МИНОБРНАУКИ РОССИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра математической экономики

# Использование библиотеки Scikit-learn для решения задач машинного обучения

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 441 группы

направления 09.03.03 Прикладная информатика

механико-математического факультета

Мызникова Александра Сергеевича

Научный руководитель
д. э. н., профессор _____ В. А. Балаш

Зав. Кафедрой
д.ф-м.н, профессор _____ С. И. Дудов

Саратов 2020

**Введение.** Пользуясь машинным обучением программисту необязательно писать различные инструкции, которые будут учитывать все возможные проблемы и которые будут содержать все решения этих проблем. Вместо этого в программу можно заложить алгоритм, который сам будет находить решения путем комплексного использования статистических данных, из которых в свою очередь будут выводиться различные закономерности и на основании которых будут делаться прогнозы.

Свое начало технология машинного обучения берет еще в 1950 году. В то время началась разработка первых программ для игры в шашки. До сегодняшнего времени общие принципы не подверглись изменениям. Но благодаря значительному росту вычислительной мощности компьютеров возросла степень сложности закономерностей и прогнозов, создаваемых на основе этих закономерностей, и значительно расширился круг задач, которые можно решить с помощью машинного обучения.

Чтобы начать процесс машинного обучения, нужно загрузить Dataset (Dataset – это обработанные и структурированные исходные данные, которые хранятся в табличном виде), на которых выбранные алгоритм будет обучаться обрабатывать запросы. Например, могут быть фотографии различных животных, а на этих фотографиях будут метки, которые будут обозначать кто к какому животному относится. По окончанию обучения программа будет сама распознавать животных на каких-то новых изображениях, на которых уже не будет соответствующих меток.

Стоит заметить, что обучение продолжится и после выданных программой прогнозов, так как с все большим количеством проанализированных данных повышается точность распознавания нужных нам изображений.

С помощью машинного обучения на рисунках можно распознавать не только лица, но и различные фигуры, предметы, текст и т. д. Проверка грамматики

присутствует в любом текстовом редакторе, будь то компьютер, телефон или другая электроника — это все тоже благодаря машинному обучению. Существуют различные ПО, которые могут писать новостные статьи на разные темы. Диагностика заболеваний, прогнозирование на финансовых рынках, поиск мест с полезными ископаемыми — примеры применения машинного обучения в реальной жизни.

Целью данной работы является изучение основ теории по машинному обучению, а также решение задач по анализу данных.

**Основное содержание работы.** Работа состоит из 3 частей:

1. Машинное обучение. Общая теория
2. Алгоритмы моделей машинного обучения. Разбор алгоритмов
3. Решение практических задач по анализу данных. Задача Титаник и Energy Star Score

Каждая глава в свою очередь состоит из нескольких частей. В первой части приведена общая теория по машинному обучению и приведены типы задач машинного обучения.

Машинное обучение (Machine Learning или просто ML) — большой подраздел искусственного интеллекта, который изучает методы для построения разных алгоритмов, способных к обучению.

Различают два типа обучения:

- обучение по прецедентам (также его называют индуктивным обучением). Оно основано на выявлении общих закономерностей по каким-либо частным эмпирическим данным

- дедуктивное обучение. Этот тип обучения предполагает формализацию и систематизацию знаний экспертов и их перенос в базу данных.

Общая постановка задачи индуктивного обучения выглядит так: дано конечное множество прецедентов (ситуаций или объектов). По каждому прецеденту собрано некоторое количество данных. Данные об объекте по-другому называются описанием объекта. Совокупность всех описаний объектов или ситуаций называется обучающей выборкой. Требуется по таким частным данным выявить общие закономерности или взаимосвязи, которые присущи не только данной конкретной выборке, но и всем остальным прецедентам, даже тем, которые еще не исследовались.

Имеются основные стандартные типы задач:

- Обучение с учителем.
  - Задача *классификации*
  - Задача *регрессии*
  - Задача *ранжирования*
  - Задача *прогнозирования*
- Обучение без учителя.
  - Задача *кластеризации*
  - Задача *фильтрации выбросов*
  - Задача *заполнения пропущенных значений*
- Метаобучение.
- Динамическое обучение.
- Обучение с подкреплением.

