Минобрнауки России

Федеральное государственное бюджетное образовательное учреждение высшего образования «Нижегородский государственный архитектурно-строительный университет» (ННГАСУ)

институт информационных технологий

кафедра прикладной информатики и статистики

Расчетно-графическая работа

по дисциплине «Анализ и обработка данных»

|  |  |  |
| --- | --- | --- |
| Выполнил студент:  ПРИ22/2 |  | Фролов Никита  Максимович |
| Проверила отчет:  к.ф.-м.н, проф. |  | Прокопенко Наталья Юрьевна |

Нижний Новгород

2024

# Содержание

[Содержание 2](#_Toc185120296)

[Введение 3](#_Toc185120297)

[Исследование распределения по данным 4](#_Toc185120298)

[Аудит 6](#_Toc185120299)

[Проверка качества данных 6](#_Toc185120300)

[Биннинг 12](#_Toc185120301)

[Корреляционный анализ 14](#_Toc185120302)

[Гипотезы 16](#_Toc185120303)

[Разделение данных по обучающую и тестовую выборку 17](#_Toc185120304)

[Ансамбль моделей 21](#_Toc185120305)

[Кластеризация 31](#_Toc185120306)

[Заключение 36](#_Toc185120307)

[Список литературы 38](#_Toc185120308)

# Введение

Анализ данных сегодня — это важный инструмент, который помогает находить полезную информацию и принимать правильные решения в самых разных областях, включая работу с жалобами населения. С помощью анализа можно выяснить, какие факторы влияют на удовлетворение жалоб, и использовать эти знания для повышения шансов добиться своего.

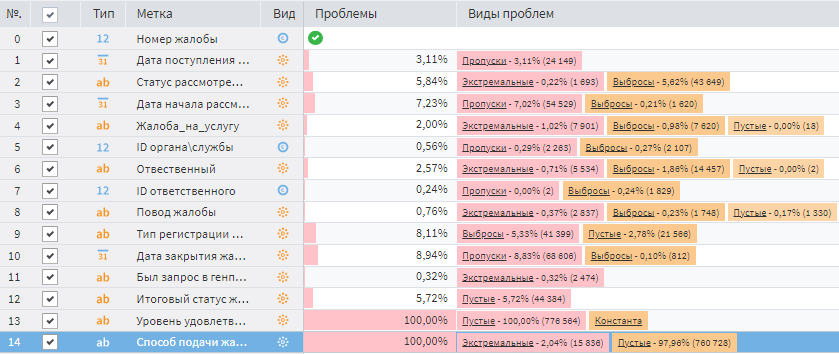
В рамках этой работы я провожу исследование с использованием методики KDD (Knowledge Discovery in Databases). Для анализа я выбрал набор данных, содержащий данные обо всех жалобах населения на государственные ведомства начиная с 2018 года, которые смог достать на сайте ГосУслуг (ГосЖалобы). В этом наборе данных представлены разнообразные данные, включая ведомство, причину обращения, форму отправки жалобы, разные даты, итог процесса жалоба-обработки и др.

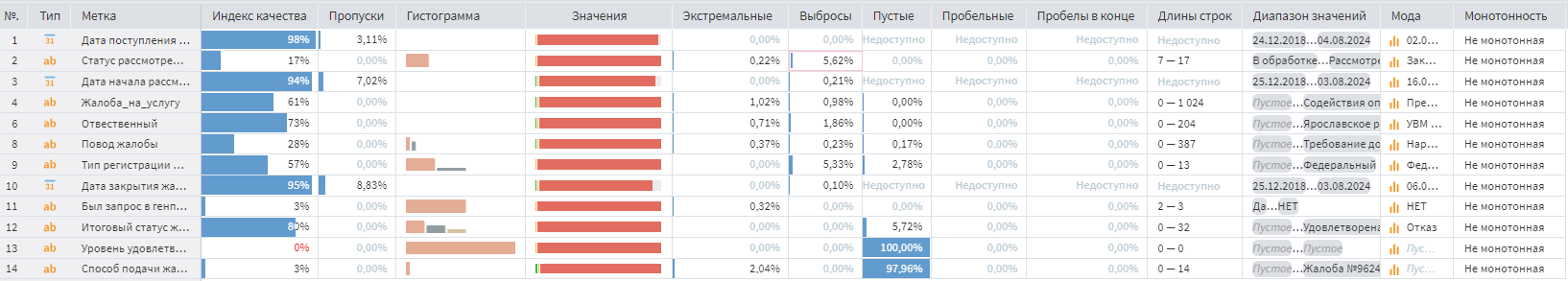
Цель моей работы — изучить, как разные факторы влияют на удовлетворение жалоб, а также выделить наиболее значимые признаки. Для выполнения задачи я буду использовать отечественную платформу Loginom, что позволит проанализировать данные на всех этапах: от их проверки и преобразования до построения моделей и анализа результатов.

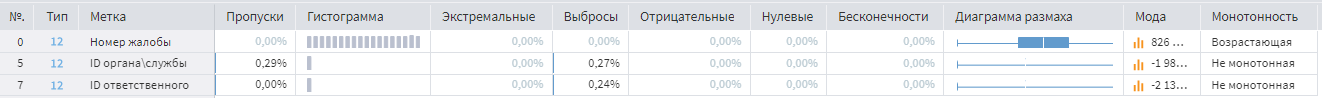
Эта работа поможет мне не только разобраться в процессе анализа данных, но и понять, как можно применять полученные результаты для решения реальных задач, например, в сфере подачи жалоб.

# Исследование распределения по данным

Для начала оценим исходные данные с помощью визуализатора «Качество данных». Результат на рисунках 1-3.

  
Рисунок 1. Качество данных

  
Рисунок 2. Дискретные

  
Рисунок 3. Непрерывные

Исходя из этих результатов видно, что данные нужно будет обработать.

