ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ «ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Высшая школа бизнеса

РАЗРАБОТКА НОВОГО ПРОДУКТА В СФЕРЕ СТРАХОВАНИЯ ИМУЩЕСТВА С ИСПОЛЬЗОВАНИЕМ ПРИНЦИПОВ ПОВЕДЕНЧЕСКОЙ ЭКОНОМИКИ НА ОСНОВЕ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Курсовая работа студента

Лухнева Игоря Дмитриевича

2 курс, направление подготовки: 38.03.05 «Бизнес-информатика»

образовательная программа «Бизнес-информатика»

Научный руководитель

канд. техн. наук, доцент

С. Г. Ефремов

Москва 2020

# СОДЕРЖАНИЕ

[СОДЕРЖАНИЕ 2](#_Toc74216224)

[ВВЕДЕНИЕ 3](#_Toc74216225)

[РАЗРАБОТКА ПРОДУКТА В СФЕРЕ СТРАХОВАНИЯ ТРАНСПОРТНЫХ СРЕДСТВ 6](#_Toc74216226)

[ТЕОРЕТИЧЕСКАЯ БАЗА ИССЛЕДОВАНИЯ 6](#_Toc74216227)

[ПОСТАНОВКА ПРОБЛЕМ И ВЫБОР СПОСОБОВ ИХ РЕШЕНИЯ 11](#_Toc74216228)

[РАЗРАБОТКА И ТЕСТИРОВАНИЕ ПРОДУКТА 14](#_Toc74216229)

[ЗАКЛЮЧЕНИЕ 31](#_Toc74216230)

[СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ 33](#_Toc74216231)

[ПРИЛОЖЕНИЕ 1. ИСХОДНЫЙ КОД АЛГОРИТМА ПРЕДСКАЗАНИЯ РИСКА 35](#_Toc74216232)

[ПРИЛОЖЕНИЕ 2. JUPYTER NOTEBOOKS 41](#_Toc74216233)

# ВВЕДЕНИЕ

Страхование как механизм является важным инструментов в экономике, оно позволяет домохозяйствам и фирмам защитить своё финансовое благополучие на случай возникновения незапланированных затрат, связанных с имущественным ущербом. Кроме того, тот же механизм работает на финансовом рынке, позволяя инвесторам уменьшить риск. Страхование очень важно, как способ управления рисками, так как даёт возможность переложить риск на других экономических агентов. В отношении имущественных рисков, которые обусловлены высокой степенью неопределенности вследствие стохастичности процессов, которые могут причинить ущерб имуществу, страхование предполагает в сущности усреднение ущерба между всеми застрахованными лицами, так как за счёт премий, которые оплачивает каждое страхуемое лицо, в случае наступления страхового случая, пострадавшему выплачивается страховая сумма, компенсирующая его потери. Во многих ситуациях страхование имущества является рациональным действием, которое позволяет защититься от банкротства. В особенности это касается страхования транспортных средств, учитывая то, что любые дорожно-транспортные происшествия являются во многом событиями, зависящими от поведения всех участников дорожного движения, качества дорожно-транспортной инфраструктуры и погодных условий, а следовательно имеется высокая степень неопределенности в отношении наступления неблагоприятного события.

В сфере страхования транспортных средств существует большая доля участия государства, которое обязует участников дорожного движения страховать определенную часть рисков, чтобы избежать ситуаций, когда виновник происшествия не может оплатить причинённый им вред имуществу пострадавшему. Такое участие государства позволяет сделать процесс страхования транспортных средств рациональнее. Однако существует множество других аспектов в страховании, которые могут препятствовать экономически-обоснованному рациональному поведению экономических агентов: от когнитивных искажений индивидов и неполноты информации до нерыночных моделей поведения фирм.

Именно наличие в системе страхования транспортных средств элементов поведения, не соотносящегося с поведением рационального экономического агента, является проблемой, рассматриваемой в данной работе. Таким образом, данная работа предполагает разработку продукта, который мог бы оптимизировать и рационализировать процесс страхования транспортных средств.

Основные проблемы в системе страхования, которые снижают его эффективность и приводят к нерациональным решениям экономических агентов заключаются именно в поведенческих отклонениях, что неоднократно исследовалось экономистами. Ошибки делятся на ошибки со стороны регулятора, спроса и предложения (Кунройтер, Полли, Макморроу, 2013 [1]). В данной работе внимание будет уделяться именно ошибкам со стороны спроса и предложения, так как, по мнению экономистов, их решение возможно посредством создания продукта с использованием определённых подходов, позволяющих исключить данные ошибки.

В поведенческой экономике существует понятие «наджинг» (в русскоязычной литературе чаще называется «теория подталкивания»), введенный А. Тверски и Д. Канеманом, который подразумевает использование психологических манипуляций для того, чтобы индивид принимал более рациональные решения. Именно с применением «теории подталкивания» в США удалось увеличить число граждан, имеющих пенсионные сбережения, путём включения в трудовой договор «по умолчанию» пункта о ежемесячных отчислениях в пенсионный фонд с заработной платы. Применение «теории подталкивания» и других поведенческих методов повышения эффективности системы страхования транспортных средств предполагается в данной работе во время разработки нового продукта.

Актуальность данной темы заключается в том, что новый продукт позволит сократить число махинаций со автострахованием, которое является обязательным для управления транспортным средством во многих регионах и государствах, за счет того, что индивид предпочтёт оформить страховку, осознавая рациональность и обоснованность данного решения. Более того, позволит применять более справедливые тарифные планы и обеспечит финансовую устойчивость страховых компаний, а также эффективная система страхования поможет сократить долю участия в данном процессе государственных органов и высвободить ресурсы на инвестиции и развитие других сфер экономики. Практические решения в рамках работы могут быть интегрированы в реальные бизнес-процессы и применяться действующими игроками на рынке.

Тема, рассматриваемая в данной работе, как уже было отмечено ранее, неоднократно исследовалась, а теоретические результаты некоторых научных работ будут использованы в процессе разработки продукта. Так, например, в работе А. Гринберга [2] и книге Г. Кунройтера, М. Полли, С. Макморроу [1], которые станут теоретической базой данной работы, отмечается, что человеку свойственно игнорировать события, вероятности которых находятся в достаточно малой правой окрестности 0. Социологическое исследование Р. Таллера [3] демонстрирует, что индивид воспринимает страхование как инвестиции, что приводит к тому, что он ожидает получить с них доход. А аналитика Accenture [4] показывает, что большинство клиентов страховых компаний испытывают трудности с изучением договора и выбором подходящих опций, из-за чего прибегают к услугам страховых агентов, что создаёт дополнительные трансакционные издержки.

Согласно другому исследованию Accenture [5], многие страховые компании уже используют методы машинного обучения для оценки вероятности наступления страхового случая и суммы ущерба. Для тех же целей алгоритмы машинного обучения будут использованы и в данной работе. Кроме того, методы анализа данных предполагается применить для расчета рисков клиента попасть в аварию на протяжении своего ежедневного маршрута.

Цель данной работы заключается в разработке инструментов, основанных на методах интеллектуального анализа данных, для страховых компаний, помогающих избежать неэффективность в процессе страхования транспортных средств, вызванных поведенческими факторами экономических агентов. Из цели вытекают задачи исследования, которые определяют структуру работы:

* Исследование существующей теоретической базы относительно поведенческой экономики страхования и путей решений проблем с экономической эффективностью страхования;
* Выявление основных проблем, которые будут решены в процессе разработки продукта;
* Выбор способов реализации предложенных в теоретических материалах проблем;
* Реализация решений проблем и проектирование продукта;
* Тестирование готового решения;
* Исследование перспектив применения разработанных решений и способов дальнейшего развития продукта.

Согласно перечисленным выше задачам будет выстроена структура работы, которая будет включать следующие разделы: введение, основную часть из 3 подразделов, заключение.

# РАЗРАБОТКА ПРОДУКТА В СФЕРЕ СТРАХОВАНИЯ ТРАНСПОРТНЫХ СРЕДСТВ

## ТЕОРЕТИЧЕСКАЯ БАЗА ИССЛЕДОВАНИЯ

Основным трудом по теме исследования страхования с точки зрения поведенческой экономики является книга Г. Кунройтера, М. Полли, С. Макморроу [2], именно она легла в основу разработки концепта нового продукта, который станет результатом данной работы. В упомянутой книге авторы приводят ситуации, когда поведение индивида, страховой компании или государственного регулятора не соответствует использованной для исследования экономической модели, определяющей формально рациональные действия. Такое поведение, называемое авторами аномалией, является объектом поведенческой экономики, изучающей и объясняющей его. Отличительной чертой поведенческой экономики по сравнению с другими направлениями экономической науки является фокус на том, что в реальности экономическое поведение и решения экономических агентов не соотносятся с ожидаемыми от них экономистами. Для выявления таких ситуация поведенческие экономисты используют стандартные модели. Так, в случае со страхованием авторы использовали модель спроса и предложения с учётом риска наступления страхового случая. Объяснения же несоответствия поведения лежит в области психологии и нейробиологии, как и возможные решения, способствующие корректировке поведения.

Авторы отмечают, что рынок страховых инструментов в США, несмотря на свои большие размеры, не отвечает важнейшему критерию стандартной экономической модели, так как во многом не является свободным добровольным выбором экономических агентов относительно покупки и продажи страховых инструментов. Это обусловлено тем, что государство и банки обязуют страховать имущество, чтобы получить некоторые услуги. Например, в большинстве штатов невозможно зарегистрировать автомобиль без наличия страхового полиса, и ни один банк не предоставит ипотечный кредит без страхования приобретаемого имущества. В целом так сложилось из-за первичного понимания, что индивид не всегда поступает рационально и верно оценивает риски, из-за чего могут возникнуть крупные убытки, которые нечем будет покрыть. В этом заключается корень проблемы – люди понимают, что не все и не всегда действуют рационально, но тем не менее, само понимание не ведёт к изучению данной проблемы и поиску её оптимального решения.

Почему и какие возникают аномалии в поведении экономических агентов в отношении страхования? Отвечая на данный вопрос, авторы классифицируют аномалии на несколько типов, как уже отмечалось ранее: аномалии со стороны спроса и аномалии предложения. Аномалии спроса также делятся на следующие:

* неадекватно низкий спрос при обоснованной страховой премии;
* высокий спрос при завышенной страховой премии;
* выбор необоснованной суммы или типа покрытия.

Аномалии предложения выделяются на следующие:

* покрытие не предлагается, когда экономически выгодно его предлагать;
* сумма покрытия отклоняется от приведенной стоимости страховых премий.

