CUP_IT_2020 DATA SCIENCE

Команда Пироги

ЗАПОЛНЕНИЕ ПРОПУСКОВ

Несложно заметить, что некоторые признаки(столбцы) сильно разрежены. Хотелось бы их заполнить актуальными данными.

Выдвинем гипотезу, что пропуски в признаках, характеризующих баланс кредитных карт, баланс ипотеки и т.д. связан с отсутствием наличия таковых счетов. Следовательно при истинности выдвинутой гипотезы, правильнее было бы заполнить пропуски данных полей нулями, а не матожиданием или медианами. Список признаков попадающих под гипотезу:

счёт депозитов : кол-во депозитов ¶

количество персональных кредитов : баланс кредитов

количество ипотечных кредитов : баланс ипотеки

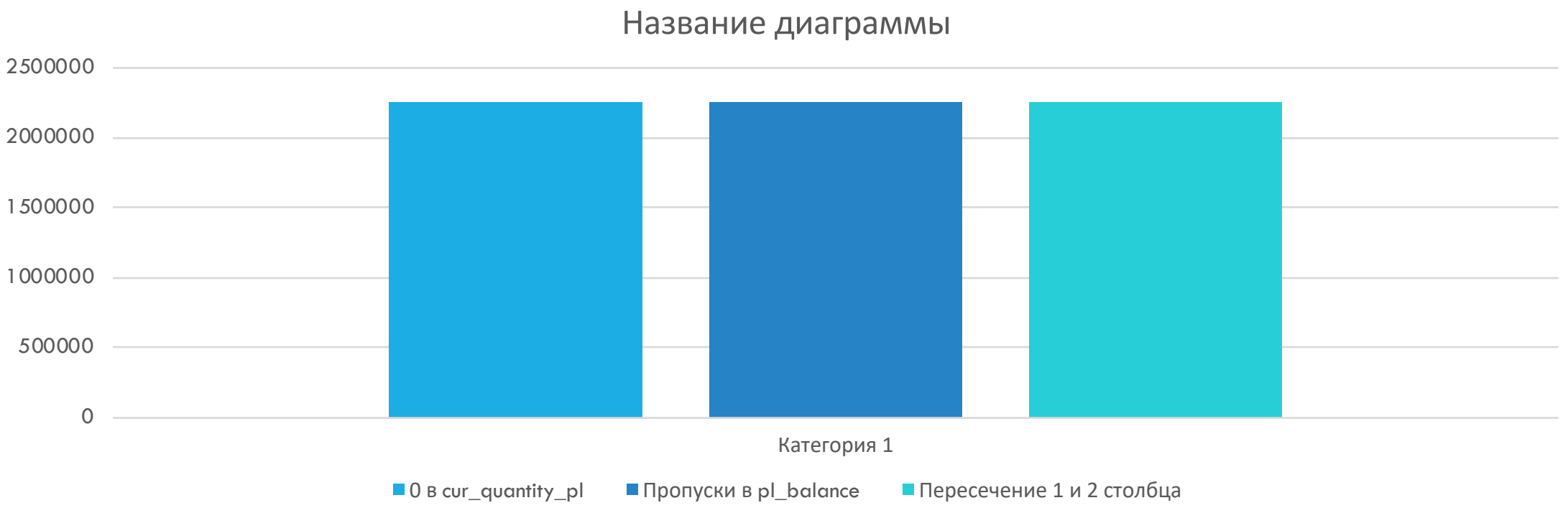
количество кредитных карт : баланс кредитных карт

