LOGINOM
ХАКАТОН 2019

Loginom

Команда ФГБОУ ВО

«Брянский государственный технический университет»

Вариант № 1

Руководитель:

к.т.н., доцент, Лагерев Д.Г.

Кузьмин С.А. Курилов А.С. Толстенок В.П.

Задание

Входные данные

Клиент микрофинансовой организации → Оценка риска просрочки или невозврата займа в течение длительного времени

27 684 записи

уникальный идентификатор клиента

текстовое поле

текстовое поле

0 – не наблюдалось;
1 – наблюдалось;
пусто – нет данных.

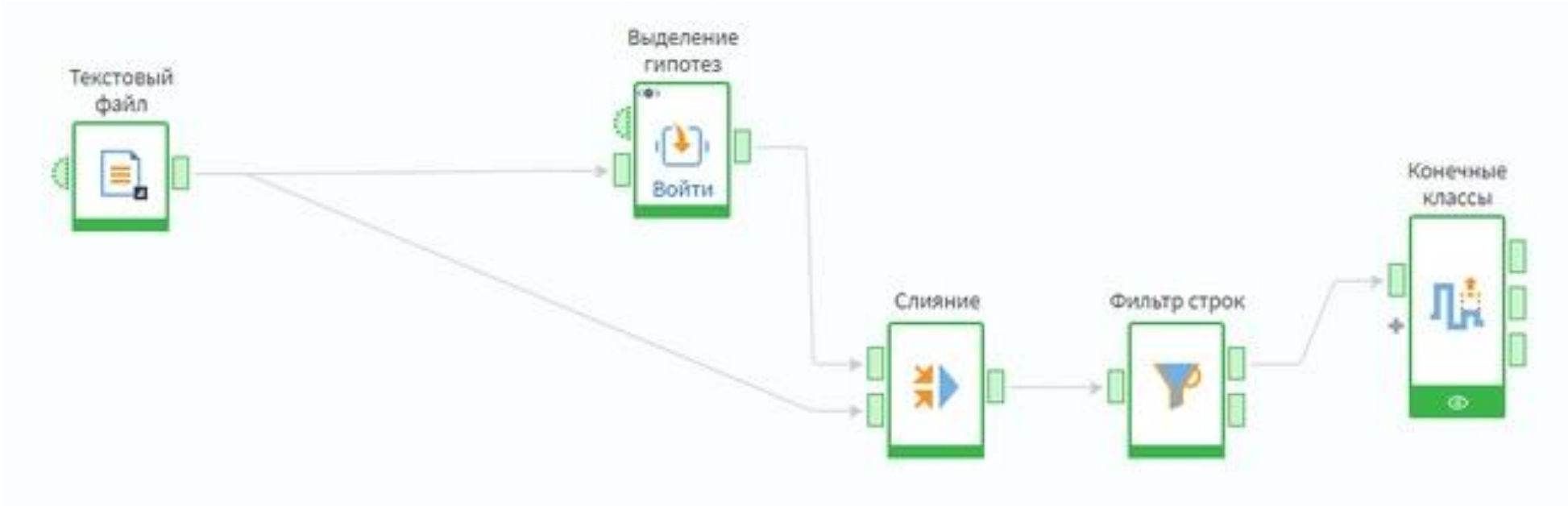
Клиент	Место работы	Метка региона	Событие
E3EDCA0F6E68BFB76EAF26A8EB6DD94B	Самарский Филиал Учреждения ООО "Инвакон-М"	Новосибирск	1
4CC42CEFB756CC4D4FC320900365E426	ООО Кафе Чилипепперс	Новосибирск	
901B2D0B152D0EEFB3AB8BEFD01E92C9	ФКУ ОК УФСИН	Новосибирск	
3E834D5CE607547397DE337CD1C4E05A	ООО "Реминдустрия"	Ярославль	
F736185C629CD6BC82CCE484528A2184	ИП Шилова Марина Евгеньевна	Ярославль	
64B08C66DFE324E822345640828669 00	ОАО "ЯЗДА"	Ярославль	
50CE3C2AEC63ADF44E47A539D231C33D	МУП ПАТП- 1	Ярославль	

Задание

Результат

- Опубликованный компонент, принимающий на вход поле Место работы.
- Автоматическое извлечение информации из поля Место работы с дальнейшим выдвижением гипотез, взаимосвязанных с риском просрочки займа.
- Производилось предварительное оценивание гипотез с помощью WoE и IV-анализа.
- Гипотезы имеют градацию значимости от «Отсутствует» до «Высокая».

