Федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»

Факультет компьютерных наук
Основная образовательная программа
«Прикладная математика и информатика»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

(Исследовательский проект) на тему

Предсказание образовательной и жизненной траектории молодых людей на данных ТрОП

Выполнил студент группы БПМИ 176, 4 курса,
Шабалина Анастасия Владимировна

Руководитель ВКР:
Преподаватель, Сивак Елизавета Викторовна

Москва 2021

Оглавление

Аннотация

Данная работа включает в себя изучение, обработку и использование данных лонгитюдного исследования ТрОП для предсказания жизненной траектории молодых людей. В данном проекте будут использоваться данные панельного исследования “Траекторий в образовании и профессии” (ТрОП). Всего в данном исследовании приняли участие 4893 ученика 8-х классов в 210 школах в 42 регионах Российской Федерации.

Предполагается использовать различные методы машинного обучения, чтобы предсказать планы участников исследования на рождение детей. Большинство исследований в данной сфере социологии обычно направлено на обнаружение связей между различными переменными в данных. Задача предсказания каких-либо жизненных событий не настолько изучена, а в исследованиях на эту тему модели не давали достаточно высокого качества. Поэтому в данной работе я планирую использовать и проверить различные модели, чтобы после этого проанализировать полученные результаты.

This paper includes the analysis, preprocessing, and using data from TrEC's longitudinal research to predict young people's life trajectories. This work includes data from the Russian panel study “Trajectories in Education and Careers” (TrEC). Support from the Basic Research Program of the National Research University Higher School of Economics is gratefully acknowledged. In total, 4,893 8th-grade students in 210 schools in 42 regions of the Russian Federation took part in this study.

Various machine learning methods are supposed to be used to predict the plans for having children of the study participants. Most research in this area of sociology usually focuses on discovering relationships between different variables in the data. The task of predicting any life events is not that well understood, and in studies on this topic, models have not performed well enough. Therefore, in this paper I plan to use and test different models in order to analyze the results afterwards.

Список ключевых слов: машинное обучение, анализ данных,

предсказательные модели, отложенное деторождение, Российская Федерация, исследование опросов, социальные науки

1. Введение

Данная выпускная квалификационная работа будет связана с использованием методов машинного обучения в области социологии для прогнозирования планов респондентов на рождение детей. В рамках данной работы я буду работать с данными исследования “Траектории в образовании и профессии” [1]. Данный проект объединяет несколько лонгитюдных исследований (научный метод, при котором одна и та же группа объектов изучается в течение длительного времени, за которое эти объекты успевают существенно изменить какие-либо свои значимые характеристики), дополненных данными из интервью, посвященных изучению траекторий в образовании, переходящих в траектории на рынке труда, и проводится с 2009 года.

Объектом изучения являлись учащихся 8 классов в 2010 - 2011 годах в Российской Федерации. Ученики принимали участие в тестированиях и анкетировании. Кроме этого, исследователи опрашивали учителей и родителей учащихся. В данном исследовании одни и те же респонденты опрашиваются несколько раз через определенные промежутки времени (примерно один раз в 1-2 года), что дает возможность проследить профессионально-образовательные траектории на индивидуальном уровне. Последняя волна опросов была проведена осенью 2017 года, когда респонденты были студентами 4 курса ВУЗов или выпускниками заведений СПО (среднее профессиональное образование).

Моими задачами в данной работе будут изучение и обработка данных опросов исследования для лучшего извлечения информации, отбор независимых переменных, предсказание зависимых переменных (планы на рождение детей), используя модели машинного обучения различной сложности, анализ полученных результатов при помощи метрик, определение модели, показывающей лучшее качество и обоснование полученных результатов.

Для изучения будут использоваться переменные, которые говорят о том, планируют ли респонденты рождение детей в ближайшем будущем и когда. Я

буду изучать, можно ли, и если можно, то насколько точно возможно предсказать эти планы по другой известной об участниках исследования.

Планируется использовать такие методы, как логистическая регрессия, деревья решений, случайный лес, бустинг и другие.

Данная работа может помочь определить причины, по которым люди «откладывают» взросление, например откладывают отъезд из родительского дома, замужество или женитьбу и рождение детей. И также, соответственно, работа может помочь в устранении данных причин или уменьшить их влияние.

Кроме этого различные попытки улучшить прогнозные характеристики могут помочь развитию теории и методов в данной сфере.

Ранее в Принстонском и Колумбийском университетах проводилось похожее исследование “Fragile Families Challenge” [2], в рамках которого изучались результаты опросов примерно 5000 детей, родившихся в крупных американских городах в период с 1998 по 2000 год, с большим количеством детей, рожденных от неженатых родителей. При помощи данных опросов и различных методов машинного обучения ученые пытались спрогнозировать различные переменные, такие как: средний балл ребенка (GPA), упорство ребенка, выселение из семьи, материальные трудности домашнего хозяйства, увольнение основного опекуна, и участие основного опекуна в профессиональном обучении.

Оказалось, что даже самые лучшие прогнозы были не очень точными. Также, стало ясно, что лучшие прогнозы, которые были сделаны при помощи сложных методов машинного обучения и использовали большое количество данных (тысячи переменных), не сильно отличались в лучшую сторону от прогнозов простых моделей, таких как линейная и логистическая регрессия, которые использовали всего 4 переменные.

В своей работе я провела подобные эксперименты, чтобы спрогнозировать выбранные переменные и использовала различные модели машинного обучения, в которых получила достаточно точные предсказания. Также, я изучила, имеет ли влияние на качество моделей ручной отбор независимых признаков для модели по социологическим факторам.

В разделе "Обзор литературы" будет проведен обзор и анализ основной литературы и исследований, связанных с данным проектом. В разделе "Анализ и предобработка данных" будут описаны основные методы, использованные в проекте для обработки предоставленных данных и отбора независимых переменных. Также будет обоснован выбор этих методов. В разделе "Проведение экспериментов и их анализ" будут описаны модели, которые были использованы для проведения экспериментов, а также результаты данных экспериментов и их анализ. А в разделе "Заключение" будет описано сравнение полученных результатов с целью и задачами данного проекта. Также в этом разделе будут описаны дальнейшие перспективы исследования.

2. Обзор литературы

В данной работе я планирую изучить, какие факторы влияют на желание людей позднее заводить семью и детей по сравнению с другими поколениями. В своей работе 2000 года [3] Джеффри Арнетт ввел понятие «emerging adulthood» (отложенное взросление – люди живут долго с родителями, откладывают замужество\женитьбу и рождение детей). И целью данного обзора литературы является поиск и изучение причин и факторов, которые обуславливают данное «отложенное взросление», а также методов, которые позволят предсказать по предоставленным данным, кто из респондентов будет откладывать рождение детей, а кто нет.

2.1. Факторы, влияющие на отложенное рождение детей

Я изучила несколько статей, в которых исследовались причины «отложенного взросления» людей и их нежелание заводить детей в более раннем возрасте.

