

Описание и обработка данных.

Для анализа выбран набор данных о подержанных автомобилях в Великобритании. Всего в датасете 6267 объявлений о продаже автомобиля марки «Skoda». Набор данных содержит информацию о модели, годе выпуска, цене, трансмиссии, пробеге, типе топлива, дорожном налоге, расходе топлива и объеме двигателя. Повторяющиеся списки и столбцы в данных не содержатся.

| model | year | price | transmission | mileage | fuelType | tax |
|--------------|------|-------|--------------|---------|----------|-----|
| Octavia | 2017 | 10550 | Manual | 25250 | Petrol | 150 |
| Citigo | 2018 | 8200 | Manual | 1264 | Petrol | 145 |
| Octavia | 2019 | 15650 | Automatic | 6825 | Diesel | 145 |
| Yeti Outdoor | 2015 | 14000 | Automatic | 28431 | Diesel | 165 |
| Superb | 2019 | 18350 | Manual | 10912 | Petrol | 150 |
| Yeti Outdoor | 2017 | 13250 | Automatic | 47005 | Diesel | 145 |
| Superb | 2019 | 15250 | Manual | 14850 | Petrol | 145 |
| Octavia | 2019 | 18950 | Automatic | 5850 | Diesel | 150 |
| Kodiaq | 2019 | 29900 | Automatic | 2633 | Petrol | 150 |

Рисунок 1. Первые 9 строк датасета.

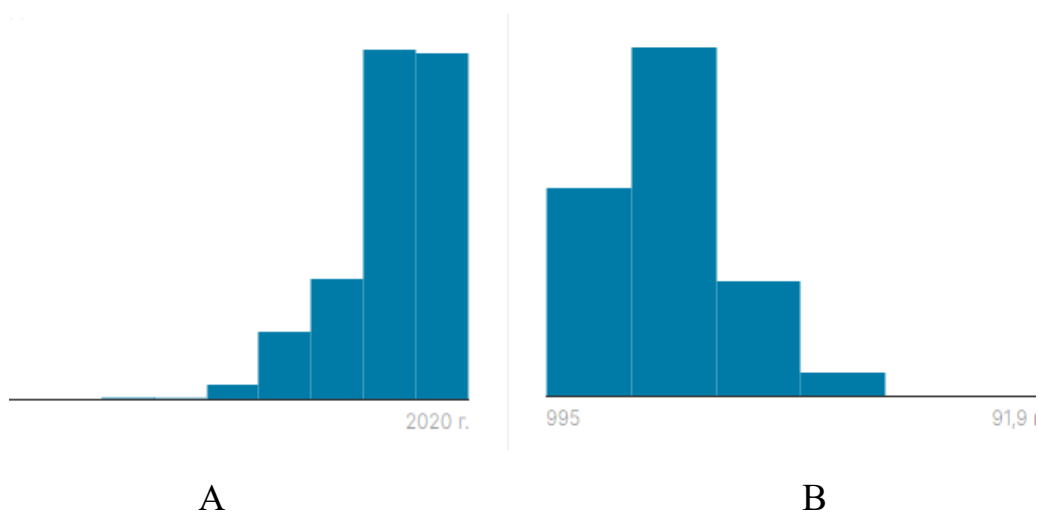


Рисунок 2. Гистограммы распределения по годам (А) и по цене (В).

Целевым признаком выбрана цена автомобиля. Все остальные признаки были выбраны входными, кроме признака «Модель». Далее все входные признаки были бинаризованы.

1. Год выпуска. Год выпуска был преобразован в два признака: >2010года, > 2017 года. Первые 10 строк получившего датасета представлены на рисунке 3.

| | price | transmission | mileage | fuelType | ... | mpg | engineSize | year_2010 | year_2017 |
|---|-------|--------------|---------|----------|-----|------|------------|-----------|-----------|
| 0 | 10550 | Manual | 25250 | Petrol | ... | 54.3 | 1.4 | 1 | 0 |
| 1 | 8200 | Manual | 1264 | Petrol | ... | 67.3 | 1.0 | 1 | 1 |
| 2 | 15650 | Automatic | 6825 | Diesel | ... | 67.3 | 2.0 | 1 | 1 |
| 3 | 14000 | Automatic | 28431 | Diesel | ... | 51.4 | 2.0 | 1 | 0 |
| 4 | 18350 | Manual | 10912 | Petrol | ... | 40.9 | 1.5 | 1 | 1 |
| 5 | 13250 | Automatic | 47005 | Diesel | ... | 51.4 | 2.0 | 1 | 0 |
| 6 | 15250 | Manual | 14850 | Petrol | ... | 40.9 | 1.5 | 1 | 1 |
| 7 | 18950 | Automatic | 5850 | Diesel | ... | 50.4 | 2.0 | 1 | 1 |
| 8 | 29900 | Automatic | 2633 | Petrol | ... | 31.4 | 2.0 | 1 | 1 |
| 9 | 18990 | Manual | 20000 | Petrol | ... | 43.5 | 2.0 | 1 | 0 |

Рисунок 3.