Разработанный компонент



Входные и выходные значения компонента

Входные значения компонента

#	ab Клиент	ab Место работы	ab Метка региона	9.0 Событие
4 967	0FA84437D89E241CA5A54C55FE69B97F	ООО СК "Союз"	Ярославль	1,00
4 968	9E33BA9CE4AA26727F286074F543AFF1	Юридическая компания ООО"Консультъ"	Новосибирск	1,00
4 969	1081617E822A16541B3A581EC2EE2BBD	ИП Щекотов А.В.	Новосибирск	
4 970	A2D203C4CB3877206C557417B5483DAA	ООО " Русьхлеб"	Ярославль	
4 971	A66CC88FBEEF053A6C9509D78184E20F	ООО Компания Холидей	Новосибирск	0,00

Выходные значения компонента

#	0/1 Англ. символы?	0/1 Указано место работы	0/1 Возможная небрежность	ab Сфера доходов	ab Нестабильный з...	ab Нестандар...	ab Сезонность	ab Категория
3 527	false	true	true	Несущественно	Стабильный	Стандартный	Не определено	Не определено
3 528	false	true	true	Несущественно	Стабильный	Стандартный	Не определено	Не определено
3 529	false	true	true	Несущественно	Торговля	Стандартный	Не определено	Торговля
3 530	false	true	true	Несущественно	Торговля	Стандартный	Не определено	Торговля
3 531	false	true	true	Несущественно	Стабильный	Стандартный	Сельское хозя...	Не определено

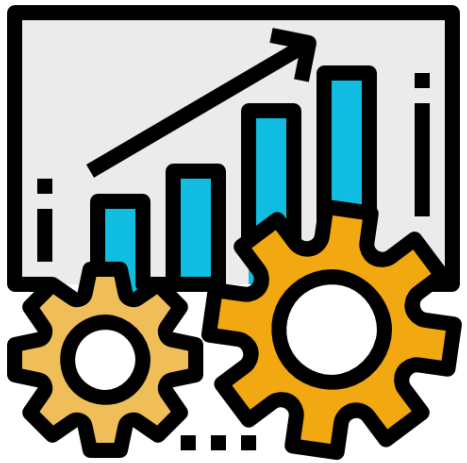
Предложенное решение

В качестве решения было предложено выделить гипотезы, после чего с помощью узла конечных классов выявить коэффициенты WoE для дальнейшей оценки их применимости и приоритетности.

Выделенные гипотезы

- Работы с нестабильным заработком
- Высокооплачиваемые работы
- Сезонные работы
- Места работы, являющиеся филиалами
- Наличие английских символов
- Присутствие нестандартных слов
- Выделение форм собственности

- Разбиение мест работы по отраслям
- Выделение крупных компаний
- Определение филиала
- Указано ли место работы?
- Возможная небрежность



Настройка

- Возможность улучшения предсказательной способности благодаря пользовательской настройке справочников.
- Пользователь может сам загрузить excel-файлы, содержимое которых наилучшим образом подходит под специфику имеющихся у него данных.

Были разработаны следующие справочники:

- Нестабильного заработка
- Сезонных работ
- Нестандартных слов
- Категорий
- Форм собственности
- Высокооплачиваемых работ
- Крупных компаний

Разработанные справочники

Справочник нестабильного заработка • Наб…

#	Ключевое слово	Сфера деятельности
1	чоп	Охрана
2	чоо	Охрана
3	охранн	Охрана
4	тд	Торговля
5	торг	Торговля
6	такси	Такси
7	ретейл	Розничные продажи
8	ритейл	Розничные продажи
9	пансион	Отдых и туризм
10	отель	Отдых и туризм
11	курорт	Отдых и туризм

Закрыть

Справочник сезонных работ • Набор данных • …

#	Ключевое слово	Сфера деятельности
1	отель	Туризм
2	турагенство	Туризм
3	туроператор	Туризм
4	ферма	Сельское хозяйство
5	кфх	Сельское хозяйство
6	агро	Сельское хозяйство
7	сельхоз	Сельское хозяйство
8	рыба	Охота и рыбалка
9	охот	Охота и рыбалка
10	геолог	Обработка ландшафта
11	мелио	Обработка ландшафта
12	нефт	Нефтедобыча
13	лес	Деревообработка
14	строй	Строительство

Закрыть

Справочник нестандартных слов • …

#	Ключевое слово	Категория
1	плюс	Плюс
2	+	Плюс
3	гранд	Англицизмы
4	офф	Суффикс -офф
5	off	Суффикс -офф
6	off	Суффикс -офф
7	off	Суффикс -офф
8	голд	Англицизмы
9	gold	Англ. Слова
10	лэйбл	Англицизмы
11	секонд	Англицизмы
12	хенд	Англицизмы
13	паблишинг	Англицизмы
14	&	Спец символы

Закрыть

Справочник крупных компаний • Набор данных • Б…

#	Компания	Сфера деятельности
1	Газпром	Нефть и газ
2	ЛУКОЙЛ	Нефть и газ
3	Роснефть	Нефть и газ
4	Сбербанк России	Банки
5	РЖД	Железнодорожный транспорт
6	ВТБ	Банки
7	Ростех	Инвестиции
8	Сургутнефтегаз	Нефть и газ
9	Магнит	Розничная торговля
10	Россети	Электроэнергетика
11	Интер РАО	Электроэнергетика
12	Транснефть	Трубопроводный транспорт
13	АФК Система	Инвестиции