Например, в исследовании Мелинды Миллс 2001 года [4] говорится, что основными причинами позднего рождения ребенка являются появление противозачаточных средств и более высокий уровень образования у женщин. В своем детальном эконометрическом анализе Голдин и Кац (2002) [5] продемонстрировали, как распространение оральных контрацептивов в конце 1960-х годов в США привело к отсрочке возраста первого вступления в брак для женщин с высшим образованием. Получение противозачаточных средств позволило им дольше оставаться в сфере образования, инвестировать в долгосрочную карьеру на рынке труда и избегать беременности, будучи сексуально активными. Но, так как в Южной и Восточной Европе в это время были распространены менее надежные методы контрацепции, а рождаемость в этих регионах все равно понизилась, то следует, что есть и другие факторы, связанные с откладыванием рождения детей.

Также, по данным различных исследований, само повышение женщинами своего уровня образования тоже влияет на более позднее рождение детей, так как и обучение, и материнство занимает много времени, и женщина, которая

желает получить образование, а потом и развивать карьеру, будет откладывать рождение ребенка. Но при различных условиях это влияние уменьшалось или даже совсем сходило на нет. Например, в Норвегии данное влияние исчезло из-за улучшения государством доступа к детским садам (т.е. мать может оставить ребенка в детском саду и продолжить учиться или работать).

В своем исследовании в 2010 году Ян Ван Бавель [6] обнаружил, что женщины, окончившие обучение по дисциплине, где преобладают стереотипные семейные отношения, значительно реже откладывают свои первые роды. Также оказалось, что чем больше женщин доминирует в данной сфере образования, тем меньше выпускники склонны откладывать материнство. Кроме этого, по результатам данного исследования, женщины, получившие образование в области, где выпускники, как ожидается, будут иметь относительно высокий заработок на момент выхода на рынок труда, значительно чаще откладывают рождение первого ребенка, по сравнению с женщинами, ожидающими иметь низкий доход в течение первых лет своей работы. Как и, если женщина имеет образование в области, где дополнительный год на рынке труда связан с сильным увеличением ежемесячного заработка, то ожидается, что она отложит первые роды на большее время (в исследовании показывается, что эффект крутизны профиля заработка сильнее, чем эффект начальной заработной платы).

Также, так как со временем сожительство перед вступлением в брак и несколько последовательных партнеров и отношений стало нормой, то этот фактор тоже повлиял на задержку рождения детей (вероятность того, что люди заведут ребенка во время совместного проживания, значительно меньше, чем у женатых людей). Кроме этого, трудности в поиске партнера могут также способствовать задержке рождаемости у женщины. Но имеет место быть и обратная связь: женщины в неравноправных обществах (например, в Японии) иногда наоборот избегают замужества, так как не хотят, чтобы их принуждали к материнству и увольняли с работы.

И последним из факторов в данном исследовании указывается жилищная и экономическая неопределенность. Данный фактор может как и оттягивать

рождение первого ребенка, так и подталкивать женщину к рождению детей (в зависимости от других условий) . Например, в исследовании Микаэлы Крейенфельд (2010) [7] говорилось, что связь между экономической неопределенностью и первым рождением варьируется в зависимости от уровня образования. Если более высокообразованные женщины откладывают материнство, когда сталкиваются с неопределенностью в сфере занятости, то женщины с низким уровнем образования чаще становятся матерями в данных ситуациях.

Рынок жилья также является примером социального структурного фактора, который может непреднамеренно влиять на возраст рождения первого ребенка. Например, в Италии для взятия ипотеки требуется большие авансовые платежи, в отличие от Нидерландов. В этих условиях молодым итальянцам труднее приобрести дом, и это, несомненно, приводит к отсрочке отцовства. В странах, где легче получить ипотеку или выйти на рынок государственной аренды, люди более способны утвердиться и вступить в семью раньше.

В достаточно большом количестве исследований экономическая неопределенность (безработица и нестабильная ситуация на рынке труда) связана с отсрочкой первых родов из-за неспособности будущих родителей брать на себя подобные долгосрочные обязательства.

Влияние экономической неопределенности на отсрочку также зависело от того, существовала ли более сильная система социальной защиты для защиты людей от экономической неопределенности. В странах с сильной системой социальной защиты, таких как Швеция и Норвегия, наблюдалось значительно более слабое влияние данного фактора на откладывание первых родов [8]. Безработица и обеспокоенность безопасностью своей работы могут также побудить женщин с более высоким образованием отложить первые роды, а менее образованных женщин - наоборот стать матерями.

Изучая тенденции изменения рождаемости в Швеции в 1980-х и 1990-х годах, Андерссон показал [9], что женщины, занимающие должности с низким уровнем дохода, и студенты имели более низкую и отложенную рождаемость. Однако, согласно соответствующим исследованиям, он делает вывод, что на

рождаемость влияют не только отдельные факторы, но и важные социальные факторы и, в частности, социальная политика, к которой мы сейчас обратимся.

В Венгрии обнаружили [10], что резкие изменения системы выплат семейного пособия с универсальной системы на систему проверки нуждаемости оказали влияние на появление первого ребенка в семье. Люди с высшим образованием и доходом лишились права на получение льгот и, следовательно, отложили рождение детей.

Кроме этого, в другом исследовании [11] было показано, что если матери родили своих дочерей в раннем возрасте, то дочери тоже, с большей вероятностью, родят своего первого ребенка в раннем возрасте. Кроме этого, была показана обратная зависимость между уровнем образования родителей ребенка и возрастом, в котором он заведет своих детей. Также было замечено, что доход родителей человека повышает возраст появления у него самого детей. Отсюда можно сделать вывод, что в образованных семьях с высоким экономическим и социальным статусом появление детей вытесняется потреблением прочих благ, и данная философия передается из поколения в поколение.

2.2. Методы машинного обучения для предсказания переменных

В данной работе планируется понять, можно ли предсказать планирует ли человек заводить детей в ближайшем будущем и насколько точно получится это предсказать. Как я уже описала выше, ранее проводилось исследование, подобное “Траектории в образовании и профессии” [1] - “Fragile Families and Child Wellbeing Study” [12]. “Fragile Families and Child Wellbeing Study” это лонгитюдное исследование, которое проводилось в тысячах американских семей более 15 лет. За это время в рамках данного исследования была собрана информация о детях, их родителях, школах и их более широком окружении. Сбор исходных данных проводился в период с 1998 по 2000 год. В него входили интервью с обоими биологическими родителями вскоре после рождения ребенка.

Последующие волны опросов проводились, когда ребенку было 1, 3, 5, 9 и 15 лет. Помимо собеседований с родителями, на третьем, пятом и девятом годах исследования проводилась оценка ухода за детьми, а также интервью с ребенком. Основные цели исследования - узнать о способностях и взаимоотношениях родителей, не состоящих в браке, а также о том, как с течением времени живут семьи и дети с учетом различных медицинских и социальных показателей.

Эти данные опросов были использованы в сотнях научных статей и десятках диссертаций для определения зависимостей между различными переменными. Но участникам конкурса “Fragile Families Challenge” [2] было предложено использовать эти данные по-новому. Им предлагалось предсказать шесть ключевых результатов из данных опроса 15-го года исследования. Для решения данной задачи они могли использовать все исходные данные от рождения до 9-го года исследования и некоторые данные за 15-й год.

В итоге, результаты данного конкурса показали, что либо удача играет значимую роль в жизни людей, либо социологические теории упускают некоторые важные переменные. Так как лучшие модели (которые показывали лучшее качество) не сильно превосходили предсказания простых моделей, которые использовали небольшое количество данных. К тому же все модели предсказывали недостаточно точно, независимо от количества данных, которые они использовали.

