Домашняя работа по курсу

Методы и средства обработки больших данных

Выполнила Клепинина К.Д.

Задание состоит из 2-ух частей:

1.Исследование данных и обработка данных для проведения последующей сегментации;

2.Составить профили клиентов на основе проведенных сегментаций

Содержание

Часть 1

Подготовка данных.

Проверка на дубликаты показала, что дубликатов нет.

Среди значений таблицы есть неудобная для обработки переменная EMPL_SIZE со значениями `'>250', '>=50', '>=100', '>=200', '< 50', nan, '>100', '>=150'`. Заменим эти значения на случайные в обозначенных интервалах (<50, от 50 до 100 и т.д.)

Рассмотрим количество уникальных и пропущенных значений и количество нулей для каждого значения.

	кол-во уникальных	кол-во нулей	доля нулей	кол-во пропущенных	доля пропущенных
Номер варианта	1	0	0	0	0
ID	3963	0	0	0	0
INCOME_BASE_TYPE	4	0	0	0	0
CREDIT_PURPOSE	10	0	0	0	0
INSURANCE_FLAG	2	1541	38,88	0	0
DTI	57	0	0	0	0
SEX	2	0	0	0	0
FULL_AGE_CHILD_NUMBER	6	2438	61,52	0	0
DEPENDANT_NUMBER	4	3955	99,8	0	0
EDUCATION	9	0	0	0	0
EMPL_TYPE	8	0	0	0	0
EMPL_SIZE	759	0	0	1	0,03
BANKACCOUNT_FLAG	3	3221	81,28	0	0
Period_at_work	264	0	0	1	0,03
age	38	0	0	0	0
EMPL_PROPERTY	5	0	0	0	0
EMPL_FORM	6	0	0	0	0
FAMILY_STATUS	7	0	0	1	0,03
max90days	22	1041	26,27	60	1,51
max60days	20	1527	38,53	60	1,51
max30days	16	1936	48,85	60	1,51
max21days	15	2319	58,52	60	1,51
max14days	16	2499	63,06	60	1,51
avg_num_delay	1132	1538	38,81	375	9,46
if_zalog	3	2400	60,56	361	9,11
num_AccountActive180	8	2561	64,62	361	9,11
num_AccountActive90	7	3073	77,54	361	9,11
num_AccountActive60	6	3263	82,34	361	9,11
Active_to_All_prc	100	475	11,99	361	9,11
numAccountActiveAll	15	446	11,25	361	9,11
numAccountClosed	24	429	10,83	361	9,11

sum_of_paym_months	325	14	0,35	361	9,11
all_credits	30	0	0	361	9,11
Active_not_cc	9	1178	29,72	361	9,11
own_closed	9	2058	51,93	361	9,11
min_MnthAfterLoan	98	153	3,86	361	9,11
max_MnthAfterLoan	129	11	0,28	361	9,11
dlq_exist	3	1552	39,16	361	9,11
thirty_in_a_year	3	3092	78,02	361	9,11
sixty_in_a_year	3	3306	83,42	361	9,11
ninety_in_a_year	3	3381	85,31	361	9,11
thirty_vintage	3	3485	87,94	361	9,11
sixty_vintage	3	3547	89,5	361	9,11
ninety_vintage	3	3555	89,7	361	9,11

Рассмотрим среднее значение, медиана, стандартное отклонение, минимум, максимум по каждому столбцу данных:

	Номер варианта	ID	INSURANCE_FLAG	DTI	FULL_AGE_CHILD_NUMBER	DEPENDANT_NUMBER	EMPL_SIZE	BANKACCOUNT_FLAG
count	10243.0	1.024300e+04	10243.000000	10120.000000	10243.000000	10243.000000	10121.000000	7908.000000
mean	7.0	1.102427e+06	0.600898	0.388604	0.562335	0.003905	187.768995	0.386950
std	0.0	5.914087e+04	0.489738	0.137372	0.780858	0.083753	88.802754	0.872694
min	7.0	1.000007e+06	0.000000	0.000000	0.000000	0.000000	20.000000	0.000000
25%	7.0	1.051217e+06	0.000000	0.280000	0.000000	0.000000	100.000000	0.000000
50%	7.0	1.102427e+06	1.000000	0.400000	0.000000	0.000000	250.000000	0.000000
75%	7.0	1.153637e+06	1.000000	0.490000	1.000000	0.000000	250.000000	0.000000
max	7.0	1.204847e+06	1.000000	0.590000	7.000000	3.000000	250.000000	3.000000

Вызывает интерес минимальное значение переменной min_MnthAfterLoan = -1, т.к. min_MnthAfterLoan означает количество месяцев, прошедших со взятия последнего кредита.

All	numAccountClosed	sum_of_paym_months	all_credits	Active_not_cc	own_closed	min_MnthAfterLoan	max_MnthAfterLoan	dlq_exist	thirty_in_a_year	s
00	3602.000000	3602.000000	3602.000000	3602.000000	3602.000000	3602.000000	3602.000000	3602.000000	3602.000000	
92	3.522210	80.861466	5.717102	1.098279	0.716269	14.004442	60.474459	0.569128	0.141588	
14	3.198632	70.737158	4.051142	1.064852	1.036146	15.103052	30.339091	0.495267	0.348675	
00	0.000000	0.000000	1.000000	0.000000	0.000000	-1.000000	0.000000	0.000000	0.000000	
00	1.000000	30.000000	3.000000	0.000000	0.000000	4.000000	34.000000	0.000000	0.000000	
00	3.000000	62.000000	5.000000	1.000000	0.000000	10.000000	64.000000	1.000000	0.000000	
00	5.000000	113.000000	8.000000	2.000000	1.000000	18.000000	86.000000	1.000000	0.000000	
00	23.000000	807.000000	30.000000	7.000000	7.000000	115.000000	177.000000	1.000000	1.000000	

Рассмотрим строки с этим значением переменной внимательнее, чтобы решить, как поступить с этим значением дальше. Есть несколько

предположений о причине возникновения значения -1 в ячейке, в которой его не должно быть. Либо у этих клиентов никогда не было кредитов и значение ошибочно, в таком случае будем удалять эти строки, либо это планируемый кредит, т.е. кредит будет взят в следующем месяце, в этом случае заменим -1 на 0.

	2997	3527	3550	6606	7709	9952
Номер варианта	7	7	7	7	7	7
ID	1059947	1070547	1071007	1132127	1154187	1199047
INCOME_BASE_TYPE	Поступление зарплаты на счет	Поступление зарплаты на счет	Форма банка (без печати работодателя)	Форма банка (без печати работодателя)	2НДФЛ	Форма банка (без печати работодателя)
CREDIT_PURPOSE	Покупка земли	Другое	Покупка бытовой техники	Ремонт	Покупка недвижимости/ строительство	Ремонт
INSURANCE_FLAG	0	1	1	1	1	1
DTI	0,58	0,54	0,22	0,44	0,51	0,2
SEX	женский	женский	мужской	женский	женский	женский
FULL_AGE_CHILD_NUMBER	0	0	1	0	0	2
DEPENDANT_NUMBER	0	0	0	0	0	0
EDUCATION	высшее	высшее	среднее	высшее	высшее	высшее
EMPL_TYPE	специалист	вспомогательный персонал	специалист	специалист	рабочий	менеджер по продажам
EMPL_SIZE	250	250	250	250	250	250
BANKACCOUNT_FLAG	1	0	0	0	0	0
Period_at_work	120	6	73	53	93	97
age	43	24	33	25	26	42
EMPL_PROPERTY	Другое	Другое	Другое	Торговля	Другое	Торговля
EMPL_FORM	ЗАО	ООО	ООО	ООО	ОАО	ООО
FAMILY_STATUS	женат / замужем	женат / замужем	женат / замужем	холост / не замужем	женат / замужем	разведен / разведена
max90days	2	2	4	2	3	10
max60days	2	2	4	2	2	1
max30days	2	2	3	2	2	1
max21days	1	2	2	2	2	0
max14days	1	0	0	0	0	0
avg_num_delay	0,004587	0,040816	0	0,01	0	0,02834
if_zalog	1	1	0	1	0	1
num_AccountActive180	1	2	4	1	2	2
num_AccountActive90	1	2	2	1	1	2
num_AccountActive60	1	2	2	1	1	2
Active_to_All_prc	0,5	0,571429	0,8	0,5	0,571429	0,583333
numAccountActiveAll	5	4	4	5	4	7
numAccountClosed	5	3	1	5	3	5