642

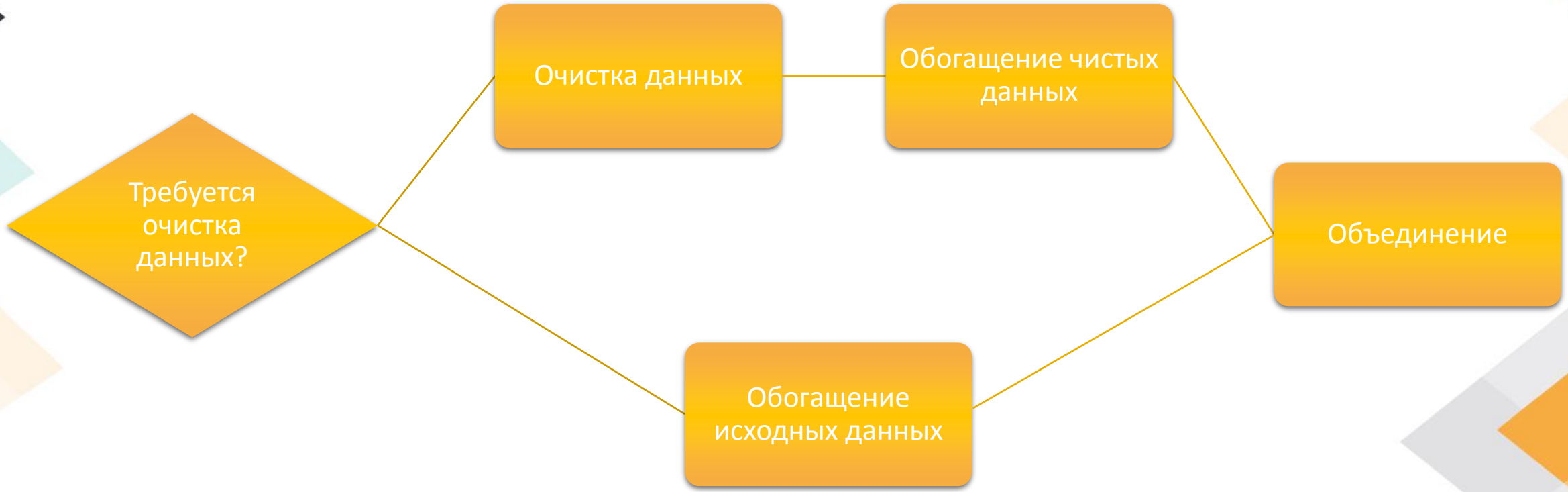
Закрыть

Структура

Компонент состоит из 2-х модулей, первый из которых содержит 3 подмодели:

- Очистка данных
- Обогащение чистых данных
- Обогащение исходных данных

Схема работы



Подмодель «Очистка данных»

Входные данные:
Данные, загруженные пользователем.

Выходные данные:
Очищенные данные.

Преобразует исходные данные в вид, наиболее подходящий для дальнейшего анализа.

Используемые методы

Токенизация	Очистка пустых значений
Перевод в нижний регистр	Выделение дополнительных гипотез (узлы js)

Используемые компоненты

- Узел JavaScript «Токенизация» – разбивает строку «место работы» на массив содержащихся в ней слов, разделенных запятыми.
- Узел Калькулятор «Очистка пустых значений».
- Узел-ссылка «Формы собственности».
- Узел Калькулятор «Нижний регистр».
- Узел-ссылка «Справочник крупных компаний».
- Узел Калькулятор «Нижний регистр».
- Узел JavaScript «Декомпозиция места работы» – выделяет из поля «место работы» организационно-правовую форму.
- Узел Слияние «Слияние».

Подмодель «Обогащение чистых данных»

Входные данные:	*Выходные данные:*
Очищенные данные	Обогащенные данные

На основе предварительно обработанных данных, выделяет некоторые признаки, информационная значимость которых проверяется в дальнейшем.

Используемые методы

Выделение основных гипотез (узлы js)	Объединение данных

Используемые компоненты

- Узлы JavaScript «Определение филиала», «Нахождение англ. символов» и «Указано место работы».
- Узел Дополнение данных «Дополнение данных».
- Узел Подмодель «Топ-10 форм собственности».
- Подмодель «Частые слова».

Подмодель «Обогащение исходных данных»

Входные данные:	*Выходные данные:*
Исходные данные (данные загруженные пользователем)	Обогащенные исходные данные.

.

Выделяет некоторые признаки, информационная значимость которых проверяется в дальнейшем, без предварительной обработки.