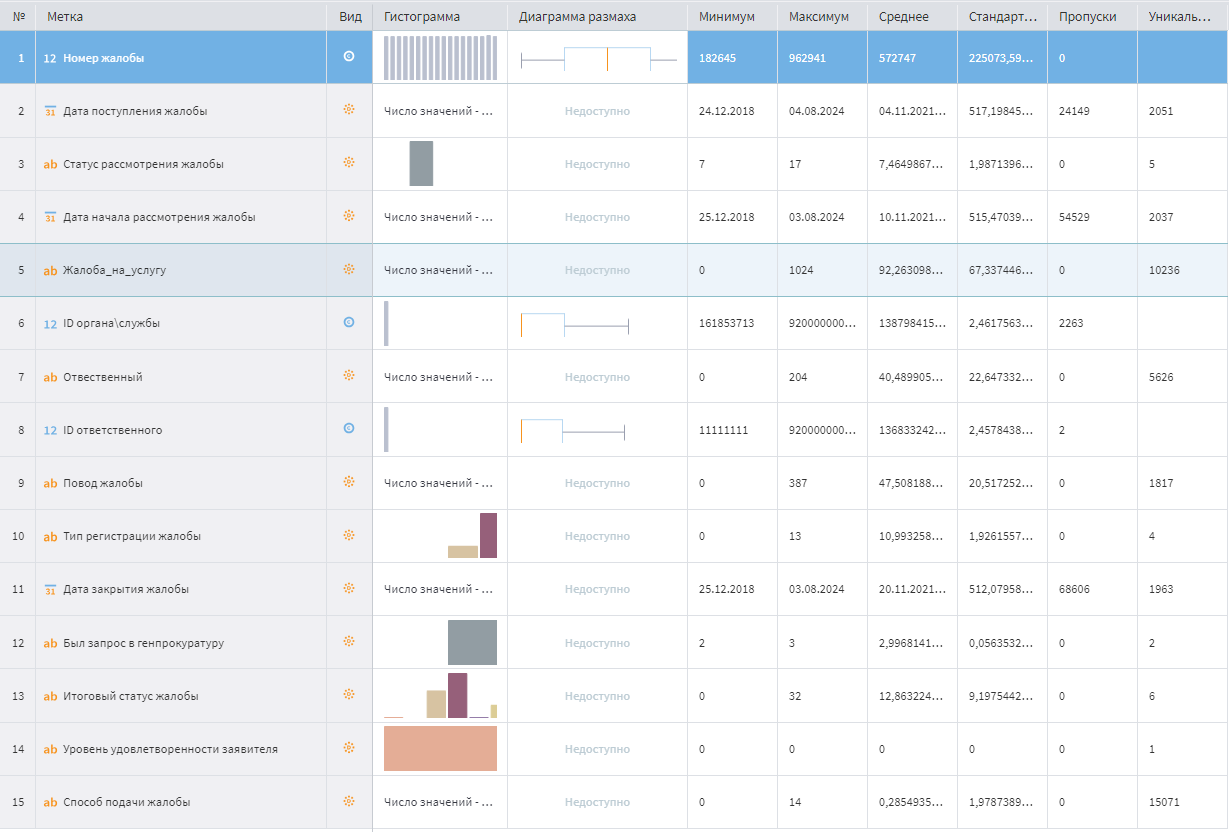


Рисунок 4. Визуализатор статистика

# Аудит

## **Проверка качества данных**

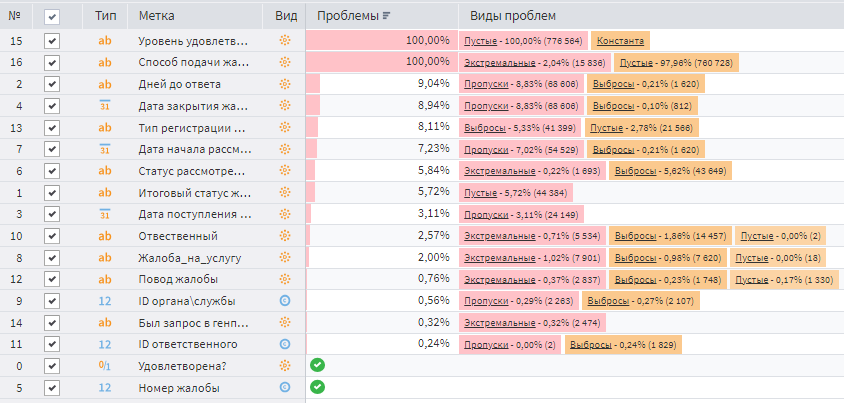
Первым делом создам параметр (выходное поле) «Удовлетворена?» с помощью компонента «Калькулятор», который отбирает те жалобы, у которых итоговый статус «Удовлетворена» (у поля «итоговый статус жалобы» есть несколько значений, а не просто да\нет). Формула представлена на рисунке 6.

Заодно сразу добавим еще один параметр «Дней до ответа». Его формула на рисунке 5.

  
Рисунок 5. Дней до ответа

  
Рисунок 6. Выходной параметр

Еще раз воспользуемся визуализатором «Качество данных» но теперь на выходе компонента «Калькулятор». Результат на рисунке 7.

  
Рисунок 7. Качество данных

Из показаний визуализатора видно, что пропуски есть, и они есть почти в каждом показателе набора данных - их нужно обработать. Также присутствуют экстремальные значения и выбросы, с которыми тоже нужно поработать, так как они будут влиять на дальнейший анализ.

Поскольку в моих данных много проблем, то я буду работать с каждым полем отдельно, начнем с пропусков и пустых значений:

1. Уровень удовлетворенности (100% пропуски) – удалить полностью столбец. Нет смысла пытаться что-любо сделать с этим столбцом, его значение – константа на пропусках. Видимо он был введен «на всякий случай».
2. Способ подачи жалобы (98% пустые) – тут есть два варианта (почта или сайт). В этих данных, если пользователь отправил жалобу по почте, то ей присваивается номер (Жалоба №…), а если через сайт – поле пустое. Я переделаю этот параметр, чтобы было именно 2 варианта значения.
3. Дней до ответа (8,83% пропуски) – это связано с тем, что это вычисляемое поле, и оно целиком зависит от поля ниже, поэтому для них будет одно и тоже действие – удалить пропуски.
4. Дата закрытия жалобы (8,83% пропуски) – как и сказано выше – удалить пропуски.
5. Тип регистрации жалобы (2,78% пустые) – пропусков немного, поэтому просто удалить.
6. Дата начала рассмотрения (7,2% пропуски) – удалить, да, пропусков много, но и данных около миллиона, потери будут не слишком весомы.
7. Итоговый статус жалобы (5,72% пустые) – удалить
8. Дата поступления жалобы (3,11% пропуски) – если посмотреть на выборку данных по этому столбцу, то можно заметить, что данные отсортированы по нему, а значит заполним наиболее вероятным.
9. Ответственный (пара пропусков) – просто удалить
10. Жалоба на услугу (аналогично 9) – просто удалить
11. Повод жалобы (0,17% пустые) – удалить
12. ID органа\службы (0,29% пропуски) – удалить
13. ID ответственного (пара штук пропусков) - удалить

Итак, ниже представлены скриншоты проделанной работы:

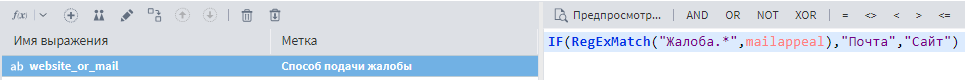
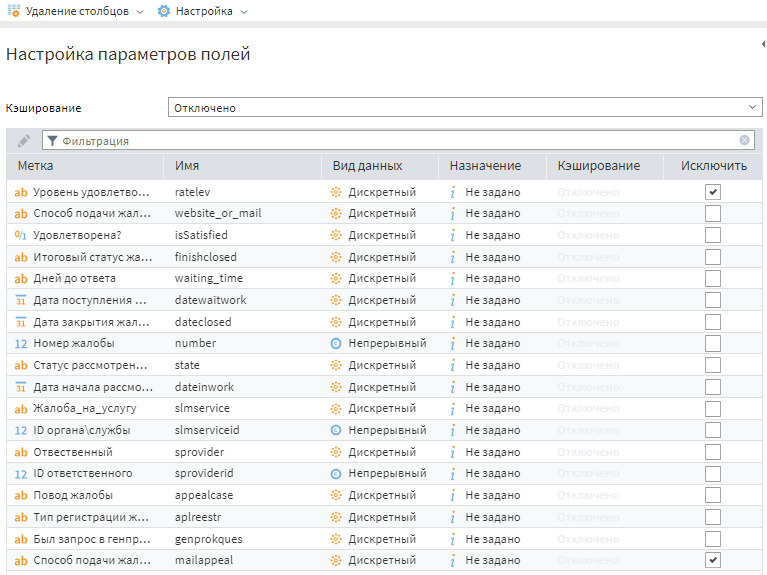
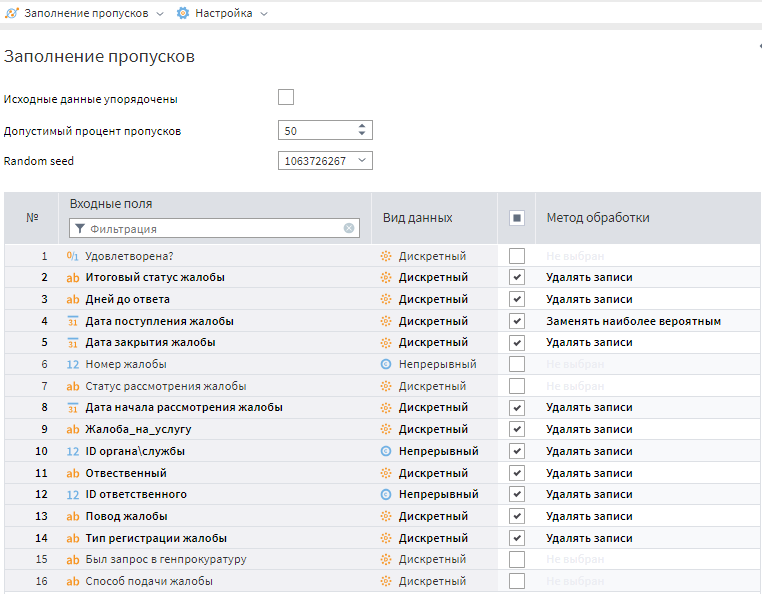


Рисунок 8. Работа с пунктом 2

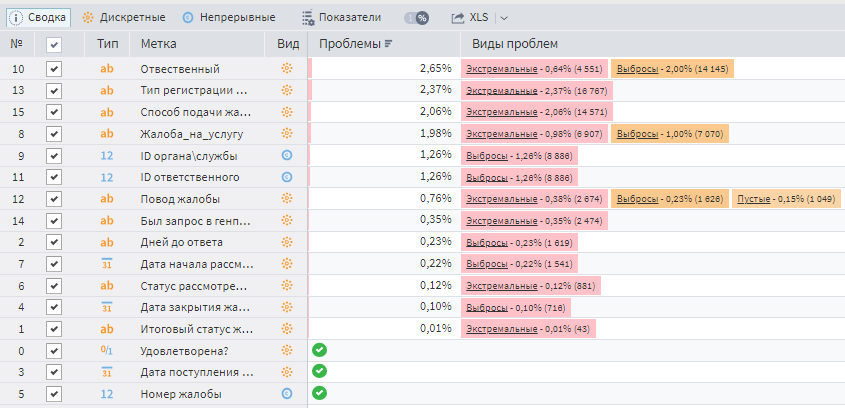
Далее, поскольку у нас теперь новый параметр «Способ подачи жалобы», то старый удалим по ненадобности.