Д. Стэнеску, Эрик Х. Ванг и С. Ямаучи сделали лучшее статистическое предсказание [13] в этом конкурсе. Они обнаружили множество вопросов, на которые респонденты не ответили. Это обстоятельство затрудняло поиск значимых прогностических переменных. Поэтому они объединили традиционные методы с методом LASSO, в результате чего было найдено 339 значимых переменных. Затем они вновь запустили LASSO, что дало им более точный прогноз финансовых трудностей 15-летних подростков.

Авторы данной работы использовали метод LASSO для предварительной обработки данных. Затем они применили алгоритм Амелии [14] для заполнения недостающих данных. И потом, снова использовали LASSO для составления

прогнозов на основе заполненных данных. Средняя квадратичная ошибка данной команды составила 0,019 (самый низкий показатель среди всех участников конкурса Fragile Families Challenge). Они увидели, что, судя по результатам работы, ответы матерей были более полезны для прогнозирования материальных трудностей.

В работе Карнеги и Ву 2019 года [15] для автоматического отбора независимых переменных были использованы методы LASSO, BGLM (Bayesian generalized linear models), and BART (Bayesian additive regression trees). Заменяли пропуски в столбцах медианой для непрерывных признаков и модой для категориальных признаков (также предлагали применить более сложные, например hot-deck) . А для предсказания также использовалась модель BART. В данном исследовании делается вывод, что выбор метода для отбора независимых признаков не сильно влияет на конечный результат (возможно из-за того, что было большое количество потенциальных предикторов, многие из которых, скорее всего, мало связаны с прогнозируемыми переменными).

В другом исследовании 2019 года [16] для отбора признаков они используют параллельно 2 метода: LASSO и Mutual information. Для предсказания переменной параллельно использовались elastic net, random forest и градиентный бустинг. Но даже самый эффективный вариант показал результат менее чем на 20 процентов лучше, чем простой бейзлайн.

Также, в работе Клаудии В. Робертс (2019) [17] исследовали автоматический отбор признаков и ручной отбор по социологическим факторам. Исследователи изучили различные варианты моделей для предсказания (Ridge регрессия, Elastic net, LASSO, дерево решений и другие), 2 стратегии заполнения пропусков (медиана и мода), а также применяли стандартизацию. Лучшие результаты показали модели LASSO и Elastic net с заполнением пропусков медианами и использованием стандартизации. Далее исследователи использовали вручную отобранные переменные для лучших 10-ти моделей с автоматическим отбором признаков и показали, что ручной отбор переменных улучшил большинство из этих моделей.

2.3. Выводы

Исходя из проведенного обзора литературных источников для отбора независимых переменных я буду использовать L1-регуляризацию, так как этот метод положительно оценивается в исследованиях с подобными задачами и данными. Для заполнения пропусков можно будет использовать медиану. А для предсказания целевых переменных я проведу эксперименты с различными моделями, такими как логистическая регрессия или деревья решений (Decision Tree, Random Forest, градиентный бустинг).

Изучение социологических факторов, влияющих на планы на рождение детей я буду использовать при ручном отборе признаков, что также может улучшить качество моделей. Например, один из значимых факторов - это уровень образования. Также, может сильно влиять на планы на рождение ребенка экономическая нестабильность человека, что также будет использовано при ручном отборе признаков. Но, так как в большинстве изученных мною статей на эту тему проводятся корреляционные исследования с парой определенных факторов. Поэтому в данной работе планируется понять можно ли предсказать подобные факторы (планы на рождение детей) и насколько точно получится их предсказать.

3. Анализ и предобработка данных

3.1. Анализ основных переменных

Данные ТрОП состоят из 8-ми отдельных файлов (по одному на каждую волну опросов соответственно). 1-ой зависимой переменной является столбец iQ49 - “Вы бы хотели или не хотели иметь детей в будущем?” с 6-ю категориями:

1. определённо хотел(-а) бы
2. скорее хотел(-а) бы
3. скорее не хотел(-а) бы
4. определённо не хотел(-а) бы
5. я пока не задумывался(-лась) об этом
6. я сейчас жду ребёнка

Поэтому я буду решать задачу многоклассовой классификации с 6-ю классами. Я буду использовать данные только последних 4-х волн, так как выборка респондентов в каждой волне немного отличается, и при пересечении всех волн сразу остается слишком мало данных. В итоге у меня остается 2725 строк и 1429 столбцов с данными.

Я изучила основные переменные: целевую переменную - планы на рождение детей, и независимые переменные - например, пол, семейное положение и размер дохода респондентов, а, также, их образование (учатся ли они?) и работу (работают ли?).

Также, я изучила взаимосвязи между зависимой переменной и различными основными независимыми переменными. Далее будут изображены гистограммы распределений нескольких основных переменных и гистограммы для некоторых взаимосвязей между переменными.

3.2. Обработка данных

Для начала я исключаю из полученных данных переменные, которые не представляют аналитического интереса. Например, различные вычисленные веса выборки, открытые варианты ответов на вопросы, переменные с нулевой

дисперсией и т. п. Далее я разделяю все признаки на категориальные и числовые. Одним из важных этапов предварительной обработки является заполнение пробелов в данных. Я заполняю их средним значением или медианой для числовых признаков и объединяю все информативных пробелы в одну новую категорию для категориальных переменных (также, можно заполнить пропуски модой). Также, я нормирую числовые переменные для того, чтобы учесть их масштаб. Кроме этого, я попробую поэкспериментировать с данными, в которых я отбросила все признаки, где слишком много пропусков (больше 80 процентов). После чего применяю binary encoding. В данном случае one-hot encoding нежелательно использовать, так как в предоставленных мне данных достаточно много независимых признаков с большим количеством категорий, и при использовании данного метода обработки получится матрица разреженная матрица большого размера .

