**ОТЧЕТ**

**о разработке модели оценки вероятности того, что клиент уйдет в просрочку на 30 дней в течении первых трех месяцев кредитования**

# Общие положения

Настоящий Отчет о разработке модели оценки вероятности того, что клиент уйдет в просрочку на 30 дней в течении первых трех месяцев кредитования (далее – Отчет) создан с целью подтверждения обоснованности и корректности разработанной модели. Для достижения этих целей Отчёт подробно описывает основные методологические подходы к моделированию, а также предпосылки и решения, которые принимались на каждом из этапов разработки модели.

Целью разработки модели является оценка вероятности того, клиент уйдет в просрочку на 30 дней в течении первых трех месяцев кредитования.

# Сбор данных.

Для построения модели выбраны заявки с 2022-11-01 до 2024-01-01.

Условия формирования выборки данных для построения модели:

В качестве сегмента субъектов, для которых будет применяться модель, выбраны заявки по следующим критериям:

Внешняя выдача: -7 +30 дней от даты заявки, только типы 6 и 1.

Целевая переменная: Флаг, обозначающий что у клиента есть просрочка на тридцать и более дней в течении первых трех месяцев на сумму большую или равную 2500р.

Общее число наблюдений в выборке – 531701, из них негативных - 7249. Доля негативных наблюдений ~ 0.014%.

Полученная выборка (Приложение к Отчету) была разбита на подвыборки:

- обучающую (~45% выборки) с 2022-11-01 по 2023-08-30, ---236711

- тестовую (~25% выборки) с 2022-11-01 по 2023-08-30 ---135616

- out-of-time (~30% выборки) c 2023-09-01 по 2023-12-30 ---159374

* Train/test 75/25 за период ноябрь 2022 –июнь 2023
* Train/test 50/50 за июль 2023
* Test 100% за август 2023

На основе тестовой выборки и out-of-time выборки производилось тестирование модели.

Описательная статистика выборки представлена в таблице «Сводное описание выборок».

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Тестовая выборка** | **Обучающая выборка** | **Out-of-time выборка** |
| Временной период для формирования выборки дефолтов | c 2022-11-01 по 2023-08-30 | | c 2023-09-01 по 2023-12-31 |
| Временной период для формирования выборки не-дефолтов |
| Общее количество клиентов в выборке | 135616 | 236711 | 159374 |
| Количество клиентов с просрочкой | 1639 | 2877 | 2733 |
| Количество клиентов без просрочки | 133977 | 233834 | 156641 |
| Соотношение в выборках | 0.0121% | | 0.0171% |

Перед обучением модели происходит WOE-преобразование значений переменных и биннинг посредством применения энкодера класса WOETransformer. В энкодере отдельно указываются категориальные переменные.

# Структура модели

В качестве статистического инструмента для построения модели (решение задачи бинарной классификации) используется логистическая регрессия.

## Теоретические основы логистической регрессии.

Результатом применения к модели к заявке является балл (или Score). Который определен как линейная комбинация преобразованных\* значений факторов данного модуля с использованием оптимальных, статистически рассчитанных весов:

где – вес *k*-го фактора, - преобразованное значение *k*-го фактора.

Модель вероятности дефолта представляет собой набор факторов с диапазонами значений этих факторов и, соответствующих этим диапазонам, баллов. Итоговый балл по сделке, характеризующий степень склонности потенциального Заемщика к дефолту, вычисляется как сумма баллов по всем факторам, входящим в модель. Чем больше величина балла, тем меньше вероятность дефолта.

где – зависимая переменная, = 1 – реализовавшееся событие дефолта, – набор независимых, объясняющих факторов, – веса объясняющих факторов, – свободный член, *PD* – вероятность дефолта (Probability of Default).

Расчет весов факторов осуществляется методом максимального правдоподобия, который максимизирует вероятность совместной реализации событий (дефолта / недефолта). Метод максимального правдоподобия заключается в максимизации функции правдоподобия. Математически оптимизационная задача для нахождения значений коэффициентов выглядит следующим образом:

где функция максимального правдоподобия (функция плотности вероятности), реализованное событие, логистическая функция, вектор коэффициентов логистической кривой.

