Определение стоимости аренды жилья

## Цели и задачи

Сервисы для аренды жилья чрезвычайно популярны, что не удивительно, свободные квадратные метры дают возможность получения прибыли для владельцев недвижимости, а также позволяют получить услуги по найму жилья для тех, кто в нем нуждается. Наличие удобных способов оценки стоимости, возможность планирования прибыли способны привлечь внимание клиентов.

В данном исследовании рассмотрена возможность определения стоимости аренды жилья для размещения на онлайн-площадке AirBnB. Для работы в качестве примера использованы данные с платформы Kaggle. Территория исследования – город Берлин, выбран как одно из популярнейших направлений для туризма, следовательно, показательное для изучения в области аренды временного жилья.

## Предварительный анализ данных

Данные, представленные в виде таблиц, содержат различную информацию о предоставляемом жилье, такую как количество комнат, количество гостей, которое данное жилье способно разместить, место расположения объекта, наличие различных удобств, влияющих на качество проживания, и стоимость. Так же имеется информация о владельце и отзывы посетителей.

В ходе работы проведен предварительный анализ данных с изучением их влияния на целевую переменную – стоимость аренды за сутки. Некоторые виды данных, такие как количество спален, стоимость различных дополнительных услуг, предоставляемых владельцем, могут использоваться в исходном виде. Другие же, такие как наличие уникальных удобств, тип предоставляемого жилья, условия отмены бронирования, нуждаются в преобразовании и представлении в другом, более понятном представлении. Для этого найдены категории по каждому виду данных, которые оказывают существенное влияние на стоимость, а сами данные далее отнесены к этим категориям.

## Обработка пропущенных значений

В данных имеются пропущенные значения, которые затрудняют работу над исследованием. Для борьбы с этим явлением применены различные подходы. Например, если отсутствует информация о дополнительной плате за уборку, то будем считать, что плата за данную услугу не взимается, или если нет данных о подтверждении личности владельца, будем считать ее неподтвержденной. Но так можно поступить не со всеми видами данных. Некоторые данные можно получить, основываясь на уже имеющихся других признаках. Так, для заполнения пропущенных значений площади жилья, а это довольно важная информация, которая может существенно повлиять на стоимость, использован алгоритм машинного обучения. Основываясь на информации о возможном количестве гостей, количестве ванных комнат, спален, стоимости залогового депозита, уборки, при помощи метода линейной регрессии заполнены недостающие значения. Таким образом мы получили примерные значения площади, но в данном исследовании важно не точное количество квадратных метров, а скорее представление о размере.

# Конструирование признаков

Проведя предварительный анализ, можно выявить зависимость целевой переменной от имеющихся признаков, определить влияние данных на результат исследования и отобрать признаки для использования в модели машинного обучения. Но также необходимо использовать неочевидные виды данных для формирования новых признаков, способных помочь в исследовании.

Такие признаки можно извлечь исходя из информации о месте расположения объекта аренды недвижимости. На первый взгляд может показаться, что на стоимость влияет удаленность конкретного места от центра города. Но так ли это в действительности? Проведя небольшой анализ, можно выявить тенденцию уменьшения стоимости по мере удаления от центра, но для получения более точных результатов этого недостаточно. Основываясь на идее о том, что определенное расположение относительно объектов, имеющих историческое или культурное значение, а также выгодное по отношению к объектам развлекательной сферы или для проведения досуга может существенно повлиять на стоимость, целесообразно использовать дополнительные источники информации. В данной работе использована информация проекта OpenStreetMap, предоставляющего данные о достопримечательностях и других объектах по всему миру.

Новые признаки, полученные на основе картографических данных представляют собой расстояние в километрах до определенных объектов городской инфраструктуры.

## Создание модели машинного обучения

На основе признаков, полученных после предварительной обработки данных, а также за счет конструирования новых в данной работе создано несколько моделей машинного обучения с применением различных алгоритмов:

1. Логистическая регрессия (Logistic Regression)
2. Метод опорных векторов (Support Vector Regression)
3. Случайный лес (Random Forest)
4. Градиентный бустинг (XGBoost)

## Результат

Результаты работы разных моделей представляют собой набор метрик из пакета ScikitLearn:

1. Коэффициент детерминации (R2)
2. Среднеквадратическая ошибка (RMSE)
3. Средняя абсолютная ошибка (MAE)

Сравнив имеющиеся показатели, можно сделать вывод, что модели, в которых использовался алгоритм градиентного бустинга показывают наилучшие результаты.

## Вывод

Модель машинного обучения, созданная на основе имеющихся признаков помогает в определении примерной стоимости объекта аренды. Чтобы повысить качество работы модели и добиться более высокой точности необходимы дополнительные данные. Немаловажно иметь представление о внешнем виде предлагаемых помещений. Дополнительные признаки, сконструированные на основе фотоснимков помогли бы создать модель с более высоким качеством работы.

## Перспективы

В области аренды временного жилья существует множество направлений, которые нуждаются в исследовании. Помимо основной задачи – определения стоимости – можно выделить и другие. Для владельцев недвижимости это могло бы быть планирование дохода, повышение эффективности за счет улучшения признаков, рекомендации по оптимизации прибыли. Нанимателей жилых помещений могли бы заинтересовать сервисы по оптимизации расходов, рекомендации по подбору наиболее выгодных предложений.