Санкт-Петербургский Государственный Университет

Направление Математическое Обеспечение и Администрирование
Информационных Систем
Профиль Информационные Системы и Базы Данных

Вахрушев Артем Андреевич

Прогнозирование уровня преступности на основе статистических данных

Магистерская диссертация

Научный руководитель:
к. ф.-м. н., доцент Графеефа Н. Г.

Рецензент:
ИТ эксперт Денисенко А. В.

Санкт-Петербург
2016

SAINT-PETERSBURG STATE UNIVERSITY

Main Field of Study Program Software and Administration of Information Systems
Area of Specialisation Information Systems and Databases

Vakhrushev Artem

Crime rate prediction based on statistics

Master's Thesis

Scientific supervisor:
docent Grafeeva Natalia

Reviewer:
IT expert Denisenko Aleksandr

Saint-Petersburg
2016

Оглавление

Введение **4**

1. Обзор существующих работ **6**

2. Постановка задачи **7**

3. Обзор методов предсказания временных рядов **8**

3.1. Регрессионные модели 8

3.2. Линейная авторегрессия 9

3.3. Модели экспоненциального сглаживания 11

3.4. Нейросетевые модели 12

3.5. Модели на основе цепей Маркова 14

4. Описание и предварительный анализ данных **16**

4.1. Описание данных 16

4.2. Визуализация аггрегированных данных 17

4.3. Связь температуры и числа преступлений 23

4.4. Географические температурные карты 25

5. Построение моделей прогнозирования **27**

5.1. Описание моделей 28

5.2. Построение моделей предсказания преступности 31

5.3. Задача классификации 32

Заключение **35**

Список литературы **36**

Введение

Сегодня проблема снижения уровня преступности стоит очень остро. С давних времен существуют государственные структуры, которые занимаются розыском и поимкой преступников, а также стремятся предотвратить преступления. В современном мире технологий очень развиты системы видеонаблюдения, которые позволяют быстро разыскать человека, совершившего преступление, а методы криминалистики практически не оставляют преступнику шансов уйти от правосудия, даже если он не попал в объективы камер. Но гораздо лучше, если благодаря действию правоохранительных органов преступление не произошло вовсе. Благодаря развитию информационных технологий уже существует несколько экспериментальных систем, позволяющих предсказывать вспышки числа преступлений в том или ином районе города. Но пока эти системы распространены только в развитых странах мира, таких, например, как США или Великобритания.

В настоящей работе планируется разработать методы предсказания преступлений. В качестве данных для построения моделей используются реальные данные о преступлениях в городе Чикаго. Эти данные содержат подробное описание каждого преступления произошедшего в период с 2001 по 2015 года. Такое детальное описание позволяет заметить закономерности, которое не так очевидны, узнать когда, где и каких преступлений происходит больше, тем самым, возможно, придумать модели, которые позволят заранее узнать о том, куда стоит направить дополнительный патруль.

Структура работы имеет следующий вид. В 1 главе приводится обзор существующих работ по предсказанию преступности. Для построения прогнозов авторы данных работ применяют методы анализа временных рядов, но в своих прогнозах они опираются только на исторические данные о преступности и не учитывают влияние различных внешних факторов. В главе 2 формируются основные цели и задачи исследования, а именно построения качественных прогнозов уровня преступности. В главе 3 приводится описание наиболее популярных моде-

лей анализа временных рядов, таких как регрессионные модели, авторегрессионные модели, модели экспоненциального сглаживания, нейросетевые модели и модели на основе марковских цепей. Глава 4 посвящена подробной визуализации данных и построению интерактивных карт преступности, которые позволяют наглядно представить данные о самых опасных районах и улицах города. В главе 5 на основе найденных закономерностей подбираются алгоритмы для построения прогнозов, строятся сами прогнозы и приводятся их оценки. Далее задача восстановления временного ряда сводится к задаче классификации для определения наиболее опасных дней с точки зрения количества преступлений. Данная задача решается поредством построения нейросетевой модели.

1. Обзор существующих работ

Основная литература, посвященная анализу данных, на основе которого строятся модели предсказания уровня преступности являются англоязычными. Вероятно, это связано с тем, что в анализ данных на современном этапе больше всего развит в США, а также именно в Штатах в открытом доступе можно найти официальные массивы данных полиции о преступности.

В современном мире очень остро стоит проблема безопасности, поэтому уровень преступности является важным фактором, например, при выборе места для переезда, выборе жилья или просто выборе ресторана для ужина в том или ином районе города. На предсказание уровня преступности влияет множество факторов разного рода, поэтому эта проблема находит свое отражение в научных работах ученых разных направлений. Так в работе "Predicting crime: a review of the research" [1] исследуется влияние демографических, экономических, психологических и других факторов на уровень преступности. Также в данной работе представлен обзор методов прогнозирования, однако самих прогнозов не делается.

В работе "Forecasting Crime: A City Level Analysis" [2] строятся линейные регрессионные модели прогнозирования числа убийств в городах США. Работа "Predicting Crime" [3] использует методы предсказания рыночных временных рядов для предсказания преступности. В этих двух работах модели основаны исключительно на исторических данных о преступности и не учитывают различные внешние факторы, которые могли бы помочь улучшить прогноз, а сами прогнозы строятся лишь по определенному виду преступления.

Как видно, основным средством для прогнозирования преступности являются методы анализа временных рядов. Так же для улучшения качества прогноза могут быть применены различные внешние факторы, которые оказывают влияние на число преступлений.

2. Постановка задачи

Целью настоящей работы является построение качественных прогнозов уровня преступности в городе Чикаго, на основании которых могут быть определы неблагоприятные дни, в которые число преступлений значительно превосходит средний уровень.

Для этого приводится описание наиболее известных моделей прогнозирования, исследуется взаимосвязь количества преступлений и различных внешних факторов (например, дня недели или погодных условий). Также на основании найденных данных приводится подробная визуализация - графики и таблицы, которые отражают наиболее опасное время суток, дни недели и месяцы в году, в которые необходимо усиливать полицейский контроль на улицах города. Помимо этого строятся интерактивные карты, которые легко позволяют найти наиболее опасные районы и улицы города. На основе полученных закономерностей строятся модели прогнозирования уровня преступности, которые учитывают как исторические данные по преступлениям так и различные внешние факторы. Приводятся оценки точности полученных результатов, которые позволяют судить о качестве прогнозов.

3. Обзор методов предсказания временных рядов

В классическом определении временным рядом называется совокупность измерений $y_1, y_2, \ldots, y_n, \ldots$, где $y_i \in \mathbb{R}$ некоторой величины через определенные (часто равные) промежутки времени.

Задача предсказания временного ряда записывается следующим образом:

$$\hat{y}_{t+d} = f_t(y_1, y_2, ..., y_t; \alpha, z),$$

где $\hat{y}_{t+d}$ - предсказанное значение искомой величины в некоторый момент в будущем, α - параметр модели, а z - другие факторы влияющие на измеряемую величину [4]. То есть требуется найти некоторую функцию f_t, которая на основании предыдущих значений временного ряда и, возможно, других факторов будет предсказывать следующие его значения.

Основные явления, которые могут наблюдаться во временных рядах:

- Тренды
- Сезонности
- Смена модели

Далее представлен обзор наиболее распространенных методов предсказания временных рядов.

3.1. Регрессионные модели

Часто на предсказываемую величину оказывают действия множество сторонних факторов. Так целью регрессионного анализа для прогнозирования временных рядов является построение некоторой функциональной зависимости между прогнозируемой величиной и этими факторами [5]. Самый простой вариант регрессионной модели - это линейная регрессия. Суть данного метода очень проста. Имеется неко-

торая переменная x, и мы пытаемся предсказать значение временного ряда y_t на основе значения этой переменной:

$$y_t = \alpha_0 + \alpha_1 x_t + \epsilon_t,$$

где α_0 и α_1 — коэффициенты регрессии; ϵ_t — ошибка модели.

