УДК 343.2.7

ПРОГНОЗИРОВАНИЕ УРОВНЯ ПРЕСТУПНОСТИ В ГОРОДАХ НА ОСНОВЕ ДЕМОГРАФИЧЕСКИХ ДАННЫХ

Лозовицкая Галина Петровна **–** док-р.юр.наук., доцент, профессор кафедры государственно-правовых и уголовно-правовых дисциплин РЭУ им. Г.В. Плеханова; lozlina@yandex.ru

***Соколинский Юрий Валерьевич* —** студент РЭУ им. Г.В. Плеханова

Шалько Виктория Александровна — студент РЭУ им. Г.В. Плеханова

yurysokolinskiy@yandex.ru**,** v_shalko@mail.ru

- ФГБОУ ВО «Российский экономический университет им. Г.В. Плеханова»; Адрес: 115054, Москва, Стремянный переулок, д. 36

Тел.: +7 (495) 800-12-00;

Аннотация. Внешние факторы, такие как обострившееся геополитическое противостояние и введенные санкции, а также внутренние факторы, среди которых широкий разрыв между слоями населения, высокий уровень коррупции в стране формируют проблему активно растущего уровня преступности в современных экономических условиях, что подтверждает актуальность разработки предиктивных математических моделей, способных определить уровень преступности с учетом социально-демографических факторов.

Ключевые слова: уголовное право, социально-демографическая статистика, уровень преступности, регрессионная модель, прогнозирование.

FORECASTING THE CRIME RATE IN CITIES BASED ON DEMOGRAPHIC DATA

Lozovitskaya Galina Petrovna – Doctor of Law

Sokolinsky Yuri Valeryevich – is a student of Plekhanov RUE.

Shalko Victoria Alexandrovna – is a student of Plekhanov RUE

Annotation. External factors, such as the escalating geopolitical tensions and imposed sanctions, as well as internal issues, including a widening gap between

different segments of society, a high level of corruption within the country, contribute to the problem of a rapidly increasing crime rate in today's economic climate. This highlights the importance of developing predictive mathematical models that can accurately determine crime rates based on socio-demographic data.

Key words: criminal law, socio-demographic statistics, crime rate, regression model, forecasting.

Введение. Преступность в городах является одной из самых серьезных социальных проблем, требующих эффективных методов прогнозирования и профилактики. В последние годы с учетом технологического развития, позволяющего оценивать более сложные наборы данных, исследователи стали связывать анализ преступности с показателями социальной, демографической и других сфер общества. Демографические показатели, к которым относится половозрастная структура, количество браков и разводов и др., социальные показатели, среди которых уровень занятости, доля населения с доходами ниже прожиточного минимума, показатели, оценивающие жилищное положение граждан и прочие, могут оказывать существенное влияние на уровень преступности. Наличие исследований в данной области подтверждает научный интерес и актуальность разработки и применения статистической модели, способной прогнозировать количество преступлений с учетом уровня жизни населения. Цель статьи — построение и описание регрессионной модели, включающей в себя 23 социально-демографических факторов, влияющих на уровень преступности.

Обзор литературы. В современном научном поле существует большое количество статей, тезисов и исследований, в которых описывается взаимосвязь социально-демографических факторов с уровнем преступности. Например, в работе «Прогнозирование преступности: обзор результатов исследования» доктора наук, профессора Школы судебных исследований Университета Райерсона (Канада), Шнайдера С. исследуется влияние демографических, экономических, психологических и других факторов на уровень преступности.