2. Пробег. Признак пробег был разделен на следующие признаки: >15000, >50000, >100000, >200000. Фрагмент полученного датасета приведен на рисунке 4.

| | price | transmission | fuelType | tax | ... | mileage_1 | mileage_2 | mileage_3 | mileage_4 |
|---|-------|--------------|----------|-----|-----|-----------|-----------|-----------|-----------|
| 0 | 10550 | Manual | Petrol | 150 | ... | 1 | 0 | 1 | 1 |
| 1 | 8200 | Manual | Petrol | 145 | ... | 0 | 0 | 0 | 0 |
| 2 | 15650 | Automatic | Diesel | 145 | ... | 0 | 0 | 0 | 0 |
| 3 | 14000 | Automatic | Diesel | 165 | ... | 1 | 0 | 1 | 1 |
| 4 | 18350 | Manual | Petrol | 150 | ... | 0 | 0 | 1 | 0 |
| 5 | 13250 | Automatic | Diesel | 145 | ... | 1 | 0 | 1 | 1 |
| 6 | 15250 | Manual | Petrol | 145 | ... | 0 | 0 | 1 | 0 |
| 7 | 18950 | Automatic | Diesel | 150 | ... | 0 | 0 | 0 | 0 |
| 8 | 29900 | Automatic | Petrol | 150 | ... | 0 | 0 | 0 | 0 |
| 9 | 18990 | Manual | Petrol | 150 | ... | 1 | 0 | 1 | 0 |

Рисунок 4.

3. Расход топлива. Данный признак разделен на следующие признаки: >57, >70. Фрагмент полученного датасета приведен на рисунке 5.

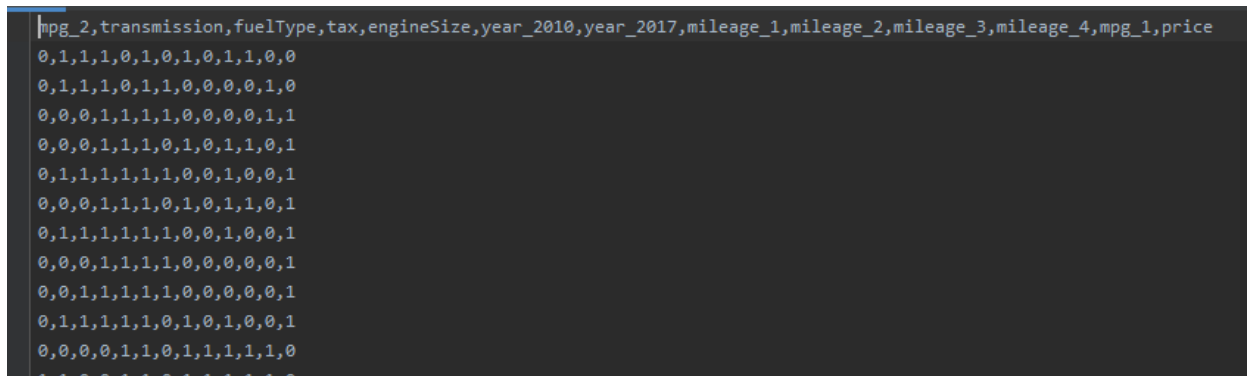
| | price | transmission | fuelType | tax | ... | mileage_3 | mileage_4 | mpg_1 | mpg_2 |
|---|-------|--------------|----------|-----|-----|-----------|-----------|-------|-------|
| 0 | 10550 | Manual | Petrol | 150 | ... | 1 | 1 | 0 | 0 |
| 1 | 8200 | Manual | Petrol | 145 | ... | 0 | 0 | 1 | 0 |
| 2 | 15650 | Automatic | Diesel | 145 | ... | 0 | 0 | 1 | 0 |
| 3 | 14000 | Automatic | Diesel | 165 | ... | 1 | 1 | 0 | 0 |
| 4 | 18350 | Manual | Petrol | 150 | ... | 1 | 0 | 0 | 0 |
| 5 | 13250 | Automatic | Diesel | 145 | ... | 1 | 1 | 0 | 0 |
| 6 | 15250 | Manual | Petrol | 145 | ... | 1 | 0 | 0 | 0 |
| 7 | 18950 | Automatic | Diesel | 150 | ... | 0 | 0 | 0 | 0 |
| 8 | 29900 | Automatic | Petrol | 150 | ... | 0 | 0 | 0 | 0 |
| 9 | 18990 | Manual | Petrol | 150 | ... | 1 | 0 | 0 | 0 |