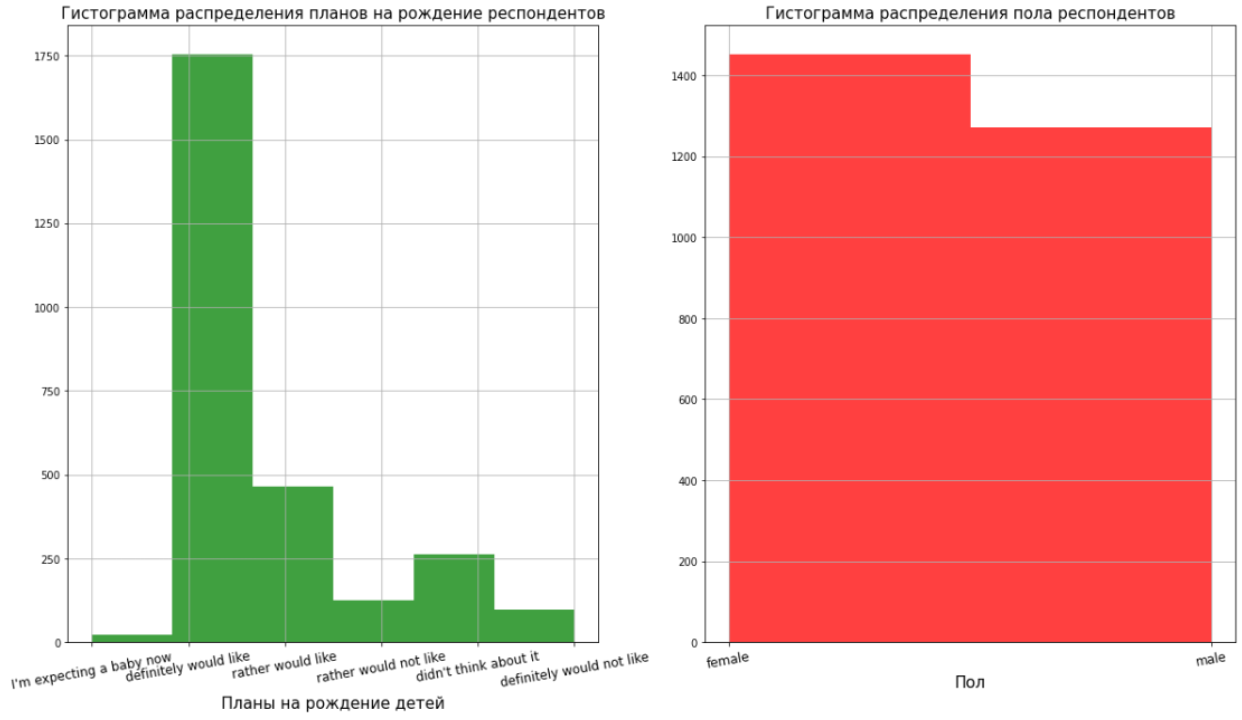

Рис. 3.1. Гистограммы распределений зависимой переменной и переменной "Пол респондента"

Для сужения набора независимых признаков для предсказания я использую L1-регуляризацию (логистическая регрессия и метод опорных векторов). Далее использую методы feature importance или permutation feature importance. После L1-регуляризации у меня остается 300-400 признаков,

которые я в дальнейшем буду использовать в моделях.

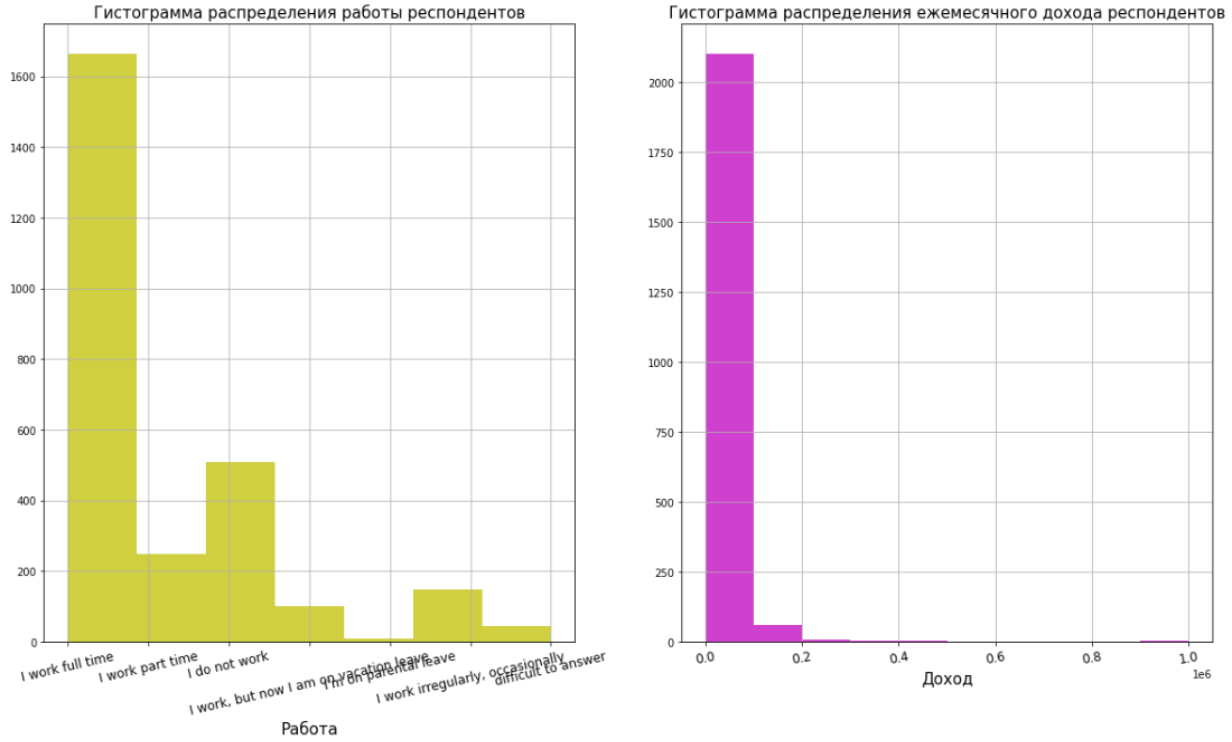

Рис. 3.2. Гистограммы распределений переменных “Работа” и “Доход”