Когда имеется несколько факторов влияющих на модель, то используется множественная регрессия:

$$y_t = \alpha_0 + \sum_{i=1}^{n} \alpha_i x_{it} + \epsilon_t$$

Если значения временного ряда имеют нелинейную зависимость от факторов, то стоит применять нелинейную регрессию:

$$y_t = F(x_t, \alpha) + \epsilon_t$$

где F - функция, которая, как мы предполагаем, хорошо описывает взаимосвязь между значениями факторов и значениями временного ряда, а $\alpha,\ x_t$ вектора коэффициентов и факторов соответсвенно.

Коэффициенты модели подбираются на основе имеющихся данных путем минимизации ошибки, например, методом наименьших квадратов [6] или методом максимального правдоподобия [7].

На практике часто бывает трудно получить значение внешних факторов в тот же момент времени, в который очуществляется прогноз, что является большим недостатком регрессионных моделей.

3.2. Линейная авторегрессия

Рассмотрим следующую формулу для предсказания значения временного ряда:

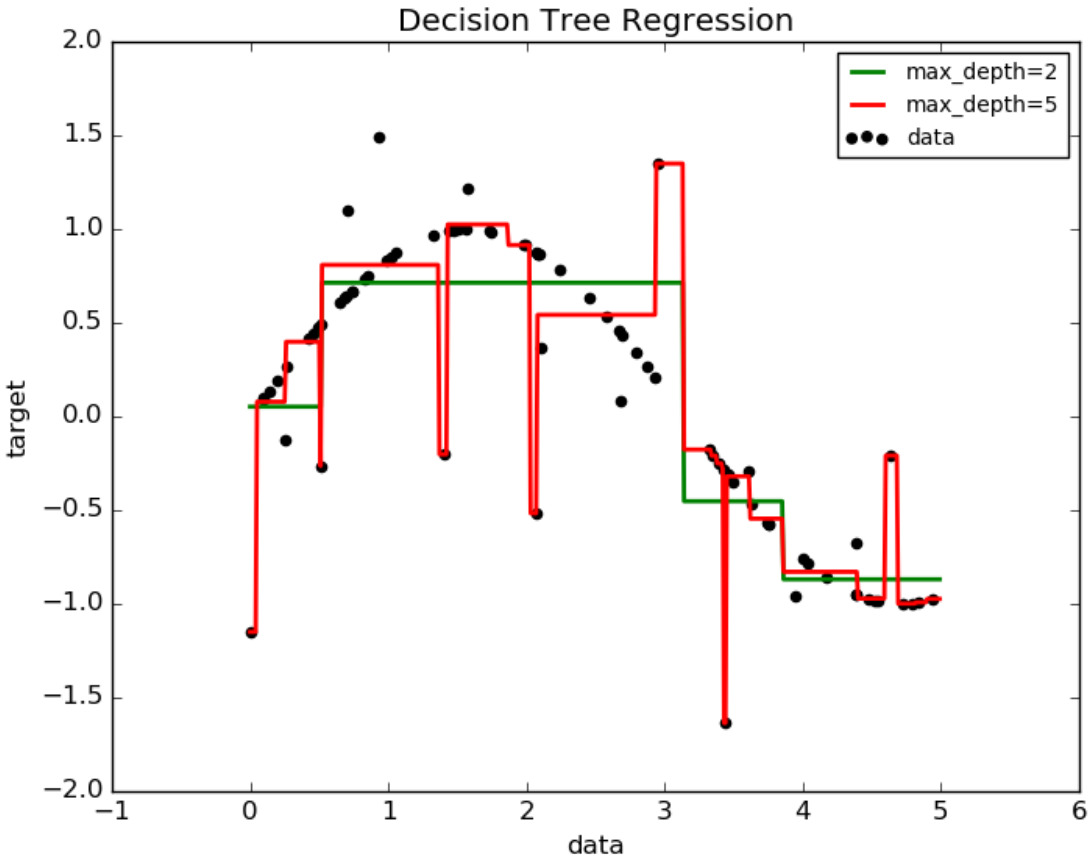
$$\hat{y}_{t+1} = \sum_{i=0}^{n} \alpha_i f(y_{t-i}),\ \alpha_i \in \mathbb{R}\ \forall i,$$

то есть для предсказания следующего значения используется n предыдущих значений ряда. Отсюда можно получить следующую матрицу

признаков и вектор ответов на для каждого набора признаков:

$$X = \begin{pmatrix} y_{t-1} & y_{t-2} & y_{t-3} & \cdots & y_{t-n} \\ y_{t-2} & y_{t-3} & y_{t-4} & \cdots & y_{t-n-1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ y_{n-1} & y_{n-2} & y_{n-3} & \cdots & y_0 \end{pmatrix}, \; y = \begin{pmatrix} y_t \\ y_{t-1} \\ \cdots \\ y_n \end{pmatrix}$$

В такой постановке задача может решаться самыми различными методами машинного обучения, поскольку входными данными для них являются матрицы признаков и ответов на них. Например, можно применить методы построения CART (Classification and Regression Tree) [8]. На рисунке представлен пример построения деревьев классификации для тестового временного ряда.



3.3. Модели экспоненциального сглаживания

Помимо авторегрессионных моделей существуют различные модели сглаживания. Одна из первых моделей сглаживания - это экспоненциальное сглаживание. Модель экспоненциального сглаживания представляет собой рекуррентное соотношение, в котором каждое последующее предсказываемое значение выражается через значение предсказанное на предыдущем шаге и истинное значение на предыдущем шаге:

$$\hat{y}_{t+1} = \hat{y}_t + \alpha(y_t - \hat{y}_t)$$

Однако, данная модель не учитывает различных явлений, которые могут наблюдаться в данных. Поэтому разрабатывались и другие модели, в которых учитываются тренды и сезонность.

Модель Хольта [9] описывает поведение временного ряда с линейным трендом по следующей формуле:

$$\hat{y}_{t+d} = a_t + db_t,$$

где $a_t,\ b_t$ - адаптивные компоненты линейного тренда, которые получаются из рекуррентных соотношений:

$$a_t = \alpha_1 y_t + (1 - \alpha_1)(a_{t-1} + b_{t-1})$$
$$b_t = \alpha_2(a_t - a_{t-1}) + (1 - \alpha_2)b_{t-1}$$

Модель Тейла-Вейджа [10] описывает данные с линейным трендом и аддитивной сезонностью:

$$\hat{y}_{t+d} = a_t + db_t\Theta_{t+(d \bmod s)-s},$$
$$a_t = \alpha_1 \left(y_t - \Theta_{t-s}\right) + \left(1 - \alpha_1\right)\left(a_{t-1} + b_{t-1}\right),$$
$$b_t = \alpha_3 \left(a_t - a_{t-1}\right) + \left(1 - \alpha_3\right) b_{t-1},$$
$$\Theta_t = \alpha_2 \left(y_t - a_t\right) + \left(1 - \alpha_2\right) \Theta_{t-s},$$

где s — период сезонности, $\Theta_i,\ i \in 0, \ldots, s-1$ — сезонный профиль, b_t — трендовый параметр, a_t — параметр прогноза, очищенный от тренда и сезонности. А $\alpha_1,\ \alpha_2,\ \alpha_3 \in (0, 1)$ параметры модели.

