|  |  |
| --- | --- |
|  | Согласовано  Заместитель Председателя Правления,  Глава риск-менеджмента  Машчык Р.А.  «\_\_\_» \_\_\_\_\_\_\_\_\_\_ 2025 г. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

**ОТЧЕТ О ПОСТРОЕНИИ PD (PROBABILITY OF DEFAULT)/СКОРИНГОВЫХ МОДЕЛЕЙ   
ПО ПРОДУКТУ «ПЕНСИОННАЯ»  
В АО «НАРОДНЫЙ БАНК КАЗАХСТАНА»**

**Алматы 2025**

**СОДЕРЖАНИЕ**

[1 Общие положения 3](#_Toc208829921)

[2 История модели 3](#_Toc208829922)

[3 Общая методика построения модели 4](#_Toc208829923)

[4 Предобработка данных 5](#_Toc208829924)

[5 Модель логистической регрессии 9](#_Toc208829925)

[6 Модель логистической регрессии с ансамблем моделей 15](#_Toc208829926)

[7 Оценка качества модели логистической регрессии с ансамблем моделей 16](#_Toc208829927)

[8 Реализация модели и операции 17](#_Toc208829928)

[9 Заключение 18](#_Toc208829929)

[Приложение 1 19](#_Toc208829930)

[Приложение 2 20](#_Toc208829931)

[Приложение 3 22](#_Toc208829932)

[Приложение 4 23](#_Toc208829933)

## 1 Общие положения

**Цель построения модели**

Построение модели вероятности дефолта (PD) по выданным займам (контрактам) с января 2022 по декабрь 2023 года по продукту Пенсионная с зависимой переменной – NPL91\_EVER. Выборка включила в себя 203 279 контрактов, из них 6 083 контрактов, когда-либо выходивших на просрочку более 90 дней за все время кредитования/получивших реструктуризацию за время кредитования/ умерших клиентов (см. таблица 1).

Таблица 1. Распределение целевой переменной NPL91\_EVER

|  |  |
| --- | --- |
| **Продукт** | **Пенсионная от 01.09.2025** |
| Хорошие | 197 196 |
| Плохие | 6 083 |
| Доля плохих | 3.0 % |

## 2 История модели

**Модель Пенсионная от 23.12.2022**

23.12.2022 была разработана скоринговая модель по продукту Пенсионная по выданным займам с 01 ноября 2016 по 30 июня 2021 года. Распределение целевой переменной NPL91\_EVER представлена в таблице 2.

Таблица 2. Распределение целевой переменной NPL91\_EVER

|  |  |
| --- | --- |
| **Продукт** | **Пенсионная от 23.12.2022** |
| Хорошие | 416 627 |
| Плохие | 14 949 |
| Доля плохих | 3.59 % |

В модель логистической регрессии вошли 11 параметров:

2 транзакционных параметра по карточкам из ХБ,

1 биографический параметр – возраст,

7 параметров из ПКБ,

1 параметр из онлайн заявки – сумма займа/среднемесячный доход.

В модель XGBoost вошли 12 параметров:

6 параметров из ПКБ,

6 параметров транзакций внутри ХБ.

Качество разработанной модели показало приемлемый уровень – Gini = 0.4 для логистической регрессии и Gini = 0.6 для модели логистической регрессии с параметром от бустинга (см. табл. 3).

Таблица 3. Качество модели Пенсионная от 23.12.2022

|  |  |  |
| --- | --- | --- |
| **Показатель** | **Логистическая регрессия** | **Логистическая регрессия с XGBoost** |
| AUC | 0.70 | 0.80 |
| GINI | 0.40 | 0.60 |

В июне 2023 года Банком была внедрена выше описанная скоринговая модель. По результатам проведенной валидации от 25.12.2024 г. было установлено снижение качества данной модели, в связи с чем было рекомендовано о необходимости пересмотра действующей модели по продукту Пенсионная.

**Модель Пенсионная от 02.07.2025**

02.07.2025 была разработана скоринговая модель по продукту Пенсионная по выданным займам с 01.01.2022 года по 31.12.2023 года. Распределение целевой переменной NPL91\_EVER представлена в таблице 4.

Таблица 4. Распределение целевой переменной NPL91\_EVER

|  |  |
| --- | --- |
| **Продукт** | **Пенсионная от 02.07.2025** |
| Хорошие | 205 629 |
| Плохие | 5 179 |
| Доля плохих | 2.46 % |

В модель логистической регрессии вошли 7 параметров:

1 параметр из ХБ – категория кредитной истории,

6 параметров из ПКБ.

В модель XGBoost вошли 50 параметров:

38 параметров из карточных операций внутри ХБ,

9 параметров из кредитных данных (активные кредиты, просроченный долг, ежемесячные платежи),

3 параметра из ПКБ (количество дней с даты последнего отчета с ПКБ, количество дней между первым контрактом и датой заявки, количество дней между последним контрактом и датой заявки).

Качество разработанной модели показало следующие результаты – Gini = 0.24 для логистической регрессии и Gini = 0.71 для XGBoost (см. табл. 3).

Таблица 5. Качество модели Пенсионная от 02.07.2025

|  |  |  |
| --- | --- | --- |
| **Показатель** | **Логистическая регрессия** | **XGBoost** |
| AUC | 0.62 | 0.86 |
| GINI | 0.24 | 0.71 |

В августе 2025 года была проведена валидация скоринговой модели по продукту Пенсионная от 02.07.2025, в результате которой на валидационных данных коэффициент Gini составил 0.04, далее было принято решение о необходимости доработки модели.

## Общая методика построения модели

Общая методика построения скоринговой модели состоит из следующих основных этапов:

1) сбор и подготовка данных с первичным анализом предоставленных данных, поиск ошибок и несоответствий в данных;

2) исправление ошибок («предобработка данных») и повторное предоставление данных (в п. 1 и 2 может быть несколько итераций);

3) повторная проверка после «предобработки» и углубленный статистический анализ данных, преобразование данных и поиск производных показателей от исходных для построения модели, исследование связей между целевой и независимыми переменными;

4) построение скоринговой модели на основе статистической модели, исследование свойств модели, поиск альтернативных моделей (с разными наборами переменных), выбор наилучшей из них;

5) тестирование и валидация выбранной итоговой модели.

Для решения задачи бинарной классификации был разработан ансамблевый подход с использованием следующего метода: логистическая регрессия с параметром от взвешенного усреднения результатов LightGBM, XGBoost и CatBoost моделей. Данная комбинация позволяет объединить интерпретируемость логистической регрессии и сложность нелинейных моделей с улучшением предсказательной способности модели.

Все этапы построения скоринговой модели, представленные выше были выполнены в Jupyter Notebook на языке программирования Python.

## 4 Предобработка данных

**Методика предобработки данных**

Предобработка данных включает в себя следующие процессы: удаление дубликатов; исправление ошибок в данных, обработка пропущенных значений; нормализация данных (обнаружение отклонений и аномалий); преобразование типов данных; рандомизация данных (стирание эффектов определенного порядка, в котором были подготовлены данные); изменение распределения зависимой переменной при необходимости; исследовательский анализ данных (Exploratory Data Analysis – выявление основных характеристик, проверка предположений); анализ мультиколлинеарности; визуализация данных (обнаружение соответствующих отношений между переменными или дисбалансами классов); категоризация количественных переменных; перекодировка текстовых переменных в категориальных шкалах в числовой тип данных.

Анализ зависимостей между целевой и независимыми переменными проводится с целью сокращения списка переменных для построения модели и отсеивания не влияющих переменных. Визуальное представление данных делают с помощью гистограмм. Категоризация количественных переменных позволяет облегчить обработку выбросов и экстремальных значений, упростить интерпретацию скоринговой карты, отразить сложные нелинейные связи. Количественная переменная разбивается на основе равных процентилей на несколько групп. Затем в каждой группе считается доля плохих и хороших кредитов, а также показатель веса категорий предиктора WOE. Веса категорий предиктора помогают найти по переменной «границы чувствительности» к появлению моделируемого события риска и провести оптимальным образом категоризацию количественных переменных.

Процесс предобработки данных начался с конвертирования переменных в числовой/стринговый формат данных при наличии ошибок формата переменных. Переменные *'FMGROUP', 'PKB\_RESIDENT\_FL', 'PKB\_HEAD\_REP\_TYPE', 'pkb\_resident\_fl', 'CategoryParametrSet', 'Branch TypeParametrSet', 'CBIsCurDelayLarger5kParametrSet', 'ClientCategoryParametrSet', 'HBIsRestrParametrSet', 'IsApplFormActualParametrSet', 'FinHasCredHistParametrSet', 'ASPFlagParametrSet', 'HBHasCreditHistParametrSet', 'IgnoreDiscountParametrSet', 'CtrlGrpParametrSet', 'IsCPPPaysInLastMonthParametrSet', 'IsCPPPaysInLast2MonthsParametrSet','Categorydssb\_ocds\_paramset','CtrlGrpdssb\_ocds\_paramset','IsRefinancedssb\_ocds\_paramset','ASPFlagdssb\_ocds\_paramset','FinHasCredHistdssb\_ocds\_paramset', 'ClientCategorydssb\_ocds\_paramset', 'VerifiedIncomedssb\_ocds\_paramset', 'IsCPPPaysInLast2Monthsdssb\_ocds\_paramset', 'verifiedincomedmparse2'* были конвертироаны в объектный формат данных. На рисунке 1 представлены функции конвертирования данных в стринговый и числовой форматы данных.

