# Государственное бюджетное общеобразовательное учреждение города Москвы "Школа № 1554"

**Анализ данных национального виктимизационного опроса**

Над проектом работали:

Литовченко Максим Евгеньевич, 11 «Б» класс, ГБОУ «Школа 1554»

Каюмов Вадим Рустамович, 11 «Б» класс, ГБОУ «Школа 1554»

Научные руководители:

Сиухина Арина Алексеевна, 3 курс, мехмат факультет МГУ им. М.В. Ломоносова

Смирнова Ирина Сергеевна, ГБОУ «Школа 1554»

**Москва, 2021 г.**

**Содержание**

1. Введение ................................................................................................................................ 1
2. Обоснование выбора темы, актуальность работы, анализ существующих решений .... 1
3. Цель, и задачи работы, гипотеза исследования ................................................................. 2
4. Методика и ход выполнения работы .................................................................................. 2
   1. Предобработка данных ............................................................................................ 3
   2. Анализ выборки лиц, являющихся жертвами преступлений .............................. 4
      1. Анализ влияния полового признака на совершение преступлений и подверженность им ...................................................................................... 4
      2. Был ли найден нападавший ......................................................................... 5
      3. Была ли необходима медицинская помощь ............................................... 5
      4. Был ли нападавший при исполнении ......................................................... 7
      5. Анализ частоты видов преступлений ......................................................... 8
      6. Анализ влияния возраста на подверженность преступлениям ................ 9
      7. Анализ влияния дохода на подверженность преступлениям ................. 11
      8. Анализ влияния образования на подверженность преступлениям ........ 12
      9. Анализ влияния социально-экономического статуса на подверженность преступлениям ............................................................................................ 15
      10. Анализ скорости реагирования правоохранителей ................................. 17
      11. Анализ материального ущерба пострадавших ........................................ 18
      12. Анализ зависимости вида преступлений от времени суток ................... 19
      13. Портрет человека, который с наибольшей вероятностью является жертвой преступления ............................................................................... 21
   3. Анализ выборки лиц, не являющихся жертвами преступлений ........................ 21
      1. Анализ полового признака ........................................................................ 21
      2. Анализ возраста .......................................................................................... 22
      3. Анализ уровня дохода ................................................................................ 22
      4. Анализ уровня образования ....................................................................... 24
      5. Анализ социально-экономического статуса ............................................ 24
      6. Портрет человека, который с наибольшей вероятностью не является жертвой преступления ............................................................................... 25
   4. Сравнение выборок из лиц, являющихся жертвами преступлений и лиц, не являющихся жертвами преступлений .................................................................. 26
      1. Сравнение по половому признаку ............................................................ 26
      2. Сравнение по возрасту ............................................................................... 26
      3. Сравнение по уровню дохода .................................................................... 28
      4. Сравнение по уровню образования .......................................................... 28
      5. Сравнение по социально-экономическому статусу ................................ 28
      6. Общий вывод из сравнения выборок ........................................................ 32
   5. Классификационная модель .................................................................................. 32
      1. Создание необходимых выборок .............................................................. 32
      2. Создание модели ......................................................................................... 32
      3. Обучение модели ........................................................................................ 32
      4. Расчет точности модели и кросс-валидация ............................................ 33
      5. Применение модели к случайным данным .............................................. 35
      6. Применение модели к портретам лиц, с наибольшей вероятностью являющихся и не являющихся жертвами преступлений ........................ 36
      7. Вывод по построению классификационной модели ............................... 36
5. Результаты проделанной работы ...................................................................................... 37
6. Список использованной литературы ................................................................................ 38

**Введение**

# Данный проект посвящен изучению преступности на территории Российской Федерации и выполнен в рамках программы поиска и разработки научно-технологических проектов для школьников «Сириус.Лето: начни свой проект» и научно-практической конференции «Инженеры будущего» при поддержке МГУ им. М.В. Ломоносова. Исходные данные для анализа собраны Институтом проблем правоприменения (ИПП) при ЕУ СПб и представляют результаты телефонного опроса респондентов не моложе 18 лет на основании простой случайной выборки телефонных номеров. Респонденты опрашивались не зависимо от гражданства, но, по понятным причинам, большинство (98.1%) были гражданами РФ. Вопросы делились на анкетные (пол, возраст, социо-демографический профиль), основной («были ли жертвой преступления») и уточняющие вопросы о преступлениях (где, когда и т.п.). Анкета составлена специальным образом профессионалами, учитывая многие особенности, например, специфику русского языка.

##### Обоснование выбора темы, актуальность работы, анализ существующих решений

##### Россия – одна из крупнейших юрисдикций мира, имеющая огромные объемы анализируемых данных о работе органов внутренних дел. Однако, ничтожно малым остается объем исследований о работе правоохранителей и о преступности в стране, которые бы проводились независимыми экспертами. До сих пор такие исследования в России проводятся либо на сверхмалых нерепрезентативных выборках, либо собирают очень мало информации (обычно только сведения о самом факте виктимизации), что вызвало в нас интерес к проведению данного исследования.

##### Результаты данной работы могут быть полезны подведомственным органам Министерства внутренних дел Российской Федерации, правозащитным организациям, поскольку подобные исследования дают возможность объективно оценить уровень преступности в стране, так как официальная статистика не только подвержена искажениям, но и не отражает тех случаев преступлений, по поводу которых пострадавшие не обращались в полицию, дают возможность охарактеризовать качество работы полиции.

**Цель и задачи**

# Цель работы – выяснить, как влияют социально-демографические характеристики человека на его возможность оказаться жертвой преступления (виктимность).

# Задачи данного исследования:

* Ознакомиться с форматом данных, осуществить предобработку.
* Сделать анализ данных опроса жителей РФ о преступлениях, жертвами которых они становились.
* Сделать визуализации найденных закономерностей.
* Исходя из выявленных зависимостей, построить классификационную модель, определяющую виктимность человека.

**Гипотеза исследования –** социально-демографические параметры человека влияют на его вероятность быть подверженным преступлению.

**Методика и ход выполнения работы**

Для того, чтобы определить признаки, влияющие на виктимность человека, был проведен анализ социально-демографических характеристик людей из двух выборок: лиц, являющихся жертвами преступлений, и лиц, не являющихся жертвами преступлений, в результате по выявленным признакам были составлены идеальные портреты людей, которые с наибольшей вероятностью являются жертвой преступления и не являются соответственно. Кроме того, в ходе анализа выборки жертв преступлений, в частности было проведено общее исследование показателей, дающих более подробную информацию о преступлениях и их обстоятельствах.

В ходе исследования использовался язык программирования Python и такие библиотеки для анализа данных, как Pandas, NumPy. Для импортирования данных в код был использован модуль utils высокоуровневого API Keras библиотеки TensorFlow. Для визуализации полученных корреляций и закономерностей (построения гистограмм, ящиков размаха, и др.) использовались библиотеки Matplotlib, Seaborn. Для построения классификационной модели использовались библиотеки Scikit-learn, классификационная модель CatBoostClassifier из библиотеки CatBoost (Открытая программная библиотека для построения моделей машинного обучения, разработанная компанией Яндекс). Также в ходе исследования были использованы методы математической статистики, в частности медианные и средние значения, межквартильный размах (для исключения выбросов данных), расчет процентных долей выборок.

Перед началом исследования была изучена техническая документация к данным, и в результате было принято решение использовать только сконструированные переменные, которые не несут в себе субъективных оценок респондентов опроса и являются необходимыми для анализа, добавив к ним переменную offender\_is\_on\_duty – логическую переменную, содержащую информацию о том, был ли преступник при исполнении своих служебных обязанностей.

После загрузки данных была осуществлена их первичная предобработка: проверка на дубликаты, удаление ненужных столбцов. Затем было определено процентное отношение в генеральной совокупности людей, которые являются и не являются жертвами преступлений, и по полученным значениям построена диаграмма (рис. 1). Таким образом, почти каждый четвертый опрошенный человек является жертвой преступления.

Для последующего анализа данные были разбиты на две ранее указанные выборки. В ходе изучения данных в обоих выборках, было замечено, что в выборке лиц, не являющихся жертвами преступлений, присутствуют столбцы, содержащие только пустые значения или нули, поэтому они были исключены.

Рис. 1

Отношение лиц, являющихся и не являющихся жертвами преступлений

Далее был проведен анализ выборки жертв преступлений.

