Модель машинного обучения для системы кредитного скоринга банка

Команда ‘Fantastic Four’
Иван, Роман, Юлия

1 - Предобработка данных

Иван, Юлия

Предобработка данных

train.csv, previous_loan

Уменьшение размерности матрицы признаков:
143 объясняющий и 1 целевой

Проверка числовых признаков на мультиколлинеарность:
59 объясняющих и 1 целевой

Проверка категориальных признаков:
52 объясняющих и 1 целевой

Обработка выбросов в числовых данных и аномалий в категориях

Обработка пропущенных значений:
38 объясняющих и 1 целевой

- Исходные данные – 261384 наблюдения, 143 зависимых признака, 1 - целевой

- Коэффициент корреляции Пирсона > 0.8 или < -0.8 между 10 группами признаков
- 84 признаков мультиколлинеарны

- Проверка на мультиколлинеарность категорий коэффициентом VIF >5 и анализатором Phi_K

- Выбросы значений по ящику с усами
- Аномальные значения в категориях

- <10% пропусков безопасно удаляем
- Числовые признаки заполняем медианой
- Категории максимально заполняем в соответствии с характером пропусков

Как заполняли пропуски в категориях в зависимости от происхождения пропуска

Признак

- **fondkapremon_mode (42%)**
['nan', 'not specified', 'reg oper account', 'org spec account', 'reg oper spec account']

- **mode_house_type_mode (49.5%)**
[nan, 'block of flats', 'terraced house', 'specific housing']

- **type_of_occupation (32.3%)**
['Sales staff', 'Managers', 'High skill tech staff', 'Laborers', nan, 'Core staff', 'Accountants', 'Low-skill Laborers', 'Medicine staff', 'Secretaries']

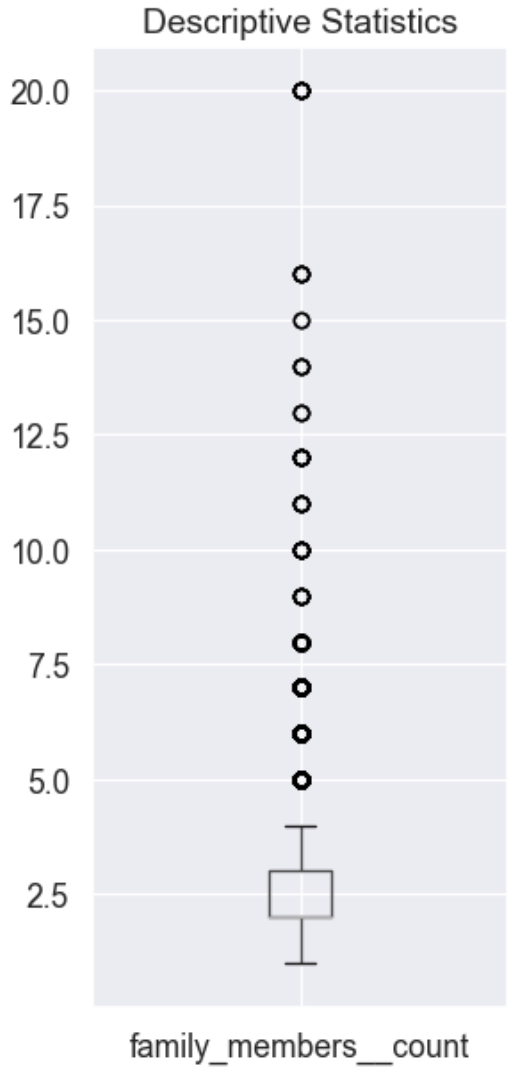
Характер происхождения пропуска

- Ручная ошибка
- Заполнили 42% пропусков как 'not specified'

- Ручная ошибка
- Заполнили 49.5% пропусков как 'not specified'

- Характер пропусков был не ясен, но для исследования было важно сохранить колонку, заполняли пропуски наиболее частой категорией

Как обработали выбросы в числовых признаках



Признаки с выбросами и порог, как обрезали

- ['income'] > 500000
- ['termination_date'] > 50000
- ['loan_body'] > 300000
- ['last_due_date'] > 50000]
- ['family_members__count'] > 6
- ['days_first_due'] > 180000
- ['down_payment_rate'] > 0.3
- ['days_first_drawing'] > 300000
- ['requests_bki_year'] > 10