Далее имеет смысл рассмотреть каждую аномалию по-отдельности и выявить их причины.

Аномалия неадекватно низкого спроса при обоснованной страховой премии выражается в том, что менее 70% потенциальных клиентов покупают страховой инструмент, несмотря на то, что его рыночная цена является справедливой. Наиболее выраженной данная аномалия оказывается на рынках, где приведенная сумма страховых премий оказывается ниже, чем ожидаемые потери из-за государственного регулирования и субсидирования, но индивиды тем не менее не торопятся совершить сделку. Для рынка автомобильного страхования данная аномалия не является частой.

Аномалия высокого спроса при завышенной страховой премии возникает, когда дополнительные трансакционные и операционные издержки составляют более 40% страховой премии, из-за чего страхование оказывается переоцененным, однако спрос остается на высоком уровне. Чаще всего такие аномалии встречаются на рынке страхования жизни и здоровья при полетах или движении на наземном транспорте (междугородние автобусы, поезда дальнего следования и др.).

Аномалия выбора необоснованной суммы или типа покрытия описывает ситуации, когда соотношение спроса и предложения в целом укладываются в предсказываемые экономическими моделями, однако сумма или тип покрытия, выбираемые потребителями противоречат экономически обоснованному выбору. Эмпирические наблюдения показывают, что индивиды отказываются от повышения франшизы, доли страховой суммы, которую страховая компания не будет оплачивать, при страховании автомобиля или домовладений, несмотря на то, что сокращение страховых премий значительно превысит увеличение суммы, которую придётся потратить в случае наступления страхового случая (Калтер и Зекхаузер 2004 [6], Синдор, 2010 [7]).

Аномалия, связанная с тем, что покрытие страховыми компаниями не предоставляется несмотря на наличие спроса, стала известна после террористического акта 11 сентября 2001 года в Нью-Йорке, тогда многие страховые компании вовсе перестали страховать данный риск, а несколько оставшихся из-за высокого спроса смогли позволить себе сильно увеличить цену такой услуги.

Аномалия, когда сумма покрытия оказывается меньше или больше, чем приведенная стоимость страховых премий, иногда встречается на конкурентных рынках, где теоретически она не могла бы быть устойчивым явлением, которым является по факту. Чаще всего такая аномалия наблюдается на рынке медицинской страховки.

Таким образом, можем заключить, что аномалии в сфере страхования существуют и объяснения, почему они возникли, можно встретить в научно-популярных изданиях и средствах массовой информации, однако стоит отметить, что порой они не отражают реальности и предлагают не самые оптимальные способы их устранения. Важно указать на действительные причины возникновения аномалий, которые были выявлены и доказаны поведенческими экономистами.

Нельзя утверждать, что индивид, принимая решение относительно страхования, не руководствуется соображениями максимизации ожидаемой полезности; ошибки, которые он совершает, имеют другую природу, обоснованной комплексностью самой задачи (Либман и Зекхаузер, 2008 [8]). Прежде всего, проблема заключается в недостатке или неправильной интерпретации информации для индивида относительно вероятностей наступления страхового случая и суммы возможного ущерба. Например, опросы показывают, что не имея информации о вероятности выхода из строя прибора, лишь малая часть покупателей принимает во внимание удачу при принятии решения о покупке гарантии (Хогарт и Кунройтер [9]).

В то же время неверно делать вывод о том, что при доступности и полноте информации индивид сможет принять решение, удовлетворяющее моделям экономистов. Исследования показывают, что индивиды придают слишком большее значение колебаниям размера страховых премий, чем ожидаемым выгодам от разных тарифных планов (Абалак и Грубер, 2011 [10]). Кроме того, значительное влияние на выбор имеет опыт индивида – если ранее он столкнулся с наступлением страхового случая, более вероятно, что он оформит страховой полис на следующие периоды и будет расценивать вероятность неблагоприятного события выше. Ещё одним важным фактором, который влияет на поведение индивидов, является то, что события, вероятность которых сводится к нулю, вне зависимости от тяжести последствий, воспринимаются как невозможные, а события, вероятность наступления которых трудно или невозможно определить, из-за неопределенности сильнее волнуют человека.

Со стороны предложения существует другая поведенческая проблема – неблагоприятный отбор (adverse selection в англоязычных источниках). Данная проблема была подробно разобрана в статье М. Ротшильда и Дж. Стиглица [11]. Она заключается в том, что если вероятность наступления потерь оценивается в среднем для всех граждан, то рациональные индивиды, понимающие свой уровень риска, будут приобретать страховку только в случае, если вероятность потерь для них выше средней, в то время как остальные посчитают страхование не имеющим смысла. В таком случае страховая компания окажется в ситуации, когда её оценка риска оказалась неверной и суммы страховых премий будет недостаточно, чтобы выплачивать страховые суммы и оставаться финансово-стабильными, а дальнейшее увеличение страховых премий будет только снижать число клиентов и повышать риск банкротства.

Ещё одной проблемой со стороны предложения заключается в моральном ущербе (moral hazard в англоязычной литературе) – ситуации, когда застрахованное лицо меняет своё поведение так, чтобы повысить вероятность потерь, а страховая компания не может отследить такое поведение. Данная проблема ведёт к неверной оценке вероятности потерь страховой компании и так же, как неблагоприятный отбор, может приводить к потере финансовой стабильности страховой компанией и даже банкротству. Но само по себе явление морального ущерба не подходит под определение аномалии, аномалией является нежелание индивидов оформлять страховой полис с ограничениями, которые позволяют страховой компании контролировать моральный ущерб, за меньшую цену, чем без данных ограничений.

Со стороны предложения возникают также проблемы, связанные с совместными событиями, например, если клиент А несёт убытки, то и клиент Б тоже понесёт убытки, потому что в таком случае нельзя рассчитать вероятность неблагоприятного события, используя законы больших чисел. Но сама по себе проблема корреляции случайных величин не является аномалией, как и в случае с моральным ущербом. Такие риски могут привести к банкротству компании, если она не увеличит сумму страховых премий и резервы, однако тогда акционеры компании будут проигрывать из-за роста трансакционных издержек и снижения выручки. Так как менеджеры компании теряют всё в случае банкротства, они стараются сильнее защититься от такого исхода, чем этого ожидают инвесторы, жертвуя при этом прибылью. Такое поведение, когда менеджмент и акционеры преследуют разные цели тоже является аномальным.

Исследованная литература демонстрирует, что рынки страховых услуг не всегда работают эффективно, причиной тому является поведение индивида и страховых фирм, которое порой отклоняется от стандартного поведения, определяемого экономическими моделями, что приводит к аномалиям на рынке. Аномалии возникают как на стороне спроса, так и на стороне предложения, однако важно поставить своей целью устранение аномалий спроса и предложения, используя инструменты, которые будут рассмотрены в дальнейших разделах работы. Поведенческая экономика, изучая такие аномалии, позволяет выявить особенности человеческого поведения, которые при текущих условиях ведут к аномалиям. Результаты исследований поведенческих экономистов способствуют тому, чтобы экономические агенты, учитывая их, могли изменить условия на рынке таким образом, чтобы поведение скорректировалось относительно экономически рационального, что позволит сделать рынки более эффективными и приведёт к росту общего благосостояния.

## ПОСТАНОВКА ПРОБЛЕМ И ВЫБОР СПОСОБОВ ИХ РЕШЕНИЯ

Ранее мы рассмотрели рынки страховых услуг в общем, выявляя конкретные аномалии, которые свойственны различным рынкам. В данном разделе внимание будет посвящено конкретному рынку и его проблемам, вызывающим аномалии, а также предлагаемым экономистами подходам к их решению. Рассматриваться будет рынок страховых полисов для транспортных средств, в основном, автомобилей.

Рынок автострахования был выбран для данной работы, так как он представляет большой интерес, учитывая его специфичность, обусловленной большой долей государственного регулирования, и большие размеры в США (около трети всего рынка страховых услуг) [12], а также наличие общих черт для рынков разных стран, из чего следует независимость основных характеристик данного рынка от местоположения и климатических условий.

На данном рынке несомненно также присутствуют аномалии, однако не все из тех, что были упомянуты ранее. Например, аномально низкий спрос на страхование при обоснованных суммах страховых премий практически не встречается на рынке страхования автомобилей из-за того, что спрос обеспечивается требованием государства о наличии страхового полиса для управления транспортным средством. Аномально высокий спрос на страхование с необоснованно высокими страховыми премиями может встречаться на рынке, однако чаще всего государство всё же регулирует размер страховых премий, устраняя возможность аномалии. Однако аномалия выбора нерациональных суммы или типа страхового покрытия возникает на рынке страхования автомобилей достаточно часто. Именно эту аномалию спроса мы рассмотрим подробнее и выстроим стратегию её устранения. Аномалии, связанные с моральным ущербом и неблагоприятным отбором, тоже присутствуют на рынке автострахования и требуют детального рассмотрения в этом контексте.

Аномалия выбора нерациональных суммы или типа покрытия на рынке страхования автомобилей заключается именно в том, что из-за обязанности застраховать автомобиль индивид принимает меньше действий для реальной оценки своих рисков и часто просто следует предложенным законодательством рекомендациям, которые основаны на усредненных данных и в каждом конкретном случае могут оказаться не оптимальными. В целом, данная аномалия может обосновывать и тем, что рекомендации государства являются в некоторых юрисдикциях обязательными, и отклоняться от них нельзя, однако даже когда государство не обязывает заключать договор на конкретных условиях, аномалия часто не исчезает из-за предубеждений и особенностей восприятия человеком вероятностей.

Аномалии предложения, связанные с моральным ущербом, очень часто встречаются именно на рынке автомобильного страхования из-за подхода к оценке риска и обязательность страхования. Индивид порой принимает решение застраховать автомобиль только с целью получить возможность им управлять, из-за чего он может воспринимать получение страховых выплат как компенсацию за данный выбор и потому меньше заботиться о состоянии автомобиля, что увеличивает вероятность наступления страхового случая. Кроме того, индивид с таким настроем склонен предоставлять недостоверные сведения, что заставляет страховые компании проводить независимые экспертизы и увеличивает трансакционные издержки, последствием чего может стать рост страховых премий для всех клиентов. Самым важным поведенческим аспектом, вызывающим аномалии, является недостаточное понимание страхуемого лица необходимости оформления страхового полиса и восприятие его в качестве инвестиции.