Модель Хольта-Уинтерса [11] описывает данные с экспоненциальным трендом и аддитивной сезонностью:

$$\hat{y}_{t+d} = a_t (r_t)^d \Theta_{t+(d \bmod s)-s},$$
$$a_t = \alpha_1 \cdot \frac{y_t}{\Theta_{t-s}} + (1-\alpha_1)\, a_{t-1} r_{t-1},$$
$$r_t = \alpha_3 \cdot \frac{a_t}{a_{t-1}} + (1-\alpha_3)\, r_{t-1},$$
$$\Theta_t = \alpha_2 \cdot \frac{y_t}{a_t} + (1-\alpha_2)\, \Theta_{t-s},$$

здесь, как и ранее, s — период сезонности, $\Theta_i,\ i \in 0, \ldots, s-1$ — сезонный профиль, r_t — параметр тренда, a_t — параметр прогноза, очищенный от влияния тренда и сезонности. Параметры $\alpha_1,\ \alpha_2,\ \alpha_3 \in (0, 1)$

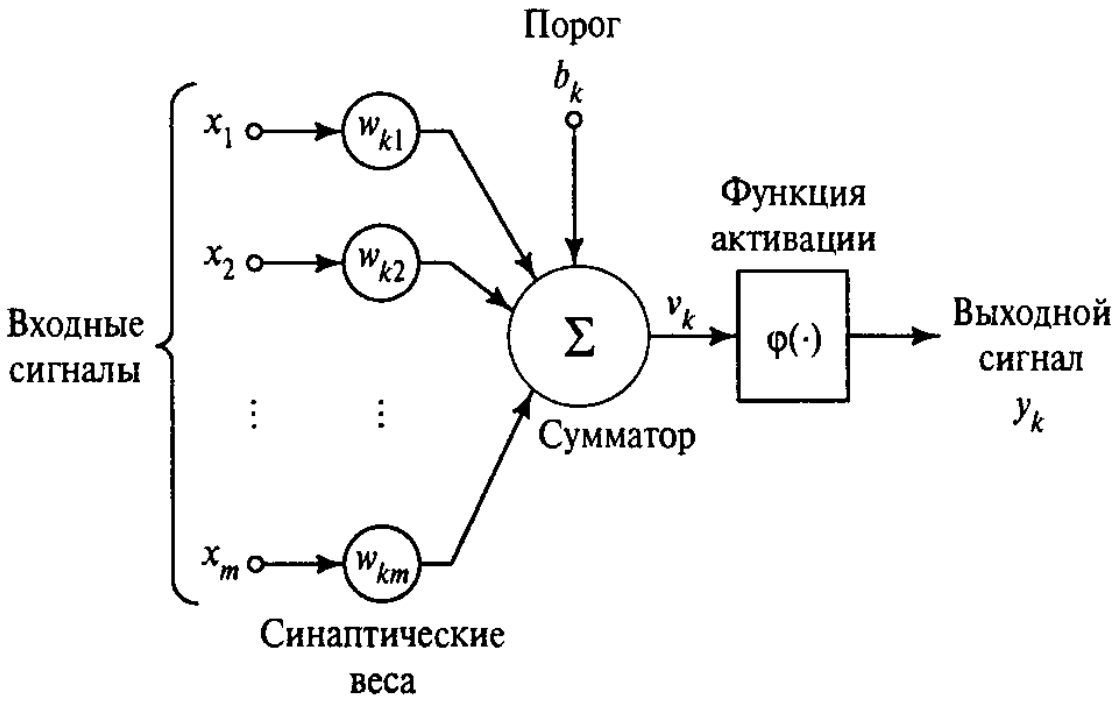
Все параметры рассмотренных моделей могут быть получены, например, минимизацией среднеквадратичной ошибки между предсказываемым значением $\hat{y}_t$ и наблюдаемым y_t:

$$\epsilon_t^2 = (y_t - \hat{y}_t)^2.$$

Здесь представлены наиболее распространенные модели сглаживания. Кроме того существует еще много моделей, различающихся видами сезонности и трендов.

3.4. Нейросетевые модели

В настоящее время нейросетевые модели довольно часто используются для предсказания временных рядов [12]. Нейросетевые модели строятся из нейронов. Стандартная модель нейрона приведена на рисунке ниже:



Как видно из рисунка модель нейрона для временного ряда можно описать следующими функциями:

$$u_t = \sum_{i=1}^{m} \omega_i x_{t-i} + \omega_0$$

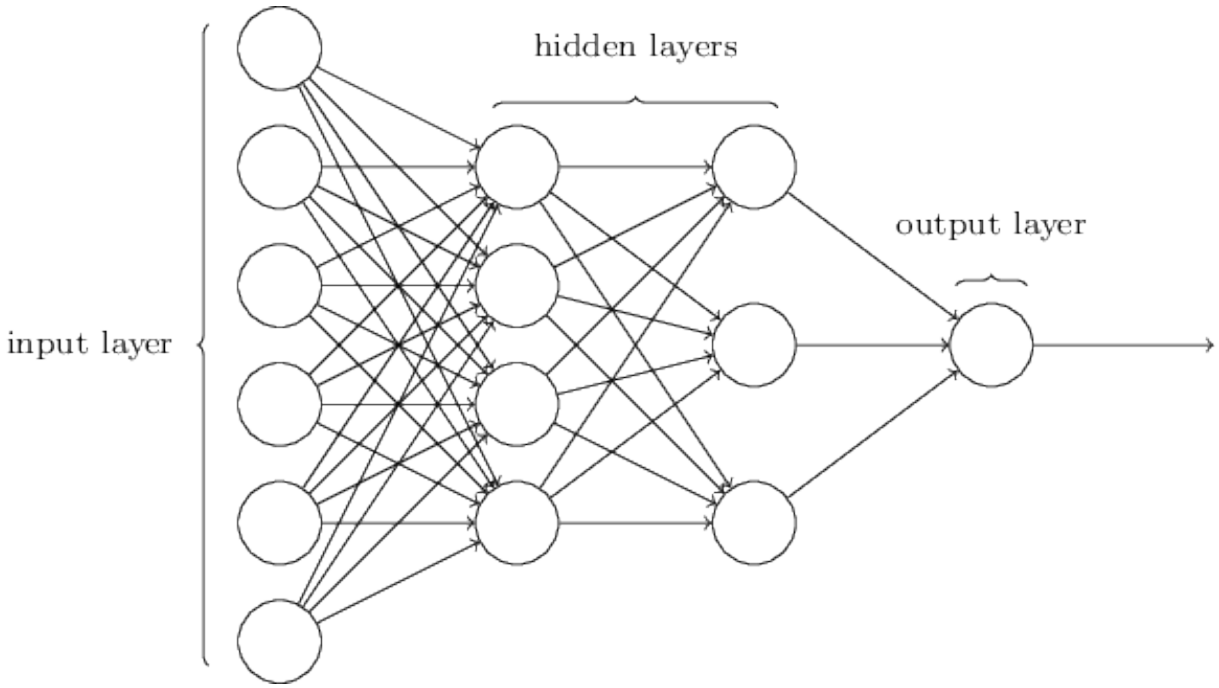
$$z_t = \phi(u_t)$$

здесь предыдущие значения ряда $y_{t-1}, y_{t-2}, \ldots, y_{t-n}$ представляются как входные сигналы, ω_i - веса сигналов, а ϕ - функция активации.

Функция активации бывает трех основных типов:

- функция единичного скачка
- кусочно-линейная функция
- сигмоидальная функция
- функция гиперболического тангенса

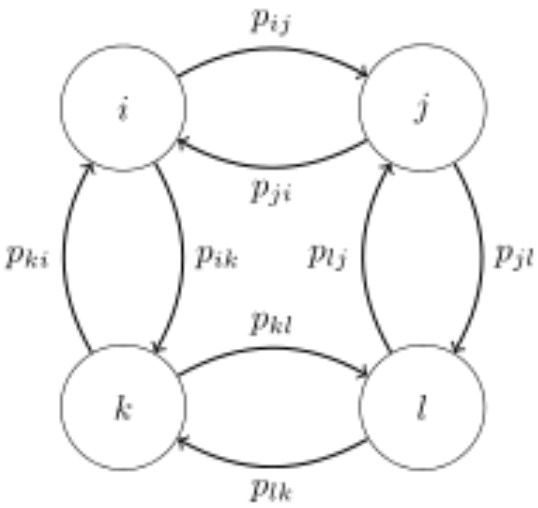
Из нейронов можно построить нейронные сети с различным числом слоев, например:



Таким образом можно строить достаточно сложные модели на основе предыдущих значений временного ряда. Модели могу быть нелинейными - это определяется числом слоев и функцией активации в каждом из них.