Используемые методы

Очистка пустых значений	Выделение дополнительных гипотез (узлы js)
Выделение основных гипотез (узлы js)	Объединение данных

Используемые компоненты

- Узлы JavaScript «Определение филиала», «Форма собственности», «Нахождение англ. символов», «Указано место работы?» и «Проверка небрежности» – на основе признака отсутствия кавычек в поле «Место работы» предполагает наличие небрежного написания.
- Узел Дополнение данных «Дополнение данных».
- Узел Подмодель «Частые слова».
- Узел Калькулятор «Очистка пустых значений.
- Узел Подмодель «Дополнительные гипотезы».
- Узел Слияние «Слияние».

Использование библиотек

- Была использована бесплатная библиотека Az.js. Данная библиотека распространяется по лицензии MIT.

```
const Az = require('azjs/az.js') //библиотека токенизации
```

- Суть токенизации очень проста: на входе принимается строка – а на выходе получаем «токены», группы символов, которые (вероятно) являются отдельными сущностями в этой строке.
- Fs.js – вспомогательная библиотека для корректной работы с файловой системой.

JavaScript

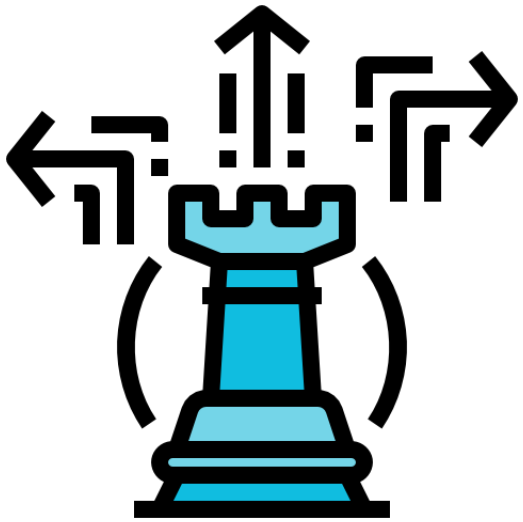
```
import { InputTable, InputTables, InputVariables, OutputTable, DataType, DataKind, UsageType }
const Az = require('azjs/az.js') //библиотека токенизации

for(let i =0; i < InputTables[0].RowCount; i++) {
    let client = InputTables[0].Get(i,0)
    let job = InputTables[0].Get(i,1)
    let region = InputTables[0].Get(i,2)
    job = Az.Tokens(job).done();
    //Az.Morph.init('azjs/dicts', function() {
        //var parses = Az.Morph('стали');
    //console.log(parses); // => 6 вариантов разбора
    //console.log(parses[0].tag.toString()); // => 'VERB,perf,intr plur,past,indc'
    //console.log(parses[1].tag.toString()); // => 'NOUN,inan,femn plur,nomn'
//});
    //var gram = Az.Morph(job).done()
    OutputTable.Append()
    OutputTable.Set(0, client)
    OutputTable.Set(1, job)
    OutputTable.Set(2, region)
    OutputTable.Set(3,InputTables[0].Get(i,1))
    //console.log(gram)
}