Неблагоприятный отбор, как уже было отмечено в прошлом разделе, также приводит к аномалиям предложения. Неблагоприятный отбор сказывается на тех тарифах, которые являются добровольными, чаще всего, касающиеся возмещения ущерба застрахованному автомобилю, а не ущерба, причинённого им. В данном случае индивид не оформляет страховой полис, если компания предлагает страховые премии, основываясь на среднем риске, так как считает, что его риски ниже средних, а рисковые водители, напротив, оформляют, осознавая, что их риски выше средних.

Таким образом, обозначены проблемы выбранного рынка, которые можно решить, используя принципы поведенческой экономики.

Самым очевидным шагом, который нужно осуществить, является предоставление клиентам достоверной, полной и понятной информации о страховании, индивидуальных рисков и размеров возможного ущерба. В данной работе будет предпринята попытка реализовать алгоритм, предсказывающий индивидуальную вероятность аварии при движении по привычным маршрутам в привычное время, а также индивидуального возможного ущерба с учётом погодных условий и других данных о водителе и автомобиле.

С точки зрения психологии, проще воспринимать вероятности не в абсолютных значениях, а в относительных. Именно поэтому разрабатываемый продукт будет представлять пользователю аналитику с относительными рисками.

Как упоминалось в прошлом разделе, люди часто воспринимают события с маленькими вероятностями как невозможные, поэтому разрабатываемый алгоритм, вычисляя вероятность аварии за одну поездку по маршруту, будет демонстрировать вероятность аварии за год.

Кроме того, продукт будет предоставлять клиенту калькулятор для подсчета выгоды от страхования с франшизой, то есть клиент сможет увидеть, сколько денег он сэкономит и на сколько уменьшится матожидание выплаты.

Перечисленные выше решения помогут устранить некоторые аномалии спроса, что в целом будет способствовать тому, что компании будет легче предсказывать спрос, а, следовательно, и максимизировать прибыль. К тому же, ответственный подход и открытость могут быть эффективно использованы для маркетинга. Однако решение проблем со стороны одного спроса будет недостаточно для эффективности рынка, необходимы также решения, устраняющие аномалии со стороны предложения.

Прежде всего, та же открытая и достоверная информация, предоставляемая клиентам, должна быть основой для расчёта тарифов компанией, в размерах которых ключевую роль должны играть индивидуальная вероятность каждого клиента и потенциальная сумма ущерба. Именно они должны лежать в основе разных размеров страховых сумм. Принимая во внимание данный факт, я разрабатывал продукт, который использует методы машинного обучения для расчёта индивидуальной вероятности и суммы ущерба, которые затем определяли сумму покрытия и страховых премий.

Выше представлены подходы, которые лягут в основу разработки продукта. Важно отметить, что это лишь та часть, которая показалась мне наиболее значимой для реализации. Необходимо также провести анализ существующих продуктов, чтобы понимать, на сколько часто встречаются похожие решения и, если их мало, то почему.

## РАЗРАБОТКА И ТЕСТИРОВАНИЕ ПРОДУКТА

Как было указано выше, для разработки продукта были выбраны определенные подходы, которые необходимо реализовать, чтобы устранить аномалии как со стороны спроса, так и со стороны предложения. В данном разделе будет описываться разработка алгоритмов, необходимых для функционирования продукта.

Важно для начала определить концепцию самого продукта и соответствующих алгоритмов. Поэтому я хочу разделить весь продукт на несколько крупных функций, которые он должен выполнять, а затем описать каждую из них, подход и реализацию.

Важнейшей функцией, которая помогла бы избежать аномалий со стороны и спроса, и предложения, является индивидуальная оценка риска клиента. Учитывая специфику данных, необходимых для наиболее точной оценки риска, в данный момент, не имея доступа к части информации, которая является конфиденциальной, в разрабатываемом продукте не получится в полной мере реализовать данный функционал, однако частично воплотить его удастся. Продукт будет рассчитывать вероятность попадания клиента в аварию на основе информации о его регулярных маршрутах в городе, используя открытые данные об авариях (местоположение, тип автомобиля) и погодных условиях в день аварии (температура, наличие осадков, скорости и направлении ветра, общей погодной обстановке). Рассчитанная вероятность будет демонстрироваться пользователю, а также использоваться для формирования размера страховых премий. Для реализации будут применяться алгоритмы машинного обучения.

Вторым значимым компонентом, который также относится к полноте информации и устраняет аномалии со стороны как спроса, так и предложения, является расчёт потенциальных потерь клиента в случае попадания в аварию. Данная сумма также будет сообщена потенциальному клиенту, а также будет лежать в основе расчёта размера страховых премий. Решение данной задачи регрессии возможно с использованием алгоритмов машинного обучения с использованием данных страховых компаний в рассматриваемом городе об обращениях клиентов и выплатах (информация о клиенте, автомобиле, тяжести аварии).

Наконец, необходимо разработать формулу для расчёта размера страховых премий, учитывая размер франшизы, вероятности страхового случая для клиента и возможного размера потери.

Начнём с первой функции. Продукт будет разрабатываться для города Нью-Йорк. Более того, учитывая его размеры, для разработки возьмём данные только для Манхэттена. Прежде всего, с официального сайта открытых данных правительства города были скачаны данные об авариях в городе в период с 2012 по 2021 год, а также данные о размере трафика на дорогах города. С сайта Национального центра климатической информации США был скачан набор данных о погоде со станции в Центральном парке на Манхэттене за период с 1 января 2012 года по 31 декабря 2020 года. Затем, перед разработкой, был проведён анализ данных для выявления значимых параметров, необходимых для предсказания вероятностей. Предполагая, что число аварий сильно зависит от дня недели и месяца, я построил соответствующий график (рисунок 1).

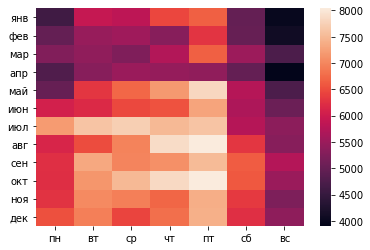


Рисунок 1. Число аварий в разрезе по месяцам и дням недели

По графику видно, что больше всего аварий случается по пятницам, по воскресеньям – меньше всего. Кроме того, в выходные в целом аварий меньше, чем в будние, что может быть связано с большим потоком в будние дни. Ещё одним интересным фактом является то, что с мая число аварий становится больше во все дни недели, что, наверное, связано с появлением на дорогах тех, кто страхует автомобиль только на несколько месяцев, а также мотоциклистов и велосипедистов. Если наблюдается зависимость числа аварий от дня недели и месяца, то нужно проверить гипотезу о распределении трафика (что по будням машин больше, чем по выходным). Подобный график может подтвердить или опровергнуть гипотезу. Построенный график приведен на рисунке 2.

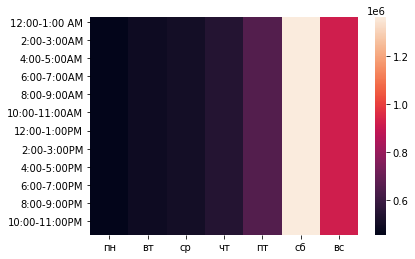


Рисунок 2. Объем трафика в разрезе дней недели и временного периода

По графику видно, что самый слабый трафик на улицах Манхэттена в любой часовой интервал в течение дня меньше в будние, чем в выходные. В понедельник – самый небольшой, а в субботу – самый большой. Вероятно, это связано с тем, что многие жители других районов Нью-Йорка на выходные приезжают на Манхэттен, где расположены крупные точки притяжения города. Данный график не подтверждает гипотезу об одинаковой частоте аварий в разные дни недели: по будням аварий больше, а трафик – меньше, что говорит о том, что поездки в разные дни неделю имеют разный риск.

Интересно было бы также посмотреть на то, как погодные условия сказываются на количестве аварий. Вероятно, во время осадков и снижения видимости увеличивается число дорожно-транспортных происшествий. Построим график распределения аварий по погодным условиям (рисунок 3).

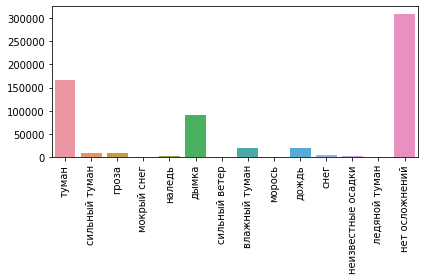


Рисунок 3. Распределение аварий по погодным условиям

Очевидно, что погодные условия могут сказываться на риске аварии. Так, судя по графику, большая часть ДТП происходит, когда ухудшается видимость: во время тумана и дымки, однако большинство аварий приходятся на погоду без осложнений для водителя, что, вероятно, объясняется тем, что такая погода преобладает на Манхэттене. Подтвердить данную гипотезу или опровергнуть может график (рисунок 4).

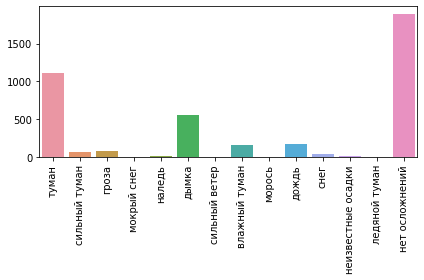


Рисунок 4. Частотность погодных условий за 9 лет на Манхэттене

Оказалось, что графики (рисунки 3 и 4) очень похожи, что может свидетельствовать о том, что корреляция между числом аварий и погодными условиями не очень высокая, однако всё же наблюдается снижение долей влажного тумана и дождя в числе аварий по сравнению с графиком распределения дней, зато доля погоды без осложнений в данном случае растёт относительно доли дней с такой погодой, что может свидетельствовать о том, что частота аварий зависит от погодных условий, при этом замеченную неочевидную корреляцию можно объяснить, например, тем, что в плохую погоду меньшее число водителей решает сесть за руль и трафик уменьшается.

Хотелось бы также рассмотреть распределение аварий по типам автомобилей. Возможно, есть типы автомобилей, аварийность которых выше других. Данную гипотезу можно проверить, построив график по имеющимся данным по авариям. Однако сразу можно сказать, что интерпретация данного графика неоднозначна, так как он не учитывает популярность того или иного класса автомобилей, поэтому также потребуется построить и второй график. Первый график представлен на рисунке 5. Мы видим, что больше всего аварий наблюдается у автомобилей типа седан с большим отрывом от второго типа по количеству аварий (универсал и внедорожник), а также есть типы транспортных средств, аварий с которыми очень мало.