Рисунок 5

4. Объем двигателя. Признак бинаризован в соответствии со средним медианным значением равным 1,45.
5. Налог. Признак бинаризован в соответствии со средним медианным значением равным 144.

6. Тип топлива. Бинаризован в соответствии: бензин(Petrol)=1, дизель (Diesel) = 0.
7. Трансмиссия. Бинаризован в соответствии: механика(Manual)=1, автомат (Automatic) = 0.
8. Цена. Признак выбран как целевой. Бинаризован в соответствии со средним медианным значением, для того что бы обучение проходило для примерно одинакового количества положительных и отрицательных исходов. Среднее медианное значение равно 12998.

На рисунке 6 представлен датасет после бинаризации всех признаков.



| mpg_2 | transmission | fuelType | tax | engineSize | year_2010 | year_2017 | mileage_1 | mileage_2 | mileage_3 | mileage_4 | mpg_1 | price |
|-------|--------------|----------|-----|------------|-----------|-----------|-----------|-----------|-----------|-----------|-------|-------|
| 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 |
| 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 |
| 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 |
| 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 |

Рисунок 6.

Первичная работа с данными.

- Для проверки качества классификации на имеющихся данных, данные были разделены случайным образом на обучающую и тестовую выборки в соотношении 80% на 20%, соответственно.
- Обучающая выборка была разделена на 2 класса: $K_{positive}$ и $K_{negative}$. в соответствии со значением целевого признака.

Алгоритм 1. (simple_classifier)

Простой и быстрый алгоритм. Описание классифицируемого объекта из тестовой выборки сравнивается по очереди с описанием каждого примера из положительного контекста. В соответствии с размером пересечения классифицируемому объекту добавляется нормированный голос за положительную классификацию. Далее каждый классифицируемый объект сравнивается по очереди с описанием каждого примера из отрицательного контекста и, в соответствии с размером пересечения, классифицируемому объекту добавляется нормированный голос за отрицательную классификацию. Последним шагом является сравнение вычисленных голосов и принятие решения.

$$S^+ = S^+ + \frac{|g' \cap g'_+|}{|g'|}$$

$$S^- = S^- + \frac{|g' \cap g'_-|}{|g'|}$$

Где g – классифицируемый объект, g_+ , g_- – объекты из положительной и отрицательной выборки.

```
Accuracy: 0.7877094972067039
ROC AUC: 0.7879507850041529
True Positive: 476
True Negative: 511
False Positive: 111
False Negative: 155
Precision: 0.8109028960817717
Recall: 0.7543581616481775
```

Рисунок 7. Вывод при работе алгоритма 1.

Алгоритм 2.

Более сложный и медленный алгоритм. Вначале формируются пересечения описания классифицируемого объекта в описания из «+» и «-» контекстов. После для каждого пересечения проверяется количество вложений данного пересечения в описания примеров противоположенного контекста, то есть для положительного контекста в отрицательном, а для отрицательного в положительном. Далее вводится порог, с учетом которого будет проходить голосование. Если количество вложений не преодолевает порог, то данный пример из положительного контекста «голосует» за классифицируемый объект. Аналогично, для примеров из отрицательного контекста. Далее суммы голосов нормируются. Классификация происходит уже по сравнению нормированных голосов.

$$if \frac{|g' \cap g'_{+/-} \subseteq g'_{+/-}|}{|G_{-/+}|} < C, then votes(+|-) += 1$$

$$\frac{votes(+)}{|G_+|} <> \frac{votes(-)}{|G_-|}$$

Так как алгоритм является трудоемким и входные данные имеют большой объем, было решено оценивать размер поддержки путем случайного выбора объектов из выборки. Также решено исследовать зависимость

точности классификации от размера выборки. Тем самым можно найти баланс между временем счета и качеством приближения. В таблице 1 представлено значение точности в зависимости от значений порога (по горизонтали) и выбранного количества объектов выборки, в процентах (по вертикали).