Как обработали аномалии в категориях

Категория	Аномалия	Как обработали
• **'portfolio_name'** (21.1% 'XNA')	• ['POS', 'Cards', 'Cash', **'XNA'**, 'Cars']	• Заполнили наиболее часто встречающимся значением
• **'reject_reason_code'**	• ['XAP', 'HC', 'LIMIT', 'CLIENT', 'SCO', **'XNA'**, 'SCOFR', 'VERIF', 'SYSTEM']	• Сделали допущение, что эти категории – внутренняя разработка банка и аномалия только одна
• **'client_type_name'** (< 1% 'XNA')	• ['Repeater', 'New', 'Refreshed', **'XNA']**	• Удалили

Аномалии в категориях – это НЕ пропуски

Финальный датасет для анализа

38 признаков

- Уменьшили размер матрицы признаков с 121 до 38 входных параметров

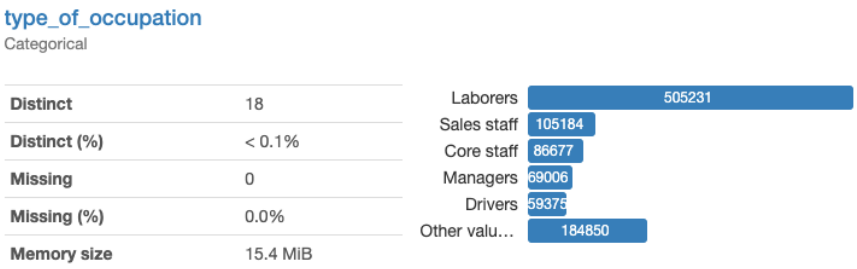
Дисбаланс таргета

- Оставили как есть ввиду большого объема данных

```
0    738372
1    64009
```

Profiling Report

- Мини-дашборды



Панельная структура данных

255614	0	Cash loans	F	0	139500.0	922500.0	27103.5	18.0	high	Cash Street: high
255614	0	Cash loans	F	0	139500.0	922500.0	27103.5	42.0	middle	Cash X-Sell: middle
255614	0	Cash loans	F	0	139500.0	922500.0	27103.5	48.0	middle	Cash X-Sell: middle
255614	0	Cash loans	F	0	139500.0	922500.0	27103.5	36.0	low_normal	Cash X-Sell: low

2 - Описательные статистики

Роман

Исследовательский анализ данных
Описательная статистика

Были выдвинуты гипотезы о большем влиянии выбранных категорий на построение моделей:

- Доход клиента – **income**
- Размер кредита - **loan_body**
- Типа занятости – **type_of_occupation**
- Возраст заёмщика - **Age**
- Уровень высшего образования - **education_type_name**

	income	loan_body	days_employed
count	261384.000000	261384.000000	261384.000000
mean	168923.794050	599186.557651	63867.624675
std	253791.189637	402330.879558	141317.267570
min	25650.000000	45000.000000	-17912.000000
25%	112500.000000	270000.000000	-2758.000000
50%	148500.000000	513531.000000	-1213.000000
75%	202500.000000	808650.000000	-289.000000
max	117000000.000000	4050000.000000	365243.000000

