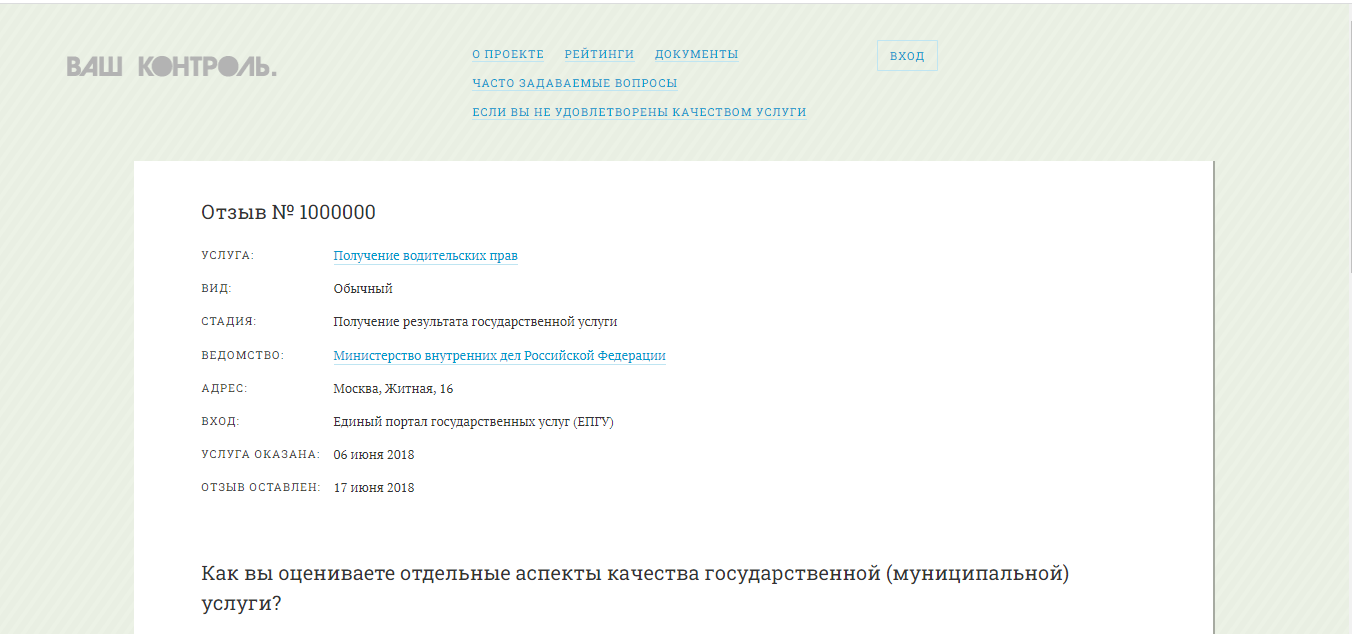
Лабораторная работа №1. Классификация текстов на реальных данных.

Задание 1. Произведите автоматический сбор информации из Интернет-ресурса по отзывам посетителей. В качестве сайта можете брать любой. Например, <https://vashkontrol.ru/about>, <https://www.restoran.ru/msk/opinions/>, Banki.ru и другие

В качестве инструмента для сбора использовался используйте язык Python, а именно, библиотеки Requests (HTTP библиотека, которая нужна для получения доступа к веб-ресурсам) и BeautifulSoup (библиотека Python для извлечения данных из файлов HTML и XML).

Далее идут примеры для сайта <https://vashkontrol.ru/about> (рис. 1)



**Рис. 1 Вид сайта**

Данный сайт был создан в качестве инструмента для оценки эффективности работы государственных и муниципальных органов власти. Пользователи сайта могут оставить отзыв о работе конкретного чиновника, описать, что они думают о работе органов власти, а также изучить мнения других пользователей.

Собранные данные включали:

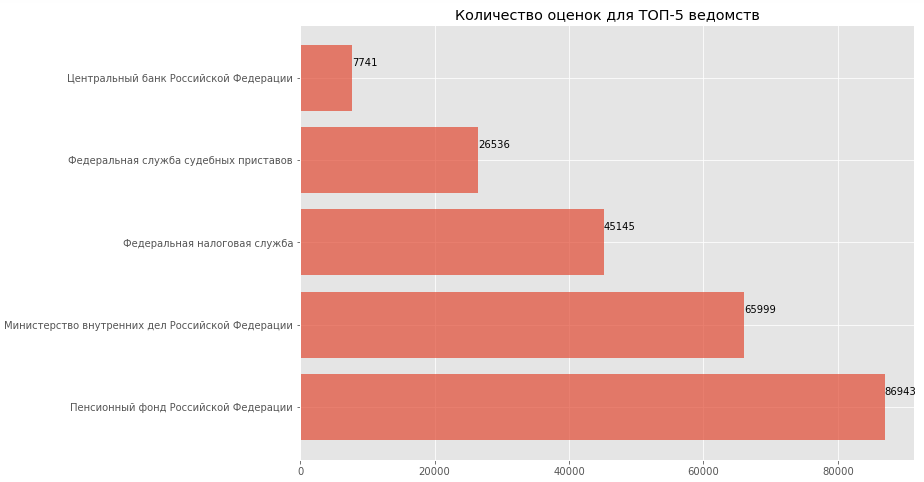
* *service* – услуга, оказанная государственным органом, чиновником,
* *type* – вид, в котором была оказана услуга (очный формат или в электронном виде),
* *stage* – стадия, на которой находится оказание услуги,
* *department* – ведомство, которое оказывает услугу,
* *address* – адрес ведомства,
* *login* – способ авторизации пользователя на сайте (через Единый портал государственных услуг или напрямую через сайт «Ваш контроль»),
* *date of service* – дата оказания услуги ведомством,
* *date of review* – дата, когда пользователем был оставлен отзыв,
* *rate* – средняя оценка качества оказания услуги (5 – «Отлично», 4 – «Хорошо», 3 – «Нормально», 2 – «Плохо», 1 – «Очень плохо»),
* *text* – текст отзыва.

Задание 2. Провидите предпроцессинг данных (очистку данных, удалите выбросов, пустые значения и т.д). Выполните описательную статистику.

Задание 3. Провидите анализ предметной области и сделайте несколько практических выводов, исходя из данных. (Визуализация и выводы)

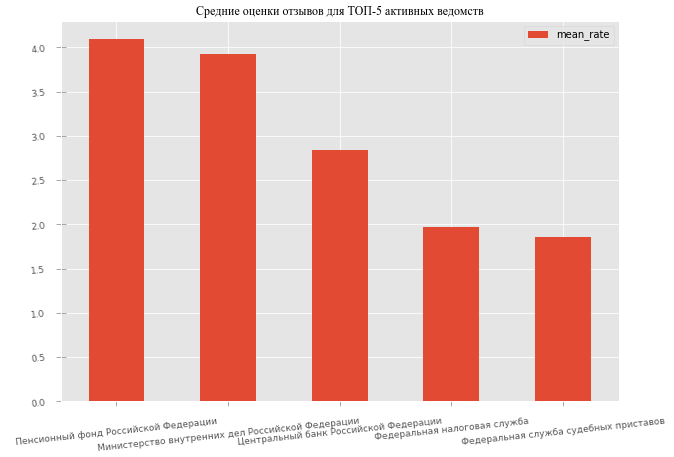
Примеры: рис 1-4.

На линейчатой диаграмме (Рис.1) видно, что наиболее активно люди отзываются о ведомствах федерального значения, а в частности лидером является Пенсионный фонд Российской Федерации.



**Рис. 2 Количество оценок для ТОП-5 ведомств**

Далее выясним, насколько пользователи удовлетворены качеством оказания услуг в каждом из этих ведомств.



**Рис. 3 Средние оценки отзывов для ТОП-5 активных ведомств**

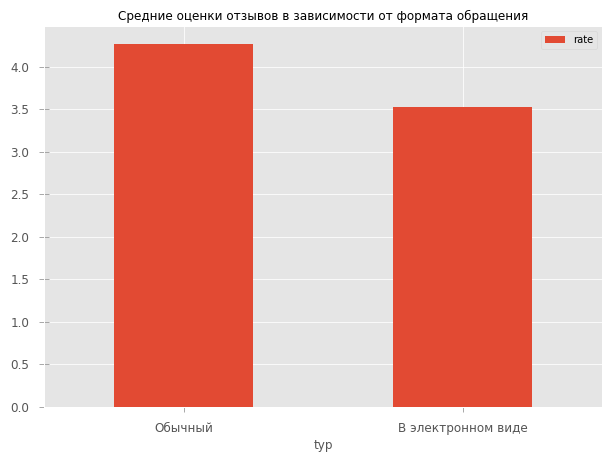
На Рис.2 видно, что самая высокая средняя оценка у Пенсионного фонда – выше 4, то есть «хорошо». Рассмотрим отдельно оценки каждой услуги, которую оказывает данное ведомство. В итоге видно, что единственная низкая оценка у услуги: *«Заявления о выборе инвестиционного портфеля или смене пенсионного фонда»*. Следовательно, данному ведомству стоит обратить дополнительное внимание на детали оказания этой услуги и улучшить работу сотрудников.

