Задание 2. Детектирование спам сообщений.

В качестве исходных данных выступают сообщения, которым в обучающей выборке проставлена маркеровка- spam/ham.

Задача- построить модель классификации, определяющая принадлежность сообщения к той или иной категории(спам, не спам).

Препроцессинг данных:

Для начала необходимо провести препроцессинг текстовых сообщений , свойственных для такого типа задач:

- Удаление стоп-слов (удаление символов, не несущих информацию для модели классификации сообщений – пробелы, запятые, числа и т.п.)
- Стемминг (приведение слов к нормальной словоформе)
- Удаление часто встречающихся в документах коротких слов(в данном случае такие слова длиной до 2 символов вносят шум в модель, поэтому их необходимо очистить)

Следующим этапом было приведение сообщений, состоящих из слов к TF-IDF числовой матрице (далее будем рассматривать только это приведение, так как приведение сообщений к CountVectorizer числовой матричной форме не принес хорошего качества в моделях)

Важным также оказался показатель длины сообщений для каждого абонента.

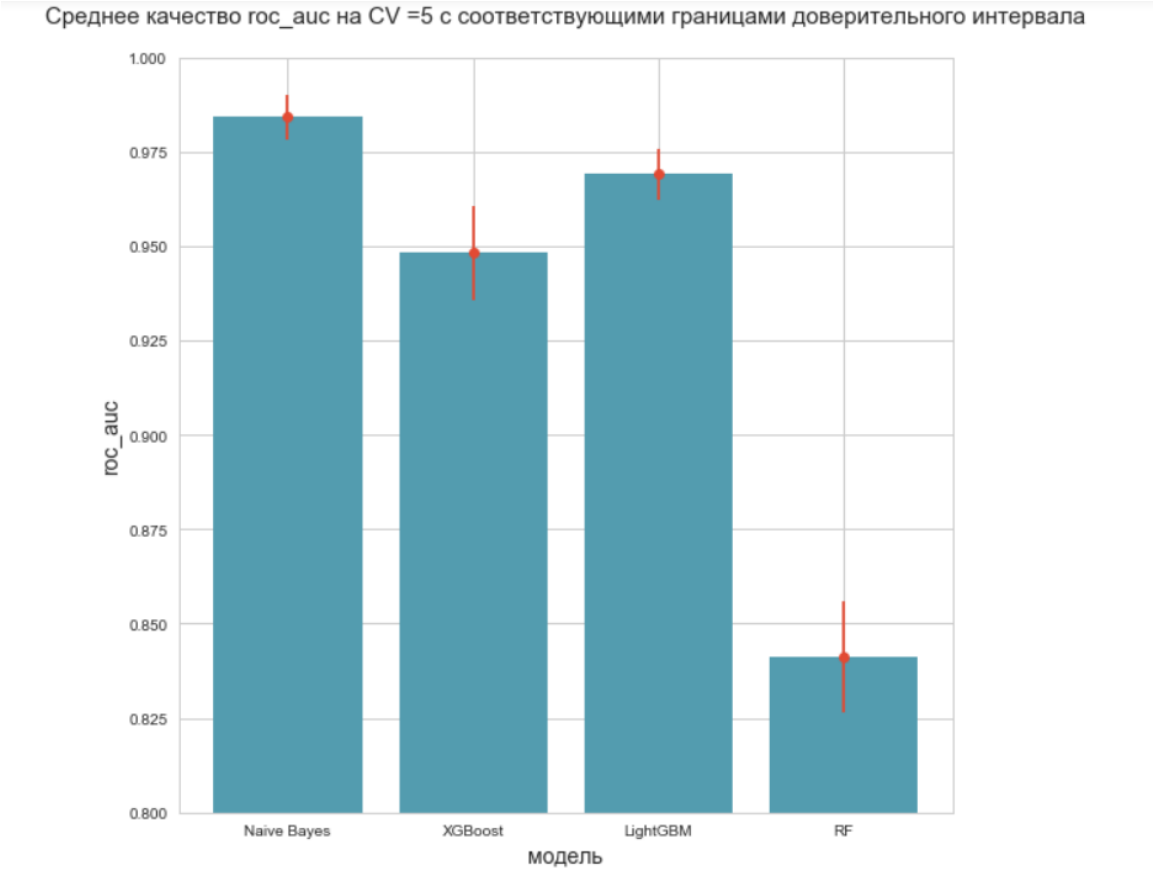
На рисунке можно легко увидеть, что длина spam –сообщений гораздо больше в среднем длины ham – сообщений. Таким образом сформировали обучающую выборку в числовом виде.

