|  |
| --- |

**Итоговый отчет**

**команда “Ноукодеры”, УЦП-22**

Малинова Ксения

Трогова Алина

Горелик Борис

Мир Илья

**Содержание**

Состав команды, распределение ролей 3

Введение и актуальность 4

Понимание бизнеса 6

Понимание данных 9

Источник данных 9

Анализ данных 11

Поиск внешних данных 16

Подготовка данных 17

Моделирование с использованием Orange Data Mining 24

Предварительная оценка экономического эффекта 28

Подготовка данных и моделирование с использованием Python 29

Рекомендации по внедрению 33

**Состав команды, распределение ролей**

|  | **PM** | **BU** | **DS+ML** | **BA** | **Sum** |
| --- | --- | --- | --- | --- | --- |
| **Малинова Ксения** |  |  | 0,4 | 0,6 | 1 |
| **Трогова Алина** | 0,2 | 0,8 |  |  | 1 |
| **Горелик Борис** | 0,2 |  | 0,6 | 0,2 | 1 |
| **Мир Илья** | 0,6 | 0,2 |  | 0,2 | 1 |
| **Sum** | 1 | 1 | 1 | 1 |  |

1. **Введение и актуальность**

Целью данной работы является разработка алгоритмов выявления мошеннических транзакций на рынке добровольного медицинского страхования в абстрактной стране Кабо-Верде. Объект изучения - страховая компания “Medicina para todos”, которая является лидером локального рынка страхования медицинских услуг. Компания испытывает трудности в связи со значимой долей мошеннических операций, которые являются следствием недобросовестности как некоторых сотрудников компании, так и контрагентов: пациентов, врачей, сотрудников лечебных учреждений. Некоторое время назад компания приняла решение создать подразделение по выявлению такого рода транзакций.

Для введения в контекст опишем, как устроена система медицинского страхования и на каком этапе возникают мошеннические транзакции, о которых пойдет речь в нашей работе.

Лица или организации покупают медицинскую страховку у страховой компании. Это может быть индивидуальное страхование или корпоративное, предоставляемое работодателями. Страховые компании разрабатывают различные полисы, определяя покрытие, лимиты и условия страхования. Эти полисы могут включать в себя различные виды медицинских услуг, такие как визиты к врачу, лабораторные исследования, госпитализация и другие.

Страховые компании могут иметь соглашения с определенными лечебными учреждениями, включая больницы, поликлиники, лаборатории и т.д (далее - ЛПУ). Эти учреждения обычно входят в сеть предоставляемых услуг и соглашаются работать с определенными страховыми компаниями. Застрахованные лица обращаются за медицинской помощью в учреждения, включенные в сеть. Учреждения предоставляют услуги в соответствии с условиями страхового полиса. Помимо этого, пациенты могут и сами частично финансировать медицинские услуги, например, выбирая более дорогие материалы и т.д. Лечебные учреждения выставляют счета за предоставленные услуги страховой компании. Оплата за медицинские услуги происходит в соответствии с условиями страхового полиса. В случае возникновения разногласий или проблем страховые компании и лечебные учреждения могут взаимодействовать для урегулирования споров и обработки претензий.

Медицинское страхование постоянно сталкивается с рядом существенных вызовов:

* старение населения, злоупотребление вредной пищей, распространенность вредных привычек - все это приводит к увеличению обращений за медицинскими услугами
* пандемии и эпидемии
* требовательность клиентов: с развитием технологий, клиенты страховых компаний становятся более требовательными к сервису
* законодательные риски: ужесточения законов о хранении и обработке персональных медицинских данных, а также растущие угрозы сохранения конфиденциальности данных

Однако самым значительным фактором риска для страховой компании является мошенничество поставщиков (ЛПУ).

Какие типы мошеннических схем существуют?

1. Подача ложных заявлений о медицинских услугах или условиях для получения несуществующей компенсации. Это могут быть:

- нерелевантные назначения исследований и визитов к врачу (злоупотребление)

- несуществующие визиты к врачам / не проведенные в реальности исследования

- недостоверные сведения о лечении

- заведомо неправильное установление диагноза или предоставление завышенной информации о состоянии пациента для получения больших страховых выплат

1. Идентификационное мошенничество - когда услуга оказывается другому человеку

В рамках нашей работы мы будем рассматривать мошеннические транзакции со стороны ЛПУ. По статистике около 5% счетов, выставляемых локальными поликлиниками по программам добровольного [медицинского](https://zdrav.expert/index.php/%D0%9C%D0%B5%D0%B4%D0%B8%D1%86%D0%B8%D0%BD%D0%B0) [страхования](https://zdrav.expert/index.php/%D0%A1%D1%82%D1%80%D0%B0%D1%85%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5), содержат услуги, которые либо не нужны пациентам, либо физически им не оказаны. Глобально на фрод в сегменте ДМС приходится от 10% до 20% от всех оказанных клиниками услуг.

Из обнаруженных случаев мошенничества 35% приходится на использование схемы расширения диагноза. Суть этой схемы заключается в том, что страховые компании обычно отклоняют услуги, не соответствующие установленному диагнозу. Для обхода этого ограничения пациентам устанавливают максимально "обобщенные" диагнозы, такие как бронхит вместо насморка. Это позволяет расширить перечень предписываемых медицинских процедур.

Дополнительные 17% случаев мошенничества связаны с добавлением услуг. При таком виде мошенничества в счете для пациента, получившего необходимые медицинские услуги, появляются услуги, которые клиника "добавила". Обычно такие услуги подбираются таким образом, чтобы не противоречить диагнозу пациента. Также 16% случаев мошенничества происходит в контексте схем, связанных с нереализованными повторными приемами и неполной посещаемостью курсов лечения.

**2. Понимание бизнеса**

**Бизнес контекст**

Страховая компания “Medicina para todos”, специализирующаяся на добровольном медицинском страховании, представляет собой ключевого участника здравоохранения в одной небольшой стране с развивающейся экономикой и стареющим населением, предлагая своим клиентам высококачественные и персонализированные медицинские страховые решения.

**Масштаб**

“Medicina para todos” обслуживает сотни тысяч клиентов по всей стране, предоставляя широкий спектр медицинских страховых продуктов и услуг.

**Инновации**

Компания постоянно внедряет новые технологии и подходы в сфере здравоохранения.

**Партнерство**

“Medicina para todos” активно сотрудничает с ведущими медицинскими учреждениями, врачами, лабораториями и другими партнерами в области здравоохранения для обеспечения своих клиентов лучшими возможными услугами.

Результаты работы сотрудников компании по выявлению фрода показывают, что от 10 до 20% от общих расходов “Medicina para todos” составляют недобросовестные претензии, которые являются следствием мошеннических действий сотрудников и/или контрагентов, имеющих цели корыстной наживы.

Дальнейшая цель состоит в том, чтобы научиться в автоматическом режиме предсказывать, является ли страховой случай потенциально мошенническим или оценить вероятность мошеннической деятельности сотрудника и/или контрагента, найти причины, стоящие за этим, а также предотвратить финансовые потери.

В зависимости от оценки вероятности и причин мошенничества, страховая компания может принять или отклонить претензию или провести расследование в отношении этого страхового случая.

Не только финансовые потери являются большой проблемой, но также и защита системы страхования в здравоохранении, нацеленная на качественную и безопасную помощь пациентам, деятельность которых остается в рамках закона.

**Юнит-экономика и финмодель**

Конечной целью внедрения модели является снижение затрат на обеспечение проверки страховых случаев, которая в настоящее время проводятся вручную с использованием труда высокооплачиваемых специалистов. Кроме того в ручном режиме невозможно проверить 100% полученных счетов: таким образом, счета проверяются рандомно, в итоге удается охватить около 50% полученных счетов. В результате внедрения модели планируется первичную проверку осуществлять автоматически и подключать специалистов только на те случаи, которые признаны первичной проверкой мошенническими или “пограничными”, что позволит сократить затраты на специалистов пропорционально доли мошеннических операций в общем объеме, то есть ориентировочно в 10 раз.

Расчет экономического эффекта от внедрения модели представлен в разделе 6 “Предварительная оценка экономического эффекта”.

**Описание задачи**

В основе работы лежит создание модели бинарной классификации на основе предварительного анализа заявок на выплаты сумм страхового вознаграждения, сумм софинансирования пациентом, вместе с данными о нахождении в стационаре, на амбулаторном лечении, данными о пациентах и данными об обнаруженном ранее фроде в привязке к конкретному сотруднику страховой компании, который составлял заявку на выплату вознаграждения по данному страховому случаю.

Важным ограничением при решении задачи является то, что в основе модели должен лежать принцип интерпретируемости, так как в случае обнаружения фрода необходима достаточная доказательная база для последующей легализации решения.

В цепочке принятия решения о видах лечения пациента, наличия тех или иных процедур и манипуляций принимает участие несколько врачей, сотрудник страховой компании и сам пациент. Соответственно, можно предположить, что все они могут быть инициаторами и/или соучастниками фрода. В условие задачи входит временное ограничение (30 дней) на оплату законной транзакции. Главной заинтересованной стороной для обнаружения фрода является страховая компания.

К предполагаемым схемам мошенничества можно отнести:

1. Увеличение суммы счета, в том числе путем увеличения срока госпитализации по документам по сравнению с фактическим сроком
2. Увеличение суммы счета путем подмены услуг на более дорогие
3. Выставление счета за неоказанные услуги

Цель создания модели - предсказание фрода на основе имеющихся данных о пациенте, параметрам стационарного лечения, параметрам амбулаторного лечения, суммах платежей.

Ограничения модели состоят также в том, что стоимость неправильной классификации фрода достаточно значима. В первую очередь для законных транзакций, которые могут быть классифицированы как незаконные (False Positive), так как это может привести к проигранным встречным искам и потере деловой репутации страховой компании, сокращению числа партнеров.

В случае пропуска незаконной транзакции (False Negative) речь идет о финансовых потерях и возможном повторении фрода в будущем со стороны участников транзакции.

Также на необходимость высокой интерпретируемости модели влияет вероятное последующее расследование уполномоченными сотрудниками со стороны агентства.

В качестве критериев оценки модели, помимо интерпретируемости, предлагаются следующие метрики:

1. AUC
2. F1
3. Accuracy
4. FPR
5. FNR

**3. Понимание данных**

**Источник данных**

Для решения поставленной задачи было подготовлено несколько датасетов. В первую очередь это данные о страховых случаях за год, отдельно для амбулаторного лечения пациентов и отдельно для стационарного лечения.

Специфика амбулаторного и стационарного лечения отражается в различном составе атрибутов. Важным моментом при составления этих датасетов является указание в них ID пациента, сотрудника страховой компании, принимавшего решение по данному случаю и участвовавших в лечении врачей.

Отдельно был собран датасет по пациентам - клиентам страховой компании с указанием их личных данных, сумм, потраченных на лечение, сумм страхового возмещения и данных о имеющихся у них хронических заболеваниях. Персональные данные пациентов были обезличены.

Также в результате работы специальной комиссии были собраны данные о выявленных со стороны сотрудников страховой компании мошеннических действиях.

В результате, в настоящее время мы располагаем набором следующих данных:

1. Данные о пациентах

*(Train\_Beneficiarydata-1542865627584.csv)*

*BeneID* - идентификатор пациента в системе

*DOB* - дата рождения пациента

*DOD* - дата смерти пациента (если пациент скончался на момент составления датасета, в случае, если пациент жив - N/A)

*Gender* - пол пациента

*Race* - раса (от 1 до 5)

*State, County* - штат, округ

*IPAnnualReimbursementAmt*, *OPAnnualReimbursementAmt* - суммы запрошенных страховых возмещений для случаев, требующих госпитализации и для случаев амбулаторного лечения

*IPAnnualDeductibleAmt*, *OPAnnualDeductibleAmt* - суммы, которые пациент оплатил самостоятельно, соответственно при госпитализации и при амбулаторном лечении

Также в датасете присутствуют категориальные данные о наличии у пациента различных хронических заболеваний.

1. Данные о госпитализациях

*(Train\_Inpatientdata-1542865627584.csv)*

*BeneID* - идентификатор пациента в системе

*ClaimID* - идентификатор заявки (страхового случая)

ClaimStartDt - дата начала страхового случая

*ClaimEndDt* - дата окончания страхового случая

*Provider* - идентификатор сотрудника страховой компании, который принимал решение по данному случаю

*InscClaimAmtReimbursed* - сумма возмещения по данному случаю

*DeductibleAmtPaid* - сумма, уплаченная пациентом

*AttendingPhysician*, *OperatingPhysician*, *OtherPhysician* - идентификаторы врачей, участвовавших на разных стадиях лечения

*AdmissionDt* - дата поступления пациента в стационар

*DischargeDt* - дата выписки пациента из стационара

Также в датасет содержатся категориальные данные о поставленных диагнозах и различных проведенных манипуляциях.