sum_of_paym_months	217	47	42	99	88	240
all_credits	10	7	5	10	7	12
Active_not_cc	4	2	2	3	2	4
own_closed	0	0	1	0	0	2
min_MnthAfterLoan	-1	-1	-1	-1	-1	-1
max_MnthAfterLoan	90	39	34	48	35	83
dlq_exist	1	1	0	1	0	1
thirty_in_a_year	0	0	0	0	0	1
sixty_in_a_year	0	0	0	0	0	1
ninety_in_a_year	0	0	0	0	0	1

Всего 6 клиентов с странным значением min_MnthAfterLoan (таблица транспонирована для удобства восприятия). Можно заметить, что у каждого из клиентов есть как минимум 5 кредитов, т.е. теория об отсутствии кредитов у таких клиентов не подтвердилась. Для проверки второй теории посмотрим на значения других переменных, таких как количество активных счетов за последние 60 дней и количество запросов в бюро кредитных историй за последний месяц. У всех клиентов есть как минимум один активный за последние 60 дней счёт, но в последние 21-14 дней запросы в бюро кредитных историй были не у всех. Насколько я понимаю, запрос кредитной истории необходим для принятия решения о выдаче кредитов. Предположу возможность ситуации, в которой между запросом в бюро кредитных историй и моментом выдачи кредита клиенту проходит >21, но <30 дней. Оставим эти строки в исходном виде, посчитав, что числа несут содержательную информацию.

Также кажется интересным минимум равный 4 в столбце **Period_at_work,** где записано количество дней работы. Стаж работы для получения кредита в крупных банках должен быть не меньше 3-4 месяцев. Посмотрим на количество клиентов со стажем меньше месяца – таких найдено 2766, что составляет достаточно большой процент от всех значений базы данных, следовательно предположить ошибочное попадание данных сложно. Максимальное значение переменной **Period_at_work** равно 489, т.е. немногим больше года. Логичным кажется предположить ошибку в описании данных, скорее всего это значение стажа в месяцах, тогда максимальное

значение переменной равно примерно 40 годам, что выглядит вполне реально при максимальном возрасте 62 года. Данные будем считать верными.

Проверим нечисловые переменные на наличие «странных» значений:

```
##проверка нечисловых переменных
df.CREDIT_PURPOSE.unique()
```

```
array(['Покупка недвижимости/ строительство', 'Ремонт',
       'Покупка бытовой техники', 'Покупка мебели', 'Другое',
       'Покупка автомобиля', 'Отпуск', 'Лечение', 'Покупка земли',
       'Обучение'], dtype=object)
```

```
df.SEX.unique()
```

```
array(['мужской', 'женский'], dtype=object)
```

```
df.EDUCATION.unique()
```

```
array(['высшее', 'Высшее/Второе высшее/Ученая степень',
       'среднее-специальное', 'второе высшее', 'среднее',
       'незаконченное высшее', 'Неполное среднее', 'ученая степень'],
      dtype=object)
```

```
df.EMPL_TYPE.unique()
```

```
array(['специалист', 'менеджер среднего звена', 'менеджер по продажам',
       'рабочий', 'вспомогательный персонал', 'торговый представитель',
       'другое', 'менеджер высшего звена', nan, 'страховой агент'],
      dtype=object)
```

```
df.EMPL_PROPERTY.unique()
```

```
array(['Наука', nan, 'Информационные технологии', 'Торговля',
       'Производство', 'Другое', 'Туризм', 'Финансы',
       'Государственная служба', 'Строительство', 'Транспорт',
       'Юридические услуги', 'Сельское и лесное хозяйство'], dtype=object)
```

```
df.EMPL_FORM.unique()
```

```
array([nan, 'ООО', 'Государственное предприятие', 'ОАО',
       'Индивидуальный предприниматель', 'Иная форма', 'ЗАО'],
      dtype=object)
```

```
df.FAMILY_STATUS.unique()
```

```
array([nan, 'женат / замужем', 'гражданский брак', 'разведен / разведена',
       'холост / не замужем', 'повторный брак', 'вдовец / вдова'],
      dtype=object)
```

Нестандартных значений не обнаружено.

Очистка данных.

Для каждой переменной было рассчитано число уникальных, нулевых и пропущенных значений:

	кол-во уникальных	кол-во нулей	доля нулей	кол-во пропущенных	доля пропущенных
Номер варианта	1	0	0	0	0
ID	10243	0	0	0	0
INCOME_BASE_TYPE	5	0	0	63	0,62
CREDIT_PURPOSE	10	0	0	0	0
INSURANCE_FLAG	2	4088	39,91	0	0
DTI	59	1	0,01	123	1,2
SEX	2	0	0	0	0
FULL_AGE_CHILD_NUMBER	8	6085	59,41	0	0
DEPENDANT_NUMBER	4	10216	99,74	0	0
EDUCATION	8	0	0	0	0
EMPL_TYPE	10	0	0	11	0,11
EMPL_SIZE	474	0	0	122	1,19
BANKACCOUNT_FLAG	4	6230	60,82	2335	22,8
Period_at_work	374	0	0	2337	22,82
age	39	0	0	2335	22,8
EMPL_PROPERTY	13	0	0	2335	22,8
EMPL_FORM	7	0	0	6280	61,31
FAMILY_STATUS	7	0	0	6281	61,32
max90days	22	1041	10,16	6340	61,9
max60days	20	1527	14,91	6340	61,9
max30days	16	1936	18,9	6340	61,9
max21days	15	2319	22,64	6340	61,9
max14days	16	2499	24,4	6340	61,9
avg_num_delay	1132	1538	15,02	6655	64,97
if_zalog	3	2400	23,43	6641	64,83
num_AccountActive180	8	2561	25	6641	64,83
num_AccountActive90	7	3073	30	6641	64,83
num_AccountActive60	6	3263	31,86	6641	64,83
Active_to_All_prc	100	475	4,64	6641	64,83
numAccountActiveAll	15	446	4,35	6641	64,83
numAccountClosed	24	429	4,19	6641	64,83
sum_of_paym_months	325	14	0,14	6641	64,83
all_credits	30	0	0	6641	64,83
Active_not_cc	9	1178	11,5	6641	64,83
own_closed	9	2058	20,09	6641	64,83
min_MnthAfterLoan	98	153	1,49	6641	64,83
max_MnthAfterLoan	129	11	0,11	6641	64,83
dlq_exist	3	1552	15,15	6641	64,83
thirty_in_a_year	3	3092	30,19	6641	64,83
sixty_in_a_year	3	3306	32,28	6641	64,83
ninety_in_a_year	3	3381	33,01	6641	64,83
thirty_vintage	3	3485	34,02	6641	64,83

