**СОДЕРЖАНИЕ**

[Описание объекта валидации, использованных методов и постановка задачи 4](#_Toc124459565)

[Описание данных для экспериментов, бенчмаркинг 4](#_Toc124459566)

[Выборки, использованные для валидации и разработки модели 4](#_Toc124459567)

[Инжиниринг факторов 5](#_Toc124459568)

[Однофакторный анализ 5](#_Toc124459569)

[Построение модели логистического регрессии 6](#_Toc124459570)

[Метрики качества 6](#_Toc124459571)

[Спецификация итоговой модели 7](#_Toc124459572)

[Результаты валидации 8](#_Toc124459573)

# Описание объекта валидации, использованных методов и постановка задачи

Целью валидационного отчета является описание подхода к разработке моделей прогноза отмены бронирования, предоставление расчетов метрик качества для моделей на выборке для валидации.

При подготовке валидационного отчета были использованы данные, предоставленные индустриальным партнером АО «НСБ».

Задача состоит в прогнозировании факта отмены бронирования до заезда в зависимости от атрибутов бронирования (базовые переменные).

Вероятность отмены бронирования равна единице минус вероятность подтверждения бронирования, поэтому далее, не умаляя общности, будет рассмотрена модель прогноза подтверждения бронирования (т.е. когда заезд действительно состоялся).

Целевой переменной является факт заезда по данном бронированию (целевая переменная «target\_go», которая принимает значение «1», если заезд состоялся и «0», если была произведена отмена бронирования).

Для моделирования целевой переменной был настроен пайплайн, реализующий инжиниринг факторов, процедуру WOE-трансформации, обучения модели логистической регрессии по данным факторам.

В итоге выбрана единая архитектура модели, которая включает отель и типа номеров как факторы, в совокупности она показала высокую точность прогноза вероятности по данным валидации на отложенной выборке.

# Описание данных для экспериментов, бенчмаркинг

## Выборки, использованные для валидации и разработки модели

Таблица «Бронирования» с >64 тыс. бронированиями была использована как исходная выборка. Поскольку оценка вероятности подтверждения бронирования требуется на уровне каждого бронирования, то агрегация данных не производилась.

При разработке и валидации моделей использовались следующие итоговые выборки:

**Обучающая выборка** (2020 год исключен)**:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Отель** | **Тип номера** | **Число наблюдений** | **Средняя доля целевого события** (совершенный заезд – 1, отмена – 0) | **Дата начала выборки** | **Дата окончания выборки** |
| Дача Винтера | Апартаменты | 604 | 76,3% | 01.01.2019 | 29.12.2021 |
| Дача Винтера | Коттедж | 409 | 75,8% | 03.01.2019 | 27.12.2021 |
| Дача Винтера | Стандарт | 3759 | 77,5% | 02.01.2019 | 31.12.2021 |
| Игора | Апартаменты | 2591 | 70,3% | 01.01.2019 | 31.12.2021 |
| Игора | Коттедж | 1732 | 78,8% | 01.01.2019 | 27.12.2021 |
| Игора | Стандарт | 7081 | 72,2% | 01.01.2019 | 31.12.2021 |
| ТНК Видлица | Студия | 1746 | 74,2% | 01.01.2021 | 31.12.2021 |
| ТНК ЛП | Стандарт | 69 | 76,8% | 30.11.2021 | 31.12.2021 |
| ТНК ЛП | Студия | 33 | 72,7% | 08.12.2021 | 28.12.2021 |
| ТНК Приозерск | Стандарт | 8536 | 68,9% | 01.01.2019 | 31.12.2021 |
| ТНК Приозерск | Студия | 3921 | 70,7% | 01.01.2019 | 30.12.2021 |
| ТНК Сортавала | Стандарт | 11654 | 77,7% | 01.01.2019 | 31.12.2021 |

**Валидационная выборка:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Отель** | **Тип номера** | **Число наблюдений** | **Средняя доля целевого события** (совершенный заезд – 1, отмена – 0) | **Дата начала выборки** | **Дата окончания выборки** |
| Дача Винтера | Апартаменты | 355 | 75,5% | 01.01.2022 | 31.10.2022 |
| Дача Винтера | Коттедж | 225 | 69,3% | 01.01.2022 | 30.10.2022 |
| Дача Винтера | Стандарт | 2094 | 77,7% | 01.01.2022 | 31.10.2022 |
| Игора | Апартаменты | 1148 | 78,8% | 01.01.2022 | 31.10.2022 |
| Игора | Коттедж | 736 | 81,7% | 03.01.2022 | 31.10.2022 |
| Игора | Стандарт | 3192 | 80,7% | 01.01.2022 | 31.10.2022 |
| ТНК Видлица | Студия | 1518 | 77,2% | 01.01.2022 | 31.10.2022 |
| ТНК ЛП | Стандарт | 2204 | 82,8% | 02.01.2022 | 31.10.2022 |
| ТНК ЛП | Студия | 947 | 84,3% | 02.01.2022 | 31.10.2022 |
| ТНК Приозерск | Стандарт | 3212 | 81,3% | 01.01.2022 | 31.10.2022 |
| ТНК Приозерск | Студия | 1664 | 80,8% | 01.01.2022 | 31.10.2022 |
| ТНК Сортавала | Стандарт | 5485 | 79,2% | 01.01.2022 | 31.10.2022 |