Построение модели.

Так как баланс классов у нас несбалансированный, **ham/ spam = 0.86591/ 0.134086**, в качестве метрики оптимизации при построении модели выбираем **roc_auc_score**

Для каждой модели классификации посмотрели на качество классификации на стратифицированной кросс-валидации и взяли среднее значение качества по CV= 5 folds.

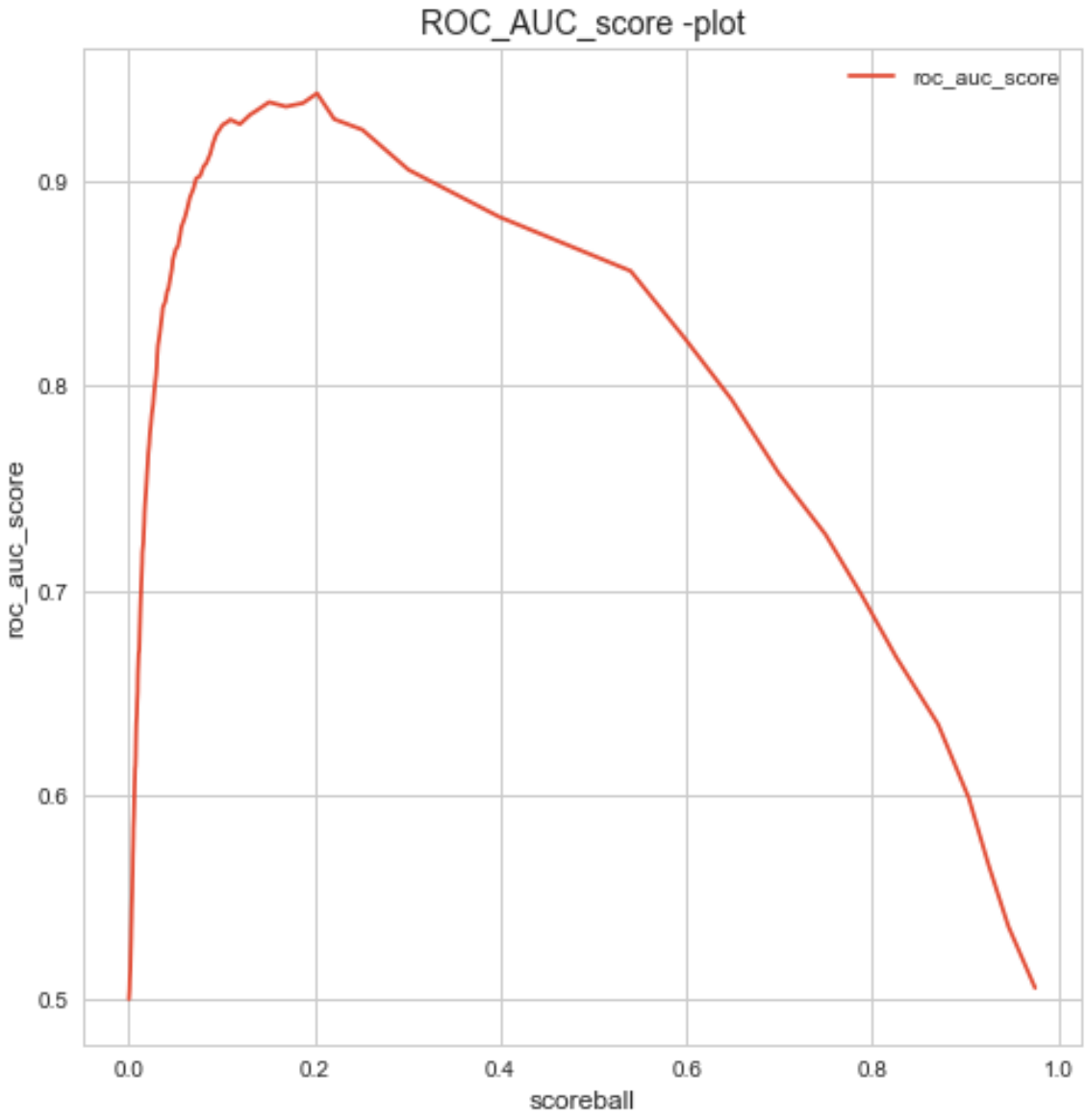
Ниже представлен график статистики среднего качества roc_auc_score алгоритмов классификации вместе с их 95%-ми доверительными интервалами(p-value=0,05).



Как видно, наилучший результат показал мультиновмиальный наивный байесовский классификатор с соответствующим средним качеством roc_auc_score = 0.9841 по 5 фолдам.