Так как логарифм – монотонно возрастающая функция, то максимизация функции правдоподобия эквивалентна максимизации ее логарифма:

где функция максимального правдоподобия (функция плотности вероятности), реализованное событие, логистическая функция, вектор коэффициентов логистической кривой.

Результатом решения данной оптимизационной задачи является – вектор коэффициентов регрессии . Так как все факторы приведены к единой зависимости от вероятности дефолта, то знаки коэффициентов регрессии должны быть отрицательными. Зависимость баллов от значений факторов соответствует следующей логике - чем больше значение фактора, тем выше балл и ниже вероятность дефолта.

Таким образом, вес показывает степень влияния каждого фактора на вероятность дефолта и рассчитывается как относительная доля фактора в сумме всех факторов:

## WOE-преобразование факторов

По тренировочным данным факторы были разбиты на диапазоны по уровню рисков и затем преобразованы в Weight of Evidence (WOE). WOE для каждого значения / диапазона значений фактора определяется по формуле:

если и не равны нулю и

 в ином случае,

где и – количество недефолтных субъектов в диапазоне и по всей выборке, соответственно, а и – количество дефолтных субъектов в диапазоне и по всей выборке, соответственно.

## Однофакторный анализ

Проведение однофакторного анализа подразумевает оценку прогнозной силы каждого WOE-фактора из Длинного списка и анализ зависимостей между факторами. Результатом однофакторного анализа является решение о целесообразности включения каждого фактора в «короткий» список для построения модели, определение диапазонов значений факторов и ранжирование по уровню риска полученных диапазонов.

Однофакторный анализ состоял из следующих этапов:

- Оценка дискриминирующей силы факторов;

- Оценка корреляций

## Оценка дискриминирующей силы факторов

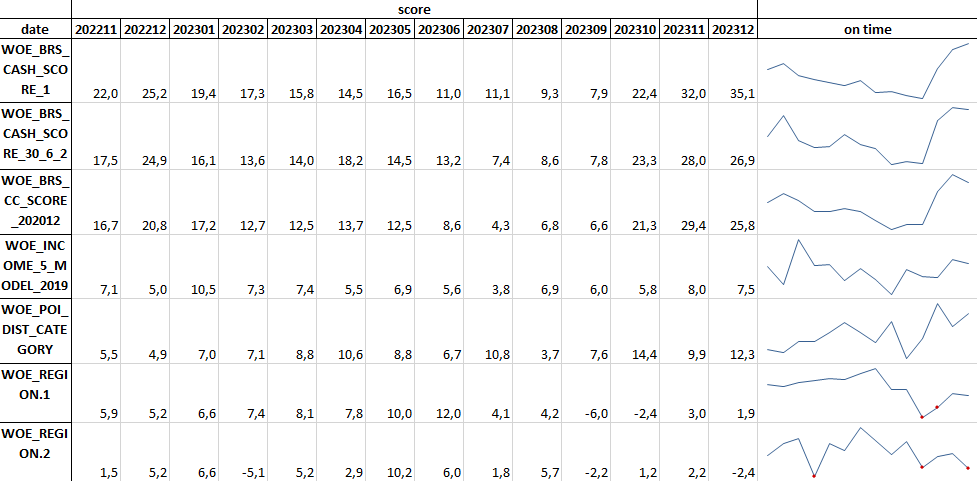
Для каждого фактора строится своя однофакторная логистическая модель,

по результатам предсказаний которой можно оценить дискриминирующую силу, а именно по коэффициенту Gini.

В результате оценки были исключены из дальнейшего рассмотрения факторы c коэффициентом Gini меньше 5.

|  |  |
| --- | --- |
| **feature** | **score** |
| WOE\_REGION.2 | 5,568109 |
| WOE\_REGION.1 | 8,986385 |
| WOE\_BRS\_CC\_SCORE\_202012 | 14,41672 |
| WOE\_BRS\_CASH\_SCORE\_30\_6\_2021\_02 | 16,33147 |
| WOE\_POI\_DIST\_CATEGORY | 7,08848 |
| WOE\_INCOME\_5\_MODEL\_201912\_SCORE | 6,830165 |
| WOE\_BRS\_CASH\_SCORE\_1 | 18,3995 |

При выборе факторов учитывалась также стабильность предсказательной способности факторов во времени. Для этого оценка Gini рассчитывалась на данных по каждому месяцу.