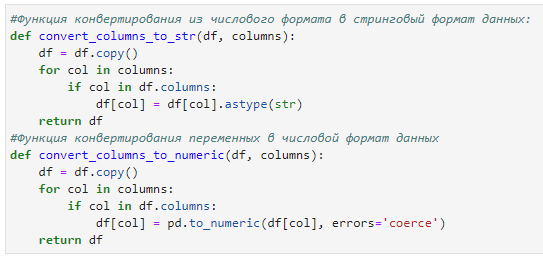


Рисунок 1. Функция конвертирования из числового формата в стринговый формат данных

Для заполнения пустых значений в стринговых переменных была реализована функция fillna\_none\_except\_list, представленная на рисунке 2.

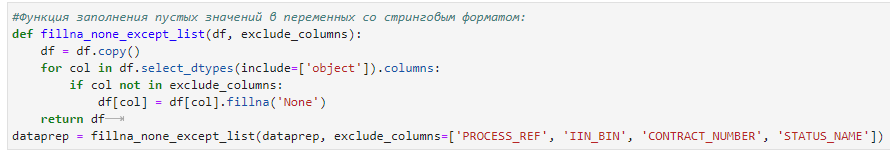
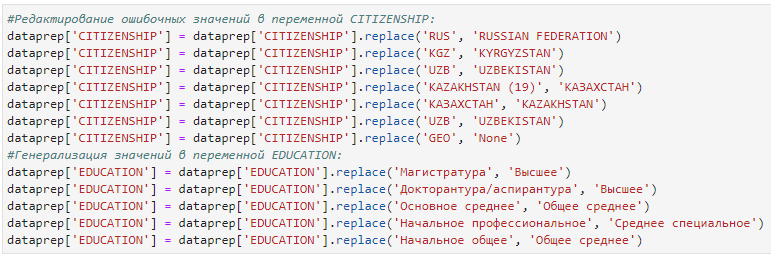
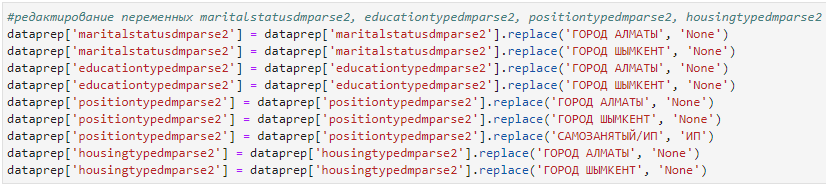


Рисунок 2. Функция заполнения пустых значений в переменных со стринговым форматом.

Переменные *'AmountInColvirsLineParametrSet','CBIsCurDelayLarger5kParametrSet', 'IsApplFormActualParametrSet','FinHasCredHistParametrSet','LowRODParametrSet','HBHasCreditHistParametrSet','IgnoreDiscountParametrSet','CtrlGrpParametrSet','GtsvpServiceResponseData\_Request\_EmpLogin','CBHistDelayMaxAmountdssb\_ocds\_paramset','CBoverdues\_90\_2Ydssb\_ocds\_paramset','VerifiedIncomedssb\_ocds\_paramset','verifiedincomedmparse2','cred\_dpd\_cnt\_avto','cred\_dpd\_cnt\_30d\_avto','cred\_dpd\_cnt\_90d\_avto','cred\_insur\_cnt\_avto','cred\_repaid\_early\_sum\_in360d\_avto','pkb\_bvu\_nominal\_rate\_pledge\_bzk\_0','pkb\_bvu\_nominal\_rate\_pledge\_cc\_0','pkb\_bvu\_nominal\_rate\_pledge\_lgot\_avto','pkb\_criminal\_status','pkb\_missing\_status','cred\_dpd\_cnt\_30d\_bnpl','cred\_dpd\_cnt\_90d\_bnpl','cred\_insur\_cnt\_bnpl','cred\_dpd\_cnt\_90d\_bzk','cred\_dpd\_sum\_ipoteka','cred\_dpd\_sum\_10d\_ipoteka','cred\_dpd\_cnt\_30d\_ipoteka','cred\_dpd\_cnt\_90d\_ipoteka','cred\_dpd\_annuity\_sum\_ipoteka','cred\_dpd\_annuity\_sum\_10d\_ipoteka','cred\_insur\_cnt\_ipoteka','cred\_req\_ipoteka\_cnt\_90d\_issued','cred\_req\_ipoteka\_cnt\_180d\_issued','depo\_in\_euro\_30d','depo\_in\_euro\_90d','depo\_in\_euro\_360d','pkb\_bvu\_loan\_cnt\_lender\_alhilal','pkb\_bvu\_monthly\_payment\_lender\_vtb','pkb\_bvu\_od\_pledge\_lgot\_avto','pkb\_bvu\_max\_delay\_day\_cnt\_2y\_pledge\_7\_20\_25','pkb\_bvu\_max\_delay\_day\_cnt\_2y\_pledge\_lgot\_avto','pkb\_bvu\_max\_delay\_amount\_2y\_pledge\_7\_20\_25','pkb\_bvu\_max\_delay\_amount\_2y\_pledge\_lgot\_avto','pkb\_bvu\_monthly\_payment\_lender\_alhilal','CBoverdues\_90ParametrSet','pkb\_bvu\_credit\_sum\_lender\_alhilal','pkb\_bvu\_od\_lender\_alhilal','pkb\_bvu\_max\_delay\_day\_cnt\_2y\_lender\_vtb','pkb\_bvu\_max\_delay\_day\_cnt\_2y\_lender\_alhilal','pkb\_bvu\_max\_delay\_day\_cnt\_2y\_lender\_shinkhan','pkb\_bvu\_max\_delay\_amount\_2y\_lender\_vtb','pkb\_bvu\_max\_delay\_amount\_2y\_lender\_alhilal','pkb\_bvu\_max\_delay\_amount\_2y\_lender\_shinkhan','pkb\_bvu\_nominal\_rate\_lender\_alhilal','pkb\_bvu\_loan\_cnt\_pledge\_lgot\_avto','pkb\_bvu\_monthly\_payment\_pledge\_lgot\_avto','pkb\_bvu\_credit\_sum\_pledge\_lgot\_avto'* были удалены из-за количества уникальных значений меньше или равно двух с использованием листа: vc\_2\_cols = [col for col in dataprep.columns if dataprep[col].nunique(dropna=False) <= 2]. Также были удалены следующие уникальные переменные: номера счетов в Colvir (*InternalDelinquencyAccounts\_Table\_stat*); номера документов (*'DOC\_NUMBER', 'documents\_document\_number');* id клиентов(*'DOC\_NUMBER', 'documents\_document\_number'*); id заявок (*'id', 'ProcessID', 'order\_id', 'process\_ref', 'PROCESS\_REF'*); статусы контрактов (*'STATUS\_NAME'*); дублирующиеся переменных на казахском языке (*'documents\_document\_type\_nameKz','birthPlace\_region\_nameKz','birthPlace\_district\_nameKz','birthPlace\_country\_nameKz', 'regAddress\_region\_nameKz', 'regAddress\_district\_nameKz', 'citizenship\_nameKz', 'documents\_document\_status\_nameKz', 'nationality\_nameKz'*); дублирующиеся переменные с большим процентом пустых значений (*'CBCreditHistoryFromServdssb\_ocds\_paramset','HBCreditHistorydssb\_ocds\_paramset','HBCreditHistoryFromServdssb\_ocds\_paramset','FinalCreditHistorydssb\_ocds\_paramset','CtrlGrpdssb\_ocds\_paramset', 'IsRefinancedssb\_ocds\_paramset', 'ASPFlagdssb\_ocds\_paramset', 'Categorydssb\_ocds\_paramset','MaritalStatusdssb\_ocds\_paramset','CBCreditHistorydssb\_ocds\_paramset','ClientCategorydssb\_ocds\_paramset','IsCPPPaysInLast2Monthsdssb\_ocds\_paramset','bindssb\_ocds\_paramset','cbcredithistorydmparse2','hbcredithistorydmparse2','finalcredithistorydmparse2'*); переменные с количеством уникальных значений более 5000 из-за сложности внедрения в кредитный конвейер (*'regAddress\_city', 'birthPlace\_city', 'birthPlace\_foreignData\_RegionName', 'InternalDelinquencyAccounts\_Table\_deaCode'*); непонятные значения переменных (*'Branch TypeParametrSet','InternalDelinquencyAccounts\_Table\_prdName','InternalDelinquencyAccounts\_Table\_dclName','ClientCategoryParametrSet','PKB\_HEAD\_REP\_TYPE','WarningCodedssb\_ocds\_paramset','pkb\_resident\_fl','FinHasCredHistdssb\_ocds\_paramset','SalaryByWordParametrSet'*).