```

Пользовательские возможности

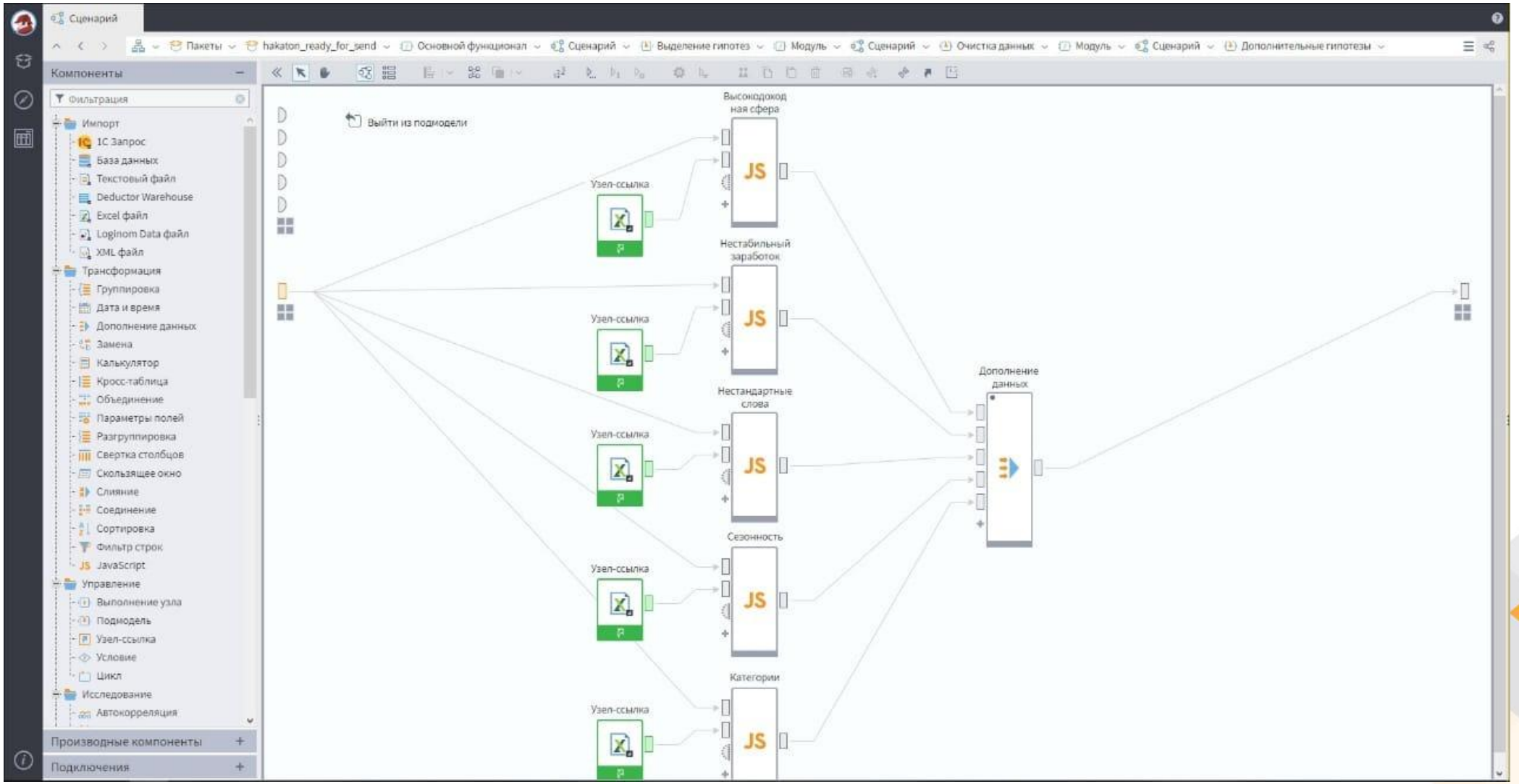
- Пользователь сам может выбрать нужна ли ему предварительная очистка данных, а также может либо использовать справочники для гипотез по умолчанию, либо настроить их в соответствии со своим видением предметной области или рекомендациями аналитика.
- Благодаря предварительной очистке данных, не имеет значения, в каком регистре будет написана форма собственности, а также количество специальных символов, отступов в названии места работы.

Расширение возможностей

- Дополняется подключением различных REST и WSDL сервисов для более точного выделения различных значимых признаков.
- Также модель может выделять наиболее часто встречающиеся формы собственности и слова. В дальнейшем планируется на их основе предложить дополнительные гипотезы.



Расширение возможностей



Сфера применения

- Компонент генерирует некоторое количество новых полей, которые в дальнейшем могут использоваться для улучшения предсказательной способности скоринговой модели.

Тестирование компонента

Гипотеза	Уровень значимости	
	Выборка 1 (Новосибирск, Ярославль)	Выборка 2 (Нижний Новгород, Самара)
Работы с нестабильным заработком	0,02	0
Высокооплачиваемые работы	0,02	0,02
Сезонные работы	**0,03**	0,01
Указано место работы	0,02	0
Наличие английских символов	0	**0,05**
Присутствие нестандартных слов	0,02	0
Выделение форм собственности	**0,12**	**0,08**
Разбиение мест работы по отраслям (категория)	**0,05**	0,02

Ключевые возможности Loginom

- Возможность создания производных компонентов
- Реализация методов ООП
- Поддержка JavaScript
- Параллельная обработка данных
- Проектирование без данных

СПАСИБО ЗА ВНИМАНИЕ!