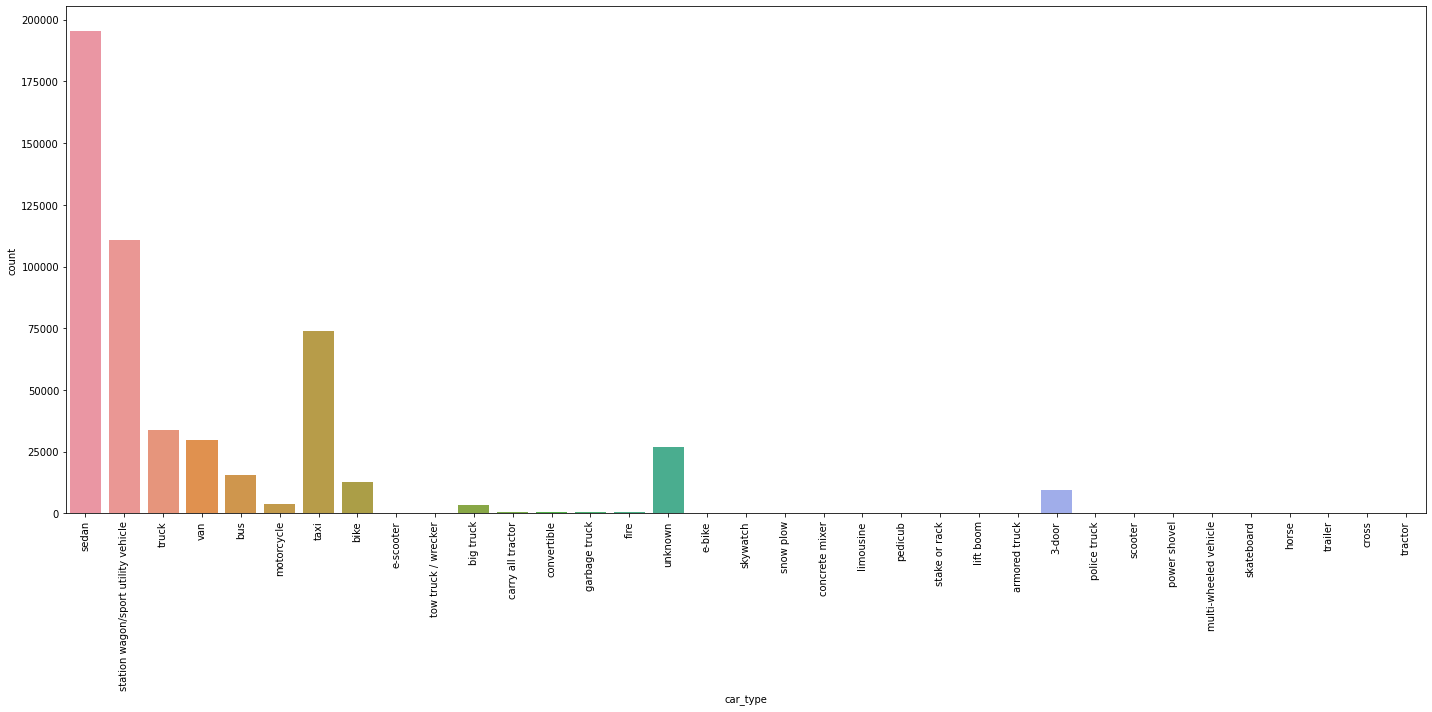


Рисунок 5. Распределение числа аварий по типу ТС

На следующем графике (рисунок 6) показано распределение количества зарегистрированных ТС в Нью-Йорке по типам транспортных средств.

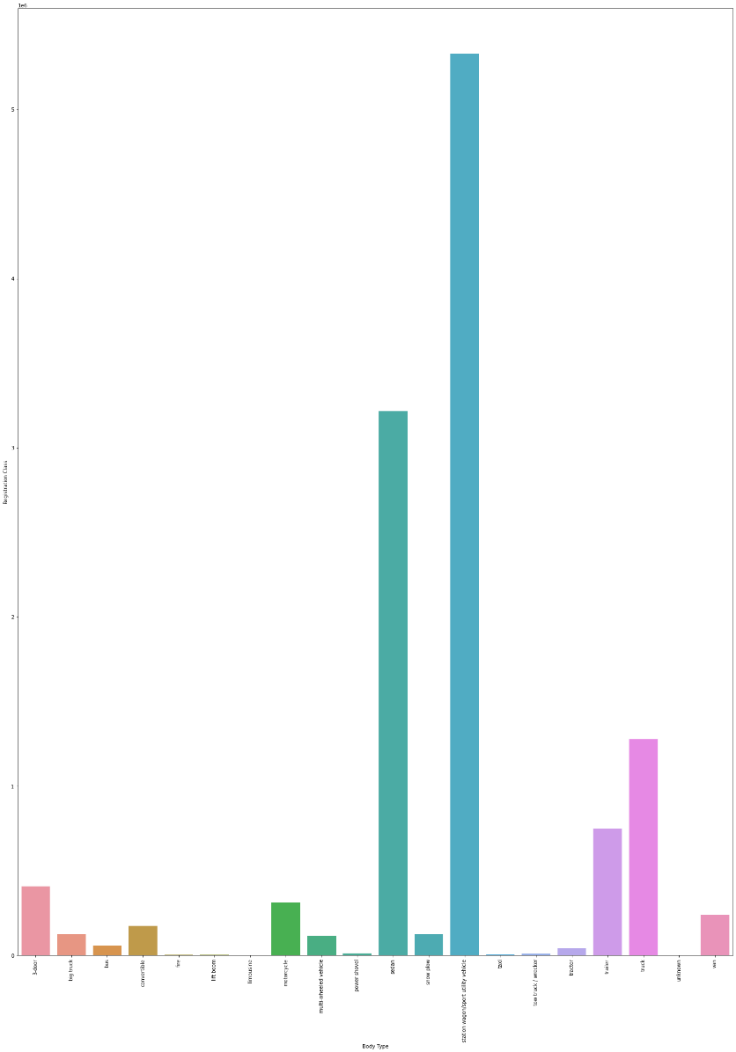


Рисунок 6. Число зарегистрированных ТС по типам ТС

Очевидно, что графики не совсем совпадают – заметно, что седанов зарегистрировано меньше, чем универсалов и внедорожников, что практически противоположно тому, что видно на графике распределения аварий. Тем не менее, доля грузовиков в авариях в целом совпадает с их долей в общем числе автомобилей. Можно сделать вывод о том, что тип автомобиля коррелирует с риском аварии и может быть важным признаком для предсказания данного риска, так как для разных типов частотность аварий различна.

Анализ имеющихся данных показывает, что есть смысл использовать данные о типе транспортного средства, месяце, дне недели, временном периоде, погодных условиях для того, чтобы предсказывать вероятность попадания в аварию клиента. В соответствии с этим, данные были собраны определенным образом: во-первых, данные об авариях были преобразованы в данные об автомобилях, попавших в аварии, то есть из набора данных о ДТП были извлечены количество автомобилей и их типы в связке с датой и местоположением автомобильного происшествия. Во-вторых, дата происшествия была разделена на месяц, день недели и временной период. В-третьих, по дате были объединены (операцией INNER JOIN) таблицы с машинами, попавшими в ДТП, и погодными условиями. Далее, так как решалась задача классификации, необходимо было создать целевую переменную – индикатор того, что автомобиль попал в аварию (для имеющихся данных - всегда 1). Готовый набор данных был преобразован в формат csv для дальнейшего использования в работе. Сейчас посмотреть на содержание данных можно на портале kaggle.com [13], где были опубликованы наборы данных.

Следующим шагом стало создание графа дорожной сети Манхэттена, что было сделано с использованием библиотеки osmnx для Python, которая использует данные Open Street Map [14] – портала с открытым кодом, содержащего различных картографические данные. Построенный граф Манхэттена, граф с добавленными на него авариями представлены на рисунках 7 и 8.

Для дальнейшей работы была необходима также информация о загрузке дорог, так как она очевидным образом сказывается на частоте аварий. Для этого были использованы данные с портала открытых данных Нью-Йорка о потоке в часовые интервалы на улицах города. Улицы города были указаны с помощью названия, что потребовало дополнительной обработки для отбора релевантных данных. Кроме того, дата в данных о потоке и погоде была также преобразована в несколько полей: месяц и день недели.



Рисунок 7. Граф дорожной сети Нью-Йорка в радиусе 10км от центра Манхэттена



Рисунок 8. Граф дорожной сети Нью-Йорка в радиусе 10км от центра Манхэттена с нанесенным авариями (красными точками) в период с 2012 по 2020 годы

Имеющиеся данные были использованы для написания модели, предсказывающей вероятность попадания в аварию для клиента в поездках на протяжении 9 лет по регулярным маршрутам.

Для реализации основной логики программы был выбран объектно-ориентированный подход, так как он позволяет создать структуру данных, хранящую атрибуты и методы в оптимальном для дальнейшего взаимодействия формате.

В рамках реализуемой задачи было создано два класса объектов: пользователь (User) и предсказывающий алгоритм (ProbaPredictor). Пользователь имеет несколько атрибутов: имя (name), тип автомобиля (car\_type) и регулярные маршруты (marchroutes). Регулярные маршруты – составной атрибут, представляющий из себя словарь (dict), в котором ключом является название маршрута (name), а значениями – словари, описывающие маршрут, ключами в которых являются точка отправления (origin), точка следования (destination), месяца, когда маршрут актуален (months), дни недели (days\_of\_week) и период времени, когда по нему осуществляется движение (time\_periods). Предсказывающий алгоритм создаётся с наборами данных, являющимися его атрибутами: данные о транспортных средствах, попавших в аварию (positive), данные о погодных условиях в рассматриваемый период (weather\_df), данные о величине траффика (aadt), данные об авариях, используемых для визуализации результата (raw\_data), данные о количестве зарегистрированных транспортных средств (car\_popularity), граф дорожной сети города (city\_graph). Кроме того, алгоритм имеет атрибут классификатора (clf) – модели, используемой для предсказания, применяемая функция потерь – логарифмическая (LogLoss), предсказывающая вероятности. Для решения данной задачи, учитывая её специфику – табличные данные, наличие категориальных признаков, необходимость предсказывать вероятность, а не класс – было принято решение использовать алгоритм градиентного бустинга, так как это один из самых мощных инструментов машинного обучения на табличных данных. Градиентный бустинг является композицией алгоритмов, называемых базовыми. Базовым алгоритмом для решаемой задачи, принимая во внимание все сложные зависимости, выявленные на этапе анализа данных, был выбран алгоритм решающего дерева. В качестве библиотеки, в которой реализован градиентный бустинг, использовалась Python версия catboost – библиотеки градиентного бустинга, умеющей работать с категориальными данными, разработанной компанией Yandex.