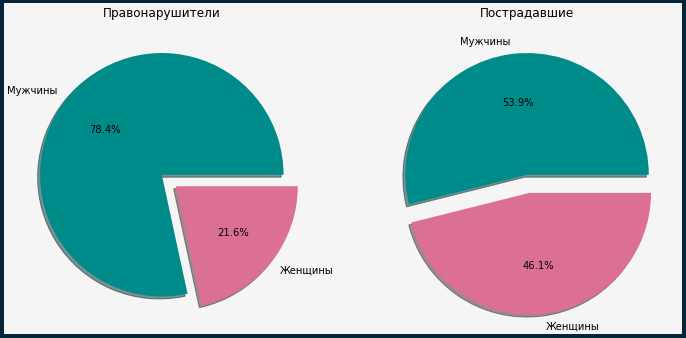
Вначале было изучено влияние полового признака на подверженность преступлениям, на участие в них. Для этого было найдено процентное отношение среди пострадавших и правонарушителей (если пол преступника был установлен), и по полученным значениям были построены графики (рис.2). Таким образом, были сделаны выводы:

* Исходя из случаев, где пол правонарушителя был установлен, мужчины чаще замешаны в преступлениях (в качестве жертвы или правонарушителя).
* Исходя из тех случаев, где пол правонарушителя был установлен, женщины являются преступниками лишь в четверти случаев правонарушений.
* Среди пострадавших процент мужчин и женщин отличается незначительно.

Далее было изучено, насколько много случаев правонарушений, когда правонарушитель был найден. Для этого был определено процентное отношение случаев, когда преступник был и не был найден, и по полученным значениям построена диаграмма (рис.3). Исходя из полученных значений, был сделан вывод, что, по данным опроса, более чем в трёх из четырёх случаев преступления правонарушитель остается ненайденным.

Рис. 2

Отношение мужчин и женщин среди правонарушителей и пострадавших



Затем было определена процентная доля жертв преступлений, которые впоследствии обратились за медицинской помощью. Для этого предварительно была отобрана выборка людей, пострадавших от физического насилия, и по полученным значениям был построен график (рис. 4). Исходя из этого, был сделан вывод: почти в половине случаев физического насилия пострадавший от преступления нуждался в медицинской помощи.

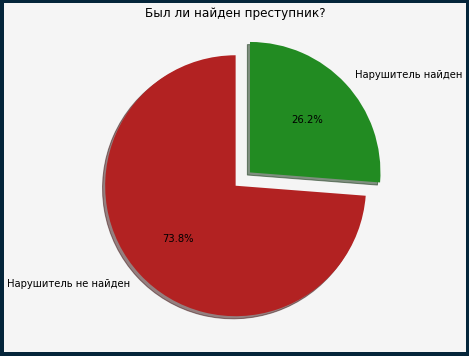
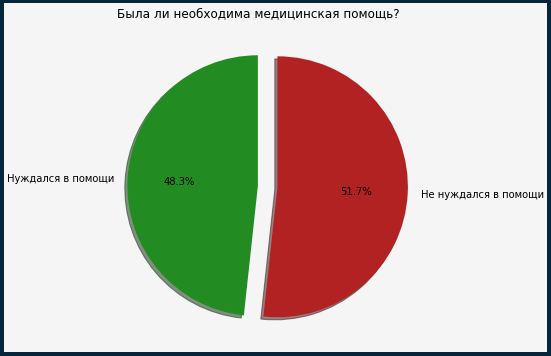


Рис. 3

Отношение случаев, когда преступник был и не был найден

Рис. 4

Отношение лиц, которые нуждались и не нуждались в мед. помощи



Далее были определены случаи, в которых преступник был при исполнении своих служебных обязанностей, а также вид его занятости. Для этого было найдено процентное отношение правонарушителей, которые были и не были при исполнении, и процентное отношение видов занятости среди преступников (рис. 5). Исходя из полученных в результате анализа соотношений, были сделаны следующие выводы:

* Более 20% злоумышленников находились при исполнении своих обязанностей.
* Большая часть правонарушителей, находящихся при исполнении обязанностей, относится к категории сотрудников коммерческих организаций.
* Более 30% правонарушителей являются чиновниками (госслужащими), среди них большую часть преступлений совершают правоприменители (17.5%), меньшую – военные (2.9%), а представители власти, по данным опроса, занимают 11.3% от всех правонарушителей при исполнении.

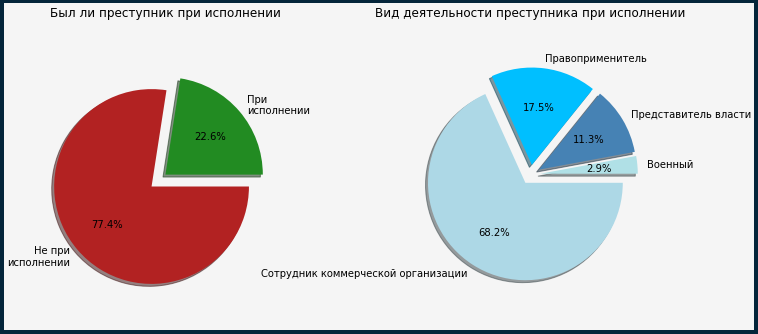


Рис. 5

Отношение лиц, которые были и не были при исполнении служебных обязанностей

Отношение видов деятельности преступников при исполнении обязанностей

Впоследствии был произведен анализ встречаемости каждого вида правонарушений. Для этого было определено процентное отношение каждого типа преступлений (рис. 6). Исходя из полученных в результате анализа соотношений, были сделаны следующие выводы:

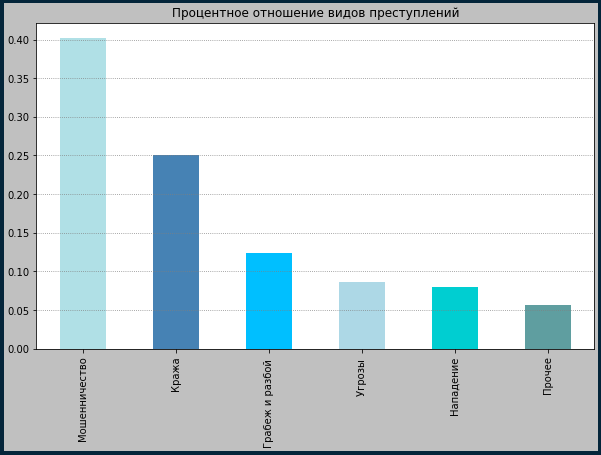
* Наиболее частый вид преступления по данным опроса – мошенничество (почти половина от всех преступлений).
* Кражи (тайные хищения) в 2 раза более распространены чем грабежи и разбои (насильственные хищения).
* Количество угроз незначительно больше количества нападений.
* Прочие правонарушения занимают чуть более 5% от всех преступлений.
* Таким образом, по данным опроса, мы можем сделать вывод, что преступления, подразумевающие использование физической силы, совершаются реже, чем преступления без насилия.

Рис. 6

Отношение видов правонарушений

Затем было проанализировано влияние возраста на подверженность преступлениям. Для этого было определено процентное соотношение каждой возрастной категории в выборке (рис. 7). Исходя из полученных в результате анализа соотношений, были сделаны следующие выводы:

* Наименее подвержены преступлениям возрастные категории "18-24" и "65+".
* Наиболее подвержены преступлениям лица 25-34 лет, далее процент жертв убывает экспоненциально.

