**Правительство Российской Федерации**

**Федеральное государственное автономное образовательное**

**учреждение высшего образования**

**Национальный исследовательский университет**

**«Высшая школа экономики»**

Дисциплина «Data Science в клиентской и текстовой аналитике»

**Домашнее задание**

Выполнила: Абдряхимова Аделя Хамитовна

Образовательная программа «Бизнес-информатика»

Высшая школа бизнеса

3-й курс, группа ББИ181

Москва

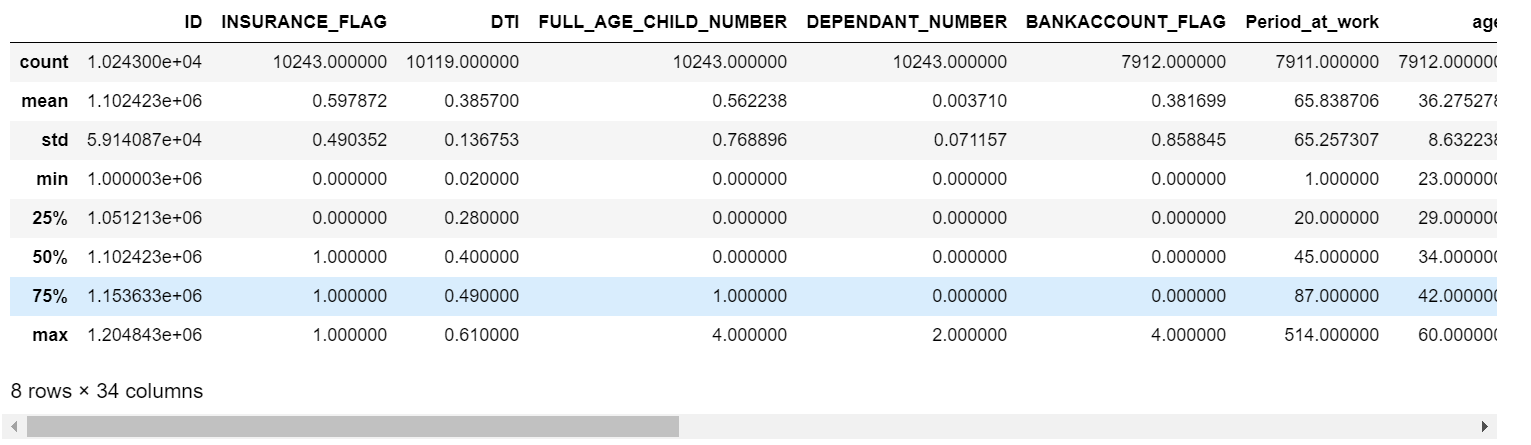
2020

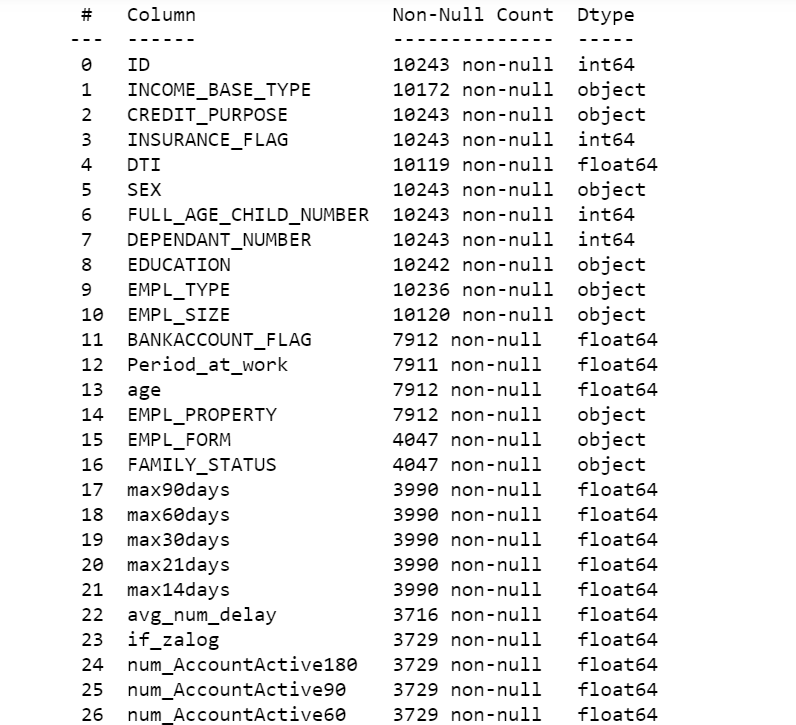
Вариант 3.

Часть 1

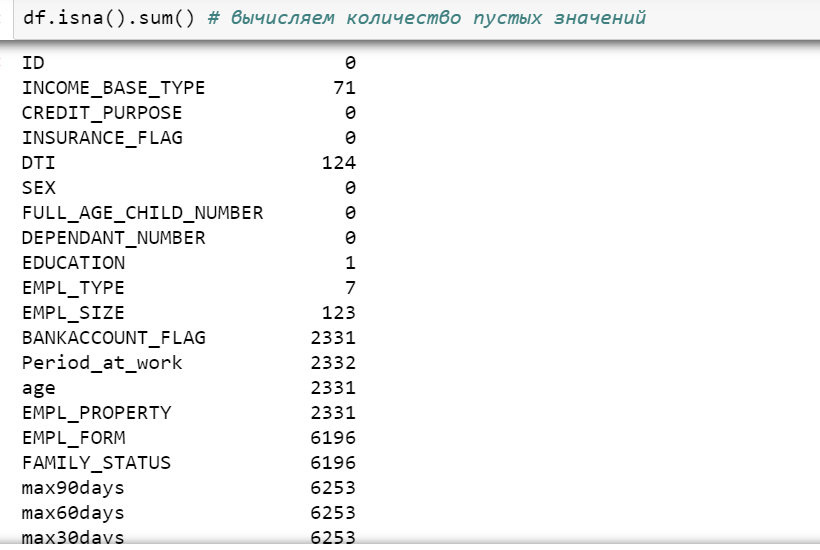
Исследуем распределения по данным (подробнее см. Jupiter Notebook)