Упомянутые ранее классы объектов также имеют методы, в которых и реализуется основная логика программы. Для пользователя реализован один метод - метод добавления нового маршрута (add\_marchroute), который, принимая на вход название маршрута (name), точку отправления (origin), точку назначения (destination), список месяцев (monthes), список дней недели (days\_of\_week), список временных периодов движения по маршруту (time\_periods), индикатор наличия обратного маршрута (is\_backward, по умолчанию – ложь), список временных периодов движения по обратному маршруту (back\_periods), добавляет данный маршрут в словарь маршрутов пользователя (marchroutes). Для предсказывающего алгоритма реализованы следующие методы:

* генерация данных без аварий (generate\_negative) – данный метод необходим, так как мы предполагаем обучение алгоритма на размеченных данных, то есть матрице с объектами, которые представляют из себя набор значений признаков и целевой переменной – наличия аварий. Кроме того, авария действительно не является частым событием и большая часть автомобилей в каждый момент времени не попадает в аварийную ситуацию, поэтому случайная генерация таких автомобилей может иметь смысл. Метод принимает на вход месяц (month), день недели (day\_of\_week), временной период (time\_period), средний размер трафика на данном участке дороги в данный часовой интервал в данный день недели в данный месяц (aadt) и название ребра графа дорожной сети, по которому проходит маршрут (edge). Далее он генерирует случайным образом, учитывая частотность типов автомобилей, а также вероятность различных погодных условий, набор машин, которые двигались в данные дни по той же дороге, но не попали в аварию.
* генерация случайной точки на дороге (generate\_random\_point) – данный метод, получая на вход координаты начала и конца дороги, находит уравнение прямой и, генерируя случайное значение абсциссы из интервала между минимума из абсцисс начал и конца до максимума между ними же, получает соответствующее значение ординаты.
* обработка пользователя (process\_user) – данный метод, получая на вход объект класса Пользователь (user), индикатор необходимости построения визуализации маршрута (to\_plot, по умолчанию – ложь) и индикатор необходимости вывода в консоль результатов для маршрутов (print\_result, по умолчанию – ложь), перебирает маршруты пользователя, вычисляя вероятность попасть в аварию на них. Метод возвращает полную вероятность пользователя попасть в аварию, учитывая характеристики его автомобиля и регулярных маршрутов и частоту движения по ним.
* обработка маршрута (process\_marchroute) – метод, рассчитывающий полную вероятность пользователя попасть в аварию на данном маршруте. Метод принимает в качестве аргументов маршрут (marchroute), тип автомобиля пользователя (car\_type), индикатор необходимости визуализации (to\_plot). Он обрабатывает маршрут, перебирая грани маршрута, месяцы, дни недели, часовые периоды, рассчитывая вероятность попадания в аварию для каждой комбинации данных параметров. Метод возвращает риск аварии для маршрута.
* Предсказание вероятности для определенного набора данных (predict\_proba) – метод, в котором применяется градиентный бустинг для предсказания вероятности аварии для клиента. Он принимает в качестве параметров сгенерированную матрицу объекты-признаки (машины на данной дороге в конкретные месяц, день недели и часовой интервал и погодные условия) (X), вектор правильных ответов для объектов (y), тип автомобиля пользователя (car\_type), месяц (month), день недели (day), часовой интервал (period). Внутри метода выборка разделается на обучающую и тестовую, на обучающей происходит обучение классификатора, на тестовой – валидация и выбор лучшей модели. Обученный алгоритм предсказывает для пользователя вероятность попадания в аварию на маршруте в случаях плохих и хороших погодных условий, а затем возвращает полную вероятность аварии.
* построение визуализации (plot\_graph) – метод, визуализирующий маршрут на графе города и аварии, случившиеся в городе в то же время.
* сохранение качества классификатора (save\_clf\_quality) – метод, оценивающий качество классификации по доли правильных ответов, точности и полноте для каждого запуска классификатора.
* просмотр качества классификатора (get\_clf\_quality) – метод, возвращающий результаты модели для каждого маршрута.

Предсказывающий алгоритм имеет сложную структуру, которая может

быть не очевидной из описания, поэтому далее будет приведена схема выполнения алгоритма (рисунок 9).

**ProbaPredictor**

process\_user

process\_marchroute

predict\_proba

call

call

call

generate\_negative

return

return

return

Рисунок 9. Схема работы алгоритма предсказания вероятности

Протестируем готовый алгоритм, создав пользователя с именем Иван и машиной типа фургон, добавив маршрут из одного из жилых домов на Манхэттене до офисного здания по понедельникам в январе и феврале в интервале с 9 до 10 часов утра, а также маршрут до магазина игрушек из того же здания в тот же временной интервал по субботам в апреле. Визуализация маршрутов представлена на рисунках 10 и 11.

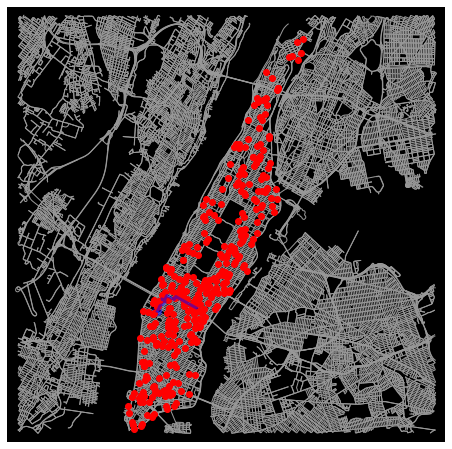


Рисунок 10. Маршрут до офисного здания по понедельникам в январе и феврале с 9 до 10 утра

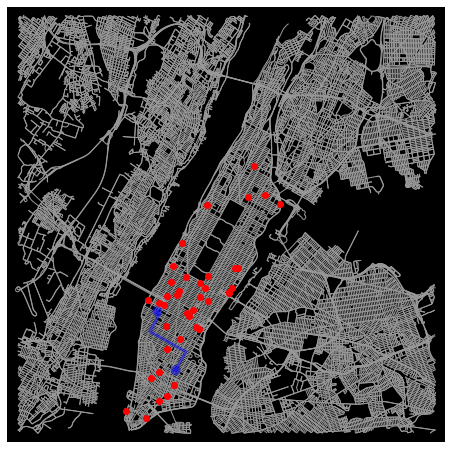


Рисунок 11. Маршрут из дома до магазина игрушек по субботам в апреле во временной промежуток с 9 до 10 утра

Результат выполнения кода представлен на рисунке 12.

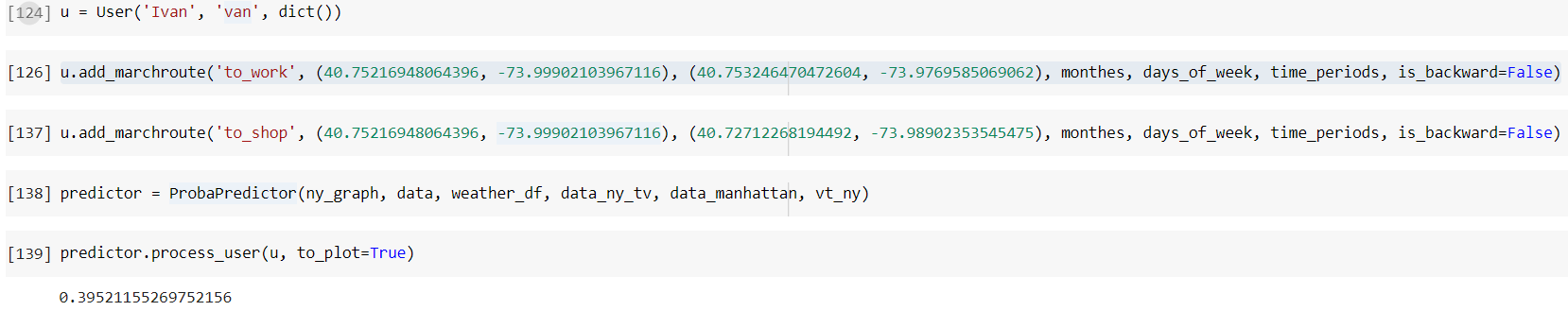


Рисунок 12. Результат работы алгоритма

Как можно увидеть, модель возвращает вероятность в почти 40%, что кажется довольно большим числом, если воспринимать его, как вероятность аварии в одной поездке, однако по факту модель выдает вероятность аварии во время поездок на протяжении 9 лет (с 2012 года по 2020 год). Нетрудно найти вероятность для одного года, считая процесс пуассоновским. Тогда, если вероятность наступления события 1 и более раз за 9 лет равна P, то вероятность того, что оно не наступит ни разу равна 1-P. По формуле Пуассона (1), где х – число наступлений события, можно вычислить p – вероятность аварии за один год.

Данный функционал реализован в классе Пользователь, метод set\_proba. Теперь получим вероятность для нашего пользователя Ивана в годовом выражении – она составляет 5.6% (рисунок 13).

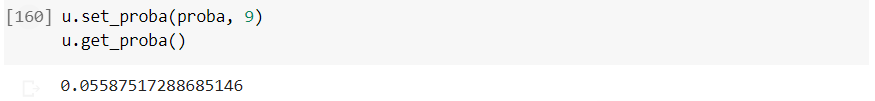


Рисунок 13. Расчёт вероятности для пользователя попасть в аварию в годовом выражении

Таким образом, разработанный алгоритм позволяет вычислить индивидуальный риск для пользователя попасть в аварию, учитывая его регулярные маршруты.

Следующим шагом является разработка алгоритма, предсказывающего сумму ущерба для клиента, при этом, вероятно, в большей степени будут важны характеристики автомобиля. Для разработки алгоритма воспользуемся датасетом, опубликованным на GitHub [15], содержащим информацию о суммах, выплаченных попавшим в аварию клиентам страховых компаний. Вероятно, для данной задачи важными будут характеристики автомобиля.