1. Данные об амбулаторном лечении *(Train\_Outpatientdata-1542865627584.csv)*

*BeneID* - идентификатор пациента в системе

*ClaimID* - идентификатор заявки (страхового случая)

*ClaimStartDt* - дата начала страхового случая

*ClaimEndDt* - дата окончания страхового случая

*Provider* - идентификатор сотрудника страховой компании, который принимал решение по данному случаю

*InscClaimAmtReimbursed* - сумма возмещения по данному случаю

*DeductibleAmtPaid* - сумма, уплаченная пациентом

*AttendingPhysician*, *OperatingPhysician*, *OtherPhysician* - идентификаторы врачей, участвовавших на разных стадиях лечения

Также в датасет содержатся категориальные данные о поставленных диагнозах и различных проведенных манипуляциях.

1. Данные проверенных ранее транзакциях

*(Train-1542865627584.csv)*

*Provider* - идентификатор сотрудника страховой компании

*PotentialFraud* - категориальный идентификатор фрода (Yes, No)

**Анализ данных**

1. Данные о пациентах

*(Train\_Beneficiarydata-1542865627584.csv)*

Присутствуют данные о 138556 пациентах (уникальный *BeneID*) с 25 атрибутами, Пропущенные данные встречаются только у атрибута *DOD* - дата смерти пациента. Предлагается в будущем вместо атрибутов *DOB* - дата рождения пациента и *DOD* - дата смерти пациента, ввести единый атрибут *Age* - возраст пациента на момент начала страхового случая. Также вместо атрибутов *IPAnnualReimbursementAmt*, *OPAnnualReimbursementAmt* - суммы запрошенных страховых возмещений для случаев, требующих госпитализации и для случаев амбулаторного лечения и соответственно атрибутов *IPAnnualDeductibleAmt*, *OPAnnualDeductibleAmt* - суммы, которые пациент оплатил самостоятельно, соответственно при госпитализации и при амбулаторном лечении, предлагается ввести атрибуты конечной оплаты соответствующих услуг:

*PHosp = IPAnnualReimbursementAmt - IPAnnualDeductibleAmt*

и

*PAmb = OPAnnualReimbursementAmt - OPAnnualDeductibleAmt*

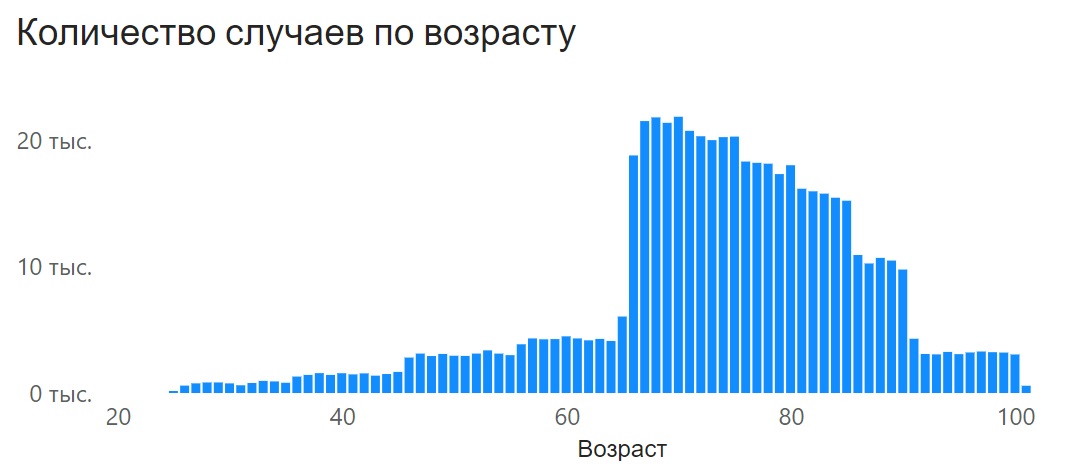
и атрибут общей оплаты:

*PTotal = PHosp + PAmb*

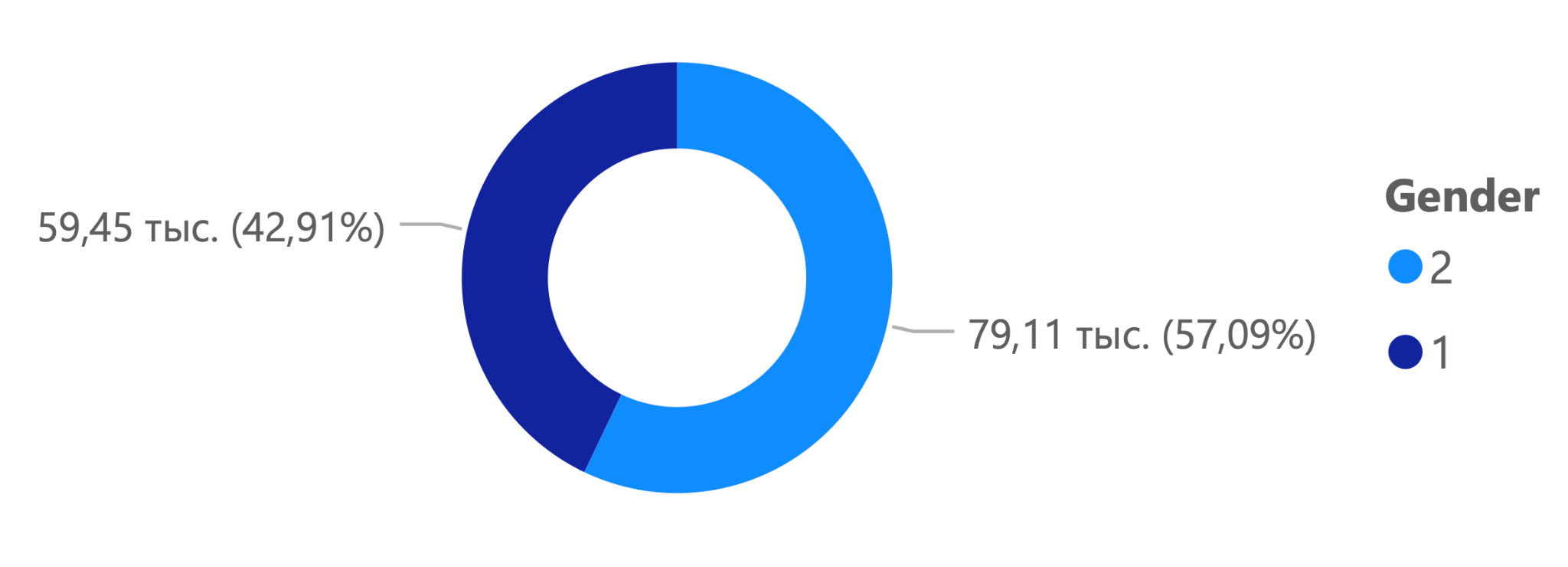
Проведенный развед анализ не выявил каких-либо проблем с данными, которые необходимо устранить на предварительном этапе.

Визуализация данных:

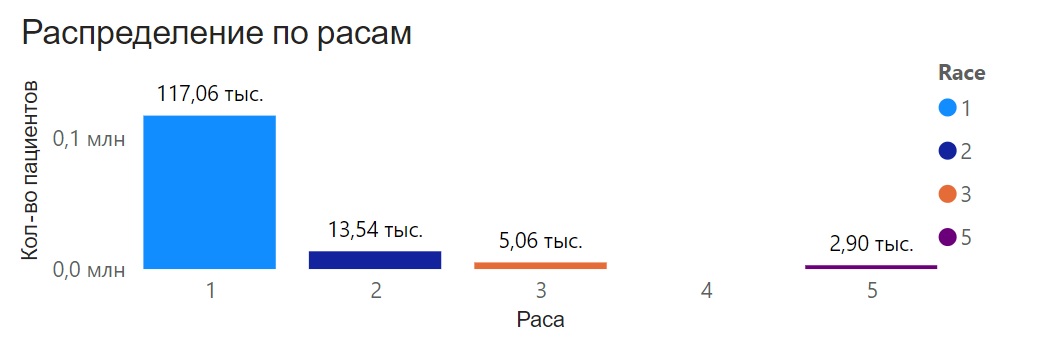
*1.1. Распределение по возрасту (на момент наступления страхового случая)*