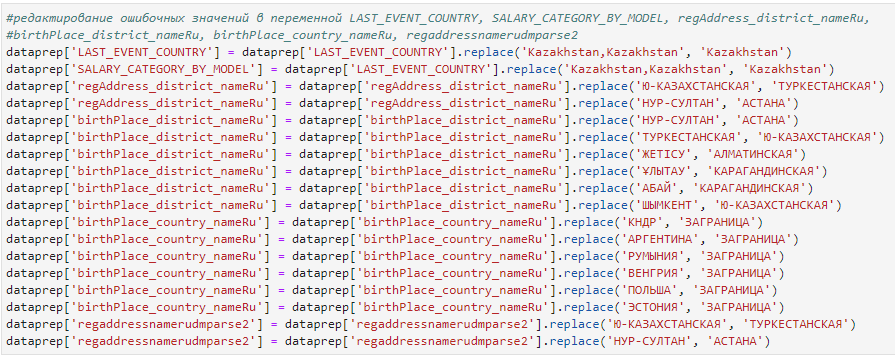


Рисунок 3. Редактирование значений в переменных CITIZENSHIP, EDUCATION, maritalstatusdmparse2, educationtypedmparse2, positiontypedmparse2, housingtypedmparse2, LAST\_EVENT\_COUNTRY,SALARY\_CATEGORY\_BY\_MODEL,regAddress\_district\_nameRu, birthPlace\_district\_nameRu, birthPlace\_country\_nameRu, regaddressnamerudmparse2

Были созданы переменные *datediff\_contropendate\_documentbegindate,datediff\_contropendate\_GtsvpServiceResponseData\_Request\_documentIssueDate,datediff\_contropendate\_regAddress\_region\_changeDate,datediff\_contropendate\_birthPlace\_region\_changeDate,datediff\_contropendate\_ColvirDepositAccount\_Fromdate,datediff\_contropendate\_documents\_document\_iIvalidityDate,datediff\_contropendate\_documents\_document\_endDate* как разница в днях между датой открытия контракта и датой начала документа, датой изменением адреса регистрации, датой изменением адреса рождения, датой открытия депозита в ХБ, датой валидности документа соответственно (см. рис. 4).

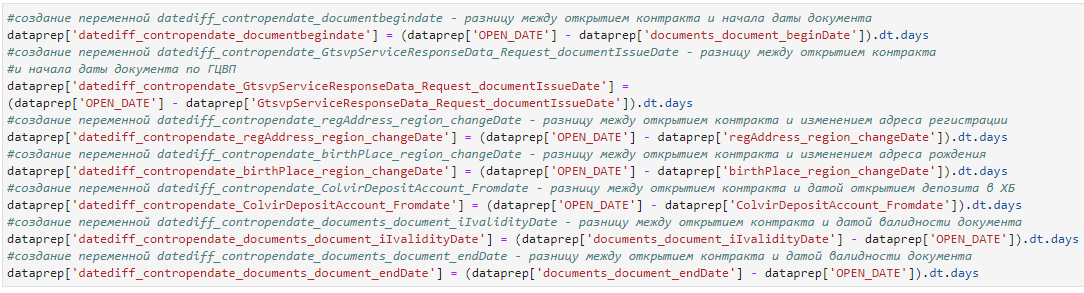


Рисунок 4. Создание временных переменных

Для исправления ошибочных значений в переменных SALARY\_CATEGORY\_BY\_MODEL, TRANSACTION\_CLIENT\_SEGMENT, MCG\_TOP1, MCG\_TOP2, MCG\_TOP3 была реализована функция extract\_lastint, извлекающая последнее значение из стринговых значений (см. рис. 5).

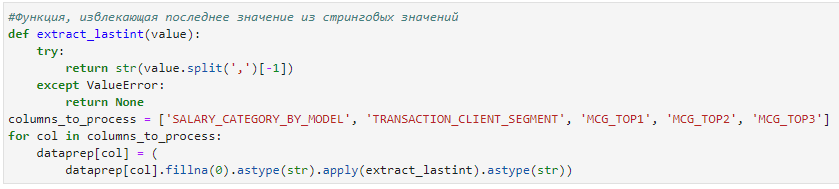


Рисунок 5. Функция. Извлекающая последнее значение из стринговых значений

Для сокращения уникальных значений в переменной ‘nationality\_nameRu’ было выполнено преобразование значений из-за пола, подающего заявку (рис. 6).



Рисунок 6. Преобразование значений в переменной ‘nationality\_nameRu’

В переменных *CBClosedContrCntParametrSet, ChildQntyParametrSet, PoVClient\_Client24hCountOrderOnlineApproved,VerIncAmountParametrSet,PoVClient\_Client0dMaxUpsellOnline, PoVClient\_Client24hMaxUpsellOnline, FinalIncomeOldParametrSet, FinalIncomeParametrSet, PoVClient\_Client72hMaxUpsellOnline, CBIncomeParametrSet, ColvirAccountAmountParametrSet, NBCardParametrSet, CB1YIncomeParametrSet, CB3YIncomeParametrSet, IncomeNPAParametrSet, MonthlyPaymentParametrSet, HalykBankCardOldParametrSet, CBClosedKZTContrSumParametrSet* выбросы были заменены на средние значения переменных, которые были выявлены при просмотре boxplot-ов.

Была создана переменная KDN как сумма обязательств в ХБ и ПКБ, разделенная на доход клиента: dataprep['KDN'] = np.where(dataprep['FinalIncomeParametrSet'] !=0, (dataprep['HBMPtsAmntParametrSet']+dataprep['CBMPtsAmntParametrSet'])/dataprep['FinalIncomeParametrSet'], np.nan).

В результате предобработки данных были отобраны 1015 независимых переменных. Перед экспортом предобработанных данных для последующего моделирования, были удалены все переменные формата datetime.

## 5 Модель логистической регрессии

Наиболее распространенными методами построения скоринговой модели являются логистическая регрессия и бустинговые методы. Первый метод имеет большие преимущества если имеют линейную структуру и выполнена качественная предварительная обработка данных: простота и интерпретируемость модели, где влияние каждого признака легко интерпретировать; легко объяснима регуляторам; менее склонна к переобучению при небольшом количестве признаков; легко выявить изменение в поведении модели со временем с использованием PSI – Population Stability Index; легко внедряется в существующие системы принятия решений в режиме реального времени. При наличии данных с нелинейной структурой лучше использовать более сложные алгоритмы.

Основные этапы построения скоринговой модели включают следующие процессы: отбор переменных, разделение на обучающую и тестовую выборки, построение модели, отбор финальных переменных, настройка параметров модели, построение финальной модели. Для построения модели формируются две выборки: обучающая — по ней строится модель, и тестовая — используется для проверки модели.

Анализ зависимостей между целевой и независимыми переменными проводился с целью сокращения списка переменных для построения модели и отсеивания не влияющих переменных. При выборе значимых количественных переменных в скоринговой модели проводилась категоризация переменных на основе оптимальной (‘optimal binning’) и визуальной категоризации. Оптимальный биннинг распределяет числовую характеристику по ячейкам для последующего использования при моделировании оценки. Для выбора значимых категориальных переменных в скоринговой модели рассчитывался показатель Information Value (IV) – Информационная ценность, которая возрастает по мере увеличения бинов/групп для независимой переменной, соответственно позволяет ранжировать переменные на основе их важности.

В результате, в качестве основного показателя предсказательной способности дефолтов был выбран показатель «IV». Чем выше информационное значение предиктора, тем больший вес он имеет с точки зрения полезности при построении модели: меньше 0,02 – предиктор не обладает прогностической способностью; от 0,02 до 0,1 – низкая прогностическая способность; от 0,1 до 0,3 – средняя прогностическая способность; от 0,3 до 0,5 – высокая прогностическая способность; более 0,5 – превосходная прогностическая способность.

При итоговом отборе переменных для построения модели логистической регрессии учитывались следующие факторы: содержательная интерпретация переменной и логичность ее использования для скоринга; отсутствие дублирования информации: не должно быть близких по содер категориальных переменных, а также сильно коррелирующих друг с другом количественных переменных; анализ и логичность трендов по WOE по образованным порядковым переменным – при переходе от большей к меньшей категории (или наоборот) риск дефолта должен возрастать или падать; Информационное значение (IV) - 0,1 и более.

Переменные с IV меньше 0,05 были удалены из выборки. Анализ значений WOE по каждой категории независимых переменных проводился с целью оптимизации количества категорий по каждой переменной и логичности тренда WOE. При расчете показателя WOE сравнивалась доля плохих к хорошим, поэтому большие значения WOE показывают, что для данной категории риск дефолта больше по сравнению с категориями с меньшим значением WOE. По всем переменным изменение значений показателя веса категорий в предсказании дефолта имеют логичную интерпретацию.

Модель строилась по методу множественной логистической регрессии методом исключения. Метод исключения представляет собой добавление всех переменных в модель, далее исключение переменных с наименьшим значением IV до получения качественной модели. При построении использовалась следующая тактика получения итоговой модели:

- незначимые переменные удалялись из модели (во множественной регрессии на фоне влияния остальных переменных они могут перестать быть значимыми);

- если переменная значимая в целом, но получены незначимые коэффициенты при некоторых фиктивных переменных, то проводилась альтернативная группировка переменной или укрупнение категорий, цель построения модели – получить все значимые коэффициенты при всех фиктивных переменных;

- если получено значимое уравнение со всеми значимыми коэффициентами, то дополнительно анализировались присвоение баллов в категориях по каждому параметру. Если баллы в категориях близки по значению, то такие категории объединялись в одну.

В приложении 3 приведены информационные значения переменных, отобранных для модели в логистической регрессии.