количество счетов : баланс счетов

количество накопительных счетов : баланс накопительных счетов

количество инвестиционных продуктов : баланс инвестиций

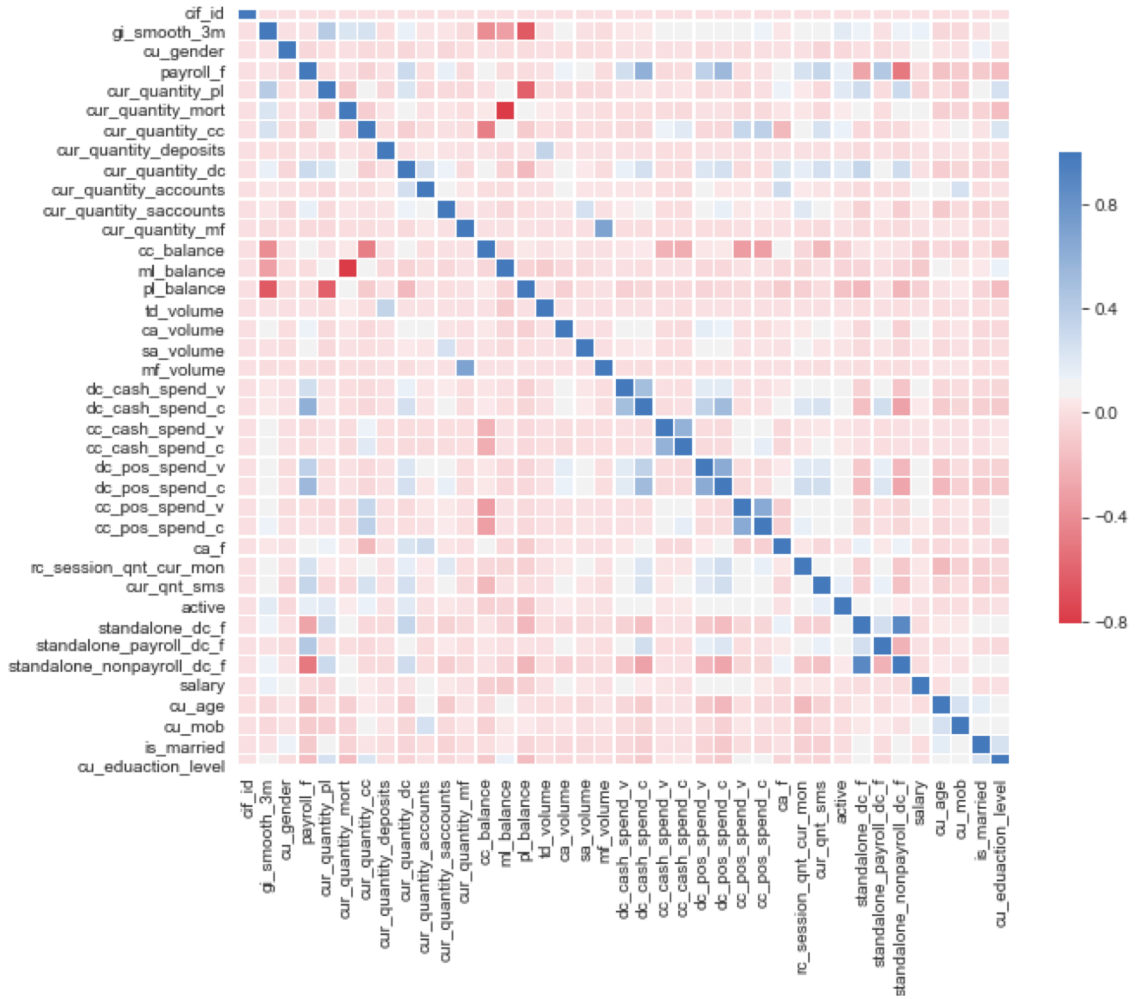
Легко заметить, что пропуски в столбце баланс кредитов указывают на значение 0 в столбце количество персональных кредитов и наоборот. То есть целесообразно заменить пропуски признака баланс кредитов на 0, так как при отсутствии кредитов их счёт не может быть другим. Аналогично поступим с оставшимися полями. Признаки (количество счетов : баланс счетов) не подчиняется гипотезе, так как нет клиентов банка без счетов. В данном случае заменяем пропуски медианами.



УСТРАНЕНИЕ КОРРЕЛИРУЮЩИХ ПРИЗНАКОВ

Список пар коррелирующих признаков:

1. pl_balance ~ cur_quantity_pl
2. ml_balance ~ cur_quantity_mort
3. cc_balance ~ cur_quantity_cc
4. mf_volume ~ cur_quantity_mf
5. standalone_dc_f ~standalone_payroll_dc_f
6. dc_cash_spend_v ~ dc_cash_spend_c
7. cc_cash_spend_v ~ cc_cash_spend_c
8. dc_pos_spend_v ~ dc_pos_spend_c
9. cc_pos_spend_v ~ cc_pos_spend_c



МОДЕЛЬ ПРЕДСТАВЛЕНИЯ ДАННЫХ

Попробуем спрогнозировать вероятность оттока клиента через 6 месяцев с помощью логистической регрессии. Прибыль от клиента будем рассчитывать как средний доход на протяжении известного периода жизни (от 1 до 6 месяцев) умноженный на вероятность того, что клиент останется в банке, умноженное на ширину прогнозируемого окна(в данном случае на 6).

В данном случае матрицей признаков будет информация за первые 6 месяцев 2018 года.
В качестве целевого параметра возьмём информацию о активности клиента за декабрь 2018 года, для прогнозирования оттока, а после для прогнозирования дохода за 6 месяцев возьмём доход за эти 6 месяцев.
Казалось бы, мы должны брать информацию об активность через 6 месяцев, а не за 12-й месяц. Например если человек в мае "ушёл из банка" и "вернулся" в августе того же года. В этом случае последний активный месяц в матрице признаков май (5-й), следовательно предсказывать мы должны отток к на ноябрь(11-й месяц), но мы предполагаем что вероятность возврата клиента слишком мала. Следовательно значение целевого признака, что для 11-го, что для 12-го месяца будет одно и то же.

ГИПЕРПАРАМЕТРЫ МОДЕЛИ

В данном случае настраивался коэффициент регуляризации.

После перебора по сетке оптимальное значение оказалось равно единице, с долей верных предсказаний равной 0.8427.

После по вышеописанной формуле были получены предсказания прибыли по каждому клиенту.

Данная модель не обладает сильной масштабируемостью и требует некоторой доработки для улучшения данного параметра.

СПАСИБО ЗА ВНИМАНИЕ