3.5. Модели на основе цепей Маркова

Так же для краткосрочного прогнозирование применяются Марковские цепи [13]. Схематичный вид такой цепи представлен ниже.



Здесь события $i,\ j,\ k,\ l$ - некоторые состояния временного ряда, а p_{ij} - соответствующие вероятности переходов. Если известна предыстория временного ряда, то моделируя процесс посредством марковской цепи можно получить вероятности нахождения системы в каждом из состояний через какой-то промежуток времени. Однако данная модель имеет явную сложность - это определение вероятностей переходов, которые должны быть заданы заранее.

4. Описание и предварительный анализ данных

Для анализа был выбран набор данных о преступлениях совершенных в Чикаго в период с 2001 по 2015 год [14].

4.1. Описание данных

Полученный набор данных имел следующий набор признаков:

1. Время совершения преступления
2. День недели
3. Район совершения преступления
4. Координаты широты
5. Координаты долготы
6. Тип преступления

Всего в наборе данных представлено более сорока типов преступлений. Однако, большая часть из них представлена очень небольшой выборкой, что делает данные типы преступлений непригодными для анализа. Поэтому для дальнейшего рассмотрения были выбраны наиболее распространённые и опасные виды преступлений, число которых в неделю превосходит 100:

- Нападение
- Побои
- Кражи со взломом
- Причинение ущерба
- Посягательство на преступление

- Мошеничество
- Похищение транспортных средств
- Преступления связанные с наркотиками
- Грабеж
- Воровство

Как видно, набор данных содержит небольшое число признаков, однако, такие признаки, как время и географичесие координаты, позволяют более подробно взглянуть на то, от чего зависит количество преступлений.

4.2. Визуализация аггрегированных данных

Посмотрим на графики различных преступлений (Рис. 1), на которых приведены аггрегированные данные об общем количестве преступлений. Сплошная линия на графике представляет собой экспоненциальное сглаженное количество преступлений.

Как видно таким преступлениям как нападения, побои, кражи со взломом, причинение ущерба, грабежи, воровство, присуща годовая сезонность. Также видно, что общее число случаев нападений, побоев, причинения ущерба , посягательство на преступление, распространения наркотиков, краж транспортных средств, грабежей, воровства снижается к 2015 году, а число случаев мошеничества, наоборот, возрастает. Однако стоит отметить, что в таких преступлениях, как кражи со взломом и кража транспортных средств, наблюдаются пики в 2005, 2011 годах соответсвенно.

Остановимся подробнее на календарной зависимоти. Ниже приведены две гистограммы для побоев и нападаний, на которых показана зависимость числа преступлений от месяца (Рис. 2). Видно, что больше преступлений совершается в теплые месяцы. Так же на каждой гистограмме видно, что в январе происходит больше преступлений, чем в другие зимние месяцы.

