ВЫПОЛНИЛА: СУЛЯГИНА АНАСТАСИЯ

РУКОВОДИТЕЛИ: АНТОН ВОЛОХОВ,ГЕОРГИЙ ЕФИМОВ

КЛАССИФИКАЦИЯ ОБЪЯВЛЕНИЙ НА ОСНОВЕ ОПИСАНИЙ

ЗАДАЧА

Научиться распознавать среди объявлений про автомобили объявления мошенников и перекупщиков.

ДАННЫЕ

- Текст объявления
- Маркер от модератора
- Жалобы пользователей

ИНСТРУМЕНТЫ

- Язык программирования: Python
- Библиотеки: pymorphy, scikit-learn, theano, lasagne
- Метрики: precision/recall, ROC, accuracy

РАБОТА С ТЕКСТОМ

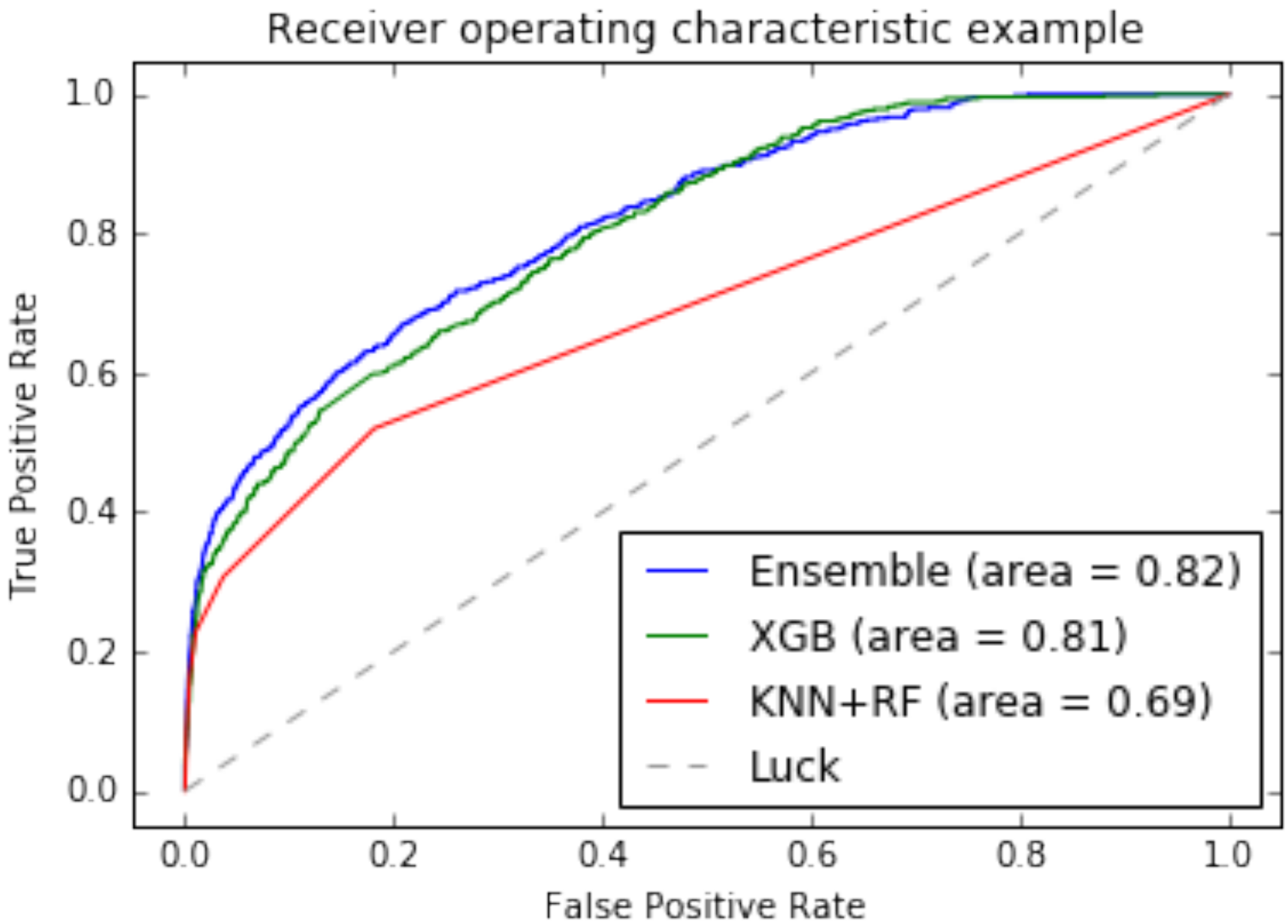
- Нормализация
- Удаление стоп-слов
- Замена незначимых слов на лейблы (42 -> _number)
- Соединение "не" с последующим словом
- Bag of words/ TF-IDF
- Подсчет встречаемости слов в хорошем и плохом словарях
- N-grams для сети

КЛАССИФИКАЦИЯ

Модель	precision	recall	AUC	accuracy
Ансамбль*	0.73	0.20	0.82	0.96
Градиентный бустинг	0.72	0.14	0.81	0.96
Ближайшие соседи + лес	0.61	0.23	0.69	0.95

*Состоит из двух RandomForest, двух ExtraTrees с разными функциями качества и одного бустинга, объединенных логистической регрессией

ROC


Receiver operating characteristic example
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Ensemble (area = 0.82)
XGB (area = 0.81)
KNN+RF (area = 0.69)
Luck

НЕЙРОННЫЕ СЕТИ

- Просто сеть с дропаутами
- CNN

СЛОЖНОСТИ

- Русский язык
- Тонкая грань между "хорошими" и "плохими" объявлениями
- Нехватка мощности компьютера
- Отсутствие внятных статей про применение нейронных сетей для классификации текстов

ОПЫТ

- NLP
- Машинное обучение
- Нейронные сети

РЕЗУЛЬТАТ

- Текст объявлений приведен к пригодному для классификации виду
- Проведена классификация с помощью различных моделей
- Достигнута точность, достаточная для внедрения классификатора