Аналогичным образом рассмотрим оценки для Министерства внутренних дел Российской Федерации, так как средняя оценка его услуг также близка к оценке «хорошо». В результате видно, что в данном ведомстве две услуги, оценки которых ниже среднего:

1. *«Миграционный учет иностранных граждан и лиц без гражданства»*
2. *«Заключения об отсутствии судимости»*

Следовательно, МВД стоит улучшить работу сотрудников при оказании этих услуг. Однако при дальнейшем изучении данных выясняется тот факт, что 30,56% оказываемых услуг имеют оценку плохо или ниже, что говорит о серьёзном недовольстве получателей услуг.

Следующим шагом была выявлена разница в качестве оценок для разных типов оказания услуг: для услуг, оказанных в очной форме, средняя оценка – 4,3, а для услуг, оказанных в электронном виде, на порядок ниже – 3,5(Рис.3)

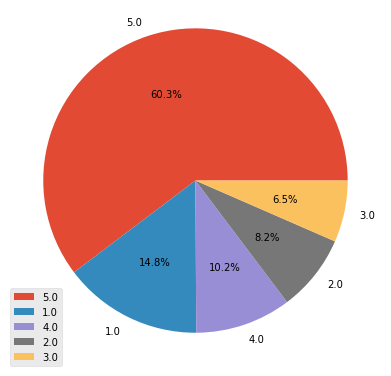


**Рис. 4 Средние оценки отзывов в зависимости от формата обращения**

Таким образом, можно сделать вывод о том, что удовлетворенность клиента услугой выше, если она была оказана в очном формате.

Анализ всей совокупности собранных данных даёт более полную картину о предметной области, что так же важно для практической стороны данного исследования.

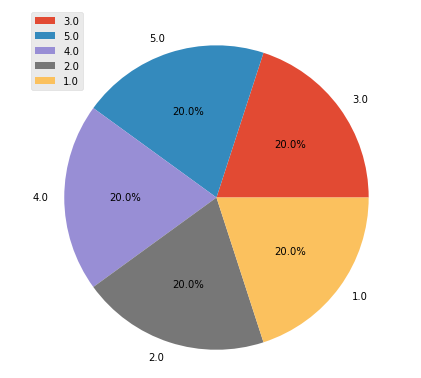
Задание 4. Провидите обработку текста. Сравните классы (рис.6).



**Рис. 6 Доли классов в несбалансированной выборке**

Если есть разбалансировка классов, то используйте методы Random Undersampling и Oversampling. Пусть в каждом классе будет столько же данных, сколько в классе с оценкой 1. Тогда класс 5 будет называться мажоритарным, так как в нем слишком много данных, а классы 2, 3, 4 – миноритарными, так как в них данных недостаточно. Следовательно, теперь необходимо применить Random Undersampling (удаление случайных примеров) к классу 5, и Random Oversampling (случайное повторение примеров) к классам 4, 3, 2.

Оцените баланс классов после обработки (Рис.6).



**Рис. 7 Доли классов в сбалансированной выборке**

Задание 5. Примените 4 метода машинного обучения и результаты оформите по примеру таблицы 1.

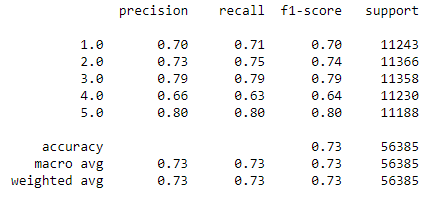
**Таб. 1. Результаты работы моделей**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Run time | Traing time | Test time | Accuracy |
| Naive Bayes | 0.9989 | 0.8519 | 0.147 | 0.51 |
| Logistic regression | 50.733 | 50.7029 | 0.0302 | 0.52 |
| Decision tree | 959.784 | 959.631 | 0.153 | 0.69 |
| k Nearest Neighbor | 941.835 | 0.077 | 941.758 | 0.52 |
| Random forest | 10590.809 | 10561.629 | 29.148 | 0.73 |
| Ridge Classifier | 88.158 | 88.112 | 0.047 | 0.54 |

Сделайте выводы!

Примеры выводов.

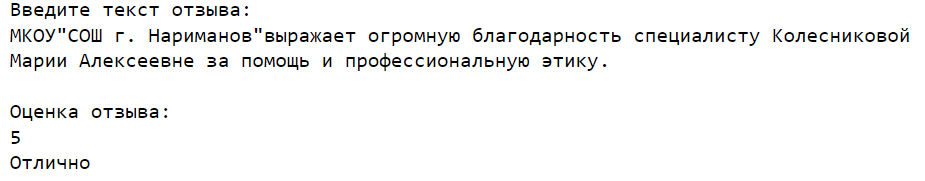
Анализируя метрики для каждого из классов, можно заключить, что модель достаточно плохо определяет класс 4.