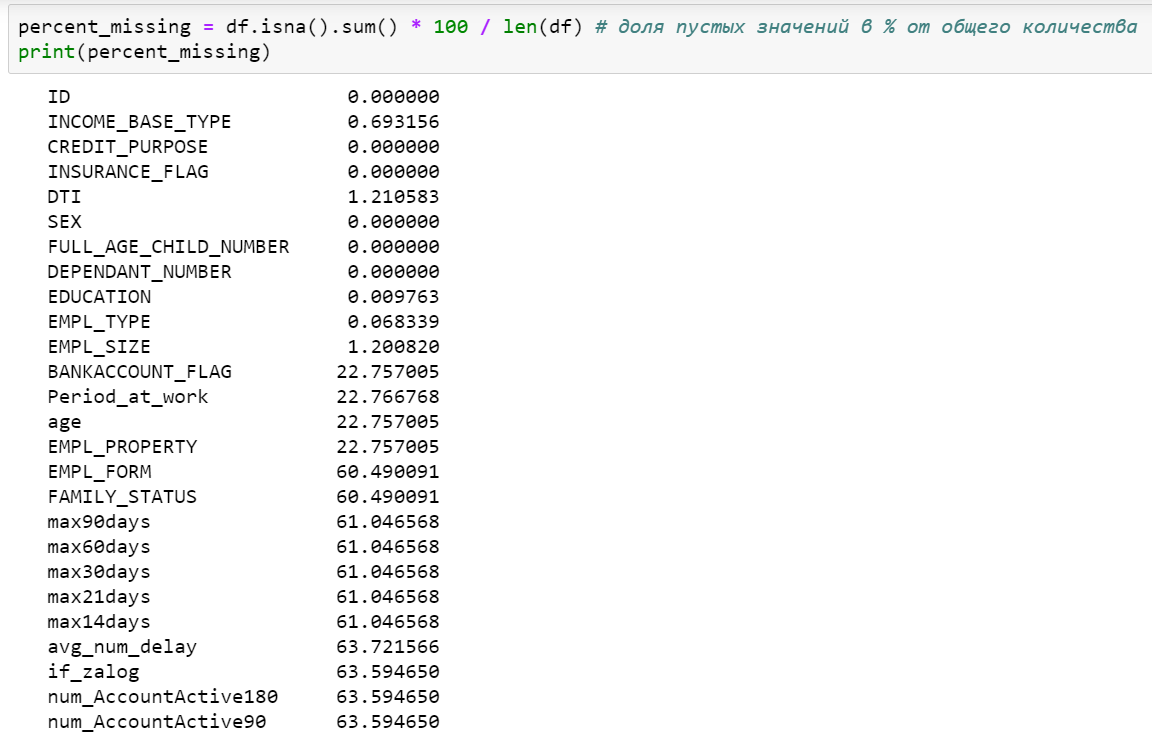
Для каждой переменной набора данных были рассчитаны среднее значение, медиана, стандартное отклонение, минимум, максимум, а также тип данных.

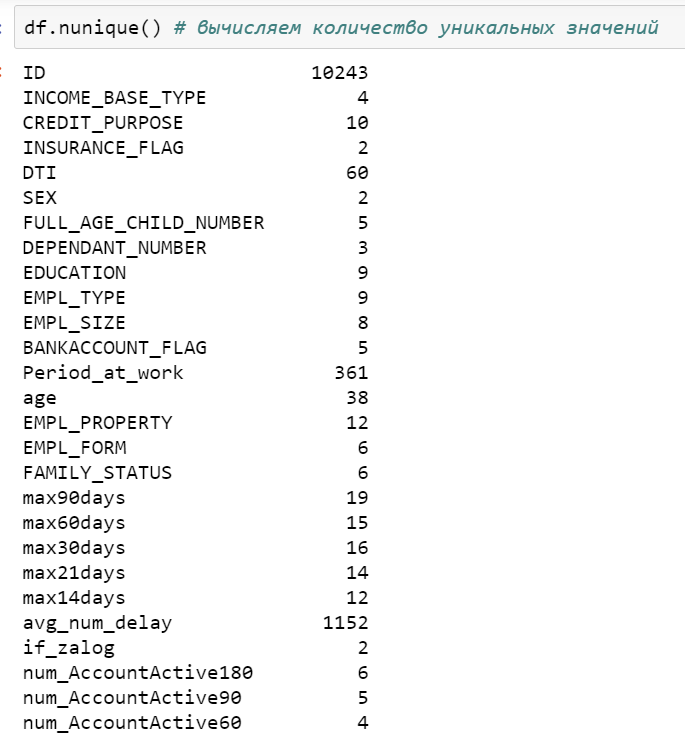


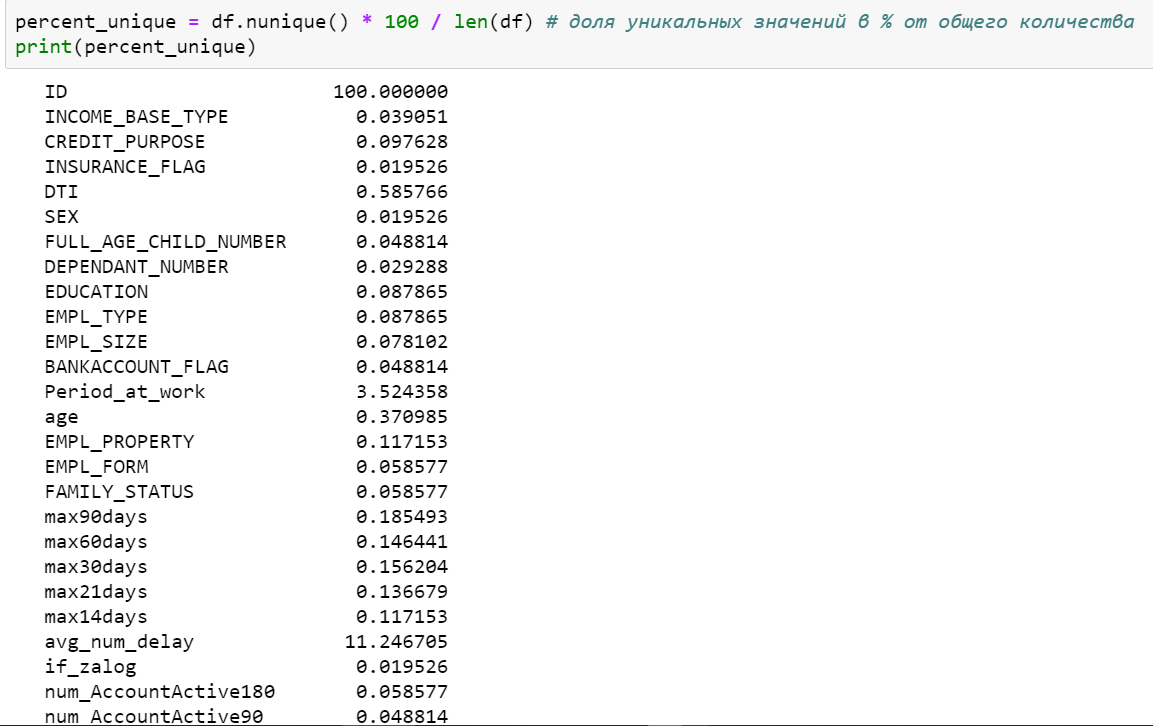


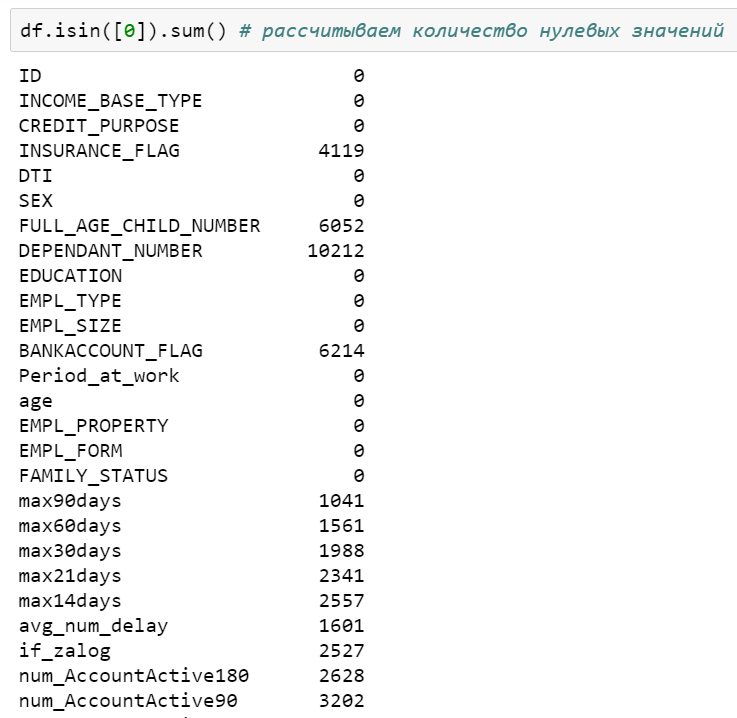
Для каждой переменной набора данных также были рассчитаны: кол-во уникальных, нулевых и пустых значений, а также доли в % от общего количества.