*1.2. Распределение по полу*



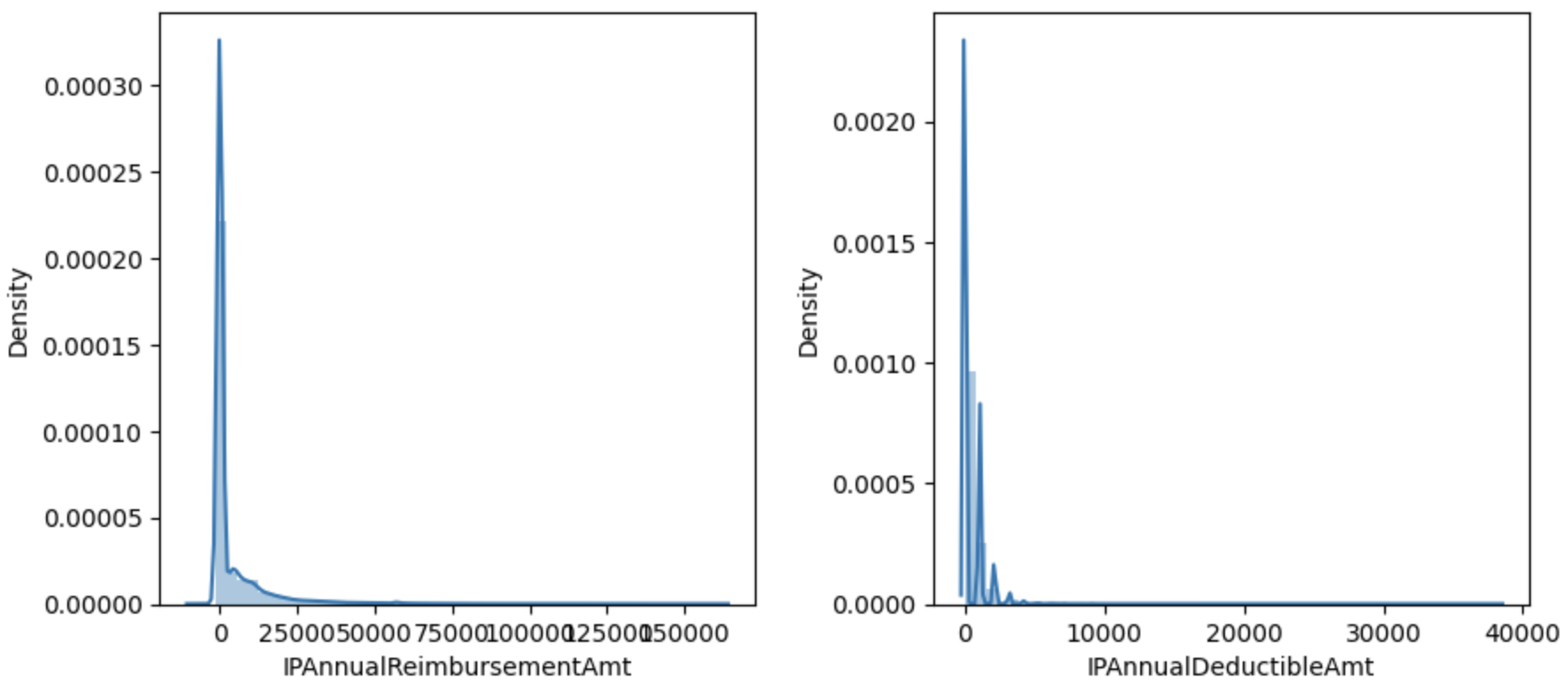
*1.3. Распределение по расам*

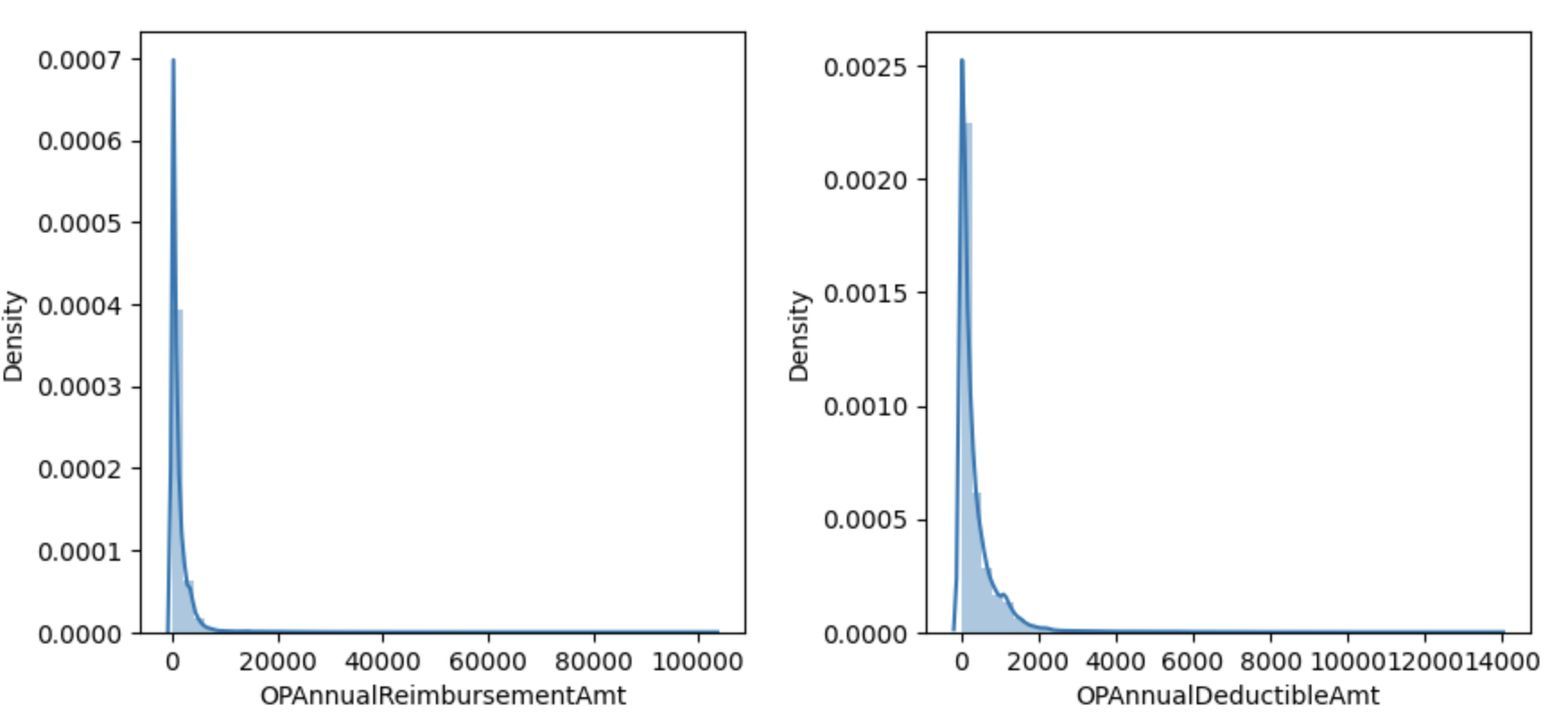
**

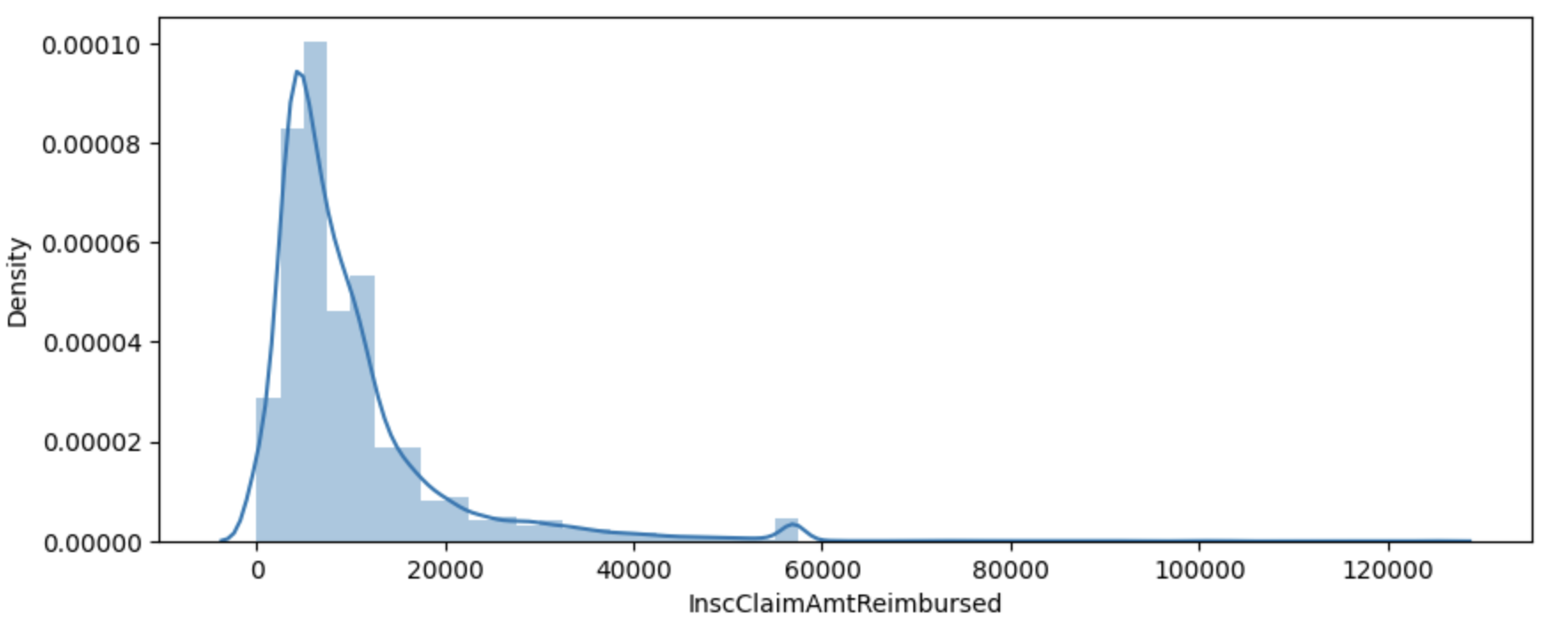
*1.4. Распределение по штатам*



*1.5. Распределение сумм возмещения и выплат пациентов*







1. Данные о госпитализациях

*(Train\_Inpatientdata-1542865627584.csv)*

Присутствуют данные о 40 474 страховых случаях с 30 атрибутами. Пропущенные данные встречаются только у атрибутов, которые являются кодами медицинских манипуляций, что логично, поскольку не все манипуляции проводились с конкретным пациентом. У атрибутов, которые описывают даты начала и конца страхового случая и даты госпитализации и выписки - пропусков нет. Обращает на себя внимание тот факт, что эти даты иногда не совпадают. Можно предположить, что это явилось следствием злонамеренных манипуляций и является признаком фрода. Предлагается в дальнейшем ввести новые атрибуты - продолжительность госпитализации:

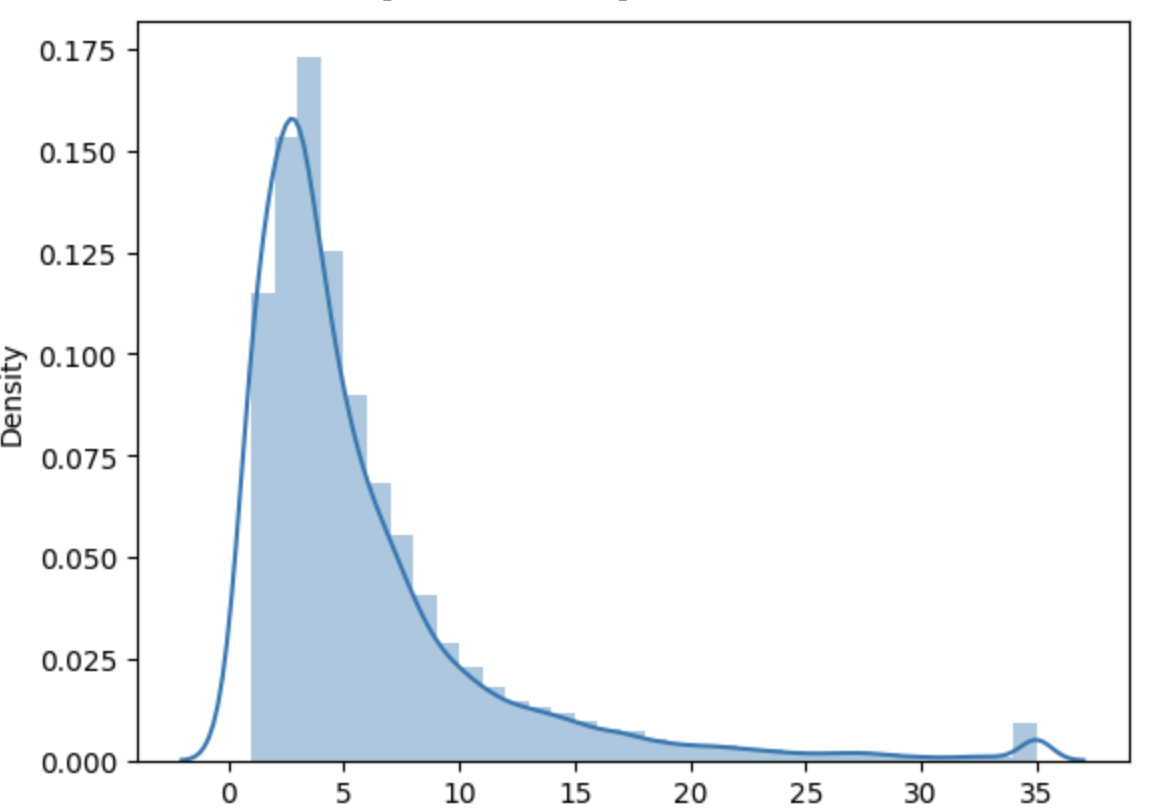
*Dur\_Hosp = DischargeDt - AdmissionDt*

разницу между этими двумя продолжительностями:

*Problem\_dur = ( ClaimEndDt - ClaimStartDt ) - Dur\_Hosp*

Визуализация данных:

*2.1. Продолжительность госпитализации*



1. Данные об амбулаторных лечениях

*(Train\_Outpatientdata-1542865627584.csv)*

Присутствует 517 737 записей о страховых случаях с 27 атрибутами. Пропущенные данные встречаются только у атрибутов с кодами медицинских манипуляций, что логично. Данных для сравнения сроков проведения манипуляций и формальных сроков страхового случая недостаточно, присутствуют только вторые.