Начнем с анализа данных. Посмотрим, какие признаки есть в данных: индикатор наличия детей в машине во время аварии (KIDSDRIV), дата рождения клиента (BIRTH), возраст клиента (AGE), число детей у клиента (HOMEKIDS), стаж работы (YOJ), доход (INCOME), индикатор того, что клиент воспитывает детей один (PARENT1), стоимость дома (HOME\_VAL), семейное положение (MSTATUS), пол (GENDER), образование (EDUCATION), занятость (OCCUPATION), расстояние до работы (TRAVTIME), применение автомобиля (CAR\_USE), стоимость автомобиля (BLUEBOOK), продолжительность рабочего дня (TIF), тип автомобиля (CAR\_TYPE), индикатор красного цвета машины (RED\_CAR), размер выплаченных страховых сумм за последние 5 лет (OLDCLAIM), количество обращений за последние 5 лет (CLM\_FREQ), дата последнего обновления прав (REVOKED), баллы водителя относительно за нарушение ПДД (MVR\_PTS), размер выплат (CLM\_AMT), возраст автомобиля (CAR\_AGE), индикатор вождения в городе (URBANCITY).

Построим матрицу корреляции между признаками и целевой переменной, чтобы посмотреть, какие признаки могут быть значимыми (рисунок 14).

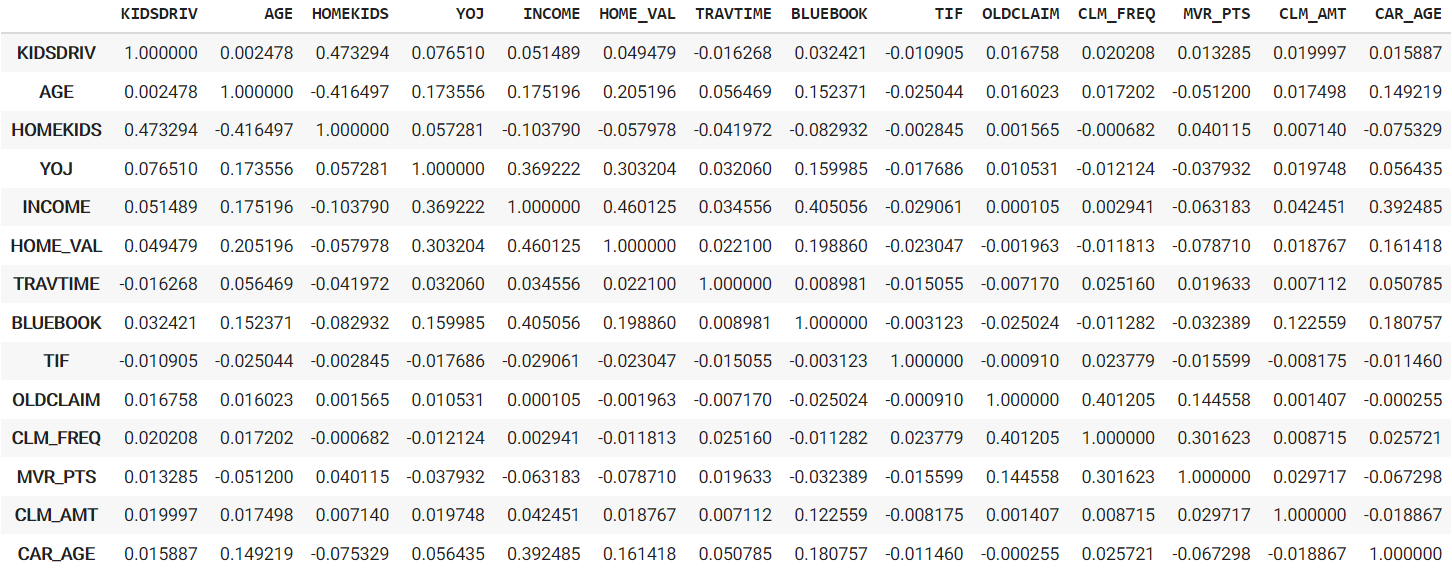


Рисунок . Матрица корреляции между вещественными признаками и целевой переменной

Вещественные признаки не сильно коррелированы между собой, наиболее коррелированные признаки (46%) – доход клиента (INCOME) и стоимости дома (HOME\_VAL). Заметно также, что вещественные признаки не имеют большой корреляции с целевой переменной (CLM\_AMT), сильнее всего коррелирует с величиной страховой выплаты признак стоимости автомобиля (BLUEBOOK). Интересно также посмотреть, какое значение имеют категориальные признаки на целевую переменную (рисунок 15).

Данный график демонстрирует среднее значение (высота цветных прямоугольников) и разброс (черная полоса) целевой переменной по значениям категориального признака. Соответственно, чем больше размер черной полосы у значений признака, тем менее разделяемы объекты по данному признаку. Интересно, что в среднем одинокие родители получают большую компенсацию в случае аварии, хотя и разброс значений у них выше. Данное наблюдение не стоит интерпретировать, как действительную взаимосвязь указанных признаков. Судя по данным, признак семейного положения и индикатор одиноких родителей имеет высокую корреляцию. При обучении важность одного из признаков может оказаться близкой к 0. Гораздо интереснее, как распределены средние суммы страховых выплат между клиентами с разным уровнем образования. Заметно, что между разными уровнями образования существует довольно значительная разница относительно средней суммы выплат, однако разброс довольно большой, из-за чего все равно разделение по данному признаку может оказаться затруднительным. Аналогичная ситуация складывается и с признаком места работы. В целом мы видим, что более образованные люди с более квалифицированными профессиями получают в среднем большие выплаты, что может говорить о том, что их транспортные средства стоят больше. Страховые выплаты коммерческим автомобилям в среднем больше, чем частным, что может быть связано с тем, что клиенты аккуратнее водят личное авто, чем коммерческое. Признак типа машины также открывает не самый очевидный факт – спортивные автомобили получают в среднем меньше страховых выплат, чем какие-либо другие типы машин. Забавно, но оказывается, что красный цвет машины оказывается не самым бесполезным признаком. А вот признак типа населенного пункта, в котором зарегистрирован автомобиль не оказывает влияния на среднюю выплату, но для машин вне города разброс значений больше.

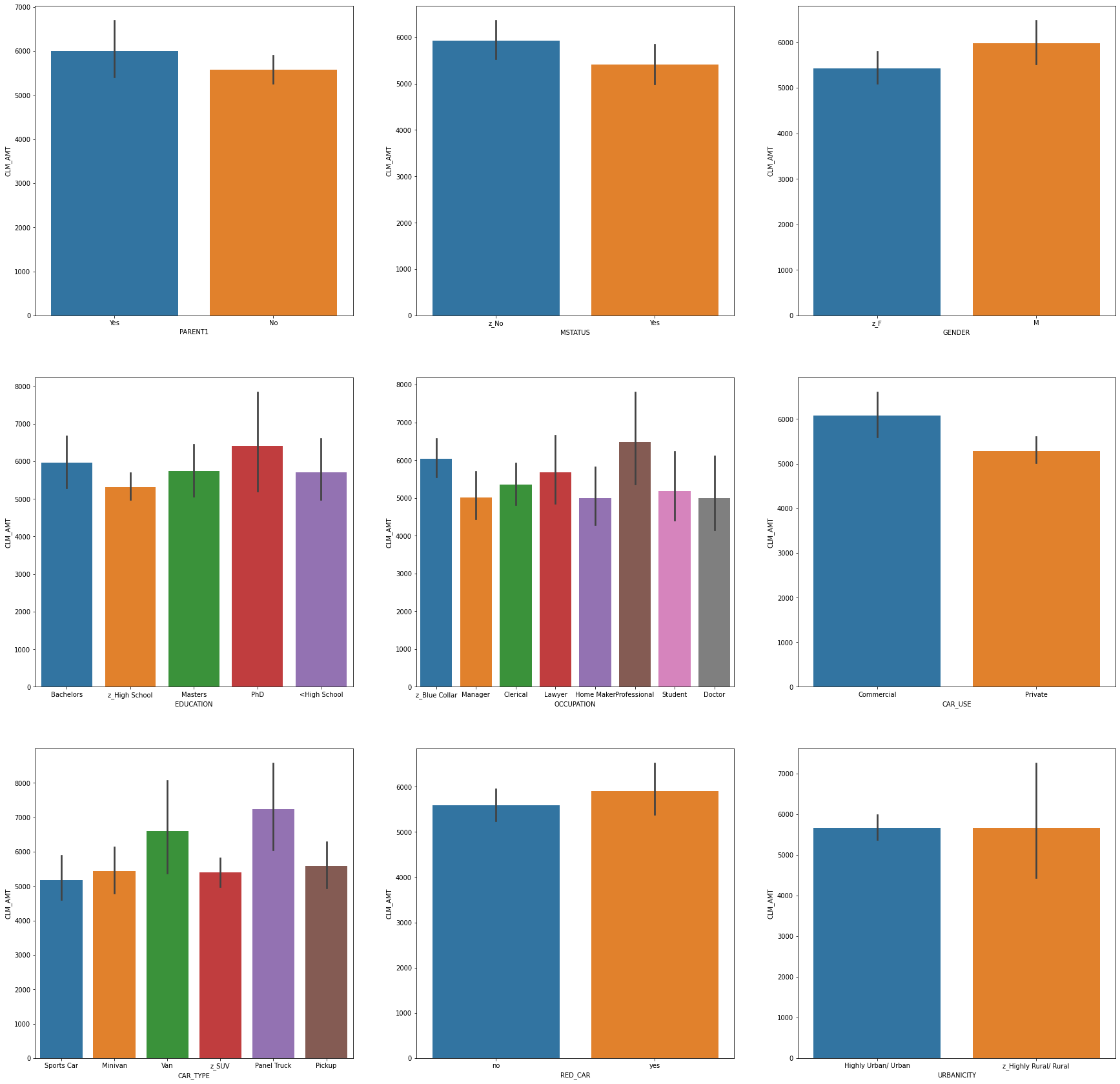


Рисунок . Графики зависимости значения целевой переменной от категориальных признаков

Учитывая специфику имеющихся данных: отсутствие признаков, по которым явно прослеживается возможность разделения объектов на группы с близкими значениями размера страховой выплаты и слабая корреляция между вещественными признаками и целевой переменной, необходимо использовать мощный алгоритм – композицию базовых слабых алгоритмов. Такими алгоритмами являются случайный лес и градиентный бустинг. Однако использование случайного леса может быть затруднительным, так как данные имеют большое количество признаков, часть из которых категориальные и при кодировании могут ещё значительно расширить количество признаков. Для предсказания будем использовать алгоритм градиентного бустинга для регрессии catboost.