В случае значительной нестабильности и наличия большого количества отрицательных значений Gini фактор не включался в многофакторную модель.

## Оценка попарных корреляций

Целью оценки корреляций является не допустить вхождения в итоговую модель факторов с высокой корреляцией, поскольку использование коррелированных факторов в регрессии повышает стандартные отклонения оценок весов в многофакторном анализе, что снижает устойчивость и надёжность моделей.

Для анализа взаимозависимости была рассчитана корреляционная матрица для WOE-факторов. Верхняя граница допустимого значения линейного коэффициента корреляции между факторами – 0.65.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **WOE\_BRS\_CASH\_SCORE\_1** | **WOE\_REGION.2** | **WOE\_BRS\_CASH\_SCORE\_30\_6\_2021\_02** | **WOE\_INCOME\_5\_MODEL\_201912\_SCORE** | **WOE\_REGION.1** | **WOE\_BRS\_CC\_SCORE\_202012** | **WOE\_POI\_DIST\_CATEGORY** |
| **WOE\_BRS\_CASH\_SCORE\_1** | 1,00 | 0,04 | 0,49 | 0,21 | 0,02 | 0,49 | 0,25 |
| **WOE\_REGION.2** | 0,04 | 1,00 | 0,01 | 0,02 | 0,53 | 0,04 | 0,00 |
| **WOE\_BRS\_CASH\_SCORE\_30\_6\_2021\_02** | 0,49 | 0,01 | 1,00 | 0,08 | 0,01 | 0,47 | 0,22 |
| **WOE\_INCOME\_5\_MODEL\_201912\_SCORE** | 0,21 | 0,02 | 0,08 | 1,00 | 0,02 | 0,10 | 0,03 |
| **WOE\_REGION.1** | 0,02 | 0,53 | 0,01 | 0,02 | 1,00 | 0,01 | 0,01 |
| **WOE\_BRS\_CC\_SCORE\_202012** | 0,49 | 0,04 | 0,47 | 0,10 | 0,01 | 1,00 | 0,18 |
| **WOE\_POI\_DIST\_CATEGORY** | 0,25 | 0,00 | 0,22 | 0,03 | 0,01 | 0,18 | 1,00 |

Результатом однофакторного анализа является Короткий список факторов. Для принятия решения по отбору факторов необходимо учитывались не только предсказательная сила факторов, но и мнение и опыт экспертов. Отбор факторов в большей степени нацелен на то, чтобы исключить факторы, которые точно не могут быть использованы в модели ввиду нелогичности результатов или очень низкой предсказательной силы.

# Многофакторный анализ

## Структура и веса итоговой модели

Многофакторный анализ являлся следующим шагом после проведения однофакторного анализа. Объектом анализа являлись преобразованные факторы из короткого списка. Целью многофакторного анализа является выбор комбинации факторов и подбор соответствующих им весов, обеспечивающих наибольшую прогнозную силу и стабильность модели. Результатом анализа является линейная комбинация факторов с оптимальными весами.

Многофакторный анализ реализовывался через применение многомерной логистической регрессии, которая позволила определить веса объясняющих факторов, обеспечивающую наибольшую предсказательную силу модели. Формула регрессии имеет вид:

Оценка коэффициентов логистической регрессии и статистическая значимость факторов, вошедших в итоговую модель.

|  |  |  |  |
| --- | --- | --- | --- |
| **feature\_name** | **chi2** | **p\_value** | **coef** |
| WOE\_BRS\_CASH\_SCORE\_1 | 524.759514 | 3.800238e-109 | -0.601834 |
| WOE\_REGION.2 | 57.346005 | 2.167951e-12 | -0.193983 |
| WOE\_BRS\_CASH\_SCORE\_30\_6\_2021\_02 | 371.895162 | 3.064883e-77 | -0.426241 |
| WOE\_INCOME\_5\_MODEL\_201912\_SCORE | 68.017689 | 5.947360e-14 | -0.598110 |
| WOE\_REGION.1 | 95.709864 | 1.647659e-21 | -0.610911 |
| WOE\_BRS\_CC\_SCORE\_202012 | 302.156102 | 3.444346e-63 | -0.341945 |
| WOE\_POI\_DIST\_CATEGORY | 89.424748 | 8.874718e-18 | -0.275247 |