  
Рисунок 9. Работа с пунктом 1 и старым полем

Далее переходим к остальным пунктам:

  
Рисунок 10. Работа с оставшимися пунктами

После проделанной обработки подведем итоги, взглянув на визуализатор «Качество данных».

  
Рисунок 10. Качество данных после предобработки

Проанализируем полученную информацию:

1. Ответственный (2% выбросы) – это нормально, ведь данные со всей страны и в какой-нибудь деревне люди тоже могут жаловаться на местное управление
2. Тип регистрации (2,37% экстремальные) – это значение «Муниципальные», так что все в порядке
3. Способ подачи жалобы (2,06% экстремальные) – это как раз значение «Почта», оставляем
4. Жалоба на услугу (1% выбросы) – предоставляемых услуг большое количество, просто в списке есть и редкие – оставляем
5. …

Далее аналогичная ситуация с оставшимися столбцами.

Перейдем к дубликатам и противоречиям:

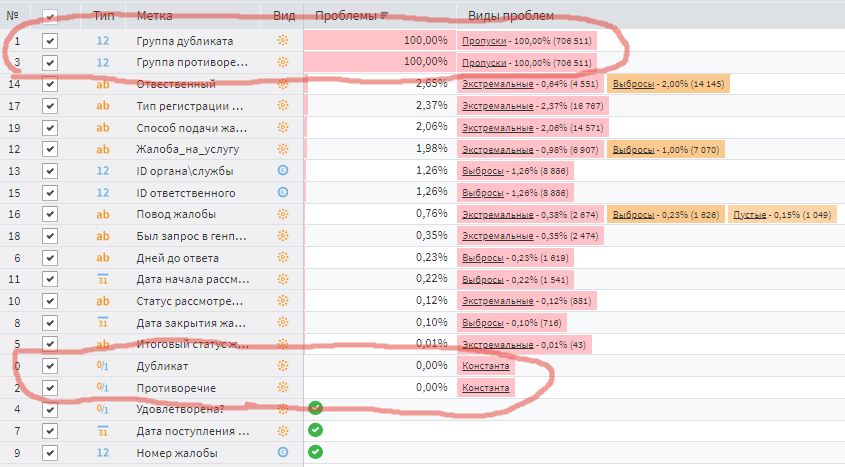


Рисунок 11. Дубликаты и противоречия

Как видно из визуализатора «Качество данных» дубликатов и противоречий тут нет. (Пояснение :)

В качестве входных данных я выбрал ВСЕ кроме результирующего поля «Удовлетворена?». Возникает вопрос, почему я также использовал «Номер жалобы» как входное поле. Ответ: когда подаешь жалобу ей присваивается отдельный Уникальный номер, по которому ты можешь отслеживать состояние поданной жалобы. Т.е. не может быть несколько ссылок на одну и ту же жалобу – разработчики об этом позаботились. Поэтому высчитывая дубликаты нужно учитывать и это поле.

## **Биннинг**

Выполним биннинг с помощью компонента «Конечные классы». Все поля кроме «Итоговый статус жалобы» (наше результирующее поле высчитывается за счет него, нет смысла проверять его значимость) назначим входными, а поле «Удовлетворена?» назначим выходным. Настройки компонента представлены на рисунке 12.

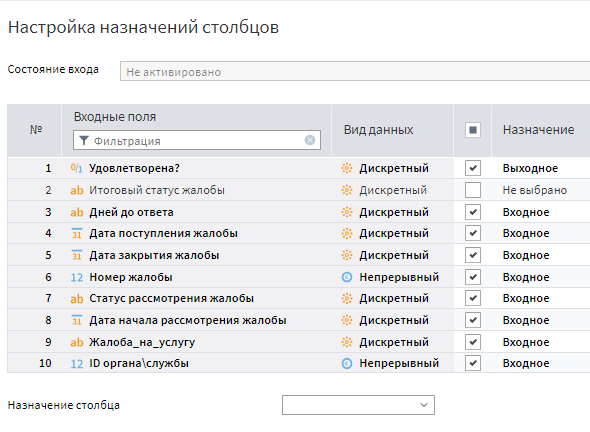


Рисунок 12. Настройка конечных классов

По информационному индексу IV определяем значимость переменных. Пример результата визуализатора представлен на рисунке 13.

Очень высокая:

-

Высокая:

* Жалоба на услугу
* ID органа\службы
* Ответственный
* ID ответственного
* Тип регистрации жалобы

Средняя:

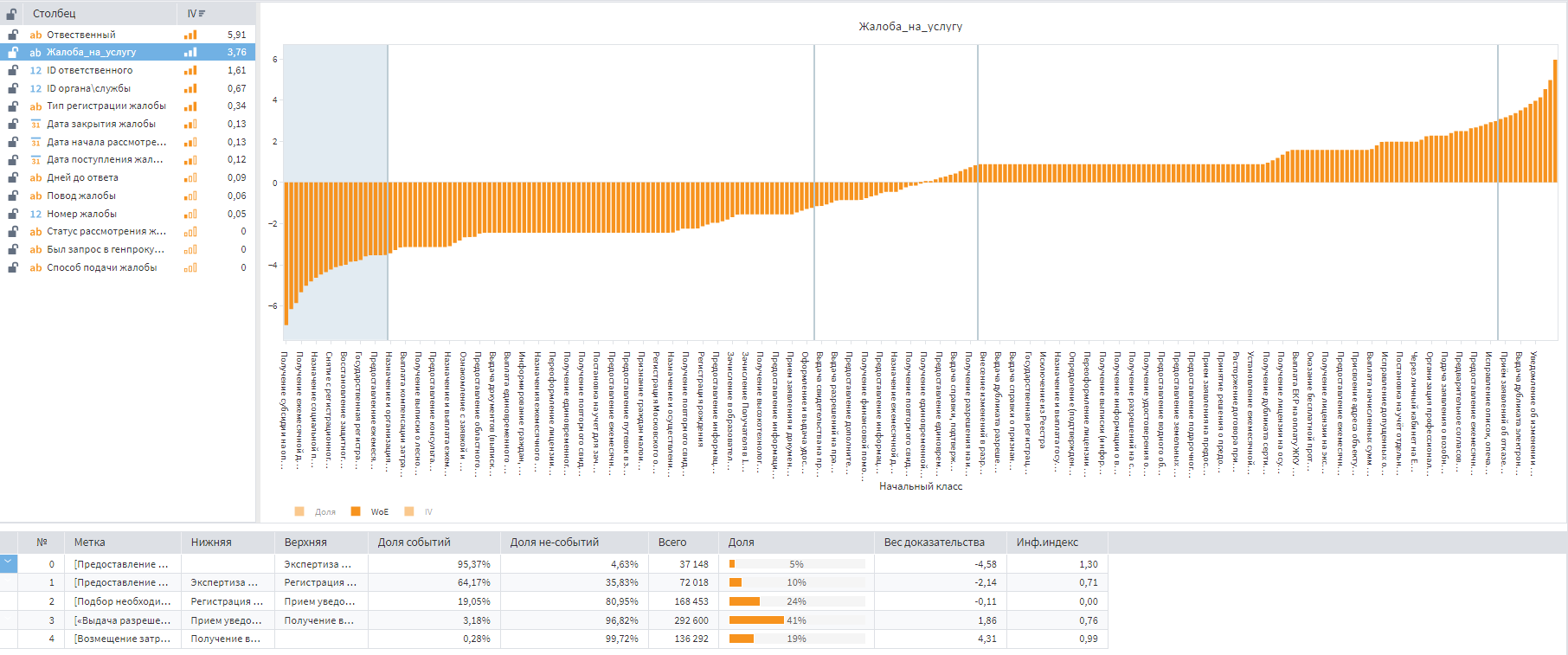
* Дата поступления жалобы
* Дата закрытия жалобы
* Дата начала рассмотрения жалобы

Низкая:

* Дней до ответа
* Номер жалобы
* Повод жалобы

Отсутствует:

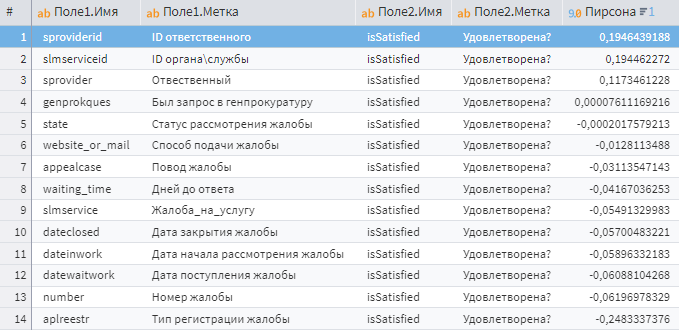
* Статус рассмотрения жалобы
* Был запрос в ген прокуратуру
* Способ подачи жалобы

  
Рисунок 13. Визуализатор «Конечные классы»

По результатам биннинга можно исключить переменные с низкой значимостью и её отсутствием.