## Инжиниринг факторов

Для инжиниринга факторов применялись преобразования следующего типа:

1. Соотношения факторов, например, «Доля предоплаты»:
2. Расчет длительностей временных периодов, таких как число дней до планируемого заезда.
3. Категориальные переменные: порядковый номер месяца, дня, дня недели,
4. Дамми-переменные: праздники, даты открытия продаж для даты заезда.

Ключевое требование к инжинирингу факторов, которое предъявлялось – расчет фактора должен быть строго на момент совершения бронирования и не использовать информацию о будущем (т.е. требование отсутствия «data leakage»). В дальнейшей работе можно расширить исходную выборку, «развернув» ее по дням с момента бронирования, чтобы дополнительно учесть факторы резкого изменения погоды, и получив возможность оценивать весь портфель бронирований с точки зрения вероятности отмены, а не только на момент самого бронирования.

## Однофакторный анализ

Процедура отбора факторов перед передачей на вход в модели включала два этапа:

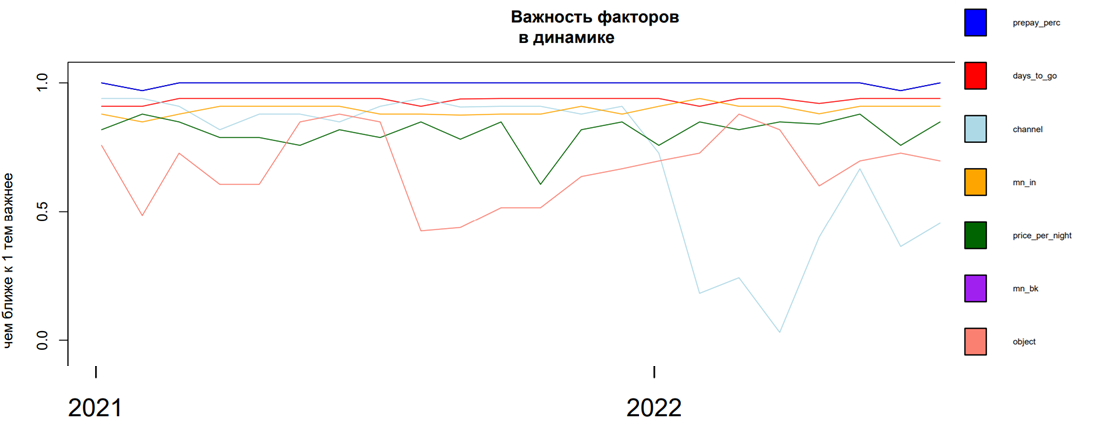
1. WOE-трансформация. Преобразование переменных, аналогичное таргет-энкодингу.
2. Отбор значимых факторов по важности на основе отсечения по показателю Information Value (не менее 0,02), отсутствию попарной корреляции (не более 0,6), а также отсутствию высоких значений Variance inflation factor (не более 3).

Результаты расчета показателя Information Value представлены ниже. Факторы отсортированы в порядке уменьшения информативности с точки зрения предсказания факта подтверждения заезда.

|  |  |  |
| --- | --- | --- |
| **Фактор** | **Кодовое название фактора** | **Information value** |
| Доля предоплаты | prepay\_perc\_woe | 0,7151 |
| Тип оплаты | pmnt\_type\_woe | 0,6747 |
| Дней до планируемого заезда | days\_to\_go\_woe | 0,2340 |
| Канал | channel\_woe | 0,0985 |
| Месяц планируемого заезда | mn\_in\_woe | 0,0452 |
| Цена за 1 ночь 1 номера | price\_per\_night\_woe | 0,0435 |
| Месяц бронирования | mn\_bk\_woe | 0,0377 |
| Отель | object\_woe | 0,0300 |
| Сумма бронирования | amount\_woe | 0,0141 |
| Дамми 1 в течение 7 дней после праздника | h\_p7\_woe | 0,0116 |
| Дамми 1 в течение 6 дней после праздника | h\_p6\_woe | 0,0111 |
| Дамми 1 в течение 4 дней после праздника | h\_p4\_woe | 0,0109 |
| Дамми 1 в течение 5 дней после праздника | h\_p5\_woe | 0,0107 |
| Дамми 1 в течение 3 дней после праздника | h\_p3\_woe | 0,0105 |
| Дамми 1 в течение 2 дней после праздника | h\_p2\_woe | 0,0105 |
| Дамми 1 за 2 дня после праздника | h\_m2\_woe | 0,0102 |
| Дамми 1 за 3 дня после праздника | h\_m3\_woe | 0,0087 |
| Дамми 1 за 7 дней после праздника | h\_m7\_woe | 0,0086 |
| Дамми 1 за 6 дней после праздника | h\_m6\_woe | 0,0083 |
| Дамми 1 за 7 дней до старта продаж | os\_m7\_woe | 0,0082 |
| Дамми 1 за 4 дня после праздника | h\_m4\_woe | 0,0082 |
| Дамми 1 в течение 7 дней после старта продаж | os\_p7\_woe | 0,0082 |
| Дамми 1 за 5 дня после праздника | h\_m5\_woe | 0,0081 |
| День заезда праздник - 1, иначе 0 | pubhol\_woe | 0,0081 |
| Дамми 1 за 1 день до праздника | h\_m1\_woe | 0,0081 |
| Дамми 1 в течение 1 дня после праздника | h\_p1\_woe | 0,0081 |
| День является днем открытия продаж - 1, иначе 0 | open\_sales\_woe | 0,0081 |
| День месяца заезда | day\_in\_woe | 0,0079 |
| Число гостей | guests\_woe | 0,0056 |
| Тип номера | room\_type\_agg\_woe | 0,0055 |
| Число ночей | nights\_woe | 0,0027 |
| День месяца заезда | day\_bk\_woe | 0,0017 |
| День недели бронирования | wd\_bk\_woe | 0,0011 |
| День недели заезда | wd\_in\_woe | 0,0007 |