Во второй части работы приведены алгоритмы моделей машинного обучения, а также представлена теория по этим алгоритмам.

Перечислим алгоритмы, которые наиболее популярны в современном машинном обучении:

- Линейная регрессия
- Логистическая регрессия
- Деревья принятия решений
- Ансамблевые методы
  - Бэггинг и случайный лес
  - Бустинг и Adaboost
- K-ближайших соседей (KNN)
- Метод опорных векторов (SVM)
- Наивный Байесовский классификатор

В третьей части работы приведено решение и анализ двух задач по анализу данных.

В решении задач по анализу данных использована библиотека Scikit-learn. Данная библиотека довольно широко используется как и в различных промышленных системах, применяющих алгоритмы и методы ML, так и в начальных этапах изучения machine learning и применения на практике полученных знаний.

Для работы библиотека Scikit-learn использует много других широко известных библиотек:

- Pandas для обработки и анализа данных
- Matpotlib для визуализации данных

- IPython для использования интерактивной консоли
- NumPy для различных математических операций
- SymPy для символьной математики

Далее приведен анализ решения для двух выбранных задач: Titanic и предсказание рейтинга энергопотребления здания.

Титаник — одна из самых известных задач на Kaggle. Датасет задачи Титаник содержит данные пассажиров корабля. Целью задачи является построение модели, которая наилучшим образом сможет сделать предсказание, остался ли произвольный пассажир в живых или нет.

Исследованы признаки и роль, которую они сыграли в выживании или гибели путешественника.

Это был случай задачи классификации, и далее была сделана попытка сделать предсказания с помощью двух алгоритмов — Случайный лес и Гауссовский Наивный Байесовский классификатор. Гауссовский Наивный Байесовский классификатор работал плохо, а Случайный лес делал предсказания с точностью более 80%.