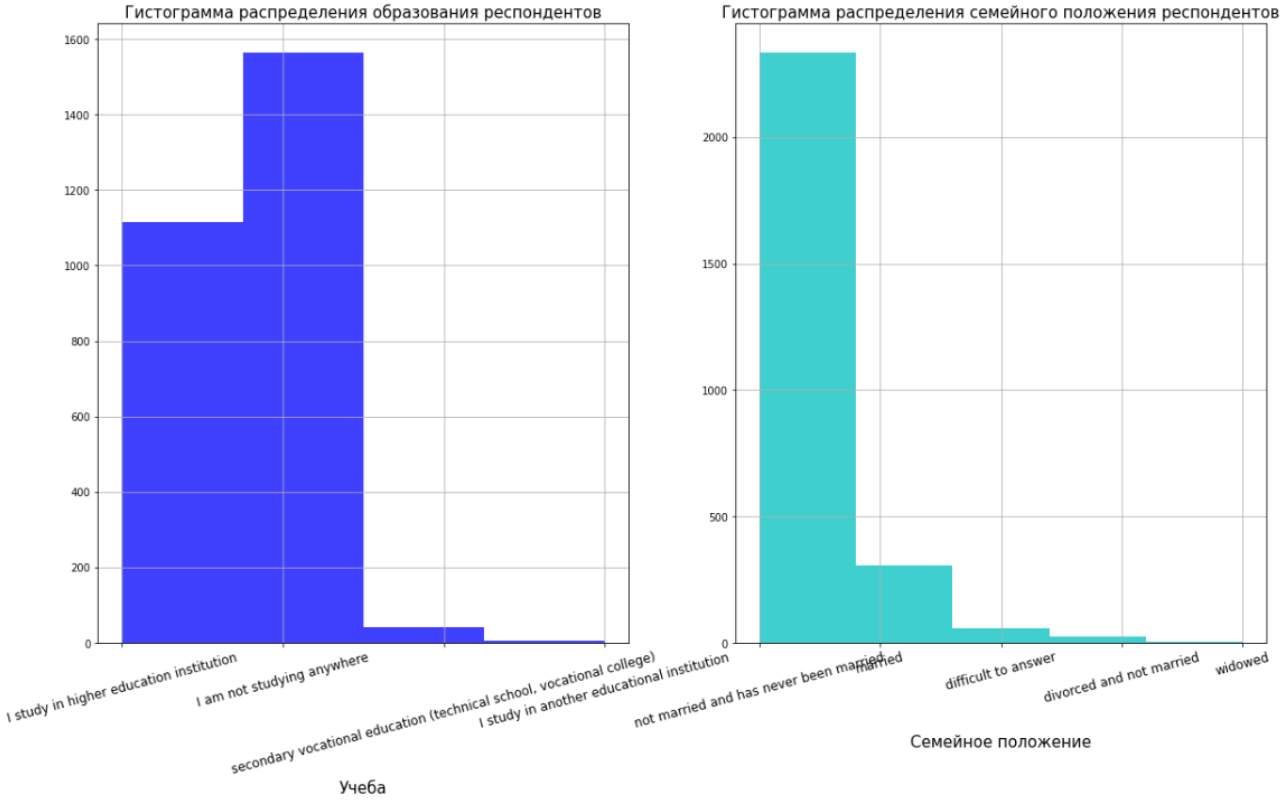

Рис. 3.3. Гистограммы распределений переменных “Учеба” и “Семейное положение”

Также, для экспериментов я дополнительно вручную отобрала 185 признаков [18], которые могут влиять на зависимую переменную, в соответствии с изученными на эту тему исследованиями, для сравнения двух методов отбора переменных применительно к данной задаче.

3.3. Выводы и результаты

В результате предобработки данных ТрОП у меня появилось несколько наборов данных за счет различных методов заполнения пропусков, выбора независимых признаков и т. п. В следующей главе я расскажу про проведение экспериментов с различными моделями и полученными датасетами. После этого мы проанализируем результаты экспериментов и определим наиболее подходящую модель машинного обучения для данной задачи.

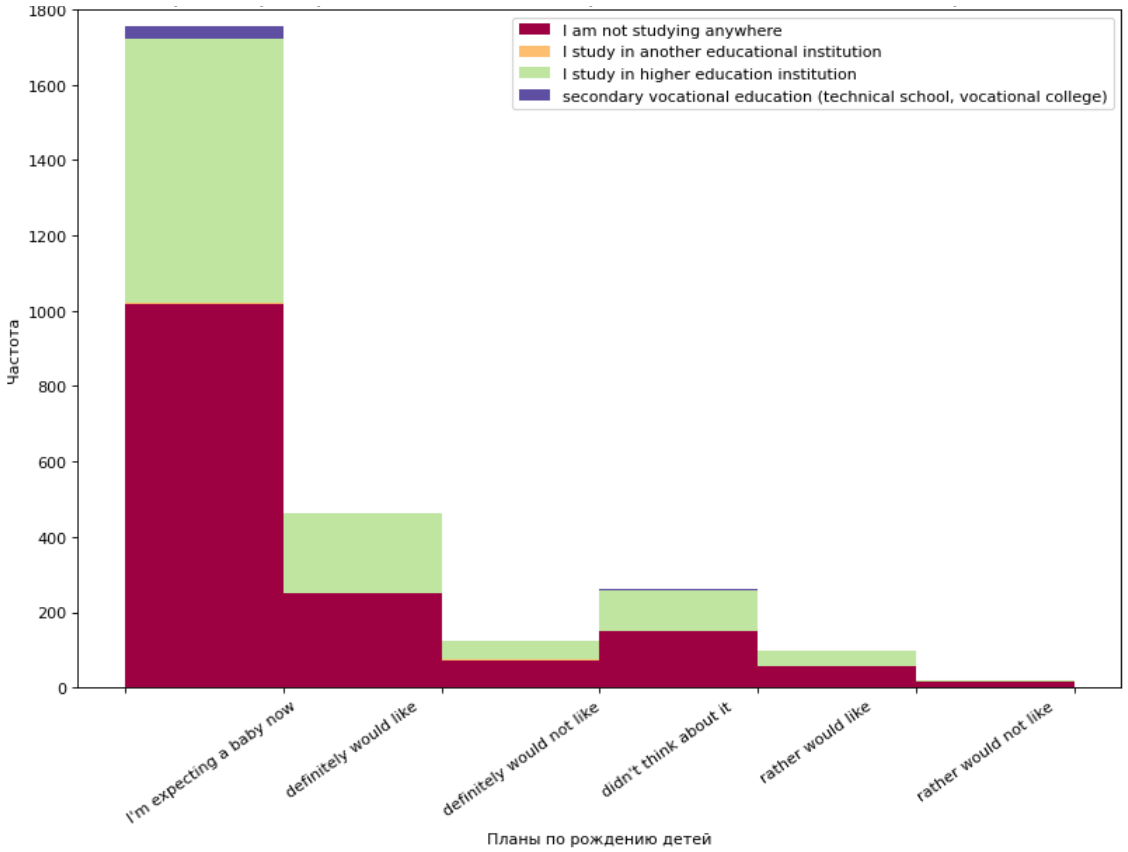

Рис. 3.4. Гистограмма распределения целевой переменной в связи с образованием респондентов

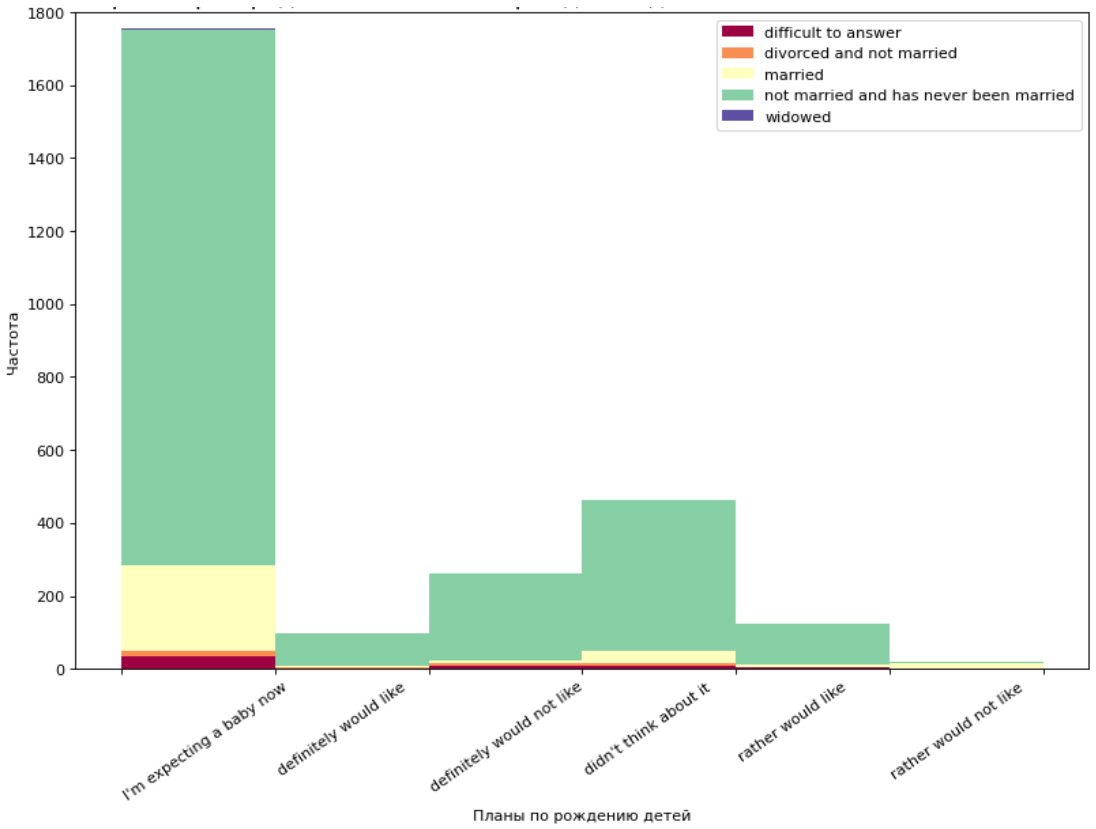

Рис. 3.5. Гистограмма распределения целевой переменной в связи с семейным положением респондентов