## Эффективность работы модели

Значение коэффициента Джини полученной модели на выборках составляет:

|  |  |  |  |
| --- | --- | --- | --- |
| Gini | | | |
| train | test | oot | score |
| 0.24 | 0.17 | 0.26 | 0.24 |

# Тестирование модели

Тестирование заключалось в сравнении прогнозной силы на тестовой выборке с прогнозной силой на выборке для разработки модели.

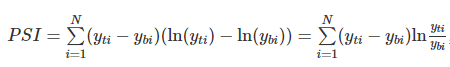
Gini ***- Оценка качества модели***

|  |  |  |  |
| --- | --- | --- | --- |
| **date** | **score\_train** | **score\_test** | **score\_oot** |
| 202211 | 25,1 | 19,9 | NaN |
| 202212 | 29,3 | 26,6 | NaN |
| 202301 | 25,1 | 25,5 | NaN |
| 202302 | 22,5 | 15,7 | NaN |
| 202303 | 21,1 | 19,6 | NaN |
| 202304 | 21,6 | 15,9 | NaN |
| 202305 | 26,6 | 18,2 | NaN |
| 202306 | 22,5 | 11,8 | NaN |
| 202307 | 14,6 | 11,4 | NaN |
| 202309 | NaN | NaN | 8,3 |
| 202310 | NaN | NaN | 24,4 |
| 202311 | NaN | NaN | 32,4 |
| 202312 | NaN | NaN | 30,9 |

## PSI

Индекс стабильности популяции (PSI) Эта метрика позволяет измерить то, насколько переменная изменилась в распределении между двумя выборками с течением времени. Она широко используется для мониторинга изменений характеристик популяции и диагностики возможных проблем с эффективностью модели.

Индекс рассчитывается следующим образом:



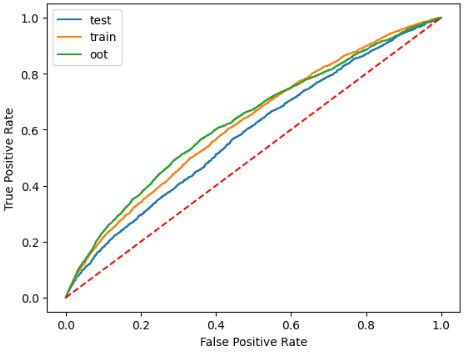
где — значение показателя целевого периода, попавшая в i-й интервал, — значение показателя базового периода.

В базовый период входят наблюдения за весь период train, а в целевой период – наблюдения за период oot.

Индекс стабильности популяции интерпретируется следующим образом:

* PSI менее 10% показывает отсутствие значимого изменения в текущей выборке;
* PSI в диапазоне от 10 до 25% свидетельствует о незначительном изменении, которые необходимо исследовать;
* PSI более 25% говорит о значительном смещении популяции и требуется перестроение модели.