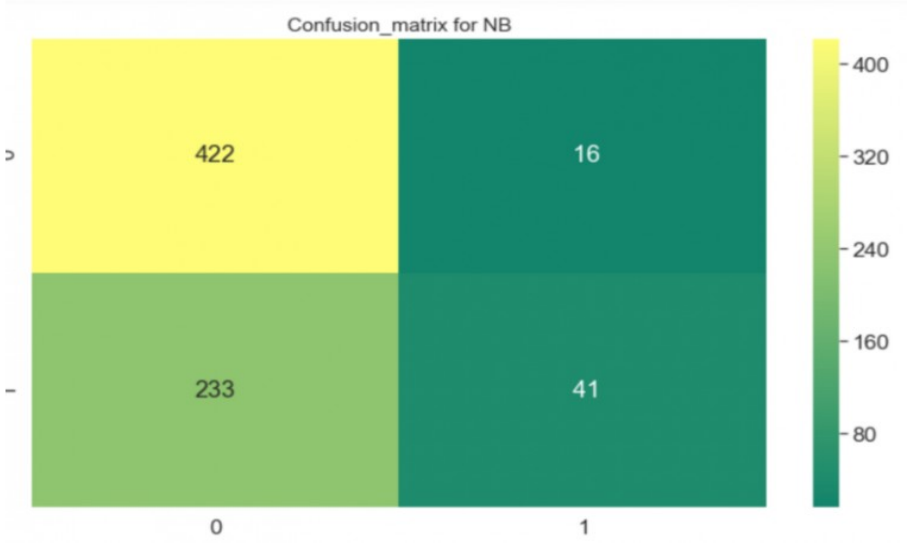


Рис. 1 — Матрица для Наивного Байеса

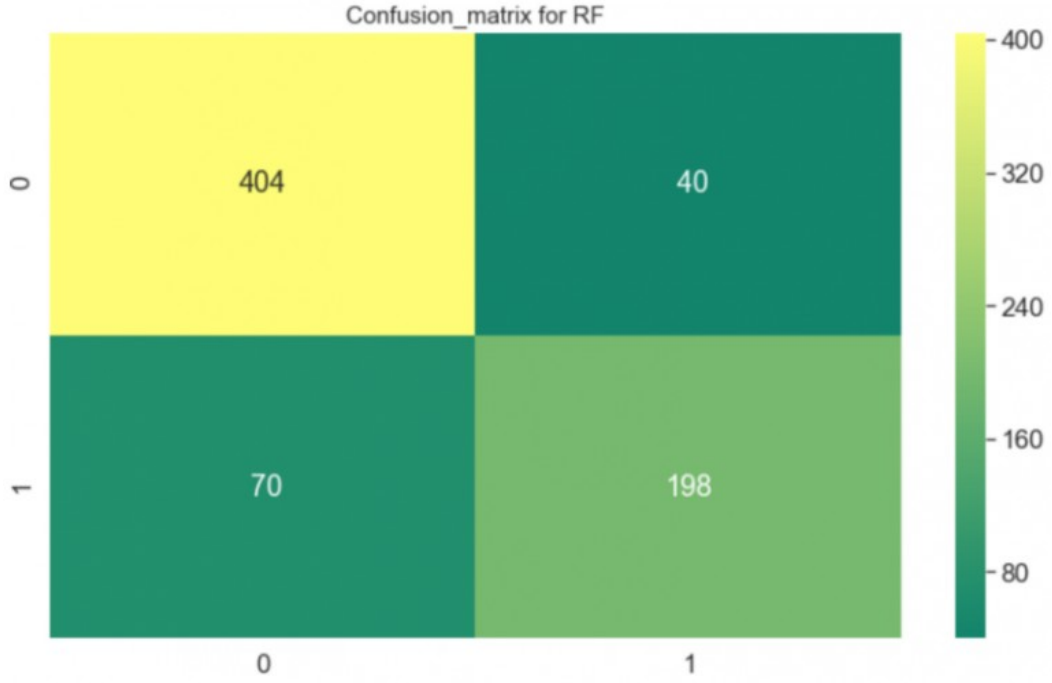


Рис. 2 — Матрица для Случайного леса

Предсказание рейтинга энергопотребления здания (Energy Star Score). Наши данные — открытые сведения об энергопотреблении зданий в Нью-Йорке.

Наша цель — предсказание рейтинга энергопотребления (Energy Star Score) здания и понять, какие признаки оказывают на него сильнейшее влияние.

Данные уже содержат в себе Energy Star Score, так что задача относится к классу задач машинного обучения с учителем, и представляет собой построение регрессии:

- Обучение с учителем: у нас есть как все необходимые признаки, на основе которых выполняется предсказание, так и сам целевой признак
- Регрессия: будем считать, что рейтинг энергопотребления — это непрерывная величина

В конечном итоге была построена наиболее точная модель, которая на выходе дает легко интерпретируемые результаты, то есть мы сможем понять на основании чего модель делает тот или иной вывод.

Для построения наиболее точной модели сначала была произведена оценка влияния значения категориальных признаков, например, к какому типу относится то или иное здание или в каком районе расположено здание