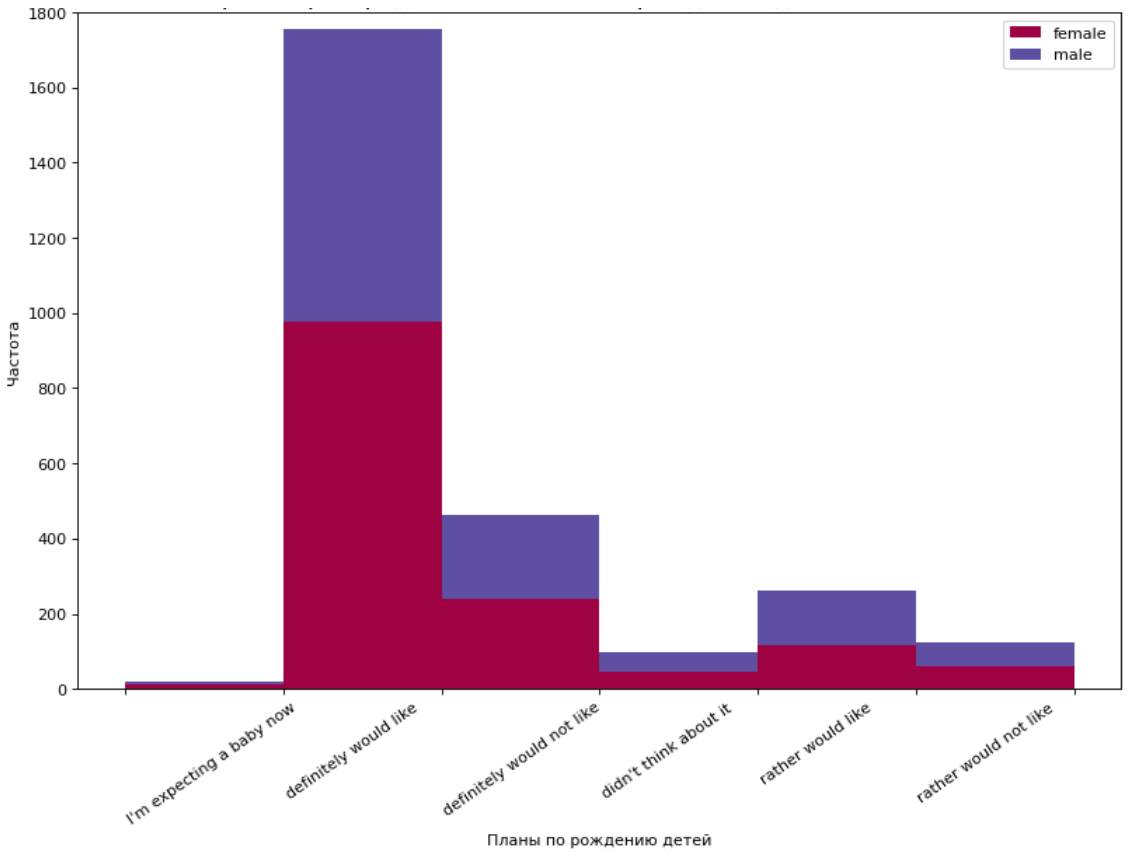

Рис. 3.6. Гистограмма распределения целевой переменной в связи с полом

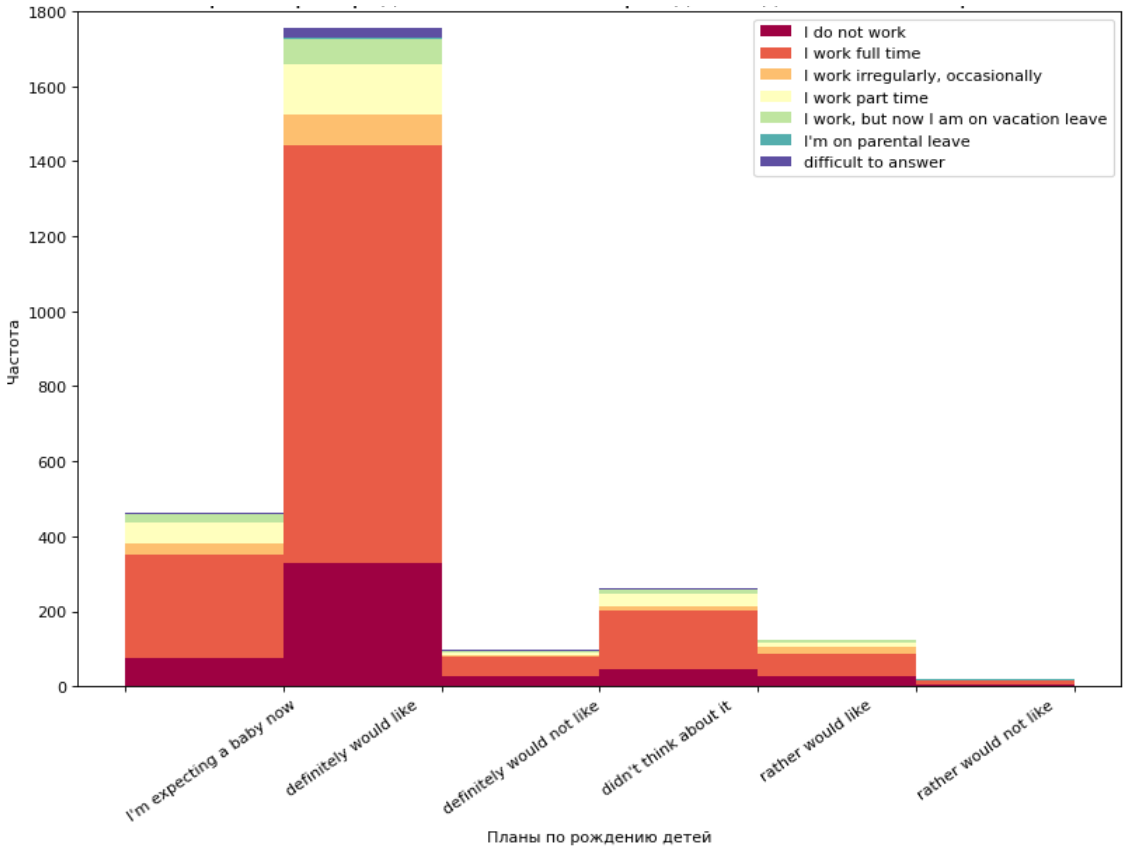

Рис. 3.7. Гистограмма распределения целевой переменной в связи с работой респондентов