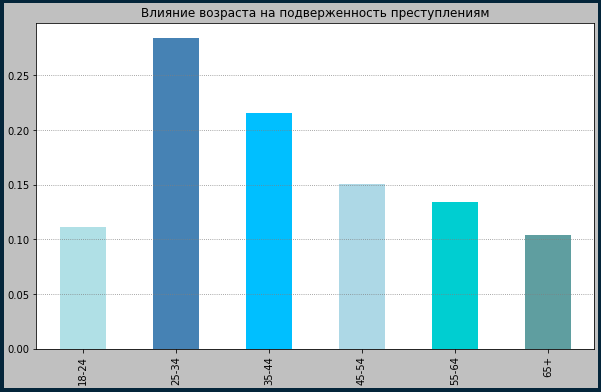


Рис. 7

Отношение возрастных категорий

Далее был произведен анализ зависимости вида преступления от возраста. Для этого были созданы выборки по каждой возрастной категории и посчитан процентное соотношение видов правонарушений в полученных выборках (рис. 8). Исходя из полученных в результате анализа соотношений, были сделаны следующие выводы:

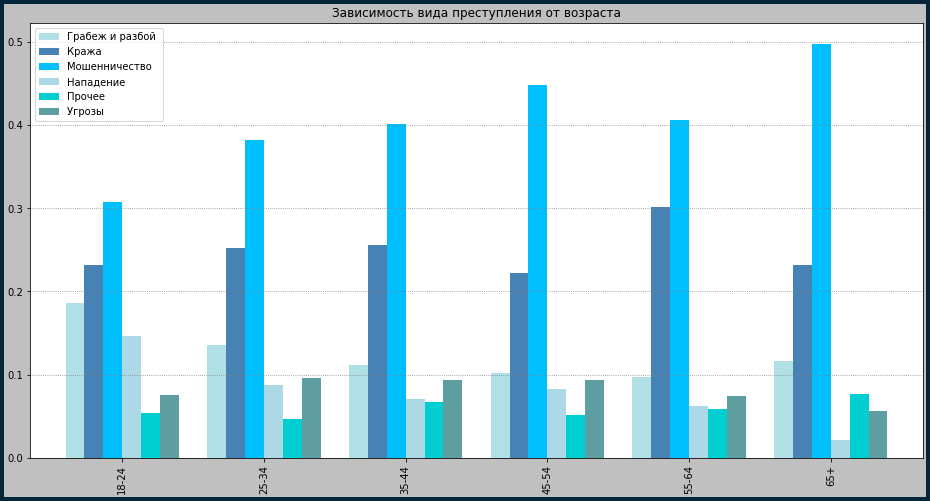
* Пожилые люди чаще других становятся жертвами мошенничества, в то время, как молодые – реже всего. Вероятно, это связанно с совершенствованием и цифровизацией методов мошенничества, что подвергает лиц, не обладающих достаточной цифровой грамотностью, наибольшим рискам стать жертвой мошенников.
* Молодые люди чаще остальных становятся жертвами грабежей и разбоев, далее процент плавно уменьшается и слегка возрастает у лиц возраста "65+".
* Наиболее подвержены кражам лица категории "55-64", остальные категории по процентному отношению отличаются незначительно.
* Наименее подвержены угрозам возрастные группы "18-24" и "65+", остальные категории по процентному отношению отличаются незначительно.
* Жертвами краж чаще всего становятся люди возрастной категории "55-64". Возможно, это связанно с тем, что люди к данному периоду жизни уже материально обеспеченны, чем вызывают интерес преступников.

Рис. 8

Отношение видов преступлений среди возрастных категорий

После этого было изучено влияние дохода жертвы на ее подверженность преступлениям. Для этого было найдено процентное отношение каждого уровня дохода среди пострадавших и посчитаны их процентные доли, по полученным значениям была построена гистрограмма (рис. 9). На основании этого были сделаны следующие выводы:

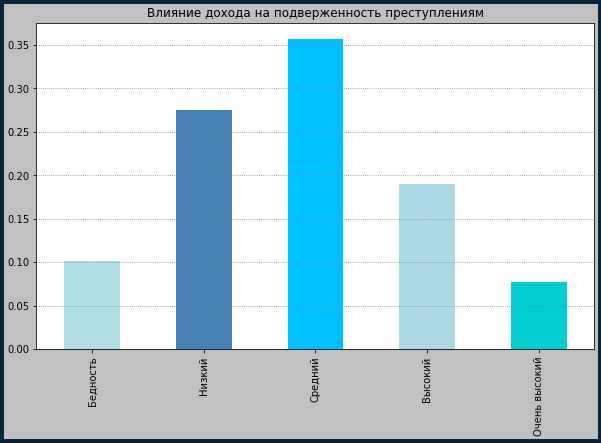
* Распределение процентного отношения подверженности преступлениям среди разных уровней дохода близко к нормальному.
* Логично предположить, что бедные менее подвержены преступлениям из-за отсутствия материальных средств и имущества, а люди с очень высоким уровнем дохода – в силу того, что их уровень достатка позволяет позаботиться о безопасности своих ценностей.

Рис. 9

Отношение категорий дохода среди жертв преступлений

Затем была определена зависимость между уровнем дохода и типом преступления. Для этого были взяты выборки по каждому уровню дохода и среди них были найдены процентные соотношения видов преступлений (рис. 10). После проведенного анализа были сделаны выводы:

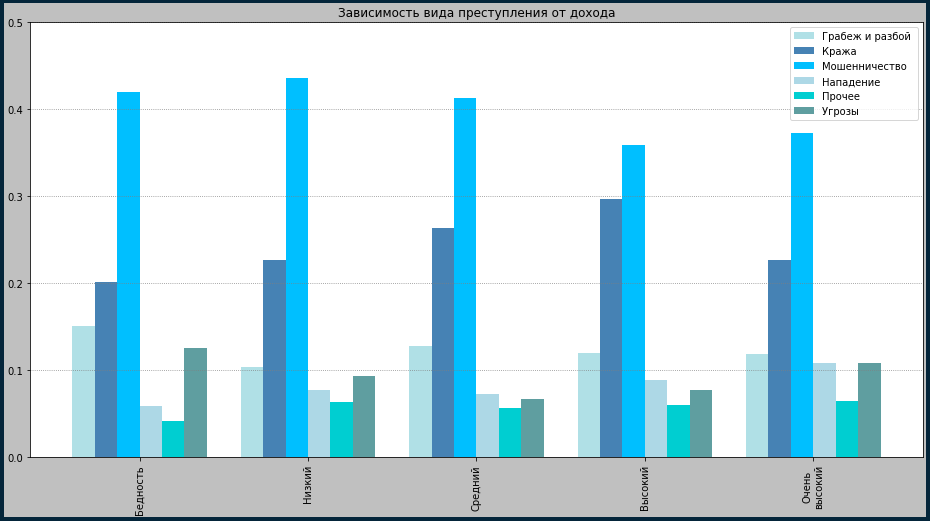
* Среди всех уровней дохода наиболее распространённым видом преступления является мошенничество. Наиболее подвержены люди с низкими и очень низкими доходами.
* Бедные больше других подвержены грабежам и разбоям, угрозам, но меньше всех – нападениям и прочим видам преступлений.
* Люди с очень высоким уровнем дохода часто становятся жертвами угроз, нападений, грабежей и разбоев, что связанно с повышенной заинтересованностью преступников в присваивании чужих материальных ценностей.
* Высокий класс больше других подвержен кражам.

Рис. 10

Отношение видов преступлений в каждой категории дохода

Далее был проведен анализ влияния уровня образования на подверженность преступлениям. Для этого были найдено процентное отношение каждого вида образования среди пострадавших (рис. 11). Основываясь на полученные значения, были сделаны выводы:

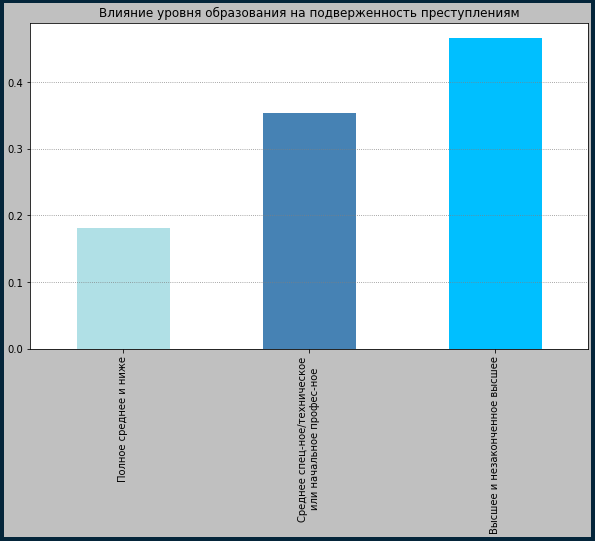
* Почти половина пострадавших респондентов имеет высшее или незаконченное высшее образование.
* Чуть больше трети пострадавших респондентов имеет среднее специальное/техническое или начальное профессиональное образование.
* Менее четверти пострадавших респондентов имеет полное среднее образование и ниже.
* Данное распределение объясняется тем, что люди категории 18-24 имеют более низкий уровень образования и менее подвержены преступлениям, чем люди более старших возрастов, которые соответственно имеют более высокий уровень образования.