Сам Шнайддер отмечает, что демографические факторы выступают в качестве сильнейших факторов, определяющих уровень преступности, и, следовательно, играют центральную роль в прогнозировании преступности [2]. До Шнайдера влияние социально-демографических факторов рассматривала исследовательская группа из Университета штата Нью-Йорк в Олбани, которые в своей работе «Жертвы личных преступлений: эмпирическая основа для теории личной виктимизации» сформулировали теорию воздействия образа жизни, сущность которой заключается в определении корреляции социально-демографических характеристик с числом совершенных преступлений. На основании проведенного исследования было выявлено, что определенные группы людей, как правило, ведут образ жизни, который подвергает их более высокому риску виктимизации по сравнению с другими группами, что и объясняется теорией воздействия образа жизни [3]. Данные исследования подтверждают актуальность и значимость исследований в сфере анализа связи социально-демографических факторов с уровнем преступности.

Также в современных условиях с развитием методов и моделей математического прогнозирования одной из наиболее точных остается регрессионная модель предсказания. Значимость использования регрессионных моделей в сфере исследования взаимосвязи социально-демографических показателей с числом совершенных преступлений подтверждается исследованиями Кадар К. и Плетикоса И., доктора наук, специалисты кафедры системного проектирования Швейцарской высшей технической школы Цюриха. Ученые в своей статье предлагают методологию оценки преступности, на основе трех различных регрессионных моделей с использованием машинного обучения [1]. Результаты построения моделей в их работе подтверждают актуальность использования метода регрессионного анализа для построения прогноза уровня преступности в регионах Российской Федерации с учетом социально-демографических факторов.

Результаты исследования. При построении регрессионной модели для анализа количества преступлений важно учитывать множество факторов, поэтому на входе были следующие показатели по регионам России за 2023 год:

1. Уровень бедности;
2. Численность населения;
3. Уровень урбанизации;
4. Уровень безработицы;
5. Среднедушевые доходы;
6. Кол-во разводов;
7. Кол-во браков;
8. Ввод жилья;
9. Кол-во мигрантов;
10. Кол-во преступлений;
11. Мужчины до труд. возраста;
12. Мужчины труд. возраста;
13. Мужчины старше труд. возраста;
14. Женщины до труд. возраста;
15. Женщины труд. возраста;
16. Женщины старше труд. возраста;
17. Плотность населения;
18. Площадь жилья на 1 жителя;
19. Доля ветхого жилья;
20. Кол-во трансп. средств;
21. Кол-во учреждений культуры;
22. Домашнее насилие;
23. Алкоголизм.

После выбора указанных показателей, было проведено их предварительное преобразование для повышения точности и устойчивости регрессионной модели:

- Все показатели были нормализованы, то есть приведены к диапазону от 0 до для того, чтобы устранить влияние различных масштабов данных и улучшит сходимость алгоритмов машинного обучения;
- После нормализации данные были логарифмированы, в связи с тем, что изначальная выборка данных имела ненормальное распределение. Логарифмирование применяется для приведения данных к более нормальному распределению, что делает их более подходящими для линейных моделей. Это позволяет снизить влияние выбросов и неравномерности в распределении данных.

Также, требуется вывить вид зависимости между предикторами и зависимой переменной, для этого были построены диаграммы рассеяния, которые показали, что большая часть зависимостей является линейной, однако, нормализация данных путем логарифмирования не дала 100% результатов, в связи с чем, часть показателей остались ненормально распределенными, данные факт, следует учесть при дальнейшем построении линейной регрессии и выбрать правильный метод построения.

Для анализа корреляции была выбрана корреляция Расстояний, так как она подходит для анализа количественных данных, в отличии от корреляций Тау Кандела и Спирмена, которые рассчитаны на ранговые данные. Использование корреляционной матрицы расстояний позволяет получить более точные и надежные оценки взаимосвязей между показателями, несмотря на ненормальное распределение данных. На Рисунке 1 представлена итоговая корреляционная матрица.