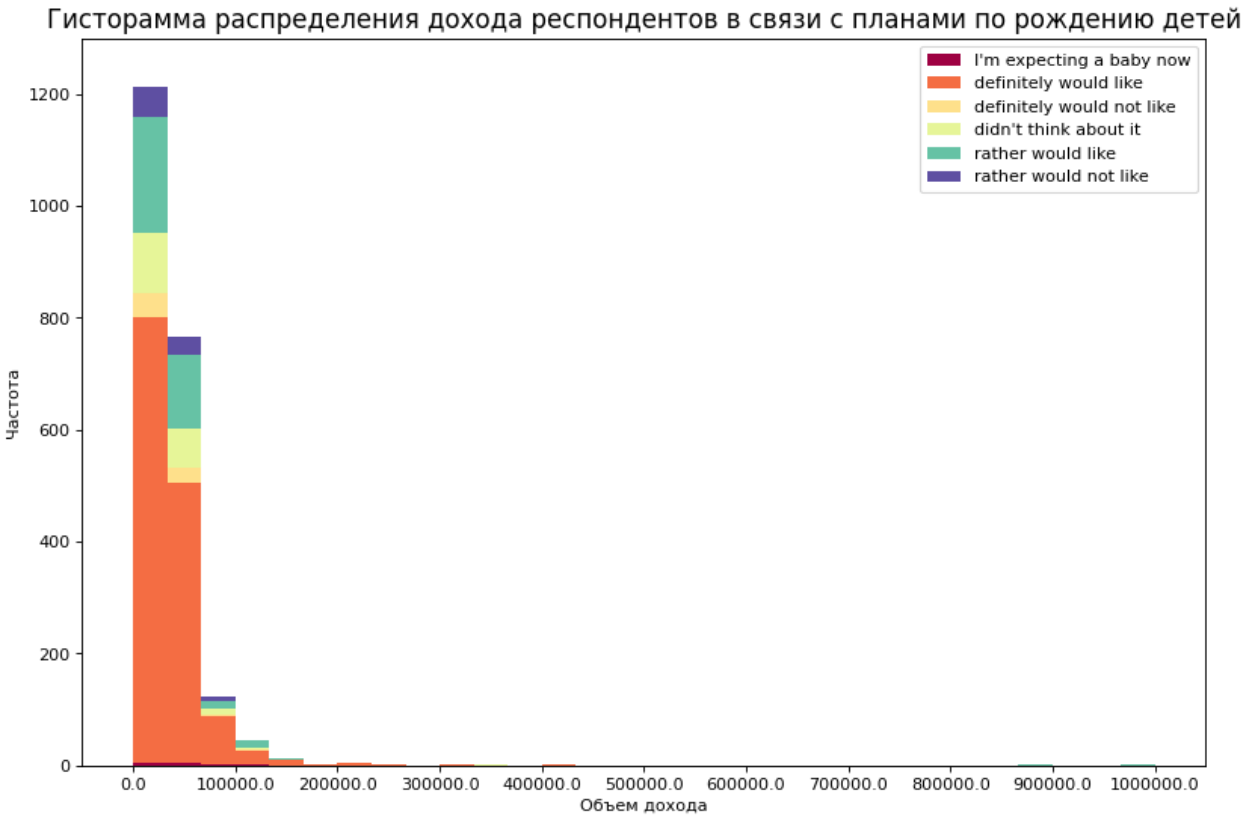

Рис. 3.8. Гистограмма распределения общего дохода респондента в связи с его планами по рождению детей

4. Проведение экспериментов и их анализ

4.1. Эксперименты для 1-ой целевой переменной (iQ49)

В качестве моделей для предсказания зависимой переменной я использую логистическую регрессию, Decision Tree, Random Forest, градиентный бустинг (различные его версии) [18]. После чего я сравню результаты моделей. В качестве метрики для оценивания качества моделей я использую кросс-валидацию (с 10-ю подвыборками) и считаю accuracy score по всем подвыборкам. В качестве baseline-модели я использую логистическую регрессию, но с маленьким набором переменных, связанных с уровнем образования респондентов, т. к. в исследованиях данный фактор обычно влияет на планы на детей больше всего. Для baseline я взяла признаки:

1. Corrected_iQ1 - учится ли респондент на данный момент
2. iQ3 - на каком курсе учится респондент
3. iQ4 - в какой стране учится респондент
4. iQ10 - респондент учится на бюджете или платно
5. iQ11 - форма обучения (очно, заочно, дистанционно и т. п.)
6. Corrected_iQ13 - законченный уровень образования респондента
7. iQ14_n - собирается ли респондент продолжать обучение (n бинарных переменных)

Accuracy score для baseline-модели получился 0.643.

Логистическая регрессия улучшила результаты для различных наборов данных (и с binary encoding, и с нормировкой признаков), но недостаточно сильно (среднее значение accuracy равно 0.645).

Методов обработки данных binary encoding улучшил качество больше всех остальных методов. Но можно заметить, что для более простых моделей (DecisionTree и RandomForest) предпочтительнее использовать ручной отбор признаков, а для более сложных моделей (GradientBoosting (sklearn), XGBoost, LightGBM) лучше использовать binary encoding. Метод feature scaling, в свою очередь, не показал каких-либо улучшений качества в использовании с простыми моделями и немного повысил качество для моделей GradientBoosting

(sklearn) и XGBoost.

Лучшие результаты в этой серии экспериментов показали для обычного датасета и датасета с binary encoding показал LightGBM и XGBoost (0.772 и 0.77 соответственно)

Таблица 4.1. Выборочные результаты проведенных экспериментов для зависимой переменной "iQ49"

Accuracy	**Decision Tree Classifier**	**Random Forest**	**Gradient Boosting (sklearn)**	**XGBoost**	**LightGBM**
Without additional processing	0.747	0.744	0.755	0.758	0.759
Feature scaling	0.747	0.745	0.761	0.764	0.760
Binary encoding	0.731	0.749	0.769	0.77	0.772
Hand-selected	0.748	0.751	0.754	0.75	0.755

4.2. Эксперименты для 2-ой целевой переменной (iQ50)

Также, я построила модели для другой целевой переменной "iQ50", так как данная переменная тоже связана с "отложенным" взрослением и рождением детей. Данная переменная обозначает вопрос "Когда Вы хотели бы завести первого ребенка?" с 8-ю категориями ответа:

1. в ближайшие год-два
2. в течение трёх лет
3. через 4-5 лет
4. через 6-8 лет
5. через 9-11 лет
6. через 12-15 лет

7. через 16 лет и позже
8. пока не знаю, не думал(-а) об этом

Я, также, построила для нее гистограмму распределения:

Рис 4.2. Гистограмма распределения переменной "Когда Вы хотели бы завести первого ребенка?"

Я превратила данную переменную в бинарную, объединив категории "в ближайшие год-два" и "в течение трех лет" в один класс, а все остальные - в другой. Таким образом, размеры классов стали практически одинаковые (1094 и 1125).

Далее я провела такую же предобработку данных, как и для переменной "iQ49", так как датасет одинаковый, и эти переменные немного коррелируют. В качестве baseline-модели я также взяла логистическую регрессию, с теми же признаками, связанными с образованием. Accuracy score для данной baseline-модели получился 0.562.

