УДК 519.24

АНАЛИЗ ДАННЫХ О ПОЕЗДКАХ В МЕГАПОЛИСЕ И ПРОГНОЗИРОВАНИЕ СТОИМОСТИ ПОЕЗДКИ

Дзержинский Р.И., Лапшин И.А., Аносов Т.Э.

МИРЭА – Российский технологический университет, 119454, Россия, г. Москва, проспект Вернадского, 78, e-mail: 9015111295@mail.ru, vanialapshin99@yandex.ru, tumyp.anosov@gmail.com

В данной статье исследуется набор данных о поездках такси в мегаполисе с целью определения факторов, влияющих на ценообразование стоимости поездки. Использованы модели машинного обучения: линейная регрессия, полиномиальная регрессия, деревья принятия решений, ансамблевое обучение и метод случайных лесов. В ходе сравнения выяснилось, что оптимальной моделью является бэггинг. Получен способ определения прогностической оценки стоимости поездки.

Ключевые слова: прогнозирование стоимости услуг такси, модель машинного обучения, линейная регрессия, полиномиальная регрессия, модель деревьев принятия решений, метод случайных лесов, ансамблевое обучение, бэггинг, бустинг.

ANALYSIS OF DATA ON TRIPS TO THE METROPOLIS AND FORECASTING THE COST OF THE TRIP

Dzerzhinskiy R.I., Lapshin I.A., Anosov T.E.

MIREA - Russian Technological University, 119454, Moscow, 78 Vernadskogo Avenue, Russia, e-mail: 9015111295@mail.ru, vanialapshin99@yandex.ru, tumyp.anosov@gmail.com

This article examines a set of data on taxi rides in the metropolis in order to determine the factors that affect the price of the trip. Machine learning models are used: linear regression, polynomial regression, decision trees, ensemble learning, and the random forest method. During the comparison, it turned out that the optimal model is begging. A method for determining the predictive estimate of the cost of a trip is obtained.

Key words: taxi service cost forecasting, machine learning model, linear regression, polynomial regression, decision tree model, random forest method, ensemble learning, begging, boosting.

Введение

Рынок такси опережает многие сектора среднего и малого бизнеса развитых и развивающихся стран мира. В Российской Федерации теневой сектор услуг такси сильно сократился. Его вытесняют высококлассные сектора бизнеса такие, как «Uber», «Яндекс.Такси», «Ситимобил». Они имеют необходимые лицензии, организованные механизмы получения и распределения заказов, а также систему ценообразования, работающую по специальным моделям. Однако моделирование прогнозирования цен на услуги такси по-прежнему несет в себе следующие задачи, без решения которых существующие модели могут приводить к слишком долгому ожиданию клиентом автомобиля или, например, высокой цене при низком спросе:

- Прогнозирование цены в городах с неравномерной территориальной нагруженностью автомобильным трафиком (Москва, Санкт-Петербург, другие мегаполисы);
- Прогнозирование цены в городах с высокой насыщенностью транспортных узлов (аэропорты, вокзалы);
- Прогнозирование цены в условиях проведения многолюдных культурно-массовых мероприятий (футбольные матчи, концерты и т.д.);
- Прогнозирование цены при различных погодных условиях (дождь, гололёд, снег).

Вышеперечисленные задачи подтверждают тот факт, что проблема моделирования прогнозирования цен на услуги такси в крупных городах и мегаполисах является актуальной.

Целью работы является поиск наиболее оптимизированного подхода к прогнозированию цен на услуги такси в мегаполисе на основе данных о поездках в Нью-Йорке.

В качестве объекта исследования используется набор данных о поездках в Нью-Йорке, информация априорного характера об инфраструктуре города не используется, число аэропортов следует из анализа задачи, строятся наилучшие модели машинного обучения, способные выдавать оптимальный прогноз. Рассматриваемая в данной работе выборка может представлять интерес для дальнейшего исследования прогнозирования цен на услуги такси в крупных российских городах (например, в Москве) из-за следующих факторов:

- неравномерная загруженность трафиком, наличие нескольких аэропортов/вокзалов как в Нью-Йорке, так и в Москве;

- развитая инфраструктура мегаполиса (загруженное метро, большие деловые центры, наличие большого количества развлекательных центров/областей для отдыха);

Обнаружение и визуализация данных для понимания их сущности

Набор данных включает в себя стоимость поездки, долготу и широту места высадки, долготу и широту места посадки, а также количество пассажиров. Среднее значение стоимости составляет 11.35$ и имеет стандартное отклонение 9.42 единиц. Гистограмма плотности распределения помогает сравнить эмпирическое распределение с предполагаемым теоретическим, выделить основные квантили и понять характер вариации данных. Такой график логарифма стоимости поездки представлен на рисунке 1.