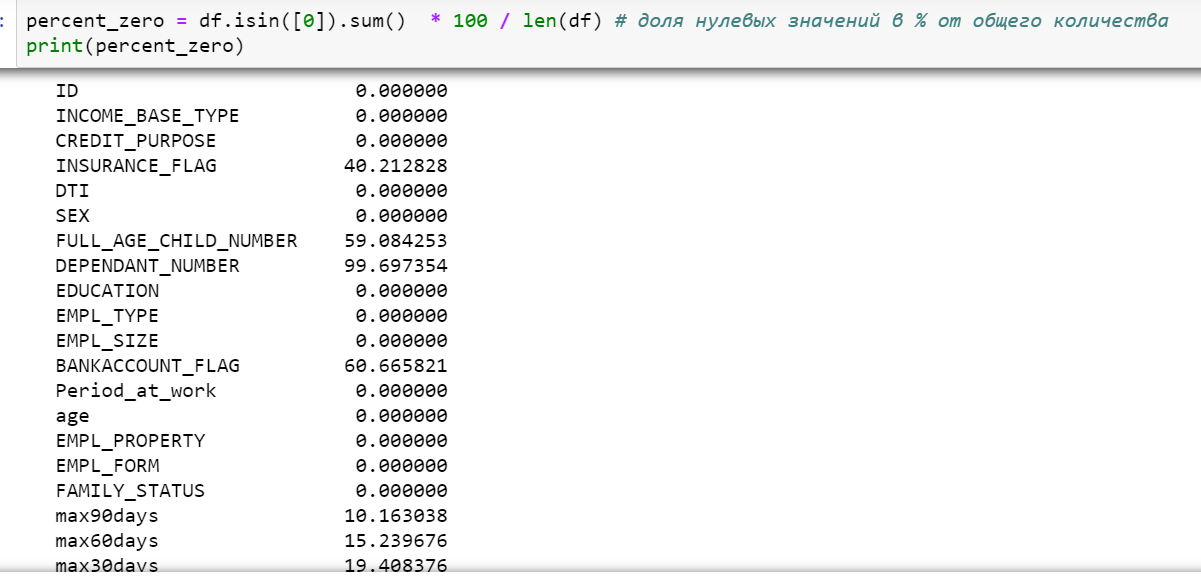






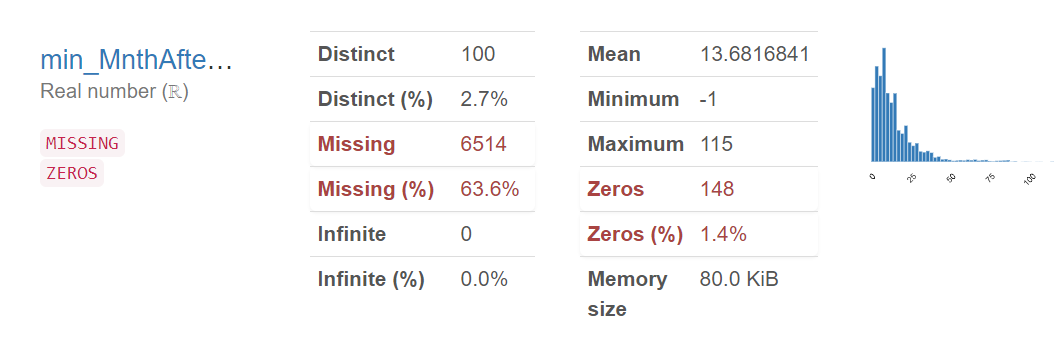




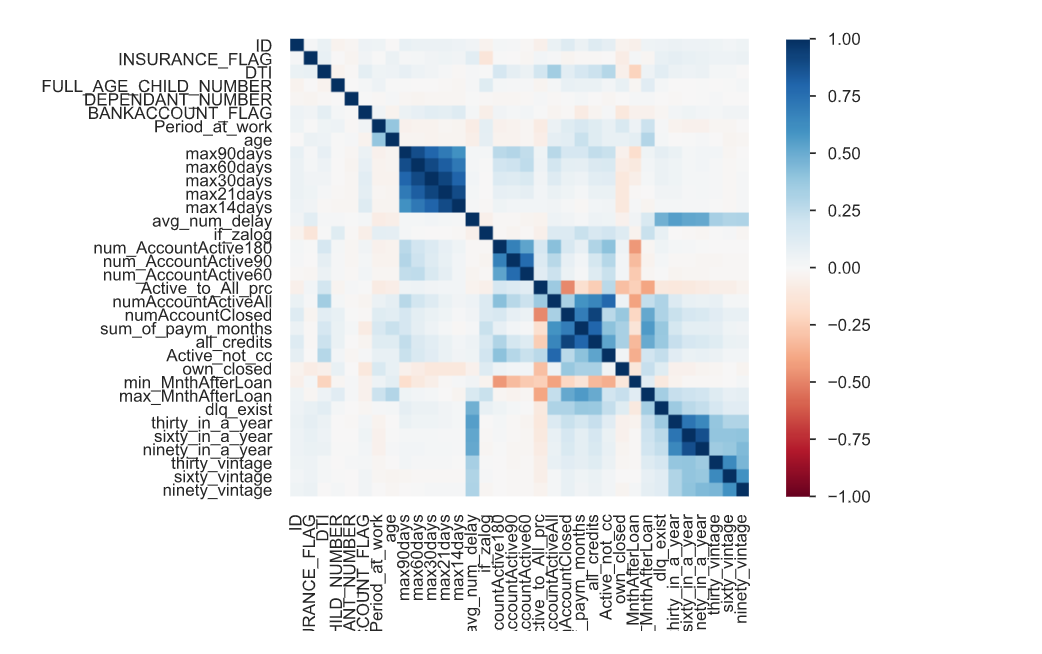


В результате анализа было найдено, что дубликаты в данных отсутствуют, все ID уникальны. Кроме того, в данных оказалось много пустых значений: у переменных Empl\_Form, Family\_Status, max90days, max60days, max30days, max21days,max14days,avg\_num\_delay,if\_zalog,num\_AccountActive180,num\_AccountActive90,num\_AccountActive60,Active\_to\_All\_prc,numAccountActiveAll,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. Поэтому было принято заменить пропущенные значения на медиану или моду. Если данные асимметричны, пропущенные значения заменялись на среднее, если симметричны – на медиану.

Также при анализе было выявлено недопустимое значение: минимальное значение min\_MnthAfterLoan (минимальное кол-во месяцев, которое прошло с момета взятия прошлого кредита) составляет -1.