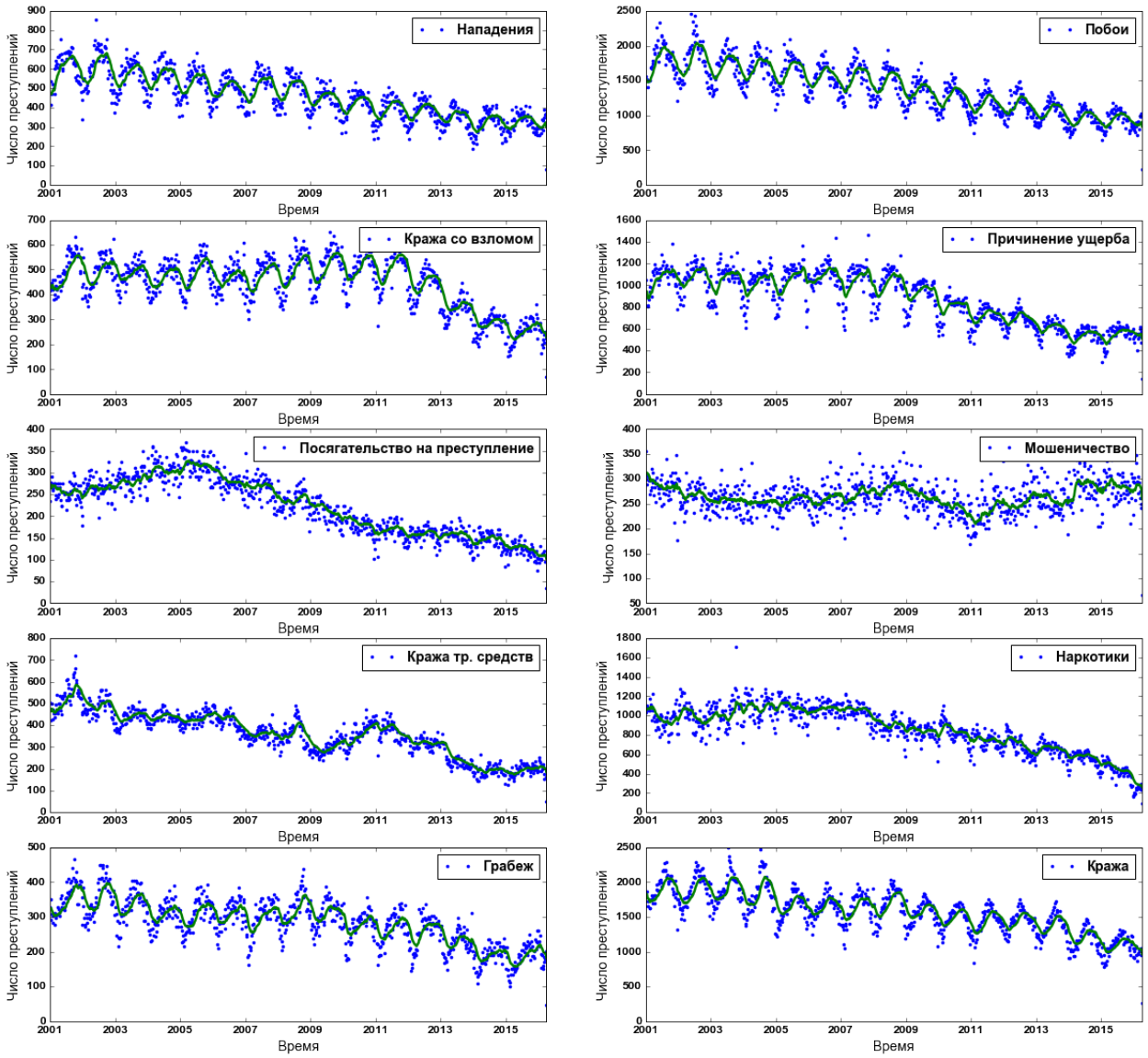


Рис. 1: Число преступлений каждого типа в разбивке по неделям

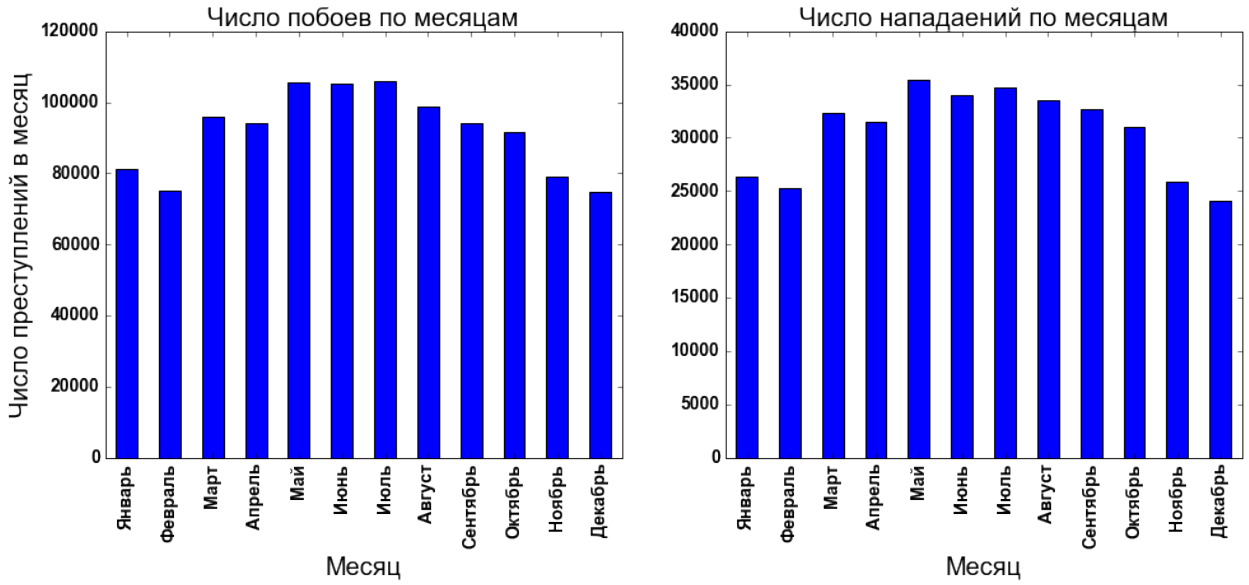


Рис. 2: Число преступлений по месяцам

Кроме того, влияние оказывает также время внутри дня и, например, номер дня в месяце (Рис. 3). Тепловая карта показывает, что намного больше преступлений совершается в первый день месяца в самом начале дня (0 часов). Так же наблюдается пик в полдень каждого дня, и, как и следовало ожидать, больше преступлений совершается вечером. Любопытно также что в середине месяца (15 числа) преступлений больше.

Следующий график дает представление о том, как распределены преступления по дням недели и часам (Рис. 4). Опять же можно заметить, что больше преступлений происходит в полдень, вечером, а так же ночью с пятницы на субботу и с субботы на восресенье. Также видно что ночью преступлений в целом совершается меньше, а на выходных этот минимум смещен.

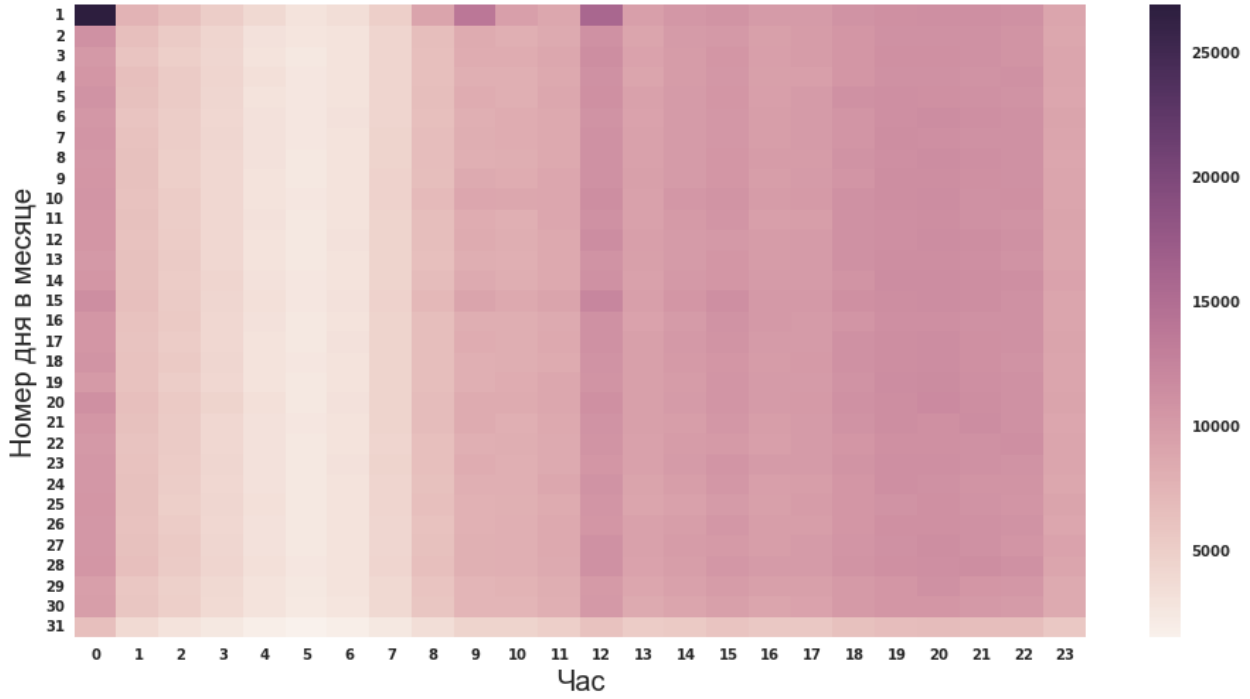


Рис. 3: Тепловая карта числа преступлений в зависимости от часа и дня месяца

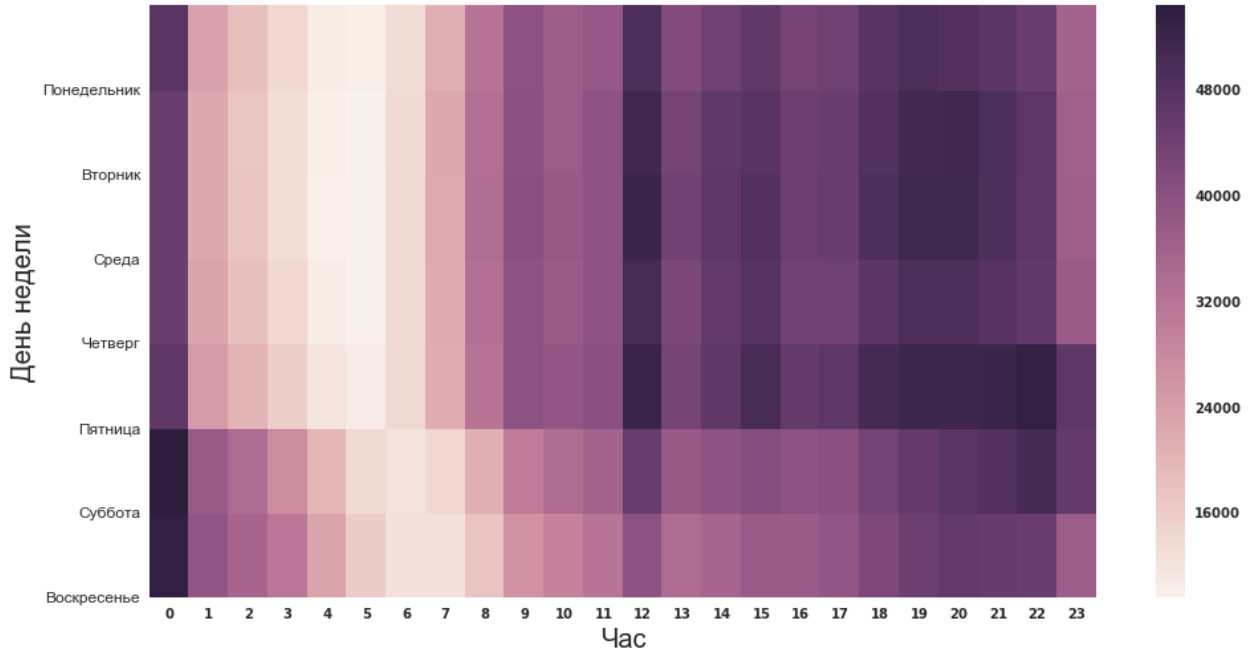


Рис. 4: Тепловая карта числа преступлений в зависимости от часа и дня недели в месяце

Теперь рассмотрим данные, по которым был построен Рис. 4, в разрезе по типам преступлений, и для каждого типа построим гистограмму (Рис. 5). Отчетливо видно, что побои, причинение ущерба чаще происходят в выходные, нападения, кражи со взломом, мошеничество, распространение наркотиков и обычные кражи, напротив, чаще происходят в будние дни и особенно в пятницу. Угон транспортных средств также выше в пятницу, чем в другие дни.