Таблица 4.3. Выборочные результаты проведенных экспериментов для зависимой переменной “iQ50”

Accuracy	**Logistic Regression**	**Decision Tree Classifier**	**Random Forest**	**Gradient Boosting (sklearn)**	**XGBoost**	**LightGBM**
Without additional processing	0.716	0.664	0.707	0.725	0.728	0.732
Feature scaling	0.714	0.664	0.708	0.734	0.734	0.733
Binary encoding	0.782	0.691	0.731	0.747	0.749	0.748
Hand-selected	0.698	0.675	0.699	0.709	0.711	0.711

В этом случае логистическая регрессия показала наилучший результат среди всех вариантов экспериментов (в эксперименте с binary encoding - 0.782).

Метод DecisionTreeClassifier показал худшие результаты (по сравнению с другими моделями) для всех вариантов методов обработки данных.

В случае с данной переменной, можно увидеть, что метод binary encoding улучшил результаты для всех моделей. А feature scaling в этот раз улучшил результаты для бустингов и никак не повлиял на более простые модели. Ручной отбор признаков в этот раз не показал ожидаемых улучшений качества для какой-либо модели.

Среди бустингов методы XGBoost и LightGBM показали примерно одинаковые результаты, и немного выше, чем результаты метода GradientBoosting (из библиотеки sklearn).

4.3. Выводы и результаты

Поскольку в ранее описанных аналогичных исследованиях модели не давали точных предсказаний, я ожидала как положительных, так и отрицательных результатов экспериментов. Но некоторые модели показали достаточно хорошие результаты (в большинстве случаев это бустинги XGBoost и LightGBM). Следовательно, над этой задачей, в перспективе, можно будет провести больше экспериментов и с более сложными моделями машинного обучения (например, нейросетями). Также стоит сказать, что при использовании ручного метода отбора переменных я не увидела достаточно сильного улучшения качества, и моя гипотеза, что ручной отбор может помочь в предсказании переменных (так как в некоторых изученных мною источниках было представлено такое мнение), была опровергнута. Возможно нужно было учитывать другие социологические факторы, так как я выбирала в основном по 2 факторам: образование и экономическая нестабильность.

5. Заключение

В результате данной выпускной квалификационной работы я изучила предоставленные мне данные ТрОП и проанализировала результаты исследований и статей в которых работали с подобными данными. Я обработала данные для лучшего прогнозирования моделями (заполнение пустых ответов в опросах, преобразование и создание новых признаков) и отобрала переменные, которые влияют на целевую переменную больше всего. После этого я провела эксперименты для двух выбранных целевых переменных, связанных с “отложенным” рождением детей, с различными методами машинного обучения и определила, какие из них наиболее подходят для данной задачи предсказания. Среди изученных мною моделей лучшие результаты в большинстве своем показали модели градиентного бустинга (XGBoost и LightGBM).

Проделанная мною работа может помочь продвинуться в изучении и использовании подобных данных, полученных в ходе лонгитюдных исследований. Также это может помочь в определении причин, из-за которых люди «откладывают» взросление.

В качестве дальнейшего развития данного проекта можно попробовать более сложные алгоритмы, например использовать нейросети. Кроме этого, можно поэкспериментировать с более сложными методами заполнения пропусков (таких как hot-deck imputation и K-Nearest Neighbour) и отбора независимых переменных. Также, можно попробовать объединить эти данные с другими подобными исследованиями (TIMSS, PISA) для получения более подробной информации о респондентах, и предсказать новые переменные не обязательно связанные с “отложенным” взрослением.

Список источников

1. Траектории в образовании и профессии [Электронный ресурс]. – Режим доступа: http://trec.hse.ru/, свободный. – (дата обращения: 16.05.2021).
2. Fragile Families Challenge [Электронный ресурс]. – Режим доступа: https://www.fragilefamilieschallenge.org/, свободный. – (дата обращения: 16.05.2021).
3. Arnett, J. J. Emerging adulthood: A theory of development from the late teens through the twenties // American Psychologist. 2000. №55(5). C. 469-480.
4. Why do people postpone parenthood? Reasons and social policy incentives / Mills M., Rindfuss R. R., McDonald P., Te Velde E., on behalf of the ESHRE Reproduction and Society Task Force // Human Reproduction Update. 2011. №17(6). C. 848-860.
5. Goldin C., Katz LF. The power of the pill: oral contraceptives and women's career and marriage decisions // J Pol Econ. 2002. №110. C.730-770.
6. Van Bavel, J. Choice of study discipline and the postponement of motherhood in Europe: The impact of expected earnings, gender composition, and family attitudes // Demography. 2010. №47. C.439-458.
7. Kreyenfeld M. Uncertainties in Female Employment Careers and the Postponement of Parenthood in Germany // European Sociological Review. 2010. №26(3). C. 351-366.
8. Adserà A. Changing fertility rates in developed countries. The impact of labor market institutions // J Popul Econ. 2004. №17. C.17-43.
9. Andersson G. The Impact of Labour-Force Participation on Childbearing Behaviour: Pro-Cyclical Fertility in Sweden during the 1980s and the 1990s // European Journal of Population. 2000. №16. C.293-333.
10. Aassve A., Billari F.C. & Spéder, Z. Societal Transition, Policy Changes and Family Formation: Evidence from Hungary // Eur J Population. 2006. №22. C.127-152.
11. Журавлева Т. Л., Гаврилова Я. А. Анализ факторов рождаемости в России: что говорят данные РМЭЗ НИУ ВШЭ? // Экономический журнал ВШЭ.

2017. №1. URL: https://cyberleninka.ru/article/n/analiz-faktorov-rozhdaemosti-v-rossii-chto-govoryat-dannye-rmez-niu-vshe (дата обращения: 05.05.2021)

12. Fragile Families and Child Wellbeing Study [Электронный ресурс]. – Режим доступа: https://fragilefamilies.princeton.edu/, свободный. – (дата обращения: 16.05.2021).

13. Stanescu D., Wang E., & Yamauchi S. Using LASSO to Assist Imputation and Predict Child Well-being // Socius. 2019.

14. Honaker J., King G., & Blackwell M. Amelia II: A Program for Missing Data // Journal of Statistical Software. 2011. №45(7). С.1-47.

15. Carnegie N. B., & Wu J. Variable Selection and Parameter Tuning for BART Modeling in the Fragile Families Challenge // Socius. 2019.

16. Winning Models for Grade Point Average, Grit, and Layoff in the Fragile Families Challenge / Rigobon D. E., Jahani E., Suhara Y., AlGhoneim K., Alghunaim A., Pentland A. “Sandy,” & Almaatouq A. // Socius. 2019.

17. Roberts C. V. Friend Request Pending: A Comparative Assessment of Engineering- and Social Science–Inspired Approaches to Analyzing Complex Birth Cohort Survey Data // Socius. 2019.

18. Predicting the Educational and Life Trajectory of Young People Using TrEP Data [Электронный ресурс]. – Режим доступа: https://github.com/Anastassiya08/Predicting-the-Educational-and-Life-Trajectory-of-Young-People-Using-TrEP-Data, свободный. – (дата обращения: 16.05.2021).