## **Корреляционный анализ**

Корреляционный анализ выполняется с помощью компонента «Корреляционный анализ». Ориентируемся на коэффициент Пирсона. Результирующие данные представлены на рисунке 14.

  
Рисунок 14. Пирсона

Анализ показал, что у следующих элементов есть связь выходным полем «Удовлетворена?», но она незначительна:

* ID ответственного
* ID органа\службы
* Ответственный
* Был запрос в генпрокуратуру

Причиной этого может быть факт того, что анализ Пирсона не всегда распознает сложность взаимоотношений меж переменными. К тому же связь может быть не линейной, а более сложной.

Следующий шаг – удаление ненужных столбцов с помощью компонента «Параметры полей».

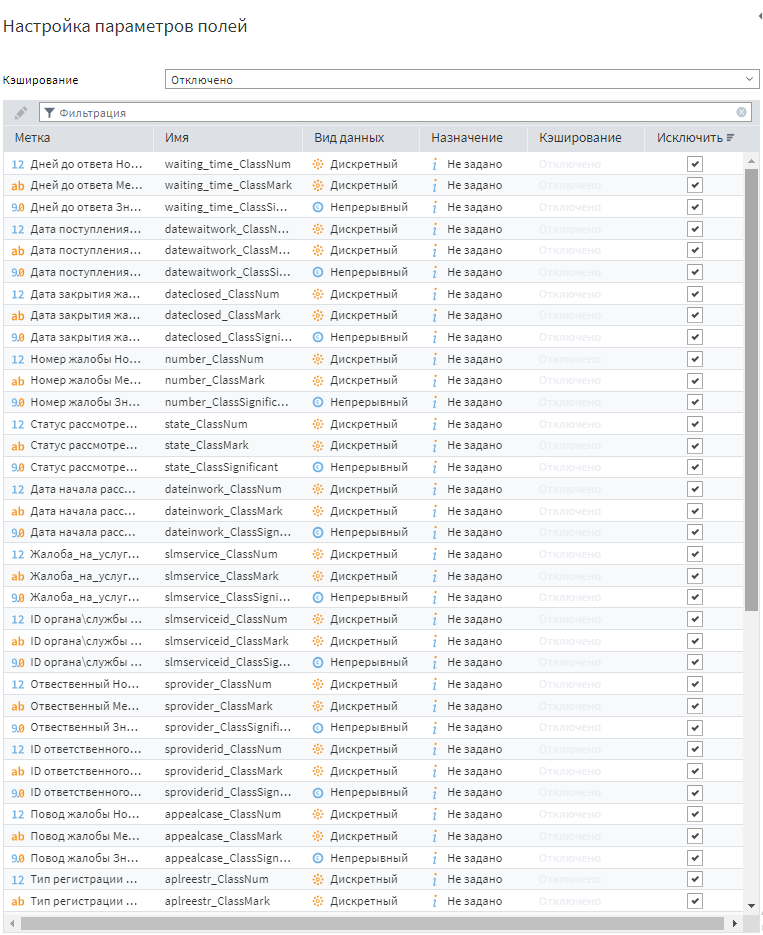


Рисунок 15. Удаление лишнего

# Гипотезы

Проведя биннинг и корреляционный анализ над набором данных можно выдвинуть гипотезы о том, что влияет на удовлетворение жалобы.

**Гипотезы:**

На положительный отклик на жалобу во многом влияет **на кого эта самая жалоба была подана** (в некоторых случаях человек может не разобраться на кого именно подавать жалобу и подает куда, как он считает, надо, а ее просто перенаправляют (в нашем случае это будет считаться как Удовлетворена = НЕТ), также возможно, какие-либо организации могут быть коррумпированы, и удовлетворять только те жалобы, которые захотят (за доп. плату) (ВОЗМОЖНО!) и т.п.)

Также крайне важна и сама **услуга**, на которую подают жалобу (некоторые люди могут до конца пытаться получить, например, меры социальной поддержки от государства, хотя сами не подходят по критериям, из-за этого может быть увеличенное число отказов и т.п.)

Также важно **на каком уровне** твоя жалоба (например, кто-то хочет, что называется, с самого начала «Дойти до Президента», а так нельзя, все должно идти поэтапно, как бы снизу-вверх)

Также вероятно, **сколько дней** рассматривалась жалоба косвенно влияет на успех (если жалоба было закрыта более чем за неделю, но менее чем за 30 дней, это значит вашим вопросом действительно занимаются, а не отклоняют его в связи с неправильно заполненной формой и т.п.).

# Разделение данных по обучающую и тестовую выборку

Обучающая выборка — структурированный набор данных, применяемый для обучения аналитических моделей. Каждая запись обучающего множества представляет собой обучающий пример, содержащий заданное входное воздействие и соответствующий ему правильный выходной (целевой) результат.

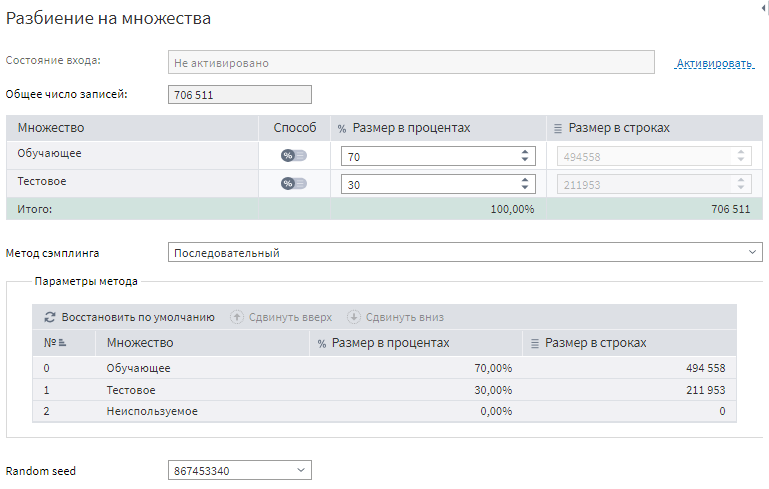
Тестовое — подмножество обучающей выборки, содержащее тестовые примеры, т.е. примеры, использующиеся не для обучения модели, а для проверки его результатов.

Существует несколько видов сэмплинга: случайный, равномерный случайный, стратифицированный, последовательный, отбор со смещением.

Для разделения набора данных на выборки воспользуюсь компонентом «Разбиение на множества» и воспользуюсь последовательным методом сэмплинга, а именно возьму первые 30% строк для тестовой выборки.

**(НО! См. след. Стр.)**

Такое процентное соотношение соответствует нормам сэмплинга, где обычно делят на 20% и 80% или 30% и 70%. Настройки компонента представлены на рисунке 16.