При окончательном отборе переменных для построения модели логистической регрессии учитываются следующие факторы: содержательная интерпретация переменной и логичность ее использования для скоринга; отсутствие дублирования информации: не должно быть близких по содержанию категориальных переменных, а также сильно коррелирующих друг с другом количественных переменных; анализ и логичность трендов по весам факторов (WOE) по образованным порядковым переменным – при переходе от большей к меньшей категории (или наоборот) риск дефолта должен возрастать или падать.

При отборе переменных учитываются следующие моменты: независимость между объясняющими переменными, итоговую прогностическую способность модели, интерпретируемость получаемых коэффициентов при значимых переменных. Для проверки адекватности и точности предсказания скоринговой модели на этапе ее разработки историческую выборку необходимо разделить на 2 группы: обучающую выборку – наблюдения, по которым будет непосредственно строится модель и тестовую выборки.

Для построения окончательной модели логистической регрессии и отбора финальных переменных используются стандартные пошаговые алгоритмы включения и исключения переменных или их сочетание для получения значимого уравнения регрессии с достаточной прогностической способностью и со всеми значимыми коэффициентами регрессии. Каждая итерация процесса является этапом обучения. Производится настройка параметров модели для повышения производительности. Анализ зависимостей между целевой и независимыми переменными проводится с целью сокращения списка переменных для построения модели и отсеивания не влияющих переменных.

По количественным переменным проводится категоризация на основе процедуры «Оптимальная категоризация» (optimal binning). Для оценки предсказательной способности независимых переменных также рассчитывается показатель Информационный критерий (IV).

Весь набор данных по выданным контрактам для разработки скоринговой модели был поделен на обучающие (70%) и тестовые (30%) наборы (см. таблица 6).

Таблица 6. Распределение целевой переменной по продукту Товары в кредит/рассрочку (зависимая переменная – NPL91\_EVER)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Наличие дефолта** | **Обучающая выборка** | | **Тестовая выборка** | |
| **Частота** | **Проценты** | **Частота** | **Проценты** |
| нет | 138 055 | 97.0 % | 59 141 | 97.0 % |
| да | 4 240 | 3.0 % | 1 843 | 3.0 % |
| Всего | 142 295 | 100.00 % | 60 984 | 100.00 % |

**Оценка качества модели**

Проверка достоверности модели заключается в ее применении и сравнении результатов на обучающей и контрольной выборках. Модель должна давать корректные прогнозы не только на обучающей совокупности, но и на практике при ее применении.

Достоверность модели логистической регрессии характеризуется ее способностью отличать «хороших» заемщиков от «плохих». Дискриминирующую способность модели можно оценить, проанализировав таблицу классификации (матрица ошибок). Для оценки качества классификации модели строят ROC-кривую, которая показывает зависимость количества верно классифицированных положительных исходов от количества неверно классифицированных отрицательных исходов. Площади под ROC-кривыми показатель называется AUC(Area under the Receiver Operating Characteristic Curve) и измеряется от 0.5 до 1. Чем больше значение площади, тем лучше модель. По значению площади под ROC-кривой можно вычислить показатель индекс Джини. Показатель Джини переводит значение площади под кривой в диапазон о 0 до 1.

Общепринятая интерпретация индекса Джини: менее 30 – не пригодна; 30-50 – нормальная; 50-60 – хорошая; 60-70 – очень хорошая; 70-85 – превосходная; 85 и выше – редко бывает, следует проверить (слишком хорошо, чтобы быть правдой).

Скоринговые карты разрабатываются для ранжирования заемщиков по шансам наступления просрочки по кредиту. Чем более явно разделены распределения скоринговых баллов для плохих и хороших кредитов, тем эффективнее будет работать скоринговая карта. Для оценки качества простроенной по модели логистической регрессии скоринговой карты также анализируют распределение скоринговых баллов по плохим и хорошим заемщикам. Для оценки качества прогнозирования модели логистической регрессии на основе этого распределения рассчитывают тест Колмогорова-Смирнова(далееКС). В тесте КС сравниваются два кумулятивных распределения скоринговых баллов «хороших» и «плохих» заемщиков. Статистика КС вычисляется как максимальная разница между кумулятивными функциями этих распределений. Диапазон значений статистики КС измеряется от 0 до 100. Чем выше значение статистики КС, тем лучше работает модель.

Общепринятая интерпретация статистики Колмогорова-Смирнова: менее 20 –не пригодна; 20-40 – нормальная; 40-50 – хорошая; 50-60 – очень хорошая; 60-75 – превосходная; 75 и выше – редко бывает, следует проверить (слишком хорошо, чтобы быть правдой).

Хорошая модель дает приемлемые результаты точности и на обучающей, и на контрольной выборке. Схожие показатели, полученные на обеих выборках — признак того, что на практике модель будет выдавать верные прогнозы. Значения оценок коэффициентов и эффективности на разных выборках не должны существенно различаться, это говорит о статистической устойчивости модели и надежности полученных коэффициентов регрессии

Для оценки стабильности модели и ее статистической устойчивости модель проверялась на тестовой и на всей выборке. Значения оценки коэффициентов на разных выборках практически не различаются и всегда статистически значимы. Были выполнены проверки на недообучение и переобучение модели. Модели статистически устойчивы и полученные коэффициенты регрессии надежны.

В финальную модель после анализа корреляции, информационных значений, просмотра бинов и значимости признаков было отобрано 14 параметров из источников ПКБ и XML заявки в соотношении 7-4-3 (ПКБ - 50%, Биографические - 29%, Внутренние ХБ -21%):

1) Пол,

2) Ежемесячный платеж последнего одобренного оффера,

3) Количество фактических платежей в ПКБ,

4) Возраст,

5) Количество платежей без просрочек,

6) Максимальное количество дней просрочки в ПКБ,

7) Оценка кредитной истории ПКБ,

8) Количество запросов в ПКБ за последний год,

9) Сумма остатка кредитов/кредитных лимитов (ПКБ),

10) Макс. кол-во дней проср. с начала действия договора (ПКБ),

11) Объем количества просроченных платежей свыше 1 дней (ПКБ),

12) Уровень образования,

13) Прожиточный минимум: (dependents\_count\*0.5+1)\*43407,

14) КДН.

Распределение уровня дефолта в бинах в обучающей и тестовой выборках приведены в Приложении 4.

После бинирования независимых переменных, данные группировки были трансформированы в WOE. Параметры логистической регрессии – penalty, solver, class\_weight, C были выбраны методом GridSearchCV (рисунок 7). Площадь под ROC-кривой модели логистической регрессии с зависимыми переменными приведены на рис. 8.



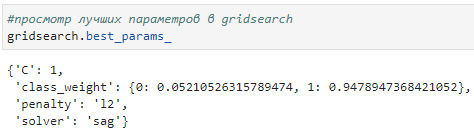


Рисунок 7. GridSearchCV, отобранные параметры для лог.регрессии

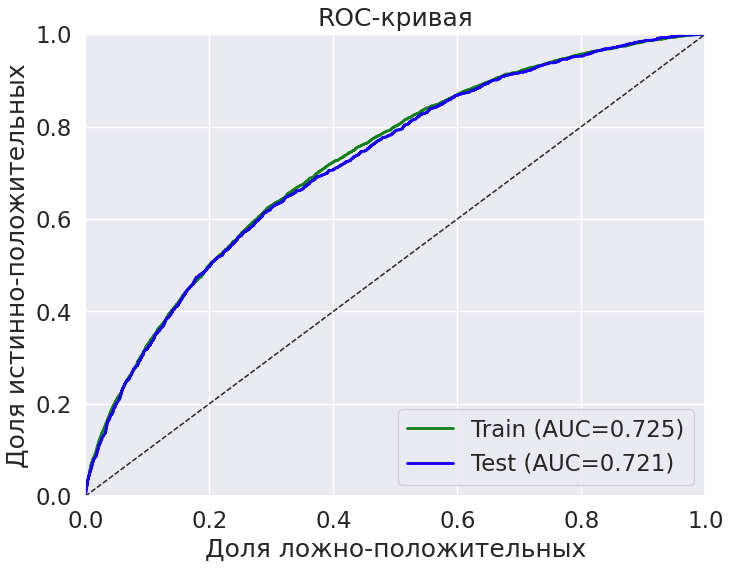


Рисунок 8. Площадь под ROC-кривой полученной модели методом лог.регрессии

Качество (Джини) модели логистической регрессии с зависимой переменной NPL91\_EVER на обучающей и тестовой выборках составило 0.45 и 0.44. Статистики KS, AUC и Gini приведены в таблицах 7 и 8.

Таблица 7. Качество модели логистической регрессии

|  |  |  |  |
| --- | --- | --- | --- |
| **Train** | | **Test** | |
| KS | 0,331 | KS | 0,325 |
| AUC | 0,725 | AUC | 0,721 |
| **Gini** | **0,45** | **Gini** | **0,44** |

Таблица 8. Матрицы ошибок в обучающей и тестовой выборках c зависимой переменной (NPL91\_EVER)

|  |  |  |  |
| --- | --- | --- | --- |
|  | |  | |
| Confusion matrix (train) | | Confusion matrix (test) | |
| Default | 65% | Default | 65% |
| Non-default | 68% | Non-default | 67% |

Для просмотра значимости переменных был построен график важности признаков, полученных из коэффициентов логистической регрессии (см. рис. 9). Согласно графику, все отобранные переменные имеют хорошую степень значимости по коэффициентам в модели – от 0.18 и более.