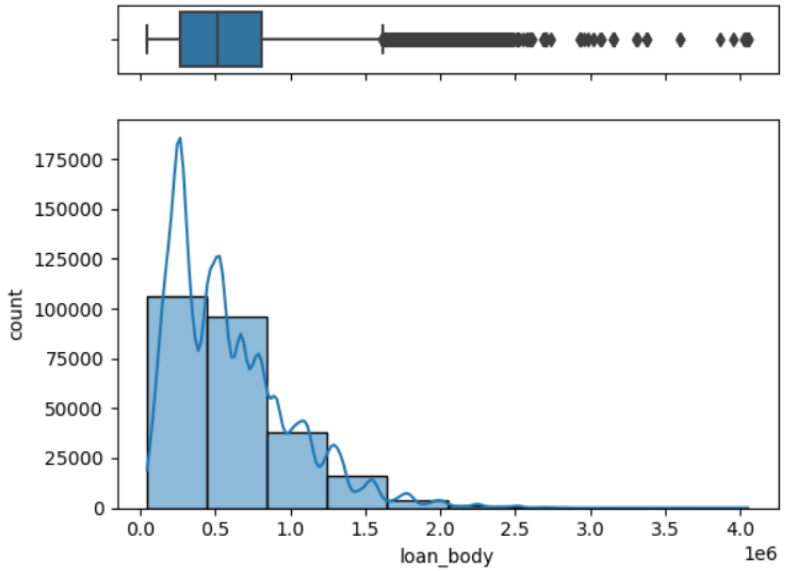
В представленных данных видны аномалии в колонках income, loan_body и days_employed. Например, максимальное значение признаков слишком выбивается из всех других значений. Что говорит скорее всего о выбросах в данных.

Исследовательский анализ данных

Описательная статистика

- Начнём с признаков размера кредита и дохода клиента

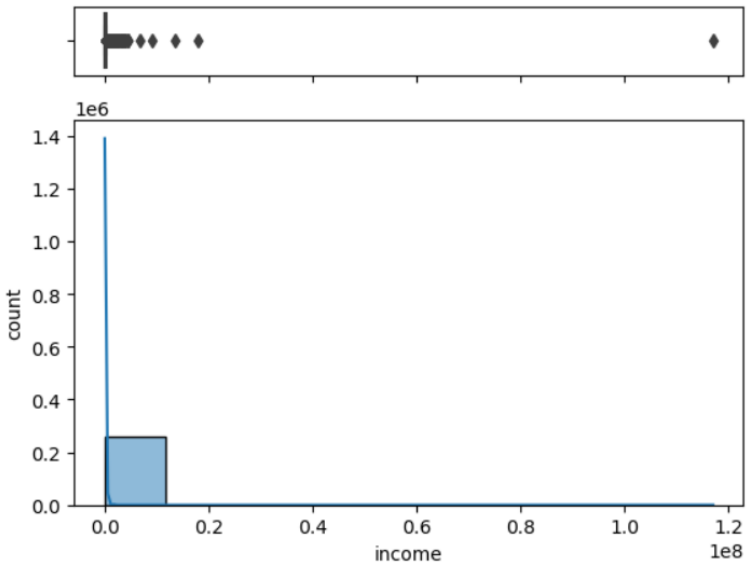
Loan_body



loan_body	
count	261384.000000
mean	599186.557651
std	402330.879558
min	45000.000000
25%	270000.000000
50%	513531.000000
75%	808650.000000
max	4050000.000000

По результатам видно, что чаще всего берут кредит на сумму 500 000. Однако выбросов довольно много, в дальнейшем их поправим.

Income



income	
count	261384.000000
mean	168923.794050
std	253791.189637
min	25650.000000
25%	112500.000000
50%	148500.000000
75%	202500.000000
max	117000000.000000

Хорошо виден выброс, который очень сильно может повлиять на данные. В дальнейшем уберём это значение.