Таблица 1.

| | 0.001 | 0.01 | 0.1 | 0.3 | 0.4 | 0.5 | 0.7 | 0.8 | 1 | 2 |
|----|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 5 | 0.565 | 0.65 | 0.771 | 0.803 | 0.814 | 0.81 | 0.789 | 0.794 | 0.796 | 0.79 |
| 10 | 0.543 | 0.62 | 0.779 | 0.82 | 0.817 | 0.806 | 0.791 | 0.793 | 0.794 | 0.79 |
| 20 | 0.521 | 0.636 | 0.776 | 0.821 | 0.817 | 0.810 | 0.79 | 0.79 | 0.794 | 0.791 |
| 30 | 0.514 | 0.632 | 0.774 | 0.825 | 0.812 | 0.806 | 0.790 | 0.79 | 0.794 | 0.791 |
| 40 | 0.513 | 0.642 | 0.775 | 0.822 | 0.815 | 0.810 | 0.789 | 0.79 | 0.793 | 0.79 |

В процессе анализа работы алгоритмов выяснилось, что наиболее оптимальным пороговым значением является порог равный 0.3 при этом случайно выбирались 30% выборки. Этот порог и был выбран для дальнейшего сравнения алгоритмов.

```

pros= 0.3 coef 0.3
Accuracy: 0.825219473264166
ROC AUC: 0.8251019108280255
True Positive: 485
True Negative: 549
False Positive: 79
False Negative: 140
Precision: 0.8599290780141844
Recall: 0.776

```

Рисунок 8. Вывод при работе алгоритма 2.

Результаты работы алгоритма 1 и алгоритма 2 представлены в таблице 2.

Таблица 2.

| | Алгоритм 1 | Алгоритм 2 при пороге 0.3 |
|---------------|------------|---------------------------|
| Accuracy | 0.787 | 0.825 |
| Roc AUC | 0.787 | 0.825 |
| True positive | 476 | 485 |
| True negative | 511 | 549 |

| | | |
|----------------|-------|-------|
| False positive | 111 | 79 |
| False negative | 155 | 140 |
| Precision | 0.81 | 0.859 |
| Recall | 0.754 | 0.776 |

Сравнение с другими известными алгоритмами.

Также в процессе выполнения задания было решено произвести сравнение со следующими алгоритмами: метод k-ближайших соседей, наивный байесовский метод, дерево решений. Все вышеперечисленные алгоритмы использовались со стандартными параметрами. Результаты работы всех алгоритмов приведены в таблице 3.

| | Алгоритм 1 | Алгоритм 2 при пороге 0,3 | k-NN | Naïve Bayes | Decision True |
|----------------|------------|------------------------------|-------|----------------|------------------|
| Accuracy | 0.787 | 0.825 | 0.889 | 0.601 | 0.892 |
| Roc AUC | 0.787 | 0.825 | 0.887 | 0.582 | 0.892 |
| True positive | 476 | 485 | 591 | 646 | 585 |
| True negative | 511 | 549 | 530 | 108 | 534 |
| False positive | 111 | 79 | 67 | 489 | 63 |
| False negative | 155 | 140 | 66 | 11 | 72 |
| Precision | 0.81 | 0.859 | 0.898 | 0.569 | 0.902 |
| Recall | 0.754 | 0.776 | 0.899 | 0.983 | 0.89 |

По итогу работы можно сделать следующие выводы. Алгоритм 2 дает большую точность предсказания, чем алгоритм 1, но в свою очередь является более ресурсоемким, то есть требует длительных вычислений и задействует большую память. Решить эту проблему удалось с помощью случайного выбора объектов из выборки, то есть проходить не по всем объектам датасета, а по случайно выбранным в меньшем количестве. Для выбранного датасета, наилучшем порогом для второго алгоритма является порог 0,3 при этом выбраны 30 % из всех объектов. Тем самым найдено не плохое приближение и сэкономлено время. Алгоритм k-ближайших соседей и алгоритм дерево решений не сильно отличаются по точности от второго алгоритма, но все же работают лучше. Наивный байесовский метод показал наименьшую точность из всех проверенных алгоритмов.