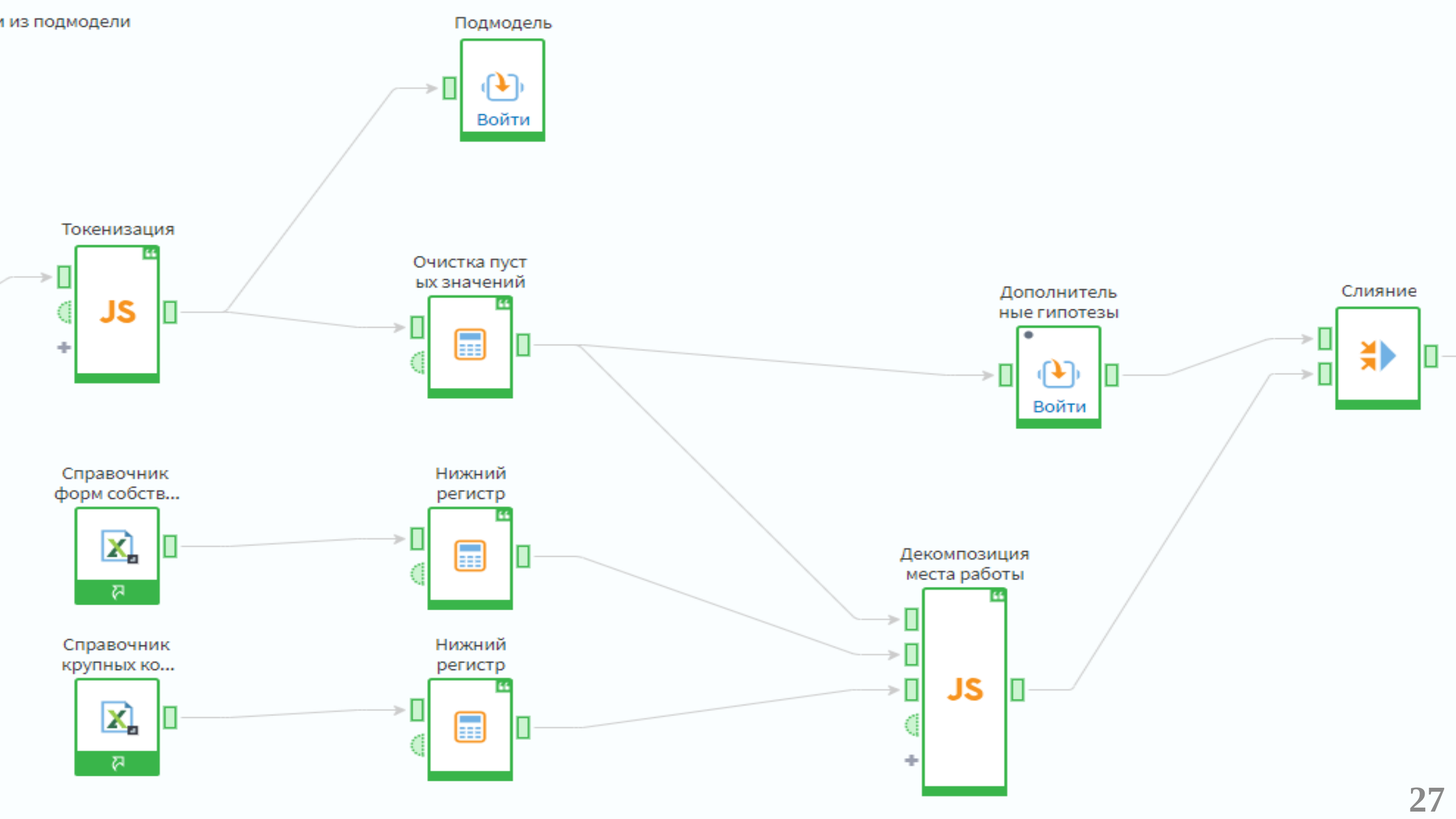
Руководитель:

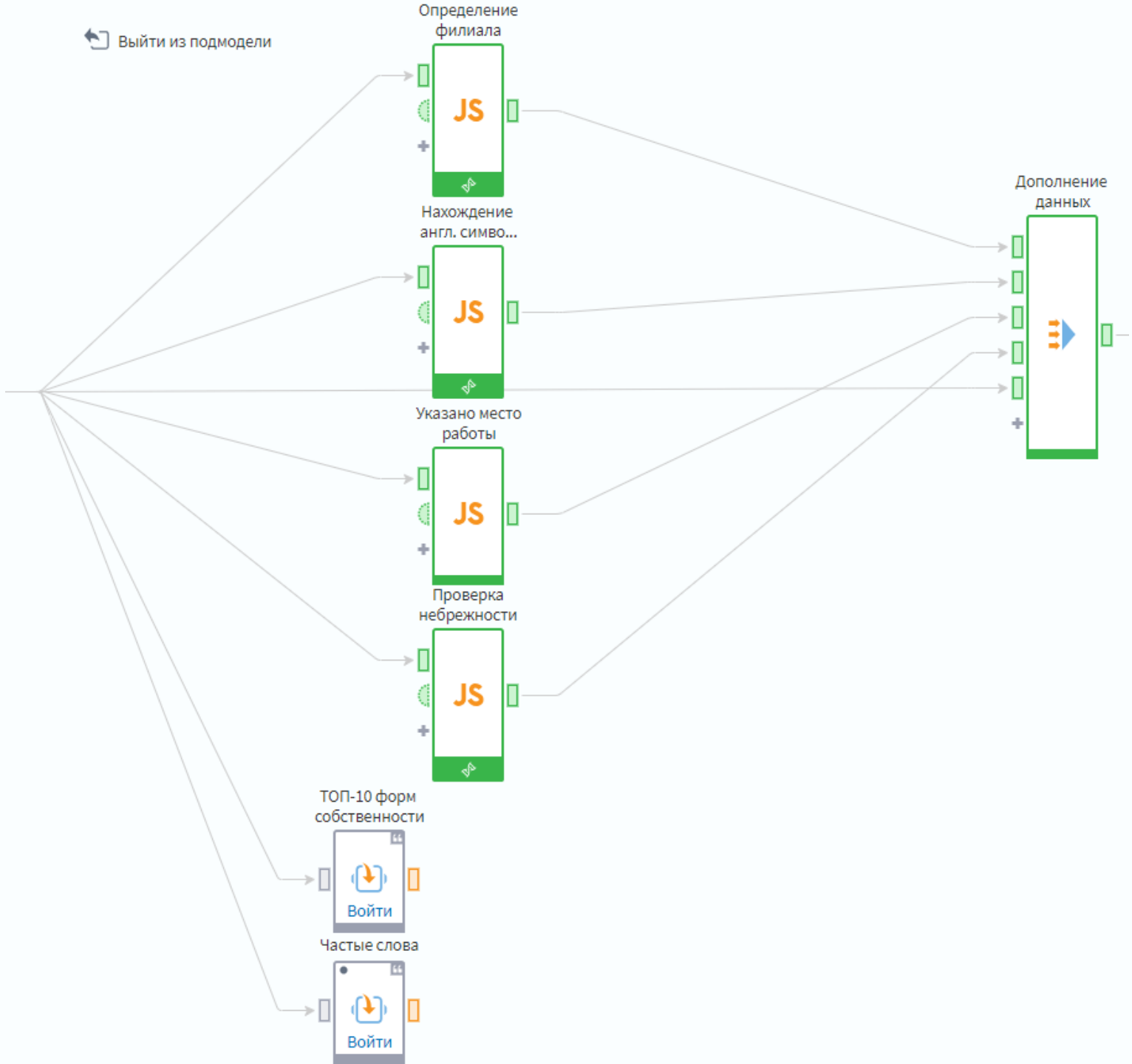
к. т.н., доцент, Лагерев Д.Г.