Последним этапом разработки продукта является разработка формулы премий. Размер премии, как было указано в прошлом разделе должен соответствовать риску клиента и размеру потенциального ущерба для того, чтобы сократить вероятность наличия аномалий на рынке. Поэтому разрабатываемый продукт будет учитывать данные рекомендации поведенческих экономистов.

Очевидно, что страховая премия является основой финансовой стабильности компании, и потому критически важно разработать формулу, сводящую к нулю вероятность банкротства компании. Суммы всех премий должно хватать на обеспечение операционной деятельности компании, выплату страховых сумм и создание резервов. Рассмотрим следующую модель страховой фирмы: у фирмы есть постоянные издержки на функционирование, не зависящие от количества застрахованных автомобилей, есть переменные издержки, связанны с обслуживанием клиентов, зависят от количества застрахованных ТС (n), и издержки, связанные с резервированием денежных средств, зависящие от размера резервов (R). Запишем формулу издержек фирмы (2).

Фирма владеет информацией об индивидуальном риске клиента, рассчитываемого разработанной моделью и размером потенциального ущерба, полученного также с помощью модели машинного обучения. Данная информация позволяет рассмотреть две случайные величины. Первая – число аварий для каждого клиента за период страхования (η), вторая – сумма страхового покрытия, которую выплатит фирма за период страхования всем клиентам (ξ).

Полагая, что число аварий распределено по пуассоновскому распределению, мы можем сказать, что вычисляемая алгоритмом вероятность аварии для клиента за поездки на протяжении года является матожиданием числа аварий за год, то есть рассматриваемый период страхования. Таким образом, мы имеем для каждого клиента матожидание количества его аварий за год, которое редко превышает 0,1. Случайная величина суммы ущерба, принимая во внимание имеющиеся данные, является дискретной, принимающей значения от суммы всех ущербов по всем клиентам за год с вероятностью, равной произведению всех вероятностей аварии клиентов (в силу независимости величин) до 0 (если ни одной аварии не произойдет). Очевидно, что также в силу независимости случайных событий при росте числа клиентов, вероятность того, что аварий не произойдет будет снижаться. Число возможных значений случайной величины равно числу всех подмножеств множества клиентов. Таким образом, мы можем рассчитать выборочное матожидание всех выплат фирмы за год, а также найти доверительный интервал для истинного значения математического ожидания.

Для того, чтобы фирма оставалась стабильной и не обанкротилась, ей необходимо иметь резервы. Величину резервов следует выбрать равной δ – точности интервальной оценки с 99% уровнем доверия. Тогда экономически целесообразно фирме максимизировать следующую функцию прибыли:

В формуле (3) – величина страховой премии i-го клиента. Максимизировать прибыль очевидно необходимо по n – числу клиентов. Учитывая уравнения (2) и (3) запишем условие 1-го порядка:

Уравнение (4) является формулой для расчёта страховых премий, которое можно произвести программным путём. Формула зависит от числа клиентов, риска клиента и суммы его покрытия. Дополнительно укажем, что математическое ожидание выплат вычисляется по следующей формуле:

В формуле (5) – величина франшизы. Соответственно, величина франшизы также влияет на размер страховой премии. Разработанная модель отвечает поставленным задачам.

# ЗАКЛЮЧЕНИЕ

В данной работе был разработан подход к расчёту риска дорожно-транспортного происшествия для отдельного водителя с учётом его регулярных маршрутов, погодных условий и вида транспортного средства, а также алгоритм, решающий задачу регрессии для нахождения суммы ущерба, а также аналитически получена формула для расчёта размера страховых премий. Всё это может использоваться для устранения аномалий на рынке автострахования: информация о риске аварии демонстрируется за 9 лет, чтобы клиент понимал, что, на самом деле, на протяжении длительного срока, его риски понести ущерб оказываются не такими незначительными, как за одну поездку. Кроме того, результат алгоритма, предсказывающего возможный ущерб, также может убедить клиента тщательнее подойти к вопросу страхования, чтобы сохранить больше денежных средств. Таким образом, можно сказать о том, что цели работы выполнены.

Однако нельзя не сказать о существующих проблемах: построенные модели машинного обучения можно улучшать, повышая их качество, посредством добавления более точных данных, что возможно сделать при наличии желания и денежных средств. К тому же, алгоритм, рассчитывающий вероятность аварии выполняется довольно долго даже для маршрутов со средней протяженностью. Это можно было бы также улучшить, избавившись от неинформативных признаков и используя более мощное оборудование.

В перспективе страховые компании могли бы предлагать клиентам, готовым установить у себя в транспортном средстве устройства, которое будет записывать маршруты, уровень вождения и другие данные о клиенте, анализ которых бы позволил с высокой точностью оценивать индивидуальные риски, более выгодные тарифы, а потребители бы имели полный доступ к информации с возможностью оценить её и принять верное решение. Уже сейчас для компаний не является проблемой оценить риски клиентов в суммах на основе анализа доступных фирмам данных, однако они не всегда имеют стимулы для этого, так как на некоторых рынках как, например, в России существуют неконкурентные условия, из-за чего рыночная власть компании довольно сильна, чтобы максимизировать прибыль иначе. Ещё одним препятствием для полного устранения аномалий на рынке страхования транспортных средств является обязательность страхования. В действительности некоторые ограничения имеет смысл отменить, обеспечив при этом полный доступ и правильное представление информации для граждан и страховых компаний, тогда рынок сам придет в равновесие и позволит повышать благосостояние не только больших около государственных страховых фирм, но и всего общества в целом. По крайней мере, технологии для этого существуют и активно развиваются.

# СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. *Kunreuther H. C., Pauly M. V., McMorrow S*. Insurance and Behavioral Economics. Improving Decisions in the Most Misunderstood Industry. – NY: Cambridge University Press, 2013 – 312 p.
2. *Greenberg A*. Applying Behavioral Economics Concepts in Designing Usage-Based Car Insurance Products. – Washington, DC: Federal Highway Administration, Office of Operations, 2012 – 26 p.
3. *Талер Р*. Новая поведенческая экономика: почему люди нарушают правила традиционной экономики и как на этом заработать. – М.: Эксмо, 2018 – 384 с.
4. *Hoying T., Spit M*. Life Insurers Secret Weapon: Behavioral Economics // Accenture strategy, 2017.
5. *Malhotra R., Sharma S*. Machine Learning in Insurance // Accenture FS Perspectives, 2018.
6. *Culter D. M. and Zeckhauser R*. Extending the theory to meet the practice of insurance. – Washington, DC: Brookings Institution Press, 2004 – 53p.
7. *Sydnor J*. (Over)insuring modest risks // American Economic Journal: Applied Economics. 2010. No. 2. P. 177-199.
8. *Liebman J. B. and Zeckhauser R*. Simple humans, complex insurance, subtle subsidies. // NBER Working Paper No. 14330. Cambridge, MA: National Bureau of Economic Research, 2008.
9. *Hogarth R. M. and Kunreuther H*. Decision making under ignorance: Arguing with yourself. // Joutnal of Risk and Uncertainty. 1995. No. 10. P. 15-36.
10. *Abaluck J*. *and Gruber J*. Choice inconsistencies among the elderly: Evidence from plan choice in the Medicare part D program. // American Economic Review. 2011. No. 101. P. 1180-1210.
11. *Rothschild M. and Stiglitz J*. Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. // Quarterly Journal of Economics. 1976. No. 90. P. 629-649.
12. IBISWorld: Atomobile Insurance in the US – Market Size 2004-2026 [Электронный ресурс] – URL: <https://www.ibisworld.com/industry-statistics/market-size/automobile-insurance-united-states/> (Дата обращения: 20.05.2021).
13. Manhattan Road Accidents – Kaggle [Электронный ресурс] – URL: <https://www.kaggle.com/igorlukhnev/mahattan-road-accidents> (Дата обращения: 08.06.2021).
14. Open Street Map [Электронный ресурс] – URL: <https://www.openstreetmap.org/#map=2/69.6/-74.9> (Дата обращения 15.05.2021).
15. GitHub – devin-norris / EDA-ML-Car-Insurance [Электронный ресурс] – URL: <https://github.com/devin-norris/EDA-ML-Car-Insurance> (Дата обращения: 09.06.2021).

# ПРИЛОЖЕНИЕ 1. ИСХОДНЫЙ КОД АЛГОРИТМА ПРЕДСКАЗАНИЯ РИСКА

TYPE = 'car\_type'

VT = 'Body Type'

AMOUNT = 'Registration Class'

MONTH = 'month'

DOW = 'day\_of\_week'

TP = 'time\_period'

EDGE = 'edge'

TARGET = 'isaccident'

OR = 'origin'

DEST = 'destination'

MONTHS = 'months'

DOWS = 'days\_of\_week'

TPS = 'time\_periods'

GW = 'without\_driver\_troubles'

WEATHER = ['AWND', 'PGTM',  'PRCP', 'SNOW', 'SNWD', 'TAVG', 'TMAX', 'TMIN', 'TSUN', 'WDF2', 'WDF5', 'WSF2', 'WSF5', 'WT01', 'WT02', 'WT03', 'WT04', 'WT06', 'WT08', 'WT11', 'WT13', 'WT14', 'WT16', 'WT18', 'WT19', 'WT22', 'without\_driver\_troubles']

TIMES = ['12:00-1:00 AM', '1:00-2:00AM', '2:00-3:00AM', '3:00-4:00AM', '4:00-5:00AM', '5:00-6:00AM', '6:00-7:00AM', '7:00-8:00AM', '8:00-9:00AM', '9:00-10:00AM', '10:00-11:00AM', '11:00-12:00PM', '12:00-1:00PM', '1:00-2:00PM', '2:00-3:00PM', '3:00-4:00PM', '4:00-5:00PM', '5:00-6:00PM', '6:00-7:00PM', '7:00-8:00PM', '8:00-9:00PM', '9:00-10:00PM', '10:00-11:00PM', '11:00-12:00AM']

LON = 'LONGITUDE'

LAT = 'LATITUDE'

CD = ['car\_type', 'month', 'day\_of\_week', 'time\_period']

Листинг 1.1. Объявление глобальных переменных

def create\_path(route\_start, route\_end, city\_graph):

    start\_node = ox.nearest\_nodes(city\_graph, route\_start[1], route\_start[0])

    end\_node = ox.nearest\_nodes(city\_graph, route\_end[1], route\_end[0])

    # Calculate the shortest path

    route = nx.shortest\_path(city\_graph, start\_node, end\_node, weight='travel\_time')

    return route

Листинг 1.2. Объявление функции создания маршрута

def add\_accidents(ax, lng\_ser, lat\_ser, clr):

    ax.scatter(lng\_ser, lat\_ser, c=clr)

def cast\_to\_float(st):

    return tuple(map(float, st.strip('()').split(', ')))