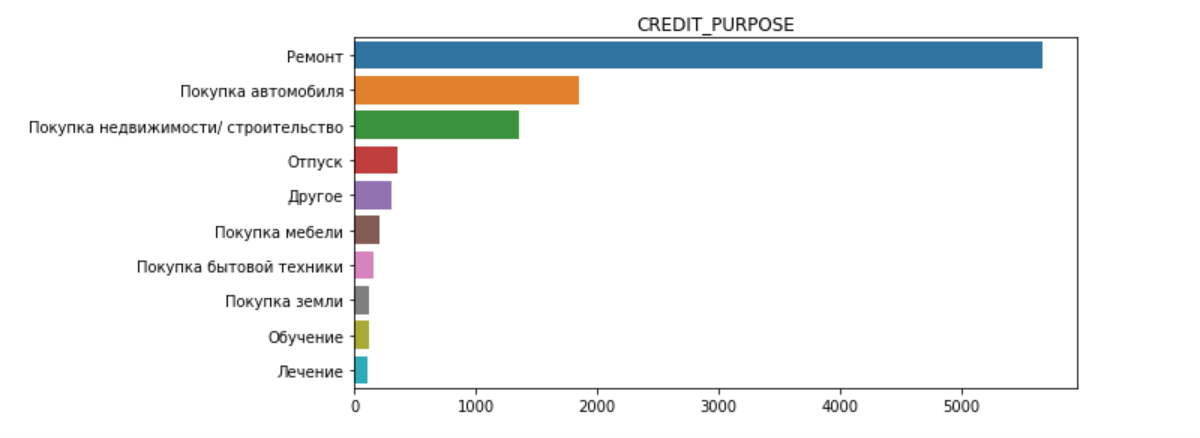


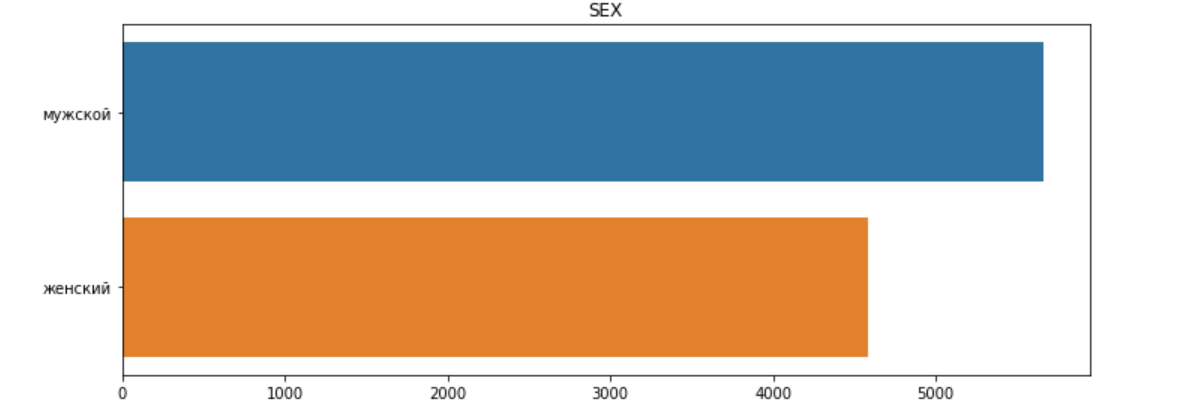
По числовым переменным была построена корреляционная карта:

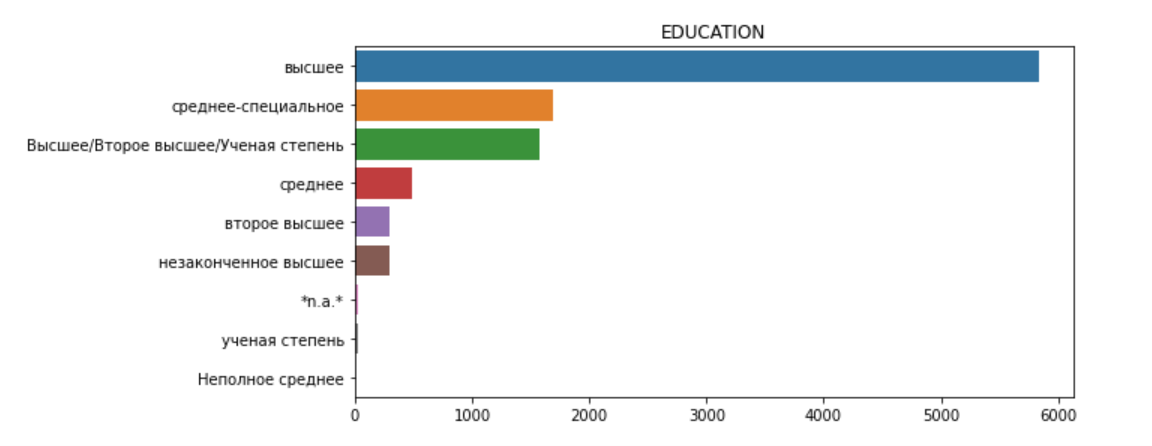


По этой карте можно определить, что в целом между переменными наблюдается слабая корреляция– линейно-регрессионный анализ в данном случае неуместен.

Затем исследуем распределение данных категориальных переменных, таких как образование, пол, цель получения кредита и т.д. Примеры диаграмм расположены ниже.







Примеры диаграмм числовых переменных приведены ниже.