Рис. 5: Диаграмма разброса для каждого типа приступления по дням недели

4.3. Связь температуры и числа преступлений

Наибольшее число преступлений, как было показано, совершается в более теплые месяцы. Для проверки этой гипотезы, были получены архивные данные о погоде в Чикаго [15] и построен сравнительный график (Рис. 6). Хорошо видно, что средняя температура и количество преступлнений, для которых ранее была показана сезонность, связаны, причем больше преступленний происходит в теплую погоду. Для каждого преступления был подсчитан коэффициент корреляции Пирсона (результаты приведены в таблице 2):

$$r_{cw} = \frac{\sum_{i=1}^{m}(c_i - \bar{c})(w_i - \bar{w})}{\sqrt{\sum_{i=1}^{m}(c_i - \bar{c})^2 \sum_{i=1}^{m}(w_i - \bar{w})^2}},$$

где c_i, w_i - i-е измерения числа преступлнений и средней температуры соответсвенно, а $\bar{c}$, $\bar{w}$ их средние значения. Любопытно что все коэффициенты получились больше 0, а все значения p-value,кроме значения для мошеничества и угонов, меньше 0.05 (p-value показывает значимость корреляции и определяется по критерию Стьюдента).

Константа 0.05 стандартный порог для p-value. Значения p-value меньшие 0.05 говорят о том, что если принять гипотезу о значимости корреляции можно ошибиться только в 5% случаев. Таким образом, корреляция средней температуры и количества преступлений статистически значима в большинстве случаев (в таблице выделены красным), и данные о температуре могут помочь в построении моделей предсказания количества преступлений.

Тип преступления	Коэффициент корреляции	p-value
Нападение	0.800	$1.865 \cdot 10^{-12}$
Побои	0.885	$6.855 \cdot 10^{-18}$
Кражи со взломом	0.531	$6.052 \cdot 10^{-5}$
Приченение ущерба	0.816	$2.860 \cdot 10^{-13}$
Посягательство на преступление	0.494	$2.293 \cdot 10^{-4}$
Мошеничество	0.200	0.160
Угон транспортных средств	0.245	0.083
Распространение наркотиков	0.401	0.132
Грабежи	0.417	$2.342 \cdot 10^{-3}$
Воровство	0.866	$2.522 \cdot 10^{-16}$

Таблица 1: Корреляция средней температуры и числа преступлений

Рис. 6: Средняя температура и количество преступлений преступлений

4.4. Географические температурные карты

Также для более наглядной демонастрации были построены географические температурные карты(Рис. 7). Видно, что много преступлений происходит вдоль побережья, а также в восточной и южной частях города.

Также была построена интерактивная версия карты, на которой можно рассмотреть наиболее преступные районы подробнее и получить описание преступления [16].

Криминальность района может влиять на цену недвижимости или выбор места проживания. Поэтому были построенны интерактивные карты, по которым можно судить о безопасности того или иного жилого района [17]. Наиболее криминальные районы находятся, как и описано выше, вдоль побережья, на востоке и юге города.

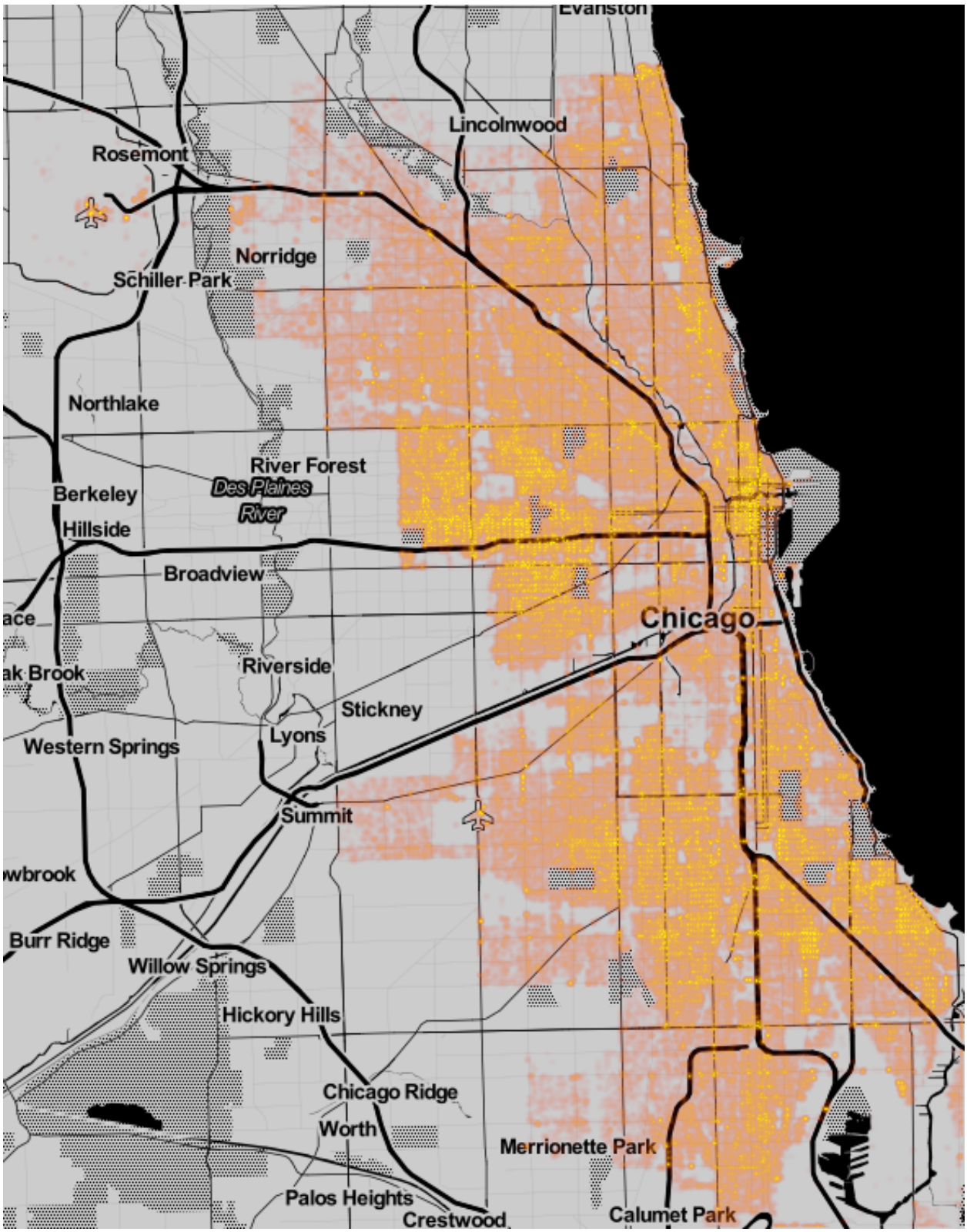


Рис. 7: Наиболее преступные улицы и районы города

5. Построение моделей прогнозирования

Для дальнейшего анализа и построение моделей были отобраны преступления следующих типов:

- Нападения
- Побои
- Кражи со взломом
- Причинение ущерба
- Мошеничество
- Грабежи
- Воровство

поскольку в предыдущие главе было показано что эти преступления имеют некоторую зависимость от временных и погодных факторов. Кроме того, если брать данные о преступности отдельно по каждому району, то их оказывается недостаточно для построения точных прогнозов.