Рис. 11

Отношение уровней образования среди жертв преступлений

Затем была определена зависимость между уровнем образования и каждым видом преступления. Для этого были созданы выборки по каждому уровню образования и среди них было найдено процентное соотношение видов преступлений (рис. 12). В результате были сделаны следующие выводы:

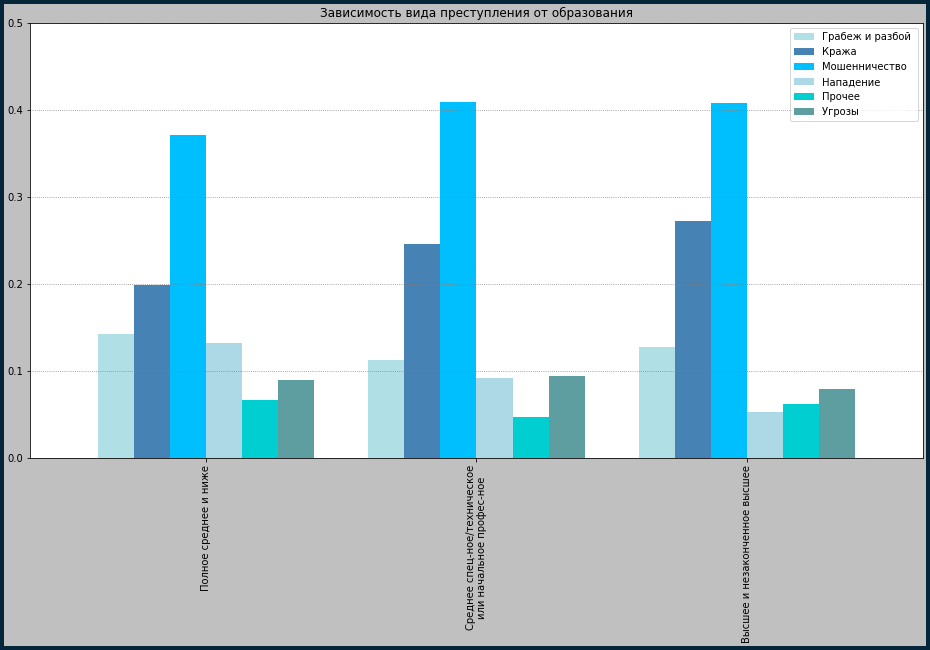
* По мере роста уровня образования респондентов растет уровень краж, что, возможно, связано с тем, что более высокообразованные люди имеют более высокооплачиваемую работу и, следовательно, обладают ценными материальными благами, что делает их более вероятными жертвами покушений на кражи.
* По мере роста уровня образования респондентов падает количество совершаемых в их сторону нападений.
* Уровни остальных преступлений не имеют прямых зависимостей от уровня образования опрошенного.

Рис. 12

Отношение видов преступлений среди разных уровней образования

После этого был проведен анализ влияния социально-экономического статуса (СЭС) на виктимность человека. Для этого было посчитано процентное содержание каждой группы в выборке (рис. 13). Впоследствии были сделаны выводы:

* Студенты меньше всего подвергаются преступлениям.
* Служащие, безработные, руководители и рабочие примерно в равной степени подвержены преступлениям.
* Пенсионеры наиболее подвержены преступлениям. Вероятно, это связано с тем, что пенсионеры из-за своего возраста зачастую становятся легкой добычей для мошенников и воров.

Затем была определена зависимость между социально-экономическим статусом жертв и типом преступления. Для этого были созданы выборки по каждому СЭС и среди них найдены процентные соотношения видов преступлений (рис. 14). Исходя из этого, были сделаны следующие выводы:

* Угрозам чаще всего подвергаются служащие и безработные.
* Кражам немногим чаще подвержены пенсионеры и служащие.
* Пенсионеры чаще всего становятся жертвами мошенников и прочих преступлений, не указанных в списке, что подтверждает вывод из анализа влияния возраста на подверженность преступлениям категории "65+", в которую входят большинство пенсионеров.
* Студенты больше других подвергаются грабежам, разбоям и нападениям, меньше других – мошенничеству, что тоже подтверждает вывод из анализа влияния возраста на подверженность преступлениям категории "18-24", к которой можно причислить большинство студентов.
* Пенсионеры реже других подвергаются нападениям и грабежам.

Рис. 14

Отношение видов преступлений среди разных социально-экономических статусов

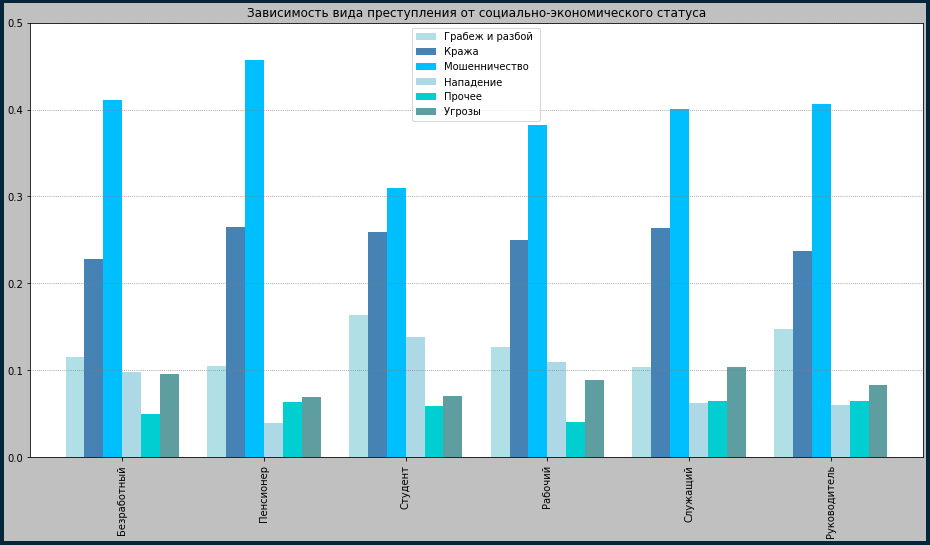
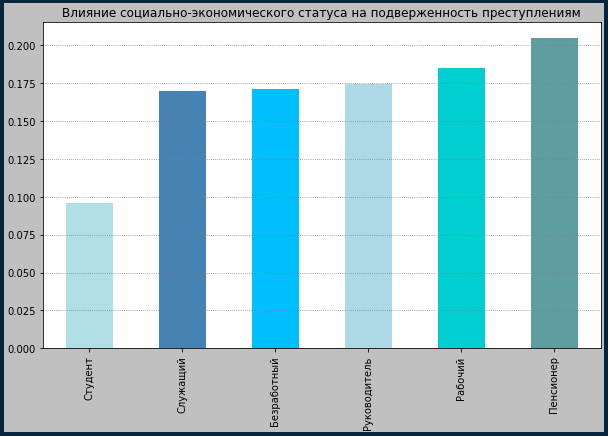


Рис. 13

Отношение социально-экономических статусов в выборке жертв



Далее была изучена скорость реагирования правоохранителей. Для этого было посчитано отношение каждой категории скорости реакции в случаях, когда совершался вызов полиции (рис. 15). Исходя из полученных значений, были сделаны выводы:

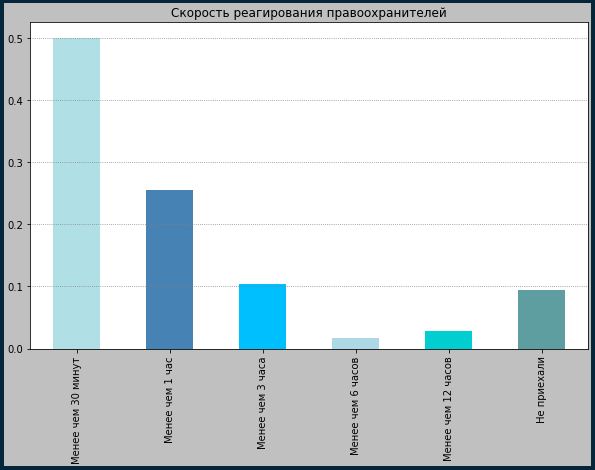
* В половине случаев правоохранители реагируют на вызов с должной скоростью (меньше 30 минут).
* В четверти случаев полицейские приезжают менее, чем за час.
* В 10% случаев полицейские приезжают в течение трех часов.
* Почти в каждом десятом случае полицейские не приезжают по вызову.