sixty_vintage	3	3547	34,63	6641	64,83
ninety_vintage	3	3555	34,71	6641	64,83

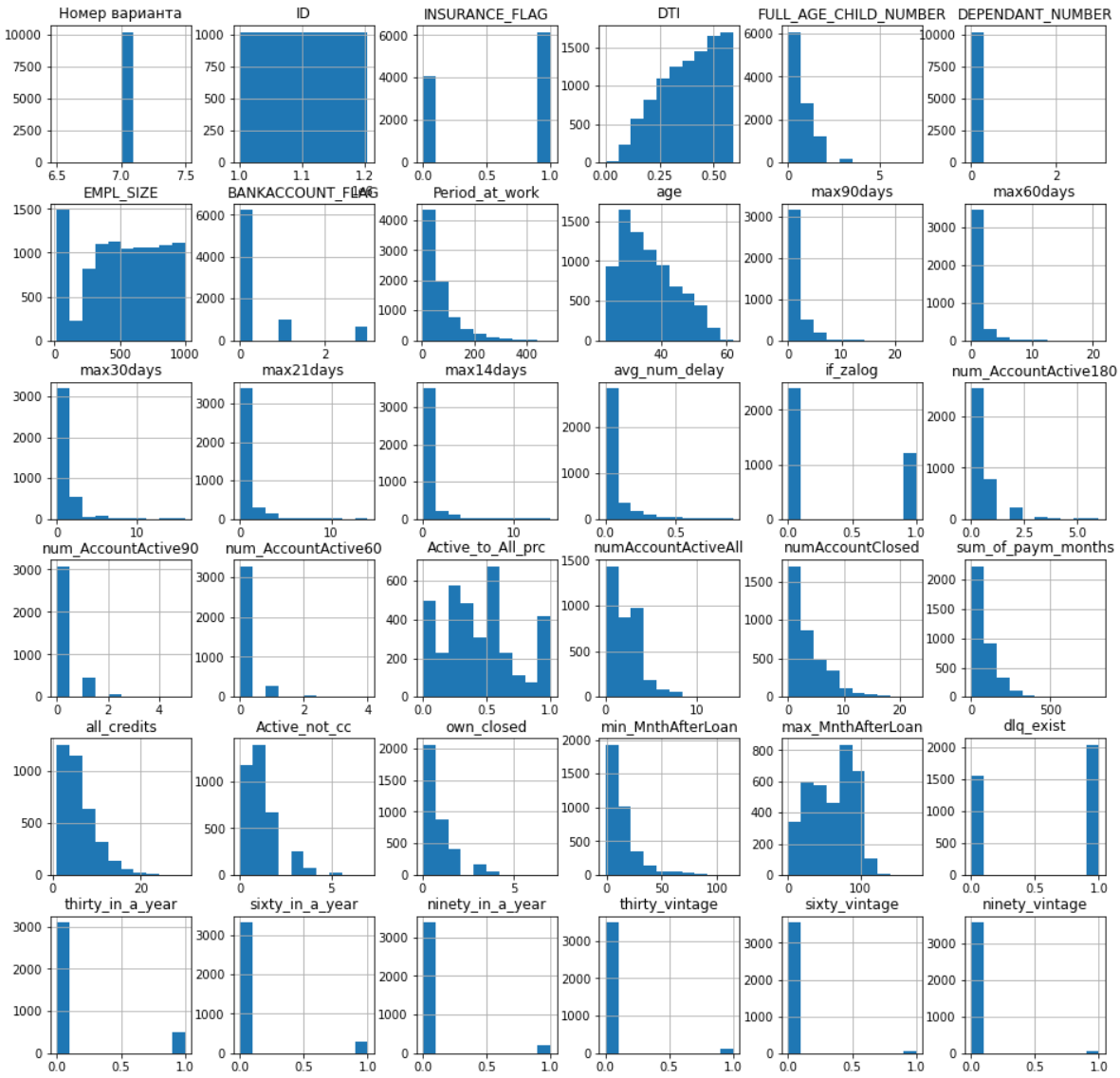
При очистке данных сохраним строки, в которых есть как минимум треть непустых значений. Пустые значения заполним медианными , категориальные переменные преобразуем в количественные при помощи OneHotEncoder.

После очистки данных максимальный процент пропущенных значений равен 9.46.

avg_num_delay	1132.0	1538.0	38.81	375.0	9.46
if_zalog	3.0	2400.0	60.56	361.0	9.11
num_AccountActive180	8.0	2561.0	64.62	361.0	9.11
num_AccountActive90	7.0	3073.0	77.54	361.0	9.11
num_AccountActive60	6.0	3263.0	82.34	361.0	9.11
Active_to_All_prc	100.0	475.0	11.99	361.0	9.11
numAccountActiveAll	15.0	446.0	11.25	361.0	9.11
numAccountClosed	24.0	429.0	10.83	361.0	9.11
sum_of_paym_months	325.0	14.0	0.35	361.0	9.11
all_credits	30.0	0.0	0.00	361.0	9.11
Active_not_cc	9.0	1178.0	29.72	361.0	9.11
own_closed	9.0	2058.0	51.93	361.0	9.11
min_MnthAfterLoan	98.0	153.0	3.86	361.0	9.11
max_MnthAfterLoan	129.0	11.0	0.28	361.0	9.11
dlq_exist	3.0	1552.0	39.16	361.0	9.11

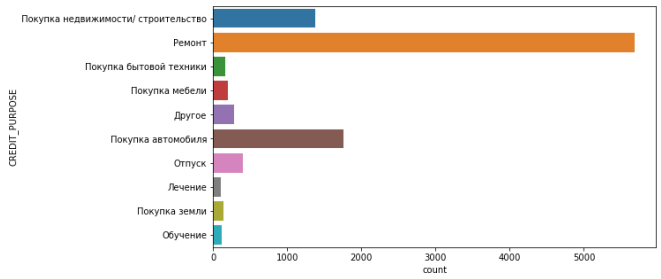
Предварительный анализ данных

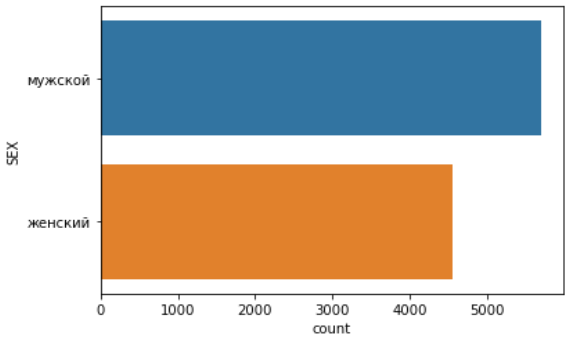
Построим гистограммы для всех некатегориальных переменных

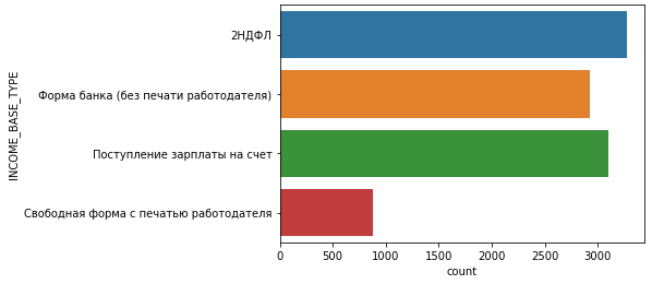


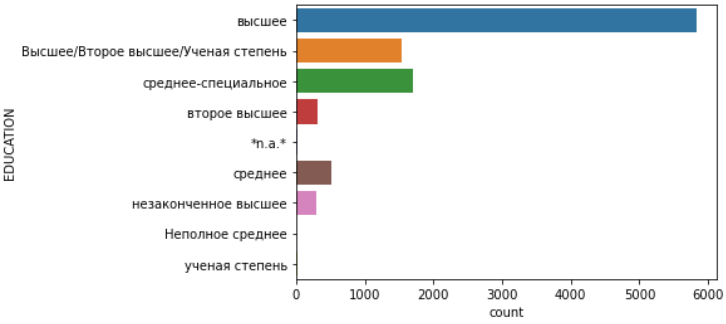
Как видно из графиков, в выборке преобладают достаточно молодые клиенты, в основном от 30 до 40 лет и без детей или с маленькими детьми. Большая часть клиентов работает меньше 4 лет и получает зарплату от 250 тыс.руб. Большинство клиентов брали кредит за последний год, не имеют аккаунта в банке, просрочки по платежам, если происходят, меньше 30 дней.