Факторы со значением IV менее 0,2 были исключены. Также был исключен фактор «Тип оплаты» (pmnt\_type\_woe), поскольку он коррелировал с фактором «Доля предоплаты» (prepay\_perc\_woe) со значениями попарной корреляции на уровне 0,85. В итоговый список WOE-преобразованных факторов вошли: «Доля предоплаты», «Дней до планируемого заезда», «Канал», «Месяц планируемого заезда», «Цена за 1 ночь 1 номера», «Месяц бронирования», «Отель».

Анализ значимости отобранных факторов в динамике подтверждает, что наиболее значимые факторы сохраняют свою значимость стабильно на всем горизонте с 2021 по 2022 год:



## Построение модели логистического регрессии

Далее, на факторах, отобранных на этапе однофакторного анализа, обучается модель логистической регрессии, которая взвешивает с коэффициентами значения WOE по переменным, предоставляя на выходе вероятность подтверждения бронирования.

Для оценки предсказательной и обобщающей способности моделей была проведена валидация out-of-time на 2022 году.

При этом прогноз на период теста производится без обновления спецификации модели, т.е. симулируется ситуация промышленного применения модели без дообучения. Полученные высокие оценки точности говорят о том, что вероятность подтверждения (или отмены) бронирования хорошо моделируемый процесс.

## Метрики качества

Использовалось две метрики:

1. Коэффициент площади под ROC кривой.
2. Калибровочная кривая, которая представляет собой сравнение средней прогнозной доли подтвержденных заездов против средней фактической доли подтвержденных заездов и расчетом метрики точности:

где – среднее фактическое значение доли целевой переменной в бакете прогноза *k*, а – соответственно среднее прогнозное значение по бакету *k*, где *K* – число бакетов, - число наблюдений в *k*-ом бакете

Дополнительно для информации приводятся показатели значимости переменных (на основе p-value в логистической регрессии).

## Спецификация итоговой модели

Первая часть модели представляет собой набор WOE-преобразованных факторов, дискриптивная статистика представлена ниже:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Фактор** | **Кодовое название фактора** | **Диапазоны** | **Число наблюдений в дипазоне** | **Доля подтвержденных заездов** | **Значение WOE** | **IV фактора** |
| Доля предоплаты | prepay\_perc | [-Inf,0.1) | 22024 | 0,5929 | -0,6554 | 0,7151 |
| [0.1,0.7) | 5687 | 0,8542 | 0,7366 | 0,7151 |
| [0.7, Inf) | 14424 | 0,9114 | 1,2992 | 0,7151 |
| Дней до планируемого заезда | days\_to\_go | [-Inf,4) | 8431 | 0,8763 | 0,9262 | 0,2340 |
| [4,10) | 7723 | 0,8126 | 0,4356 | 0,2340 |
| [10, Inf) | 25981 | 0,6697 | -0,3249 | 0,2340 |
| Канал | channel | Официальный сайт, integration.com, OTA | 12116 | 0,6391 | -0,4603 | 0,0985 |
| Сайт, booking.com | 30019 | 0,7768 | 0,2158 | 0,0985 |
| Месяц планируемого заезда | mn\_in | [-Inf,4) | 9528 | 0,7622 | 0,1330 | 0,0452 |
| [4,5) | 2597 | 0,7289 | -0,0425 | 0,0452 |
| [5,7) | 6931 | 0,7941 | 0,3183 | 0,0452 |
| [7,10) | 12670 | 0,7421 | 0,0251 | 0,0452 |
| [10, Inf) | 10409 | 0,6727 | -0,3112 | 0,0452 |
| Цена за 1 ночь 1 номера | price\_per\_night | [-Inf,4500) | 5314 | 0,6658 | -0,3424 | 0,0435 |
| [4500,7500) | 14010 | 0,7090 | -0,1411 | 0,0435 |
| [7500,8500) | 3469 | 0,7374 | 0,0008 | 0,0435 |
| [8500, Inf) | 19342 | 0,7773 | 0,2182 | 0,0435 |
| Месяц бронирования | mn\_bk | [-Inf,6) | 16424 | 0,7783 | 0,2243 | 0,0377 |
| [6,7) | 4124 | 0,7449 | 0,0400 | 0,0377 |
| [7,10) | 11925 | 0,7225 | -0,0746 | 0,0377 |
| [10, Inf) | 9662 | 0,6823 | -0,2674 | 0,0377 |
| Отель | object | ТНК Приозерск | 12457 | 0,6948 | -0,2090 | 0,0300 |
| Игора, ТНК Видлица | 13150 | 0,7296 | -0,0391 | 0,0300 |
| ТНК ЛП,Дача Винтера | 4874 | 0,7716 | 0,1860 | 0,0300 |
| ТНК Сортавала | 11654 | 0,7768 | 0,2156 | 0,0300 |