Рис. 15

График скоростей реагирования правоохранителей на вызов

Затем был проведен анализ материального ущерба жертв, которые понесли жертвы (если таковой был). Для этого была создана выборка, содержащая материальный ущерб, найдены среднее и медианное значения. Однако, после их получения (3116267 руб. и 12000 руб. соответственно) было замечено, что данные значения сильно различаются, что происходит из-за выбросов в данных. Для того, чтобы их исключить, был определен вид распределения данных – экспоненциальный (рис. 16). Для определения выбросов был использован межквартильный размах. Было найдено расстояние (IQR) между третьим (Q3) и первым (Q1) квартилями, затем посчитана верхняя и нижняя граница для выбросов. Для этого расстояние между квартилями (IQR) было умножено на 1,5, и вычтено из третьего (Q3) и первого (Q1) квартилей соответственно. После исключения выбросов была построена диаграмма размаха (рис. 17), демонстрирующая, что большинство выбросов были удалены из выборки, и посчитаны медианное и среднее значения – 9000 руб. и 13243 руб. соответственно. Исходя из этого, следует, что:

* Существуют случаи, когда ущерб в результате преступления значительно превышает ущерб в большинстве других случаев. Из-за значительного влияния на среднее и медианное значение, данные случаи были исключены.
* Медианный материальный ущерб после исключения выбросов данных, полученный пострадавшими в результате преступления, составляет 9 000 рублей.
* Средний материальный ущерб после исключения выбросов данных, полученный пострадавшими в результате преступления, составляет 13 243 рублей.

Рис. 17

Гистограмма размаха материального ущерба после удаления выбросов

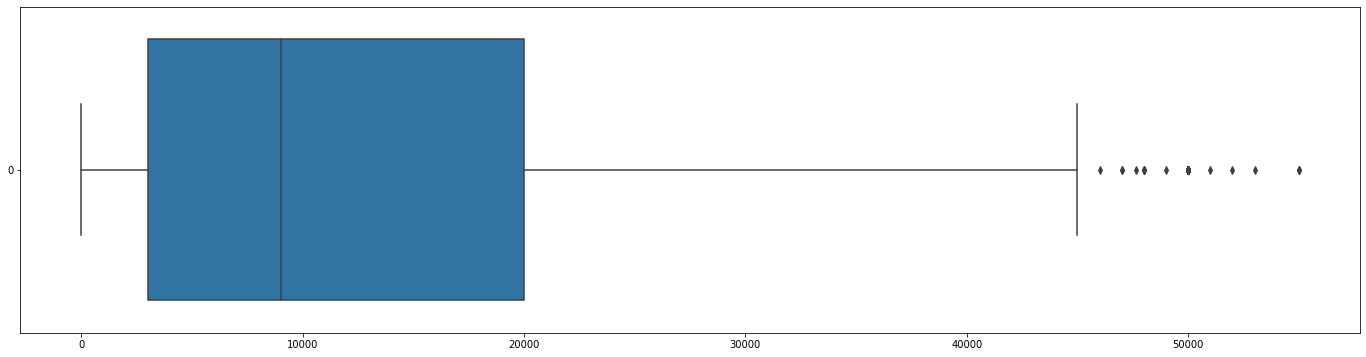
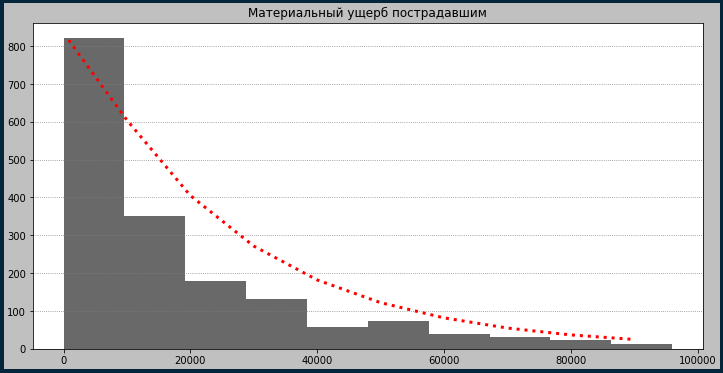


Рис. 16

График распределения материального ущерба пострадавших



После этого был проведен анализ зависимости вида преступлений от времени суток, в которое оно было совершено. Для этого вначале была определена процентная зависимость количества совершаемых преступлений от времени суток (рис. 18). Исходя из этого был сделан вывод, что большая часть преступлений совершается в дневное время суток.



Рис. 18

Процентное отношение кол-ва преступлений в разное время суток

Затем была найдена зависимость между видами преступлений и временем суток. Для этого были созданы выборки по каждому виду преступлений и найдено процентное соотношение преступлений ночью и днем (рис. 19). Основываясь на полученные значения, были сделаны следующие выводы:

* Наиболее тяжкие преступления, такие как нападения, угрозы, грабежи и разбои, кражи, совершаются в большинстве случаев в тёмное время суток.
* Преступления, связанные с мошенничеством и удаленным мошенничеством, чаще всего совершаются днём.
* Таким образом, несмотря на то, что большинство преступлений осуществляется днём, наиболее тяжкие из них осуществляются ночью. Процент дневных преступлений больше, поскольку значительная часть от них занимает мошенничество, осуществляемое в большинстве случаев в светлое время суток.

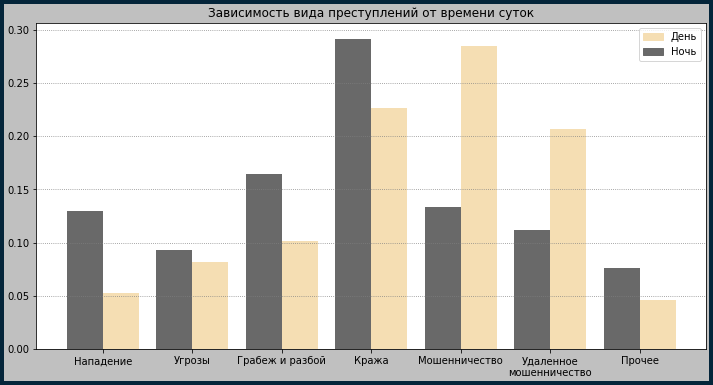
После осуществленного анализа выборки лиц, являющихся жертвами преступлений, исходя из найденных зависимостей виктимности человека от его социально-демографических характеристик, был составлен портрет человека, который с наибольшей долей вероятности будет являться жертвой преступления:

Рис. 19

Соотношение видов преступлений в разное время суток

* Пол: мужчина
* Возрастная группа: 25-34 года
* Уровень дохода: средний
* Уровень образования: высшее или незаконченное высшее
* Социально-экономический статус: рабочий

Далее был проведен анализ выборки лиц, не являющихся жертвами преступлений.

Для начала был произведен анализ по половому признаку. Для этого были найдены процентное соотношение мужчин и женщин в выборке (рис. 20). Исходя из полученных в результате анализа соотношений, был сделан следующий вывод:

* Среди лиц, не являющихся жертвами преступлений, женщин немного больше, чем мужчин.

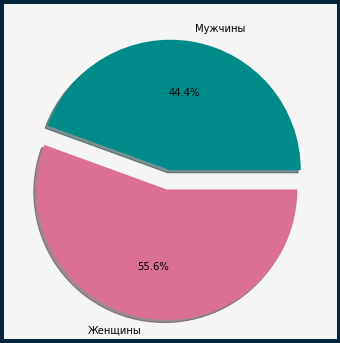


Рис. 20

Соотношение мужчин и женщин

Далее был произведен анализ возрастных категорий. Для этого были найдены процентные соотношения возрастных категорий в выборке (рис. 21). Исходя из полученных в результате анализа соотношений, были сделаны следующие выводы:

* Меньшую часть данной выборки составляют лица категории "18-24"
* Большую часть выборки составляют лица категории "25-34", далее процент каждой группы убывает.

Затем был произведен анализ дохода. Для этого были найдены процентные соотношения категорий уровня дохода в выборке (рис. 22). Исходя из полученных в результате анализа соотношений, были сделаны следующие выводы:

* Распределение разных уровней дохода близко к нормальному.