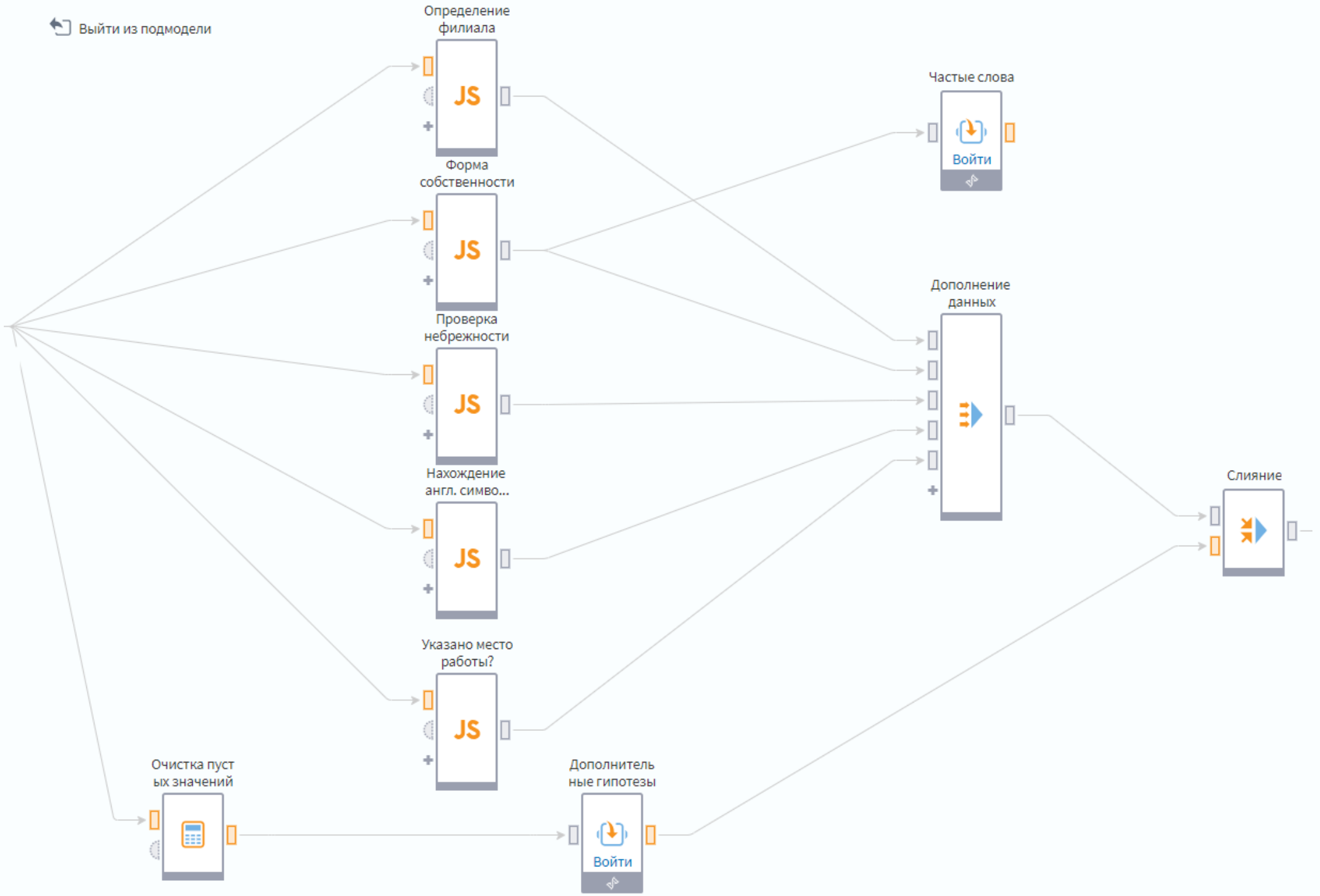
Кузьмин С.А.
wolv3333@mail.ru

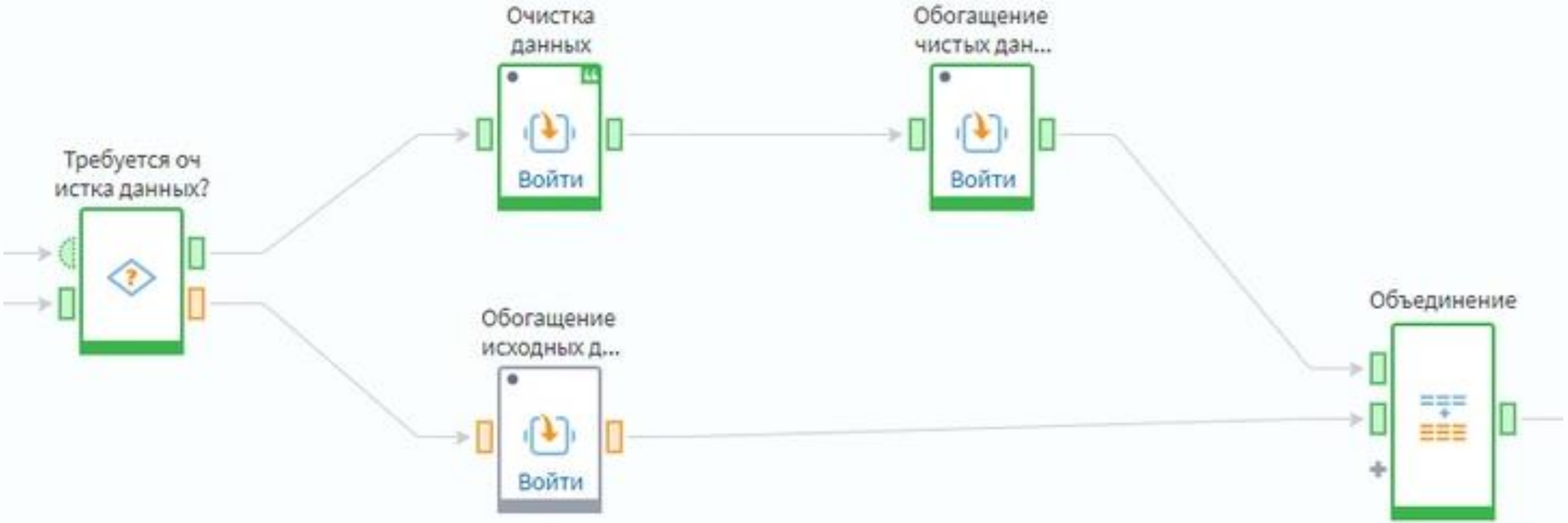
Курилов А.С.
kac.kurilov@yandex.ru

Толстенок В.П.
tolstenok21@yandex.ru


и из подмодели
Подмодель
Войти
Токенизация
JS
Очистка пуст
ых значений
Дополнитель
ные гипотезы
Войти
Слияние
Справочник
форм собств...
Нижний
регистр
Декомпозиция
места работы
Справочник
крупных ко...
Нижний
регистр
JS


Выйти из подмодели
Определение филиала
JS
Нахождение англ. симво...
JS
Указано место работы
JS
Проверка небрежности
JS
Дополнение данных
ТОП-10 форм собственности
Войти
Частые слова
Войти


Выйти из подмодели
Определение филиала
JS
Частые слова
Войти
Форма собственности
JS
Дополнение данных
Проверка небрежности
JS
Нахождение англ. симво...
JS
Слияние
Указано место работы?
JS
Очистка пустых значений
Дополнительные гипотезы
Войти


Очистка данных
Обогащение чистых дан...
Требуется оч истка данных?
Войти
Войти
Объединение
Обогащение исходных д...
Войти