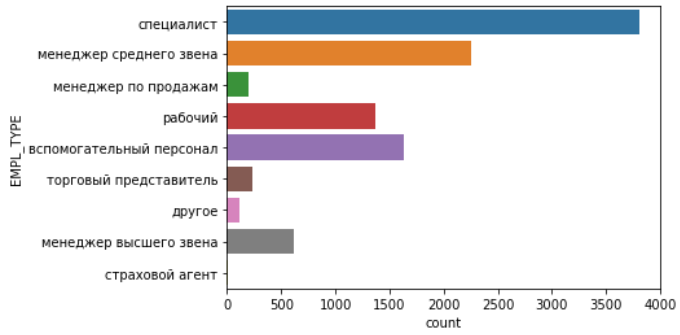
Также рассмотрим диаграммы по некатегориальным переменным.

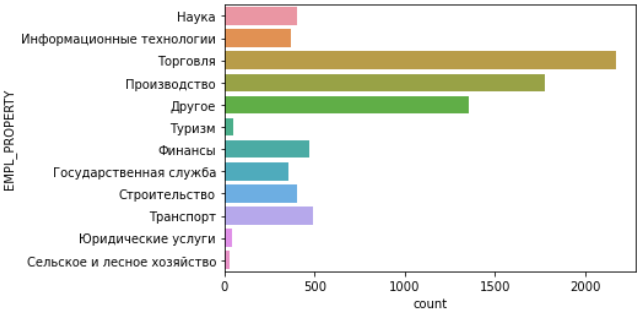

CREDIT_PURPOSE
Покупка недвижимости/ строительство
Ремонт
Покупка бытовой техники
Покупка мебели
Другое
Покупка автомобиля
Отпуск
Лечение
Покупка земли
Обучение
0
1000
2000
3000
4000
5000
count

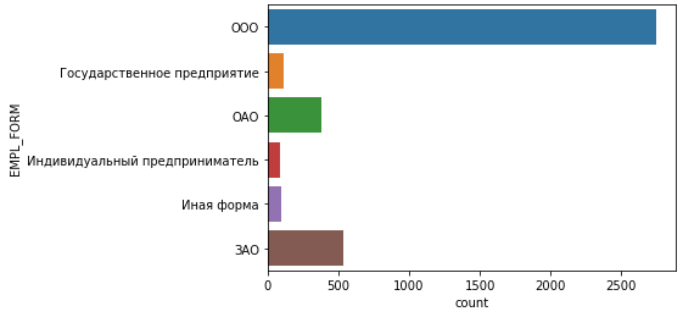

SEX
мужской
женский
0
1000
2000
3000
4000
5000
count

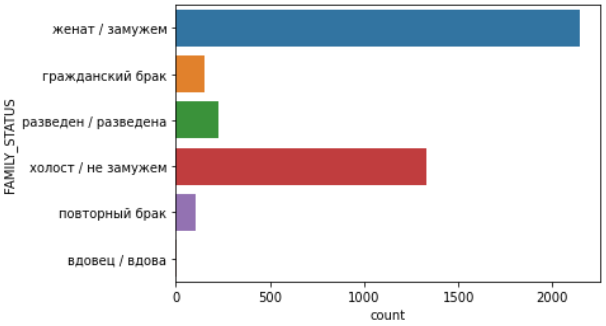

INCOME_BASE_TYPE
2НДФЛ
Форма банка (без печати работодателя)
Поступление зарплаты на счет
Свободная форма с печатью работодателя
0
500
1000
1500
2000
2500
3000
count


EDUCATION
высшее
Высшее/Второе высшее/Ученая степень
среднее-специальное
второе высшее
n.a.
среднее
незаконченное высшее
Неполное среднее
ученая степень
0
1000
2000
3000
4000
5000
6000
count


EMPL_TYPE
специалист
менеджер среднего звена
менеджер по продажам
рабочий
вспомогательный персонал
торговый представитель
другое
менеджер высшего звена
страховой агент
0
500
1000
1500
2000
2500
3000
3500
4000
count


EMPL_PROPERTY
Наука
Информационные технологии
Торговля
Производство
Другое
Туризм
Финансы
Государственная служба
Строительство
Транспорт
Юридические услуги
Сельское и лесное хозяйство
0
500
1000
1500
2000
count

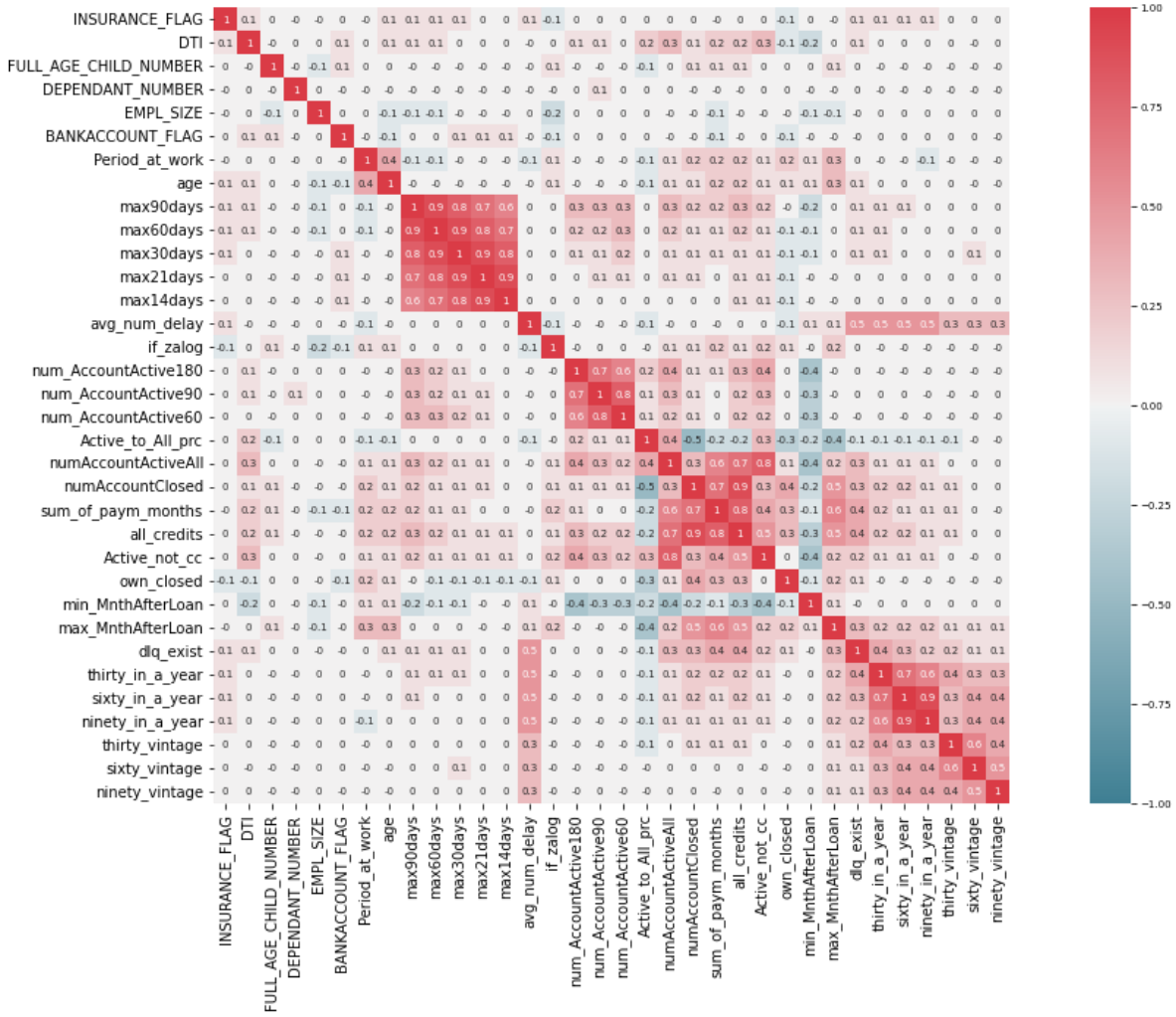

EMPL_FORM
ООО
Государственное предприятие
ОАО
Индивидуальный предприниматель
Иная форма
ЗАО
0
500
1000
1500
2000
2500
count


FAMILY_STATUS
женат / замужем
гражданский брак
разведен / разведена
холост / не замужем
повторный брак
вдовец / вдова
0
500
1000
1500
2000
count

Часть 2.