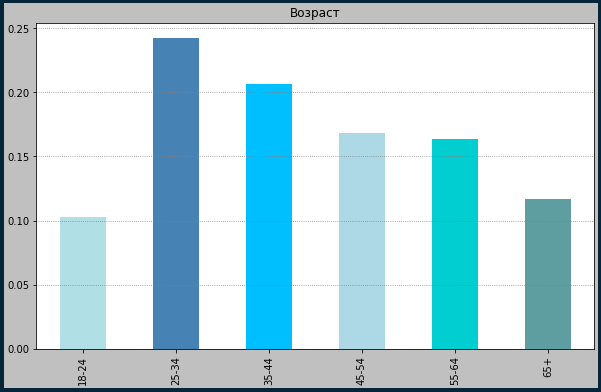


Рис. 21

Отношение возрастных категорий

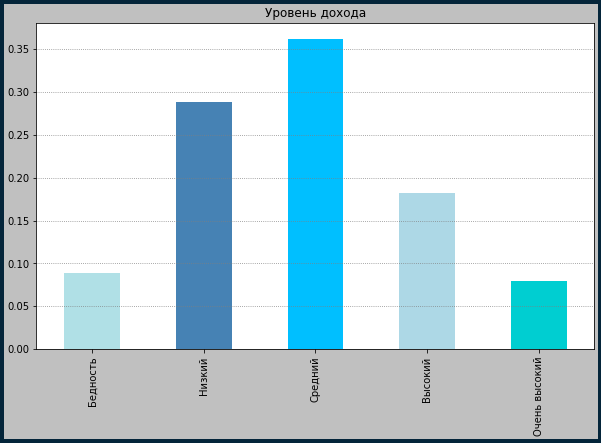


Рис. 22

Отношений уровней дохода

Далее был осуществлен анализ уровня образования. Для этого были найдены процентные соотношения различных уровней образования в выборке. (рис. 23). Исходя из полученных в результате анализа соотношений, были сделаны следующие выводы:

* Среди лиц, не подвергавшихся преступлениям, меньше всего лиц с полным средним и более низким уровнем образования.
* Людей со средним специальным/техническим или начальным профессиональным образованием чуть больше, чем людей с высшим и незаконченным высшим образованием.

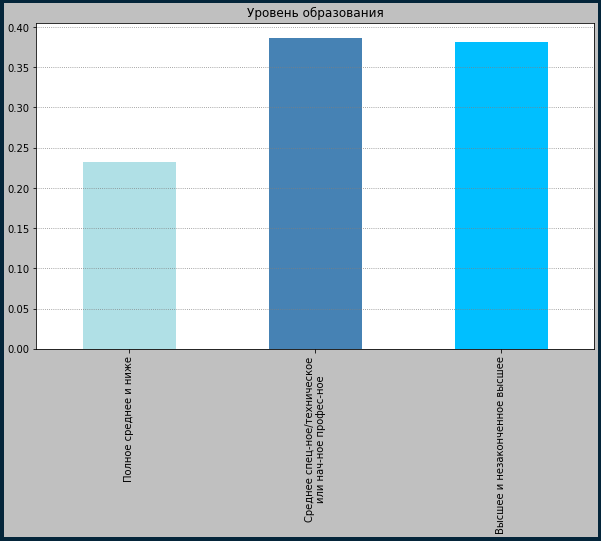


Рис. 23

Отношение уровней образования

Затем был произведен анализ социально-экономического статуса. Для этого были найдены процентные соотношения каждого социально-экономического статуса в выборке (рис. 24). Исходя из полученных в результате анализа соотношений, были сделаны следующие выводы:

* Среди лиц, которые не подвергались преступлениям, меньше всего студентов, далее процентная составляющая каждой группы увеличивается.

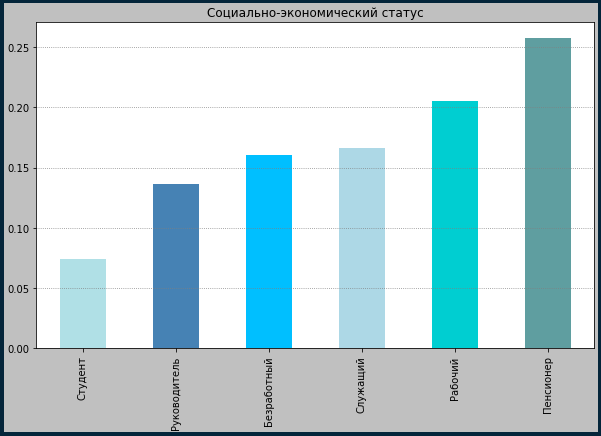


Рис. 24

Отношение социально-экономических статусов

Таким образом, исходя из анализа социально-демографических характеристик жертв преступлений, составим портрет человека, который с наибольшей вероятностью не является жертвой преступления:

* Пол: женщина
* Возрастная группа: 25-34 года
* Уровень дохода: средний
* Уровень образования: среднее специальное\техническое или начальное профессиональное
* Социально-экономический статус: рабочий

Сравним выборки лиц, являющихся жертвами преступлений и лиц, не являющихся жертвами преступлений.

Для начала произведем сравнение по половому признаку, для этого используем найденные ранее процентные отношения (рис. 25). Исходя из сравнения полученных в результате анализа соотношений были сделаны следующие выводы:

* Среди жертв преступлений и лиц, не являющихся жертвами преступлений проценты мужчин и женщин приблизительно равны.
* Так как процент женщин больше в обоих выборках, мы можем сделать вывод, что среди респондентов опроса женщин больше, чем мужчин.

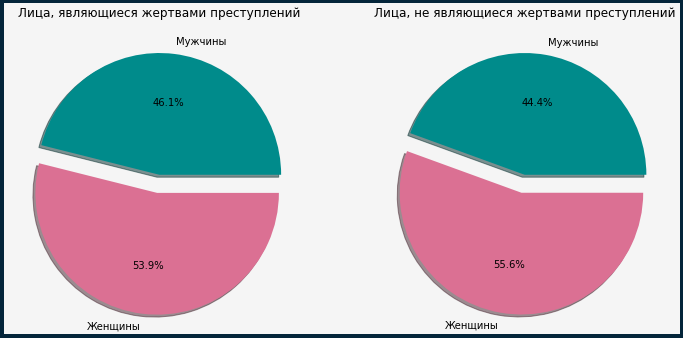


Рис. 25

Сравнение отношений полового признака среди лиц, являющихся и не являющихся жертвами преступлений

Затем сравним выборки по возрастному признаку, для этого используем найденные ранее процентные отношения (рис. 26). Исходя из сравнения полученных в результате анализа соотношений, были сделаны следующие выводы:

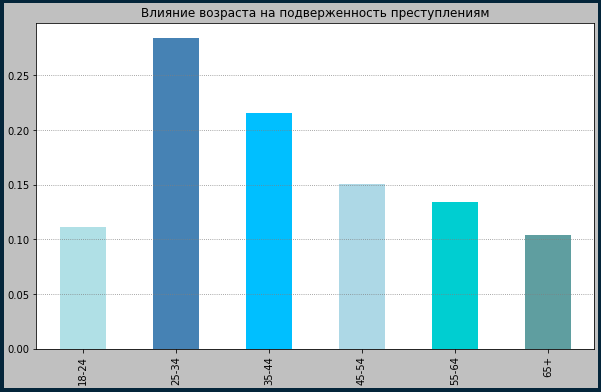
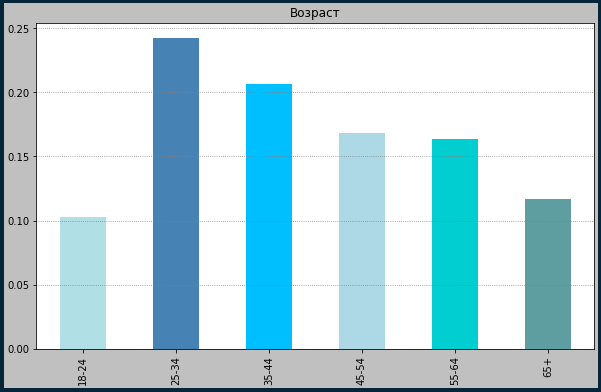
* Обе выборки имеют схожее распределение.
* В выборке жертв преступлений значительно больше процент лиц категории "25-34".
* В выборке людей, не являющихся жертвами преступлений, больше процент лиц категорий "45-54", "55-64" и "65+".

Рис. 26

Сравнений отношений возрастных категорий среди лиц, являющихся и не являющихся жертвами преступлений

Далее произведем сравнение по уровню дохода, для этого используем найденные ранее процентные отношения (рис. 27). Исходя из сравнения полученных в результате анализа соотношений, были сделаны следующие выводы:

* Процентное содержание каждой категории уровня дохода примерно схоже в обоих выборках.
* Так как распределение в обоих выборках одинаково, мы можем сделать вывод что в генеральной совокупности распределение уровней дохода является нормальным.