Распределение имеет колоколообразную форму с дополнительными вершинами на хостах. Такое поведение свидетельствует о неоднородности выборки. Непрерывная случайная величина имеет логарифмически-нормальное (сокращенно логонормальное распределение), если ее логарифм подчинен нормальному закону [1]. Из приведенного определения, можно сделать вывод, что распределение стоимости поездки может подчиняться логонормальному закону.

В наборе данных присутствуют географическая информация (долгота и широта) о местах начала и окончания поездки. Человеческий мозг очень хорошо выявляет паттерны на рисунках, поэтому для получения лучшего представления необходимо построить графики рассеивания таких точек.

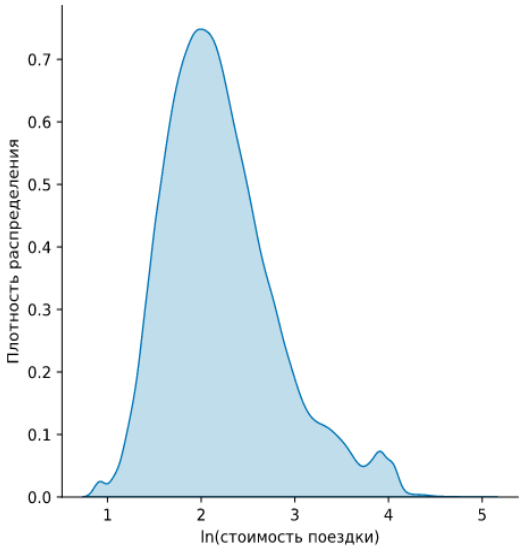

Рисунок 1. Эмпирический закон распределения

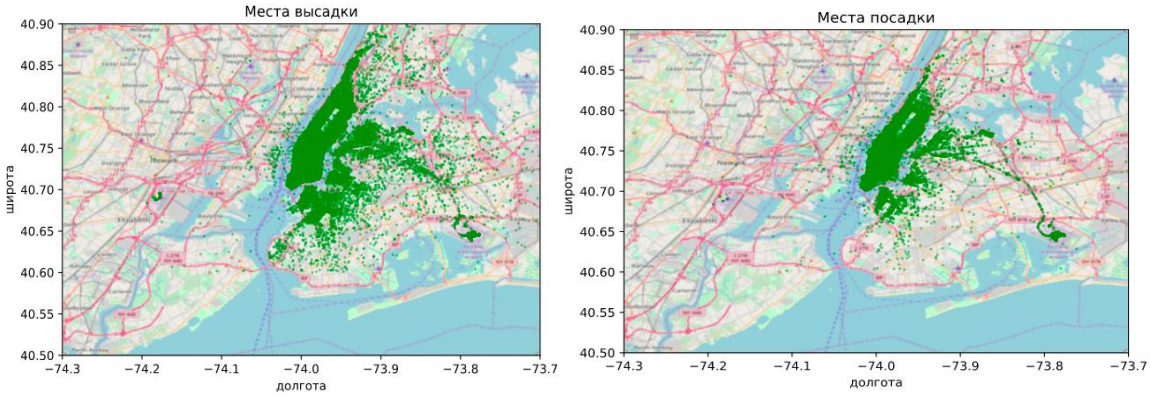

Рисунок 2. Точки посадок и высадок

На рисунке 2 представлены точки посадок и высадок. Не трудно заметить участки с высокой плотностью - район Манхэттен, побережье Бруклина и Квинс. Интерес представляет обособленный участок, расположенный на юге Квинс. В этом месте находится международный аэропорт имени Джона Кеннеди.

На востоке карты тоже присутствуют такие участки - в районе международного аэропорта Ньюарк Либерти и аэропорта помощи общей авиации в районах Тетерборо. Также нельзя не заметить, что места высадки имеют более хаотичное распределение вне вышеупомянутых районов, это говорит о том, что жители заказывают поездки из центра чаще чем в центр.