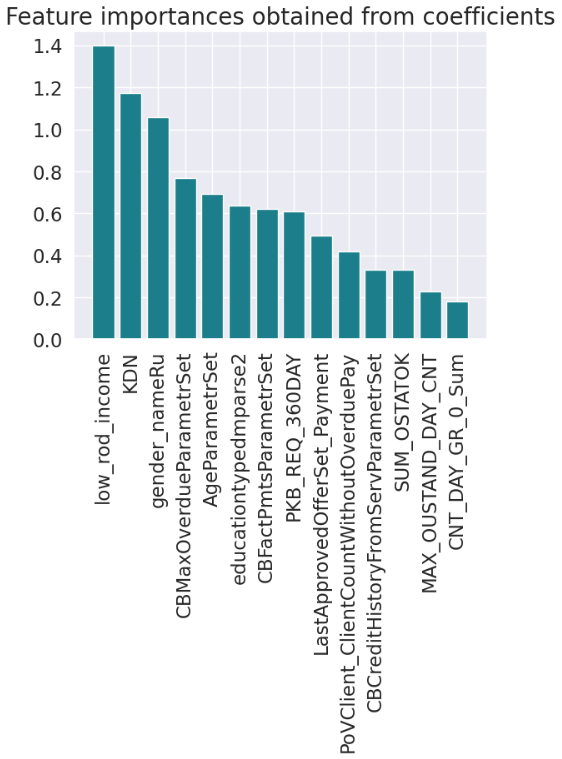


Рисунок 9. Важность признаков, полученные из коэффициентов логистической регрессии

**Расчет скоринговых баллов**

Классификацию заемщиков можно проводить также на основе баллов, а не по вероятности дефолта.

Расчет баллов – это масштабирование коэффициентов регрессии в новую шкалу. При этом расчет баллов можно проводить без учета константы модели. Выбор шкалы баллов не влияет на предсказательную силу скоринговой карты. Индекс Джини и площадь под ROC кривой при переводе вероятностей дефолта в баллы от этого не изменится.

Масштабирование – это перевод баллов из шкалы натуральных логарифмов в удобный для пользователя вид, исходя из критериев возможности реализации в конкретном программном обеспечении, простоты понимания для сотрудников, преемственности и совместимости с существующими в компании скоринговыми картами.

Балл в шкале натуральных логарифмов рассчитывается по формуле:

,

где  – оценки коэффициентов модели;

 – значения независимых переменных, представленные фиктивными переменными.

Для перевода балла из шкалы натуральных логарифмов необходимо выбрать следующие параметры:

1. скорость изменения положительных шансов по мере роста баллов – число баллов для удвоения шансов быть плохим (PDO);
2. точку, в которой должно быть заданное соотношение шансов быть плохим.

Далее был рассчитан балл по следующему стандарту: PDO = 50 – каждые 50 баллов удваивают шанс быть плохим.

Формула для перевода баллов из шкалы натуральных логарифмов в линейную шкалу, при которой больший балл присваивается более надежным заемщикам:

В приложении 1 приведены модель логистической регрессии для продукта Пенсионная.

Для интерпретации можно использовать сами коэффициенты в Приложении 1 – коэффициент PD положительный означает увеличение риска дефолта для соответствующей категории переменной, а коэффициент PD с отрицательным знаком – уменьшении риска.

В целом интерпретация коэффициентов регрессии логична для всех переменных.

## 6 Модель логистической регрессии с ансамблем моделей

Для повышения качества модели логистической регрессии было решено внедрить 15-м параметром в модель логистической регрессии вероятность дефолта, полученной взвешенным усреднением от трех моделей: LightGBM, XGBoost и CatBoost. Использование взвешенного ансамбля позволяет повысить точность и стабильность прогнозов за счет объединения сильных сторон каждой модели и минимизации их слабых мест. Совместное использование моделей позволяет усреднить ошибки, снизить влияние шума и случайных ошибок отдельных моделей, делая прогноз более стабильным. Используя преимущества каждой из моделей – CatBoost (обработка категориальных переменных), LightGBM (большой набор данных), XGBoost (стабильность на хорошо структурированных данных) возможна компенсация слабых сторон моделей. Также используя веса можно уменьшить влияние моделей, склонных к переобучению или менее устойчивых на данных разной структуры – численные, категориальные и бинарные.

Для построения ансамбля моделей были проанализированы все параметры из исходного датасета, переменные с IV менее 0.05 были удалены из выборки, также были удалены переменные, использующиеся в логистической регрессии, переменные из заявок в режиме реального времени, переменные, коррелирующиеся с переменными логистической регрессии.

Для каждого обученного бустингового классификатора на датасете была вычислена важность признаков, откуда были выбраны топ-25 признаков для обучения моделей по основным признакам. Также были рассчитаны веса для каждой модели на основе точности классификатора и рассчитана вероятность дефолта на основе взвешенного объединения предсказаний от трех моделей.

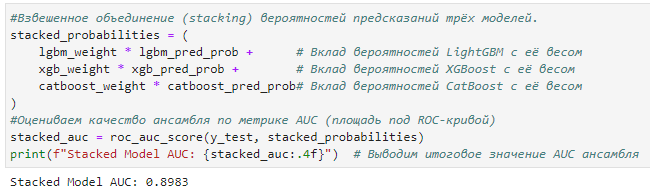


Рисунок 10. Взвешенное объединение вероятностей предсказаний трех моделей

Спрогнозированная вероятность дефолта была разделена на 6 бинов от 0 до 1. Качество ансамбля моделей (AUC) составило 0.83 и 0.80 на обучающей и тестовой выборках соответственно. Признаки значимости отобранных переменных в ансамбле моделей приведены на рисунке 12. В приложении 2 приведена построенная скоринговая модель на основе логистической регрессии и ансамбля моделей.

## 7 Оценка качества модели логистической регрессии с ансамблем моделей

Для проверки качества скоринговой модели проводилась оценка площади под ROC-кривой (AUC) (см. рис. 11). Далее по AUC рассчитывался индекс Gini ((AUC-0,5)\*2) и тест Колмогорова-Смирнова (см. табл. 9). Полученная модель на обучающей и на тестовой выборках показала схожие отличные результаты - Gini более 0,80 и КС более 0,69.

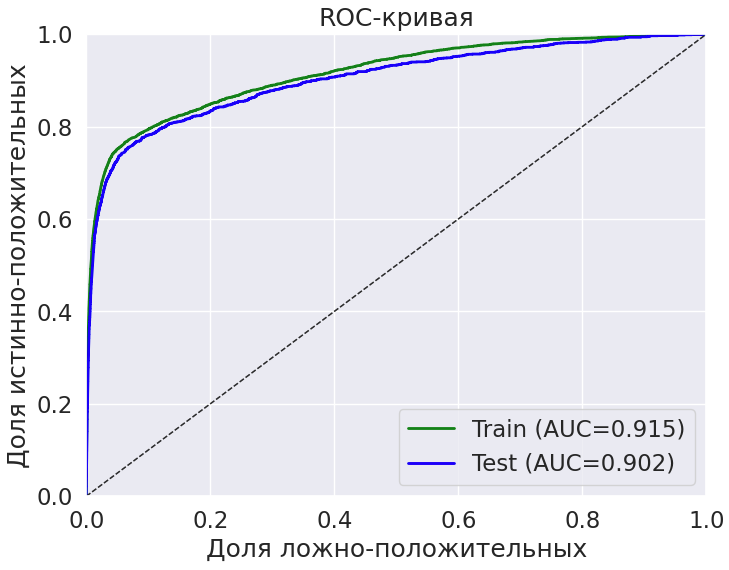


Рисунок 11. Площадь под ROC-кривой полученной модели логистической регрессии с параметром из ансамбля моделей

Таблица 9. Качество построенной скоринговой модели методом логистической регрессии с добавлением параметра из ансамбля моделей

|  |  |  |  |
| --- | --- | --- | --- |
| **Train** | | **Test** | |
| KS | 0,704 | KS | 0,687 |
| AUC | 0,915 | AUC | 0,902 |
| **Gini** | **0,83** | **Gini** | **0,80** |

Матрица ошибок является показателем успешности классификации с двумя и более классами. В таблице 10 приведены матрицы ошибок в обучающей и тестовой выборках. На обучающей выборке 3 196 плохих заемщиков были определены правильно, 1 044 заемщиков были ошибочно оценены как плохие – качество определения плохих заемщиков 75%, 130 855 хороших заемщиков были определены правильно, 7 200 заемщиков были ошибочно оценены как плохие заемщики, качество определения хороших заемщиков 95%.

Таблица 10. Матрицы ошибок в обучающей и тестовой выборках c зависимой переменной (NPL91\_EVER)

|  |  |  |  |
| --- | --- | --- | --- |
|  | |  | |
| Confusion matrix (train) | | Confusion matrix (test) | |
| Default | 75% | Default | 72% |
| Non-default | 95% | Non-default | 95% |

Для просмотра значимости переменных был построен график важности признаков, полученных из коэффициентов логистической регрессии (см. рис. 12).

Качество определения плохих и хороших клиентов на тестовой выборке составило 72% и 95% соответственно (табл.10).