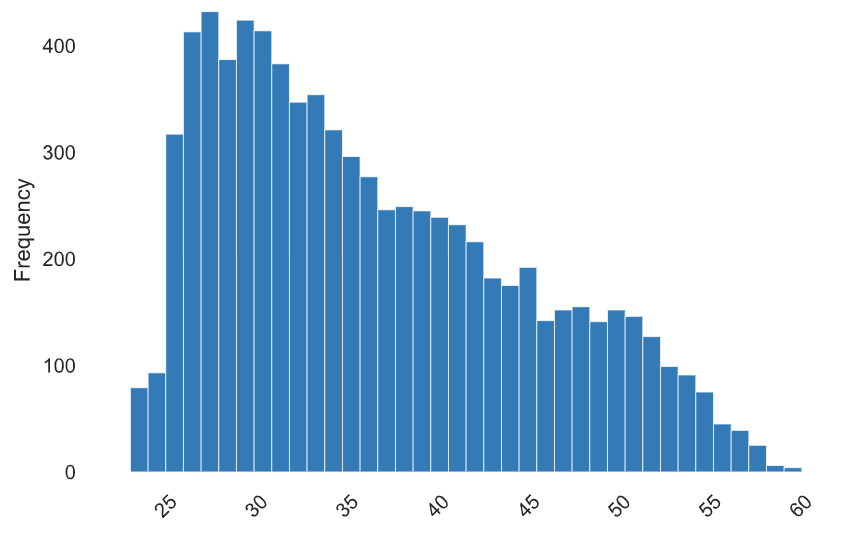


Рисунок . Age

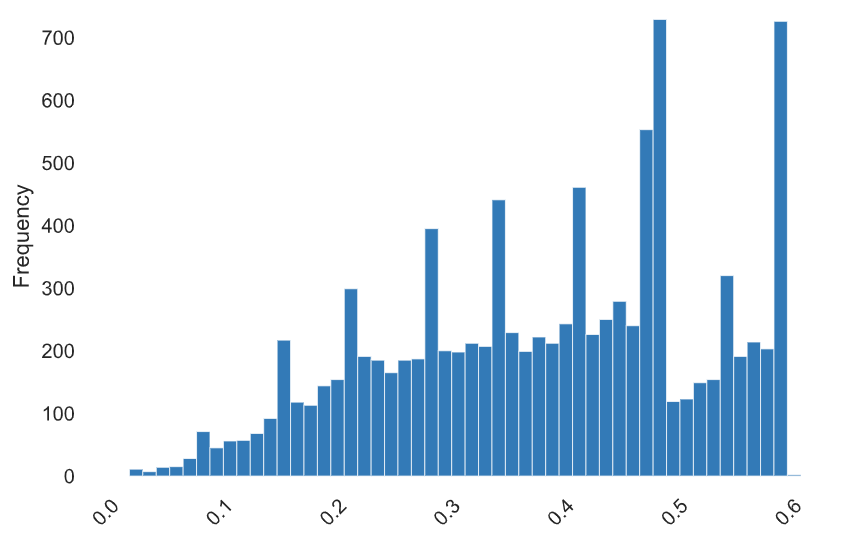


Рисунок . DTI

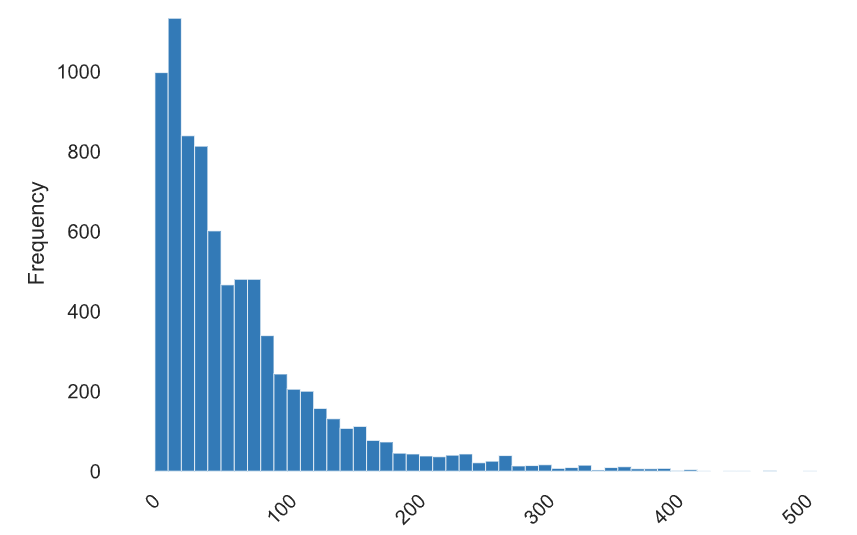


Рисунок . Period at work

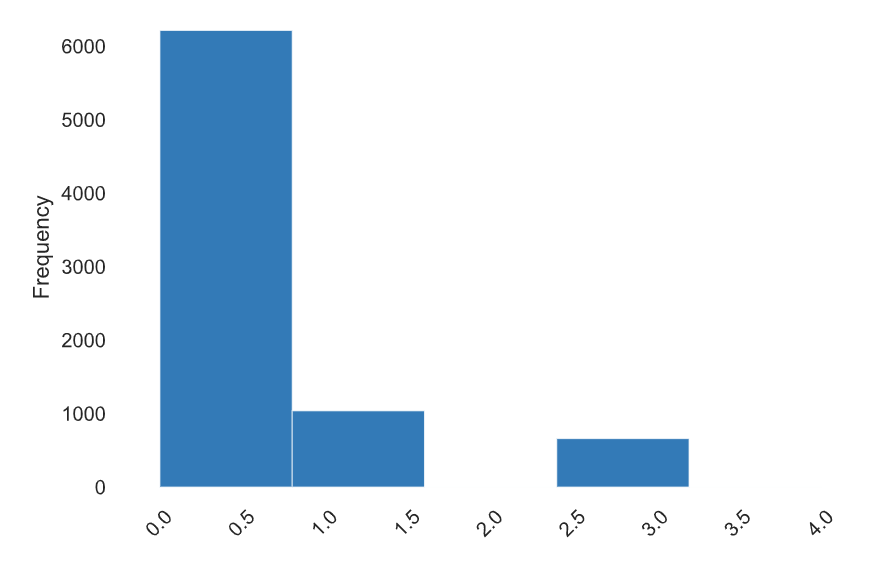


Рисунок . Back account flag

Из анализа диаграмм можно сделать вывод, что в основном в выборке присутствуют в основном молодые люди (до 40 лет), равно, как мужчины, так и женщины. Также большую часть выборки составляют люди, имеющие высшее образование/второе-высшее образование/ученую степень (75,5%). Было найдено, что основной целью получения кредита (55,3%) является получение ремонта, у достаточно большой части выборки высокий DTI (отношение долга к доходам), у большей части людей нет иждивенцев (DEPENDANT\_NUMBER, 99,7%). Кроме того, у большей части выборки небольшой опыт работы (до 100 дней), низкое количество активных кредитов (all\_credits), низкое количество платежей за последний месяц (sum\_of\_paym\_months) и низкое количество активных кредитных счетов (numAccountActiveAll).

Из финальной выборки были исключены колонки sixty\_in\_a\_year,sixty\_vintage,thirty\_in\_a\_year,thirty\_vintage,ninety\_in\_a\_year,ninety\_vintage (переменные, у которых слишком большое количество пропусков и которые не несут ценности для дальнейшего анализа). Также была исключена колонка ID, так как для дальнейшего анализа она не несет ценности.

Также были исключены колонка min\_MnthAfterLoan, man\_MnthAfterLoan из-за большого количества пропущенных значений и странностей в данных.

Выводы по части 1:

1. У переменных наблюдается слабая корреляция, линейно-регрессионный анализ неуместен.
2. В данных достаточно много пропущенных значений (43,2%). В связи с этим было решено заменить пропущенные значения на моду или медиану.
3. Из колонки Education были удалены значения \*n.a.\* и пропущенные значения. Из колонки Empl\_Size также было решено удалить пропущенные значения.
4. В финальную витрину не вошли переменные ID, min\_MnthAfterLoan, max\_MnthAfterLoan,sixty\_in\_a\_year,sixty\_vintage,thirty\_in\_a\_year,thirty\_vintage,ninety\_in\_a\_year,ninety\_vintage.

С учетом всех корректировок была составлена итоговая витрина.