Листинг 1.3. Объявление других вспомогательных функций

def get\_path\_edges(path, city\_graph):

    edges\_list = []

    for i in range(len(path) - 1):

        node1 = path[i]

        node2 = path[i+1]

        for e in city\_graph.edges:

            if node1 in e and node2 in e:

                edges\_list.append(e)

                break

    return edges\_list

Листинг 1.4. Объявление функции, возвращающей рёбра маршрута

class User:

    def \_\_init\_\_(self, name, car\_type, marchroutes):

        self.car\_type = car\_type

        self.name = name

        self.marchroutes = marchroutes

        self.proba = 0

    def add\_marchroute(self, name, origin, destination, monthes, days\_of\_week, time\_periods, is\_backward=False, back\_periods=None):

        self.marchroutes[name] = {OR: origin, DEST: destination, MONTHS: monthes, DOWS: days\_of\_week, TPS: time\_periods}

        if is\_backward:

            self.marchroutes[f'{name}\_back'] = {OR: origin, DEST: destination, MONTHS: monthes, DOWS: days\_of\_week, TPS: back\_periods}

    def set\_proba(self, proba, n):

        self.proba = -np.log(1 - proba) / n

    def get\_proba(self):

        return self.proba

    def get\_one\_trip\_proba(self):

        self.cnt\_trips = 0

        for m in self.marchroutes:

            self.cnt\_trips += len(self.marchroutes[m][MONTHS])\*len(self.marchroutes[m][DOWS])\*len(self.marchroutes[m][TPS]\*4

        return self.proba/self.cnt\_trips

Листинг 1.5. Объявление класса Пользователь

class ProbaPredictor:

    def \_\_init\_\_(self, city\_graph, positive, weather\_df, aadt, raw\_data, car\_popularity):

        self.positive = positive

        self.weather\_df = weather\_df

        self.clf = catboost.CatBoostClassifier(n\_estimators=3000, use\_best\_model=True

        self.aadt = aadt

        self.city\_graph = city\_graph

        self.raw\_data = raw\_data

        self.car\_popularity = car\_popularity.groupby(VT).sum(AMOUNT)

        self.model\_quality = ''

        np.random.seed(123)

    def generate\_negative(self, month, day\_of\_week, time\_period, aadt, edge):

        sample = []

        p = (self.car\_popularity/self.car\_popularity.sum())

        a = self.car\_popularity.index

        aadt = 36\*int(round(aadt, 0))

        gwp = self.weather\_df[self.weather\_df[MONTH] == month][GW].sum()/self.weather\_df[self.weather\_df[MONTH] == month][GW].shape[0]

        gwaadt = int(gwp\*aadt)

        bwaadt = aadt - gwaadt

        weather\_good\_id = np.random.choice(np.array(self.weather\_df[(self.weather\_df[MONTH] == month) & (self.weather\_df[GW] == 1)].index), size=int(gwaadt), replace=True)

        weather\_bad\_id = np.random.choice(np.array(self.weather\_df[(self.weather\_df[MONTH] == month) & (self.weather\_df[GW] == 0)].index), size=int(bwaadt), replace=True)

        weather\_id = np.concatenate((weather\_good\_id, weather\_bad\_id))

        # line = self.city\_graph.edges[edge]['geometry']

        for i in range(aadt):

            car\_type = np.random.choice(a, size=1, p=p[AMOUNT])[0]

            # edge = self.generate\_random\_point(line.coords[0], line.coords[1])

            sample.append([car\_type, month, day\_of\_week, time\_period, edge])

        sample = pd.DataFrame(sample, columns=[TYPE, MONTH, DOW, TP, EDGE])

        sample = pd.concat((sample, self.weather\_df.loc[weather\_id, WEATHER].set\_index(sample.index)), axis=1)

        sample[TARGET] = (sample[TYPE].isna()).astype(int)

        self.negative = sample

    def generate\_random\_point(self, point1, point2):

        a = (point2[1] - point1[1])/(point2[0] - point1[0])

        b = point1[1] - a\*point1[0]

        x = random.uniform(point1[0], point2[0])

        y = a\*x + b

        return (x, y)

    def process\_user(self, user, to\_plot=False, print\_result=False):

        proba\_list = []

        drives = 0

        proba\_m = []

        for m in user.marchroutes:

            cnt\_drives = len(user.marchroutes[m][MONTHS])\*len(user.marchroutes[m][DOWS])\*len(user.marchroutes[m][TPS])

            drives += cnt\_drives

            proba\_m.append(cnt\_drives)

            proba = self.process\_marchroute(user.marchroutes[m], user.car\_type, to\_plot)

            if print\_result:

                print(f'For {user.name} on route {m} proba is: {proba}')

            proba\_list.append(proba)

        return (np.array(proba\_list)\*np.array(proba\_m)/drives).sum()

def process\_marchroute(self, marchroute, car\_type, to\_plot):

        proba = 0

        path = create\_path(marchroute[OR], marchroute[DEST], self.city\_graph)

        path\_edges = get\_path\_edges(path, self.city\_graph)

        if to\_plot:

            self.plot\_graph(path, marchroute[MONTHS], marchroute[DOWS], marchroute[TPS])

        proba\_month = []

        for month in marchroute[MONTHS]:

            proba\_day = []

            for day in marchroute[DOWS]:

                for period in marchroute[TPS]:

                    proba\_route = [0]

                    for e in path\_edges:

                        aadt = self.aadt[(self.aadt[EDGE] == e) & (self.aadt[MONTH] == month) & (self.aadt[DOW] == day)][period].mean()

                        if aadt is np.nan:

                            aadt = self.aadt[(self.aadt[MONTH] == month) & (self.aadt[DOW] == day)][period].mean()

                            if aadt is np.nan:

                                aadt = self.aadt[self.aadt[DOW] == day][period].mean()

                        self.generate\_negative(month, day, period, aadt,  e)

                        X\_p = self.positive[(self.positive[MONTH] == month) & (self.positive[DOW] == day) & (self.positive[TP] == period)]

                        if 'geometry' in self.city\_graph.edges[e]:

                            X\_p = X\_p[X\_p[EDGE].apply(lambda x: self.city\_graph.edges[e]['geometry'].distance(Point(x)) <= 250)]

                            if X\_p.shape[0] == 0:

                                proba\_route.append(0)

                            else:

                                X = pd.concat([X\_p, self.negative])

                                X.sample(frac=1)

                                X.reset\_index(inplace=True

                                X.drop(columns=[EDGE], inplace=True)

                                y = X[TARGET]

                                X.drop([TARGET, 'index'], axis=1, inplace=True)

                                proba\_route.append((1-sum(proba\_route))\*self.predict\_proba(X, y, car\_type, month, day, period, e))

                proba\_day.append(sum(proba\_route))

            proba\_month.append(sum(proba\_day)/len(proba\_day))

        proba = sum(proba\_month)/len(proba\_month)

        return proba

    def predict\_proba(self, X, y, car\_type, month, day, period, edge):

        X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, random\_state=17, test\_size=0.05)

        self.clf.fit(X\_train, y\_train, cat\_features=[0, 3], eval\_set=(X\_test, y\_test), silent=True)

        weather\_id\_good = random.choice(self.weather\_df[(self.weather\_df[MONTH] == month) & (self.weather\_df[GW].astype(bool))].index)

        weather\_id\_bad = random.choice(self.weather\_df[(self.weather\_df[MONTH] == month) & (~self.weather\_df[GW].astype(bool))].index)

        good\_weather\_proba = self.weather\_df[self.weather\_df[MONTH] == month][GW].mean()

        user\_sample = pd.DataFrame([[car\_type, month, day, period]], columns=CD)

        user\_sample\_good = pd.concat([user\_sample, self.weather\_df.loc[weather\_id\_good, WEATHER].to\_frame().transpose().set\_index(user\_sample.index)], axis=1)

        user\_sample\_bad = pd.concat([user\_sample, self.weather\_df.loc[weather\_id\_bad, WEATHER].to\_frame().transpose().set\_index(user\_sample.index)], axis=1)

        self.save\_clf\_quality(X\_test, y\_test)

        # Посчитаем полную вероятность для клиента в случае разных погодных условий

        proba = good\_weather\_proba \* self.clf.predict\_proba(user\_sample\_good) + (1 - good\_weather\_proba) \* self.clf.predict\_proba(user\_sample\_bad)

        return proba[0, 1]

    def plot\_graph(self, path, months, days\_of\_week, time\_periods):

        fig, ax = ox.plot\_graph\_route(self.city\_graph, path, route\_linewidth=3, route\_color='b', node\_size=0, bgcolor='k', show=False, close=False)

        self.raw\_data[MONTH] = pd.to\_datetime(self.raw\_data['CRASH DATE']).dt.month

        self.raw\_data[DOW] = pd.to\_datetime(self.raw\_data['CRASH DATE']).dt.dayofweek

        self.raw\_data[TP] = pd.to\_datetime(self.raw\_data['CRASH TIME']).dt.hour.apply(lambda x: TIMES[x])

        for m in months:

            for d in days\_of\_week:

                for t in time\_periods:

                    accidents = self.raw\_data[(self.raw\_data[MONTH] == m) & (self.raw\_data[DOW] == d) & (self.raw\_data[TP] == t)]

                    add\_accidents(ax, accidents[LON], accidents[LAT], clr='r')

    def save\_clf\_quality(self, X\_test, y\_test):

        min\_accuracy = y\_test.sum()/y\_test.shape[0]

        y\_pred = self.clf.predict(X\_test)

        s1 = f'Min expected accuracy: {1 - min\_accuracy} - Model result is {accuracy\_score(y\_test, y\_pred)}\nModel predicts crash for {1 - precision\_score(y\_test, y\_pred, zero\_division=0)} cars without crashes\nModel found {recall\_score(y\_test, y\_pred)} crashes'

        s2 = f'Best model iteration is: {self.clf.get\_best\_iteration()}\n'

        self.model\_quality += '\n'.join([s1, s2])

    def get\_clf\_quality(self):

        return self.model\_quality

Листинг 1.6. Код класса предсказывающего алгоритма.

# ПРИЛОЖЕНИЕ 2. JUPYTER NOTEBOOKS

Все данные, используемые в работе и блокноты сохранены здесь:

<https://github.com/IgorLukhnev/Designing-New-Car-Insurance-Product>