1. Данные об обнаруженном фроде

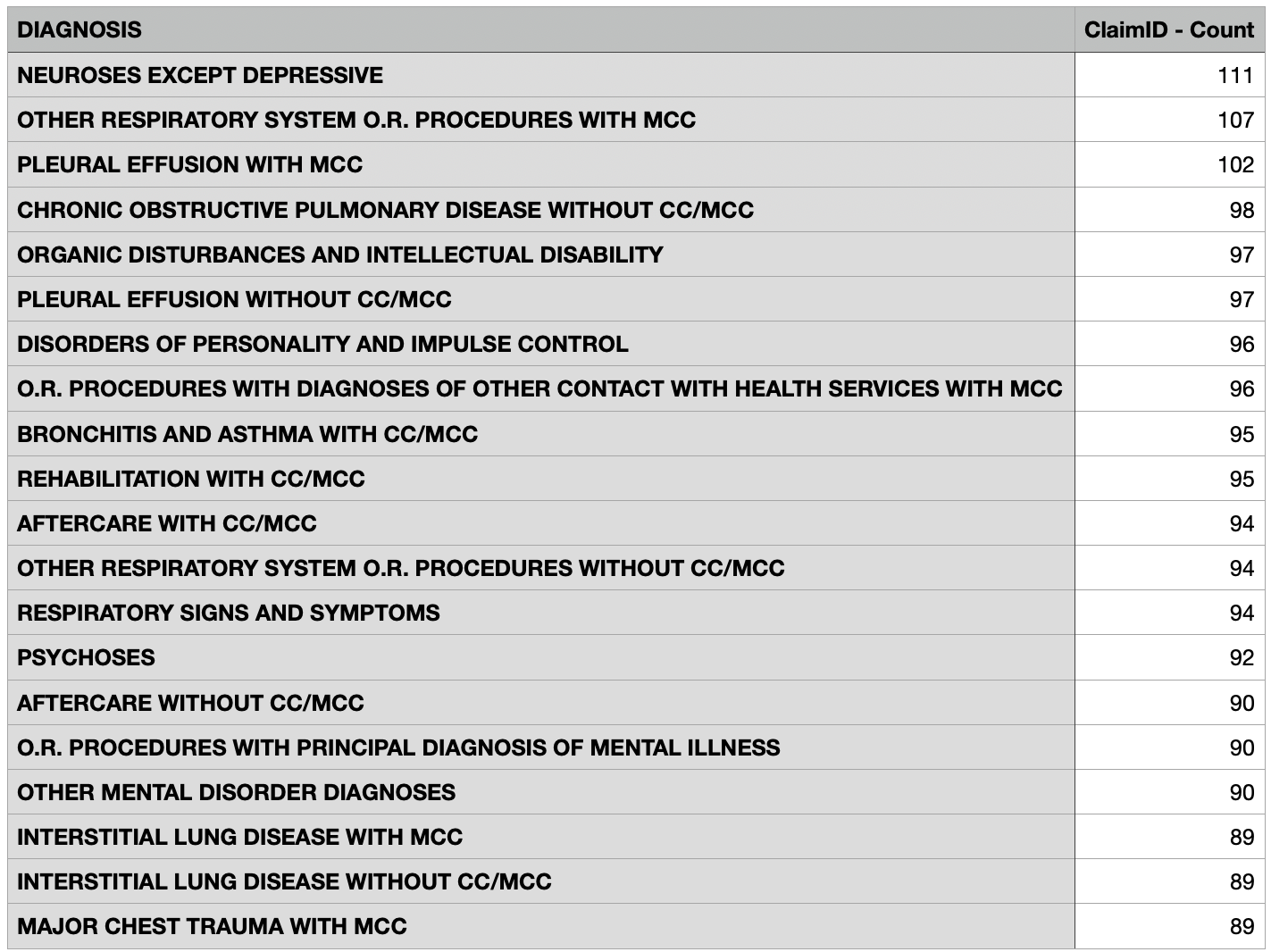
*(Train-1542865627584.csv)*

Присутствует 5 410 записей о сотрудниках страховой компании (провайдерах) с указанием обнаруженного ранее фрода. Данных о том, в каком конкретном страховом случае был обнаружен фрод в датасете нет. Всего описано 506 случаев фрода (9,35% от общей базы агентов).

Можно предположить, что злонамеренные действия сотрудника страховой компании по увеличению сумм страховых выплат являются неоднократными. В таком случае анализ среднего показателя сумм выплаченных вознаграждений по каждому сотруднику может быть информативен для обнаружения фрода в том случае, если эти показатели отличаются от средних. В качестве дополнительного атрибута предлагается ввести среднюю сумму страхового возмещения по каждому агенту *PTotal-mean* по всей выборке его страховых случаев. Возможно, что наличие средней суммы выше среднего является основанием подозревать фрод.

**Поиск внешних данных**

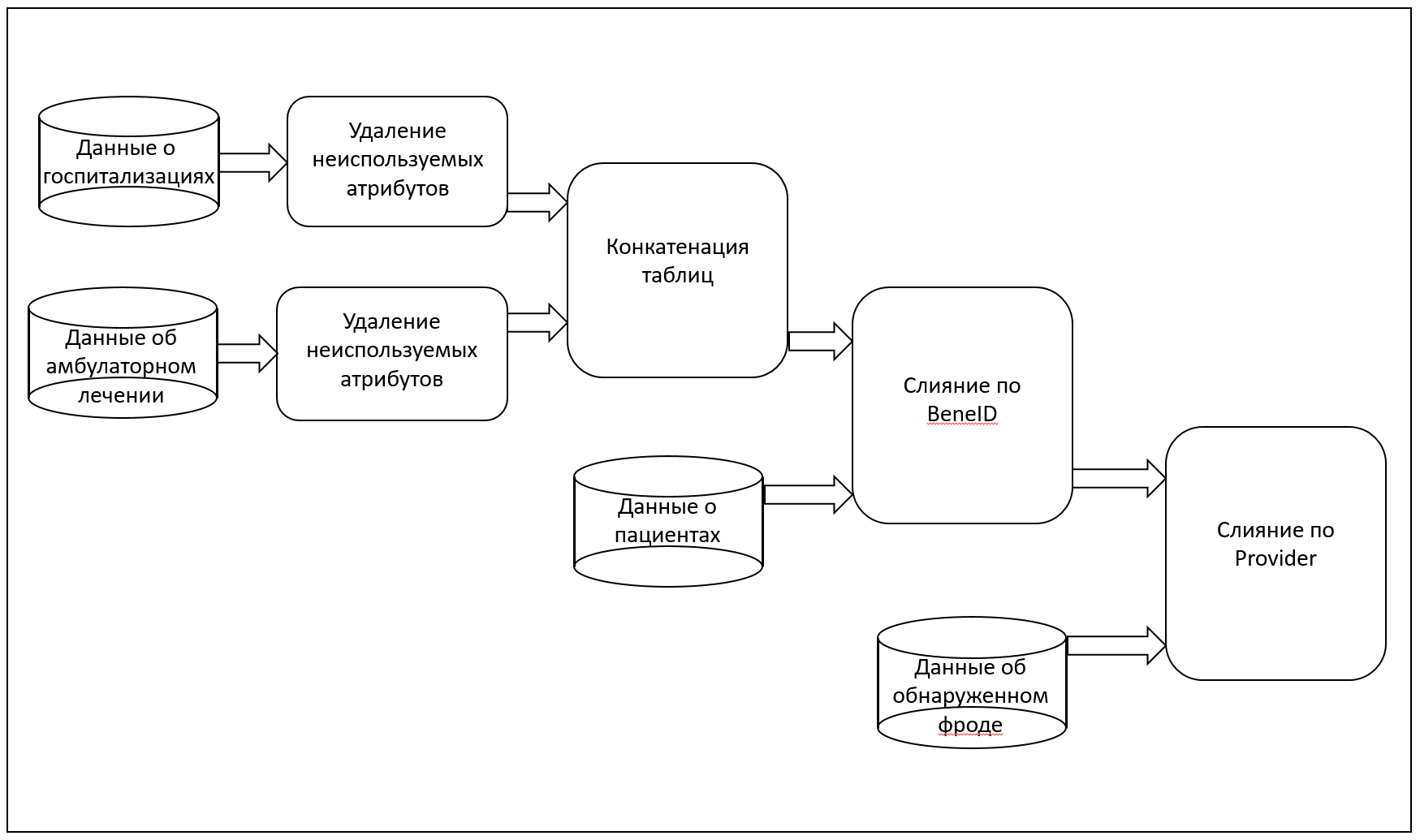
В открытых источниках удалось найти данные о расшифровке кодов диагностики, опубликованные Министерством здравоохранения Кабо-Верде (файл *Codes.csv*). В имеющихся датасетах коды диагностики описаны для пациентов, проходивших лечение в стационаре (атрибут *DiagnisisGroupCode*). Для амбулаторных пациентов такой атрибут не определен. После процесса подготовки данных, описанного в следующем разделе удалось сопоставить поставленные диагнозы и случаи фрода. Топ-20 диагнозов, по которым было обнаружено максимальное количество фрода представлены в таблице:



**4. Подготовка данных**

Для подготовки данных к созданию модели, помимо введения новых атрибутов, описанных выше, необходимо сделать слияние таблиц с целью создать единый датасет для последующего использования в обучении модели.

Последовательность слияния изображена на схеме:



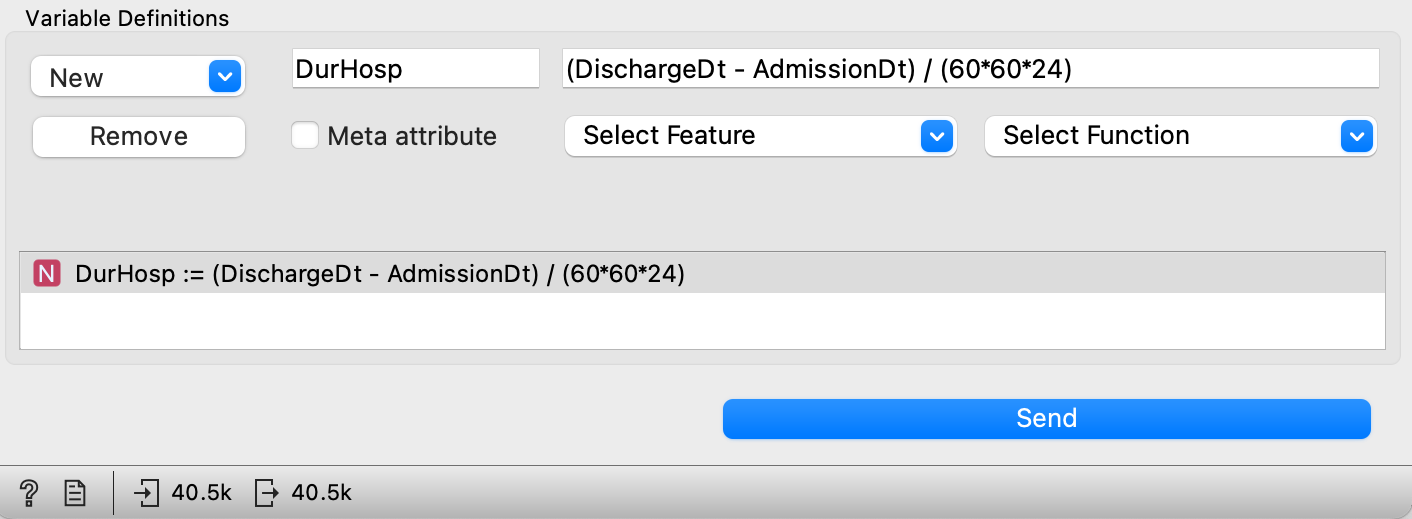
Для проведения предварительных операций по подготовке данных и валидации гипотез о выборе модели команда приняла решение использовать nocode платформу Orange Data Mining.

На первом этапе необходимо было провести конкатенацию таблиц по описанию страховых случаев, возникших в результате стационарного и амбулаторного лечения пациентов. Вводим новые атрибуты, определяющие продолжительность лечения

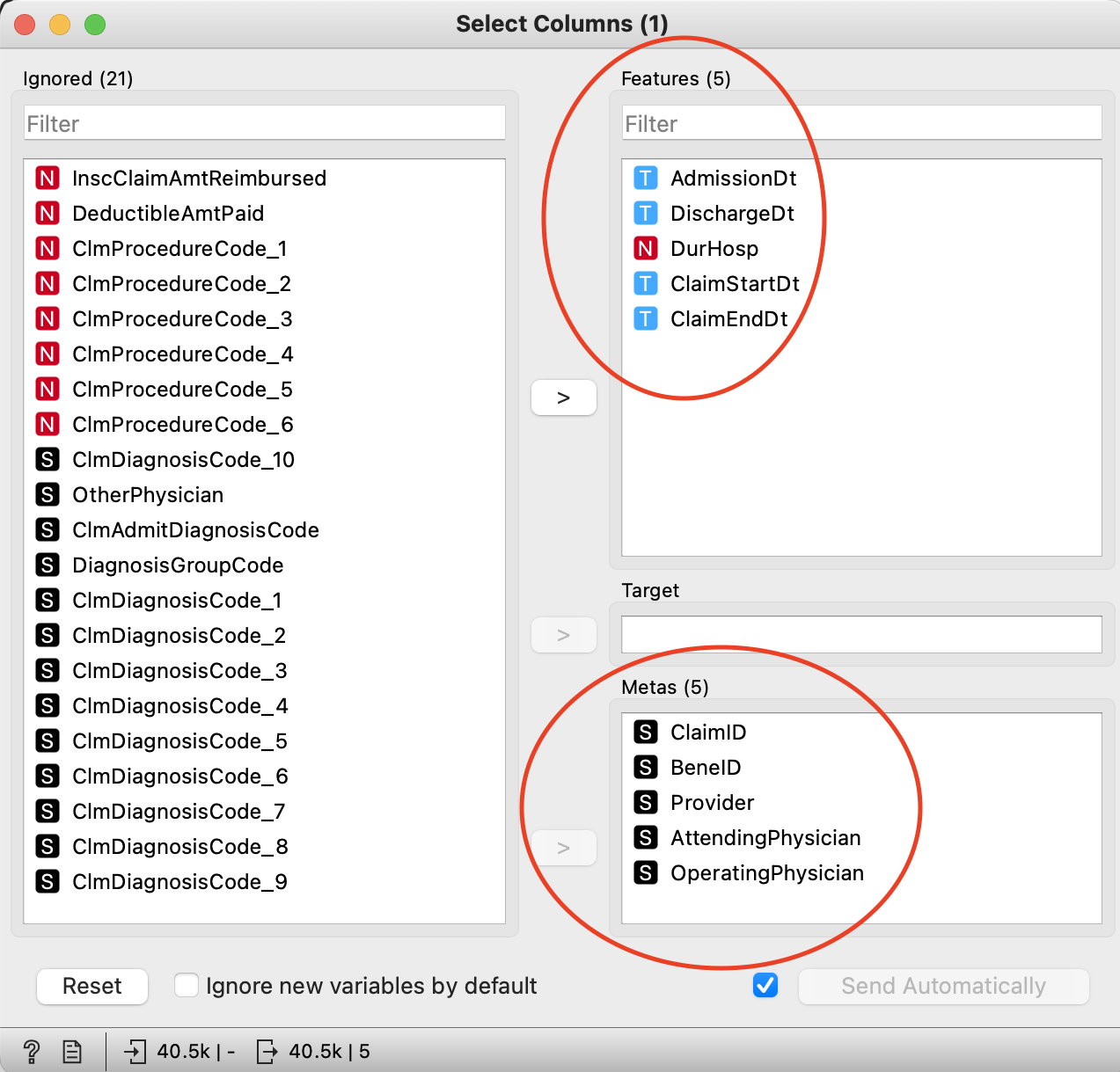
1. Для стационарного решения *(Train\_Inpatientdata)*