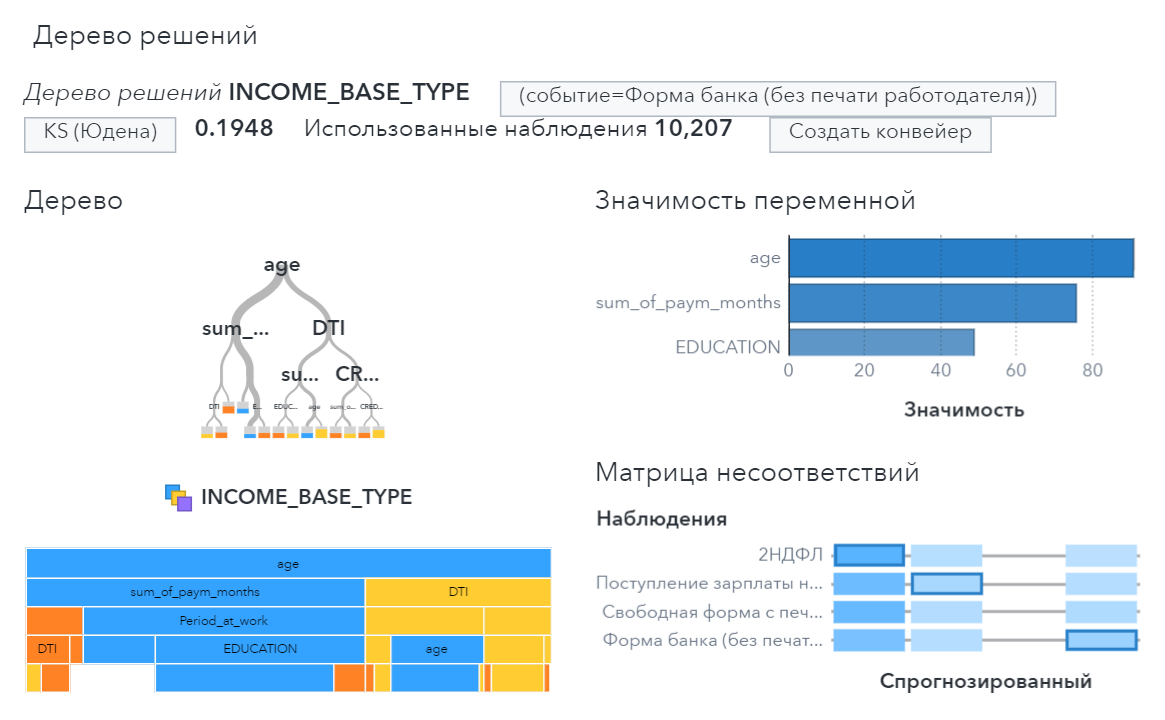
Часть 2

**Сегментация с учителем в SAS VA. Дерево решений.**

Ссылка на отчет в SAS: http://tiger.cs.msu.ru/links/resources/report?uri=%2Freports%2Freports%2F958930a7-38c9-4513-abd8-0d3ca44d18f7&page=vi4042

Для построения дерева решений было решено взять в качестве целевой переменной INCOME\_BASE\_TYPE, так как у этой переменной нет перекосов в какую-либо сторону (коэффициент асимметрии skewness близок к нулю). В качестве предикторов были взяты наиболее важные переменные: age, Sex, Education, DTI, Period\_at\_work, all\_credits, sum\_of\_paym\_months, dlq\_exists, Credit\_purpose.

Максимальную глубину дерева было решено ограничить 5, а максимум ветвей –2.



Дерево решений выявило 14 сегментов (отобранные сегменты выделены):

* **Молодые люди, от 23 до 34 лет, с маленьким или средним отношением долгов к доходам (до 0.34), цель получения кредита которых – покупка бытовой техники**
* **Молодые люди, от 23 до 34 лет, с маленьким или средним отношением долгов к доходам (до 0.34), цель получения кредита которых – отпуск или покупка мебели**
* Молодые люди, от 23 до 34 лет, с маленьким или средним отношением долгов к доходам (до 0.34), с маленькими или средними тратами в месяц (до 74 тыс. в месяц), цель получения кредита которых – ремонт, покупка недвижимости, покупка земли, покупка автомобиля, обучение, лечение. Слишком широкий сегмент (см. файл SAS), также маловероятно, что люди, получающие кредит на ремонт/лечение, будут тратить небольшие суммы денег в месяц
* **Молодые люди, от 23 до 34 лет, с маленьким или средним отношением долгов к доходам (до 0.34), с большими тратами в месяц (от 74 тыс. в месяц), цель получения кредита которых – ремонт, покупка недвижимости, покупка земли, покупка автомобиля, обучение, лечение. Этот сегмент не широкий и возможен, в отличие от предыдущего, так как траты соответствуют цели получения кредита**
* Молодые люди, от 23 до 25 лет, со средним или большим отношением долгов к доходам (от 0.34), с ежемесячными тратами до 64 тыс. Довольно мало информации об этом сегменте, нет информации об образовании и/или цели получения кредита.
* Молодые люди, от 25 до 34 лет, со средним или высоким отношением долгов к доходам (от 0.34), с маленькими или средними тратами в месяц (до 64 тыс. в месяц)
* **Молодые люди, от 23 до 34 лет, имеющие высшее образование/второе высшее или ученую степень, со средним или высоким отношением долгов к доходам (от 0.34), с ежемесячными тратами от 64 тыс.**
* Молодые люди, от 23 до 34 лет, имеющие высшее/второе высшее/ученую степень или незаконченное высшее, или среднее, или средне-специальное, со средним или высоким отношением долгов к доходам (от 0.34), с ежемесячными тратами от 64 тыс., со средним или большим отношением к долгам (от 0.34). Маловероятно, что в одном сегменте могут находиться люди со средним образованием и с ученой степенью, тем более маловероятно, что люди со средним образованием будут тратить большие суммы денег.
* Взрослые люди старше 34 лет, имеющие второе/второе высшее/ученую степень, со средними тратами в месяц (до 64 тыс.), работающие меньше 2 месяцев.
* Взрослые люди старше 34 лет, имеющие высшее/второе высшее/ученую степень или незаконченное высшее, или среднее, или средне-специальное, работающие меньше 75 дней, тратящие меньше 64 тыс. в месяц. Слишком широкий сегмент.
* **Взрослые люди старше 34 лет, тратящие меньше 64 тыс. в месяц, работающие больше 75 дней**
* **Взрослые люди, имеющие высшее/второе высшее/ученую степень или незаконченное высшее, тратящие больше 64 тыс. в месяц**
* Взрослые люди старше 34 лет, имеющие высшее, или второе высшее, или среднее, или средне-специальное, тратящие больше 64 тыс. в месяц, с отношением долгов к доходам до 0.48. Маловероятно, что люди, имеющие только среднее образование, будут тратить больше 64 тыс. в месяц. Также слишком велик разброс переменной DTI.
* Взрослые люди старше 34 лет, имеющие высшее, или второе высшее, или среднее, или средне-специальное, тратящие больше 64 тыс. в месяц, с высоким отношением долгов к доходам (от 0.48). DTI соответствует большим ежемесячным тратам. Однако, как и с предыдущим сегментом, маловероятно, что люди, имеющие только среднее образование, будут тратить больше 64 тыс. в месяц.

Таким образом, можно выделить следующие портреты клиентов:

1. Молодые люди от 23 до 30-34 лет, имеющие высшее образование/второе высшее или ученую степень, со средним или высоким отношением долгов к доходам (от 0.34), ежемесячно тратящие большие суммы денег (от 64 тыс.). Довольно рисковый сегмент, учитывая коэффициент DTI. Рекомендация – не выдавать такому сегменту кредит, так как вероятность выплаты маловероятна. Доля сегмента в общем обороте клиентов: 3%
2. Молодые люди, от 23 до 34 лет, с маленьким или средним отношением долгов к доходам (до 0.34), цель получения кредита которых – отпуск/ покупка мебели/покупка бытовой техники. Скорее всего, эти люди недавно закончили ВУЗ, обладающие стабильным доходом, создающие семьи, уже имеющие собственную недвижимость. Довольно стабильный сегмент, давать кредит можно. Доля сегмента в общем обороте клиентов: 1.3%
3. Молодые люди, от 23 до 34 лет, с маленьким или средним отношением долгов к доходам (до 0.34), с большими тратами в месяц (от 74 тыс. в месяц), цель получения кредита которых – ремонт, покупка недвижимости, покупка земли, покупка автомобиля, обучение, лечение. Возможно, эти люди уже закончили ВУЗ, обладают высоким заработком и работают на крупных должностях. Доля сегмента в общем обороте клиентов: 1.5%
4. Взрослые люди старше 34 лет, тратящие меньше 64 тыс. в месяц, работающие больше 75 дней. Возможно, это люди с небольшим или средним заработком, занимающие средние или маленькие должности. Для такого сегмента не следует подбирать очень рисковые кредиты. Доля сегмента в общем обороте клиентов: 13%
5. Взрослые люди старше 34 лет, имеющие высшее/второе высшее/ученую степень или незаконченное высшее, тратящие больше 64 тыс. в месяц. Люди, которые имеют стабильные должности и тратящие ежемесячно крупные суммы денег. Можно давать рисковые кредиты. Доля сегмента в общем обороте клиентов: 2.7%