Приняв во внимание полученную информацию, данные разделяются на несколько классов: поездки, связанные с каждым из аэропортов, другие поездки. На рисунке 3 показано распределение логарифма стоимости поездок, с учетом принадлежности к тому или иному аэропорту.

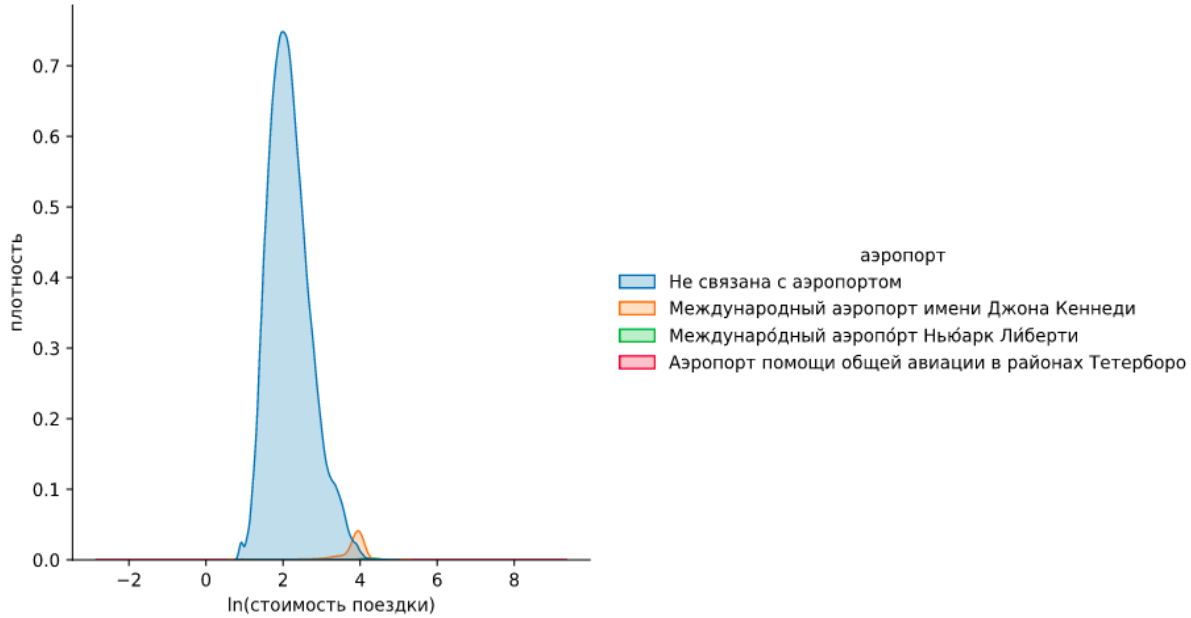

Рисунок 3. Распределение логарифма стоимости поездок с учетом аэропортов

На гистограмме видно, что правый пик обусловлен именно поездками, связанными с международным аэропортом имени Джона Кеннеди. Из этого можно сделать вывод, что модель ценообразования для таких поездок существенно отличается от остальных.

Далее проверим влияние пройденного расстояния на стоимость поездки. За показатель тесноты линейной связи отвечает коэффициент корреляции τ, значение которого лежит в диапазоне от -1 до +1. В данном случае $\tau = 0.81$, что говорит о функциональной зависимости стоимости от расстояния. Диаграмма рассеивания для упомянутой зависимости с учетом разбиения на классы представлена на рисунке 4. Как видно из графика, она действительно напоминает линейную: с увеличением расстояния растет стоимость в областях, которые не связаны с поездками в какой-либо аэропорт или из аэропорта. Отметки, отвечающие за такие поездки, расположены горизонтально и параллельно оси абсцисс, следовательно, стоимость никак не зависит от расстояния, она постоянна.