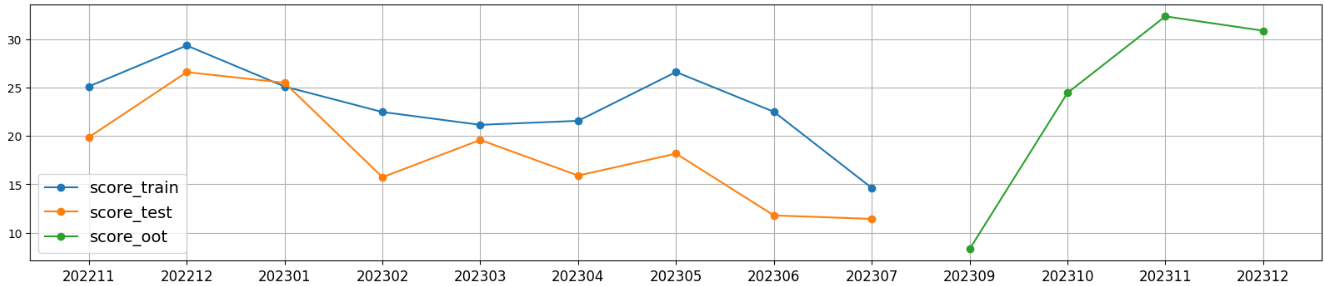
## Кривая roc auc



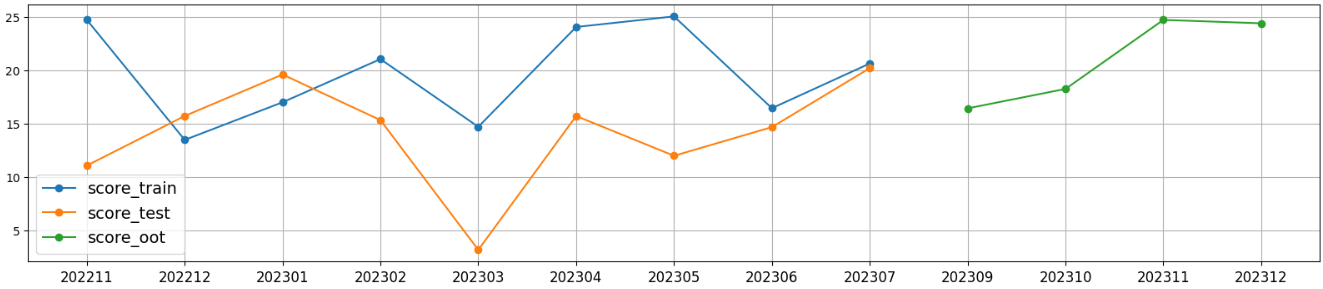
## Коэффициент Джини

Коэффициент Джини на тестовой и обучающих выборках во времени отображен на рисункaх

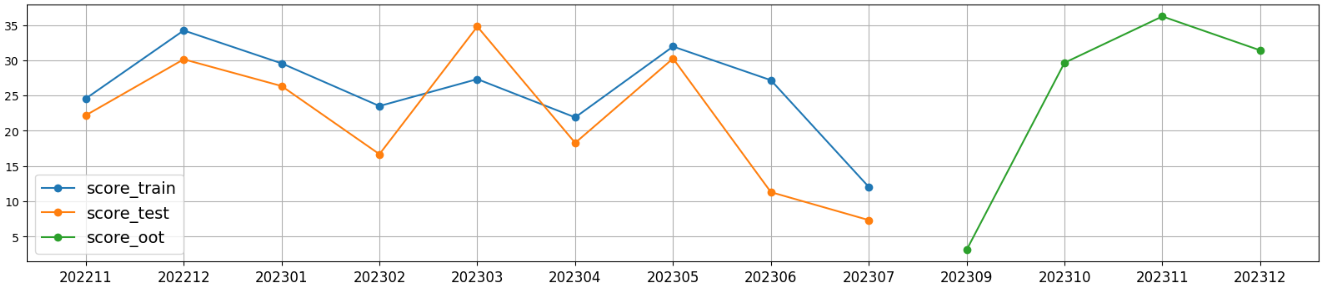
Весь пул заявок:



Внутренняя выдача:



Внешняя выдача:



# Итоговая модель

## PSI переменных, вошедших в модель:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **202212** | **202301** | **202302** | **202303** | **202304** |
| **WOE\_BRS\_CASH\_SCORE\_1** | 0,15544 | 0,2259 | 0,673949 | 2,408067 | 0,919757 |
| **WOE\_REGION.2** | 0,022673 | 0,164902 | 0,020767 | 0,03414 | 0,05532 |
| **WOE\_BRS\_CASH\_SCORE\_30\_6\_2021\_02** | 0,125137 | 0,164981 | 0,218453 | 0,122369 | 0,150776 |
| **WOE\_INCOME\_5\_MODEL\_201912\_SCORE** | 0,288497 | 7,358702 | 7,057773 | 6,725489 | 6,322891 |
| **WOE\_REGION.1** | 0,05873 | 0,276332 | 0,060554 | 0,155883 | 0,216988 |
| **WOE\_BRS\_CC\_SCORE\_202012** | 0,031944 | 0,089635 | 0,464193 | 0,259538 | 1,267477 |
| **WOE\_POI\_DIST\_CATEGORY** | 0,057193 | 0,030739 | 0,401404 | 0,012544 | 0,278406 |
|  | **202305** | **202306** | **202307** | **202308** |  |
| **WOE\_BRS\_CASH\_SCORE\_1** | 0,786423 | 1,778599 | 1,435785 | 0,854662 |  |
| **WOE\_REGION.2** | 0,08768 | 0,104037 | 0,003241 | 0,383921 |  |
| **WOE\_BRS\_CASH\_SCORE\_30\_6\_2021\_02** | 0,227228 | 0,602773 | 0,620214 | 0,88009 |  |
| **WOE\_INCOME\_5\_MODEL\_201912\_SCORE** | 1,58567 | 3,525551 | 6,733119 | 8,061495 |  |
| **WOE\_REGION.1** | 0,158757 | 0,437892 | 0,747424 | 1,036379 |  |
| **WOE\_BRS\_CC\_SCORE\_202012** | 1,784704 | 2,232192 | 6,201982 | 9,683495 |  |
| **WOE\_POI\_DIST\_CATEGORY** | 1,001439 | 0,881398 | 1,188974 | 1,021078 |  |
|  | **202309** | **202310** | **202311** | **202312** |  |
| **WOE\_BRS\_CASH\_SCORE\_1** | 1,092771 | 0,885012 | 1,208243 | 1,586379 |  |
| **WOE\_REGION.2** | 0,160721 | 0,204973 | 0,086507 | 0,057987 |  |
| **WOE\_BRS\_CASH\_SCORE\_30\_6\_2021\_02** | 2,074849 | 3,260446 | 1,197562 | 0,810108 |  |
| **WOE\_INCOME\_5\_MODEL\_201912\_SCORE** | 5,89677 | 4,72437 | 2,548986 | 3,188344 |  |
| **WOE\_REGION.1** | 0,845401 | 1,101854 | 0,732366 | 0,654749 |  |
| **WOE\_BRS\_CC\_SCORE\_202012** | 6,676048 | 6,957964 | 3,750023 | 3,610975 |  |
| **WOE\_POI\_DIST\_CATEGORY** | 1,333107 | 1,986492 | 3,053384 | 5,563093 |  |