Для сегментации данных будем использовать метод k-means и дерево решений.

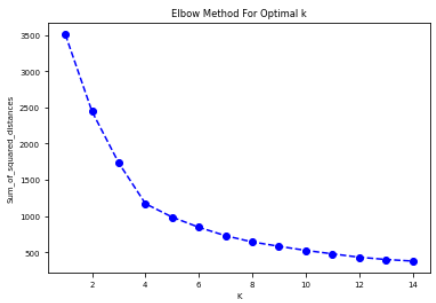
Сначала построим матрицу корреляций:

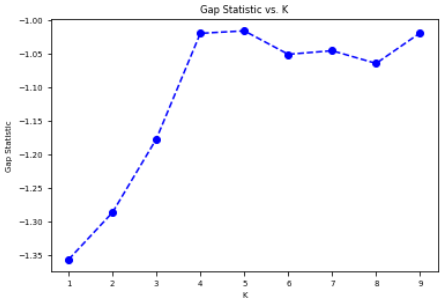


Как видно из матрицы, корреляция в целом достаточно слабая, но есть несколько очагов сильной корреляции между переменными, значения линейно зависимы, таких как, например, количество запросов в бюро кредитной истории за последние 14, 21, 30, 60 и 90 дней. Можно заменить существующие переменные, например, их отношением, но в рамках задания и используемых методов сегментации, в этом нет необходимости.

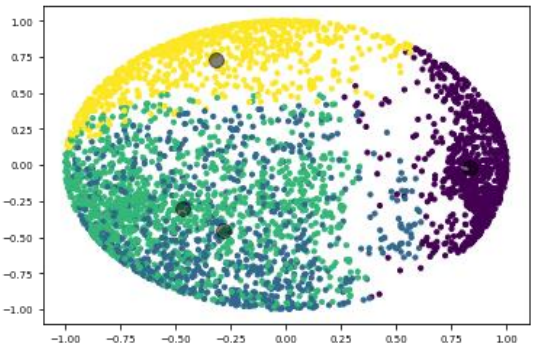
K-means

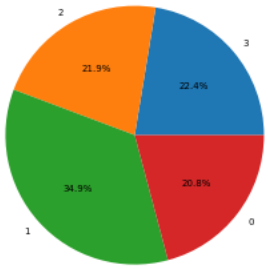
Воспользуемся методом главных компонент для понижения размерности и определим оптимальное количество кластеров при помощи gap statistics и elbow method.

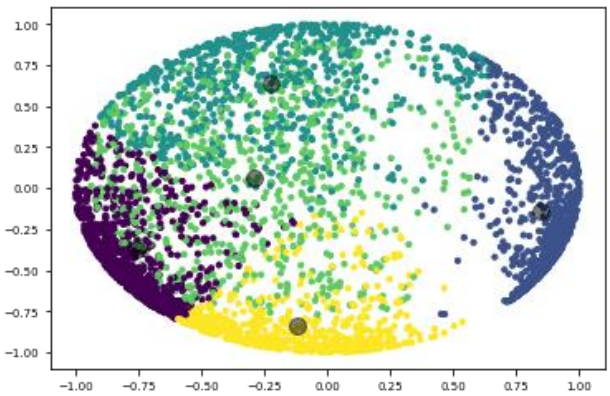


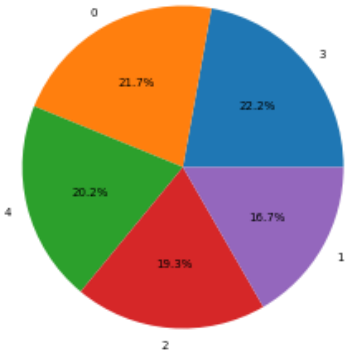


Исходя из графиков, предположим оптимальное количество кластеров равное 4. Однако, при выборе 5 кластеров, результат работы алгоритма выглядит лучше.