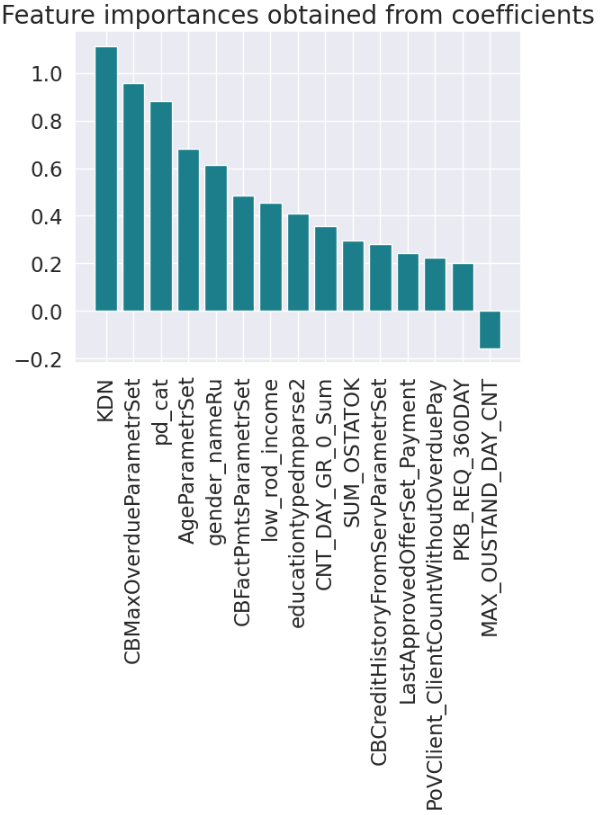


Рисунок 12. Значимость признаков логистической регрессии с параметром из ансамбля моделей

## 8 Реализация модели и операции

Проверка эффективности скоринговой модели (валидация) проводится после ее внедрения и осуществляется мониторинг в течение всего срока эксплуатации. Валидация осуществляется посредством одного или нескольких следующих методов: проверка дискриминационной способности модели; оценка прогнозной точности модели; анализ миграции рейтингов; сравнительный анализ рейтингов.

Результатом проверки эффективности скоринговой модели (валидации) является следующее: модель в хорошем состоянии; необходимо внесение корректировок/доработка модели для улучшения ее производительности; необходима полная переработка модели.

Чтобы смягчить влияние смещений распределения, непрерывный мониторинг и перекалибровка модели оценки необходимы для поддержания ее предсказательной валидности в изменяющихся условиях. Регулярные обновления модели могут обеспечить ее адаптивность и эффективность в различных сценариях. Адаптивность моделей оценки имеет решающее значение для реагирования на рёазвивающиеся распределения данных и обеспечения продолжительной предсказательной точности в различных приложениях.

Валидация осуществляется не реже 1 (одного) раза в 1 (один) год. Частота проведения валидации зависит от текущей рыночной ситуации, стратегии, объема активов, уровня сложности операций Банка, увеличивается в случае существенных изменений в экономике или во внутренних процессах кредитования Банка. Результаты валидации предоставляются Комитету по вопросам управления рисками.

В случае если текущая скоринговая модель признана устаревшей и/или неэффективной, осуществляется рекалибровка существующей модели. Процесс рекалибровки аналогичен процессу построения скоринговой модели. В случае если параметры в отдельных переменных текущей скоринговой модели не актуальны или являются неэффективными, осуществляется их изменение.

Для автоматизации рекалибровки модели следует внедрить самообучение моделей, при котором модель изначально обучается на небольшом количестве размеченных данных, затем итеративно улучшается методом многоступенчатого самообучения с постепенным включением большого количества неразмеченных данных.

## 9 Заключение

Качество скоринговых моделей – модель логистической регрессии и комбинированная модель логистической регрессии с ансамблем моделей показывают отличные результаты. Учитывая полученные результаты по качеству и значимые/логичные параметры в уравнениях моделей, для принятия решений ДРЦК считает целесообразным внедрение разработанных моделей с зависимой переменной NPL91\_EVER в кредитную стратегию в процессе кредитования по продукту Пенсионная.

## Приложение 1

**Скоринговая модель с зависимой переменной NPL91\_EVER (логист. регрессия)**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **No** | **Переменная** | **Variable** | **Binning** | **Count** | **Default** | **Non-default** | **Dist\_Bads** | **Dist\_Goods** | **Score** | **WOE** | **Coef (-0,0082)** | **PD** |
| 1 | Пол (XML) | gender\_nameRu | Женский | 128646 | 3054 | 125592 | 50,21% | 63,69% | 19 | -0,246 | 1,059 | 0,435 |
| Мужской | 74633 | 3029 | 71604 | 49,79% | 36,31% | -25 | 0,324 | 1,059 | 0,585 |
| 2 | Ежемесячный платеж последнего одобренного оффера (XML) | LastApprovedOfferSet\_Payment | >=30000 | 66818 | 1484 | 65334 | 24,40% | 33,13% | 11 | -0,313 | 0,496 | 0,461 |
| nan | 56169 | 1678 | 54491 | 27,59% | 27,63% | 0 | 0,000 | 0,496 | 0,500 |
| 5000-30000 | 68834 | 2237 | 66597 | 36,77% | 33,77% | -3 | 0,090 | 0,496 | 0,511 |
| <5000 | 11458 | 684 | 10774 | 11,24% | 5,46% | -26 | 0,718 | 0,496 | 0,588 |
| 3 | Количество фактических платежей в ПКБ (XML) | CBFactPmtsParametrSet | >=100 | 49440 | 1043 | 48397 | 17,15% | 24,54% | 17 | -0,368 | 0,623 | 0,443 |
| 40-99 | 90136 | 2286 | 87850 | 37,58% | 44,55% | 8 | -0,172 | 0,623 | 0,473 |
| 24-39 | 29291 | 984 | 28307 | 16,18% | 14,35% | -5 | 0,105 | 0,623 | 0,516 |
| 13-23 | 10902 | 509 | 10393 | 8,37% | 5,27% | -22 | 0,484 | 0,623 | 0,575 |
| <=12 | 23510 | 1261 | 22249 | 20,73% | 11,28% | -28 | 0,620 | 0,623 | 0,595 |
| 4 | Возраст (XML) | AgeParametrSet | <58 | 12006 | 206 | 11800 | 3,39% | 5,98% | 27 | -0,550 | 0,692 | 0,406 |
| >=68 | 19574 | 406 | 19168 | 6,67% | 9,72% | 20 | -0,402 | 0,692 | 0,431 |
| 58-62 | 38835 | 1018 | 37817 | 16,74% | 19,18% | 6 | -0,117 | 0,692 | 0,480 |
| 63-67 | 132864 | 4453 | 128411 | 73,20% | 65,12% | -6 | 0,114 | 0,692 | 0,520 |
| 5 | Количество платежей без просрочек (XML) | PoVClient\_ClientCountWithoutOverduePay | >=70 | 51791 | 981 | 50810 | 16,13% | 25,77% | 14 | -0,452 | 0,418 | 0,453 |
| 30-69 | 47572 | 1108 | 46464 | 18,21% | 23,56% | 9 | -0,283 | 0,418 | 0,471 |
| 7-29 | 36028 | 1200 | 34828 | 19,73% | 17,66% | -4 | 0,121 | 0,418 | 0,513 |
| nan | 56351 | 2109 | 54242 | 34,67% | 27,51% | -7 | 0,216 | 0,418 | 0,523 |
| <= 6 | 11537 | 685 | 10852 | 11,26% | 5,50% | -23 | 0,756 | 0,418 | 0,578 |
| 6 | Максимальное количество дней просрочки (XML) | CBMaxOverdueParametrSet | <=9 | 179949 | 4987 | 174962 | 81,98% | 88,72% | 4 | -0,075 | 0,767 | 0,486 |
| >=10 | 23330 | 1096 | 22234 | 18,02% | 11,28% | -25 | 0,446 | 0,767 | 0,585 |
| 7 | Оценка кредитной истории ПКБ (XML) | CBCreditHistoryFromServParametrSet | Положительная | 135939 | 3100 | 132839 | 50,96% | 67,36% | 7 | -0,291 | 0,332 | 0,476 |
| Приемлемая 2','Приемлемая 1', 'Удовлетворительная' | 61722 | 2595 | 59127 | 42,66% | 29,98% | -9 | 0,369 | 0,332 | 0,531 |
| Отрицательная | 5618 | 388 | 5230 | 6,38% | 2,65% | -21 | 0,862 | 0,332 | 0,571 |
| 8 | Количество запросов в ПКБ за последний год (ПКБ) | PKB\_REQ\_360DAY | <=2 | 93605 | 1985 | 91620 | 32,63% | 46,46% | 16 | -0,356 | 0,610 | 0,446 |
| 3-7 | 52315 | 1420 | 50895 | 23,34% | 25,81% | 6 | -0,128 | 0,610 | 0,480 |
| nan | 31906 | 1254 | 30652 | 20,61% | 15,54% | -13 | 0,300 | 0,610 | 0,546 |
| 8-15 | 17353 | 776 | 16577 | 12,76% | 8,41% | -18 | 0,403 | 0,610 | 0,561 |
| >=16 | 8100 | 648 | 7452 | 10,65% | 3,78% | -48 | 1,085 | 0,610 | 0,660 |
| 9 | Сумма остатка кредитов/кредитных лимитов (ПКБ) | SUM\_OSTATOK | < 100000 | 44146 | 842 | 43304 | 13,84% | 21,96% | 12 | -0,483 | 0,331 | 0,460 |
| 100000-499999 | 51778 | 1374 | 50404 | 22,59% | 25,56% | 3 | -0,122 | 0,331 | 0,490 |
| nan | 53341 | 1694 | 51647 | 27,85% | 26,19% | -2 | 0,064 | 0,331 | 0,505 |
| >=500000 | 54014 | 2173 | 51841 | 35,72% | 26,29% | -7 | 0,310 | 0,331 | 0,526 |
| 10 | Макс. кол-во дней проср. с начала действия договора (ПКБ) | MAX\_OUSTAND\_DAY\_CNT | <=11 | 123432 | 2860 | 120572 | 47,02% | 61,14% | 4 | -0,259 | 0,229 | 0,485 |
| 12-27 | 28392 | 946 | 27446 | 15,55% | 13,92% | -1 | 0,091 | 0,229 | 0,505 |
| nan | 31906 | 1254 | 30652 | 20,61% | 15,54% | -5 | 0,300 | 0,229 | 0,517 |
| >=28 | 19549 | 1023 | 18526 | 16,82% | 9,39% | -9 | 0,566 | 0,229 | 0,532 |
| 11 | Объем количества просроченных платежей свыше 1 дней (ПКБ) | CNT\_DAY\_GR\_0\_Sum | 0 | 95086 | 2186 | 92900 | 35,94% | 47,11% | 3 | -0,254 | 0,182 | 0,488 |
| 1-3 | 45158 | 1154 | 44004 | 18,97% | 22,31% | 2 | -0,172 | 0,182 | 0,492 |
| 4-10 | 19146 | 745 | 18401 | 12,25% | 9,33% | -3 | 0,211 | 0,182 | 0,510 |
| nan | 31906 | 1254 | 30652 | 20,61% | 15,54% | -4 | 0,300 | 0,182 | 0,514 |
| >=11 | 11983 | 744 | 11239 | 12,23% | 5,70% | -10 | 0,759 | 0,182 | 0,534 |
| 12 | Уровень образования (XML) | educationtypedmparse2 | ВЫСШЕЕ','УЧЁНАЯ СТЕПЕНЬ' | 99884 | 2288 | 97596 | 37,61% | 49,49% | 13 | -0,290 | 0,639 | 0,454 |
| СРЕДНЕЕ СПЕЦИАЛЬНОЕ | 63638 | 2064 | 61574 | 33,93% | 31,22% | -5 | 0,102 | 0,639 | 0,516 |
| None', 'СРЕДНЕЕ', 'НЕПОЛНОЕ СРЕДНЕЕ' | 39757 | 1731 | 38026 | 28,46% | 19,28% | -18 | 0,386 | 0,639 | 0,561 |
| 13 | Прожиточный минимум: (dependents\_count\*0.5+1)\*43407 | low\_rod\_income | > 65000 | 7260 | 123 | 7137 | 2,02% | 3,62% | 57 | -0,569 | 1,398 | 0,311 |
| <= 65000 | 75664 | 1669 | 73995 | 27,44% | 37,52% | 29 | -0,287 | 1,398 | 0,401 |
| nan | 120355 | 4291 | 116064 | 70,54% | 58,86% | -17 | 0,170 | 1,398 | 0,559 |
| 14 | КДН из XML (HBMPtsAmnt + CBMPtsAmnt )/ FinalIncome) | KDN | <0.1 | 75668 | 1788 | 73880 | 29,39% | 37,47% | 21 | -0,251 | 1,173 | 0,427 |
| 0.1-0.39 | 74554 | 2039 | 72515 | 33,52% | 36,77% | 7 | -0,083 | 1,173 | 0,476 |
| >=0.4 | 53046 | 2255 | 50791 | 37,07% | 25,76% | -31 | 0,361 | 1,173 | 0,604 |
| nan | 11 | 1 | 10 | 0,02% | 0,01% | -109 | 1,286 | 1,173 | 0,819 |