Набор данных был преобразован для анализа, и для каждого вида преступления был представлен следующими признаками: число преступлений данного типа, совершенных в текущий день (целевая переменная); год; месяц; число месяца; номер дня в году; день недели; температура воздуха; влажность воздуха; давление воздуха и 14 предыдущих измерений числа преступлений.

Как видно, для построения наряду с предыдущими значениями временного ряда будут использоваться внешние факторы, такие как день недели или погодные условия. Поэтому исключительно авторегрессионные модели здесь не подходят, и необходимо использовать алгоритмы регрессии и авторегрессии совместно. Для этого отлично подходят модели на основе алгоритмов машинного обучения, которые принимают

на вход простую матрицу признаков, а на выходе дают значение целевой переменной.

В нашем случае матрица признаков получается достаточно просто, строка такой матрицы будет представлять совокупность описанных признаков для одного дня (год, месяц, день недели и т.д.), а ответом для одного такого набора признаков будет значение числа преступлений, произошедших в этот день. Таким образом, весь набор данных был преобразован к матрице признаков X, имеющей следующий вид:

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} & ... & x_{1m} \\ x_{21} & x_{22} & x_{23} & ... & x_{2m} \\ ... & ... & ... & ... & ... \\ x_{n1} & x_{n2} & x_{n3} & ... & x_{nm} \end{pmatrix}$$

.

И вектору ответов Y:

$$Y = (y_1, y_2, y_3, ...y_n)^T$$

5.1. Описание моделей

Для построения моделей использовался язык Python и пакет sklearn. Так как прогнозы строились для 7 типов преступлений, то, вполне ожидаемо, что одна модель не в состоянии описать каждый из типов преступлений, поэтому были выбраны 4 алгоритма для построения регрессии:

- Ридж регрессии [18]
- Регрессия на базе случайных лесов [8]
- Регресиия на базе решающих деревьев [19]
- Регрессия на базе SVM [20]

Ранее нами была получена матрица признаков X и вектор ответов Y для каждого типа преступления. Строки матрицы X называются объектами, а соответвующие элементы вектора Y ответами на этих

объектах. Каждый из представленных алгоритмов пытается восстановить неизвестную зависимоть $y : X \to Y$ путем построения некоторого приближения $a : X \to Y$, именно методом построения приближения a алгоритмы отличаются друг от друга. Остановимся на каждом из алгоритмов подробнее.

Чтобы описать алгоритм ридж регрессии стоит обратиться к алгоритму линейной регресси. Алгоритм линейной регрессии предполагает линейную связь каждого из параметров матрицы признаков и вектора ответов, то есть $a = (X, \theta)$, где θ - неизвестный параметр, а скобки означают скалярное произведение. Параметр θ ищется методом наименьших квадратов, то есть путем минимизации функционала:

$$Q = \|(X, \theta) - Y\|^2$$

Часто случается так, что признаки матрицы X оказываются мультиколлинеарными, то есть между ними наблюдается линейная зависимость. Что приводит к тому что решение задачи линейной регресси получается неустойчивым (элементы вектора θ получаются большими и разных знаков) . Алгоритм ридж регрессии пытается устранить данный недостаток путем добавления к функционалу Q регуляризатора:

$$Q = \|(X, \theta) - Y\|^2 + \tau\|\theta\|^2$$

где $\tau > 0$ - параметр регуляризации. Большим параметрам τ соответсвует более сильная регуляризация, то есть коэффициенты должны не сильно отклоняться от 0. Настраиваемым параметром в данной модели является параметр регуляризации τ.

Решающее дерево представляет собой дерево в листьях которого находятся значения апроксимирующей функции a, а узлы представляют собой условия перехода по ребрам (к примеру "температура больше 20 градусов"). На каждом шаге алгоритма по каждому признаку строится разделяющая плоскость, и для каждой части выборки, разделенных плоскостью, предсказывается среднее значение ответов объектов, попавших в эту часть пространства, и на основе этого считается средне-

квадратичная ошибка для каждого из полупространств. В итоге выбирается плоскость, которая разбивает пространство так, что суммарная среднеквадратичная ошибка минимальна. Вообще разделяющая плоскость может строиться по нескольким признакам, но это ведет к росту вычислительной сложности алгоритма. Критерием остановки алгоритма является либо достижение им максимальной начально заданной глубины дерева, либо наперед заданнай точности прогноза. Настраиваемыми параметрами модели являются глубина дерева, число измерений, по которым выбирается лучшее разбиение в узле, минимальное количество объектов в листе.

Алгоритм регрессии на базе случайных лесов представляет собой композицию множества решающих деревьев, которые представляют собой сильно усеченные, и построенные лишь на подмножестве признаков деревьев, обученных как описано выше. Дело в том, что в отдельности эти деревья не несут никакой ценности, поскольку делают очень неточные прогнозы, однако если взять большое число таких деревьев (больше 20) и усреднить все их показания, то прогнозы получаются достаточно точными. Настраиваемыми параметрами в данной модели являются число деревьев, количество признаков для обучения одного дерева, а так же параметры, используемые при обучении обычных деревьев.

Алгоритм регрессии на базе SVM ищет функция a в следующем виде:

$$a(x) = (\omega,\ x) + \omega_0$$

где в скобках указано скалярное произведение. Для нахождения вектора $\Omega = (\omega_0, \omega_1, \ldots, \omega_n)$ решается задача мимнимизации функционала:

$$Q = \sum_{i=1}^{n} (|(\omega,\ x_i) + \omega_0 - y_i| - \delta)_+ + \frac{1}{2C}\|\omega\|^2$$

где δ некоторая допустимая ошибка предсказания, функция $(z)_+ = z$ если $z < 0$, в противном случае равна 0, $\frac{1}{2C}$ как и ранее параметр регуляризации. заменной переменных $\zeta_i^+ = ((\omega, x_i) + \omega_0 - y_i - \delta)_+$ и $\zeta_i^- = (-(\omega, x_i) - \omega_0 + y_i - \delta)_+$ задача сводится к задаче квадратичного

программирования с линейными ограничениями:

$$\begin{cases} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\zeta_i^+ + \zeta_i^-) \to min_{\Omega,\zeta_i^+,\zeta_i^-} \\ y_i - \delta - \zeta_i^- \leq (\omega,\ x_i) + \omega_0 \leq y_i + \delta + \zeta_i^+ \ \forall i \\ \zeta_i^-, \zeta_i^+ \geq 0 \ \forall i \end{cases}$$

Решение данной задачи может быть найдено в следующем виде:

$$a(x) = \sum_{i=1}^{n}(\lambda_i^+ + \lambda_i^-)(x_i,\ x) + \omega_0,$$

где $\lambda_i^-,\ \lambda_i^+$ двойственные переменные, возникающие при решении задачи квадратичного программирования. Скалярное произведение $(x_i,\ x)$ может представлять собой не только классическое скалярное произведение, но и любую другую функцию от двух векторов обладающую всеми свойствами скалярного произведения, такая функция называется ядром, тогда выражение для $a(x)$ может быть записано в следующем виде:

$$a(x) = \sum_{i=1}^{n}(\lambda_i^+ + \lambda_i^-)K(x_i,\ x) + \omega_0,$$

, где $K(x,\ y)$-ядро.

Настраиваемыми параметрами описанного алгоритма являются ядро, параметры самого ядра, коэффициент при регуляризаторе C и допустимая ошибка δ.

5.2. Построение моделей предсказания преступности

Для поиска оптимальной модели часть данных за 2015 год была отложена для потроения и оценки самого прогноза. На оставшейся части данных методом перекрестной проверки по сетке параметров подбирался наилучший алгоритм и его параметры. Результат каждого алгоритма с каждым набором параметров оценивался по метрике MAPE (mean

absolute percentage error), которая позволяет судить о средней ошибке прогноза:

$$MAPE = \left(\sum_{i=1}^{k} \frac{|y_i - \hat{y}_i|}{y_i} \right) \cdot 100\%$$

.