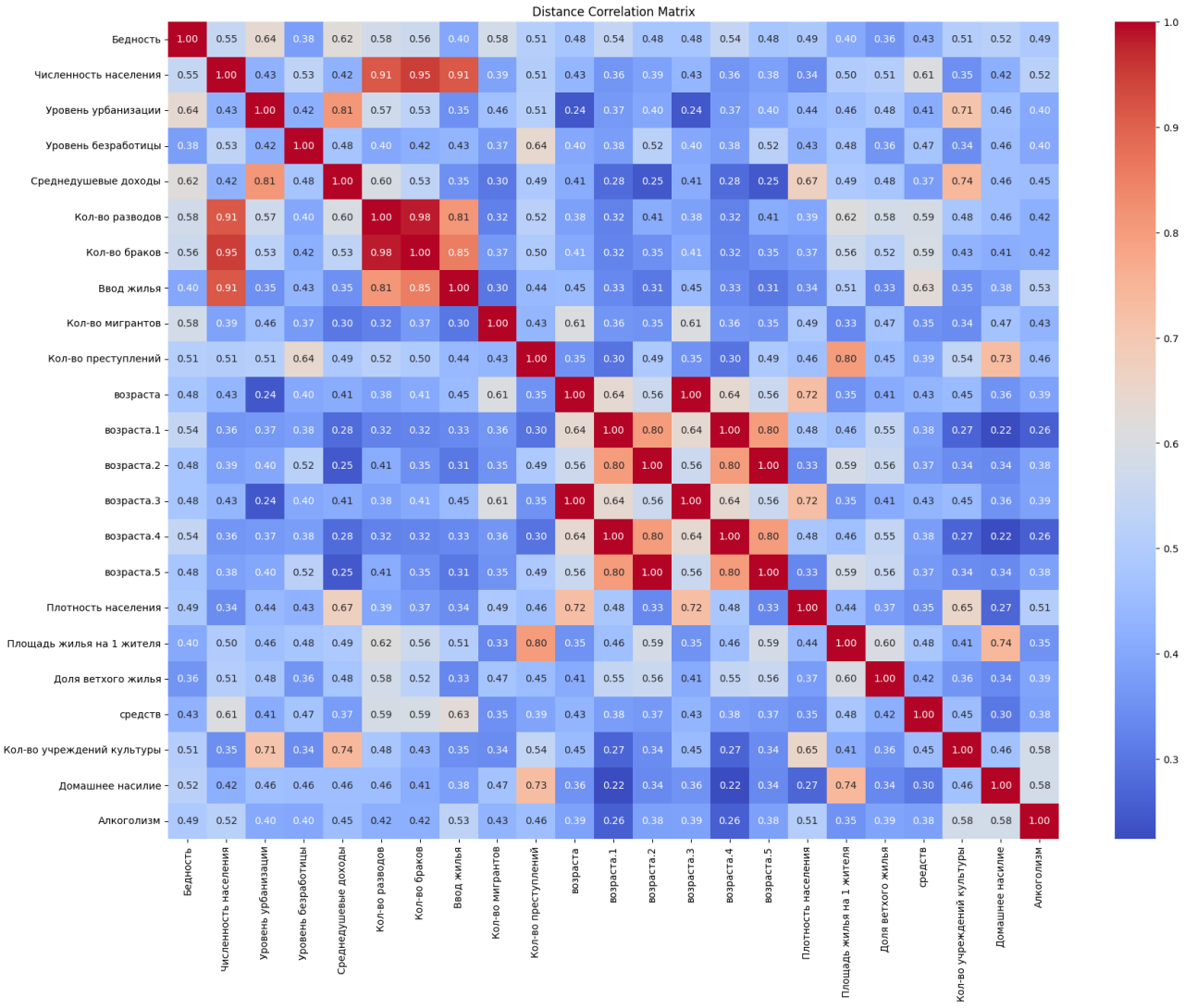


Рисунок 1 — Корреляционная матрица Расстояний (Distance Correlation)

На основе данной корреляционной матрицы были выбраны предикторы, которые наибольшим образом коррелируют с зависимой переменной (Кол-во преступлений) и имеют наименьшую корреляцию между собой.