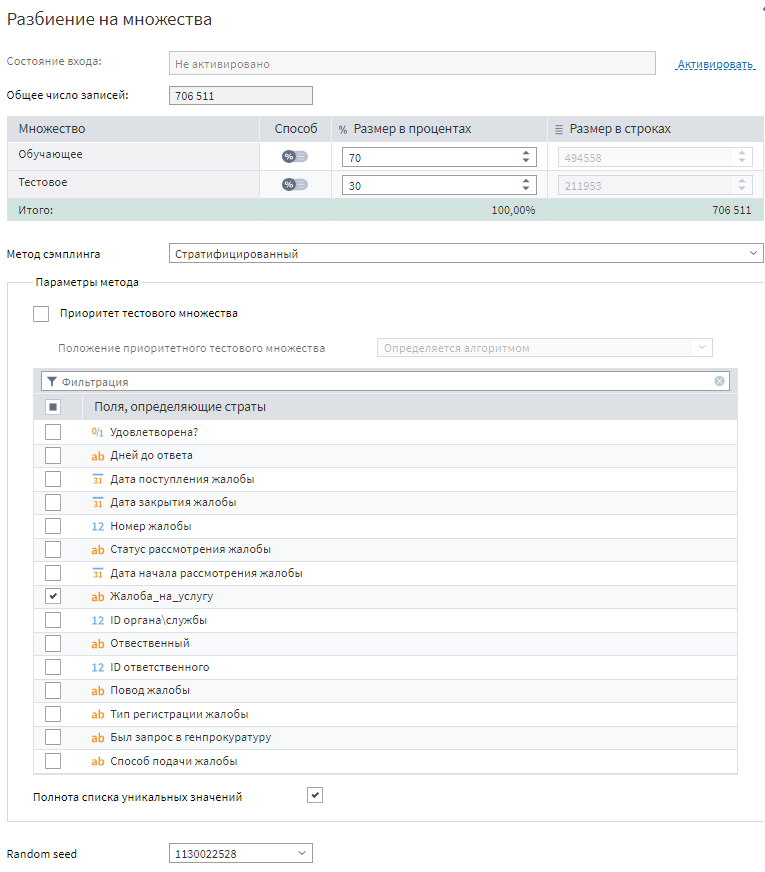
  
Рисунок 16. Разбиение на множества

На этом моменте начинаются основные трудности моей работы. Возникает ситуация: при последовательном сэмплинге некоторые модели попросту не могут работать с тестовой выборкой ввиду появления в ней новых, не попавших в обучающее множество, значений. На этом этапе я провел достаточное количество времени и пришел к выводу:

Метод сэмплинга должен быть СТРАТИФИЦИРОВАННЫЙ. В моих данных большое количество уникальных значений у некоторых полей. Причем, для каждого такого поля нужно подготовить свою выборку (тестовую и обучающую), т.е. в работе будет 2 разбиения на множества. Это наилучший вариант, хотя и он не без изъянов.

Ввиду того, что данных около 700 000, компонент «Разбиение на множества» делает свою работу ощутимое количество времени, это при учете, что у меня очень мощное аппаратное обеспечение на ПК.

Поэтому прошу проверяющего в случае необходимости набраться немного терпения и чуть-чуть подождать (в моем случае это время было   
~17 секунд).

  
Рисунок 17. Разбиение на множества №1 (Все жалобы на услугу)

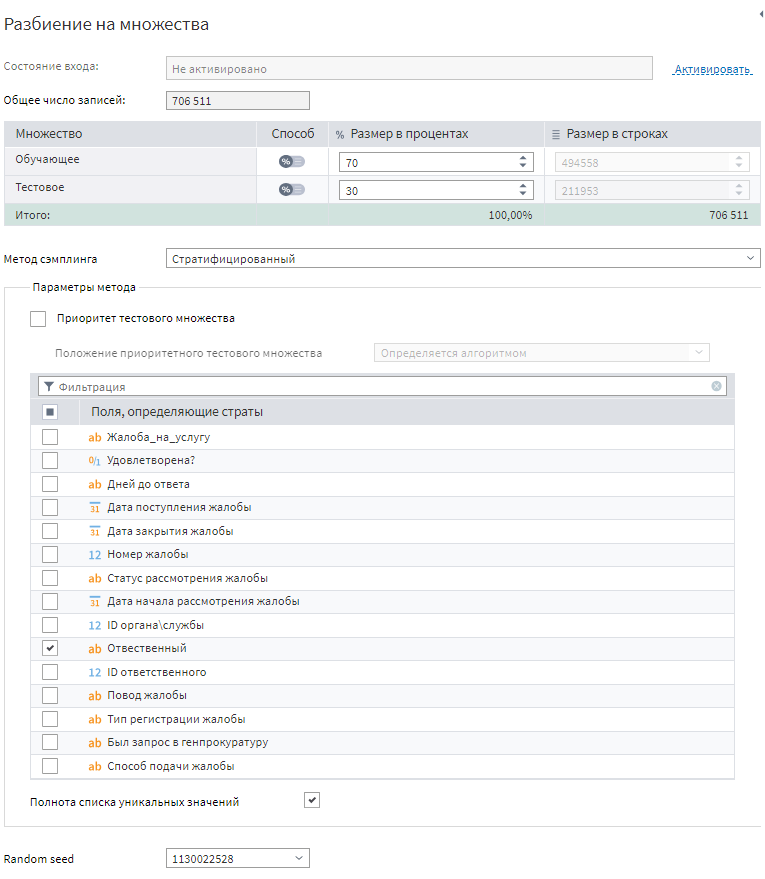


Рисунок 18. Разбиение на множества №2 (Все типы ответственных)

# Ансамбль моделей

К тестовым выходным наборам применим несколько компонентов «Логистическая регрессия».

Компоненты с большим временем загрузки:

* Жалоба на услугу (~60 сек)
* Все типы ответственных (~15 сек)

**1. Логистическая регрессия 1 (жалоба на услугу)**

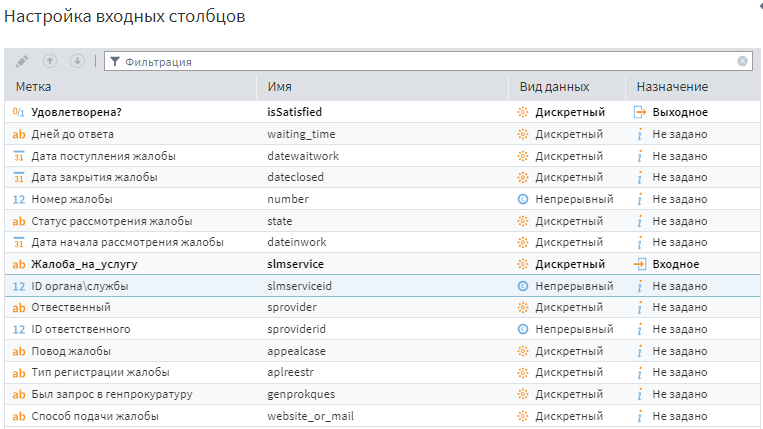


Рисунок 19. Логистическая регрессия 1 (жалоба на услугу)

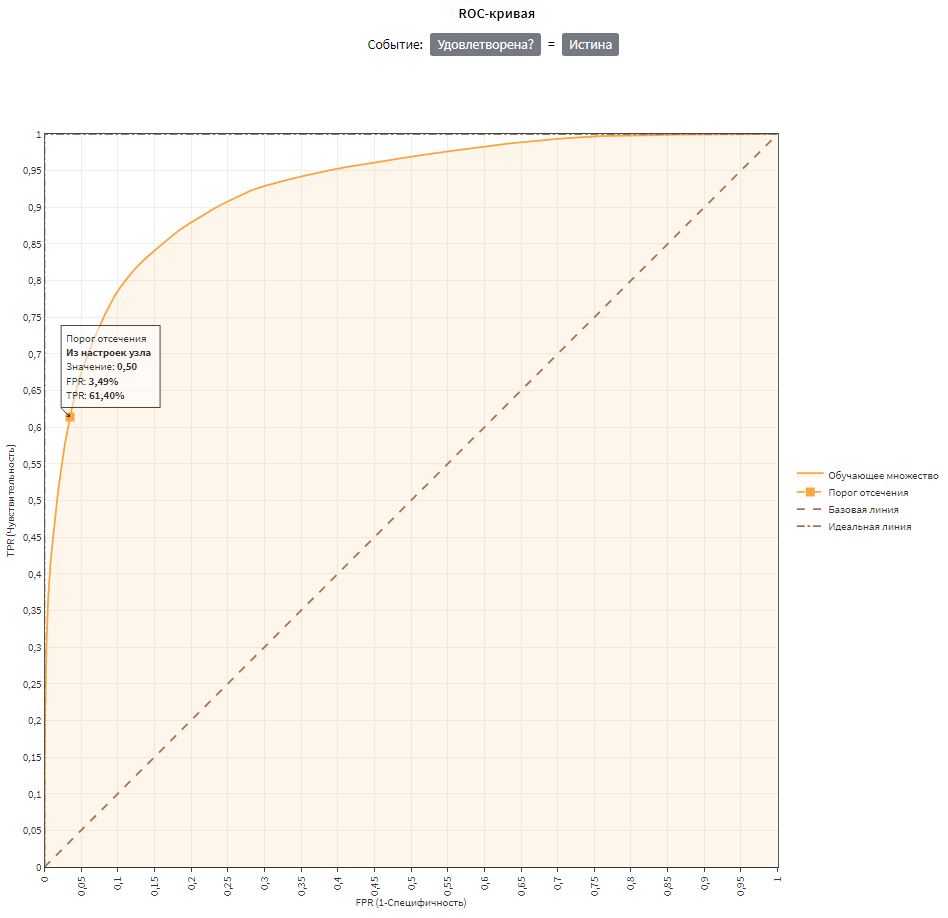


Рисунок 20. Логистическая регрессия 1 (жалоба на услугу)