## Приложение 2

**Скоринговая модель с зависимой переменной NPL91\_EVER (лог. регрессия + ансамбль моделей)**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **No** | **Переменная** | **Variable** | **Binning** | **Count** | **Default** | **Non-default** | **Dist\_Bads** | **Dist\_Goods** | **Score** | **WOE** | **Coef (-0,0260)** | **PD** |
| 1 | Пол (XML) | gender\_nameRu | Женский | 128646 | 3054 | 125592 | 50,21% | 63,69% | 11 | -0,246 | 0,603 | 0,463 |
| Мужской | 74633 | 3029 | 71604 | 49,79% | 36,31% | -14 | 0,324 | 0,603 | 0,549 |
| 2 | Ежемесячный платеж последнего одобренного оффера (XML) | LastApprovedOfferSet\_Payment | >=30000 | 66818 | 1484 | 65334 | 24,40% | 33,13% | 4 | -0,313 | 0,168 | 0,487 |
| nan | 56169 | 1678 | 54491 | 27,59% | 27,63% | 0 | 0,000 | 0,168 | 0,500 |
| 5000-30000 | 68834 | 2237 | 66597 | 36,77% | 33,77% | -1 | 0,090 | 0,168 | 0,504 |
| <5000 | 11458 | 684 | 10774 | 11,24% | 5,46% | -9 | 0,718 | 0,168 | 0,530 |
| 3 | Количество фактических платежей в ПКБ (XML) | CBFactPmtsParametrSet | >=100 | 49440 | 1043 | 48397 | 17,15% | 24,54% | 17 | -0,368 | 0,633 | 0,442 |
| 40-99 | 90136 | 2286 | 87850 | 37,58% | 44,55% | 8 | -0,172 | 0,633 | 0,473 |
| 24-39 | 29291 | 984 | 28307 | 16,18% | 14,35% | -5 | 0,105 | 0,633 | 0,517 |
| 13-23 | 10902 | 509 | 10393 | 8,37% | 5,27% | -22 | 0,484 | 0,633 | 0,576 |
| <=12 | 23510 | 1261 | 22249 | 20,73% | 11,28% | -28 | 0,620 | 0,633 | 0,597 |
| 4 | Возраст (XML) | AgeParametrSet | <58 | 12006 | 206 | 11800 | 3,39% | 5,98% | 31 | -0,550 | 0,790 | 0,393 |
| >=68 | 19574 | 406 | 19168 | 6,67% | 9,72% | 23 | -0,402 | 0,790 | 0,421 |
| 58-62 | 38835 | 1018 | 37817 | 16,74% | 19,18% | 7 | -0,117 | 0,790 | 0,477 |
| 63-67 | 132864 | 4453 | 128411 | 73,20% | 65,12% | -6 | 0,114 | 0,790 | 0,523 |
| 5 | Количество платежей без просрочек (XML) | PoVClient\_ClientCountWithoutOverduePay | >=70 | 51791 | 981 | 50810 | 16,13% | 25,77% | 6 | -0,452 | 0,186 | 0,479 |
| 30-69 | 47572 | 1108 | 46464 | 18,21% | 23,56% | 4 | -0,283 | 0,186 | 0,487 |
| 7-29 | 36028 | 1200 | 34828 | 19,73% | 17,66% | -2 | 0,121 | 0,186 | 0,506 |
| nan | 56351 | 2109 | 54242 | 34,67% | 27,51% | -3 | 0,216 | 0,186 | 0,510 |
| <= 6 | 11537 | 685 | 10852 | 11,26% | 5,50% | -10 | 0,756 | 0,186 | 0,535 |
| 6 | Максимальное количество дней просрочки (XML) | CBMaxOverdueParametrSet | <=9 | 179949 | 4987 | 174962 | 81,98% | 88,72% | 3 | -0,075 | 0,623 | 0,488 |
| >=10 | 23330 | 1096 | 22234 | 18,02% | 11,28% | -20 | 0,446 | 0,623 | 0,569 |
| 7 | Оценка кредитной истории ПКБ (XML) | CBCreditHistoryFromServParametrSet | Положительная | 135939 | 3100 | 132839 | 50,96% | 67,36% | 2 | -0,291 | 0,108 | 0,492 |
| Приемлемая 2', 'Приемлемая 1', 'Удовлетворительная' | 61722 | 2595 | 59127 | 42,66% | 29,98% | -3 | 0,369 | 0,108 | 0,510 |
| Отрицательная | 5618 | 388 | 5230 | 6,38% | 2,65% | -7 | 0,862 | 0,108 | 0,523 |
| 8 | Количество запросов в ПКБ за последний год (ПКБ) | PKB\_REQ\_360DAY | <=2 | 93605 | 1985 | 91620 | 32,63% | 46,46% | 4 | -0,356 | 0,157 | 0,486 |
| 3-7 | 52315 | 1420 | 50895 | 23,34% | 25,81% | 1 | -0,128 | 0,157 | 0,495 |
| nan | 31906 | 1254 | 30652 | 20,61% | 15,54% | -3 | 0,300 | 0,157 | 0,512 |
| 8-15 | 17353 | 776 | 16577 | 12,76% | 8,41% | -5 | 0,403 | 0,157 | 0,516 |
| >=16 | 8100 | 648 | 7452 | 10,65% | 3,78% | -12 | 1,085 | 0,157 | 0,543 |
| 9 | Сумма остатка кредитов/кредитных лимитов (ПКБ) | SUM\_OSTATOK | < 100000 | 44146 | 842 | 43304 | 13,84% | 21,96% | 17 | -0,483 | 0,479 | 0,442 |
| 100000-499999 | 51778 | 1374 | 50404 | 22,59% | 25,56% | 4 | -0,122 | 0,479 | 0,485 |
| nan | 53341 | 1694 | 51647 | 27,85% | 26,19% | -2 | 0,064 | 0,479 | 0,508 |
| >=500000 | 54014 | 2173 | 51841 | 35,72% | 26,29% | -11 | 0,310 | 0,479 | 0,537 |
| 10 | Макс. кол-во дней проср. с начала действия договора (ПКБ) | MAX\_OUSTAND\_DAY\_CNT | <=11 | 123432 | 2860 | 120572 | 47,02% | 61,14% | 3 | -0,259 | 0,165 | 0,489 |
| 12-27 | 28392 | 946 | 27446 | 15,55% | 13,92% | -1 | 0,091 | 0,165 | 0,504 |
| nan | 31906 | 1254 | 30652 | 20,61% | 15,54% | -4 | 0,300 | 0,165 | 0,512 |
| >=28 | 19549 | 1023 | 18526 | 16,82% | 9,39% | -7 | 0,566 | 0,165 | 0,523 |
| 11 | Объем количества просроченных платежей свыше 1 дней (ПКБ) | CNT\_DAY\_GR\_0\_Sum | 0 | 95086 | 2186 | 92900 | 35,94% | 47,11% | 5 | -0,254 | 0,252 | 0,484 |
| 1-3 | 45158 | 1154 | 44004 | 18,97% | 22,31% | 3 | -0,172 | 0,252 | 0,489 |
| 4-10 | 19146 | 745 | 18401 | 12,25% | 9,33% | -4 | 0,211 | 0,252 | 0,513 |
| nan | 31906 | 1254 | 30652 | 20,61% | 15,54% | -5 | 0,300 | 0,252 | 0,519 |
| >=11 | 11983 | 744 | 11239 | 12,23% | 5,70% | -14 | 0,759 | 0,252 | 0,548 |