Вторая часть модели представляет собой веса логистической регрессии, которые присвоены каждому WOE-преобразованному фактору.

где– факт заезда, , – вероятность подтверждения брони при условии параметров брони *score*.

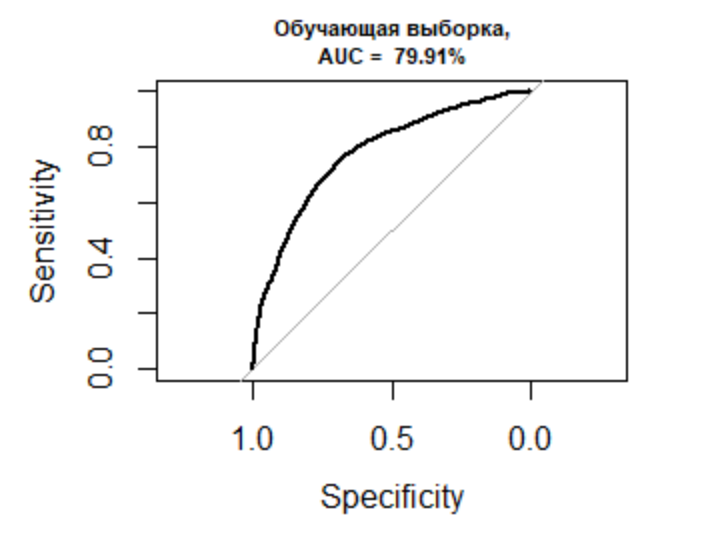
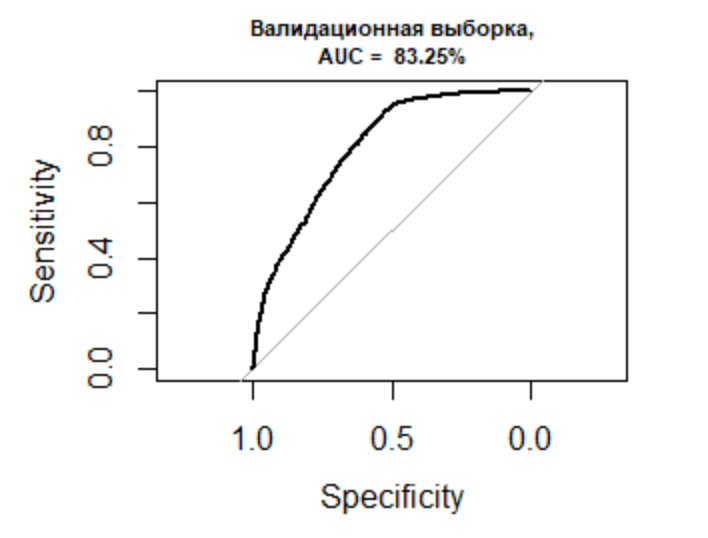
Финальным выходом модели является вероятность подтверждения бронирования.

Спецификация коэффициентов логистической регрессии приведена ниже:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Фактор** | **Кодовое название фактора** | **Коэффициент** | **Стандартная ошибка** | **Z-значение** | **Значимость на основе p-value** |
| Константа | (Intercept) | 1,0155 | 0,0128 | 79,6110 | Очень высокая |
| Доля предоплаты | prepay\_perc\_woe | 1,3208 | 0,0205 | 64,4930 | Очень высокая |
| Дней до планируемого заезда | days\_to\_go\_woe | 1,2075 | 0,0270 | 44,7640 | Очень высокая |
| Канал | channel\_woe | -0,6512 | 0,0435 | -14,9770 | Очень высокая |
| Месяц планируемого заезда | mn\_in\_woe | 0,2364 | 0,0775 | 3,0490 | Высокая |
| Цена за 1 ночь 1 номера | price\_per\_night\_woe | -1,0847 | 0,0662 | -16,3750 | Очень высокая |
| Месяц бронирования | mn\_bk\_woe | 0,1018 | 0,0854 | 1,1920 | Умеренная |
| Отель | object\_woe | 1,0447 | 0,0701 | 14,9010 | Очень высокая |