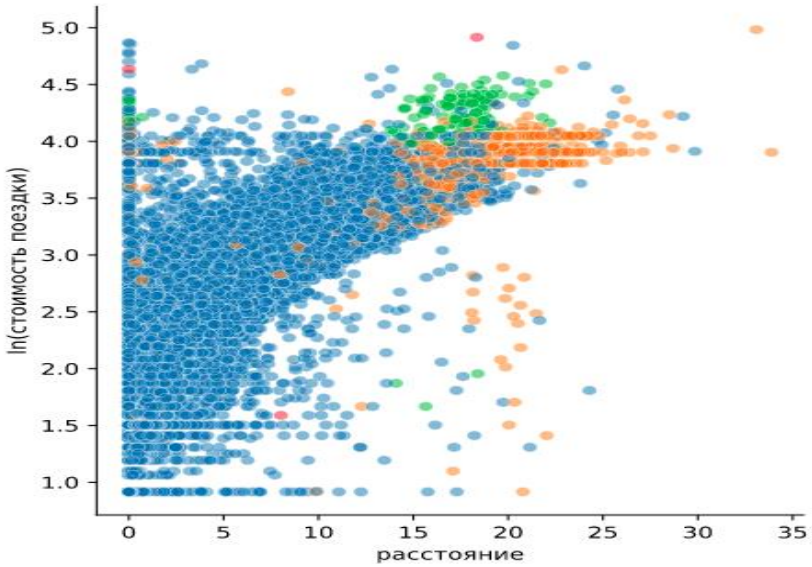

Рисунок 4. Зависимость логарифма стоимости поездки от расстояния.

На рисунке 5 показано влияние времени суток на среднюю стоимость. Судя по графику, цена увеличивается с каждым годом, но паттерн остается одинаковым. Характерными временными интервалами увеличения цены являются часы перед началом рабочего дня, после его окончания и вечернее время.

Важным фактором, влияющим на стоимость поездки, является обстановка в городе. Чем выше трафик, тем больше времени понадобится для того, чтобы добраться из точки А в точку Б. Также важно, чтобы спрос клиентов

и предложение таксистов были в оптимальном соотношении, тем самым, минимизируя случаи, когда в районе имеются заказы, но нет ни одной свободной машины.

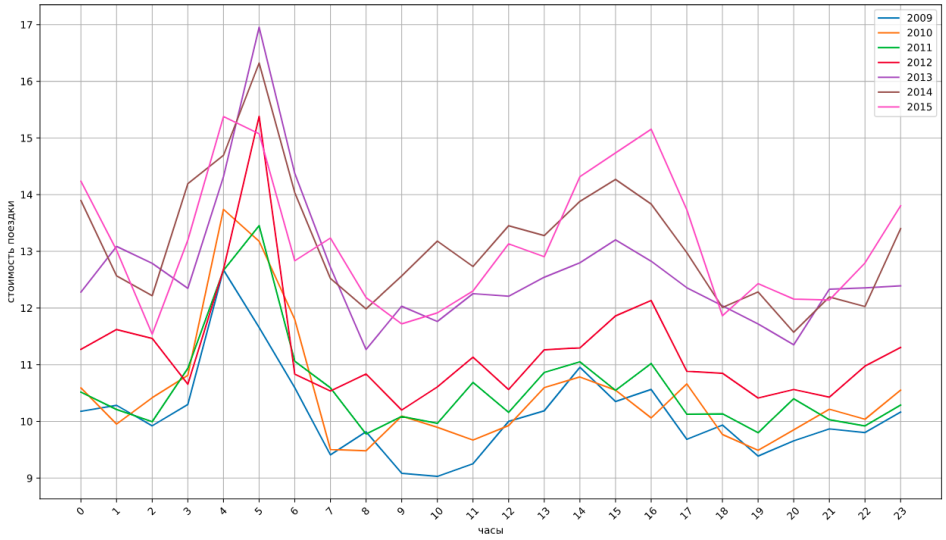

Рисунок 5. Влияние времени суток на среднюю стоимость

Не имея доступа к информации о трафике в Нью-Йорке, пришлось воспользоваться имеющимися данными. Идея состояла в том, чтобы посчитать количество заказов, сделанных в течение часа, которые при этом начинались/заканчивались в области текущего заказа. Такого результат удалось достичь путем группировки данных по дням и подсчете количества заказов в соответствующих географических зонах. Далее полученный результат еще раз сгруппировался, но уже по месяцам, дням недели, времени суток. Результатом стало среднее значение для каждой такой группы. Таким образом, в наборе данных появился еще один признак - среднее количество машин рядом с заказом. На рисунке 6 представлена зависимость логарифма стоимости от «трафика».

На графике видно, что существуют данные, у которых цена не зависит от показателя «трафика». Но также существует тенденция к понижению стоимости с увеличением количества машин в пределах заказа.

После окончания исследования набора данных и получения представления об общей картине, можно приступить к построению прогноза.