| 12 | Уровень образования (XML) | educationtypedmparse2 | ВЫСШЕЕ', 'УЧЁНАЯ СТЕПЕНЬ' | 99884 | 2288 | 97596 | 37,61% | 49,49% | 9 | -0,290 | 0,410 | 0,470 |
| СРЕДНЕЕ СПЕЦИАЛЬНОЕ | 63638 | 2064 | 61574 | 33,93% | 31,22% | -3 | 0,102 | 0,410 | 0,510 |
| None', 'СРЕДНЕЕ', 'НЕПОЛНОЕ СРЕДНЕЕ' | 39757 | 1731 | 38026 | 28,46% | 19,28% | -11 | 0,386 | 0,410 | 0,539 |
| 13 | Прожиточный минимум: (dependents\_count\*0.5+1)\*43407 | low\_rod\_income | > 65000 | 7260 | 123 | 7137 | 2,02% | 3,62% | 18 | -0,569 | 0,427 | 0,440 |
| <= 65000 | 75664 | 1669 | 73995 | 27,44% | 37,52% | 9 | -0,287 | 0,427 | 0,469 |
| nan | 120355 | 4291 | 116064 | 70,54% | 58,86% | -5 | 0,170 | 0,427 | 0,518 |
| 14 | КДН из XML (HBMPtsAmnt + CBMPtsAmnt )/ FinalIncome) | KDN | <0.1 | 75668 | 1788 | 73880 | 29,39% | 37,47% | 18 | -0,251 | 0,983 | 0,439 |
| 0.1-0.39 | 74554 | 2039 | 72515 | 33,52% | 36,77% | 6 | -0,083 | 0,983 | 0,480 |
| >=0.4 | 53046 | 2255 | 50791 | 37,07% | 25,76% | -26 | 0,361 | 0,983 | 0,588 |
| nan | 11 | 1 | 10 | 0,02% | 0,01% | -91 | 1,286 | 0,983 | 0,780 |
| 15 | Категория PD (Ансамбль моделей) | pd\_cat | 0-0.1 | 56005 | 136 | 55869 | 2,24% | 28,33% | 189 | -2,758 | 0,949 | 0,068 |
| 0.11-0.2 | 133866 | 1467 | 132399 | 24,12% | 67,14% | 71 | -1,038 | 0,949 | 0,272 |
| 0.21-0.3 | 7286 | 919 | 6367 | 15,11% | 3,23% | -105 | 1,540 | 0,949 | 0,812 |
| 0.31-0.4 | 2424 | 894 | 1530 | 14,70% | 0,78% | -204 | 2,975 | 0,949 | 0,944 |
| 0.41-0.5 | 1899 | 1183 | 716 | 19,45% | 0,36% | -273 | 3,995 | 0,949 | 0,978 |
| >0.5 | 1799 | 1484 | 315 | 24,40% | 0,16% | -355 | 5,182 | 0,949 | 0,993 |

## Приложение 3

**Результаты влияния независимых переменных на зависимую NPL91\_EVER**

|  |  |  |
| --- | --- | --- |
|  | **Переменная** | **IV** |
| 1 | Пол (XML) | 0,075 |
| 2 | Ежемесячный платеж последнего одобренного оффера (XML) | 0,071 |
| 3 | Количество фактических платежей в ПКБ (XML) | 0,112 |
| 4 | Возраст (XML) | 0,039 |
| 5 | Количество платежей без просрочек (XML) | 0,119 |
| 6 | Максимальное количество дней просрочки (XML) | 0,037 |
| 7 | Оценка кредитной истории ПКБ (XML) | 0,123 |
| 8 | Количество запросов в ПКБ за последний год (ПКБ) | 0,155 |
| 9 | Сумма остатка кредитов/кредитных лимитов (ПКБ) | 0,071 |
| 10 | Макс. кол-во дней проср. с начала действия договора (ПКБ) | 0,096 |
| 11 | Объем количества просроченных платежей свыше 1 дней (ПКБ) | 0,108 |
| 12 | Уровень образования (XML) | 0,071 |
| 13 | Прожиточный минимум: (dependents\_count\*0.5+1)\*43407 | 0,062 |
| 14 | КДН из XML (HBMPtsAmnt + CBMPtsAmnt )/ FinalIncome) | 0,064 |
| 15 | Категория PD (Ансамбль моделей) | 3,674 |

**Процент пустых значений в переменных**

|  |  |  |
| --- | --- | --- |
|  | **Переменная** | **NA, %** |
| 1 | Пол (XML) | 0 |
| 2 | Ежемесячный платеж последнего одобренного оффера (XML) | 27,6 |
| 3 | Количество фактических платежей в ПКБ (XML) | 0 |
| 4 | Возраст (XML) | 0 |
| 5 | Количество платежей без просрочек (XML) | 27,7 |
| 6 | Максимальное количество дней просрочки (XML) | 0 |
| 7 | Оценка кредитной истории ПКБ (XML) | 0 |
| 8 | Количество запросов в ПКБ за последний год (ПКБ) | 15,7 |
| 9 | Сумма остатка кредитов/кредитных лимитов (ПКБ) | 26,2 |
| 10 | Макс. кол-во дней проср. с начала действия договора (ПКБ) | 15,7 |
| 11 | Объем количества просроченных платежей свыше 1 дней (ПКБ) | 15,7 |
| 12 | Уровень образования (XML) | 0,27 |
| 13 | Прожиточный минимум: (dependents\_count\*0.5+1)\*43407 | 59,2 |
| 14 | КДН из XML (HBMPtsAmnt + CBMPtsAmnt )/ FinalIncome) | 0,01 |
| 15 | Категория PD (Ансамбль моделей) | 0 |

## Приложение 4

|  |  |  |
| --- | --- | --- |
| **Графики распределений переменных в обучающей (слева) и тестовой (справа) выборках с зависимой переменной NPL91\_EVER** | | |
|  | |  |
| **Пол (XML)** | | |
|  | |  |
| **Ежемесячный платеж последнего одобренного оффера (XML)** | | |
|  | |  |
| **Количество фактических платежей в ПКБ (XML)** | | |
|  | |  |
| **Возраст (XML)** | | |
|  | |  |
| **Количество платежей без просрочек (XML)** | | |
|  | |  |
| **Максимальное количество дней просрочки (XML)** | | |
|  | |  |
| **Оценка кредитной истории ПКБ (XML)** | | |
|  | |  |
| **Количество запросов в ПКБ за последний год** | | |
|  | |  |
| **Сумма остатка кредитов/кредитных лимитов (ПКБ)** | | |
|  | |  |
| **Макс. кол-во дней проср. с начала действия договора (ПКБ)** | | |
|  | |  |
| **Объем количества просроченных платежей свыше 1 дней (ПКБ)** | | |
|  | |  |
| **Уровень образования (XML)** | | |
|  | |  |
| **Прожиточный минимум: (dependents\_count\*0.5+1)\*43407** | | |
|  |  | |
| **КДН из XML (HBMPtsAmnt + CBMPtsAmnt )/ FinalIncome)** | | |