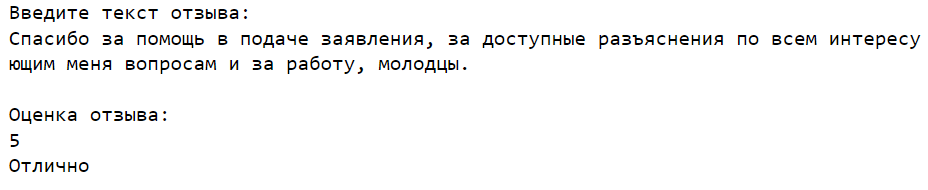


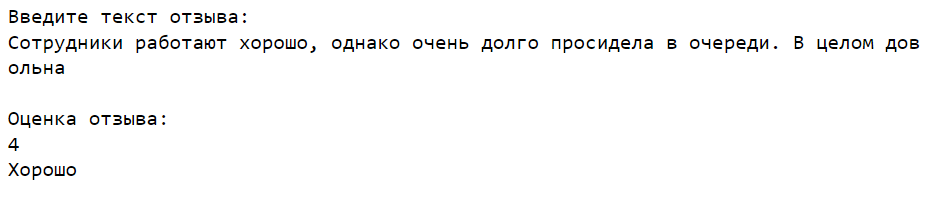
Дополнительное задание!!

Проверьте работу модели на отдельных текстовых отзывах об оказании государственных услуг, введенных вручную.

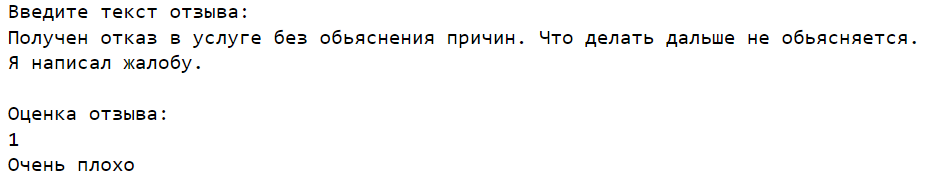
Примеры:

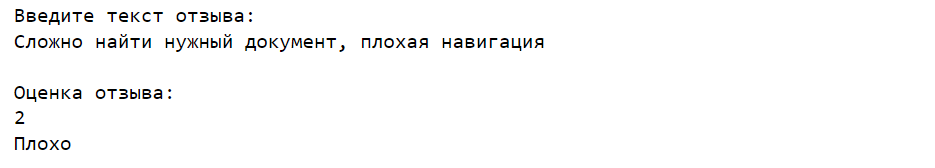




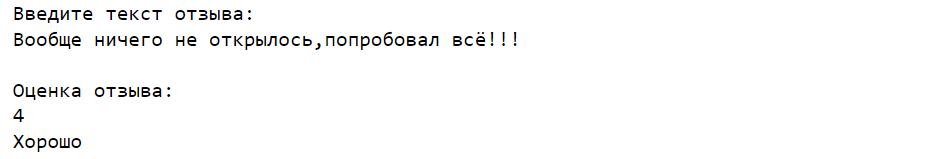


Модель достаточно хорошо определяет положительные оценки пользователей.

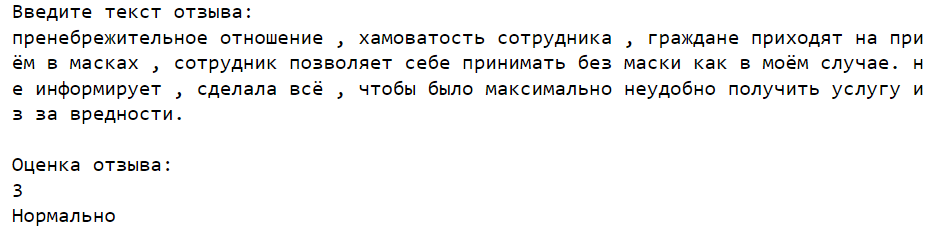




Явное неудовлетворение качеством работы ведомств также определяется моделью, однако модель всё же работает недостаточно точно.



На данном примере видно, что хотя пользователь явно не доволен, результат работы модели говорит о том, что оценка данного отзыва хорошая, что противоречиво. В следующем примере также существует вероятность, что модель сработала неправильно.



Пользователь рассказывает лишь о явных недочетах работы сотрудников, и, хотя он мог поставить оценку «Нормально», это маловероятно.