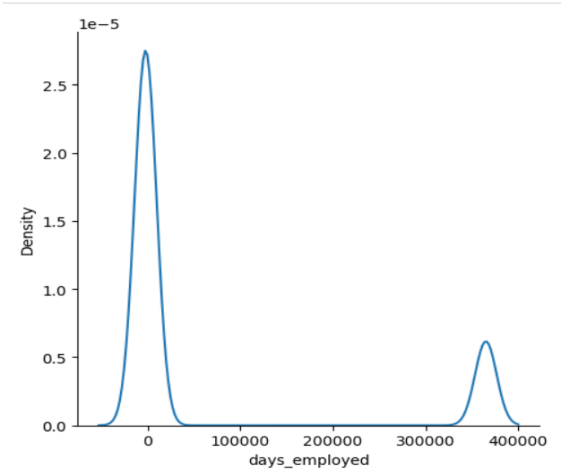
Исследовательский анализ данных

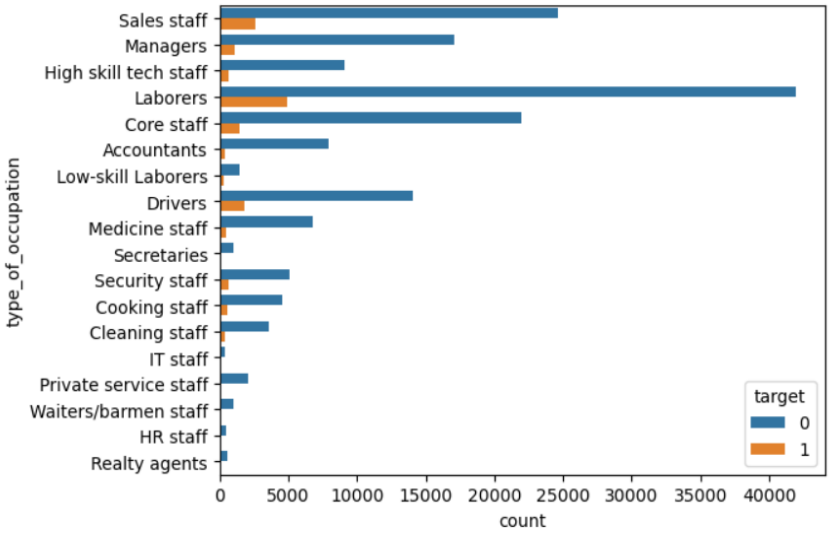
Описательная статистика

- Далее разберём признаки days_employed и `type_of_occupation`

days_employed

	days_employed
count	261384.000000
mean	63867.624675
std	141317.267570
min	-17912.000000
25%	-2758.000000
50%	-1213.000000
75%	-289.000000
max	365243.000000





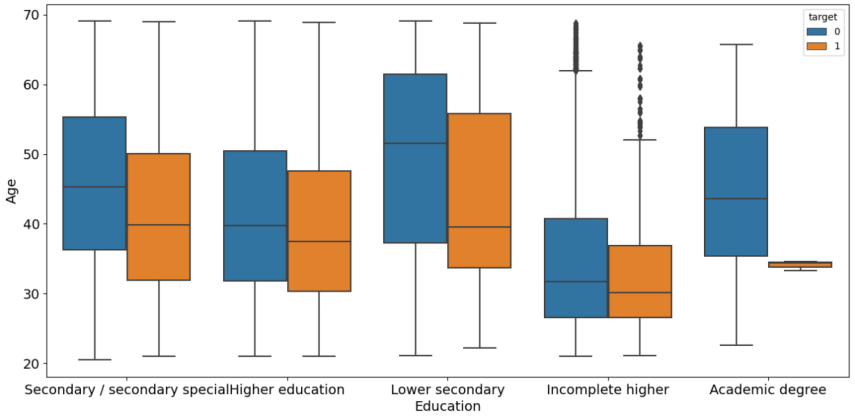
Обратим внимание, что среднее составляет 64000 положительное, но при этом 3-й квартиль отрицательный, т.е. по факту основная часть данных отрицательна. И среднее таким большим делает максимальное значение. Скорее всего положительный выброс возможно объяснить, что пропуски заполняли константой.

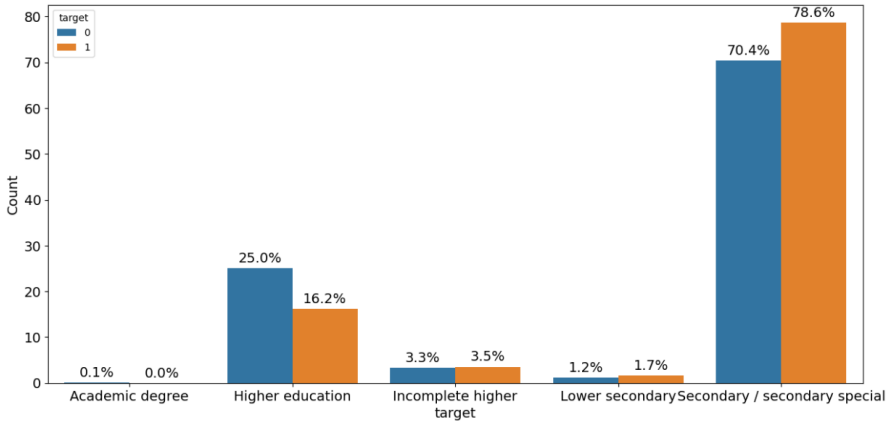
По этим данным можно сказать, что активнее всего берут кредиты рядовые специалисты, среди них же - больше всего невозвратов; в целом, по выборке число заемщиков, которые НЕ вернули кредит составляет не более 5-7%