Преимущества дерева решений при задаче сегментации клиентов:

1. Решающие деревья легко интерпретировать, даже для технически неподготовленной аудитории. Удобное графическое представление результатов.
2. Решающее дерево позволяет генерировать правила в области, где специалисту самому тяжело формализовать свои знания.
3. Не требуется много параметров для модели. Многие параметры, отобранные аналитиком, могут не потребоваться.
4. В общем случае метод может работать даже на неподготовленных данных. Прочие методы требуют обязательной предварительной обработки данных.
5. Сегменты можно сопоставить с определенной бизнес-задачей

Недостатки дерева решений при задаче сегментации клиентов:

1. Возможно “переобучение” решающего дерева, поэтому необходимо устанавливать максимальную глубину дерева, минимальное число элементов в листьях дерева и максимально допустимое количество ветвей при разделении узла.
2. Результат сильно зависит от выбора целевой переменной и признаков.
3. Решающее дерево чувствительно к шумам во входных данных. Небольшие изменения в обучающей выборке могут отразиться на интерпретируемости модели и привести к ее глобальной корректировке.
4. Необходимо контролировать точность модели в связи с изменением во времени.

**Бизнес-правила.**

Бизнес-задача: сгруппировать клиентов исходя из цели получения кредита и образования по следующим параметрам: Age, DTI, Sum of payments, Period at work.

При выборе переменных также были учтены результаты, полученные в результате работы метода решающих деревьев. Применяем бизнес-правила и группируем клиентов: по цели получения кредита и по образованию.

По цели получения кредита:

* Крупные покупки: покупка недвижимости, покупка земли, покупка автомобиля
* Покупка бытовой техники, покупка мебели
* Ремонт
* Отпуск, лечение, образование

По образованию:

* Ученая степень, высшее образование, второе высшее, неполное высшее
* Среднее-специальное, среднее, неполное среднее

Получаем 8 предварительных сегментов в соответствии с бизнес-правилами.

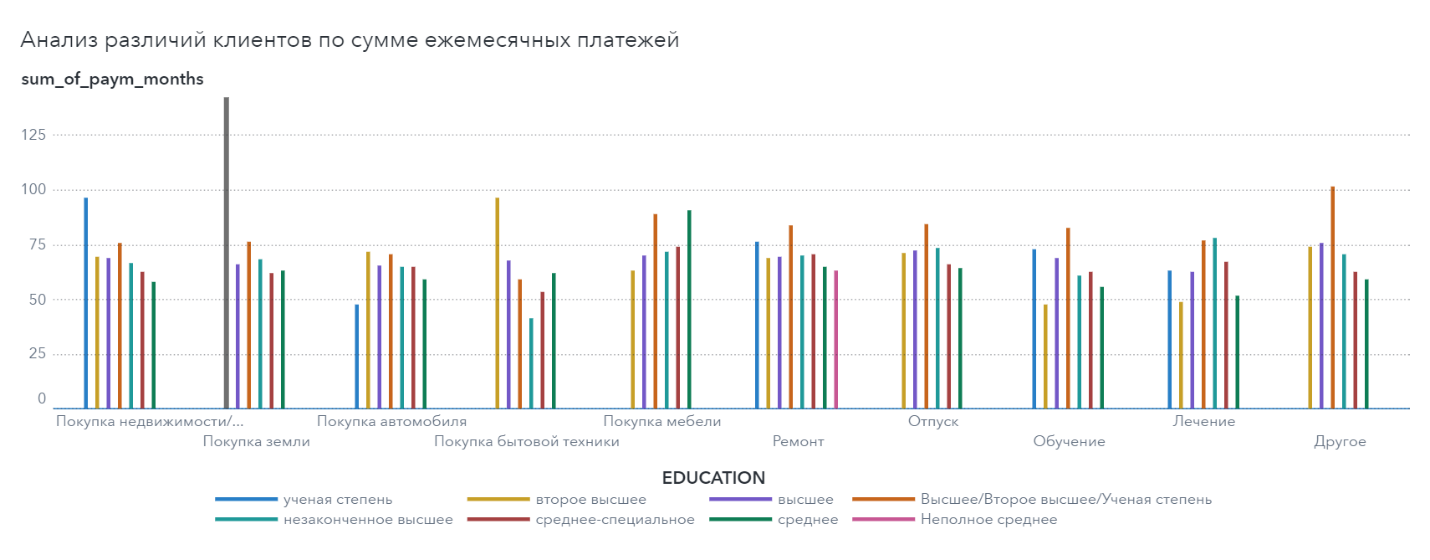
Анализ различий клиентов по показателям приведен ниже. Также проанализируем получившиеся результаты.

**Age**

****

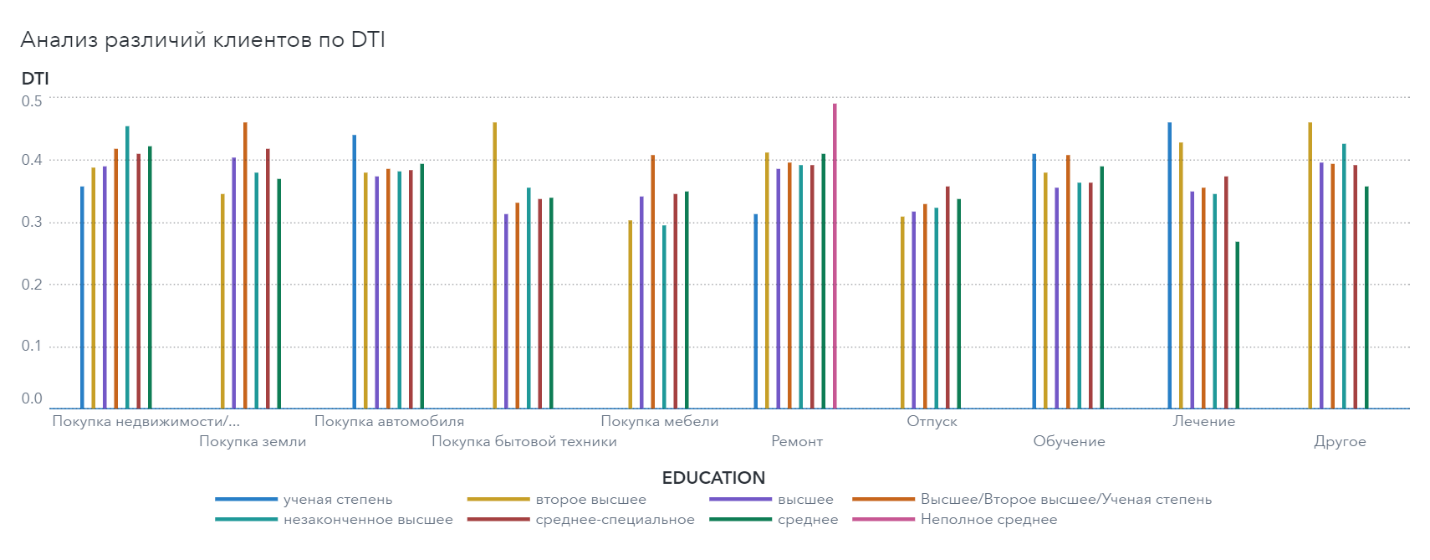
* В категориях лечение и обучения преобладают больше всего людей с ученой степенью или высшим образованием около 40 лет и старше
* Такие категории, как отпуск, ремонт и покупка автомобиля, одинаково распределены
* Покупку земли в молодом возрасте могут позволить в основном люди с высшим образованием или выше
* Покупку недвижимости в основном осуществляют люди с высшим образованием