Затем произведем сравнение по уровню образования, для этого используем найденные ранее процентные отношения (рис. 28). Исходя из сравнения полученных в результате анализа соотношений, были сделаны следующие выводы:

* В выборке лиц, ставших жертвами преступлений, значительно меньше лиц с образованием уровней "полное среднее и ниже" и "среднее специальное/техническое или начальное профессиональное".
* В выборке лиц, не являющихся жертвами преступлений, значительно меньше людей с высшим и незаконченным высшим образованием.

После этого сравним выборки по социально-экономическому статусу, для этого используем найденные ранее процентные отношения (рис. 29). Исходя из сравнения полученных в результате анализа соотношений, были сделаны следующие выводы:

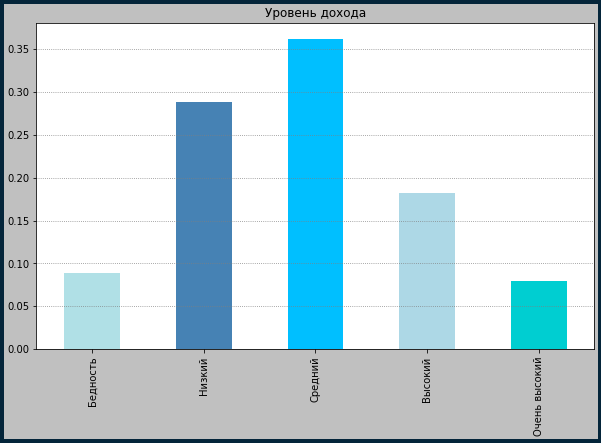
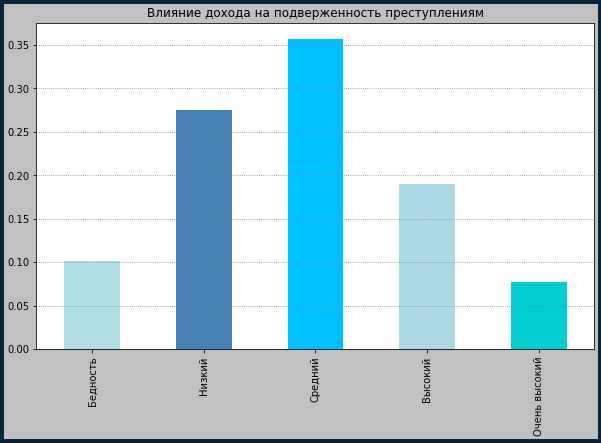
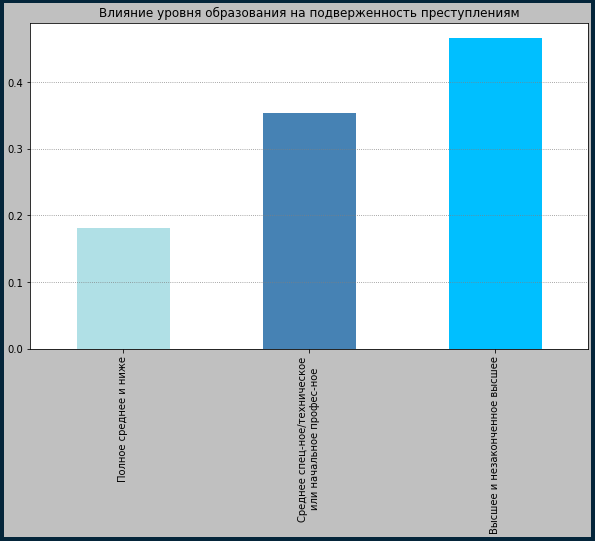
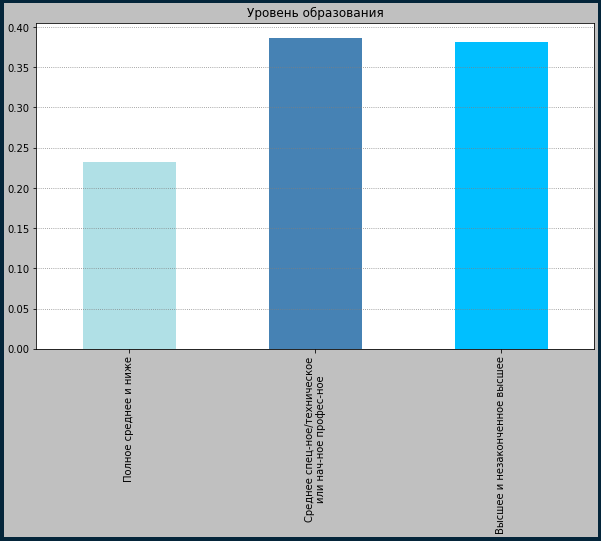
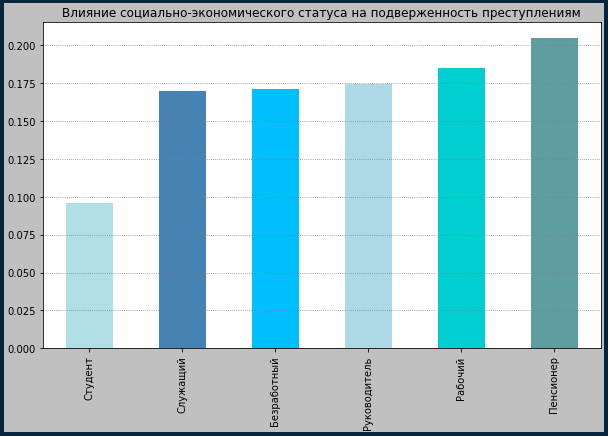
* В выборке жертв преступлений процентное содержание студентов меньше, чем в выборке лиц, которые не становились жертвами преступлений.
* В выборке жертв преступлений процентное содержание служащих, безработных, руководителей и рабочих примерно одинаково и больше, чем в другой выборке.
* В выборке людей, не являющихся жертвами преступлений, значительно больше пенсионеров.

Рис. 27

Сравнение отношений уровней дохода среди лиц, являющихся и не являющихся жертвами преступлений

Рис. 28

Сравнение отношений уровней образования среди лиц, являющихся и не являющихся жертвами преступлений



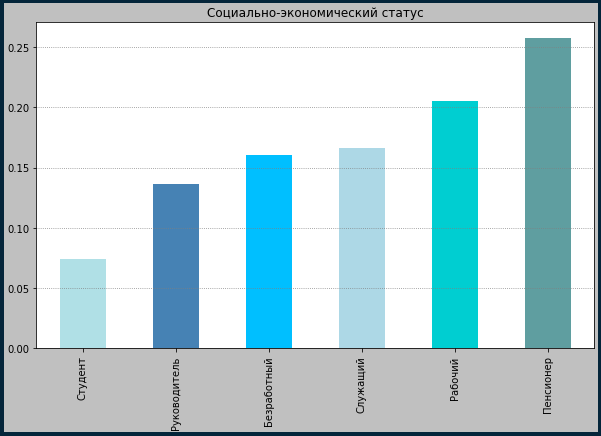


Рис. 29

Сравнение отношений социально-экономических статусов среди лиц, являющихся и не являющихся жертвами преступлений

Таким образом мы можем выделить признаки, которые оказывают наибольшее влияния на вероятность стать жертвой преступления:

* Пол
* Образование
* Социально-экономический статус

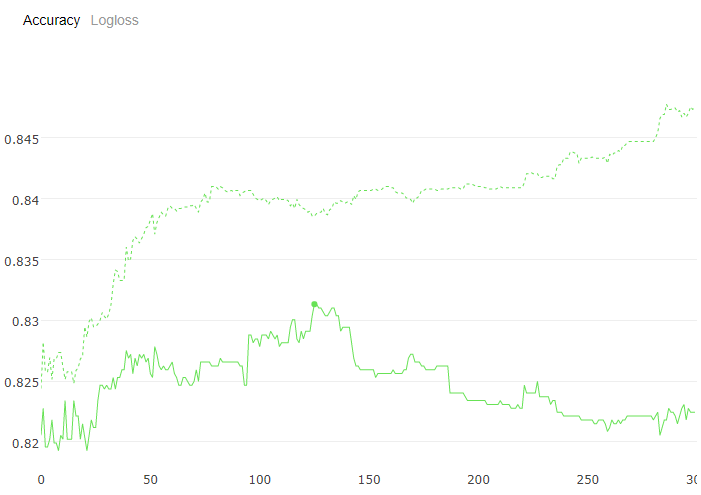
В заключении нашего анализа построим классификационную модель, способную предсказать по параметрам индивидуума, вероятность оказаться жертвой преступления. Исходя из полученных нами данных в результате исследования влияния социально-демографических характеристик человека на его подверженность преступлению, было принято решение построить модель на основании пяти категориальных признаков:

* Пол
* Возраст
* Уровень образования
* Уровень дохода
* Социально-экономический статус

Для этого была выбрана библиотека **CatBoost** – открытая программная библиотека, разработанная компанией **Яндекс** и реализующая уникальный патентованный алгоритм построения моделей машинного обучения, использующий одну из оригинальных схем градиентного бустинга, включающая в себя классификационную модель **CatBoostClassifier**.