Вводим атрибут *DurHosp.*

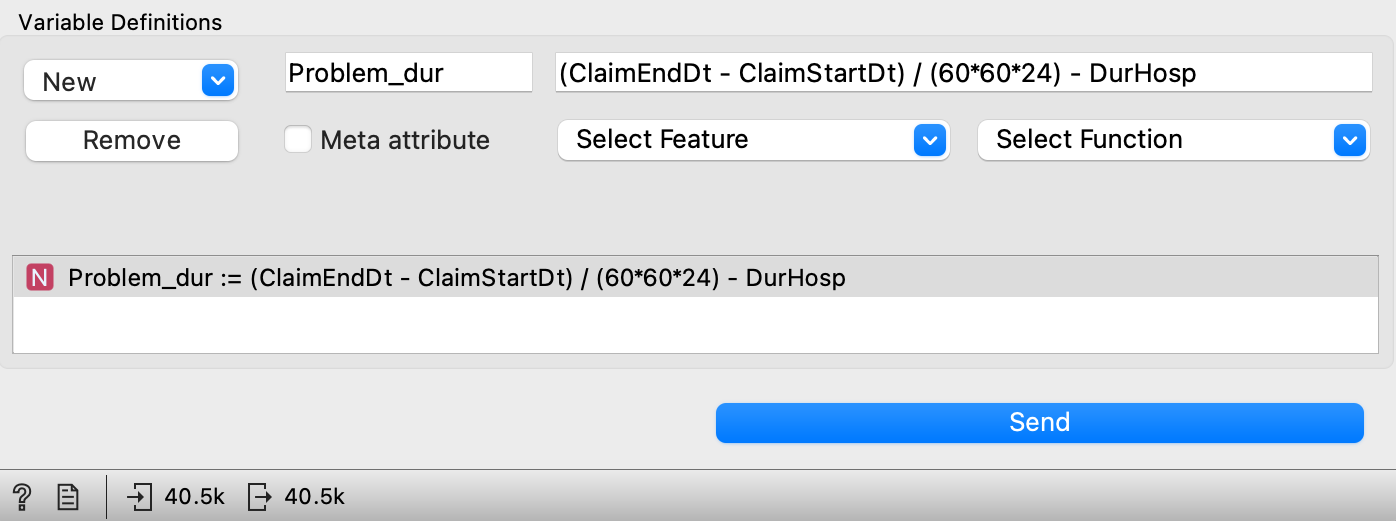
В датасете временные интервалы выражены в секундах, поэтому переводим их в сутки.

**

Далее оставляем только значимые атрибуты:

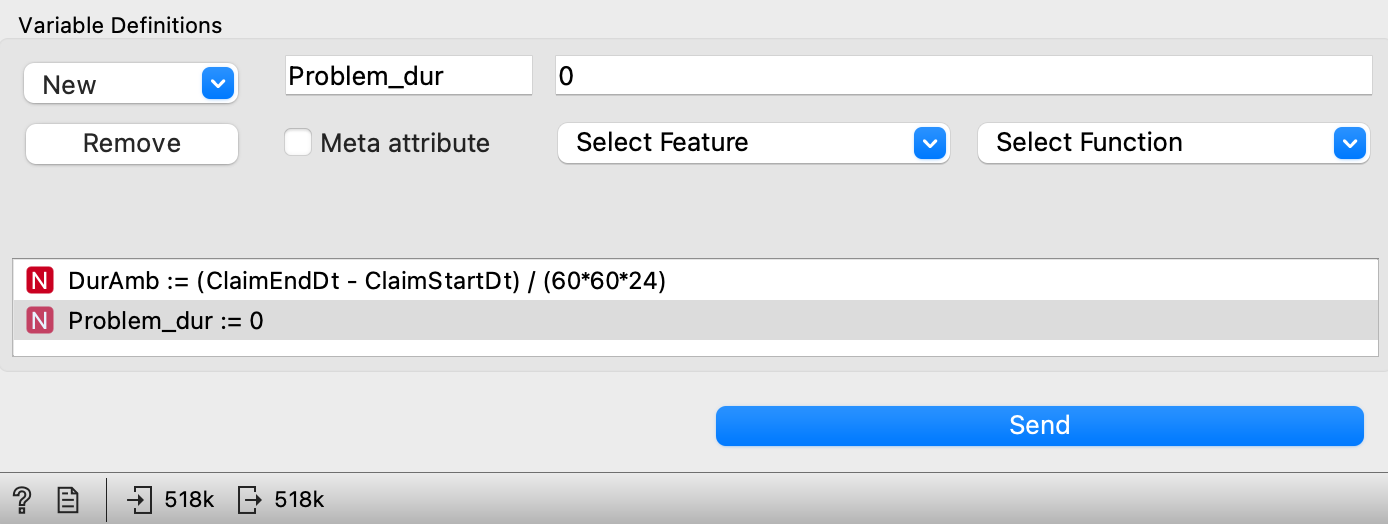


И вводим атрибут *Problem\_dur*, показывающий количество “проблемных дней”:

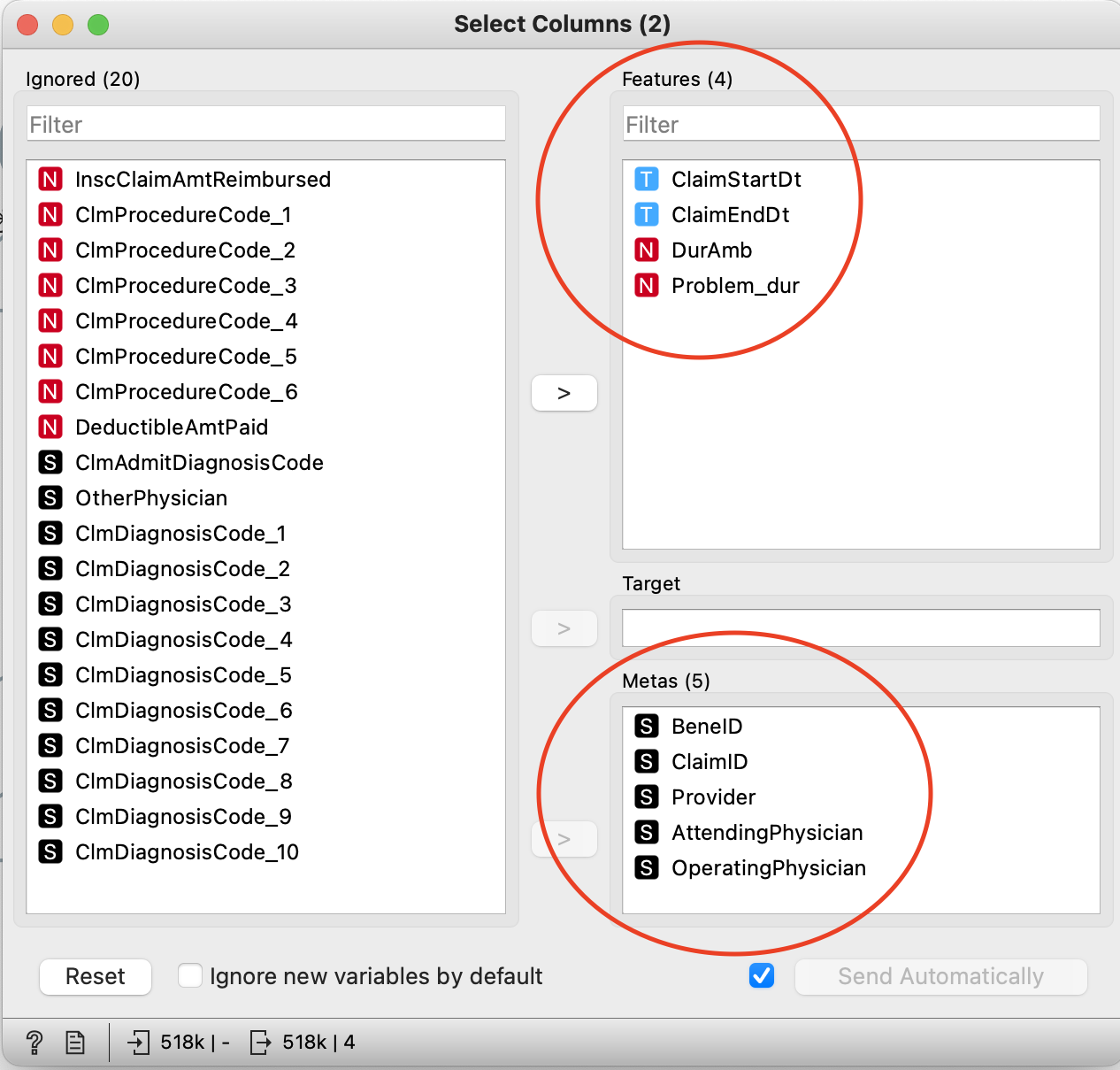


1. Для амбулаторного лечения (*Train\_Outpatientdata)*

Вводим аналогичный атрибуты, но в силу отсутствия данных о периоде амбулаторного лечения принимаем везде *Problem\_dur = 0*



Оставляем идентичный предыдущему датасету набор атрибутов:

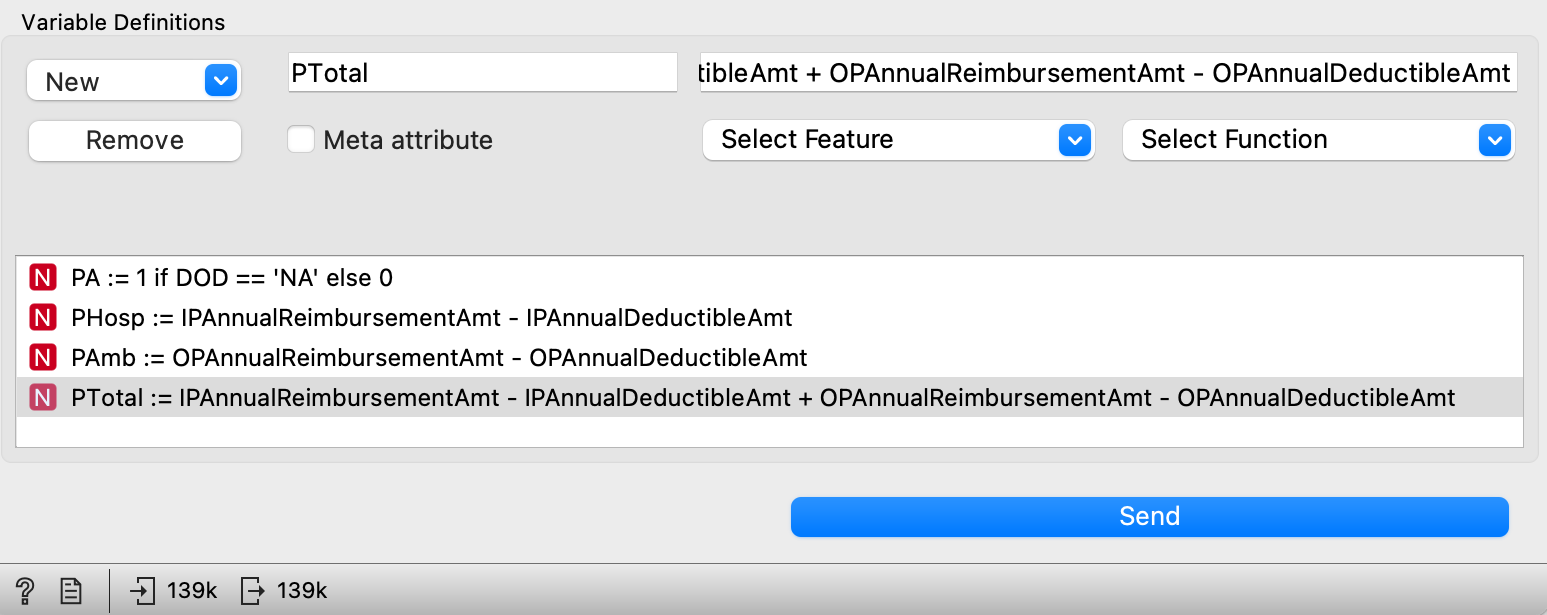


Таблицы готовы к конкатенации. В результате получаем единую таблицу по всем страховым случаям *(Train\_Inpatientdata+Outpatientdata)* со следующими параметрами: 558211 строк, 12 атрибутов.