Для построения матрицы Расстояний был использован следующий фрагмент кода:

```
def calculate_distance_correlation_matrix(df):
    columns = df.columns
    n = len(columns)
    dist_corr_matrix = np.zeros((n, n))
    for i in range(n):
        for j in range(n):
            dist_corr_matrix[i, j] = dcor.distance_correlation(df[columns[i]], df[columns[j]])

    return pd.DataFrame(dist_corr_matrix, index=columns, columns=columns)

dist_corr_matrix = calculate_distance_correlation_matrix(data)
```

Листинг 1 —корреляционная матрица Расстояний

На основе проведенного анализа, включающего диаграммы рассеяния и корреляционную матрицу расстояний, было принято решение использовать метод наименьших модулей (Least Absolute Deviations, LAD) для построения линейной регрессии. Он был выбран в силу того, что метод наименьших модулей более устойчив к данным, не следующими нормальному распределению. В отличие от метода наименьших квадратов, который чувствителен к выбросам и отклонениям от нормальности, LAD минимизирует сумму абсолютных отклонений, что делает его более подходящим для анализа наших данных. Для определения LAD был использован фрагмент кода:

```
def on_button_clicked(b):
    with output:
        output.clear_output()
        selected_features = [cb.description for cb in checkboxes if cb.value]
        if not selected_features:
            print("Выберите предикторы.")
            return
        X = data[selected_features]
        y = data['Кол-во преступлений']
        X = sm.add_constant(X)
        model = sm.QuantReg(y, X).fit(q=0.5)
        print(model.summary())
```

Листинг 2 — Линейная регрессия методом наименьших модулей

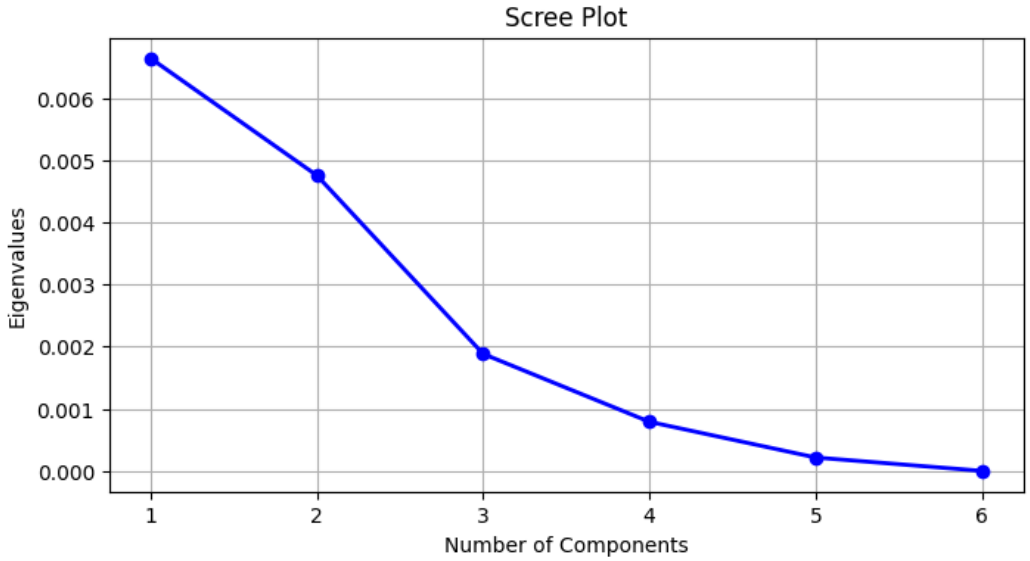
На основе построенной корреляционной матрицы, были выбраны предикторы, которые имеют высокую корреляцию с зависимой переменной и наименьшую друг с другом:

- Бедность (X1);
- Уровень урбанизации (X2);
- Уровень безработицы (X3);
- Площадь жилья на 1 жителя (X4);
- Случаи домашнего насилия (X5).