## Итоговая модель

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **feature** | **bin** | **cnt\_bad** | **cnt\_good** | **cnt\_all** | **pcnt\_of\_all** | **bad\_rate** |
| WOE\_BRS\_CASH\_SCORE\_1 | -0,623 | 1869 | 67516 | 69385 | 0,130 | 0,027 |
| WOE\_BRS\_CASH\_SCORE\_1 | -0,114 | 1008 | 64487 | 65495 | 0,123 | 0,015 |
| WOE\_BRS\_CASH\_SCORE\_1 | -0,067 | 928 | 63769 | 64697 | 0,122 | 0,014 |
| WOE\_BRS\_CASH\_SCORE\_1 | -0,051 | 869 | 64163 | 65032 | 0,122 | 0,013 |
| WOE\_BRS\_CASH\_SCORE\_1 | 0,200 | 710 | 64343 | 65053 | 0,122 | 0,011 |
| WOE\_BRS\_CASH\_SCORE\_1 | 0,301 | 651 | 64744 | 65395 | 0,123 | 0,010 |
| WOE\_BRS\_CASH\_SCORE\_1 | 0,320 | 616 | 66195 | 66811 | 0,126 | 0,009 |
| WOE\_BRS\_CASH\_SCORE\_1 | 0,458 | 598 | 69235 | 69833 | 0,131 | 0,009 |
| WOE\_REGION.2 | -0,405 | 4076 | 288239 | 292315 | 0,550 | 0,014 |
| WOE\_REGION.2 | -0,127 | 1125 | 77701 | 78826 | 0,148 | 0,014 |
| WOE\_REGION.2 | 0,018 | 1514 | 109497 | 111011 | 0,209 | 0,014 |
| WOE\_REGION.2 | 0,510 | 534 | 49015 | 49549 | 0,093 | 0,011 |
| WOE\_BRS\_CASH\_SCORE\_30\_6\_2021\_02 | -0,535 | 1654 | 67207 | 68861 | 0,130 | 0,024 |
| WOE\_BRS\_CASH\_SCORE\_30\_6\_2021\_02 | -0,217 | 1199 | 67514 | 68713 | 0,129 | 0,017 |
| WOE\_BRS\_CASH\_SCORE\_30\_6\_2021\_02 | -0,004 | 936 | 66354 | 67290 | 0,127 | 0,014 |
| WOE\_BRS\_CASH\_SCORE\_30\_6\_2021\_02 | 0,024 | 827 | 66392 | 67219 | 0,126 | 0,012 |
| WOE\_BRS\_CASH\_SCORE\_30\_6\_2021\_02 | 0,167 | 1448 | 130369 | 131817 | 0,248 | 0,011 |
| WOE\_BRS\_CASH\_SCORE\_30\_6\_2021\_02 | 0,309 | 626 | 63245 | 63871 | 0,120 | 0,010 |
| WOE\_BRS\_CASH\_SCORE\_30\_6\_2021\_02 | 0,416 | 559 | 63371 | 63930 | 0,120 | 0,009 |
| WOE\_INCOME\_5\_MODEL\_201912\_SCORE | -0,138 | 2396 | 151083 | 153479 | 0,289 | 0,016 |
| WOE\_INCOME\_5\_MODEL\_201912\_SCORE | -0,021 | 1348 | 96684 | 98032 | 0,184 | 0,014 |
| WOE\_INCOME\_5\_MODEL\_201912\_SCORE | -0,009 | 1165 | 84150 | 85315 | 0,160 | 0,014 |
| WOE\_INCOME\_5\_MODEL\_201912\_SCORE | 0,047 | 1435 | 110632 | 112067 | 0,211 | 0,013 |
| WOE\_INCOME\_5\_MODEL\_201912\_SCORE | 0,283 | 905 | 81903 | 82808 | 0,156 | 0,011 |
| WOE\_REGION.1 | -0,374 | 3678 | 248834 | 252512 | 0,475 | 0,015 |
| WOE\_REGION.1 | -0,005 | 2859 | 209211 | 212070 | 0,399 | 0,013 |
| WOE\_REGION.1 | 0,643 | 712 | 66407 | 67119 | 0,126 | 0,011 |
| WOE\_BRS\_CC\_SCORE\_202012 | -0,535 | 1715 | 69966 | 71681 | 0,135 | 0,024 |
| WOE\_BRS\_CC\_SCORE\_202012 | -0,158 | 1168 | 71008 | 72176 | 0,136 | 0,016 |
| WOE\_BRS\_CC\_SCORE\_202012 | -0,018 | 947 | 68860 | 69807 | 0,131 | 0,014 |
| WOE\_BRS\_CC\_SCORE\_202012 | 0,066 | 823 | 66942 | 67765 | 0,127 | 0,012 |
| WOE\_BRS\_CC\_SCORE\_202012 | 0,210 | 2026 | 188450 | 190476 | 0,358 | 0,011 |
| WOE\_BRS\_CC\_SCORE\_202012 | 0,290 | 570 | 59226 | 59796 | 0,112 | 0,010 |
| WOE\_POI\_DIST\_CATEGORY | -0,288 | 1441 | 73460 | 74901 | 0,141 | 0,019 |
| WOE\_POI\_DIST\_CATEGORY | -0,149 | 992 | 61643 | 62635 | 0,118 | 0,016 |
| WOE\_POI\_DIST\_CATEGORY | 0,020 | 552 | 40664 | 41216 | 0,078 | 0,013 |
| WOE\_POI\_DIST\_CATEGORY | 0,067 | 1466 | 118455 | 119921 | 0,226 | 0,012 |
| WOE\_POI\_DIST\_CATEGORY | 0,084 | 685 | 54909 | 55594 | 0,105 | 0,012 |
| WOE\_POI\_DIST\_CATEGORY | 0,109 | 2113 | 175321 | 177434 | 0,334 | 0,012 |