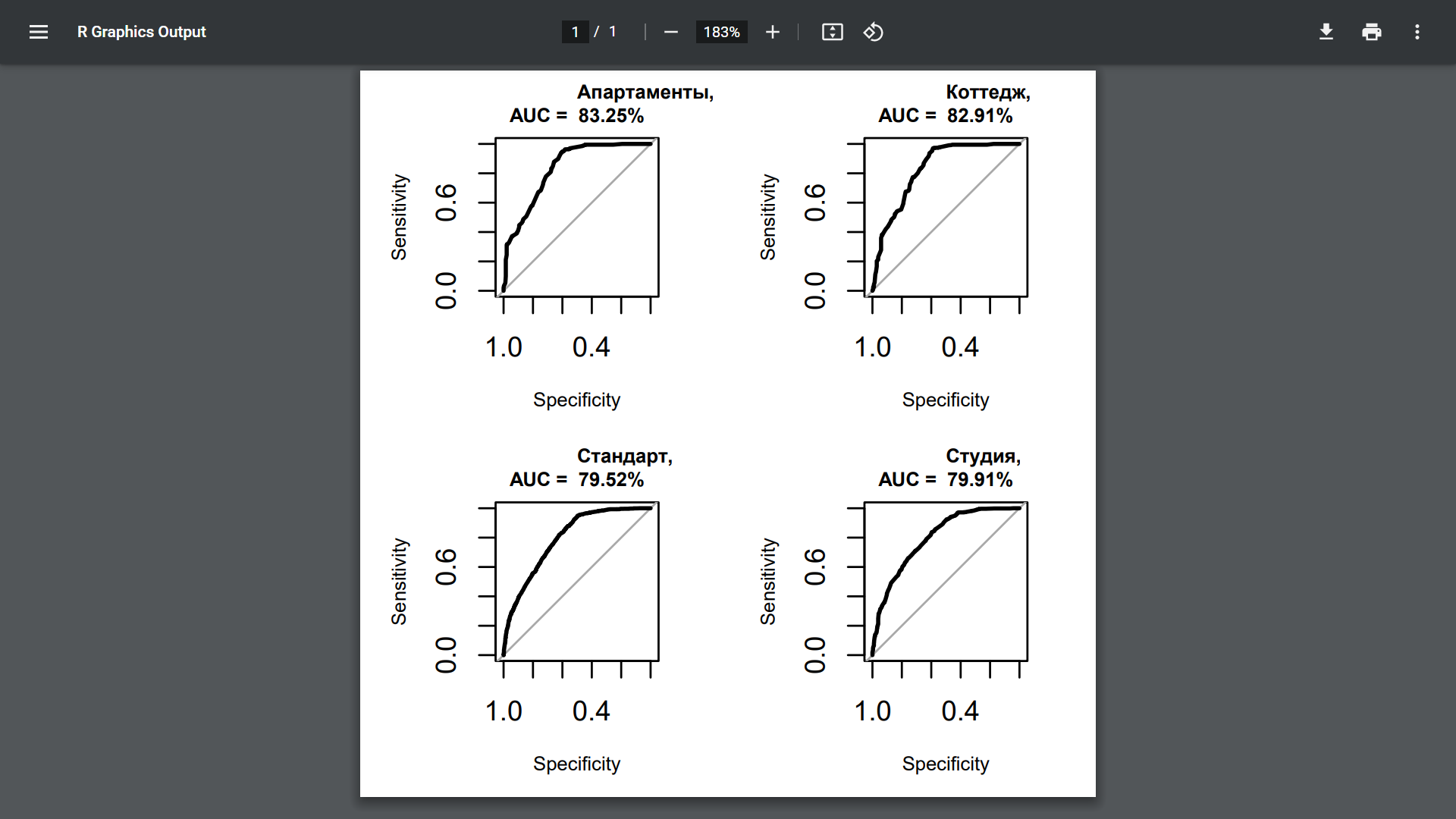
## Результаты валидации

1. Были рассчитаны метрики ROC-AUC для агрегированной обучающей и тестовой выборок по всем отелям и типам номеров:

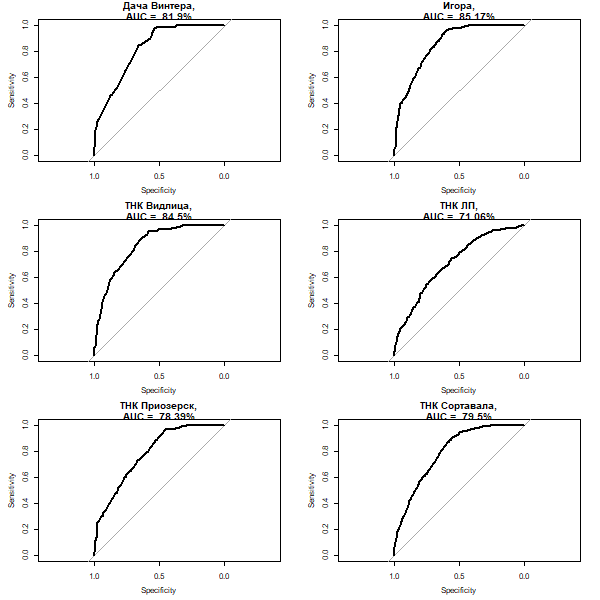
 

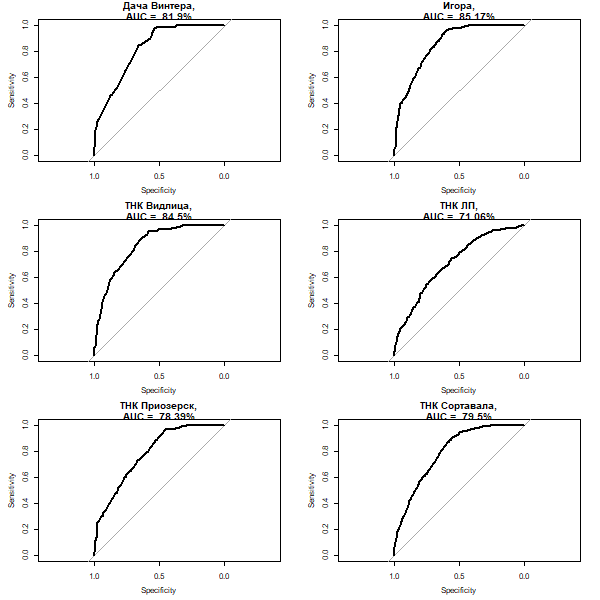
Показатели дает высокую оценку ранжирующей способности модели >79%. При этом показатель стабильно высокий, как на обучающей, так и на тестовой выборке.

Дополнительные графики ROC-AUC в разрезе типов номеров для валидационной выборки, на основе которых можно заключить, что ранжирующая способность стабильная и на высоком уровне на различных сегментах:

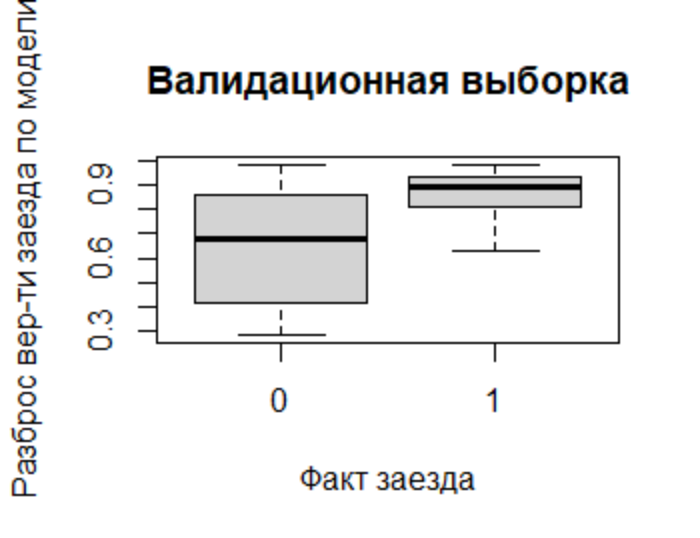
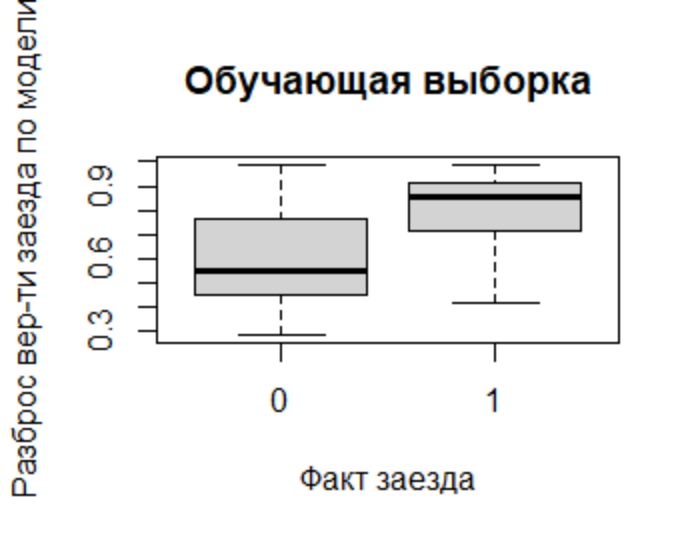


Графики ROC-AUC в разрезе отелей для валидационной выборки, также со стабильной ранжирующей способностью ниже:



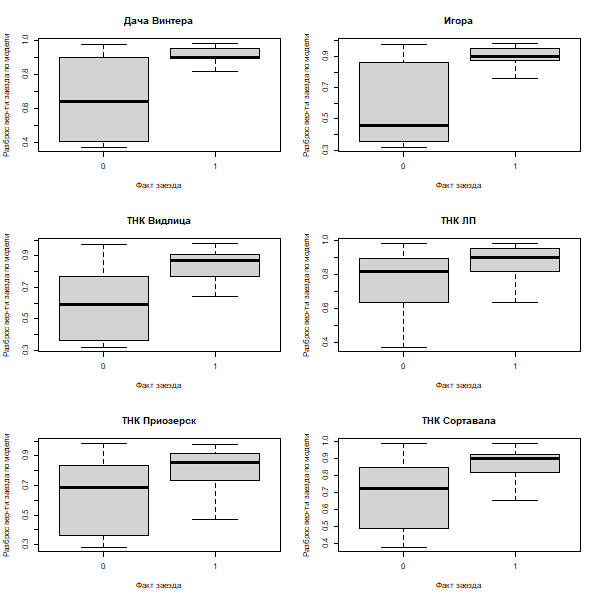


Ниже представлено распределение вероятности прогноза подтверждения брони (вероятности заезда) в двух группах агрегированно: среди тех клиентов, которые отменили свою бронь (факт заезда – 0), и среди тех, которые заехали (факт заезда – 1):

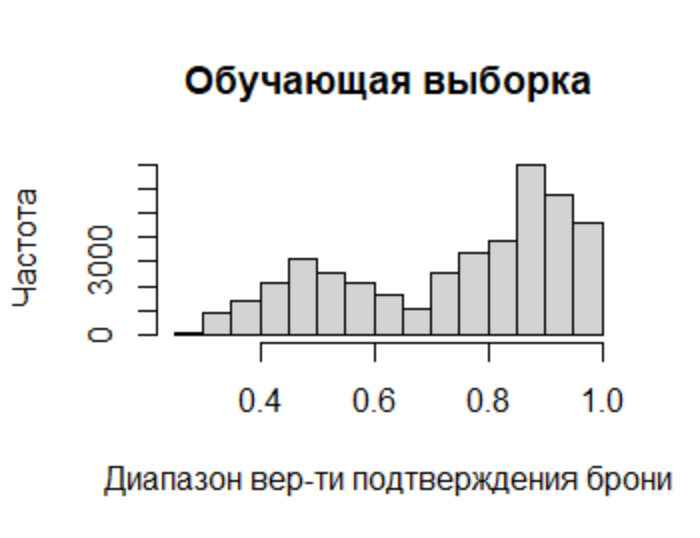
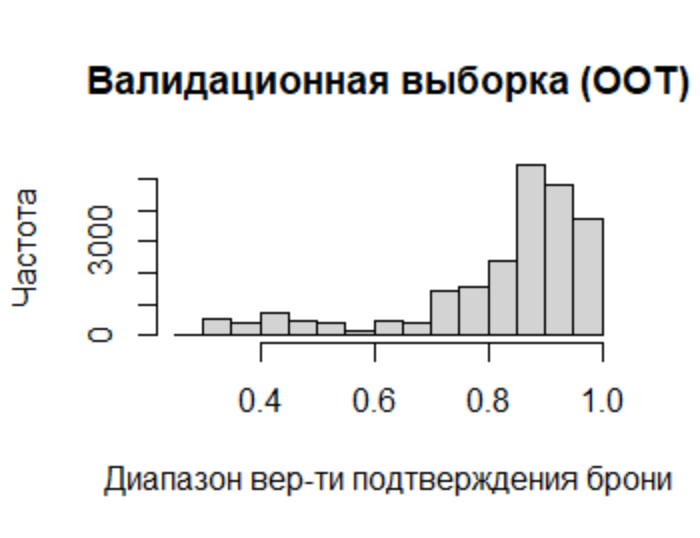


Видно, что для группы клиентов, которые заехали, значимо распределение прогнозов смещено вверх, что демонстрирует корректную работу модели.