На втором этапе готовим данные о пациентах *(Train\_Beneficiarydata)* к последующему слиянию с единым датасетом по идентификатору *BeneID*.

Процедура подготовки следующая:

1. Обращаем внимание на тот факт, что некоторые пациенты скончались на момент составления датасета. Вводим категориальный атрибут *PA*, который принимает значение 1, если пациент жив и 0 если нет.
2. Также вводим атрибуты PHosp, PAmb и PTotal, которые определяют сальдо пациента по расчетам за стационарное, амбулаторное лечение и общее сальдо:



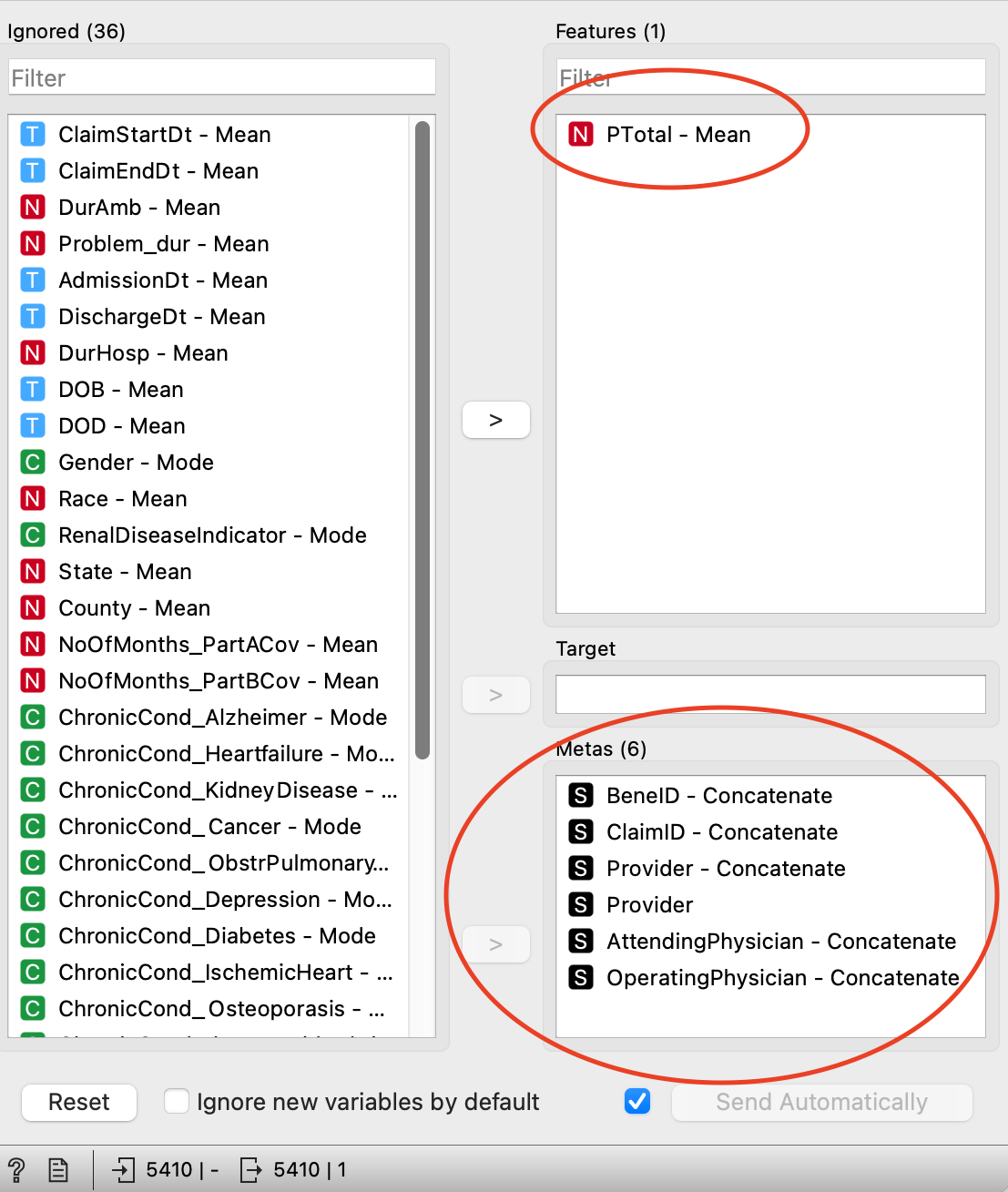
Теперь производим слияние датасетов *Train\_Beneficiarydata* и *Train\_Inpatientdata+Outpatientdata* по идентификатору *BeneID*. В результате получаем общую таблицу *Train\_Inpatientdata+Outpatientdata+Beneficiarydata* со следующими параметрами: 558211 строк, 40 атрибутов.

На третьем этапе производим слияние полученной таблицы *Train\_Inpatientdata+Outpatientdata+Beneficiarydata* с данными об обнаруженном фроде по идентификатору *Provider*. После этого работаем с итоговой таблицей.

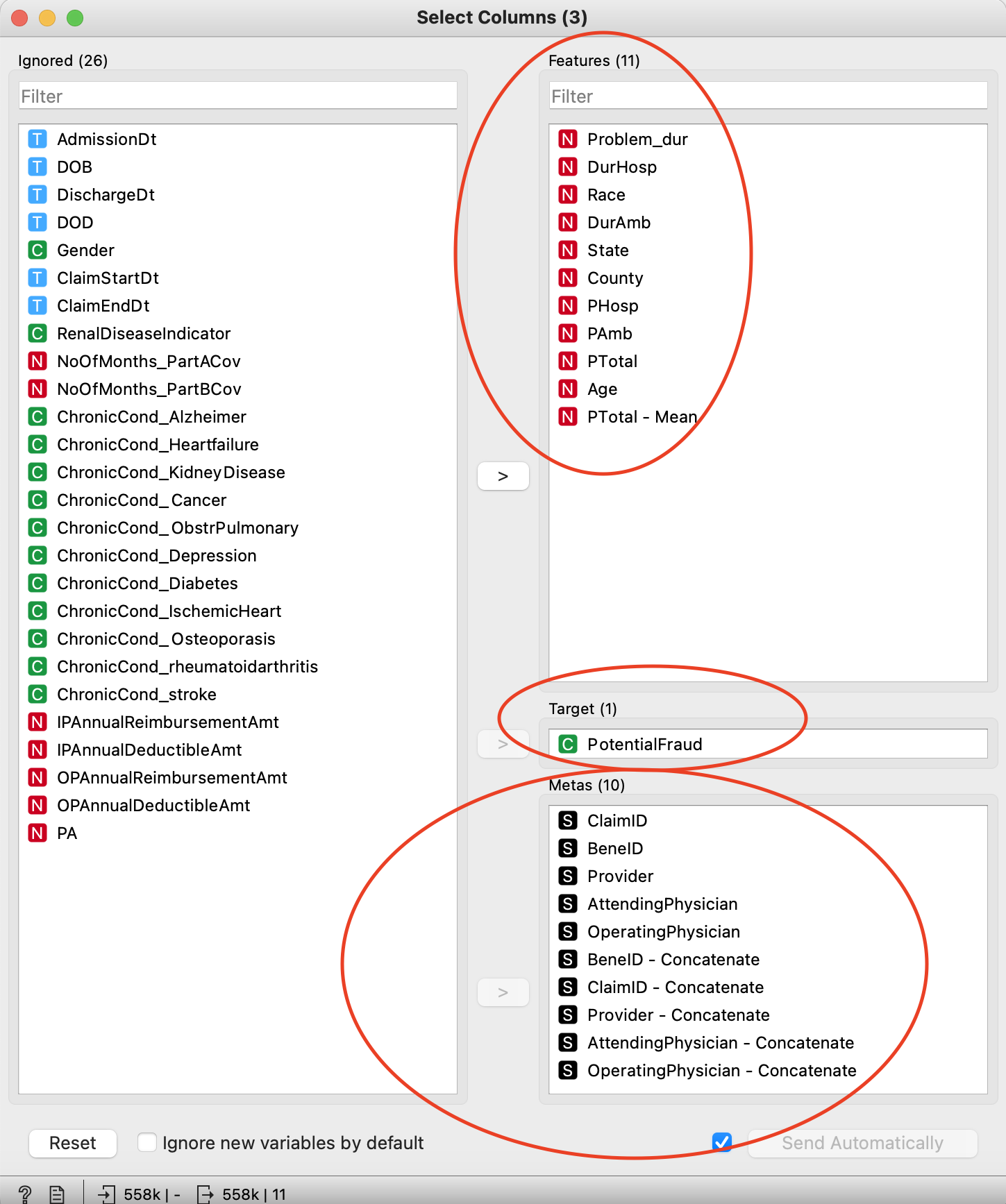
Вводим атрибут *Age*, который определяет возраст пациента на момент начала страхового случая:



Далее делаем группировку по атрибуту Provider, создаем новый атрибут PTotal-mean, который определяет среднее значение выплат по каждому сотруднику страховой компании:



После этого определяем окончательный набор атрибутов для продолжения работы с моделью:



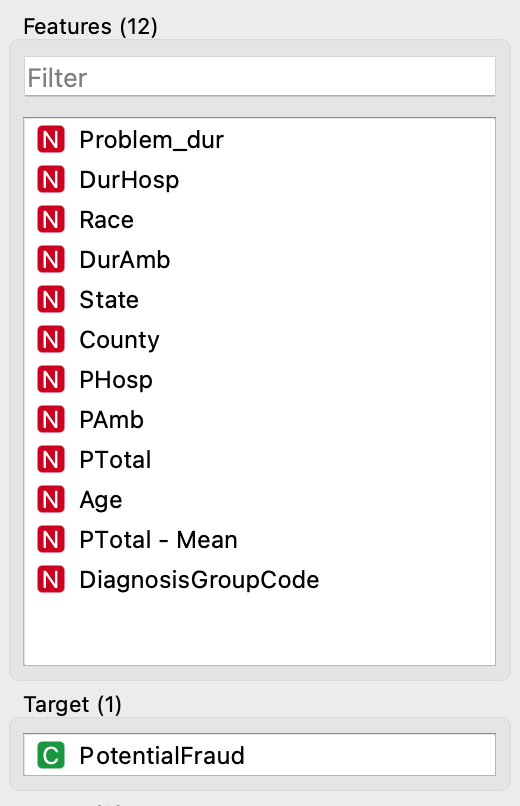
В полученной таблице 558211 строк и 22 атрибута.

Подготовка данных завершена.

**5. Моделирование с использованием Orange Data Mining**

Важным ограничением для бизнес-контекста является интерпретируемость модели. Полученные данные должны быть настолько наглядны, чтобы их можно было использовать для последующей легализации при предъявлении встречных претензий или защиты своей позиции в судебных инстанциях. Поэтому командой было принято решение для первичной валидации модели отказаться от ансамблей и нейросетей. На предварительном этапе с использованием Orange Data Mining изучались два основных варианта: логистическая регрессия и простое дерево решений.