**Sum of payments**

****

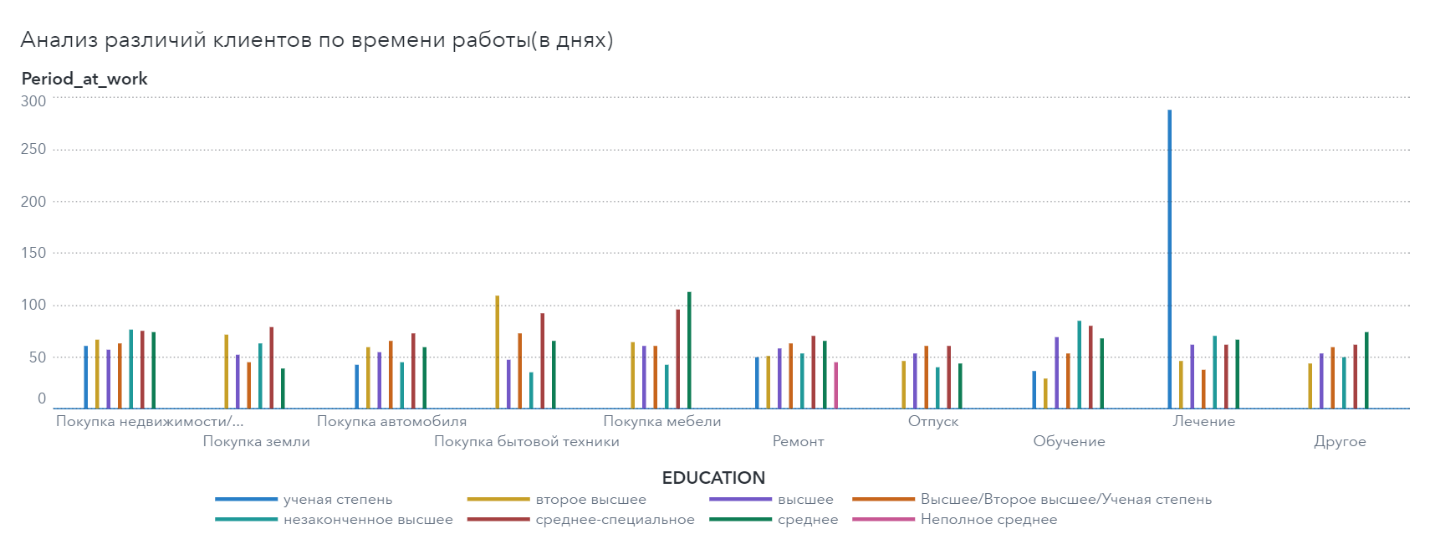
* Люди с высшим образованием и выше, с целью кредита покупки недвижимости/земли имеют самые высокие ежемесячные траты
* Люди со средним образованием имеют самые высокие ежемесячные траты с целью покупки мебели или ремонта

**DTI**

****

* У людей со средним образованием достаточно высокий DTI – цель получения кредита – ремонт
* У людей с высшим образованием высокий DTI при выборе лечения и покупки недвижимости

**Period at work**

****Таким образом, можно выделить следующие сегменты клиентов (наиболее удачные сегменты выделены):

* **Высшее образование – крупные покупки. И молодые, и взрослые люди. Высокие ежемесячные траты, выше 70 тыс., высокий DTI.**
* **Высшее образование – средние покупки. В основном взрослые люди с ежемесячными тратами до 90 тыс. (средние или высокие ежемесячные траты), средний DTI.**
* **Высшее образование – ремонт. Средние ежемесячные траты, около 60 тыс., в основном молодые люди, средний DTI.**
* **Высшее образование – персональные услуги. И молодые, и взрослые люди, средние ежемесячные траты, высокий или средний DTI.**
* **Среднее образование - крупные покупки. Преобладают взрослые люди, средние ежемесячные траты, высокий DTI.**
* Среднее образование - средние покупки. И молодые, и взрослые люди, со средними или высокими ежемесячными тратами, средний DTI.
* **Среднее образование – ремонт. В основном взрослые люди, со средними ежемесячными тратами, очень высокий DTI.**
* **Среднее образование - персональные услуги. И взрослые, и молодые. Небольшие ежемесячные траты, средний DTI (около 0.3)**

На основе полученных сегментов можно сформировать следующие профили клиентов:

* “Универсальный”. Цель получения кредита – покупка недвижимости/покупка земли/покупка бытовой техники/покупка мебели/ремонт. И молодые, и взрослые люди. Однако может быть рисковым, исходя из DTI, поэтому необходимо тщательно проверять данные клиента, если у индивида среднее образование, то это может послужить отказом в кредите.
* “Эконом”. Цель получения кредита – ремонт, персональные траты. Нацелен на индивидов со средними ежемесячными тратами. В основном взрослые люди.
* “Стандарт”. Цель получения кредита – средние покупки, ремонт. Нацелен на индивидов со средними ежемесячными тратами. В основном молодые люди. Есть высокая вероятность дефолта у индивидов со средним образованием.

Преимущества выбора бизнес-правил при задаче сегментации клиентов:

1. Самый простой метод сегментации
2. Метод легок в интерпретации, легко выполнить
3. Соответствует бизнес-задачам

Недостатки выбора бизнес-правил при задаче сегментации клиентов:

1. Может не отражать действительность, если бизнес-правила неверно определены
2. Тяжелее определить сегменты, чем при других методах
3. Между некоторыми сегментами может не быть значимого различия

**Выбор метода сегментации**

Было решено выбрать метод сегментации с учителем (дерево решений), так как является алгоритмическим способом кластеризации и поэтому предполагает меньшее участие человека, что обеспечивает большую точность.

**Рекомендованная кампания**

|  |  |  |  |
| --- | --- | --- | --- |
| **Название кредита** | **Цель получения кредита** | **Целевая аудитория** | **Описание** |
| **Кредит “Эконом”** | **Ремонт, покупка автомобиля, покупка бытовой техники, лечение** | **Взрослые люди от 34 лет, работающие, имеющие средний или небольшой доход** | **Самый простой кредит, предназначен для людей со скромными доходами** |
| **Кредит “Премиум”** | **Покупка недвижимости, покупка автомобиля, обучение** | **Взрослые люди от 34 лет, имеющие высшее образование и выше, обладающие высокими доходами** | **Направлен на взрослых людей с высокими доходами (руководители компаний, топ-менеджеры), есть дополнительные услуги** |
| **Кредит “Стандарт”** | **Покупка бытовой техники, мебели, отпуск** | **Молодые люди до 34 лет, работающие, имеющие высшее образование** | **Простота получения кредита, направленный на сегмент молодежи, которые уже имеют стабильную работу и собственную недвижимость** |
| **Кредит “Молодежный”** | **Ремонт, покупка недвижимости, образование, лечение** | **Молодые люди до 34 лет, недавно закончившие вуз, создающие семьи** | **Кредит направлен на людей, начинающих самостоятельную жизнь и обособленных материально от родителей** |