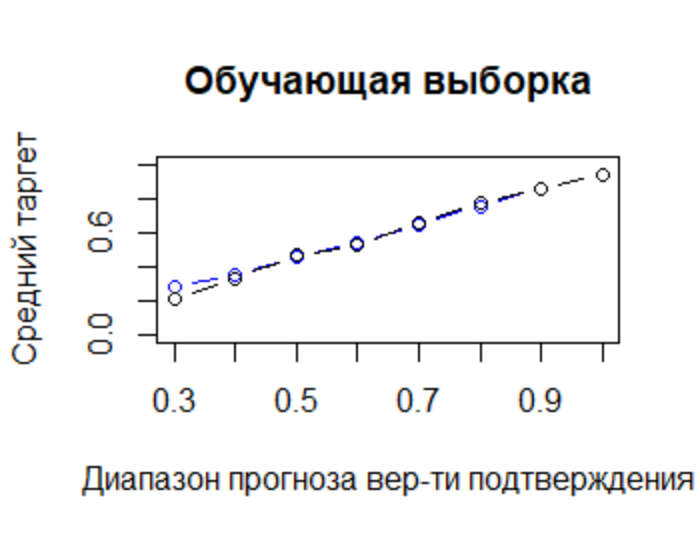
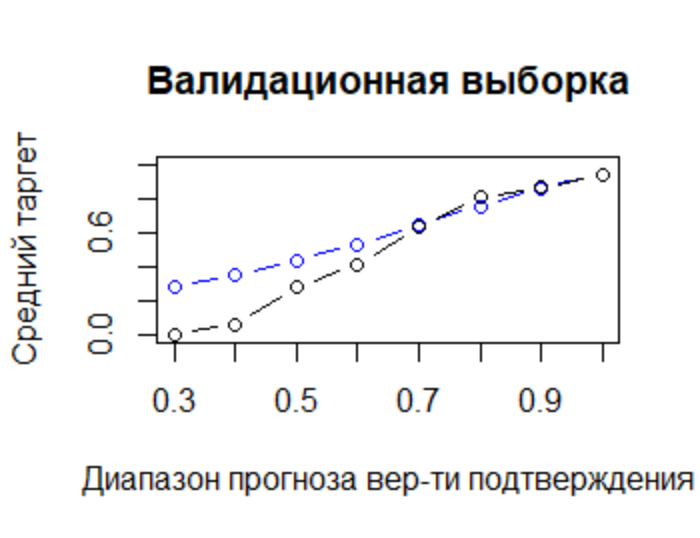
Ниже дополнительно приведены графики разброса вероятности заезда в разрезе отелей. Видно, что в разрезе отелей так же стабильная разделяющая способность: средние прогнозные доля заезда значительно отличаются в группах подтвержденных и отмененных броней.



Ниже, дополнительно, приведены гистограммы распределения прогнозов по модели как для обучающей (до 2022 года), так и для валидационной выборки (2022 год). Наблюдается ярковыраженная бимодальность на выборке для обучения, что косвенно подтверждает, что модель устойчиво обучилась и показывает стабильную разделяющую способность, при этом на валидационной выборке бимодальность менее выражена и представлена только «тяжелым хвостом» распределения. Это отражает тот факт, что в целом в 2022 г. люди заезжали чаще (в среднем в 80% случаев, а до 2022 г. – лишь в 73,7%):

1. Ниже приведены расчеты калибровочных кривых и метрик точности. Синяя линия – прогноз, черная линия – факт доли подтвержденных заездов (броней).

*Синяя линия – прогнозные доли (средний таргет), черная линия – фактические доли   
подтвержденных заездов (броней).*

Ось абсцисс на данном графике представляет собой бакеты, в рамках которых рассчитывается среднее прогнозное и среднее фактическое значение доли подтвержденных броней (заехавших клиентов). На графике справа (валидационная выборка) наблюдается более резкое разделение групп клиентов: фактическая доля клиентов, подтвердивших заезд практически равна нулю в бакетах с самой низкой вероятностью заезда.

График калибровочной кривой демонстрирует также и сохранение монотонности, т.е. для бакетов с более высокой прогнозной оценкой средние фактические доли подтвердивших бронь так же более высокие.

Расхождения по бакетам наблюдаются для бакетов со сравнительно меньшим числом наблюдений:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Обучающая выборка** | | | |  | **Валидационная выборка** | | | |
| **Диапазон значения прогноза** | **Число наблюдений** | **Среднее прогнозное значение** | **Среднее фактическое значение** |  | **Диапазон значения прогноза** | **Число наблюдений** | **Среднее прогнозное значение** | **Среднее фактическое значение** |
| [0.2; 0.3) | 79 | 0,2834 | 0,2152 |  | [0.2; 0.3) | 19 | 0,2834 | 0,0000 |
| [0.3; 0.4) | 2271 | 0,3562 | 0,3320 |  | [0.3; 0.4) | 906 | 0,3475 | 0,0640 |
| [0.4; 0.5) | 5257 | 0,4580 | 0,4638 |  | [0.4; 0.5) | 1174 | 0,4374 | 0,2845 |
| [0.5; 0.6) | 4695 | 0,5435 | 0,5278 |  | [0.5; 0.6) | 525 | 0,5311 | 0,4076 |
| [0.6; 0.7) | 2741 | 0,6442 | 0,6611 |  | [0.6; 0.7) | 836 | 0,6476 | 0,6364 |
| [0.7; 0.8) | 5968 | 0,7545 | 0,7733 |  | [0.7; 0.8) | 2957 | 0,7557 | 0,8076 |
| [0.8; 0.9) | 10831 | 0,8583 | 0,8599 |  | [0.8; 0.9) | 7851 | 0,8601 | 0,8617 |
| [0.9; 1.0) | 10293 | 0,9431 | 0,9361 |  | [0.9; 1.0) | 8512 | 0,9412 | 0,9327 |

Значения метрик точности составляют для обучающей выборки: 0.9901, а для валидацинной: 0.9668.