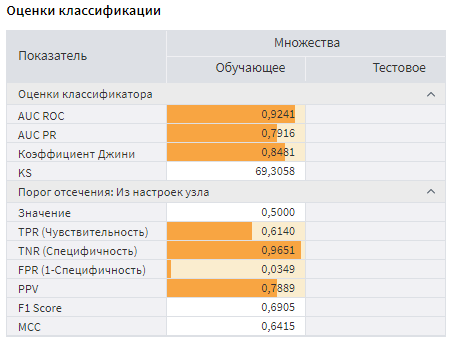


Рисунок 21. Логистическая регрессия 1 (жалоба на услугу)

**2. Логистическая регрессия 2 (Дней до ответа)**

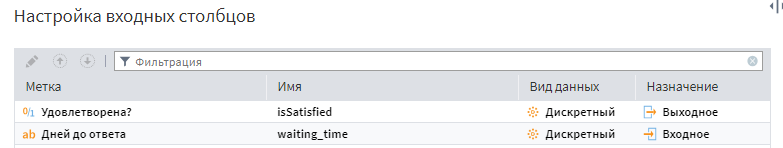
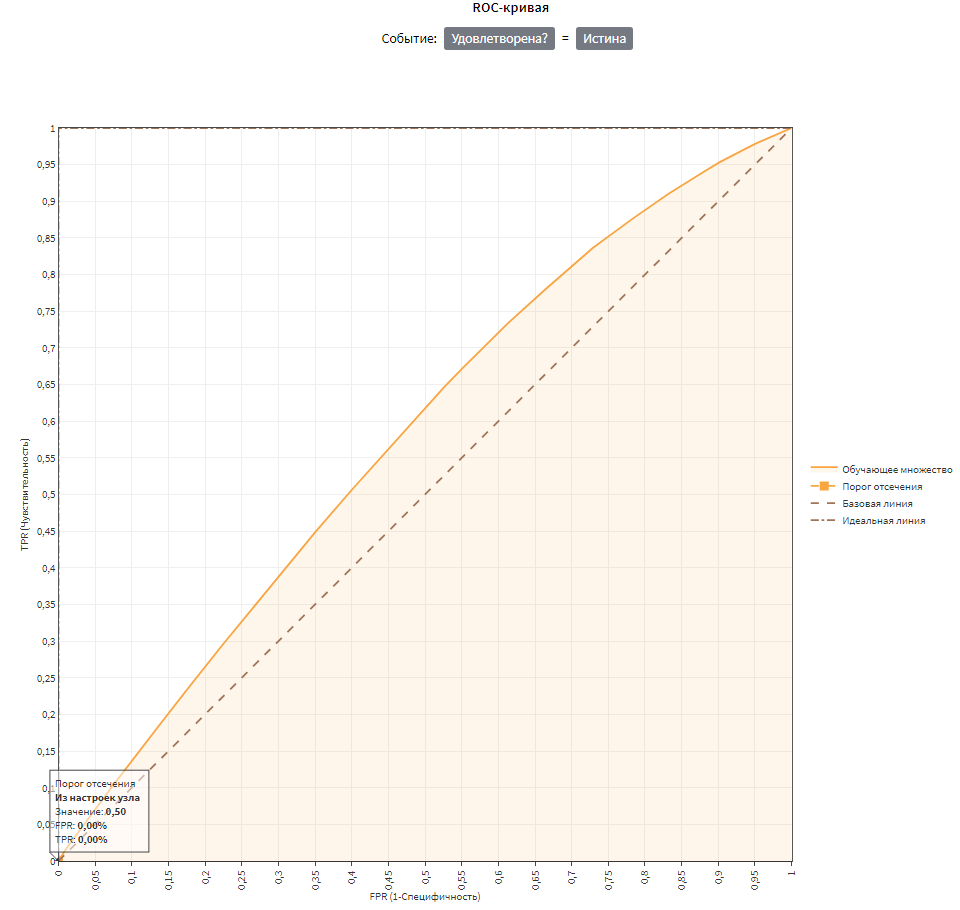
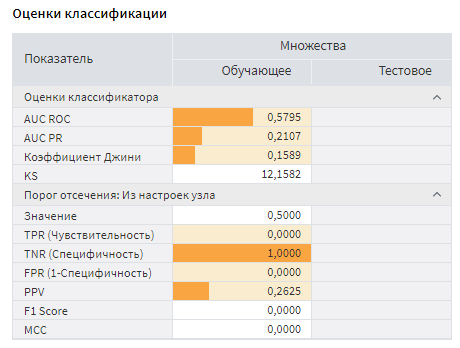
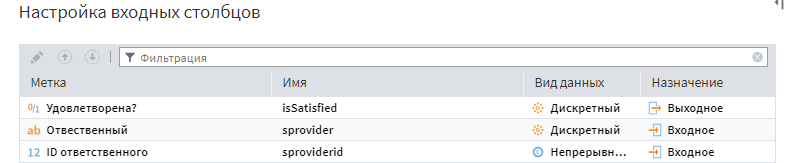


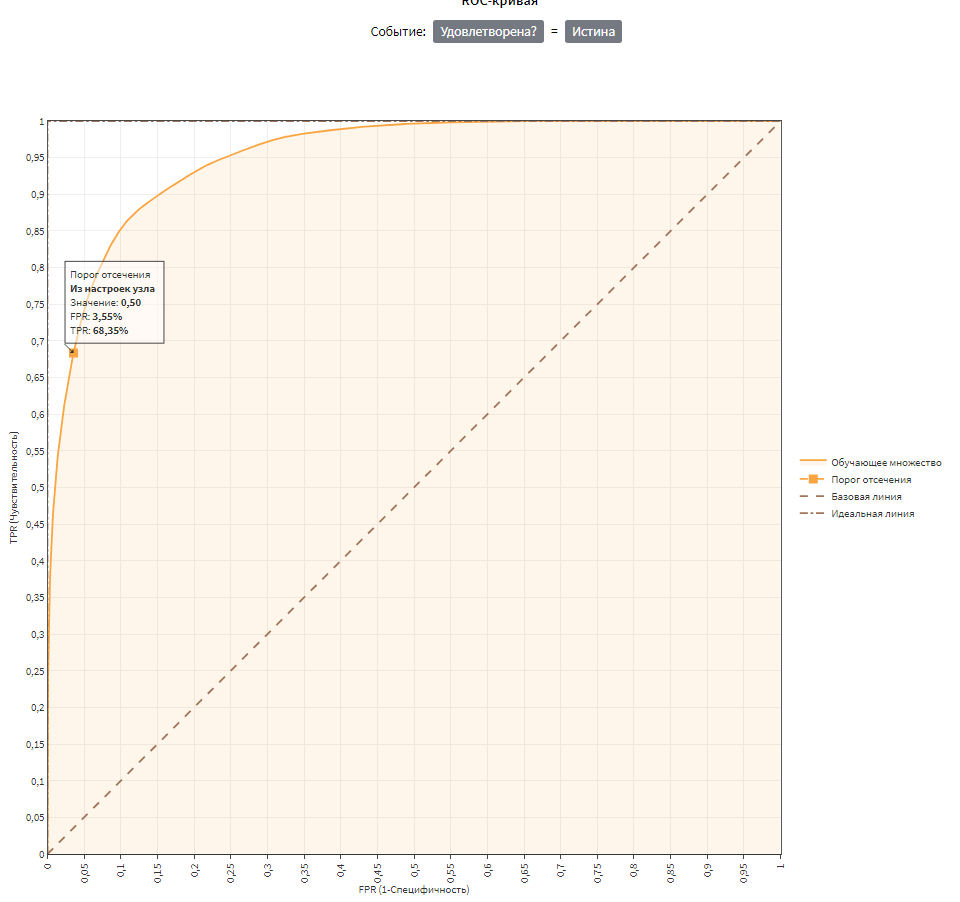
Рисунок 22. Логистическая регрессия 2 (Дней до ответа)