Результаты работы наилучших алгоритмов для каждого типа преступлений приведены в таблице.

Тип преступления	Ридж Регрессия	Дерево решений	Случайный лес	SVM регрессия
Нападение	14.94 %	14.41 %	7.92 %	25.32 %
Побои	17.01 %	10.06 %	11.55 %	26.19 %
Кражи со взломом	24.33 %	14.15 %	14.90 %	9.98 %
Приченение ущерба	19.09 %	17.54 %	12.40 %	25.72 %
Мошеничество	10.77 %	18.28 %	9.09 %	12.08 %
Грабежи	12.93 %	17.57 %	12.29 %	20.93 %
Воровство	8.73 %	16.52 %	9.65 %	17.47 %

Таблица 2: Результаты построенных моделей

Результаты работы наилучшего для каждого преступления алгоритма выделены красным. Наилучшие результаты показал алгоритм регрессии на основе случайного леса.

5.3. Задача классификации

Построенные прогнозы кажутся весьма точными, однако, более полезным является предсказание аномальных дней, когда число преступлений гораздо выше чем в среднем. Для того чтобы показать как можно классифицировать преступления, были выбраны данные о побоях. Если рассмотреть данные о побоях в пределах любых 30-100 дней, то распределение числа побоев окажется нормальным. Поэтому можно, например, классифицировать число побоев следующим образом: если число побоев превосходит среднее число побоев за предыдущие 30 дней больше чем на одно среднеквадратичное отклонение, то этот день считается аномальным и относится к первому классу, остальные дни относятся к нулевому классу. Таким образом, задача восстановления регрессии может быть сведена к задаче бинарной классификации. В среднем оказывается, что на один год приходится порядка 70 аномальных дней.

Задача классификации решалась путем построения нейронной сети с тремя скрытыми слоями. Первый скрытый слой содержал 100 нейронов, второй 50, третий 10, данные цифры были подобраны экспериментально путем сравнения большого числа различных нейросетей. Входной слой имел линейную структуру, а все три скратых слоя были представлены сигмоидальными нейронами [21]. Выходной слой состоял из одного сигмоидального нейрона.

Сигмоидальная функция имеет значение от 0 до 1, поэтому будем интерпретировать значение выходного слоя как вероятность принадлежности к классу, если она больше 0.5, то относим объект (день) к класу 1, иначе к классу 0.

Нейронная сеть обучалась алгоритмом обратного распространения ошибки путем мимимизации среднеквадратичной ошибки [22]. Для оценки результатов работы нейросети использовались метрики точности (precision):

$$precicion = \frac{P}{N}$$

, где P число объестов данного класса, которые классифицируются правильно, а N общее число объестов классифицированных как данный класс. Так же используется мера полноты (recall):

$$recall = \frac{P}{M}$$

, где P как и выше, число объестов данного класса, которые классифицируются правильно, а M число объектов в тестовой выборке которые реально принадлежат нашему классу. Оценивать точность и полноту будем для первого класса, то есть класса, который описывает аномальные дни. В данной задаче классификации более важной является метрика полноты. Поскольку мы стремимся находить все опасные дни, которые присутствуют в тестовом наборе данных.

В резудьтате работы нейросети на тестовом наборе данных за 2015

год, были получены следущие значения точности и полноты:

$$precision = 0.70$$
$$recall = 0.55$$

Таким образом можно утверждать, что мы предсказываем порядка 55% плохих дней с точностью прогноза 70%. Что является неплохим результатом.

Такой подход сведения задачи регрессии к задаче классификации может быть применен так же и к другим типам преступлений, поскольку число преступлений в пределах 30-100 дней оказывается распределенным по нормальному закону для каждого из них. Далее теми же методами может быть построена нейронная сеть которая позволит находить аномальные дни. Однако эта задача требует больших временных затрат, так как расчет сложной нейросети вычислительно сложная задача.

Заключение

В данной работе приведен подробный анализ уровня преступности в Чикаго за 2001-2015 года. Была проанализирована связь уровня преступности и различных внешних факторов, таких как погодные условия, день недели, номер дня в году. Приведена визуализация данных, которая позволяет увидеть интересные закономерности. На основе географических данных о преступлениях, были построены интерактивные карты, на которых можно легко найти наиболее опасные улицы и районы города.

На основании выявленных закономерностей построены модели для прогнозирования уровня преступности для таких преступлений как нападения, побои, кражи со взломом, причинение ущерба, мошеничество, грабежи и воровство. Модели строились на данных за 2001-2014 года, а оценка проводилась на данных за 2015 год. Для построения моделей использовались алгоритмы машинного обучения, которые учитывают как исторические данные о преступлениях, так и другие внешние факторы. Приведен сравнительный анализ алгоритмов. Построеные модели показали достаточно высокую точность. Но наилучший результат показал алгоритм регресси на основе случайных лесов.

Далее задача прогнозирования числа преступлений сведена к задаче классификации для выявления наиболее опасных дней с точки зрения количества преступлений. Данная задача решалась путем построения нейросетевой модели. Полученные результаты так же кажутся достаточно удволетворительными.

Результаты полученные в рамках данного исследования могут оказаться полезными для прогнозирования уровня преступности и в других городах.

Данные методы прогнозирования временных рядов могут быть полезны для снижения уровня преступности, так как позволяют предсказывать неблагоприятные периоды, в которые необходимо усиливать патрулирование улиц города.

Список литературы

[1] Schneider S. Predicting crime: a review of the research. page 37, 2002.

[2] John V. Forecasting crime: A city level analysis. page 33, 2007.

[3] Henderson T., Wolfers J., and Zitzewitz E. Predicting crime. page 63, 2008.

[4] Бокс Дж. and Дженкинс Г.М. *Анализ временных рядов, прогноз и управление.* M., 1974.

[5] Draper N. and Smith H. *Applied regression analysis.* N.Y., 1981.

[6] Айвазян С.А. *Прикладная статистика. Основы эконометрики.* M., 2001.

[7] Andersen and Erling B. Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society*, pages 283–301, 1970.

[8] Meek C., Chickering D.M., and Heckerman D. Autoregressive tree models for time-series analysis. 2002.

[9] Holt C.C. Forecasting trends and seasonals by exponentially weighted moving averages, 1957.

[10] Theil H. and Wage S. Some observations on adaptive forecasting. 1964.

[11] Winters P.R. Forecasting sales by exponentially weighted moving averages. *Management Science*, (6):324–342, 1960.

[12] Ginzburg I. and Horn.D. Combined neural networks for time series analysis. pages 1–2.

[13] Liu T. Application of markov chains to analyze and predict the time series. 2009.

[14] City of Chicago. Crimes - 2001 to present. https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2, 2016.

[15] Historical weather data. Weather in chicago. https://www.wunderground.com/history/airport/ORD, 2016.

[16] Interactive geo heatmap. https://drive.google.com/uc?id=0BzDO_U5QbqF5U1poZjR1UUJwbWs, 2016.

[17] Neighborhood geo heatmap. https://drive.google.com/uc?id=0BzDO_U5QbqF5aldVUXY3M2taTXc, 2016.

[18] Тихонов А.Н. О решении некорректно поставленных задач и методе регуляризации. *Доклады Академии Наук СССР 151*, page 4, 1963.

[19] Breiman L. Random forests. page 33, 2001.

[20] Drucker H., Burges C., and Kaufman L. Support vector regression machines. page 9, 1997.

[21] Martin A. Discrete mathematics of neural networks: Selected topics. page 3, 2001.

[22] Хайкин. С. *Нейронные сети. Полный курс.* М., 2006.