Наилучших показателей удалось добиться при использовании дерева решений со следующими предикторами:

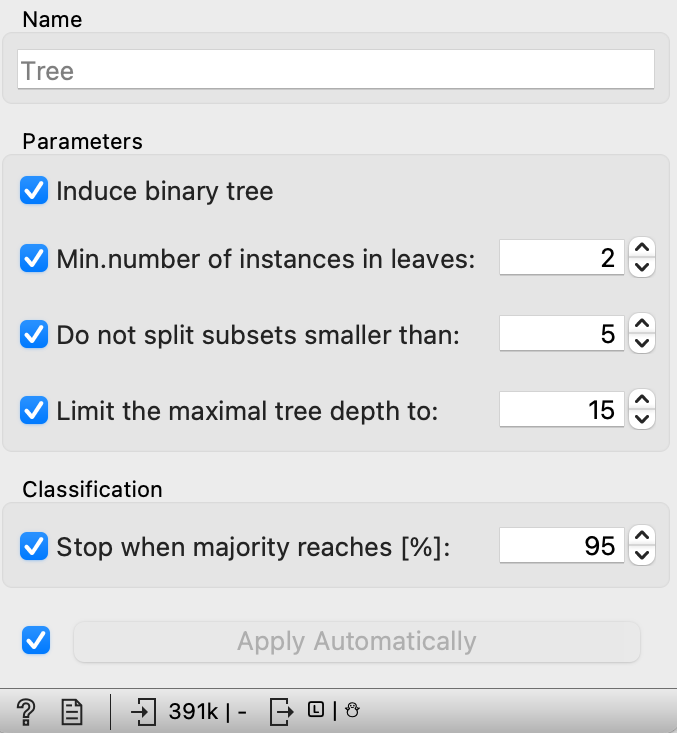


Общая выборка была разделена на обучающую и тестовую в пропорции 70/30.

Модель была исследована на переобучение. Для этого были построены варианты с различными параметрами глубины. Результаты представлены в таблице:

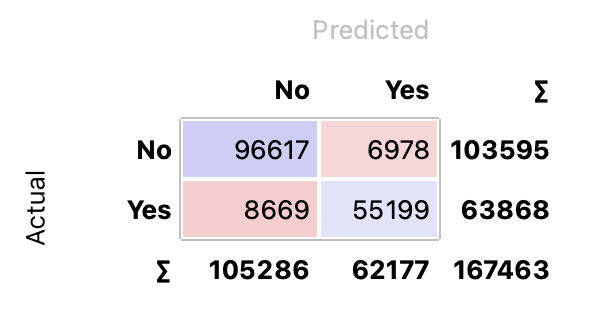
| **Tree depth** | **AUC** | **F1** | **Prec** | **Recall** |
| --- | --- | --- | --- | --- |
| 10 | 0,902 | 0,727 | 0,804 | 0,663 |
| 11 | 0,922 | 0,764 | 0,823 | 0,714 |
| 12 | 0,941 | 0,811 | 0,834 | 0,790 |
| 13 | 0,954 | 0,833 | 0,864 | 0,804 |
| 14 | 0,964 | 0,855 | 0,887 | 0,826 |
| 15 | 0,970 | 0,876 | 0,888 | 0,864 |
| 16 | 0,976 | 0,890 | 0,916 | 0,860 |
| 20 | 0,990 | 0,949 | 0,959 | 0,940 |
| 30 | 0,991 | 0,978 | 0,982 | 0,973 |

Из полученных результатов можно сделать вывод, что модель начинает переобучаться, начиная с параметра глубины равном 14-15. Окончательный выбор параметров дерева следующий:

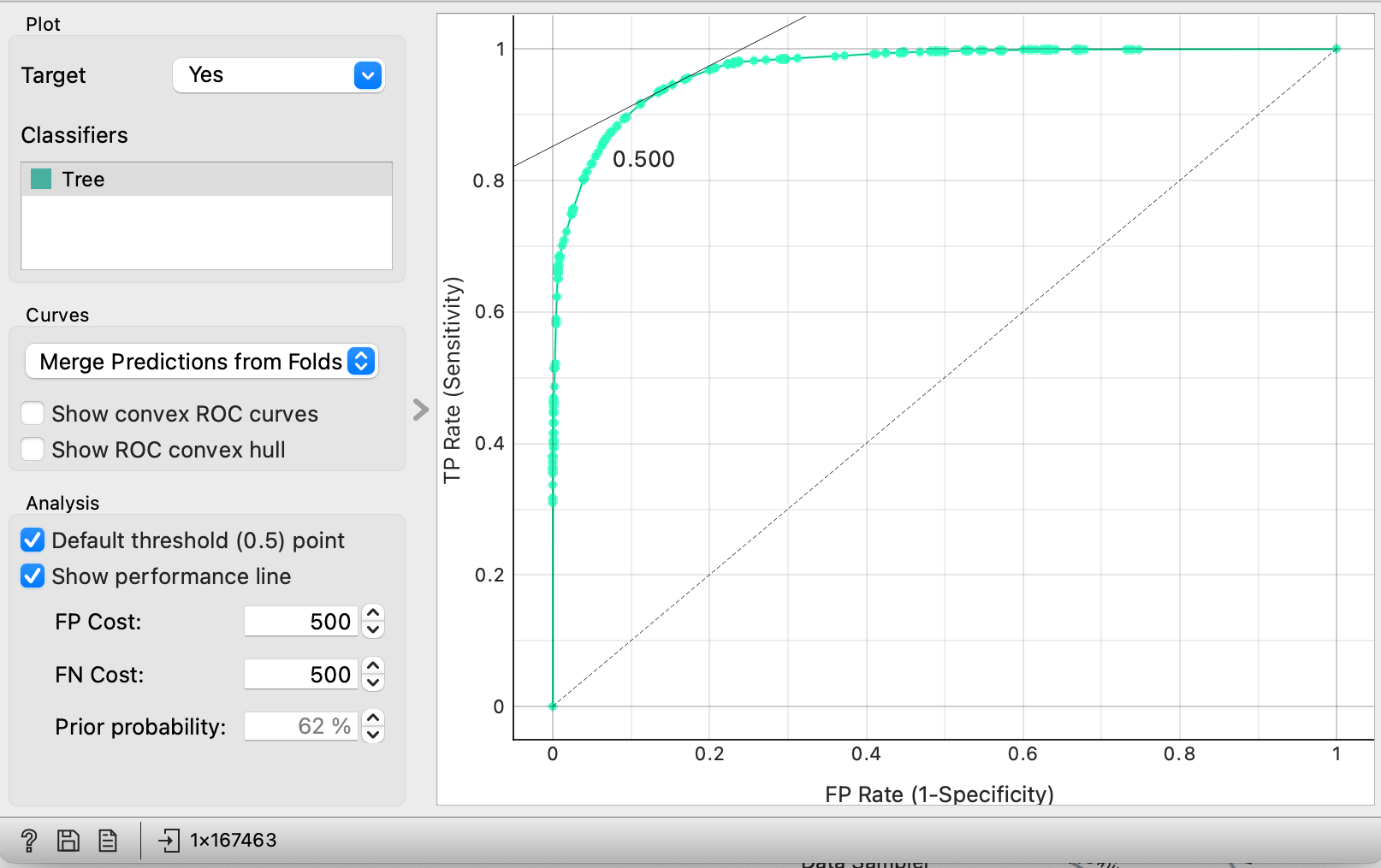


Полученные при этом результаты:

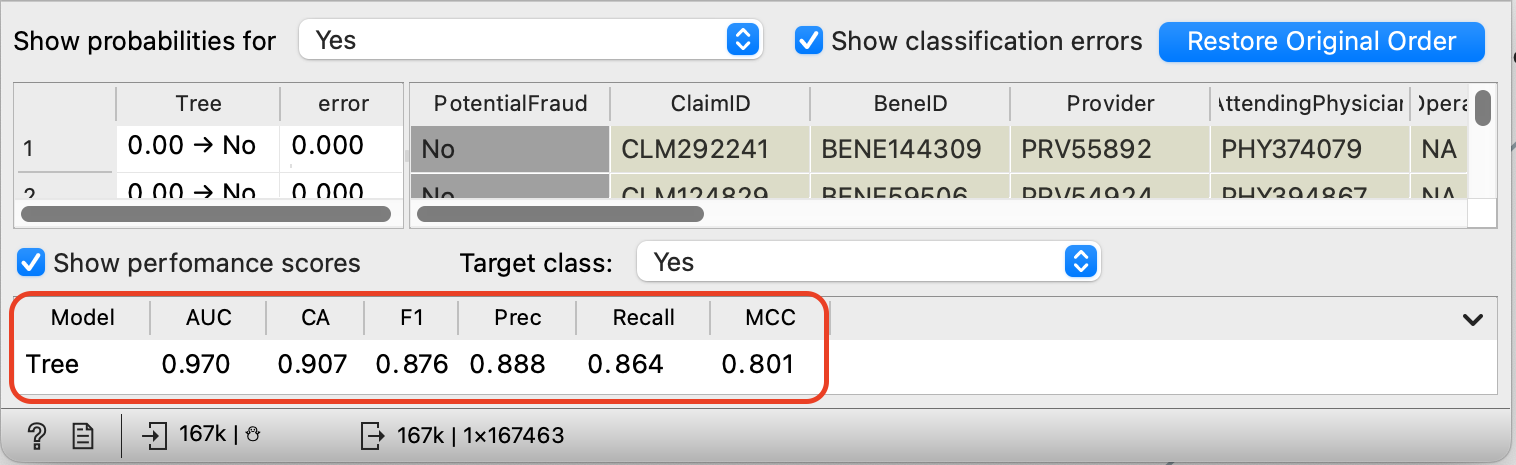
Матрица ошибок



ROC-кривая



Основные метрики



Рассчитываемые метрики модели:

*Accuracy = (TP + TN) / (TP + TN + FP + FN) = 0,9101*

*FPR = FP / (FP + TN) = 6978 / (6978 + 96617) = 0,0674*

*FNR = FN / (FN + TP) = 8669 / (8669 + 55199) = 0,1357*

Полученная точность видится достаточной для дальнейшего изучения модели.

**6. Предварительная оценка экономического эффекта**

Для того, чтобы принять решение о дальнейших шагах по внедрению модели необходимо предварительно оценить экономический эффект. Для расчета юнит-экономики принимаем следующие допущения:

1. В настоящее время средний мировой уровень обнаруженного фрода составляет 20%, что известно из открытой статистики . Полученные в результате данного исследования значения потенциального фрода составляют около 33%. Таким образом при значении Prec = 0,89, внедрение модели позволит дополнительно сэкономить сумму в размере 9,3% от зафиксированных страховых случаев.
2. Введение машинной проверки страховых случаев приведет к ложноположительному обнаружению фрода в размере 6,7% от легальных транзакций, которые составляют как минимум 67% от всего количества случаев, т.е. 4,5% от всех страховых случаев.
3. Средняя сумма выплат по каждому страховому случаю составляет 6290 эскудо.
4. Средняя сумма судебных издержек по проигранному иску составляет 10000 эскудо (оплата адвокатов, экспертиз, пошлин и т.п.)
5. Общее количество зафиксированных за год страховых случаев 558 тысяч.

Таким образом, даже в случае максимальной доли проигранных исков при ложноположительном определении фрода, экономический эффект составит:

*0,093 x 558000 x 6290 - 0,045 x 558000 x 10000 = 75,3 млн эскудо*

Кроме этого, должны высвободиться рабочие места специалистов, занятых на проверке фрода в ручную, пусть и не в полном объеме, но в достаточном для выделения ФОТ программистов, обслуживающих новый софт.

Проведенный этап предварительного моделирования с помощью nocode решения Orange Data Mining и проведенный предварительный расчет экономического эффекта позволяет перейти к более глубокому моделированию с использованием Python.

**7. Подготовка данных и моделирование с использованием Python**

В целом для написания кода была использована та же схема обработки данных, что и на предварительном этапе моделирования. Наличие дополнительных возможностей, доступных при использовании библиотек Python, позволило ввести больше атрибутов и усложнить модель путем использования градиентного бустинга. Градиентный бустинг использует деревья решений в качестве базовых алгоритмов. Этот метод хорошо показал себя при работе с табличными неоднородными данными. Общая идея алгоритма заключается в том, что предикторы применяются последовательно таким образом, что каждая последующая модель (дерево решений) снижает ошибку предыдущей.

Было принято решение реализовать метод градиентного бустинга Light Gradient Boosted Machine (LightGBM). Это реализация градиентного бустинга с открытым кодом, библиотека ПО и алгоритм машинного обучения. Метод был разработан в 2016 году и подробно описан в различных источниках.