Исследовательский анализ данных

Описательная статистика

- Рассмотрим связь возраста и образования на погашение кредита





По графику видно, что среднее значение возраста у клиентов без трудностей оплаты выше, значит после 40 лет люди скорее возвращают, чем не возвращают кредиты. Скорее всего возраст оказывает важное значение на целевой признак.

Чаще всего кредит берут люди со средним образованием, причём среди них больше проблем с возвратом кредита. Это будет влиять на модель больше, чем все остальные типы образования.

3 - Модели, финальные выводы

Юлия

Обучение моделей

Кодирование и масштабирование признаков

- После деления на train и test, провели **кодирование** и **масштабирование**, чтобы избежать утечки данных
- `OrdinalEncoder()`
- Масштабирование `MinMaxScaler()`

В идеале надо было бы кодирование обернуть в контейнер

Подбор гиперпараметров

- Поиск по сетке
- Параметров немного
- Деревья неглубокие
- Criterion: `'gini'`
- `n_estimators < 50`

Деревья неглубокие, чтобы не пришлось использовать регуляризацию и избежать переобучения

- **Дерево решений**
- **Случайный лес**
- **Градиентный бустинг**
- **CatBoostClassifier**

- **По AUC-ROC победил лес**
- (локально) обучили **CatBoostClassifier –** удобно, что не надо кодировать переменные!, но нужен GPU для обучения, иначе долго

- **Gini получился низкий – грустим =((**
- **Живые тестовые данные остались без предсказаний – тоже грустим =((**

Финальные выводы

***Вывод:* Больше всего на целевой признак - *вернет ли клиент кредит банку* - влияют следующие признаки:**

- **Возраст заемщика** `age` - средний возраст в выборке составляет 44.6 лет; описательные статистики показывают, после 40 лет люди скорее возвращают, чем не возвращают кредиты:

```
0    43.480 лет
1    39.115 лет
```

- **Размер кредита** `loan_body` - средний размер займа 603717 руб.; описательные статистики показывают, что меньшие по размеру кредиты (менее 500 тыс. руб) возвращаются хуже, чем займы свыше 500 тыс. руб.

```
0    517788.0 руб
1    497520.0 руб
```

- **Тип занятости** `type_of_occupation` - описательные статистики показывают, что активнее всего берут кредиты рядовые специалисты, среди них же - больше всего невозвратов; в целом, по выборке число заемщиков, которые НЕ вернули кредит составляет не более 5-7%;
- **Доход клиента** `income` - средний доход заемщика - 168-175 тыс. руб в год; также изучение описательных статистик показывает, что клиенты с меньшим доходом скорее НЕ возвращают кредиты:

```
0    148500.0 руб
1    135000.0 руб
```

Также на вероятность возврата кредита банку влияют следующие признаки:

- **Тип дохода клиента (businessman, working, maternity leave,...)** `income_type_name`
- **Ставка первоначального взноса, нормированная по предыдущему займу** `down_payment_rate`;
- **За сколько дней до оформления займа человек начал текущую работу** - `days_employed`;
- **Когда была первая и последняя выплаты по предыдущему займу относительно даты оформления текущего займа** `first_due_date`, `last_due_date`;
- **Когда было ожидаемое закрытие предыдущего займа относительно даты оформления текущего займа** `termination_date`;
- **Когда должна была быть первая выплата по предыдущему займу относительно даты оформления текущего займа** `days_first_due`

Влияние в меньшей степени на возврат кредита оказывают количество детей, тип образования, лояльность клиента к банку, согласие получать маркетинговые материалы по электронной почте.

Спасибо!

Команда ‘Fantastic Four’

Иван, Роман, Юлия