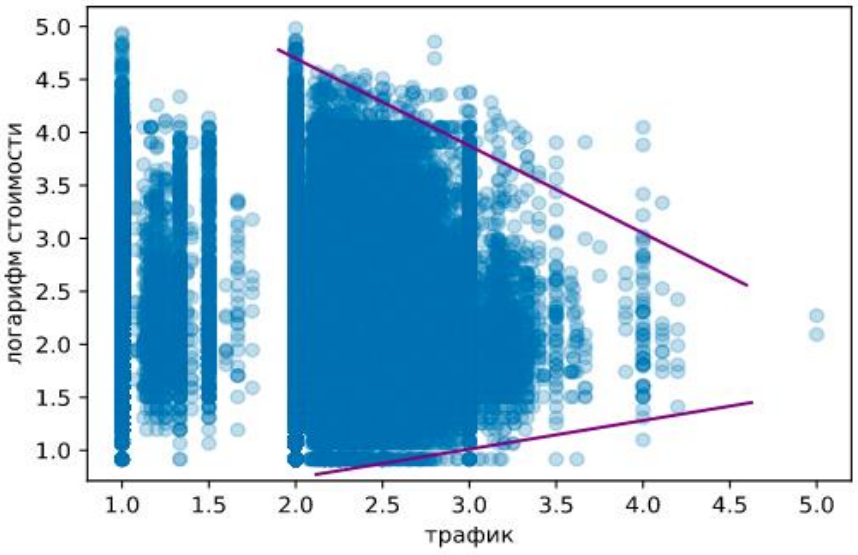

Рисунок 6. Зависимость логарифма стоимости от трафика.

Прогнозирование

Чтобы дать оценку стоимости поездки, необходимо решить задачу регрессии. Для этого можно воспользоваться моделями машинного обучения, в число которых входят: модель линейной регрессии, считающая взвешенную сумму входных признаков; модель полиномиальной регрессии, которая в отличие от линейной регрессии, подходит для аппроксимации нелинейных данных; деревья принятия решений, которые строят специальный граф, выдающий прогноз путем маршрутизации входящих признаков через свои вершины и другие.

Для оценки качества работы этих алгоритмов применяется среднеквадратическая ошибка, которая показывает, на сколько в среднем прогноз отличается от ожидаемого значения.

Альтернативой представленному выше способу является перекрестная проверка. Ее алгоритм разбивает обучающий набор на n независимых подмножеств набора, после чего обучает их и оценивает модель n-1 раз, каждый раз выбирая для оценки другое подмножество и проводя обучение на оставшихся n-1 подмножествах.

В таблице 1 приведено сравнение нескольких моделей, обучающихся на исследуемых в этой статье данных.

Таблица 1. Сравнение методов машинного обучения

Модель	Объем обучающей выборки	Объем проверочной выборки	Время обучения (сек)	Среднеквадратическая ошибка	Результат перекрестной проверки (среднее)	Результат перекрестной проверки (стандартное отклонение)
Линейная регрессия	7803	9451	0	4.711	4.413	0.12
Полиномиальная регрессия второй степени	7803	19451	0	4.458	4.241	0.129
Дерево принятия решений для задач регрессии	7803	9451	0	4.43	4.09	0.13
Метод случайных лесов	7803	9451	29	4.18	3.9	0.1
Бэггинг	7803	9451	169	4.09	3.97	0.1
Бустинг	7803	9451	193	4.05	3.95	0.09

Основным критерием для оценки эффективности метода в рамках исследования является минимизация среднеквадратической ошибки. Нетрудно заметить, что наиболее эффективными методами в рамках описанного критерия являются методы бэггинга[1] и бустинга[2]. В данном случае для обучения с помощью методов бэггинга и бустинга использовался ансамбль из линейной регрессии, полиномиальной регрессии и деревьев принятия решений.

Дополнительным критерием оценки эффективности выбирается время, затрачиваемое на обучение модели тем или иным методом. В данном эксперименте выяснилось, что наиболее затратными по данному критерию методами также являются методы бэггинга и бустинга, при этом метод бустинга, несмотря на параллельный подход к обучению моделей, не сильно уступает методу бэггинга на исследуемой выборке, тем самым показывая наибольшую эффективность среди всех исследуемых методов.

Заключение

В данной статье был рассмотрен набор реальных данных о поездках такси в Нью-Йорке, получено представление о ценообразующих факторах.