Сначала была создана выборка из генеральной совокупности, включающая в себя необходимые столбцы. Затем полученная выборка была разделена на тренировочную, валидационную и тестовую выборки.

Далее модель была создана и обучена на тренировочной выборке, по мере обучения были построены графики, показывающие точность (рис. 30) и логическую ошибку (рис. 31) модели на каждом из этапов обучения.



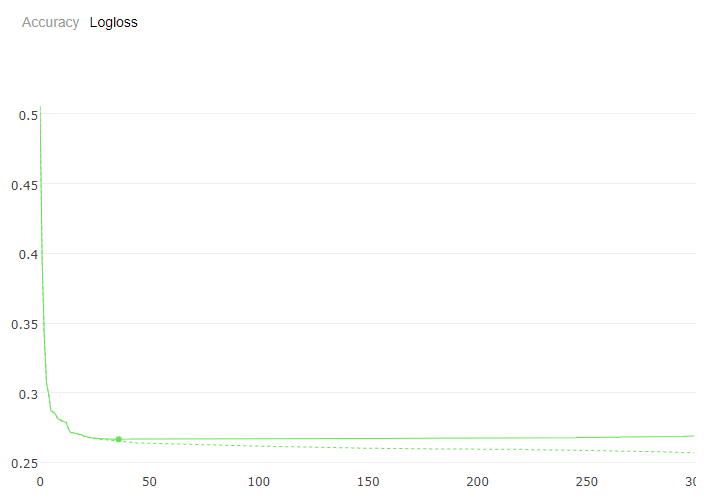


Рис. 30, 31

График точность модели по мере обучения

График логической ошибки модели по мере обучения

Получив модели, мы приступили к оценке её точности, для этого была проведена кросс-валидация — метод оценки аналитической модели и её поведения на независимых данных. При оценке модели имеющиеся в наличии данные разбиваются на k частей. Затем на k-1 частях данных производится обучение модели, а оставшаяся часть данных используется для тестирования. Процедура повторяется k раз; в итоге каждая из k частей данных используется для тестирования. В результате получается оценка эффективности выбранной модели с наиболее равномерным использованием имеющихся данных. В итоге, мы получим график точности модели по мере её применения к тренировочной выборке (рис. 32).

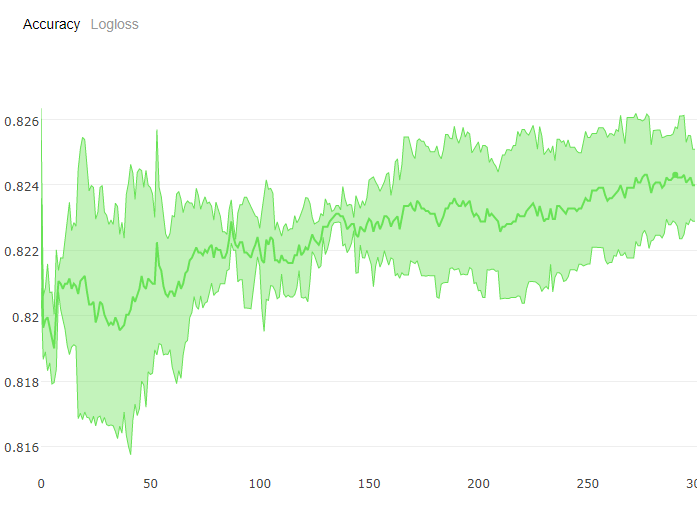


Рис. 32

График кросс-валидации модели

Теперь определим истинное значение точности модели, для этого применим её к валидационной выборке и построим график точности (рис. 33). Для определения нашего истинной точности необходимо взять среднее от всех значений. В итоге истинное значение точности составляет 82,431%.

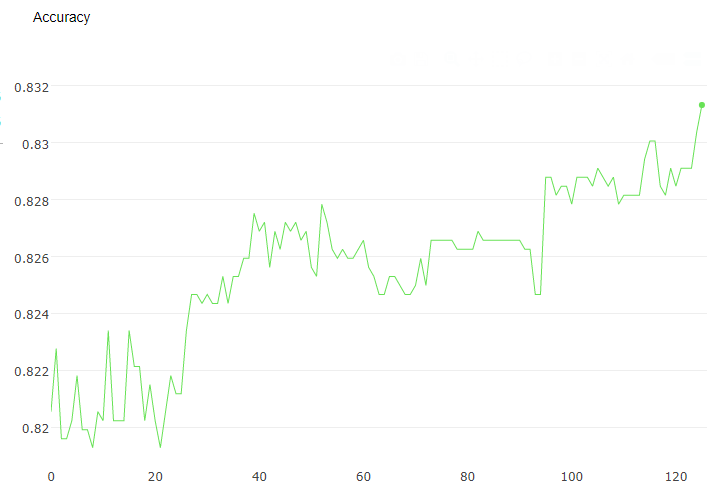


Рис. 33

График точности по мере применения модели к валидационной выборке

Продолжим тестирования нашей модели, посмотрим наглядно на формат работы нашей модели. Для этого применим её к случайно-сгенерированным данным. В результате получим таблицу, содержащую характеристики лица, предсказание модели и вероятность каждого предсказания (рис. 34).



Рис. 34

Таблица, полученная в результате применения модели к случайным данным

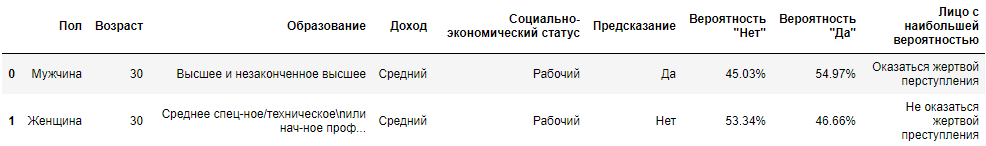
Для того чтобы окончательно убедиться в исправности классификационной модели применим её к ранее составленным портретам лиц, с наибольшей и наименьшей вероятностью стать жертвой преступления. В итоге получим таблицу, аналогичную предыдущей (Рис. 35). **Как мы видим, наша модель сделала правильное предсказание в обоих случаях, что ещё раз доказывает её точность.**

Рис. 35

Таблица, полученная в результате применения модели к портретам лиц, с наибольшей и наименьшей вероятностью стать жертвой преступления

Таким образом, была построена модель, способная предсказать, окажется ли человек жертвой преступления, с точностью 82,431%

В результате проделанной работы были выполнены задачи и достигнута поставленная цель, а именно:

* **Была выполнена предварительная обработка данных.**
* **Определены и визуализированы закономерности и корреляции признаков.**
* **Найдены параметры, по которым можно определить группу людей, наиболее подверженных преступлениям.**
* **Найдены параметры, по которым можно определить группу людей, наименее подверженных преступлениям.**
* **Были определены признаки, которые оказывают наибольшее влияние на вероятность оказаться жертвой преступления.**
* **Построена классификационная модель, с помощью которой можно предсказать вероятность оказаться жертвой преступления для определенного лица.**
* Опираясь на сделанные в ходе работы выводы, гипотеза исследования была доказана.

**Список литературы**

1. Уэс Маккини. Python и анализ данных. ДМК-Пресс, 2020 г. ISBN: 978-5-97060-590-5.
2. Владимир Савельев. Статистика и котики. АСТ, 2018 г. ISBN: 978-5-17-108287-1
3. Donald Michie. Machine Learning, Neural and Statistical Classification. 1995 г. ISBN:978-0-13-106360-0
4. Себастьян Рашка. Python и машинное обучение. ДМК-Пресс, 2017 г. ISBN: 978-5-97060-409-0