  
Рисунок 23. Логистическая регрессия 2 (Дней до ответа)

  
Рисунок 24. Логистическая регрессия 2 (Дней до ответа)

**3. Логистическая регрессия 3 (Жалоба на какое ведомство)**

  
Рисунок 25. Логистическая регрессия 3 (Жалоба на какое ведомство)

  
Рисунок 26. Логистическая регрессия 3 (Жалоба на какое ведомство)

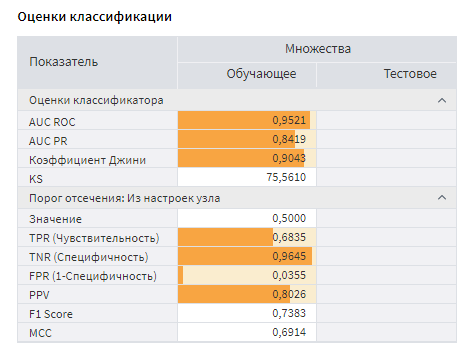
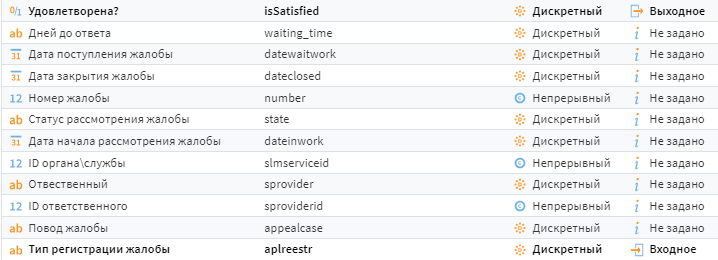
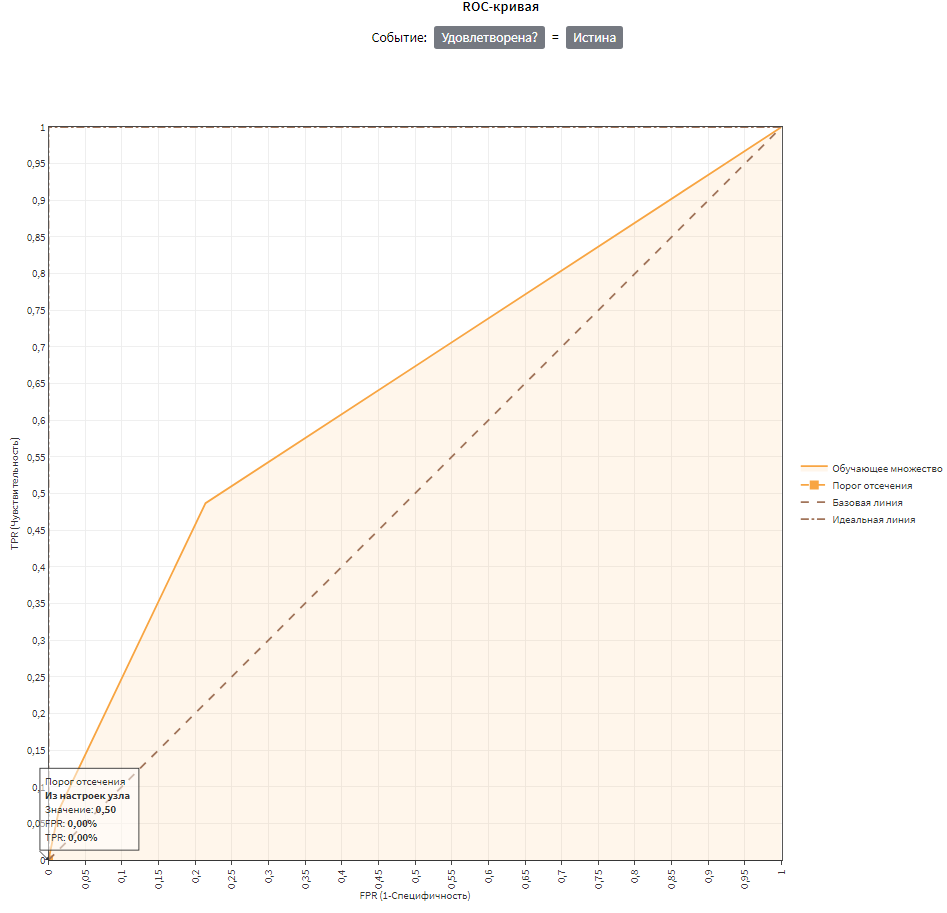
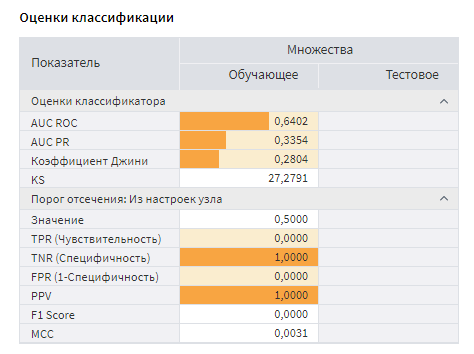


Рисунок 27. Логистическая регрессия 3 (Жалоба на какое ведомство)

**4. Логистическая регрессия 4 (Тип регистрации жалобы)**

  
Рисунок 28. Логистическая регрессия 4 (Тип регистрации жалобы)

  
Рисунок 29. Логистическая регрессия 4 (Тип регистрации жалобы)

  
Рисунок 30. Логистическая регрессия 4 (Тип регистрации жалобы)

Итак, все модели построены, пройдемся же по их предсказательности:

1. **Жалоба на услугу (AUC ROC = 0,9241) – отличная модель**
2. **Дней до ответа (AUC ROC = 0,5795) – случайная модель**
3. **Жалоба на какое ведомство (AUC ROC = 0,9521) – почти идеальная модель**
4. **Тип регистрации жалобы (AUC ROC = 0,6402) – случайная модель**

У моделей 1 и 3 очень высокая предсказательная способность, что говорит о том, что в тестовых данных должны происходить очень хорошие прогнозы результата.

Для каждой модели я сделал «Выполнение узла» на тестовых данных:

1. **AUC ROC = 0,9127 – отличная модель**
2. **AUC ROC = 0,5802 – случайная модель**
3. **AUC ROC = 0,9479 – отличная модель**
4. **AUC ROC = 0,6417 – случайная модель**

**Сравнение ошибок:**

Ошибки моделей на обучающем наборе данных варьируются от 42,05% до 4,79%, а на тестовом – от 41,98% до 5,21%. Это говорит о том, что отдельно взятые модели показывают очень хороший результат как на обучении, так и на тестах.

Разница в AUC ROC меж обучающими и тестовыми моделями крайне низка

(~1%), это говорит о стабильной работе моделей.

Вывод:

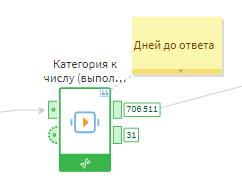
Модели 1 и 3 отлично справляются с поставленной задачей.

# Кластеризация

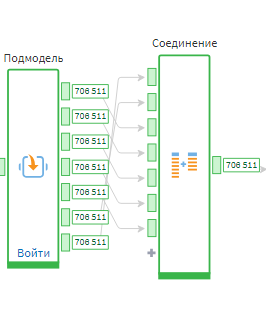
По требованию преподавателя, я должен сделать это задание. Результат моей работы очень спорный, но изучив достаточное количество информации, прихожу к выводу, что мои данные в целом тяжело подставить под условие этой задачи.  
Итак, результат:

Для выполнения кластеризации, было принято решение, исключить некоторые поля из набора данных, а оставшиеся преобразовать к непрерывному типу.

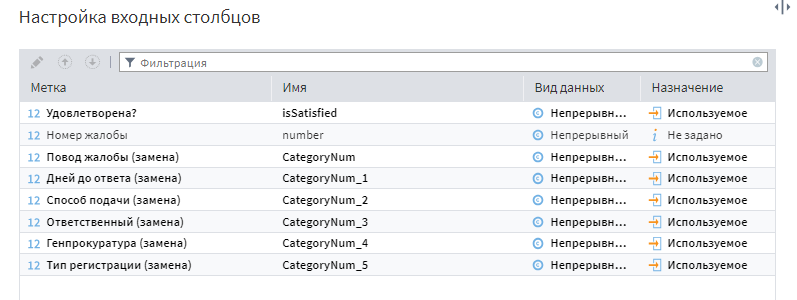
Для этой задачи, я решил воспользоваться сторонней библиотекой loginom, а именно Loginom\_silver\_kit. С помощью ее компонента «Категория к числу» я преобразовал оставшиеся данные к числам.