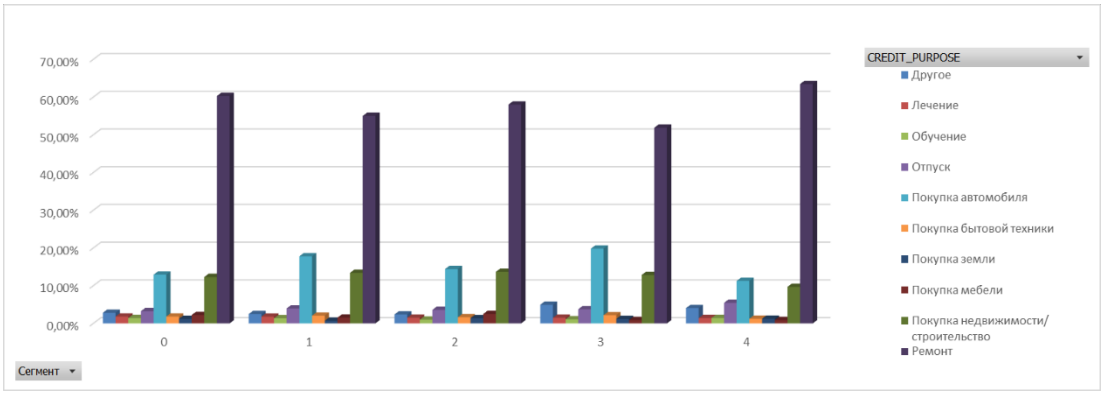


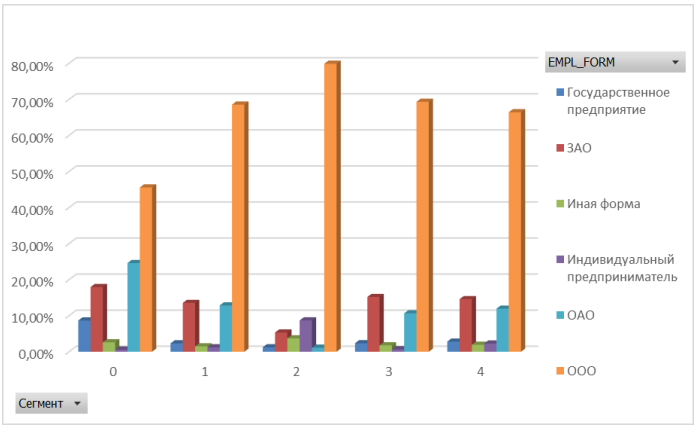


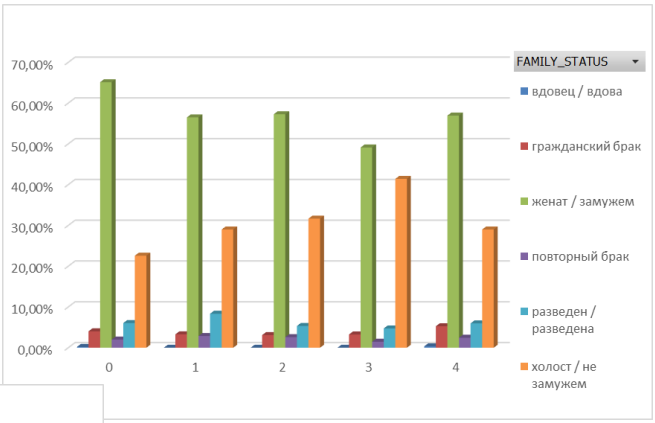


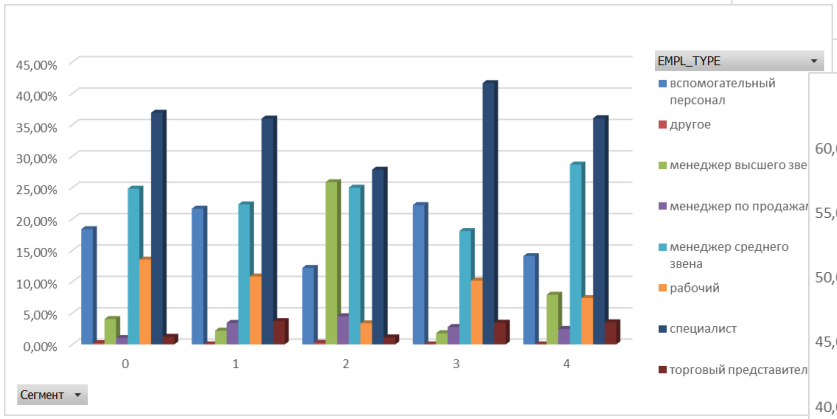
Результаты работы алгоритма

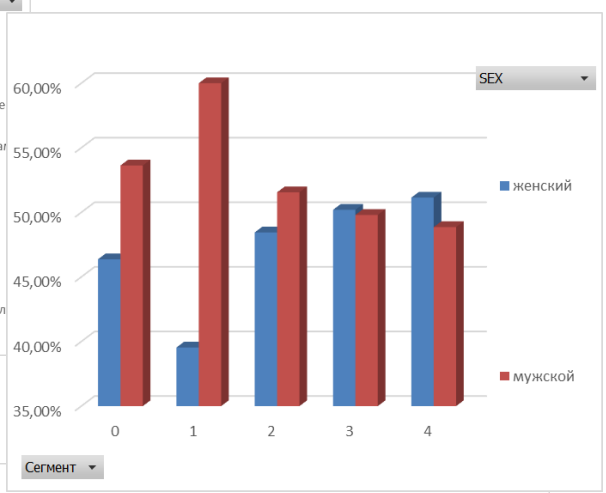
Теперь рассмотрим гистограммы признаков для сегментов:

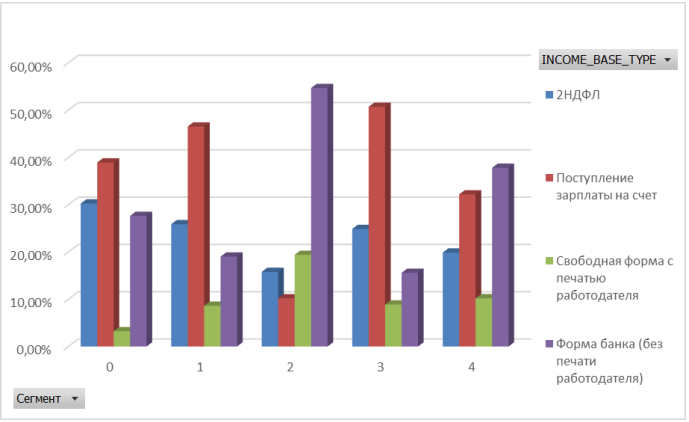


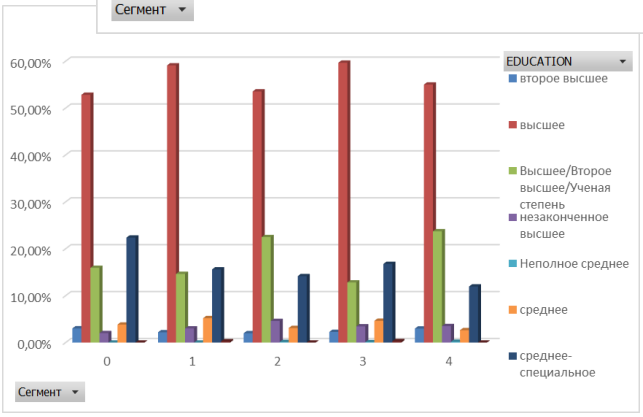






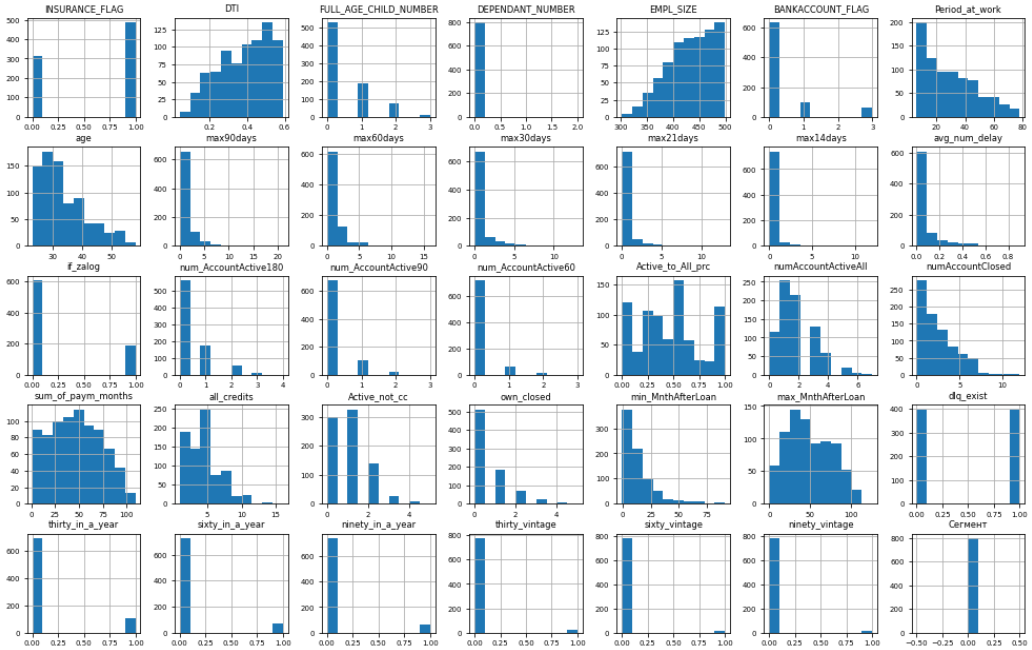




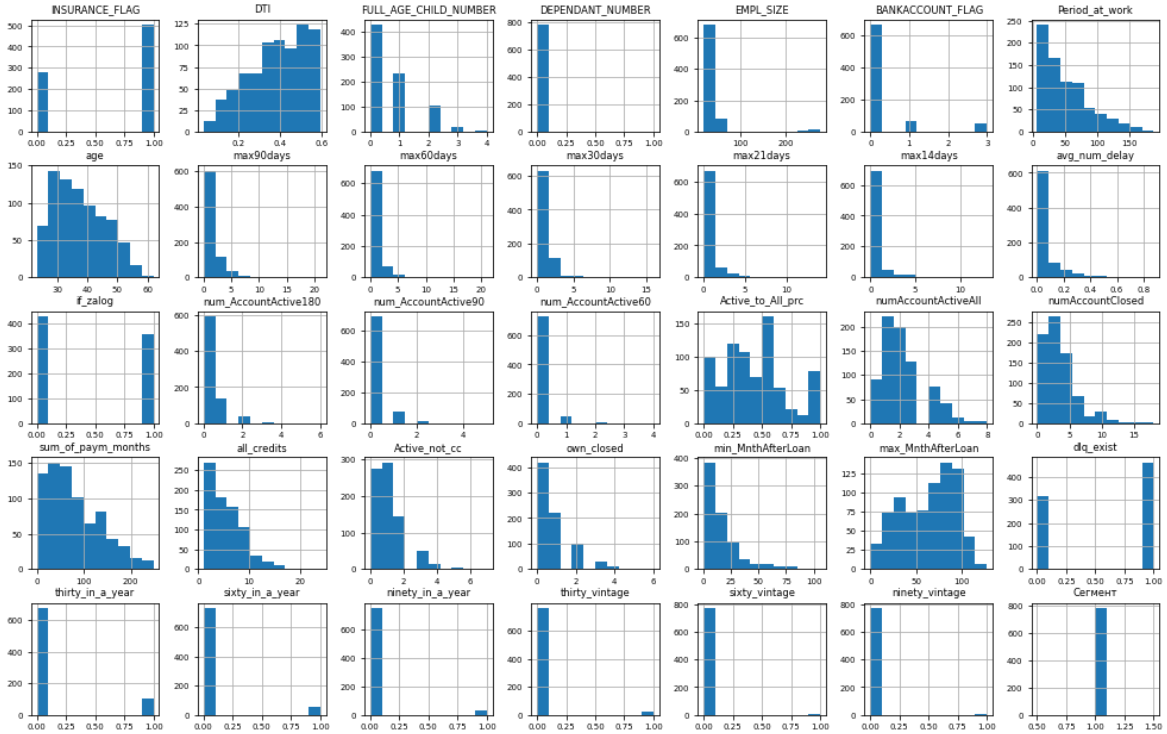


И отдельно гистограммы для числовых признаков по каждому из сегментов.

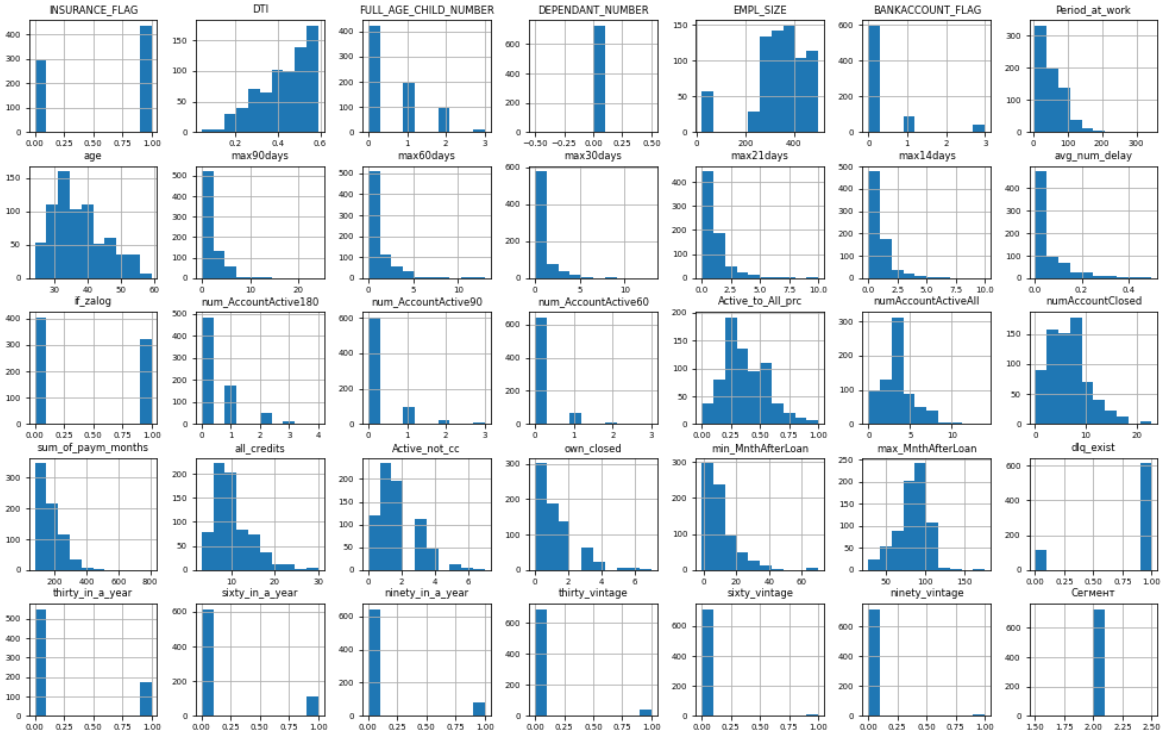
Сегмент 1:



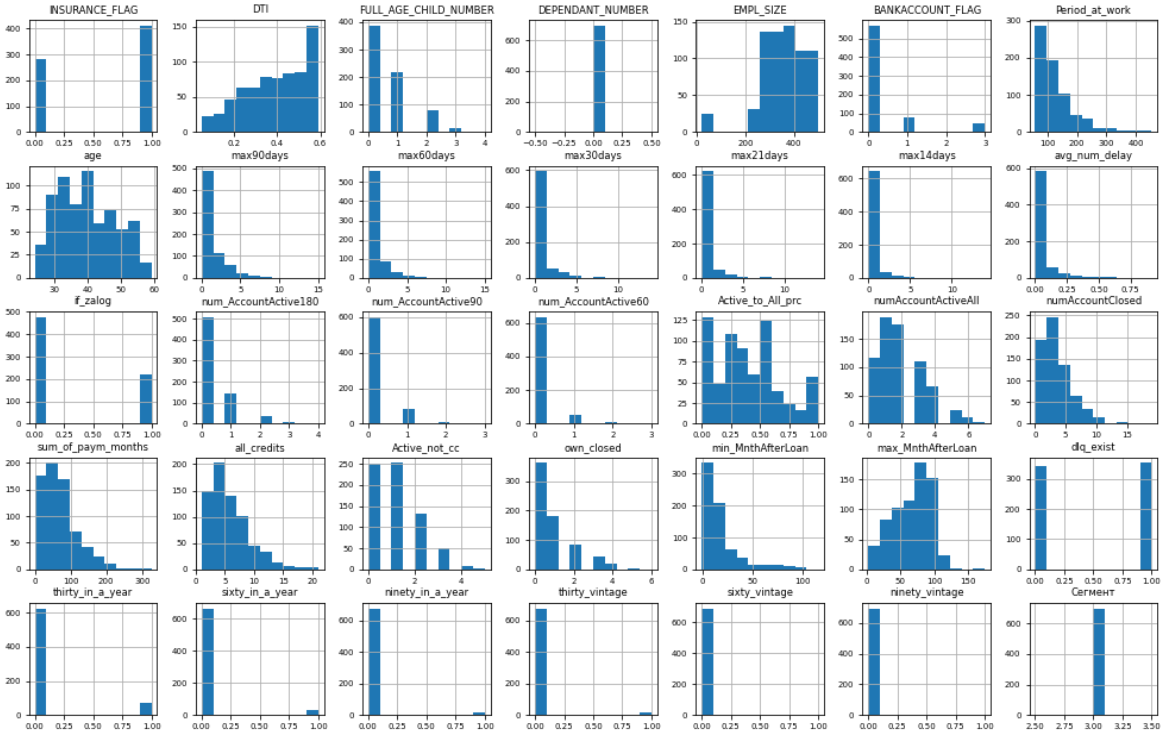
Сегмент 2



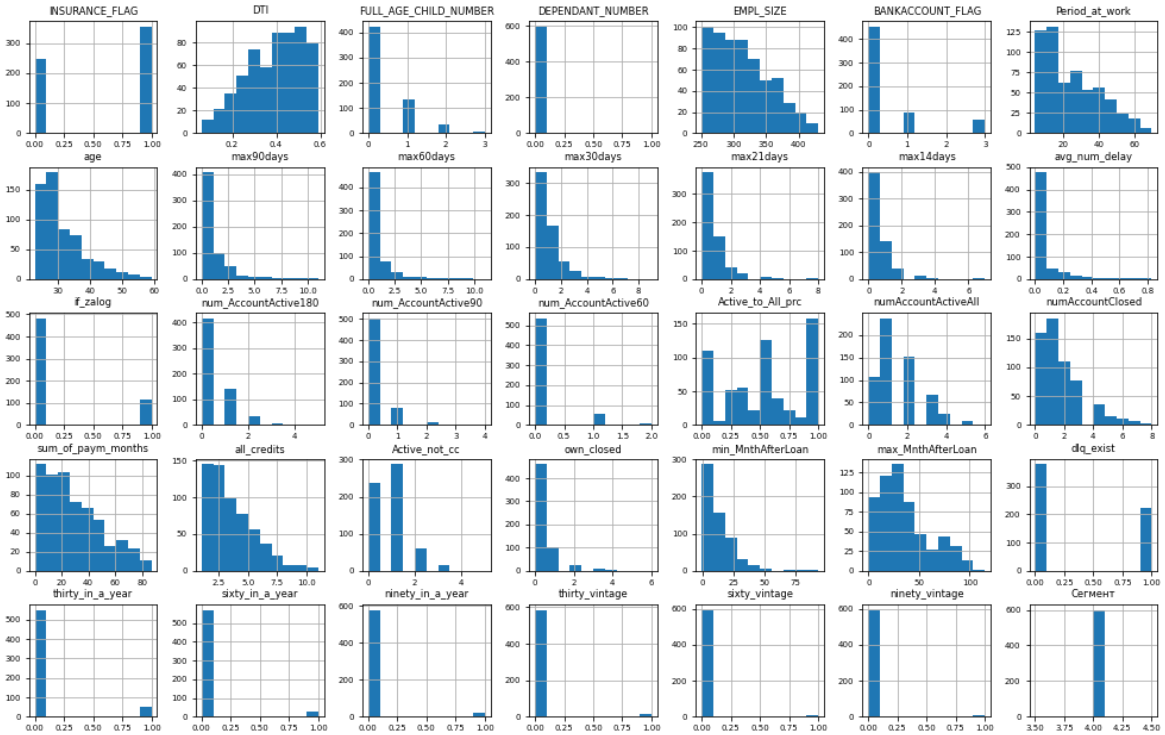
Сегмент 3



Сегмент 4



Сегмент 5


INSURANCE_FLAG
DTI
FULL_AGE_CHILD_NUMBER
DEPENDANT_NUMBER
EMPL_SIZE
BANKACCOUNT_FLAG
Period_at_work
age
max90days
max60days
max30days
max21days
max14days
avg_num_delay
if_zalog
num_AccountActive180
num_AccountActive90
num_AccountActive60
Active_to_All_prc
numAccountActiveAll
numAccountClosed
sum_of_paym_months
all_credits
Active_not_cc
own_closed
min_MnthAfterLoan
max_MnthAfterLoan
dlq_exist
thirty_in_a_year
sixty_in_a_year
ninety_in_a_year
thirty_vintage
sixty_vintage
ninety_vintage
Сегмент

Описание сегментов

Сегмент 1: В этом сегменте самая высокая доля замужних/женатых людей и клиентов со средне-специальным образованием, рабочих, наибольшее количество кредитных карт,

Сегмент 2: Чаще всех берут кредит под залог, наиболее давние клиенты банка, чаще всех оформляют страховки, преимущественно мужчины, в основном специалисты.

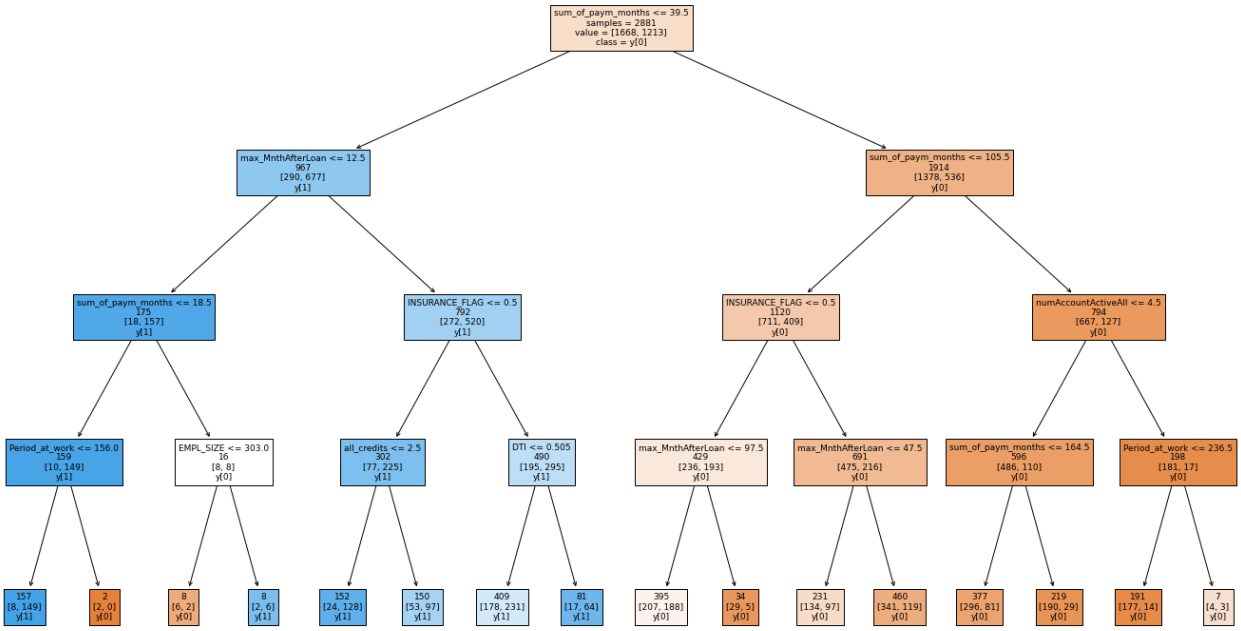
Сегмент 3: Наибольшее отношение долга к доходам, почти все имеют просрочку выплат по кредиту на данный момент, наибольшее количество кредитов. В этом сегменте примерно равное количество специалистов, менеджеров высшего и среднего звена

Сегмент 4: Наибольшее количество детей у клиентов, большая часть клиентов брали кредит недавно, примерно равное количество клиентов с просрочками выплат и без, наибольший стаж работы, примерно равное количество клиентов женского и мужского пола.

Сегмент 5: Наибольшее количество людей до 30, чаще всего не имеют залога, наименьшая сумма выплат по кредитам за месяц, преобладают клиенты женского пола

Решающее дерево

Построим решающее дерево для определение сегментов клиентов, которые с наименьшей вероятностью будут просрочивать платежи.



Дерево выделило следующие сегменты клиентов, которые скорее всего будут вовремя вносить платежи по кредитам:

- Недавние (меньше года) клиенты банка, стаж работы меньше 13 лет, с суммой платежей по кредиту в месяц не больше 18 тысяч рублей.
- Давние клиенты банка со страховками и активной кредитной историей, суммой платежей не больше 39,5 тысяч рублей.
- Давние (больше года) клиенты, у которых не больше 2 кредитов, не берут страховки. Ежемесячная выплата по кредитам не больше 39,5 тысяч рублей.

Обоснование выбора методов
Дерево решений

Плюсы	**Минусы**
Легко интерпретируется в бизнес-правило	Разделяющая граница, которую строит дерево, не всегда проходит наиболее оптимальным способом
Быстро учится	Очень легко переобучается
Поддерживает числовые и категориальные признаки	Нестабильно, очень зависит от малейших изменений в данных

K means

Плюсы	**Минусы**
Прост в реализации	Необходимо подбирать количество кластеров
Легко масштабируется	Плохо справляется с задачей, когда элемент принадлежит сразу к нескольким кластерам в равной степени
Быстро обучается	Чувствителен к выбросам