В результате построения модели было получено уравнение регрессии, где Y — количество преступлений. Полученное уравнение регрессии имеет вид:

$$Y = 0.615 - 0.352 \cdot X1 - 0.422 \cdot X2 - 0.524 \cdot X3 - 0.466 \cdot X4 + 0.237 \cdot X5$$

Также, проведенный анализ главных компонент показал, что четыре компоненты регрессии объясняют более 95% модели, однако, именно при 5 предикторах, p-value показателей и константы достигают максимального значения, также именно при 5 предикторах значение Pseudo R наивысшее. На Рисунке 2 представлен график «каменистой осыпи», который описывает



зависимости собственных значений от числа факторов в порядке их выделения.

Рисунок 2 — График «каменистой осыпи»

В полученной модели Pseudo R-квадрат равен 0,7641, что указывает на хорошее соответствие модели, объясняющее около 76,41% вариации зависимой переменной. Bandwidth (0,01085) и Sparsity (0,02488) находятся в допустимых пределах. Bandwidth выбирается таким образом, чтобы обеспечить достаточное сглаживание и избежать переобучения. Значение Sparsity указывает на низкую разреженность в ковариационной матрице остатков, что способствует точным оценкам стандартных ошибок и доверительных интервалов. Эти параметры подтверждают адекватность модели и точность полученных результатов. Также, стоит отметить низкое значение p-value предикторов и константы (<0.05), что значит статистическую значимость каждого из показателей, что подтверждает их важность в модели (см. Рис. 3).

QuantReg Regression Results

Dep. Variable:	Кол-во преступлений	Pseudo R-squared:	0.7641
Model:	QuantReg	Bandwidth:	0.01085
Method:	Least Squares	Sparsity:	0.02488
Date:	Wed, 29 May 2024	No. Observations:	12
Time:	15:56:14	Df Residuals:	6
		Df Model:	5

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.6154	0.066	9.338	0.000	0.454	0.777
Бедность	-0.3529	0.081	-4.361	0.005	-0.551	-0.155
Уровень урбанизации	-0.4223	0.096	-4.397	0.005	-0.657	-0.187
Уровень безработицы	-0.5249	0.135	-3.897	0.008	-0.854	-0.195
Площадь жилья на 1 жителя	-0.4655	0.074	-6.265	0.001	-0.647	-0.284
Домашнее насилие	1.0317	0.237	4.351	0.005	0.452	1.612

Рисунок 3 — Результаты линейной регрессии методом наименьших модулей

Выводы. Проведенное исследование показало, что уровень преступности в существенно зависит от социальных факторов больше, чем от демографических. Среди наиболее значимых предикторов были выявлены: уровень бедности, урбанизация, безработица, площадь жилья на одного жителя и случаи домашнего насилия. Регрессионная модель, построенная с

использованием метода наименьших модулей (LAD), продемонстрировала высокую точность в прогнозировании уровня преступности, объясняя 76,41% его вариации. Таким образом, для эффективного управления и профилактики преступности необходимо учитывать данные ключевые социально-демографические факторы.

Литература:

1. Kadar C., Pletikosa I. Mining large-scale human mobility data for long-term crime prediction. EPJ Data Sci. 7, 26 (2018). [Электронный ресурс]. — Режим доступа: https://doi.org/10.1140/epjds/s13688-018-0150-z
2. Schneider S. Predicting crime: a review of the research. Summary Report // Department of Justice Canada. 2004. [Электронный ресурс]. — Режим доступа: https://www.justice.gc.ca/fra/pr-rp/index.html
3. Hindelang M. J., Gottfredson M. R., Garofalo J. Victims of personal crime — an empirical foundation for a theory of personal victimization // U.S. Department of Justice. [Электронный ресурс]. — Режим доступа: https://clck.ru/3AtTZi