  
Рисунок 31. Данные к числу

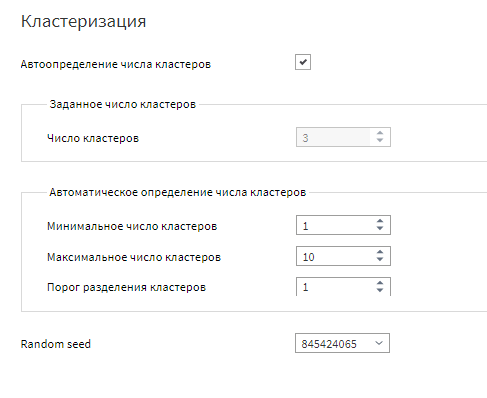
Этот компонент способен преобразовать лишь одно поле за раз, т.е. этих компонентов несколько. В итоге все они сходятся в новый компонент «Соединение»:

  
Рисунок 32. Соединение

Далее я удаляю ненужные данные, а точнее те данные, что были до преобразования, с помощью «Параметры полей». Теперь приступим к настройке кластеров.

  
Рисунок 33. Настройка входа

Для первого варианта, я выбираю вариант – g-means:

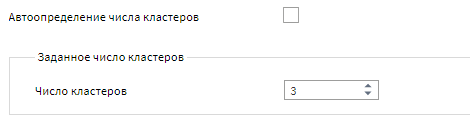
  
Рисунок 34. G-means

Результат был не из лучших:



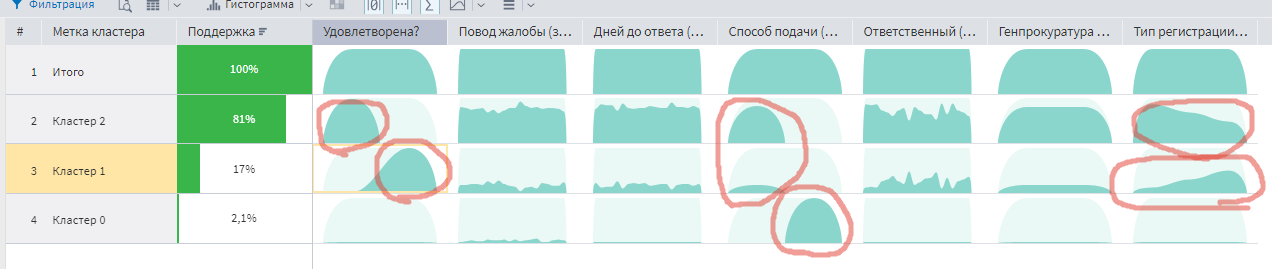
Рисунок 35. G-means визуализация

Далее я стал применять k-means:

  
Рисунок 36. K-means

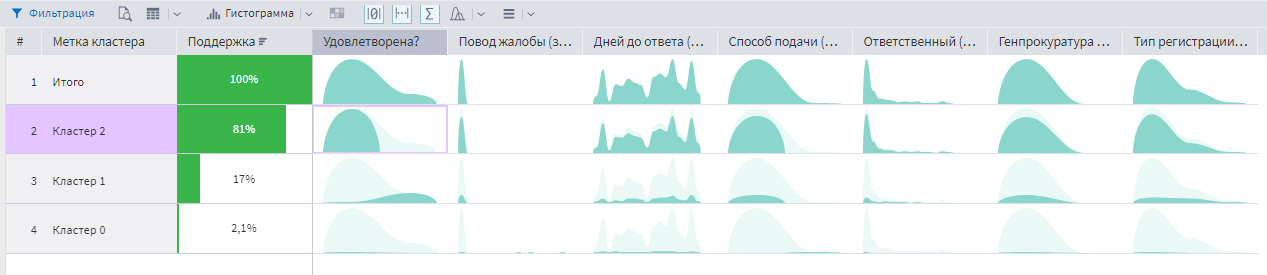
Я задавал вручную количество кластеров для нескольких моделей. Нужно выбрать лучшую. При оценке качества разбиения, я опирался на графики, т.к. это лучший вариант для меня – визуальный.

Если ориентироваться на графики, то лучший результат дает модель, где 3 кластера:

  
Рисунок 37. Модель с 3 кластерами

Т.к. при таком разбиении хотя бы заметны отличительные признаки отдельных кластеров. Остальные разбиения не показывают нужный мне результат.

На тестовых данных модель повела себя «приемлемо»:

  
Рисунок 38. Тестовые данные

Вывод: я провел кластеризацию, где лучшей моделью оказалось разбиение на 3 кластера, на тестовых данных модель повела себя ожидаемо.

# Заключение

В ходе выполнения расчетно-графической работы был проведен полный цикл анализа данных с использованием методики KDD. Работа охватывала этапы аудита данных, их трансформации, формирования моделей, анализа результатов, выдвижения и проверки гипотез. Основное внимание уделялось изучению факторов, влияющих на положительный ответ на жалобу. Анализ был выполнен с использованием платформы Loginom.

**Ключевые этапы:**

**Аудит данных:**

Проведена проверка исходного набора данных на наличие пропусков, дубликатов и некорректных значений.

Осуществлена статистическая оценка признаков (средние значения, медианы, стандартные отклонения), а также анализ распределений данных.

**Трансформация данных:**

Были выполнены вычисления новых полей, полезных для анализа.

Удалены некоторые ненужные столбцы.

**Построение моделей и проверка гипотез:**

Проведена проверка гипотез с помощью логистической регрессии. Результаты показали, что такие факторы, как «Адресат жалобы» и «Жалоба на какую услугу» Существенно влияют на вероятность события.

**Ансамбль моделей:**

Было построено несколько моделей с различными настройками, включая логистическую регрессию.

На основании метрики AUC ROC было проведено сравнение качества моделей.

Анализ ошибок выявил стабильную работу моделей, где точность их прогнозов крайне высока.

**Практическая ценность:**

В результате анализа были выделены значимые признаки, которые могут использоваться для предсказания удовлетворенности жалобы. Эти данные позволяют:

Использовать себя для выбора лучшего места (по критерию отзывчивость на жалобы) получения государственных услуг.

**Итог:**

Использование методики KDD позволило изучить влияние различных факторов на удовлетворяемость жалоб.  
Полученные результаты подчеркивают значимость систематического подхода к обработке и интерпретации данных и могут быть применены для решения аналогичных задач в других областях.

# Список литературы

1. Прокопенко Н.Ю. Анализ и обработка данных: учебно-методическое пособие для обучающихся по дисциплине «Анализ и обработка данных» по направлению подготовки 09.03.04 Программная инженерия профиль 09.03.04 Разработка программно-информационных систем. / Н.Ю. Прокопенко; Нижегородский государственный архитектурно-строительный университет – Нижний Новгород: ННГАСУ, 2021–Текст : непосредственный.
2. Loginom Документация «Коэффициенты регрессии». – URL: <https://help.loginom.ru/userguide/processors/datamining/logistic-regression/coef-regression.html> (дата обращения: 10.12.2024). – Текст: электронный.
3. Loginom Документация «Логистическая регрессия». – URL: <https://help.loginom.ru/userguide/processors/datamining/logistic-regression/index.html> (дата обращения: 10.12.2024). – Текст: электронный.
4. Loginom Блог «Логистическая регрессия: AUC и ROC-кривая». – URL: <https://loginom.ru/blog/logistic-regression-roc-auc> (дата обращения: 10.12.2024). – Текст: электронный.
5. Loginom Wiki «Биннинг». – URL: <https://wiki.loginom.ru/articles/binning.html> (дата обращения: 10.12.2024). – Текст: электронный.
6. Loginom Документация «Корреляционный анализ». – URL: <https://help.loginom.ru/userguide/processors/scrutiny/correlation-analysis.html> (дата обращения: 10.12.2024). – Текст: электронный.
7. Loginom Документация «Ассоциативные правила». – URL: <https://help.loginom.ru/userguide/processors/datamining/association-rules.html> (дата обращения: 10.12.2024). – Текст: электронный.
8. Loginom Документация «Анализ качества данных». – URL: <https://help.loginom.ru/userguide/visualization/data-quality/index.html> (дата обращения: 10.12.2024). – Текст: электронный.
9. Loginom Wiki «AUC». – URL: <https://wiki.loginom.ru/articles/auc.html> (дата обращения: 10.12.2024). – Текст: электронный.