Выбор отсечки

В качестве отсечки по вероятности предсказаний была выбрана отсечка по скорбаллу, равная 0.2025955. Ниже приведена динамика показателя roc_auc_score в зависимости от скорбалла.

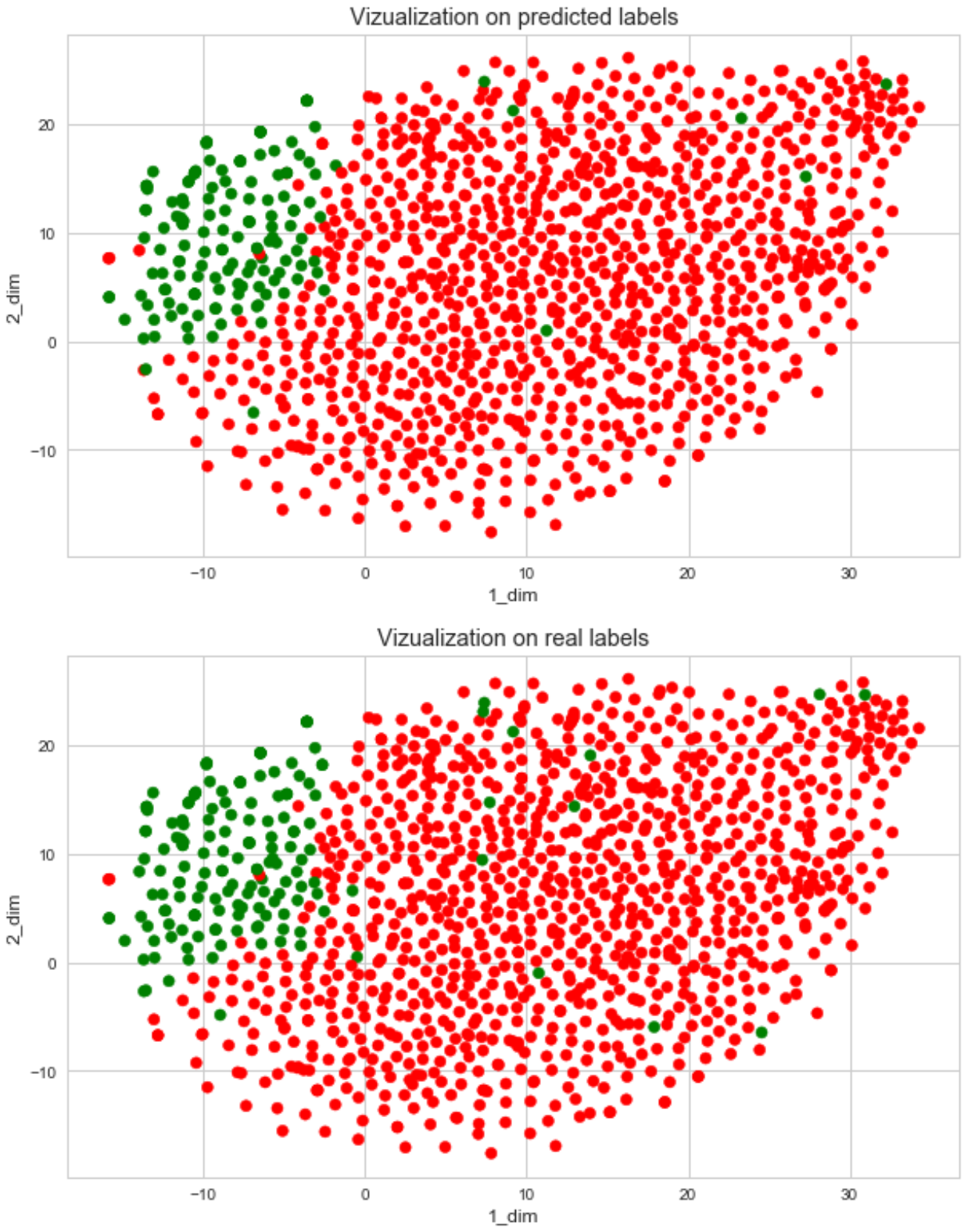


Визуализация качества алгоритма

Для того, чтобы визуально оценить качество модели, применим TSNE понижение размерности на плоскость.

Далее сделаем разделение one-hold на train и test. Обучимся на train выборке и предскажем spam/ham на test.

Сделаем визуализацию настоящих и предсказанных значений spam/ham на test выборке. Spam сообщения обозначены красными точками, соответственно ham- красными.



Видно, что классификация данных происходит хорошо.