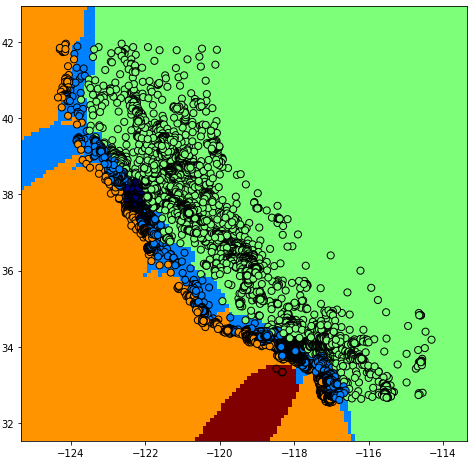
Сразу видна отличительная особенность работы алгоритма случайного леса от knn: случайный лес разбивает выборку линиями, параллельными осям longitude и latitude, что позволяет ему определить границы классов четче, чем knn.

Попробуем улучшить модель, сбалансировав классы:



Как мы можем видеть, оба алгоритма показывают 100 процентную точность классификации, но все-таки в нашем проекте будет использоваться алгоритм случайного леса, потому что его разбиение уменьшит ошибку на новых данных: например, рассмотрим результат разбиения knn в диапозоне от 38 до 40 по latitude и от 0 до -124 по longitude. Если взглянуть на карту распределения классов, то с человеческой точки зрения данный фрагмент карты(красный квадрат) должен быть оранжевого цвета, как на разбиении случайного леса.

Из-за этого может получиться неверное определение класса



Также посмотрим на карту распределения классов, к сожалению, мы не можем решить проблему с неправильным определением класса квартир на острове, выделенном маленьким красным квадратом, такие объекты будут отнесены к классу NEAR OCEAN, а не к ISLAND. Решить данную проблему не представляется возможным для наших данных из-за нехватки информации для данного региона. Хотя стоит отметить, что проблема могла бы быть решена, если бы вручную было добавлено несколько красных точек на остров, выделенный маленьким красным квадратом.