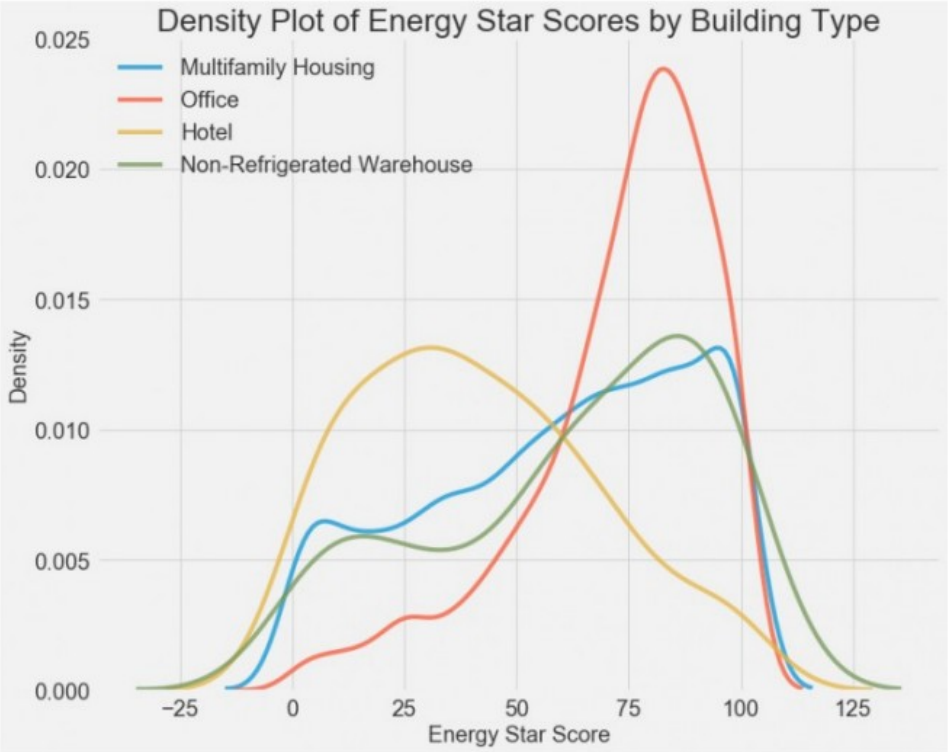


Рис. 3 — Energy Star Score в зависимости от тип здания

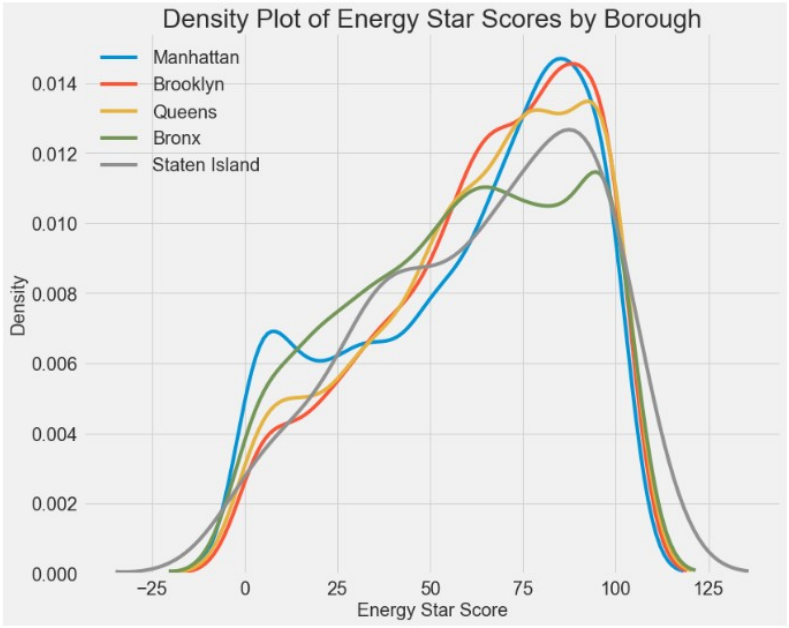


Рис. 4 — Energy Star Score в зависимости от района города

Далее была измерена средняя абсолютная ошибка в прогнозах (MAE) и на основе этого показателя сравнены методы машинного обучения, после чего выбран наиболее лучший:

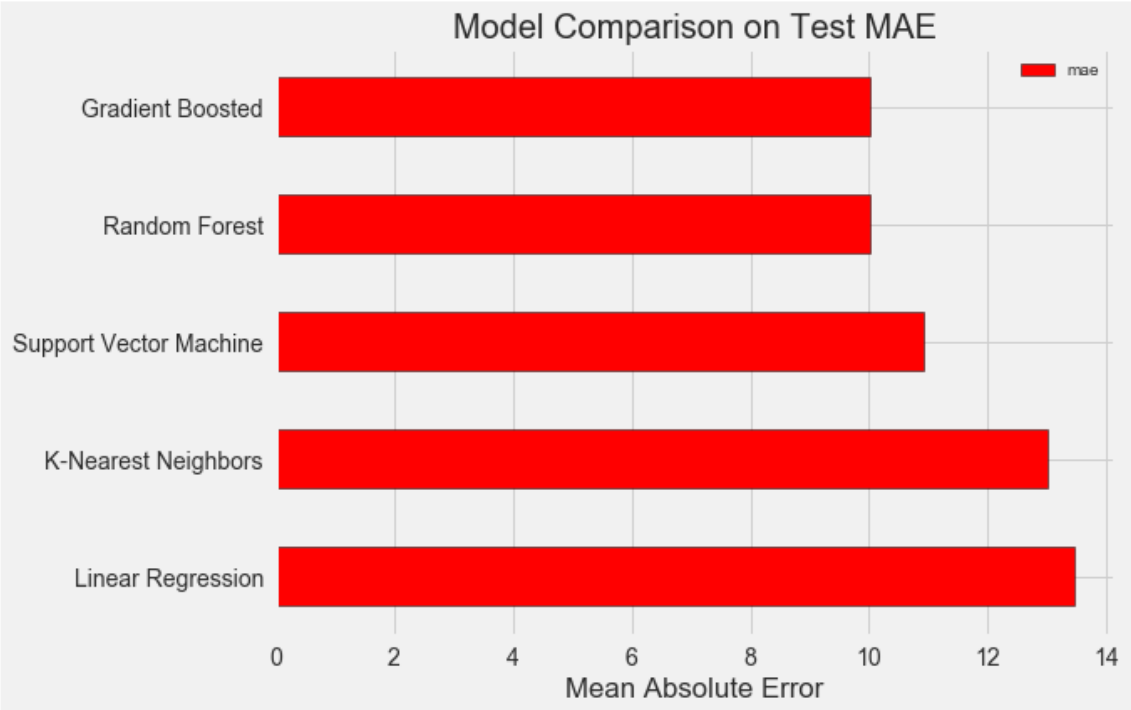


Рис. 5 — Сравнение методов и алгоритмов

Видим, что значение MAE у метода градиентного бустинга и случайного леса практически равны (10.013 против 10.014). Результат у GB (бустинга) чуть меньше, поэтому выбираем для оптимизации модели именно его.

Как только были получены окончательные прогнозы, мы проверили их, чтобы увидеть, нет ли у них заметных искажений.

Прогнозы модели, по-видимому, следуют распределению фактических значений, хотя пик плотности происходит ближе к медианному значению (66) в тренировочном наборе, чем к истинному пику плотности (который составляет около 100). Остатки почти нормально распределены, хотя было заметно несколько больших отрицательных значений, где предсказания модели были намного ниже истинных значений.