В реалиях современного мира не может быть построена аналитически достоверная модель, потому в качестве такого механизма используются различные модели машинного обучения. Каждый день по всему миру миллионы людей пользуются услугами такси. Для обеспечения баланса, между удовлетворенностью ценами клиентами и прибылью компании, необходимо иметь механизм, способный выдавать оценку адекватной стоимости поездки, которая учитывает все возможные факторы. В статье было проведено сравнение наиболее популярных регрессионных моделей и представлены результаты прогнозирования по данным моделям.

Основным критерием оценки эффективности метода выбиралась минимизация среднеквадратической ошибки. Исходя из результатов, наиболее эффективным по данному критерию методом оказался метод бустинга, показавший наименьшую ошибку как при стандартном расчете среднеквадратической ошибки, так и при перекрестной проверке.

---

[1] **Бэггинг** - технология классификации, использующая композиции алгоритмов, каждый из которых обучается независимо. Результат классификации определяется путем голосования. Бэггинг позволяет снизить процент ошибки классификации в случае, когда высока дисперсия ошибки базового метода.

[2] **Бустинг** - метод построения ансамбля моделей, при котором базовые модели обучаются последовательно, и каждая последующая модель ансамбля применяется к результатам на выходе предыдущей.

Список литературы

1. Кремер Н. Ш. Теория вероятности и математическая статистика. - 2-е изд., перераб. и доп. - М.: ЮНИТИДАНА, 2004. - 573 с.
2. Орельен Жерон. Прикладное машинное обучение с помощью Scikit-learn и TensorFlow. - Пер. с англ. - СпЮ: ООО «Альфа-книга», 2018. - 688 с
3. Уэс Маккинни. Python и анализ данных. - Пер. с англ. Слинкин А. А. - М.: ДМК Пресс, 2015. - 482 с.: ил.
4. Под капотом “Яндекс такси” [Электронный ресурс] https://vc.ru/yandex.go/40971-pod-kapotom-yandeks-taksi
5. Прогнозирование: какие методы использует Uber [Электронный ресурс] https://neurohive.io/ru/osnovy-data-science/prognozirovanie-uber/
6. Динамическое ценообразование, или как Яндекс такси прогнозирует высокий спрос [Электронный ресурс] https://habr.com/ru/company/yandex/blog/429226/
7. Продвинутый уровень визуализации данных [Электронный ресурс] https://habr.com/ru/company/skillfactory/blog/510320/
8. Машинное обучение[Электронный ресурс] https://www.machinelearningmastery.ru/a-tour-of-machine-learning-algorithms/
9. Оценка качества прогнозируемых моделей [Электронный ресурс] https://forecasting.svetunkov.ru/etextbook/forecasting_toolbox/models_quality/
10. Набор данных о поездках такси [Электронный ресурс] https://www.kaggle.com/c/new-york-city-taxi-fare-prediction

References

1. Kremer N. Sh. Probability theory and mathematical statistics. - 2nd ed., Rev. and add. - M.: YUNITIDANA, 2004.- 573 p.
2. Aurelien Geron. Applied Machine Learning with Scikit-learn and TensorFlow. - Per. from English - Sleep: LLC "Alfa-kniga", 2018. - 688 p: ill. - Parallel. tit. English
3. Wes McKinney. Python and data analysis. - Per. from English Slinkin A.A. - M.: DMK Press, 2015 .-- 482 p.:
4. Under the hood "Yandex taxi" [Electronic resource] https://vc.ru/yandex.go/40971-pod-kapotom-yandeks-taksi
5. Forecasting: what methods does Uber use [Electronic resource] https://neurohive.io/ru/osnovy-data-science/prognozirovanie-uber/
6. Dynamic pricing, or how Yandex taxi predicts high demand [Electronic resource] https://habr.com/ru/company/yandex/blog/429226/
7. Advanced level of data visualization [Electronic resource] https://habr.com/ru/company/skillfactory/blog/510320/
8. Machine learning [Electronic resource] https://www.machinelearningmastery.ru/a-tour-of-machine-learning-algorithms/
9. Assessment of the quality of predicted models [Electronic resource] https://forecasting.svetunkov.ru/etextbook/forecasting_toolbox/models_quality/
10. A set of data on taxi trips [Electronic resource] https://www.kaggle.com/c/new-york-city-taxi-fare-prediction