Процесс написания кода подробно описан в прилагаемом блокноте Jupyter Notebook. Здесь остановимся на основных этапах и результатах.

После загрузки данных, была проведена их предварительная обработка, в том числе приведение категориальных переменных к бинарному виду. Далее было проведено более глубокое визуальное исследование атрибутов.

После слияния датасетов: данных о пациентах, данных о госпитализациях и данных об амбулаторном лечении, были введены атрибуты *Age* - возраст пациента на момент начала страхового случая, *AllCaseNumBene* - количество страховых случаев у пациента и *TypeCaseNumBene* - классификатор типа страхового случая.

Предварительно полученные данные о фроде были приведены к бинарному виду и объединены с общим датасетом по атрибуту *Provider*.

Далее данные были сгруппированы по атрибуту *Provider.* Были взяты средние величины по суммам возмещений, но в отличие от этапа предварительного моделирования, эти счетчики и средние были подсчитаны в разрезе каждого значения атрибута *Race*, что предположительно должно повысить точность модели. Итого были введены атрибуты *Mean1Race, …, Mean5Race* и *Count1Race, …, Count5Race*.

Далее были введены атрибуты, определяющие с каким количеством врачей взаимодействовал каждый провайдер и какое количество страховых случаев он рассмотрел с классификацией их типов: A*ttendingPhysicianCount, AllCaseNum, AmbNumRatio*.

Далее полученные счетчики и средние величины были объединены с основным датасетом и определен список *FEATURES* из 37 атрибутов, которые предполагается использовать для обучения модели:

*0 InscClaimAmtReimbursed 558211 non-null int64*

*1 DeductibleAmtPaid 557312 non-null float64*

*2 DurAmb 558211 non-null float64*

*3 ProblemDur 40474 non-null float64*

*4 Gender 558211 non-null int64*

*5 RenalDiseaseIndicator 558211 non-null int64*

*6 NoOfMonths\_PartACov 558211 non-null int64*

*7 NoOfMonths\_PartBCov 558211 non-null int64*

*8 ChronicCond\_Alzheimer 558211 non-null int64*

*9 ChronicCond\_Heartfailure 558211 non-null int64*

*10 ChronicCond\_KidneyDisease 558211 non-null int64*

*11 ChronicCond\_Cancer 558211 non-null int64*

*12 ChronicCond\_ObstrPulmonary 558211 non-null int64*

*13 ChronicCond\_Depression 558211 non-null int64*

*14 ChronicCond\_Diabetes 558211 non-null int64*

*15 ChronicCond\_IschemicHeart 558211 non-null int64*

*16 ChronicCond\_Osteoporasis 558211 non-null int64*

*17 ChronicCond\_rheumatoidarthritis 558211 non-null int64*

*18 ChronicCond\_stroke 558211 non-null int64*

*19 IPAnnualReimbursementAmt 558211 non-null int64*

*20 IPAnnualDeductibleAmt 558211 non-null int64*

*21 OPAnnualReimbursementAmt 558211 non-null int64*

*22 OPAnnualDeductibleAmt 558211 non-null int64*

*23 Age 558211 non-null int64*

*24 AllCaseNumBene 558211 non-null int64*

*25 TypeCaseNumBene 558211 non-null int64*

*26 Mean1Race 558211 non-null float64*

*27 Mean2Race 558211 non-null float64*

*28 Mean3Race 558211 non-null float64*

*29 Mean5Race 558211 non-null float64*

*30 Count1Race 558211 non-null float64*

*31 Count2Race 558211 non-null float64*

*32 Count3Race 558211 non-null float64*

*33 Count5Race 558211 non-null float64*

*34 AttendingPhysicianCount 558211 non-null int64*

*35 AllCaseNum 558211 non-null float64*

*36 AmbNumRatio 558211 non-null float64*

Далее датасет был разделен на обучающую и тестовую выборки в пропорции 70/30. Обучение проводилось с разными параметрами. В качестве основной метрики модели был выбран параметр *auc*. Также подбирались параметры *num\_leaves, max\_depths, min\_data\_in\_leaf и num\_boost\_round*. Параметры подбирались таким образом, чтобы в обучении участвовало максимально возможное количество предикторов и при этом не наблюдалось признаков переобучения. В итоге были выбраны следующие параметры:

*num\_leaves = 50*

*max\_depths = 5*

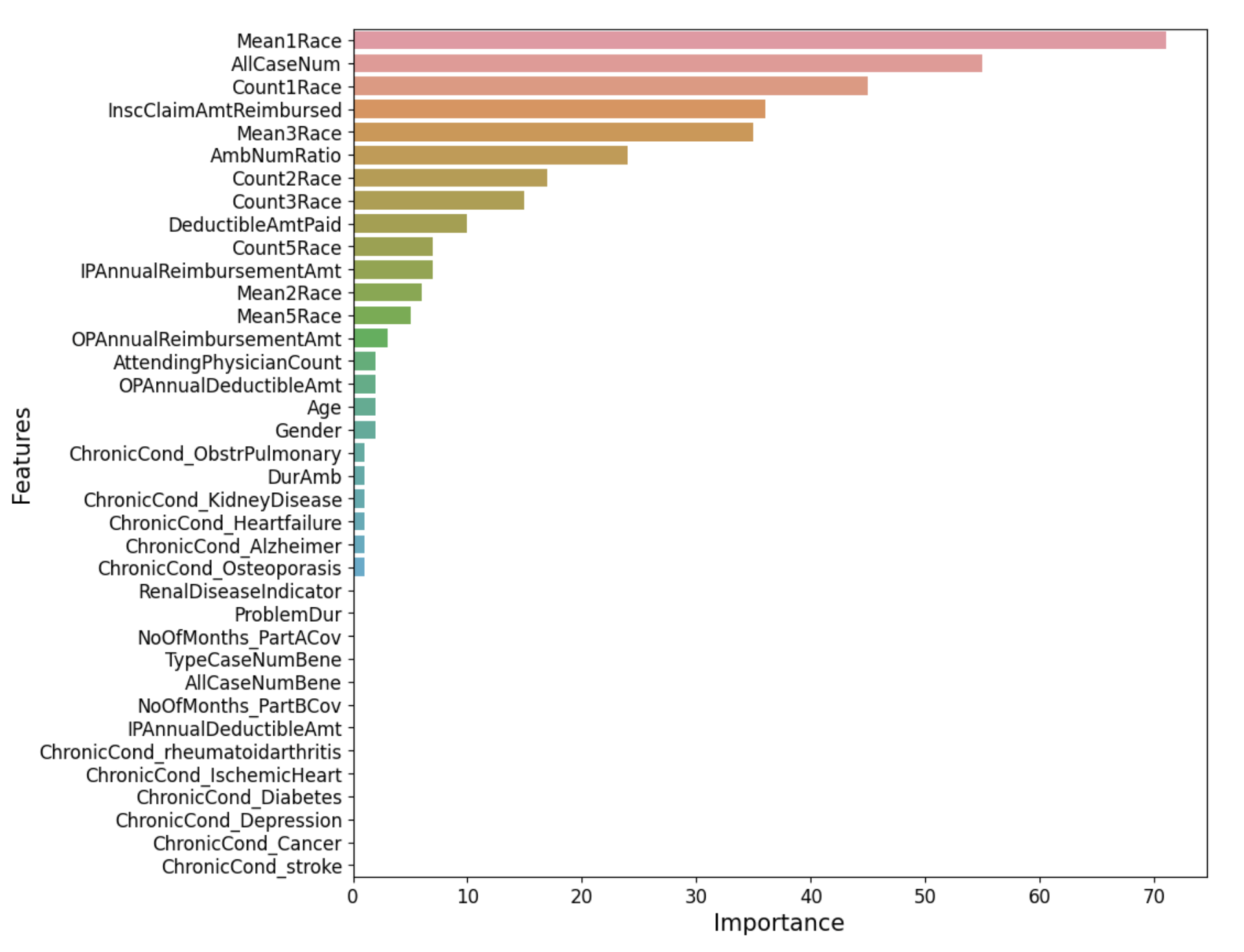
*min\_data\_in\_leaf = 50*

*num\_boost\_round = 15*

При этом была достигнута метрика

*roc\_auc\_score = 0.9730121679587646*

Для данной модели была построена диаграмма значимости предикторов:



Далее была определена пороговая вероятность принятия решения о фроде *prob*, при которой значение F1 максимально:

*prob = 0.38, f1\_max = 0.8946394976860855*

и для этой пороговой вероятности построена матрица ошибок:

*array([[97839, 6087],*

*[ 7186, 56352]])*

После этого были рассчитаны основные метрики модели:

*accuracy = 0.9207411742225194*

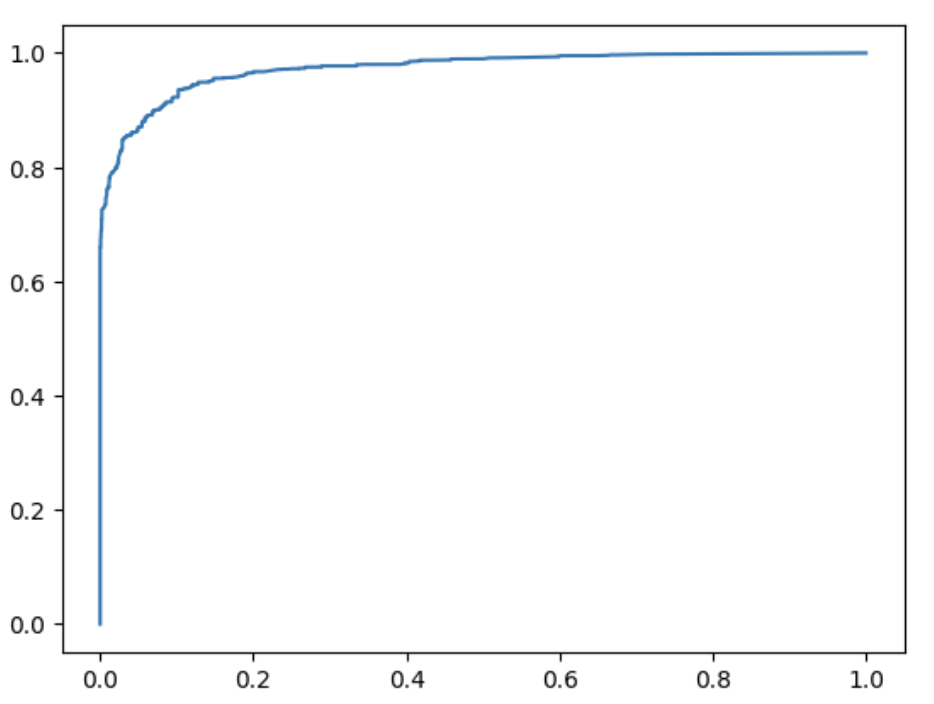
*precision = 0.902512852544083*

*recall = 0.8869023261670181*

*FNR = 0.11309767383298183*

*FPR = 0.05857052133248658*

и построена ROC-кривая:



**8. Рекомендации по внедрению**

В результате выполнения проекта было реализовано две модели автоматического определения мошеннических операций.

Модель на основе алгоритма градиентного бустинга обладает более высокой точностью, учитывает больше параметров, но не обладает свойством наглядной интерпретируемости, необходимой в случае представления доказательной базы при возникновении судебных прецедентов.

Модель на основе простого дерева решений, разработанная при помощи Orange Data Mining обладает сопоставимой точностью и может быть наглядно интерпретируема сотрудниками, не имеющими навыков в области машинного обучения.

Кроме этого были определены значимые триггеры для ручного анализа фрода, такие как наиболее часто используемые при мошеннических операциях диагнозы, прибавление несанкционированных дней для стационарного лечения и наиболее значимые предикторы, показавшие себя при машинном обучении. Также была рассчитана пороговая вероятность отнесения транзакции к фроду, при которой достигается максимальная точность модели.

Экономический эффект от внедрения модели был также рассчитан и является значимым.



Все это позволяет переходить к этапу практического внедрения, который по мнению разработчиков проекта будет включать в себя последовательное использование обеих моделей и, в случае необходимости, ручную проверку только тех транзакций, где рассчитанная вероятность отнесения операции к фроду будет близка к граничной или модели будут давать противоречивые результаты. Также предполагается ежеквартальный аудит новых датасетов на предмет изучения изменений в значимости предикторов. Обслуживание моделей и аудит будет проводиться вновь приглашенными сотрудниками IT подразделения, при этом сотрудников, занятых на ручной проверке данных предполагается соразмерно сократить.