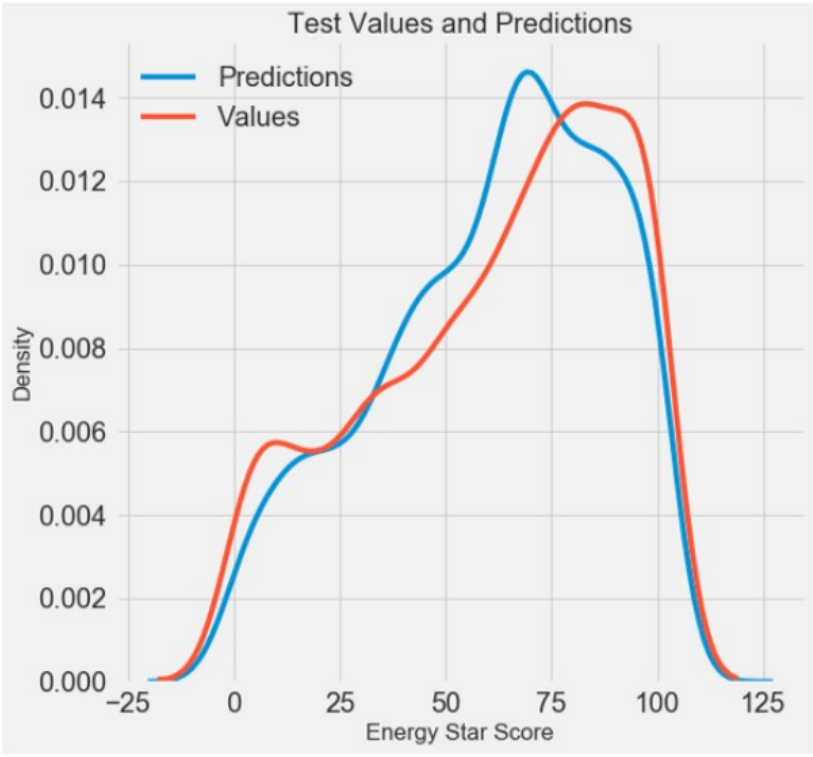


Рис. 6 — График плотности прогнозируемых и фактических значений

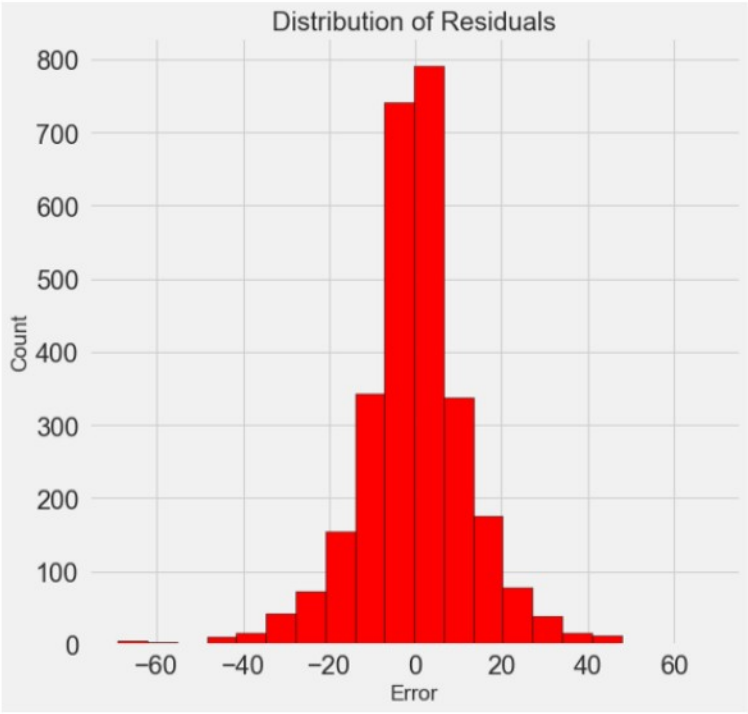


Рис. 7 — Гистограмма остатков

Далее было визуализировано дерево решений. Оно получилось довольно большим, поэтому стоит показать лишь пример одного из блоков дерева.

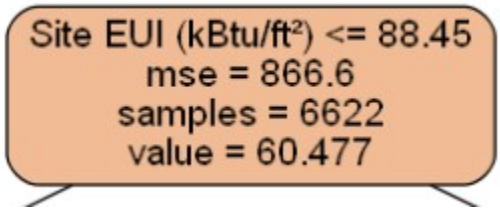


Рис. 8 — Пример одного из блоков дерева

Ансамбли на основе дерева решений объединяют предсказания многих отдельных деревьев решений, чтобы создать более точную модель с меньшими отклонениями. Такие ансамбли очень точны и понятны.

**Заключение.** В практической части дипломной работы были сделаны и проанализированы две конкретные задачи по анализу данных. В общем случае процесс решения задач возникающих в машинном обучении состоит из следующих этапов:

1. Очистка и форматирование данных
2. Предварительный анализ данных
3. Создание и выбор наиболее полезных признаков
4. Сравнение качества работы нескольких моделей
5. Оптимизация гиперпараметров для модели
6. Проверка моделей на тестовых выборках
7. Интерпретация результатов

Также была предоставлена вся необходимая теория по всему, что использовалось в практической части: от фундаментальных определений до описания принципов работы используемых библиотек.